İSTATİSTİK BÖLÜMÜ

KAZANAN TAKIMLARIN İLERİ BASKETBOL İSTATİSTİKLERİ KULLANILARAK BELİRLENMESİ: MAKİNE ÖĞRENMESİ YAKLAŞIMI

Ahmet Ekrem YILDIZ 20023074

Danışman: Doç. Dr. Erhan ÇENE

ÖZET

Bu çalışmada 1997-98, 2014-15, 2021-22, 2022-23, 2023-24 ve 2024-25 sezonlarına ait takım istatistiklerini kullanarak maç sonuçlarını tahmin etmeyi amaçlar. Lojistik regresyon, rassal orman ve yapay sinir ağı modelleriyle yapılan analizlerde, her sezon için doğruluk oranları karşılaştırılmıştır. Kullanılan değişkenler hücum ribaundu (OREB_PCT), savunma ribaundu(DREB_PCT),takım top kaybı oranı (TM_TOV_PCT),40 dakikadaki tempo (PACE_PER40) ve gerçek şut yüzdesi (TS_PCT) gibi ileri düzey performans metrikleri yer almaktadır. En yüksek doğruluk 2022 – 2023 sezonunda %81,6 ile lojistik regresyon modeli elde etmiştir. Model doğrulukları sezona göre değiştiği gözlemlenmiş, bu farklar NBA'deki oyun anlayışındaki evrimle ilişkilendirilmiştir. Üçlük odaklı hücum sistemleri, savunma çeşitliliği ve tempodaki artış gibi faktörler algoritmaların tahmin gücünü doğrudan etkilemiştir. Bulgular, spor analizlerinde makine öğrenmesi modellerinin yalnızca teknik performans ölçümünde değil, oyunun yapısal dinamiklerini anlamada da önemli katkı sağladığını göstermektedir.

Giriş

Spor analitiği, istatistiksel modelleme ve yapay zeka uygulamalarıyla birlikte hızla yükselen ve kümülatif bir şekilde bilgi birikiminin oluştuğu bir alan olmuştur. Özellikle basketbol, oyuncu veya takım bazlı istatistiklerin çok fazla değişkenlik gösterdiği ve performanslarının ölçüldüğü, maç sonuçlarının tahminlemesi gibi aksiyonların profesyonel spor organizasyonlarının ve akademik çalışmaların ilgi noktası olmuştur. Bu çalışmada, belirli NBA sezonlarına ait takım istatistikleri kullanılarak maç sonuçları tahmin edilmeye çalışılmıştır. Lojistik Regresyon, Rasgele Orman ve Yapay Sinir Ağı (ANN) gibi farklı sınıflandırma algoritmaları ile model performansları karşılaştırılmış; doğruluk oranları sezonsal bazda değerlendirilmiştir. Basketboldaki oyun anlayışının yıllar içinde değişmesi —örneğin tempo, üç sayılık atış oranı, ribaund etkinliği gibi faktörlerdeki dönüşüm— model performansları üzerinde etkili olmuş ve bu da analizde önemli bir boyut olarak ele alınmıştır.

Yöntemler

Lojistik regresyon, bir olayın gerçekleşip gerçekleşmeyeceğini olasılık temelli olarak tahmin eden denetimli bir makine öğrenmesi algoritmasıdır. Girdi olarak sayısal ya da kategorik değişkenleri alır ve çıktıyı 0 ile 1 arasında bir olasılık değeri olarak üretir; bu değer genellikle belirli bir eşik değere göre sınıflandırılır. Rassal Orman, sınıflandırma ve regresyon görevleri için yaygın olarak kullanılan bir makine öğrenmesi algoritmasıdır ve topluluk öğrenme kategorisine aittir. Yapay sinir ağları, insan beynindeki sinir hücrelerinden esinlenerek tasarlanmış, birbiriyle bağlantılı yapay nöronlardan oluşan öğrenme sistemleridir. Girdi verilerinden örüntüleri öğrenerek sınıflandırma, regresyon gibi görevleri yerine getirir.

NBA API ve Veriseti

NBA API, NBA'in resmi veri altyapısına dayanan, maçlar, oyuncular, takımlar, istatistikler ve daha fazlasına erişim sağlayan bir programlama arayüzüdür. Python için özel olarak geliştirilen nba_api adlı kütüphane, bu verileri otomatik olarak çekip analiz etmeye olanak tanır. Bu çalışmadaki veriseti yaklaşık 5000 maç istatistiklerini içermektedir. NBA API, 2015 yılından itibaren daha düzenli ve sık bir şekilde bu API'yi kullanıcılarına hizmeti sağlamaktadır.

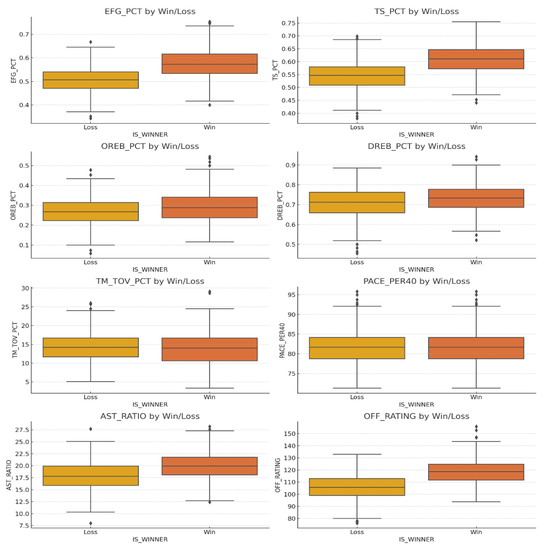
Değişken	Açıklama	Formül	Tür
OREB_PCT	Takımın hücum ribaundlarını alma yüzdesi	OREB / (OREB + Opp DREB)	Bağımsız Değişken
DREB_PCT	Takımın savunma ribaundlarını alma yüzdesi	DREB / (DREB + Opp OREB)	Bağımsız Değişken
TM_TOV_PCT	Top kaybı yüzdesi	TOV / (FGA + 0.44*FTA + TOV)	Bağımsız Değişken
PACE_PER40	40 dakika başına pozisyon sayısı	Possessions * (40 / Minutes Played)	Bağımsız Değişken
TS_PCT	Gerçek şut yüzdesi	PTS / (2 * (FGA + 0.44 * FTA))	Bağımsız Değişken
IS_WINNER	Maçı kazanıp kazanmadığını belirten ikili değişken (1: galibiyet, 0: mağlubiyet)	-	Bağımlı Değişken

Tablo 1: Değişkenler listesi

```
In [15]: df[['OREB_PCT', 'DREB_PCT', 'TM_TOV_PCT', 'PACE_PER40', 'TS_PCT', 'IS_WINNER']]
Out[15]:
      OREB_PCT  DREB_PCT  TM_TOV_PCT  PACE_PER40  TS_PCT  IS_WINNER
0        0.333     0.679        13.1       82.08   0.585          1
1        0.321     0.667        17.3       82.08   0.537          0
4        0.311     0.667         7.9       73.75   0.599          1
5        0.333     0.689        15.9       73.75   0.597          0
6        0.320     0.767        16.5       85.42   0.585          1
...        ...       ...         ...         ...     ...        ...
1821     0.333     0.654        20.2       86.67   0.606          0
1824     0.375     0.706        14.1       79.22   0.557          1
1825     0.294     0.625        16.5       79.22   0.556          0
1828     0.234     0.731        13.7       85.00   0.599          1
1829     0.269     0.766        11.8       85.00   0.578          0
```

Şekil 1: Verisetinden örnek istatistikler

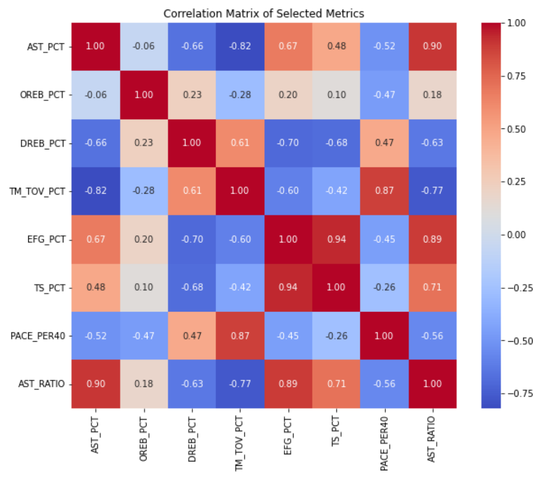
Birleştirilmiş verisetinin maç sonu istatisikleridir.



Şekil 2: 2023-2024 Sezonu İstatistikleri Kutu Grafiği

2023-2024 sezonuna ait bu kutu grafiği görselinde Bu sezona ait bazı ileri istatistiklerin medyanını, Kartillerini, ortalamalarını ve varsa aykırı değerlerini tespit edebiliyoruz.

Efektif şut yüzdesi değişkeni (eFG), üçlüklerin, ikilik-erden 1 sayı daha fazla verdiği için katsayısının daha fazla olduğu bir istatistiktir. Tabloda görüldüğü üzere eFG ve TS_PCT istatistikleri kazanmada en etken faktör. Hücum ribaundu ve savunma ribaundları da tamamlayıcı faktör olarak karşımıza çıkmaktadır.



2024-25 sezonunun hesaplanan Pearson korelasyon katsayıları, özellikle asist, şut verimliliği ve top kaybı gibi hücum performans metrikleri arasında güçlü doğrusal ilişkiler olduğunu göstermektedir. Özellikle AST_PCT ile AST_RATIO (r = 0.90) ve EFG_PCT ile TS_PCT (r = 0.94) gibi yüksek korelasyonlar, ilgili metriklerin birbirini tamamladığını; TM_TOV_PCT ile AST_PCT (r = -0.82) ve PACE_PER40 (r = 0.87) arasındaki ilişkiler ise oyun tarzının top kontrolü üzerindeki etkisini ortaya koymaktadır.

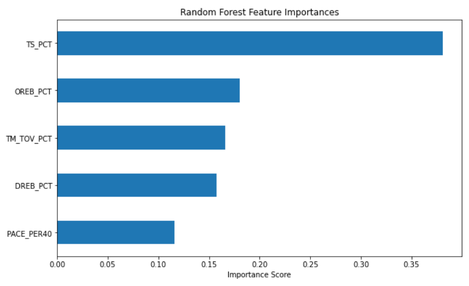
Şekil 3: 2024-25 Sezonu İstatisikleri Korelasyon Matrisi

Uygulama

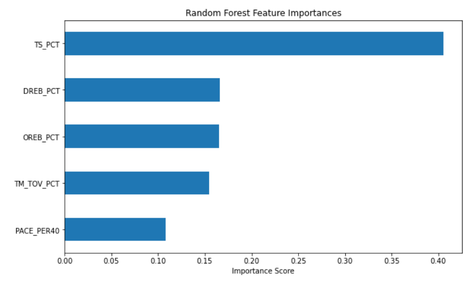
Sezon	Lojistik Regresyon Doğruluk	Rassal Orman Doğruluk	Sinir Ağları Doğruluk
2014–2015	0.70	0.707	0.721
2021–2022	0.737	0.752	0.77
2022–2023	0.816	0.717	0.747
2023–2024	0.75	0.78	0.763
2024–2025	0.7027	0.73	0.73

Tablo 2: Modellerin Karşılaştırması

Bu tablo, NBA maç sonucu tahmini üzerine uygulanan üç farklı makine öğrenmesi algoritmasının (Lojistik Regresyon, Rassal Orman, Sinir Ağları) sezonlara göre doğruluk (accuracy) performanslarını karşılaştırmalı olarak sunmaktadır.



2024-25 sezonunun farkı: Hücum ribaundu ve top kaybı oranı.



2023-24 sezonunun farkı: Savunma ribaundu ve hücum ribaundu.

2022-23 sezonunun farkı: Savunma ribaundu ve hücum ribaundu.

2021-22 sezonunun farkı: Hücum ribaundu ve takım top kaybı oranı.

Şekil 4: Değişken Önem Dereceleri Ölçümü

SONUÇ

Bu çalışmada, belirli NBA sezonlarına ait takım istatistikleri kullanılarak maç sonuçlarını tahmin etmek amacıyla lojistik regresyon, rassal orman ve sinir ağı algoritmaları uygulanmıştır. Her modelin doğruluk oranları sezona göre farklılık göstermiş; bu da verinin yapısının ve oyun tarzının model başarısını etkilediğini göstermiştir. *2022–2023 sezonunda lojistik regresyon en yüksek doğruluğu sağlamış*, bu da daha tutarlı istatistiksel yapıların basit modellerle daha iyi temsil edilebileceğini ortaya koymuştur. Bu çalışmada galibiyet tahmininde en etkili metrikler hücum verimliliği, pas paylaşımı ve top kontrolüyle doğrudan ilişkili olarak öne çıkmıştır. Özellikle TS_PCT ve OREB_PCT ve DREB_PCT kazanan takımlarda istikrarlı bir şekilde öne çıkmaktadır. Sonuç olarak, model seçiminin başarısı sadece teknik faktörlere değil, aynı zamanda basketbolun dönemsel dinamiklerine de bağlıdır.

KAYNAKÇA

[1] Kumar, V., Saini, H., & Arora, A. (2019). Demand Forecasting Using Machine Learning Algorithms. *International Journal of Engineering Research & Technology (IJERT)*, 8(6), 2278-0181. https://www.ijert.org/demand-forecasting-using-machine-learning-algorithms

[2] Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3). https://doi.org/10.2202/1559-0410.1070

[3] Loeffelholz, B., Readdy, T., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1). https://doi.org/10.2202/1559-0410.1150

2024-2025 BAHAR YARIYILI İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE İSTANBUL TRAFİK YOĞUNLUĞU TAHMİNİ

Ali Emin Aslan 20023025

Danışman: Prof. Dr. Fatma Noyan Tekeli

Bu çalışma, İstanbul Büyükşehir Belediyesi'nin yayınladığı saatlik trafik yoğunluğu verileri kullanılarak, İstanbul'da trafik hacmini tahmin etmeye yönelik bir makine öğrenmesi uygulamasını kapsamaktadır.

Ekim 2024 - Ocak 2025 tarihleri arasını kapsayan veri seti, 6.5 milyondan fazla satırdan oluşmakta ve sabit ölçüm noktalarında her saat başı ölçülen araç sayısı, minimum–maksimum hız gibi bilgileri içermektedir.

Tahmin için üç farklı makine öğrenmesi algoritması eğitilmiş ve karşılaştırılmıştır: **Doğrusal Regresyon**, **Rastgele Orman (Random Forest)** ve **XGBoost**.

Model değerlendirmesinde MAE (Ortalama Mutlak Hata), RMSE (Kök Ortalama Kare Hata) ve R^2 (Determination Coefficient) metrikleri kullanılmıştır.

En başarılı performans, R^2 değeri 0.93 olan **Random Forest** algoritması ile elde edilmiştir. Doğrusal regresyon modeli ise düşük R^2 değeriyle trafik gibi karmaşık sistemleri modellemede yetersiz kalmıştır.

Giriş

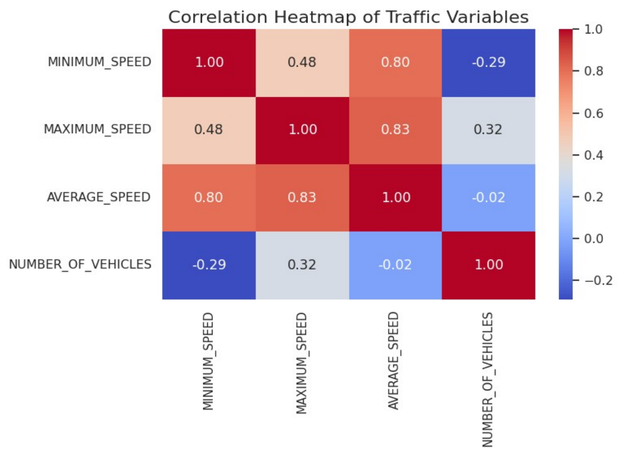
İstanbul gibi büyük metropollerde trafik, günlük yaşamı doğrudan etkileyen en önemli sorunlardan biridir. Bu çalışmada, İstanbul Büyükşehir Belediyesi'nin yayınladığı trafik yoğunluğu verileri kullanılarak saatlik araç sayısının makine öğrenmesi algoritmaları ile tahmin edilmesi amaçlanmıştır.

Makine Öğrenmesi Yöntemleri

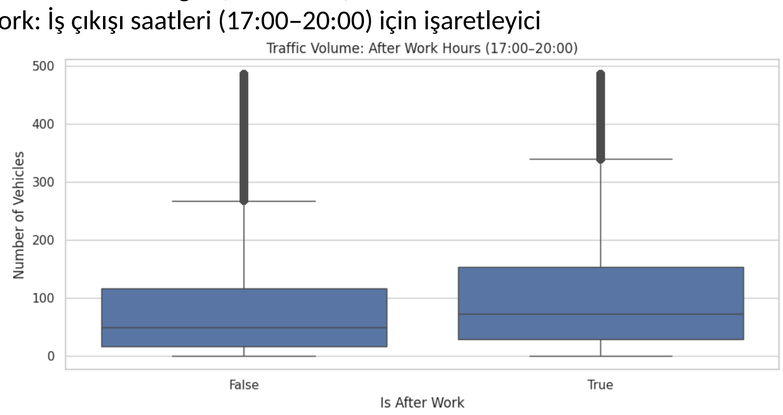
- **Doğrusal Regresyon**, bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkiyi modellemeye yarayan temel bir regresyon yöntemidir. Model, en küçük kareler yöntemiyle (OLS – Ordinary Least Squares) tahmin edilen doğrusal katsayılar sayesinde veriye en iyi uyan doğrusal doğrultuyu belirler. Basitliği nedeniyle yorumlanabilirliği yüksektir, ancak değişkenler arasında doğrusal olmayan ilişkileri açıklamakta yetersiz kalabilir.
- **Random Forest**, çok sayıda karar ağacının birleşiminden oluşan bir topluluk öğrenme (ensemble learning) yöntemidir. Her bir ağaç, rastgele seçilmiş örneklem ve değişkenlerle eğitilir. Son tahmin, regresyon problemlerinde bu ağaçların ortalaması alınarak yapılır. Bu yöntem, overfitting riskini azaltır, değişkenler arası karmaşık ilişkileri başarılı şekilde öğrenebilir ve yüksek veri çeşitliliğinde bile sağlam sonuçlar üretebilir.
- **XGBoost (Extreme Gradient Boosting)**, boosting prensibine dayanan ve tahmin hatalarını ardışık ağaçlarla iyileştiren güçlü bir makine öğrenmesi algoritmasıdır. Her yeni ağaç, bir önceki modelin hata kalıntılarına odaklanarak öğrenmeyi sürdürür. Ağaçların katkısı öğrenme oranı (learning rate) gibi parametrelerle dengelenir. XGBoost, düzenlileştirme (regularization), eksik veri toleransı ve paralel işlem kabiliyeti gibi gelişmiş özellikleriyle diğer boosting algoritmalarına göre daha verimli ve kararlı sonuçlar sunar.

Veri Kümesi

- **Veri Sağlayıcısı:** İstanbul Büyükşehir Belediyesi Açık Veri Portalı
- **Tarih Aralığı:** 01.10.2024 – 31.01.2025
- **Toplam Gözlem Sayısı:** 6.510.140
- **Zaman Aralığı:** Saatlik ölçümler
- **Veri Tipi:** Noktasal trafik ölçümleri (birden fazla sabit lokasyondan alınmış)
- **Temel Değişkenler:**
  - DATE_TIME: Zaman bilgisi
  - LATITUDE, LONGITUDE: Konum bilgisi
  - MINIMUM_SPEED, MAXIMUM_SPEED, AVERAGE_SPEED: Hız değerleri (km/saat)
  - NUMBER_OF_VEHICLES: Saatlik ölçülen toplam araç sayısı



- **Türetilen Ek Özellikler:**
  - hour: Gün içindeki saat (0–23)
  - day_of_week: Haftanın günü (0 = Pazartesi)
  - month: Ay bilgisi
  - is_weekend: Hafta sonu bilgisi (True/False)
  - is_after_work: İş çıkışı saatleri (17:00–20:00) için işaretleyici



Veri Ön İşleme Adımları

Ham veri kümesi, İstanbul'daki sabit noktalardan saatlik olarak toplanmış toplam 6.510.140 ölçüm içermektedir.

İlk olarak, modelleme sürecine uygun olmayan ve MINIMUM_SPEED, MAXIMUM_SPEED, AVERAGE_SPEED ile NUMBER_OF_VEHICLES sütunlarında eksik ya da sayısal format hatası içeren gözlemler temizlenmiştir. Ardından bu sütunlar sayısal tipe dönüştürülerek bütünlük sağlanmıştır. Ayrıca uç değerlerin modele etkisini azaltmak amacıyla **z-score yöntemi** kullanılarak istatistiksel olarak aykırı kabul edilen gözlemler veri setinden çıkarılmıştır.

Bu işlemler sonucunda kullanılabilir veri sayısı 6.339.245 satıra düşürülmüştür. Ayrıca modelin öğrenme gücünü artırmak amacıyla tarih-saat bilgisi üzerinden saat, hafta günü, ay gibi zaman değişkenleri üretilmiş; hafta sonu ve iş çıkış saatleri gibi kategorik özellikler de veri setine dahil edilmiştir. Bu ön işleme süreci, modelin trafik desenlerini daha doğru yakalayabilmesini sağlamıştır.

Yöntem ve Modeller

Bu çalışmada trafik yoğunluğu tahmini için üç farklı regresyon modeli kullanılmıştır:

1. Linear Regression

En basit modeldir; doğrusal ilişkileri varsayar. Karmaşık trafik dinamiklerini yansıtmakta yetersiz kalmıştır.

2. Random Forest

Birden fazla karar ağacının birleşimiyle oluşturulmuş topluluk (ensemble) modelidir. Aşırı öğrenmeye karşı dirençlidir, doğruluğu yüksektir. n_estimators artırıldıkça model performansı da artmıştır.

3. XGBoost (Extreme Gradient Boosting)

Boosting algoritmalarının güçlü bir türüdür. Daha hassas öğrenme ve hata azaltımı yapar. Ancak bu veri setinde Random Forest kadar başarılı olmamıştır.

Model Performansı Değerlendirme Metrikleri:

MAE (Mean Absolute Error): Ortalama mutlak hata

RMSE (Root Mean Square Error): Hataların karesel ortalaması

R^2 (R-Kare): Modelin açıklayıcılık oranı (1'e yakınsa daha iyidir)

Sonuçlar (n_estimators = 20)

Makine Öğrenmesi Yöntemi	MAE	RMSE	R-Kare
Doğrusal Regresyon	68.58	92.56	0.0942
Random Forest	14.37	26.32	0.9267
XGBoost	49.45	71.93	0.4530

Random Forest için n_esitmatiors Karşılaştırması

n_estimators	MAE	RMSE	R-Kare
5	14.65	26.87	0.9237
10	14.46	26.50	0.9257
15	14.40	26.38	0.9264
20	14.37	26.32	0.9267

- Artan estimator sayısı, modelin doğruluğunu artırmakta; ancak işlem süresi de uzamaktadır.

Sonuç

- En yüksek performans, **Random Forest** modeli ile elde edilmiştir ($R^2 \approx 0.93$).
- Zaman bilgileri ve hafta içi/sonu ayrımı gibi özellikler tahmin gücünü önemli ölçüde artırmaktadır.
- Model çıktıları, şehir içi ulaşım planlaması ve trafik yönetimi için potansiyel destek sunmaktadır.

Öneriler

Çalışmanın mevcut sonuçları oldukça başarılı olmakla birlikte, modelin daha geniş zaman aralıklarını kapsayan veri setleriyle eğitilmesi önerilmektedir. Böylece mevsimsel etkiler ve uzun dönemli trafik desenleri de modele yansıtılabilir. Ayrıca, trafik yoğunluğunu etkileyen diğer dışsal faktörlerin — özellikle hava durumu, resmî tatiller ve özel günler gibi — veri setine entegre edilmesi, modelin açıklayıcılığını artırabilir.

Aykırı değerlerin daha gelişmiş yöntemlerle (örneğin görsel analiz destekli) temizlenmesi de modelin doğruluğunu olumlu yönde etkileyebilir.

Son olarak, elde edilen tahmin modeli uygun şekilde optimize edilerek şehir içi ulaşım sistemlerine entegre edilebilir ve gerçek zamanlı trafik yoğunluğu tahminlerinde kullanılabilir.

KAYNAKÇA

[1] https://data.ibb.gov.tr/dataset/hourly-traffic-density-data-set

[2] Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). *Short-term traffic forecasting: Where we are and where we're going*. Transportation Research Part C: Emerging Technologies, 43, 3–19.

[3] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

2024-2025 BAHAR Yarıyılı İSTATİSTİK BÖLÜMÜ

Makine Öğrenmesi ve Zaman Serisi Yöntemleri ile BIST Banka Hisselerinde Fiyat Tahmini: Karşılaştırmalı Analiz

20023082 ASYA BÜYÜKTUNA

Danışman: Prof. Dr. ALİ HAKAN BÜYÜKLÜ

ÖZET

Bu çalışmada, Borsa İstanbul'da işlem gören banka hisseleri için fiyat tahmini yapılmıştır. ARIMA ve Prophet gibi zaman serisi yöntemleri ile XGBoost, LightGBM, Random Forest ve LSTM gibi makine öğrenmesi algoritmaları kullanılarak farklı tahmin modelleri geliştirilmiştir. Ayrıca USD/TRY ve altın fiyatlarının hisse fiyatları üzerindeki etkisi regresyon analizi ile incelenmiştir. Risk, getiri ve volatilite analizleri gerçekleştirilmiş ve elde edilen bulgular, model karşılaştırmalarıyla birlikte değerlendirilmiştir.

İstatistiksel Ön İşleme ve Tanımlayıcı Analiz

	Ortalama Fiyat	Standart Sapma	Min Fiyat	Max Fiyat	Toplam Getiri
YKBNK	10.144412	10.178799	1.48	37.96	13.511848
ISCTR	4.591615	4.721590	0.68	17.36	12.612245
GARAN	31.769384	35.004963	5.32	128.55	11.954644
TSKB	4.331639	3.729662	0.86	13.13	9.268908
AKBNK	18.871519	19.208489	3.77	68.28	8.654490
SKBNK	2.080176	1.499271	0.50	6.01	4.386364
ALBRK	2.570616	1.607305	0.80	6.45	4.252101
VAKBN	9.170392	6.386668	3.22	25.52	3.201802
HALKB	9.155004	4.655364	4.18	19.82	1.642741
ICBCT	8.539241	4.037717	2.91	19.54	1.617702
KLNMA	5.425308	2.750212	2.26	18.85	0.933649

2020–2024 dönemine ait fiyat verilerine göre, YKBNK, ISCTR ve GARAN hisseleri en yüksek toplam getiriye sahip bankalar olarak öne çıkmaktadır. Bu hisseler aynı zamanda yatırımcıya uzun vadede daha yüksek kazanç potansiyeli sunmuştur. GARAN ve AKBNK yüksek ortalama fiyat ve volatilite değerleriyle dikkat çekerken, KLNMA, ICBCT ve HALKB gibi hisseler daha düşük fiyat seviyeleri ve sınırlı getiri oranlarıyla görece zayıf bir performans sergilemiştir. Bu göstergeler, bankalar arası risk-getiri farklılıklarını ortaya koymak ve portföy tercihlerine yön vermek açısından önemlidir.

ZAMAN SERİLERİ

Zaman serisi modelleme sürecinde ARIMA ve Prophet yöntemleri kullanılmıştır. Seriler öncelikle durağanlaştırılmış, ardından uygun model parametreleri belirlenerek tahminleme yapılmıştır. ARIMA modeli otoregresif ve hareketli ortalama bileşenlerini temel alırken, Prophet modeli trend ve mevsimsellik yapısını ayrı bileşenlerle ele almıştır. Her iki yöntem ileriye dönük fiyat tahminlerinde kullanılmış ve performansları istatistiksel hata metrikleriyle değerlendirilmiştir.

ARIMA ve Prophet modellerinin tahmin performansları, test verisi üzerinde hesaplanan RMSE ve MAE hata metrikleri ile karşılaştırılmıştır. Elde edilen sonuçlara göre, düşük volatiliteye sahip hisselerde (örneğin KLNMA, SKBNK, ALBRK) her iki model de düşük hata düzeyleri ile başarılı tahminler üretmiştir. Öte yandan, GARAN ve AKBNK gibi yüksek dalgalanma içeren hisselerde her iki modelde de hata oranları belirgin şekilde yükselmiştir. Genel değerlendirmede, Prophet modeli bazı hisselerde görsel açıklanabilirlik avantajı sunmasına rağmen, ARIMA modeli ortalama olarak daha düşük RMSE ve MAE değerleri ile daha tutarlı ve istatistiksel olarak daha başarılı bir performans sergilemiştir.

Banka	ARIMA_RMSE	ARIMA_MAE	Prophet_RMSE	Prophet_MAE
KLNMA	0.2634	0.2379	0.6546	0.6120
SKBNK	0.5357	0.4227	0.6668	0.5504
HALKB	0.6594	0.5339	1.6106	1.4716
ALBRK	0.8179	0.7199	0.7572	0.6947
TSKB	0.8611	0.7912	1.3025	1.2752
ICBCT	1.0803	0.9317	0.6659	0.5481
ISCTR	1.9266	1.7160	2.4912	2.3702
VAKBN	1.9993	1.7860	2.3299	1.9540
YKBNK	4.7086	4.0224	5.7924	5.2877
AKBNK	11.4504	9.8021	12.8425	11.6430
GARAN	13.1741	11.0375	19.6759	18.3758

MAKİNE ÖĞRENMESİ

Bu çalışmada, hisse senedi fiyat tahmini amacıyla çeşitli makine öğrenmesi algoritmaları kullanılmıştır. XGBoost, LightGBM ve Random Forest gibi ağaç tabanlı yöntemler; verideki karmaşık ilişkileri, değişken önem düzeylerini dikkate alarak tahmin üretir. Bu algoritmalar, çok sayıda karar ağacını bir araya getirerek yüksek doğruluk sağlar. Zaman serisi verisiyle çalışabilen LSTM (Long Short-Term Memory) ağı ise, geçmiş verilere dayalı bağımlılıkları öğrenerek zaman boyutlu tahminleme yapar. Tüm modeller, teknik analiz göstergeleri ve makroekonomik değişkenlerle eğitilmiş; test verisi üzerinde performansları istatistiksel hata metrikleriyle değerlendirilmiştir.

	AKBNK	GARAN	ISCTR	YKBNK	VAKBN	HALKB	TSKB	ALBRK	SKBNK	ICBCT	...	lag_3	lag_4	lag_5	ma_5	ma_20	log_return	weekly_return	RSI	bollinger_upper	bollinger_lower
2020-01-30	6.79	9.94	1.09	2.53	6.75	6.91	1.35	1.31	0.88	5.67	...	6.86	7.03	7.07	6.820	6.8425	0.020834	-0.039604	41.121495	7.464851	6.220149
2020-01-31	6.63	9.78	1.09	2.48	6.60	6.92	1.34	1.31	0.89	5.65	...	6.77	6.86	7.03	6.740	6.8550	-0.023846	-0.056899	34.710744	7.447568	6.262432
2020-02-03	6.70	9.72	1.11	2.45	6.59	6.88	1.33	1.31	0.89	5.54	...	6.65	6.77	6.86	6.708	6.8815	0.010503	-0.023324	31.896552	7.385934	6.377066
2020-02-04	6.91	10.08	1.14	2.53	6.79	7.41	1.36	1.34	0.91	5.62	...	6.79	6.65	6.77	6.736	6.9140	0.030862	0.020679	44.274809	7.324925	6.503075
2020-02-05	6.96	10.17	1.13	2.51	6.77	7.21	1.34	1.32	0.91	5.71	...	6.63	6.79	6.65	6.798	6.9425	0.007210	0.046617	44.274809	7.271254	6.613746

5 rows × 23 columns

Makine öğrenmesi modelleri için oluşturulan veri seti, bankaların günlük kapanış fiyatlarının yanı sıra teknik göstergeler (RSI, log_return, weekly_return, hareketli ortalamalar, Bollinger bantları) ve gecikmeli fiyat bilgileri (lag_3, lag_4, lag_5) ile zenginleştirilmiştir. Bu değişkenler, fiyat hareketlerinin yapısını modellemek ve geleceğe yönelik tahminler üretmek amacıyla algoritmalara girdi olarak sunulmuştur.

	Banka	Model	RMSE	MAE	R^2
8	SKBNK	LSTM	0.1320	0.0976	0.8760
7	ALBRK	LSTM	0.2626	0.2112	0.8772
6	TSKB	LSTM	0.3888	0.3012	0.9409
9	ICBCT	LSTM	0.5235	0.4156	0.8812
10	KLNMA	LSTM	0.5295	0.3917	0.6967
5	HALKB	LSTM	0.5923	0.4422	0.8879
2	ISCTR	LSTM	0.7254	0.6012	0.8638
4	VAKBN	LSTM	0.8826	0.6767	0.9403
3	YKBNK	LSTM	1.1801	0.9208	0.9199
0	AKBNK	LSTM	2.8617	2.2660	0.9054
1	GARAN	LSTM	4.1863	3.2577	0.9685

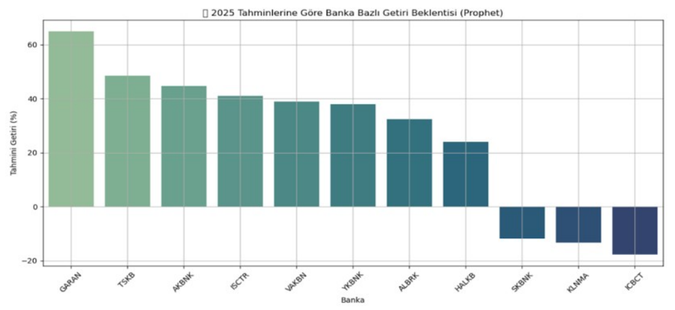
Tablodaki sonuçlara göre, LSTM modeli genel olarak düşük hata metrikleri ve yüksek R^2 değerleri ile banka hisselerinin fiyat hareketlerini başarılı bir şekilde tahmin edebilmiştir. Özellikle SKBNK, ALBRK ve TSKB hisselerinde RMSE ve MAE değerleri oldukça düşük, R^2 skorları ise yüksek çıkmıştır. Bu durum, modelin bu hisselerde geçmiş fiyat hareketlerinden öğrenerek daha doğru tahminler üretebildiğini göstermektedir. Buna karşın, AKBNK ve GARAN hisselerinde tahmin hatalarının daha yüksek olduğu, dolayısıyla bu hisselerde modelin fiyat dinamiklerini tam olarak yakalayamadığı görülmektedir. RMSE ve MAE değerlerinin düşük olması, modelin tahminlerinde daha az hata yaptığını; R^2 değerinin yüksek olması ise modelin fiyat hareketlerini daha iyi açıklayabildiğini göstermektedir. Genel olarak, LSTM modeli birçok banka için başarılı tahminler üretmiş olsa da bazı hisselerde daha karmaşık fiyat yapısı nedeniyle performansın sınırlı kaldığı söylenebilir.

	Banka	Model	Best RMSE	Params
8	SKBNK	XGBoost	0.4190	{'subsample': 1.0, 'n_estimators': 100, 'max_d...
7	ALBRK	XGBoost	0.5438	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
10	KLNMA	XGBoost	0.7606	{'subsample': 0.7, 'n_estimators': 100, 'max_d...
6	TSKB	XGBoost	1.0305	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
9	ICBCT	XGBoost	1.0380	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
5	HALKB	XGBoost	1.1295	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
2	ISCTR	XGBoost	1.7801	{'subsample': 1.0, 'n_estimators': 100, 'max_d...
4	VAKBN	XGBoost	1.9988	{'subsample': 1.0, 'n_estimators': 100, 'max_d...
3	YKBNK	XGBoost	3.8792	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
0	AKBNK	XGBoost	7.7773	{'subsample': 0.7, 'n_estimators': 200, 'max_d...
1	GARAN	XGBoost	15.0478	{'subsample': 0.7, 'n_estimators': 200, 'max_d...

	Banka	Model	Best RMSE	Params
8	SKBNK	LightGBM	0.4216	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
7	ALBRK	LightGBM	0.5716	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
10	KLNMA	LightGBM	0.8677	{'num_leaves': 31, 'n_estimators': 100, 'max_d...
9	ICBCT	LightGBM	1.1241	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
6	TSKB	LightGBM	1.1630	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
5	HALKB	LightGBM	1.2634	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
2	ISCTR	LightGBM	1.8549	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
4	VAKBN	LightGBM	2.1472	{'num_leaves': 20, 'n_estimators': 100, 'max_d...
3	YKBNK	LightGBM	4.0368	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
0	AKBNK	LightGBM	7.9451	{'num_leaves': 31, 'n_estimators': 200, 'max_d...
1	GARAN	LightGBM	15.8363	{'num_leaves': 50, 'n_estimators': 100, 'max_d...

	Banka	Model	Best RMSE	Params
8	SKBNK	Random Forest	0.4178	{'n_estimators': 200, 'min_samples_split': 5, ...
7	ALBRK	Random Forest	0.5483	{'n_estimators': 100, 'min_samples_split': 5, ...
10	KLNMA	Random Forest	0.8183	{'n_estimators': 100, 'min_samples_split': 2, ...
9	ICBCT	Random Forest	1.0554	{'n_estimators': 200, 'min_samples_split': 5, ...
6	TSKB	Random Forest	1.0737	{'n_estimators': 200, 'min_samples_split': 5, ...
5	HALKB	Random Forest	1.1939	{'n_estimators': 100, 'min_samples_split': 2, ...
2	ISCTR	Random Forest	1.7931	{'n_estimators': 100, 'min_samples_split': 2, ...
4	VAKBN	Random Forest	1.9877	{'n_estimators': 100, 'min_samples_split': 5, ...
3	YKBNK	Random Forest	3.8824	{'n_estimators': 200, 'min_samples_split': 2, ...
0	AKBNK	Random Forest	7.8459	{'n_estimators': 200, 'min_samples_split': 2, ...
1	GARAN	Random Forest	15.1620	{'n_estimators': 200, 'min_samples_split': 2, ...

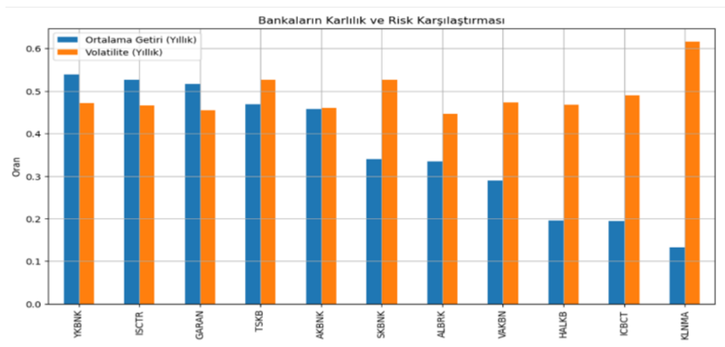
Makine öğrenmesi modelleri karşılaştırıldığında, XGBoost çoğu banka için en düşük RMSE değerini vererek en başarılı model olmuştur. SKBNK, ALBRK ve KLNMA gibi hisselerde en iyi sonuçlar XGBoost ile elde edilmiştir. Random Forest bazı hisselerde rekabetçi sonuçlar sunsa da genel başarı XGBoost'tadır. LightGBM ise daha yüksek hata değerleriyle görece zayıf kalmıştır.

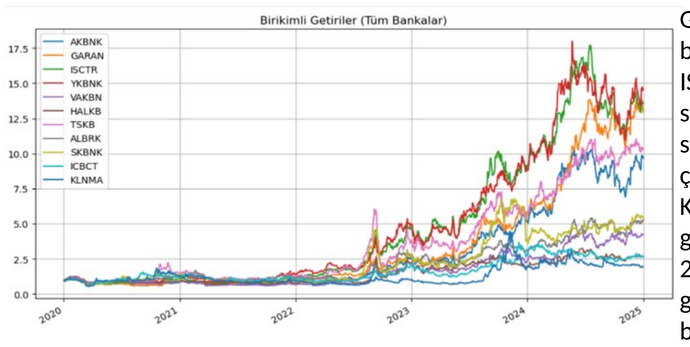


Prophet modeli ile yapılan 2025 yılı tahminlerinde, GARAN, TSKB ve AKBNK hisseleri en yüksek getiri potansiyeline sahip bankalar olarak öne çıkmıştır. ISCTR, VAKBN ve YKBNK hisselerinde de pozitif beklentiler gözlemlenmiştir. Öte yandan, SKBNK, KLNMA ve ICBCT için negatif getiriler tahmin edilmiş, bu hisselerde düşüş riski öne çıkmıştır.

Risk ve Getiri Analizi

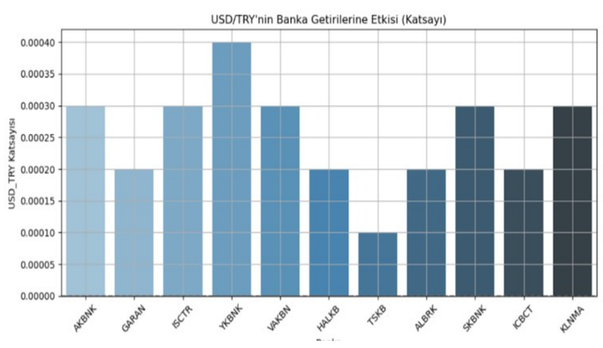
Grafikte bankaların yıllık ortalama getirileri ve volatilite oranları birlikte değerlendirilmiştir. YKBNK, ISCTR ve GARAN en yüksek yıllık getiri sunan bankalardır; YKBNK ve ISCTR daha dengeli risk profiline sahipken, GARAN yüksek riskle birlikte yüksek getiri potansiyeli taşımaktadır. KLNMA ve ICBCT ise düşük getiri ve yüksek volatilite ile olumsuz bir risk-getiri dengesi sergilemiştir.





Grafikte, 2020–2024 döneminde bankaların birikimli getiri performansları karşılaştırılmıştır. ISCTR ve YKBNK en yüksek birikimli getiriyi sağlayarak güçlü uzun vadeli trendler sergilemiştir. GARAN ve AKBNK de dikkat çeken pozitif eğilim göstermiştir. Buna karşılık KLNMA ve ICBCT gibi hisseler düşük birikimli getirilerle daha durağan bir profil sunmuştur. 2022 sonu itibariyle bankacılık hisselerinde genel bir yükseliş gözlenmiş, ancak bu artışın banka bazında farklılaştığı görülmüştür.

2020–2024 döneminde bankaların USD/TRY değişimlerine duyarlılıkları regresyon analiziyle incelenmiştir. Tüm bankaların beta katsayıları oldukça düşük (0.0001–0.0004) seviyelerde bulunmuş, bu da döviz kurunun hisse getirileri üzerindeki etkisinin sınırlı olduğunu göstermektedir. YKBNK, AKBNK ve VAKBN görece daha yüksek beta değerleriyle kur değişimlerinden biraz daha fazla etkilenmiştir. P-değerlerinin 0.05'in üzerinde ve R^2 değerlerinin çok düşük olması, bu ilişkinin istatistiksel olarak anlamlı olmadığını ortaya koymaktadır. Sonuçlar, banka getirilerinin dövizden çok yerel ekonomik faktörlerle şekillendiğine işaret etmektedir.



KAYNAKÇA

[1] İş Yatırım Menkul Değerler A.Ş. (2024). Hisse senedi veri arşivi. https://www.isyatirim.com.tr/tr

[2] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.

[3] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks... *European Journal of Operational Research, 270(2)*, 654–669.

2024-2025 Yarıyılı İstatistik BÖLÜMÜ

Uçuş Verilerinin İstatiksel Analizi

Ayşe Nur ERMAN/ 21023606

Danışman: Prof. Dr. Filiz KARAMAN

ÖZET

Bu tez, 2013 yılına ait 21 değişkenli ve yaklaşık 336000 satırlık uçuş verileri analiz ederek, havacılık endüstrisindeki çeşitli faktörlerin etkisini incelemeyi amaçladı. Analizde özet istatistikler, veri görselleştirme(histogram, boxplot, scatter plot), korelasyon analizi ve makine öğrenmesi(random forest modeli) kullanılmıştır.

Veri Kümesinin Tanımlanması

id	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour	name
0	2013	1	1	517.0	515	2.0	830.0	819	11.0	UA	1545	N14228	EWR	IAH	227.0	1400	5	15	2013-01-01 05:00:00	United Air Lines Inc.
1	2013	1	1	533.0	529	4.0	850.0	830	20.0	UA	1714	N24211	LGA	IAH	227.0	1416	5	29	2013-01-01 05:00:00	United Air Lines Inc.
2	2013	1	1	542.0	540	2.0	923.0	850	33.0	AA	1141	N619AA	JFK	MIA	160.0	1089	5	40	2013-01-01 05:00:00	American Airlines Inc.
3	2013	1	1	544.0	545	-1.0	1004.0	1022	-18.0	B6	725	N804JB	JFK	BQN	183.0	1576	5	45	2013-01-01 05:00:00	JetBlue Airways

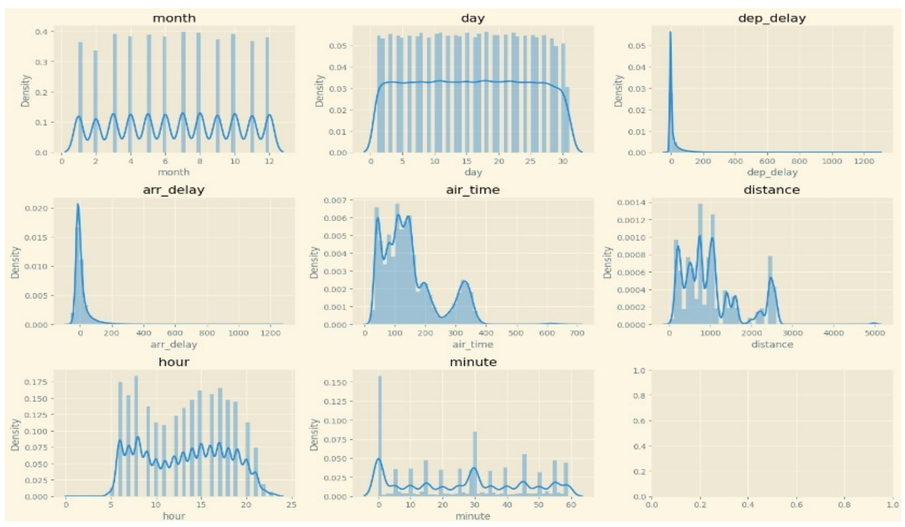
Uçuşların 2013 yılı içerisindeki bilgileri içeren bu veri kümesi 336775 satırdan ve yukarıdaki tabloda yer alan 21 değişkenden oluşmaktadır. Her bir satır ilgili uçuşun değerlerini içermektedir.

	count	unique	top	freq
carrier	336776	16	UA	58665
flight	336776	3844	15	968
tailnum	334264	4043	N725MQ	575
origin	336776	3	EWR	120835
dest	336776	105	ORD	17283
name	336776	16	United Air Lines Inc.	58665
arr_time_timebased	328063	1411	10:08	485
dep_time_timebased	328521	1318	05:55	834
sched_arr_time_timebased	336776	1163	10:25	1324
sched_dep_time_timebased	336776	1021	06:00	7016

.

Yukarıdaki tablodan çıkartılan sonuçlar:

- Veri.describe() fonksiyonuna göre, en fazla uçuş UA havayolu tarafından gerçekleştirilmiştir.
- EWR kalkış noktasından en fazla uçuş gerçekleşmiştir.
- United Air Lines en popüler havayolu taşıyıcısıdır.
- ORD en çok uçuşun varış noktasıdır.
- Uçuş ve tailnum özellikleri sırasıyla uçuşun uçuş numarasını ve benzersiz tanımlayıcısını gösterir. Bu sütunlar çok sayıda benzersiz değer içerdiğinden, analiz için pek fazla bilgi sağlamazlar.
- Çoğu uçuşun sabah saatlerinde planlandığını görebiliriz.



Yukarıdaki dağılım grafiği, uçuşların yılın her günü(day) ve ayında(month) eşit bir şekilde dağıldığını gösteriyor. Dep_delay ve arr_delay sağa çarpık (right-skewed). Bu, bazı uçuşların diğer uçuşlara göre yüksek gecikmeye sahip olduğu anlamına gelir.

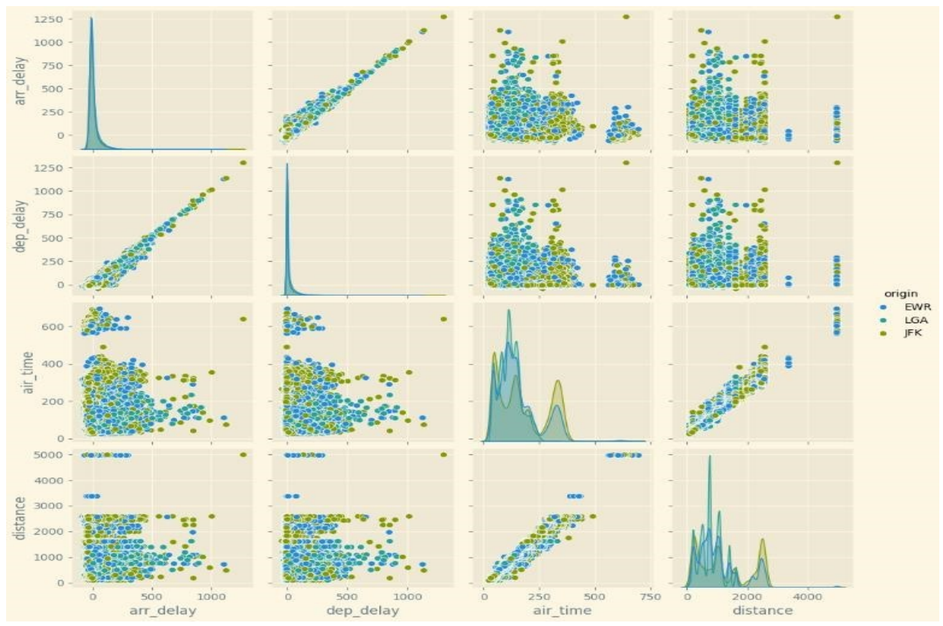
Air_time ve mesafe özellikleri asimetriktir. Bu, veriyi daha iyi analiz etmek istiyorsak, veriyi air_time ve mesafe kategorilerine ayırmalı ve sonra her bir kısmı ayrı ayrı analiz edebilmemiz gerektiği anlamına gelir.
Örneğin, air_time'ı dört kategoriye ayırabiliriz:

100 dakikadan az
100 - 200 dakika
200 - 400 dakika
400 dakikadan fazla

Ayrıca, mesafe özelliği için şu şekilde yapabiliriz:

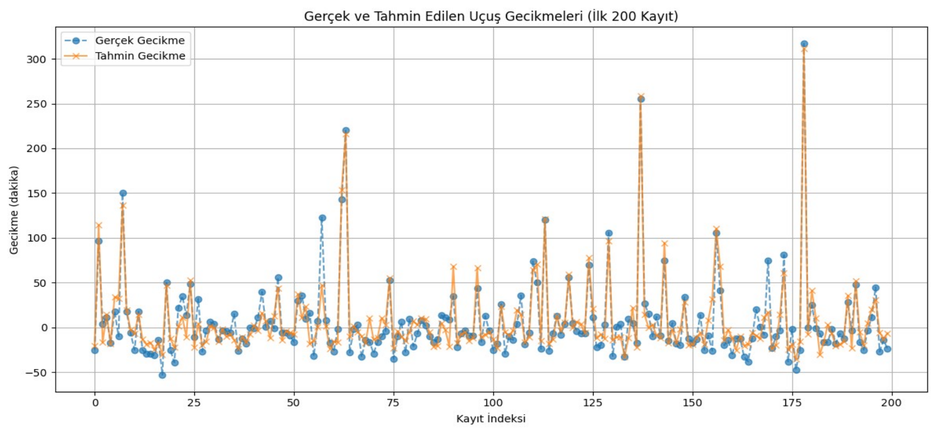
1000 milin altında
1000 - 2000 mil
2000 - 3000 mil
3000 milin üstünde

Saat özelliği, uçuşların çoğunun sabah 5 ile 8 arasında ve ayrıca öğleden sonra 3 ile 5 arasında gerçekleştiğini gösteriyor.
Dakika özelliği de, çoğu uçuşun saat başlarında ve saatlerin ortasında gerçekleştiğini gösteriyor.



Yukarıdaki grafik ise pairplot, varis_gecikme ve kalkis_gecikme arasında lineer bir ilişki gösteriyor. Ayrıca, bu durum mesafe ve havada_kalis arasında da geçerlidir.
Ayrıca, yukarıdaki grafik, 500 dakikadan fazla süren uçuşların daha az gecikmeye sahip olduğunu göstermektedir.
Dahası, 3000 milin üzerinde bir mesafeye sahip uçuşların daha az gecikmeye sahip olduğu görülmektedir

Makine Öğrenmesi



Veri setine Random Forest Regressor Modeli uygulanmıştır. Buna göre;

- RMSE(Root Mean Squared Errror):Tahmini hataların karekök ortalamasıdır. Model gerçek gecikme süresinden ortalama 16.44 dakika sapıyor demektir.
- MAE(Mean Absolute Error): Modelin ortalama olarak tahminlerinde yaptığı mutlak hatadır. Yani model gecikmeyi ortalama 11.67 dakika sapmayla tahmin ediyor.

Bu sonuçlara göre:

- **MAE'nin 11.67 dakika olması**, modelin pratikte orta derecede iyi tahmin yaptığını gösteriyor. 11.67 dakikalık hata, hava yolu planlaması gibi uygulamalarda kabul edilebilir olabilir ama çok hassas uygulamalarda yetersiz sayılır.
- **RMSE'nin MAE'den büyük olması**, bazı uçuşlarda **çok büyük tahmin hataları** (aykırı değerler) olduğunu gösteriyor. Örneğin bazı uçuşlar 200–300 dakika gecikmiş olabilir ve model bu ekstrem gecikmeleri yeterince iyi tahmin edememiş olabilir.

SONUÇ

Bu çalışma sonucunda;

- Uçuş gecikmelerinde en etkili faktörlerin kalkış saati, hava yolu şirketi, mesafe ve havaalanı olduğu görülmüştür.
- Makine öğrenmesi ile ortalama 11 dakikalık hata payı ile gecikme tahmini yapılabilmiştir.
- Bu analizler, havayolu şirketlerinin operasyonel planlamalarına katkı sağlayabilir.
- Gelecekte veri seti genişletilerek farklı değişkenlerle (hava durumu gibi) daha hassas tahminler yapılabilir.

KAYNAKÇA

[1] Chowdhury ve Rahman, 2012
[2] Belobaba, 2002
[3] ABBAS SEIFOSSADAT - Exploratory Data Analysis on Flights Dataset

İSTATİSTİK BÖLÜMÜ

DV Lottery Verileri Analizi ve Green Card Vizesi Alımının Sınıflandırma Modeliyle Tahmini

Barış Köse 20023051

Danışman: Prof.Dr.Ersoy Öz

ÖZET

Bu çalışmada, Diversity Visa (DV) Lottery sürecinde vize alım durumunu tahmin etmek amacıyla 2014–2024 yıllarına ait CEAC (Consular Electronic Application Center) verileri ve aylık vize sıralaması eşik değerleri (cut-off) kullanılarak bir sınıflandırma modeli geliştirilmiştir. Kullanılan veri seti; case number (dosya numarası), DS-260 formunun işlenme tarihi (submitDate), mülakat tarihi (2nlDate), bölge (region) ve cut-off kapsamında olup olmama durumu (in_cutoff) gibi değişkenleri içermektedir. Makine öğrenmesi algoritmalarından XGBoost ile oluşturulan modelin başarımı, accuracy, recall, precision, f1-score ve AUC metrikleri ile değerlendirilmiştir. Model çıktıları, yalnızca dosya numarası değil; formun işlenme zamanı ve kişinin açıklanan cut-off değerleri kapsamında olup olmamasının da vize alma ihtimalinde belirleyici olduğunu göstermektedir.
Ayrıca, 2014–2024 yılları arasında yayımlanan resmi DV Lottery istatistikleri kullanılarak, yıllık seçilen kişi sayıları ve bölgesel dağılımlar analiz edilmiştir.

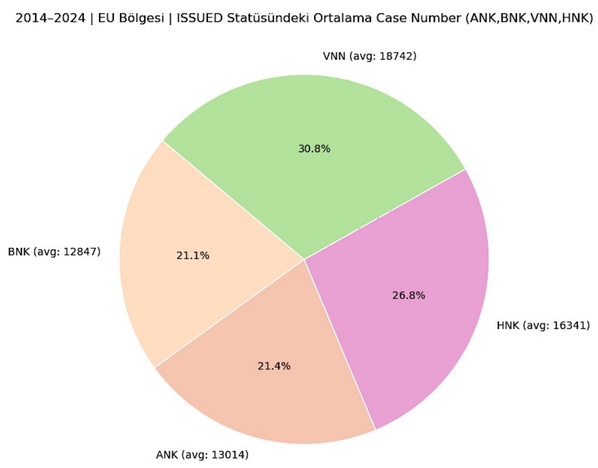
ABD Çeşitlilik Vizesi Sürecine Veri Odaklı Bir Bakış

Diversity Visa (DV) Lottery Programı, her yıl dünya genelinden 50.000'e yakın kişiye ABD'de kalıcı oturma ve çalışma hakkı tanıyan önemli bir göçmenlik sistemidir. Süreç, rastgele seçimle başlamakta; ancak vize hakkı kazanmak, sadece seçilmekle sınırlı kalmayıp, dosya numarası sıralaması ve vize kotasına bağlı olarak mülakat alma şansına da dayanmaktadır.
Bu belirsizlik, özellikle yüksek dosya numarasına sahip başvuru sahipleri için önemli bir karar problemi yaratmaktadır. Dolayısıyla, başvuru sürecinde veriye dayalı tahmin araçlarının geliştirilmesi, başvuru sahiplerine süreç hakkında öngörü sunmak açısından oldukça değerlidir.
Bu bağlamda, bu çalışma DV Lottery sürecindeki vize alma olasılığını tahmin etmek amacıyla makine öğrenmesi tabanlı bir model geliştirmekte ve süreci etkileyen temel faktörleri analiz etmektedir.

VERİ SETİNİ TANIYALIM

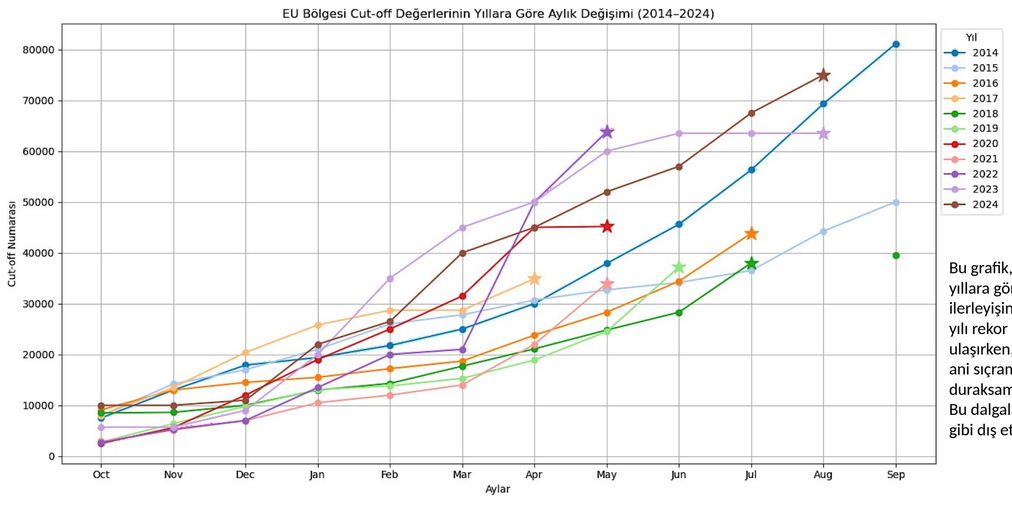
Bu çalışmada üç temel veri kaynağı kullanılmıştır:

1. **CEAC Verileri :** ABD Dışişleri Bakanlığı'na bağlı **CEAC (Consular Electronic Application Center)** sistemi üzerinden elde edilen bu veri seti, DV Lottery kazananlarına ait bireysel başvuru süreçlerini içermektedir. Dosya numarası, formun işlenme tarihi, mülakat bildirimi ve vize sonucu gibi değişkenler yer almaktadır.



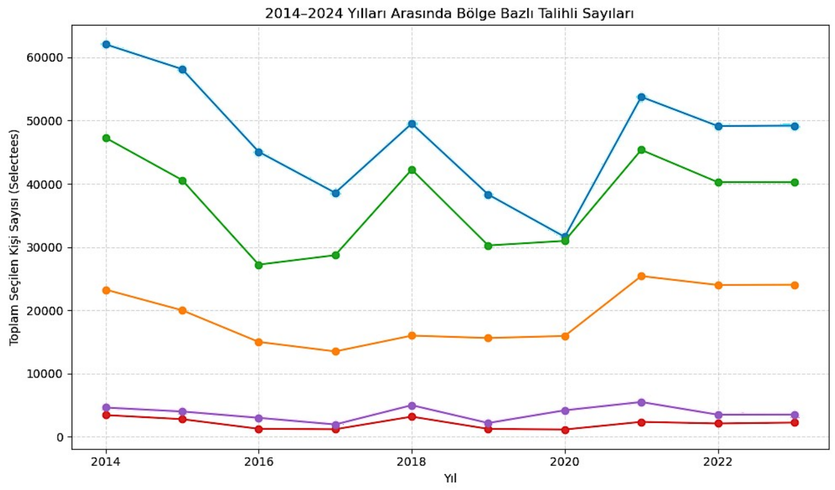
Grafik, farklı konsoloslukların Avrupa bölgesinden gelen başvuruları nasıl işlediğini göstermektedir.
Viyana (VNN) ve Hong Kong (HNK), daha yüksek ortalama case number'larla öne çıkarak, ileri sıradaki başvuruları da değerlendirdiklerini göstermektedir.
Ankara (ANK) ve Bangkok (BNK) ise daha düşük ortalama değerlerle, genellikle daha erken çağrılan başvurulara odaklanmıştır.
Bu farklar, konsoloslukların işlem kapasitesi, yoğunluğu ve önceliklendirme yaklaşımlarından kaynaklanabilir.

2. **Visa Bulletin Aylık Cut-off Verileri:** Her ay yayımlanan bu duyurular, başvurular arasından **hangi dosya numaralarının mülakata çağrılabileceğini** gösteren cut-off (eşik) değerlerini içermektedir. Bu değerler modelde CEAC verisiyle birleştirilerek **in_cutoff** değişkeni oluşturulmuştur.



Bu grafik, EU bölgesinde yıllara göre cut-off ilerleyişini gösterir. 2024 yılı rekor seviyeye ulaşırken, 2020 ve 2022'de ani sıçramalar veya duraksamalar görülmüştür. Bu dalgalanmalar, pandemi gibi dış etkenlerle ilişkilidir.

3. **Selectee (Talihli) Sayıları:** Resmi kaynaklardan derlenen bu veri seti, yıllara ve ülkelere göre DV Lottery kapsamında **seçilen kişi sayılarını** içermektedir. Bu sayılar kullanılarak **yıllık değişim**, **bölgesel yoğunluk** ve **programın zaman içindeki ölçeği** analiz edilmiştir.



Bu grafikte 2014–2024 arasında en çok talihli AFRICA ve EUROPE bölgelerinden seçildiği gözlemlenir. ASIA istikrarlı seyrederken, OCEANIA ve SOUTH AMERICA'nın kontenjanı düşüktür.

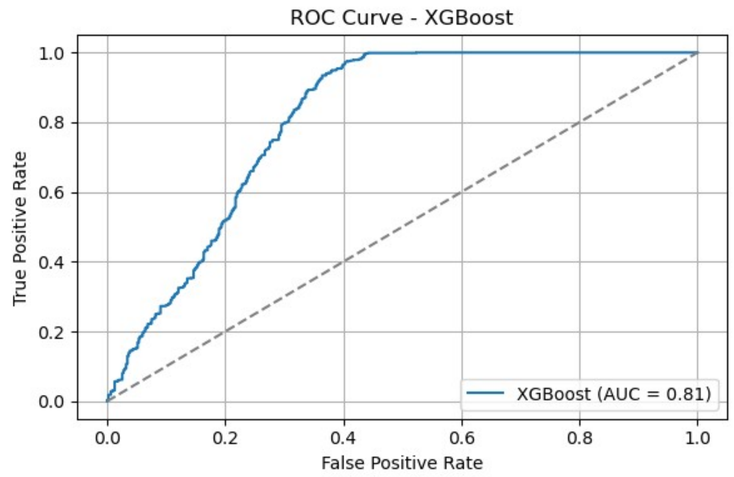
Makine Öğrenmesi ile Vize Sonucunun Tahmini (XGBoost Modeli)

Modelin Amacı ve Yöntemi

Bu çalışmada, Diversity Visa (DV) Lottery başvurularında bir kişinin vize alıp almayacağını tahmin etmek amacıyla bir sınıflandırma modeli geliştirilmiştir.
Tahminleme için **XGBoost algoritması** kullanılmıştır; bu yöntem, yüksek doğruluk oranı ve esnekliğiyle bilinen güçlü bir gradient boosting algoritmasıdır.

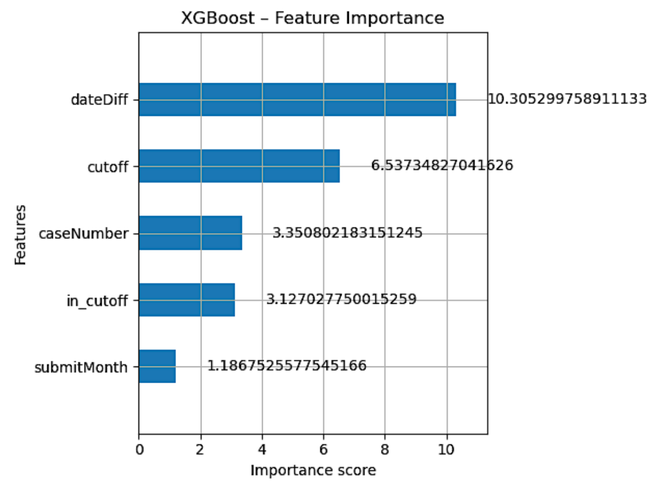
Model Performansı – ROC Eğrisi ve AUC Skoru

ROC (Receiver Operating Characteristic) eğrisi, modelin farklı eşik değerleri altında pozitif ve negatif sınıfları ne kadar iyi ayırt ettiğini gösterir. Bu eğrinin altında kalan alan olan **AUC (Area Under the Curve)**, modelin genel ayırt edicilik başarısını ölçer.
AUC skoru: 0.81; Bu, modelin bir başvurunun vize alıp alamayacağını **%81 doğrulukla ayırt edebildiğini** göstermektedir. Eğri, rastgele tahmine göre (gri kesikli çizgi) anlamlı ölçüde yukarıda seyretmektedir.



Öznitelik Etkileri – Feature Importance Analizi

Modelin karar verirken hangi değişkenleri ne kadar kullandığını gösteren **feature importance** grafiği aşağıda verilmiştir. Bu analiz, başvurunun sonucu üzerinde en etkili faktörleri sıralamak açısından önemlidir.



Değişken	Açıklama	Etki
dateDiff	DS-260 işlenme (gönderim) tarihi ile 2NL tarihi arasındaki süre (gün bazında)	En yüksek
cutoff	Başvurunun ait olduğu ayın cutoff numarası	Yüksek
caseNumber	Başvurunun sıra numarası	Orta
in_cutoff	Case number'ın o ay cutoff altında olup olmadığı	Orta
submitMonth	DS-260 formunun gönderildiği ay	Düşük

Genel Değerlendirme

Bu çalışma, DV Lottery başvurularında vize alma olasılığını tahmin eden bir sınıflandırma modeli (XGBoost) geliştirmiştir. Model, test verisi üzerinde %83 doğruluk, %95 recall, %81 precision, %87 F1-score ve 0.81 ROC-AUC skoru ile yüksek performans sergilemiştir. Özellikle başvuru ile 2NL tarihi arasındaki sürenin kısa olması (dateDiff), vize alınma ihtimalini artıran en güçlü değişken olarak öne çıkmıştır. Model çıktıları, başvuru stratejilerini optimize etmek isteyen adaylara veri temelli bir yol haritası sunmaktadır.

KAYNAKÇA

[1] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794
[2] Kulkarni, A., Patel, S., & Mehta, R. (2018). *H1B Visa Prediction Using Machine Learning Algorithms*. International Journal of Engineering Research and Technology (IJERT), 7(7), 394–397.
[3] Kumar, R., & Naresh, E. (2018). *H1B Visa Application Status Prediction Using Machine Learning Algorithms*. International Journal of Computer Sciences and Engineering, 6(5), 871–875.

İSTATİSTİK BÖLÜMÜ

ÜNİVERSİTE ÖĞRENCİLERİNDE PSİKOLOJİK ETKENLERİN AKADEMİK BAŞARIYA ETKİSİ

Burhan Yasa 19023028

Danışman: Dr. Öğr. Üyesi Tugay KARADAĞ

ÖZET

Bu çalışmanın amacı, yaygın olarak görülen psikolojik semptomların üniversite öğrencilerinde akademik başarı üzerindeki etkisini incelemektir. Güney Afrika'daki bir devlet üniversitesine kayıtlı 506 birinci sınıf öğrencisinden elde edilen anket verileri kullanılmıştır. Öğrencilerin dikkat eksikliği ve hiperaktivite bozukluğu (DEHB), depresyon ve anksiyete düzeyleri; Dikkat Eksikliği ve Hiperaktivite Bozukluğu Öz Bildirim Ölçeği (ASRS-V1.1), Beck Depresyon Envanteri-II (BDI-II) ve Beck Anksiyete Envanteri (BAI) ölçekleriyle ölçülmüştür. Akademik başarı göstergesi olarak; üniversitenin öğrenci bilgi sisteminden alınan ilk dönem not ortalaması (GPA), lise mezuniyet notu ve üniversiteye girişte kullanılan sınavların ortalamasıyla elde edilen puan kullanılmıştır. İlk olarak betimsel istatistikler sunulmuş, ardından değişkenler arası ilişkiler korelasyon analiziyle incelenmiştir. Son olarak, psikolojik semptomların akademik başarı üzerindeki etkisini değerlendirmek amacıyla regresyon modelleri uygulanmıştır. Analiz sonuçları doğrultusunda ilgili psikolojik faktörlerin akademik performansla ilişkisi çok yönlü olarak tartışılmıştır.

GİRİŞ

Üniversite öğrencilerinin akademik başarıları yalnızca bilişsel yeterlilikleriyle değil psikolojik faktörlerle de yakından ilişkilidir. Bu çalışmada depresyon, anksiyete ve dikkat eksikliği hiperaktivite bozukluğu (DEHB) gibi değişkenlerin başarı üzerindeki etkileri incelenmiştir. Depresyon ruh halinde çöküntü, anksiyete yoğun kaygı, DEHB ise dikkat azalması ve dürtüsellikle karakterize edilen bozukluklardır. Bu tür psikolojik etkenlerin başarıya olan katkısı öğrenci destek hizmetlerinin planlanması açısından önemlidir.

YÖNTEM VE METOD

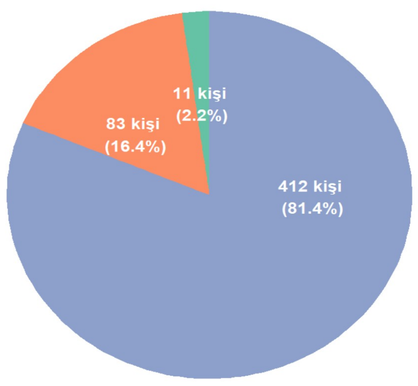
Çalışmada öncelikle betimleyici analizler yapılarak değişkenlerin ortalama, medyan ve standart sapma gibi merkezi eğilim ölçüleri hesaplanmış ve verilerin genel dağılımı hakkında bilgi verilmiştir. Normallik varsayımı Shapiro-Wilk testi ile değerlendirilmiştir. Değişkenlerin normal dağılmadığı tespit edildikten sonra gruplar arası farkları test etmek amacıyla parametrik olmayan Mann-Whitney U testi kullanılmıştır. Bu test iki bağımsız grup arasında medyan farkının anlamlı olup olmadığını değerlendirir.
Değişkenler arası ilişkiyi belirlemek için Spearman korelasyon analizi uygulanmıştır. Spearman korelasyonu, sıralı veriler için kullanılan ve değişkenler arasındaki yönlü ilişkiyi ortaya koyan parametrik olmayan bir testtir. Korelasyon katsayıları −1 ile +1 arasında değişir; değer +1'e yaklaştıkça pozitif yönlü güçlü bir ilişki, −1'e yaklaştıkça negatif yönlü güçlü bir ilişki olduğu anlaşılır.
Son olarak, psikolojik değişkenlerin akademik başarı (GPA) üzerindeki etkilerini incelemek amacıyla çoklu doğrusal regresyon analizi gerçekleştirilmiştir. Bu analiz, birden fazla bağımsız değişkenin bir bağımlı değişken üzerindeki etkisini aynı anda değerlendirir ve hangi değişkenlerin istatistiksel olarak bağımlı değişkeni açıklamada anlamlı katkı sağladığını belirlemeye yardımcı olur.

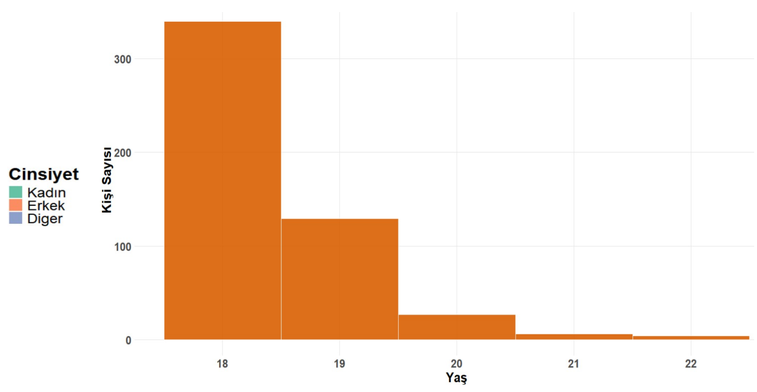
Çalışmada üç psikolojik değişken için standart öz bildirim ölçekleri kullanılmıştır. Beck Depresyon Envanteri (BDI-II) depresif belirtileri ölçen 21 maddelik bir öz bildirim ölçeğidir [2,3]. Beck Anksiyete Envanteri (BAI) anksiyete semptomlarını değerlendiren 21 maddelik bir ölçektir [3]. Dikkat Eksikliği ve Hiperaktivite Bozukluğu Öz Bildirim Ölçeği (ASRS-V1.1) dikkat eksikliği ve hiperaktivite belirtilerini ölçen 18 maddelik bir öz bildirim ölçeğidir [4].

Depresyon Ölçeği (BDI II) | Anksiyete Ölçeği (BAI) | Dikkat Eksikliği ve Hiperaktivite Bozukluğu Öz Bildirim Ölçeği (ASRS-V1.1)

Veri Seti ve Tanımlayıcı İstatistikler

Bu araştırmanın verileri, Güney Afrika'daki bir devlet üniversitesine kayıtlı 506 Psikoloji bölümü birinci sınıf öğrencilerinden elde edilmiştir. Katılımcıların psikolojik değerlendirmeleri, öz bildirim yoluyla uygulanan anket formları aracılığıyla toplanmıştır [1].





Çalışmaya katılan bireylerin yaşları 18 ile 22 arasında değişmekte olup yaş ortalaması 18.43'tür. Cinsiyet dağılımına bakıldığında katılımcıların %81.4'ü kadın (412), %16.4'ü erkek (83) iken katılımcıların %2.2'si cinsiyet sorusunda diğer seçeneğini doldurmuşlardır. Cinsiyet seçeneğini boş bırakan veya farklı şekilde dolduran 11 katılımcı analiz dışı bırakılmıştır ve analizler 495 kişi üzerinden gerçekleştirilmiştir.

Değişken	Ortalama	Standart Sapma	Medyan	Aralık
Depresyon Düzeyi (BDI-II)	16.06	10.15	14.00	0 – 58
Anksiyete Düzeyi (BAI)	20.90	12.51	20.00	0 – 62
DEHB Düzeyi (ASRS)	35.35	11.25	35.00	0 – 70
Genel Not Ortalaması (GPA)	60.63	11.11	61.00	1.5 – 88
Üniversiteye Giriş Puanı	61.73	11.21	61.50	33.5 – 88
Lise Mezuniyet Notu	74.33	6.35	74.00	58 – 92

Katılımcıların depresyon düzeyleri Depresyon Envanteri puanlarına göre ortalama **16.06** olarak belirlenmiştir. Anksiyete düzeyleri ortalama **20.90** olarak bulunmuştur. Dikkat Eksikliği ve Hiperaktivite Bozukluğu semptomlarını ölçen ASRS puan ortalaması **35.35** olarak bulunmuştur. GPA ortalaması **60.63** olarak hesaplanmıştır. Üniversiteye giriş skoru ortalaması **61.73** ve lise bitirme notu ortalaması ise **74.33** olarak tespit edilmiştir. Bu değerler 100 üzerinden değerlendirilen sistemde örneklemin genel akademik başarısının **orta düzeyde** olduğunu göstermektedir.

SONUÇLAR

Değişkenlerin büyük çoğunluğu normal dağılım göstermediğinden dolayı ikili grup karşılaştırmalarında parametrik olmayan testlerden faydalanılmıştır. Cinsiyet değişkeni ile psikolojik değişkenler arasındaki farklar **Mann-Whitney U Testi** ile değerlendirilmiştir.
Analiz sonuçları özetle şu bulguları ortaya koymuştur:

- **Depresyon düzeyi** kadınlarda erkeklere kıyasla anlamlı düzeyde daha yüksektir (U = 13237.50; Z = −3.25; p <.05).
- **GPA** kadın öğrencilerde anlamlı şekilde daha yüksektir (U = 14548.50; Z = −2.15; p <.05).
- **Üniversiteye giriş puanı** erkek öğrencilerde anlamlı olarak daha yüksektir (U = 13151.00; Z = −3.32; p <.05).
- **Anksiyete düzeyi** kadınlarda anlamlı şekilde daha yüksektir (U = 10913.50; Z = −5.20; p <.05).

Değişkenler arasındaki ikili ilişkiler **Spearman rho katsayıları** ile değerlendirilmiş ve aşağıdaki heatmap ile görselleştirilmiştir.

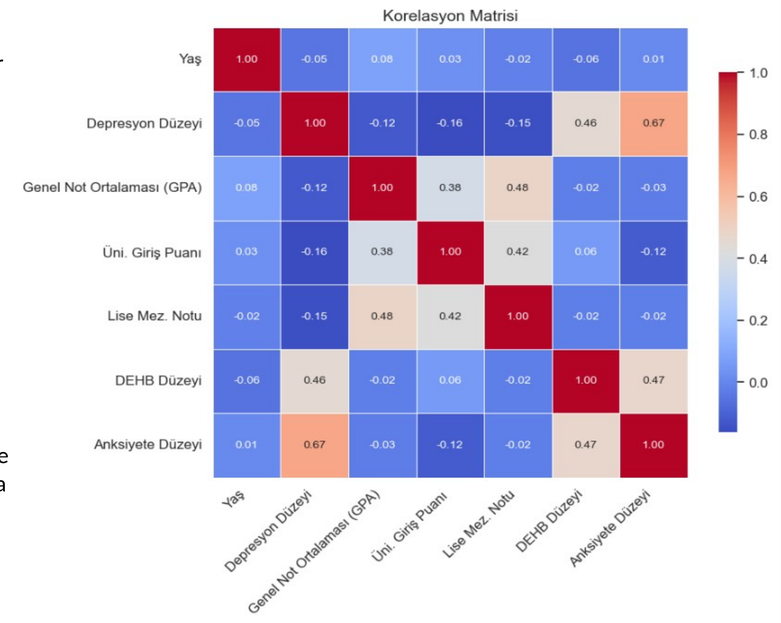
GPA ile depresyon düzeyi arasında zayıf düzeyde negatif ve anlamlı bir ilişki bulunmuştur (r = – 0.12, p < 0.05).

Depresyon ile Anksiyete arasında çok güçlü pozitif ve anlamlı ilişki bulunmaktadır (r = 0.67, p < 0.05).

Depresyon ile DEHB semptomları arasında da orta düzeyde pozitif ve anlamlı bir ilişki tespit edilmiştir (r = 0.46, p < 0.05).

Anksiyete ile DEHB arasında ile anlamlı pozitif ilişki gözlenmiştir (r = 0.47, p < 0.05).

GPA ile lise mezuniyet notu (r = 0.48, p < 0.05) ve üniversite giriş puanı (r = 0.38, p < 0.05) arasında anlamlı ve pozitif yönlü ilişkiler gözlenmiştir.



Hipotezler

H_0: Anksiyete, depresyon ve DEHB düzeyleri; yaş ve cinsiyet kontrol edildiğinde öğrencilerin akademik başarısı (GPA) üzerinde anlamlı bir etkiye sahip değildir.
H_1: Anksiyete, depresyon ve DEHB düzeyleri; yaş ve cinsiyet kontrol edildiğinde öğrencilerin akademik başarısı (GPA) üzerinde anlamlı bir etkiye sahiptir.

~~$GPA = \beta_0 + \beta_1(\text{Anksiyete Düzeyi}) + \beta_2(\text{Depresyon Düzeyi}) + \beta_3(\text{DEHB Düzeyi}) + \beta_4(\text{Yaş}) + \beta_5(\text{Cinsiyet}) + \varepsilon$~~

Değişken Adı	β (Regresyon Katsayısı)	Standart Hata	T (Test İstatistiği)	p (Anlamlılık Düzeyi)
(Sabit)	29.223	13.266	2.203	0.028
Anksiyete Düzeyi	0.033	0.054	0.604	0.546
Yaş	1.587	0.696	2.281	**0.023**
Cinsiyet	3.942	1.386	2.844	**0.005**
DEHB Düzeyi	0.052	0.051	1.021	0.308
Depresyon Düzeyi	-0.227	0.065	-3.490	**0.001**

Depresyon düzeyi, öğrencilerin GPA'sını açıklamada %5 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunmuştur.

Anksiyete ve DEHB düzeyleri, öğrencilerin GPA'sını açıklamada istatistiksel olarak anlamlı bulunmamıştır.

TARTIŞMA

Psikolojik etkenlerin akademik başarı üzerindeki rolü bu araştırmada çok yönlü biçimde değerlendirilmiş ve dikkat çekici sonuçlar elde edilmiştir.
Kadın öğrencilerin erkek öğrencilere kıyasla hem depresyon hem de anksiyete düzeylerinde anlamlı biçimde daha yüksek puanlar aldığı ancak buna rağmen GPA'larının daha yüksek olduğu görülmüştür. Bu bulgu kadın öğrencilerin psikolojik zorluklara rağmen akademik başarılarını koruyabildiklerini düşündürmektedir.

Depresyon ile anksiyete düzeyleri arasında çok güçlü bir pozitif ilişki saptanmıştır. Bu anksiyete düzeyi yüksek bireylerin depresyon riski taşıyabileceğini ve ruh sağlığı semptomlarının iç içe geçerek bireyin genel işlevselliğini olumsuz etkileyebileceğini göstermektedir. Bu durumun akademik performans üzerinde dolaylı yollardan etkili olabileceği değerlendirilmektedir.

Depresyon ile DEHB semptomları arasında orta düzeyde, anksiyete ile DEHB arasında ise anlamlı pozitif ilişkiler gözlenmiştir. Bu da dikkat eksikliği ve dürtüsellik gibi belirtiler taşıyan bireylerin aynı zamanda duygusal zorluklar yaşama olasılıklarının yüksek olduğunu ve bu durumun akademik süreci olumsuz etkileyebileceğini ortaya koymaktadır.

Lise not ortalaması ve üniversite giriş puanı ile GPA arasında güçlü pozitif ilişkiler bulunmuştur. Bu bireyin geçmiş akademik performansının üniversite düzeyindeki başarıyı öngörmede etkili olduğunu göstermektedir.

Yapılan regresyon analizi sonucunda yalnızca depresyon düzeyinin akademik başarı üzerinde anlamlı ve negatif bir etkisi olduğu görülmüştür. Buna karşılık anksiyete ve DEHB düzeyleri aynı modelde anlamlı çıkmamıştır. Ayrıca cinsiyet ve yaş değişkenlerinin de akademik başarı üzerinde anlamlı etkileri olduğu saptanmıştır. Kadın öğrencilerin GPA puanlarının erkek öğrencilere kıyasla istatistiksel olarak anlamlı biçimde daha yüksek olduğu görülmüştür. Yaş değişkeni açısından ise yaş arttıkça GPA'nın da anlamlı biçimde arttığı gözlemlenmiştir. Bu bulgu öğrencilerin yaşları ilerledikçe daha fazla akademik olgunluk ve sorumluluk geliştirdiklerini ve bu sayede akademik performanslarının da iyileşebileceğini göstermektedir. Sonuç olarak akademik başarıyı etkileyen çok boyutlu bir yapı söz konusudur. Bu bağlamda duygusal destek sunan yapıların öğrenci başarısını artırmada etkili olabileceği değerlendirilmektedir.

KAYNAKÇA

[1]Mohamad, M. H., Wagener, D., & McLachlan, M. H. (2024). Symptoms of ADHD and other common mental disorders influence academic success in South African undergraduates. *BMC Psychology*, 12.
[2]Hisli, N. (1989). Beck Depresyon Envanteri'nin üniversite öğrencileri için geçerliği ve güvenirliği. *Psikoloji Dergisi*, 7(23),313.
[3]Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II (BDI-II) and Beck Anxiety Inventory (BAI)*. San Antonio, TX: Psychological Corporation.
[4]Kessler, R. C. et al. (2005). Adult ADHD Self-Report Scale (ASRS). *Psychological Medicine*, *35*(2), 245–256.

2024.-2025. 8.Yarıyılı İSTATİSTİK BÖLÜMÜ

FİNANSAL PİYASALARDA RİSK ANALİZİ

Büşra ÇETİNKAYA 21023609

Danışman: Prof. Dr. Serpil Kılıç Depren

ÖZET

Bu çalışma, finansal piyasaların içsel oynaklığını ve aşırı olay riskini anlamak amacıyla, TÜPRAŞ (BIST: TUPRS) hisse senedinin 2020-2024 dönemi günlük getirileri üzerinde kapsamlı bir ampirik analiz sunmaktadır. Geleneksel risk ölçüm modellerinin sınırlarının ötesine geçilerek, getirilerdeki volatilite kümelenmesini modellemek için GARCH (Genelleştirilmiş Otoregresif Koşullu Değişen Varyans) modeli ve nadir ancak şiddetli kayıplara yol açabilen aşırı olayları analiz etmek için EVT (Aşırı Değerler Teorisi) kullanılmıştır. Yahoo Finance'ten temin edilen verilerle yapılan analizlerde, TÜPRAŞ hissesinde yüksek oynaklık ve volatilite kümelenmesi gözlemlenmiştir. GARCH ve EVT modelleri, potansiyel riskleri farklı perspektiflerden değerlendirerek kalın kuyrukların risk yönetimindeki önemini vurgulamıştır. Çalışma, elde edilen bulgularla risk yöneticilerine ve yatırımcılara daha sağlam karar alma süreçleri için değerli içgörüler sunmayı hedeflemektedir.

Risk Kavramı ve Risk Yönetimi Süreci

Risk, gelecekteki olayların belirsizliği nedeniyle ortaya çıkan ve beklenen bir sonuçtan sapma potansiyeli taşıyan durumu ifade eder. Bu sapma, genellikle olumsuz sonuçlara odaklanılır. Bu belirsizlik ortamında, olası olumsuz etkileri en aza indirmek veya fırsatları optimize etmek amacıyla risk yönetimi süreci uygulanır. Bu süreç, sistematik bir yaklaşımla dört ana aşamadan oluşur.

Risk Yönetimi Süreci

i. **Risk Tanımlama:** Potansiyel risk kaynakları belirlenir.
ii. **Risk Ölçümü/Değerlendirme:** Risklerin olasılık ve etkilerinin analizi uygulanır.
iii. **Risk Kontrolü/Azaltma:** Stratejiler belirlenir ve uygulanır.
iv. **Risk İzleme/Gözden Geçirme:** Stratejilerin etkinliği sürekli takip edilir.

Finansal piyasalar özelinde, bu süreç çok daha karmaşık ve dinamik bir hal alabilir. Piyasaların kendine özgü özellikleri, finansal riskin ölçümünü ve yönetimini özel modeller gerektiren bir disiplin haline getirmiştir



Finansal Risk ve Piyasa Analizindeki Önemi

Finansal risk, bir yatırımın beklenen getiriden daha az getiri sağlama veya sermayede kayıp yaşama olasılığıdır. Finansal piyasaların kendine özgü dinamikleri, yatırımcıları ve kurumları çeşitli risk türlerine maruz bırakır.

Başlıca Finansal Risk Türleri

- **Piyasa Riski:** Piyasa fiyatlarındaki değişimlerden kaynaklanır.
- **Kredi Riski:** Karşı tarafın yükümlülüklerini yerine getirememe ihtimalini içerir.
- **Operasyonel Risk:** Operasyonel süreçlerdeki aksaklıklardan doğar.
- **Likidite Riski:** Likidite sıkıntıları nedeniyle ortaya çıkar.

Bu çalışma, spesifik olarak, finansal varlıkların piyasa değerindeki dalgalanmalardan kaynaklanan piyasa riskine odaklanmaktadır.

