Kredi Kartı Dolandırıcılığı Tespiti Analizi ve Model Karşılaştırma Raporu

Grup üyeleri:
Can Şensoy – 2022280151
Nazife Boharalı – 2022280143
Feyza Koç – 2023280119
Nehir Akca – 2023280085

1. Giriş

Bu rapor, verilen kredi kartı dolandırıcılığı tespit veri setinin sezgisel analizini, veri ön işleme aşamasını ve uygulanan makine öğrenimi modellerinin performans karşılaştırmasını sunmaktadır. Temel amaç, dolandırıcılık işlemlerini yüksek doğrulukla tespit edebilen, özellikle **Geri Çağırma (Recall)** metriğini maksimize eden bir model geliştirmektir.

2. Veri Seti Analizi ve Sezgisel Yaklaşımlar

2.1 Veri Seti Özellikleri

Veri seti, toplam 284.807 işlemden oluşmaktadır. Önemli gözlemler şunlardır:

- **Sınıf Dengesizliği:** Veri seti, yalnızca 492 dolandırıcılık vakası () içerdiği için aşırı derecede dengesizdir (highly skewed). Bu, model eğitiminde ciddi bir zorluk teşkil eder.
- **Anonimleştirilmiş Özellikler:** V1'den V28'e kadar olan 28 özellik, Temel Bileşen Analizi (PCA) ile dönüştürülmüştür ve anonimdir.
- **Anonim Olmayan Özellikler:** Sadece 'Time' (Zaman) ve 'Amount' (Miktar) özellikleri dönüştürülmemiştir.
- **Eksik Değerler:** Veri setinde eksik değer bulunmamaktadır.

2.2 Sınıf Dengesizliğinin Etkisi

Sınıf dengesizliği, modellerin çoğunluk sınıfına (**dolandırıcılık olmayan**) odaklanmasına neden olur. Problem açıklamasında belirtildiği gibi, amacımız dolandırıcılık işlemlerinin **%100'ünü tespit etmektir (Recall'ı maksimize etmek)**. Dengeli olmayan bir veri setinde eğitilen bir model, yüksek genel doğruluk (Accuracy) sağlasa bile, dolandırıcılık işlemlerini (azınlık sınıfı) tespit etmede başarısız olabilir.

2.3 Anonim Olmayan Özelliklerin Keşfedilmesi

İstatistik	Dolandırıcılık (Fraud) Miktarı	Normal (Normal) Miktarı
Count	492.00	284315.00
Mean	122.21	88.29
Std	256.68	250.11
Max	2125.87	25691.16

- **Miktar ('Amount'):** Dolandırıcılık işlemlerinin ortalama miktarı (122.21), normal işlemlerin ortalama miktarından (88.29) biraz daha yüksektir. Ancak maksimum normal işlem miktarı (25691.16), maksimum dolandırıcılık işlem miktarından (2125.87) çok daha fazladır.
- **Zaman ('Time'):** Dolandırıcılık işlemlerinin ve normal işlemlerin zaman dağılımları incelendiğinde, işlemin günün hangi saatinde yapıldığının dolandırıcılık tespiti için tek başına güçlü bir gösterge olmadığı görülmüştür.

2.4 Korelasyon Analizi

Korelasyon ısı haritası, PCA bileşenlerinin (V1-V28) birbirleriyle korelasyonunun olmadığını, ancak bazı V bileşenlerinin 'Amount' ve 'Time' ile orta düzeyde ilişki gösterdiğini ortaya koymuştur:

- |Time§V3| ≈ 0.42
- |Amount§V2| ≈ 0.53
- |Amount§V4| ≈ 0.40

Önemli olarak, 'Class' (Dolandırıcılık/Normal) değişkeni, V bileşenlerinin bazılarıyla pozitif ve negatif korelasyonlara sahiptir, bu da bu özelliklerin model için değerli olduğunu göstermektedir.

3. Veri Ön İşleme ve Model Kurulumu

3.1 Ölçeklendirme (Scaling)

'Time' ve 'Amount' özellikleri, diğer PCA dönüşümü görmüş özelliklerle benzer bir ölçeğe getirilmek için **Standard Scaler** kullanılarak ölçeklenmiştir.

