FET 445 Veri Madenciliği

- US Crime Data
- Grup:Olay Yeri İnceleme
- Youtube Link:
- Github Link:https://github.com/KarakusBaran/Advanced-Crime-ML-Analysis

Üyeler:
Enes Bahadır Salman 22040101041
Baran Karakuş 22040101046
Kerem Oğuz 22040101039
Yiğit Kutluğ 2204010156

Problemin Açıklaması

1. Operasyonel Kaynak Sorunu:

- Her yıl binlerce cinayet vakası işlenmektedir. Ancak polis departmanlarının dedektif sayısı, bütçesi ve zamanı sınırlıdır.
- **Kritik Soru:** Mevcut kaynakları hangi dosyaya aktarmalıyız? Hangi vaka çözülmeye daha yakın?

2. Teknik Engel: "Accuracy Tuzağı" ve Dengesizlik:

- Suç verileri doğası gereği Dengesizdir (Imbalanced).
- Eğer bir şehirde suçların %70'i çözülemiyorsa, hiçbir şey öğrenmeyen "Aptal Model" bile her dosyaya "Çözülemez" diyerek %70 Accuracy (Doğruluk) alabilir.
- **Bizim Farkımız:** Biz bu tuzağa düşmedik. Vize döneminde gördüğümüz bu sorunu, Final döneminde F1-Score ve ROC-AUC odaklı stratejilerle aştık.

3. Projenin Amacı (Goal):

- Olay yeri verilerinden (Silah türü, Bölge, Kurban profili vb.) yola çıkarak; bir dosyanın kapanma ihtimalini öngören "Akıllı Karar Destek Sistemi" geliştirmek.

Veri Seti Açıklaması

1. Veri Seti Künyesi:

- Kaynak: US Crime Data (1980-2014 Cinayet Raporları).
- Boyut: Yaklaşık 630.000 Satır.
- Hedef Değişken (Target): Crime Solved
- 1 (Yes): Suç Çözüldü
- 0 (No): Suç Çözülemedi

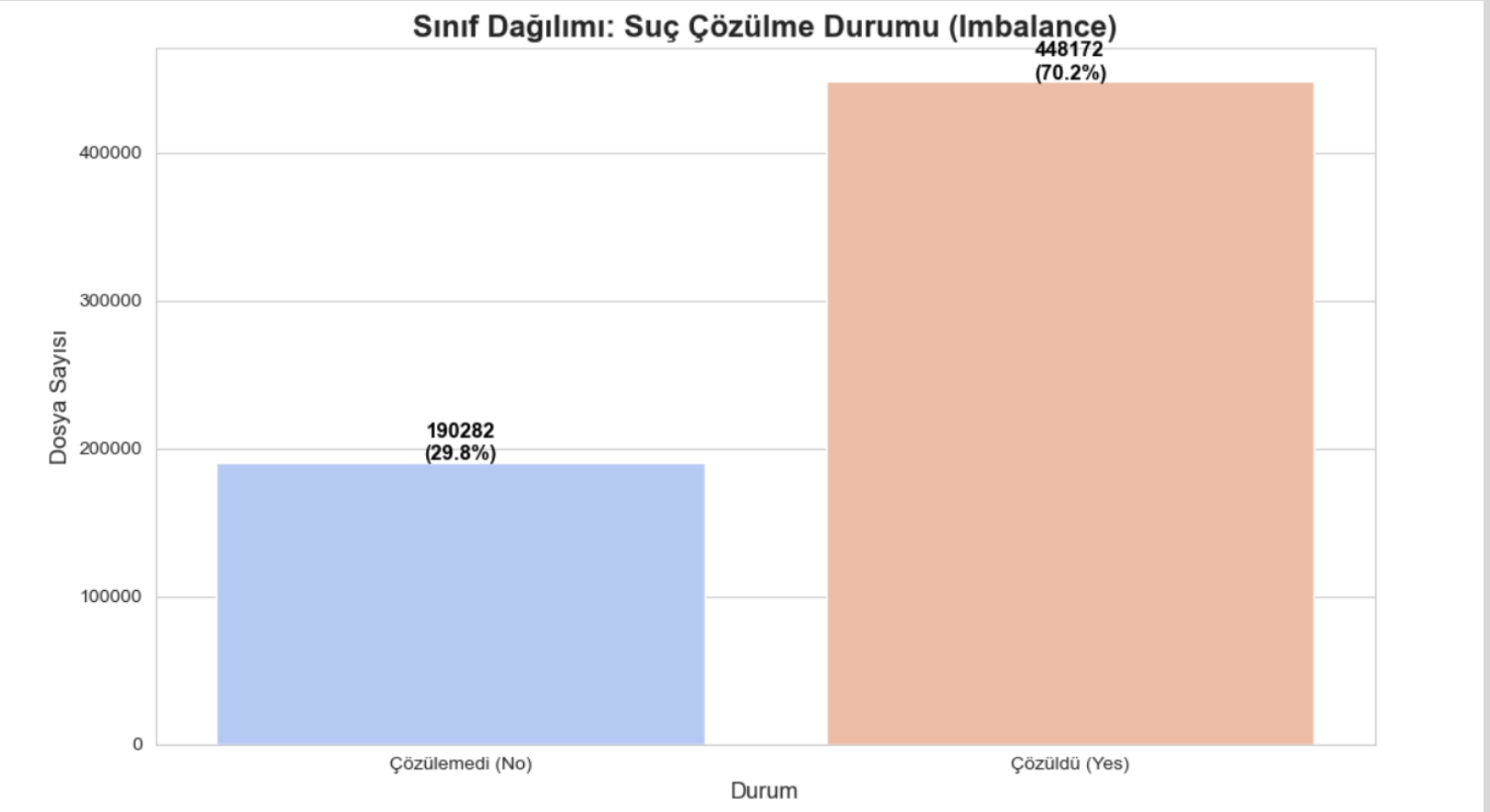
2. Veri Sızıntısı (Data Leakage) Önlemi :