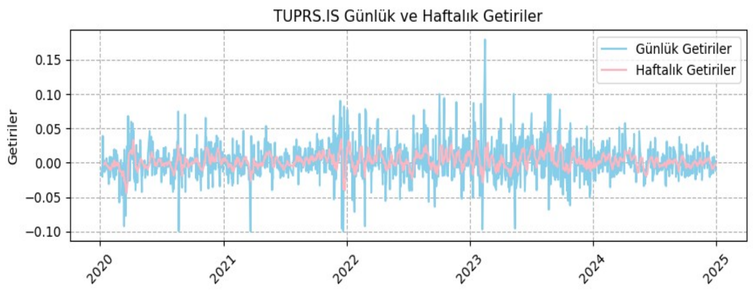
Modelleme Metodolojisi

TÜPRAŞ hisse senedi getirilerinin piyasa riskini analiz etmek amacıyla iki temel ileri düzey modelleme yaklaşımı olan GARCH ve EVT benimsenmiştir. GARCH(1,1) modeli, getirilerdeki oynaklığın zamanla değişen koşullu varyansını ve geçmiş şoklara olan bağımlılığını modelleyerek volatilite kümelenmesini etkin bir şekilde yakalamak için kullanılmıştır. Öte yandan, EVT, getiri dağılımlarının kalın kuyruklarındaki davranışı modellemek için kullanılmıştır. Bu iki modelleme yaklaşımı, TÜPRAŞ hisse senedi özelinde, piyasa riskinin hem dinamik oynaklık hem de aşırı olayların potansiyel etkisi açısından kapsamlı bir analizine olanak tanımaktadır.

Veri Yapısı ve Temini

TÜPRAŞ hisse senedinin finansal performans ve risk analizini gerçekleştirmek amacıyla, 2020-2024 dönemine ait günlük fiyat verileri kullanılmıştır. Veri Yahoo Finance'ten temin edilmiş olup, hisse senedinin açılış, kapanış, en yüksek ve en düşük fiyatları ile işlem hacmi gibi temel bilgileri içermektedir. Analizler için logaritmik günlük getiriler hesaplanmıştır.

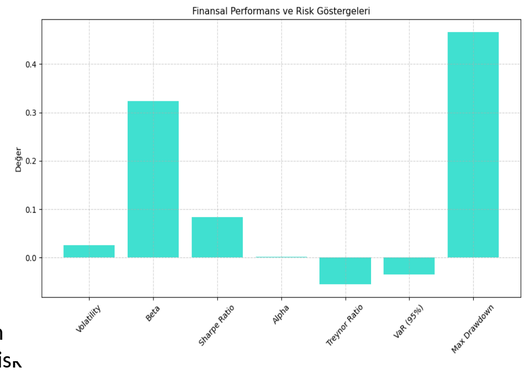
Price	Adj Close	Close	High	Low	Open	Volume
Ticker	TUPRS.IS	TUPRS.IS	TUPRS.IS	TUPRS.IS	TUPRS.IS	TUPRS.IS
Date						
2020-01-02	13.402785	18.285713	18.414286	18.114285	18.185715	10267544
2020-01-03	13.161955	17.957142	18.271427	17.671429	18.228571	13947619
2020-01-06	12.952538	17.671429	17.785713	17.614285	17.700001	9460164
2020-01-07	12.868770	17.557142	17.900000	17.542856	17.757141	14402885
2020-01-08	12.606997	17.200001	17.457142	17.142857	17.385714	17810471



Getiri Metrikleri

- **Voltalite:** 0.02615 olarak hesaplanmıştır.

TÜPRAŞ hissesinin günlük getirilerinde ortalama olarak %2.615'lik bir dalgalanma sergilediğini gösterir.



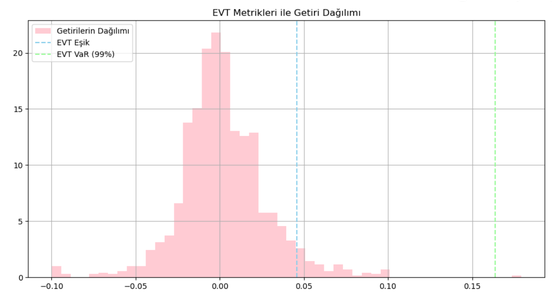
- **Beta:** 0.324 olarak bulunmuştur. Hissenin piyasa ile aynı yönde ancak piyasadan daha az agresif hareket etme eğiliminde olduğunu ve piyasa riskin karşı defansif bir yapı sergilediğini ifade eder.
- **Alfa:** 0.00200 olarak tahmin edilmiştir.

TÜPRAŞ hissesinin piyasanın getirisine ve kendi beta'sına göre ayarlandıktan sonra, günlük bazda ortalama %0.20'lik ek bir getiri sağladığını gösterir.

- **Treynor Oranı:** −0.0550 olarak hesaplanmıştır.

TÜPRAŞ hissesinin üstlendiği her bir birim sistematik risk (beta) için risksiz getirinin altında bir getiri sağladığını ima ederek sistematik risk açısından verimsizliğe işaret eder.

- **VaR (%95):** −0.0350 olarak elde edilmiştir.TÜPRAŞ hissesinde %95 olasılıkla bir sonraki işlem gününde en fazla %3.50'lik bir kayıp yaşanabileceğini öngörür.
- **Maksimum Düşüş:** 0.4662 olarak bulunmuştur. Maksimum Düşüş, TÜPRAŞ hissesinin incelenen dönemde bir zirve noktasından sonra en fazla %46.62'lik bir değer kaybı yaşadığını gösteren tarihsel bir risk ölçütüdür.



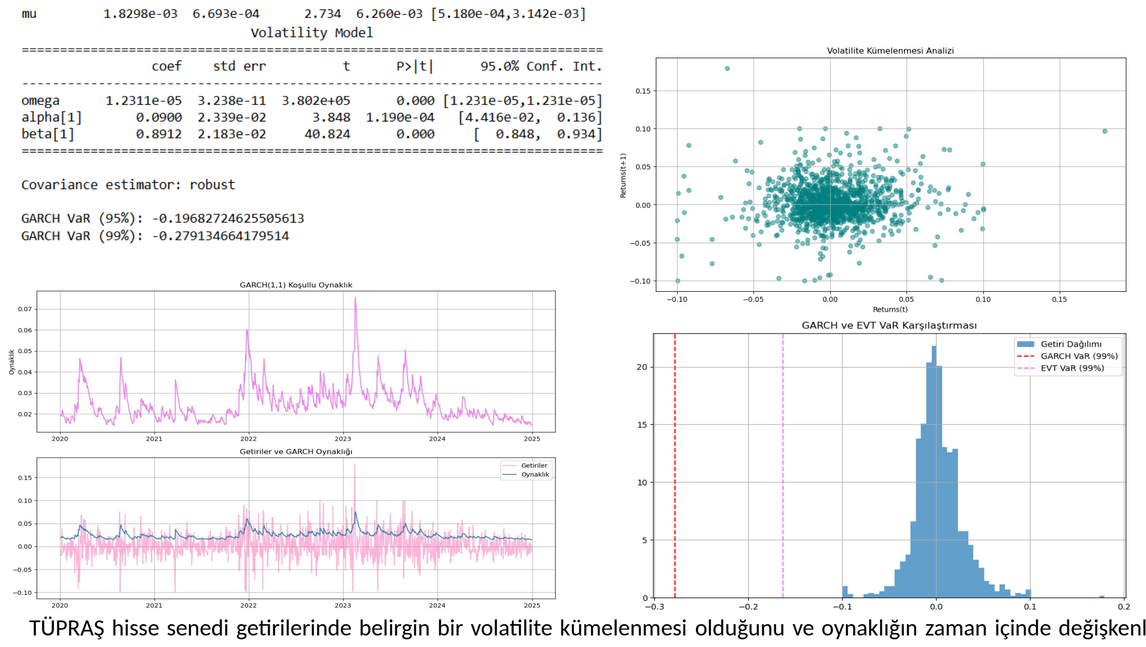
Bu histogram getirilerin merkezde yoğunlaştığını ancak kuyruklarda da gözlemlerin bulunduğunu göstermektedir. Yaklaşık %4.60'taki EVT Eşiği ve %16.35'teki EVT VaR (%99) çizgileri, EVT'nin ekstrem olaylara odaklanarak potansiyel kayıp riskini nasıl değerlendirdiğini ortaya koyar.

```
GARCH Model Results:
                  Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                Returns   R-squared:                       0.000
Mean Model:             Constant Mean   Adj. R-squared:                  0.000
Vol Model:                      GARCH   Log-Likelihood:               2872.26
Distribution:                  Normal   AIC:                          -5736.52
Method:            Maximum Likelihood   BIC:                          -5716.00
                                        No. Observations:                 1249
Date:                Sun, Jun 15 2025   Df Residuals:                     1248
Time:                        22:19:51   Df Model:                            1
                            Mean Model
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
mu         1.8298e-03  6.693e-04      2.734  6.260e-03 [5.180e-04,3.142e-03]
                          Volatility Model
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
omega      1.2311e-05  3.238e-11  3.802e+05      0.000 [1.231e-05,1.231e-05]
alpha[1]       0.0900  2.339e-02      3.848  1.190e-04   [4.416e-02,  0.136]
beta[1]        0.8912  2.183e-02     40.824      0.000     [  0.848,  0.934]
==============================================================================

Covariance estimator: robust

GARCH VaR (95%): -0.19682724625505613
GARCH VaR (99%): -0.279134664179514
```

GARCH(1,1) model sonuçları, TÜPRAŞ getirilerinin ortalamasının pozitif ve anlamlı (1.83×10 −3) olduğunu göstermiştir. Oynaklık model parametreleri (α 1 =0.0900,β 1 =0.8912) istatistiksel olarak anlamlı olup, piyasa şoklarının oynaklığı artırdığını ve oynaklığın yüksek kalıcılıkla kümelendiğini kanıtlamıştır



TÜPRAŞ hisse senedi getirilerinde belirgin bir volatilite kümelenmesi olduğunu ve oynaklığın zaman içinde değişkenlik gösterdiğini ortaya koymuştur. GARCH modeli, hissenin dinamik oynaklığını getirilerle birlikte takip ederek, piyasa hareketliliğinin kalıcılığını doğrulamıştır. Risk ölçümlerinde ise, GARCH ve EVT modellerinden elde edilen %99 VaR tahminleri, piyasa riskini ölçen iki farklı modelin TÜPRAŞ hissesi için potansiyel kaybı ne kadar farklı tahmin ettiklerini görmek amacıyla yapılmıştır. VaR değerlerinin getiri dağılımının solunda yer alması, modellerin potansiyel kayıpları doğru yerde gösterdiğinin bir işaretidir.

Sonuçlar

Bu tez, finansal piyasalardaki volatilite kümelenmesi ve kalın kuyruklar gibi belirgin piyasa davranışları nedeniyle, geleneksel yöntemlerin yetersiz kaldığı piyasa riskini analiz etmek amacıyla TÜPRAŞ hisse senedi getirileri üzerinde GARCH ve EVT modellerini uygulamıştır. Uygulama sonucunda, TÜPRAŞ hissesinde yüksek oynaklık ve güçlü volatilite kümelenmesi tespit edilmiştir. GARCH (%99 VaR: -%27.91) ve EVT (%99 VaR: -%16.35) modelleri, farklı varsayımlarla riski farklı nicelendirse de, piyasadaki ekstrem potansiyel kayıpları daha gerçekçi bir şekilde ortaya koymuştur. Elde edilen bulgular, risk yönetiminde geleneksel yöntemlere kıyasla daha güvenilir tahminler sunan bu gelişmiş yaklaşımların önemini vurgulamaktadır. Yatırımcılara ve risk yöneticilerine, piyasanın karmaşık yapısını daha iyi anlamak ve daha sağlam kararlar almak için bu tür dinamik ve ekstrem olay odaklı modelleri kullanmaları önerilmektedir.

KAYNAKÇA

[1] Alexander, C. (2008). Market risk analysis, practical financial econometrics (Vol. 2). John Wiley & Sons.
[2] Christoffersen, P. (2011). Elements of financial risk management. Academic press.
[3] Echaust, K., & Just, M. (2020). Value at risk estimation using the GARCH-EVT approach with optimal tail selection. Mathematics, 8(1), 114.

2024.-2025. Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

Otomobil Fiyatlarının Belirleyicilerinin Veri Analizi Yöntemleriyle İncelenmesi

Candaş YÜKSEK - 18023022

Danışman: Doç. Dr. Fatma Sevinç Kurnaz

ÖZET

Bu çalışmada otomobil piyasasında ikinci el araç fiyatlarını etkileyen faktörleri tespit etmektir. Çalışmada bağımlı değişken araç fiyatı, bağımsız değişkenler ise vergi, üretici, model ve üretim yılı. Her araç için şu veriler derlenmelidir: üretim yılı, kategori, deri iç mekan, yakıt tipi, motor hacmi, kilometre, silindir sayısı, vites kutusu tipi, çekiş tipi, kapı sayısı, tekerlek tipi, araç rengi ve hava yastığı sayısı. Veri analizi sürecinde, veri ön işlemenin ilk adımı eksik ve hatalı verilerin elenmesini gerektirmiştir. Bunu, keşifsel veri analizi (EDA) ve korelasyon analizi kullanılarak değişkenler arasındaki ilişkilerin incelenmesi takip etmiştir. Modelleme aşamasında, fiyat tahmin modelleri oluşturmak için çoklu doğrusal regresyon, karar ağaçları ve rastgele orman gibi makine öğrenimi yöntemleri kullanılmıştır. Çalışmanın bulguları, üretim yılı, motor hacmi, kilometre ve araç kategorisi gibi değişkenlerin araç fiyatı üzerinde istatistiksel olarak anlamlı ve önemli bir etkiye sahip olduğunu göstermektedir. Ayrıca, modellerin doğruluk oranlarını karşılaştırmak ve en etkili tahmin yöntemini belirlemek için bir analiz yapılmıştır

Python Veri Analizi

Python, veri analizi ve modelleme uygulamaları için kapsamlı ve etkili kütüphaneler sunmaktadır. Bu tezde, pandas, NumPy, matplotlib, seaborn, scikit-learn, XGBoost ve plotly dahil olmak üzere veri temizleme, görselleştirme ve tahmine dayalı modellemeyi başarıyla gerçekleştirmek için bir dizi araç kullanılmıştır.

Python, hem geleneksel istatistiksel analiz tekniklerini hem de gelişmiş makine öğrenimi algoritmalarını destekleyen bir programlama dilidir. Random Forest ve XGBoost gibi modeller Python'da başarıyla eğitilmiş ve değerlendirilmiştir

Veri

Bu çalışmada analiz edilen veriler, Kaggle adlı açık veri platformundan temin edilmiştir. Veri seti, ikinci el otomobil piyasasına ilişkin çeşitli teknik ve donanımsal özelliklerin yanı sıra fiyat bilgilerini de içermektedir. Bu çalışmada kullanılan veri seti, ikinci el otomobil piyasasına yönelik çok sayıda değişken içermektedir. Veri temizleme, ön işleme ve analiz aşamalarında Python programlama dili ve ilgili veri analizi kütüphaneleri (pandas, numpy, matplotlib, seaborn vb.) kullanılmıştır.

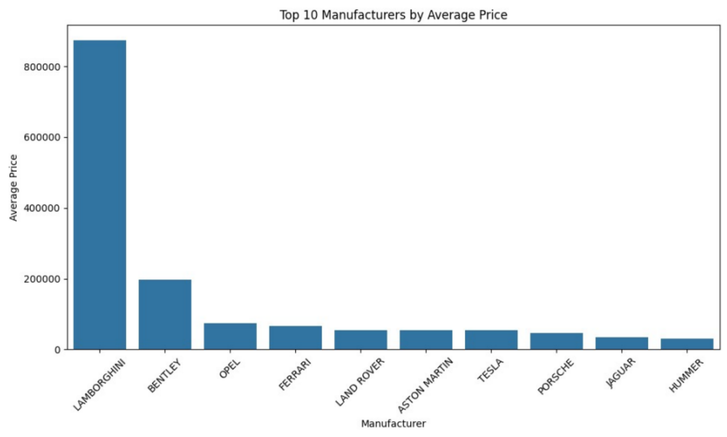
	ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color	Airbags
0	45654403	13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005 km	6.0	Automatic	4x4	04-May	Left wheel	Silver	12
1	44731507	16621	1018	CHEVROLET	Equinox	2011	Jeep	No	Petrol	3	192000 km	6.0	Tiptronic	4x4	04-May	Left wheel	Black	8
2	45774419	8467	-	HONDA	FIT	2006	Hatchback	No	Petrol	1.3	200000 km	4.0	Variator	Front	04-May	Right-hand drive	Black	2
3	45769185	3607	862	FORD	Escape	2011	Jeep	Yes	Hybrid	2.5	168966 km	4.0	Automatic	4x4	04-May	Left wheel	White	0
4	45809263	11726	446	HONDA	FIT	2014	Hatchback	Yes	Petrol	1.3	91901 km	4.0	Automatic	Front	04-May	Left wheel	Silver	4

Tezin Amacı

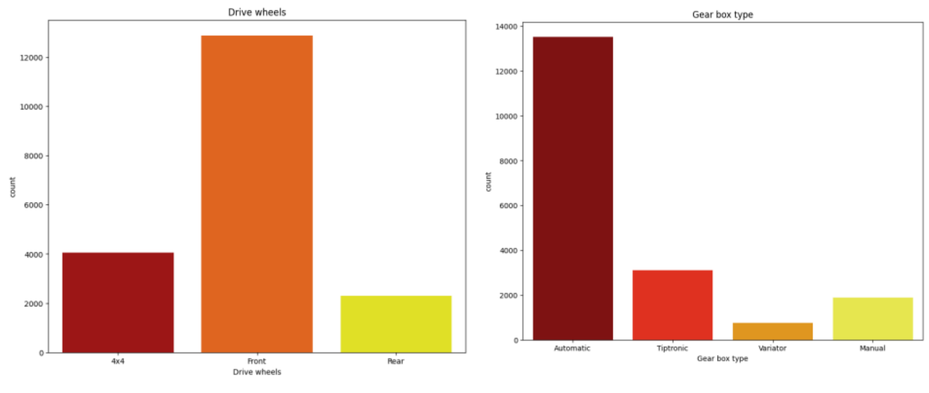
Bu tezin temel amacı, ikinci el otomobil piyasasında araç fiyatlarını etkileyen belirleyici değişkenleri niceliksel olarak analiz etmektir. Amaç, makine öğrenimi tabanlı regresyon modelleri aracılığıyla fiyat tahmin performansını optimize etmektir. Pazar dinamiklerinin karmaşık yapısı ve otomotiv araçlarının kendine has özellikleri ışığında, fiyat tahmini sürecini kolaylaştırmak için sistematik ve veri odaklı metodolojilerin kullanılması zorunludur.

Değişkenler arasındaki yapısal ilişkileri ortaya çıkarmak için keşifsel veri analizi (EDA), görselleştirme ve korelasyon analizleri takip etmiştir. Modelleme aşamasında, regresyon problemini çözmek için çoklu doğrusal regresyon, karar ağaçları ve rastgele orman algoritmaları uygulanmıştır. Model performansları doğruluk ölçütleri (MAE, RMSE, R^2) kullanılarak karşılaştırılmıştır

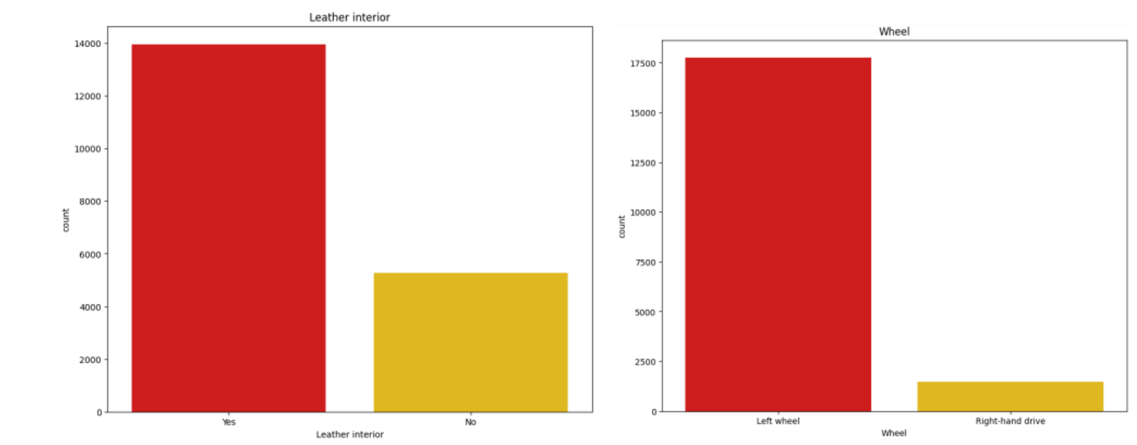
Veri Görselleştirme



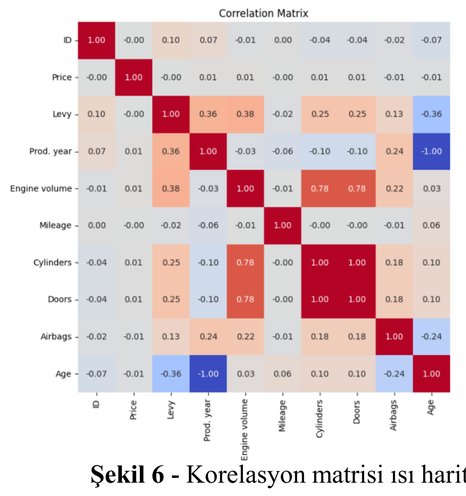
Sekil 1 - En yüksek ortalamaya sahip ilk 10 marka grafiği



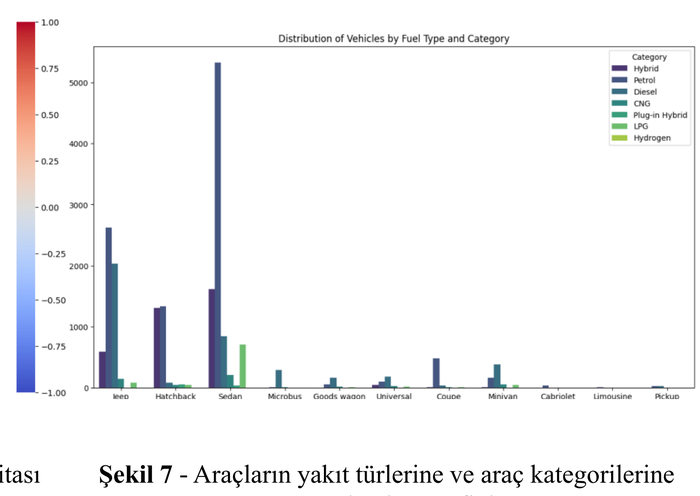
Şekil 2,3 - Çekiş sistemi ve vites kutusu tipi grafiği



Şekil 4,5 - Deri iç donanım ve direksiyon tipi (left-right) grafiği



Şekil 6 - Korelasyon matrisi ısı haritası



Şekil 7 - Araçların yakıt türlerine ve araç kategorilerine göre dağılım grafiği

Model

RandomForestRegressor: Rastgele Orman Regresörü, çok sayıda karar ağacı oluşturmak ve ardından tahminler yapmak için sonuçlarının ortalamasını almak için bir tür topluluk yöntemi olan torbalama yaklaşımını kullanarak çalışan oldukça etkili bir regresyon algoritmasıdır. Bu yapılandırmanın aşırı uyum eğilimini azalttığı ve böylece modelin genellenebilirliğini artırdığı gösterilmiştir

XGBRegressor (Extreme Gradient Boosting): XGBRegressor, yüksek doğruluk seviyelerine ulaştığı kanıtlanmış çağdaş bir regresyon algoritmasıdır. Güçlendirme yöntemi üzerine kurulmuştur ve karar ağaçlarının artımlı eğitimini kullanır. Model tarafından kullanılan güçlendirme tekniği, yeni ağaçlar oluşturarak önceki tahmin hatalarını geriye dönük olarak ele alma ve düzeltme yeteneğinde etkilidir, böylece modelin genel sağlamlığını artırır. Mevcut algoritmanın Rastgele Orman algoritmasından daha optimize ve daha hızlı olduğu gösterilmiştir.

Model Sonuçları

```
Random Forest Sonuçları:
MAE : 6610.57026488701
RMSE: 124176.9100163909
R²   : -48.48675388192922

XGBoost Sonuçları:
MAE : 5871.033203125
RMSE: 11109.20735246219
R²   : 0.6039286851882935
```

Modelin başarımını değerlendirmek amacıyla **MAE (Mean Absolute Error)**, **RMSE (Root Mean Squared Error)** ve **R^2 (Determination Coefficient)** performans metrikleri kullanılmıştır. **MAE**, tahmin edilen değerlerin gerçek değerlere olan farklarının mutlak ortalamasını alarak, modelin ortalama hata düzeyini doğrudan yansıtır. **RMSE** ise hataların karelerini alarak büyük sapmalara daha fazla ağırlık verir ve bu yönüyle uç değerlerin model performansına etkisini daha güçlü biçimde yansıtır. Son olarak, **R^2 katsayısı**, modelin bağımlı değişkenin toplam varyansını ne ölçüde açıkladığını gösterir; 1'e yaklaştıkça modelin açıklayıcılığının arttığı, 0'a yaklaştıkça ise yetersiz olduğu anlaşılır.

Random Forest modeli, yaklaşık **6610** MAE değeriyle ortalama düzeyde bir hata üretmiş olsa da, **124176** gibi oldukça yüksek bir **RMSE** değeri ve **-48** gibi negatif bir **R^2** sonucu ile başarısız bir performans sergilemiştir. Bu, modelin tahminlerinin, ortalama değeri tahmin etmekten bile daha kötü olduğunu göstermektedir.

Buna karşılık **XGBoost** modeli, yaklaşık **5871** MAE ve **11109** RMSE değeriyle daha düşük hata oranları sunmuş, **0.60** R^2 değeriyle ise veri varyansının %60'ını açıklayarak başarılı ve güvenilir bir performans ortaya koymuştur.

KAYNAKÇA

[1] McKinney, W. (2018). Python for Data Analysis (2nd ed.). O'Reilly Media.
[2] VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly
[3] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.)

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

Duygu Analizi ve Makine Öğrenmesi Yöntemleri ile Havayolu Müşterilerinin Churn Olasılığının Tahmini

DİLARA ÖZDİLEK 20023010

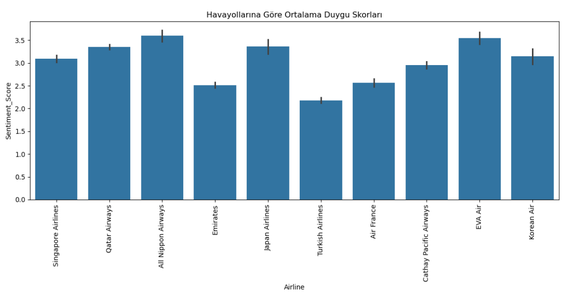
Danışman:Doç. Dr. Erhan ÇENE

ÖZET

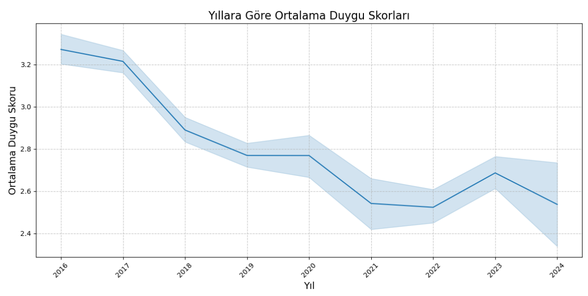
Bu çalışmada, duygu analizi ve makine öğrenmesi yöntemleri kullanılarak havayolu müşterilerinin churn (terk) etme olasılığı tahmin edilmiştir. 10 farklı havayolu şirketine ait müşteri yorumlarından duygu skorları elde edilmiş; koltuk konforu, yiyecek ve içecek, uçak içi eğlence, kabin ekibi hizmeti , paranın karşılığı, genel değerlendirme ve duygu skoru gibi bağımsız değişkenlerle, tavsiye etme değişkeni bağımlı değişken olarak modellenmiştir. Veri keşfi sürecinde üç farklı hipotez testi yapılmış (t-testi,ANOVA), korelasyon analiziyle en ilişkili değişkenler belirlenmiş, müşterilerin tavsiye etme durumlarına göre anlamlı fark olup olmadığı analiz edilmiş ve havayolu şirketlerinin ortalama duygu skorları karşılaştırılmıştır. En düşük ortalama duygu skoru 2.17 ile Türk Hava Yolları'na ait olup,en yükseği ise 3.58 ile EVA Air'e aittir. Makine öğrenmesi sürecinde XGBoost, Karar Ağacı ve Gaussian Naive Bayes algoritmaları uygulanmış; her bir algoritma üç farklı modelle denenmiş ve toplamda dokuz model oluşturulmuştur. Birinci modelde sadece duygu skoru, ikinci modelde müşteri deneyimi değişkenleri, üçüncü modelde ise hem deneyim değişkenleri hem duygu skoru birlikte bağımsız değişken olarak kullanılmıştır. Modeller karmaşıklık matrisi, F1 skoru, doğruluk, kesinlik, duyarlılık, ROC AUC, kappa skoru, MCC, log kaybı ve ortalama kesinlik skoru gibi birçok metrikle değerlendirilmiştir. Dokuz modelin karşılaştırılması sonucunda en yüksek başarı %94.85 doğruluk oranıyla tüm bağımsız değişkenlerin birlikte kullanıldığı üçüncü modeldeki Gaussian Naive Bayes algoritması ile elde edilmiş; bu modelin, müşteri memnuniyeti ve tavsiye davranışının öngörülmesinde güçlü ve dengeli bir araç olduğu sonucuna varılmıştır.

1.Problem Tanımı

Bu projenin temel amacı, havayolu şirketlerine ait müşteri yorumlarını analiz ederek müşteri memnuniyeti ile tavsiye davranışı arasındaki ilişkiyi ortaya koymaktır. BERT tabanlı duygu analizi yöntemiyle yorumlar 1 ile 5 arasında puanlanarak duygu skorları elde edilmiştir. Araştırma kapsamında, yolcuların koltuk konforu, personel hizmeti, yiyecek-içecek kalitesi, uçak içi eğlence, paranın karşılığı ve genel değerlendirme gibi deneyim parametreleri analiz edilmiş; istatistiksel testler (T-test, ANOVA) ve makine öğrenmesi algoritmaları (Karar Ağaçları, Gaussian Naive Bayes, XGBoost) ile desteklenmiştir. Proje, müşteri memnuniyetine etki eden faktörleri belirlemeyi, havayolu şirketleri arasındaki farkları ortaya koymayı ve şirketlerin hizmet kalitesini müşteri yorumları üzerinden değerlendirmeyi amaçlamaktadır. Elde edilen veri odaklı iç görüler, havayolu sektöründe müşteri deneyimini geliştirmeye yönelik stratejik kararların alınmasına katkı sağlayacaktır.



Şekil(1).Havayollarına Göre Ortalama Duygu Skorları **All Nippon Airways** ve **EVA Air**, en yüksek ortalama duygu skoruna sahipken, **Turkish Airlines** en düşük duygu skoruyla öne çıkmaktadır.



Şekil(2).Yıllara Göre Ortalama Duygu Skorları Genel eğilim düşüş yönündedir; özellikle **2018 sonrası ve pandemi yıllarında (2020–2021)** duygu skorlarında belirgin azalma görülmektedir.

2.Veri Kaynağı ve Özellikler

Bu çalışmada kullanılan veri seti, https://www.kaggle.com/datasets/elijahconnectng/airlines-review-dataset/data 'den alınmıştır. 8.100 yolcu yorumundan oluşmakta olup havayolu deneyimlerine dair hem sayısal hem metinsel değerlendirmeleri içermektedir. Toplam 17 tam dolu değişkenden oluşan veri seti, sınıflandırma problemleri için etiketli ve eksiksiz veri sunmaktadır.

Kategori	Özellik Adı – Veri Türü	Açıklama
1.Temel Bilgiler	Title - String	Yorumun başlığı
	Name - String	Yorumu yapan kişinin adı
	Review Date - Datetime	Yorumun yapıldığı tarih
	Airline - String	Havayolu şirketi adı
	Verified - Boolean	Yorumun doğrulanmış olup olmadığı (True/False)
	Reviews - String	Detaylı yorum metni
2. Yolculuk Bilgileri	Type of Traveller - String	Yolcu tipi (Solo Leisure, Family Leisure)
	Month Flown - String	Uçuşun gerçekleştiği ay
	Route - String	Uçuş rotası (kalkış-varış noktaları)
	Class - String	Uçuş sınıfı (Economy Class, Business Class vb.)
3. Değerlendirme Kriterleri	Seat Comfort - Nümerik	Koltuk konforu (1-5 arası puan)
	Staff Service - Nümerik	Personel hizmeti (1-5 arası puan)
	Food & Beverages - Nümerik	Yiyecek ve içecek hizmetleri (1-5 arası puan)
	Inflight Entertainment - Nümerik	Uçak içi eğlence sistemleri (1-5 arası puan)
	Value For Money - Nümerik	Fiyat-performans değerlendirmesi (1-5 arası puan)
	Overall Rating - Nümerik	Genel değerlendirme (1-10 arası puan)
4. Tavsiye Bilgisi	Recommended - Kategorik	Havayolunun tavsiye edilip edilmediği (yes/no)
5. Hesaplanan Özellik	Sentiment_Score - Nümerik	BERT modeli ile yorumdan elde edilen duygu skoru (1-5 arası)

Tablo(1). Veri Kaynağı ve Özellikleri Özet Tablosu

3. Duygu Analizi Süreci

•Model Seçimi ve Yükleme:
nlptown/bert-base-multilingual-uncased-sentiment modeli ve tokenizer, *Transformers* kütüphanesi kullanılarak yüklenmiştir.
•Metin İşleme ve Tokenizasyon:
Yorum metinleri 512 token sınırıyla işlenmiş, modele uygun tensör yapısına dönüştürülmüştür.
•Duygu Skoru Hesaplama:
Her yorum için 1–5 arasında duygu skoru elde edilmiştir (1: çok negatif, 5: çok pozitif).
Skorlar softmax ile normalize edilmiş, ağırlıklı ortalama alınarak final skor üretilmiştir.

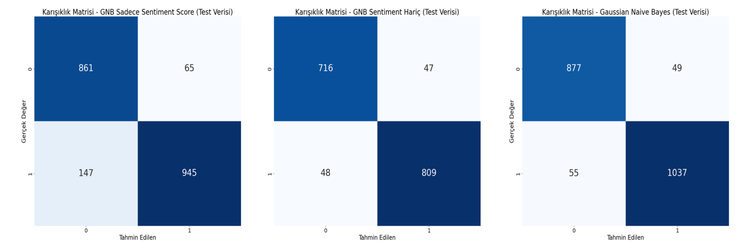
•Genel İstatistikler:
- Ortalama skor: **2.83**
- Standart sapma: **1.25**
- Min–Max skor: **1.02 – 4.98**
- Medyan: **2.62**

- **Tavsiye Durumuna Göre Farklılıklar:**
- **Tavsiye edenler (4.287 kişi):** Ortalama: **3.78** | Std sapma: **0.86**
- **Tavsiye etmeyenler (3.813 kişi):** Ortalama: **1.77** | Std sapma: **0.57**

➔ Bu süreç, müşteri yorumlarındaki duygusal tonun otomatik olarak sayısallaştırılmasını sağlayarak, memnuniyet analizinde güçlü bir temel sunmaktadır.

4. Churn Tahmin Modeli

Projede, havayolu müşterilerinin hizmet sonrası şirketi tavsiye edip etmeyeceği üzerinden churn tahmini yapılmıştır. Gaussian Naive Bayes, Karar Ağacı ve XGBoost algoritmaları ile toplam 6 model test edilmiştir. Modeller farklı değişken kombinasyonlarıyla (sadece duygu skoru, sadece hizmet puanları, tüm özellikler) değerlendirilmiş; en yüksek başarı, tüm değişkenleri kullanan **Gaussian Naive Bayes modeli** ile elde edilmiştir. Modellerin performansı **doğruluk, kesinlik, duyarlılık, F1 skoru ve ROC AUC** gibi metriklerle ölçülmüştür. Özellik önem sıralamasında **Genel Değerlendirme, Fiyat-Performans ve Duygu Skoru** öne çıkmıştır. 5 katlı çapraz doğrulama ile güvenilirlik artırılmış; sonuçlar, havayolu şirketlerinin müşteri kaybını önceden tahmin ederek stratejik önlemler almasına katkı sağlamaktadır.



Şekil(3). GNB Modeli Karmaşıklık Matrisleri

5. Model Performansları

```
Model & Features                  Accuracy  Precision  Recall   F1 Score  ROC AUC
-------------------------------------------------------------------------------
XGBoost (All Features)            0.9444    0.9451     0.9440   0.9445    0.9849
XGBoost (Without Sentiment)       0.9414    0.9451     0.9440   0.9445    0.9849
XGBoost (Only Sentiment)          0.8949    0.9356     0.8654   0.8991    0.9558
-------------------------------------------------------------------------------
Decision Tree (All Features)      0.9414    0.9451     0.9440   0.9445    0.9849
Decision Tree (Without Sentiment) 0.9400    0.9435     0.9431   0.9433    0.9846
Decision Tree (Only Sentiment)    0.8955    0.9314     0.8710   0.9002    0.9555
-------------------------------------------------------------------------------
GNB (All Features)                0.9485    0.9549     0.9496   0.9522    0.9875
GNB (Without Sentiment)           0.9414    0.9451     0.9440   0.9445    0.9849
GNB (Only Sentiment)              0.8949    0.9356     0.8654   0.8991    0.9558
===============================================================================
```

Şekil(4). Model Performansları Terminal Çıktısı

En yüksek doğruluğa (%94.85) **Gaussian Naive Bayes (GNB)** modeli tüm özelliklerle birlikte ulaşmıştır. **XGBoost** (%94.44) ve **Karar Ağacı**(%94.14) modelleri de benzer seviyelerde başarılı sonuçlar üretmiştir. Tüm modellerin **ROC AUC** skorlarının 0.98'in üzerinde olması, sınıflar arasında güçlü ayrım yapabildiklerini göstermektedir.

Duygu skoru tek başına kullanıldığında model başarısı düşmekte (%89–90), ancak diğer değerlendirme kriterleriyle birlikte kullanıldığında doğruluk artmaktadır. Bu durum, duygu analizinin destekleyici rolünü ortaya koymaktadır.
GNB modeli, tüm metriklerde (precision, recall, F1, ROC AUC) en dengeli ve yüksek performansı göstererek öne çıkmıştır. Karar ağacı modeli ise yüksek yorumlanabirliğiyle pratik uygulamalarda tercih edilebilir bir alternatiftir.
Sonuç olarak, duygu skorları ve değerlendirme parametrelerinin birlikte kullanımı, havayolu müşterilerinin tavsiye davranışını tahmin etmede güçlü sonuçlar sağlamaktadır.

6. Öne Çıkan Bulgular

- **Duygu Skoru ile Tavsiye Arasındaki İlişki:** Duygu analizine dayalı skorlar ile müşterilerin tavsiye etme davranışı arasında güçlü bir bağ bulunmuştur. Sadece duygu skoru ile %89.5 doğruluk elde edilmiş; olumlu duygu içeren yorumların büyük kısmı tavsiye ile sonuçlanmıştır.
- **Model Performans Karşılaştırması:** Gaussian Naive Bayes modeli %94.85 doğrulukla en yüksek başarıyı göstermiştir. XGBoost ve Karar Ağacı modelleri de %94+ doğruluk ve 0.98+ ROC AUC değerleriyle yüksek ayırt ediciliğe sahiptir.
- **Özelliklerin Etkisi:** En etkili değişkenler: Fiyat-Performans değerlendirmesi, Genel Değerlendirme ve Personel Hizmeti olarak öne çıkmıştır. Duygu skoru, tek başına sınırlı etkili olsa da, diğer özelliklerle birlikte modeli anlamlı şekilde güçlendirmiştir.
- **Zamana Bağlı Değişim (Trend Analizi):** Yıllara göre memnuniyet seviyelerinde değişim gözlemlenmiş; özellikle Pandemi döneminde düşüşler kaydedilmiştir. Bu havayolu hizmet kalitesine dair önemli stratejik sinyaller sunmaktadır.
- **Model Güvenilirliği ve Pratik Kullanım:** Yüksek kesinlik ve duyarlılık değerleri sayesinde modeller güvenilir ve dengelidir. GNB modeli hızlı ve verimli çalıştığı için uygulamada avantaj sağlar; Karar Ağacı modeli ise yorumlanabilirliği ile öne çıkar.
- **İşlevsellik ve Katkı:** Modeller müşteri kaybı riskini öngörmede kullanılabilir; duygu analizi + değerlendirme metriklerinin birleşimi, etkili ve bütüncül bir tahmin sistemi sunar. Çalışma, benzer araştırmalar için metodolojik örnek teşkil etmektedir.

7. Sonuç

Bu çalışma, havayolu müşterilerinin churn olasılığını tahmin etmede duygu analizi ve makine öğrenmesi yöntemlerinin birlikte kullanımının etkili olduğunu göstermiştir. %94.85'e varan doğruluk oranlarıyla, özellikle Gaussian Naive Bayes modelinin başarılı sonuçlar verdiği kanıtlanmıştır. Müşteri yorumlarından elde edilen duygu skorlarının, geleneksel değerlendirme metrikleriyle birleştirilmesi, daha kapsamlı ve güvenilir tahminler yapılmasını sağlamıştır.

KAYNAKÇA

[1] Chung, S. H., Ma, H. L., Hansen, M., & Choi, T. M. (2020). Data science and analytics in aviation. In *Transportation Research Part E: Logistics and Transportation Review* (Vol. 134). Elsevier Ltd. https://doi.org/10.1016/j.tre.2020.101837
[2] Ling, Y., Wang, Z., & Wu, C. (2021). Application in Aviation Customer Churn Prediction Using Classification Algorithm Based on Machine Learning. *Communications in Computer and Information Science*, *1424*, 254–264. https://doi.org/10.1007/978-3-030-78621-2_20/FIGURES/2
[3] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11856 LNAI*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16/TABLES/6

2024-2025 Bahar Yarıyılı İSTATİSTİK

BÖLÜMÜ

EtP DEĞERİ BULUNUMU VE PUAN ANALİZİ

Doruk BAYCAN 19023046

Danışman: Arş. Gör. Dr. Egemen Özkan

ÖZET

Bu çalışmada, 2024-2025 Türkiye Süper Ligi'nde üç büyük takımın oynadığı toplam 102 maç ve bu maçlarda tespit edilen 120 büyük hakem hatası analiz edilmiştir. Bu verilerden yola çıkılarak, hakem kararlarının maçlara olan etkisini ölçmek amacıyla Etki Puanı (EtP) adında bir gösterge geliştirilmiştir. EtP değeri pozitif veya negatif olabilmekte; Pozitif EtP, ilgili takımın lehine yapılan hataları, Negatif EtP, ilgili takımın aleyhine yapılan hataları göstermektedir. Bu EtP değerleri hesaplandıktan sonra, mevcut resmi puan tablosuna ek olarak iki alternatif puan durumu daha oluşturulmuştur: Toplam Hatalara Göre Puan Durumu , İlk Hatalara Göre Puan Durumu . Bu alternatif tablolar, hakem hatalarının lig sıralamasına olası etkisini daha net şekilde ortaya koymak amacıyla tasarlanmıştır. Kullanılan veri setlerinin nasıl ve neye göre bulundukları belirtilmiş olup bu veriler efektif bir şekilde kullanılmıştır. Bunun dışında belirtilen formüller açıklanmış, ardından da örneklerle desteklenmiştir. 1 örnek maçla beraber de EtP değerinin hangi durumlarda nasıl işlediği ortaya konmak istenmiştir. Çalışmanın sonunda ise 3 tane farklı grafik elde edilmiş olup bu grafikler üstünden, her biri için ayrı olarak yorumlandırmalar yapılmış, önemli olduğu düşünülen tespitler de yazılmıştır.

Hakem Hatalarına Göre EtP Değeri Oluşturulması ve Tahmine Dayalı Yeni Puanlar Belirlenmesi

Bu projede kullanılan veri seti, 2024–2025 Süper Lig sezonunda Beşiktaş, Fenerbahçe ve Galatasaray'ın oynadığı toplam 102 maçı kapsamaktadır. Bu maçlarda tespit edilen toplam 120 kritik hakem hatası, beIN Sports'ta yayınlanan "TRİO" programında yer alan eski hakemler Deniz Çoban, Bülent Yıldırım ve Bahattin Duran'ın değerlendirmeleri esas alınarak oluşturulmuştur. Üç hakem arasında ortak görüş oluşmayan pozisyonlarda ise kamuoyunda öne çıkan farklı uzman yorumlarına da başvurulmuştur. Her pozisyon için; maçın dakikası, pozisyonun türü ve ciddiyeti, skor durumu, hatadan etkilenen takım ve maç öncesi bahis oranları gibi bilgiler detaylı olarak kaydedilmiştir. Bu veriler üzerinden her pozisyona bir Etki Puanı (EtP) atanarak, hakem hatalarının maç sonucuna olan etkisi ölçülmüş ve bu puanlar kullanılarak alternatif puan tabloları oluşturulmuştur.

EtP Değerini Bulmak İçin Belirlenmiş Olan Formüller

Bileşen	Formül	Açıklama
Dakika Katsayısı (Gol & Penaltı)	(dk / 90)^1.8 × 0.75 + 0.25	Pozisyon maçın ilerleyen dakikalarında ise daha
Dakika Katsayısı (Kırmızı Kart)	((90 - dk) / 90)^1.5 × 0.9 + 0.05	Maçın erken dakikalarında daha yüksek etki
Güç Dengesi Katsayısı (GDK)	Güçlü takım: 1 Zayıf takım: 1 + (Güç Farkı)^0.8782 × 3	Bahis oranlarından elde edilen güç farkına göre
Etki Puanı (EtP)	Pozisyon Katsayısı × GDK × Dakika Katsayısı	Hatanın toplam etkisini hesaplar

Dakika Katsayıları

Gol ve penaltı gibi pozisyonlar maçın sonlarına doğru çok daha belirleyici hale gelir. Bu nedenle dakika ilerledikçe artan bir fonksiyon (üs formülü) tercih edildi. Kırmızı kartlar ise maçın erken bölümlerinde takım yapısını bozduğundan, dakika ilerledikçe etkisi azalan bir formül tercih edildi. Bu nedenle farklı türdeki pozisyonlara farklı üs katsayıları (örneğin 1.8 ya da 1.5) uygulanmıştır.

Güç Dengesi Katsayısı:

Her maç için iddia oranlarından galibiyet, beraberlik ve mağlubiyet yüzdeleri çıkarılıp normalleştirildi.
Bu sayede takımlar arasındaki güç farkı hesaplandı. Güçlü takımın lehine yapılan hatalar için GDK = 1 sabit alınırken, zayıf takıma yapılan hatalarda bu farkın etkisini artırmak adına özel bir güç katsayısı formülü kullanıldı.

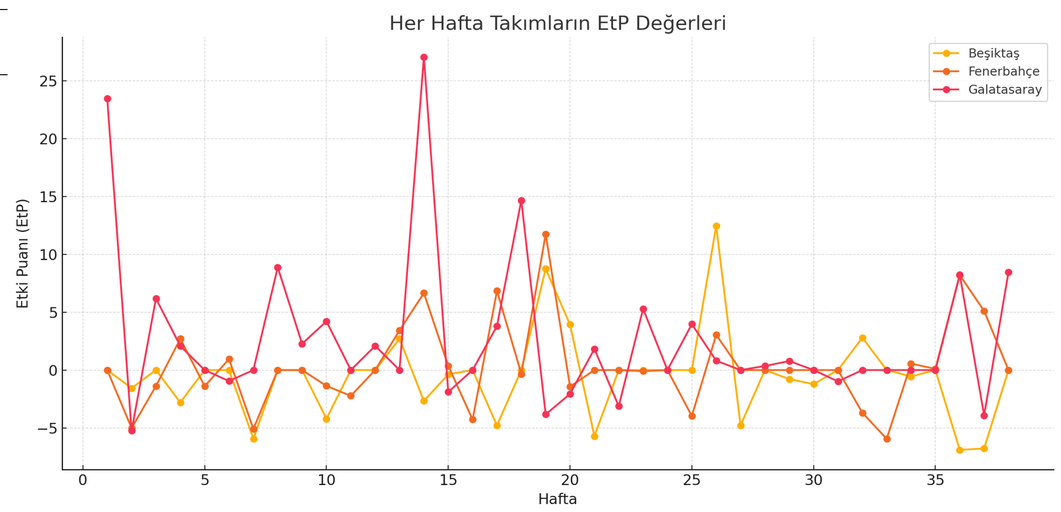
Pozisyon Önemi Katsayısı:

Her pozisyonun oyuna etkisinin eşit olmadığından yola çıkılarak, 48 farklı skor ve pozisyon türüne göre katsayılar belirlendi. Örneğin; beraberlik durumunda verilmeyen bir gol 7.00 gibi yüksek bir değer alırken, 2 farklı önde olunan bir maçtaki verilmeyen gole 1.80 katsayısı gibi daha az etkili bir katsayı verildi. Bu katsayılar, deneysel gözlem, maç senaryoları ve hakem değerlendirmeleri dikkate alınarak oluşturuldu.

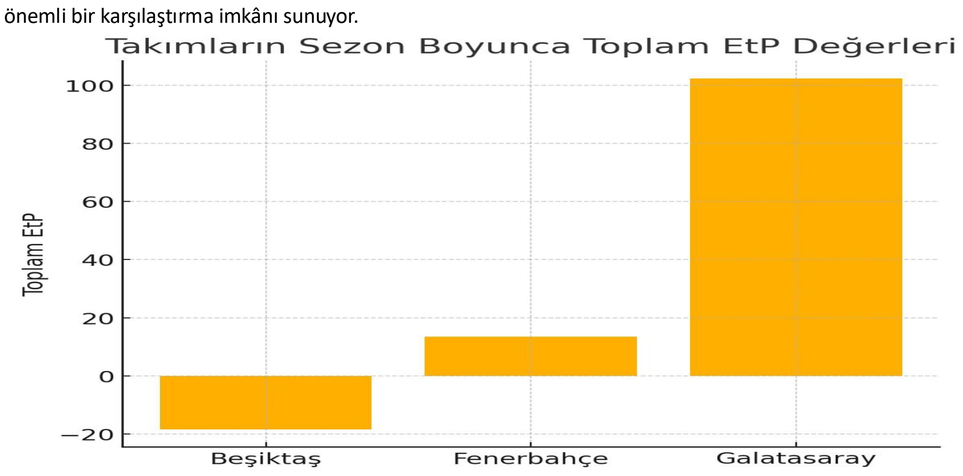
Örnek Bir Maç:

Dakika	Pozisyon	Skor	EtP Değeri
12	Eyüpspor'un faul nedeniyle iptal edilmeyen ilk golü	0 - 0	-2.02
18	Davinson Sanchez'in Thiam'a tekmesi sonucu verilmeyen kırmızı kart	0 - 1	10.8
26	Bruno'nun Barış Alper'e sarılması sonucu verilmeyen penaltı	0 - 1	-1.87
56	Davinson Sanchez'in ayağa basması sonucu ikinci sarıdan verilmeyen kırmızı	2 - 1	4.45
84	Ampem'in düşürülmesinde verilmeyen penaltı	2 - 2	15.69
GP	Gerçek Puan		1
THGP	Toplam Hatalara Göre Puan		0
İHGP	İlk Hataya Göre Puan		3

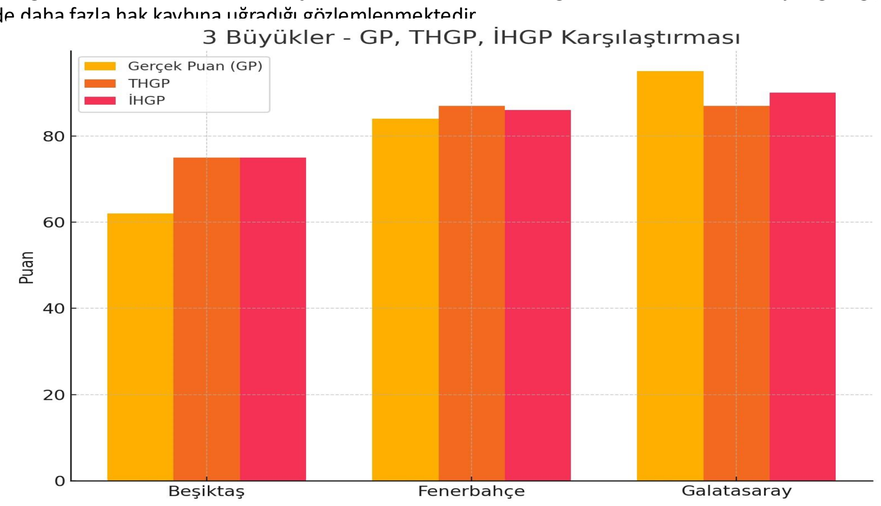
Örnek maç olarak Galatasaray'ın 14. haftada oynadığı Eyüpspor karşılaşması seçilmiştir. Bu tercihin temel nedeni, maçta tam 5 büyük hakem hatasının bulunması ve bu hataların dakikaları itibarıyla maçın gidişatına ciddi şekilde etki etmesidir. Her pozisyona ait EtP puanları hesaplandığında, Galatasaray lehine yapılan hataların toplamda ne kadar etkili olduğu açıkça görülmektedir. Ancak dikkat çeken bir diğer unsur, ilk büyük hatanın Galatasaray'ın aleyhine gerçekleşmiş olmasıdır. Bu durum Galatasaray'ın Gerçek Puanı'nı (GP) 1, Toplam Hatalara Göre Puanı'nı (THGP) 0 ve İlk Hataya Göre Puanı'nı (İHGP) 3 yapmaktadır. Bu çelişkili yapı, Eyüpspor maçını veri setindeki diğer maçlardan oldukça farklı ve istisnai bir konuma getirmektedir.



Grafikte üç büyük takımın sezon boyunca haftalık EtP dağılımları yer almakta. Galatasaray'ın özellikle sezonun ilk yarısında üst üste pozitif EtP alarak ciddi avantaj sağladığı gözlemleniyor. Fenerbahçe'nin grafiği daha dalgalı giderken, hem lehine hem de aleyhine hataların dengeli dağıldığı haftalar dikkat çekiyor. Beşiktaş ise çoğu hafta negatif EtP alarak sezon boyunca hakem hatalarından en fazla zarar gören takım konumunda. Buna ek Beşiktaş adına ilk 0'dan büyük EtP değerinin 13. haftada gelmiş olması da dikkat çeken bir olay. Bu dağılım, sadece toplam hata sayısını değil, bu hataların zamanlaması ve etkisini de öne çıkartarak önemli bir karşılaştırma imkânı sunuyor.



Bu grafikte ilk göze çarpan detay, Galatasaray'ın toplam EtP değerinin diğer iki takıma kıyasla dikkat çekici biçimde yüksek olmasıdır. Değerin 100'ün üzerine çıkması, tabloyu son derece çarpıcı hale getirmektedir. Fenerbahçe ve Beşiktaş'ın EtP seyirleri ise daha dengeli görünmekte; ancak Fenerbahçe'nin hakem kararlarından görece daha fazla avantaj sağladığı, Beşiktaş'ın ise sezon genelinde daha fazla hak kaybına uğradığı gözlemlenmektedir.



Bu tabloya baktığımızda, hakem hatalarının puan durumuna ne denli etki edebildiği açıkça görülmektedir. Beşiktaş her iki alternatif puan tablosunda da +13 puan kazanmasına rağmen, EtP değerinin çok da uç seviyelerde negatif olmaması; Güç Dengesi Katsayısı'nın varlığına bağlanabilir. Fenerbahçe ise Toplam Hataya Göre Puan tablosunda Galatasaray'ı yakalamış olsa da, İlk Hataya Göre Puan tablosunda 4 puan geride kalmıştır. Bu durum, Galatasaray'ın sezon boyunca genel olarak daha fazla hakem desteği görmesine rağmen, bazı maçlarda ilk hatanın aleyhlerine yapılmasından kaynaklanmaktadır.
Üç takım özelinde değerlendirildiğinde, hakem hatalarından en fazla olumsuz etkilenen takımın Beşiktaş olduğu her iki puan tablosunda da açıkça ortaya çıkmaktadır. Fenerbahçe ise Galatasaray'ı geçebilecek bir konumdayken, Galatasaray'ın toplamda 100'ün üzerinde EtP puanı almasına rağmen sadece -8 ve -5 puanla karşılık bulması dikkat çekicidir. Bu durum, Galatasaray lehine yapılan bazı hataların çok daha kritik anlarda gerçekleşmesiyle, yani EtP katsayılarının oldukça yüksek olmasıyla açıklanabilir.

KAYNAKÇA

[1]Appiah, D., & Amoah-Oppong, R. (2024). *The impact of refereeing decisions on match outcomes in soccer: Perceptions and realities*. Journal of Sports Sciences.
[2] Garcia, J., López, M., & Ruiz, C. (2023). *The effects of referees on final score in football: Evidence from Argentina's first division*. International Journal of Sport Policy and Politics, 15(2), 145–161.Carson, F., Lawson, M., & Brown, R. (2023). *Decision-making accuracy of soccer referees in relation to physiological load*. European Journal of Sport Science, 23(4), 502–510.
[3] Carson, F., Lawson, M., & Brown, R. (2023). *Decision-making accuracy of soccer referees in relation to physiological load*. European Journal of Sport Science, 23(4), 502–510.

2024-2025 Bahar Yarıyılı
İSTATİSTİK BÖLÜMÜ

Panel Veri Tabanlı Makine Öğrenmesi Yaklaşımı ile Çatışma Riski Tahmini: Ekonomik ve Siyasal Belirleyiciler Üzerine Bir Uygulama

Ecenur KIRAY 20023069
Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

ÖZET

Bu çalışma, 1989–2020 yılları arasında 20 ülkeye ait panel veriler kullanılarak, makine öğrenmesi algoritmaları ile çatışma riski tahmini yapmayı amaçlamaktadır. UCDP çatışma verisi, Polity IV rejim tipi, COW-NMC askeri kapasite göstergeleri ve IMF-Dünya Bankası ekonomik verileri birleştirilmiştir. Veri ön işleme, eksik değer giderimi, normalizasyon ve SMOTE ile sınıf dengesi sağlandıktan sonra; Naive Bayes, Lojistik Regresyon, Random Forest ve XGBoost modelleri uygulanmıştır. Random Forest ve XGBoost modelleri yüksek doğruluk oranı ile öne çıkmıştır. Bu çalışma, karar vericilere erken uyarı sistemleri için katkı sunmayı hedeflemektedir.

VERİ

Bu çalışmada, 1989–2020 yılları arasında 20 ülkeye ait panel veriler kullanılmıştır. Veri setleri UCDP (çatışma bilgisi), Polity IV (rejim tipi), COW-NMC (askeri kapasite), IMF ve Dünya Bankası (ekonomik göstergeler) kaynaklıdır.

Kaynak	Açıklama
UCDP	Devletlerarası çatışma verisi
Polity IV	Rejim tipi ve rejim değişimi verisi
COW-NMC	Askeri kapasite: `milper`, `milex`, `cinc`
IMF/WB	Enflasyon ve askeri harcama/GSYH oranı

Veri ön işleme aşamalarında:

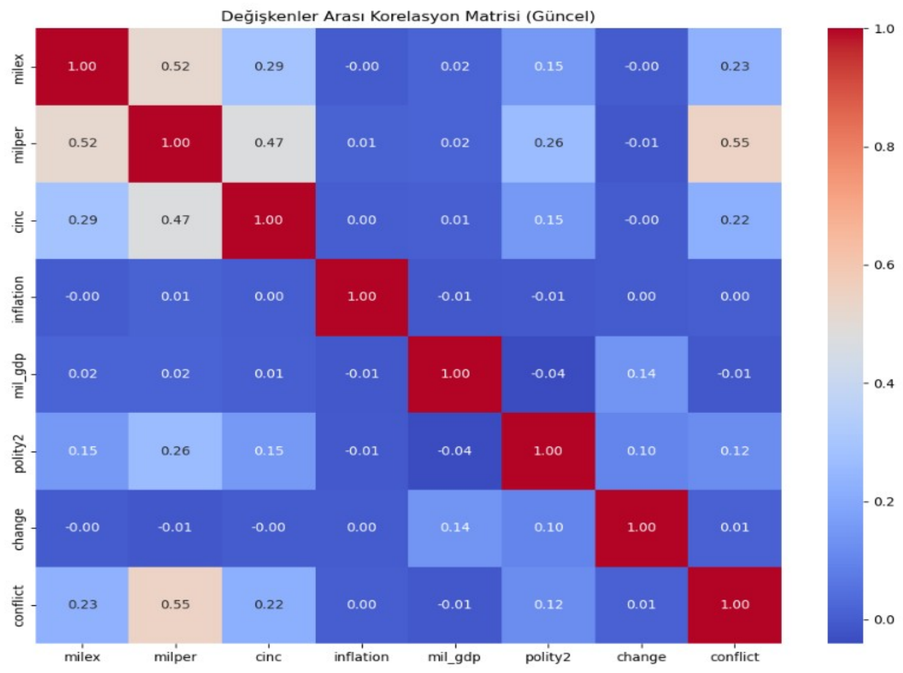
- Eksik değerler sütun ortalamalarıyla doldurulmuş, Değişkenler StandardScaler yöntemiyle normalize edilmiş, Sınıf dengesizliği SMOTE yöntemiyle giderilmiştir.
- Bağımlı değişken olarak ülke-yıl bazında çatışma varlığı (0: yok, 1: var) kullanılmıştır.
- Bağımsız değişkenler:milex, milper, inflation, mil_gdp, cinc, polity2, change

	country_code	year	milex	milper	cinc	inflation	mil_gdp	polity2	change	conflict
0	USA	1950.0	14559000.0	1460.0	0.284443	0.000059	0.000382	9.0	1.579906	0
1	USA	1951.0	33398000.0	3249.0	0.319499	0.000059	0.000382	9.0	1.579906	0
2	USA	1952.0	47852000.0	3636.0	0.311367	0.000059	0.000382	9.0	1.579906	0
3	USA	1953.0	49621008.0	3555.0	0.311158	0.000059	0.000382	9.0	1.579906	0
4	USA	1954.0	42786000.0	3302.0	0.280940	0.000059	0.000382	9.0	1.579906	0

- Veri setinde sınıflar arasında ciddi bir dengesizlik gözlenmiştir.
- Ülkelerin yaklaşık %80'inde çatışma yaşanırken, %20'sinde yaşanmamıştır. Bu durum, modelleme sürecinde **SMOTE** yöntemiyle dengelenmiştir.

Sınıf Durumu	Oran (%)
Çatışma Var (1)	%80.1
Çatışma Yok (0)	%19.9

Değişkenler Arası Korelasyon Matrisi



- milex ile milper: 0.52 → Daha fazla bütçe, daha fazla askerle birlikte.
- milper ile cinc: 0.47 → Asker sayısı arttıkça ülke gücü de artıyor.
- cinc ile milex: 0.29 → Maddi güç ve askeri harcama ilişkili.

Korelasyon analizine göre, çatışma riski en çok askeri personel (milper), askeri harcama (milex) ve ulusal güç (cinc) ile pozitif yönde ilişkilidir. Buna karşın ekonomik göstergeler (örneğin inflation, mil_gdp) doğrudan ilişki göstermemektedir. Siyasal rejim tipi (polity2) ile ilişki zayıftır.

YÖNTEM

- Modelleme sürecinde dört sınıflandırma algoritması kullanılmıştır:
  Naive Bayes, **Lojistik Regresyon**, **Random Forest** ve **XGBoost**.
- Analizler Python programlama dili kullanılarak Jupyter Notebook ortamında gerçekleştirilmiştir. Değerlendirme için **Accuracy**, **F1-Score** ve **AUC** gibi performans metrikleri kullanılmıştır.

SMOTE uygulanmadan önceki durum:

- Veri seti dengesizdir; yani bazı sınıflar çok fazla örneğe sahipken, bazı sınıflar oldukça azdır.
- Tüm modellerin hatırlama (recall) ve F1-skorları oldukça yüksek ve birbirine eşit çıkmış.
- Bu, modellerin ağırlıklı olarak çoğunluk sınıfa odaklanarak azınlık sınıfı doğru tahmin etme konusunda fazla çaba harcamadığını gösteriyor.

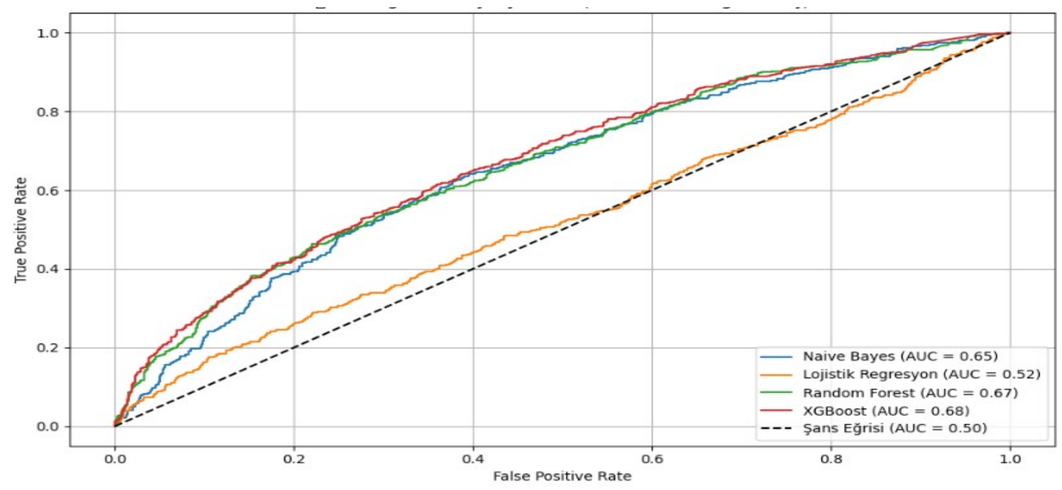
Model	Precision	Recall	F1-Score
Naive Bayes	0.794	1.0	0.885173
Lojistik Regresyon	0.794	1.0	0.885173
Random Forest	0.794	1.0	0.885173
XGBoost	0.794	1.0	0.885173

SMOTE uygulandıktan sonra

- SMOTE ile veri dengelenmiş, yani azınlık sınıftan da yeterli miktarda örnek oluşturularak sınıflar arasındaki fark azaltılmış.
- Dengeli veri setinde modellerin hatırlama oranları düşmüş ve artık birbirinden farklı performans göstermeye başlamışlar.
- Bazı modeller yeni veri dağılımına uyum sağlayarak daha iyi sonuçlar verirken, bazıları performans kaybı yaşamış.

Model	Precision (Çatışma)	Recall (Çatışma)	F1-Score (Çatışma)
Naive Bayes	0.636499	0.540302	0.584469
Lojistik Regresyon	0.515072	0.494962	0.504817
Random Forest	0.641994	0.535264	0.583791
XGBoost	0.637131	0.570529	0.601993

Özetle, ilk grafikte model sınıflar arasındaki dengesizliğe uyum sağlamış gibi görünüyor ve bu yüzden performans çok yüksek çıkıyor. Ancak ikinci grafik, dengeli veri setinin gerçekte daha zorlu bir sınıflandırma problemi sunduğunu ve model performanslarında farklılaşmalara neden olduğunu gösteriyor.



Temel Çıkarımlar

- XGBoost (AUC = 0.68) en iyi performansı gösteriyor; bu model sınıflar arasında en iyi ayrımı yapabiliyor.
- Random Forest (AUC = 0.67) rekabetçi bir performans gösteriyor ve oldukça başarılı.
- Naive Bayes (AUC = 0.65) fena değil ama önceki iki modele kıyasla bir miktar daha düşük kalıyor.
- Lojistik Regresyon (AUC = 0.52) zayıf bir performans sergiliyor, yani sınıfları iyi ayıramıyor.

SONUÇ

SMOTE uygulaması, sınıf dengesizliğini azaltarak modellerin azınlık sınıfı daha iyi öğrenmesini sağlamıştır. SMOTE sonrası bazı modellerde performans düşüşü gözlenmiş, bu da dengesiz veri setlerinin yanıltıcı biçimde yüksek doğruluk verebildiğini göstermiştir.

Sonuç olarak, XGBoost ve Random Forest modelleri, çatışma riski tahmininde karar vericiler için **erken uyarı sistemleri** geliştirme potansiyeline sahiptir.

KAYNAKÇA

[1] Cederman, L. E., Gleditsch, K. S., & Buhaug, H. (2013). *Inequality, grievances, and civil war*. Cambridge University Press.
[2] Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, 51(1), 5-18.
[3] Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford Economic Papers*, 56(4), 563-595

İSTATİSTİK BÖLÜMÜ

EVCİL HAYVANLARIN SAHİPLENİLME DURUMLARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE TAHMİNİ

Elif Ezgi ERSOY 22023604

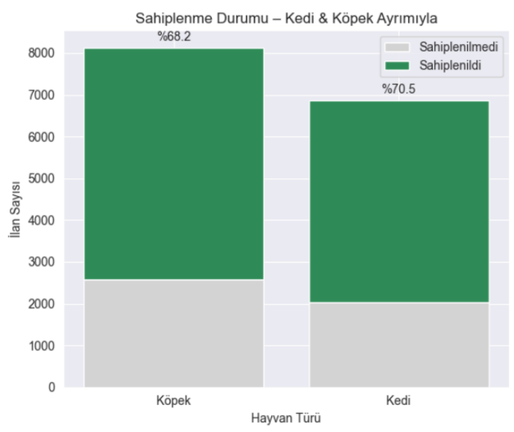
Danışman: Prof. Dr. Serpil KILIÇ DEPREN

ÖZET

Bu çalışma, bir çevrimiçi evcil hayvan sahiplendirme platformunun veri seti kullanılarak gerçekleştirilmiştir. Çalışmanın temel amacı, bir hayvanın sahiplendirilip sahiplendirilmeyeceğini etkileyen faktörleri belirlemek ve bu doğrultuda bir tahmin modeli geliştirmektir.
Veri setinde hayvanların türü, cinsiyeti, yaşı, kısırlaştırılma durumu, rengi, açıklama uzunluğu ve kurtarıcı bilgileri gibi çeşitli demografik ve davranışsal özellikler yer almaktadır. Bu bilgilerden yola çıkarak değişken mühendisliği adımları uygulanmış, görselleştirme ve ön işleme süreçleri gerçekleştirilmiştir.
Makine öğrenmesi algoritmalarıyla sınıflandırma modelleri oluşturulmuş ve elde edilen performans ölçütleri karşılaştırılmıştır. Çalışma, evcil hayvanların sahiplendirilmesini etkileyen dinamikleri ortaya koymayı ve bu konuda veri temelli öngörüler sunmayı amaçlamaktadır.

VERİ

Bu çalışmada, Malezya merkezli evcil hayvan sahiplendirme platformu PetFinder.my'a ait, Kaggle üzerinden paylaşılan ve yaklaşık 15.000 evcil hayvan ilanından oluşan açık veri seti kullanılmıştır. Her bir gözlem bir hayvan ilanını temsil etmektedir ve sahiplendirme süreciyle ilgili geniş kapsamlı bilgiler içermektedir.
Veri setinde hayvanlara ait cins, yaş, cinsiyet, kısırlaştırma durumu, sağlık durumu, açıklama uzunluğu, renk sayısı, fotoğraf sayısı, barınak aktivitesi gibi birçok açıklayıcı değişken bulunmaktadır.



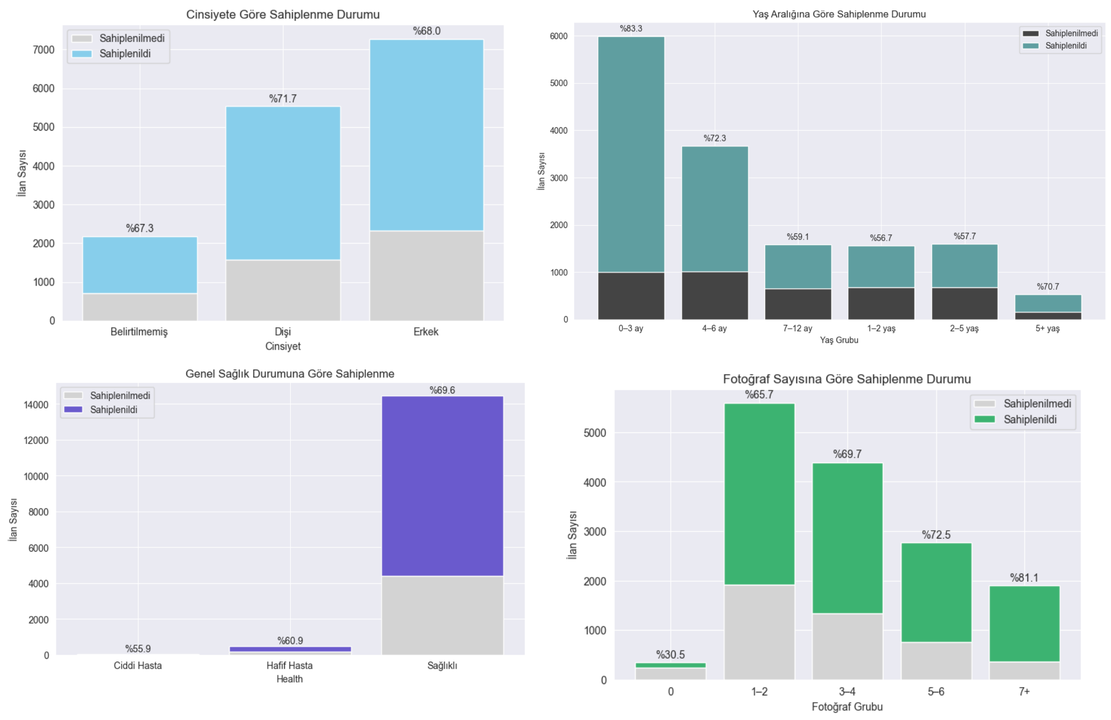
Veri setindeki ilanların 8132'si köpek, 6861'i ise kedidir. Sahiplenilme oranları incelendiğinde, köpek ilanlarının %68.2'si, kedi ilanlarının ise %70.5'i sahiplenilmiştir.

Değişken	Açıklama
Tür	Hayvanın türü (kedi / köpek)
Yaş	Hayvanın yaşı (ay cinsinden)
Cinsiyet	Hayvanın cinsiyeti (erkek / dişi)
Renkler	Hayvanın sahip olduğu 1–3 renk bilgisi
Kısırlaştırma	Hayvanın kısırlaştırılmış olup olmadığı (evet / hayır / emin değil)
Aşı / Parazit	Aşılanma ve iç-dış parazit tedavisi durumu
Sağlık Durumu	Hayvanın genel sağlık durumu (sağlıklı / hafif yaralı / ağır yaralı)
Açıklama Uzunluğu	İlan açıklamasının karakter uzunluğu (türetilmiş değişken)
Fotoğraf Sayısı	İlanda yer alan fotoğraf sayısı
Ücret	Sahiplenme için istenen ücret (0 = ücretsiz)
Kurtarıcı ID	İlanı açan kurtarıcının anonim kimlik numarası
Sahiplendirme (AdoptionSpeed)	Sahiplenme süresi

VERİ ÖN İŞLEME

Veri setindeki AdoptionSpeed değişkeni, sahiplendirme hızını 0–4 arasında sayısal olarak ifade etmektedir (0: aynı gün, 1: 1–7 gün, 2: 8–30 gün, 3: 31–90 gün, 4: 90-100 gün içinde sahiplenilmedi). Bu çalışmada 4 değeri "sahiplenilmedi", diğerleri "sahiplenildi" olarak gruplanmış ve ikili hedef değişken (Adopted) oluşturulmuştur. Veri setinde Breed, Color ve State gibi kimlik değerleri, karşılık gelen isim etiketlerine dönüştürülmüştür. Description değişkeni dışında veri setinde dikkate değer eksik gözlem bulunmadığı için doğrudan modelleme sürecine geçilmiştir. Model performansını artırmak amacıyla açıklama uzunluğu, video/fotoğraf varlığı, renk sayısı, yaş durumu ve kurtarıcı ilan sayısına dayalı çeşitli türetilmiş değişkenler oluşturulmuştur. Ayrıca modelleme öncesinde tüm kategorik değişkenler LabelEncoder kullanılarak sayısal forma dönüştürülmüştür.

KEŞİFSEL VERİ ANALİZİ



Dişi hayvanların sahiplenilme oranı %71.7 ile erkeklerden (%68.0) biraz daha yüksektir.

0–3 aylık hayvanlar %83.3 oranında sahiplenilirken, yaş arttıkça bu oran düşmektedir. Bu durum, ilanlarda genç hayvanların daha çok tercih edildiğini göstermektedir.

Kısırlaştırılmamış hayvanların sahiplenilme oranı %75.9 ile en yüksek seviyededir. Ancak bu durum doğrudan bir tercih göstergesi olmayabilir. Genç hayvanların genellikle henüz kısırlaştırılmamış olması, yaşla bağlantılı olarak bu oranın yükselmesine neden olmuş olabilir.

İlanlarda kullanılan görsel sayısı sahiplenme oranlarını belirgin şekilde etkilemektedir. Hiç fotoğrafı olmayan ilanlarda sahiplenilme oranı %30.5 iken, 7 veya daha fazla fotoğraf bulunan ilanlarda bu oran %81.1'e kadar çıkmaktadır.

MODELLEME SÜRECİ

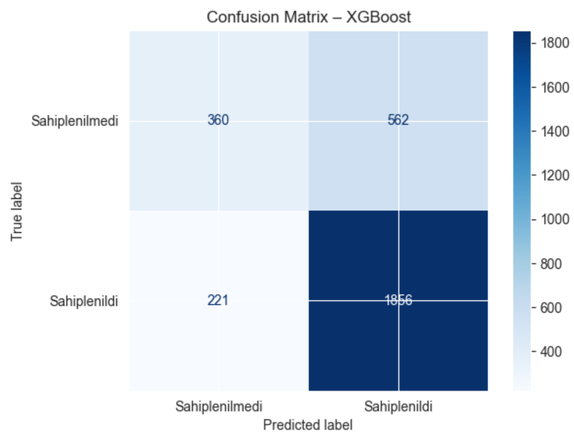
Bu çalışmada sahiplenme durumunu tahmin etmek amacıyla XGBoost algoritması kullanılmıştır. Modelin hiperparametreleri GridSearchCV yöntemi ile optimize edilmiştir. Eğitim ve test verisi %80–%20 oranında ayrılmış, modelin başarısı F1 skoru, doğruluk (accuracy) ve ROC AUC gibi ölçütlerle değerlendirilmiştir. Veri setindeki dengesiz dağılım dikkate alınarak sınıf ağırlıkları da hesaba katılmıştır.

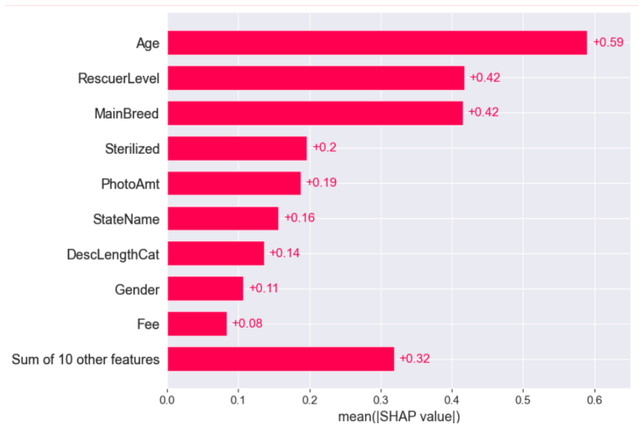
Model eğitiminde toplam 20 değişken kullanılmıştır. Eğitim ve test verisi ayrımında sınıf dengesizliği göz önüne alınarak XGBoost sınıf ağırlığı parametresi etkinleştirilmiştir. Yanda modelin en iyi hiperparametre kombinasyonu verilmiştir:

Parametre	Değer
max_depth	5
learning_rate	0.1
n_estimators	100

Model Performansı

XGBoost modeli test verisinde başarılı sonuçlar üretmiştir. F1 skoru 0.74, doğruluk (accuracy) 0.78 ve ROC AUC değeri 0.80 olarak hesaplanmıştır. Confusion matrix sonuçları modelin sahiplenilen ve sahiplenilmeyen sınıfları dengeli şekilde ayırabildiğini göstermektedir.







Modelin kararlarını daha iyi yorumlayabilmek amacıyla SHAP (SHapley Additive exPlanations) analizi uygulanmıştır. Analiz sonucuna göre modelin sahiplenme tahminlerinde en etkili değişkenin hayvanın yaşı olduğu görülmektedir. Bunun yanı sıra kurtarıcı ilan geçmişi (RescuerLevel), birincil ırk (MainBreed), kısırlaştırma durumu ve fotoğraf sayısı (PhotoAmt) da önemli katkı sağlamaktadır. Açıklama uzunluğu (DescLengthCat) gibi metin tabanlı öznitelikler de modelin kararlarında belirleyici rol oynamıştır.

SONUÇ VE ÖNERİLER

Bu çalışmada evcil hayvan sahiplendirme sürecini etkileyen faktörler analiz edilmiş ve sahiplenilip sahiplenilmeme durumunu tahmin etmek için XGBoost algoritması uygulanmıştır. Model, özellikle sahiplenilen hayvanları doğru tahmin etmede başarılı sonuçlar üretmiştir. Elde edilen sonuçlara göre yaş, açıklama uzunluğu, fotoğraf sayısı, ücret bilgisi ve kurtarıcı profili gibi değişkenler sahiplenme kararında önemli rol oynamaktadır.
Model çıktılarının SHAP analizi ile yorumlanabilir hale getirilmesi, sistemin kullanıcı davranışlarını daha iyi anlamasını sağlamıştır. Bu tür tahmin modelleri, barınakların ve platformların sahiplendirme süreçlerini iyileştirmelerinde destekleyici bir araç olarak kullanılabilir.

KAYNAKÇA

[1] https://www.kaggle.com/competitions/petfinder-adoption-prediction
[2] https://xgboost.readthedocs.io/
[3] https://shap.readthedocs.io/en/latest/

İSTATİSTİK BÖLÜMÜ

LLM Tabanlı PDF Chatbot'larının Karşılaştırmalı Performans Analizi:

LLaMA 3, Gemma 3 ve DeepSeek R1 Örneği

ELİF GÜRDEMİR 21023039

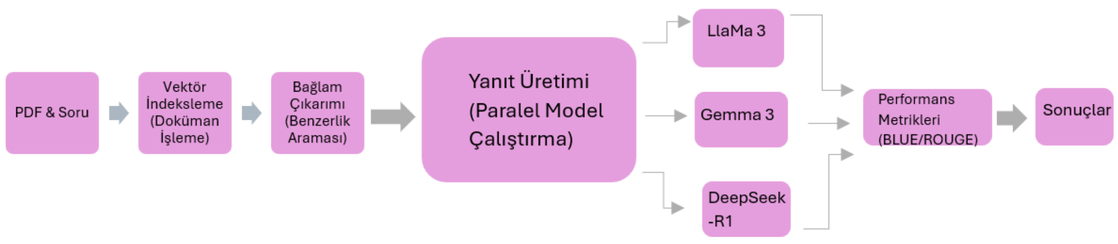
Danışman: Doç. Dr. Erhan ÇENE

ÖZET

Bu bitirme çalışması, büyük dil modelleri (LLM) kullanılarak geliştirilen PDF tabanlı soru-cevap (QA) chatbot sistemlerinin performans karşılaştırmasını ele almaktadır. Retrieval-Augmented Generation (RAG) mimarisi kullanılarak, LLaMA 3, Gemma 3 ve DeepSeek R1 modelleriyle oluşturulan yanıtlar, sabit bir PDF veri seti ve sabit sorular üzerinden alınmış ve BLEU ile ROUGE metrikleriyle değerlendirilmiştir. Amaç; farklı LLM'lerin belge odaklı soru-cevap süreçlerindeki doğruluk, tutarlılık ve performanslarını analiz etmektir. Genel olarak Gemma3 modelinin daha iyi performans gösterdiği gözlemlenmiştir.

LLM TABANLI BELGE SORGULAMA

Geleneksel bilgiye erişim yöntemleri, özellikle PDF gibi uzun ve karmaşık belgeler söz konusu olduğunda çoğu zaman yetersiz kalmaktadır. Bu noktada devreye giren Büyük Dil Modelleri (LLM), kullanıcıların bu belgeler üzerinden doğal dilde sorular sormasını ve içerikten anlamlı yanıtlar almasını mümkün kılmaktadır. RAG (Retrieval-Augmented Generation) mimarisi, LLM'leri belge içeriğine dayalı olarak daha bağlamsal ve tutarlı yanıtlar üretebilecek hale getirir. Bu proje kapsamında geliştirilen sistem, kullanıcıların PDF yükleyerek doğrudan içerikle etkileşime geçebileceği bir chatbot arayüzü sunar. Kullanıcılar sorularını girdikçe sistem belgeleri analiz eder, ilgili bölümleri çıkarır ve seçili LLM modeliyle yanıtlar üretir. Elde edilen çıktılar hem kullanılabilirlik hem de doğruluk açısından metriklerle değerlendirilmiştir, PDF dosyaları üzerinden soru-cevap yapan bir chatbot uygulaması geliştirilmiş ve üç farklı açık kaynak LLM (LLaMA 3, Gemma 3, DeepSeek R1) karşılaştırılmıştır.



Şekil 1. PDF tabanlı chatbot sisteminin genel mimarisi

KULLANILAN YÖNTEMLER

Bu çalışma Python programlama dili ile yapılmıştır. PDF tabanlı soru-cevap sistemi oluşturulurken PDF belgelerinden içerik çıkartma, belge içeriğini bölme, vektörleştirme, model üzerinden yanıt üretme ve metrik değerlendirme işlemleri için Streamlit, LangChain, Ollama, PyMuPDF, BLEU, ROUGE yöntem ve araçlarından yararlanılmıştır.

```
import streamlit as st
import os
import numpy as np

from langchain_community.document_loaders import PyMuPDFLoader
from langchain_text_splitters import RecursiveCharacterTextSplitter
from langchain_ollama import OllamaEmbeddings, OllamaLLM
from langchain_community.vectorstores import InMemoryVectorStore
from langchain.prompts import PromptTemplate
```

VERİ HAZIRLIK VE BELGE İŞLEME AŞAMALARI

Yüklenen PDF belgesi, PyMuPDF kullanılarak metne dönüştürülmüştür. Ardından bu metin, RecursiveCharacterTextSplitter ile küçük parçalara ayrılmış ve vektör temsili oluşturulmuştur. Tüm parçalar, Ollama ile embedding işlemi sonrası vektör veritabanına (InMemoryVectorStore) aktarılmıştır.

Kullanılan Modeller:

LLaMA 3 (Meta): Düşük kaynakla yüksek performans hedeflenmiş, çok sayıda veriyle eğitilmiştir.
Gemma 3 (Google): Multilingual destekli, verimlilik odaklı bir modeldir.
DeepSeek R1: Kod ve belge anlama konusunda optimize edilmiştir.

MODEL ENTEGRASYONU VE SORU-CEVAP



Şekil 2. Uygulama önyüzü

Chatbot sistemi, kullanıcıdan gelen soruya karşılık en uygun belge parçalarını bulmakta ve bu parçaları seçilen LLM modeline (LLaMA 3, Gemma 3, DeepSeek R1) bağlam olarak vermektedir. Sistem, sorulara verilen yanıtları kullanıcıya sunmakta ve aynı zamanda referans yanıtlarla karşılaştırmak üzere kaydetmektedir.

METRİKLERLE PERFORMANS DEĞERLENDİRME

Çıktıların kalitesini değerlendirmek amacıyla şu metrikler kullanılmıştır:
BLEU (Bilingual Evaluation Understudy): Yanıtın referansla kelime düzeyinde ne kadar örtüştüğünü ölçer.
ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Özellikle ROUGE-1 ve ROUGE-L metrikleri kullanılarak içerik benzerliği değerlendirilmiştir.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

p_n: n-gram doğruluk oranı (n-gram precision)
w_n: her n-gram için ağırlık (genelde eşit alınır: $\frac{1}{N}$)
BP: brevity penalty (kısa cümlelere ceza uygular)
N: n-gram seviyesi (genelde 4)

$$\text{ROUGE-N}_{F1} = \frac{2 \cdot P \cdot R}{P + R}$$

P (Precision) = $\frac{\text{Ortak n-gram sayısı}}{\text{Model çıktısındaki toplam n-gram}}$
R (Recall) = $\frac{\text{Ortak n-gram sayısı}}{\text{Referanstaki toplam n-gram}}$

ROUGE-1 → Unigram
ROUGE-2 → Bigram
ROUGE-L → En uzun ortak alt dizi (LCS)

Şekil 3. BLUE ve ROUGE metriklerinin hesaplanma formülleri

MODEL KARŞILAŞTIRMASI

Yıldız Teknik Üniversitesi'nin resmi internet sitesinde yer alan 'YTÜ Öğrenci Rehberi' PDF belgesi yüklenip aşağıda yer alan üç örnek soru kullanılarak modeller karşılaştırılmıştır:

SORULAR
OBS'ye giriş adımları ve şifre oluşturma bilgilerini çıkar, Türkçe olarak.
YTÜ öğrenci kulüpleri ve 'Kulüpler Vadisi' hakkında bilgi ver, Türkçe olarak
Belgenin ana bölümlerini ve temel içerik özetini çıkar, Türkçe olarak.