3.2 Veri Bölme

Veri, eğitim (%56), doğrulama (%14) ve test (%30) kümelerine ayrılmıştır.

3.3 Değerlendirme Metrikleri

Model performansını değerlendirmek için sadece genel doğruluk (Accuracy) yerine, sınıf dengesizliği nedeniyle daha uygun olan metrikler kullanılmıştır: **Recall (Geri Çağırma), Precision (Hassasiyet), F1-Score** ve **Karmaşıklık Matrisi (Confusion Matrix)**. Özellikle, dolandırıcılığı kaçırmamak adına **Recall'ı maksimize etmek** en kritik hedeftir.

4. Model Karşılaştırması

Aşağıdaki modeller eğitilmiş ve test veri kümesi üzerinde değerlendirilmiştir:

- 4.1. Yapay Sinir Ağı (ANN)
- 4.2. XGBoost
- 4.3. Random Forest

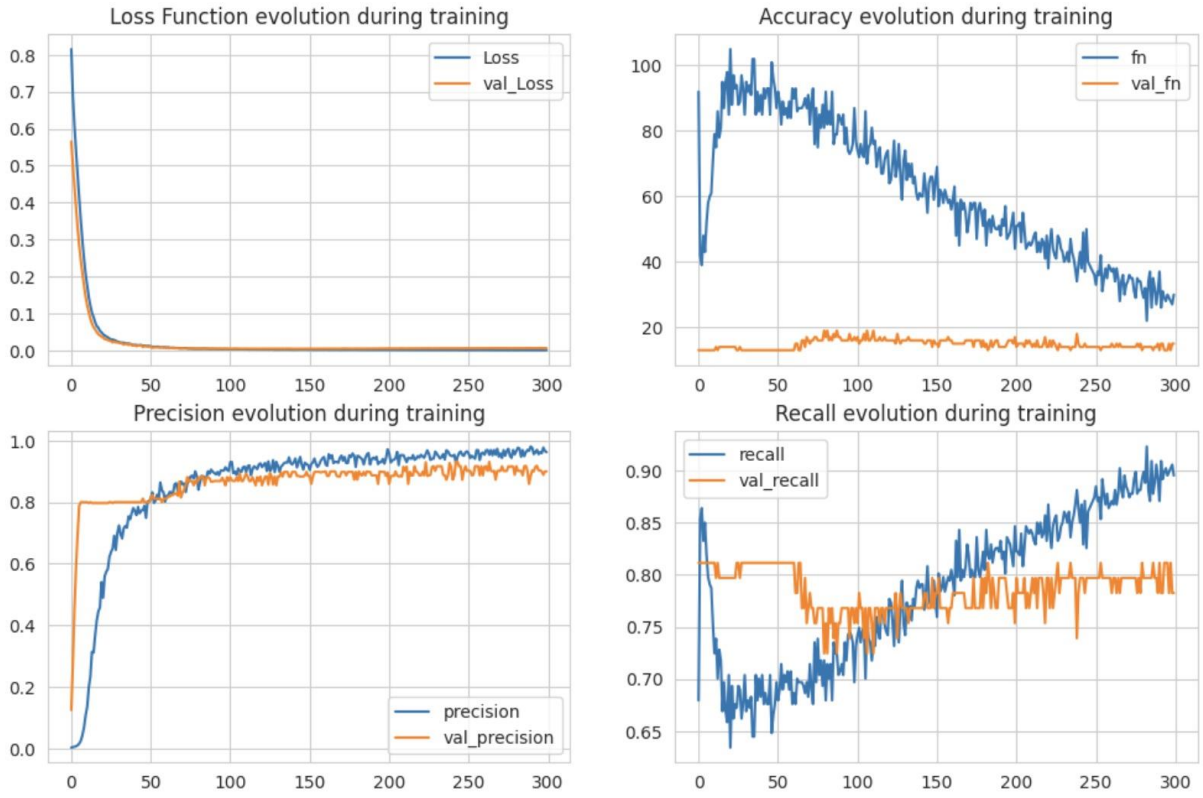
Aşağıdaki tablo, modellerin test veri seti üzerindeki temel performans metriklerini özetlemektedir:

