THE ROOKİES

Aleyna Aydoğdu – 22040301015

Miray Merve Durmuş – 22040301075

Ali Serhat Aslan – 22040301137

Melih Can – 22040301067

Beyzanur Kaya - 22040301033

YouTube Linki:

PROBLEMİN TANIMI

Polis teşkilatları son yıllarda tutuklama ve müdahale verilerini kamuya açmaya başlamıştır. Bu veriler, yalnızca suç oranlarını incelemek için değil, aynı zamanda tutuklama kararlarını etkileyen faktörleri anlamak açısından da büyük önem taşımaktadır.

Bu projede temel problem, geçmiş tutuklama kayıtları kullanılarak bir olay anında bir kişinin tutuklanma olasılığının tahmin edilmesidir. Bu çalışma, yalnızca tahmin performansına odaklanmakla kalmayıp, aynı zamanda polis uygulamalarındaki olası önyargıları ve eşitsizlikleri görünür kılmayı da hedeflemektedir.

Problem: Tutuklama verileri genellikle yüksek boyutlu, dengesiz ve çarpık dağılımlar içeren yapılara sahiptir. Bu durum, klasik analiz yöntemlerinin ve basit sınıflandırma yaklaşımlarının yetersiz kalmasına neden olmaktadır. Ayrıca demografik bilgilerin varlığı, modelin tarihsel önyargıları öğrenme ve bu önyargıları pekiştirme riskini beraberinde getirmektedir. Bu nedenle hem tahmin performansı yüksek hem de dikkatle değerlendirilmiş bir model geliştirmek önemli bir zorluktur.

Çözüm: Bu çalışmada, geçmiş tutuklama verileri kullanılarak bir olayın tutuklama ile sonuçlanma olasılığını tahmin eden makine öğrenmesi tabanlı bir sınıflandırma yaklaşımı kullanılmıştır. Veri temizleme, özellik mühendisliği ve özellik seçimi adımlarının ardından farklı model aileleri karşılaştırılmış ve özellikle ensemble yöntemlerin daha başarılı olduğu gözlemlenmiştir. Seçilen değerlendirme metrikleri ve dengeleme stratejileri sayesinde, modelin hem ayırt ediciliği hem de genelleme yeteneği artırılmıştır.

Problem Türü:

- Denetimli öğrenme
- İkili sınıflandırma

Hedef:

- Tutuklanma olasılığını tahmin etmek
- Tutuklanma kararlarında etkili değişkenleri belirlemek

VERİ SETİ

- **Veri Seti**
  - *Police Transparency – Arrests* (Denormalized)
- **Kaynak**
  - Data.gov
- **Boyut**
  - 47.445 satır
  - 26 sütun

Projede kullanılan veri seti, Tempe Polis Departmanı'nın geçmiş tutuklama kayıtlarını içermektedir. Veri seti toplamda 47.445 satır ve 26 sütundan oluşmaktadır. Her bir satır, gerçekleşmiş bir tutuklama olayını temsil etmektedir. Veri setinde suç türü, tutuklama zamanı, konum bilgileri, şüphelinin yaşı, cinsiyeti ve ırk bilgileri gibi hem sayısal hem de kategorik değişkenler yer almaktadır. Verinin denormalize edilmiş olması, analiz sürecinde çok sayıda bilgiyi tek tablo üzerinden değerlendirme imkânı sunmuştur.

Veri Seti Özellikleri

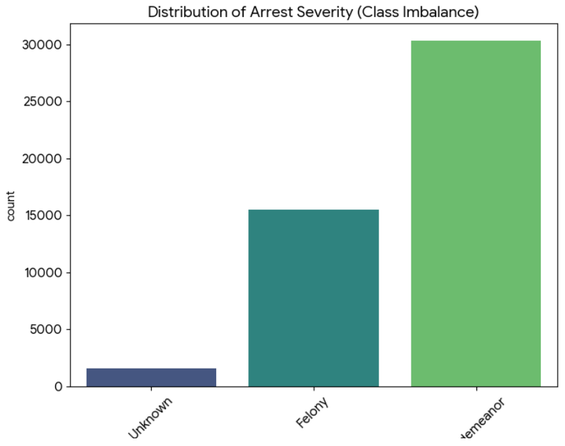
Veri seti incelendiğinde üç temel özellik öne çıkmaktadır.