1.Soru için modellerin verdiği cevaplar:

LLaMa 3

(**llama3**): OBS'ye giriş adımları ve şifre oluşturma bilgilerini çıkarabiliriz:

- Geçici şifreniz T.C Kimlik Kartınızın seri numarasıdır.
- Şifreyi Giriş Ekranında "Şifremi unuttum" seçeneğine tıklayıp, istenen bilgileri girerek oluşturabilirsiniz.
- Giriş yaptığınızda otomatik olarak şifre değiştirmeniz istenecektir. Daha sonra şifre değiştirmek için Kullanıcı İşlemleri menüsünden "Şifre Değiştir" seçeneğine tıklayınız ve yeni oluşturmak istediğiniz şifreyi giriniz.

Türkçe olarak özetleme:

OBS'ye giriş adımları şu şekilde çalışır: Geçici şifreniz T.C Kimlik Kartserial numarasınızdır. Bu seri numaranın yanı sıra, "Şifremi unuttum" seçeneğini tıklayıp istenen bilgileri girerek oluşturabilirsiniz. Giriş yaptığınızda otomatik olarak şifre değiştirmeniz istenecektir.

Gemma 3

(**gemma3**): OBS'ye giriş yaparken aşağıdaki adımları izleyebilirsiniz:

1. OBS giriş sayfasına gitmek için https://obs.yildiz.edu.tr/oibs/ogrenci/login.aspx adresinden ulaşabilirsiniz.
2. Geçici şifreniz T.C. Kimlik Kartınızın seri numarasıdır. Nüfus Cüzdanı seri numarasını girerken araya boşluk veya "-" karakteri eklenmemelidir.
3. Giriş ekranında "Şifremi Unuttum" seçeneğine tıkladıktan sonra istenen bilgileri girerek geçici şifreyi oluşturabilirsiniz.
4. Giriş yaptığınızda otomatik olarak şifreniz değiştirilecektir. Şifrenizi değiştirmek için "Kullanıcı İşlemleri" menüsünden "Şifre Değiştir" seçeneğine tıklayın ve yeni şifreyi girin.

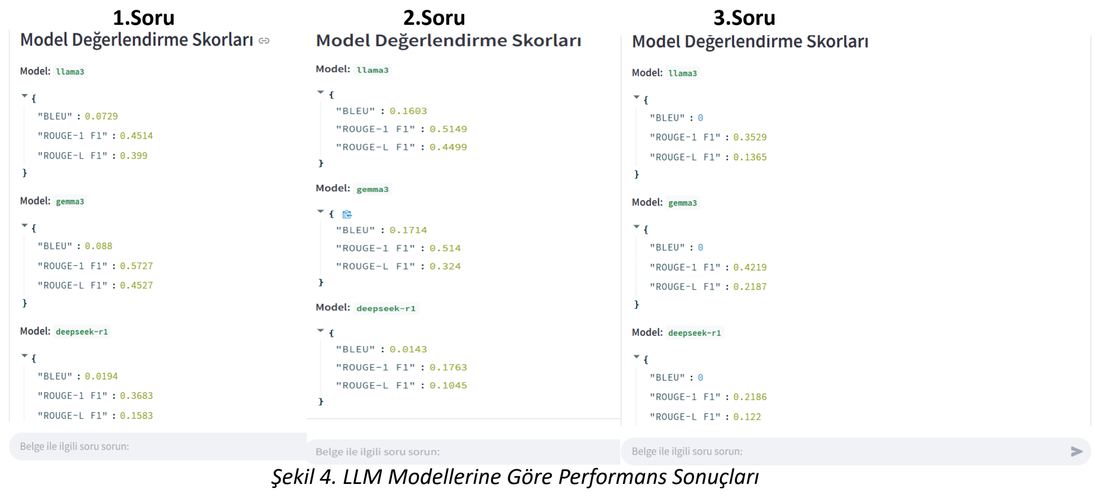
Önemli: Şifrenizi kimseyle paylaşmayın ve öğrenim süreniz boyunca güvenli bir şekilde saklayın. Şifreniz, OBS'de ders ekleme-silme, ders seçim tarihleri ile birlikte aktif olacaktır.

DeepSeek R1

OBS'ye giriş yapmak için aşağıdaki adımları izleyerek kullanıcı adı ve geçici şifre oluşturabilirsiniz:

1. **Web Sayfalarından Uygulama Açama:**
   - Üniversite web sitesinden "Uygulamalar" menüsülere gideyin.
   - Uygulamalar listesinde "Obs" uygulamasını seçin.
2. **Giriş Yapma:**
   - Açılan giriş sayfasına geçin.
   - Kullanıcı adı: sekiz haneli öğrenci numarası
   - Geçici şifre ekleyerek giriş yapın.
3. **Şifrenin Oluşturulması:**
   - OBS'te belge ettiğiniz bir metotla veya menüden seçerek geçici şifreyi oluşturun.
   - Otomatik olarakObs tesisindeki bilgiler ve konuşturmaları yapabilmelisiniz.
4. **Giriş Anakademi:**
   - After giriş, akademik işlemler performeden takip edebilirsiniz:
     - Ders ekleme-silme
     - Ders dikkat edimi
     - Not ortalaması görüntüleme
     - Staj bilgileri ve özlük bilgileri görüntüleme
     - Adres bilgisi güncelleme

Bunlarla ilgili detaylı bilgilere göreObs tesisindeki adımları takip ederek kullanıcı ve şifreyi oluşturabilirsiniz.



Şekil 4. LLM Modellerine Göre Performans Sonuçları

Verilen model yanıtları referans cevapla karşılaştırıldığında **Gemma 3**'ün hem içerik benzerliği (ROUGE-1) hem de yapısal benzerlik (ROUGE-L) bakımından en iyi performansı gösterdiğini ortaya koymaktadır. **LLaMA 3**, referans içeriğe göre anlamlı bir ikinci sırada yer almakta, ancak cümle yapısında daha düşük örtüşme sunmuştur. **DeepSeek R1**, bu sorularda en düşük benzerlik skorlarına sahiptir. BLEU metriklerinin tüm modellerde 0' yakın veya 0 olması, modellerin doğru içerik vermiş olsalar bile referans metinle birebir örtüşmeyen alternatif ifadeler kullandığını göstermektedir. Bu da BLEU'nun özellikle açık uçlu metinlerde sınırlı bir gösterge olduğunu düşündürmektedir.

SONUÇ VE YORUMLAR

LLM'ler ile PDF içeriklerine dayalı etkili chatbotlar geliştirilebilmektedir. Performans farkları, görev türüne göre değişmektedir. Metrik tabanlı değerlendirmeler, nesnel karşılaştırma sunar. Gelecek çalışmalarda daha iyi performans sonuçları için Türkçe veriyle eğitilmiş yerli modeller denenebilir, GPT-4 gibi kapalı kaynak modellerle kıyaslama yapılabilir, gerçek kullanıcı testi ile kalite doğrulanabilir. Kullanıcıdan geri bildirim toplayan mekanizmaların entegrasyonuyla model yanıtlarının insan değerlendirmesi yapılması ve daha büyük ve çeşitli PDF koleksiyonları ile testler yürütülmesi iyileştirmeye yönelik olarak tartışılabilir.

KAYNAKÇA

[1]Chen et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", 2020.
[2]Wang et al., "GLM-130B: An Open Bilingual Pre-trained Model",2022.
[3] Zhang et al. "BERTScore: Evaluating Text Generation with BERT" ,2020.
[4] Yıldız Teknik Üniversitesi Öğrenci Rehberi, YTÜ Rektörlüğü, 2024.

İSTATİSTİK BÖLÜMÜ

Kripto Para Piyasasında Risk ve Getiri Analizi Sharpe Oranı Maksimizasyonu ile Portföy Optimizasyonu

EMİRCAN KOÇ 20023060

Danışman: Doç. Dr. Gülder KEMALBAY

ÖZET

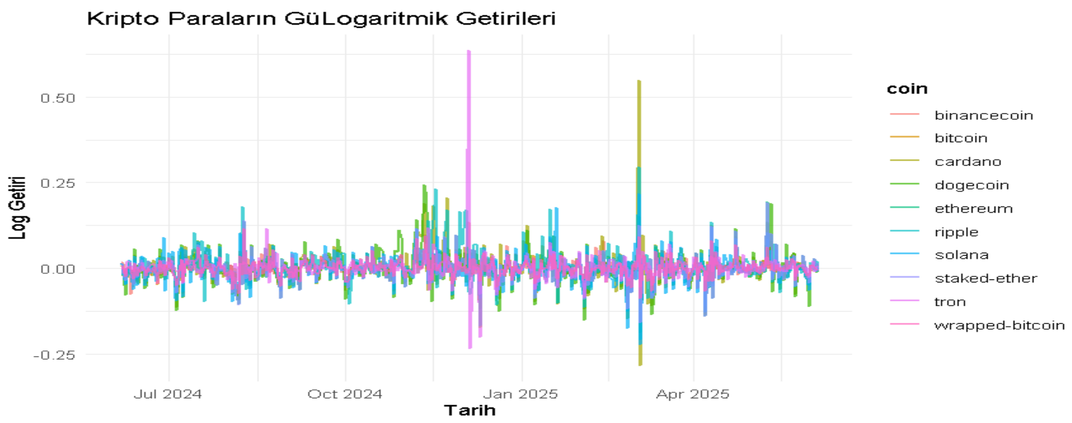
Bu çalışmada, piyasa değeri en yüksek ilk 10 kripto paranın günlük kapanış verileri kullanılarak risk-getiri analizi gerçekleştirilmiş ve yatırımcıların farklı risk tercihlerine uygun portföy yapılarını belirlemek amacıyla iki farklı optimizasyon modeli uygulanmıştır. Öncelikle coinlerin logaritmik getirileri hesaplanmış; varyans, standart sapma, Value at Risk (VaR) ve Conditional Value at Risk (CVaR) gibi risk ölçütleri ile volatilite yapıları analiz edilmiştir.Coinler arası ilişki düzeyi Pearson korelasyon katsayısı ile incelenmiş ve bu ilişkiler portföy çeşitlendirmesi açısından değerlendirilmiştir. Ardından, R programlama dilinde geliştirilen modellerle minimum varyans portföyü ve Sharpe oranını maksimize eden portföyler oluşturulmuştur. Sharpe optimizasyonu sonucu Ripple ve Tron gibi yüksek getirili coinler ön plana çıkarken; minimum varyans portföyü Wrapped-Bitcoin ve Binance Coin gibi düşük oynaklı varlıklara yönelmiştir. Elde edilen bulgular, yatırımcının risk toleransına bağlı olarak portföy kompozisyonunun önemli ölçüde değişebileceğini göstermektedir. Bu bağlamda çalışma, kripto para piyasasında veri temelli yatırım kararlarının alınmasında yol gösterici bir kaynak sunmaktadır.

Veri Toplama

Bu çalışmada kullanılan veriler, kripto para piyasalarının güncel ve güvenilir veri sağlayıcılarından biri olan CoinGecko platformu aracılığıyla elde edilmiştir. Veri çekimi, R programlama dili kullanılarak CoinGecko'nun halka açık API yapısı üzerinden otomatikleştirilmiş şekilde gerçekleştirilmiş; veriler günlük kapanış fiyatları bazında ve Amerikan Doları (USD) cinsinden alınmıştır. Analize dahil edilecek kripto paralar, 2024 yılı itibarıyla piyasa değeri en yüksek ilk 10 coin arasından seçilmiş; ancak sabit değerli kripto paralar (Tether, USDC gibi stablecoin'ler), değer oynaklığı taşımadıkları ve yatırım riski analizine uygun olmadıkları gerekçesiyle çalışma dışında tutulmuştur. Bu kapsamda analizlerde yer alan coinler şunlardır: Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB), Ripple (XRP), Solana (SOL), Dogecoin (DOGE), Cardano (ADA), Tron (TRX), Wrapped Bitcoin (WBTC) ve Staked Ether (STETH). Her bir coin için son 365 güne ait günlük kapanış fiyatları kullanılarak logaritmik getiriler hesaplanmış ve analizlerde bu seriler temel alınmıştır.

Getiri Hesaplamaları ve Zaman Serisi Yapısı

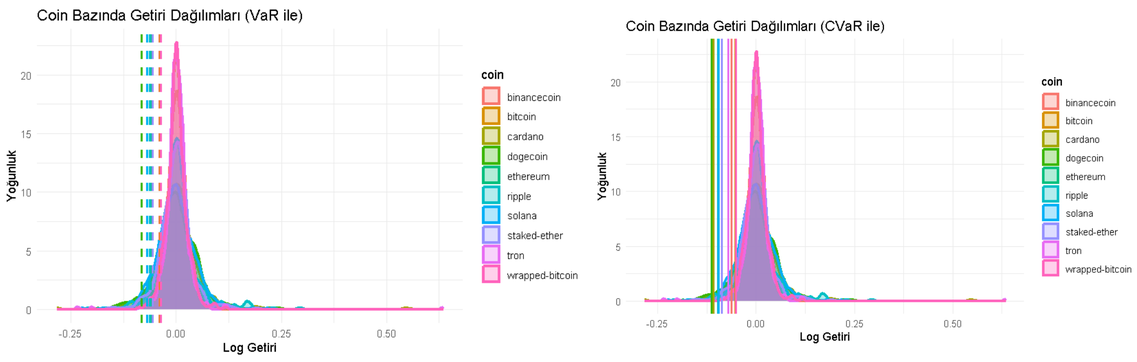
Kripto varlıkların fiyat serileri üzerinden yapılan analizlerde getiri kavramı, yatırımcı açısından beklenen kazançların ölçülmesinde temel bir unsurdur. Bu çalışmada, seçilen her bir kripto paranın logaritmik getirileri hesaplanmıştır. Logaritmik getiri, zaman serilerinin istatistiksel özelliklerini (örneğin istasyonerlik) koruyarak analiz edilmesine olanak sağladığı için tercih edilmiştir. Her bir kripto paranın Pt gün sonu kapanış fiyatı üzerinden hesaplanan logaritmik getiri formülü şu şekildedir; $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$



Bu zaman serisi grafiği, her coinin 1 yıl boyunca sergilediği günlük logaritmik getiri oynaklıklarını göstermektedir. Grafikte özellikle bazı dönemlerde yaşanan ani sıçramalar piyasa şoklarına verilen tepkileri ortaya koymaktadır. Ocak 2025 civarındaki ani yükseliş ve düşüşler, bu dönemde piyasada yaşanan volatilitenin yoğunluğunu yansıtır.

Risk Analiz Yöntemleri

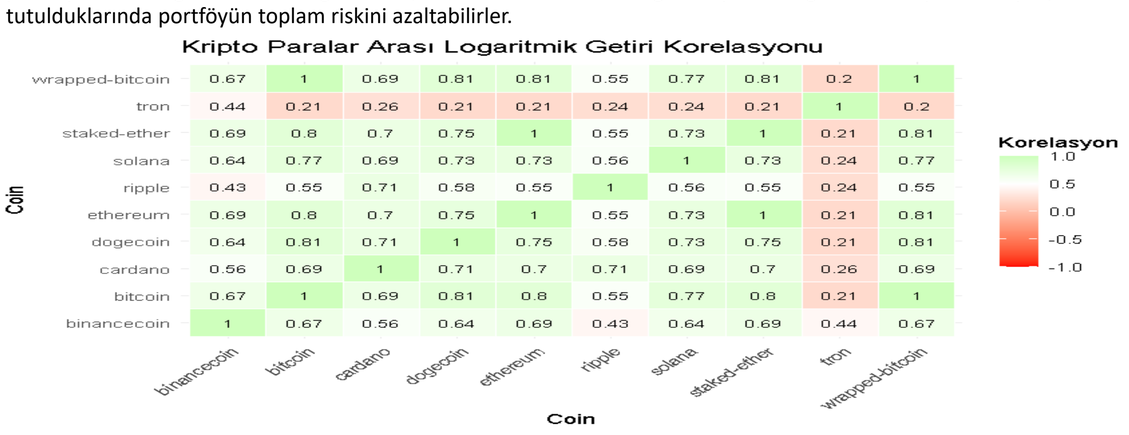
Finansal analizde risk, gelecekteki belirsizliği ve olası kayıpları ifade eder. Kripto para piyasası, yüksek oynaklık ve ani fiyat değişimleri nedeniyle geleneksel finansal enstrümanlara kıyasla daha yüksek risk barındırmaktadır.Bu nedenle, yatırım kararlarında yalnızca getiriler değil, aynı zamanda bu getirilerin riskleri de dikkate alınmalıdır. Bu çalışmada üç temel risk ölçütü kullanılmıştır.Standart sapma, Sharpe oranı, Value at Risk **(VaR)** ile Conditional Value at Risk **(CVaR)**.



Grafiklerde, analiz kapsamındaki coinlerin günlük logaritmik getirilerine göre hesaplanan %5 güven düzeyindeki Value at Risk (VaR) ve Conditional Value at Risk (CVaR) değerleri görselleştirilmiştir. VaR değerleri, her bir coin için belirli bir güven aralığında karşılaşılabilecek en yüksek beklenen zararı ifade ederken; CVaR ise bu seviyenin ötesinde gerçekleşebilecek ortalama kayıpları temsil etmektedir. Grafikler incelendiğinde, Dogecoin, Cardano ve Solana gibi bazı coinlerin hem VaR hem de CVaR değerlerinin daha sola kaydığı, yani potansiyel kayıp risklerinin daha yüksek olduğu görülmektedir. Buna karşın, Ripple ve Wrapped Bitcoin gibi coinler daha sınırlı bir kayıp riski taşımaktadır. Bu farklar, coinlerin uç risk davranışlarını karşılaştırmak ve portföyde yer alma düzeylerini belirlemek açısından öneme sahiptir.

Korelasyon Analizi

Portföy teorisinin temel varsayımlarından biri, yatırım araçlarının getirileri arasındaki ilişkilerin (korelasyonların) portföy riskini doğrudan etkilediğidir. Yani, iki varlık birbirinden bağımsız ya da negatif korelasyona sahipse, birlikte tutulduklarında portföyün toplam riskini azaltabilirler.



Bu grafik, analiz kapsamındaki coinlerin birlikte hareket etme düzeylerini sezgisel olarak sunmaktadır. Özellikle Bitcoin**,** Wrapped-Bitcoin ve Ethereum gibi büyük hacimli coinlerin kendi aralarında yüksek korelasyona sahip olduğu; buna karşın **Tron,** Cardano ve Dogecoin gibi coinlerin görece daha düşük korelasyon seviyeleriyle çeşitlendirme potansiyeli taşıdığı gözlemlenmektedir.

Portföy Optimizasyonu

Bu bölümde, önceki kısımlarda elde edilen risk ve getiri metrikleri doğrultusunda kripto para portföyü oluşturulması amaçlanmıştır. Portföy optimizasyon süreci kapsamında iki farklı yaklaşım benimsenmiştir:Minimum Varyans Portföyü ve Sharpe Oranı Maksimizasyonu.Bu yaklaşımlar doğrultusunda her bir coinin portföydeki ağırlıkları belirlenmiş, ardından portföylerin getiri ve risk performansları karşılaştırmalı olarak analiz edilmiştir.

Sharpe Oranına Göre Portföy Optimizasyonu

Sharpe oranı, yatırımcının birim risk başına en yüksek getiriyi hedeflemesini sağlayan temel bir performans ölçütüdür. Bu çalışmada, coin ağırlıklarını belirlemek amacıyla Sharpe oranını maksimize eden bir optimizasyon modeli oluşturulmuş ve R programlama dilinde nloptr fonksiyonu kullanılarak çözülmüştür. Modelde, ağırlıkların toplamının 1 olması, negatif ağırlığa izin verilmemesi ve risksiz getiri oranının sıfır kabul edilmesi gibi kısıtlar uygulanmıştır. Optimizasyon çıktısı, portföyün büyük oranda yalnızca iki coine (Ripple ve Tron) yoğunlaştığını ortaya koymuş; bu durum, analiz döneminde bu coinlerin en yüksek risk-getiri dengesini sunduğunu göstermiştir.
Model çıktısı, portföyün yalnızca iki coin üzerinde yoğunlaştığını göstermiştir:

Coin	Ağırlık (%)
Ripple	62.61
Tron	37.39
Diğerleri	0.00

Bu sonuç, analiz döneminde en iyi risk/getiri oranını sunan iki varlığın Ripple ve Tron olduğunu göstermektedir. Diğer coinlerin optimize edilen Sharpe oranı kapsamında portföye dahil edilmediği görülmektedir.

Minimum Varyans Portföyü

Minimum varyans portföyü, toplam risk düzeyini en düşük seviyeye indirmeyi hedefleyen ve özellikle temkinli yatırımcılar için uygun olan klasik bir optimizasyon yaklaşımıdır. Bu çalışmada, coinler arası kovaryans matrisi temel alınarak portföy varyansı minimize edilmiştir. `nloptr` fonksiyonu kullanılarak yapılan optimizasyonda, ağırlıkların toplamının 1 olması ve negatif pozisyona izin verilmemesi kısıtları uygulanmıştır. Sonuçta, düşük volatiliteye sahip coinlerin (örneğin Wrapped-Bitcoin ve Binance Coin) portföyde daha yüksek ağırlıklarla yer aldığı gözlemlenmiştir.

Coin	Ağırlık (%)
Wrapped BTC	59.5
Binance Coin	26.9
Tron	13.6
Diğerleri	0.00

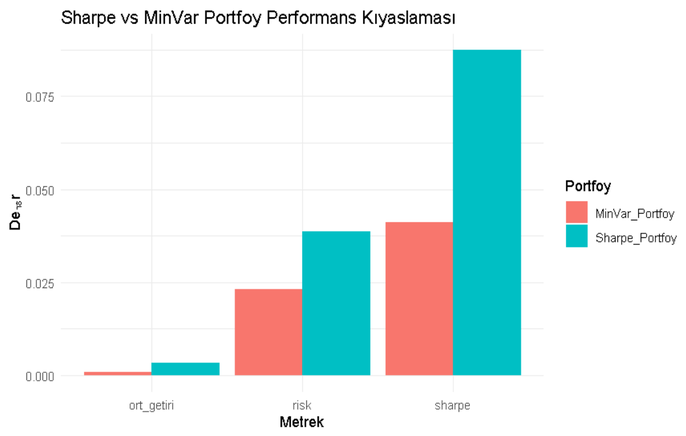
Bu portföy yapısı, daha düşük getiriler hedeflese de, oynaklığın minimize edilmesi sayesinde istikrarlı bir getiri profili sunmayı amaçlamaktadır. Yatırım tercihlerinde riskten kaçınan bireyler için daha uygun bir seçenek oluşturmaktadır.

Portföy Performans Karşılaştırması

Sharpe Oranı Maksimizasyonu ve Minimum Varyans yaklaşımlarıyla oluşturulan iki portföyün risk-getiri performansları karşılaştırmalı olarak değerlendirilmiştir. Bu değerlendirme, hem portföylerin istatistiksel metrikleri hem de görsel analizleri üzerinden yapılmıştır. Her iki portföy için ortalama günlük getiri, günlük risk (standart sapma) ve Sharpe oranı hesaplanarak aşağıdaki tabloya yansıtılmıştır:

Portföy	Ortalama Getiri	Günlük Risk	Sharpe Oranı
Sharpe Portföyü	0.00030	0.0308	0.5266
Minimum Varyans Portföyü	0.00009	0.0157	0.0585

Aşağıdaki grafikte bu iki sonuçların karşılaştırılması görsel olarak histogram çeşidinde verilmiştir.



Sharpe portföyü, günlük bazda yaklaşık **0.00030** ortalama getiri ile minimum varyans portföyüne göre daha yüksek bir kazanç sağlamaktadır. Ancak bu getiri, **0.0308** düzeyindeki görece yüksek bir oynaklıkla elde edilmektedir. Buna karşın, **Minimum Varyans portföyü** ortalama **0.00009** getiri ile daha düşük kazanç sunsa da, yalnızca **0.0157** düzeyinde bir riskle işlemektedir. Her iki portföyün Sharpe oranları da bu farkı net biçimde ortaya koymaktadır. Sharpe portföyü **0.5266** ile birim risk başına en yüksek getiriyi sağlarken, minimum varyans portföyü yalnızca **0.0585** ile oldukça düşük bir performans sunmaktadır.

Bu fark, yatırımcının risk tercihlerine göre portföy seçimini doğrudan etkilemektedir. Özellikle kısa vadeli ve daha agresif getiriler arayan yatırımcılar için Sharpe portföyü daha cazip olurken, riskten kaçınan ve sermaye korunmasını önceliklendiren yatırımcılar için minimum varyans portföyü uygun bir alternatif olarak öne çıkmaktadır.

KAYNAKÇA

[1] Markowitz, H. (1952). *Portfolio Selection*. The Journal of Finance, 7(1), 77–91
[2] Sharpe, W. F. (1966). *Mutual Fund Performance*. The Journal of Business, 39(1), 119–138
[3] Ekinci, A. (2020). *Finans Ekonomi ve Sosyal Araştırmalar Dergisi*, 5(3), 655–672

İSTATİSTİK BÖLÜMÜ

TENİS MAÇLARINDA OYUNCU PERFORMANSINI ETKİLEYEN FAKTÖRLERİN MAKİNE ÖĞRENMESİYLE ANALİZİ

EMİRHAN ÇINAR 20023037

Danışman: Prof.Dr.Ersoy ÖZ

ÖZET

Bu çalışma, 2013–2023 yılları arasında oynanan 494 erkek tenis maçının verilerini kullanarak oyuncu performansını etkileyen faktörleri makine öğrenmesi yöntemleriyle analiz etmektedir. Zemin türü, sıcaklık, nem, yaş ve BMI gibi değişkenlerin maç süresi ve sonucu üzerindeki etkileri incelenmiştir. Veriler üzerinde ön işleme ve SMOTE yöntemiyle dengeleme yapılmış; farklı makine öğrenmesi algoritmalarıyla performans tahmin modelleri oluşturulmuştur. En yüksek doğruluğa SVM, en yüksek AUC değerine ise Gradient Boosting modeli ulaşmıştır. Bulgular, çim zeminin servis oyuncuları için avantaj sağladığını, sıcaklık ve nemin maç süresini uzattığını, fiziksel özelliklerin ise sonuç üzerinde etkili olduğunu göstermektedir. Bu sonuçlar, spor stratejisi ve organizasyon süreçlerinde rehberlik edebilir.

VERİ SETİ

	Turnuva	Oyuncular	Sonuç	Ace Sayısı	Çift Hata Sa	Servis Yü	Sıcaklı	Nem	Zemin	Turnuva Si	Sonuç Set	Kazandığı S	Maç Süreler	Yaşı	BMI
2	Australian Open	Novak Djokovic	1	9	2	65	31	39	Sert Zemin	5	4	3	220 dakika	25	21,79
3	Australian Open	Andy Murray	0	8	3	62	31	39	Sert Zemin	5	4	1	220 dakika	25	23,03
4	Australian Open	Stan Wawrinka	1	19	3	68	31	32	Sert Zemin	5	4	3	141 dakika	28	24,19
5	Australian Open	Rafael Nadal	0	1	2	74	31	40	Sert Zemin	5	4	1	141 dakika	27	24,84
6	Australian Open	Novak Djokovic	1	8	2	74	29	45	Sert Zemin	5	4	3	219 dakika	27	21,79
7	Australian Open	Andy Murray	0	12	4	60	29	45	Sert Zemin	5	4	1	219 dakika	27	23,03
8	Australian Open	Novak Djokovic	1	7	2	69	33	38	Sert Zemin	5	3	3	173 dakika	28	21,79
9	Australian Open	Andy Murray	0	8	5	64	33	38	Sert Zemin	5	3	0	173 dakika	28	23,03
10	Australian Open	Roger Federer	1	20	3	63	30	42	Sert Zemin	5	5	3	218 dakika	35	24,84
11	Australian Open	Rafael Nadal	0	4	2	73	30	42	Sert Zemin	5	5	2	218 dakika	30	24,84
12	Australian Open	Roger Federer	1	24	3	63	35	36	Sert Zemin	5	5	3	183 dakika	36	24,84
13	Australian Open	Marin Cilic	0	16	5	61	35	36	Sert Zemin	5	5	2	183 dakika	29	22,7
14	Australian Open	Novak Djokovic	1	9	2	68	34	31	Sert Zemin	5	3	3	124 dakika	31	21,79
15	Australian Open	Rafael Nadal	0	5	3	65	34	31	Sert Zemin	5	3	0	124 dakika	32	24,84
16	Australian Open	Novak Djokovic	1	9	3	64	31	44	Sert Zemin	5	5	3	239 dakika	32	21,79
17	Australian Open	Dominic Thiem	0	8	5	67	31	44	Sert Zemin	5	5	2	239 dakika	26	23,08
18	Australian Open	Novak Djokovic	1	9	2	67	26	50	Sert Zemin	5	3	3	113 dakika	33	21,79
19	Australian Open	Daniil Medvedev	0	6	4	64	26	50	Sert Zemin	5	3	0	113 dakika	25	21,17
20	Australian Open	Rafael Nadal	1	3	2	64	25	52	Sert Zemin	5	5	3	324 dakika	35	24,84
21	Australian Open	Daniil Medvedev	0	23	5	69	25	52	Sert Zemin	5	5	2	324 dakika	25	21,17
22	Australian Open	Novak Djokovic	1	12	3	65	33	39	Sert Zemin	5	3	3	176 dakika	35	21,79
23	Australian Open	Stefanos Tsitsipas	0	8	4	62	33	39	Sert Zemin	5	3	0	176 dakika	24	23,89
24	French Open	Rafael Nadal	1	4	3	69	30	41	Toprak Zemin	5	3	3	136 dakika	27	24,84
25	French Open	David Ferrer	0	2	4	65	30	41	Toprak Zemin	5	3	0	136 dakika	31	23,84
26	French Open	Rafael Nadal	1	3	2	71	24	55	Toprak Zemin	5	4	3	211 dakika	28	24,84
27	French Open	Novak Djokovic	0	2	3	67	24	55	Toprak Zemin	5	4	1	211 dakika	27	21,79

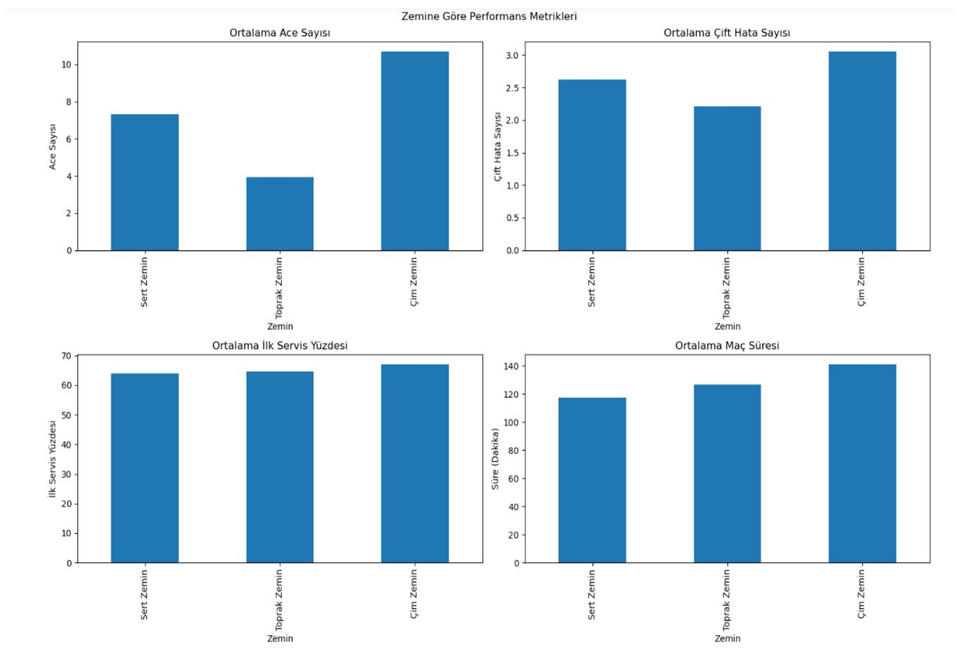
Veri setinde, tenis maçlarına ilişkin oyuncu, maç ve çevresel faktörlere dair değişkenler yer almaktadır. Turnuva ve Oyuncular, maç bilgilerini belirtirken; Sonuç, oyuncunun maçı kazanıp kazanmadığını gösterir. Ace Sayısı, Çift Hata Sayısı ve Servis Yüzdesi, servis performansını yansıtır. Sıcaklık ve Nem, ortam koşullarını; Zemin, kort yüzey tipini belirtir. Turnuva Set Sayısı, Sonuç Seti ve Kazandığı Set, maçın set bazlı sonuçlarını sunar. Maç Süresi, karşılaşmanın toplam süresini; Yaş ve BMI, oyuncunun fiziksel özelliklerini ifade eder.

MODEL PERFORMANSLARI

1.Maç Sonucu Tahmincisi (SVM)

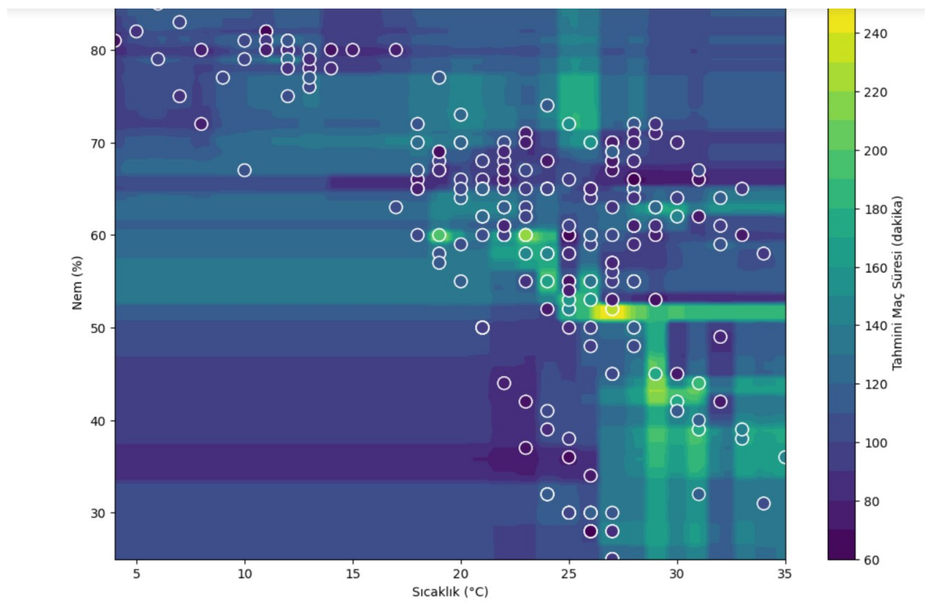
SVM modeli kullanılarak, maç sonucu tahmini yapan bir sistem geliştirilmiştir. Sistem; yaş, BMI, ace sayısı, çift hata, servis yüzdesi, sıcaklık, nem ve zemin türünü dikkate alarak kazanma olasılığını hesaplamaktadır. Örnek senaryolarda, yüksek servis yüzdesi ve düşük hata oranı olan oyuncuların kazanma ihtimalinin daha yüksek olduğu görülmüştür. Ayrıca, zemin türü ve çevresel faktörlerin de performansı etkilediği belirlenmiştir.
Örneğin; 25 yaşında, BMI'si 23.5 olan, %65 ilk servis yüzdesiyle oynayan bir oyuncunun sert zeminde kazanma olasılığı %45.95 olarak hesaplanmıştır. Diğer yandan, 30 yaşında, BMI'si 24.0 olan ve %70 servis yüzdesine sahip bir oyuncunun çim zeminde kazanma olasılığı %72.19'dur. Bu örnekler, teknik becerilerin ve zemin türünün kazanma ihtimali üzerinde belirleyici olduğunu göstermektedir.

2.Zemin-Performans İlişkisi



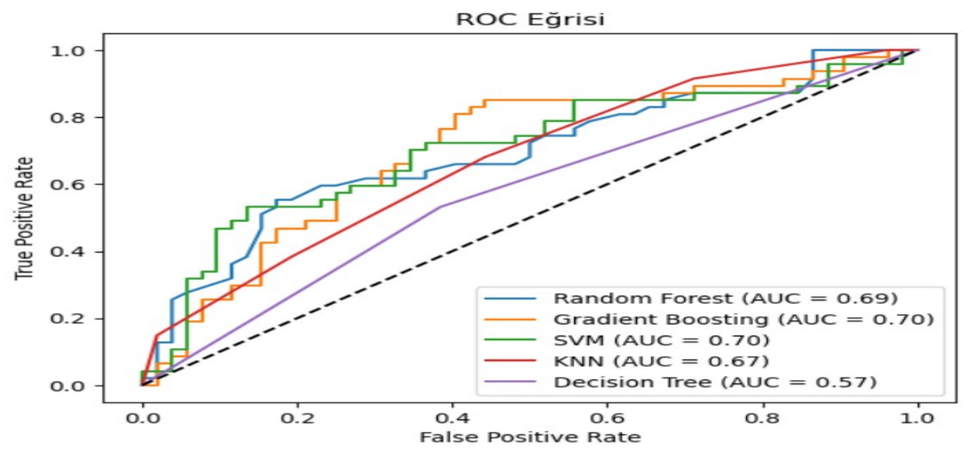
Zemin türü, oyuncu performans metriklerini belirgin şekilde etkilemektedir. Örneğin, çim zeminde ace sayısı ortalama 10.68 ile en yüksek değerdeyken, sert zeminde 7.32, toprakta ise 3.94 olarak ölçülmüştür. İlk servis yüzdesi de çim zeminde en yüksek olup %67.08'dir; bu oran toprakta %64.68, sert zeminde %63.94'tür. Çift hata sayısı en düşük 2.21 ile toprak zeminde görülmüş, en yüksek ise 3.05 ile çim zeminde kaydedilmiştir. Maç süresi bakımından ise çim zemin 141 dakika ile en uzun, sert zemin ise 117.14 dakika ile en kısa ortalamaya sahiptir. Bu veriler, zemin türlerinin oyuncu davranışlarını ve maç dinamiklerini anlamlı biçimde etkilediğini göstermektedir.

3.Sıcaklık ve Nem'in Maç Süresine Etkisi

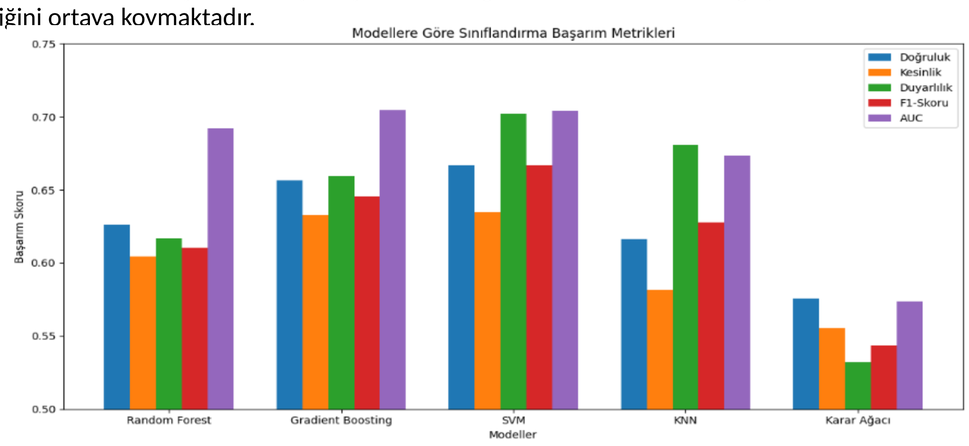


Analiz sonuçları, sıcaklık ve nemin tenis maçlarının süresi üzerinde belirleyici bir etkiye sahip olduğunu göstermektedir. Özellikle 30°C üzeri sıcaklık ve %70'ten yüksek nem koşullarında maçların belirgin şekilde uzadığı görülmüştür. Özellik önemliliği grafiği, nemin maç süresi üzerindeki etkisinin sıcaklıktan biraz daha fazla olduğunu ortaya koymuştur. Model karşılaştırmalarında ise Random Forest, iklim değişkenlerinin karmaşık etkilerini daha başarılı şekilde yansıtarak Karar Ağacı modeline üstünlük sağlamıştır. Bu bulgular, zorlu hava koşullarının oyuncu dayanıklılığı üzerinde doğrudan etkili olduğunu desteklemektedir.

PERFORMANS SONUÇLARI



Makine öğrenmesi modellerinin sınıflandırma performansları doğruluk, kesinlik, duyarlılık, F1-skoru ve AUC metrikleriyle değerlendirilmiştir. SVM, %66.7 doğruluk ile en iyi sonucu verirken, Gradient Boosting modeli 0.705 AUC değeri ile pozitif-negatif ayrımında en başarılı model olmuştur.
İlk grafikteki ROC eğrileri, Gradient Boosting ve SVM modellerinin en yüksek eğri altında kalan alana (AUC) sahip olduğunu, Decision Tree'nin ise sınıflandırma açısından zayıf kaldığını göstermektedir.
İkinci grafikte ise, SVM ve Gradient Boosting modellerinin tüm metriklerde genel olarak üstün olduğu, Random Forest ve KNN'nin orta seviyede, Karar Ağacı'nın ise düşük performans sergilediği açıkça görülmektedir. Bu sonuçlar, karmaşık ilişkilerin olduğu veri setlerinde daha güçlü modellerin tercih edilmesi gerektiğini ortava kovmaktadır.



KAYNAKÇA

[1] A. Rajkumar, The Effect of Different Court Surfaces on Tennis Player Performance and Injury Risk, 2020.
[2] T. Barnett, G. Pollard, Tennis Surface Study - Information Visualization Final Report, University of California, Berkeley, 2016.
[3] N. D. Clarke, D. Fewtrell, R. Macrae, The influence of environmental parameters on tennis match characteristics, Journal of Sports Sciences, vol. 29, no. 14, pp. 1531–1539, 2011.

İSTATİSTİK BÖLÜMÜ

Müşteri Terk Oranının Makine Öğrenmesi Yöntemleriyle Tahmini

ENES COŞKUN 21023005

Danışman: Doç.Dr.Elif Tuna

ÖZET

Bu çalışmada, e-ticaret sektöründe müşteri terk oranını tahmin etmek amacıyla Random Forest, XGBoost ve LightGBM gibi makine öğrenmesi yöntemleri kullanılmıştır. Veri seti bir e-ticaret firmasına ait 5630 müşteri verisinden oluşmakta olup, ön işleme ve SMOTE gibi teknikler uygulanmıştır. Model performansları çeşitli metriklerle değerlendirilmiş ve LightGBM algoritması genel başarı açısından en iyi sonucu vermiştir. SHAP analizleri ile müşteri şikayeti, memnuniyet skoru ve son sipariş tarihi gibi davranışsal değişkenlerin churn üzerinde etkili olduğu belirlenmiştir .

CHURN NEDİR VE NEDEN ÖNEMLİDİR

CHURN, bir müşterinin belirli bir zaman dilimi içinde bir hizmeti ya da ürünü kullanmayı bırakmasıdır.
Özellikle e-ticaret gibi rekabetin yoğun olduğu sektörlerde churn, müşteri sadakatinin en kritik göstergesidir.
Churn oranını doğru analiz edebilmek; müşteri kaybını azaltmak, gelir sürekliliğini sağlamak ve pazarlama stratejilerini optimize etmek açısından büyük önem taşır.

VERİ SETİNİ TANIYALIM

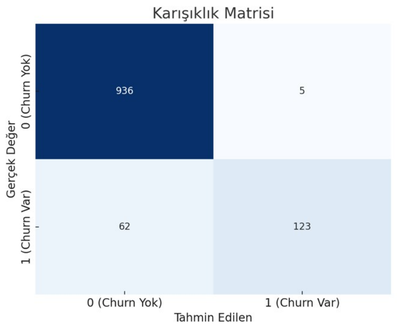
Bu çalışma için kullanılan veri seti, bir e-ticaret platformundan elde edilen 5630 müşteriye ait işlem ve davranış verilerini içermektedir. Veri setinde müşterilerin alışveriş geçmişi, memnuniyet düzeyleri, kampanya kullanım bilgileri ve terk durumu gibi toplam 32 değişken bulunmaktadır. Veri seti, eğitim ve test olmak üzere %80-%20 oranında ikiye ayrılmış; dengesiz sınıflar SMOTE yöntemi ile dengelenmiştir. Veriler analiz öncesi eksik gözlemlerden arındırılmış ve kategorik değişkenler sayısallaştırılmıştır.

Değişken	Açıklama
MüşteriNo	Benzersiz müşteri kimliği
Terk	Terk (churn)
Tercih Edilen Giriş Cihazı	Müşterinin tercih ettiği giriş cihazı
Şehir Büyüklüğü	Şehir büyüklüğü
Depodan Eve Mesafe	Depodan müşterinin evine olan mesafe
Tercih Edilen Ödeme Yöntemi	Müşterinin tercih ettiği ödeme yöntemi
Cinsiyet	Müşterinin cinsiyeti
Uygulamada HarcananS aat	Mobil uygulama veya web sitesinde harcanan saat
Kayıtlı Cihaz Sayısı	Belirli bir müşteriye kayıtlı cihaz sayısı
Tercih Edilen Sipariş Kategorisi	Müşterinin geçen ay en çok sipariş verdiği kategori
Memnuniyet Skoru	Hizmetle ilgili müşteri memnuniyet puanı
Medeni Durum	Müşterinin medeni durumu
Adres Sayısı	Müşteri adına kayıtlı adres sayısı
Değişken	Açıklama
Şikayet	Geçen ay şikayet oluşturulup oluşturulmadığı
Geçen Yıla Göre SiparişArtışı	Geçen yıla göre sipariş miktarındaki artış yüzdesi
Kullanılan Kupon	Geçen ay kullanılan kupon sayısı
Sipariş Sayısı	Geçen ay verilen toplam sipariş sayısı
Son Siparişten Sonra Geçen Gün	Müşterinin son siparişinden bu yana geçen gün sayısı
Aylık Geri Ödeme Miktarı	Geçen ayki ortalama geri ödeme (cashback) miktarı

Kullanılan Makine Öğrenmesi Modelleri

Random Forest

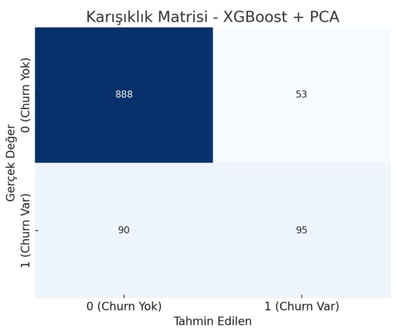
Birçok karar ağacından oluşan ve bu ağaçların tahminlerini birleştirerek daha kararlı sonuçlar üreten bir topluluk (ensemble) yöntemidir. Ağaçlar farklı veri alt kümeleriyle eğitilir ve sonuçlar ortalama alınarak belirlenir, bu sayede aşırı öğrenme riski azaltılır.



Model, churn olmayan müşterileri neredeyse hatasız tahmin etmiş (936 TN, 5 FP) ve churn eden müşterilerin büyük kısmını da doğru sınıflamıştır (123 TP, 62 FN).

XGBoost

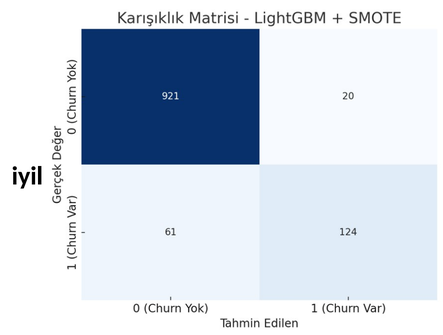
Ardışık şekilde karar ağaçları oluşturarak önceki modellerin hatalarını gidermeye odaklanan güçlü bir gradyan artırma yöntemidir. Özellikle hız ve doğruluk açısından büyük veri setlerinde etkin sonuçlar verir.



Bu karışıklık matrisine göre model, churn olmayan müşterileri büyük oranda doğru sınıflandırmıştır (888 TN, 53 FP).Ancak churn eden müşterilerin neredeyse yarısı hatalı şekilde sınıflanmıştır (90 FN, 95 TP). Bu durum, PCA uygulaması sonrası churn sınıfında duyarlılığın düştüğünü, modelin ağırlıklı olarak çoğunluk sınıfına odaklandığını göstermektedir.

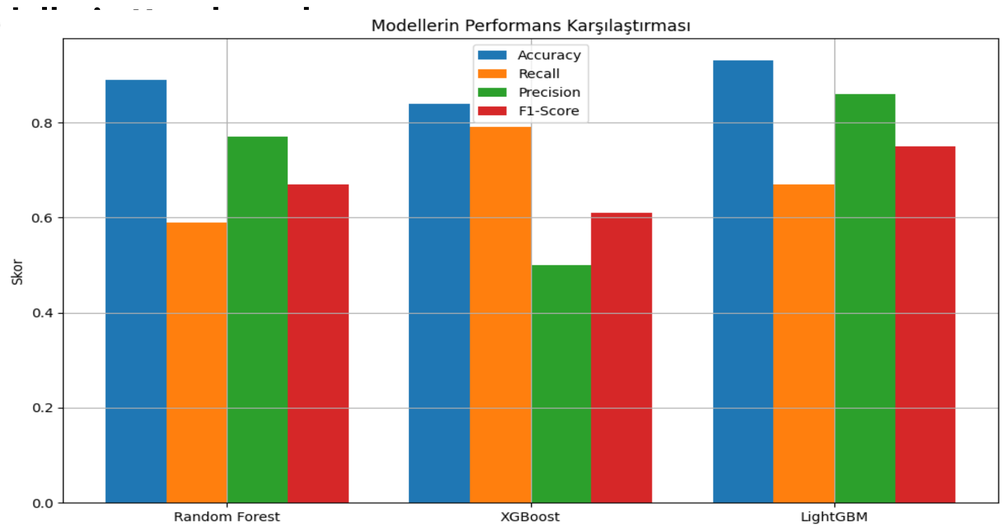
LightGBM

Daha az bellek kullanarak yüksek doğruluk sağlayan ve özellikle kategorik değişkenleri hızlı işleyebilen bir gradyan artırma yöntemidir. XGBoost'a benzer şekilde çalışsa da, genellikle daha yüksek verimlilik ve performans sunar.



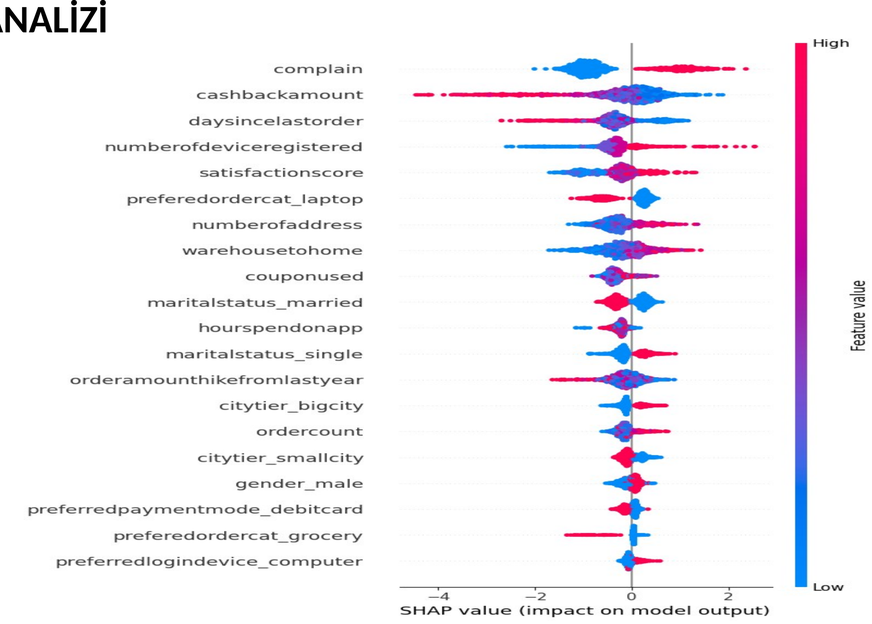
Model, churn olmayan müşterilerin neredeyse tamamını doğru sınıflandırmıştır (921 TN, 20 FP).Churn eden müşterilerin ise büyük çoğunluğu doğru tahmin edilmiştir (124 TP, 61 FN), bu da churn sınıfında anlamlı bir işaret eder.

Mo



Modeller, accuracy, recall, precision ve f1-score gibi temel performans metrikleri üzerinden karşılaştırılmıştır. LightGBM, tüm metriklerde dengeli ve yüksek performans göstererek en başarılı model olmuştur. Özellikle yüksek precision ve f1-score değerleriyle churn tahmininde isabetli ve dengeli sonuçlar vermiştir. XGBoost ise yüksek recall değeri sayesinde churn eden müşterileri en iyi yakalayan model olarak öne çıkmıştır. Random Forest genel doğruluk açısından güçlü olsa da, churn sınıfını tanıma konusunda diğer modellere göre zayıf kalmıştır. Genel olarak LightGBM, dengesiz veri yapısına rağmen en tutarlı ve güvenilir performansı sergilemiştir.

SHAP ANALİZİ



Modelin karar sürecini açıklamak amacıyla SHAP yöntemi kullanılmış ve değişkenlerin churn tahminine etkisi analiz edilmiştir. SHAP özet grafiğine göre en belirleyici değişken müşteri şikayetidir; şikayet eden müşterilerin churn olasılığı anlamlı şekilde artmaktadır. Ayrıca, düşük cashback miktarı ve uzun süredir sipariş vermemiş olmak churn riskini artıran önemli faktörlerdir. Müşteri memnuniyeti, cihaz kaydı sayısı ve kupon kullanımı gibi davranışsal değişkenler modelin tahmininde önemli rol oynamaktadır. Demografik özelliklerin etkisi ise daha sınırlı kalmıştır. Sonuç olarak, modelin churn tahminleri büyük ölçüde müşteri davranışlarına dayanmaktadır ve stratejilerin bu alana odaklanması önerilmektedir.

KAYNAKÇA

[1] Manzoor, A., Qureshi, M. A., & Kidney, E. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners.
[2] Oğuz, K., & Yasin, G. (2017). *Makine öğrenmesi yöntemleriyle müşteri kaybı analizi*.
[3] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

2024-2025 Bahar Yarıyılı

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KREDİBİLİTE TAHMİNİ

Eren DEMİR 20023062

Danışman: Doç. Dr. Selçuk ALP

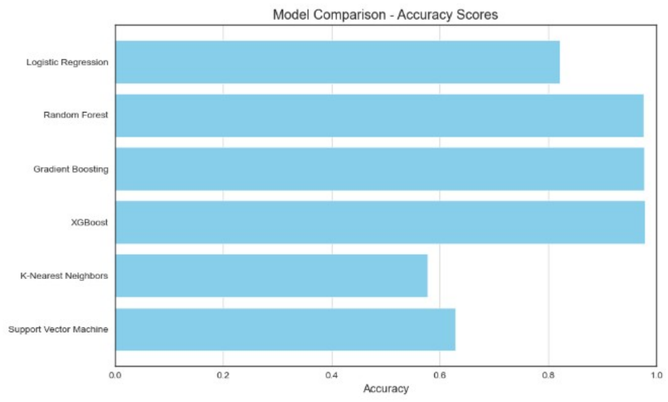
ÖZET

Bu çalışmada, bireylerin kredi başvurularının onaylanma durumları; gelir, varlıklar, kredi tutarı gibi finansal bilgiler temel alınarak makine öğrenmesi algoritmaları ile tahmin edilmiştir. XGBoost, Random Forest gibi yöntemler uygulanmış; en yüksek başarı, %98 doğrulukla CIBIL kredi skoru kullanıldığında elde edilmiştir. CIBIL Skoru olmadan performans ciddi bir düşüş göstererek %65 seviyesine düşmüştür. Bu durum, kredi skorunun model açıklayıcılığı açısından kritik bir değişken olduğunu göstermektedir.

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KREDİBİLİTE TAHMİNİ

Kullanılan Metodoloji

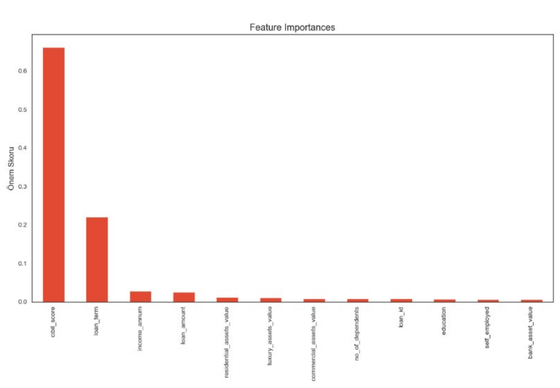
- Kredi başvurularına ait demografik ve finansal veriler incelendi.
- Verideki eksik ve kategorik bilgiler düzenlendi.
- Kredi skoru içeren ve içermeyen iki ayrı model senaryosu oluşturuldu.
- Altı farklı makine öğrenmesi algoritması kullanılarak tahmin modelleri kuruldu.
- Modeller doğruluk, kesinlik ve ROC eğrisi MCC gibi performans kriterlerine göre karşılaştırıldı.
- En başarılı modelin XGBoost olduğu belirlendi.
- Özelliklerin önem dereceleri analiz edilerek sonuçlar yorumlandı.

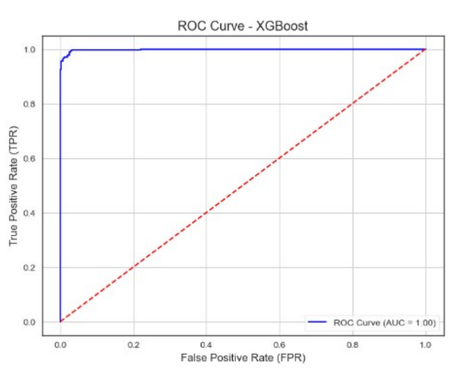


Makine öğrenmesiyle kredi başvuru kararlarını öngörebilmek için farklı yapısal özelliklere sahip altı güçlü algoritma test edildi. Bu modeller:

- **XGBoost**
- **Random Forest**
- **Gradient Boosting**
- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**

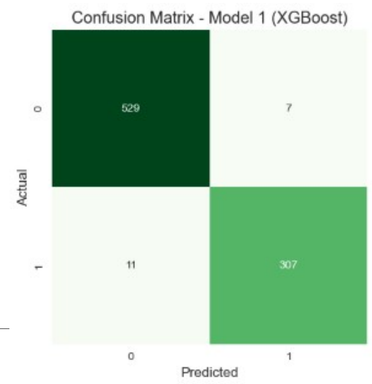
Her biri farklı öğrenme stratejilerine sahip bu algoritmalar, hem doğrusal hem de karmaşık ilişkileri öğrenme kabiliyetleri açısından test edildi. Bu çeşitlilik, hangi algoritmanın bu problem için en başarılı sonuçları verdiğini objektif olarak değerlendirebilmemizi sağladı.

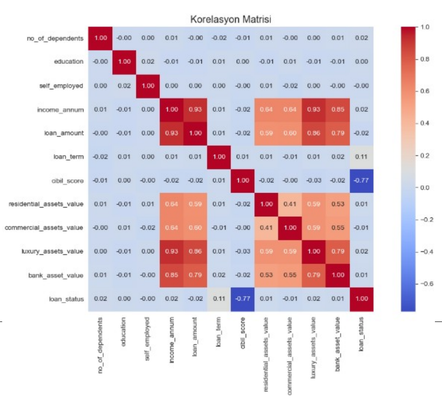


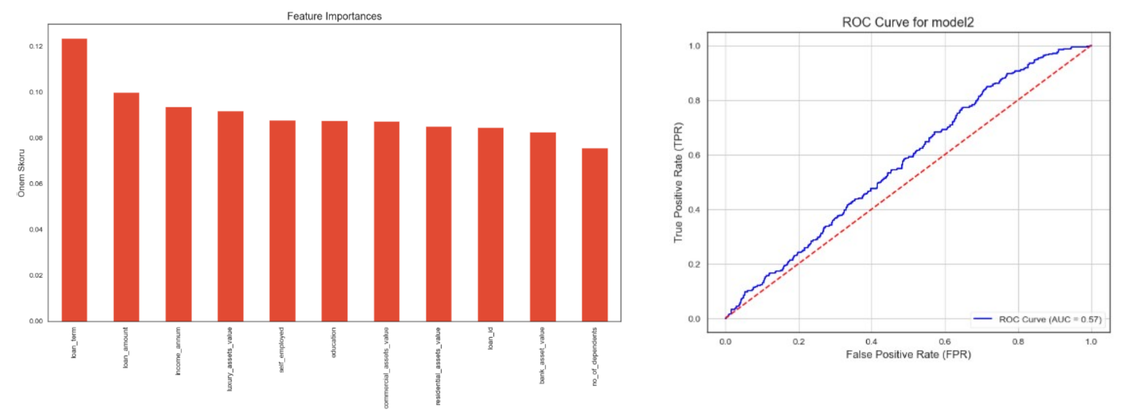


CIBIL Skoru Her Şeyi Değiştiriyor

Modelleme sürecine CIBIL skoru dahil edildiğinde, bu değişken tek başına modelin kararlarının büyük kısmını yönlendirmiştir. Özellik önem analizine göre %65'in üzerinde katkı sağlayarak diğer tüm değişkenleri gölgede bırakmıştır.
Modelin sınıflar arasında ayırt ediciliği de bu katkıyla birlikte dramatik şekilde artmıştır. AUC skoru 1.00'e ulaşarak, XGBoost algoritmasının neredeyse kusursuz bir sınıflandırma yaptığı görülmüştür. Bu durum, kredi skorunun sadece bir değişken değil; adeta bir tahmin motoru gibi çalıştığını göstermektedir.

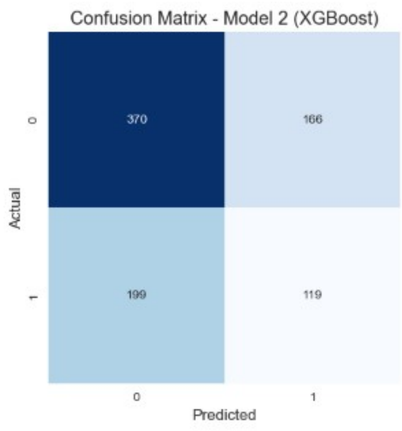


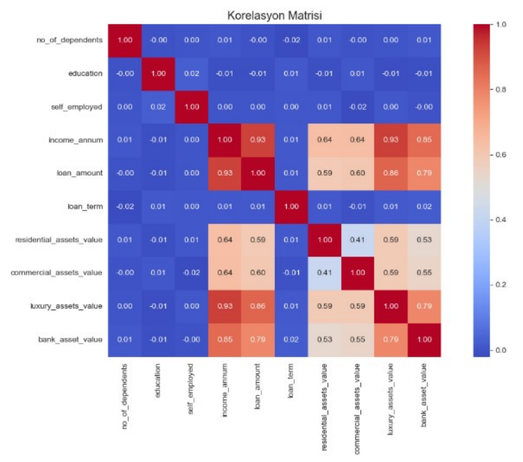




CIBIL Skoru Olmadan: Model Ne Söylüyor?

CIBIL skoru hariç bırakıldığında model, kararlarını kredi vadesi, tutarı ve yıllık gelir gibi daha zayıf göstergeler üzerinden vermek zorunda kalmıştır. Özellik önem grafiği yüzeyde dengeli görünse de, aslında modelin hangi değişkene güveneceğini bilemediğini ve "her şeyden biraz" öğrenmeye çalıştığını göstermektedir.
Bu belirsizlik, ROC eğrisine de doğrudan yansımıştır. AUC skoru sadece **0.57** ile neredeyse rastgele tahmin seviyesine inmiş; modelin ayırt etme gücü ciddi oranda zayıflamıştır. Güçlü bir skor verisi olmadan kurulan modellerin, karar destek sistemlerinde yeterince güvenilir sonuçlar sunamayabileceği bu çıktılarla açıkça görülmektedir.





CIBIL Skoru Olmadan Model Ne Kadar Zorlanıyor?

CIBIL skoru hariç bırakıldığında modelin performansı belirgin şekilde düşmüştür. Confusion matrix, modelin özellikle riskli başvuruları (1) doğru sınıflandıramadığını göstermektedir: 199 riskli başvuru yanlışlıkla onaylanmış, bu da güvenlik açısından ciddi bir açık oluşturmuştur.
Korelasyon matrisi incelendiğinde ise hiçbir değişkenin kredi onay durumu ile güçlü bir ilişkisi olmadığı görülmektedir. Bu durum, CIBIL skorunun yokluğunda modelin tahmin gücünü destekleyecek güçlü sinyallerin eksikliğini ortaya koymaktadır.

Sonuç

Bu çalışmada, bireylerin kredi başvuru sonuçları makine öğrenmesi yöntemleriyle tahmin edilmeye çalışılmıştır. CIBIL skoru modele dahil edildiğinde XGBoost algoritması %98 doğruluk ve güçlü ayırt etme kabiliyeti ile öne çıkmıştır. Ancak skor verisi hariç bırakıldığında modelin performansı belirgin biçimde düşmüş; doğruluk, AUC ve tutarlılık kayda değer şekilde azalmıştır. Bu da kredi skoru gibi bileşik göstergelerin, karar destek sistemlerinde ne kadar kritik olduğunu ortaya koymuştur. Gelecekte, skor verisine sahip olmayan bireyler için alternatif öngörücü değişkenlerin zenginleştirilmesi yönünde çalışmalar yapılması faydalı olacaktır.

KAYNAKÇA

[1] Matplotlib, Visualization with Matplotlib, https://matplotlib.org/stable/gallery/index.html, 12 Temmuz 2024.
[2] Python Software Foundation, Python 3.11.3 Documentation, https://docs.python.org/3/, 12 Temmuz 2024.
[3] XGBoost Documentation, Introduction to Boosted Trees, https://xgboost.readthedocs.io/en/stable/tutorials/model.html, 12 Temmuz 2024

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

MÜŞTERİ ALIŞVERİŞ EĞİLİMLERİ ANALİZİ

Eylül DEMİR 21023072

Danışman: Prof. Dr. Filiz KARAMAN

ÖZET

Bu tez çalışması, müşteri alışveriş eğilimlerini anlamak ve tüketici davranışlarına yönelik bulgular elde etmek amacıyla kapsamlı bir veri analizi sunmayı hedeflemektedir. Müşteri demografisi, satın alma alışkanlıkları ve harcama davranışlarını içeren veri seti üzerinde Kategori Bazında Satın Alma Tutarları ve RFM analizleri gibi temel incelemelerin yanı sıra, K-means kümeleme ile müşteri segmentasyonu ve Random Forest gibi makine öğrenimi algoritmaları kullanılarak değişken önem derecesi analizleri gerçekleştirilmiştir. Elde edilen bulgular, müşteri segmentlerinin belirlenmesi ve satın alma davranışlarını etkileyen kritik faktörlerin ortaya konulmasıyla, işletmelerin pazarlama stratejilerini optimize etmesine ve müşteri memnuniyetini artırmasına katkı sağlayabileceğini göstermektedir.

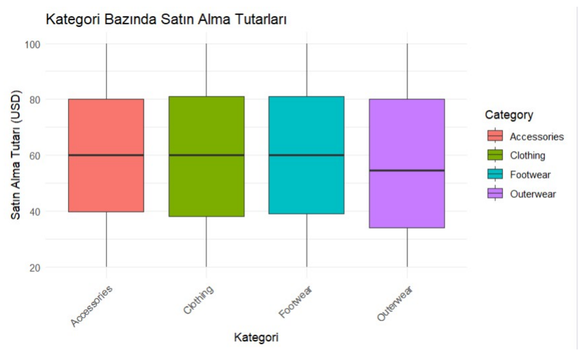
Müşteri Alışveriş Eğilimleri Analizi ve Pazarlama Stratejileri

Müşteri Alışveriş Eğilimlerinin Analizi, işletmelerin müşteri davranışlarını daha iyi anlayarak pazarlama stratejilerini veri odaklı şekilde şekillendirmelerine yardımcı olmak amacıyla gerçekleştirilmiş bir çalışmadır. Bu bitirme projesinde, Kaggle üzerinden alınan müşteri alışveriş eğilimleri veri seti kullanılarak demografik özellikler, satın alma tercihleri gibi çeşitli faktörler analiz edilmiştir. Bu analizler, veri odaklı müşteri ilişkileri yönetimi, hedefli kampanya geliştirme ve müşteri sadakat stratejileri açısından işletmelere destek sağlamayı amaçlamaktadır.

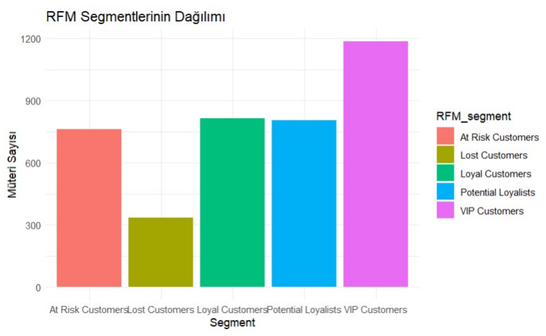
Veri Seti

Kullanılan veri seti tüketici davranışları ve satın alma kalıpları hakkında bilgiler içermektedir. Veri setinde 3900 adet müşteri bulunmaktadır ve her müşteri için 19 farklı özellik bulunmaktadır. Bu özellikler arasında müşteri ID, yaş, cinsiyet, satın alınan ürün, kategori, satın alma tutarı, lokasyon, beden, renk, sezon, inceleme puanı, abonelik durumu, ödeme yöntemi, gönderim türü, indirim uygulandı (evet/hayır), promosyon kodu kullanıldı (evet/hayır), önceki satın almalar, tercih edilen ödeme yöntemi ve satın alma sıklığı verileri yer almaktadır.

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (U	Location	Size	Color	Season	Review Rating	Subscription Status	Payment Method	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Preferred Payment Me	Frequency of Purchase
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Credit Card	Express	Yes	Yes	14	Venmo	Fortnightly
2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Bank Transfer	Express	Yes	Yes	2	Cash	Fortnightly
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Cash	Free Shipping	Yes	Yes	23	Credit Card	Weekly
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	PayPal	Next Day Air	Yes	Yes	49	PayPal	Weekly
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Cash	Free Shipping	Yes	Yes	31	PayPal	Annually
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes	Venmo	Standard	Yes	Yes	14	Venmo	Weekly
7	63	Male	Shirt	Clothing	85	Montana	M	Gray	Fall	3.2	Yes	Debit Card	Free Shipping	Yes	Yes	49	Cash	Quarterly
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes	Debit Card	Free Shipping	Yes	Yes	19	Credit Card	Weekly
9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes	Venmo	Express	Yes	Yes	8	Venmo	Annually
10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8	Yes	PayPal	2-Day Shipping	Yes	Yes	4	Cash	Quarterly



Kategori bazında satın alma tutarları analizi, farklı ürün kategorilerindeki harcama dağılımlarını incelemek için gerçekleştirilmiştir. Kutu grafiği (box plot), her bir kategori için satın alma tutarlarının medyan, çeyrekler arası aralık ve aykırı değerlerini göstermektedir. Grafik incelendiğinde, Aksesuar (Accessories), Giyim (Clothing) ve Ayakkabı (Footwear) kategorilerinin medyan satın alma tutarları birbirine oldukça yakın olup yaklaşık olarak 60 USD civarındadır. Bu, üç kategorideki alışverişlerde harcanan ortanca miktarın benzer olduğunu gösterir. Dış Giyim (Outerwear) kategorisinin medyan satın alma tutarı ise diğerlerine göre biraz daha düşük ve yaklaşık 55 USD civarındadır. Bunların yanı sıra kategorilerde satın alma tutarları genellikle 40-80 USD aralığında yoğunlaşmaktadır. Bu bulgu, kategoriler arasında fiyat algısı ve değer önerisinin nispeten tutarlı olduğunu ve müşterilerin farklı kategorilerde benzer tutarlarda harcama yaptığını göstermektedir.



Grafik, RFM analizine göre oluşturulan müşteri segmentlerinin dağılımını göstermektedir. Segmentler, VIP Müşteriler, Potansiyel Sadık Müşteriler, Sadık Müşteriler, Kaybedilen Müşteriler ve Risk Altındaki Müşteriler olarak sınıflandırılmıştır.

•VIP Müşteriler (VIP Customers): Müşteri sayısının en fazla olduğu segmenttir. Bu durum, işletmenin yüksek değerli ve sık alışveriş yapan geniş bir müşteri tabanına sahip olduğunu göstermektedir. Bu grup işletmenin gelirlerinin büyük bir kısmını oluşturur ve elde tutulmaları öncelikli olmalıdır.

-Özel Ayrıcalıklar ve Sadakat Programları - Özel Müşteri Hizmetleri

•Sadık Müşteriler (Loyal Customers) ve Potansiyel Sadık Müşteriler (Potential Loyalists): Bu iki segment de önemli sayıda müşteriyi barındırmaktadır. Sadık müşteriler, düzenli olarak alışveriş yapan ve işletmeye bağlılık gösteren gruptur. Potansiyel sadık müşteriler ise, doğru stratejilerle sadık müşterilere dönüştürülebilecek potansiyele sahip müşterilerdir.

-Düzenli İletişim ve Değer Odaklı İçerik - Tekrarlayan Satın Almaları Teşvik Etme - Kişiselleştirilmiş Ürün Önerileri

•Risk Altındaki Müşteriler (At Risk Customers): İşletmeyle ilişkisi azalmaya başlayan müşterileri temsil eder. Bu müşterilerin kaybedilmesini önlemek için bu gruba özel hedef kitleli kampanyalar ve yeniden etkileşim stratejileri geliştirilmelidir.

-Yeniden Etkileşim Kampanyaları - Özel İndirimler ve Teşvikler

•Kaybedilen Müşteriler (Lost Customers): En düşük müşteri sayısına sahip segmenttir. Bu müşteriler, uzun süredir alışveriş yapmayan ve geri kazanılması zor olabilecek müşterilerdir. Yine de geri kazanma potansiyeli olanlar için hedefli kampanyalar yapılabilir. Ayrıca en düşük segmentin bu segment olması işletmenin başarılı olduğunun bir kanıtıdır.

-Düşük Maliyetli Geri Kazanım Kampanyaları



Müşteri segmentasyonu analizi, K-means kümeleme algoritması kullanılarak müşterileri benzer özelliklere sahip gruplara ayırmak için gerçekleştirilmiştir. Dağılım grafiği, müşterilerin yaş ve satın alma tutarlarına göre üç farklı kümeye ayrıldığını göstermektedir.

Grafik incelendiğinde, üç belirgin müşteri segmenti ortaya çıkmaktadır:

1. Küme 1 (Kırmızı): Yüksek satın alma tutarına sahip, geniş yaş aralığındaki müşteriler. Bu segment, premium müşteri grubu olarak değerlendirilebilir.

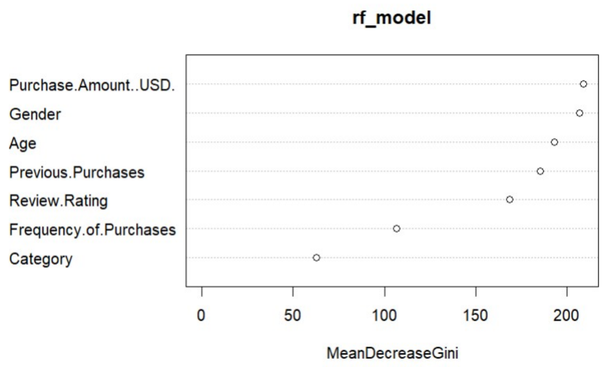
-Değer Odaklı İletişim – Lüks Marka Pazarlaması

2. Küme 2 (Yeşil): Düşük-orta satın alma tutarına sahip, genellikle orta yaş ve üzeri müşteriler. Bu segment, bütçe odaklı müşteri grubu olarak tanımlanabilir.

-Sezonluk indirimler -Toplu alım fırsatları – «İkinci ürüne %X indirim» gibi cazip promosyonlar

3. Küme 3 (Mavi): Değişken satın alma tutarlarına sahip, genellikle genç müşteriler. Bu segment, genç ve dinamik müşteri grubu olarak düşünülebilir.

-Instagram, TikTok gibi sosyal medya platformları - Influencer pazarlaması -Trendlere Uygun Ürünler



Değişken önem derecesi analizi, bir makine öğrenimi modeli olan Random Forest kullanılarak gerçekleştirilmiştir. Bu analiz, modelin tahmin performansına her bir değişkenin ne kadar katkıda bulunduğunu anlamamızı sağlar. Grafikte gösterilen "MeanDecreaseGini" değeri, bir değişkenin modeldeki karar ağaçlarının saflığını ne kadar azalttığını ifade eder. Yüksek "MeanDecreaseGini" değeri, o değişkenin modelin tahmin gücüne daha fazla katkıda bulunduğunu göstermektedir.

Grafiği incelediğimizde, değişkenlerin önem dereceleri aşağıdaki gibidir:

1. Purchase.Amount..USD. (Satın Alma Tutarı - USD): Satın alma tutarının, modelin müşteri davranışlarını tahmin etmede en etkili değişken olduğunu görülmektedir.
2. Gender (Cinsiyet): Cinsiyet odaklı pazarlama stratejilerinin dikkate alınması gerektiğini görülmektedir.
3. Age (Yaş): Farklı yaş gruplarının farklı ihtiyaçları ve satın alma alışkanlıkları olabilmektedir.
4. Previous.Purchases (Önceki Satın Almalar): Bir müşterinin geçmişte yaptığı satın alma sayısı, gelecekteki davranışlarını tahmin etmede önemli bir göstergedir.
5. Review.Rating (İnceleme Puanı): Ürün inceleme puanlarının, müşteri memnuniyeti veya ürünün kalitesi gibi faktörler üzerinde etkili olduğu ve modelin tahminlerine katkıda bulunduğu görülmektedir.
6. Frequency.of.Purchases (Satın Alma Sıklığı): Diğer değişkenlere göre düşük kalsa bile müşterinin ne sıklıkla alışveriş yaptığı, belirli davranışları tahmin etmede hala değerli olabilir.
7. Category (Kategori): En düşük değere sahiptir. Bu modelin tahmin gücüne diğer değişkenlere kıyasla en az katkıda bulunan değişken olduğunu göstermektedir.

SONUÇ

Gerçekleştirilen analizler ile, müşteri alışveriş eğilimlerine dair bulgular sunulmuştur. Kategori bazında satın alma tutarları benzer harcama eğilimlerini gösterirken, RFM ve K-means segmentasyonları müşterileri farklı gruplara ayırarak kişiselleştirilmiş pazarlama stratejilerinin oluşturulmasına ortam sağlamıştır. Değişken önem derecesi analizi ile ise satın alma tutarı, cinsiyet ve yaş gibi faktörlerinin müşteri davranışlarını tahmin etmede en etkili değişkenler olduğu ortaya çıkmıştır. Bu bulgular, işletmelerin veri odaklı yaklaşımlarla müşteri memnuniyetini, sadakatini ve satış performansının arttırılmasına olanak sağlamaktadır.

KAYNAKÇA

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[2] Wedel, M., & Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media.

[3] Kumar, V., & Reinartz, W. (2018). Customer relationship management: Concept, strategy, and tools. Springer.

İSTATİSTİK BÖLÜMÜ

EEG Verileriyle Epileptik Nöbet Tespiti için Makine Öğrenmesi Yöntemleri

Eylül Deniz Hülakü 20023032

Danışman:Prof. Dr. Ali Hakan BÜYÜKLÜ

ÖZET

Bu çalışmada, EEG sinyalleri kullanılarak epileptik nöbetlerin otomatik tespiti amaçlanmıştır. Özellik çıkarımı için DWT ve AR modelleme yöntemleri kullanılarak çeşitli makine öğrenmesi algoritmalarıyla sınıflandırma modelleri oluşturulmuştur. Lojistik Regresyon, SVM, k-NN, ANN ve Naive Bayes algoritmaları test edilmiş; DWT ile elde edilen özniteliklerin AR modeline göre daha başarılı olduğu görülmüştür. DWT öznitelikleriyle ANN modeli %98.7 doğrulukla en iyi sonuçları verirken, CNN + LSTM tabanlı derin öğrenme modeli %97.81 doğruluk sağlamıştır. Sonuçlar, epilepsi tanısında hem geleneksel hem de derin öğrenme yöntemlerinin etkili karar destek araçları olabileceğini göstermektedir.

VERİ HAKKINDA

Kullanılan veri seti, epileptik nöbet tespiti amacıyla EEG (Elektroensefalografi) sinyallerinden oluşturulmuş ve Kaggle üzerinden temin edilmiştir. Veri seti, 500 bireye ait EEG kayıtlarının her biri yaklaşık 23.6 saniye uzunluğunda olup 4097 zaman noktasından oluşmaktadır. İşlem kolaylığı açısından bu kayıtlar, her biri 178 veri noktası içeren 23 parçaya bölünmüş ve toplamda 11.500 satır veri elde edilmiştir.
Her satır yaklaşık 1 saniyelik EEG aktivitesini temsil ederken, 179. sütunda ilgili sınıf etiketi (y) yer almaktadır.

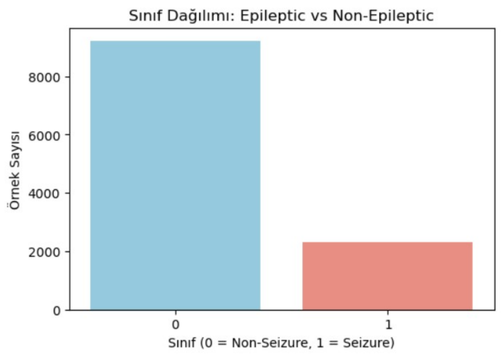
	Unnamed	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	X21.V1.791	135	190	229	223	192	125	55	-9	-33	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
1	X15.V1.924	386	382	356	331	320	315	307	272	244	...	164	150	146	152	157	156	154	143	129	1
2	X8.V1.1	-32	-39	-47	-37	-32	-36	-57	-73	-85	...	57	64	48	19	-12	-30	-35	-35	-36	5
3	X16.V1.60	-105	-101	-96	-92	-89	-95	-102	-100	-87	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	X20.V1.54	-9	-65	-98	-102	-78	-48	-16	0	-21	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

5 rows × 180 columns

Yalnızca sınıf 1 epileptik nöbet durumunu temsil ederken, 2, 3, 4 ve 5 numaralı sınıflar nöbet dışı (sağlıklı veya farklı durumlarda) beyin aktivitelerine karşılık gelir. Bu nedenle veri seti hem çok sınıflı sınıflandırma hem de çoğu çalışmada yapıldığı gibi sınıf 1'e karşı diğer sınıfların birleştirildiği ikili sınıflandırma için uygundur.

VERİ ÖN İŞLEME

Ham EEG verisinde, hedef değişken olan y sütunu beş farklı sınıftan oluşmaktadır. Bu çalışma epileptik nöbet tespiti amacı taşıdığından, çok sınıflı yapı ikili sınıflandırma problemine dönüştürülmüştür. Bu amaçla, sadece "1" etiketli sınıf epileptik nöbeti temsil edecek şekilde tutulmuş; "1" dışındaki tüm sınıflar (2, 3, 4 ve 5) epileptik olmayan durumlar olarak gruplanmış ve "0" olarak yeniden etiketlenmiştir.

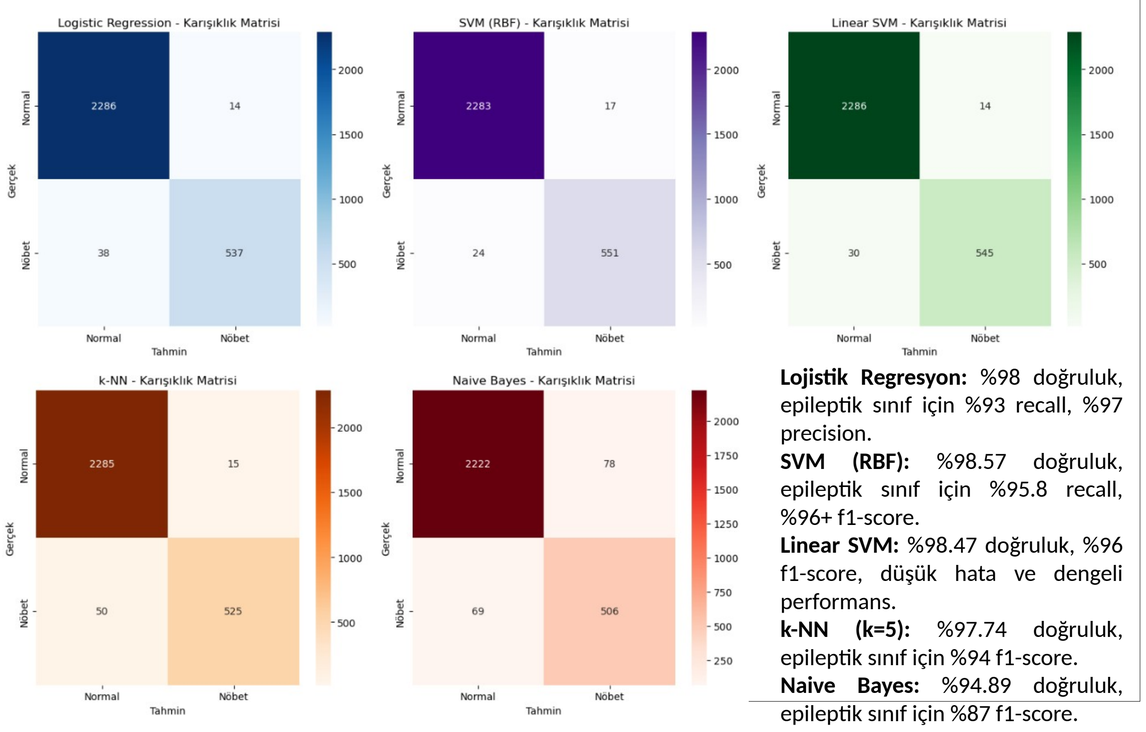


Yeniden etiketleme sonrası, veri kümesinde 9200 adet nöbetsiz (etiket = 0) ve 2300 adet epileptik (etiket = 1) örnek olduğu tespit edilmiştir.

DWT Tabanlı Özellik Çıkarımı ve Sınıflandırma

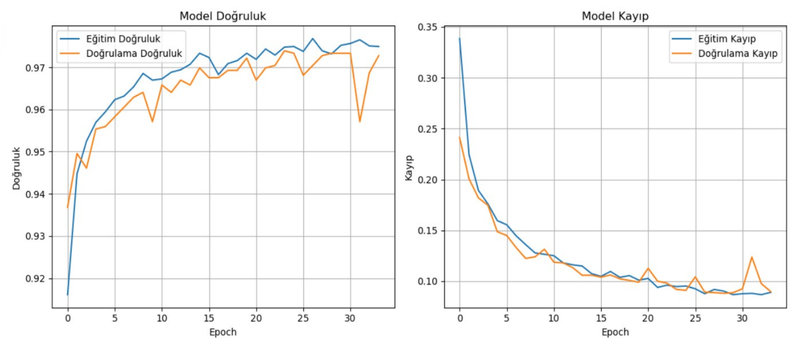
EEG sinyallerinden Discrete Wavelet Transform (DWT) yöntemiyle zenginleştirilmiş öznitelikler çıkarılmıştır. Her sinyal, db4 dalgacığı ile 4 seviyede ayrıştırılarak katsayılar üzerinden ortalama, medyan, standart sapma, maksimum, minimum, enerji, entropi, çarpıklık ve basıklık gibi istatistiksel özellikler hesaplanmıştır. Elde edilen bu öznitelikler sabit uzunlukta bir vektöre dönüştürülerek sınıflandırmaya uygun hale getirilmiştir. Etiketler ikili formata (epileptik = 1, diğer = 0) dönüştürülmüş ve veri kümesi %75 eğitim – %25 test olarak ayrılmıştır. Ardından z-score standardizasyonu uygulanarak tüm öznitelikler normalize edilmiştir. DWT tabanlı özniteliklerle eğitilen sınıflandırma modelleri accuracy, recall ve specificity metrikleri ile değerlendirilmiş ve performans karşılaştırmaları yapılmıştır.

Makine Öğrenmesi Modelleri



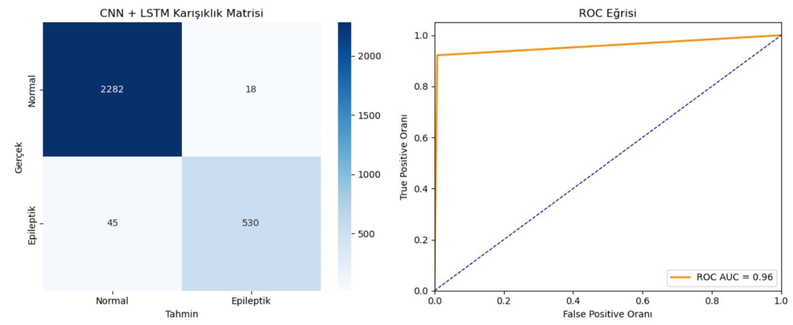
Lojistik Regresyon: %98 doğruluk, epileptik sınıf için %93 recall, %97 precision.
SVM (RBF): %98.57 doğruluk, epileptik sınıf için %95.8 recall, %96+ f1-score.
Linear SVM: %98.47 doğruluk, %96 f1-score, düşük hata ve dengeli performans.
k-NN (k=5): %97.74 doğruluk, epileptik sınıf için %94 f1-score.
Naive Bayes: %94.89 doğruluk, epileptik sınıf için %87 f1-score.

Yapay Sinir Ağı (ANN) Modeli



İlk denemelerde modelin overfitting yaptığı gözlemlenmiştir. Bu sorunu azaltmak amacıyla, yapay sinir ağına %50 ve %30 oranlarında iki Dropout katmanı eklenmiş ve EarlyStopping tekniği uygulanmıştır. Bu stratejiler sayesinde modelin ezber yapma eğilimi azaltılmış, eğitim süresi optimize edilmiş ve %98.7 doğrulama doğruluğu ile yüksek genelleme başarısı elde edilmiştir.