Model	Accuracy (Doğruluk)	Precision (Sınıf 1)	Recall (Sınıf 1)	F1-Score (Sınıf 1)	TP (Gerçek Pozitif)	FP (Yanlış Pozitif)	FN (Yanlış Negatif)
ANN	99.96%	0.92	0.81	0.86	110	9	26
XGBoost	99.96%	0.95	0.82	0.88	111	6	25
Random Forest	99.96%	0.91	0.82	0.86	111	11	25

Not: Sınıf 1=Dolandırıcılık (Fraud).

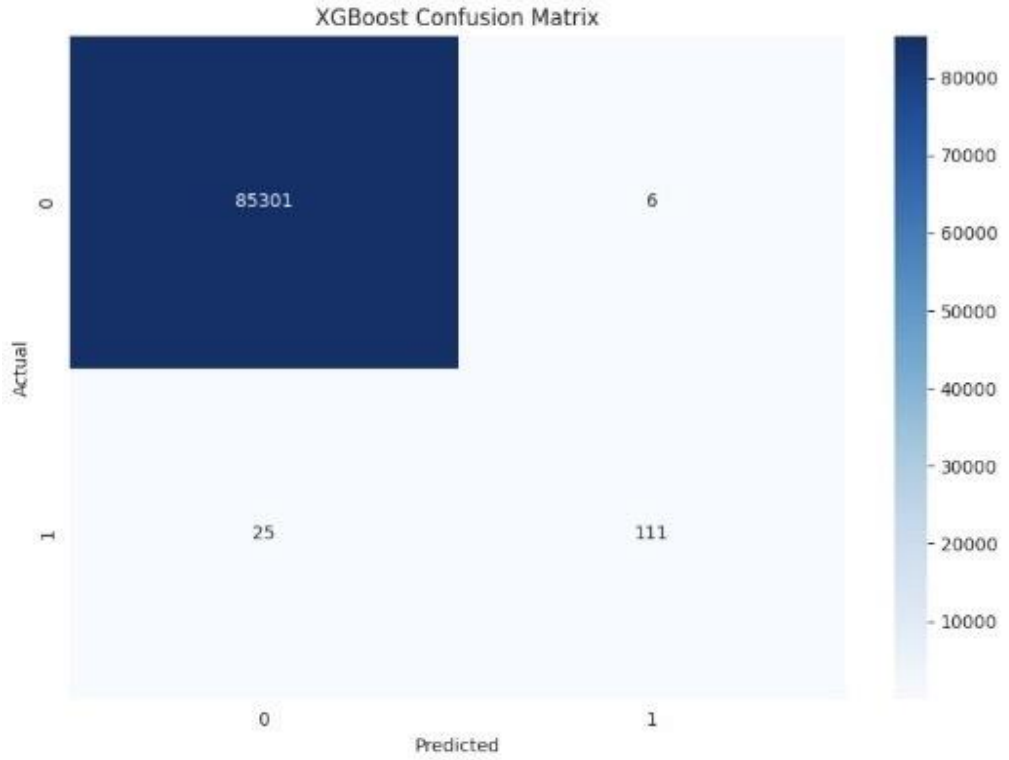
4.1. Yapay Sinir Ağı (ANN) Sonuçları

ANN, Recall ve F1-Score ile başarılı bir performans sergiledi. Sadece 26 dolandırıcılık işlemini kaçırdı (FN).