- Birincisi, sınıf dengesizliğidir. Suç ciddiyeti ve tutuklama sonuçları incelendiğinde, “Misdemeanor” sınıfının büyük ölçüde baskın olduğu görülmüştür. Bu durum, doğrudan accuracy metriğine dayalı modellerin yanıltıcı sonuçlar üretmesine neden olabilir.
- İkinci olarak, bazı sayısal değişkenlerde çarpık dağılımlar bulunmaktadır. Özellikle charge_count değişkeni sağa çarpık bir dağılım göstermektedir. Bu, az sayıda olayda çok yüksek suçlama sayısı olduğunu ve dönüşüm gerektirdiğini göstermektedir.
- Üçüncü olarak, veri setinde eksik değerler mevcuttur. Eksik değerler özellikle konum ve memur bilgileri gibi alanlarda yoğunlaşmıştır ve veri hazırlama aşamasında özel olarak ele alınmıştır.

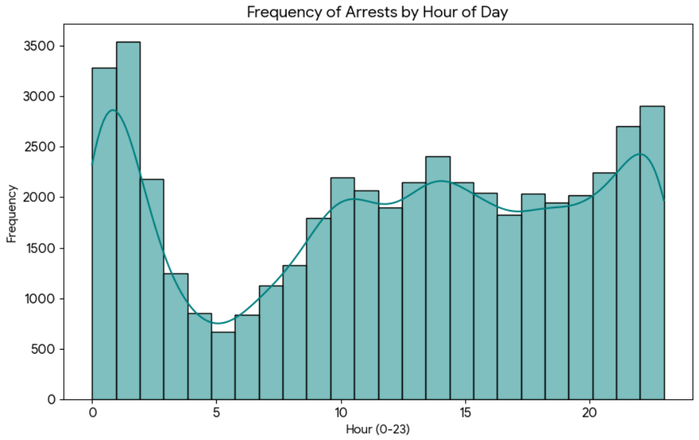
	X	Y	rin	primary_key	arrest_type	arrest_translation	arrest_dt	arrest_time	arrest_hour_of_day	location	...
0	693985	882475	110528	TE20226631	O	On-View: Arrested when first observed/investi...	2022/09/24 21:17:00+00	2117	21	2XX E 5TH ST	...
1	710663	880237	110739	TE20226842	T	Taken Into Custody: Arrest on warrant or PC f...	2022/10/04 17:30:00+00	1730	17	9XX S ACORN AVE	...
2	704259	878441	110587	TE20226690	T	Taken Into Custody: Arrest on warrant or PC f...	2022/09/27 21:52:00+00	2152	21	1XXX E APACHE BLVD	...
3	693160	868321	110521	TE20226624	O	On-View: Arrested when first observed/investi...	2022/09/24 21:00:00+00	2100	21	4XXX S MILL AVE	...
4	708569	883842	110789	TE20226892	T	Taken Into Custody: Arrest on warrant or PC f...	2022/10/07 03:54:00+00	354	3	2XXX W RIO SALADO PKWY	...

KEŞİFSEL VERİ ANALİZİ (EDA)

Veri seti, 40'tan fazla öznitelik barındıran ve hem kategorik hem de sürekli değişkenlerin bir arada bulunduğu heterojen bir yapıya sahiptir. Yapılan keşifsel veri analizi (EDA) sonucunda, hedef değişkende ciddi bir sınıf dengesizliği (class imbalance) olduğu belirlenmiştir; verilerin yaklaşık %75'i belirli bir sınıfta toplanırken, geri kalan %25'lik kısım azınlık sınıfını oluşturmaktadır. Ayrıca, arrest_hour_of_day gibi zaman bazlı değişkenlerin suç yoğunluğu ile korelasyonu incelenmiş, belirli saat aralıklarında varyansın arttığı gözlemlenmiştir. Sayısal değişkenler arasındaki multikolinearite (çoklu doğrusallık) durumunu analiz etmek için korelasyon matrisleri kullanılmıştır.