- Tespit: Veri setinde Perpetrator Age (Fail Yaşı), Perpetrator Race (Fail Irkı) gibi sütunlar mevcuttu.
- Problem: Eğer failin yaşı biliniyorsa, fail zaten yakalanmış demektir. Bu durumda suç otomatikman "Çözüldü" olur.
- Müdahale: Modelin cevabı "kopya çekerek" bulmasını engellemek için fail ile ilgili tüm sütunlar eğitimden önce silindi.
- Silinenler: Perpetrator Age, Perpetrator Sex, Perpetrator Race, Record ID.

3. Sınıf Dağılımı (Imbalance):

- Veri setinde "Çözülen" suçlar çoğunlukta, "Çözülemeyenler" azınlıktadır. (Buraya ufak bir pasta grafiği veya bar grafiği koyacağız).

Class Distribution



Yöntem Ve Ön İşleme(Methodology)

1.Eğitim Stratejisi:

- **Train-Test Split:** %80 Eğitim - %20 Test.
- **Dengesizlik Çözümü** :Cost-Sensitive Learning uygulandı.

2. Ön işleme(Feature Engineering):

- **Mevsimsellik:** Month --> Season (Suçların mevsimsel döngüsünü yakalamak için)
- **Yaş Grupları:** Victim Age --> Age_Group (Çocuk/Yetişkin/Yaşlı ayrımı için)
- **Coğrafi Bölge:** State --> Region (50 Eyaleti 4 ana bölgede genellemek için)

3. Performans Metrikleri:

- **F1 Score:**Dengesiz veri olduğu için en kritik metrik
- **Recall(Duyarlılık):**Suçluyu yakalamak için önemli
- **ROC-AUC:**Modelin ayırt etme gücü.
- **Accuracy(Doğruluk):**Modelin genel başarısı.
- **Precision(Kesinlik):**Yanlış olma oranı

En İyi Model Stacking Classifier

1. Neden Bu Model Kazandı?

- Tek bir modelin (Örn: Sadece MLP veya Sadece Ağaç) kararına güvenmek yerine, 3 farklı "Uzman Modelin" görüşünü birleştiren Yığınlama (Stacking) mimarisini kurduk.
- Lineer modeller (SGD, SVC) bu karmaşık veride yetersiz kalırken, Stacking modeli %74 ROC-AUC skoruyla suçluları masumlardan ayırmada en yüksek başarıyı gösterdi.

2. Modelin Mimarisi (Kadro):

- 1. Katman (Uzmanlar):
- ExtraTrees & HistGradient: Verideki karmaşık karar sınırlarını yakalamak için.
- MLP (Sinir Ağı): Lineer olmayan derin ilişkileri görmek için.

2. Katman (Karar Verici - Final Estimator):

- Logistic Regression: Uzmanlardan gelen tahminleri ağırlıklandırarak nihai kararı verdi.

3. Performans Karnesi:

- ROC AUC: 0.74 (En Kritik Başarı - Ayırt Etme Gücü)
- F1 Score: 0.69 (Dengesiz veriye rağmen yüksek)
- Accuracy: %67.9 (Gerçekçi ve güvenilir).

Dönem	Model	Accuracy	F1 Score	Precision	Recall	ROC AUC
Vize	Random Forest	%73.2	0.72	0.71	0.73	0.72
Vize	Decision Tree	%61.2	0.63	0.68	0.61	0.66
Vize	Naive Bayes	%58.1	0.60	0.65	0.58	0.60
Vize	Logistic Regression	%55.3	0.57	0.65	0.55	0.60
Final	Stacking (Kazanan)	%67.9	0.69	0.72	0.68	0.74
Final	HistGradientBoosting	%67.5	0.69	0.72	0.68	0.73
Final	MLP(Neural Network)	%72.6	0.68	0.70	0.73	0.71
Final	Extra Trees	%69.3	0.70	0.70	0.69	0.68
Final	GradientBoosting	%72.1	0.65	0.70	0.72	0.70
Final	Ada Boost	%70.5	0.60	0.66	0.71	0.67
Final	LinearSVC	%70.4	0.58	0.68	0.70	0.61
Final	SGD Classifier	%70.4	0.58	0.68	0.70	0.61

MODELLERIN KARŞILAŞTIRILMASI

Sonuç ve Değerlendirme

1. "Accuracy Tuzağı" Aşıldı:

- Vize dönemindeki temel modellerde gördük ki; dengesiz veri setlerinde model hiçbir şey öğrenmese bile yüksek doğruluk (Accuracy) verebiliyor.
- Final projesinde ise Precision, Recall ve ROC-AUC (0.74) metriklerine odaklanarak, sadece puanı değil modelin güvenilirliğini maksimize ettik.

2. Şampiyon Model: Stacking Classifier:

- Tekil modellerin (LinearSVC, SGD) zayıflıklarını, Ensemble (Topluluk) öğrenmesiyle kapattık.
- HistGradientBoosting ve ExtraTrees modellerini birleştirerek kurduğumuz bu "Ortak Akıl" (Stacking) yapısı, suçluyu ayırt etmede en kararlı ve en yüksek performansı sergileyen mimari oldu.