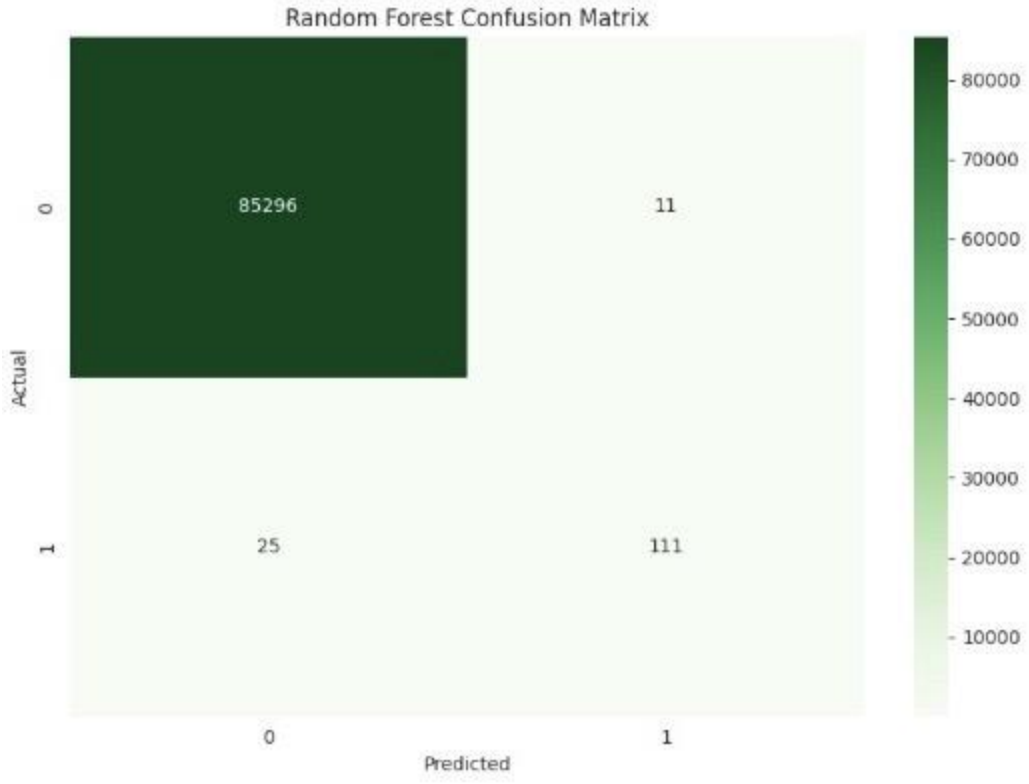
4.2. XGBoost Sonuçları

XGBoost, en iyi F1-Score () ve en düşük Yanlış Pozitif (FP=6) sayısını elde etti. Recall değeri ile CatBoost ve Random Forest ile aynıdır.