CNN + LSTM Modeli



Test verisi üzerinde %97.81 doğruluk oranı elde etmiştir. Epileptik nöbetleri %92 (recall) oranında doğru tespit edildi. Ek olarak, ROC eğrisi altında kalan alan (AUC) değeri 0.96 olarak hesaplanmıştır.

AR Modellemesi ile Özellik Çıkarımı ve Sınıflandırma

AR modele dayalı öznitelikler çıkarılmıştır. Her sinyal için lag (gecikme) sayısı 6 olarak belirlenmiş ve AutoReg modeli ile geçmiş değer ilişkilerine dayalı katsayılar elde edilmiştir. Bu katsayılar, sinyalin yapısal özelliklerini temsil eden öznitelikler olarak kullanılmıştır. Veri kümesi, %75 eğitim – %25 test oranında train_test_split ile ayrılmış ve sınıf dengesi stratify parametresi ile korunmuştur.

ANN	Accuracy:%93.70		
	True Epilepsy	True Normal	Class Precision
Prediction Epilepsy	470	76	%86.08
Prediction Normal	105	2224	%95.51
Class Recall	%81.01	%96.69	

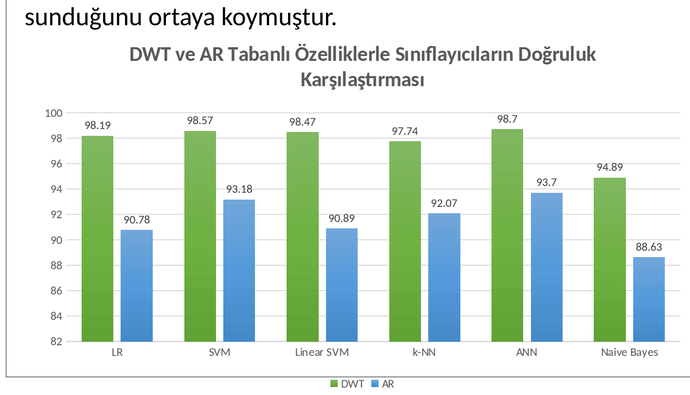
SVM	Accuracy:%93.18		
	True Epilepsy	True Normal	Class Precision
Prediction Epilepsy	446	67	%86.94
Prediction Normal	129	2233	%94.53
Class Recall	%77.56	%97.08	

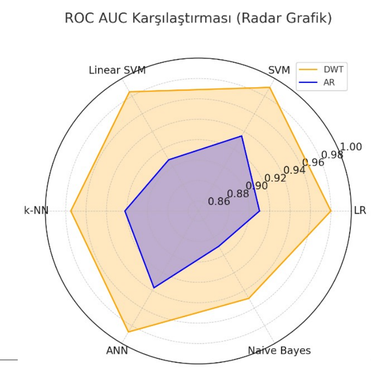
k-NN	Accuracy:%92.07		
	True Epilepsy	True Normal	Class Precision
Prediction Epilepsy	436	89	%83.05
Prediction Normal	139	2211	%94.08
Class Recall	%75.83	%96.13	

En yüksek doğruluk oranı %93.70 ile Yapay Sinir Ağı (ANN) modeline aittir. Bu model, epileptik sınıf için ayrıca %86.08 precision ve %81.01 recall değerleriyle güçlü bir denge sağlamıştır. SVM (RBF) ve k-NN modelleri sırasıyla %93.18 ve %92.07 doğruluk ile etkili sonuçlar vermiştir. LR ve Linear SVM modelleri ise yaklaşık %91 doğruluk sağlamış, ancak epileptik sınıfı tanımada başarıları görece düşük kalmıştır.
Naive Bayes, %88.63 doğruluk ve epileptik sınıfta %70 precision, %75 recall ile en düşük başarıyı göstermiştir.
Genel olarak, AR tabanlı en iyi performans, ANN ve çekirdekli SVM gibi karmaşık yapıları öğrenebilen modellerle elde edilmiştir. Bu sonuç, AR özelliklerinin zamansal bilgi taşımasıyla birlikte güçlü sınıflandırıcılarla etkili bir şekilde değerlendirilebildiğini göstermektedir.

SONUÇ

Sonuçlar, DWT tabanlı modellerin AR tabanlı modellere kıyasla daha yüksek doğruluk sağladığını göstermiştir. Özellikle ANN, DWT ile %98.7 doğruluk elde ederken, AR ile %93.7'de kalmıştır. Benzer şekilde SVM, k-NN ve Lojistik Regresyon da DWT ile daha başarılı sonuçlar vermiştir.
Bu başarı farkı, DWT'nin EEG sinyalini hem zaman hem frekans boyutunda analiz edebilmesine bağlanmaktadır. Öte yandan AR modeli yalnızca zamansal bağımlılıkları temel aldığından sınırlı bilgi sunmaktadır. Alternatif bir yöntem olarak, derin öğrenme tekniklerinden yararlanılan CNN + LSTM tabanlı hibrit bir model de çalışmaya dahil edilmiştir. DWT ile elde edilen öznitelikleriin girdi olarak verildiği bu mimari, test verisi üzerinde %97.81 doğruluk oranına ulaşarak geleneksel yöntemlere güçlü bir alternatif sunduğunu ortaya koymuştur.





KAYNAKÇA

[1] Karlık, B., & Bayrak Hayta, Ş. (2014). Comparison Machine Learning Algorithms for Recognition of Epileptic Seizures in EEG.
[2] Chadwick, D. (1994). Epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*
[3] Niedermeyer, E., & da Silva, F. L. (2004). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (5th ed.).

2024-2025 Bahar Yarıyılı

İSTATİSTİK BÖLÜMÜ

Banka Müşteri Kaybı Analizi ve Tahmini

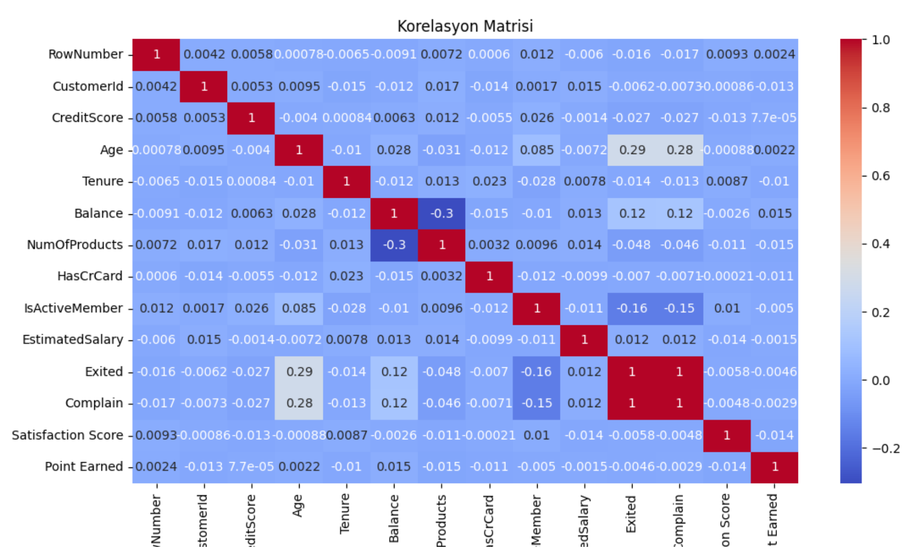
Ezgi BULUT 20023077

Doç. Dr. Atıf Ahmet EVREN

ÖZET

Bu çalışma, banka müşterilerinin ayrılma (churn) davranışlarını analiz etmek ve bu davranışı etkileyen temel faktörleri belirlemek amacıyla gerçekleştirilmiştir. Bankacılık sektöründe müşteri kaybı, gelir kaybına ve rekabet avantajının azalmasına neden olabilmektedir. Bu bağlamda, churn analizleri hem müşteri bağlılığını artırmak hem de hedef odaklı pazarlama stratejileri geliştirmek açısından büyük önem taşımaktadır.
Çalışmada "Customer Churn Records" adlı veri seti kullanılmıştır. Veri setinde yer alan demografik ve finansal bilgiler çeşitli ön işleme adımlarından geçirilmiş, kategorik değişkenler yeniden düzenlenmiş ve bazı sürekli değişkenler kategorilere ayrılarak anlamlı sınıflar oluşturulmuştur. Analitik sürecin ikinci aşamasında, lojistik regresyon algoritması kullanılarak bir sınıflandırma modeli oluşturulmuştur. Modelin performansı doğruluk (accuracy), sınıflandırma raporu, karışıklık matrisi (confusion matrix) ve ROC eğrisi gibi metrikler ile değerlendirilmiştir. Elde edilen sonuçlar, müşteri kaybının en çok etkilendiği faktörlerin yaş, ürün sayısı, aktif müşteri olma durumu ve bölge gibi değişkenler olduğunu göstermektedir.

KORELASYON MATRİSİ



Churn ile Korelasyonlar

Age-Exited : 0.29

Yaş arttıkça müşteri kaybı olasılığı da artıyor. Bu,yaşlı bireylerin bankadan daha çok ayrıldığını gösteriyor olabilir.Bu ilişki orta düzeyde anlamlıdır.

Complain-Exited : 0.28

Şikayet sayısı ile churn arasında güçlü bir pozitif korelasyon var. Yani şikayet eden müşterilerin bankayı terk etme olasılığı artıyor.

IsActiveMember-Exited : -0.16

Negatif ve orta düzeyde anlamlı korelasyon. Aktif müşteriler churn etme ihtimali düşüktür. Banka ile sık etkileşimde bulunan müşteriler daha sadıktır.

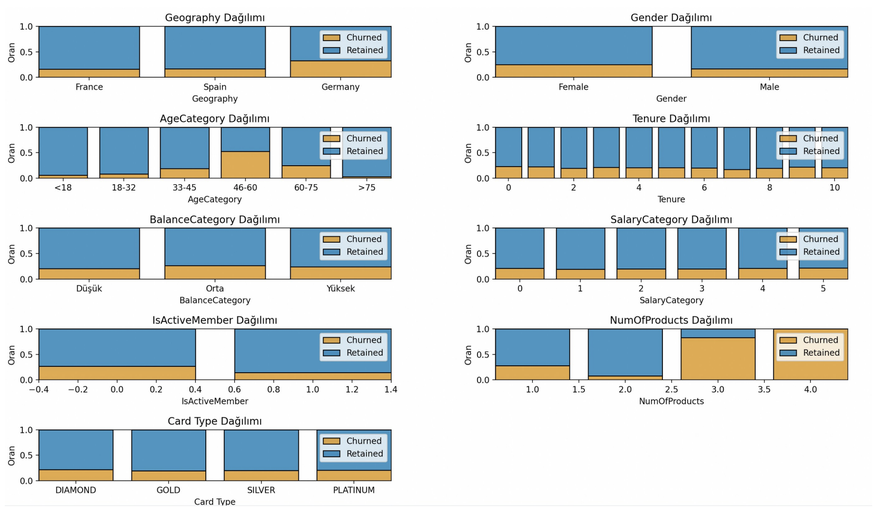
Satisfaction Score-Exited : -0.058

Memnuniyet arttıkça churn etme ihtimali düşüyor. Korelasyon zayıf olsa da beklenen yönde.

Age-NumOfProducts : -0.031

Yaş ilerledikçe kullanılan ürün sayısı azalıyor olabilir.Bu korelasyon zayıftır.

KATEGORİK DEĞİŞKENLERİN CHURN'E ETKİSİ



-Geography Dağılımı (Ülke Bazlı)

Almanya müşterileri arasında churn oranı %35 civarındayken, Fransa ve İspanya'da bu oran sırasıyla %15 ve %12 civarındadır.

Bu durum, Almanya'daki müşterilerin hizmetlerden daha az memnun olduğu ya da rekabetin daha yoğun olduğu anlamına gelebilir.

-Gender Dağılımı

Kadın müşterilerde churn oranı %20, erkek müşterilerde ise %18'dir.

Aradaki fark düşük olsa da, kadın müşterilerde hafifçe daha fazla müşteri kaybı gözlenmektedir.

-AgeCategory Dağılımı

Yaş arttıkça churn oranı da artıyor gibi görünmektedir.

60-75 yaş grubunda churn oranı %40'a kadar çıkarken, 18-32 yaş grubunda bu oran %10 civarındadır.

Bu durum, yaşlı müşterilerin hizmetleri daha kolay terk ettiğini veya hedef kitlenin genç kullanıcılar olduğunu düşündürebilir.

-Tahmini Maaş (SalaryCategory Dağılımı)

Maaş dağılımına göre churn oranları arasında büyük farklar görünmemektedir.

Bu değişkenin etkisi diğerlerine göre daha düşüktür.

-Kart Türü (Card Type Dağılımı)

Kart türleri arasında churn farkı görece daha azdır.

Ancak GOLD ve DIAMOND kart sahiplerinde churn oranı biraz daha yüksektir.

-Genel Sonuçlar-

Almanya, yaşlı müşteriler, aktif olmayan üyeler ve çok ürün kullananlar, churn riski en yüksek gruplardır.

Bankanın bu gruplara özel politikalar, kampanyalar veya destek hizmetleri geliştirmesi churn oranını azaltabilir.

Örneğin:

-Almanya'daki müşteri temsilciliği sistemini gözden geçirmek,

-3 veya daha fazla ürün kullanan müşterilerin deneyimlerini iyileştirmek,

-Yaşlı bireyler için sadeleştirilmiş dijital kanallar sunmak stratejik fayda sağlayabilir.

BAŞARI METRİKLERİ TABLOSU

Başarı Ölçütü	Değer
Doğruluk (Accuracy)	0.81
Hassasiyet (Precision)	0.72
Geri Çağırma (Recall)	0.65
F1 Skoru	0.68
AUC Skoru	0.76

Accuracy (0.81): Modelin genel olarak %81 doğrulukla tahmin yaptığı anlaşılmaktadır. Bu, modelin tüm örnekler üzerindeki genel başarısını yansıtır.

Precision (0.72): Churn eden müşteriler olarak tahmin edilenlerin %72'si gerçekten churn etmiştir. Yani, yanlış pozitiflerin (hatalı churn tahminlerinin) oranı oldukça düşüktür.

Recall (0.65): Gerçekte churn eden müşterilerin %65'i doğru tahmin edilmiştir. Bu oran, modelin churn edenleri ne kadar iyi yakaladığını gösterir.

F1 Skoru (0.68): Precision ve recall dengesini gösterir. Değerin 0.68 olması, modelin hem hassas hem de duyarlı olduğunu göstermektedir.

AUC (0.76): ROC eğrisi altındaki alanın genişliği, modelin iki sınıfı (churn eden / etmeyen) ayırt etme becerisini ifade eder. 0.76 değeri, kabul edilebilir düzeyde iyi bir ayırt edicilik anlamına gelir.

CONFUSİON MATRİX

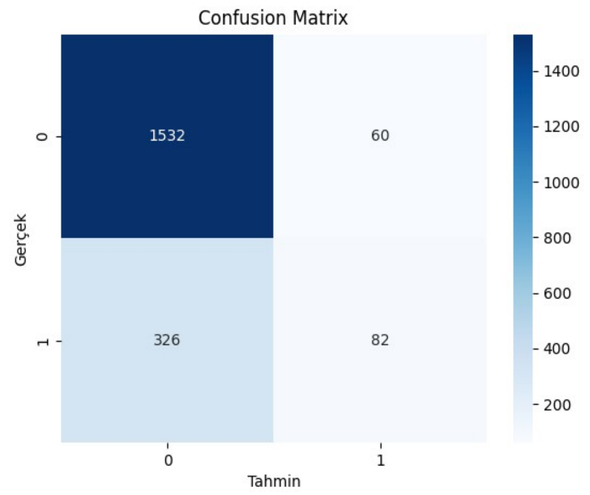
-1532 müşteri doğru şekilde churn etmedi diye tahmin edilmiş.(doğru negatif).

-82 müşteri churn etti ve model doğru bildi. (doğru pozitif).

-60 müşteri churn etmedi ama model yanlışlıkla 'edecek'dedi.(yanlış pozitif).

-326 müşteri churn etti ama model bunu bilemedi.(yanlış negatif).

Doğru negatif oranı çok yüksek;Model,sadık müşterileri doğru şekilde yanıyor.Bu,bankanın güvenli müşteri grubunu iyi ayırt ettiğini gösterir.

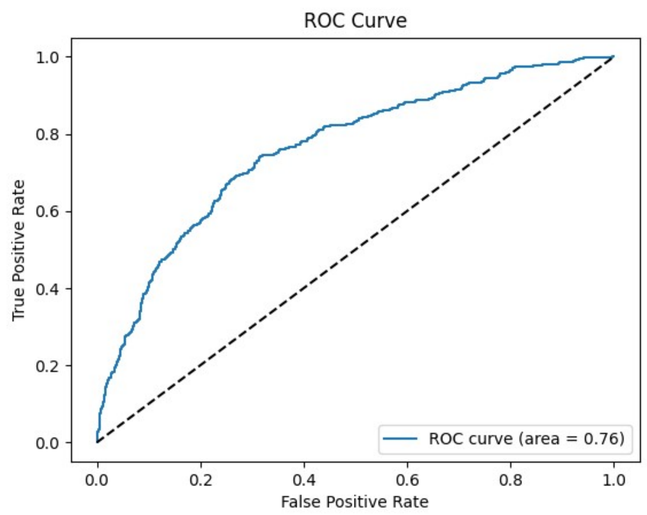


ROC EĞRİSİ

Bu grafik,lojistik regresyon modelinin churn eden ve etmeyen müşterileri ne kadar iyi ayırt ettiğini gösterir.

-Eğri altında kalan alan (AUC = 0.76) modelin %76 oranında ayırt edici başarıya sahip olduğunu göstermektedir.

-Bu değer,modelin pratikte anlamlı sonuçlar verdiğini ve müşteri kaybı tahmini için kullanılabilir olduğunu desteklemektedir.



KAYNAKÇA

[1] Kaggle. (2020). *Bank Customer Churn Prediction Dataset*.

[2] Tuncel, E. (2021). *Python ile Veri Analizi ve Makine Öğrenmesi*. Pusula Yayıncılık.

[3] Veri Bilimi Türkiye. (2023). *Churn Analizi Nedir?*

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

2025 DÜNYA BÖLGESEL MUTLULUK SKORUNUN TAHMİNİ: PANEL VERİ YÖNTEMLERİYLE SABİT ETKİLER MODELİ

Feride Nergis Çelenk 21023050

Doç. Dr. Öyküm Esra Yiğit

Bu çalışmanın amacı 165 ülkenin 2005-2024 yılları arasındaki mutluluk skoru analizini yaparak 2025 yılındaki mutluluk skorlarını bölge bazında tahmin etmektir. Bu doğrultuda GSYİH, sağlıklı yaşam beklentisi, sosyal destek, seçim yapma özgürlüğü, cömertlik, yolsuzluk algısı, pozitif etki, negatif etki ve ulusal hükümete güven bağımsız değişkenlerinin 20 yıllık bir süreçte yıllık olarak toplanan verileri dengeli panel veri formatında düzenlenmiştir. 2025 yılı için tahminleme yapmak için veri setindeki verilen bağımsız değişkenler kullanılarak Havuzlanmış OLS, Sabit Etkiler ve Rasgele Etkiler Modelleri karşılaştırılmıştır. Karşılaştırılan bu 3 modelin testlerinin yapılmasıyla Sabit Etkiler Modeli ile tahminlemenin yapılmasına karar verilmiştir. 2025 yılının mutluluk skoru tahminlemesi yapılırken 2025 yılının bilinmeyen bağımsız değişkenleri ARIMA yöntemi ile tahmin edilmiştir.

KAVRAM VE GİRİŞ

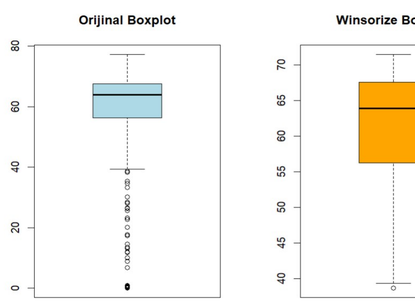
Mutluluk kavramı, psikoloji ve ekonomi gibi farklı disiplinler doğrultusunda araştırmalara konu olmuştur. Toplumların refah seviyelerinin ölçülmesindeki önemli göstergelerden birisi mutluluk endeksidir. Mutluluk endeksini oluşturan 9 değişken vardır: GSYİH, sağlıklı yaşam beklentisi, sosyal destek, seçim yapma özgürlüğü, cömertlik, yolsuzluk algısı, pozitif etki, negatif etki ve ulusal hükümete güven.

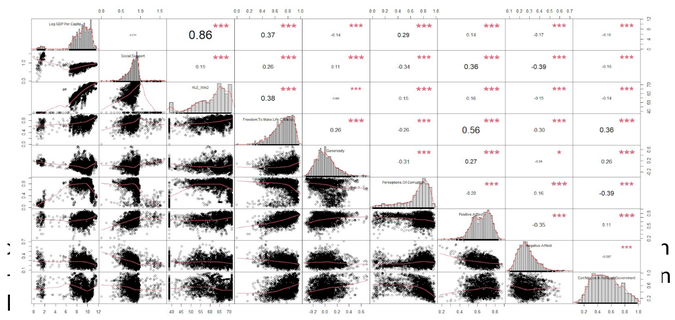
Mutluluk skoru araştırmalarına öncülük eden isimlerden birisi Hollandalı sosyolog Ruut Veenhoven'dır. Veenhoven'un mutluluk skoru çalışması, "World Database of Happiness" adlı çevrimiçi bir veri tabanının oluşturulmasına öncülük etmiştir. Dünyanın farklı bölgelerinden mutluluk üzerine veriler toplayan bu çalışma, mutluluğun aynı zamanda bölgesel olarak da araştırılabileceğini ve dolayısıyla analizler yapılabilir olduğunu göstermiştir.

VERİ SETİ

Bu çalışmada model oluşturmak için kullanılan veri seti için açık kaynak olarak paylaşılan 3 ayrı veri seti 2005-2024 yıllarının verilerini kapsayacak şekilde birleştirildi. Birleştirme işleminden sonra oluşan 1706 NA değer lineer interpolasyon yöntemiyle her ülke kendi içinde değerlendirilerek doldurduldu. 165 ülkenin 2005-2024 yılları arasındaki 9 bağımsız değişken panel_df nesne ismiyle panel veri formatına dönüştürüldü. Panel veri dengelidir. Her ülke için 2005-2024 yılları arasını kapsayan veriler bulunmaktadır.

Veri hazırlandıktan sonra ön tanı adımlarına geçildi. İlk olarak describe() fonksiyonu uygulandı. Bağımsız değişken Healthy.Life.Expectancy.At.Birth için çarpıklık ve basıklık değerleri incelendi ve "Çarpıklık= -2.84" , "Basıklık = 8.64" değerleri elde edildi. Bu sonuçlar doğrultusunda ilgili değişkenin sola çarpık ve kuyruklu (aykırı değer içeren) olduğu yorumu yapılır. Bu nedenle modelleme aşamasından önce winsorize dönüşümü yapılır. %5'in altındaki ve %95'in üstündeki değerler "Winsorize" edilir. Winsorize dönüşümü yapıldıktan sonra ilgili değişkenin uç değerlerinden arındırıldığı aşağıdaki kutu grafiği ile gösterilmiştir.





Yukarıda verilen korelasyon matrisinde GSYİH ve sağlıklı yaşam beklentisi değişkenleri için korelasyon katsayısının 0.86 olduğu görülmektedir. Bu değer ilgili iki değişken arasında yüksek pozitif korelasyon olduğunun göstergesidir. Modelden sırayla sadece biri kullanıldığı durumda da değişkenleri anlamlılığı önemli ölçüde değişmediği ve literatürde bu değişkenler birlikte kullanıldığı için bu çalışmada da iki değişken birlikte modele dahil edildi.

MODELLER

Bu panel veri tahminleme çalışmasının model kurma aşamasında "Havuzlanmış OLS, Sabit Etkiler ve Rasgele Etkiler Modeli"; kurulmuştur ve F testi, Breusch-Pagan LM testi ve Hausman Testi ile sıraysıyla test edilmiştir. Sonuçlar aşağıdadır.

Test	Değer	Karar
(F testi)	F = 28.748, p < 2.2e-16	FE modeli
Breusch-Pagan LM testi	chisq = 6693.6, p < 2.2e-16	RE modeli
Hausman testi	chisq = 257.02, p < 2.2e-16	FE modeli

Veri seti üzerinde yapılan Hausman Testi sonuçlarına bakılarak Sabit Etkiler (FE) Modeli seçildi. Son karar aşamasında yapılan Hausman Testi'nin hipotezleri:

H0(Null Hipotez): Rasgele etkiler modeli ile sabit etkiler modeli ile arasında anlamlı fark yoktur.(Rasgele etkiler modeli tutarlıdır.)

H1(Alternative Hipotez): Rasgele etkiler modeli ile sabit etkiler modeli ile arasında anlamlı fark vardır. (Rasgele etkiler modeli tutarsızdır. Sabit etkiler modeli kullanılmalıdır.)

P-değeri < 2.2e-16 olduğu için H0 reddedilir ve sabit etkiler modeli(FE) kabul edilir.

Sabit etkiler modeli, her bir birimin (ülke, şirket, kişi vb.) kendine özgü sabit bir terim (αi) eklenerek zamanla sabit kalan özelliklerin etkisini ayrıştırır. Ülkeler arasında gözlemlenemeyen farklılıklar olduğunda tercih edilebilir. Sabit etkiler modelinin formülasyonun aşağıdaki gibidir:

$$Yit=\alpha i+\beta 1X1it+\beta 2X2it+\beta 3X3it+\beta 4X4it+\beta 5X5it+\beta 6X6it+\beta 7X7it+\beta 8X8it+\beta 9X9it+uit$$

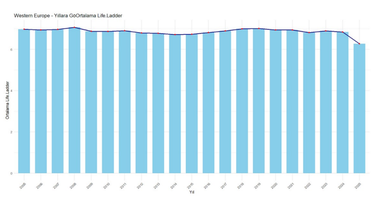
Verilen formülasyondaki "Yit", bağımlı değişken(mutluluk skoru)'dur; "β1Xit+β2Xit+……..+β9Xit" ise bağımsız değişkenlerdir(GSYİH, sağlıklı yaşam beklentisi, sosyal destek, seçim yapma özgürlüğü, cömertlik, yolsuzluk algısı, pozitif etki, negatif etki, ulusal hükümete güven). Sabit etkili model kurulduktan sonra klasik artık tabanlı testler uygulandı. Bu aşamada durağanlık, seri korelasyon ve heteroskedastisite varsayımları kontrol edildi. Model tanı testleri sonucunda heteroskedastisite ve seri korelasyon gözlemlenmiştir. Bu nedenle panel sabit etkiler modeli robust standart hatalar ile raporlanmalıdır.

Katsayılar robust standart hatalar ile tekrar analiz edilmiştir. Anlamlı çıkan bağımsız değişkenler son aşama olan tahmin aşamasında yorumlanırken dikkate alınacaktır. Tüm değişkenler tahminleme aşamasında da modele dahil edilmiştir.

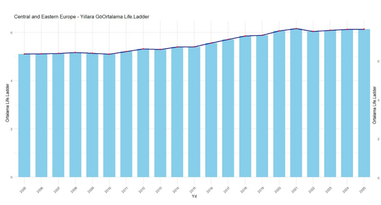
Sabit etkiler modelinin robust standart hatalardan elde edilen sonuçlarının yorumu şu şekildedir:

En etkisiz değişken ulusal hükümete güven, etkili değişken sosyal destek olmuştur.

İlgili sabit etkiler modeli kullanılarak mutluluk skoru tahminleme aşamasına geçildiğinde 2025 yılına ait X bağımsız değişkenleri ARIMA yöntemi ile doldurulmuş ve 2025 yılına ait bağımlı değişken olan mutluluk skoru her ülke için ayrı olarak tahmin edilmiştir. Tahmin aşamasında 2024 yılına ait veriler test verisi olarak seçilmiştir. 2005-2023 yılları arasındaki bağımsız değişkenler 2025 yılı tahminleri için eğitim verisi olarak ayrılmıştır. Ülke bazında tahmin edilen mutluluk skorlarını bölgesel incelemek için aynı bölge ismine sahip ülkeler gruplandırılarak ortalaması bulundu. Tahmin sonuçlarına istinaden en mutlu ilk 3 bölgenin grafikleri yukarıda sağdadır.



2025 Tahmini En Mutlu 1.Bölge



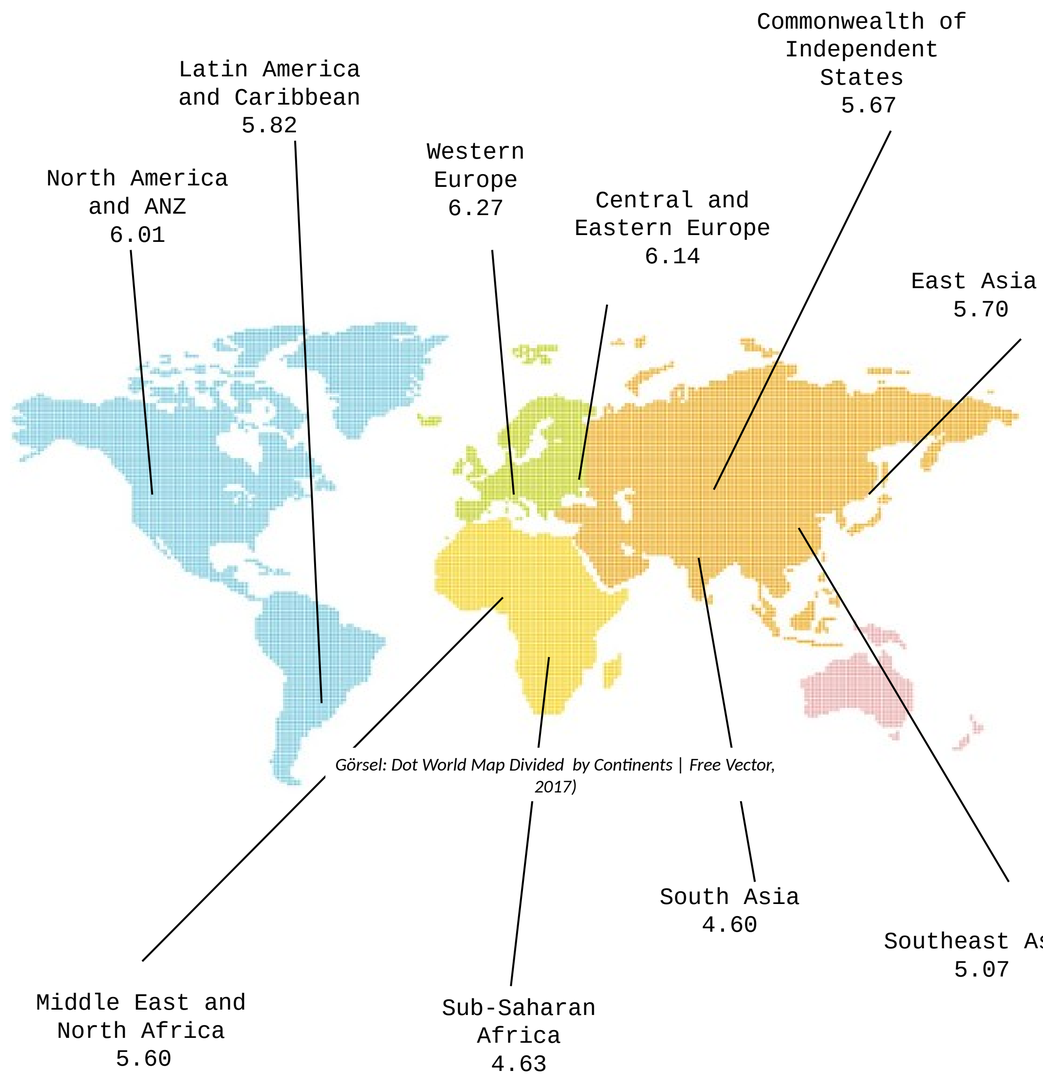
2025 Tahmini En Mutlu 2.Bölge



2025 Tahmini En Mutlu 3.Bölge

Bu tahmin sonuçlarının güvenirliği MAE, RMSE ve MAPE hata ölçütleriyle incelendi. MAPE değeri sabit etkiler modelinin gerçek değerleri ile tahminleri arasındaki sapmayı gösterir. Literatürde tek bir doğru sınır olmamakla birlikte MAPE %20-25% arası baremde olması modelin sapmasını makul olarak kabul eder. Model için hata ölçütleri hesaplandığında sonuçlar şu şekildedir: MAE , 1.282503 ; RMSE , 1.602927 ; MAPE, 23.741607 . Model 2024 yılı için mutluluk skoru tahminini makul bir seviye olarak değerlendirilen %23.7 sapma ile yapmıştır. Modelin güvenirliği test edildikten sonra bölge ortalamaları alınarak sonuca ulaşıldı.

BÖLGE BAZINDA 2025 YILI MUTLULUK SKORU TAHMİNLERİ



Görsel: Dot World Map Divided by Continents | Free Vector, 2017)

(Görsel üzerinde işaretlenen bölgeler %100 kesinlik ifade etmez. Görsel olarak tahmini işaretlemelerdir.)

SONUÇ

Yukarıdaki dünya haritası üzerinde gösterildiği üzere 2025 yılı için bölgesel mutluluk skoru tahmini en yüksek olan bölge 6.27 ile Batı Avrupa'dır. 2. ve 3. sırada; 6.14 ile Orta ve Doğu Avrupa, 6.01 ile Kuzey Amerika ve ANZ bölgesi yer almaktadır. 2025 yılı tahminine göre 10 bölgeden yalnızca 4'ü için görsel olarak belirgin bir şekilde kendi 2024 yılı bölge ortalamasının üzerinde bir skor beklendiği tahmin edilmiştir.

(Orta Doğu ve Avrupa Bölgesi 2024 yılı gerçekleşen ile 2025 tahmini skorları arasındaki 0.01 olduğundan dolayı belirgin artış olarak adlandırılmadı)

KAYNAKÇA

[1]https://www.kaggle.com/datasets/usamabuttar/world-happiness-report-2005-present?resource=download

[2] https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2023

[3] https://www.kaggle.com/datasets/jainaru/world-happiness-report-2024-yearly-updated

[4]*10.3 Fixed Effects Regression*. (n.d.). Introduction to Econometrics with R. Retrieved June 17, 2025, from https://www.econometrics-with-r.org/10.3-fixed-effects-regression.html

[5] https://link.springer.com/article/10.1007/s11482-013-9260-8?utm_source=chatgpt.com

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

MAKROEKONOMİK DEĞERLERİN TÜRKİYE'DEKİ TURİZME ETKİSİ

Furkan ÇAKAR 20023033

Danışman: Doç. Dr. Atıf Ahmet EVREN

ÖZET

Bu çalışmada, 2010-2025 yılları arasında Türkiye'ye gelen bölgesel turist sayıları ile temel makroekonomik göstergeler (faiz oranları, GSYİH, enflasyon ve işsizlik oranı) arasındaki ilişkiler incelenmiştir. Avrupa ve Asya bölgelerinden gelen turist sayılarının, makroekonomik göstergelere olan tepkileri bölgesel farklılıklar göstermektedir. Avrupa'da işsizlik oranı, Asya'da ise enflasyon ve işsizlik oranı turizm talebini anlamlı şekilde etkilemiştir. Çoklu regresyon analizleri ve korelasyon analizleri ile elde edilen bulgular, Türkiye'nin turizm sektöründe sürdürülebilir büyüme sağlanabilmesi için bölgesel bazda makroekonomik politikaların geliştirilmesi gerektiğini ortaya koymaktadır.

Çalışmanın Amacı ve Kapsamı

Bu çalışma, Türkiye'nin turizm sektörüne yönelik makroekonomik göstergelerin etkilerini incelemeyi amaçlamaktadır. Özellikle 2010-2025 yılları arasında Türkiye'ye gelen bölgesel turist sayılarının, temel makroekonomik değişkenlerle olan ilişkisi ele alınmıştır. Çalışmanın temel amacı, Avrupa ve Asya bölgelerinden gelen turist sayıları ile faiz oranları, gayri safi yurtiçi hasıla (GSYİH), enflasyon ve işsizlik oranı gibi makroekonomik göstergeler arasındaki etkileşimi araştırmaktır. Bu doğrultuda, Türkiye İstatistik Kurumu (TÜİK) ve diğer resmi kaynaklardan alınan aylık zaman serisi verileri kullanılmıştır.

Çalışma kapsamında, farklı coğrafi bölgelerden gelen turist sayıları ile ekonomik göstergeler arasındaki ilişki çoklu regresyon analizi ile modellenmiştir. Avrupa ve Asya bölgelerinden gelen turist sayıları üzerinden yapılan karşılaştırmalı analizler, bölgesel dinamiklerin ve ekonomik faktörlerin turizm talebini nasıl şekillendirdiğini ortaya koymaktadır. Ayrıca, makroekonomik değişkenlerin turizm sektörüne olan etkilerinin daha iyi anlaşılabilmesi için zaman serisi analizleri ve korelasyon testleri yapılmıştır.

Sonuç olarak, bu çalışma, Türkiye'nin turizm sektörü için politika yapıcılar ve sektör temsilcileri için önemli çıkarımlar sağlamayı hedeflemektedir. Makroekonomik göstergelerin etkilerinin bölgesel farklılıklar gösterdiği ve ekonomik stratejilerin bu farklara göre şekillendirilmesi gerektiği vurgulanmaktadır.

Makroekonomik Göstergeler ve Turizm

Makroekonomik göstergeler, bir ülkenin ekonomik sağlığını gösteren ve turizm sektörünü etkileyen temel verilerdir. Faiz oranları, enflasyon, işsizlik ve GSYİH gibi faktörler, turistlerin harcama kararları ve sektörün büyüme potansiyeli üzerinde doğrudan etkiye sahiptir. Örneğin, yüksek faiz oranları ve enflasyon, turist harcamalarını kısıtlayabilirken, işsizlik oranlarındaki artış ekonomik belirsizliği artırarak turizm talebini olumsuz yönde etkileyebilir. Diğer yandan, GSYİH'nın büyümesi turizm talebinde artışa yol açabilir. Bu göstergeler, turizm sektörü için önemli bir yönlendirici olmuştur.

Veri Seti ve Kullanılabilir Yöntemler

Bu çalışmada kullanılan veri seti, Türkiye'ye gelen turist sayıları ve makroekonomik göstergelere ait 2010-2025 yılları arasındaki aylık zaman serisi verilerini içermektedir. Veriler, Birleşmiş Milletler Dünya Turizm Örgütü (UNWTO), Dünya Seyahat ve Turizm Konseyi (WTTC), Türkiye İstatistik Kurumu (TÜİK) ve Türkiye Cumhuriyet Merkez Bankası (TCMB) gibi güvenilir resmi kurumlardan temin edilmiştir. Veri seti, özellikle Avrupa ve Asya bölgelerinden gelen turist sayıları ile faiz oranları, gayri safi yurtiçi hasıla (GSYİH), enflasyon ve işsizlik oranı gibi temel makroekonomik göstergeleri kapsamaktadır.

Veri ön işleme sürecinde, eksik veriler doğrusal interpolasyon yöntemi ile tamamlanmış, aykırı değerler z-skoru analiziyle tespit edilip düzeltilmiştir. Ayrıca, zaman serisi analizlerinin güvenilirliğini artırmak amacıyla Augmented Dickey-Fuller (ADF) testi kullanılarak durağanlık testleri yapılmış ve gerekirse veriler logaritmik dönüşüm ve fark alma gibi yöntemlerle stabilize edilmiştir.

Analizlerde kullanılan yöntemler arasında çoklu regresyon analizi, korelasyon matrisi ve zaman serisi analizleri yer almaktadır. Bu yöntemler, makroekonomik göstergelerin turizm sektörü üzerindeki etkilerini anlamak ve bölgesel farklılıkları değerlendirmek amacıyla kullanılmıştır. Ayrıca, değişkenler arasındaki çoklu bağlantı (multicollinearity) riski göz önünde bulundurularak Varyans Enflasyon Faktörü (VIF) testi uygulanmıştır.

Çoklu Regresyon Analizleri

Bu çalışmada, Türkiye'ye gelen turist sayıları ile temel makroekonomik göstergeler arasındaki ilişkiyi incelemek için çoklu regresyon analizi kullanılmıştır. Bağımsız değişkenler olarak **faiz oranı**, **gayri safi yurtiçi hasıla (GSYİH)**, **enflasyon** ve **işsizlik oranı** yer alırken, bağımlı değişken olarak **bölgesel turist sayıları** kullanılmıştır. Analiz, **Avrupa** ve **Asya** bölgelerinden gelen turist sayılarının, makroekonomik göstergelere olan etkilerini ölçmek amacıyla yapılmıştır.

Modeldeki sonuçlar, Avrupa ve Asya bölgelerinin farklı ekonomik dinamiklere sahip olduğunu ve makroekonomik göstergelere farklı tepkiler verdiğini göstermektedir. Avrupa bölgesinde **işsizlik oranı**, turist sayıları üzerinde **pozitif ve anlamlı** bir etkiye sahipken, **faiz oranı**, **enflasyon** ve **GSYİH** anlamlı bir etki yaratmamıştır. Asya bölgesinde ise **enflasyon** ve **işsizlik oranı**, turist sayıları üzerinde **pozitif** etkiler yaratırken, **GSYİH** ile **negatif** bir ilişki gözlemlenmiştir.

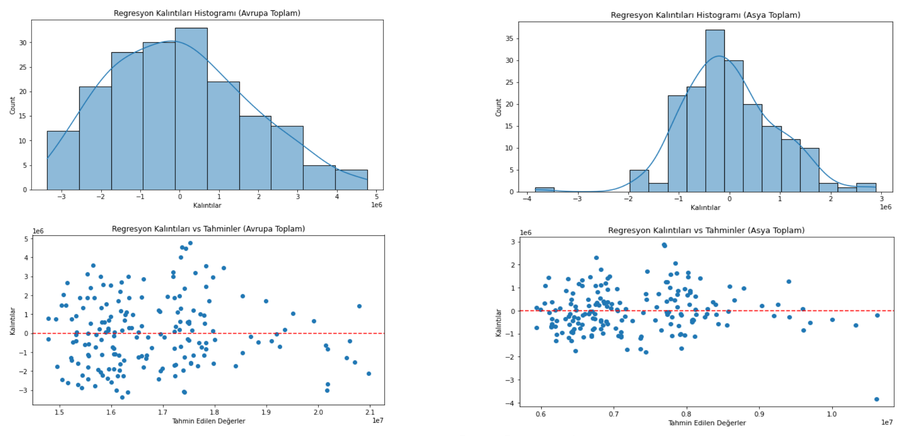
```
--- Avrupa Toplam için Çoklu Regresyon Sonuçları ---
                            OLS Regression Results
==============================================================================
Dep. Variable:          Avrupa Toplam   R-squared:                       0.365
Model:                            OLS   Adj. R-squared:                  0.350
Method:                 Least Squares   F-statistic:                     25.54
Date:                Mon, 26 May 2025   Prob (F-statistic):           9.86e-17
Time:                        22:26:56   Log-Likelihood:                -2893.1
No. Observations:                 183   AIC:                             5796.
Df Residuals:                     178   BIC:                             5812.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.192e+07   9.07e+05     13.144      0.000    1.01e+07    1.37e+07
interest_rate  -267.0490   2.71e+04     -0.010      0.992   -5.37e+04    5.31e+04
gdp               0.0005      0.001      0.481      0.631      -0.002       0.003
inflation      1227.4541   1762.230      0.697      0.487   -2250.096    4705.005
unemployment   3.586e+05   8.18e+04      4.384      0.000    1.97e+05    5.2e+05
==============================================================================
Omnibus:                        5.856   Durbin-Watson:                   1.201
Prob(Omnibus):                  0.054   Jarque-Bera (JB):                5.564
Skew:                           0.373   Prob(JB):                       0.0619
Kurtosis:                       2.583   Cond. No.                     8.92e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.92e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
Durbin-Watson Test İstatistiği: 1.2005426920458357
```

```
--- Asya Toplam için Çoklu Regresyon Sonuçları ---
                            OLS Regression Results
==============================================================================
Dep. Variable:          Asya Toplam   R-squared:                       0.516
Model:                            OLS   Adj. R-squared:                  0.505
Method:                 Least Squares   F-statistic:                     47.42
Date:                Mon, 26 May 2025   Prob (F-statistic):           4.30e-27
Time:                        22:26:56   Log-Likelihood:                -2773.4
No. Observations:                 183   AIC:                             5557.
Df Residuals:                     178   BIC:                             5573.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.874e+06   4.71e+05      8.217      0.000    2.94e+06    4.8e+06
interest_rate -1.324e+04   1.41e+04     -0.942      0.348    -4.1e+04    1.45e+04
gdp              -0.0021      0.001     -3.623      0.000      -0.003      -0.001
inflation      4968.7552    915.958      5.425      0.000    3161.220    6776.290
unemployment   2.068e+05   4.25e+04      4.863      0.000    1.23e+05    2.91e+05
==============================================================================
Omnibus:                        8.336   Durbin-Watson:                   1.157
Prob(Omnibus):                  0.015   Jarque-Bera (JB):               14.722
Skew:                           0.142   Prob(JB):                     0.000635
Kurtosis:                       4.360   Cond. No.                     8.92e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.92e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
Durbin-Watson Test İstatistiği: 1.1574036680876845
```

Çoklu Regresyon Analiz Sonuçları



Regresyon Kalıntıları Histogramı

Sonuçlar

Bu çalışma, Türkiye'de 2010-2025 döneminde aylık bazda toplanan bölgesel turist sayıları ile makroekonomik göstergeler arasındaki ilişkileri kapsamlı bir şekilde incelemiştir. Avrupa ve Asya bölgelerinden gelen turist sayıları, faiz oranı, GSYİH, enflasyon ve işsizlik gibi temel makroekonomik değişkenler kullanılarak çoklu regresyon modelleri oluşturulmuştur.

Analiz sonuçları, makroekonomik göstergelerin turizm performansı üzerinde bölgesel farklılıklar içermekle birlikte anlamlı etkiler yarattığını göstermektedir. Avrupa bölgesinde işsizlik oranı turizm talebini anlamlı biçimde etkilerken, faiz oranı, enflasyon ve GSYİH anlamlı bulunmamıştır. Buna karşın, Asya bölgesinde enflasyon ve işsizlik pozitif, GSYİH ise negatif etkiler göstermiştir. Faiz oranları ise her iki bölgede de anlamlı bir etkiye sahip olmamıştır.

Verilerin durağanlık analizleri, makroekonomik değişkenlerin çoğunun durağan olduğunu; turist sayılarının ve işsizlik oranının ise durağan olmadığını ortaya koymuştur. Bu nedenle analizlerde gerekli dönüşümler yapılmış ve model varsayımlarının sağlanmasına özen gösterilmiştir.

Korelasyon analizi, faiz oranı, GSYİH ve enflasyon arasında yüksek pozitif korelasyon olduğunu, işsizlik ile diğer değişkenler arasında ise zayıf negatif korelasyon bulunduğunu ortaya koymuştur. Model kalıntılarının normal dağılıma yakın olması ve hata terimlerinin homoskedastik olması, analizlerin güvenilirliğini artırmaktadır.

KAYNAKÇA

[1] United Nations World Tourism Organization (UNWTO). (2020). International Tourism Highlights, 2020 Edition. https://www.e-unwto.org/doi/book/10.18111/9789284422456

[2] Türkiye İstatistik Kurumu (TÜİK). (2020). *Turizm İstatistikleri*. https://data.tuik.gov.tr/Bulten/Index?p=Turizm-Istatistikleri-2020-37405

[3] Türkiye Cumhuriyet Merkez Bankası (TCMB). (2023). *Ekonomik Veriler Raporu*. https://www.tcmb.gov.tr/wps/wcm/connect/tr/tcmb+tr/main+page+site+area/raporlar

2024-2025 BAHAR Yarıyılı İSTATİSTİK BÖLÜMÜ

İstatistik Bölümü Öğrencilerinin Yapay Zekaya Yönelik Kariyer Algısı: Lojistik Regresyon ve Hipotez Testleriyle Analiz

Gökalp ŞAHİN Öğrenci No:21023012

Danışman: Dr.Öğr.Üyesi Doğan YILDIZ

ÖZET

Bu çalışmada, Ankara'daki devlet üniversitelerinde öğrenim gören istatistik bölümü öğrencilerinin yapay zekâya yönelik kariyer algıları incelenmiştir. Veriler, öğrencilerin anket yanıtları doğrultusunda geliştirilen ve 0–100 aralığında skor üreten AI_Career_Impact_Score ile sayısallaştırılmıştır. Elde edilen skorlar, lojistik regresyon analizi ile incelenmiş; cinsiyet, üniversite, sınıf düzeyi, not ortalaması, bilgi düzeyi ve staj deneyimi gibi değişkenlerin etkileri test edilmiştir. Ayrıca, Tek Yönlü Varyans Analizi (ANOVA), Kruskal-Wallis Testi ve Bağımsız İki Örneklem t-Testi kullanılarak belirli değişken grupları arasında algı farklılıkları değerlendirilmiştir.

Veri Setinin Detayları

.

	Gender	Univ	Year	GPA	Intern	AI_Knowledge	AI_CareerImpact	AI_JobImpact	AI_Concern	AI_Anxiety	AI_EduPrepared	AI_StatsChoice	AI_InternshipIssue	AI_Career_Impact_Score	AI_Career_Positive_Impact
1	Female	Hacettepe University	4	2.01 - 2.50	Yes	High knowledge	Yes	Positive impact	Not concerned	No anxiety	Partially agree	No influence	Partially	82	1
2	Male	Hacettepe University	4	2.01 - 2.50	Yes	Moderate knowledge	Yes	Positive impact	Moderately concerned	No anxiety	Agree	No influence	Yes	80	1
3	Female	Hacettepe University	4	2.51 - 3.00	Yes	Little knowledge	Yes	No impact	Not concerned	No anxiety	Disagree	No influence	No	70	1
4	Female	Hacettepe University	4	3.01 - 3.50	Yes	Little knowledge	Yes	Positive impact	Moderately concerned	Slight anxiety	Disagree	No influence	Partially	65	1
5	Male	Hacettepe University	4+	2.51 - 3.00	Yes	High knowledge	No	Positive impact	Not concerned	No anxiety	Disagree	High influence	No	70	1
6	Male	Hacettepe University	4	2.51 - 3.00	No	Little knowledge	No	Negative impact	Slightly concerned	High anxiety	Partially agree	No influence	No	17	0
7	Male	Hacettepe University	4	2.01 - 2.50	No	Moderate knowledge	Yes	Positive impact	Not concerned	No anxiety	Partially agree	No influence	No	87	1
8	Female	Gazi University	1	2.01 - 2.50	No	No knowledge	Yes	No impact	Slightly concerned	Slight anxiety	Disagree	Slight influence	No	50	1
9	Female	Hacettepe University	4	3.01 - 3.50	Yes	Moderate knowledge	No	Positive impact	Not concerned	No anxiety	Disagree	No influence	Partially	55	1
10	Female	Ankara University	3	2.51 - 3.00	Yes	Moderate knowledge	Yes	No impact	Moderately concerned	High anxiety	Disagree	Moderate influence	Yes	40	0

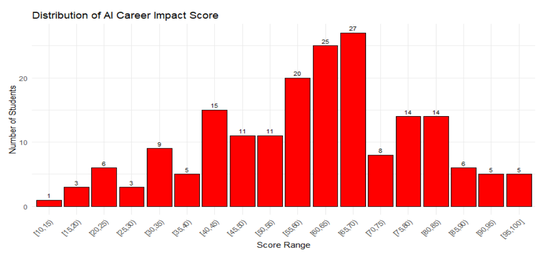
Bu çalışmada kullanılan veri seti, Ankara'daki devlet üniversitelerinde öğrenim gören istatistik bölümü öğrencilerine ait çeşitli demografik ve akademik bilgileri içermektedir. Veri setinde yer alan değişkenler şunlardır:

Gender (Cinsiyet), Univ (Üniversite), Year (Sınıf), GPA (Not Ortalaması), Intern (Staj Durumu)
AI_Knowledge (Yapay Zeka Bilgi Düzeyi), AI_CareerImpact (Kariyer Planına Etki)
AI_JobImpact (İş Fırsatlarına Etki), AI_Concern (Kaygı Düzeyi), AI_Anxiety (Anksiyete Düzeyi)
AI_EduPrepared (Eğitim Hazırlık Algısı), AI_StatsChoice (Bölüm Seçiminde AI Etkisi)
AI_InternshipIssue (Staj Mülakatında Zorluk)

AI_Carrer_Impact_Score Değişkeninin Oluşturulması

- Veriler, öğrencilerin yapay zekaya dair algı ve deneyimlerini değerlendirmeye yönelik hazırlanmıştır. AI_Career_Impact_Score, öğrencilerin yapay zekaya yönelik algılarını ölçmek için 7 sorudan oluşan anket yanıtlarına puan verilerek hesaplanmıştır. Her soruya verilen yanıta göre 0–20 arası puanlar atanmış ve toplam skor 0 ile 100 arasında olacak şekilde belirlenmiştir.

Histogram grafiğinde öğrencilerin AI_Career_Impact_Score değerleri ağırlıklı olarak 50–70 aralığında toplanmakta bu da çoğunluğun yapay zekaykariyer açısından olumlu değerlendirdiğini göstermektedir. Skorlar uç değerlerde daha seyrek olup algı genel olarak orta-yüksek düzeyde yoğunlaşıyor.



Hipotez Testleri

Çalışmada sonuçları güçlendirmek için hipotez testleri ve lojisitik regresyon analizi yapılmıştır.Bu kısımda üç tane hipotez kurup bu hipotezleri test ederek bazı değişkenlerin nasıl etkilendiğini anlamak amaçlanmıştır.

Hipotez 1: Üniversiteye Göre AI Algısı
Araştırılan: Öğrencilerin yapay zekaya yönelik kariyer algısı öğrenim gördükleri üniversiteye göre farklılık gösteriyor mu?
H0: Üniversiteler arasında AI_Career_Impact_Score ortalamaları arasında fark yoktur.
H1: En az bir üniversitenin ortalaması diğerlerinden farklıdır.
Test: Tek Yönlü Varyans Analizi (**One-Way ANOVA**)

```
> summary(anova_model)
             Df Sum Sq Mean Sq F value Pr(>F)
Univ          3   1740   579.9   1.617  0.187
Residuals   184  65973   358.5
```

Hipotez 2: AI Bilgi Düzeyine Göre AI Algısı
Araştırılan: AI bilgi düzeyine göre öğrencilerin AI kariyer algısı farklılık gösteriyor mu?
H0: AI bilgi düzeyine göre AI_Career_Impact_Score ortalamaları arasında fark yoktur.
H1: Bilgi düzeyine göre en az bir grup diğerlerinden farklıdır.
Test: Kruskal-Wallis H Testi

```
        Kruskal-Wallis rank sum test

data:  AI_Career_Impact_Score by AI and Automation Knowledge Level
Kruskal-Wallis chi-squared = 3.2946, df = 3, p-value = 0.3484
```

Hipotez 3: Staj Deneyimine Göre AI Algısı
Araştırılan: Staj yapmış olmak, AI algısını etkiliyor mu?
H0: Staj yapan ve yapmayan öğrencilerin AI_Career_Impact_Score ortalamaları arasında fark yoktur.
H1: Staj yapma durumuna göre ortalamalar arasında fark vardır.
Test: Bağımsız İki Örneklem t-Testi

```
        Two Sample t-test

data:  AI_Career_Impact_Score by Internship Experience
t = 0.0093094, df = 186, p-value = 0.9926
alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
95 percent confidence interval:
 -5.686454  5.740376
sample estimates:
 mean in group No mean in group Yes
         58.54167          58.51471
```

- Üniversite, bilgi düzeyi ve staj durumu gibi faktörlerin öğrencilerin yapay zekaya yönelik kariyer algısı üzerinde istatistiksel olarak anlamlı bir etkisi bulunmamıştır. Elde edilen p-değerleri, bu değişkenler açısından **gruplar arasında anlamlı bir farklılık olmadığını göstermektedir.**

Lojistik Regresyon Analizi Modellemesi

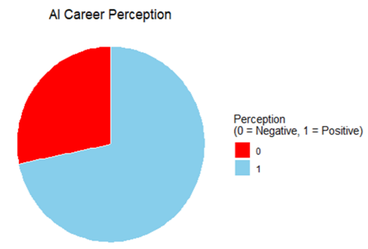
Bu çalışmada, öğrencilerin yapay zekaya yönelik kariyer algılarını belirleyen faktörleri ortaya koymak amacıyla lojistik regresyon analizi kullanılmıştır. Bu yöntem, bağımlı değişkenin ikili (0 veya 1) olduğu durumlarda tercih edilen güçlü bir istatistiksel modelleme tekniğidir.
Çalışmada kullanılan bağımlı değişken, öğrencilerin yapay zekanın kariyerlerine etkisine yönelik genel algısını yansıtan ve AI_Career_Impact_Score üzerinden türetilen AI_Career_Positive_Impact değişkenidir. Bu değişken:
0 = Yapay zekayı olumsuz değerlendirenler
1 = Yapay zekayı olumlu değerlendirenler şeklinde sınıflandırılmıştır.

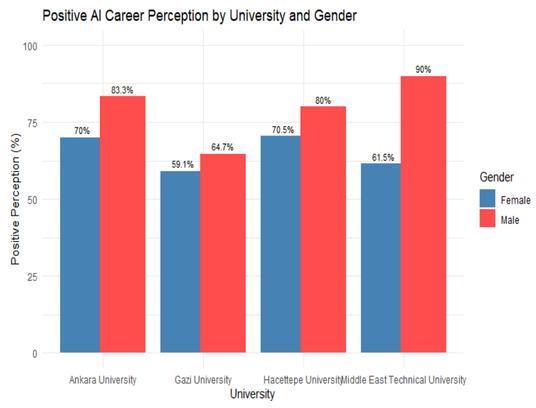
- **Bağımsız değişkenler (açıklayıcı faktörler)** ise şunlardır: Cinsiyet (Kadın/Erkek),Üniversite (öğrencinin bağlı olduğu devlet üniversitesi), Sınıf düzeyi (1. – 4+ sınıf arası),Not ortalaması (GPA),Yapay zeka ve otomasyon bilgi düzeyi (düşük/orta/yüksek), Staj deneyimi (var/yok)

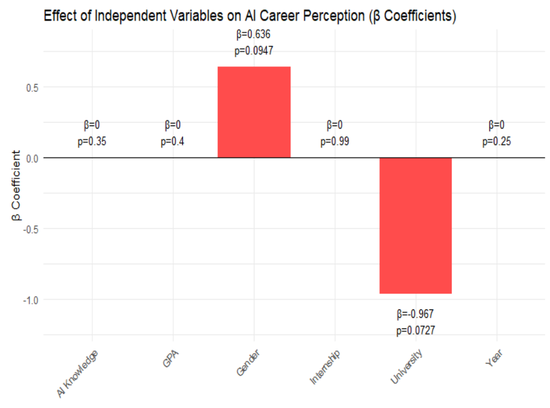
- Bağımlı değişkenin dağılımı:

Sonuç 1 - 135 öğrenci (%71.8) yapay zekayı **olumlu**
Sonuç 0- 53 öğrenci (%28.2) yapay zekayı **olumsuz**

olarak algılamıştır.Bu sonuç, öğrencilerin büyük çoğunluğununYapay zekanın ilerde kariyerlerine olumluEtkileri olacağını düşündüklerini belirtiyor.







- Anlamlı Değişkenler:

Cinsiyet (β = 0.636, p = 0.0947): Erkek öğrenciler, kadınlara kıyasla yapay zekayı kariyer açısından daha olumlu değerlendirmektedir. Bu durum, erkek öğrencilerin teknolojiyle daha sık etkileşimde bulunmaları ya da teknoloji temelli meslekleri daha fazla hedeflemeleriyle ilişkilendirilebilir.
Üniversite (β = -0.967, p = 0.0727): Gazi Üniversitesi öğrencileri, diğer üniversitelere göre yapay zekaya karşı daha temkinli veya olumsuz bir tutum sergilemektedir. Bu fark, üniversiteler arası müfredat yapıları, teknolojiye erişim imkanları veya akademik yönlendirmelerdeki farklılıklardan kaynaklanabilir.

- Anlamsız Değişkenler:

Sınıf düzeyi, GPA, AI bilgi düzeyi ve staj deneyimi değişkenleri istatistiksel olarak anlamlı bulunmamıştır (p > 0.1). Bu değişkenlerin etkisiz kalması, öğrencilerin yapay zekaya dair kariyer algılarının daha çok kişisel değerler, medya etkisi ve genel teknoloji kültürüne olan duyarlılıkla şekillendiğini düşündürmektedir.

- Modelin Performansı ve Uyum Değerlendirmesi:Modelin genel sınıflama gücü ROC eğrisi üzerinden hesaplanmış ve AUC değeri 0.672 olarak bulunmuştur; bu da modelin pozitif ve negatif AI algılarını %67.2 doğrulukla ayırt edebildiğini göstermektedir. Hosmer-Lemeshow uyum testi sonucunda p > 0.05 değeri elde edilerek modelin veriyle istatistiksel olarak uyumlu olduğu doğrulanmıştır. Ayrıca modelin açıklayıcılığı McFadden's Pseudo R^2 istatistiği ile ölçülmüş ve sosyal bilimler bağlamında kabul edilebilir düzeyde bir açıklama gücü ortaya konmuştur. Bu sonuçlar, modelin sınırlı sayıda değişkenle çalışmasına rağmen tutarlı ve geçerli çıktılar ürettiğini göstermektedir.

Özetle
istatistik bölümü öğrencilerinin yapay zekaya yönelik kariyer algıları nicel bir yaklaşımla analiz edilmiştir. Öğrencilerin AI'a dair tutumları, geliştirilen AI_Career_Impact_Score ile sayısal hale getirilmiş ve lojistik regresyon modeliyle incelenmiştir. Elde edilen bulgular, özellikle cinsiyet ve üniversite gibi sosyo-demografik değişkenlerin AI algısını etkileyebildiğini; bilgi düzeyi ve staj gibi değişkenlerin ise anlamlı fark yaratmadığını ortaya koymuştur. Bu durum, bireysel anlam dünyasının ve teknolojiye yönelik kişisel tutumların akademik değişkenlerden daha belirleyici olabileceğini göstermektedir. Yapılan hipotez testleri de bu sonuçları destekler niteliktedir. Genel olarak öğrenciler yapay zekayı bir fırsat olarak değerlendirmekte; bu nedenle üniversitelerin eleştirel düşünme ve kariyer vizyonunu destekleyici ortamlar sunması önerilmektedir.

KAYNAKÇA

[1] Kandemir, Y. (2023). *AI Impact on Statistics Students in Ankara* [Data set]. Kaggle. https://www.kaggle.com/datasets/yigitkandemir/ai-impact-on-statistics-students-in-ankara
[2] **Demirtaş, M., & Yıldız, D.** (2021). Üniversite öğrencilerinin yapay zekâya yönelik kariyer algıları üzerine bir değerlendirme. *Sosyal Bilimler Dergisi*, 12(2), 101–116.
[3] **Brynjolfsson, E., & McAfee, A.** (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W.W. Norton & Company.

İSTATİSTİK BÖLÜMÜ

GELİŞMİŞ VE GELİŞMEKTE OLAN ÜLKELERDE CO_2 EMİSYONLARININ PANEL VERİ ANALİZİ İLE KARŞILAŞTIRILMASI

Görkem Güven 20023026

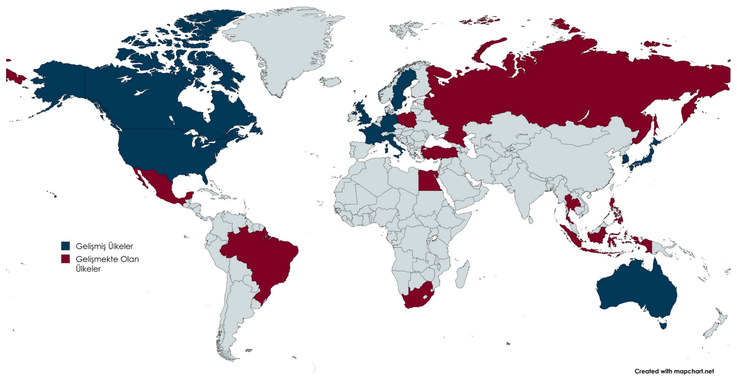
Danışman: Dr. Öğr. Üyesi Doğan Yıldız

ÖZET

Bu çalışma, gelişmiş ve gelişmekte olan ülkelerdeki kişi başı karbondioksit emisyonlarının ekonomik büyüme, enerji tüketimi, şehirleşme oranı ve yenilenebilir enerji kullanımı gibi makroekonomik göstergelerle ilişkisini panel veri ekonometrik yöntemleriyle karşılaştırmalı olarak incelemektedir. 1990–2022 dönemine ait yıllık verilerden oluşan dengeli panel veri seti kullanılarak her iki ülke grubuna özgü modeller tahmin edilmiştir.
Analiz sürecinde önce durağanlık kontrolü yapılmış, ardından uygun model sabit/rastgele etkiler üzerinden Hausman testi ile belirlenmiştir. Model varsayımlarının ihlal edildiği durumlarda White ve Driscoll-Kraay robust tahmincileri uygulanmıştır.

VERİ SETİ VE DEĞİŞKENLER

Sanayileşme, kentleşme ve ekonomik büyüme süreçleri, küresel ölçekte çevresel sorunları beraberinde getirmiştir. Bu sorunların başında gelen karbondioksit emisyonları, sürdürülebilir kalkınma politikaları açısından kritik öneme sahiptir. Bu çalışma, CO_2 emisyonlarının ekonomik büyüme, enerji kullanımı, şehirleşme ve yenilenebilir enerji ile olan ilişkisini, gelişmiş ve gelişmekte olan ülkeler özelinde karşılaştırmalı olarak incelemektedir.
Analiz, 1990–2022 dönemine ait, eksiksiz ve dengeli panel yapıdaki bir veri seti ile gerçekleştirilmiştir. Her iki ülke grubunda 10'ar ülke yer almakta olup ülkeler ve temel değişkenler şu şekildedir:

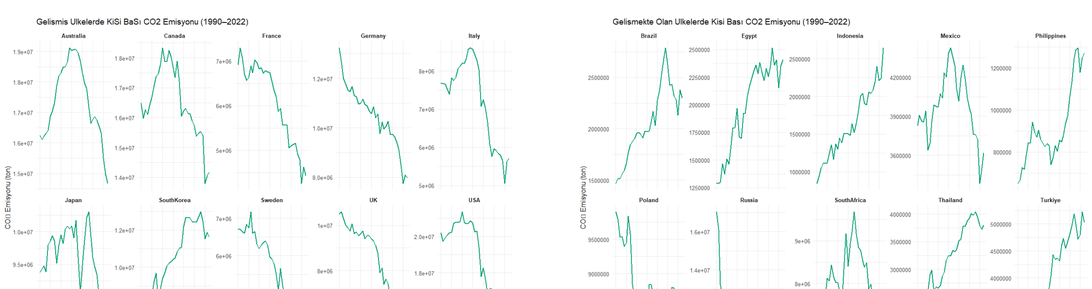


Değişken	Açıklama
CO_2 Emission (per capita)	Kişi başına düşen karbondioksit emisyonlarıdır
GDP (per capita)	Kişi başı Gayri Safi Yurtiçi Hasıla düzeyini temsil eder.
Urban population (% of total population):	Şehirleşme oranını gösterir.
Energy use (per capita)	Kişi başına düşen enerji tüketimini gösterir.
Renewable energy consumption (% of total energy use):	Yenilenebilir enerji kullanım oranını gösterir.

YÖNTEM

Bu çalışmada panel veri analizi kapsamında öncelikle değişkenlerin durağanlığı Im-Pesaran-Shin (IPS) panel birim kök testi ile incelenmiştir. Durağan olmayan serilere birinci fark alma yöntemi uygulanmış, yalnızca gelişmekte olan ülkelerdeki şehirleşme oranı değişkeni için ikinci fark alma gerekmiştir. Uygun modelin belirlenmesi amacıyla sabit ve rassal etkiler modelleri tahmin edilmiş, Hausman testi ile her ülke grubu için en tutarlı model seçilmiştir. Analiz öncesinde bağımsız değişkenler arasında çoklu doğrusal bağlantı olup olmadığı VIF testiyle kontrol edilmiştir. Tahmin edilen modellerde, klasik varsayımlar heteroskedastisite, serisel korelasyon ve kesit bağımlılığı testleriyle sınanmıştır. Bu varsayımlarda ihlal tespit edilen durumlarda ise, gelişmiş ülkeler için White, gelişmekte olan ülkeler için ise Driscoll-Kraay robust tahmin yöntemleri kullanılarak daha güvenilir sonuçlara ulaşılması sağlanmıştır.

UYGULAMA



Şekil 1: Ülke Gruplarına Göre Kişi Başı CO_2 Emisyonu Zaman Serileri

```
Fixed Effects Model (White Robust SE)
=========================================
                   Dependent variable:
               ----------------------------
                          d_co2
-----------------------------------------
d_gdp                   84.6006***
                         (18.6581)

d_energy              1,462.4110***
                       (346.2268)

d_urban_pop          17,981.8600***
                       (5,569.4370)

d_renewable             366.8381
                      (20,913.2800)

-----------------------------------------
Observations              320
R2                       0.6001
Adjusted R2              0.5831
F Statistic     114.7960*** (df = 4; 306)
=========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

```
Random Effects Model with Driscoll-Kraay Robust Standard Errors
=========================================
                   Dependent variable:
               ----------------------------
                          d_co2
-----------------------------------------
d_gdp                    8.3947
                        (21.9098)

d_energy              2,414.7420***
                       (442.8571)

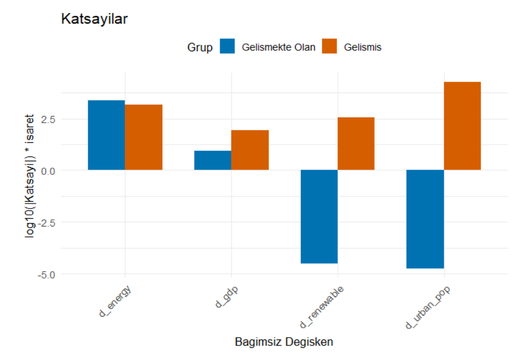
d2_urban_pop          -55,851.6300
                      (35,313.2400)

d_renewable           -33,221.4500***
                       (4,944.8690)

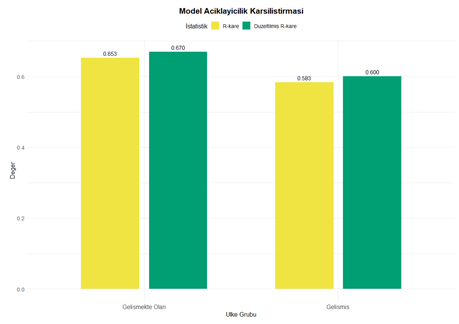
Constant              -29,446.0800***
                       (8,648.3650)

-----------------------------------------
Observations              310
R2                       0.6696
Adjusted R2              0.6653
F Statistic           618.1212***
=========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Yapılan panel veri analizleri sonucunda, gelişmiş ülkeler için White robust, gelişmekte olan ülkeler için ise Driscoll-Kraay robust tahmin yöntemleri kullanılmıştır. Tablolarda, her iki ülke grubu için uygulanan farklı panel veri modellerinin regresyon katsayılarını ve standart hatalarını karşılaştırmalı olarak göstermektedir.



Şekil 2

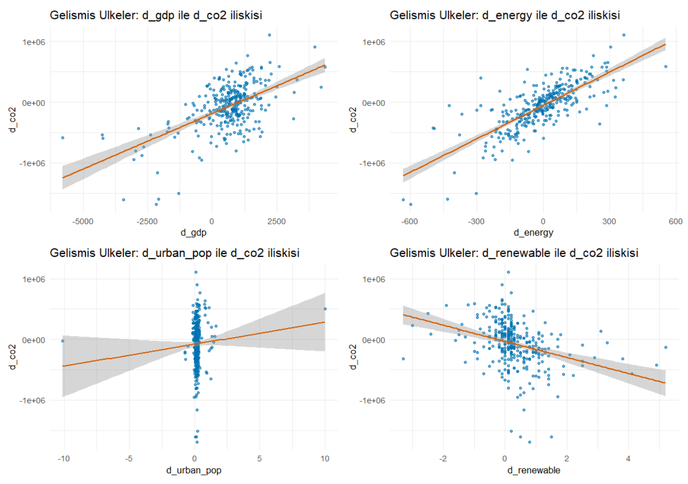


Şekil 3

Şekil 2, CO_2 emisyonları üzerindeki etkileri karşılaştırmak amacıyla gelişmiş ve gelişmekte olan ülkelerde oluşturulan panel veri modellerinin regresyon katsayılarını göstermektedir. Şekil 3, her iki modelin R^2 ve düzeltilmiş R^2 değerlerini yani modelin açıklayıcılığını sunmaktadır. Her iki model de veri setini güçlü biçimde açıklamakta, ancak politika öncelikleri ülke grubuna göre farklılaşmalıdır.

Gelişmiş ülkelerde;

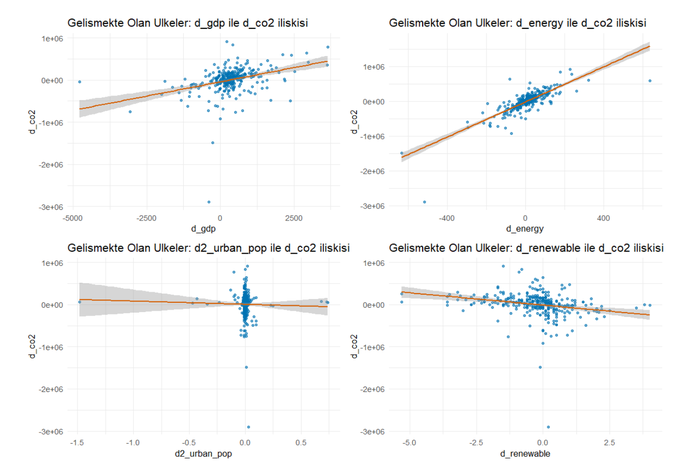
- Kişi başına gdp artışı CO_2 emisyonlarını artırma eğiliminde. Bu ilişki pozitif olsa da, veriler heterojen dağılmış, yani etki güçlü ama çok belirgin değil.
- Enerji tüketimindeki artış CO_2 emisyonlarını açık şekilde artırıyor. Bu, gelişmiş ülkelerde enerji talebi arttıkça emisyonların da arttığını gösteriyor.
- Şehirleşme oranındaki değişim çok düşük olduğu için CO_2 üzerindeki etkisi istatistiksel olarak anlamlı olmadığı görülüyor. Bu da gelişmiş ülkelerde şehirleşmenin zaten doygunluğa ulaşmış olmasından kaynaklanabilir.
- Yenilenebilir enerji kullanımındaki artış, CO_2 emisyonlarını azaltma yönünde bir etki oluşturuyor. Ancak regresyon çizgisinin eğiminin fazla olmaması, etkinin sınırlı ya da başka faktörlerden etkileniyor olabileceğini gösteriyor.



Şekil 4: Gelişmiş Ülkelerde CO_2 Belirleyicileri

Gelişmekte olan ülkelerde;

- Gdp'deki artış emisyonları artırıyor gibi görünse de ilişki zayıf ve çok yayılmış.
- Enerji tüketimindeki artış CO_2 emisyonlarını doğrudan ve güçlü şekilde artırıyor. Bu, gelişmekte olan ülkelerde enerji üretiminin büyük ölçüde fosil kaynaklara dayanmasıyla uyumludur.
- Şehirleşme değişkeninin ikinci farkı alınmıştı, yani oldukça düzleştirilmiş haliyle modelde kullanıldı. Bundan dolayı istatistiksel anlamlı bir etki olmadığı görülüyor. Hem şehirleşme değişimi az, hem etki çok zayıf.
- Yenilenebilir enerji kullanımındaki artış CO_2 emisyonlarını azaltıcı yönde etkili.



Şekil 5: Gelişmekte Olan Ülkelerde CO_2 Belirleyicileri

DEĞERLENDİRME

- Bu bulgular, çevre politikalarının ülkelerin gelişmişlik düzeyine göre farklı önceliklerle kurgulanması gerektiğini göstermektedir. Gelişmekte olan ülkelerde enerji tüketiminin CO_2 emisyonlarını güçlü biçimde artırması, bu ülkelerde fosil kaynaklara dayalı üretimin hâlâ baskın olduğunu ortaya koymaktadır. Bu nedenle, enerji dönüşüm politikaları ve yenilenebilir kaynak yatırımları bu grup için acil bir ihtiyaçtır.
- Gelişmiş ülkelerde ise kişi başı gelir artışı ve enerji tüketimi emisyonları artırmakla birlikte, yenilenebilir enerji kullanımının azaltıcı etkisi daha sınırlı kalmaktadır. Bu durum, mevcut sistemlerde enerji verimliliğini artırıcı önlemlerin ve düşük karbonlu büyüme stratejilerinin ön plana çıkması gerektiğini göstermektedir. Model sonuçları, politika tasarımlarının her ülke grubunun yapısal özelliklerine göre şekillendirilmesinin önemini vurgulamaktadır.

KAYNAKÇA

[1] Apergis, N., & Payne, J. E. (2010). The emissions, energy consumption, and growth nexus: Evidence from the commonwealth of independent states. *Energy Economics*, 32(3), 558–564.
[2] Al-Mulali, U., & Sab, C. N. B. C. (2012). The impact of energy consumption and CO_2 emission on the economic growth of 19 selected countries. *Renewable and Sustainable Energy Reviews*, 16(7), 4365–4369.
[3] Stern, D. I. (2004). The rise and fall of the environmental Kuznets curve. *World Development*, 32(8), 1419–1439.

İSTATİSTİK BÖLÜMÜ

BİREYLERİN RUH SAĞLIĞINA YÖNELİK TEDAVİ ALIP ALMAMA DURUMUNUN TAHMİN EDİLMESİ

İrem Sude UZUN 20023049

Danışman: Prof. Dr. Filiz KARAMAN

ÖZET

Bu çalışmada, bireylerin ruh sağlığına yönelik olarak tedavi alıp almadıklarını tahmin edebilmek amacıyla bir makine öğrenmesi modeli geliştirilmiştir. Veri kümesi, anket yöntemiyle toplanmış bireysel demografik ve psikososyal değişkenleri içermektedir. Modelleme sürecinde, eksik veriler temizlenmiş, kategorik değişkenler dönüştürülmüş ve eğitim-test ayrımı yapılmıştır. Ardından Lojistik Regresyon ve Rastgele Orman (Random Forest) gibi denetimli öğrenme algoritmaları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Elde edilen sonuçlar, ruh sağlığı alanında veri temelli karar destek sistemleri geliştirmek adına bulgular sunmaktadır.

GENEL BİLGİ

Ruh sağlığı, bireylerin yaşam kalitesini ve toplumsal işlevselliğini doğrudan etkileyen temel bir sağlık bileşenidir. Günümüzde özellikle çalışan bireylerde ruhsal destek alma ihtiyacı artmakta, ancak her birey bu desteği alma yoluna gitmemektedir. Bu nedenle bireylerin tedavi alıp almayacağına etki eden faktörlerin belirlenmesi ve bu durumun tahmin edilmesi önem kazanmıştır.

VERİ KÜMESİ VE PROBLEM TANIMI

Bu çalışmada, bireylerin ruh sağlığına yönelik olarak tedavi alıp almadıkları durumu, çeşitli demografik ve psikososyal değişkenler aracılığıyla tahmin etmeye yönelik bir sınıflandırma modeli geliştirilmiştir. Bu amaçla, Kaggle platformunda yer alan “Mental Health in Tech Survey” veri seti kullanılmıştır.

Veri seti, özellikle teknoloji sektöründe çalışan bireylerden toplanmış ve bireylerin yaş, cinsiyet, ruh sağlığı geçmişi, iş yerindeki destek sistemleri, anonimlik algısı, çalışanların ruh sağlığı hizmetlerine ulaşımı gibi çok sayıda değişkeni içermektedir.

Tahmin edilmek istenen değişken, bireyin ruh sağlığına yönelik tedavi alıp almadığını belirten treatment değişkenidir (0 = Hayır, 1 = Evet). Bu nedenle çalışma, **ikili sınıflandırma (binary classification)** problemine dayanmaktadır.

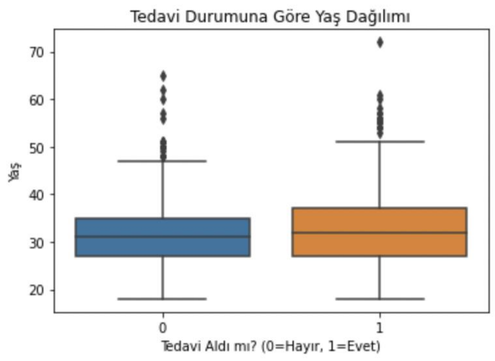
Amacımız; bireylerin sahip olduğu özelliklere bakarak, ruh sağlığı tedavisi alıp almadıklarını yüksek doğrulukla öngörebilen bir model inşa etmektir. Bu modelin oluşturulmasında **denetimli makine öğrenmesi algoritmaları** kullanılmış, algoritmaların performansları çeşitli metrikler ile değerlendirilmiştir.

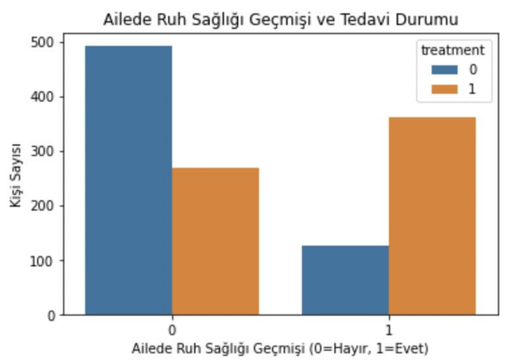
YÖNTEM VE MODELLEME

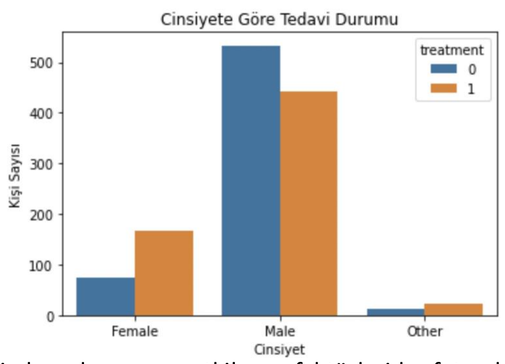
Çalışmada sınıflandırma problemi olarak ele alınan “tedavi görme durumu”, denetimli makine öğrenmesi algoritmaları ile tahmin edilmeye çalışılmıştır. Öncelikle veri kümesi üzerinde eksik veriler temizlenmiş, kategorik değişkenler etiketleme (label encoding) ve tekil temsil (one-hot encoding) yöntemleri ile sayısal hale getirilmiştir. Ardından veri, %80 eğitim ve %20 test olacak şekilde rastgele ikiye ayrılmıştır.

Modelleme sürecinde Lojistik Regresyon ve Rastgele Orman (Random Forest) algoritmaları uygulanmıştır. Logistic Regression modeli, sınıflar arasındaki olasılıkları hesaplamak için kullanılırken; Random Forest algoritması çok sayıda karar ağacından oluşan topluluk (ensemble) yöntemine dayanmaktadır ve sınıflandırma performansı açısından oldukça güçlü bir yöntemdir.

Modellerin başarıları doğruluk oranı (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi istatistiksel ölçütler aracılığıyla değerlendirilmiştir. Elde edilen sonuçlara göre, Random Forest modeli diğer yöntemlere kıyasla daha yüksek performans göstermiştir.

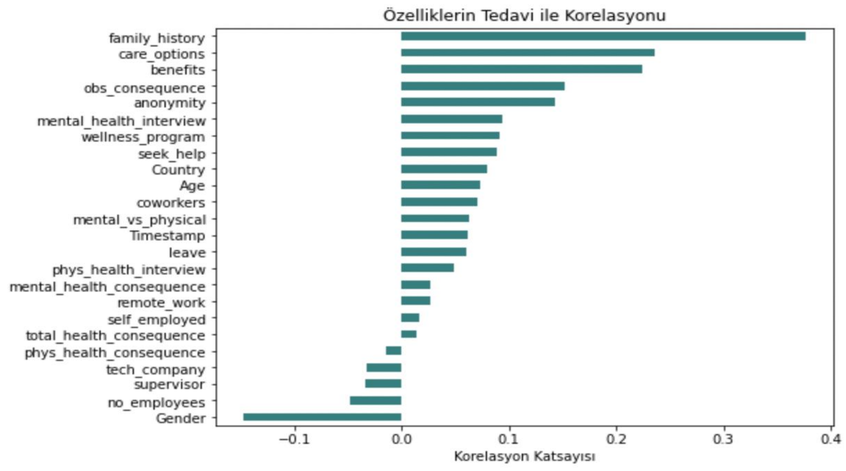






Bu görseller, bireylerin tedavi alma durumunu etkileyen faktörleri keşfetmek amacıyla oluşturulmuştur. Yaş dağılımına bakıldığında, tedavi alan bireylerin yaş ortalamasının biraz daha yüksek olduğu ve bazı aykırı değerlerin bulunduğu görülmektedir. Ailede ruh sağlığı geçmişi olan bireylerin tedavi alma olasılığı belirgin şekilde daha yüksekken, cinsiyet değişkeninde ise erkeklerin tedaviye başvurma oranı kadınlara göre daha düşüktür. Bu analizler, yaş, cinsiyet ve aile geçmişi gibi değişkenlerin modelleme sürecinde önemli açıklayıcı değişkenler olabileceğini göstermektedir.

MENTAL SAĞLIK DESTEĞİ ALMA DAVRANIŞININ BELİRLEYİCİLERİ: KORELASYON MATRİSİ ANALİZİ



Bu korelasyon matrisi, bireylerin mental sağlık desteği alıp almama kararlarını hangi faktörlerin ne yönde etkilediğini analiz etmektedir. "Seek_help" değişkeni ile en güçlü pozitif ilişkiye sahip faktörler, bireylerin destek arama eğilimini anlamada kritik rol oynamaktadır. "Anonymity" ve "mental_health_interview" gibi değişkenler, mahremiyet kaygılarının ve açık iletişim imkânlarının destek alma davranışı üzerindeki etkisini ortaya koymaktadır. "Coworkers" ve "supervisor" ile olan korelasyonlar, sosyal desteğin ve yönetici tutumlarının bu süreçteki önemini yansıtmaktadır. Son olarak, düşük veya negatif korelasyon gösteren değişkenler "self_employed", hangi grupların daha az destek aradığını belirlemede kullanılabilir.

MENTAL SAĞLIK DESTEĞİ ALMA DAVRANIŞININ TAHMİNİNDE MODEL KARŞILAŞTIRMASI

Random Forest:

	precision	recall	f1-score	support
0	0.61	0.68	0.64	113
1	0.71	0.64	0.67	138
accuracy			0.66	251
macro avg	0.66	0.66	0.66	251
weighted avg	0.66	0.66	0.66	251

Accuracy: 0.6573705179282868

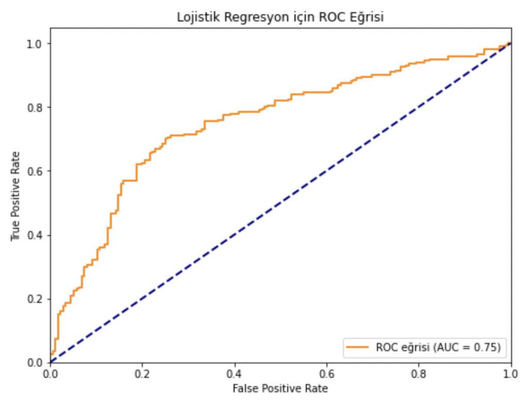
Logistic Regresyon:

	precision	recall	f1-score	support
0	0.69	0.73	0.71	176
1	0.75	0.71	0.73	200
accuracy			0.72	376
macro avg	0.72	0.72	0.72	376
weighted avg	0.72	0.72	0.72	376

Accuracy: 0.7180851063829787

Yapılan bu çalışmada, bireylerin mental sağlık desteği alma davranışlarını tahmin etmede lojistik regresyon ve Random Forest algoritmalarının performansları karşılaştırılmıştır. Analiz sonuçlarına göre lojistik regresyon modeli, test veri setinde %71.8'lik doğruluk (accuracy) oranıyla Random Forest'ın %65.7'lik performansına kıyasla daha başarılı sonuçlar üretmiştir.

Model performans metrikleri detaylı incelendiğinde, lojistik regresyonun özellikle pozitif sınıf (destek alanlar) için %75'lik precision değeriyle dikkat çekici bir performans sergilediği gözlemlenmiştir. Bu sonuç, modelin Tip I hata oranını (false positive) daha etkin kontrol edebildiğini göstermektedir. F1-skor değerlerinin 0.71 (negatif sınıf) ve 0.73 (pozitif sınıf) olarak hesaplanması, modelin sınıflar arasında dengeli bir performans gösterdiğine işaret etmektedir.



Bu grafik, lojistik regresyon modelinin ROC eğrisini göstermektedir. ROC eğrisi, modelin doğru pozitif oranı ile yanlış pozitif oranı arasındaki ilişkiyi ifade eder. Grafikteki turuncu çizgi, modelin performansını; mavi kesikli çizgi ise rastgele tahmin düzeyini temsil eder. AUC değeri 0.75 olup modelin sınıfları %75 doğrulukla ayırt edebildiğini gösterir. Bu sonuç, modelin ortalamanın üzerinde ve başarılı bir sınıflandırma performansı sergilediğini gösterir.

