



Veri Madenciliği - Final Projesi

Efe İngin 21040301056

Umut Özkan 22040301039

Enes Çakır 22040301053

Hüseyin Balık 22040301076

Semih Altun 22040301077

Muhammet Emir Aydoğan 23040301026

Problem Tanımı

Trafik kazalarında kazaların şiddeti tam olarak bilinemez. Bu projenin temel amacı; trafik kazası verilerini ve kaza oluşumundaki kritik faktörleri makine öğrenmesi ve derin öğrenme algoritmalarıyla analiz ederek, gerçekleşen bir kazanın olası şiddetini yüksek doğrulukla tahmin etmektir.

Nihai hedef ise; acil durum müdahale ekiplerinin kaynaklarını en kritik olaylara yönlendirmesini sağlayacak, insan hayatını korumayı ve müdahale süresini optimize etmeyi hedefleyen akıllı bir karar destek mekanizması oluşturmaktır.

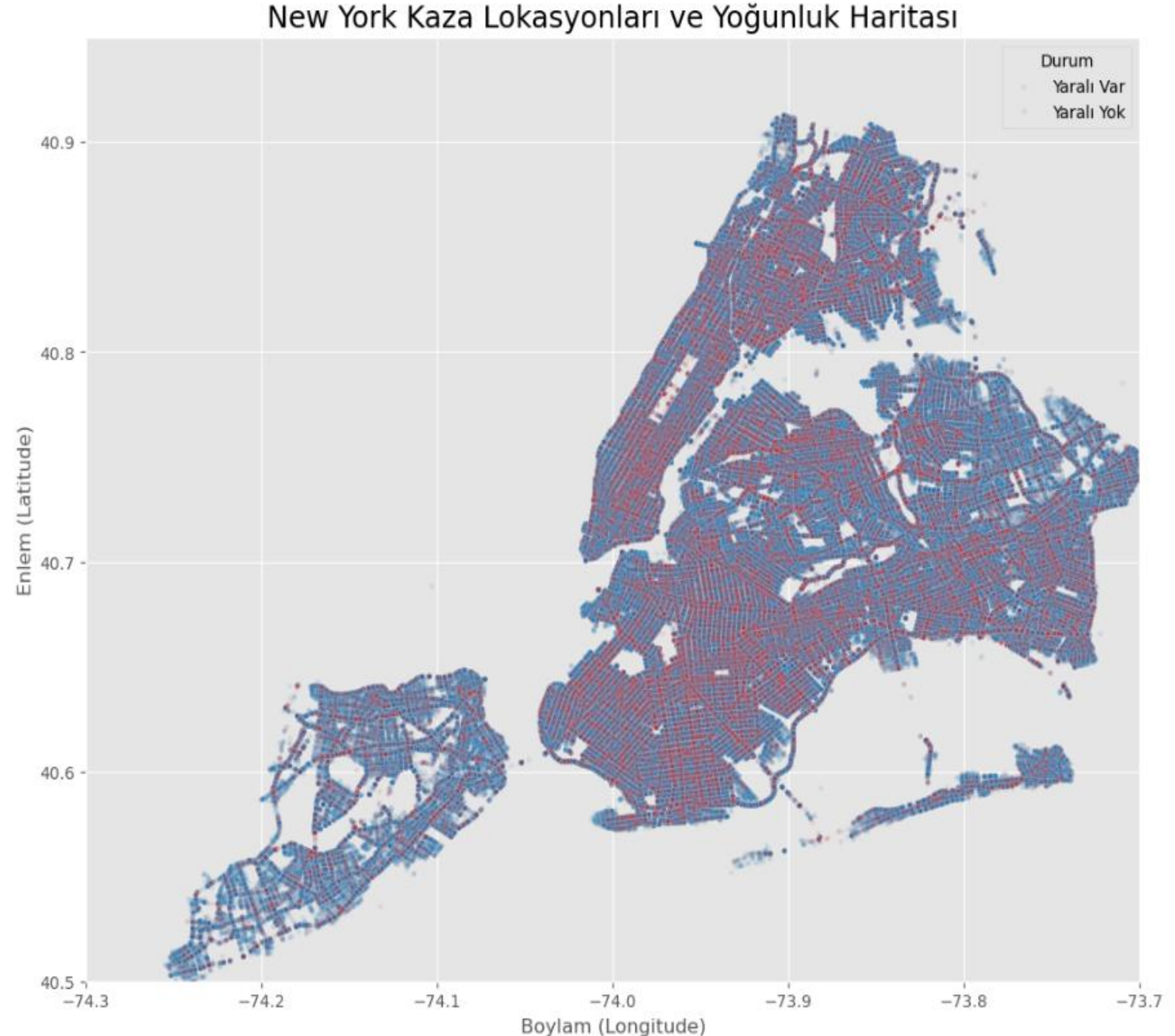
Veri Seti Linki: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data