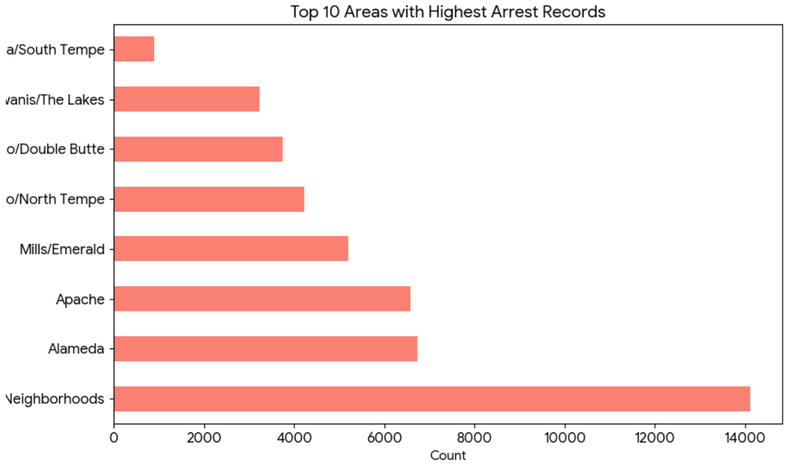


Bazı bölgelerde suç yoğunluğu artarken, aynı zamanda bu bölgelere ait adres bilgilerinde eksiklikler olduğunu fark ettik, Bu eksik alanları silmek yerine, koordinat verilerini kullanarak tamamladık ve modelin konum bilgisini doğru işlemesini sağladık.

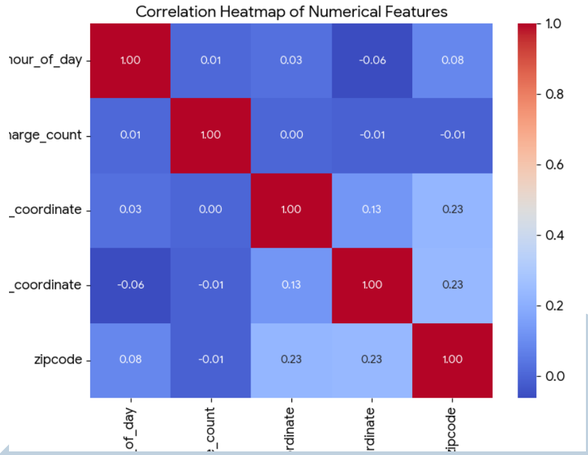


Isı haritası, birbiriyle çok benzer olan ve modeli yanıltabilecek (gereksiz) verileri ayıklamamızı sağlamıştır. Örneğin, suç sayısı ile suçun ciddiyeti arasındaki güçlü bağı bu tablo sayesinde netleştirip modele doğru şekilde öğrettik.

Verilerimizde hafif suçlar (Misdemeanor), ağır suçlara (Felony) göre yaklaşık 2 kat daha fazladır. Bu durum, modelin sadece çoğunluğu öğrenmemesi için 'Balanced' (Dengeli) eğitim yöntemini kullanmamızın temel sebebidir.



Tutuklamalar akşam 21.00 ile gece 03.00 arasında en yüksek seviyeye ulaşmaktadır. Bu belirgin artış, ‘zaman’ değişkeninin kişinin tutuklanıp tutuklanmayacağını tahmin etmede en güçlü ipuçlarından biri olduğunu göstermektedir.