KAYNAKÇA

[1] Kaggle, "Mental Health in Tech Survey",

[2] Oscherwitz, T., & Huang, Z. (2019). "Mental Health in the Workplace: Identifying and Addressing the Challenges." Journal of Occupational Health Psychology, 24(2), 95-105

[3] Almeida, D. M., & Sadeghi, N. (2019). "Mental Health Prediction Using Machine Learning: A Review."

2024-2025 Bahar Yarıyılı İstatistik BÖLÜMÜ

DOĞALGAZ YAYGINLIĞININ TÜRKIYE'DE TÜPGAZ TALEBINE ETKISI: ZAMAN SERISI YAKLAŞIMIYLA 5 YILLIK ULUSAL VE BÖLGESEL ÖNGÖRÜ ANALIZI

Kadir Cengiz Bahar - 21023603

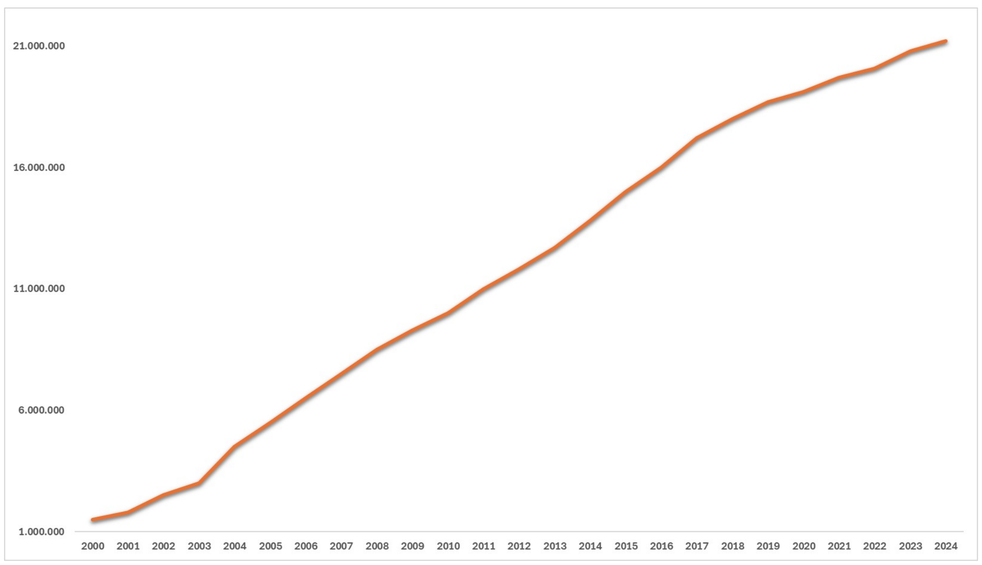
Danışman: Arş. Gör. Dr. Egemen ÖZKAN

ÖZET

Bu çalışma, Türkiye'de doğalgaz altyapısının yaygınlaşmasının Tüpgaz (LPG) pazarı üzerindeki etkilerini incelemekte ve geçmiş eğilimler doğrultusunda geleceğe yönelik öngörüler sunmayı amaçlamaktadır. 2019–2024 yılları arasında iller bazında doğalgaz abone sayıları ile tüpgaz satış miktarları karşılaştırılmış; konut sayısındaki artış, altyapı yaygınlığı ve teknolojik dönüşüm gibi ek faktörler de değerlendirilmiştir. Zaman serisi analiz yöntemlerinden ARIMA ve ETS modelleri kullanılarak, 2025–2030 dönemi için hem Türkiye geneli hem de bölgesel bazda tüpgaz tüketim tahminleri üretilmiştir. Tahmin sonuçlarına göre, 2024 yılında 573 bin ton olan yıllık tüpgaz tüketiminin, 2030 yılına kadar yaklaşık 470 bin tona kadar gerilemesi öngörülmektedir. Bu düşüş eğilimi Marmara ve İç Anadolu gibi doğalgaz altyapısının yoğun olduğu bölgelerde daha belirgin iken, Güneydoğu Anadolu ve Akdeniz bölgelerinde göreli olarak daha yavaş bir azalma beklenmektedir. Çalışma, özellikle hızlı tüketim düşüşü gösteren veya stabil kalan iller üzerinden stratejik analizler yaparak, enerji planlaması ve LPG sektörünün geleceği açısından karar vericilere yol gösterici veriler sunmayı hedeflemektedir.

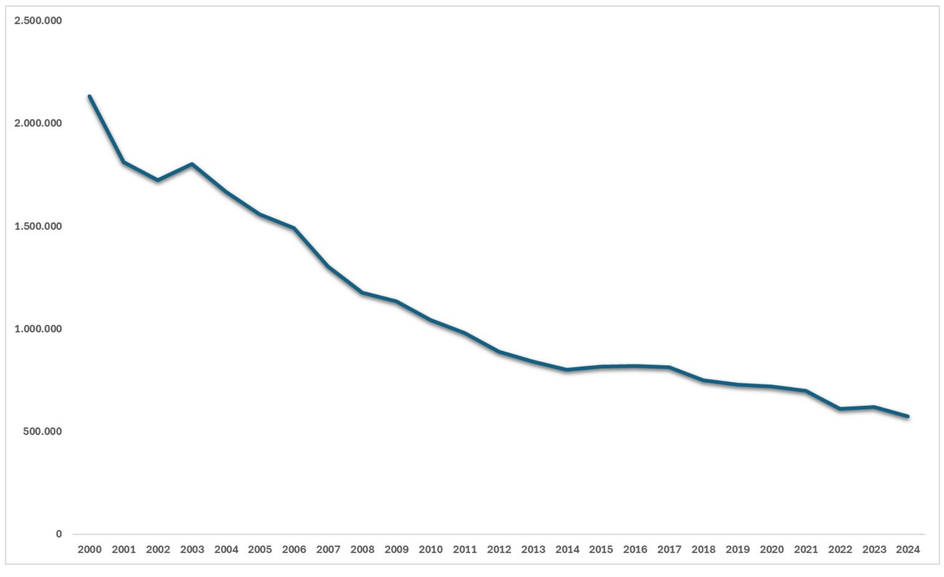
Doğalgaz Aboneliği Trendi

Türkiye'de doğalgaz kullanımı 2000'li yılların başından itibaren enerji politikalarının merkezine yerleşmiş ve hızlı bir yaygınlaşma süreci başlamıştır. 2000 yılında yaklaşık 1,5 milyon olan doğalgaz abone sayısı, altyapı yatırımları ve konutlaşmadaki artışla birlikte 2010 yılında 10 milyon sınırına ulaşmıştır. Bu dönemde her yıl ortalama 800.000 ila 1 milyon yeni abone sisteme dâhil olmuştur.
2010 sonrasında abone artış hızı görece azalsa da büyüme devam etmiş; 2022 sonunda abone sayısı yaklaşık 19,7 milyona, 2023 yılı ortasında ise 20 milyon seviyesinin üzerine çıkmıştır. 2024 yılı itibarıyla Türkiye genelinde doğalgaz abone sayısının 21 milyona yaklaştığı tahmin edilmektedir. Bu artış, özellikle büyükşehirler dışındaki orta ve küçük ölçekli kentlerin doğalgaz şebekesine dahil edilmesiyle gerçekleşmiştir.
Doğalgaz aboneliğinin bu hızlı artışı, LPG (tüpgaz) pazarını doğrudan etkilemektedir. Abone sayısındaki büyüme, özellikle konut segmentinde tüpgaz talebinde ciddi bir azalmaya neden olmuş, sektörel dönüşüm sürecini hızlandırmıştır. Bu nedenle doğalgaz abonelik trendi, yalnızca altyapı yayılımını değil, aynı zamanda enerji tüketim alışkanlıklarındaki dönüşümü de temsil etmektedir.



Tüpgaz Satış Trendi

Türkiye'de Tüpgaz (LPG) satışı 2000 yılında 2.133.000 ton iken, 2024 yılına gelindiğinde bu rakam 573.142 tona kadar düşmüştür. Bu, 24 yılda yaklaşık %73 oranında bir azalma anlamına gelmektedir.
2000'li yılların başında yaygın olarak kullanılan tüpgaz, doğalgaz altyapısının hızla genişlemesiyle birlikte özellikle konut kullanımında ciddi biçimde gerilemiştir. 2010 sonrası dönemde düşüş daha da hızlanmış, 2020'li yıllarda ise tüketim 700 bin tonun altına inmiştir. Özellikle 2021 sonrası yıllarda tüketim, 700 bin tonun altına gerilemiş; 2024 yılında ise satış miktarı 573 bin ton düzeyinde gerçekleşmiştir. Bu dönem, sektör açısından doygunluk noktasına yaklaşan ve geleneksel kullanım alanlarının daraldığı bir dönüşüm sürecine işaret etmektedir. Bu genel eğilim, doğalgaz altyapı yatırımlarının LPG pazarına olan etkisini net biçimde yansıtmakta olup; zaman serisi modellemeleriyle gelecekte bu düşüş eğiliminin devam edeceği öngörülmektedir.

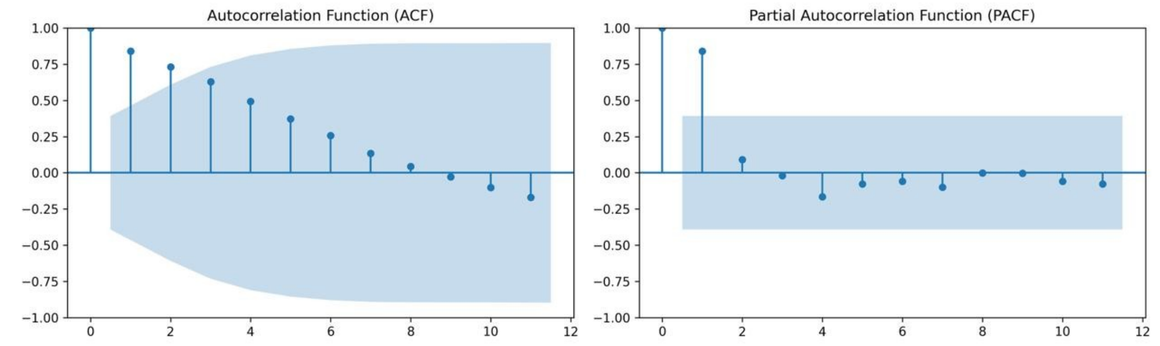


Tezin Genel Sonuçları

Bu tezde, doğalgaz kullanımının yaygınlaşmasının Türkiye'de tüpgaz pazarına etkisi analiz edilmiştir. 2000–2024 yılları arasında tüpgaz satışları istikrarlı bir düşüş gösterirken, doğalgaz abone sayılarında sürekli artış gözlemlenmiştir. Yapılan zaman serisi analizleri, bu düşüş eğiliminin 2025–2030 döneminde de süreceğini göstermektedir. Ancak bazı illerde doğalgaz altyapısının sınırlı olması nedeniyle tüpgaz kullanımının devam ettiği belirlenmiştir. Ayrıca konut sayısı artsa da yeni konutların doğalgaz ve elektrikle uyumlu olması, tüpgaz talebini artırmamaktadır. Bu bulgular, enerji politikalarının bölgesel farklılıkları gözeterek şekillendirilmesi gerektiğini ortaya koymaktadır.

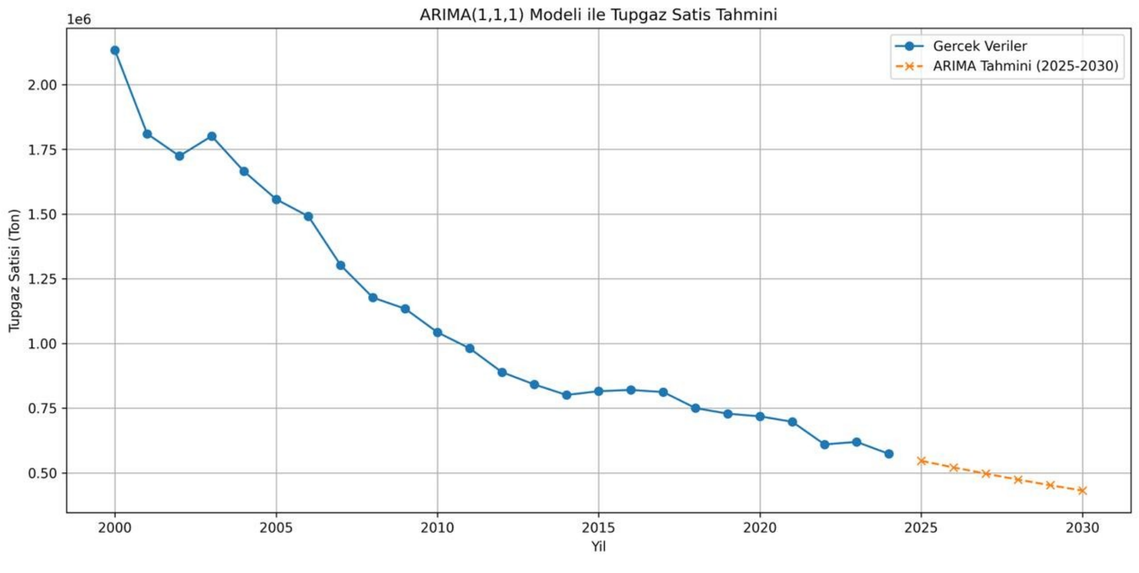
Autocorrelation (ACF) & Partial Autocorrelation (PACF) Grafikleri Açıklaması

Bu iki grafik, zaman serisi modellemesinde kullanılacak ARIMA modelinin uygun parametrelerinin belirlenmesinde temel istatistiksel araçlardır:
•ACF (Autocorrelation Function): Verinin kendi gecikmeli (lag) değerleriyle olan ilişki düzeyini ölçer. İlk 3–4 gecikmede pozitif otokorelasyon görülmektedir, bu da veride zayıflayan ama kalıcı bir trend etkisi olduğunu göstermektedir. Ayrıca, ilk birkaç lag sonrası otokorelasyon değerlerinin sıfıra yaklaşması, verinin durağanlaştırılabileceğini (d = 1) göstermektedir.
•PACF (Partial Autocorrelation Function): Her lag'deki bağımsız etkileri ölçer. Lag 1'de anlamlı bir sıçrama gözlemlenmiştir, bu da AR(1) bileşeninin modelde yer alması gerektiğini gösterir. Diğer gecikmeler anlamlı çıkmadığından modelin daha yüksek düzeyde AR bileşeni içermesine gerek yoktur.
Bu çıktılar, ARIMA modelinin (p=1, d=1, q=1) şeklinde kurulması için güçlü bir temel sunmuştur.



ARIMA Modeli ile Türkiye Tüpgaz Satış Tahmini (2000–2030)

Bu grafik, 2000–2024 yılları arasındaki yıllık tüpgaz satışlarına dayalı olarak oluşturulan ARIMA(1,1,1) modelinin 2025–2030 dönemine ilişkin tahminlerini içermektedir.
Model bileşenleri:
•AR(1) – AutoRegressive: Satış verisi, bir önceki yıla bağlı olarak ilerlemektedir.
•I(1) – Integrated: Serinin trend içerdiği belirlenmiş ve birinci dereceden fark alma işlemi uygulanarak durağanlaştırılmıştır.
•MA(1) – Moving Average: Önceki dönem hata terimleri modele dahil edilerek dalgalanmalar minimize edilmiştir.
Tahmin Özeti:
•2024 satış değeri: 573.142 ton
•2025 tahmini: ~555.000 ton
•2030 tahmini: ~470.000 ton
Model, geçmiş verilerdeki azalışı temel alarak tüpgaz tüketiminin önümüzdeki beş yıl içinde ortalama %3 oranında azalmaya devam edeceğini göstermektedir.
Bu tahminler, doğalgaz altyapısının yaygınlaşması, elektrikli mutfak cihazlarının kullanım oranının artması ve konut politikalarındaki dönüşüm gibi etkenlerle büyük ölçüde örtüşmektedir.



KAYNAKÇA

[1] Enerji Piyasası Düzenleme Kurumu (EPDK). *LPG Piyasası Sektör Raporu*.
[2] Uluslararası Enerji Ajansı (IEA). (2023). *World Energy Outlook 2023*
[3] Enerji Piyasası Düzenleme Kurumu (EPDK). *Doğalgaz Piyasası Sektör Raporu*.

İSTATİSTİK BÖLÜMÜ

Konut Fiyatları Tahmininde Makine Öğrenmesi Algoritmalarının Karşılaştırılması

Kağan ALTUN 20023007

Danışman: Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

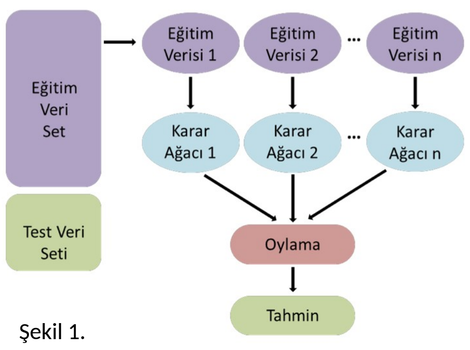
Konut fiyatlarını tahmin etmek için genellikle regresyon yöntemleri kullanılır ve model oluşturmak için çok sayıda değişken kullanılır. Literatür taraması bulgularımız, çeşitli geleneksel makine öğrenimi algoritmalarının çalışıldığını göstermektedir; ancak konut fiyatı tahmini için en uygun metodolojinin belirlenmesine ihtiyaç vardır.[1] Bu çalışmada, Türkiye konut piyasasındaki fiyatları tahmin etmek amacıyla makine öğrenmesi algoritmaları kullanılmıştır. Öncelikle, veri setindeki eksik ve hatalı değerler uygun yöntemlerle temizlenmiş ve sayısal-kategorik değişkenler modellemeye uygun hale getirilmiştir. Daha sonra, Random Forest, XGBoost ve CatBoost gibi güçlü regresyon modelleri, hiperparametre optimizasyonu ile en iyi performans verecek şekilde eğitilmiştir. Model sonuçları, doğruluk ve hız açısından karşılaştırılarak, en etkili konut fiyatı tahmin modeli belirlenmiştir. Bu sayede, gayrimenkul sektöründeki karar süreçlerine katkı sağlayacak, güvenilir ve uygulanabilir bir fiyat tahmin sistemi geliştirilebilir. Ayrıca alıcılar da mülklerin satın alınabilirliğini tahmin etmek ve bilinçli satın alma kararları vermek için modeli kullanabilir.

MAKİNE ÖĞRENMESİ HAKKINDA

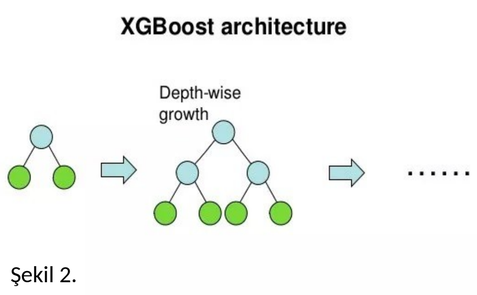
Makine öğrenmesi, karmaşık örüntüyü algılayarak veriye dayalı karar verebilme özellikleriyle ele alınan problemin çözümünü kendi kendine öğrenebilen bilgisayar algoritmalarının genel adıdır (Gültepe, 2019). Bu algoritmalar, büyük miktarda veriyi işleyerek örüntü ve ilişkiler keşfetme yetenekleri sayesinde öğrenme süreçlerini geliştirme potansiyeli taşırlar (Erbay Mermer, 2023).

RANDOM FOREST

Breiman ve ark. tarafından tanıtılan rastgele orman (RF), çeşitli veri kümeleri için kesin sonuçlar sunan sağlam ve çok yönlü bir topluluk öğrenme tekniğidir. Çeşitli veri kümeleri için sonuçları tahmin etmedeki hassasiyeti ile yaygın olarak tanınan RF, yüksek boyutlu verileri işleme, karmaşık ilişkileri yakalama ve aşırı uyumu azaltma konusunda mükemmeldir. Zayıf özelliklere daha az ağırlık vererek benzersiz bir şekilde çalışır ve diğer yöntemlere kıyasla daha hızlı işleme sağlar. Bu özelliği onu eksik veya gürültülü verilerle ve aykırı değerlerle başa çıkmak için güvenilir bir seçim haline getirir. RF, hem sınıflandırma hem de regresyon görevlerinin üstesinden gelebildiği için çok yönlüdür (Harsora vd., 2024). Modelde kararlar, bir dizi alt karara göre üretilir. Her karar ağacı, sırasıyla özellikleri ve kararları temsil eden düğümlere ve yapraklara sahiptir. Düğümlere, geçerli ağacı eğitmek için uygulanan özelliklerden en iyilerinin seçilmesiyle karar verilir. RF algoritması Şekil 1'de şematik gösterilmiştir (Yalçın Kuzu, 2022).



Şekil 1.



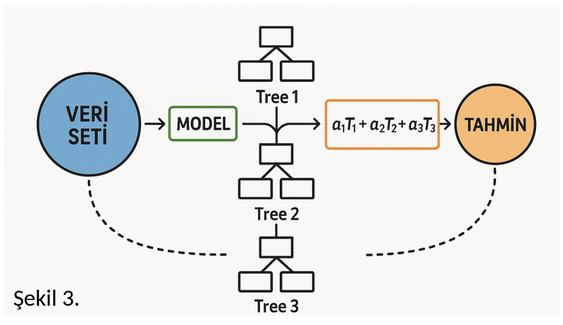
Şekil 2.

XGBOOST

XGBoost (Extreme Gradient Boosting), gradyan artırmalı ağaç tabanlı bir makine öğrenmesi algoritmasıdır. Büyük veri setlerinde yüksek doğrulukla tahmin yapabilmek için tasarlanmış, hızlı ve verimli bir modeldir. Temelde, birden çok zayıf karar ağacını ardışık olarak eğitir; her yeni ağaç, önceki ağaçların yaptığı hataları azaltmaya odaklanır (Chen & Guestrin, 2016). Bu sayede, hem regresyon hem de sınıflandırma problemlerinde üstün performans sağlar. XGBoost, düzenlileştirme (regularization), paralel hesaplama ve hata azaltma gibi gelişmiş tekniklerle aşırı öğrenmeyi engeller ve genelleme kabiliyetini artırır (Sharma vd., 2024). XGBoost algoritması Şekil 2'de şematik gösterilmiştir.

CATBOOST

CatBoost, Yandex tarafından geliştirilmiş bir gradient boosting algoritması olup özellikle kategorik verilerin etkili işlenmesi için tasarlanmıştır. Klasik one-hot encoding yönteminin yüksek kategorili değişkenlerde patlama yaşattığını, buna karşın CatBoost'un kategorik değişkenleri özel bir kodlama yöntemiyle dönüştürerek veriyi daha etkin hazırladığını ve bu sayede model eğitimini hızlandırdığını belirtmiştir. Ordered boosting (sıralı örnekleme) tekniğini kullanarak tahmin sapmalarını azaltan CatBoost, her iki türlü (sayısal ve kategorik) özelliği de istatistiksel kombinasyonlarla değerlendirir. Ayrıca simetrik karar ağaçları inşa ederek yüksek doğruluk sağlarken aşırı öğrenmeyi engeller ve uygun donanımla GPU üzerinde çalışarak büyük veri setlerinde dahi hızlı sonuçlar üretebilir (Çetin & Koçak, 2023). CatBoost algoritması Şekil 3'de şematik gösterilmiştir.



Şekil 3.

HİPERPARAMETRE OPTİMİZASYONU (GridShearchCV)

Model performansını en üst düzeye çıkarmak için hiperparametreler dikkatlice ayarlanmalıdır. Model hiperparametrelerinin optimizasyonu için GridSearchCV yöntemi kullanılmıştır. GridSearchCV, farklı parametre kombinasyonlarını sistematik olarak deneyerek en iyi sonucu veren ayarları belirleyen bir yöntemdir. Bu sayede modellerin doğruluk ve genelleme kabiliyeti artırılır.

UYGULAMA

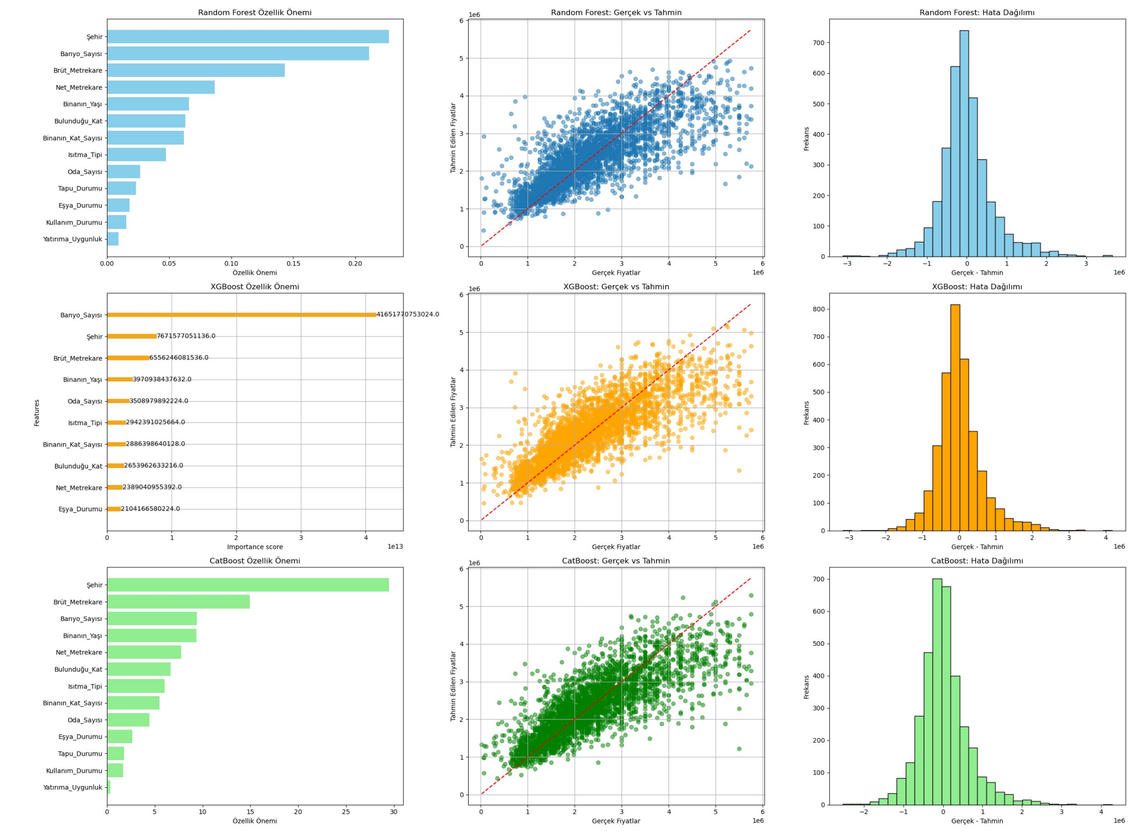
VERİ HAKKINDA

Bu çalışma için kullanılan veri seti, Kaggle platformundan temin edilmiştir. Veri seti, konut fiyat tahmini amacıyla hazırlanmış olup 15 sütun ve 20.326 kayıt içermektedir. Konutun fiziksel özellikleri (Net Metrekare, Brüt Metrekare, Oda Sayısı, Banyo Sayısı, Binanın Yaşı, Bulunduğu Kat vb.), konutun bulunduğu şehir, ısıtma tipi gibi kategorik özellikler ve konutun satış fiyatı gibi değişkenler yer almaktadır.

Veri setindeki eksik sayısal değerler medyan ile dolduruldu. Kategorik değişkenlerdeki eksik değerler ise "Bilinmiyor" etiketiyle tamamlandı. Fiyat ve metrekare gibi değişkenlerde uç değerler IQR yöntemiyle filtrelendi. Bazı kategorik değişkenler sayısala çevrildi. Tekrarlayan kayıtlar temizlendi ve sayısal veriler standartlaştırıldı. Bu adımlar, modellerin daha sağlıklı öğrenmesi için yapıldı.

Model Uygulaması ve Hiperparametre Optimizasyonu

Çalışmada Random Forest, XGBoost ve CatBoost algoritmaları kullanılmıştır. Her model için GridSearchCV ile hiperparametre optimizasyonu yapılarak en iyi parametreler belirlenmiştir. Optimizasyon sonrası modeller, test verisi üzerinde değerlendirilmiş ve performansları karşılaştırılmıştır.




```
Modellerin Eğitim ve Tahmin Süreleri ile Performans Karşılaştırması:
               Eğitim Süresi (sn) Tahmin Süresi (sn)        RMSE      R2
Model
Random Forest               5.59                0.09  656,658.65  0.6051
XGBoost                     0.22                0.01  628,830.75  0.6379
CatBoost                    1.10                0.01  618,517.00  0.6497
```

SONUÇ

Analizimizde, Random Forest, XGBoost ve CatBoost modellerinin konut fiyat tahmin performansları kapsamlı şekilde incelenmiştir. Performans metrikleri ve grafikler, CatBoost modelinin en yüksek doğruluk (en düşük RMSE, en yüksek R^2) ile öne çıktığını göstermektedir. XGBoost, eğitim ve tahmin hızında üstünlük sağlarken, Random Forest daha uzun eğitim süresiyle diğer modellerin gerisinde kalmıştır.
Özellik önemleri analizinde, şehir, net metrekare ve banyo sayısı gibi değişkenlerin fiyat tahminindeki etkisi baskın olarak ortaya çıkmıştır. Gerçek ve tahmin edilen fiyatların karşılaştırıldığı dağılım grafikleri, modellerin genel doğruluğunu ve tutarlılığını net biçimde desteklemektedir.

Bu sonuçlar, özellikle CatBoost ve XGBoost modellerinin Türkiye konut piyasasında hem doğruluk hem de işlem hızı açısından tercih edilebilir olduğunu göstermekte olup, karar destek sistemlerinde etkin şekilde kullanılabilir.

KAYNAKÇA

[1] Harsora H., Ogunleye B., Sharma H. (2024). An Optimal House Price Prediction Algorithm: XGBoost. Analytics, 3, 30-45.
[2] Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[3] https://www.kaggle.com/code/omerdasc/xgboost-randomforest-house-price#Aykırı-Fiyatları--Sil

YTÜ FEN EDEBİYAT FAKÜLTESİ

2024-2025 Bahar Yarıyılı

İSTATİSTİK BÖLÜMÜ

PEKİŞTİRMELİ ÖĞRENME (REINFORCEMENT LEARNING) TABANLI PORTFÖY OPTİMİZASYONU

Öğrenci Adı : Kemal Can ÖZMETLİ 22023614

Danışman: Doç. Dr. Selçuk ALP

ÖZET

Bu çalışmada, Borsa İstanbul'da işlem gören 10 farklı hisse senedi kullanılarak portföy optimizasyonu problemi pekiştirmeli öğrenme yöntemleriyle ele alınmıştır. PPO (Proximal Policy Optimization) ve A2C (Advantage Actor-Critic) algoritmaları uygulanarak iki farklı pekiştirmeli öğrenme ajanı eğitilmiş, bu modellerin portföy performansları karşılaştırmalı olarak analiz edilmiştir. Model performansları; eşit ağırlıklı portföy stratejisi (equal weighted portfolio) ve rastgele aksiyon üreten ajanlar (random agents) ile karşılaştırmalı backtest yoluyla değerlendirilmiştir. Değerlendirme sürecinde logaritmik getiri, volatilite, Sharpe oranı ve maksimum kayıp gibi risk-getiri metrikleri kullanılmıştır. Çalışmada 2010–2024 dönemine ait veri kullanılmış; bu dönem içinde ekonomik yapının kırılıma uğrayıp uğramadığı analiz edilmiştir. Yapısal rejim değişimi testi sonucunda 2020 yılında belirgin bir konjonktürel kırılma olduğu istatistiksel olarak saptanmıştır. Bu doğrultuda 2010–2020 ve 2020–2024 dönemleri ayrı ayrı ele alınarak her iki dönemde model performansları arasındaki farklar Mann-Whitney U testi ve Levene's test gibi istatistiksel yöntemlerle analiz edilmiştir. Elde edilen bulgular, pekiştirmeli öğrenme algoritmalarının özellikle kriz sonrası dönemde daha etkili portföy dağılımı öğrenebildiğini ortaya koymuştur.

Portföy Optimizasyonu

Portföy optimizasyonu, finansal piyasalarda çeşitli varlıklar arasındaki ağırlıkların, yatırımcının risk-getiri tercihleri doğrultusunda en uygun şekilde dağıtılmasını amaçlayan hesaplamalı bir süreçtir. Bu süreç, çok değişkenli finansal zaman serilerinin analizi, risk ölçütlerinin tanımlanması ve kısıtlar altında optimal çözümün bulunması gibi adımları içerir. Modern yaklaşımlar, durağan olmayan piyasa koşullarını ve yatırımcı davranışlarını modelleyebilmek adına, klasik deterministik yöntemlerin ötesine geçerek, yapay zekâ ve pekiştirmeli öğrenme gibi algoritmik teknikleri kullanmaktadır. Bu sayede, dinamik ve adaptif portföy stratejileri geliştirilebilmektedir.

Reinforcement Learning (Pekiştirmeli Öğrenme)

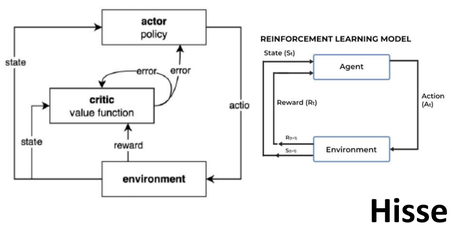
Reinforcement Learning, bir ajanın bir ortamla etkileşime girerek deneyimlerinden öğrenmesini ve ödül sinyalleri aracılığıyla optimal bir politika geliştirmesini hedefleyen, denetimsiz öğrenme ile kontrol teorisi arasında köprü kuran bir makine öğrenmesi alanıdır.

Advantage Actor-Critic (A2C):

A2C, politika tabanlı ve değer tabanlı yöntemleri birleştiren, aktör-eleştirmen mimarisi üzerinde çalışan bir pekiştirmeli öğrenme algoritmasıdır. Aktör (actor), eylem seçiminden sorumlu politikayı güncellerken; eleştirmen (critic), bu politikanın değerini tahmin ederek öğrenme sürecini yönlendirir.

Proximal Policy Optimization (PPO):

PPO, politika gradyanı yöntemlerinin kararsızlığını azaltmak amacıyla geliştirilen, klipslenmiş amaç fonksiyonu kullanan modern bir algoritmadır. Politika güncellemelerinde temkinli değişikliklere izin vererek hem istikrarlı hem de verimli öğrenme sağlar ve yüksek boyutlu ortamlarda yaygın olarak kullanılır.



Value Function:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S},$$

Action-Value Function:

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$$Policy\ gradient: E_\pi[\nabla_\theta(log\pi(s,a,\theta))R(\tau)]$$

(Policy function, Score function)

$$Update\ rule: \Delta\theta = \alpha * \nabla_\theta(log\pi(s,a,\theta))R(\tau)$$

(Change in parameters, Learning rate)

PPO's Clipped surrogate objective function:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

The Ratio Function:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

Hisse Seçimi Ve Veri Çekme

Çalışmada kullanılacak hisse senetleri, Borsa İstanbul'da uzun dönemli işlem geçmişine sahip, yüksek likiditeye ve sektörel çeşitliliğe sahip varlıklar arasından seçilmiştir. Seçim sürecinde, temsil gücü yüksek ve endeks üzerindeki ağırlığı belirgin olan şirketler tercih edilerek portföy çeşitlendirmesi sağlanmıştır.

Finansal veriler, Python programlama dili kullanılarak yfinance kütüphanesi aracılığıyla 2010–2024 dönemini kapsayacak şekilde günlük frekansta elde edilmiştir. Her bir hisse için kapanış fiyatları temel alınmış ve zaman serisi analizine uygun formatta işlenmiştir. Bu veriler daha sonra teknik analiz göstergeleri ile zenginleştirilmiştir.

Teknik Göstergelerin Eklenmesi

Modelin finansal dinamikleri daha iyi öğrenebilmesi için her hisse senedi için teknik analiz göstergeleri hesaplanarak veri setine dahil edilmiştir. Bu göstergeler; RSI (göreceli güç endeksi), MACD (hareketli ortalama yakınsama sapması), Momentum, ROC (değişim oranı), logaritmik getiri, hareketli volatilite (standart sapma), Bollinger Bantları ve işlem hacmini kapsamaktadır. Bu göstergeler Python ortamında ta, numpy ve pandas kütüphaneleri kullanılarak elde edilmiş ve ajanların gözlem uzayının bir parçası hâline getirilmiştir.

Veri Analizi

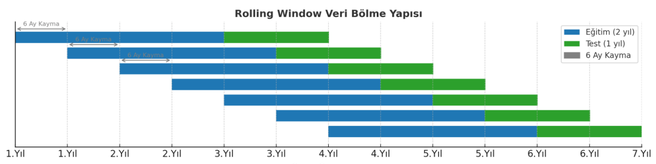
Modelleme öncesinde değişkenlerin dağılımını ve aykırı değerleri belirlemek amacıyla temel istatistiksel analizler gerçekleştirilmiştir. Kapanış fiyatları, getiriler ve teknik göstergeler; ortalama, varyans, medyan ve çeyrek değerler gibi tanımlayıcı istatistiklerle özetlenmiştir.

Eksik veriler kontrol edilmiş, teknik göstergeler zaman serisi üzerinde görselleştirilmiştir. Ortalama fiyat serisi yapısal kırılmalar açısından test edilmiş, hisse getirileri arasında korelasyon ve varyans analizleri yapılmıştır. Bu analizler, çevre tasarımı ve modelin öğrenme sürecinde temel girdi olarak kullanılmıştır.

Veri Bölme Stratejisi

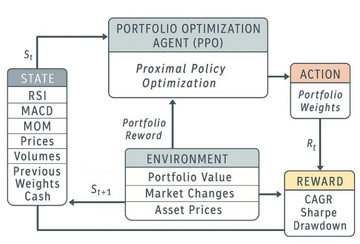
Finansal zaman serisi verilerinde, geleneksel rastgele örnekleme yöntemleri zaman bağımlılığını ihmal ettiğinden, modelin performansını gerçekçi bir şekilde değerlendiremez. Bu çalışmada, modelin geçmişe dayalı öğrenmesini ve geleceğe yönelik genelleme kabiliyetini test edebilmek amacıyla **rolling window (kayan pencere)** veri bölme yöntemi tercih edilmiştir. Her bir iterasyonda:

Eğitim dönemi (Training window): 2 yıl **, Test dönemi (Testing window):** 1 yıl **, Adım uzunluğu (Step size):** 6 ay olarak belirlenmiştir. Bu yapı sayesinde model, her döngüde güncellenen veriyle yeniden eğitilmekte ve farklı tarih aralıklarındaki test verileriyle değerlendirilerek zamana duyarlı bir doğrulama süreci gerçekleştirilmektedir. Bu yöntem, finansal piyasalardaki rejim değişimlerini yakalama ve modelin istikrarını ölçme açısından oldukça elverişlidir.



Model Tasarımı

Portföy optimizasyonu problemi, PPO (Proximal Policy Optimization) ve A2C (Advantage Actor-Critic) algoritmalarıyla pekiştirmeli öğrenme çerçevesinde modellenmiştir. Özel olarak tasarlanan Gym tabanlı ortamda ajan, teknik göstergelere (ör. RSI, MACD, Momentum, Volatilite) ve önceki portföy dağılımına dayalı olarak her adımda yeni bir portföy ağırlığı belirlemiştir. Aksiyonlar sürekli değerler olup, toplam portföy ağırlığı normalize edilmiştir ve short-selling'e izin verilmemiştir. Ödül fonksiyonu olarak log-getiri esas alınmış; belirli senaryolarda Sharpe oranı, drawdown ve işlem maliyeti gibi metrikler de ödül hesaplamasına dahil edilerek risk ayarlı öğrenme sağlanmıştır. Bu yapı, ajanın yalnızca getiri değil, aynı zamanda istikrar gözeterek karar vermesini teşvik etmiştir.




```python
class PortfolioEnvWithPenaltyV2(PortfolioEnv):
    def step(self, action):
        action = np.clip(action, 0, 1)
        action_sum = np.sum(action)
        if action_sum == 0:
            action = np.ones_like(action) / len(action)
        else:
            action /= action_sum

        if self.current_step + 1 >= len(self.data):
            obs = self._get_state()
            return obs, 0.0, True, False, {"portfolio_value": self.portfolio_value}

        next_state = self.data.iloc[self.current_step + 1]
        returns = [next_state[i * 11 + log_return_index()] for i in range(self.num_assets)]

        # Temel ödül
        reward = float(np.dot(action, returns))
```

```python
        # 🔻 Daha yumuşak işlem maliyeti cezası
        if self.action_history:
            cost = np.sum(np.abs(action - self.action_history[-1]))
            reward -= 0.0002 * cost

        # 🔒 Hafif clipping (büyük pozitif/negatifleri kırpmak için)
        reward = np.clip(reward, -0.5, 0.15)

        # Portföy güncelle
        self.portfolio_value *= np.exp(reward)
        self.current_step += 1

        obs = self._get_state()
        info = {"portfolio_value": self.portfolio_value}

        self.action_history.append(action)
        self.value_history.append(self.portfolio_value)
        self.date_history.append(self.data.index[self.current_step])

        terminated = self.current_step >= len(self.data) - 2
        truncated = False

        return obs, reward, terminated, truncated, info
```

Hiperparametre Optimizasyonu

Model başarımının artırılması amacıyla PPO algoritmasına ait hiperparametreler (öğrenme oranı, batch boyutu, n_steps, entropi katsayısı, clip_range) için Optuna kütüphanesi ile hiperparametre optimizasyonu gerçekleştirilmiştir.
Bayesyen optimizasyona dayalı TPE algoritması ile çalışan Optuna, farklı parametre kombinasyonlarını deneyerek en iyi sonuç veren yapılandırmayı belirlemiştir. Optimizasyon sürecinde performans kriteri olarak test dönemindeki portföy değeri ve risk metrikleri dikkate alınmıştır. Elde edilen optimal parametrelerle model yeniden eğitilerek nihai sonuçlar elde edilmiştir.

```python
def objective(trial):
    # Hiperparametre aralıkları
    learning_rate = trial.suggest_loguniform("learning_rate", 1e-5, 1e-3)
    n_steps = trial.suggest_categorical("n_steps", [128, 256, 512])
    batch_size = trial.suggest_categorical("batch_size", [64, 128, 256])
    ent_coef = trial.suggest_loguniform("ent_coef", 1e-5, 0.01)
    clip_range = trial.suggest_uniform("clip_range", 0.1, 0.3)
```

```
[I 2025-05-18 20:01:40,591] Trial 18 finished with value: -1.1599110196727607 and parameters: {'learning_rate': 1.9549578009487196e-05, 'n_steps': 256, 'batch_size': 128, 'ent_coef': 0.002080225982195203, 'clip_range': 0.19710539922926978}. Best is trial 1 with value: -1.5114213850448766.
[I 2025-05-18 20:07:00,154] Trial 19 finished with value: -1.4319720389954045 and parameters: {'learning_rate': 0.0003799780205913803, 'n_steps': 128, 'batch_size': 64, 'ent_coef': 0.4781571200644459e-05, 'clip_range': 0.17846261391521185}. Best is trial 1 with value: -1.5114213850448766.
🔍 En iyi Sharpe için parametreler:
{'learning_rate': 0.0002578473405734196, 'n_steps': 512, 'batch_size': 128, 'ent_coef': 0.0004584286438707457, 'clip_range': 0.2857476918832308}
```

Model Eğitimi Ve Değerlendirmesi

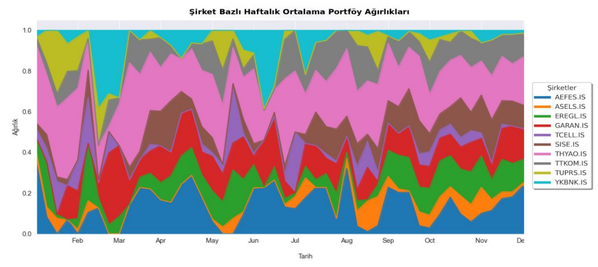
Portföy optimizasyon ajanı, PPO (Proximal Policy Optimization) algoritması ile Gym ortamında yaklaşık 200.000 zaman adımı boyunca eğitilmiştir. Eğitim sürecinde aşağıdaki metrikler izlenerek öğrenme başarımı değerlendirilmiştir:

Policy Loss: Politika ağının optimizasyon başarısı.
Value Loss: Kritik değer tahmini hatası.
Entropy: Ajanın keşif düzeyi.
Average Reward: Her adımda alınan ortalama ödül.

Eğitim istikrarı bu metriklerin seyriyle sürekli olarak izlenmiş, ani dalgalanmalar hiperparametre ayarlamalarıyla dengelenmiştir. Modelin aşırı öğrenmesini önlemek amacıyla eğitim süresi ve batch boyutları dikkatle seçilmiştir.

Aksiyon Uzayı

Ajanın çıktısı, her bir hisse senedi için [0, 1] aralığında sürekli değerlerden oluşan portföy ağırlıklarıdır.
Toplam ağırlıklar her zaman 1'e normalize edilmiştir.
Aksiyonlar: $\boldsymbol{at} = [\boldsymbol{w1}, \boldsymbol{w2}, \ldots, \boldsymbol{wn}]$
Kısıtlar: Short-selling yok, pozitif ağırlıklar ve tam yatırım.
Aksiyonlar doğrudan portföy dağılımına karşılık gelir ve her adımda yeniden dengelenir.

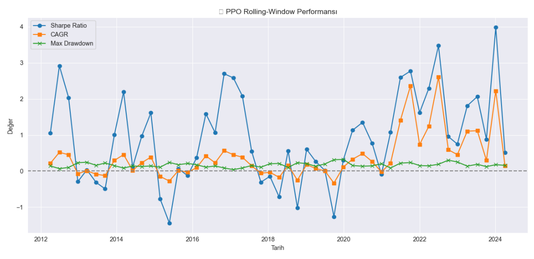


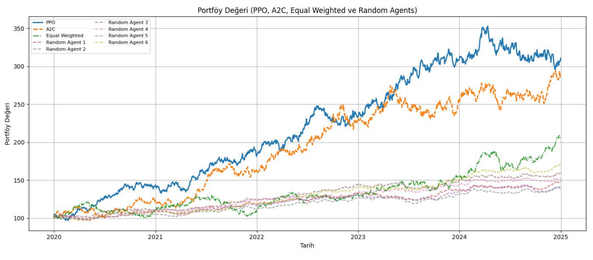
Test ve Backtest Performansı

Eğitim sonrası elde edilen modeller, 2020–2024 test döneminde değerlendirilmiştir. Her ajan (PPO, A2C) için portföy getirileri hesaplanmış ve aşağıdaki performans ölçütleriyle analiz edilmiştir:

- **CAGR (Compound Annual Growth Rate)**
- **Max Drawdown**
- **Sharpe Oranı**
- **Volatilite**

Model performansı ayrıca eşit ağırlıklı portföy, BIST100 endeksi ve rastgele ajanlarla karşılaştırılmıştır. Analiz sonucunda PPO'nun risk-getiri açısından daha istikrarlı sonuçlar verdiği gözlemlenmiştir. Ek olarak, 2010–2019 ve 2020–2024 dönemleri arasında rejim farkı istatistiksel olarak test edilmiştir.





Ekonomik Rejim Kırılımı ve Dönemsel Performans Farkı

Portföy performansı 2010–2020 ve 2020–2024 dönemleri için karşılaştırılmıştır. 2020 yılı civarında yapısal kırılma tespit edilmiş ve analiz aşağıdaki istatistiksel testlerle yapılmıştır:

Mann–Whitney U Testi: Medyan getiri, Sharpe oranı ve drawdown gibi metriklerde iki dönem arasında **anlamlı fark** olup olmadığı test edilmiştir.

Levene Testi: Dönemler arası varyans homojenliği (örneğin risk oynaklığı) incelenmiş, bazı metriklerde **varyans farklılıkları** tespit edilmiştir.

Bu testler, portföy dinamiklerinin sadece ortalama değil, aynı zamanda risk yapısında da dönemsel değişimler içerdiğini göstermektedir.

```python
import pandas as pd
from scipy.stats import mannwhitneyu
# Rejim etiketlerini ekle
results_df["regime"] = 0  # ilk dönem (2010-2020)
results2_df["regime"] = 1  # ikinci dönem (2020-2024)
# İki sonucu birleştir
full_results = pd.concat([results_df, results2_df], ignore_index=True)
# Test edilecek metrikler
metrics = ["sharpe", "drawdown", "cagr"]
# Sonuçları saklamak için liste
stat_results = []
for metric in metrics:
    group1 = full_results[full_results["regime"] == 0][metric]
    group2 = full_results[full_results["regime"] == 1][metric]
    # Mann-Whitney U testi (non-parametrik)
    stat, p_value = mannwhitneyu(group1, group2, alternative='two-sided')
    # Hangi rejim daha yüksekse yön oku ile belirt
    direction = "↑" if group2.mean() > group1.mean() else "↓"
    # Sonuçları ekle
    stat_results.append({
        "metric": metric,
        "regime1_mean": group1.mean(),
        "regime2_mean": group2.mean(),
        "p_value": p_value,
        "direction": direction,
        "significant": p_value < 0.05
    })
```

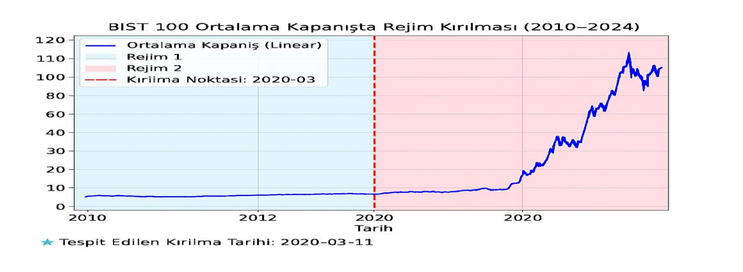
	metric	regime1_mean	regime2_mean	p_value	direction	significant
0	sharpe	0.843907	2.151480	0.027348	↑	True
1	drawdown	0.178708	0.234165	0.027348	↑	True
2	cagr	0.202675	1.327601	0.000516	↑	True

🔍 Levene's Test (Varyans Karşılaştırması)
Test statistic: 82.6740, p-değeri: 0.0000

```python
from scipy.stats import levene

# Levene's Test: varyans eşitliği testi
stat, p_value = levene(resid_1, resid_2, center='median')

# Sonuç yazdır
print(f"🔍 Levene's Test (Varyans Karşılaştırması)")
print(f"Test statistic: {stat:.4f}, p-değeri: {p_value:.4f}")
```



KAYNAKÇA

[1] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
[2] Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.
[3] Ye, Y., & Zhang, Z. (2020). Reinforcement learning for portfolio optimization with risk aversion. In *International Conference on Neural Information Processing* (pp. 370–382). Springer.

2024-2025 Bahar Yarıyılı İstatistik BÖLÜMÜ

Doğal Dil İşleme Yöntemleri ile Global Popüler Sanatçıların Şarkı Sözlerinde Duygu Analizi

Kenan Arda AĞANER 20023008

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

Bu çalışmada, globalde en popüler ilk 20 sanatçının Spotify platformunda en çok dinlenen ilk 10 şarkısının şarkı sözleri incelenmiştir. Çalışmada doğal dil işleme (NLP) yöntemleri kullanılarak şarkı sözleri üzerinden duygu analizi yapılmıştır. Sanatçıların cinsiyet bilgileri, şarkı türleri ve şarkıların yayımlandıkları dönemler de veri setine eklenerek farklı değişkenler arasında karşılaştırmalar gerçekleştirilmiştir. Analiz kapsamında duyguların müzik türleri, sanatçı cinsiyetleri ve dönemler ile ilişkisi görselleştirilmiştir. Çalışma, popüler müzikte duygu profillerini inceleyen bir yaklaşım sunmaktadır.

MÜZİĞİN DUYGUSAL HARİTASI

Müzik, evrensel bir dil olarak bireylerin duygularını ifade etme ve başkalarına aktarma konusunda güçlü bir araçtır. Günümüzde popüler müzik, yalnızca eğlence amacıyla değil; dinleyicilerin duygusal dünyasına hitap eden bir anlatım biçimi olarak da öne çıkmaktadır. Bu çalışmada, globalde en popüler sanatçıların şarkı sözleri analiz edilerek müzikte hangi duyguların ön plana çıktığı ve bu duyguların müzik türleri ile nasıl bir ilişki içerisinde olduğu incelenmiştir. Çalışmanın amacı, müzikteki duygusal eğilimleri haritalandırmak ve sanatçılar ile şarkı türleri üzerinden bir duygu profili ortaya koymaktır.

ŞARKI SÖZLERİNDE DUYGU YOĞUNLUĞU

Bu çalışmada, popüler sanatçıların şarkı sözlerinde yer alan temel duygular (öfke, iğrenme, korku, neşe, nötr, üzüntü ve sürpriz) analiz edilmiştir. Yapılan duygu analizi sonuçları, koşullu biçimlendirme yöntemi ile görselleştirilmiş; yüksek oranlar yeşil, düşük oranlar kırmızı renk ile belirtilmiştir.
Elde edilen bulgulara göre:
Korku (fear) ve Üzüntü (sadness) duyguları genel olarak sanatçılarda en yüksek oranlarda gözlemlenmiştir. Neşe (joy) oranı tüm sanatçılarda oldukça düşük seviyelerde kalmıştır. Bu durum, günümüz popüler müziğinde karamsarlık ve karmaşık duyguların daha sık işlendiğini göstermektedir.

Sanatçı	anger	disgust	fear	joy	neutral	sadness	suprise
Bruno Mars	0,13226	0,07952	0,15124	0,06264	0,23252	0,27259	0,06922
The Weeknd	0,12128	0,07877	0,28386	0,02186	0,16518	0,16812	0,10715
Lady Gaga	0,14355	0,06936	0,37093	0,05898	0,10099	0,1777	0,0584
Billie Eilish	0,08162	0,09853	0,28473	0,01344	0,13013	0,26659	0,12496
Ed Sheeran	0,09568	0,08964	0,15033	0,06567	0,24315	0,31602	0,03953
Rihanna	0,12823	0,0911	0,24427	0,11651	0,19864	0,18012	0,0411
Coldplay	0,10363	0,07255	0,36688	0,03077	0,163	0,20755	0,05563
Kendrick Lamar	0,27665	0,17503	0,14483	0,04532	0,18567	0,14559	0,0269
Taylor Swift	0,08302	0,04017	0,35280	0,03351	0,18372	0,10920	0,19758
Bad Bunny	0,18635	0,24384	0,08754	0,02117	0,01731	0,0079	0,43494
Justin Bieber	0,13014	0,08206	0,24338	0,04346	0,14114	0,29212	0,06771
Drake	0,21549	0,1014	0,29614	0,02427	0,18145	0,13466	0,04658
Ariana Grande	0,17303	0,08413	0,29173	0,04265	0,13632	0,15457	0,11752
David Guetta	0,17719	0,10018	0,1501	0,1017	0,22443	0,20571	0,04067
SZA	0,21389	0,0709	0,27063	0,06946	0,19177	0,16363	0,01973
Calvin Harris	0,06179	0,03792	0,24574	0,07453	0,17636	0,33909	0,06455
Maroon 5	0,09998	0,06976	0,25821	0,0427	0,11364	0,30327	0,11244
Dua Lipa	0,11875	0,05751	0,33336	0,03883	0,19597	0,17416	0,08143
Post Malone	0,22049	0,10102	0,20239	0,02503	0,17587	0,15508	0,12012
Katy Perry	0,1994	0,07277	0,33697	0,03116	0,1721	0,10498	0,08259

VERİ TOPLAMA SÜRECİ

Bu çalışmada kullanılacak veri seti, çeşitli güvenilir kaynaklardan manuel olarak derlenmiştir. Globalde en popüler sanatçıları belirlemek amacıyla **kworb.net** sitesinden dünya genelinde en çok dinlenen ilk 20 sanatçının verileri toplanmıştır. Her sanatçının Spotify üzerinden en çok dinlenen ilk 10 şarkısı listelenmiş ve ardından şarkı sözleri, **azlyrics.com** platformundan temin edilmiştir. Elde edilen veriler, duygu analizi ve görselleştirme için uygun hale getirilmiş ve sistematik bir şekilde düzenlenerek analiz sürecine aktarılmıştır.

DÖNEMLERE GÖRE ŞARKI DAĞILIMI

Analiz edilen şarkıların dönemsel dağılımı incelendiğinde, **2010'lar (%47,6)** ve **2020'ler (%39)** dönemlerinin çalışmada baskın olduğu görülmektedir. **2000'ler** dönemi ise daha sınırlı bir temsil oranına **(%8,5)** sahiptir. Bu sonuç, verilerin büyük ölçüde güncel müzik dönemlerine odaklandığını ve özellikle son yıllarda üretilen şarkıların duygu analizinde daha fazla yer bulduğunu göstermektedir. Ayrıca bu dağılım, son iki dönemde müzik üretiminde ve tüketiminde bir artış olduğunu destekler niteliktedir.

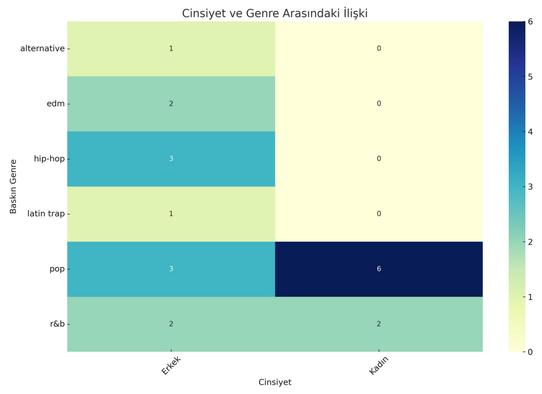
SANATÇILARIN DUYGUSAL EĞİLİMLERİ

Çalışma kapsamında analiz edilen sanatçıların baskın duygularına bakıldığında, **korku (fear)** duygusunun öne çıktığı görülmektedir. İncelenen 20 sanatçıdan 11'inin şarkılarında baskın duygunun korku olduğu belirlenmiştir. **Üzüntü (sadness)** ise 5 sanatçıda ön planda yer alırken, **öfke (anger)** ve diğer duyguların her biri yalnızca 2 sanatçıda baskın duygu olarak tespit edilmiştir. Bu sonuçlar, günümüz popüler müziğinde özellikle korku ve üzüntü gibi olumsuz duyguların sanatçıların müzikal anlatımlarında sıklıkla yer bulduğunu göstermektedir.

MÜZİK TÜRÜ; DUYGU VE CİNSİYET ARASINDAKİ İLİŞKİLER

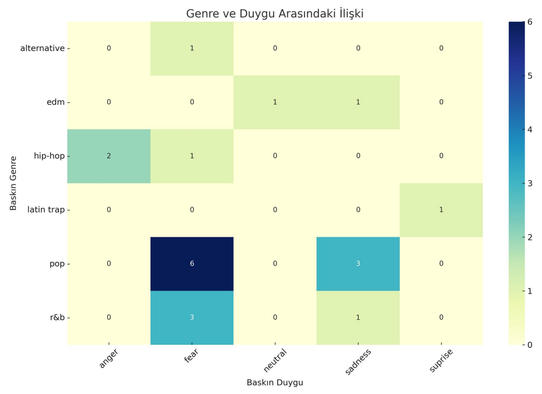
Bu çalışmada sanatçıların cinsiyetleri ile baskın müzik türleri arasındaki ilişki incelenmiştir. Görselde yer alan ısı haritasında, cinsiyetler (kadın/erkek) ile müzik türleri arasındaki frekans dağılımı gösterilmektedir. Sonuçlara göre: **Pop** müzik kadın sanatçılar arasında en baskın tür olup toplamda **6 kadın sanatçı** pop türünde yer almaktadır. Bu durum, kadın sanatçıların pop müzikte daha yoğunlaştığını göstermektedir. Erkek sanatçılar ise **daha çeşitli türlerde** dağılmıştır. Özellikle **hip-hop ve r&b** türlerinde erkek sanatçıların ağırlığı olduğu gözlemlenmiştir. Bu bulgular, müzik türü tercihinde cinsiyet farklılıklarının bulunduğunu ve özellikle pop müziğin kadın sanatçılar için öne çıkan bir alan olduğunu ortaya koymaktadır.

MÜZİK TÜRLERİNDE CİNSİYET EĞİLİMLERİ



TÜR VE DUYGU ARASI İLİŞKİ

Sanatçıların baskın müzik türleri ile şarkılarındaki baskın duygular incelenmiştir.
Pop müzik, genellikle **korku** ve **üzüntü** duyguları öne çıkmaktadır.
R&B, en çok **korku** ile eşleşmiştir.
Hip-hop, ağırlıklı olarak **öfke** ve **iğrenme** duyguları ile ilişkilidir.
EDM, alternative ve latin trap, daha çeşitli ve nadir duygularla bağlantılıdır.
Sonuçlar, müzik türlerinin şarkılardaki baskın duyguları etkileyebileceğini göstermektedir.



SANATÇILARIN CİNSİYET DAĞILIMI

Çalışmada incelenen globaldeki en popüler 20 sanatçının cinsiyet dağılımına bakıldığında, **erkek sanatçıların %65** ile çoğunlukta olduğu, **kadın sanatçıların ise %35** oranında yer aldığı görülmektedir. Bu dağılım, günümüz müzik endüstrisinde erkek sanatçıların daha yüksek oranda temsil edildiğini göstermektedir. Analiz edilen veri kümesi bu durumu yansıtmakla birlikte, cinsiyet temsiliyetine dair genel bir tablo da sunmaktadır.

SONUÇ

Bu çalışma kapsamında, globalde en popüler sanatçıların şarkı sözleri üzerinden gerçekleştirilen duygu analizinde en baskın duyguların **korku**, **üzüntü** ve **öfke** olduğu belirlenmiştir. Elde edilen bulgular, müzik türleri ile duygular arasında belirgin bir ilişki olduğunu göstermektedir. Özellikle pop, R&B ve hip-hop türlerinde belirli duyguların yoğunlaştığı dikkat çekmektedir. Ayrıca cinsiyet ve müzik türü arasındaki dağılım incelendiğinde, kadın sanatçılarda pop müziğin, erkek sanatçılarda ise hip-hop ve R&B türlerinin daha ön planda olduğu görülmüştür. Çalışmada yer alan şarkıların büyük bir kısmının 2010'lar ve 2020'ler dönemine ait olması ise araştırmanın güncel müzik eğilimlerini yansıttığını ortaya koymaktadır. Genel olarak elde edilen sonuçlar, müzik türleri ve sanatçıların şarkı sözleri aracılığıyla belirli duyguları daha sık yansıttığını ve müziğin güçlü bir kültürel ve duygusal ifade aracı olduğunu desteklemektedir.

KAYNAKÇA

[1] : Kworb. " Global Top 20 Artists by Spotify Listeners. "
https://kworb.net/spotify/listeners.html (Erişim Tarihi: 15 Haziran 2025)
[2] : Durahim, A. O., Çoşkun-Setirek, A., Başarır-Özel, B., & Kebapcı, H. (2018). *Music emotion classification for Turkish songs using lyrics*. Pamukkale University Journal of Mathematics and Engineering Sciences, 24(2), 292–301.
https://dergipark.org.tr/tr/pub/pajes/issue/36922/419724#article_cite
[3] : AZLyrics. " Şarkı sözleri "
https://www.azlyrics.com (Erişim Tarihi: 15 Haziran 2025)
[4] : Spotify. " Sanatçıların en popüler ilk 10 şarkısı "
https://open.spotify.com/ (Erişim Tarihi: 15 Haziran 2025)

İSTATİSTİK BÖLÜMÜ

LSTM ile HİSSE SENEDİ FİYAT TAHMİNİ

Mehmet Kaan GÜLER 20023002

Danışman: Doç. Dr. Gülder KEMALBAY

ÖZET

Bu çalışmada, ODAS Elektrik Üretim Sanayi Ticaret A.Ş.'nin günlük hisse senedi kapanış fiyatları kullanılarak, zaman serisi tahmini amacıyla LSTM (Long Short-Term Memory) derin öğrenme modeli uygulanmıştır. Bu çalışmada, ODAS hisse senedinin günlük kapanış fiyatları dikkate alınarak LSTM modeli ile fiyat tahmini yapılmıştır. Modelin eğitimi için veri ön işleme adımları uygulanmış, ardından zaman serisi tahmin süreci gerçekleştirilmiştir. Modelin performansı MAE ölçütüyle değerlendirilmiş ve elde edilen sonuçlar, tahminlerde başarılı bir performans ortaya koymuştur.

LSTM

LSTM (Uzun Kısa Süreli Bellek) ağı, 1997 yılında Sepp Hochreiter ve Jürgen Schmidhuber tarafından, geleneksel tekrarlayan sinir ağlarında ortaya çıkan kaybolan gradyan sorununu çözmek amacıyla geliştirilmiştir. Standart RNN yapıları, zaman içinde meydana gelen olaylar arasındaki uzun vadeli bağımlılıkları öğrenmekte yetersiz kalmakta ve bu da özellikle ardışık yapıdaki verilerin işlenmesinde ciddi performans sorunlarına yol açmaktadır. LSTM ise, özel olarak tasarlanmış hafıza hücreleri ve kapı mekanizmaları sayesinde hem kısa vadeli hem de uzun vadeli bilgiyi etkin şekilde saklayabilmekte ve gerektiğinde bu bilgileri yeniden kullanabilmektedir.

LSTM mimarisi, yalnızca teorik açıdan değil, pratik uygulamalarda da derin öğrenme alanına önemli katkılar sunmuştur. Gürültülü ve düzensiz verilerde dahi istikrarlı performans sergileyebilmesi, onu zaman serisi tahminleme gibi uygulamalarda güçlü bir araç hâline getirmiştir. Bu bağlamda, LSTM modelleri, özellikle ekonomik göstergelerin, hisse senedi fiyatlarının ve diğer finansal enstrümanların gelecekteki hareketlerinin öngörülmesinde etkili ve güvenilir bir yöntem olarak öne çıkmaktadır.

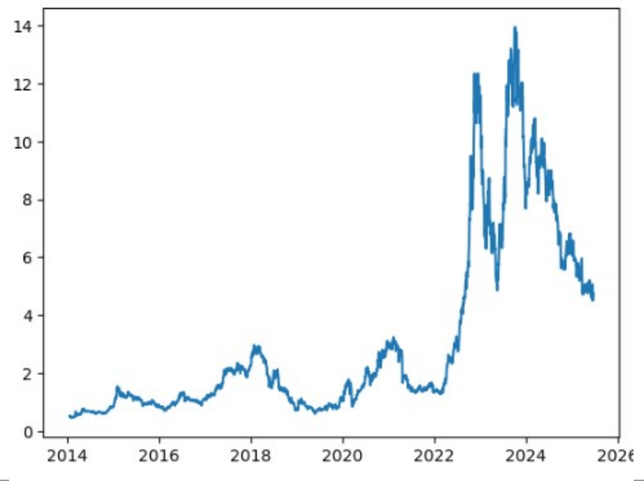
ÇALIŞMANIN AMACI

Bu çalışmanın temel amacı, LSTM (Uzun Kısa Süreli Bellek) modelinin finansal zaman serisi verileri üzerinde ne derece etkili tahminler üretebildiğini değerlendirmektir. LSTM mimarisinin özellikle gürültülü ve düzensiz yapıya sahip verilerde dahi istikrarlı performans gösterebilmesi, kısa ve uzun vadeli bağımlılıkları öğrenebilme kapasitesi gibi özellikleri, bu modeli tercih etmemde etkili olmuştur. Bu doğrultuda, ODAS hisse senedine ait günlük kapanış fiyatları kullanılarak bir LSTM modeli geliştirilmiş; modelin, geçmiş verilerden yola çıkarak gelecekteki fiyat hareketlerini tahmin edip edemeyeceği test edilmiştir.

KULLANILAN VERİ HAKKINDA

Bu çalışmada kullanılan veri seti, ODAS Elektrik Üretim Sanayi Ticaret A.Ş.'nin günlük hisse senedi fiyatlarından oluşmaktadır. Veri seti, belirli bir zaman aralığında borsa işlem günlerine ait kapanış, açılış, en yüksek, en düşük fiyatlar ile işlem hacmi ve günlük değişim yüzdesi gibi bilgileri içermektedir. Çalışmada yalnızca kapanış fiyatları dikkate alınmıştır çünkü modelin temel amacı, geçmiş kapanış değerlerine bakarak bir sonraki günün kapanış fiyatını tahmin etmektir.

Veriler zaman sırasına göre düzenlenmiş, eksik veya hatalı gözlemler kontrol edilmiştir. Modelin sağlıklı öğrenebilmesi için tüm fiyat değerleri Min-Max yöntemiyle 0 ile 1 arasında olacak şekilde ölçeklendirilmiştir. Ardından, geçmiş 60 işlem gününe ait kapanış fiyatı bilgisi kullanılarak giriş verileri (X), 61. günün kapanış fiyatı ise hedef değer (Y) olarak yapılandırılmıştır. Bu pencereleme yöntemi sayesinde model, geçmişten gelen ardışık bilgiyi kullanarak tahmin yapabilecek şekilde eğitilmiştir.

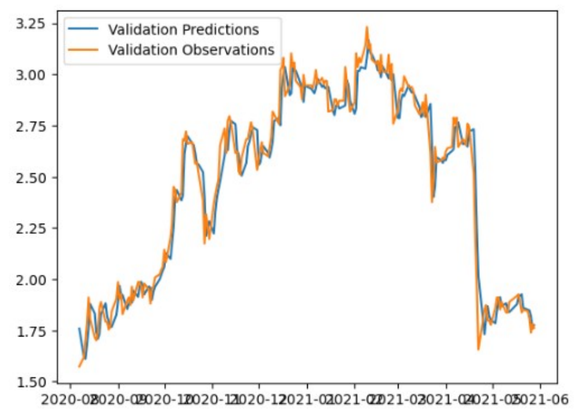


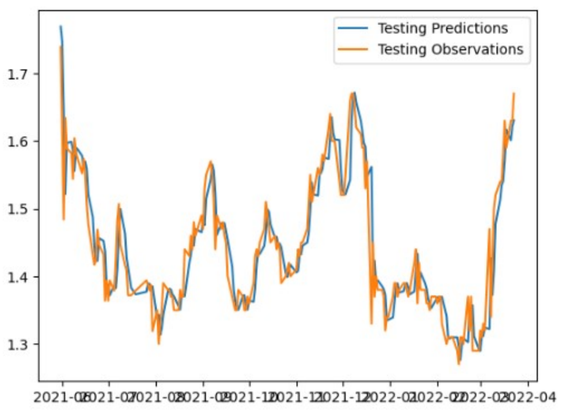
Sonuç ve Değerlendirme

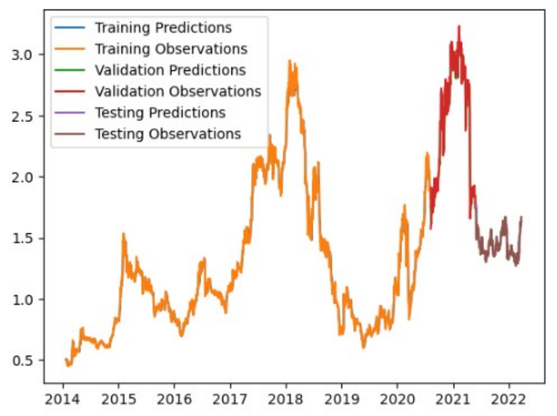
Modelin eğitim, doğrulama ve test verileri üzerinde elde ettiği tahmin sonuçları incelendiğinde, LSTM mimarisinin zaman serisi verileri üzerinde güçlü bir öğrenme performansı sergilediği gözlemlenmiştir. Eğitim verisi üzerinde tahminler gerçek gözlemlerle yüksek oranda örtüşmekte, doğrulama ve test verilerinde ise modelin genel eğilimleri ve ani değişimleri büyük ölçüde yakalayabildiği görülmektedir.

Bu bulgular, LSTM modelinin finansal piyasa verilerindeki örüntüleri öğrenme ve gelecek dönemlere dair tahminlerde bulunma konusundaki potansiyelini doğrulamaktadır. Modelin grafiksel başarımı, kısa vadeli fiyat değişimlerini anlamlandırmak ve belirli seviyelerde öngörüde bulunmak için kullanılabilir bir yapı sunduğunu ortaya koymaktadır.









Ancak elde edilen bu olumlu sonuçlara rağmen, modelin başarısı yalnızca kullanılan veri seti ve hisse senediyle sınırlıdır. ODAS hisse senedi verisi üzerinde yüksek uyum göstermiş olsa da, aynı performansın farklı sektörlerdeki ya da daha volatil finansal enstrümanlarda elde edilip edilemeyeceği kesin olarak söylenemez. Bu nedenle, bu çalışmadan çıkan sonuçlar, yalnızca örneklem üzerinde yapılan uygulama açısından geçerlidir. LSTM mimarisinin genel anlamda hisse senedi tahmininde ne derece etkili olduğu, farklı veri kümeleri, parametre ayarları ve senaryolar altında yapılacak daha kapsamlı analizlerle desteklenmelidir.

KAYNAKÇA

[1] https://www.bioinf.jku.at/publications/older/2604.pdf
[2] https://www.statisticshowto.com/absolute-error/
[3] https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

2024-2025 BAHAR YARIYILI İSTATİSTİK BÖLÜMÜ

DISNEYLAND KULLANICI YORUMLARININ LOKASYON BAZLI DUYGU VE ZAMAN SERİSİ ANALİZİ

MİNA NUR YILDIRIM, 20023022

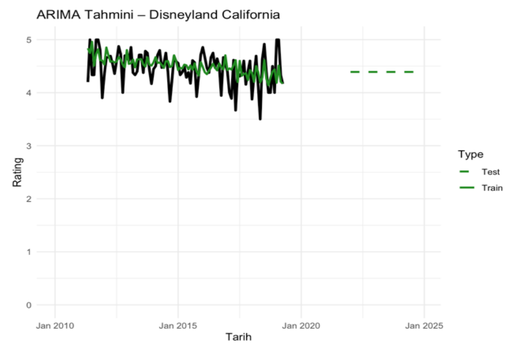
Danışman: Doç. Dr. ELİF TUNA

ÖZET

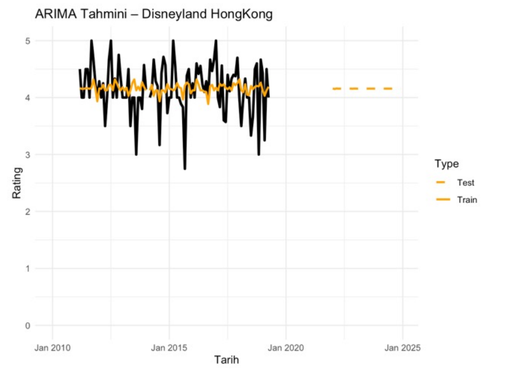
Bu tezde, Paris, Hong Kong ve California'daki Disneyland parklarına ait çevrimiçi kullanıcı yorumları; doğal dil işleme, duygu analizi, zaman serisi modelleme ve makine öğrenmesi teknikleriyle incelenmiştir. Kaggle'dan elde edilen 42.000 yorumluk veri kümesinden rastgele seçilen 2.500 örnek üzerinde yapılan analizlerde, California en yüksek memnuniyeti sunarken, Paris daha düşük skorlarla öne çıkmıştır. Duygu sınıflandırmasında CNN modeli en yüksek doğruluğa ulaşmıştır. Elde edilen sonuçlar, müşteri memnuniyetinin dijital izler üzerinden veri temelli biçimde izlenebileceğini ortaya koymaktadır.

ARIMA

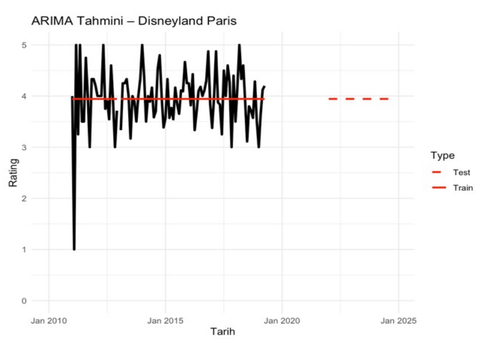
ARIMA (Otokorelasyonlu Bütünleşik Hareketli Ortalama), geçmiş verilere dayanarak gelecekteki değerleri tahmin etmeye yarayan bir zaman serisi modelidir. Serideki trend ve dalgalanmaları modelleyerek ileriye dönük öngörülerde bulunur. Özellikle düzenli yapı gösteren verilerde güçlü sonuçlar verir ve birçok alanda yaygın şekilde kullanılır (Hyndman & Athanasopoulos, 2018).



Memnuniyet puanları uzun vadede istikrarlı seyretmiş; ARIMA tahmini, gelecekte de benzer seviyelerin korunacağını öngörüyor. Zaman içinde genel olarak yüksek ve dengeli seyreden puanlar, ARIMA modeliyle 2025'e kadar 4.2 civarında sabit bir memnuniyet düzeyinin korunacağını göstermektedir. California, en istikrarlı şube olarak öne çıkmıştır.



Zaman içinde dalgalanmalar görülse de model, memnuniyetin 4 civarında sabitleneceğini tahmin ediyor. Hong Kong'da geçmişte belirgin dalgalanmalar görülmüş olsa da model, gelecekte memnuniyetin 4 civarında dengeleneceğini öngörmektedir. Dalgalı yapı gözlense de genel eğilim pozitiftir.

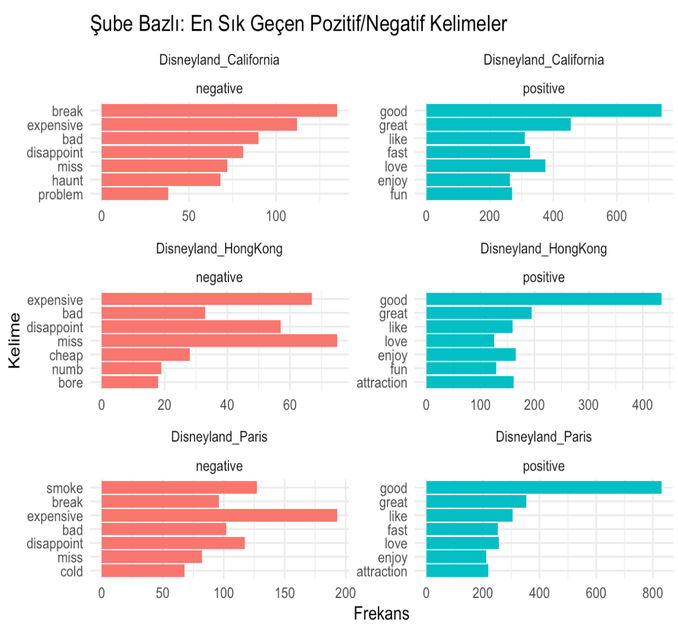


Geçmiş yıllarda en çok dalgalanma gösteren lokasyon olan Paris için model, son dönemlerde dengelenmiş bir seyir öngörmekte ve memnuniyetin yaklaşık 4 puan seviyesinde kalacağını göstermektedir.

```
> print(results)
                Branch   RMSE    MAE MAPE
1      Disneyland_Paris 0.0200 0.0200 0.57
2 Disneyland_California 0.1338 0.1233 3.08
3   Disneyland_HongKong 0.0807 0.0798 1.99
>
```

Sonuçlara göre, Disneyland Paris en düşük hata oranlarına sahip olup en başarılı tahminlerin yapıldığı şubedir. Disneyland Hong Kong orta seviyede bir performans gösterirken, Disneyland California en yüksek hata oranlarına sahiptir ve modelin en zayıf performansı bu şubede görülmektedir. Genel olarak model, Paris şubesinde oldukça isabetli sonuçlar verirken, California için iyileştirme gerektirmektedir.

DUYGU ANALİZİ



California: "Good", "great" ve "love" gibi pozitif ifadeler öne çıkarken; "expensive" ve "break" gibi kelimeler fiyat ve teknik sorunlara işaret etmektedir.

Hong Kong: Olumlu kelimeler yoğun; ancak "expensive" ve "disappoint" gibi ifadeler memnuniyetin yanında bazı beklenti sorunlarını da yansıtmaktadır.

Paris: "Good" kelimesi baskın olsa da "smoke", "cold" ve "break" gibi çevresel olumsuzluklar dikkat çekmektedir.
Bu analiz, her lokasyonda öne çıkan memnuniyet ve şikâyet temalarını belirleyerek kullanıcı deneyimlerinin temel dinamiklerini göstermektedir.

CNN Ve Diğer Derin Öğrenme Algoritmaları

CNN, verilerdeki lokal örüntüleri (kelime grupları, motifler, desenler) tanımada uzmandır. Metin analizinde, kullanıcı yorumları gibi cümleleri kelime vektörleri ile temsil eder, ardından bu vektörler üzerinde evrişim filtreleri(convolutional filters) uygulanarak anlamlı kalıplar yakalanır. Kısa ve yoğun bilgi içeren kullanıcı yorumlarında, CNN'in düşük overfitting riski ve kontekst bağımsız güçlü temsili, onu diğer modeller karşısında öne çıkarmıştır.

```
--- Disneyland_HongKong ---

Model: Simple RNN
4/4 [==============================] - 0s 2ms/step - loss: 0.2031 - accuracy: 0.9905
     loss  accuracy
0.2030604 0.9904762

Model: LSTM
4/4 [==============================] - 0s 6ms/step - loss: 0.1146 - accuracy: 0.9905
     loss  accuracy
0.1146480 0.9904762

Model: GRU
4/4 [==============================] - 0s 4ms/step - loss: 0.0706 - accuracy: 0.9905
      loss   accuracy
0.07062325 0.99047619

Model: CNN
4/4 [==============================] - 0s 1ms/step - loss: 0.1063 - accuracy: 0.9905
     loss  accuracy
0.1063038 0.9904762

Model: Dense
4/4 [==============================] - 0s 847us/step - loss: 0.0749 - accuracy: 0.9905
      loss   accuracy
0.07492465 0.99047619
>
```

```
--- Disneyland_California ---

Model: Simple RNN
7/7 [==============================] - 0s 2ms/step - loss: 0.2331 - accuracy: 0.9303
     loss  accuracy
0.2330603 0.9303483

Model: LSTM
7/7 [==============================] - 0s 5ms/step - loss: 0.2462 - accuracy: 0.9303
     loss  accuracy
0.2462297 0.9303483

Model: GRU
7/7 [==============================] - 0s 5ms/step - loss: 0.2439 - accuracy: 0.9303
     loss  accuracy
0.2439179 0.9303483

Model: CNN
7/7 [==============================] - 0s 1ms/step - loss: 0.2454 - accuracy: 0.9303
     loss  accuracy
0.2454045 0.9303483

Model: Dense
7/7 [==============================] - 0s 720us/step - loss: 0.2630 - accuracy: 0.9303
     loss  accuracy
0.2629526 0.9303483


--- Disneyland_Paris ---

Model: Simple RNN
5/5 [==============================] - 0s 2ms/step - loss: 0.7490 - accuracy: 0.5597
     loss  accuracy
0.7489539 0.5597015

Model: LSTM
5/5 [==============================] - 0s 5ms/step - loss: 0.3563 - accuracy: 0.8657
     loss  accuracy
0.3563379 0.8656716

Model: GRU
5/5 [==============================] - 0s 5ms/step - loss: 0.4503 - accuracy: 0.8433
     loss  accuracy
0.4503482 0.8432836

Model: CNN
5/5 [==============================] - 0s 1ms/step - loss: 0.4238 - accuracy: 0.8358
     loss  accuracy
0.4237882 0.8358209

Model: Dense
5/5 [==============================] - 0s 780us/step - loss: 0.3505 - accuracy: 0.8358
     loss  accuracy
0.3504826 0.8358209
```

Bu çalışmada, Disneyland tema parklarına ait kullanıcı yorumları üzerinde beş farklı derin öğrenme modeli (Simple RNN, LSTM, GRU, CNN, Dense) uygulanarak duygu sınıflandırma başarımı değerlendirilmiştir. Analizler üç lokasyon (California, Hong Kong, Paris) üzerinden yürütülmüş ve modellerin doğruluk (accuracy) ile kayıp (loss) değerleri karşılaştırılmıştır.

California lokasyonunda tüm modeller %93.03 doğruluk oranına ulaşmış, bu durum veri setinin düzenli yapısını ve sınıflar arası belirginliği yansıtmıştır. Kayıp değerlerinde ufak farklılıklar gözlense de CNN modeli, yüksek doğruluk ve düşük kayıpla birlikte etkili bir performans göstermiştir.
Hong Kong şubesinde ise modeller %99 doğruluğun üzerine çıkarak olağanüstü bir başarı sergilemiştir. GRU ve Dense modelleri en düşük kayıp değerlerine ulaşırken, CNN modeli yine güçlü ve dengeli sonuçlar vermiştir.

En dikkat çekici fark **Paris** lokasyonunda ortaya çıkmıştır. Bu şubeye ait veriler, diğerlerine göre daha fazla dalgalanma ve içerik çeşitliliği barındırdığından modellerin başarı düzeyleri düşmüştür. En iyi sonuç %86.57 doğruluk ile LSTM modelinden elde edilirken, CNN bu karmaşık yapıya rağmen %83.58 doğruluk ile oldukça rekabetçi bir performans göstermiştir. Özellikle Simple RNN modelinin başarısının %56 gibi düşük bir seviyede kalması, klasik yapının bu tarz veri setlerinde yetersiz kaldığını göstermektedir.

Genel olarak değerlendirildiğinde, **CNN modeli**, hızlı eğitim süresi, metindeki lokal örüntüleri etkin şekilde yakalaması ve genelleme gücü sayesinde üç lokasyon arasında en istikrarlı performansı sunmuştur.

XGBOOST

XGBoost (Extreme Gradient Boosting), karar ağaçlarını ardışık olarak eğiten ve hataları minimize etmeye odaklanan gelişmiş bir gradyan artırma algoritmasıdır. Hızlı, esnek ve yüksek doğruluk sağlayan yapısı sayesinde sınıflandırma ve regresyon problemlerinde sıkça tercih edilir.

Disneyland Hong Kong

```
--- XGBoost ---
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  0  0
         1  7 97

               Accuracy : 0.9327
                 95% CI : (0.8662, 0.9725)
    No Information Rate : 0.9327
    P-Value [Acc > NIR] : 0.59877

                  Kappa : 0

 Mcnemar's Test P-Value : 0.02334

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.93269
             Prevalence : 0.06731
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 0
```

Disneyland California

```
--- XGBoost ---
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0   0   0
         1  11 189

               Accuracy : 0.945
                 95% CI : (0.9037, 0.9722)
    No Information Rate : 0.945
    P-Value [Acc > NIR] : 0.579299

                  Kappa : 0

 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.945
             Prevalence : 0.055
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : 0
```

Disneyland Paris

```
-- XGBoost ---
onfusion Matrix and Statistics

          Reference
rediction   0   1
         0  10   3
         1  12 108

               Accuracy : 0.8872
                 95% CI : (0.8208, 0.9355)
    No Information Rate : 0.8346
    P-Value [Acc > NIR] : 0.05962

                  Kappa : 0.5114

 Mcnemar's Test P-Value : 0.03887

            Sensitivity : 0.45455
            Specificity : 0.97297
         Pos Pred Value : 0.76923
         Neg Pred Value : 0.90000
             Prevalence : 0.16541
         Detection Rate : 0.07519
   Detection Prevalence : 0.09774
      Balanced Accuracy : 0.71376

       'Positive' Class : 0
```

California & Hong Kong:
Doğruluk oranı yüksek (%94.5 ve %93.2) olmasına rağmen model sadece pozitif sınıfı tahmin etmiş, negatif sınıfı hiç tanıyamamıştır (Sensitivity = 0, Kappa = 0). Bu da modelin sınıflar arası ayrım gücünün olmadığını ve sadece yüzeyde başarılı göründüğünü göstermektedir.
Paris:
Doğruluk daha düşük olsa da (%88.7), model her iki sınıfı da ayırt edebilmiş ve Kappa değeri 0.51 ile anlamlı sınıflandırma başarısı göstermiştir. Dengeli doğruluk (%71.3) ve duyarlılık (%45) değerleriyle XGBoost Paris lokasyonunda en dengeli sonuçları vermiştir.

KAYNAKÇA

[1] Chillar, A. (2020). *Disneyland reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/arushchillar/disneyland-reviews

[2] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

[3] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

TÜRKİYE CUMHURİYETİ MERKEZ BANKASI PARA MÜDAHALELERİNİN BIST100 ENDEKSİ ÜZERİNDEKİ ETKİLERİNİN ANALİZİ VE ZAMAN SERİSİ MODELLERİ İLE TAHMİNLEME

MUHAMMED BESİM ARABALI - 21023611

Arş. Gör. Dr. Hülya Yürekli

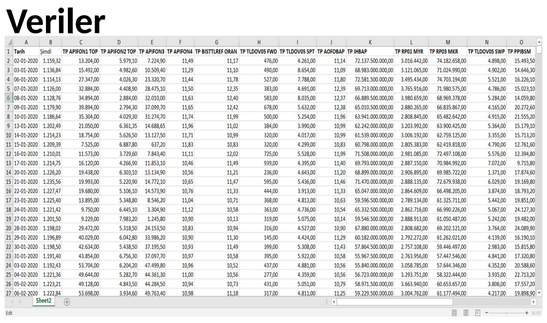
ÖZET

Finansal piyasalardaki belirsizliklerin öngörülebilmesi, yatırımcılar ve politika yapıcılar açısından kritik öneme sahiptir. Bu doğrultuda, zaman serisi analizleri finansal göstergelerin gelecekteki seyrini tahmin etmede etkili araçlardır. "Türkiye Cumhuriyet Merkez Bankası Para Müdahalelerinin BIST100 Endeksinin Analizi ve Zaman Serisi Modelleri ile Tahmini" başlıklı bu çalışma, TCMB'nin para politikası araçları ile BIST100 endeksi arasındaki ilişkiyi inceleyerek endeksin gelecekteki değerlerini tahmin etmeyi amaçlamaktadır. Veri seti, TCMB Elektronik Veri Dağıtım Sistemi'nden indirmiştir; faiz oranları, toplam fonlama, likidite oranı, swap, repo, döviz rezervleri gibi makroekonomik verilerden oluşmaktadır. İlk aşamada regresyon analizi yapılmış, anlamlı değişkenler belirlenmiş ve çoklu bağlantılar korelasyon matrisi ile değerlendirilmiştir. Daha sonra SARIMA ve SARIMAX modelleri kullanılarak zaman serisi analizlerine geçilmiştir. Model parametreleri ''Grid Search'' yöntemiyle optimize edilmiş, serilerin trend, mevsimsellik ve durağanlık özellikleri dikkate alınarak en uygun tahmin modeli belirlenmiştir. Python programlama dili kullanılarak yürütülen analizler, TCMB'nin para müdahalelerinin BIST100 üzerindeki etkisinin belirli ölçüde öngörülebilir olduğunu ortaya koymuştur.

