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
- Regresyon

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

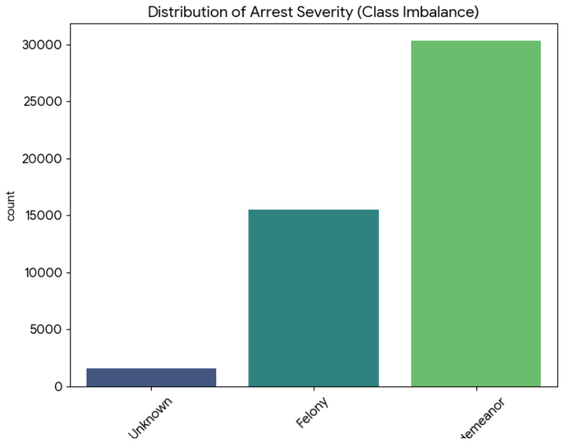
Veri seti incelendiğinde üç temel özellik öne çıkmaktadır.

- 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.

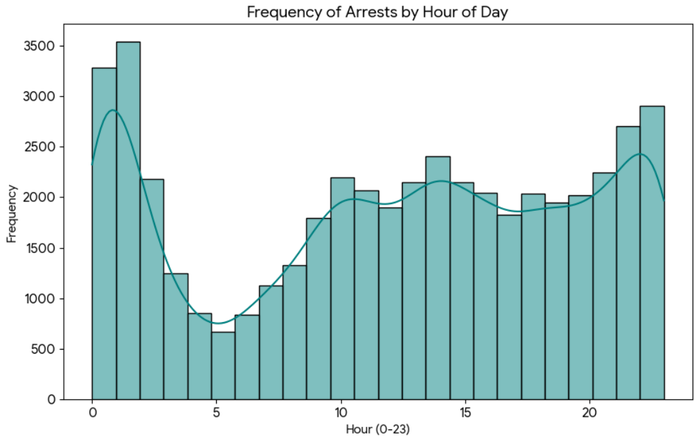
	X	Y	rin	primary_key	arrest_type	arrest_translation	arrest_dt	arrest_time	arrest_hour_of_day	location	...
0	693985	882475	110528	TE20226631	O	On-View: Arrested when first observed/investi...	2022/09/24 21:17:00+00	2117	21	2XX E 5TH ST	...
1	710663	880237	110739	TE20226842	T	Taken Into Custody: Arrest 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.

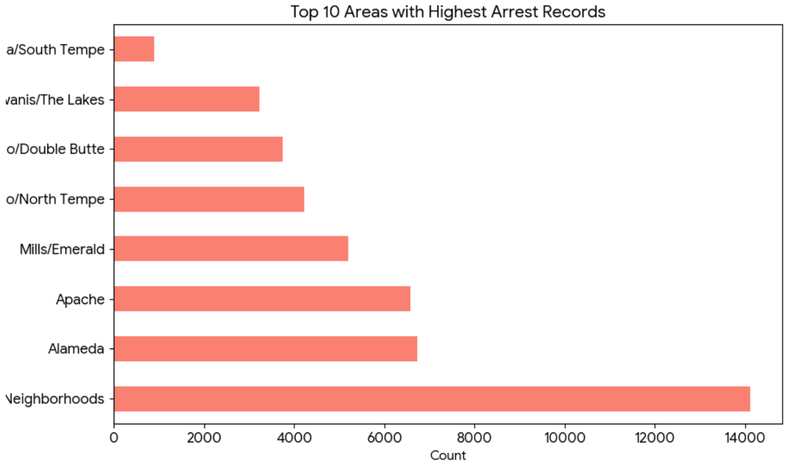


Bazı bölgelerde suç yoğunluğu artarken, aynı zamanda bu bölgelere ait adres bilgilerinde eksiklikler olduğunu fark ettik, Bu eksik alanları silmek yerine, koordinat verilerini kullanarak tamamladık ve modelin konum bilgisini doğru işlemesini sağladık.

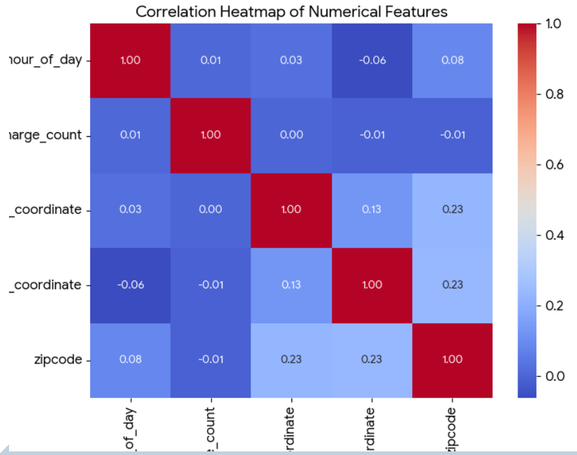


Isı haritası, birbiriyle çok benzer olan ve modeli yanıltabilecek (gereksiz) verileri ayıklamamızı sağlamıştır. Örneğin, suç sayısı ile suçun ciddiyeti arasındaki güçlü bağı bu tablo sayesinde netleştirip modele doğru şekilde öğrettik.

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

- **ROC-AUC (Classification):** Modelin pozitif/negatif sınıfları ayırt edebilme başarısını ölçer. Değer 1’e yaklaştıkça performans artar.

- **Brier Score (Probability Calibration):** Olasılık tahminlerinin doğruluğunu/kalibrasyonunu ölçer. Daha düşük değer daha iyidir.