Github Linki: <https://github.com/hopepi/NYS-Crasher-Machine-Learning/tree/master>

Veri Seti

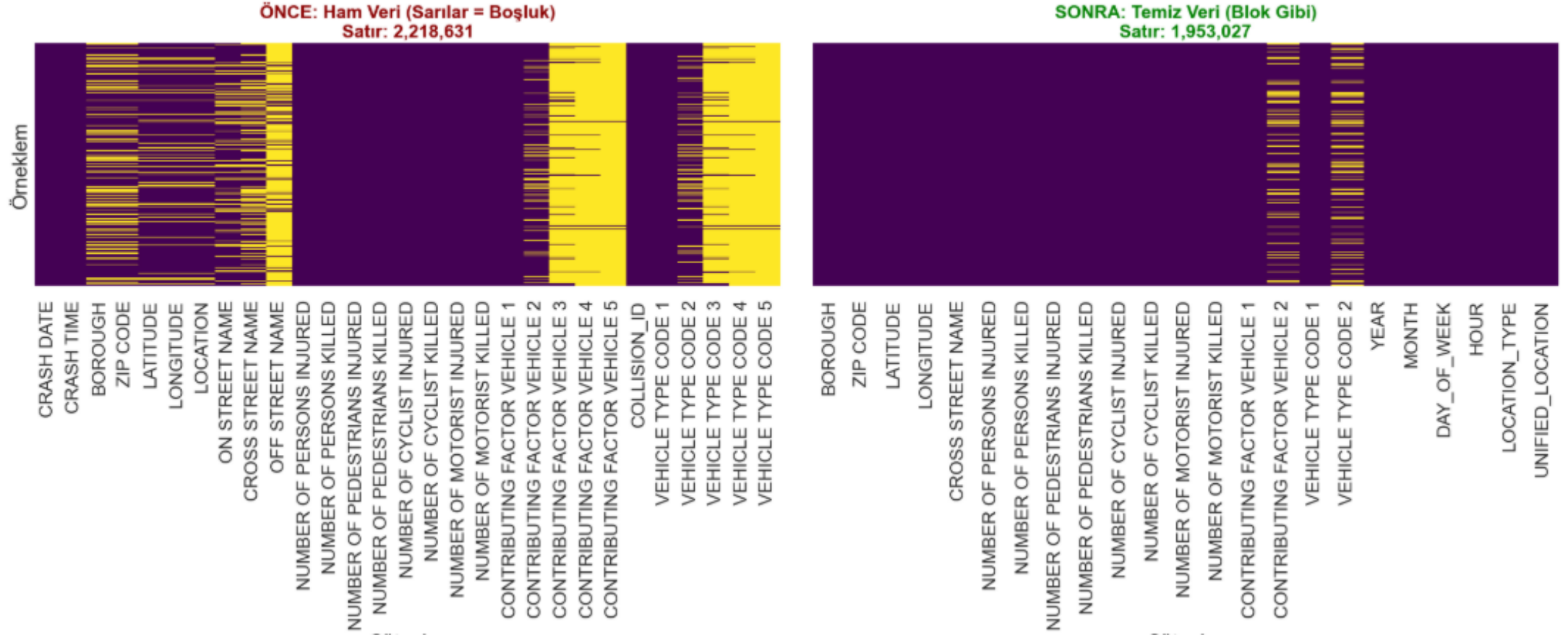
Projede kullanılan veri seti, New York Polis Departmanı tarafından kaydedilen ve NYC Open Data platformu üzerinden erişime açılan resmi trafik kaza kayıtlarından oluşmaktadır. Temmuz 2012'den günümüze uzanan ve 2 milyonu aşkın kaydı içerir. Veri seti, polis memurları tarafından doldurulan raporlara dayanmaktadır. Veri seti; kazanın gerçekleştiği tarih, saat, koordinat bilgileri, araç tipleri, yaralı ve ölü sayılarını da barındırmaktadır.