Çalışmanın Amacı

Bu çalışmada, Türkiye Cumhuriyet Merkez Bankası'nın para müdahalelerinin BIST100 endeksi üzerindeki etkileri analiz edilmiştir. Para politikalarının finansal piyasalara yansımaları, zaman serisi teknikleriyle incelenmiş ve endeksin gelecekteki değerleri tahmin edilmeye çalışılmıştır. Veriler, TCMB'nin Elektronik Veri Dağıtım Sistemi üzerinden indirilmiştir; politika faizi, gecelik faiz oranları, toplam fonlama, likidite oranı, swap, repo, döviz rezervleri gibi veriler kullanılmıştır. Veri ön işleme sürecinde zaman bilgisi uygun biçimde düzenlenmiş, değişken seçimi korelasyon analizi ve p-değerleri ile yapılmıştır. İlk olarak regresyon modeli denenmiş, temel varsayımların sağlanamaması üzerine zaman serisi modellerine geçilmiştir. Serilerin trend, mevsimsellik ve durağanlık özellikleri incelenmiş; SARIMA ve SARIMAX modelleri uygulanmıştır. SARIMAX modelinde parametre seçimi Grid Search yöntemiyle optimize edilmiştir. Elde edilen sonuçlarla, TCMB'nin para müdahalelerinin BIST100 endeksi üzerindeki etkileri belirli bir düzeyde öngörülebilir bulunmuştur.

Veriler



Tarih	Şimdi	TP APIFON1 TOP	TP APIFON2 TOP	TP APIFON3	TP APIFON4	TP BISTTLREF ORAN	TP TLDOV05 FWD	TP TLDOV05 SPT	TP AOFOBAP	TP IHBAP	TP RPG1 MYR	TP RPG5 MKR	TP TLDOV05 SWP	TP PPIBSM
02-01-2020	1.159,32	13.204,00	5.979,10	7.224,90	11,49	11,17	476,00	4.261,00	11,14	72.137.500.000,00	3.016.442,00	74.182.658,00	4.898,00	15.493,50
03-01-2020	1.134,13	27.347,00	4.026,30	23.320,70	11,44	11,78	527,00	7.788,00	11,80	72.581.500.000,00	3.495.434,00	74.703.194,00	5.521,00	16.226,50
07-01-2020	1.126,00	32.884,00	4.408,90	28.475,10	11,50	12,35	383,00	4.691,00	12,39	69.713.000.000,00	3.765.916,00	71.980.575,00	4.786,00	15.023,50
08-01-2020	1.128,78	34.894,00	2.884,00	32.010,00	11,63	12,40	583,00	8.035,00	12,37	66.889.500.000,00	3.980.659,00	68.969.378,00	5.284,00	14.059,80
09-01-2020	1.179,90	39.888,00	2.794,30	37.099,70	11,65	12,42	678,00	5.932,00	12,38	65.010.500.000,00	3.885.285,00	66.835.867,00	4.583,00	20.272,60
10-01-2020	1.186,64	35.304,00	4.029,30	31.274,70	11,74	11,99	500,00	5.254,00	11,96	63.941.000.000,00	2.808.845,00	65.482.642,00	4.915,00	21.555,20
13-01-2020	1.202,49	21.050,00	6.361,35	14.688,65	11,96	11,02	384,00	3.990,00	10,99	62.242.000.000,00	3.203.992,00	63.900.425,00	5.364,00	15.179,50
14-01-2020	1.214,23	18.754,00	5.626,50	13.127,50	10,99	10,99	320,00	4.017,00	10,99	61.518.000.000,00	3.006.192,00	62.759.125,00	3.955,00	15.713,20
15-01-2020	1.209,39	7.525,00	6.887,80	637,20	11,83	10,83	320,00	4.299,00	10,83	60.798.000.000,00	2.805.383,00	62.419.818,00	4.790,00	12.761,60
16-01-2020	1.210,01	11.573,00	3.729,00	7.843,40	11,11	12,02	725,00	5.528,00	11,99	71.508.000.000,00	2.981.085,00	72.497.108,00	5.576,00	12.394,80
17-01-2020	1.214,75	14.120,00	12.853,10	1.266,90	11,49	11,49	939,00	4.395,00	11,49	69.793.000.000,00	2.887.150,00	70.984.992,00	7.072,00	9.715,80
20-01-2020	1.226,20	19.438,00	6.303,10	13.134,90	10,56	11,21	236,00	4.643,00	11,20	68.899.000.000,00	2.906.895,00	69.985.722,00	1.371,00	17.874,60
21-01-2020	1.235,56	19.993,00	5.220,90	14.772,10	10,65	11,47	595,00	5.436,00	11,46	71.470.000.000,00	2.888.135,00	72.679.938,00	6.029,00	19.169,80
22-01-2020	1.227,47	18.680,00	5.159,10	14.573,90	10,76	11,31	444,00	3.913,00	11,31	65.047.000.000,00	2.864.609,00	66.498.205,00	4.450,00	18.793,20
23-01-2020	1.225,40	13.895,00	5.348,80	8.546,20	11,04	10,71	368,00	4.813,00	10,63	59.596.500.000,00	2.789.134,00	61.325.731,00	5.442,00	19.851,00
24-01-2020	1.221,42	9.750,00	6.445,10	3.304,90	11,12	10,58	363,00	4.736,00	10,54	65.332.500.000,00	2.862.716,00	66.990.226,00	5.067,00	24.127,30
27-01-2020	1.201,50	9.229,00	7.983,20	1.245,80	10,90	10,13	319,00	5.075,00	10,14	59.546.500.000,00	2.888.913,00	61.050.487,00	5.242,00	19.482,00
28-01-2020	1.198,02	29.472,00	5.318,50	24.153,50	10,83	10,94	316,00	4.527,00	10,90	67.880.000.000,00	2.808.682,00	69.202.121,00	3.764,00	24.089,90
29-01-2020	1.196,89	40.029,00	6.042,80	33.986,20	10,90	11,30	145,00	4.424,00	11,29	60.182.000.000,00	2.792.272,00	61.262.021,00	4.139,00	16.190,50
30-01-2020	1.198,50	42.634,00	5.518,50	37.105,50	10,90	11,43	399,00	5.308,00	11,43	57.864.500.000,00	2.757.338,00	59.446.497,00	2.983,00	15.833,80
31-01-2020	1.191,40	43.854,00	6.756,30	37.097,70	10,97	10,58	395,00	5.922,00	10,58	55.967.500.000,00	2.763.956,00	57.447.546,00	4.841,00	17.320,80
03-02-2020	1.192,41	33.704,00	6.204,20	27.499,80	10,96	10,52	437,00	4.881,00	10,56	55.840.000.000,00	3.058.785,00	57.644.346,00	4.352,00	20.588,60
04-02-2020	1.221,36	49.644,00	5.282,70	44.361,30	11,00	10,56	277,00	4.359,00	10,56	56.723.000.000,00	3.293.750,00	58.322.644,00	3.935,00	22.713,20
05-02-2020	1.223,21	49.128,00	4.843,50	44.284,50	10,94	10,73	431,00	5.055,00	10,79	58.971.500.000,00	3.663.940,00	60.653.657,00	3.806,00	17.557,20
06-02-2020	1.222,84	53.698,00	3.934,60	49.763,40	10,98	11,18	317,00	4.811,00	11,25	59.229.500.000,00	3.004.762,00	61.177.494,00	4.217,00	19.898,90

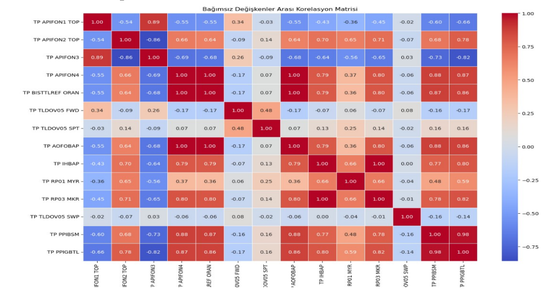
Tarih	TP PR ARZ22	TP TRY MT01 K	TP API REP ORT G1	TP KREDİ L001	TP AB TOPLAM	TP KTF10
03-01-2020	2.569.584.940,00	13,00	10,65	2.562.176.144,00	105.421,00	15,18
10-01-2020	2.555.949.159,80	12,00	10,10	2.552.662.302,00	106.214,00	15,46
17-01-2020	2.575.240.452,40	11,50	9,25	2.560.827.013,00	106.455,00	15,39
24-01-2020	2.595.304.908,00	11,70	8,56	2.572.741.710,00	104.351,00	15,13
31-01-2020	2.593.993.417,10	10,25	7,70	2.595.585.732,00	102.375,00	14,79
07-02-2020	2.610.610.356,00	9,25	7,26	2.602.096.285,00	102.759,00	14,56
14-02-2020	2.627.395.235,10	9,35	7,29	2.617.000.530,00	104.056,00	14,63
21-02-2020	2.648.212.182,50	13,50	9,50	2.627.969.087,00	102.670,00	14,14
28-02-2020	2.668.567.402,80	13,50	10,08	2.671.319.794,00	107.823,00	14,10
06-03-2020	2.671.354.776,60	14,95	11,76	2.672.260.904,00	106.834,00	14,50
13-03-2020	2.702.441.040,50	17,50	14,71	2.708.516.560,00	104.164,00	14,46
20-03-2020	2.768.840.190,60	20,25	16,50	2.749.744.247,00	93.447,00	14,60
27-03-2020	2.779.689.082,30	18,80		2.768.395.332,00	95.406,00	15,35
03-04-2020	2.837.123.166,70	18,25	18,50	2.807.290.412,00	89.553,00	16,10
10-04-2020	2.874.292.259,70	20,25		2.833.365.382,00	88.918,00	15,75
17-04-2020	2.961.826.430,50	19,80	20,50	2.899.082.661,00	87.909,00	9,45
24-04-2020	3.024.993.102,10	19,65		2.954.654.689,00	86.446,00	8,27
01-05-2020	3.041.072.501,70	19,35	20,50	2.998.580.516,00	86.242,00	8,99
08-05-2020	3.098.626.658,60	19,35		3.047.472.697,00	85.831,00	10,08
15-05-2020	3.090.886.212,70	19,30		3.036.170.973,00	84.428,00	10,79
22-05-2020	3.072.844.731,30	19,30		3.028.738.129,00	92.406,00	11,24
29-05-2020	3.058.185.606,10	19,10		3.052.801.812,00	91.025,00	11,94
05-06-2020	3.058.905.035,50	18,30		3.067.306.917,00	92.219,00	12,21
12-06-2020	3.091.988.755,20	30,00	15,51	3.097.535.724,00	93.061,00	12,65
19-06-2020	3.134.671.645,30	23,00		3.114.677.339,00	91.592,00	12,93
26-06-2020	3.146.336.530,20	22,00		3.141.042.957,00	90.237,00	12,87

Veri toplama sürecinde, Türkiye Cumhuriyet Merkez Bankası'nın Elektronik Veri Dağıtım Sistemi üzerinden araştırmalar kapsamında doğrudan müdahale verileri indirilmiştir. İlgili veriler, 01 Ocak 2020 ile 31 Aralık 2024 tarihleri arasında günlük ve haftalık frekansta, Excel(.xlsx) formatında toplanarak analiz için hazırlanmıştır. Bu değişkenler, para politikalarının piyasa üzerindeki etkilerinin incelenebilmesi amacıyla seçilmiştir.

Veri Hazırlığı

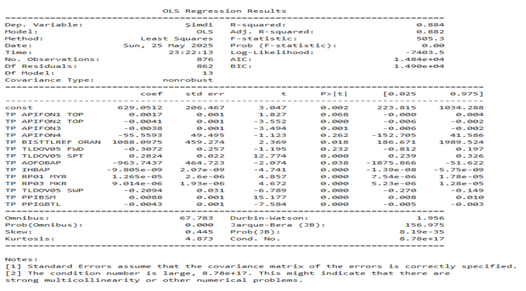
Bağımlı ve bağımsız değişkenler Python ortamına aktarıldıktan sonra, "Tarih" değişkeni datetime formatına dönüştürülerek veri çerçevesinin indeksine atanmış ve zaman serisi analizlerine uygun hale getirilmiştir. Bu sayede analiz, görselleştirme ve modelleme süreçleri zaman temelli yapıya göre düzenlenmiştir. Ayrıca çalışmanın konusunun asıl amacı ''Müdahale'' olduğu için boş değerlere herhangi bir değişiklikte bulunulmamıştır.

Korelasyon Matrisi



Bağımsız değişkenler arasındaki ilişkileri belirlemek amacıyla bir korelasyon matrisi oluşturulmuştur. Bu yöntem, yüksek korelasyon katsayısına sahip değişken çiftleri incelenmiş; benzer bilgiyi içeren değişkenlerden yalnızca günlük frekanslı veriler seçilmiştir. Böylece hem boş değeri olan haftalık frekanslı veriler çıkarılmıştır, hem çoklu doğrusal bağlantı riski azaltılmıştır.

Regresyon Modeli

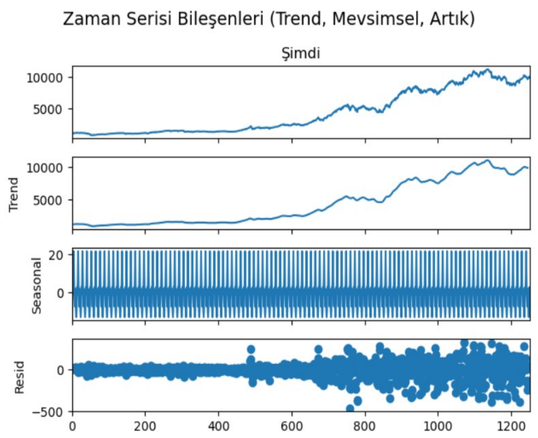


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Simdi   R-squared:                       0.884
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     585.3
Date:                Sun, 25 May 2025   Prob (F-statistic):               0.00
Time:                        23:22:13   Log-Likelihood:                -7403.5
No. Observations:                 876   AIC:                         1.486e+04
Df Residuals:                     862   BIC:                         1.493e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        629.0512    206.467      3.047      0.002     223.815    1034.288
TP APIFON1 TOP  -0.0017      0.001     -1.827      0.068      -0.000      0.004
TP APIFON2 TOP  -0.0041      0.001     -3.552      0.000      -0.006      -0.002
TP APIFON3      -0.0038      0.001     -3.494      0.001      -0.006      -0.002
TP APIFON4     -55.5593     49.495     -1.123      0.262    -152.705     41.586
TP BISTTLREF ORAN 1088.4975 459.274     2.369      0.018     186.671    1989.324
TP TLDOV05 FWD  -0.3072      0.257     -1.195      0.232      -0.812      0.197
TP TLDOV05 SPT   0.2824      0.022     12.774      0.000       0.239      0.326
TP AOFOBAP    -963.7437    464.723     -2.074      0.038   -1875.866    -51.622
TP IHBAP      -9.005e-09  2.07e-09     -4.341      0.000  -1.31e-08  -4.75e-09
TP RPG1 MYR    1.265e-05  2.6e-06      4.857      0.000    7.54e-06   1.78e-05
TP RPG3 MKR    9.014e-06  1.93e-06     4.672      0.000    5.23e-06   1.28e-05
TP TLDOV05 SWP  -0.2094     0.031     -6.789      0.000      -0.270     -0.149
TP PPIBSM       0.0008     0.001     15.177      0.000       0.000      0.010
TP PPIGBTL     -0.0043     0.001     -7.584      0.000      -0.005     -0.003
==============================================================================
Omnibus:                       67.783   Durbin-Watson:                   1.956
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              156.795
Skew:                           0.445   Prob(JB):                     8.19e-35
Kurtosis:                       4.873   Cond. No.                     8.78e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.78e+17. This might indicate that there are
strong multicollinearity or other numerical problems.
```

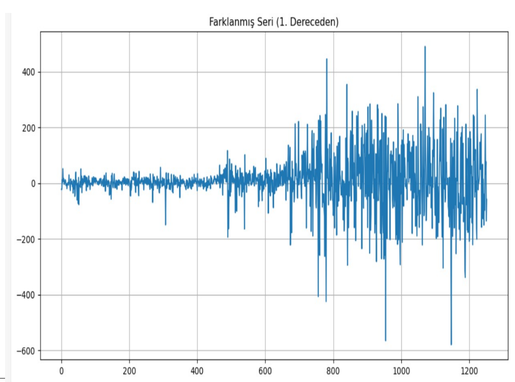
Bağımsız değişkenlerin, bağımlı değişken üzerindeki etkilerini istatistiksel olarak değerlendirmek amacıyla bir regresyon modeli oluşturulmuştur. Modelde yer alan değişkenlerin p-değerleri incelenmiş; 0.05'in üzerinde anlamlılık düzeyine sahip değişkenler modelden çıkarılmıştır.

Zaman Serisi Analizi



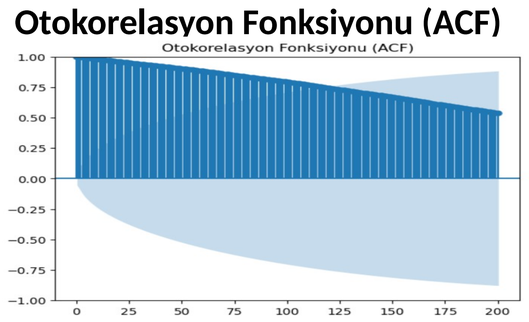
Orijinal (Şimdi) Zaman Serisi: Ham zaman serisinde genel olarak pozitif bir trend gözlemlenmekte olup, özellikle 600. gözlemden sonra belirgin bir artış eğilimi söz konusudur.
Trend Bileşeni:Seri başlangıçta yatay seyretmekle birlikte, 600. gözlem sonrasında belirgin bir yükseliş göstermekte ve uzun vadeli artan bir trendin varlığına işaret etmektedir.
Mevsimsel Bileşen: Seride sabit genlikte ve düzenli aralıklarla tekrarlayan mevsimsel döngüler mevcuttur; bu durum düzenli bir mevsimsel yapı olduğunu göstermektedir.
Artık Bileşen: Artık bileşen, serinin başlangıcında düşük varyanslı iken sonlara doğru varyansı artmakta ve bu durum olası heteroskedastisite ile birlikte bazı uç değerlerin varlığını ortaya koymaktadır.

Durağanlaştırma (Fark Alma) İşlemi



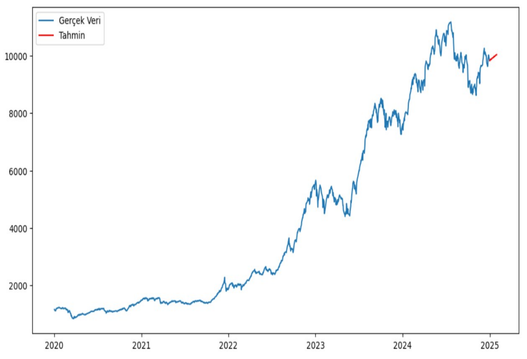
Seri üzerinde uygulanan ADF testi sonucunda, ADF test istatistiği yaklaşık olarak 0.0898, p-değeri ise 0.9654 olarak hesaplanmıştır. Bu yüksek p-değeri, serinin durağan olmadığına işaret etmektedir. Dolayısıyla, zaman serisinin analizinde sağlıklı sonuçlar elde edebilmek adına serinin durağanlaştırılması gerekliliği ortaya çıkmıştır. Bu amaçla uygulanan birinci dereceden fark alma işlemi sonrasında, ADF test istatistiği yaklaşık olarak -6.44'e düşmüş, p-değeri ise 0.0000000162 (1.62×10^{-8}) olarak bulunmuştur. Bu sonuçlar, serinin istatistiksel açıdan anlamlı düzeyde durağan hale getirildiğini güçlü bir şekilde ortaya koymaktadır.

Otokorelasyon Fonksiyonu (ACF)



ACF grafiği, zaman serisinin otokorelasyon yapısını görsel olarak ortaya koymakta ve ilk gecikmelerde otokorelasyon katsayılarının bire yakın değerler alması, seride belirgin bir trendin varlığına ve dolayısıyla durağan olmama durumuna işaret etmektedir. ACF değerlerinin %95 güven aralığının dışında kalması, istatistiksel olarak anlamlı otokorelasyonların bulunduğunu göstermektedir. Bu yapı, seride geçmiş gözlemlerin mevcut değerler üzerinde etkili olduğunu ortaya koymaktadır.

Sarıma Modeli



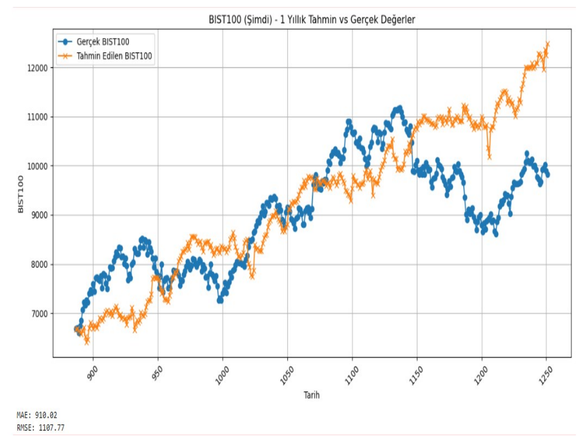
Bağımlı değişken kullanılarak oluşturulan SARIMA modeline ilişkin sonuçlar incelendiğinde, grafikte mavi çizgi gerçek BIST100 endeksi değerlerini, kırmızı çizgi ise model tarafından tahmin edilen değerleri temsil etmektedir. Tahmin sonuçları, önümüzdeki dönemde BIST100 endeksinde belirgin bir artış eğilimi öngörüldüğünü ortaya koymaktadır.

Grid Search

Zaman serisi modellemesinde parametrelerin veri odaklı ve sistematik şekilde belirlenmesi için Grid Search yöntemi uygulanmıştır. Bu yöntemle test edilen kombinasyonlar arasında en düşük hata değerini veren SARIMAX(1, 1, 1, 90) modeli en uygun parametre seti olarak seçilmiştir.

SARIMAX Modeli

"SARIMAX (Genişletilmiş Mevsimsel Otoregresif Entegre Hareketli Ortalama model) modeli, klasik ARIMA modelinin mevsimsel bileşenler ve dışsal değişkenler eklenerek genişletilmiş halidir. Bu model, zaman serilerinde hem trend, hem mevsimsellik hem de diğer dışsal etkilerin analizini sağlar. SARIMAX, (p,d,q)×(P,D,Q)m parametreleri ile tanımlanır; burada p,d,q sırasıyla otoregresif, fark alma ve hareketli ortalama terimlerinin derecelerini, P,D,Q ise mevsimsel bileşenlerin derecelerini gösterir ve mmm mevsimsel periyodu belirtir. Ayrıca model, dışsal değişkenler ile serinin açıklayıcısı olarak kullanılarak tahmin performansı artırılır. SARIMAX modelleri özellikle ekonomik, finansal ve çevresel zaman serilerinde geniş uygulama alanı bulmuştur."[3]



Modelin bir yıllık tahmin performansına ilişkin grafik, tahmin edilen BIST100 değerlerinin gerçek verilerle genel eğilim açısından büyük ölçüde örtüştüğünü göstermektedir. Özellikle yükseliş ve düşüş trendlerinde modelin piyasa dinamiklerini başarılı biçimde yansıttığı görülmektedir. Ortalama mutlak hata (MAE) 910.02, kök ortalama kare hata (RMSE) ise 1107.77 olarak hesaplanmış; bu değerler, finansal zaman serilerinin doğasında bulunan oynaklık dikkate alındığında makul düzeydedir. RMSE'nin MAE'den yüksek olması bazı dönemlerde sapmaların yaşandığını gösterse de genel tahmin başarısı güçlüdür. Sadece geçmiş verilere dayalı olarak yapılan tahminler, gerçek verilerle karşılaştırıldığında modelin öngörü gücünü ortaya koymuştur.

```
                            SARIMAX Results
==========================================================================================
Dep. Variable:                          Simdi   No. Observations:                 1252
Model:             SARIMAX(1, 1, 1)×(1, 1, 1, 90)   Log Likelihood           -6593.541
Date:                            Mon, 26 May 2025   AIC                      13225.082
Time:                                    16:14:22   BIC                      13319.597
Sample:                                         0   HQIC                     13260.888
                                             - 1252
Covariance Type:                              opg
==========================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
TP APIFON1 TOP -3.037e-05  7.02e-05     -0.433      0.665     -0.000      0.000
TP APIFON2 TOP     0.0002   5.58e-05      3.550      0.000    8.88e-05      0.000
TP APIFON3     -8.824e-05   3.89e-05     -2.267      0.023     -0.000     -1.2e-05
TP APIFON4      -1.2467   5.58e-09  -2.23e+08      0.000      -1.247      -1.247
TP BISTTLREF ORAN  27.6953   4.16e-08   6.66e+08      0.000      27.695      27.695
TP TLDOV05 FWD  -0.0099   6.36e-05   -155.656      0.000      -0.010      -0.010
TP TLDOV05 SPT      0.002   3.45e-06      0.585      0.559      -0.004       0.002
TP AOFOBAP     -26.2481   3.85e-08  -6.82e+08      0.000     -26.248     -26.248
TP IHBAP       1.199e-10   2.73e-10      0.439      0.660   -4.15e-10    6.55e-10
TP RPG1 MYR    1.599e-06   1.3e-06       1.225      0.220   -9.58e-07    4.16e-06
TP RPG3 MKR    -1.045e-07   2.93e-07     -0.356      0.722    -6.8e-07    4.71e-07
TP TLDOV05 SWP     0.0025      0.000      6.427      0.000       0.002       0.003
TP PPIBSM      -0.0003   5.64e-05     -5.403      0.000      -0.000      -0.000
TP PPIGBTL      0.0002   6.25e-05      2.595      0.009    3.97e-05       0.000
ar.L1           0.0194   3.13e-06   6216.009      0.000       0.019       0.019
ma.L1           0.0248   2.79e-06   8906.844      0.000       0.025       0.025
ar.S.L90       -0.4005   9.52e-06   -4.21e+04     0.000      -0.401      -0.400
ma.S.L90       -0.3227   7.04e-06   -4.58e+04     0.000      -0.323      -0.323
sigma2         1.751e+04   6.11e-11   2.87e+14     0.000    1.75e+04    1.75e+04
==========================================================================================
Ljung-Box (L1) (Q):                   0.25   Jarque-Bera (JB):               371.97
Prob(Q):                              0.62   Prob(JB):                         0.00
Heteroskedasticity (H):              27.88   Skew:                            -0.12
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.88
==========================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 2.4e+30. Standard
```

SARIMAX modeli, zamansal ve mevsimsel bileşenleri başarıyla içeren yapısıyla zaman serisine güçlü bir uyum sağlamıştır. İstatistiksel olarak anlamlı bulunan model parametreleri, kısa vadeli dalgalanmalar ve dönemsel etkilerin etkili biçimde yansıtıldığını göstermektedir. Ljung - Box testi artıkların otokorelasyondan arındığını doğrularken, düşük AIC ve BIC değerleri modelin uyum ve yalınlık açısından güçlü olduğunu ortaya koymaktadır. Bu sonuçlar, modelin yüksek öngörü gücüne sahip güvenilir bir araç olduğunu göstermektedir.

Sonuçlar

Bu çalışmada, BIST100 endeksine ait zaman serisi verileri kullanılarak SARIMA ve SARIMAX modelleri ile tahminleme yapılmıştır. Serinin durağanlaştırılmasının ardından, uygun parametreler Grid Search yöntemiyle belirlenmiş ve modelleme sürecine geçilmiştir. SARIMA ve SARIMAX modelleri genel olarak serinin yapısal özelliklerini başarılı şekilde yansıtmış; özellikle SARIMAX modeli, bağımsız değişkenlerin etkisini de dikkate alarak daha güçlü ve açıklayıcı sonuçlar sunmuştur. Her iki modelin de otoregresif ve mevsimsel bileşenleri anlamlı bulunmuş, düşük AIC ve BIC değerleri ise modellerin uyum ve basitlik açısından yeterli olduğunu ortaya koymuştur. Bir yıllık tahmin sonuçları, gerçek değerlerle yüksek düzeyde örtüşmüş; RMSE ve MAE gibi performans ölçütleri, finansal seriler için kabul edilebilir hata düzeylerinde kalmıştır. Elde edilen bulgular, SARIMA ve SARIMAX modellerinin BIST100 endeksinin gelecekteki eğilimlerini öngörmede etkili ve güvenilir araçlar olduğunu göstermektedir.

KAYNAKÇA

[1] Türkiye Cumhuriyet Merkez Bankası, Resmi Web Sitesi, "Tarihçe
[2] Borsa İstanbul Resmi Web Sitesi
[3] Yılmaz, M. (2019). Zaman Serileri Analizinde SARIMAX Modelleri ve Uygulamaları (Doktora Tezi). İstanbul Üniversitesi Fen Bilimleri Enstitüsü

2024-2025Yarıyılı İSTATİSTİK BÖLÜMÜ

Kredi Kartı Dolandırıcılığının Makine Öğrenmesi Yöntemleri İle Tespiti

Muhammed Efe YAMANOĞLU 20023024

Prof. Dr. Ali Hakan BÜYÜKLÜ

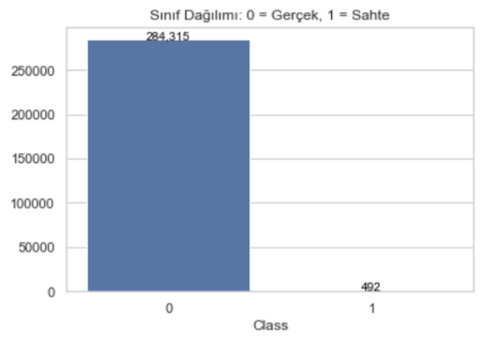
ÖZET

Dijital teknolojilerin yaygınlaşmasıyla birlikte günlük hayatta kredi kartı kullanımı önemli ölçüde artmış, bu durum beraberinde kredi kartı dolandırıcılığı gibi ciddi finansal güvenlik sorunlarını da getirmiştir. Bu ihtiyaçtan yola çıkılarak gelişen makine öğrenmesi yöntemleri, özellikle büyük veri ortamlarında dolandırıcılık tespiti alanında etkili sonuçlar sunmaktadır. Bununla birlikte, sınıf dengesizliği, model açıklanamamazlığı ve aşırı öğrenme gibi sorunlar, bu yöntemlerin uygulanabilirliğini sınırlayabilmektedir. Özellikle "kara kutu" niteliğindeki modellerin karar mekanizmalarının anlaşılamaması, bu sistemlere duyulan güveni azaltmaktadır. Bu çalışmanın temel amacı, kredi kartı dolandırıcılığını etkili ve açıklanabilir bir şekilde tespit edebilecek bir makine öğrenmesi modeli tasarlamaktır. Bu kapsamda XGBoost algoritması tercih edilmiş, modelin hiperparametreleri Optuna aracılığıyla optimize edilmiş ve SHAP (SHapley Additive exPlanations) yöntemiyle modelin öğrenme kararları açıklanabilir hale getirilmiştir.

VERİ HAKKINDA

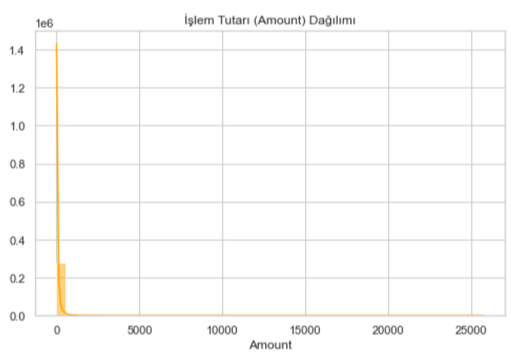
Bu çalışmada kullanılan veri seti Eylül 2013 tarihinde Avrupa merkezli bir bankanın sistemlerinden anonimleştirilerek toplanmıştır. Veri *Machine Learning Group – Université Libre de Bruxelles (ULB)* tarafından büyük veri madenciliği ve dolandırıcılık tespiti üzerine yürütülen bir araştırmada kullanılmak üzere toplanmıştır ve Kaggle platformunda "Credit Card Fraud Detection" ismi ile yayınlanmıştır.

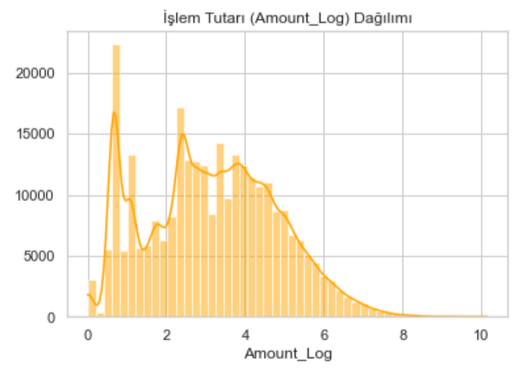
Veri toplamda 284.807 gözlemden oluşmaktadır ve bu gözlemlerden 492 adeti dolandırıcılık verisidir. Bu durum veri setinin son derece dengesiz (Imbalanced) bir sınıf dağılımına sahip olduğunu göstermektedir. Dolandırıcılık vakalarının oranı yalnızca %0.127'dir. Veri setinde yer alan toplam 31 değişkenden 28'i, temel bileşen analizi (Principal Component Analysis - PCA) uygulanarak elde edilmiş ve gizlilik amacıyla anonimleştirilmiştir. Bu işlem yüzünden bileşenleri tek başına inceleyerek herhangi bir çıkarımda bulunmak mümkün değildir.



VERİ ÖN İŞLEME

Veri seti incelendiğinde eksik yada kayıp veriye rastlanmamıştır. Time (Zaman) ve Amount (İşlem Tutarı) değerlerinde aykırı değerler gözlenmiştir. Time değişkeni ilk bankacılık işleminden itibaren geçen saniyeyi gösterdiği için bu değer dönüştürülerek saat haline getirilmiş ve veri setine 'Hour' bileşeni olarak tekrar eklenmiştir. Amount değişkeni gerçekleşen işlem tutarını göstermektedir. Büyük işlemler çok az sayıda gerçekleşmiştir ve bu değerler genellikle uç değerler (Outlier) olarak değerlendirilir. Veriyi normalize hale getirmek için logaritmik dönüşüm uygulanmış ve veri tekrardan Amount_Log olarak veri setine eklenmiştir. Bu işlemler modelin daha iyi sonuçlar vermesi ve yorumlanabilirliğinin artması için önemlidir. Bu çalışmada veri %70 eğitim verisi, %15 test verisi ve %15 doğrulama (Validation) verisi olmak üzere 3 parçaya ayrılmıştır.

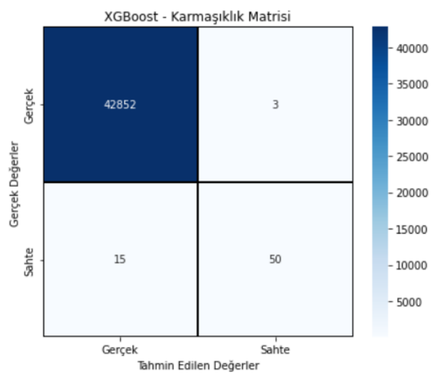




XGBOOST ALGORİTMASI UYGULANMASI

Bu çalışmada XGBoost modeli Python programı üzerinden xgboost kütüphanesi kullanılarak oluşturulmuştur. Model ilk olarak varsayılan olarak kurulmuş ve herhangi bir düzeltme yapılmamıştır.

Bu uygulamada, kredi kartı dolandırıcılığı olabilecek işlemlerin tespiti için XGBoost algoritması kullanılarak model kurulmuştur. Standart hale getirilmiş veriye ait eğitim verisi ve doğrulama verisi kullanılmıştır.

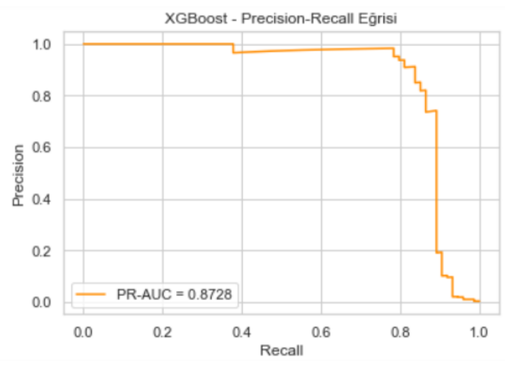


Karmaşıklık matrisine göre gerçekleşen işlemlerden 42852 tanesi ''gerçek'' işlem olarak doğru tahmin edilmiştir. Gerçekleşen 3 gerçek işlem ''sahte'' olarak değerlendirilmiştir ve yanlış tespit edilmiştir. Dolandırıcılık olduğu bilinen 15 işleme ''gerçek'' denerek yanlış tahmin edilmiştir. Gerçekte dolandırıcılık olduğu bilinen 50 işlem doğru tahmin edilmiştir. Matris incelendiğinde gerçek işlemleri tespit etme oranı çok yüksek gözükse bile modelin durumuna karar vermek için erkendir.

Modelin tüm tahminlerinin doğru tahmin edildiğini gösteren Accuracy (Doğruluk) değeri %99.96 olarak bulunmuştur. Bu oran tek başına incelendiğinde modelin mükemmel çalıştığını söyleyebiliriz. Ama çok büyük ve dengesiz veri setlerinde tek başına yeterli ve doğru bir veri metriği değildir. Doğru tahmin edilen işlemlerin gerçekten doğru olduğunu gösteren Precision (Keskinlik) değeri %95.16 olarak bulunmuştur. Bu sonuç modelin doğru tahminlerinin yüksek seviyede güvenilir olduğunu göstermektedir. F1 skoru ise %86.76 olarak bulunmuştur. Bu değer ile modelin ortalama üstü performans gösterdiği söylenebilir.

PR AUC (Precision-Recall Area Under Curve) değeri, modelin pozitif sınıfı ne kadar etkili tespit ettiğini ölçer. Bu metrik, modelin Precision ve Recall değerleri arasındaki ilişkiyi tüm eşik değerleri boyunca gösteren eğrinin altında kalan alanı temsil eder. PR AUC değeri 0.8728 olarak bulunmuştur. ROC AUC (Receiver Operating Characteristic - Area Under Curve) değeri, modelin pozitif ve negatif sınıfları ne kadar başarılı şekilde ayırt edebildiğini ölçer. Bu metrik, modelin tüm sınıflandırma eşik değerleri boyunca gösteren ROC eğrisinin altında kalan alanı ifade eder. ROC AUC değeri 0.9815 olarak bulunmuştur.

Metrik	Değer	Yüzde%
Accuracy (Doğruluk)	0.9996	99.96%
Precision (Kesinlik)	0.9516	95.16%
Recall (Duyarlılık)	0.7973	79.73%
F1 Score	0.8676	86.76%
ROC AUC	0.9815	
PR AUC	0.8728	



MODELİN OPTİMİZE EDİLMESİ

Çalışmada kullanılan veri setinin sınıf dağılımı oldukça dengesiz olduğundan, sınıf dengesini sağlamak amacıyla ilk aşamada SMOTE (Synthetic Minority Over-sampling Technique) yöntemi uygulanmış ve test verisine sentetik gözlemler eklenmiştir. Ancak, bu yöntemin uygulanmasının ardından oluşturulan modeller beklenen performansı sağlayamamıştır. Bu nedenle alternatif bir yöntem olarak, XGBoost modelinde yer alan scale_pos_weight parametresi kullanılmış ve pozitif sınıfın ağırlığı artırılarak modelin dengesiz sınıflara karşı duyarlılığı iyileştirilmeye çalışılmıştır.Bu sürecin ardından, modelin performansını artırmak amacıyla hiperparametre optimizasyonu gerçekleştirilmiş ve bu amaçla Optuna kütüphanesi kullanılarak mevcut model sistematik şekilde yeniden yapılandırılmıştır. Modelin karar verme sürecinin yorumlanabilirliğini artırmak ve hangi değişkenlerin tahminler üzerinde etkili olduğunu belirlemek amacıyla SHAP (SHapley Additive exPlanations) analizi uygulanmıştır. Bu yöntemin tercih edilme nedeni, veri setinin oldukça büyük olması ve değişkenlere uygulanan PCA (Principal Component Analysis) yöntemi ile boyut indirgemeye gidilmiş olmasıdır. Ayrıca, müşterilere ait kişisel verilerin korunması amacıyla değişkenlerin anonimleştirilmiş olması da model yorumlamasını güçleştirmiştir. Bu nedenle, model üzerinde yüksek etkiye sahip değişkenlerin tespit edilmesi ve yalnızca bu değişkenler kullanılarak modelin yeniden eğitilmesi, çalışmanın güvenilirliği ve açıklana bilirliği açısından önem arz etmektedir. Yapılan SHAP analizi sonucunda, model üzerinde etkisi yüksek olan değişkenler belirlenmiş ve bu değişkenlerle oluşturulan yeni bir veri seti ile model yeniden eğitilerek performansı değerlendirilmiştir. Yeni modelde kullanılan değişkenler 'V14', 'V4', 'V12', 'V10', 'V8', 'V7', 'V16', 'V19', 'V3', 'V11', 'V18', 'Hour'.

MODELİN TEKRARDAN KURULMASI

Metrik	Değer	Yüzde%
Accuracy	0.9994	99.94%
Precision	0.9023	90.23%
Recall	0.8108	81.08%
F1 Score	0.8540	85.40%
ROC AUC	0.9695	
PR AUC	0.8529	

Yeni kurulan modelin metrik değerleri incelenmiştir. Yeni modele ait Accuracy değerinde küçük bir düşüş (-0.0002) gözlemlenmiştir. Precision değerinde düşüş (-0.0493) gözlemlenmiştir. Recall değerinde artış(+0.0135) gözlemlenmiştir. F1 skorunda düşüş (-0.0136) gözlemlenmiştir. ROC AUC değerinde düşüş (-0.012) gözlemlenmiştir. PR AUC değerinde düşüş (-0.0199) gözlemlenmiştir. Burada iki modeli kıyaslarken dikkat etmemiz gereken en önemli metrik Recall metriğidir. Bunun sebebi modellerin çok büyük veri seti içerisinde bulunan çok az sayıdaki dolandırıcılık verisini ne kadar iyi tespit ettiklerini kıyaslamaktır.

SONUÇ

Bu çalışmada, makine öğrenmesi algoritmalarından XGBoost kullanılarak, ULB-MLG (2015) Credit Card Fraud veri seti üzerinden kredi kartı dolandırıcılığını tespit etmeye yönelik modeller geliştirilmiştir. Model performansını artırmak amacıyla Optuna ile hiperparametre optimizasyonu yapılmış; ayrıca veri dengesizliğini gidermek için SMOTE ve scale_pos_weight yöntemlerinden yararlanılmıştır. Başlangıçta varsayılan parametrelerle eğitilen modelin Recall (duyarlılık) değeri %79.73 olarak hesaplanmıştır. Daha sonra, Optuna ile optimize edilmiş parametreler ve SMOTE yöntemiyle belirlenen önemli değişkenler kullanılarak alternatif bir model oluşturulmuştur. Bu yeni model, %81.08 düzeyinde Recall değeri ile önceki modele kıyasla daha yüksek performans göstermiştir.Sonuç olarak, XGBoost algoritması ve uygun veri işleme teknikleriyle, ULB-MLG veri seti üzerinde gerçekçi ve güvenilir tahminler üretebilen etkili bir model elde edilmiştir.

KAYNAKÇA

[1] Breiman, L. (2001). *Random forests*. **Machine Learning**, *45*(1), 5–32.

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). *Calibrating probability with undersampling for unbalanced classification*. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE.

İSTATİSTİK BÖLÜMÜ

ANKARA'DA KONUT FİYATLARININ MAKİNE ÖĞRENMESİYLE TAHMİNLENMESİ

Münir Göktan Soykurt 20023048

Danışman: Prof.Dr.Ersoy Öz

ÖZET

Bu çalışma Ankara'daki konut fiyatlarını tahmin etmek amacıyla Derin Öğrenme, Random Forest ve Lineer Regresyon makine öğrenmesi modelleri kullanılmıştır. Veri seti, çeşitli emlak ilanlarından elde edilen **m², oda sayısı, bina yaşı, mahalle, ilçe** gibi özellikleri içermektedir. Model eğitimi sonrasında **Doğrusal Regresyon**, **Random Forest** ve **Derin Sinir Ağı (DNN)** sonuçları karşılaştırılmış; en yüksek başarı **DNN modeli** ile elde edilmiştir. Performans karşılaştırmaları R^2 ve RMSE metrikleriyle yapılmıştır. **SHAP analizi** ile en etkili değişkenlerin **m² (Net)**, **ilçe**, **bina yaşı** ve **oda sayısı** olduğu belirlenmiştir. Elde edilen bulgular, modellemenin hem tahmin gücünü hem de yorumlanabilirliğini göstermektedir.

Veriyle Konut Fiyatlarını Anlamak: Neden Gerekli?

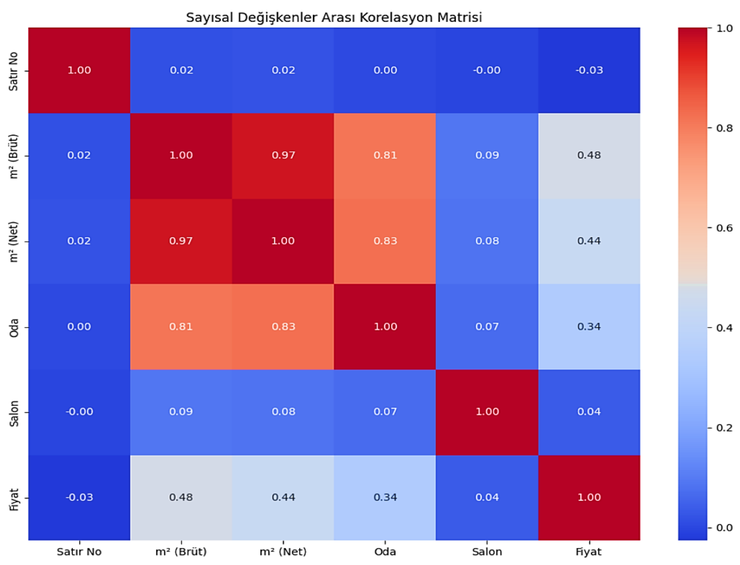
Konut fiyat tahmini, bir evin piyasa değerini belirlemek amacıyla çeşitli fiziksel ve konumsal özelliklere dayanarak yapılan öngörüdür.
Özellikle büyük şehirlerde konut fiyatlarındaki değişkenlik, yatırım kararları ve alım-satım süreçlerinde belirsizlik yaratmaktadır.
Makine öğrenmesiyle doğru fiyat tahmini yapılması; satıcılar için rekabetçi fiyat belirleme, alıcılar için gerçekçi bütçeleme ve yatırımcılar için kârlı bölgelerin tespiti açısından büyük önem taşır.

Veri Seti

Bu çalışma için kullanılan veri seti, çevrim içi emlak ilan platformlarından elde edilen Ankara iline ait konut verilerini içermektedir. Veri seti, konutların fiziksel, konumsal ve ilan özellikleri olmak üzere toplam 10,925 gözlem ve 21 değişkenden oluşmaktadır.
Veri setinde; m² (brüt/net), oda sayısı, bina yaşı, ilçe, mahalle, ısıtma tipi, kat durumu, site içi bilgisi ve fiyat gibi değişkenler yer almaktadır.
Veriler analiz öncesinde eksik ve aykırı gözlemlerden arındırılmış, kategorik değişkenler one-hot encoding yöntemi ile sayısallaştırılmıştır. Sayısal değişkenler ise StandardScaler ile ölçeklendirilmiştir. Veri seti, modelleme sürecinde eğitim ve test olmak üzere %80-%20 oranında ikiye ayrılmıştır.

Değişken	Açıklama
Satır No	Her konut ilanına ait benzersiz sıra numarası.
İlan Tarihi	İlanın sisteme yüklendiği tarih.
Emlak Tipi	Konutun türü (Daire, Müstakil, Dubleks vb.).
m² (Brüt)	Konutun toplam brüt alanı (duvarlar dahil).
m² (Net)	Konutun yaşanabilir net iç alanı.
Oda	İlanda belirtilen toplam oda sayısı.
Salon	Salona sahip olup olmadığı veya salon sayısı.
Bina Yaşı	Binanın inşa yılından bu yana geçen yıl sayısı.
Bulunduğu Kat	Konutun binadaki bulunduğu kat bilgisi.
Kat Sayısı	Binanın toplam kat sayısı.
Isıtma	Kullanılan ısıtma sistemi (Doğalgaz, Sobalı vb.).
Banyo Sayısı	Konutta bulunan banyo sayısı.
Mutfak	Mutfak tipi (Amerikan, Kapalı vb.).
Balkon	Balkonun olup olmadığı (evet/hayır).
Asansör	Asansör bulunup bulunmadığı.
Otopark	Otopark imkânının olup olmadığı.
Site İçerisinde	Konutun site içerisinde yer alıp almadığı.
İl	İlanın bulunduğu şehir (tüm veriler Ankara).
İlçe	Konutun bulunduğu ilçe.
Mahalle	Konutun bulunduğu mahalle.
Fiyat	Konutun satış fiyatı (hedef değişken).

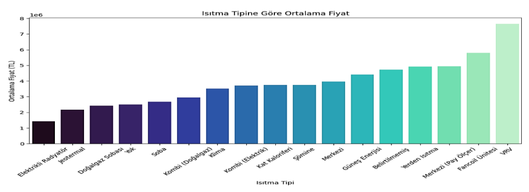
Değişken	Tür
Satır No	int64
İlan Tarihi	object
Emlak Tipi	object
m² (Brüt)	float64
m² (Net)	float64
Oda	float64
Salon	int64
Bina Yaşı	object
Bulunduğu Kat	object
Kat Sayısı	object
Isıtma	object
Banyo Sayısı	object
Mutfak	object
Balkon	object
Asansör	object
Otopark	object
Site İçerisinde	object
İl	object
İlçe	object
Mahalle	object
Fiyat	int64

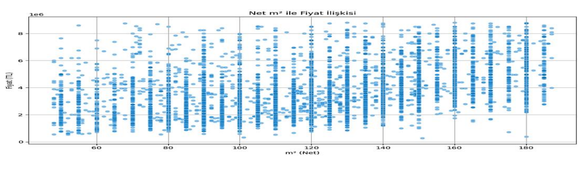


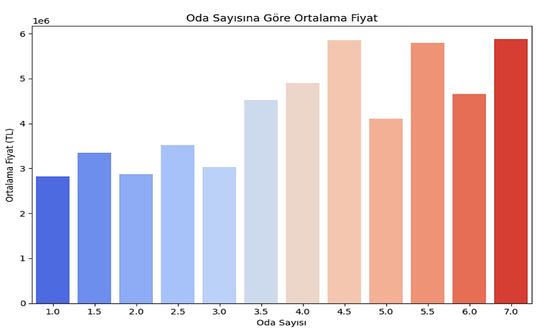
Yandaki korelasyon matrisi, veri setinde yer alan sayısal değişkenler arasındaki doğrusal ilişkiyi göstermektedir. Brüt ve net metrekare değerleri arasında çok yüksek korelasyon (r = 0.97) gözlenmiştir. m² (Brüt) ve oda sayısı gibi değişkenlerin fiyatla olan korelasyonu orta seviyededir (r ≈ 0.48 ve r ≈ 0.34). Salon sayısı ise fiyatla oldukça zayıf bir ilişki göstermektedir (r ≈ 0.04). Bu analiz, fiyat tahmininde kullanılacak değişkenlerin belirlenmesi açısından önem taşımaktadır.

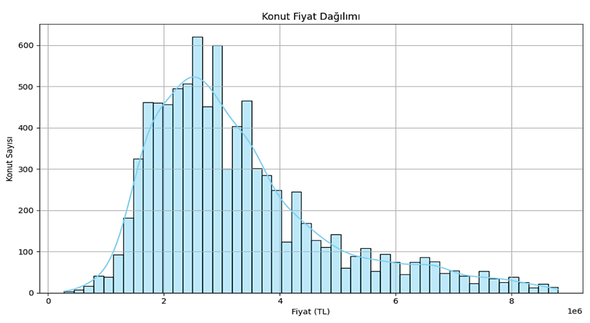
Betimsel İstatistik Grafikleri

Aşağıda sunulan betimsel istatistik grafikleri, bağımlı değişkenimiz olan fiyat (TL) ile bağımsız değişkenler arasındaki ilişkileri göstermektedir. Bu grafikler aracılığıyla, fiyatın konum, metrekare (brüt/net) ve oda sayısı gibi değişkenlerden nasıl etkilendiği gözlemlenebilmektedir. Sayfanın alt kısmında yer alan dağılım grafiği ise, fiyatların hangi aralıkta yoğunlaştığını ortaya koyarak, veri setindeki fiyat dağılımının sağa çarpık bir yapı sergilediğini göstermektedir.

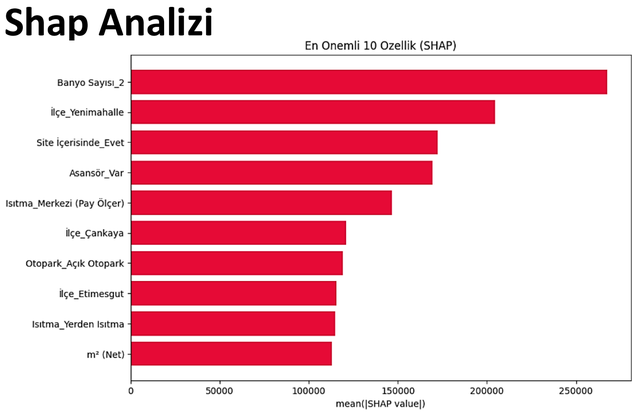








Shap Analizi

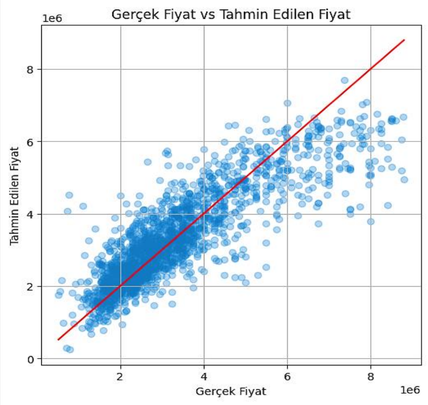


Makine öğrenmesi modellerinin kararlarını açıklamak amacıyla kullanılan bir yöntemdir. Her bir özelliğin model tahminine ne kadar ve hangi yönde katkı sağladığını sayısal olarak gösterir. Bu çalışmada SHAP analizi ile konut fiyatlarını etkileyen en önemli 10 değişken belirlenmiştir. Aşağıdaki grafikte görüldüğü üzere, banyo sayısı, ilçe bilgisi (Yenimahalle, Çankaya, Etimesgut), site içinde olup olmama durumu, asansör ve ısıtma tipi gibi özellikler, modelin fiyat tahmininde en yüksek etkiye sahip değişkenlerdir. Bu analiz, modelin yalnızca yüksek doğruluk sağlamadığını, aynı zamanda yorumlanabilir olduğunu da göstermektedir.

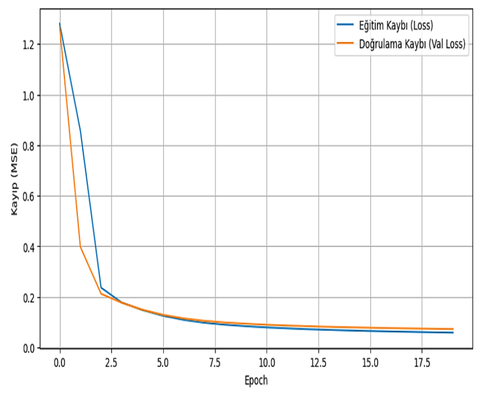
Makine Öğrenmesi Modelleri

Derin Öğrenme Modeli (DNN)

Bu çalışmada kullanılan Derin Sinir Ağı modeli, konut fiyatları gibi karmaşık ve doğrusal olmayan ilişkileri öğrenme konusunda oldukça başarılı olmuştur. Model, farklı sayısal ve kategorik değişkenleri katmanlar aracılığıyla işleyerek yüksek doğrulukta tahminler üretmiştir.

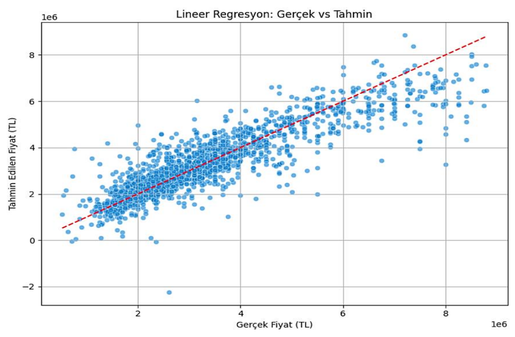


Yukardaki grafikte, DNN modelinin tahmin ettiği fiyatlarla gerçek fiyatlar karşılaştırılmıştır. Noktalar diyagonale (kırmızı çizgi) ne kadar yakınsa, tahminler o kadar doğrudur. Model, yüksek fiyat aralıklarında biraz sapma gösterse de genel olarak başarılı bir tahmin performansı sergilemiştir.



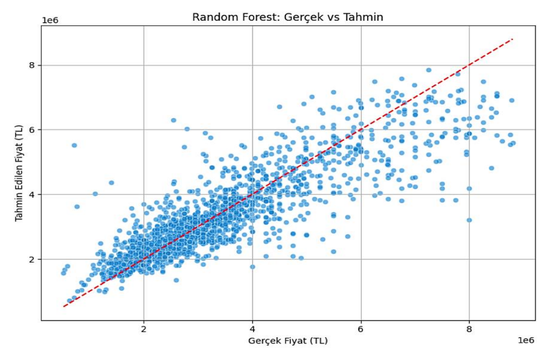
Eğitim ve Doğrulama Kayıpları (MSE)
Modelin eğitim ve doğrulama kayıpları hızlıca düşerek düşük bir seviyede dengelenmiştir. Eğitim ve doğrulama eğrilerinin birbirine yakın olması, modelin aşırı öğrenme yapmadan konut fiyatlarını başarıyla tahmin ettiğini göstermektedir.

Lineer Regresyon



Grafik, Lineer Regresyon modelinin konut fiyatlarını ne derece doğru tahmin ettiğini göstermektedir. Noktaların büyük kısmı kırmızı referans çizgisine yakın seyretse de, özellikle yüksek fiyatlı konutlarda sapmalar dikkat çekmektedir. Bu durum, modelin doğrusal varsayımının karmaşık ilişkileri tam olarak yansıtamadığını göstermektedir.

Random Forest



Grafikte, modelin tahmin ettiği fiyatlar ile gerçek konut fiyatları karşılaştırılmıştır. Noktaların kırmızı kesikli referans çizgisine yakın seyretmesi, modelin yüksek doğrulukta tahminler ürettiğini göstermektedir. Özellikle orta fiyat aralıklarında tahmin performansı güçlüdür; ancak yüksek fiyat segmentlerinde sapmalar gözlemlenmektedir.

Konut Fiyat Tahmin Modellerinin Karşılaştırması

Model	R^2	MAE	RMSE	MAPE (%)
Derin Öğrenme	0.7854	515122	747672	16.96
Linear Regression	0.7779	519913	760678	17.29
Random Forest	0.75	538924	807120	17.03

KAYNAKÇA

[1] Li, X., & Li, L. (2020). A housing price prediction model based on ensemble learning. IEEE Access, 8, 20785–20793
[2] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), 30.
[3] Koç, E., & Durdu, A. (2020). Derin öğrenme yöntemleri ile konut fiyatlarının tahmini: İstanbul örneği. Afyon Kocatepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 22(2), 355–372.

2024-2025 Bahar Yarıyılı İstatistik BÖLÜMÜ

İSTANBUL'DA PM10 HAVA KİRLİLİĞİ SEVİYELERİNİN ZAMAN SERİSİ ANALİZİ

NİLSU GÜNCE GÜRSOY , 20023059

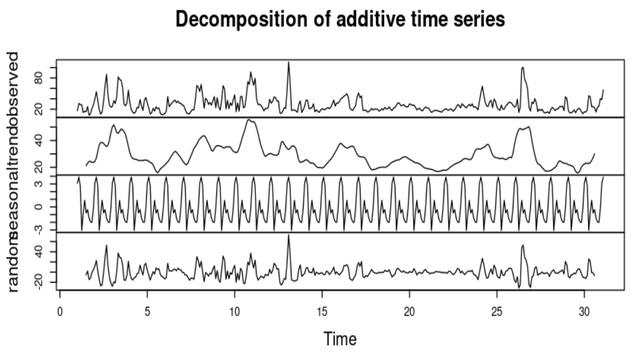
Danışman: Doç. Dr. ELİF TUNA

ÖZET

Bu çalışmada, İstanbul'daki bir hava kalitesi izleme istasyonundan elde edilen günlük PM10 verileri zaman serisi yöntemleriyle analiz edilmiştir. SARIMA, ETS ve SES modelleriyle tahminler üretilmiş; ayrıca modelleme sürecinde eğitim ve test veri setleri kullanılarak tahmin performansı değerlendirilmiştir. En başarılı sonuç SARIMA modeliyle elde edilmiştir.

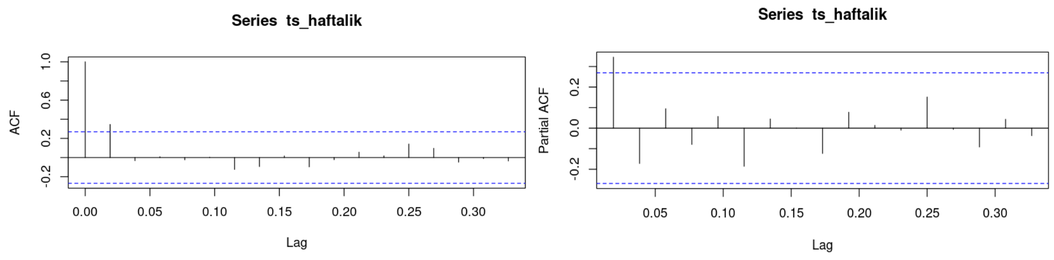
ZAMAN SERİLERİ

Zaman serileri, belirli aralıklarla toplanan veri noktalarının ardışık yapısını ifade eder. Bu analiz türü, geçmiş verilere dayanarak eğilimleri belirlemeyi ve geleceği öngörmeyi amaçlar. Durağanlık, zaman serisi analizinin temel koşullarındandır. Yaygın kullanılan modeller arasında ARIMA, SARIMA ve ETS yer alır.



Zaman serisinin yapısal bileşenlerini incelemek amacıyla additif ayrıştırma yöntemi uygulanmıştır. Grafik, gözlemlenen seriyi dört bileşene ayırarak analiz etmeyi sağlar. Üst panelde, PM10 değerlerinde dalgalı yapı ve zamanla artan yoğunluk dikkat çeker. İkinci panelde, uzun dönemli eğilim belirgin şekilde izlenmektedir.

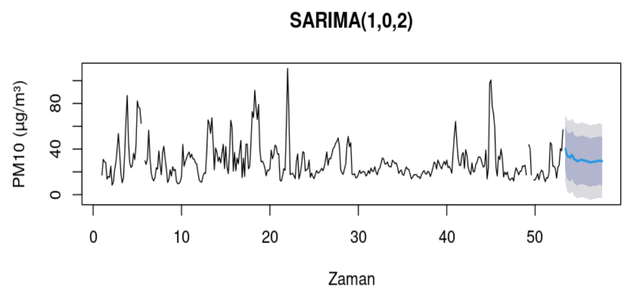
Üçüncü panelde güçlü ve düzenli bir mevsimsel döngü gözlemlenir; bu, haftalık tekrar eden desenlerin varlığına işaret eder. Alt panelde ise model tarafından açıklanamayan rastgele bileşen yer alır. Bu ayrıştırma, serideki yapısal desenlerin anlaşılmasını ve uygun model seçimlerini destekler.



ACF grafiğinde ilk birkaç gecikmede anlamlı otokorelasyon gözlemlenmiş, bu durum seride hareketli ortalama (MA) bileşeninin varlığına işaret etmiştir. PACF grafiğinde yalnızca birinci gecikme anlamlı çıkmış, bu da AR(1) yapısının uygun olabileceğini göstermektedir. Bu grafikler, SARIMA model bileşenlerinin belirlenmesinde rehberlik etmiştir.

SARIMA

SARIMA, zaman serilerine mevsimsel bileşenler ekleyerek klasik ARIMA modelini geliştirir. Hem kısa vadeli dalgalanmaları hem de mevsimsel desenleri modelleyebilir. Model; otoregresif, fark ve hareketli ortalama bileşenlerinin hem normal hem de mevsimsel düzeyde uyarlanmasıyla oluşturulur. PM10 serisinde haftalık döngü olduğu için SARIMA kullanımı uygun bulunmuştur.



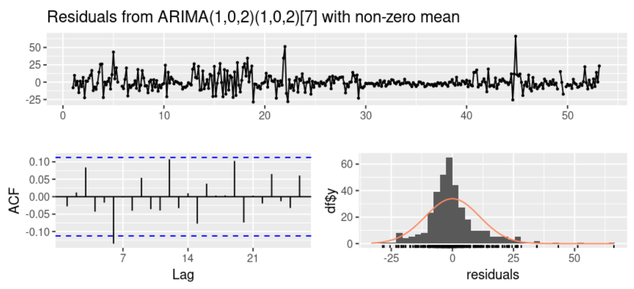
Grafikte SARIMA(1,0,2) modeliyle elde edilen 30 dönemlik PM10 tahminleri ve güven aralıkları gösterilmektedir. Model, geçmiş verilerin yapısını dikkate alarak dengeli bir öngörü sunmuş ve güven aralıklarının dar olması tahminlere olan güveni artırmıştır.

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,2)(1,0,2)[7] with non-zero mean
Q* = 12.064, df = 8, p-value = 0.1484

Model df: 6.   Total lags used: 14
```

SARIMA(1,0,2)(1,0,2)[7] modeline ait artıklar için uygulanan Ljung-Box testinde p-değeri 0.1484 bulunmuştur. Bu sonuç, artıkların otokorelasyon içermediğini ve modelin veri yapısını başarıyla temsil ettiğini göstermektedir.



Model artıklarının zaman içindeki dağılımı rastgele ve ortalama etrafında dengelidir. ACF grafiğinde anlamlı otokorelasyon gözlenmemiş, histogram ise artıkların yaklaşık normal dağıldığını göstermiştir. Bu bulgular, SARIMA modelinin geçerliliğini desteklemektedir.

Model Performansının Eğitim ve Test Verisi Üzerinden Değerlendirilmesi

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.009755352	12.79789	8.240336	-7.6048942	29.29942	1.0092836	0.02070014
Test set	3.785882353	11.87333	8.120000	-0.1212009	30.48828	0.9945447	NA

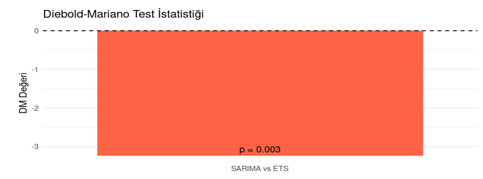
Veri seti eğitim ve test olarak ayrılarak modelin genellenebilirliği test edilmiştir. Test verisindeki hata oranlarının eğitim setine yakın kalması, SARIMA modelinin dengeli ve tutarlı tahminler ürettiğini göstermektedir.

MAPE değerlerinin %29.3 ve %30.5 gibi yakın çıkması, farklı veri bölümlerinde benzer doğrulukta sonuçlar elde edildiğini ortaya koymaktadır. ACF1 değerinin düşük olması da artıkların bağımsızlık varsayımını büyük ölçüde sağladığını göstermektedir.

ETS ve SARIMA Modellerinin Karşılaştırılması: Diebold-Mariano Testi Sonuçları

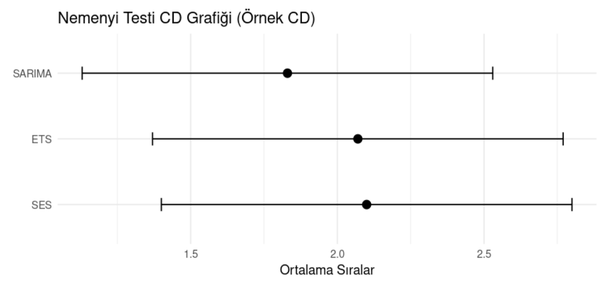
```
        Diebold-Mariano Test

data:  residuals(f1.out)residuals(f2.out)
DM = -3.2372, Forecast horizon = 1, Loss function power = 2, p-value = 0.003017
alternative hypothesis: two.sided
```



SARIMA ve ETS modellerinin tahmin hataları karşılaştırıldığında, DM testi sonucu p = 0.003 ile anlamlı fark olduğunu göstermektedir. DM istatistiğinin negatif olması, SARIMA modelinin ETS'ye göre daha düşük hata ürettiğini ve daha başarılı tahminler sağladığını ortaya koymaktadır.

Nemenyi Testi CD Grafiği ile Üç Modelin Karşılaştırılması



Grafikte SARIMA, ETS ve SES modellerinin ortalama sıralamaları Nemenyi testi ile karşılaştırılmıştır. Her bir modelin ortalama performansı ve kritik mesafeleri yatay çizgilerle gösterilmiştir. izgilerin birbirini kapsaması, modeller arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir.

Bu sonuç, tahmin başarılarının genel olarak benzer düzeyde olduğunu ve modeller arasında belirgin bir üstünlük bulunmadığını ortaya koymaktadır.

KAYNAKÇA

[1] https://www.kaggle.com/datasets/pinuto/marmara-air-quality-data-2024-istanbul-kandilli

[2] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts. https://otexts.com/fpp2/

[3] ETRI Journal. (2023). Air quality index prediction using seasonal autoregressive integrated moving average model. Wiley Online Library https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2023-0283

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

Yüksek Lisans Kabulü İçin Regresyon Tabanlı Tahmin ve Veri Analizi

Ömer BULUT 20023076

Danışman: Doç. Dr. Atıf Ahmet EVREN

ÖZET

Bu çalışma, öğrencilerin yüksek lisans programlarına kabul edilme olasılıklarını tahmin etmek amacıyla gerçekleştirilmiştir. Adayların sınav puanları, not ortalamaları ve araştırma deneyimleri gibi değişkenler kullanılarak çoklu doğrusal regresyon modeli kurulmuştur. Model sonuçları, özellikle CGPA, TOEFL puanı ve araştırma tecrübesinin kabul olasılığı üzerinde anlamlı etkileri olduğunu göstermektedir. Bu tez, başvuru sürecine dair daha bilinçli ve veriye dayalı değerlendirmeler yapılmasına katkı sunmayı hedeflemektedir.

Çoklu Doğrusal Regresyon ile Kabul Olasılığı Tahmini

Bu çalışmada, öğrencilerin yüksek lisans programlarına kabul edilme olasılıkları çoklu doğrusal regresyon (Multiple Linear Regression) yöntemiyle tahmin edilmiştir. Modelin genel formu şu şekildedir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

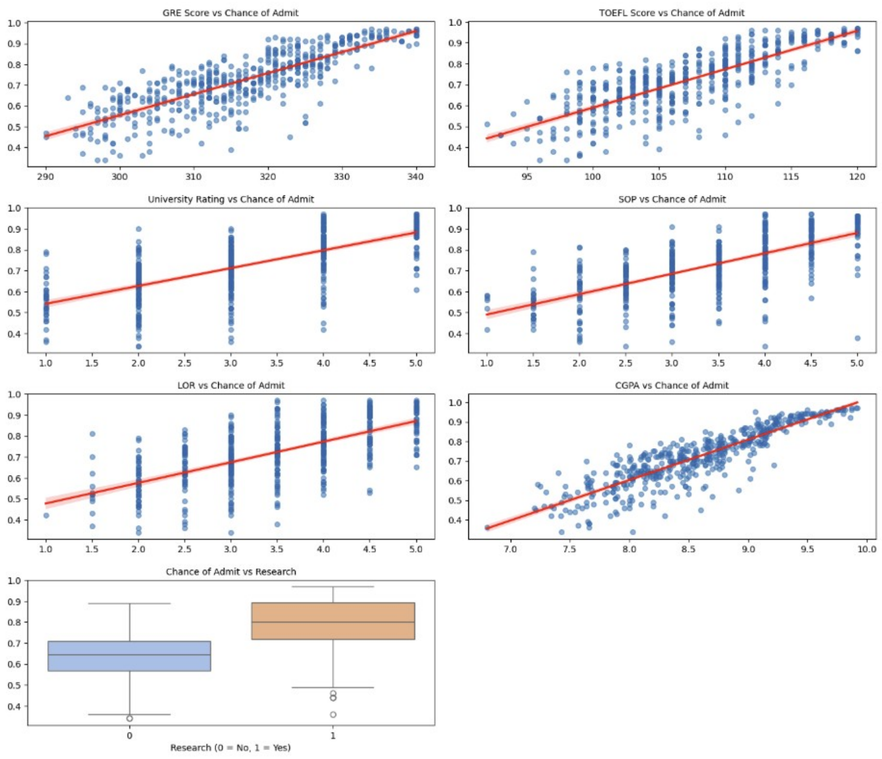
Y = Bağımlı değişken - β_0 = Sabit terim - β_n = Regresyon Katsayıları - X_n = Bağımsız Değişkenler - ε = Hata

Veri Seti Hakkında Genel Bilgi

Veri seti toplam, 500 başvuruya ait bilgilerden oluşmaktadır ve eksik veri bulunmamaktadır. Veri türleri, modelleme için uygundur. Tüm değişkenler sayısaldır ve doğrudan regresyon analizine dahil edilebilir yapıdadır.

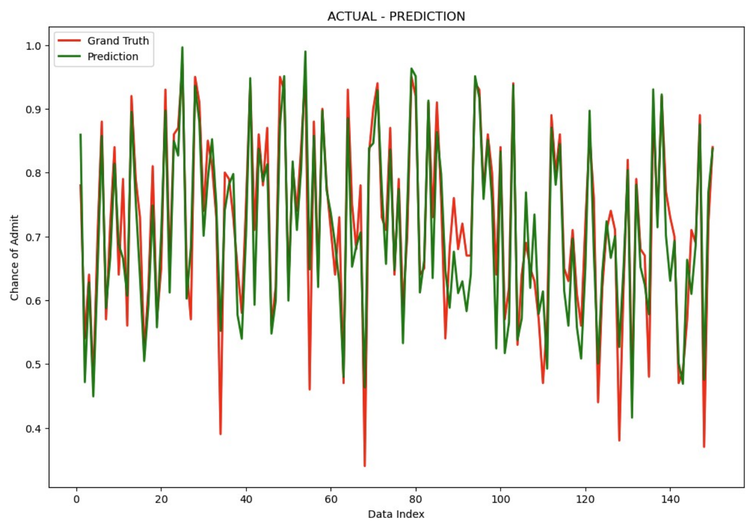
#	Column	Non-Null Count	Dtype
0	GRE Score	500 non-null	int64
1	TOEFL Score	500 non-null	int64
2	University Rating	500 non-null	int64
3	SOP	500 non-null	float64
4	LOR	500 non-null	float64
5	CGPA	500 non-null	float64
6	Research	500 non-null	int64
7	Chance of Admit	500 non-null	float64

Bağımlı Değişken ile Diğer Değişkenler Arasındaki İlişki



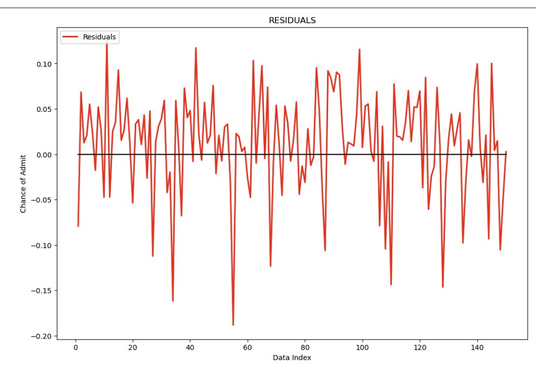
Kabul olasılığı CGPA, GRE ve TOEFL gibi akademik başarılar ile araştırma deneyimiyle pozitif ilişkilidir. Üniversite derecesi, SOP ve LOR'un etkisi daha düşüktür. Akademik başarı ve araştırma, kabulde belirleyici faktörlerdir.

Model Tahminlerinin Yapılması ve Değerlendirilmesi



Düşük olasılıklarda bazı farklar olsa da, model genel eğilimi doğru yakalamaktadır. Bu görsel, modelin tahmin performansını hızlıca değerlendirmeye yardımcı olur.

Artıkların dengeli ve rastgele dağılması, modelin sistematik hata yapmadığını ve doğrusal varsayımlara uyduğunu gösterir. Bazı aykırı değerler modelin uç durumlarda daha az başarılı olduğunu işaret eder. Genel olarak model, güvenilir ve tutarlı tahminler yapmaktadır.



OLS Yöntemi ile Model Analizi

OLS Regression Results

Dep. Variable:	y	R-squared:	0.813
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	212.8
Date:	Wed, 21 May 2025	Prob (F-statistic):	1.71e-120
Time:	00:47:42	Log-Likelihood:	485.93
No. Observations:	350	AIC:	-955.9
Df Residuals:	342	BIC:	-925.0
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.2027	0.128	-9.372	0.000	-1.455	-0.950
GRE Score	0.0011	0.001	1.788	0.075	-0.000	0.002
TOEFL Score	0.0035	0.001	3.340	0.001	0.001	0.006
University Rating	0.0052	0.005	1.153	0.250	-0.004	0.014
SOP	0.0006	0.006	0.099	0.921	-0.010	0.012
LOR	0.0172	0.005	3.389	0.001	0.007	0.027
CGPA	0.1275	0.012	10.575	0.000	0.104	0.151
Research	0.0276	0.008	3.433	0.001	0.012	0.043

Omnibus:	84.653	Durbin-Watson:	2.105
Prob(Omnibus):	0.000	Jarque-Bera (JB):	202.880
Skew:	-1.181	Prob(JB):	8.81e-45
Kurtosis:	5.887	Cond. No.	1.32e+04

OLS Regression Results

Dep. Variable:	y	R-squared:	0.810
Model:	OLS	Adj. R-squared:	0.808
Method:	Least Squares	F-statistic:	368.8
Date:	Wed, 21 May 2025	Prob (F-statistic):	3.53e-123
Time:	00:49:16	Log-Likelihood:	483.34
No. Observations:	350	AIC:	-956.7
Df Residuals:	345	BIC:	-937.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.0664	0.068	-15.581	0.000	-1.201	-0.932
TOEFL Score	0.0046	0.001	4.935	0.000	0.003	0.006
LOR	0.0181	0.005	3.905	0.000	0.009	0.027
CGPA	0.1414	0.010	13.928	0.000	0.121	0.161
Research	0.0334	0.008	4.410	0.000	0.018	0.048

Omnibus:	77.243	Durbin-Watson:	2.125
Prob(Omnibus):	0.000	Jarque-Bera (JB):	169.874
Skew:	-1.115	Prob(JB):	1.30e-37
Kurtosis:	5.583	Cond. No.	2.26e+03

İlk modelde tüm değişkenler yer almakta olup $R^2 = 0.813$ ve Adjusted $R^2 = 0.809$ değerleriyle yüksek bir açıklayma sunmuştur. Ancak yalnızca 4 değişken (CGPA, TOEFL, LOR, Research) anlamlı çıkmıştır. GRE, SOP ve University Rating değişkenlerinin etkisi sınırlı bulunmuştur.

İkinci model, yalnızca anlamlı değişkenleri içererek sadeleştirilmiş ve $R^2 = 0.810$, Adjusted $R^2 = 0.808$ ile benzer bir açıklama gücünü korumuştur. Tüm değişkenler anlamlı bulunmuş, özellikle CGPA en yüksek etkiye sahip olmuştur. Model varsayımlarını da büyük ölçüde sağlamaktadır.

Sonuç olarak, ikinci model daha yalın, anlamlı ve güvenilir bir yapı sunmakta; yorumlanabilmesi artırarak tahmin başarısını sürdürmektedir.

Sonuç ve Değerlendirme

Bu çalışmada, öğrencilerin yüksek lisans programlarına kabul olasılıklarını tahmin etmek amacıyla çoklu doğrusal regresyon modeli uygulanmıştır. GRE, TOEFL, CGPA, LOR gibi adaylara ait çeşitli veriler kullanılmıştır.

Veri ön işleme ve keşifsel analiz sürecinde değişkenler arası pozitif ilişkiler gözlemlenmiş; özellikle CGPA, kabul olasılığını en çok etkileyen faktör olarak öne çıkmıştır.

İlk modelde tüm değişkenler kullanılmış; ancak bazıları istatistiksel olarak anlamlı bulunmamıştır. Bu nedenle yalnızca anlamlı değişkenlerle ikinci bir model kurulmuş ve benzer tahmin başarısı korunarak daha sade bir yapı elde edilmiştir.

Elde edilen sonuçlar, başvuru sürecinde akademik başarı ve araştırma deneyiminin belirleyici olduğunu göstermektedir. Bu model, üniversiteler için karar destek aracı olabileceği gibi, adaylar için de başvuru sürecine ışık tutabilir.

KAYNAKÇA

[1] Statsmodels, Statistical Models in Python, https://www.statsmodels.org/
[2] Seaborn, Statistical Data Visualization, https://seaborn.pydata.org/
[3] Scikit-learn, Machine Learning in Python, https://scikit-learn.org/

2024.-2025. 8. Yarıyılı İSTATİSTİK BÖLÜMÜ

DİRİCHLET DAĞILIMI VE UYGULAMALARI

Özgür Irmak 20023001

Danışman: Doç. Dr. Atıf Ahmet Evren

ÖZET

Bu çalışma, Beta ve Gamma dağılımlarının teorik temellerini açıklamakta ve bu iki dağılımın çok değişkenli genellemesi olan Dirichlet dağılımı ile ilişkisini ortaya koymaktadır. Dirichlet dağılımı, oran tipi verilerin modellenmesinde özellikle tercih edilmekte ve bağımsız Gamma değişkenlerinin normalize edilmesiyle elde edilmektedir.

Uygulama kısmında, 2013–2023 dönemine ait IMF COFER verileriyle oluşturulan USD, EUR, Altın ve diğer rezervlerden oluşan bileşik oranlı veri seti analiz edilmiştir. FED politika faiz oranı, bağımsız değişken olarak ele alınmış ve Dirichlet regresyon modeli ile rezerv dağılımının bu faiz oranına duyarlılığı incelenmiştir.

Modelde log-oran dönüşümü kullanılarak "diğer rezervler" referans alınmış ve elde edilen katsayılarla 2024–2027 dönemine ilişkin senaryo tahminleri yapılmıştır. Bu tahminler, faiz oranlarının düşmesi durumunda **USD'nin rezervler içindeki payının artabileceğini**, buna karşın **Euro ve Altın rezerv oranlarının hafif azalabileceğini** göstermiştir.

Sonuçlar, merkez bankalarının çeşitlendirme eğilimi sınırlı olsa da, doların rezerv para statüsünün halen güçlü şekilde sürdüğünü ortaya koymaktadır. Ancak modelde yalnızca faiz oranı dikkate alındığından, jeopolitik ve döviz kuru gibi faktörlerin dışarıda bırakılması önemli bir sınırlılık olarak değerlendirilmektedir.

Dirichlet Dağılımı ve Dirichlet Regresyonu

Dirichlet Dağılımı:

Bileşenleri 0 ile 1 arasında değişen ve toplamları 1'e eşit olan oran tipi (kompozisyonel) verilerin modellenmesinde kullanılan çok değişkenli sürekli bir olasılık dağılımıdır. Bu dağılım, özellikle her bir gözlem biriminde belirli kategorilere ait oranların yer aldığı durumlarda tercih edilmektedir. Örneğin, döviz rezervlerinin farklı para birimleri arasındaki yüzdesel dağılımı gibi finansal kompozisyonel veriler Dirichlet dağılımı ile uygun şekilde temsil edilebilmektedir.

Dirichlet Regresyonu:

Kompozisyonel yapıdaki bağımlı değişkenlerin bir veya daha fazla bağımsız değişken (örneğin faiz oranı, büyüme oranı vb.) ile ilişkisini modellemeye olanak tanıyan parametrik bir regresyon tekniğidir. Bu model, log-oran dönüşümüne dayanmakta olup, oran bileşenleri arasındaki ilişkileri dikkate alarak istatistiksel çıkarım yapmaya imkân sağlamaktadır.
Bu bağlamda Dirichlet regresyonu, klasik doğrusal regresyon modellerinin varsayımlarını ihlal eden oran veriler üzerinde daha uygun ve güvenilir sonuçlar üretmektedir.

Dirichlet Dağılımının Tarihsel Gelişimi ve Kısa Bir Uygulama Örneği:

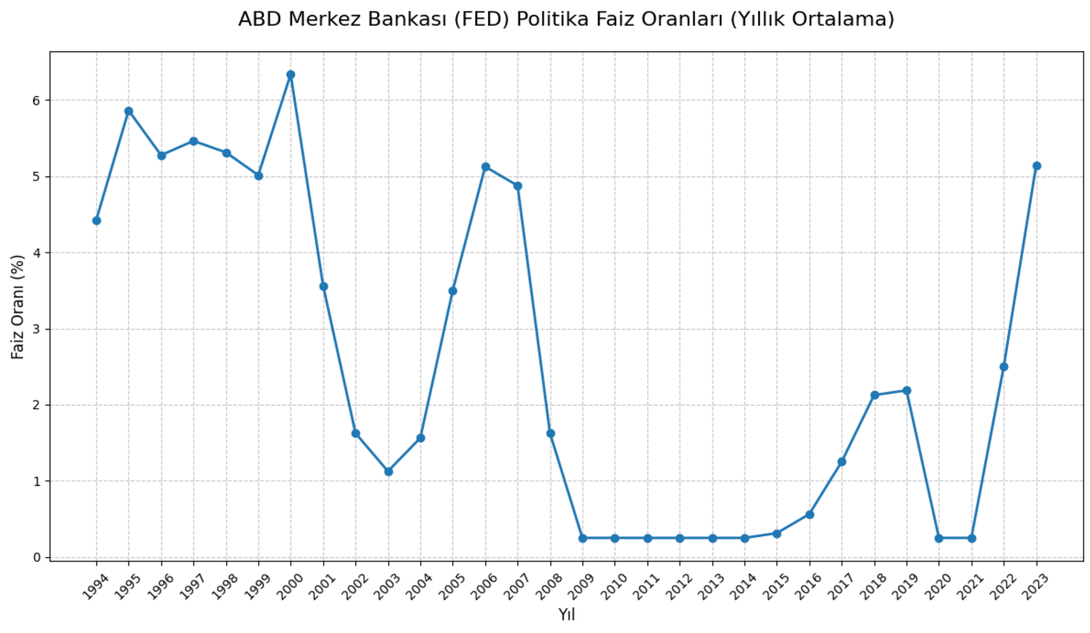
Dirichlet dağılımı, adını 19. yüzyılda yaşayan Alman matematikçi Johann Peter Gustav Lejeune Dirichlet'ten almaktadır. Tek değişkenli Beta dağılımının çok değişkenli bir genellemesi olarak ilk kez 1930'lu yıllarda tanımlanmış, daha sonra özellikle Bayesci istatistik ve kompozisyonel veri analizinde önemli bir rol oynamaya başlamıştır. 20. yüzyılın sonlarına doğru, özellikle bilgisayarların gelişimiyle birlikte, Dirichlet dağılımı çok boyutlu oran verilerinin modellenmesinde yaygın biçimde kullanılmaya başlanmıştır. Dirichlet regresyon modeli ise, bu dağılımın parametrik regresyon çerçevesine entegre edilmesiyle ortaya çıkmış, Hijazi ve Jernigan (2009) gibi araştırmacılar tarafından istatistiksel uygulamalarda sistematik olarak ele alınmıştır.

Çalışma Özeti:

Bir ülkenin döviz rezervleri dört ana kategoriye ayrılmış olsun: USD (%), EUR (%), XAU (altın, %) ve Other (%). Bu oranların toplamı her zaman 100'dür. Eğer bu dört bileşenin FED politika faizi gibi bir makroekonomik değişkene göre nasıl değiştiğini incelemek istersek, Dirichlet regresyon modeli kullanılır. Böylece her bir bileşenin, faiz oranındaki değişimle birlikte rezervler içindeki payının nasıl evrildiği tahmin edilebilir.

Veri Temini ve Metodolji:

Bu çalışma, küresel döviz rezervlerinin kompozisyonel analizini yapmak amacıyla **Dirichlet regresyon modeli** kullanmıştır. Metodoloji olarak, uluslararası rezerv bileşenleri (Amerikan Doları, Euro, Altın ve Diğer Para Birimleri) bağımlı değişkenler olarak ele alınmış; bu kompozisyonel veriler **log-oran dönüşümü** ile işlendikten sonra **tek bağımsız değişken** olarak belirlenen ABD Merkez Bankası'nın (FED) politika faiz oranı ile **lineer regresyon modelleri** kurulmuştur. Çalışmada kullanılan veriler, 2013-2023 dönemine ait çeyreklik **uluslararası rezerv verileri (IMF COFER veri tabanından)** ve yine aynı döneme ait **FED politika faiz oranları**ndan oluşmaktadır.