YAKLAŞIM ve TASARIM KARARLARI

İzlenen Adımlar

1. Veri temizleme
2. EDA
3. Feature engineering
4. Feature selection
5. Model eğitimi
6. Performans değerlendirme

Tasarım Kararları

- Dengesiz veri – class_weight
- Yüksek boyut – feature seletion

Projenin genel yaklaşımı; veri temizleme, özellik mühendisliği, özellik seçimi, çoklu model eğitimi ve karşılaştırmalı değerlendirme adımlarından oluşmaktadır. Tasarım aşamasında en kritik kararlar, veri setinin dengesiz yapısı ve yüksek boyutluluğu göz önünde bulundurularak alınmıştır.

Bu nedenle, modellerde ‘class_weight=balanced’ gibi parametreler kullanılmış, ayrıca yüksek boyut problemini azaltmak için çeşitli feature selection yöntemleri uygulanmıştır. Model başarısı değerlendirilirken tek bir metrik yerine, birden fazla metrik birlikte ele alınmıştır.

KULLANILAN METRİKLER

- Model değerlendirme sürecinde ana metrik olarak **ROC-AUC** seçilmiştir. ROC-AUC metriği, farklı karar eşikleri altında modelin pozitif ve negatif sınıfları ayırt etme gücünü ölçmesi nedeniyle bu problem için oldukça uygundur.

- Buna ek olarak, **F1-Score** ve **Accuracy** metrikleri de raporlanmıştır. Ancak veri setindeki sınıf dengesizliği nedeniyle Accuracy metriği tek başına yeterli görülmemiş; özellikle F1-Score'un yorumlanmasına daha fazla önem verilmiştir. Bu yaklaşım, modellerin yalnızca çoğunluk sınıfına yönelmesini engellemeyi amaçlamaktadır.

EN BAŞARILI MODEL

En Başarılı Model: ExtraTrees

- Accuracy: 0.9181
- Precision: 0.8120
- F1 Skoru: 0.8325
- ROC-AUC: 0.9630

Yapılan çalışmalar sonucunda, veri kümesi üzerinde özellikle **ROC-AUC ve F1-Score** metrikleri açısından en yüksek performansı gösteren model **ExtraTreesClassifier** olmuştur. Modelin sınıf dengesizliğini dikkate alan yapısı ve rastgeleliğe dayalı karar mekanizması sayesinde, azınlık sınıf üzerindeki performansı diğer modellere kıyasla daha dengeli bir şekilde artmıştır.

Feature Engineering Yaklaşımları

Veri setindeki ham bilgileri modeller için daha anlamlı hale getirmek amacıyla, grup üyeleri arasında yapılan iş bölümü doğrultusunda aşağıdaki feature engineering yöntemleri uygulanmıştır:

- Veri Temizleme ve Ön İşleme: Eksik değerlerin tespiti yapılmış, özellikle %5'in altında eksik veriye sahip sütunlar analiz edilmiştir. Tekrarlayan satırlar temizlenerek veri bütünlüğü sağlanmıştır.
- Ölçeklendirme (Scaling): KNN ve LinearSVC gibi mesafe tabanlı modellerin doğru çalışabilmesi için StandardScaler kullanılarak sayısal özellikler aynı ölçeğe getirilmiştir. Ağaç tabanlı modellerde ölçeklendirme uygulanmamıştır.
- Boyut İndirgeme (PCA): PCA tekniği, Logistic Regression ve Ridge Classifier gibi doğrusal modellerde veri karmaşıklığını azaltmak ve varyansı koruyarak eğitim süresini iyileştirmek amacıyla kullanılmıştır.
- Özellik Seçimi (Mutual Information): Hedef değişken ile en yüksek bilgi kazancına sahip özellikleri belirlemek amacıyla Mutual Information tabanlı özellik seçimi uygulanmıştır.