Veri Ön İşleme

Orijinal veri setinde "Crash Date" adlı sütun, 4 tane farklı sütuna ayrılmıştır. Bu şekilde, trafik kazalarının günü, ayı, yılı ve saati tam olarak belirlenmiştir ve model daha doğru şekilde eğitilmiştir. Bir başka işlem ise, "Borough" adlı sütundaki boş verilerin "Unknown" şeklinde doldurulmasıdır. Bu şekilde veri kaybı engellenmiştir. Aynı zamanda "Vehicle Type 3", "Vehicle Type 4" ve "Vehicle Type 5" isimli sütunlar silinmiştir. Bunun sebebi, Bu sütunların neredeyse tamamının boş olmasıdır. Bu şekilde eğitim daha hızlı sürmüştür ve gürültü engellenmiştir. Sınıflandırma olarak Çoklu Sınıflandırma ile başlanmış, daha sonra ikili sınıflandırmaya geçiş yapılmıştır. Severity değeri ölen insan ve yaralanan insan sayısı olarak belirlenmiştir.

VERİ TEMİZLİĞİ RAPORU (Before vs After)



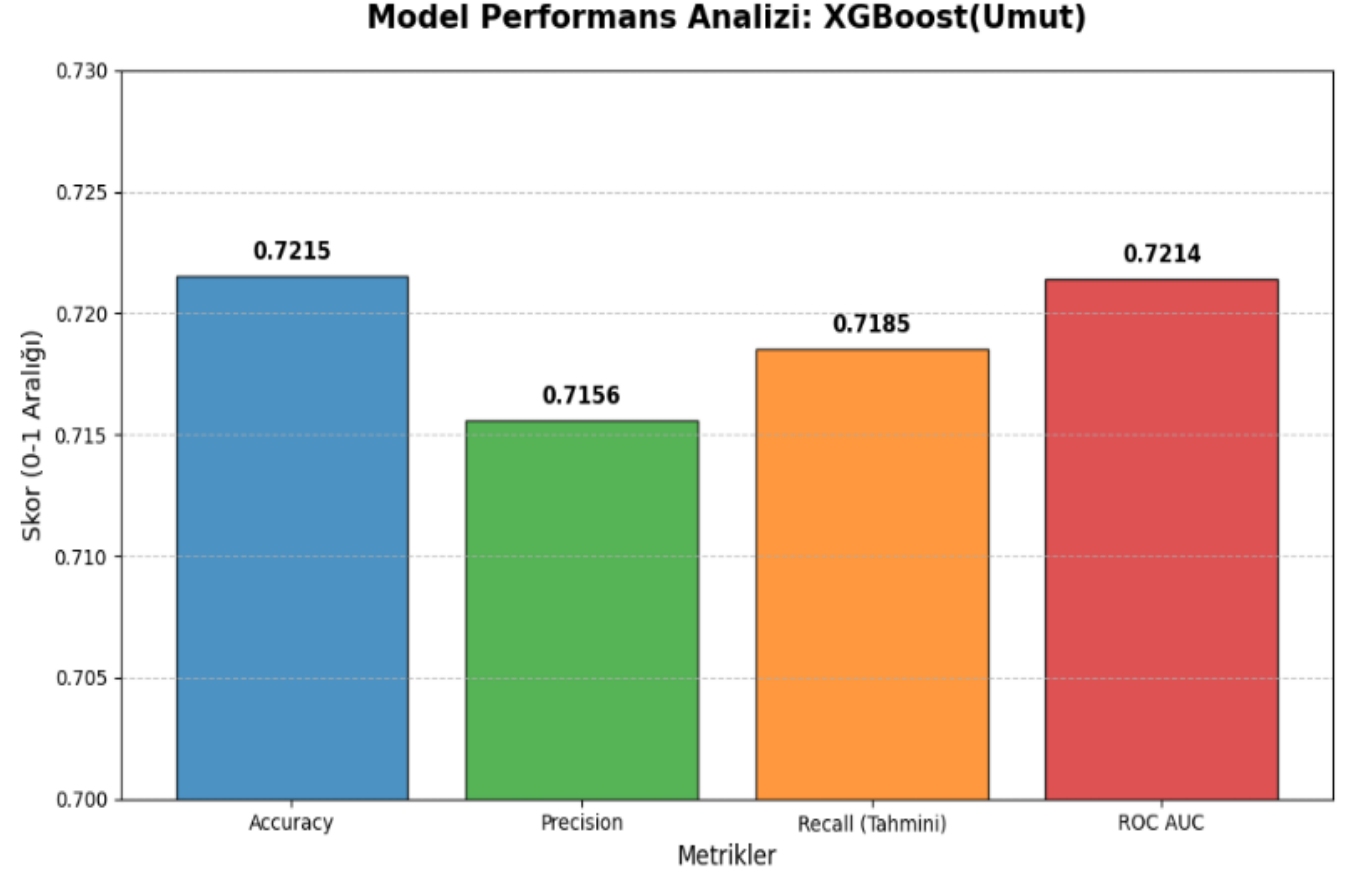
Grup Üyelerinin Kullandığı Modeller

Öğrenci Adı	Gelişmiş Modeller	HiperParametreler
Muhammet Emir Aydoğan	ML: Random Forest,XGBoost,Stacking /Voting Classifier DL : 1D-CNN, RNN(GRU)	RandomizedSearchCV
Hüseyin Balık	ML : AdaBoost,Balanced Random Forest DL : Bi-LSTM,TabNet	Optuna, RandomizedSearchCV
Semih Altun	ML : XGBoost,Random Forest, DL : MLP,RNN	Manuel,GridSearchCV
Enes Çakır	ML : CatBoost,HistGradientBoosting DL : GRU,MLP with BatchNorm	GridSearchCV