Çalışmaya Genel Bakış:

Bu çalışma, küresel döviz rezervlerinin kompozisyonel yapısını ve Amerika Birleşik Devletleri Merkez Bankası (FED) politika faiz oranlarının bu bileşim üzerindeki etkisini ampirik olarak incelemektedir. Çalışmanın temel amacı, çok değişkenli oran tipi verilerin modellenmesine uygun bir istatistiksel yaklaşım olan Dirichlet regresyonunu kullanarak, 2013-2023 dönemine ait çeyreklik uluslararası rezerv verileri ile FED faiz oranları arasındaki tarihsel ilişkiyi analiz etmek ve bu bulgulara dayanarak 2024-2027 dönemi için rezerv dağılımına dair ileriye dönük projeksiyonlar sunmaktır.

Yöntem ve Metodoloji: Çalışma, kompozisyonel veri yapısına sahip uluslararası rezerv paylarını modellemek amacıyla **Dirichlet regresyon modelini** kullanmaktadır. Bu metodoloji, bileşenleri pozitif değerli ve toplamları birime eşit olan oran vektörlerinin analizi için özel olarak tasarlanmıştır, böylece klasik regresyon tekniklerinin sınırlılıkları aşılmaktadır. Metodolojik süreç şu adımları içermektedir:

Log-Oran Dönüşümü: Rezerv bileşenleri (Dolar, Euro, Altın, Diğer Para Birimleri) doğrudan regresyona tabi tutulmadan önce, veri uzayının kısıtlarını gidermek amacıyla log-oran dönüşümü uygulanmıştır. Bu dönüşümde, Other_Currencies (Diğer Para Birimleri) kategorisi referans alınarak her bir bileşenin logaritmik oranı hesaplanmıştır.

Lineer Regresyon Modelleri: Dönüştürülen bu log-oranlar üzerinde, **tek bağımsız değişken** olarak belirlenen FED politika faiz oranıyla lineer regresyon modelleri kurulmuştur. Her bir rezerv bileşeni için ayrı bir model eğitilmiştir.

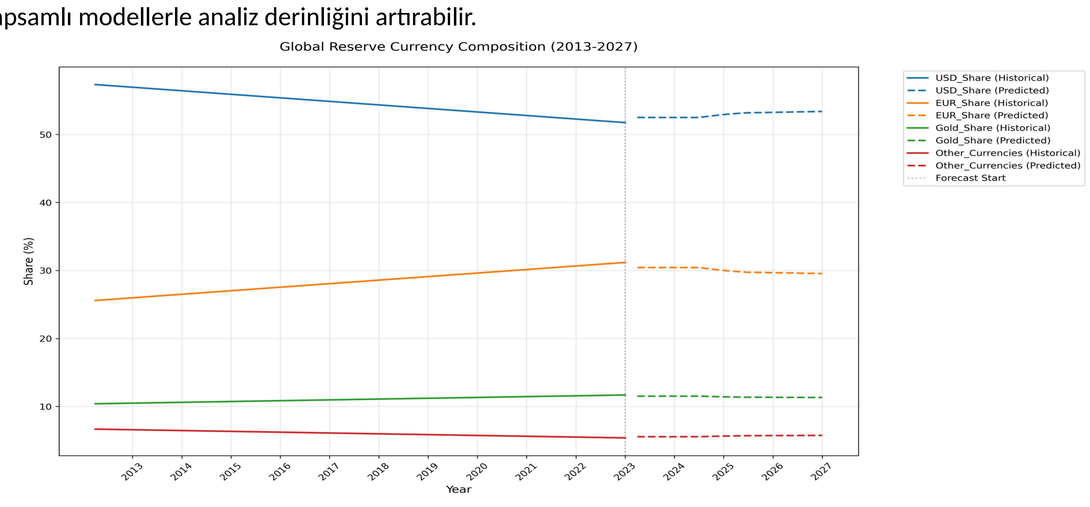
Gelecek Projeksiyonları: Modelden elde edilen regresyon katsayıları kullanılarak, geleceğe yönelik varsayımsal FED faiz tahminlerine dayalı olarak her bir rezerv bileşeninin gelecek dönemdeki payları öngörülmüştür. Tahmin edilen log-oranlar, orijinal yüzde paylarına geri dönüştürülerek toplamlarının %100 olması sağlanmıştır.

Uluslararası Rezerv Kompozisyonu Verileri: Uluslararası Para Fonu'nun (IMF) **COFER (Currency Composition of Official Foreign Exchange Reserves)** veri tabanından temin edilmiş olup, Amerikan Doları, Euro ve Diğer Para Birimleri paylarını içermektedir. Altın rezervlerinin payları için **World Gold Council (WGC)** verileri entegre edilmiştir.

FED Politika Faiz Oranları: Amerika Birleşik Devletleri Merkez Bankası'nın (FED) ilgili çeyrekteki politika faiz oranları, **Federal Reserve Economic Data (FRED)** veri tabanından elde edilmiştir. Tüm veri seti, analize uygun hale getirilmek üzere ön işleme süreçlerinden geçirilmiş, tarih formatları düzenlenmiş ve yüzdesel paylar oransal değerlere dönüştürülmüştür.

Sonuç ve Bulgular: Model çıktıları, FED faiz oranlarındaki değişimlerin küresel rezerv para bileşiminde istatistiksel olarak anlamlı etkiler yarattığını ortaya koymaktadır. Özellikle, analiz edilen senaryolarda faiz oranlarının düşüş eğilimi göstermesi durumunda dahi **Amerikan Doları'nın rezervler içindeki payının azalmak yerine artabileceği** gözlemlenmiştir. Bu bulgu, "de-dolarizasyon" süreçlerinin faiz indirimleriyle otomatik olarak ortaya çıkmadığını ve doların rezerv varlığı olarak yüksek likidite ve güven avantajlarıyla konumunu sürdürdüğünü düşündürmektedir. Öte yandan, **Euro ve altın gibi varlıkların paylarında hafif bir azalış eğilimi** öngörülürken, "diğer para birimleri" kategorisinin payında sınırlı bir artış kaydedilmiştir. Bu sonuçlar, küresel para sisteminin belirli bir çeşitlenmeye açık olmakla birlikte, doların uluslararası rezerv para statüsünde halen güçlü ve dirençli bir konumda olduğunu vurgulamaktadır.

Sınırlılıklar: Çalışmanın temel sınırlılığı, modelin yalnızca FED politika faiz oranını içeren tek değişkenli bir yapıda kurgulanmış olmasıdır. Küresel rezerv kompozisyonunu etkileyen jeopolitik riskler, döviz kuru hareketleri, ticaret dengeleri veya diğer büyük merkez bankalarının para politikaları gibi potansiyel belirleyici faktörler modelin kapsamı dışında bırakılmıştır. Gelecek araştırmalar, bu çok boyutlu değişkenleri içeren daha kapsamlı modellerle analiz derinliğini artırabilir.



KAYNAKÇA

[1] https://data.imf.org/en/datasets/IMF.STA:COFER
[2] https://fred.stlouisfed.org/series/FEDFUNDS
[3] https://journal.r-project.org/archive/2014-2/RJournal_2014-2_Maier.pdf

2024-2025 Bahar Yarıyılı

İSTATİSTİK BÖLÜMÜ

BİREYLERİN DİJİTAL OYUN OYNAMA ALIŞKANLIKLARININ PSİKOLOJİLERİNE ETKİSİ

Rabia İnci ERYILMAZ

Danışman: Arş. Gör. Dr. Hülya YÜREKLİ

ÖZET

Bu çalışma, dijital oyun oynama alışkanlıklarının bireylerin psikolojik durumları üzerindeki etkilerini incelemeyi amaçlamaktadır. Araştırma kapsamında 252 katılımcıdan çevrim içi anket yoluyla veri toplanmış; anksiyete, depresyon, yalnızlık, pozitiflik, özsaygı ve oyun bağımlılığı değişkenleri değerlendirilmiştir. Veriler, betimsel istatistiklerin yanı sıra korelasyon analizi, ANOVA, kümeleme ve çoklu doğrusal regresyon yöntemleriyle analiz edilmiştir. Bulgular, oyun sırasında pozitif duyguların artmasına rağmen, oyun sonrasında depresyon ve anksiyete düzeylerinde bir yükseliş olduğunu göstermektedir. Kümeleme analizi ile "yüksek psikolojik belirti gösteren" ve "dengeli" olmak üzere iki oyuncu profili tanımlanmıştır. Regresyon analizi sonuçlarına göre, depresyonun en güçlü yordayıcıları anksiyete, yalnızlık, düşük özsaygı ve yaştır. Elde edilen sonuçlar, dijital oyunların bireyler üzerindeki psikolojik etkilerinin kişisel ve bağlamsal faktörlere bağlı olarak değiştiğini ortaya koymaktadır.

AMAÇ

Bu araştırmanın temel amacı, bireylerin dijital oyun oynama alışkanlıklarının psikolojik durumları üzerindeki etkilerini çok boyutlu olarak incelemektir.

Bu kapsamda araştırmada:

- Oyun öncesi, sırası ve sonrası dönemlerde bireylerin duygu durumlarındaki değişimlerin analiz edilmesi,
- Dijital oyun oynama davranışlarının depresyon, anksiyete, yalnızlık, pozitiflik, özsaygı ve oyun bağımlılığı gibi psikolojik değişkenlerle ilişkilerinin ortaya konulması,
- Oyuncuların psikolojik profillerine göre gruplandırılması (kümeleme analizi),
- Depresyonu anlamlı düzeyde yordayan faktörlerin belirlenmesi (regresyon analizi),
- Oyun türü, oynama zamanı ve sosyal bağlam gibi değişkenlerin psikolojik etkiler üzerindeki rolünün değerlendirilmesi

amaçlanmaktadır.

YÖNTEM

Katılımcılardan veri toplamak amacıyla Google Forms üzerinden oluşturulan çevrim içi bir anket uygulanmıştır. Araştırma örneklemi, yaşları 12 ile 67 arasında değişen ve çoğunluğu genç yetişkinlerden oluşan 252 bireyden oluşmaktadır. Anket formu, katılımcıların demografik bilgilerine ek olarak depresyon, anksiyete, yalnızlık, pozitiflik, özsaygı ve oyun bağımlılığı gibi psikolojik değişkenleri ölçen ölçekleri içermektedir.

Veri analizi sürecinde, betimleyici istatistiklerle birlikte Pearson korelasyon analizi, tek yönlü varyans analizi (ANOVA) ve Tukey HSD testi uygulanmıştır. Oyuncuların psikolojik özelliklerine göre sınıflandırılabilmesi için K-means kümeleme analizi yapılmış, depresyon düzeyini etkileyen faktörleri belirlemek amacıyla çoklu doğrusal regresyon analizi gerçekleştirilmiştir. Verilerin analizinde Python programlama dili kullanılmış, elde edilen bulgular hem görsel hem istatistiksel olarak değerlendirilmiştir.

BULGULAR

🔹 Zamana Bağlı Ruh Hâli Değişimi

Oyun oynama sürecinin farklı evrelerinde (öncesi, sırası, sonrası) katılımcıların duygu durumlarında anlamlı değişiklikler gözlenmiştir.

- **Anksiyete düzeyi**, oyun öncesine kıyasla oyun sırasında ve sonrasında artmıştır ($\bar{x}$ = 2.13 → 2.66 → 2.81).
- **Depresyon düzeyi**, oyun sırasında düşmüş ($\bar{x}$ = 1.94) fakat oyun sonrasında önceki seviyenin de üzerine çıkmıştır ($\bar{x}$ = 3.23).
- **Pozitiflik düzeyi**, oyun sırasında en yüksek seviyeye ulaşmıştır ($\bar{x}$ = 3.40) ancak oyun sonrasında hafif bir düşüş göstermiştir.

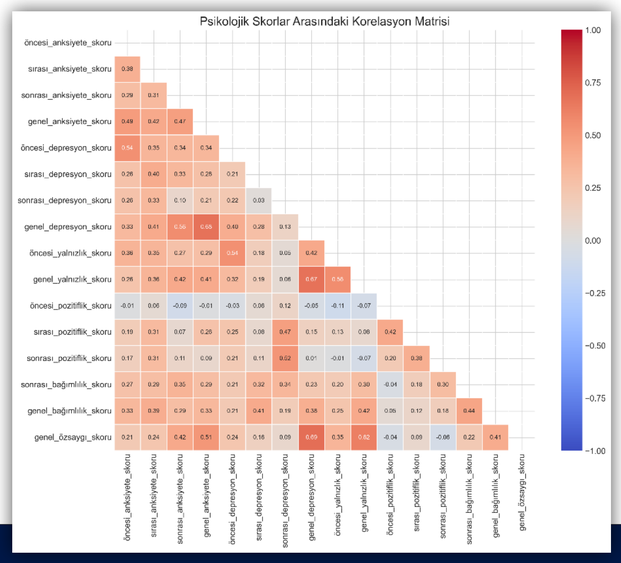
Bu bulgular, dijital oyunların anlık olarak olumlu etkiler yaratabildiğini; ancak bazı bireylerde bu etkinin oyun sonrasında tersine dönebileceğini göstermektedir.



🔹 Korelasyon Analizi Sonuçları

Psikolojik değişkenler arasında anlamlı düzeyde ilişkiler tespit edilmiştir:

- **Depresyon**, en güçlü şekilde **anksiyete** (r = 0.68), **yalnızlık** (r = 0.67) ve **düşük özsaygı** (r = 0.69) ile ilişkilidir.
- **Oyun bağımlılığı**, orta düzeyde depresyon (r = 0.38) ve yalnızlık (r = 0.42) ile pozitif korelasyon göstermektedir.
- **Pozitiflik** ile depresyon, yalnızlık ve anksiyete arasındaki ilişkiler zayıf düzeydedir; bu durum, pozitif duyguların geçici olduğunu düşündürmektedir.



🔹 Kümeleme Analizi (Oyuncu Profilleri)

K-means algoritmasıyla katılımcılar iki kümeye ayrılmıştır:

- **Küme 0 – Yüksek psikolojik belirti gösteren oyuncular:** Bu grupta depresyon, anksiyete ve yalnızlık düzeyleri yüksektir; özsaygı düşüktür. Oyun sonrası pozitiflik azalmaktadır.
- **Küme 1 – Duygusal açıdan dengeli oyuncular:** Bu bireylerde oyun sonrası pozitiflik korunmakta; anksiyete, depresyon ve yalnızlık düşüktür. Özsaygı yüksektir.

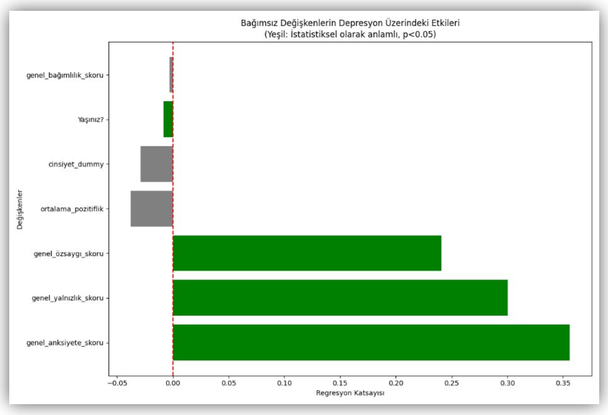
Bu bulgular, oyuncuların psikolojik tepkilerinin kişisel yapılarıyla doğrudan ilişkili olduğunu göstermektedir.

🔹 Regresyon Analizi (Depresyonun Yordayıcıları)

Çoklu doğrusal regresyon modeli, depresyon değişkeninin %68.5'ini açıklamaktadır (R^2 = 0.685).

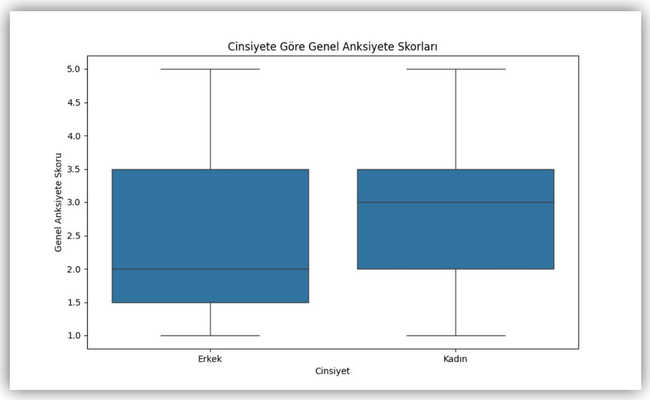
- **Anksiyete, yalnızlık, düşük özsaygı** ve **yaş** anlamlı yordayıcılar olarak belirlenmiştir.
- Oyun bağımlılığı, pozitiflik ve cinsiyet değişkenleri modelde anlamlı etkide bulunmamıştır.

Bu bulgular, bireylerin psikolojik alt yapılarının depresif belirtiler üzerinde daha belirleyici olduğunu ortaya koymaktadır.



🔹 Demografik Değişkenlere Göre Farklılıklar

- **Kadın katılımcılar**, erkeklere kıyasla daha yüksek depresyon ve anksiyete skorlarına sahiptir.
- **Genç bireyler** (12–24 yaş), daha yüksek düzeyde depresyon, yalnızlık ve anksiyete bildirmiştir.
- **Öğrenciler**, çalışan bireylere göre daha yüksek psikolojik hassasiyet taşımaktadır.



🔹 Oyun Türü ve Sosyal Bağlamın Etkisi

- **Rekabetçi oyunlar**, pozitiflik artışı sağlarken, oyun sonrası anksiyete düzeyini de yükseltebilmektedir.
- **Takım oyunları**, sosyal bağ ve iş birliği nedeniyle olumlu etkiler yaratmakla birlikte, başarısızlık durumunda depresif duyguları tetikleyebilir.
- **Rahatlatıcı oyunlar**, yalnız bireyler tarafından daha çok tercih edilmekte, ancak pozitiflik üzerinde sınırlı etki göstermektedir.
- **Sosyal bağlamda (arkadaşlarla oynanan oyunlar)** pozitiflik skorları daha yüksek; tek başına oynayanlarda yalnızlık düzeyi anlamlı olarak fazladır.

Bu bulgular doğrultusunda, dijital oyunların etkilerinin sadece oynama süresi ya da bağımlılık düzeyiyle değil, oyuncunun psikolojik profili, demografik özellikleri ve bağlamsal faktörlerle birlikte değerlendirilmesi gerektiği anlaşılmaktadır.

SONUÇ

Araştırma, dijital oyunların bireyler üzerinde hem kısa vadeli pozitif etkiler (örneğin oyun sırasında artan pozitiflik) hem de oyun sonrasında artan depresyon ve anksiyete gibi olumsuz duygular oluşturabileceğini göstermektedir.

Özellikle genç yaş grubu, düşük özsaygı ve yüksek yalnızlık düzeyine sahip bireylerde oyun sonrası psikolojik kırılganlık daha belirgin şekilde ortaya çıkmaktadır. Anksiyete, yalnızlık ve özsaygı düzeyi, depresyonun güçlü yordayıcılarıdır.

Oyun türü, oynama zamanı ve sosyal bağlam gibi faktörler de psikolojik etkiyi anlamlı düzeyde şekillendirmektedir. Dijital oyunlar, bireysel psikolojik yapılarla etkileşim içinde değerlendirilmelidir.

KAYNAKÇA

- Kowert, R. (2024). The Real Benefits of Video Games. *Built In*
- Wartberg, L., Kriston, L., & Thomasius, R. (2020). The Association of Problematic Online Gaming Behavior With Psychosocial Problems and Health-Related Quality of Life in Adolescents
- Kowert, R., Vogelgesang, J., Festl, R., & Quandt, T. (2015). Psychosocial causes and consequences of online video game play

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

KEDİ SESLERİNİN MAKİNE ÖĞRENİMİ YÖNTEMLERİYLE SINIFLANDIRILMASI: ÖZELLİK SETLERİ VE MODEL PERFORMANSLARININ KARŞILAŞTIRMALI ANALİZİ

Rana KARABIYIK 20023070

Danışman: Doç. Dr. Gülder KEMALBAY

ÖZET

Bu tez çalışması, kedilerin çeşitli duygu ve davranış durumlarını yansıtan vokalizasyonlarını makine öğrenimi yöntemleriyle sınıflandırmayı amaçlamıştır. Çalışmada, 10 farklı kategoriye ait 100 orijinal kedi sesi dosyasından oluşan bir veri seti kullanılmış ve veri artırma teknikleriyle bu sayı 300'e çıkarılmıştır. MFCC'ler, Spektral ve Chroma özellikleri gibi kapsamlı akustik özellikler çıkarılmış, bu özellik setleri SVM, Random Forest, K-NN, MLP, XGBoost ve LightGBM gibi modellerle test edilmiştir. En iyi sonuç, "%92.33 ortalama çapraz doğrulama doğruluğu" ile LightGBM modeli ve "MFCC'ler + Temel Spektral Özellikler" kombinasyonuyla elde edilmiştir. Bulgular, otomatik kedi sesi sınıflandırmasında yüksek doğruluğa ulaşılabileceğini ve hayvan refahı ile insan-hayvan etkileşimini anlamada potansiyel uygulamaları olduğunu göstermektedir.

GİRİŞ & AMAÇ

Evcil hayvanların artan önemi: Kedilerin günlük yaşamdaki yeri ve sesli iletişimin önemi.

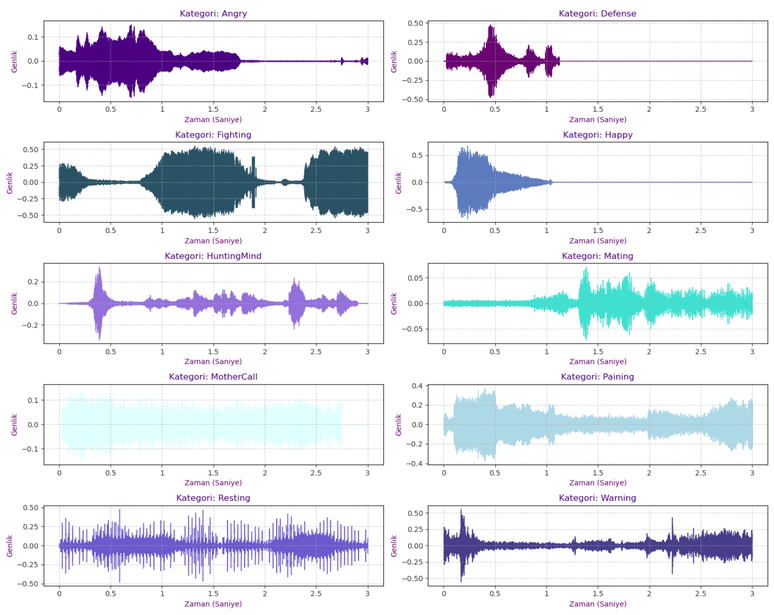
Neden sınıflandırma? Hayvan refahını artırma, davranış analizi, insan-hayvan iletişmini anlama.

Çalışmanın Amacı: Farklı özellik setleri ve makine öğrenimi modellerinin kedi sesi sınıflandırma performanslarını karşılaştırmak ve en uygun sistemi belirlemek.

METODOLOJİ



CatSound Veri Setinden Her Kategori İçin Örnek Dalga Formları



BULGULAR

Performans Karşılaştırması

En yüksek doğruluk: **LightGBM** modeli, **"MFCC'ler + Temel Spektral Özellikler"** ile %92.33 (ortalama çapraz doğrulama doğruluğu).

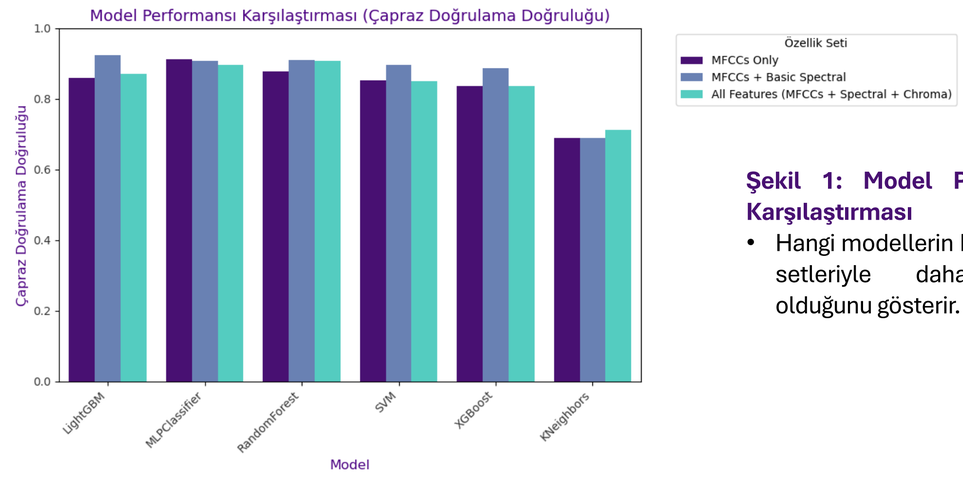
LightGBM, MFCC'lerin ve temel spektral özelliklerin birleşimiyle en dengeli ve güçlü performansı sergilemiştir.

Diğer modeller de kayda değer sonuçlar elde etse de, LightGBM hem doğruluk hem de verimlilik açısından öne çıktı.

Kedi Sesi Sınıflandırma Modelleri ve Özellik Setleri Karşılaştırma Sonuçları

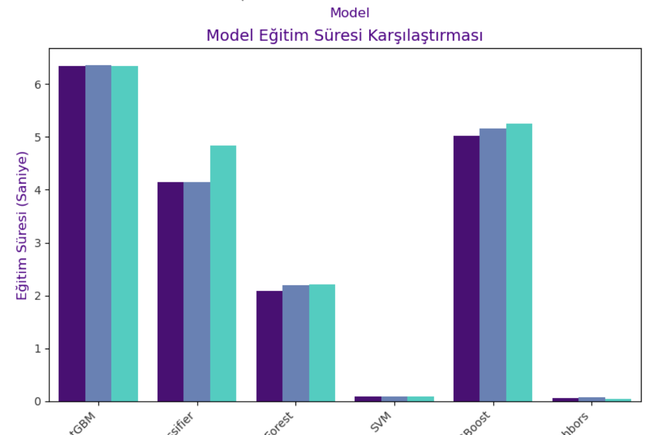
	Feature_Set	Model	Test_Accuracy_Split	CV_Mean_Accuracy	CV_Std_Accuracy	Training_Time_Sec
0	MFCCs + Basic Spectral	LightGBM	0.9067	0.9233	0.0327	6.37 s
1	MFCCs Only	MLPClassifier	0.8933	0.9133	0.0452	4.14 s
2	MFCCs + Basic Spectral	RandomForest	0.9467	0.9100	0.0249	2.20 s
3	All Features (MFCCs + Spectral + Chroma)	RandomForest	0.8667	0.9067	0.0490	2.21 s
4	MFCCs + Basic Spectral	MLPClassifier	0.8667	0.9067	0.0416	4.15 s
5	MFCCs + Basic Spectral	SVM	0.8667	0.8967	0.0562	0.09 s
6	All Features (MFCCs + Spectral + Chroma)	MLPClassifier	0.8133	0.8967	0.0839	4.84 s
7	MFCCs + Basic Spectral	XGBoost	0.8933	0.8867	0.0267	5.17 s
8	MFCCs Only	RandomForest	0.8533	0.8767	0.0478	2.09 s
9	All Features (MFCCs + Spectral + Chroma)	LightGBM	0.8533	0.8700	0.0245	6.35 s
10	MFCCs Only	LightGBM	0.8667	0.8600	0.0133	6.34 s
11	MFCCs Only	SVM	0.8800	0.8533	0.0562	0.10 s
12	All Features (MFCCs + Spectral + Chroma)	SVM	0.8133	0.8500	0.0983	0.10 s
13	MFCCs Only	XGBoost	0.8133	0.8367	0.0323	5.02 s
14	All Features (MFCCs + Spectral + Chroma)	XGBoost	0.8533	0.8367	0.0323	5.26 s
15	All Features (MFCCs + Spectral + Chroma)	KNeighbors	0.6533	0.7133	0.0464	0.05 s
16	MFCCs + Basic Spectral	KNeighbors	0.7333	0.6900	0.0478	0.08 s
17	MFCCs Only	KNeighbors	0.7467	0.6900	0.1123	0.06 s

Tablo 1: Özellik Setleri ve Makine Öğrenimi Modellerinin Performans Karşılaştırma Sonuçları



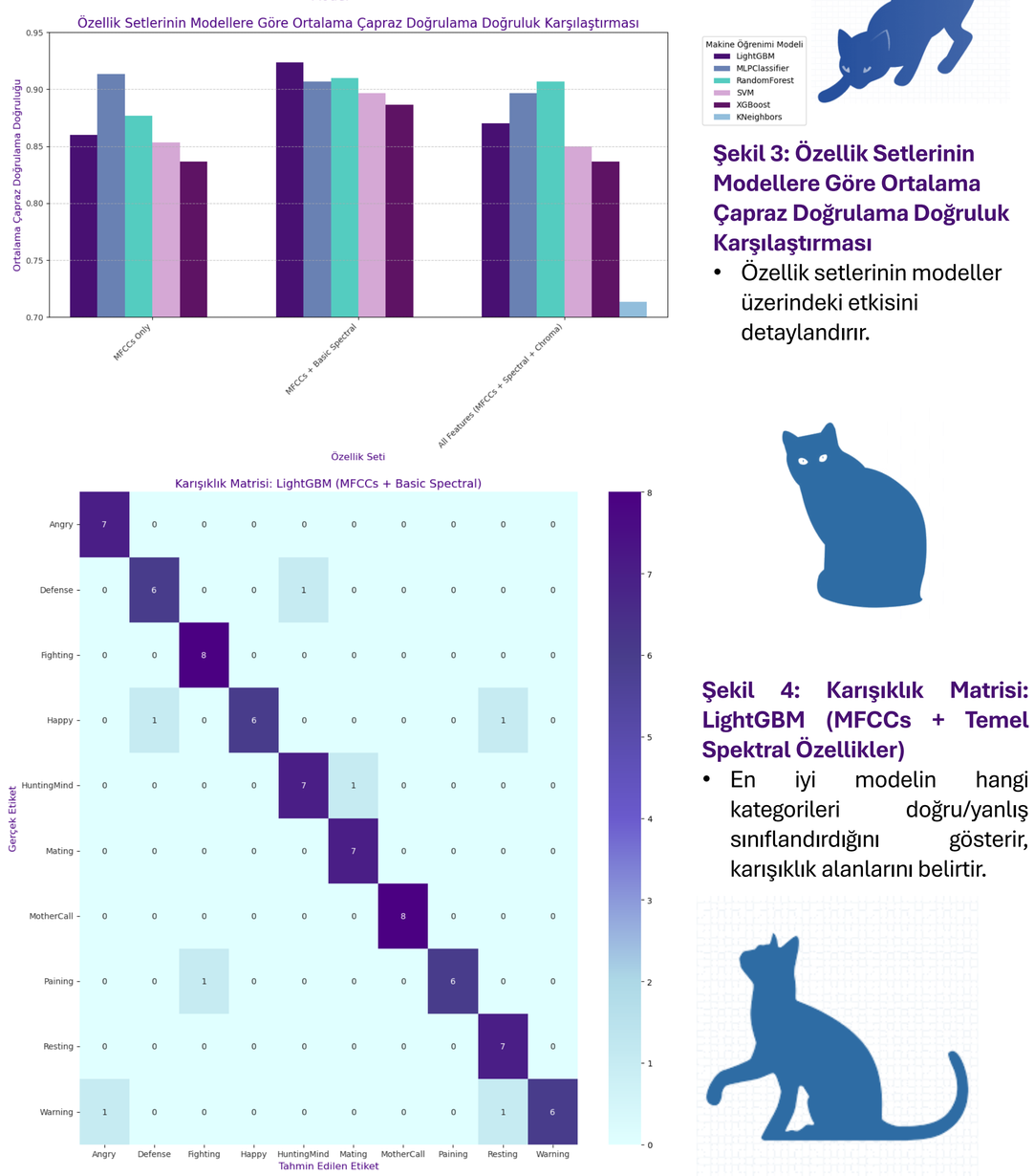
Şekil 1: Model Performansı Karşılaştırması

- Hangi modellerin hangi özellik setleriyle daha başarılı olduğunu gösterir.



Şekil 2: Model Eğitim Süresi Karşılaştırması

- Modellerin eğitim sürelerini karşılaştırır, verimli olanları vurgular.



Şekil 3: Özellik Setlerinin Modellere Göre Ortalama Çapraz Doğrulama Doğruluk Karşılaştırması

- Özellik setlerinin modeller üzerindeki etkisini detaylandırır.

Şekil 4: Karışıklık Matrisi: LightGBM (MFCCs + Temel Spektral Özellikler)

- En iyi modelin hangi kategorileri doğru/yanlış sınıflandırdığını gösterir, karışıklık alanlarını belirtir.

SONUÇ & GELECEK ÇALIŞMALAR

Sonuç: Kedi seslerinin makine öğrenimi ile yüksek doğrulukta sınıflandırılabileceği gösterildi (%92.33 ile LightGBM ve MFCC'ler + Temel Spektral). Bu, hayvan refahı ve insan-hayvan iletişimi araştırmaları için güçlü bir temel sunmaktadır.

Gelecek Çalışmalar:

- Daha büyük ve çeşitli veri setleri ile derin öğrenme modellerinin entegrasyonu.
- Daha detaylı hiperparametre optimizasyonu ve yeni özellik mühendisliği teknikleri.
- Farklı hayvan türlerinin sesleri üzerinde benzer analizlerin uygulanması.
- Geliştirilen sistemin gerçek zamanlı uygulamalara dönüştürülmesi.

KAYNAKÇA

[1] Y.R. Pandeya, D. Kim, J. Lee. Domestic cat sound classification using learned features from deep neural nets. *Appl Sci (Switzerland)*, 8 (2018).
[2] Y.R. Pandeya, J. Lee
Domestic cat sound classification using transfer learning International Journal of Fuzzy Logic and Intelligent Systems, 18 (2018), pp. 154-160,
[3] Y. Pandeya. *Cat Sound Classification Dataset*. Kaggle. (Erişim: 16 Haziran 2025).

2024-2025 BAHAR YARIYILI İSTATİSTİK BÖLÜMÜ

Dizi Seçimi İçin Mobil Uygulama Geliştirilmesi

Receb ŞEN 20023083

Danışman: Dr. Öğr. Üyesi Coşkun PARİM

ÖZET

Proje, arkadaş gruplarının birlikte dizi seçme sürecini hem hızlandırmak hem de daha keyifli hale getirmek amacıyla geliştirilecek bir mobil uygulamayı kapsamaktadır. Uygulama, kullanıcıların dizi önerilerini sağa veya sola kaydırarak değerlendirdiği ve grup üyelerinin ortak beğenileri doğrultusunda bir dizinin seçildiği etkileşimli bir sistemi temel almaktadır. Uygulamanın ön yüzü, React Native ve Expo Framework kullanılarak geliştirilecek; kullanıcı verileri ve etkileşimler ise Firebase altyapısı ile yönetilecektir. Basit, anlaşılır ve kullanıcı dostu bir arayüz hedeflenmektedir. Proje sonucunda, kullanıcıların ortak ilgi alanlarını keşfetmelerini sağlayan, grup içi iletişimi güçlendiren ve etkinlik planlamasını kolaylaştıran bir mobil çözüm sunulacaktır. Bunun yanı sıra uygulamanın reklam, sponsorluk gibi gelir modelleriyle desteklenerek ticari potansiyele sahip olması da öngörülmektedir.

Giriş

Bu proje, arkadaş gruplarının birlikte dizi seçme sürecini hem eğlenceli hem de pratik hale getirmeyi hedefleyen bir mobil uygulamadır. Kullanıcılar, dizileri sağa (beğen) veya sola (beğenme) kaydırarak değerlendirme yapar. Grup üyelerinin aynı diziyi beğenmesi durumunda sistem bu eşleşmeyi bildirerek ortak bir karar alınmasını sağlar. Böylece karar verme süreci hem keyifli bir deneyime dönüşür hem de zaman kaybının önüne geçilir. Uygulama yalnızca eğlenceyi değil, aynı zamanda sosyalleşmeyi de teşvik eder. Grup içi iletişimi güçlendirir, ortak ilgi alanlarının keşfedilmesine olanak tanır. Pandemi sonrası dijital eğlenceye olan ilginin arttığı [1] günümüzde, bu tarz bir uygulamanın kullanıcıların dikkatini çekeceği ve yaygın şekilde kullanılabileceği öngörülmektedir.

Metodoloji

Uygulama, dizileri türlerine göre filtreleyerek kullanıcılara kişiselleştirilmiş seçim imkânı sunar. Kullanıcı deneyimini ön planda tutan ara yüz tasarımı, sade ve etkileşimli bir kullanım sunarken; backend tarafında sunucu performansı ve veri yönetimine önem verilmiştir. Temel hedef, teknik açıdan hızlı ve sorunsuz bir deneyim sağlarken kullanıcı etkileşimini artırmaktır. İlerleyen süreçte yapay zeka destekli öneri sistemleriyle entegre edilerek, kullanıcıların geçmiş tercihlerine ve benzer profillere sahip diğer kullanıcıların seçimlerine dayalı daha isabetli dizi önerileri sunulması planlanmaktadır [2].

Amaçlar ve Hedefler

Projenin temel amacı, arkadaş gruplarının birlikte dizi seçme sürecini hızlandırarak keyifli bir sosyal deneyime dönüştürmektir. Uygulama, tüm grup üyelerinin eş zamanlı etkileşimde bulunmasını sağlayarak ortak bir dizi üzerinde daha hızlı anlaşmalarını kolaylaştırır. Bu sayede hem zaman kazanılır hem de grup içi sosyalleşme deneyimi güçlenir.

Amaçlar:

- Kullanıcı dostu ve sade bir arayüz ile kolay, hızlı bir kullanım deneyimi sunmak.
- Dizi zevklerini kategorilere ayırarak ortak ilgi alanlarını keşfetmeyi sağlamak.
- Grup içi iletişimi artırarak eğlenceli ve etkileşimli bir ortam oluşturmak.
- Dizi seçimini hızlandırarak etkinlik süresini daha verimli kullanmak.

Hedefler:

- İlk 3 ay içinde temel işlevleri kapsayan bir prototip geliştirmek.
- Tür bazlı filtreleme imkânı sunan bir dizi veri tabanı oluşturmak.
- Performans önceliği veren bir backend altyapısı kurmak.
- En az 50 veya daha fazla kullanıcıyla testler yaparak geri bildirim toplamak.
- Bu geri bildirimler doğrultusunda geliştirilen kararlı sürümü proje sonunda yayınlamak.

Planlama

Gözlemlenen gereksinimler sonucunda yapılması hedeflenen proje, arkadaşların birlikte dizi seçip izleme sürecini kolaylaştırmak amacıyla React Native ve Expo Framework kullanılarak geliştirilecek bir mobil uygulamayı kapsamaktadır [3]. Uygulama, arkadaşların grup içindeki etkileşimini artırmayı ve karar verme süreçlerini daha eğlenceli ve pratik hale getirmeyi hedeflemektedir. Proje süreci dört temel aşamada yürütülecektir: Literatür Taraması, Kullanılacak Programların Detaylı İncelemesi, Uygulama Geliştirme ve Test & Değerlendirme.

Prototip ve Sonuçlar



Prototip görsel sayfası görsellerden görüldüğü üzere 2 kısımdan oluşacak. Uygulama ilk açıldığında görünecek giriş yapmayı gerektiren kısım ve sonraki adımda oluşturulan odanın bilgilerinin olduğu kısım.

Dizi seçmeye başlamadan önce dizinin türünün seçileceği soldaki görsele benzer bir sayfa ile karşılaşılacak.





Seçim tamamlandıktan sonra karşımıza sağdaki gibi bir sayfa çıkacak ve iki tarafta beğendiklerini ve beğenmediklerini seçmeye başlayacak. Sağa kaydırma beğenme sola kaydırma beğenmeme şeklinde kabul edilecek.

KAYNAKÇA

[1] Erkılıç, H., & Duruel Erkılıç, S. (2021). COVID-19 Pandemisi Sürecinde Dijital Platformların Yükselişi: Sinema Değer Zincirindeki Değişim Sinema Endüstrisini Nasıl Etkiler?. Ankara Üniversitesi İlef Dergisi(2021 Özel Sayı: 2. Uluslararası Dijital Çağda İletişim Sempozyumu, Süreklilikler ve Kesintiler: Kültürler, Pazarlar ve Siyaset), 99-126. https://doi.org/10.24955/ilef.1037981

[2] Veenstra, A., Meers, P., & Biltereyst, D. (2020). Structured film-viewing preferences and practices: a quantitative analysis of hierarchies in screen and content selection among young people in Flanders. Digital Peripheries: The Online Circulation of Audiovisual Content from the Small Market Perspective, 227-244.

[3] Liberty, J., Juarez, R., & Montaquila, M. (2023). . NET MAUI for C# Developers: Build cross-platform mobile and desktop applications. Packt Publishing Ltd.

2024.-2025. Bahar Yarıyılı
İSTATİSTİK BÖLÜMÜ

BRENT PETROL FİYATLARININ LSTM İLE ZAMAN SERİSİ TAHMİNİ

Sabri Umutcan KONCA 19023024

Danışman: Prof. Dr. Serpil KILIÇ DEPREN

ÖZET

Bu çalışmada amaç, Brent petrol fiyatlarını kısa vadeli olarak öngörebilecek bir zaman serisi tahmin modeli geliştirmektir. Zaman bağımlı örüntüleri öğrenmede başarılı olan LSTM (Long Short-Term Memory) mimarisi tercih edilmiştir. LSTM, klasik istatistiksel modellere göre daha esnek ve güçlü bir öğrenme yapısına sahip olması nedeniyle seçilmiştir.
Modelin başarısı, geçmiş verilerle test edilerek değerlendirilmiş; sonuçlar, önerilen yöntemin kısa vadeli fiyat tahminleri için güçlü bir araç olduğunu göstermiştir.

GİRİŞ

Petrol, küresel ölçekte en çok işlem gören emtialardan biri olup, fiyatları ekonomik büyüme, jeopolitik riskler ve arz-talep dengesi gibi pek çok değişkenden etkilenmektedir. Bu oynak yapı, yatırımcıların ve karar vericilerin geleceğe yönelik belirsizliklerle karşılaşmasına neden olur. Özellikle Brent petrolü, dünya genelinde referans alınan önemli bir fiyat göstergesidir.

Bu çalışmada, Brent petrol fiyatlarının yakın vadede (30 gün) tahmin edilmesi amaçlanmaktadır. Klasik zaman serisi analiz yöntemleri, lineer varsayımlar nedeniyle karmaşık örüntüleri yakalamada yetersiz kalabilmektedir. Bu nedenle, geçmiş verilerdeki zamansal ilişkileri öğrenebilen, hafıza temelli bir yapı olan **LSTM (Long Short-Term Memory)** modeli tercih edilmiştir. LSTM, özellikle finansal zaman serilerinde etkili sonuçlar vermesiyle bilinen bir derin öğrenme mimarisidir.

Çalışma kapsamında veriler temizlenmiş, korelasyon analizi ile girdi değişkenleri optimize edilmiş ve eğitim/tahmin süreci boyunca modelin performansı detaylı şekilde değerlendirilmiştir.

YÖNTEM

Bu çalışmada Brent petrol fiyatlarının tahmini için geçmiş verilere dayalı bir LSTM modeli oluşturulmuştur. Modelleme süreci aşağıdaki adımlarla gerçekleştirilmiştir:

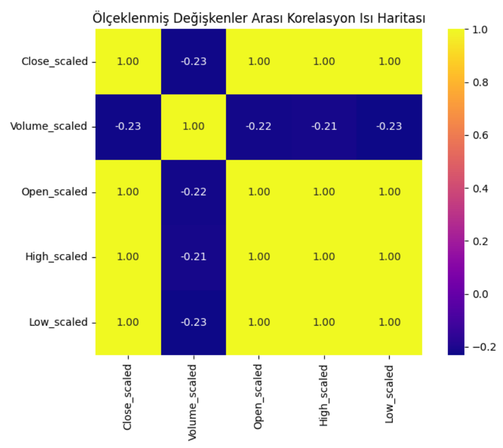
- **Veri Kaynağı:** 2013–2024 yılları arasındaki Brent petrol vadeli işlem fiyatları kullanılmıştır.

- **Ön İşleme:**

Verilerdeki eksik gözlemler temizlenmiştir.
Sayısal olmayan hacim verileri dönüştürülmüş ve tüm sütunlar uygun formata getirilmiştir.
Tarih sütunu sıralı hale getirilerek zaman serisi formatı oluşturulmuştur.

- **Özellik Seçimi:**

Korelasyon matrisi analiz edilmiştir.
Close fiyatı ile yüksek korelasyon gösteren Open, High, Low değişkenleri dışlanmış, sadece Volume değişkeni modele girdi olarak alınmıştır.



- **Ölçekleme:**

Close ve Volume sütunları MinMaxScaler ile [0, 1] aralığına normalize edilmiştir.

LSTM Yöntemi ve Model Yapısı

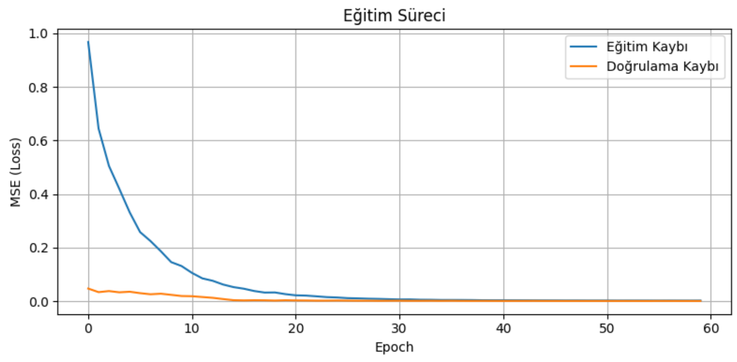
LSTM (Long Short-Term Memory), zaman serisi verilerde geçmişe dair önemli bilgileri uzun süre boyunca hatırlayabilen, tekrarlayan yapay sinir ağı (RNN) mimarisidir. Geleneksel RNN'ler zaman içinde "hafıza kaybı" yaşarken, LSTM bu sorunu özel olarak tasarlanmış hücre yapısıyla aşar.
Her LSTM hücresi üç temel yapı içerir:
Giriş kapısı (input gate): Yeni gelen bilginin ne kadarının hücreye aktarılacağını belirler.
Unutma kapısı (forget gate): Önceki bilgilerin hangilerinin silineceğini kontrol eder.
Çıkış kapısı (output gate): Hücrenin çıktısında hangi bilginin kullanılacağını belirler.
Bu mekanizma sayesinde LSTM, geçmişteki kritik olayların etkisini unutmaz; örneğin petrol fiyatlarında haftalar önce yaşanan bir değişimi model belleğinde tutabilir. Bu yetenek, özellikle ani sıçramalar, trend dönüşleri gibi finansal verilerde sık rastlanan durumlar için büyük avantaj sağlar.

Bu çalışmada kullanılan LSTM modeli şu şekilde yapılandırılmıştır:
Zaman penceresi: 90 günlük geçmiş veriye dayalı
Girdiler: Close_scaled ve Volume_scaled
Model mimarisi:
- 2 katmanlı LSTM (64 → 32 birim)
- BatchNormalization ve Dropout (0.3)
- Çıkış katmanı: Dense(1)

Optimizasyon: Adam algoritması
Kayıp fonksiyonu: Ortalama Kare Hata (MSE)
Bu yapı sayesinde model hem kısa vadeli fiyat hareketlerini öğrenmiş hem de düşük hata oranlarıyla tahminler üretmiştir.

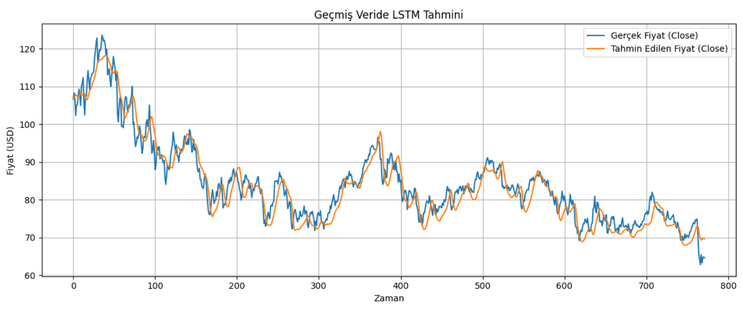
- **Eğitim Süreci:**

Eğitim %80, test %20 oranında bölünmüştür.
EarlyStopping ve ModelCheckpoint ile en iyi model korunmuştur.
60 epoch boyunca eğitim yapılmıştır.



- **Tahmin:**

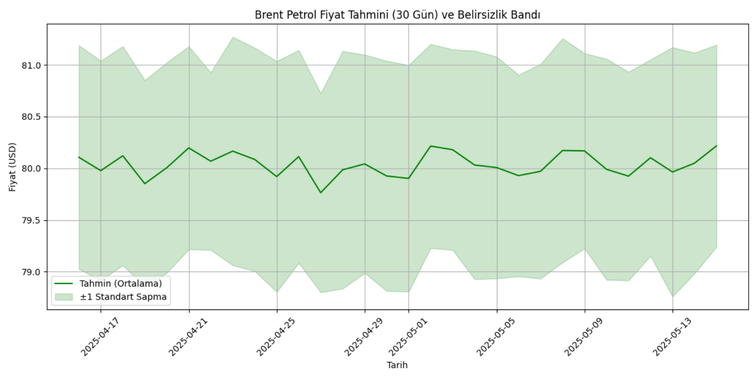
Monte Carlo yöntemi ile 100 kez örnekleme yapılmış ve ortalama tahminler ile ±1 standart sapma belirsizlik bandı hesaplanmıştır.
Sonuçlar görselleştirilmiş ve 30 günlük tahmin CSV dosyasına kaydedilmiştir.



SONUÇLAR

Model Performansı (Test Verisi):				
MSE	RMSE	MAE	MAPE	R^2
10.83	3.29	2.65	%3.16	0.9128

30 günlük ortalama fiyat tahmini yapılmış, ±1 std sapmalı belirsizlik bandı çizilmiştir.



Bu metrikler, modelin hata oranlarının düşük ve açıklayıcılığının yüksek olduğunu göstermektedir. Özellikle %3.16'lık MAPE oranı, finansal zaman serileri için oldukça tatmin edici bir tahmin gücüne işaret etmektedir.
Gelecek 30 gün için yapılan çoklu tahmin örneklemesi sayesinde ortalama fiyatlar hesaplanmış ve ±1 standart sapma ile belirsizlik bandı oluşturulmuştur. Bu yaklaşım, klasik nokta tahminlerine kıyasla daha güvenilir ve karar destekleyici bir yapıdır.
Tartışma:
Modelin başarısında, korelasyon analizi ile sadeleştirilmiş veri seti ve zaman penceresi yapısı etkili olmuştur. Fakat model yalnızca geçmiş fiyat ve hacim bilgisine dayanmakta, jeopolitik olaylar veya makroekonomik değişkenler gibi dışsal faktörler dahil edilmemiştir. Bu durum, ani şokları içermeyen koşullarda modelin başarısını artırırken, olağanüstü durumlar için sınırlayıcı olabilir. Ayrıca, kısa vadeli tahmin performansı güçlü olsa da uzun vadeli öngörüler için farklı mimariler veya melez modeller tercih edilmelidir.

KAYNAKÇA

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
[2] Brownlee, J. (2017). Deep Learning for Time Series Forecasting. Machine Learning Mastery.
[3] Zhang, K., & Hong, M. (2022). Forecasting crude oil price using LSTM neural networks. *Data Science in Finance and Economics*, 2(3), 163–180.

YTÜ FEN EDEBİYAT FAKÜLTESİ

2024-2025 Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

İçerik Bazlı Filtreleme Sistemi ile Film Öneri Uygulaması

SUDE NUR UÇAR - 18023061 MELİSNUR TURAN - 18023080

Doç. Dr. Selçuk ALP

Dijital medya platformlarının yaygınlaşması, görsel-işitsel içeriklerin, özellikle filmlerin sayısında artışa yol açmış ve bu da kullanıcıların seçim yapmasını zorlaştırmıştır. Bu soruna çözüm olarak geliştirilen öneri sistemleri, kullanıcıların tercihlerine uygun içeriklere daha hızlı ulaşmalarını sağlar. İçerik bazlı filtreleme yöntemi, kullanıcıların beğendiği filmlerin özelliklerini analiz ederek benzer filmler önerir. Bu çalışmada, öneri sistemlerinin çalışma mantığı ele alınmış ve içerik bazlı filtreleme kullanan bir film öneri uygulaması geliştirilmiştir. Amaç, kişiye özel önerilerle kullanıcı deneyimini iyileştirmektir. Ayrıca, öneri sistemlerinin dijital içerik tüketimi ve bilgiye erişim üzerindeki etkisi somut bir örnekle değerlendirilmiştir.

Öneri Sistemleri Nedir?

Öneri sistemleri, bir kullanıcının geçmiş davranışlarını, ilgi alanlarını, demografik özelliklerini ve benzer kullanıcıların tercihlerine dair verileri analiz ederek, ilgili kullanıcıya en uygun içerik veya ürünü sunmayı amaçlayan algoritmalardır. Bu sistemler, kullanıcıların tercihlerini tahmin etmek ve karar verme süreçlerini kolaylaştırmak amacıyla geliştirilmiştir. Günümüzde öneri sistemleri; film, müzik, kitap, haber, sosyal medya gönderileri, akademik içerikler ve e-ticaret ürünleri gibi pek çok alanda yaygın şekilde kullanılmaktadır.

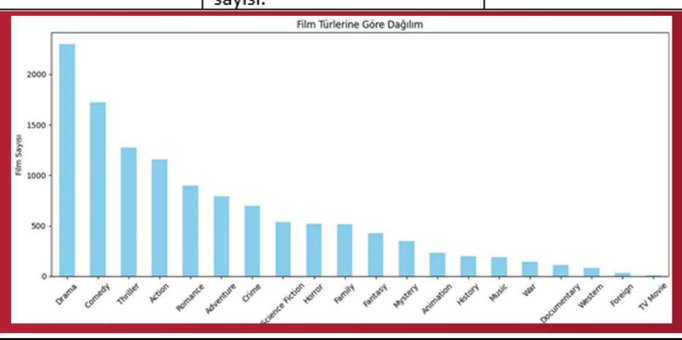
İçerik Bazlı Film Öneri Sistemi

Film öneri sistemlerinin amacı, kullanıcının ilgi alanlarına göre filmleri önermektir. Bu, kullanıcının mevcut binlerce film arasından interneti tarayarak zaman kaybetmesini önler. İçerik tabanlı öneri sistemi, kullanıcılara önerilebilecek öğelerin özelliklerini tanımlar. Bir veri setine dayanarak, daha önce beğenilen filmlerdeki özellikleri dikkate alarak kullanıcının beğeneceği filmleri tahmin eder. Öneri sistemleri, bir veya birden fazla özelliğe dayanarak film önerebilir. Bir film öneri sistemi tasarlarken, film türü, yönetmen veya oyuncular gibi çeşitli faktörler göz önünde bulundurulur. Bu çalışmada, öneri sistemi oyuncu kadrosu, anahtar kelimeler, ekip ve türler üzerine inşa edilmiştir. Bu dört özelliğin toplamından oluşan tek bir sütun yaratılmış ve bu, film öneri sistemi için baskın bir faktör olarak kullanılmıştır.

Veri Seti

Bu tez çalışmasında kullanılan veri seti, film endüstrisine yönelik en kapsamlı ve açık kaynaklı veri platformlarından biri olan The Movie Database (TMDB) tarafından sağlanan TMDB 5000 Movie Dataset'tir. Bu veri seti, film öneri sistemleri üzerine yapılan akademik ve uygulamalı çalışmalarda yaygın olarak tercih edilen, zengin içerikli ve çok boyutlu bir veri kaynağıdır. Veri seti, toplam 4.803 film ve bu filmlerle ilişkili oyuncu, yönetmen, tür, anahtar kelime, özet, dil, popülerlik gibi çeşitli bilgileri içermektedir. Veri, iki farklı .csv (Comma-Separated Values) formatındaki dosya olarak sunulmaktadır: credits.csv ve movies.csv.

Değişken Adı	Açıklama	Değişken Türü
Budget	Filmin yapım bütçesi (USD cinsinden). 0 değeri eksik veri olarak değerlendirilir.	Sayısal (Numerik)
Genres	Filmin tür veya türleri. JSON formatında listelenmiştir.	Kategorik (Çoklu Etiketli)
Homepage	Filmin resmi internet sitesi URL'si.	Metinsel (String)
ID	Her filme ait benzersiz kimlik numarası.	Sayısal (Tam sayı)
Keywords	Filmi içeriksel olarak tanımlayan anahtar kelimeler. JSON biçiminde.	Kategorik (Çoklu Etiketli)
Original_language	Filmin orijinal dili (örnek: 'en', 'fr').	Kategorik (Tek Etiketli)
Original_title	Filmin orijinal adı.	Metinsel (String)
Overview	Filmin kısa özeti veya tanıtımı.	Metinsel (String)
Popularity	TMDB tarafından hesaplanan popülerlik puanı.	Sayısal (Ondalıklı)
Production_companies	Filmin yapımında görevli şirket(ler). JSON formatında listelenmiştir.	Kategorik (Çoklu Etiketli)
Production_countries	Yapımın gerçekleştirildiği ülke veya ülkeler.	Kategorik (Çoklu Etiketli)
Release_date	Filmin vizyon tarihi (tarih formatında).	Tarih (Datetime)
Revenue	Filmin gişe geliri (USD cinsinden).	Sayısal (Numerik)
Runtime	Filmin süresi (dakika olarak).	Sayısal (Numerik)
Spoken_languages	Filmde konuşulan dillerin listesi.	Kategorik (Çoklu Etiketli)
Status	Filmin üretim durumu ('Released', 'Post Production' vb.).	Kategorik
Tagline	Filmin tanıtım sloganı.	Metinsel (String)
Title	Filmin genel adı.	Metinsel (String)
Vote_average	Filmin ortalama kullanıcı puanı (0–10 arası).	Sayısal (Ondalıklı)
Vote_count	Oylamaya katılan kullanıcı sayısı.	Sayısal (Tam sayı)



Film Öneri Sistemi Uygulaması

```
[1]: import numpy as np
     import pandas as pd
```

Python programlama dilinde, bilimsel hesaplamalar ve veri analizi süreçlerinde sıklıkla kullanılan numpy ve pandas kütüphaneleri, programın başında import ifadesiyle projeye dahil ediyoruz.

```
[2]: credits = pd.read_csv(r"C:\Users\user\OneDrive\Desktop\Tez\tmdb_5000_credits.csv")
     movies = pd.read_csv(r"C:\Users\user\OneDrive\Desktop\Tez\tmdb_5000_movies.csv")
```

Sonrasında, pandas kütüphanesinin veri okuma işlevlerinden biri olan read_csv() fonksiyonu kullanılarak, iki ayrı .csv dosyasında bulunan veri setleri içeri aktarılıyor.

```
[3]: print("1 :" , credits.info(),"\n")
     print("2 :" , movies.info(),"\n")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   homepage              1712 non-null   object
 3   id                    4803 non-null   int64
 4   keywords              4803 non-null   object
 5   original_language     4803 non-null   object
 6   original_title        4803 non-null   object
 7   overview              4800 non-null   object
 8   popularity            4803 non-null   float64
 9   production_companies  4803 non-null   object
 10  production_countries  4803 non-null   object
 11  release_date          4802 non-null   object
 12  revenue               4803 non-null   int64
 13  runtime               4801 non-null   float64
 14  spoken_languages      4803 non-null   object
 15  status                4803 non-null   object
 16  tagline               3959 non-null   object
 17  title                 4803 non-null   object
 18  vote_average          4803 non-null   float64
 19  vote_count            4803 non-null   int64
dtypes: float64(3), int64(4), object(13)
memory usage: 750.6+ KB
1 : None

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4803 entries, 0 to 4802
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   budget  4803 non-null   int64
 1   id      4803 non-null   int64
 2   title   4803 non-null   object
 3   cast    4803 non-null   object
dtypes: int64(2), object(2)
memory usage: 150.2+ KB
2 : None
```

Bu işlemden sonra, pandas kütüphanesinin bir metodu olan info() fonksiyonu uygulanarak veri setlerinin yapısal özellikleri inceleniyor. Bu metod, veri çerçevesinde yer alan sütunların isimlerini, veri tiplerini, eksik veri durumlarını ve bellekte kapladıkları alanı özet olarak sunuyor. Elde edilen bilgiler, print() komutu kullanılarak konsola yazdırılıyor.

İki veriyi de inceledikten sonra fark ediliyor ki filmlerin kimlik numaraları movies veri setinde 'id' ile isimlendirilirken, credits veri setinde ise 'movie_id' olarak isimlendiriliyor. Bu iki veri setini filmlerin kimlik numaraları üzerinden birleştirebilmek için credits'deki 'movie_id' değişkeninin ismini movies veri setindeki değişkenle aynı, yani 'id' olarak değiştiriliyor. Bunu yapmak için yine bir pandas özelliği olan rename metodu kullanılıyor.

```
[4]: credits = credits.rename(columns={"movie_id": "id"})
     df_merge = pd.merge(credits, movies, left_on=['title','id'], right_on=['title','id'])
```

Bu aşamada, metinlerdeki kelimelerin önemini ölçmek için TF-IDF yöntemi kullanılan TfidfVectorizer sınıfından bir nesne oluşturuluyor. TF-IDF, bir terimin belgede ne kadar sık geçtiğini (TF) ve kaç farklı belgede yer aldığını (IDF) değerlendirerek, o terimin önemini belirler. Ardından, fit_transform() fonksiyonu ile df_merge veri çerçevesindeki film özetlerini içeren "overview" sütunu işleniyor. values.astype('U') ifadesi, bu metinleri Unicode formatına çevirerek tutarlı şekilde analiz edilmesini sağlıyor. Sonuç olarak, her belgenin sayısal vektörlerle temsil edildiği bir TF-IDF matrisi elde ediliyor. Bu matris, belgeler arası benzerlik hesaplamalarında ve doğal dil işleme uygulamalarında temel veri olarak kullanılıyor.

```
[5]: from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.metrics.pairwise import cosine_similarity
```

Burada TfidfVectorizer, her filmin özetini sayısal vektörlere dönüştürmek için kullanılır.
Böylece her film metinsel bir içerikten sayısal bir yapıya çevrilmiş olur.

```
[6]: tfidf_vectorizer = TfidfVectorizer()
     overview_tfidf_matrix = tfidf_vectorizer.fit_transform(df_merge["overview"].values.astype('U'))
```

Gözlemler arasındaki benzerlik hesaplandıktan sonra kullanıcıdan bir film adı girmesi isteniyor. Kullanıcının girdiği film ismi film_adi değişkeninde tutulduktan sonra data frame içerisinde bu filmin olup olmadığını kontrol ediliyor ve eğer varsa bu filmin adı film değişkenine atanıyor.

```
[ ]: film_adi = input("Lütfen Filmin Adını Giriniz")
     film = df_merge[df_merge["title"] == film_adi]
```

Kullanıcı tarafından girilen filmin veri setinde bulunup bulunmadığı, if koşulu içerisinde empty yöntemi kullanılarak kontrol ediliyor.

Eğer aranan film veri setinde mevcut değilse, yani ilgili değişken empty değeri döndürüyorsa, print fonksiyonu aracılığıyla kullanıcıya filmin veri setinde bulunmadığına dair bilgilendirici bir uyarı mesajı gösteriliyor.

```
[ ]: if film.empty:
         print("Üzgünüm istediğiniz filmi bulamadım :( " )
     else:
         film_turu = film["genres"].values[0]
         film_tfidf = tfidf_vectorizer.transform(film["overview"].values.astype('U'))
         benzerlik_skorlari = cosine_similarity(film_tfidf, overview_tfidf_matrix)
         oneriler = df_merge.iloc[np.argsort(benzerlik_skorlari[0])[-6:-1][::-1]

         print("Seçilen film: ",film_adi)
         film_turu_str = ", ".join([genre["name"] for genre in eval (film_turu)])
         print("Film Türü:", film_turu_str,"\n")
         print("Benzer filmler:")
         print(oneriler["title"])
```

```
Lütfen Filmin Adını GirinizTwilight
Seçilen film:  Twilight
Film Türü: Adventure, Fantasy, Drama, Romance

Benzer filmler:
898     The Twilight Saga: New Moon
612      The Twilight Saga: Eclipse
2563          Because of Winn-Dixie
2730                    Black Swan
938                     Unfaithful
Name: title, dtype: object
```

Kullanıcının seçtiği Twilight filmi, vampirler ve insanlarla geçen romantik bir gençlik drama-macera türü yapımdır. Önerilen filmler incelendiğinde ise benzer temalar doğrultusunda, doğaüstü öğeler içeren, gençlik odaklı romantik drama veya macera türündeki filmlerin listelendiği gözlemlenir. Bu sonuçlar, içerik tabanlı öneri algoritmasının hem tematik hem de anlamsal bazda doğru ve anlamlı öneriler sunduğunu göstermektedir. Böylece kullanıcı, seçtiği filmle yüksek oranda benzerlik taşıyan diğer yapımlara kolaylıkla erişebilir.

Kullanıcı, uygulamaya veri setinde bulunmayan Barbie filmini giriş olarak verdiğinde, sistem öncelikle bu filmin veri setinde mevcut olup olmadığını kontrol eder. Veri setinde Barbie filmi hakkında herhangi bir kayıt bulunmadığı için, ilgili koşul bloğunda yer alan if kısmı devreye girer. Bu durumda, kullanıcıya filmin veri setinde bulunmadığını belirten anlamlı ve bilgilendirici bir hata mesajı gösterilir.

```
Lütfen Filmin Adını Giriniz Barbie
Üzgünüm istediğiniz filmi bulamadım :(
```

Kaynakça

[1]https://www.kaggle.com/code/vuslatsena/bitirme-tezi-movie-recommendation-system

[2]https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=aNSqBTe5CUltH9zSOeFA7w&no=PQKoj_T7Wuvfbn1Irzdjuw

[3]https://dergipark.org.tr/en/download/article-file/1640671

2024-2025 2. Yarıyılı

İstatistik BÖLÜMÜ

PLAYOFF PULSE

Taha Aziz Koca 20023042

Danışman: Doç. Dr. REŞİT ÇELİK

ÖZET

Bu çalışmada, NBA oyuncularının normal sezon ve playoff maçlarındaki performans farklılıkları karşılaştırmalı olarak incelenmiştir. Hazırlanan web sitesi aracılığıyla, kullanıcılar oyuncu bazlı istatistikleri dönemsel olarak görüntüleyebilmekte ve kritik anlarda performans değişimlerini analiz edebilmektedir. Çalışmanın amacı, oyuncu verimliliğindeki değişimleri görsel ve interaktif bir şekilde sunarak basketbol analizine katkı sağlamaktır.

PLAYOFF PULSE

BASKETBOLDA İSTATİSTİK NEDİR

Basketbolda istatistik, oyuncuların ve takımların maç içinde gerçekleştirdiği performansların sayısal olarak kaydedilmesi ve analiz edilmesidir. Bu veriler; sayı, asist, ribaund, top çalma, blok, top kaybı, şut yüzdesi gibi temel unsurları içerir.

AMAÇ

Oyuncuların katkı düzeyini objektif bir şekilde ölçmek, taktik analiz yapmak ve performansları karşılaştırmaktır. Basketbol istatistikleri, antrenörler için oyun planı oluşturma; yöneticiler için transfer kararları alma; araştırmacılar için veri analizi yapma gibi birçok alanda önemli rol oynar. Aynı zamanda izleyicilere ve taraftarlara da oyun hakkında daha derin bir anlayış kazandırır.

BASKETBOLDA İSTATİSTİĞİN ÖNEMİ

Basketbolda istatistikler, oyunu anlamak, analiz etmek ve geliştirmek için vazgeçilmez bir araçtır. Oyuncuların ve takımların performanslarını sayısal verilerle ortaya koyar. Bu sayede:

- **Antrenörler** strateji geliştirirken hangi oyuncunun ne kadar etkili olduğunu görebilir, oyun planlarını buna göre şekillendirebilir.
- **Yöneticiler** istatistikler üzerinden transfer kararları alabilir, oyuncu seçimlerini veriye dayalı yapabilir.
- **Sporcular** kendi güçlü ve zayıf yönlerini belirleyip gelişim alanlarını tespit edebilir.
- **Analistler ve araştırmacılar** performans karşılaştırmaları, takım verimliliği ve maç tahminleri üzerine detaylı çalışmalar yapabilir.
- **Taraftarlar ve medya** oyuncu ve takım performanslarını daha objektif değerlendirme şansı elde eder.

Kısacası, istatistikler modern basketbolun bilimsel temele oturmasını sağlar ve oyunun her alanında bilinçli kararlar alınmasına katkı sunar.

NORMAL SEZON VE PLAYOFFLARDA OYUNCULARIN YÜK FARKI

Basketbolda normal sezon, uzun süren ve daha düşük tempolu bir dönemdir. Bu süreçte oyuncuların yükü genellikle dakika yönetimi, istikrar ve sakatlıktan kaçınma üzerine kuruludur. Takımlar rotasyonu daha geniş tutar ve oyunculara zaman zaman dinlenme fırsatları sunulur.

Playofflar ise yüksek tempolu, daha fiziksel ve mental olarak daha yoğun geçen maçlardan oluşur. Her maçın önemi arttığı için takımların yıldız oyuncularına daha fazla sorumluluk verilir. Dakika süreleri artar, savunma baskısı yükselir ve bireysel hataların affedilmediği bir atmosfer oluşur. Oyuncular hem fiziksel olarak daha fazla zorlanır hem de baskı altında karar verme becerilerini üst seviyeye çıkarmak zorundadır.

Bu nedenle, aynı oyuncunun normal sezon ve playoff performansı arasında önemli farklar görülebilir. Playofflar, gerçek liderlik ve dayanıklılık becerilerinin öne çıktığı dönemlerdir.

PLAYOFF'LARDA İYİ OYNAYAN OYUNCULAR HAKKINDA BİLGİ

Bu oyuncular, sezon boyunca gösterdikleri performansı kritik anlarda daha da yükselten oyunculardır. Genellikle:

- Baskı altında soğukkanlı kalabilirler.
- Takımlarının en çok güvendiği skor opsiyonlarıdır.
- Kritik anlarda doğru kararlar verirler.
- Liderlik vasıfları güçlüdür ve maçın gidişatını değiştirebilirler.
- Örnek: Jimmy Butler, Kawhi Leonard, Lebron James gibi isimler kariyerlerinin çoğu döneminde playofflarda iyi oynamış isimlerdir.

PLAYOFF'LARDA KÖTÜ OYNAYAN OYUNCULAR HAKKINDA BİLGİ

Bazı oyuncular, normal sezonda çok iyi istatistikler sunsa da playofflarda:

- Artan savunma sertliği karşısında zorlanabilirler.
- Fiziksel ve mental baskıyla başa çıkmakta güçlük çekebilirler.
- Dakika ve rol değişimlerinden olumsuz etkilenebilirler.
- Takımlar playofflarda daha fazla analiz yaptığı için zayıf yönleri daha fazla açığa çıkar.
- Örnek: Ben Simmons(2021 Playoffları), Demar Derozan, Julius Randle (2021 ve 2023 playoffları) gibi İsimler kariyerlerinin belli dönemlerinde playoffları kötü oynayan oyuncular.dır.

2023-2024 NORMAL SEZON – PLAYOFF PERFORMANS ÖRNEKLERİ(SAYI İLK 10)

Rank	Player	Team	GP	MIN	Sayılar
#1	1	Joel Embiid	6	41.3	33
#2	2	Jalen Brunson	13	39.8	32.4
#3	3	Shai Gilgeous-Alexander	10	39.9	30.2
#4	4	Tyrese Maxey	6	44.5	29.8
#5	5	Donovan Mitchell	10	38.2	29.6
#6	6	Luka Dončić	22	40.9	28.9
#7	7	Nikola Jokić	12	40.2	28.7
#8	8	Anthony Davis	5	41.6	27.8
#9	8	LeBron James	5	40.8	27.8
#10	10	Anthony Edwards	16	40.6	27.6

Rank	Player	Team	GP	MIN	Sayılar
#1	1	Joel Embiid	6	41.3	33
#2	2	Jalen Brunson	13	39.8	32.4
#3	3	Shai Gilgeous-Alexander	10	39.9	30.2
#4	4	Tyrese Maxey	6	44.5	29.8
#5	5	Donovan Mitchell	10	38.2	29.6
#6	6	Luka Dončić	22	40.9	28.9
#7	7	Nikola Jokić	12	40.2	28.7
#8	8	Anthony Davis	5	41.6	27.8
#9	8	LeBron James	5	40.8	27.8
#10	10	Anthony Edwards	16	40.6	27.6

2023–2024 sezonunda normal sezon ve playoff sayı ortalamaları karşılaştırıldığında, normal sezonda daha fazla oyuncunun istikrarlı skorer olduğu görülürken, playofflarda rol oyuncularının geri maç başına oynama süresi artmış ve bu da sayı ortalamalarına doğrudan yansımış.

Bazı oyuncular (örneğin Jalen Brunson ve Nikola Jokić) playofflarda performanslarını daha da yukarı taşıyarak yüksek baskı altında liderlik sergilemişlerdir. Bu tür oyuncular, "playoff oyuncusu" tanımını karşılamaktadır.

Diğer yandan, Luka Dončić gibi yıldızlar, yüksek dakika ve sorumlulukla oynamalarına rağmen playofflarda bir miktar düşüş yaşamıştır. Bu da playoff savunmalarının sertliği ve rakip analizlerinin derinliğiyle ilişkilidir.

Ayrıca normal sezon skorerleri arasında yer almayan Lebron James, Anthony Davis ve Anthony Edwards gibi isimlerin playofflarda öne çıkması, tecrübe ve fiziksel güç faktörlerinin bu dönemde ne kadar belirleyici olduğunu göstermektedir.

SONUÇ

- Bu çalışma, istatistiksel verilerin yalnızca rakamlarla değil, oyuncuların üzerindeki yük, sorumluluk ve oyun bağlamı ile birlikte değerlendirilmesi gerektiğini ortaya koymaktadır.
- Bazı oyuncular playofflarda yükselirken, bazıları normal sezondaki formunu sürdürememekte; bu da oyuncu profillerinin esneklik ve baskı yönetimi açısından farklılaştığını göstermektedir.
- Bu tür karşılaştırmalar, hem teknik ekiplerin stratejik kararlarını hem de analistlerin ve istatistikçilerin değerlendirmelerini daha anlamlı hale getirmektedir.
- İleri düzey istatistiksel analizler, oyuncu değerlemesinde sadece sayı değil, verimlilik, savunma katkısı, clutch performans gibi karmaşık metriklerin de dikkate alınmasını gerektirir.
- Bu bağlamda istatistiksel performans analizleri,, aynı zamanda gelecekteki performans beklentileri ve kadro planlamaları için de temel oluşturur. sadece geçmişi yorumlamakla kalmaz

KAYNAKÇA

[1] National Basketball Association. (2024). *NBA player stats: 2023–2024 regular season and playoffs*. Retrieved from https://www.nba.com/stats

[2] Basketball Reference. (2024). *2023–24 NBA season summary*. Retrieved from https://www.basketball-reference.com

[3] StatMuse. (2024). *NBA player performance comparisons*. Retrieved from https://www.statmuse.com

2024-2025 Bahar Yarıyılı İstatistik Bölümü

HETEROJEN VERİLERDE JEOPOLİTİK RİSKİN BELİRLENMESİ:

QRMIX MODELİ İLE AMPİRİK BİR İNCELEME

Tuna ÇİLEK 22023601

Danışman: Prof. Dr. Serpil KILIÇ DEPREN

ÖZET

Bu çalışmada, Amerika Birleşik Devletleri ve Rusya örnekleminde jeopolitik riskin makroekonomik belirleyicileri, parametrik olmayan ve veri içi heterojenliği dikkate alan Kantil Regresyon Karma Modeli (QRMIX) aracılığıyla incelenmiştir. QRMIX modeli, klasik ortalama odaklı yöntemlerin aksine, farklı risk düzeylerinde değişkenlerin etkilerini kantil bazında ayrı ayrı analiz ederek daha esnek ve dağılımın tamamına duyarlı sonuçlar sunmaktadır. Bu çalışmanın bulguları, özellikle yüksek jeopolitik risk dönemlerinde ekonomik değişkenlerle olan ilişkilerin daha güçlü ve istatistiksel olarak anlamlı hale geldiğini göstermektedir. Bu tez, QRMIX modelinin jeopolitik risk analizine uygulanabilirliğini ortaya koyarak, hem teorik literatüre katkı sunmakta hem de ekonomik karar alıcılar için politika önerileri geliştirmeyi amaçlamaktadır.

Jeopolitik Risk

Jeopolitik risk, bir ülkenin veya bölgenin siyasi istikrarsızlığı, uluslar arası ilişkilerinde yaşadığı gerilimler, yaşanan askeri çatışmalar, ticari anlaşmazlıklar ve benzeri olaylardan kaynaklanan ve ülkelerin ekonomik faaliyetlerini, bu ülkede yaşayan veya yabancı yatırımcıların aldığı yatırım kararlarını ve finansal piyasaları doğrudan etkileyen risklerdir. Jeopolitik riskin ülke ekonomileri üzerindeki etkisinin anlamlılığı literatürdeki çalışmalar sayesinde kanıtlanmıştır.

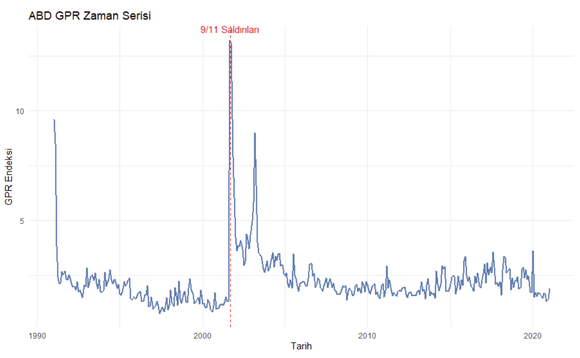


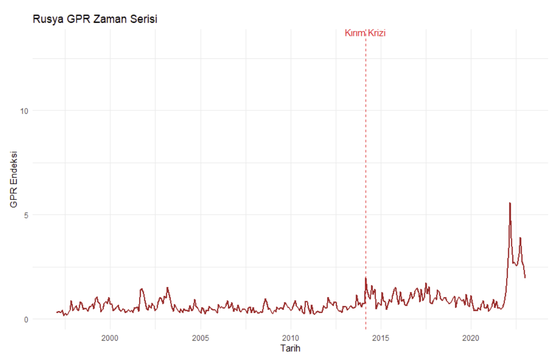


Veri Seti ve Analizi

Çalışmada pilot ülke olarak **Amerika Birleşik Devletleri** ve **Rusya** seçilmiş; bu ülkelere ilişkin **1991–2021 (ABD)** ve **1997–2023 (Rusya)** dönemlerine ait **aylık veriler** kullanılmıştır. Veri setindeki yıllık frekanstaki değişkenleri aylığa dönüştürmek için lineer enterpolasyon yöntemi uygulanırken, günlük verileri aylığa dönüştürmek adına aylık ortalama alınmıştır. Veri seti bağımsız değişken olarak; işsizlik oranı, faiz oranı, döviz kuru, brent petrol fiyatı, politika belirsizlik endeksi ve volalite endeksini içermektedir. Veri setinde, **ekonomik krizler, savaş dönemleri ve politika belirsizlikleri** gibi yüksek risk içeren dönemler de bulunmaktadır. Bu sayede model, **uç değerler ve heterojen yapılar** açısından zengin bir analiz alanı sunmuştur.

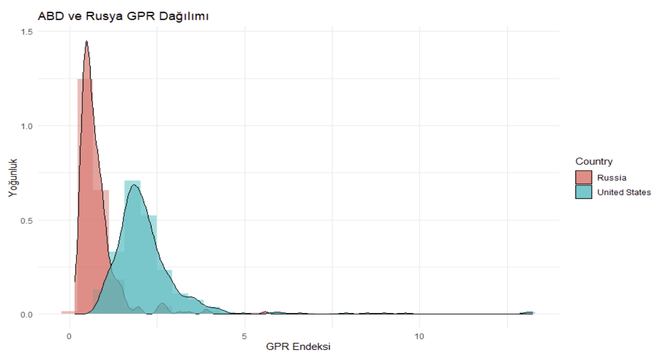
Jeopolitik Risk Endeksi Değerlerinin Ülkelere Göre Değişimi





Bu grafikler, ABD (1991–2021) ve Rusya (1997–2023) için jeopolitik risk endeksinin (GPR) zaman içindeki değişimini göstermektedir. ABD'de, 1991 yılının başındaki risk seviyelerindeki artış soğuk savaş döneminin bitişi ve Körfez krizinin etkilerini ortaya koymaktadır. 2001'deki 11 Eylül saldırıları sonrasında ise keskin bir yükseliş görülmektedir. Rusya'da ise 2014 Kırım Krizi ve 2022'de başlayan Ukrayna Savaşı, belirli bir süre boyunca stabil ilerleyen endekste belirgin sıçramalara neden olmuştur. Her iki ülke için de siyasi kriz dönemleri, GPR'ın ani ve yüksek artışlarla tepki verdiği kritik noktaları oluşturmaktadır.

Jeopolitik Risk Endeksi Dağılımlarında Ülke Bazlı Farklılıklar



ABD ve Rusya'nın GPR dağılımları incelendiğinde, her iki ülkenin de dağılımı sağa çarpık olmakla birlikte, ABD'nin risk değerlerinin daha geniş bir aralıkta yayıldığı ve yüksek risk dönemlerinin daha sık gözlendiği görülmektedir. Rusya'nın GPR değerleri ise genellikle düşük seviyelerde yoğunlaşmakta, yalnızca Kırım Krizi ve Ukrayna Savaşı gibi sınırlı dönemlerde ani sıçramalar göstermektedir.

Jeopolitik Risk Seviyelerine Göre İstatistiksel Farklılıkların Analizi

Veri setindeki dağılım özelliklerini incelemek için öncelikle Shapiro-Wilk testiyle GPR değişkeninin normalliği test edilmiştir. Hem ABD hem de Rusya için dağılımın normal olmadığı görülmüş, bu nedenle analizde parametrik olmayan testler tercih edilmiştir. Mann-Whitney U testi, ABD ve Rusya'nın GPR dağılımlarının farklı olduğunu göstermiştir. GPR seviyelerine göre değişkenlerin farklarını belirlemek amacıyla Kruskal-Wallis testi uygulanmış ve anlamlı farklar tespit edilmiştir. Farklılıkların hangi gruplar arasında olduğunu görmek için yapılan Dunn testi, özellikle yüksek risk grubunun diğer gruplardan ayrıştığını ortaya koymuştur. Rusya için işsizlik oranı ve ekonomik büyüme oranlarında, ABD için ise volatilite endeksi ve faiz oranlarında yüksek risk dönemleri ile diğer dönemler arasında anlamlı farklar bulunmuştur. Bu bulgular, ekonomik değişkenlerin GPR üzerindeki etkilerinin risk düzeyine göre değiştiğini ve QRMIX modelinin kullanılmasının daha sağlıklı olacağını göstermiştir.

Yöntem

Bu çalışmada, **verideki heterojen yapıları ve uç değerleri modellemek amacıyla** QRMIX (Kantil Regresyon Karma Modeli) yöntemi kullanılmıştır. QRMIX, **kantil regresyonun uç değerleri yakalayabilme gücü** ile **karışım modellerinin alt grup belirleyebilme yetisini** birleştirerek, farklı risk seviyelerinde değişkenlerin etkisini analiz etmeye olanak sağlar.
Model, veri setinde **gizli alt gruplar** tespit ederek her grup için ayrı kantil regresyon denklemleri kurar. Böylece, **düşük, orta ve yüksek jeopolitik risk dönemlerinde** ekonomik değişkenlerin GPR üzerindeki etkileri ayrı ayrı analiz edilebilmektedir. QRMIX, geleneksel ortalama odaklı yöntemlere kıyasla, **dağılımın tamamına duyarlı ve uç değerlere karşı dayanıklı** bir çözüm sunduğu için jeopolitik riskin yüksek seviyelerinde ortaya çıkan belirgin ekonomik etkilerin daha sağlıklı ve detaylı bir şekilde incelenmesine olanak tanımıştır.

Modelleme Süreci

Modelleme aşamasında, bağımlı değişken olan GPR değerlerinin sağa çarpık dağılım göstermesi nedeniyle logaritmik dönüşüm uygulanmıştır. Ardından, QRMIX (Kantil Regresyon Karma Modeli) yöntemiyle her iki ülke için ayrı modeller kurulmuştur. Bu model, dağılım içindeki gizli alt grupları belirleyerek her grup için farklı kantil regresyon doğruları üretmektedir. Modelde kullanılacak kantil sayısı farklı alternatifler denenerek (K = 2–6) belirlenmiş ve en iyi R^2 değerini sağlayan model en iyi model olarak seçilmiştir. Sonuçlar, ABD ve Rusya için yüksek risk dönemlerinde ekonomik değişkenlerin GPR üzerindeki etkisinin daha belirgin olduğunu ortaya koymuştur. Böylece QRMIX modeli, jeopolitik riskin sadece ortalama değil, tüm dağılım boyunca nasıl farklı şekillendiğini göstermiştir.

QRMIX Model Performansları

ABD – QRMIX Sonuçları

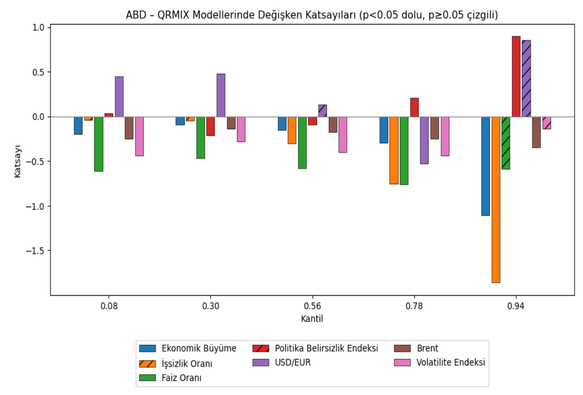
KANTİL SAYISI	RMSE	MAE	R^2
2	0.3608155	0.3106297	0.2006686
3	0.1974852	0.1727627	0.6542960
4	0.1805847	0.1556979	0.6814374
5	0.1304997	0.1115662	0.7575100
6	0.1289560	0.1022736	0.7536992

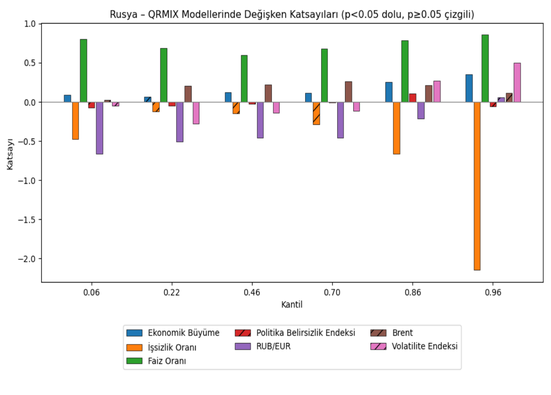
Rusya – QRMIX Sonuçları

Kantil Sayısı	RMSE	MAE	R^2
2	0.2616512	0.2048689	0.8397990
3	0.2135384	0.1530200	0.9036020
4	0.2290411	0.1587110	0.9026411
5	0.1883886	0.1358940	0.9241869
6	0.1461728	0.1098635	0.9518311

Her iki ülke için 2–6 arasında değişen kantil sayıları denenmiş, model performansı RMSE, MAE ve R^2 kriterlerine göre değerlendirilmiştir. ABD verisi için en iyi sonuç 5 kantilli modelle elde edilirken, Rusya verisi için 6 kantilli model en yüksek başarıyı sağlamıştır. Bu durum iki ülke içerisinden Rusya'nın jeopolitik risk açısından daha heterojen bir dağılım yapısına sahip olduğunu işaret etmektedir.

Değişkenlerin Etkileri ve Anlamlılık Düzeyleri





QRMIX modellerinde **atanan kantiller**, verideki gizli gruplar tanımlanırken her bir gözlemi en iyi temsil eden kantil düzeyini belirtir. Örneğin, ABD için modelin belirlediği 5 gizli alt grup, 0.08, 0.30, 0.56, 0.78 ve 0.94 kantillerine karşılık gelmektedir. Bu kantiller, GPR dağılımı boyunca farklı risk seviyelerine karşılık gelir. Rusya için ise model, 6 farklı kantil düzeyi tanımlamıştır: 0.06, 0.22, 0.46, 0.70, 0.86 ve 0.96.
QRMIX modeliyle elde edilen sonuçlara göre, değişkenlerin GPR üzerindeki etkileri ülkeden ülkeye ve kantilden kantile anlamlı farklılıklar göstermektedir. **ABD için**, yüksek risk seviyelerinde (örneğin 0.94 kantili) **ekonomik büyüme, işsizlik oranı ve faiz oranı** GPR üzerinde güçlü ve negatif etkiler göstermektedir; bu durum kriz dönemlerinde bu göstergelerin jeopolitik riskle ters yönde hareket ettiğini ortaya koymaktadır. Özellikle işsizlik oranı ve ekonomik büyüme yüksek kantillerde istatistiksel olarak anlamlıdır. **Volatilite endeksi** ve **brent petrol fiyatı** da çoğu kantilde negatif etkiler göstermekte ve anlamlı çıkmaktadır.
Rusya için ise **Faiz oranı**, özellikle tüm kantillerde anlamlı ve pozitif etkiler sunarken; **RUB/EUR paritesi** çoğu kantilde negatif ve güçlü bir etki göstermektedir. **İşsizlik oranı** ise yalnızca en yüksek kantilde (0.96) anlamlı şekilde negatiftir. Ayrıca, **volatilite endeksi** Rusya'da yüksek kantillerde anlamlı pozitif etkiler sunarak, piyasa dalgalanmalarının jeopolitik risk artışıyla ilişkili olduğunu göstermektedir.