Feature Selection Stratejileri

Modellerin aşırı öğrenmesini engellemek ve en belirleyici öznitelikleri tespit etmek amacıyla, iki temel feature selection yaklaşımı kullanılmış ve bu yöntemler keşifsel veri analizi (EDA) çıktılarıyla desteklenmiştir:

- İstatistiksel Yaklaşım (SelectKBest – Mutual Information): mutual_info_classif yöntemi kullanılarak hedef değişken (is_onview_arrest) ile en yüksek bilgi kazancına sahip özellikler seçilmiştir.
- Yinelemeli Özellik Eleme (RFE): Recursive Feature Elimination (RFE) yöntemi ile, özellikle doğrusal ve klasik makine öğrenmesi modellerinde, model performansına en az katkı sağlayan özellikler adım adım elenerek optimum özellik seti belirlenmiştir.
- Destekleyici Görsel Analiz (EDA): Pair plot ve çapraz tablolar (crosstab) aracılığıyla suç ciddiyeti, tutuklama saati ve demografik değişkenler arasındaki ilişkiler incelenmiş; seçilen özelliklerin anlamlılığı görsel olarak doğrulanmıştır.

En Başarılı 4 Modelin Karşılaştırması

Model	Accuracy	F1 Score	Recall	Precision	ROC-AUC
ExtraTrees	0.9181	0.8325	0.8540	0.8120	0.9630
XGBoost	0.8559	0.9059	0.9688	0.8507	0.8966
HistGradient Boosting	0.8795	0.7409	0.7227	0.7600	0.9225
Random Forest	0.8151	0.7783	0.8151	0.8202	0.8328

Extra Trees Classifier: Genel Performans
Tüm modeller arasında 0.9181 ile en yüksek Accuracy ve 0.9630 ile en yüksek ROC-AUC değerine sahiptir. Modelin sınıflandırma kapasitesi çok güçlüdür ve verideki karmaşık desenleri en iyi çözen algoritma olmuştur.

XGBoost Classifier: En Dengeli ve Hassas Model
F1-Score (0.9059) ve Recall (0.9688) değerlerinde tüm rakiplerini geride bırakmıştır. Bu durum, modelin sadece doğru tahmin yapmakla kalmadığını, aynı zamanda hedef sınıfları kaçırmadan yakaladığını (düşük False Negative oranı) gösterir. Gerçek dünya senaryoları için en güvenilir adaydır.

En Başarılı 4 Modelin Karşılaştırması

Model	Accuracy	F1 Score	Recall	Precision	ROC-AUC
ExtraTrees	0.9181	0.8325	0.8540	0.8120	0.9630
XGBoost	0.8559	0.9059	0.9688	0.8507	0.8966
HistGradient Boosting	0.8795	0.7409	0.7227	0.7600	0.9225
Random Forest	0.8151	0.7783	0.8151	0.8202	0.8328

HistGradient Boosting: Yüksek Ayrıştırma Gücü
0.8795 doğruluk oranıyla Extra Trees'ten sonra ikinci sırada yer alır. Özellikle 0.9225 olan ROC-AUC değeri, modelin pozitif ve negatif sınıfları birbirinden ayırma konusunda oldukça başarılı ve istikrarlı olduğunu kanıtlamaktadır.

Random Forest: Kararlı ve Klasik Yaklaşım
Modern boosting algoritmalarına karşı 0.8151 doğruluk ve dengeli Precision/Recall değerleriyle oldukça tutarlı bir performans sergilemiştir. Model, aşırı öğrenmeye (overfitting) karşı dirençli ve genel geçer bir başarı sunmaktadır.

Model Performans Değerlendirmesi ve Tecrübeler

Bu veri seti ve suç ciddiyeti sınıflandırması problemi özelinde, verideki doğrusal olmayan ilişkileri yakalayabilen ExtraTreesClassifier ve Random Forest gibi ağaç tabanlı toplu öğrenme yöntemlerinin daha uygun olduğu gözlemlenmiştir. Özellikle ExtraTreesClassifier, yüksek boyutlu özellik uzayında rastgele özellik seçimi sayesinde varyansı daha iyi yönetmiş ve genelleme kabiliyeti yüksek bir performans sergilemiştir.Buna karşılık, LinearSVC ve SGDClassifier gibi doğrusal modeller, sınıflar arasındaki ayrım çizgisinin net olmaması nedeniyle Recall değerlerinin düşük kalması sonucunda bu problem için sınırlı bir performans göstermiştir.

Teşekkürler