Umut Özkan	ML : XGBoost,Random Forest DL : Tabular Transformer,Bi-Lstm	RandomizedSearchCV
Efe İngin	ML : LightGBM,ExtraTrees DL : LSTM,DeepMLP	RandomizedSearchCV

Grup üyelerinin çoklu model denemeleri sonucunda, makine ve derin öğrenmeyi birleştiren hibrit bir mimari kurgulanmıştır. Hız odaklı LightGBM ve varyans yönetimindeki başarısıyla Extra Trees gibi ana sınıflandırıcılar olurken; bu yapı, sıralı ve karmaşık ilişkileri çözümleyen LSTM ve DeepMLP ağlarıyla güçlendirilmiştir. Nihai performans ise RandomizedSearchCV optimizasyonu ile maksimize edilmiştir.

Base Model

Projenin referans noktası olarak, L1 ve L2 regülasyon yeteneğiyle aşırı öğrenmeye direnç sağlayan XGBoost mimarisi seçilmiştir. Dengesiz veri setimizde önceki hataları öğrenerek ilerleyen bu yapı; eksik verileri işleyebilme kapasitesi ve yüksek hesaplama hızıyla, çalışmanın başarı alt sınırını belirleyen en kritik standart olmuştur.



Sonuçlar

Trafik kazalarının çok boyutlu yapısını çözmek adına, her biri farklı bir veri zorluğuna odaklanan çoklu model mimarisi kurgulanmıştır. Farklı algoritmaların ortak karar mekanizmasıyla çalıştırılması, hata payını minimize etmiş ve hayati risk taşıyan kazaların tespitinde yüksek doğruluk oranlarına ulaşılmasını sağlamıştır. Nihai çıktı, acil müdahale süreçlerini optimize edecek akıllı bir karar destek sistemidir.