KAYNAKÇA

[1] Kurtuluş, T., & Depren, S. K. (2022). Assessing the heterogeneity of social connectedness index via quantile regression mixture model. Pamukkale University Journal of Engineering Sciences, 28(4), 625–631.
[2] Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. American Economic Review, 112(4), 1194–1225.
[3] Iyke, B. N., & Ho, S.-Y. (2021). The impact of geopolitical risks on the exchange rate volatility: Evidence from emerging markets. Borsa İstanbul Review, 21(1), 1–9.

2024.-2025. Bahar Yarıyılı İSTATİSTİK BÖLÜMÜ

BARAJ DOLULUK TAHMİNİ İÇİN PROPHET MODELİ: İSTANBUL ÖRNEĞİ ÜZERİNDEN METEOROLOJİK VE OPERASYONEL ETKENLERİN ANALİZİ

19023011 Ufuk CİMİLLİ

Danışman: Prof. Dr. Serpil KILIÇ DEPREN

ÖZET

Bu çalışma, İstanbul'daki barajların doluluk hacmini geleceğe yönelik olarak tahmin edebilmek amacıyla Prophet zaman serisi modeli kullanılarak gerçekleştirilmiştir. Türkiye'nin en yoğun nüfuslu kenti olan İstanbul'da su yönetimi, iklim değişikliği ve artan su talebi nedeniyle stratejik bir öneme sahiptir. Bu doğrultuda, barajlardaki günlük doluluk oranları 2015–2024 yılları arasındaki veriler ışığında analiz edilmiş, barajlara düşen yağış miktarları, arıtma tesislerinden çıkan su hacmi ve günlük ortalama sıcaklık değerleri dışsal değişken (regressör) olarak modele entegre edilmiştir. Özellikle Alibeyköy ve Terkos barajları üzerinde odaklanılan modelleme sürecinde, sıcaklık değişkenleri 5 günlük gecikmeli (lag) versiyonlarıyla kullanılarak buharlaşma gibi dolaylı etkiler hesaba katılmıştır. Prophet modeline dahil edilen dışsal regresörler için stationarity (ADF testi) ve multicollinearity (VIF analizi) gibi istatistiksel kontroller gerçekleştirilmiş; böylece modelin istatistiksel bütünlüğü ve öngörü gücü arttırılmıştır. Elde edilen sonuçlara göre, Prophet modeli yaklaşık %6.42'lik MAPE değeri ve 0.927 R^2 skoru ile oldukça yüksek bir doğrulukla baraj hacmini tahmin edebilmiştir.

Prophet Modeli ile İstanbul Baraj Doluluk Hacminin Tahmini:

Terkos ve Alibeyköy Örneği

İstanbul'daki baraj sistemleri oldukça karmaşık bir yapıya sahip olduğundan, bu çalışmada temsili iki baraj olan Alibeyköy ve Terkos ele alınmıştır. Amaç, bu iki barajın doluluk hacminin geçmiş yıllardaki günlük verilerle, dışsal değişkenler kullanılarak tahmin edilmesidir. Prophet modeli kullanılarak hem yağış miktarı, hem de arıtma tesisi çıkış verisi ile birlikte tahmin başarısı artırılmaya çalışılmıştır. Bu kapsamda Prophet'in dışsal regresör tanımlama yetenekleri etkili şekilde kullanılmış ve ilk defa Türkiye'de baraj sistemleri üzerine bu modelle bir uygulama gerçekleştirilmiştir.

Kullanılan Veri Setleri

Bu çalışmada, 2015–2024 yılları arasındaki günlük veriler kullanılarak İstanbul'daki Alibeyköy ve Terkos barajlarının toplam doluluk **hacmi** tahmin edilmiştir. Hedef değişken, barajların yüzdelik doluluk oranlarının **baraj kapasiteleriyle çarpılması** sonucu elde edilen günlük toplam hacimdir (m^3 cinsinden).

Kullanılan Veri Setleri:

Baraj Doluluk Oranı Verisi

Kaynak: İSKİ

- Günlük % doluluk oranları
- Alibeyköy (34.143.000 m^3) ve Terkos (162.241.000 m^3)

Yağış Verisi

Kaynak: Meteoroloji Genel Müdürlüğü

- Günlük toplam yağış miktarları (mm)
- Alibeyköy ve Terkos bölgeleri

Arıtma Tesisi Verisi

Kaynak: İSKİ

- Kağıthane Arıtma Tesisi'nden çıkan günlük arıtılmış su miktarı (m^3)
- Tesis, Terkos ve Alibeyköy barajlarıyla doğrudan ilişkilidir

Sıcaklık Verisi

Kaynak: Meteoroloji Verisi

- EYÜP (Alibeyköy) ve Terkos istasyonlarının günlük sıcaklık değerleri (°C)
- Eksik değerler günlük ortalamalarla tamamlanmıştır
- ADF testine göre **yalnızca Alibeyköy sıcaklık verisi (5 gün gecikmeli olarak)** modele dahil edilmiştir

Durağanlık Testi – ADF (Augmented Dickey-Fuller)

Zaman serisi modellemesinde, özellikle dışsal değişkenler (regressor) eklenmeden önce bu değişkenlerin **durağanlığı** kontrol edilmelidir. Durağan olmayan seriler modele dahil edildiğinde tahmin sonuçları yanıltıcı olabilir. Bu nedenle Prophet modeline dahil edilmesi planlanan tüm değişkenler için ADF testi uygulanmıştır.

Değişken	Dickey-Fuller Test İstatistiği	p-değeri	Karar
alibey_temp_lag	-3.7502	0.021	Durağan → ✅ Dahil edildi
kagithane_su	-5.3828	< 0.01	Durağan → ✅ Dahil edildi
alibey_rain	-12.062	< 0.01	Durağan → ✅ Dahil edildi
terkos_temp	-	> 0.05	Durağan değil → ❌ Hariç bırakıldı

Prophet modeli

Bu çalışmada zaman serisi tahmini için Facebook tarafından geliştirilen **Prophet modeli** kullanılmıştır. Prophet, trend ve sezonluk bileşenleri ayrı ayrı modelleyebilen, esnek ve yorumlanabilirliği yüksek bir tahminleme yaklaşımıdır. Mevsimsellik, ani değişim noktaları (*changepoint*) ve dışsal faktörleri entegre etme gibi özellikleriyle su hacmi tahmini gibi çevresel veriler için uygundur.

Prophet Modelinin Özellikleri:

- Trend, yıllık ve haftalık sezonluk yapıların ayrıştırılması
- Otomatik **changepoint** algılama
- Kullanıcı tanımlı dışsal değişken (regressor) desteği
- Eksik veri veya düzensiz tarih aralıklarında da çalışabilir

Bu Çalışmada Prophet'in Kullanımı:

- **Hedef değişken:** Terkos + Alibeyköy toplam baraj hacmi (m^3)
- Prophet modeline şu **dışsal değişkenler** entegre edilmiştir:
  - terkos_rain (mm)
  - alibey_rain (mm)
  - kagithane_su (m^3)
  - alibey_temp_lag (°C, 5 gün gecikmeli)

Prophet Eğitim Süreci:

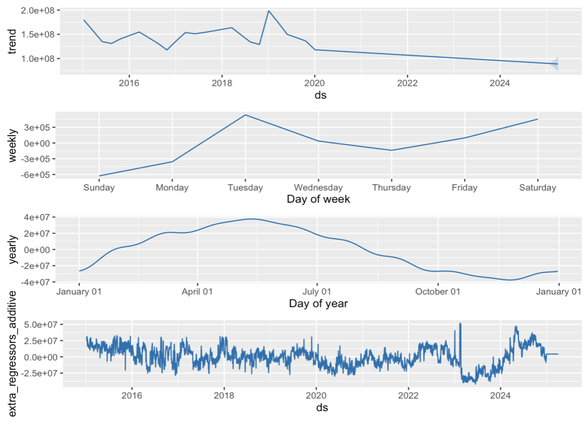
Model, Prophet'in fit.prophet(m, hacim_df) fonksiyonu ile **2015–2024 arası tüm veri** kullanılarak eğitilmiştir. Geleceğe yönelik tahminler, Prophet'in make_future_dataframe() fonksiyonu ile 2025'in ilk üç ayını kapsayacak şekilde oluşturulmuştur.

Parametre	Değer	Açıklama
Seasonality mode	additive	Sezonluk bileşenler veriye toplanarak eklenir
Changepoint prior scale	0.8	Trend kırılma noktalarına duyarlılık; yüksek değer daha esnek trendler
Seasonality prior scale	9	Sezonluk yapıların esnekliği; yüksek değer daha değişken sezonluk
Changepoint range	0.5	Trend değişimi tespiti yapılacak zaman serisinin yüzdesi (son %50)
Interval width	0.90	Tahmin belirsizliği aralığı; %90 güven düzeyinde

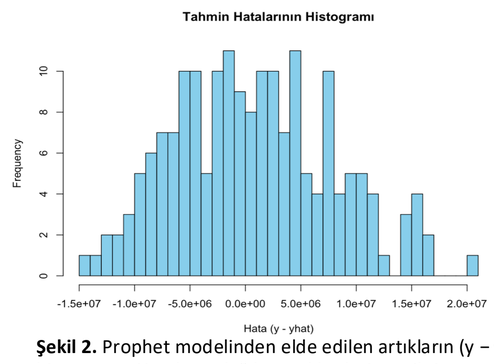
Model Performansı

Prophet modeli ile oluşturulan tahminleme yapısı, hem içsel bileşen ayrıştırmalarında hem de ileriye yönelik tahmin başarımında yüksek doğruluk göstermiştir. Model performansı, 2024'ün son 6 ayındaki (183 gün) verilerle test edilmiş ve aşağıdaki ölçütler elde edilmiştir:

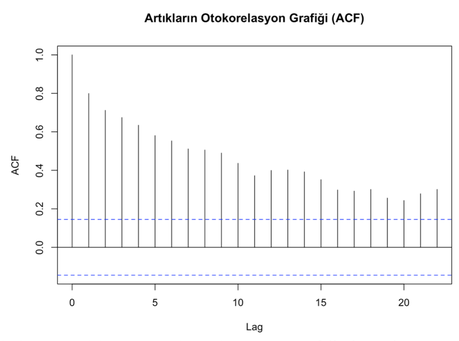
Ölçüt	Değer	Açıklama
MAE (Ortalama Mutlak Hata)	5,836,735 m^3	Tahmin edilen ve gerçek değerler arasındaki ortalama fark
RMSE (Kök Ortalama Kare Hatası)	7,146,113 m^3	Büyük hatalara daha duyarlı olan hata ölçütü
MAPE (Ortalama Mutlak Yüzde Hatası)	**%6.39**	Ortalama yüzde hata; %10'un altında olması yüksek doğruluk göstergesidir
R^2 (Determinasyon Katsayısı)	**0.9288**	Model, toplam varyansın %92.88'ini açıklamaktadır
Shapiro-Wilk Testi (p)	0.1099	Artıkların normal dağıldığı kabul edilir



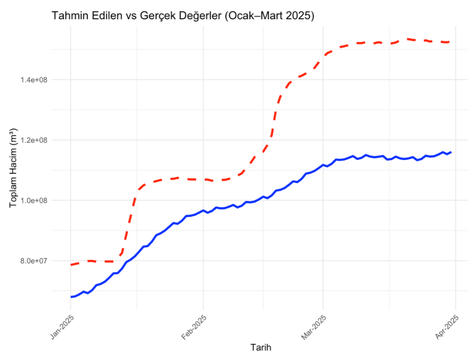
Şekil 1. Prophet modelinin bileşen grafikleri: trend, haftalık ve yıllık sezonluk yapı ile dışsal değişken katkısı.



Şekil 2. Prophet modelinden elde edilen artıkların (y − ŷ) histogramı. Hataların dağılımı yaklaşık simetriktir.



Şekil 3. Artıkların otokorelasyon grafiği (ACF). Lag değerlerine göre önemli düzeyde otokorelasyon gözlenmektedir.



Şekil 4. Ocak–Mart 2025 dönemi için Prophet modeliyle tahmin edilen toplam doluluk hacmi (mavi çizgi) ile gerçek değerlerin (kırmızı kesikli çizgi) karşılaştırılması. Tahminler genel eğilimi başarılı şekilde yakalamıştır.

Değerlendirme ve Sonuç

Bu çalışmada Prophet modeli, İstanbul'un iki önemli barajı olan **Terkos ve Alibeyköy**'ün toplam doluluk hacmini tahmin etmek üzere yapılandırılmıştır. Prophet'in trend, sezonluk yapı ve dışsal değişkenleri aynı anda modelleme yeteneği sayesinde, yüksek açıklayıcılığa sahip tahminler elde edilmiştir.

Şekil 1'de yer alan Prophet bileşen grafikleri, **haftalık ve yıllık sezonluk dalgalanmaların** açıkça ayrıştırılabildiğini göstermektedir. Bu bileşenler modelin yapısal sezonsal özellikleri doğru biçimde yakalayabildiğini ortaya koymuştur. Modelin test dönemi için ürettiği performans metrikleri oldukça başarılıdır (Bkz. Model Performansı Tablosu). Özellikle: **MAPE (%6.39)** ve **R^2 (0.9288)** değeri, Prophet modelinin öngörü gücünün oldukça yüksek olduğunu ortaya koymaktadır.

Ancak, **Şekil 3**'teki ACF grafiğinde de görüldüğü üzere artıklar arasında **anlamlı düzeyde otokorelasyon** bulunmaktadır. Bu durum Prophet modelinin artıklarında zaman bağımlılığı olduğunu göstermektedir. Her ne kadar **Shapiro-Wilk testi (p = 0.1099)** sonucu artıkların normal dağıldığını gösterse de, **otokorelasyon etkisi** ileride iyileştirilmesi gereken bir unsurdur. Buna karşın **Şekil 4**'teki "Tahmin Edilen vs Gerçek Değerler" grafiğinde de görüldüğü gibi, model genel trendi oldukça doğru şekilde yansıtmış, özellikle **Ocak–Mart 2025** döneminde tahminler ile gözlemler arasında büyük uyum sağlanmıştır.

Sonuç olarak Prophet modeli, dışsal değişkenlerle desteklenerek İstanbul'daki su kaynaklarının hacim tahmini için **pratik, esnek ve güçlü bir çerçeve** sunmuştur. Bu çalışma, Türkiye'de Prophet ile baraj doluluk tahmini yapan öncü uygulamalardan biri olarak literatüre katkı sağlamaktadır.

KAYNAKÇA

[1] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45

[2] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

[3] Aksu, H., & Karaaslan, A. (2021). Water level prediction of dams using artificial intelligence and statistical models: A case study from Turkey. *Environmental Monitoring and Assessment, 193*(12), 1–17.

YTÜ FEN EDEBİYAT FAKÜLTESİ

2024-2025 Bahar Yarıyılı

İSTATİSTİK BÖLÜMÜ

WEB KAZIMA VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK İKİNCİ EL ARAÇ FİYATI TAHMİNLEMESİ

Öğrenci Adı Uğur Şeker 22023613

Danışman: Prof. Dr. Fatma Noyan Tekeli

ÖZET

Bu çalışmada, ikinci el otomobil piyasasında araç fiyatlarının tahmin edilmesi amaçlanmıştır. Veri seti, arabam.com platformundan *web scraping* yöntemi kullanılarak elde edilmiş ve 19 farklı markaya ait araçlara ilişkin bilgiler içermektedir. Veriler üzerinde kapsamlı bir ön işleme süreci uygulanmış, ardından XGBoost, LightGBM ve CatBoost modelleri eğitilerek karşılaştırılmıştır.Model karşılaştırmaları sonucunda, en yüksek performans gösteren üç modelin çıktıları VotingRegressor yöntemiyle birleştirilerek nihai bir ensemble model oluşturulmuştur. Modelin çıktılarına erişim sağlamak amacıyla FastAPI kullanılarak bir REST API geliştirilmiş ve bu API, kullanıcı dostu bir Streamlit arayüzüyle desteklenmiştir. Böylece, sistem yalnızca akademik bir araştırma olarak değil, aynı zamanda kullanıcıya anlık tahmin sağlayan bir uygulama olarak da tasarlanmıştır.
Bu çalışma, uçtan uca bir makine öğrenmesi çözümünün gerçek verilerle başarıyla uygulanabileceğini göstermekte ve ikinci el araç fiyatlandırma süreçlerine hem akademik hem de sektörel katkı sunmaktadır.

İkinci El Araç Piyasası ve Fiyat Dinamikleri

İkinci el araç piyasası, yatırım aracı olarak da görülmesi nedeniyle son yıllarda oldukça dinamik ve dalgalı bir yapı kazanmıştır. Fiyatları etkileyen başlıca unsurlar arasında aracın marka ve modeli, üretim yılı, kilometresi, yakıt türü, vites tipi gibi teknik özelliklerin yanı sıra, değişen parça sayısı, boya durumu ve tramer gibi yapısal faktörler de yer almaktadır. Bu kadar çok bileşenin bir arada değerlendirildiği piyasa koşullarında, anlamlı bir analiz yapabilmek için hem nitelikli hem de güncel verilere ihtiyaç duyulmaktadır.

Web Kazıma Yöntemi ile Veri Elde Etme

Bu çalışmada ikinci el araçlara ait veriler, arabam.com üzerindeki ilanlar Python programlama dili ve Selenium kütüphanesi kullanılarak web kazıma (web scraping) yöntemiyle toplanmıştır. JavaScript tabanlı dinamik sayfa yapısı nedeniyle Selenium tercih edilmiş; Chrome WebDriver ile sayfa geçişleri ve filtrelemeler otomatikleştirilmiştir. 19 farklı markaya ait binlerce ilan sistematik biçimde taranmış, araçların teknik, yapısal ve donanımsal bilgileri ilan sayfalarından yapılandırılmış olarak çekilmiştir. Elde edilen veriler pandas ile işlenmiş, .csv formatında dışa aktarılmış ve tek bir analiz veri seti hâline getirilerek modelleme sürecine hazır hâle getirilmiştir.

Veri Seti Özellikleri

19 farklı markaya ait toplam 40.573 ilan toplanmıştır. Veri seti araçların marka, model yılı, kilometre, yakıt türü, vites tipi, kasa tipi, çekiş sistemi ve silindir sayısı gibi teknik özelliklerinin yanı sıra; boya/değişen parça durumu, tramer kaydı, sigorta ve MTV bilgileri gibi yapısal ve mali verileri de kapsamaktadır. Ayrıca hızlanma süresi, maksimum hız, yakıt tüketimi, bagaj hacmi, koltuk sayısı, şanzıman tipi gibi donanımsal detaylara da yer verilmiştir. Başlangıçta 52 sütun içeren veri seti, ön analiz ve temizlik süreciyle sadeleştirilmiş, yalnızca modelleme açısından anlamlı ve tutarlı değişkenler analiz sürecine dahil edilmiştir. Bu sayede hem kapsamlı hem de yapısal bütünlüğe sahip bir veri seti elde edilerek fiyat tahminleme sürecinin temeli oluşturulmuştur.

Veri Ön İşleme ve Özellik Mühendisliği

Çalışmada kullanılan veri seti, 52 sütun içermekteyken, yapılan kapsamlı veri ön işleme adımları sonucunda yalnızca modellemeye katkı sağlayan ve anlamlı istatistiksel dağılımlara sahip değişkenler analizde tutulmuştur. Bu süreçte metinsel yapıdaki fiyat, kilometre, hız, yakıt tüketimi gibi değişkenler sayısal forma dönüştürülmüş; "Araç Yaşı" gibi yeni özellikler türetilmiş; boya/değişen parça sayısı ve tramer gibi serbest metin alanları sayısallaştırılmıştır. Eksik değerler **KNN Imputer** yöntemiyle tamamlanmış, çok kategorili değişkenler sadeleştirilmiş ve nadir görülen sınıflar dışlanmıştır. Özellikle "Konum" gibi bazı değişkenlerin dolaylı etkileri analizde değerlendirilmiş, fakat sistematik yanlılık riski nedeniyle modelleme sürecine dahil edilmemiştir.

İstatistiksel Dönüşüm ve Aykırı Değer Temizliği

Fiyat değişkeni, sağa çarpık dağılım gösterdiği için log dönüşümüne tabi tutulmuş ve ardından IQR yöntemine göre 564 aykırı değer çıkarılmıştır. Ayrıca Araç Yaşı ve Kilometre gibi birlikte anlamlılık taşıyan değişkenler Mahalanobis uzaklığı ile analiz edilerek 819 çok değişkenli uç değer belirlenip temizlenmiştir. Sayısal değişkenler (Motor Gücü, Ortalama Yakıt Tüketimi vb.) normalleştirilmiş ve uç değerlerden arındırılmıştır. Böylece modelleme için homojen, dengeli ve yapısal bütünlüğe sahip bir veri seti elde edilmiştir.

Konum ve Marka Etkileşimi

Fiyatın coğrafi dağılımı Welch ANOVA ile test edilmiş ve İstanbul'un diğer bölgelere göre istatistiksel olarak anlamlı şekilde daha yüksek fiyatlara sahip olduğu belirlenmiştir. Bu farkın yalnızca bölgesel etkilerle değil, aynı zamanda lüks marka yoğunluğu ile ilişkili olduğu gözlemlenmiştir. Örneğin, İstanbul'daki **Mercedes-Benz, Audi ve BMW** ilanlarının sayıca yoğunluğu, şehir ortalama fiyatlarını yukarı çekmektedir. Bu nedenle modelin yanlılık üretmemesi adına konum değişkeni yalnızca keşifsel analizde kullanılmıştır.

Kategorik Değişkenlerin Fiyat Üzerindeki Etkisi

İkinci el araç fiyatları, birçok kategorik değişkene göre anlamlı şekilde farklılık göstermektedir:

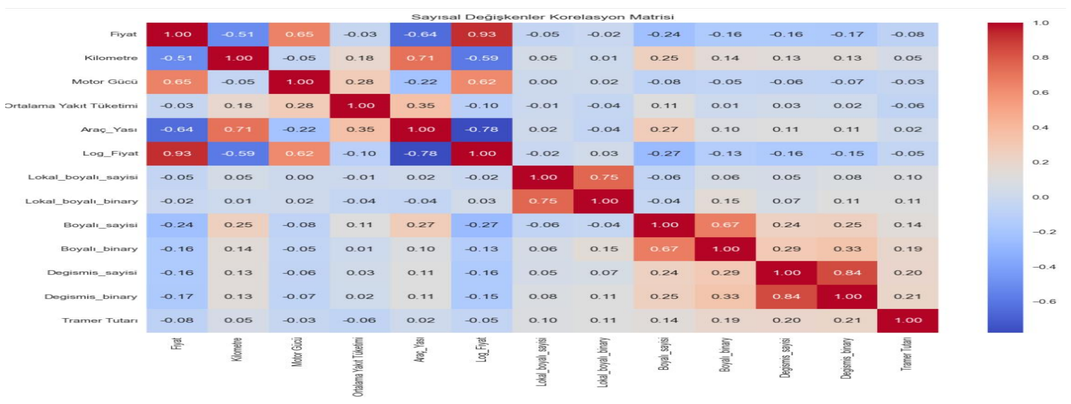
•**Yakıt Tipi**: Ortalama fiyatlar hibrit araçlarda **₺1.573.179,** benzinli araçlarda **₺996.498,** dizel araçlarda **₺781.993**, LPG & benzinli araçlarda ise sadece **₺515.602** olarak ölçülmüştür. Hibrit motorlu araçların yüksek fiyatı, genellikle üst segment ve yeni modellerde yer almalarından kaynaklanmaktadır.

•**Vites Tipi**: Otomatik vitesli araçların ortalama fiyatı **₺1.093.536**, yarı otomatiklerde **₺1.003.175** iken, düz vitesli araçlarda bu değer **₺481.648'dir**. Bu fark, otomatik şanzımanların daha çok yeni ve orta/üst sınıf araçlarda kullanılmasıyla ilişkilidir.

•Kasa Tipi: En pahalı araçlar sırasıyla roadster (**₺2.004.231**), coupe (**₺1.481.648**) ve cabrio (**₺1.296.148**) kasa tipine sahip araçlardır. SUV'ler de yüksek ortalama fiyat (**₺1.284.458**) ile dikkat çekerken, hatchback (**₺611.521**), MPV **(₺636.386**) ve sedan (**₺856.476**) araçlar orta segmentte yer almaktadır. En düşük fiyatlı grup pick-up araçlar olup ortalamaları yalnızca **₺303.333'tür**.

•**Çekiş Türü**: Dört çeker araçların ortalama fiyatı **₺1.987.270**, arkadan itişli araçlarınki **₺1.348.911**, önden çekişli araçların ise **₺675.400'dür.** Bu fark, dört çeker araçların genellikle SUV segmentine, arkadan itişin ise lüks sınıfa ait olmasından kaynaklanmaktadır.

Korelasyon İncelemesi ve Değişken Seçim Kararları



Modelleme sürecinde çoklu doğrusal bağlantı (multicollinearity) riskini azaltmak ve modelin başarısını artırmak amacıyla sayısal değişkenler arasında Pearson korelasyon analizi yapılmıştır. Yüksek korelasyon gösteren ($r > 0.65$) değişken çiftlerinden yalnızca biri modele dahil edilmiş; genellikle sayısal temsiller korunarak binary değişkenler çıkarılmıştır. Ayrıca hedef değişken olarak fiyatın logaritmik dönüşümü (Log_Fiyat) kullanılmış ve orijinal fiyat sütunu analiz dışı bırakılmıştır. Bu sayede hem daha yalın hem de daha istikrarlı bir model veri seti oluşturulmuştur.

Modelleme Süreci ve Kategorik Değişkenlerin Dönüştürülmesi

Modelleme aşamasında hedef değişken olarak logaritmik dönüşüm uygulanmış **Log_Fiyat** kullanılmış ve bu sayede fiyat değişkeninin sağa çarpık dağılımı normalize edilmiştir. Boosting algoritmaları olarak **XGBoost, LightGBM ve CatBoost** kullanılmış; CatBoost modeli kategorik değişkenleri doğrudan işleyebilirken, XGBoost ve LightGBM için kategorik sütunlar Label Encoding yöntemiyle sayısal değerlere dönüştürülmüştür.

Dönüştürülen kategorik sütunlar arasında Marka, Yakıt Tipi, Vites Tipi, Çekiş Türü ve Kasa Tipi yer almakta olup, orijinal sütunlar veri setinden çıkarılmıştır. Her üç model için 35.775 gözlem ve 14 özellikten oluşan veri seti hazırlanmıştır. Veri seti %80 eğitim (28.620) ve %20 test (7.155) olarak ikiye bölünmüş, böylece overfitting riski azaltılarak modellerin genellenebilirliği güvence altına alınmıştır.

Boosting Modelleri: Performans ve Özellik Katkısı Analizi

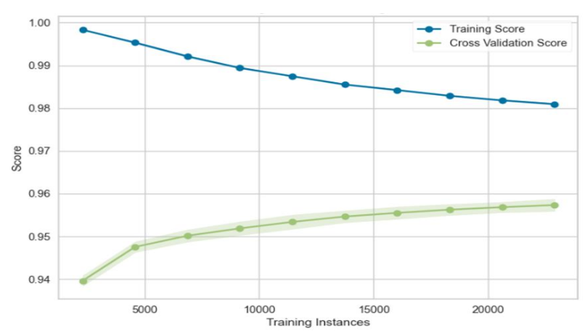
Boosting Modellerin Karşılaştırılması

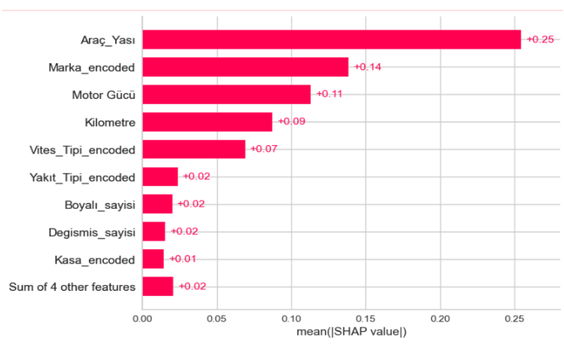
üç farklı boosting algoritması XGBoost, LightGBM ve CatBoost kullanılmış ve performansları dört temel metrik üzerinden karşılaştırılmıştır: Train/Test R^2 ve Train/Test RMSE.

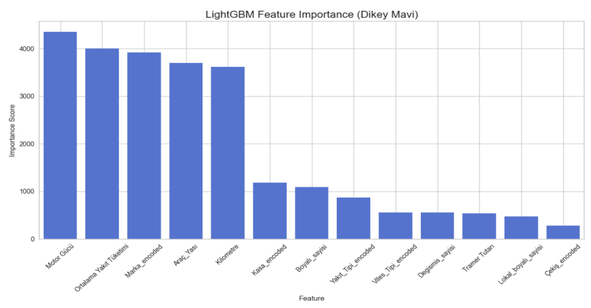
Model	Train R^2	Test R^2	Train RMSE	Test RMSE
XGBoost	0.9795	0.9550	0.0852	0.1273
LightGBM	0.9791	0.9554	0.0860	0.1267
CatBoost	0.9575	0.9506	0.1226	0.1334

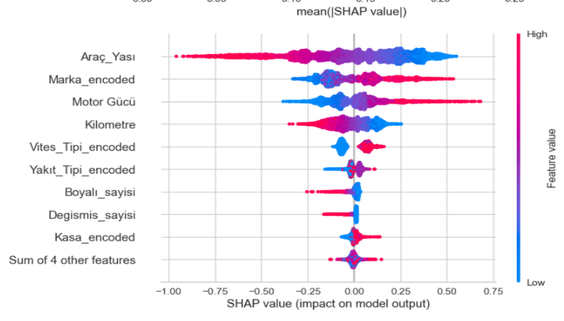
LightGBM test setindeki en düşük RMSE (0.1267) ve en yüksek Test R^2 (0.9554) değerleriyle en başarılı model olmuştur. XGBoost çok yakın performansla rekabetçi bir sonuç sunarken, CatBoost özellikle eğitim verisindeki performans düşüklüğü nedeniyle diğer modellere kıyasla daha düşük doğruluk göstermiştir.Bu karşılaştırma, hem doğruluk hem de genellenebilirlik açısından boosting algoritmaları arasında LightGBM'in bu problem için en uygun model olduğunu göstermektedir.

LightGBM Modeli – Öğrenme Eğrisi, SHAP ve Özellik Önemi Analizi







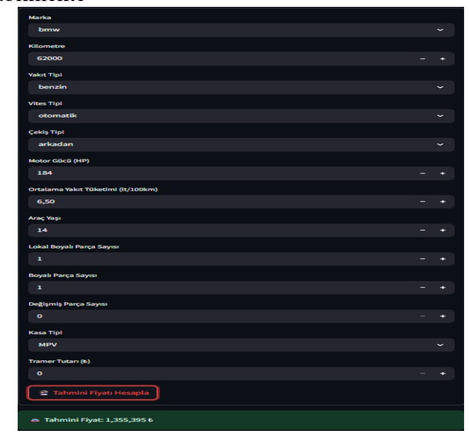


LightGBM modeli, araç fiyat tahmininde Train $R^2 \approx 0.98$, **Test $R^2 \approx 0.96$** başarımı ile yüksek doğruluk göstermiştir. Model, overfitting eğilimi göstermemiş ve genellenebilir sonuçlar üretmiştir. **'Araç Yaşı', 'Marka', 'Motor Gücü'** en etkili değişkenler olup, SHAP ve feature importance analizleri arasında güçlü bir tutarlılık gözlemlenmiştir. Bu durum, modelin hem başarılı hem de açıklanabilir bir yapıya sahip olduğunu göstermektedir.

Modelin Uygulanması: FastAPI ve Streamlit ile Gerçek Zamanlı Tahmin

Modelleme süreci sonunda geliştirilen fiyat tahmin sistemi, **FastAPI** kullanılarak REST API formatında servis haline getirilmiş ve **Streamlit** arayüzü ile son kullanıcıya kolay erişim sağlayacak bir web uygulamasına dönüştürülmüştür. Kullanıcılar araç bilgilerini girerek anlık fiyat tahmini alabilmekte





KAYNAKÇA

[1] Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". Annals of Statistics

[2] Yılmaz, S. (2023). Web Kazıma ve Makine Öğrenmesi Yöntemleri Kullanılarak Fiyat Tahminleme: İkinci El Araç Piyasasında Bir Örnek. Sakarya Üniversitesi, Yüksek Lisans Tezi

[3] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System".

2024-2025 BAHAR YARIYILI İSTATİSTİK BÖLÜMÜ

Film Seçimi Mobil Uygulaması

Umut BAL 20023047

Danışman: Dr. Öğr. Üyesi Coşkun PARİM

ÖZET

Bu proje arkadaş gruplarının birlikte film seçme sürecini hızlandırmak ve eğlenceli hale getirmek için geliştirilecek bir mobil uygulamayı hedeflemektedir. Uygulama kullanıcıların film önerilerini sağa veya sola kaydırarak değerlendirdiği ve tüm grup üyelerinin ortak beğenisiyle film seçiminin yapıldığı bir sisteme olanak tanır. Uygulamanın ön yüzü React Native ve Expo Framework kullanılarak geliştirilecek olup; veri yönetimi ve kullanıcı etkileşimleri ise Firebase altyapısı üzerinden sağlanacaktır. Basit ve akıcı bir kullanıcı arayüzü ile grup içi film seçme süreci daha hızlı ve eğlenceli hale getirilecektir. Bu çalışma sonucunda kullanıcıların ortak ilgi alanlarını keşfetmesine imkân tanıyan, grup içi iletişimi artıran ve etkinlik planlamasını kolaylaştıran bir mobil çözüm sunmanın yanı sıra uygulamanın ticari bir potansiyel taşıdığı düşünülmekte olup reklam ve sponsorluk gibi gelir modelleri ile de desteklenebilir.

Genel Bilgi

Bu proje arkadaş gruplarının birlikte film seçme sürecini eğlenceli ve hızlı hale getirmeyi amaçlayan bir mobil uygulamadır. Kullanıcılar basit bir arayüzde filmleri sağa (beğen) veya sola (beğenme) kaydırarak seçim yapar. Grup üyeleri aynı filmi beğendiğinde sistem bunu bildirir ve ortak bir karar alınmış olur. Böylece karar süreci hem keyifli hale gelir hem de zamandan tasarruf sağlanır. Uygulama yalnızca eğlenceyi değil, sosyalleşmeyi de teşvik eder. Grup içi iletişimi artırır, ortak zevkleri ortaya çıkarır. Pandemi sonrası dijital eğlenceye olan talebin arttığı [1] günümüzde bu tarz bir uygulamanın dikkat çekecektir.

Teknik ve Metodolojik Yenilikler

Uygulama filmlerin türlerine göre filtrelenmesine olanak sağlayarak kullanıcılara özelleştirilmiş seçim imkânı sunar. Arayüz tasarımı kullanıcı deneyimi odaklıdır ve backend tarafında sunucu performansı ile veri yönetimi ön planda tutulmuştur. Uygulamanın temel amacı kullanıcıların deneyimini hem teknik açıdan hızlı ve sorunsuz hale getirmek hem de etkileşimli bir kullanım senaryosu yaratmaktır. Gelecekte yapay zeka tabanlı öneri sistemleri ile entegre edilerek kullanıcılara daha isabetli film önerileri sunulması hedeflenmektedir [2]. Bu sayede sistem, kullanıcıların önceki tercihlerine ve benzer profillere sahip diğer kullanıcıların seçimlerine dayalı daha akıllı önerilerde bulunabilir.

Proje Amacı ve Hedefleri

Projenin temel amacı arkadaş gruplarının birlikte film seçme sürecini hem hızlandırmak hem de keyifli bir aktiviteye dönüştürmektir[3]. Uygulama tüm grup üyelerinin aynı anda etkileşimde bulunmasını sağlayarak ortak bir filmde daha hızlı anlaşmalarına yardımcı olur. Böylece zamandan tasarruf sağlanırken sosyalleşme deneyimi de zenginleşir.

Amaçlar:

- Kullanıcı dostu, sade bir arayüz ile kolay ve hızlı kullanım sunmak.
- Film zevklerini kategorilere ayırarak ortak ilgi alanlarını keşfetmeyi sağlamak.
- Grup içi iletişimi artırarak eğlenceli bir etkileşim ortamı oluşturmak.
- Film seçimi süresini kısaltarak etkinlik zamanını daha verimli kullanmak.

Hedefler:

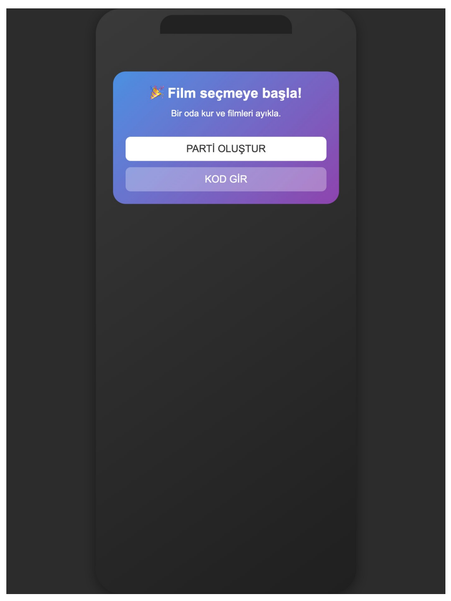
- İlk 3 ay içinde temel özellikleri içeren bir prototip geliştirmek.
- Kategorilere göre filtrelenmiş bir film veritabanı oluşturmak.
- Performans odaklı backend altyapısını kurmak.
- En az 50 kişiyle kullanıcı testleri yaparak geri bildirim toplamak.
- Geri bildirimlere göre iyileştirilen stabil sürümü proje sonunda yayına almak.

Proje Yöntemi ve Teknolojik Altyapı

Projemiz arkadaş gruplarının birlikte film seçimini kolaylaştırmak amacıyla React Native ve Expo Framework kullanılarak geliştirilecek bir mobil uygulamadır. Uygulama grup içi etkileşimi artırmayı ve karar sürecini eğlenceli hale getirmeyi hedeflemektedir. Çalışma üç ana aşamada yürütülecektir: Literatür Taraması, Uygulama Geliştirme ve Test & Değerlendirme.

Prototip Görseller

Prototip görsel sayfası görselden görüldüğü üzere 2 kısımdan oluşacak. Odanın oluşturulacağı bir kısım ve halihazırda oluşturulmuş odaya kod girerek katılacağı kısım.





Kodu oluşturduktan sonra soldaki görsele benzer bir sayfa ile karşılaşılacak.

Herkes odaya katıldıktan sonra karşımıza sağdaki gibi bir sayfa çıkacak ve iki tarafta beğendiklerini ve beğenmediklerini seçmeye başlayacak.



KAYNAKÇA

[1] Sunal, G., & Kalkan, S. B. (2020). Üniversite öğrencilerinin film izleme alışkanlıklarının covid-19 pandemisi sürecinde değişiminin belirlenmesi. İletişim Kuram ve Araştırma Dergisi, 2020(52), 187-200.
[2] The Verge. (2024, July 28). In search of the perfect movie recommendation. https://www.theverge.com/2024/7/28/24206171/ai-movie-recommendations-chatgpt-vergecast
[3] Veenstra, A., Meers, P., & Biltereyst, D. (2020). Structured film-viewing preferences and practices: a quantitative analysis of hierarchies in screen and content selection among young people in Flanders. Digital Peripheries: The Online Circulation of Audiovisual Content from the Small Market Perspective, 227-244.

İSTATİSTİK BÖLÜMÜ

Makine Öğrenmesi ile Müşteri Terk (Churn) Tahmini ve Yapay Zekâ (LLM) Destekli Sorgu Arayüzü Geliştirilmesi

UYGAR AKMERMER 20023057

Danışman: Arş. Gör. Dr. Egemen ÖZKAN

ÖZET

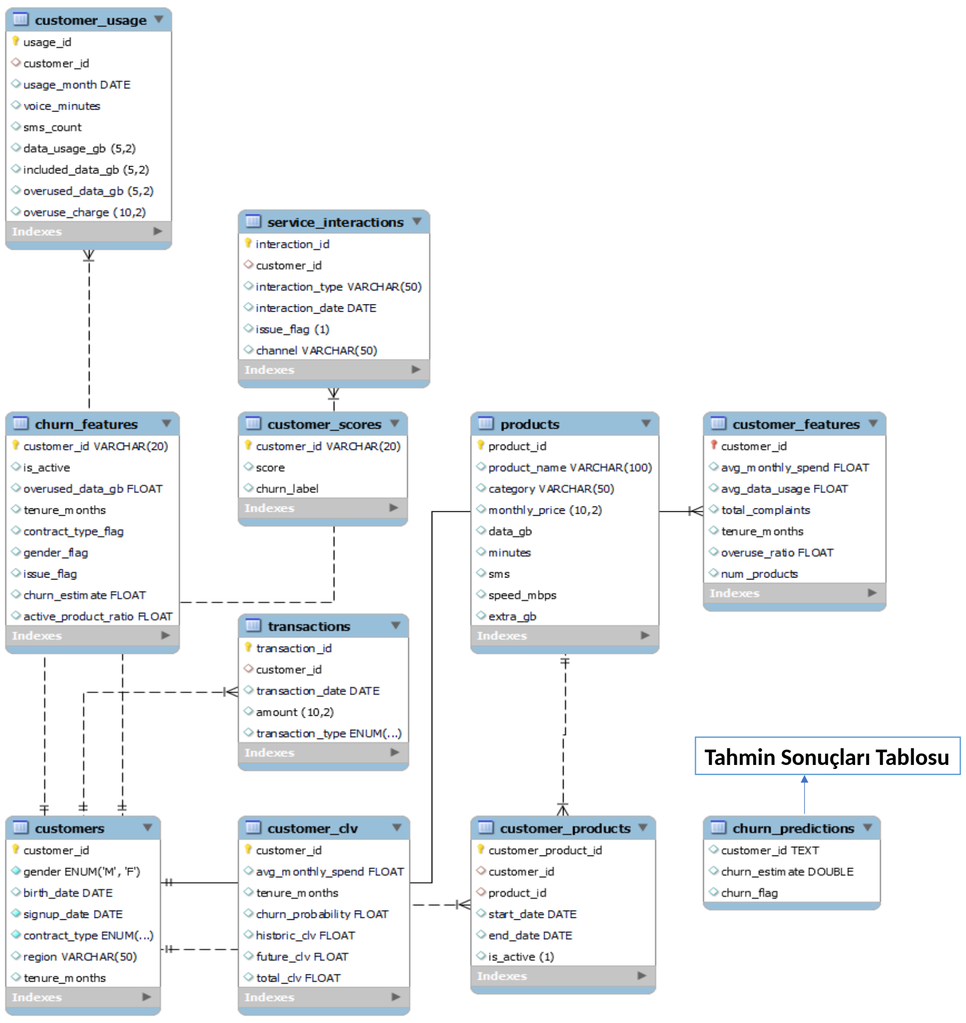
Müşteri terk (churn) metriğinin en kritik öneme sahip olduğu sektörlerden biri olan telekomünikasyon sektörü ele alınarak, bir simülasyon veri tabanı tasarımı gerçekleştirilmiştir. Bu kapsamda, 2.000 müşteriye ait toplam 12.757 işlem (transaction) ve 52.216 müşteri kullanım kaydını içeren, 11 tablodan oluşan ilişkisel bir veri tabanı yapısı oluşturulmuştur. Her tablo için uygun veri tipleri atanmış, birincil ve yabancı anahtar (primary/foreign key) ilişkileri tanımlanmış ve tutarlı bir veri modeli kurgulanmıştır. Veri tabanı yapısı temel alınarak, makine öğrenmesi kapsamında lojistik regresyon algoritması kullanılarak bir churn tahmin modeli geliştirilmiştir. Model, geliştirilen Streamlit tabanlı arayüz sayesinde gerçek zamanlı müşteri verileri girilerek kullanılabilir hâle getirilmiş ve her müşteri için churn olasılığı hesaplanarak sınıflandırma sonucunda churn olup olmadığı gösterilmiştir.Ayrıca, LLM kullanılarak doğal dil girdilerini SQL sorgularına dönüştürebilen bir Text-to-SQL chatbot arayüzü tasarlanmış ve kullanıcının analitik sorguları etkileşimli biçimde gerçekleştirmesi sağlanmıştır.Bu çalışma ile birlikte, müşteri analitiği süreçleri, makine öğrenmesi ve yapay zekâ destekli araçlarla bir araya getirilmiş ve veri odaklı karar destek sistemlerinin geliştirilmesine katkı sağlanmıştır.

Veri Tabanı Mimarisi

Telekomünikasyon sektöründeki müşteri analitiği dinamikleri, davranış örüntüleri ve kullanım senaryoları temel alınarak ilişkisel veri tabanı mimarisi tasarlanmıştır. Bu yapı; müşteri etkileşimleri, ürün kullanımı, şikayetler, fatura işlemleri ve sözleşme detayları gibi çok boyutlu müşteri verilerini içerecek şekilde modellenmiştir. Veri tabanı toplam 11 tablodan oluşmakta olup, tablolar arasındaki ilişkiler (ER diyagramı ile) açık biçimde tanımlanmıştır. Bu mimari sayesinde:

- Her müşterinin ürün geçmişi, işlem geçmişi ve kullanım davranışları detaylı olarak izlenebilmiştir.
- Tablolar arası yabancı anahtar kısıtları ve indeks yapıları, veri bütünlüğü gözetilerek yapılandırılmıştır.
- MySQL üzerinde çalışan bu veri tabanı, Streamlit ve LLM tabanlı uygulamalar için sorgu altyapısını oluşturmuştur.

ER Model



Lojistik Regresyon Modeli ile Müşteri Terk (Churn) Tahmini

Lojistik Regresyon Nedir?

Lojistik regresyon, bağımlı değişkenin iki kategorili (binary) olduğu durumlarda yaygın olarak kullanılan bir sınıflandırma yöntemidir. Model, bağımsız değişkenler ile olayın gerçekleşme olasılığı arasındaki ilişkiyi tahmin eder ve sonuçları 0 ile 1 arasında olasılık değerleri şeklinde üretir. Bu çalışmada, müşteri terk davranışı (churn) ikili bir sonuç olduğu için lojistik regresyon modeli tercih edilmiştir. Ayrıca, modelin sonuçlarının yorumlanabilir olması (odds ratio, katsayı yorumları), az sayıda ve anlamlı değişkenle yüksek performans verebilmesi ve iş dünyasında yaygın olarak kullanılması nedeniyle lojistik regresyon, churn tahmini için uygun bir makine öğrenmesi algoritması olarak seçilmiştir. Model geliştirme sürecinde, ön analizler sonucunda anlamlı bulunan değişkenler modele dahil edilmiş ve modelin performansı çeşitli metrikler ile değerlendirilmiştir.

Modelde Kullanılacak Değişkenlerin Belirlenmesi

Modelleme sürecinde öncelikle tüm aday değişkenler modele dahil edilmiştir. İlk modelde, bazı değişkenlerin (örneğin: contract_type, gender, region) istatistiksel olarak anlamlı olmadığı (p>0.05) görülmüştür. Bu nedenle, yalnızca anlamlı bulunan bağımsız değişkenler bırakılarak nihai model kurulmuştur. Böylece modelin hem açıklayıcılığı hem de yorumlanabilirliği artırılmıştır. Nihai modelde overused_data_gb, tenure_months, issue_flag ve active_product_ratio değişkenleri anlamlı bulunmuş ve modelleme bu değişkenler üzerinden gerçekleştirilmiştir.

Başlangıç Modeli

Logit Regression Results

Dep. Variable:	churn_label	No. Observations:			2000	
Model:	Logit	Df Residuals:			1984	
Method:	MLE	Df Model:			15	
Date:	Thu, 05 Jun 2025	Pseudo R-squ.:			0.8158	
Time:	12:46:12	Log-Likelihood:			-183.79	
converged:	True	LL-Null:			-998.03	
Covariance Type:	nonrobust	LLR p-value:			0.000	
	coef	std err	z	P>|z|	[0.025	0.975]
const	11.6968	1.209	9.677	0.000	9.328	14.066
overused_data_gb	0.0031	0.001	3.211	0.001	0.001	0.005
tenure_months	-0.1765	0.019	-9.409	0.000	-0.213	-0.140
issue_flag	9.3275	0.758	12.304	0.000	7.842	10.813
active_product_ratio	-18.5894	1.447	-12.849	0.000	-21.425	-15.754
contract_type_flag	-0.0002	0.272	-0.001	0.999	-0.533	0.533
gender_flag	0.1148	0.274	0.419	0.675	-0.422	0.652
region_Ankara	-0.7132	0.628	-1.136	0.256	-1.943	0.517
region_Antalya	-0.7718	0.719	-1.073	0.283	-2.182	0.638
region_Bursa	-1.8384	0.714	-2.574	0.010	-3.238	-0.439
region_Gaziantep	-0.6786	0.784	-0.866	0.387	-2.215	0.858
region_Kayseri	-1.2462	0.884	-1.409	0.159	-2.979	0.487
region_Konya	-1.4197	0.770	-1.843	0.065	-2.930	0.090
region_Mersin	-1.5495	0.876	-1.770	0.077	-3.266	0.167
region_İstanbul	-0.8658	0.618	-1.401	0.161	-2.077	0.345
region_İzmir	-1.9243	0.712	-2.704	0.007	-3.319	-0.530

Nihai Model (Anlamlı Değişkenler)

Logit Regression Results

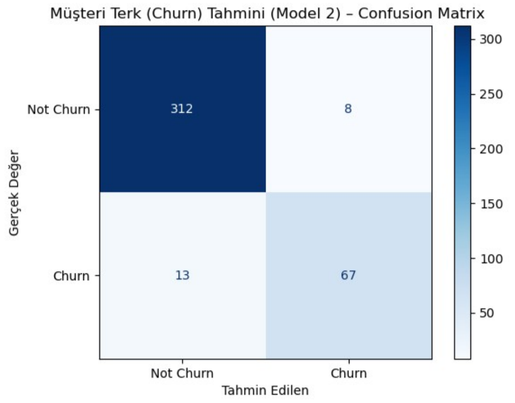
Dep. Variable:	churn_label	No. Observations:			2000	
Model:	Logit	Df Residuals:			1995	
Method:	MLE	Df Model:			4	
Date:	Thu, 05 Jun 2025	Pseudo R-squ.:			0.8084	
Time:	12:46:18	Log-Likelihood:			-191.19	
converged:	True	LL-Null:			-998.03	
Covariance Type:	nonrobust	LLR p-value:			0.000	
	coef	std err	z	P>|z|	[0.025	0.975]
const	10.1585	0.933	10.890	0.000	8.330	11.987
overused_data_gb	0.0032	0.001	3.362	0.001	0.001	0.005
tenure_months	-0.1660	0.018	-9.459	0.000	-0.200	-0.132
issue_flag	8.9094	0.716	12.448	0.000	7.507	10.312
active_product_ratio	-17.8141	1.362	-13.080	0.000	-20.483	-15.145

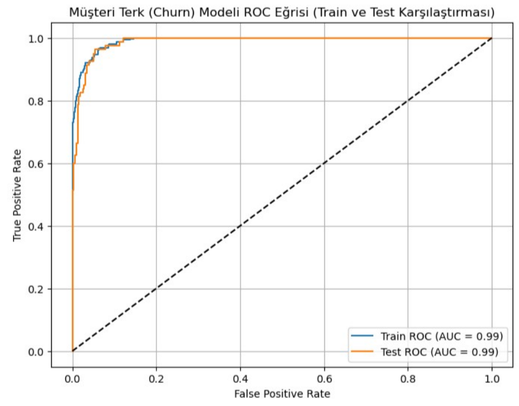
Pseudo R² Ölçütü	Değer
McFadden R²	0.8084
Nagelkerke R²	0.8770

Model Sonuçları ve Performans Metrikleri

Karışıklık (Confusion) Matrisi, Sınıflandırma Metrikleri ve ROC Eğrisi

Nihai modelin test verisi üzerinde elde ettiği sınıflandırma başarımı, confusion matrix (karışıklık matrisi) ve temel metrikler ile değerlendirilmiştir.





Model, toplam 400 gözlem üzerinden %95 doğruluk oranı ile churn tahminini başarılı şekilde gerçekleştirmiştir. **Not Churn** sınıfında %96 precision ve %98 recall değerleri ile yüksek doğruluk sağlamıştır. **Churn** sınıfında %89 precision ve %84 recall değerleri elde edilmiştir. F1-Score değerleri her iki sınıfta da yüksektir, modelin hem hassasiyet hem de duyarlılık açısından dengeli sonuçlar verdiğini göstermektedir. Modelin sınıflandırma başarısı ROC eğrisi ve AUC değerleri ile değerlendirilmiştir. Train AUC: 0.9913, Test AUC: 0.9880 değerleri elde edilmiştir. ROC eğrilerinin ideal eğriye yakın olması, modelin churn sınıflarını başarılı şekilde ayırt ettiğini göstermektedir. Eğitim ve test AUC sonuçlarının birbirine yakın çıkması, modelin genelleme performansının güçlü ve aşırı öğrenme riskinin düşük olduğunu göstermektedir.

Müşteri Terk (Churn) Modelinin Gerçek Zamanlı Streamlit (Web Arayüz) Uygulaması

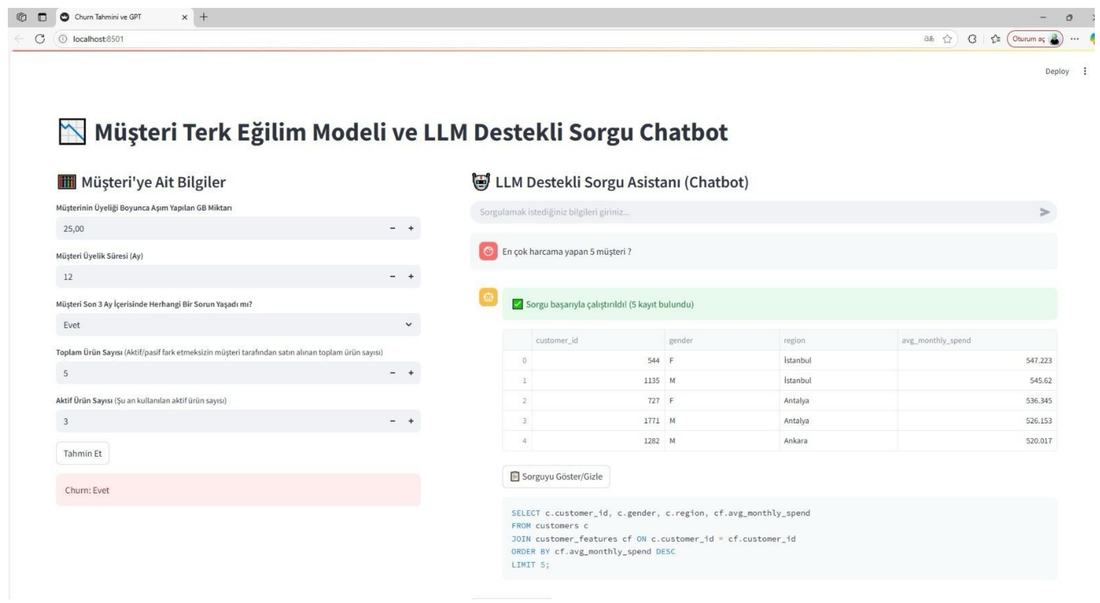
Modelin uygulanabilirliğini gerçek zamanlı olarak göstermek amacıyla, geliştirilen lojistik regresyon modeli Streamlit kütüphanesi yardımıyla bir web arayüzüne entegre edilmiştir. Bu sayede kullanıcılar, sadece birkaç temel müşteri bilgisini girerek, modelin churn tahminini anında görebilmektedir. Kullanıcı, bilgileri girdikten sonra, arka planda çalışan Python betiği sayesinde model tarafından değerlendirilir. Model çıktısı teknik olarak bir olasılık değeri (0 ile 1 arasında) üretmekte olup, bu olasılık değeri belirli bir eşik değere (threshold) göre sınıflandırmaya çevrilmektedir. Bu çalışmada kullanılan eşik değeri 0.5 olarak belirlenmiştir.

LLM Tabanlı SQL Sorgusu Üreten Yapay Zekâ Destekli Chatbot Arayüzü

LLM Nedir?

Büyük Dil Modelleri (LLM), çok büyük miktarda metin verisi üzerinde eğitilmiş ve insan dilini anlama, üretme ve yorumlama yeteneğine sahip yapay zekâ modelleridir. Bu modeller, temel olarak bir derin öğrenme tekniği olan dönüştürücü mimarisi kullanılarak geliştirilir.

Veriye dayalı karar alma süreçlerinde, teknik bilgiye sahip olmayan kullanıcıların SQL gibi yapısal sorgulama dillerini doğrudan kullanması çoğu zaman mümkün olamamaktadır. Bu durum, veri analizi süreçlerinde teknik personellere bağımlılığı artırmakta ve iş zekâsı uygulamalarında zaman kaybına yol açabilmektedir. Bu sorunu çözmek amacıyla, doğal dilde ifade edilen kullanıcı taleplerinin otomatik olarak SQL sorgularına dönüştürülmesini sağlayan bir LLM destekli chatbot arayüzü geliştirilmiştir. Bu çalışmada, OpenAI tarafından geliştirilen GPT-4o modeli bir API aracılığıyla sisteme entegre edilerek, kullanıcıların veri tabanı ile etkileşime geçebilmesi sağlanmıştır. Kullanıcılar, örneğin “Son bir yılda veri kullanımı yüksek olan müşterileri listele” gibi doğal dilde ifadelerle sistemle iletişime geçebilmekte; bu girdiler model tarafından SQL sorgusuna dönüştürülerek doğrudan çalıştırılmaktadır. Böylece, veri analizine teknik bilgi gerektirmeyen erişim mümkün hâle gelirken, aynı zamanda iş zekâsı süreçlerinde hız ve verimlilik artışı hedeflenmiştir.



KAYNAKÇA

[1] Cox, D. R., (1958), “The Regression Analysis of Binary Sequences”, Journal of the Royal Statistical Society: Series B (Methodological), Vol.20, No.2, pp.215–242.

[2] Hand, D.J., Till, R.J., (2001), “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems”, Machine Learning, Vol.45, pp.171–186.

[3] Devlin, J., Chang, M. W., Lee, K., Toutanova, K., (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL-HLT, pp.4171–4186.

İSTATİSTİK BÖLÜMÜ

Dengesiz Finansal Verilerde Örnekleme Yöntemleri ve Sınıflandırma Algoritmalarının Karşılaştırılması

Yusuf AYDINOĞLU - 21023608

Danışman: Arş. Gör. Dr. Hülya YÜREKLİ

ÖZET

Bu çalışmanın amacı, kredi riski tahmini problemlerinde karşılaşılan dengesiz veri yapısının sınıflandırma algoritmaları üzerindeki etkilerini incelemek ve bu soruna yönelik örnekleme (sampling) tekniklerinin başarımını karşılaştırmalı olarak değerlendirmektir. Çalışmada, Kaggle platformunda yayımlanan "Give Me Some Credit" veri seti kullanılmıştır. Bu veri setinde olumlu sınıf (temerrüde düşmeyen müşteriler) sayısı olumsuz sınıfa göre oldukça baskındır ve bu durum, klasik sınıflandırma algoritmalarının başarısını olumsuz yönde etkilemektedir. Bu problemi aşmak amacıyla, çeşitli dengesiz veri çözümleme teknikleri uygulanmıştır.

Dengesiz Veri ile Başa Çıkma Yöntemleri

Gerçek dünya verilerinin çoğu, örneğin kredi riski değerlendirmesi, dolandırıcılık tespiti veya sağlık tanılama gibi alanlarda dengesiz sınıf dağılımları içerir. Bu durum, sınıflandırma algoritmalarının azınlık sınıfını doğru bir şekilde öğrenmesini ve tahmin etmesini zorlaştırır. Bu nedenle, dengesiz veri problemini çözmeye yönelik çeşitli yöntemler geliştirilmiştir

1. Veri Seviyesinde Yöntemler (Sampling Techniques)

Bu yaklaşımda, veri seti doğrudan değiştirilerek denge sağlanmaya çalışılır. En yaygın yöntemler:

Over-sampling: Azınlık sınıfın örnek sayısı artırılır. En popüler tekniklerden biri SMOTE (Synthetic Minority Over-sampling Technique)'dur.

Under-sampling: Çoğunluk sınıfından rastgele örnekler çıkarılır. Ancak bu yöntem, bilgi kaybı riskini beraberinde getirebilir. En temel under-sampling yaklaşımı olan **Random Under Sampling (RUS)**, çoğunluk sınıfından rastgele örnekler silerek denge sağlar. Bu yöntem oldukça basit olmasına rağmen, bilgi kaybına yol açabileceği için dikkatle uygulanmalıdır.

SMOTE Türevleri: (SMOTEENN, Borderline-SMOTE, ADASYN): SMOTE'nin geliştirilmiş versiyonlarıdır. Gürültülü verileri temizler ya da daha kritik bölgelerde örnekler üretir.

Sentetik Azınlık Aşırı Örnekleme (SMOTE)

SMOTE, dengesiz veri kümelerini dengeli hale getirmek için etkili bir yaklaşımdır. Azınlık sınıfı verilerini doğrudan kopyalamak yerine belirli bir örnek için en yakın komşularını temel alarak sentetik veriler üretmektedir. SMOTE yöntemi, azınlık sınıfına ait sentetik örnekler üreterek sınıf dengesizliğini azaltmakta ve böylece modelin bu sınıfa yönelik öğrenme kapasitesini artırmaktadır. Bu sayede sınıflandırma modelinin genelleme yeteneği geliştirilebilir. Ancak, veri setinin boyutunun küçük olması ya da azınlık sınıfındaki örneklerin birbirine oldukça benzer olması durumunda, SMOTE modelin aşırı uyum göstermesine neden olabilir ve bu durum genelleme performansını olumsuz etkileyebilir.

Adaptif Sentetik Örnekleme (ADASYN)

ADASYN, özellikle azınlık sınıfa odaklanarak yeniden örnekleme işlemini yapmaktadır. ADASYN, SMOTE ile benzer şekilde çalışmaktadır ancak farklı ağırlıklar vererek daha hassas bir sentetik örnek oluşturma stratejisini benimsemektedir. Azınlık sınıfa ait örneklerin yoğun bölgelerine daha fazla sentetik örnek eklerken, yoğunluğun az olan bölgelere daha az sentetik örnek eklemektedir. Bu yöntem sınıf dengesizliğinin karmaşık olduğu durumlarda diğer örneklem artırma yöntemlerinden iyi performans sağlamaktadır.

Borderline SMOTE (BLSMOTE)

Borderline SMOTE algoritmasının temelinde gözlem birimlerinin bir sınır çizgisi ile ayrılması vardır. Sentetik veri üretilirken kullanılabilecek iki algoritma türü bulunur bunlar: Borderline smote1 ve Borderline smote2 olarak geçer. Her iki algoritma içinde ilk olarak azınlık sınıfında yer alan her gözlem (Xi) biriminin en yakın komşu uzaklığı bulunur. Burada önemli olan gözlem birimlerinin en yakın komşularının bulunmasında çoğunluk sınıfınında kullanılıyor olmasıdır. (Yani azınlık sınıfına dahil gözlem birimlerinin her birinin en yakın komşunun bulunmasında tüm sınıflar dikkate alınır.)

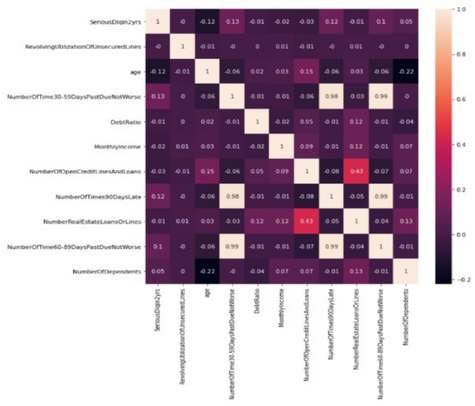
SMOTEENN

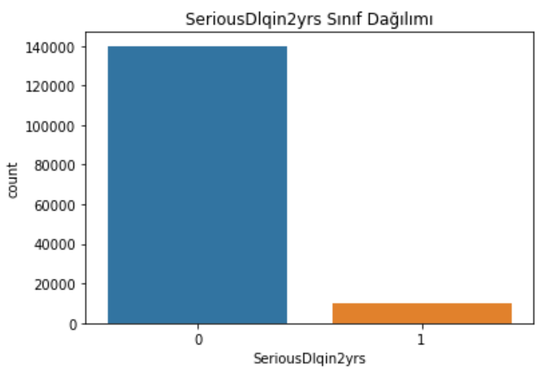
Sınıf dengesizliği sorununu çözmek amacıyla geliştirilmiş hibrit bir örnekleme yöntemidir. Bu yaklaşım, azınlık sınıfı örneklerini sentetik olarak artıran SMOTE tekniği ile, veri setinden hatalı ve gürültülü örnekleri temizleyen ENN (Edited Nearest Neighbors) algoritmasını birleştirir.

Bu yöntemde ilk olarak SMOTE algoritması kullanılarak azınlık sınıfına ait sentetik veriler üretilir. Bu sayede azınlık ve çoğunluk sınıfları arasında daha dengeli bir dağılım sağlanır

VERİ ÖN İŞLEME VE VERİ SETİNE BAKIŞ

Veri seti 150.000 bireysel gözlem ve 11 değişken içermekte olup, her gözlem için müşteriye ait çeşitli finansal ve demografik bilgiler yer almaktadır. Hedef değişken, müşterinin önümüzdeki 2 yıl içerisinde temerrüde düşüp düşmeyeceği (SeriousDlqin2yrs) olup ikili (binary) yapıdadır. Hedef değişken olan SeriousDlqin2yrs, müşterinin önümüzdeki iki yıl içinde temerrüde düşüp düşmeyeceğini ifade eder ve ikili bir sınıflandırma yapısına sahiptir. Ancak veri setindeki gözlemler incelendiğinde, temerrüde düşmeyen bireyler (sınıf 0) toplam örneklerin yaklaşık %93'ünü, temerrüde düşen bireyler (sınıf 1) ise yalnızca %7'sini oluşturmaktadır.





Veri setindeki iki değişken olan "MonthlyIncome" ve "NumberOfDependents" değişkenlerinde eksik gözlemler tespit edilmiştir. Bu eksik değerler, verinin dağılımını fazla bozmadan medyan (ortanca) değeri ile doldurulmuştur.

Tahmin modellerinde çoklu doğrusal bağlantı (multicollinearity) sorununu önlemek için değişkenler arasındaki korelasyon matrisi hesaplanmıştır. "Number Of Time 30-59 Days Past Due Not Worse", "Number Of Time 60-89 Days Past Due Not Worse" ve "Number Of Times 90 Days Late" değişkenleri arasında %99 oranında çok yüksek korelasyon olduğu tespit edilmiştir. Bu nedenle ilk değişken dışındaki diğer iki değişken veri setinden çıkarılmıştır.

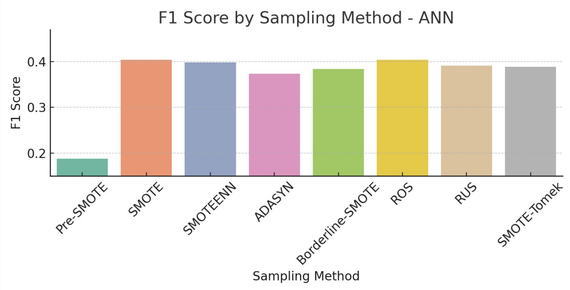
Uç değerler sınırlandırılarak veriler daha normal dağılıma yaklaştırılmıştır.

Veri ön işleme sürecinin sonunda, elde kalan 8 değişkenin hedef değişken ile olan ilişkisini ölçmek amacıyla Bilgi Değeri (Information Value - IV) hesaplanmıştır. Bu hesaplama sırasında değişkenler Weight of Evidence (WOE) yöntemiyle dönüştürülmüştür. IV değeri 0.1 ile 1 arasında olan değişkenler, modele dahil edilmiştir.

Modelleme Süreci ve Performans Değerlendirmesi

Veri dengesizliğini gidermek amacıyla eğitim verisi üzerinde farklı örnekleme yöntemleri uygulanmış ve her bir örnekleme tekniği ile **Artificial Neural Networks (ANN)**, **Random Forest (RF)**, **XGBoost** ve **LightGBM** algoritmaları ayrı ayrı eğitilmiştir. Elde edilen çıktılar şu şekildedir:

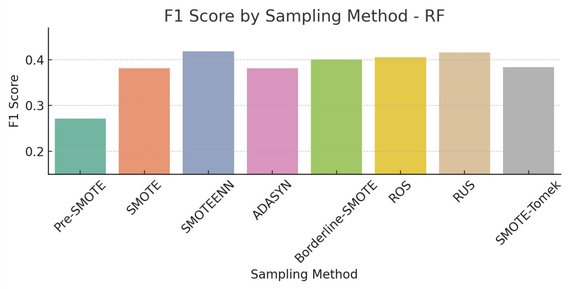
1. Artificial Neural Networks (ANN):



Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	AUC	MCC
Pre-SMOTE	0.9358	0.6084	0.1110	0.1878	0.8543	0.2410
SMOTE	0.9116	0.3676	0.4474	0.4036	0.7927	0.3584
SMOTEENN	0.9104	0.3613	0.4439	0.3984	0.8018	0.3527
ADASYN	0.9066	0.3383	0.4165	0.3734	0.7967	0.3255
Borderline-SMOTE	0.8998	0.3257	0.4663	0.3835	0.8144	0.3372
ROS	0.9116	0.3676	0.4474	0.4036	0.7927	0.3584
RUS	0.8909	0.3118	0.5237	0.3908	0.8341	0.3488
SMOTE-Tomek	0.9062	0.3444	0.4459	0.3886	0.7933	0.3421

ANN modelinde Pre-SMOTE verisi ile F1-Skor yalnızca %18 iken, SMOTE sonrası %40,36'ya ulaşarak önemli bir iyileşme sağlamıştır. RUS ile ANN modeli %52,37 duyarlılık elde ederek azınlık sınıfını en başarılı tanımlayan model konumuna gelmiştir.

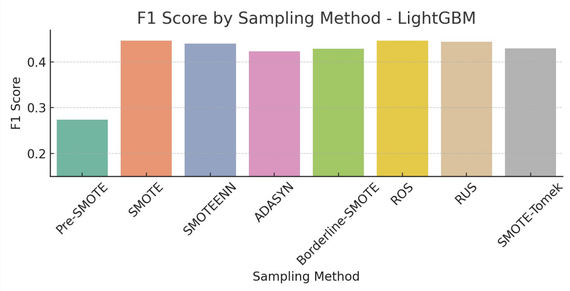
2. Random Forest (RF)



Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	AUC	MCC
Pre-SMOTE	0.9356	0.5556	0.1795	0.2714	0.8393	0.2907
SMOTE	0.8904	0.3063	0.5057	0.3815	0.8326	0.3377
SMOTEENN	0.9083	0.3633	0.4938	0.4186	0.8467	0.3752
ADASYN	0.8904	0.3063	0.5057	0.3815	0.8326	0.3377
Borderline-SMOTE	0.8894	0.3143	0.5541	0.4011	0.8457	0.3621
ROS	0.8982	0.3325	0.5197	0.4055	0.8375	0.3633
RUS	0.9082	0.3616	0.4888	0.4157	0.8545	0.3720
SMOTE-Tomek	0.8902	0.3072	0.5117	0.3839	0.8344	0.3407

SMOTEENN ve RUS yöntemleri RF modelinde yüksek performans sağlamıştır. RUS ile F1-Skor %41,57, AUC %85,45 ve MCC %37,20 seviyesine ulaşmıştır. Pre-SMOTE durumunda ise duyarlılık sadece %17,95 olarak kalmış, bu durum dengesiz verinin etkisini doğrudan göstermektedir.

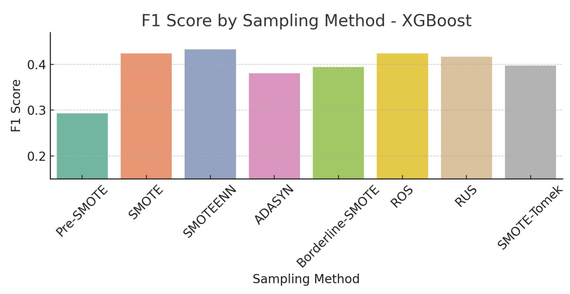
3. LightGBM Modeli



Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	AUC	MCC
Pre-SMOTE	0.9366	0.5840	0.1792	0.2742	0.8645	0.2996
SMOTE	0.9124	0.3862	0.5282	0.4462	0.8545	0.3813
SMOTEENN	0.9185	0.4070	0.4788	0.4400	0.8626	0.3979
ADASYN	0.9067	0.3607	0.5127	0.4235	0.8545	0.3813
Borderline SMOTE	0.9135	0.3838	0.4868	0.4292	0.8610	0.3860
ROS	0.9124	0.3862	0.5282	0.4462	0.8676	0.4056
RUS	0.9184	0.4077	0.4878	0.4441	0.8648	0.4024
SMOTE-Tomek	0.9083	0.3676	0.5172	0.4298	0.8569	0.3880

Pre-SMOTE durumunda en yüksek doğruluk (%93,66) elde edilmiş olsa da, F1-Skor (%27,42) ve duyarlılık (%17,92) gibi azınlık sınıfa duyarlı metrikler oldukça düşüktür. SMOTE ve ROS yöntemleriyle F1-Skor (%44,62) seviyesine çıkarak ciddi iyileşme gözlemlenmiştir. Ayrıca, ROS ile AUC (%86,76) ve MCC (%40,56) değerleri en yüksek düzeye ulaşmıştır.

4.XGBoost Modeli



Veri Seti	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	AUC	MCC
Pre-SMOTE	0.9355	0.5472	0.2005	0.2934	0.8561	0.3047
SMOTE	0.9132	0.3810	0.4788	0.4243	0.8475	0.3809
SMOTEENN	0.9191	0.4072	0.4618	0.4328	0.8575	0.3903
ADASYN	0.8974	0.3190	0.4713	0.3805	0.8385	0.3340
Borderline-SMOTE	0.8943	0.3197	0.5152	0.3946	0.8510	0.3518
ROS	0.9132	0.3810	0.4788	0.4243	0.8475	0.3809
RUS	0.9099	0.3672	0.4808	0.4164	0.8526	0.3725
SMOTE-Tomek	0.9003	0.3334	0.4918	0.3974	0.8455	0.3530

SMOTEENN ile eğitilen model, en yüksek F1-Skor (%43.28) ve AUC (%85.75) değerlerine ulaşmıştır. ADASYN ve Borderline-SMOTE gibi yöntemlerle de tatmin edici sonuçlar elde edilmiş, Pre-SMOTE durumuna kıyasla tüm metriklerde belirgin artış sağlanmıştır.

Sonuç

Bu çalışma, kredi riski tahmini gibi hayati bir karar sürecinde, sınıf dengesizliği probleminin sınıflandırma modelleri üzerindeki etkisini incelemiş ve bu probleme yönelik farklı veri dengeleme yöntemlerinin etkinliğini değerlendirmiştir. Uygulanan örnekleme teknikleri; model başarımını sadece doğruluk düzeyinde değil, aynı zamanda azınlık sınıfını tanıma kapasitesi açısından da gözle görülür biçimde etkilemiştir. Özellikle Pre-SMOTE durumunda, tüm modeller yüksek doğruluk skorları üretmiş olsa da, azınlık sınıfa dair duyarlılık ve F1 skoru ciddi ölçüde düşük kalmıştır.

Her dengeleme tekniği her algoritmayla aynı düzeyde uyumlu sonuçlar vermemektedir. Örneğin LightGBM modeli, ROS ve SMOTEENN ile çok başarılı sonuçlar üretmişken, ANN daha çok RUS ile etkili çalışmıştır. Bu nedenle, "her veri seti + model + sampling kombinasyonu ayrı test edilmelidir" anlayışı benimsenmelidir.

KAYNAKÇA

[1] Chen, L.-P. (2021). *Statistical Learning for Analysis of Credit Risk Data*. IOSR Journal of Mathematics, 17(2), 45–51. DOI: 10.9790/5728-1702044551

[2] Demiröz, S. (2024). *Dengesiz Veri Setlerinde Sınıflandırma Performansını Etkileyen Yaklaşımların İncelenmesi*. Yüksek Lisans Tezi, İstanbul Üniversitesi.

[3] Chawla NV, 2010. Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, 875-886

İSTATİSTİK BÖLÜMÜ

DİJİTAL PAZARLAMADA (FOMO) KAÇIRMA KORKUSU ETKİSİNİN İSTATİSTİKSEL ANALİZİ

Zeliha Eda KARATAŞ

Danışman: Doç. Dr. Fatma Sevinç KURNAZ

ÖZET

Bu bitirme tezinin temel amacı, dijital reklam stratejilerinde sıkça başvurulan psikolojik bir unsur olan FOMO (Fear of Missing Out – Kaçırma Korkusu) etkisinin, kampanya başarısı üzerindeki istatistiksel etkisini incelemektir. Özellikle FOMO'nun, kullanıcıları daha hızlı ve içgüdüsel kararlar almaya yönlendirerek dönüşüm oranlarını artırıp artırmadığı araştırılmıştır.
Yapay olarak oluşturulan 500 reklam örneğinden oluşan veri setinde, reklamların FOMO içeriği taşıyıp taşımadığına göre gruplandırılmış ve temel performans metrikleri (CTR, etkileşim puanı, dönüşüm oranı vb.) analiz edilmiştir. Bağımsız örneklem t-testleri ve eğitim-test ayrımlı lojistik regresyon modeli ile FOMO etkisinin reklam başarısına katkısı istatistiksel olarak test edilmiştir.
Elde edilen bulgular, FOMO içeriği barındıran reklamların yüksek dönüşüm üretme olasılığının anlamlı şekilde arttığını göstermekte ve psikolojik etki temelli stratejilerin dijital pazarlamada kayda değer bir rol oynayabileceğini ortaya koymaktadır.

FOMO (FEAR OF MISSING OUT) – KAÇIRMA KORKUSU

FOMO (Fear of Missing Out – Kaçırma Korkusu), bireylerin sınırlı süreli veya erişimi kısıtlı fırsatları kaçırma endişesiyle karar alma süreçlerinde hızlı ve çoğu zaman duygusal tepkiler vermesine neden olan bilişsel bir eğilimdir. Bu psikolojik unsur, dijital pazarlama stratejilerinde giderek daha fazla yer bulmakta; “sadece bugün”, “son şans”, “stokta az kaldı” gibi mesajlarla kullanıcı üzerinde aciliyet duygusu oluşturarak etkileşimi ve dönüşüm oranlarını artırmayı hedeflemektedir.
Son yıllarda yapılan çalışmalar, FOMO etkisinin çevrim içi tüketici davranışları üzerinde ölçülebilir düzeyde etkili olduğunu ve bu stratejilerin reklam başarısı açısından önemli bir belirleyici haline geldiğini ortaya koymaktadır.

VERİ SETİ

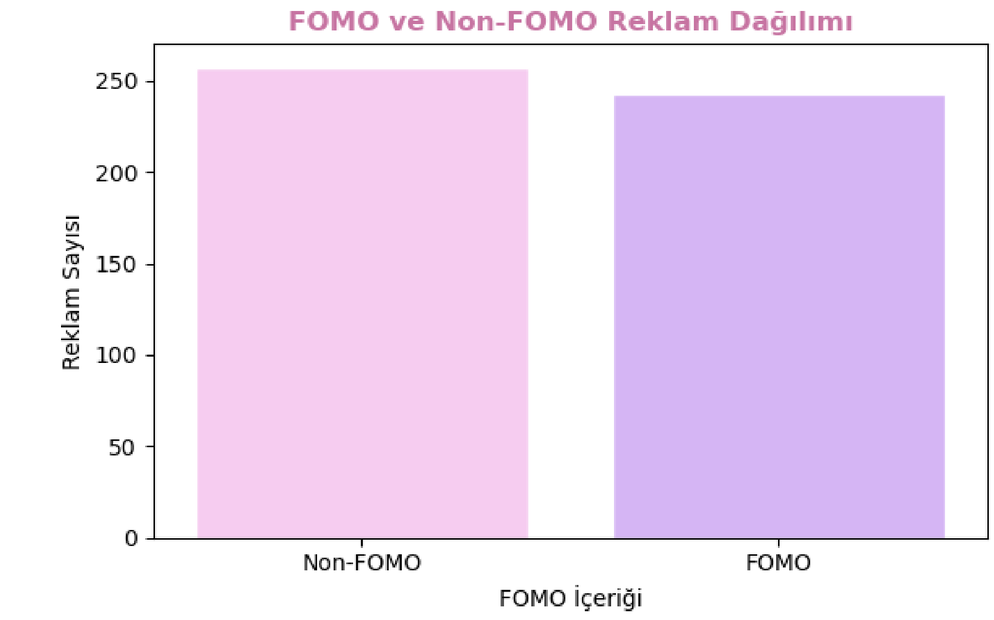
Çalışmada kullanılan veri seti, yapay olarak oluşturulmuş 500 dijital reklam örneğinden oluşmaktadır. Her bir reklam kaydı; reklam metni, FOMO içeriği taşıyıp taşımadığı (fomo_flag), gösterim (impressions), tıklama (clicks), etkileşim puanı (engagement_score), tıklanma oranı (CTR) ve dönüşüm oranı (conversion_rate) gibi performans değişkenlerini içermektedir.

Exploratory Data Analysis — EDA

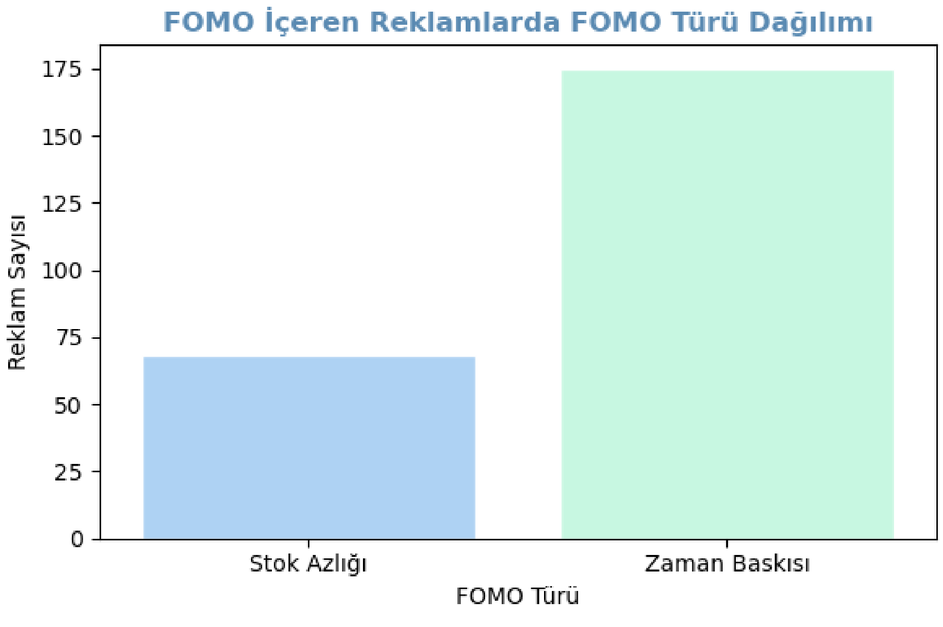
Temel İstatistikler

Metric	Mean	Std. Dev.	Min	25%	50%	75%	Max
CTR (%)	1.93	0.27	1.35	1.75	1.92	2.09	3.08
Engagement Score	2739.75	2708.79	29.92	887.05	1684.25	3393.64	10254.83
Conversion Rate (%)	3.45	1.66	0.52	2.24	3.32	4.56	6.99

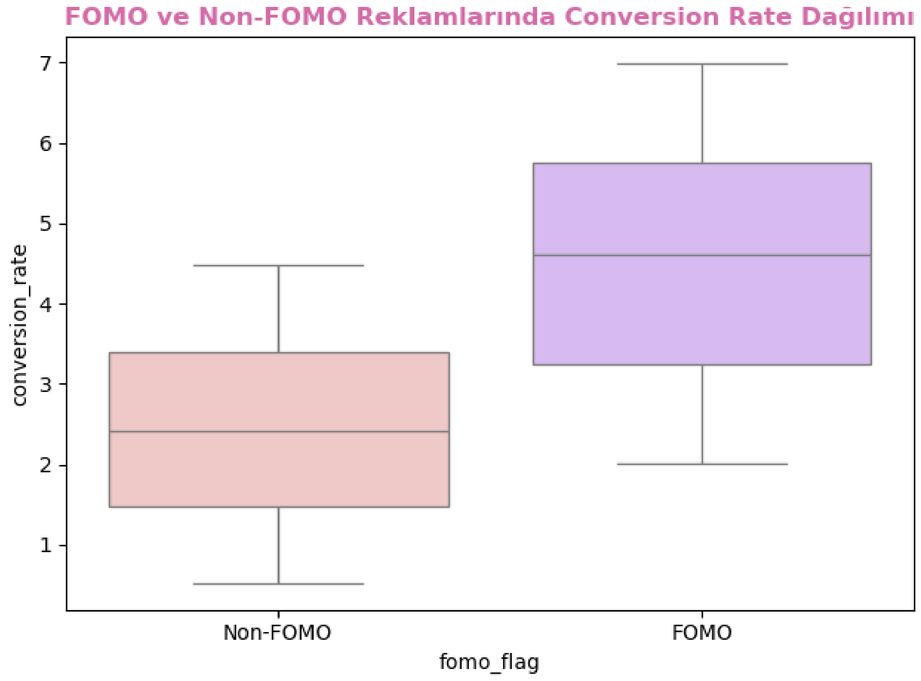
Veri Görselleştirme



FOMO içeren ve içermeyen reklamlar dengeli dağılmıştır. Bu durum, karşılaştırmalı analiz için uygundur.



FOMO temalı reklamlarda zamanla sınırlı tekliflerin, stok azlığına dayalı mesajlardan daha fazla kullanıldığını ortaya konulmuştur. Bu, reklamverenlerin kullanıcıda acil aksiyon alma dürtüsünü öncelikli bir strateji olarak benimsediğini göstermektedir



FOMO içeren reklamların dönüşüm oranı, içermeyenlere göre anlamlı şekilde daha yüksektir.

Uygulama

Bağımsız Örneklem t-Testi

Bu test, iki grup arasında ilgili performans metriklerinin (CTR ve conversion_rate) ortalamalarının anlamlı biçimde farklı olup olmadığını test eder.

Metric	t-statistic	p-value
CTR (%)	0.082	0.9345
Conversion Rate (%)	17.946	0.0

T-testi sonuçlarına göre:
CTR (%) için FOMO ve Non-FOMO grupları arasında istatistiksel olarak anlamlı bir fark bulunmamıştır (p = 0.9345).
Conversion Rate (%) için fark son derece anlamlıdır ($p < 0.001$). FOMO içeren reklamların dönüşüm oranı belirgin şekilde daha yüksektir.
Bu sonuç, FOMO stratejisinin reklam başarısı üzerindeki etkisini istatistiksel olarak doğrulamaktadır.

Lojistik Regresyon

Lojistik regresyon, ikili sonuçlara sahip durumlarda (örneğin başarılı/başarısız) olayların gerçekleşme olasılığını tahmin etmeye yönelik bir sınıflandırma yöntemidir. Bu çalışmada, reklamların yüksek performans gösterip göstermediği bu yöntemle analiz edilmiştir.
Bağımlı değişken olarak, conversion_rate ≥ 4 koşulunu sağlayan reklamlar "başarılı" (1), diğerleri "başarısız" (0) olarak etiketlenmiş ve high_conversion değişkeni oluşturulmuştur. Bağımsız değişkenler arasında ise fomo_flag, engagement_score, clicks ve impressions yer almıştır.

Değişken	Katsayı (β)	p-değeri	Anlamlılık
fomo_flag	1.9524	0.0	✔
engagement_score	0.0011	0.0	✔
clicks	0.0001	0.66	✖
impressions	-4.843e-06	0.474	✖

Lojistik regresyon analizine göre:
-FOMO içeriği (fomo_flag) değişkeni, yüksek dönüşüm olasılığı üzerinde **pozitif ve istatistiksel olarak anlamlı** bir etkiye sahiptir ($\beta = 1.95$, $p < 0.001$). Bu, FOMO içeren reklamların başarı şansını anlamlı düzeyde artırdığını göstermektedir.
-Etkileşim puanı (engagement_score) da anlamlı ve pozitif etkiye sahiptir ($\beta = 0.0011$, $p < 0.001$), yani kullanıcı etkileşimi arttıkça dönüşüm ihtimali de artmaktadır.
-Clicks ve **impressions** değişkenleri modelde yer almasına rağmen, **istatistiksel olarak anlamlı değildir** ($p > 0.05$). Bu, yalnızca tıklama ya da gösterim miktarının başarıyı tek başına açıklamakta yetersiz kaldığını gösterir.

KAYNAKÇA

[1] Przybylski, A. K. et al. (2013). *Fear of Missing Out. Computers in Human Behavior, 29(4)*, 1841–1848
[2] Bright, L. F. et al. (2015). *Social Media Fatigue. Computers in Human Behavior, 44*, 148–155.
[3] Hosmer, D. W. et al. (2013). *Applied Logistic Regression*. Wiley.