- **RMSE (Regression):**Tahmin hatalarının karelerinin ortalamasının kareköküdür. Büyük hataları daha fazla cezalandırır. Daha düşük değer daha iyidir.

- **MAE (Regression):**Tahmin hatalarının mutlak değerlerinin ortalamasıdır. Daha yorumlanabilir ve uç değerlere RMSE’ye göre daha az hassastır. Daha düşük değer daha iyidir.

EN BAŞARILI MODEL

En Başarılı Model: Random Forest Regressor (RFR)

- **RMSE:** 0.357
- **MAE:** 0.275

Random Forest Regressor, diğer regresyon modellerine kıyasla en düşük RMSE değerine sahip olması nedeniyle genel hata seviyesini en aza indirmiştir. Ağaç tabanlı ve ensemble yapısı sayesinde verideki doğrusal olmayan ilişkileri etkili bir şekilde öğrenebilmiş, bu da tahminlerin daha dengeli ve kararlı olmasını sağlamıştır. Ayrıca farklı karar ağaçlarının ortalamasını alması, modelin gürültüye ve uç değerlere karşı daha dayanıklı olmasına katkı sağlamış ve bu yönüyle regresyon probleminde en güvenilir sonuçları üretmiştir.

Feature Engineering Yaklaşımları

Veri setindeki ham bilgileri modeller için daha anlamlı hale getirmek amacıyla şu yöntemler izlenmiştir:

- Veri Temizleme ve Ön İşleme: Eksik değerlerin tespiti yapılmış, özellikle %5'in altındaki eksik verilere sahip sütunlar analiz edilmiştir. Tekrarlayan satırlar temizlenerek veri bütünlüğü sağlanmıştır.
- Ölçeklendirme (Scaling): StandardScaler kullanılarak tüm sayısal özellikler aynı ölçeğe getirilmiştir.
- Boyut İndirgeme (PCA): Veri karmaşıklığını azaltmak ve varyansı koruyarak hızı artırmak amacıyla Principal Component Analysis (PCA) tekniği uygulanmıştır.

Feature Selection Stratejileri

Modellerin aşırı öğrenmesini engellemek ve en belirleyici öznitelikleri bulmak için üç temel yaklaşım kullanılmıştır:

- İstatistiksel Yaklaşım (SelectKBest): mutual_info_classif yöntemi kullanılarak, hedef değişken (tutuklama ciddiyeti vb.) ile en yüksek bilgi kazancına sahip özellikler seçilmiştir.
- Yinelemeli Özellik Eleme (RFE): Recursive Feature Elimination (RFE) yöntemiyle, model performansı üzerinde en az etkisi olan özellikler adım adım elenerek optimum özellik seti belirlenmiştir.
- Görsel Analiz (Exploratory Data Analysis): Pair Plot ve çapraz tablolar (crosstab) kullanılarak suç ciddiyeti (severity_trans), tutuklama saati ve etnik köken gibi değişkenler arasındaki korelasyonlar incelenmiş, modelin odaklanacağı kritik değişkenler görselleştirme yoluyla teyit edilmiştir.

En Başarılı 4 Modelin Karşılaştırması

Model	Problem Türü	RMSE	MAE
CatBoost	Regression	0.375	0.294
Random Forest	Regression	0.357	0.275
KNN	Regression	0.362	0.241
LightGBM	Regression	0.374	0.300

Bu dört model, regresyon probleminde en düşük hata değerlerini üreten ve farklı algoritma ailelerini temsil eden modeller arasından seçilmiştir. Random Forest modeli en düşük RMSE değeri ile genel hata seviyesini minimize ederek en dengeli performansı göstermiştir. k-NN modeli, en düşük MAE değeri sayesinde tahmin hatalarına karşı en hassas model olarak öne çıkmıştır. LightGBM ve CatBoost modelleri ise boosting tabanlı yaklaşımlar olarak benzer ve kararlı hata değerleri üretmiş, özellikle karmaşık ve doğrusal olmayan ilişkilerin öğrenilmesinde etkili olduklarını göstermiştir. Bu modellerin birlikte değerlendirilmesi, ensemble ve mesafe tabanlı yöntemlerin regresyon problemindeki performans farklarının karşılaştırılmasını mümkün kılmıştır.

Model Performans Değerlendirmesi ve Tecrübeler

Bu veri seti ve tutuklanma olasılığı tahmin problemi için en uygun yaklaşımın, doğrusal olmayan ilişkileri etkili biçimde öğrenebilen Random Forest, LightGBM ve CatBoost gibi ağaç tabanlı topluluk öğrenme yöntemleri olduğu görülmüştür. Random Forest modeli, 0.357 RMSE değeri ile genel hata seviyesini minimize ederek regresyon probleminde en başarılı model olmuştur. Çok sayıda karar ağacının birleşimi sayesinde veri setindeki özelliklerin oluşturduğu varyansı dengeli şekilde öğrenebilmiş ve gürültüye karşı dayanıklı sonuçlar üretmiştir. Buna karşılık, k-NN ve SVR gibi mesafe tabanlı modeller, bazı örneklerde düşük hata değerleri üretmelerine rağmen genel tahmin kararlılığını sağlamakta zorlanmıştır. Elde edilen sonuçlar, bu problem için ensemble ve boosting tabanlı modellerin daha güvenilir ve genellenebilir çözümler sunduğunu göstermektedir.

Teşekkürler