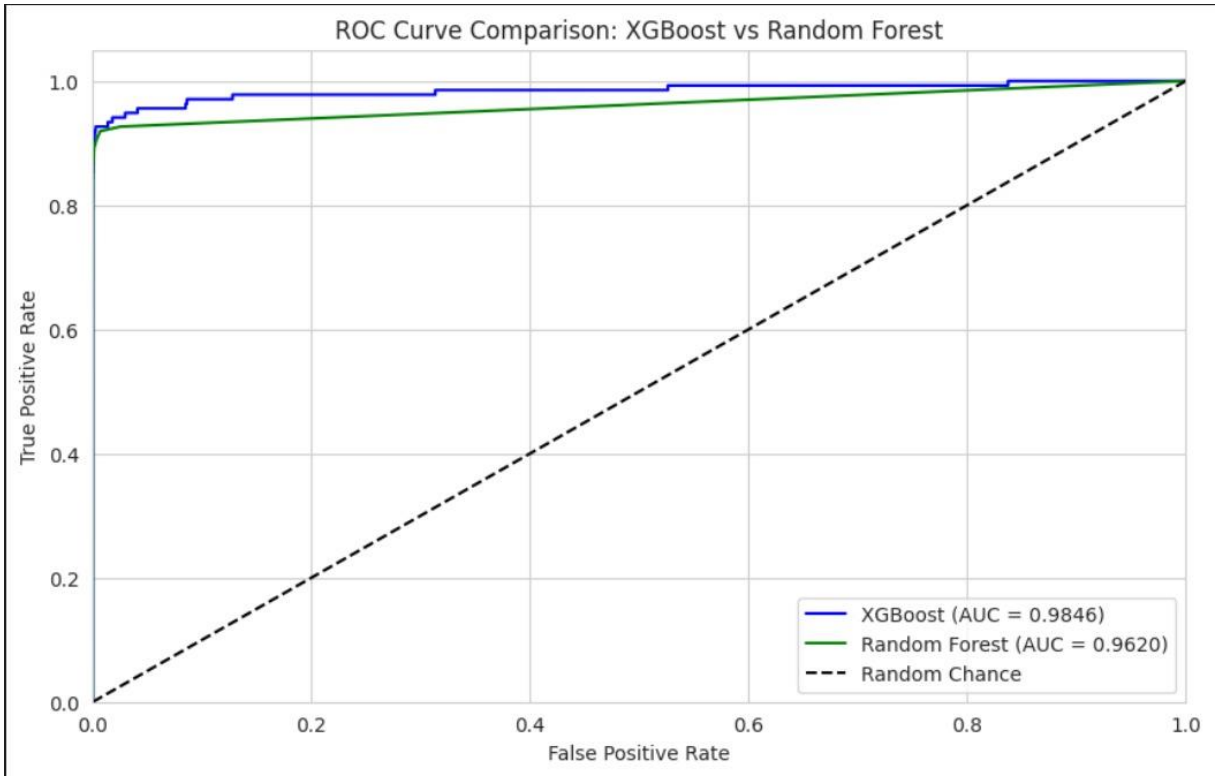
4.3. Random Forest Sonuçları

Random Forest, XGBoost ve CatBoost ile aynı Recall () değerini almasına rağmen, daha fazla Yanlış Pozitif (FP=11) üretti.

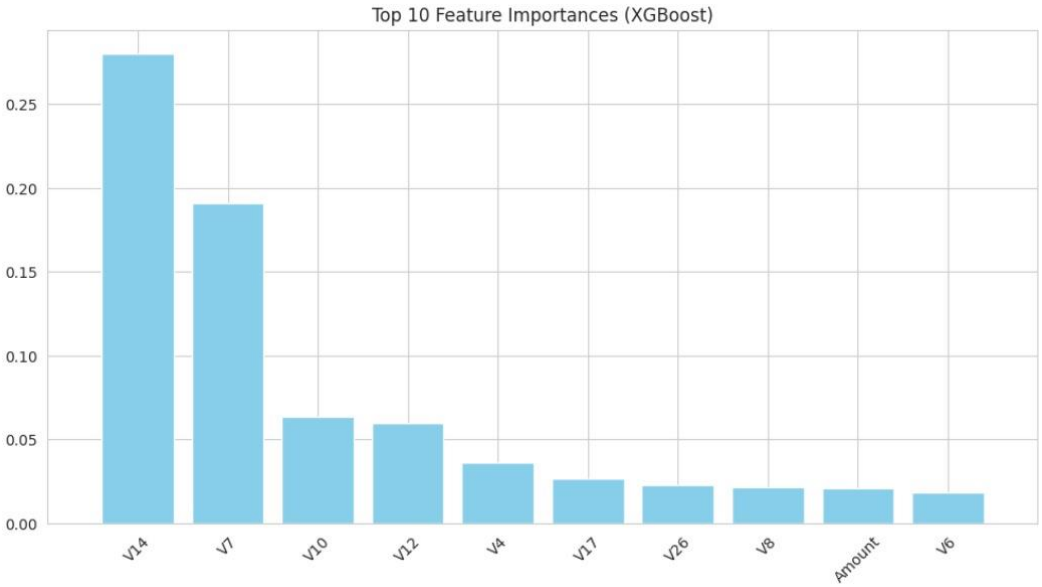


Karmaşıklık Matrisi (Heatmap): Her iki modelin de kaç tane dolandırıcılığı doğru bildiğini, kaç tanesini kaçırdığını yan yana, renkli ve okunabilir bir şekilde gösterir.

ROC Eğrisi: Dengesiz veri setlerinde (Fraud gibi) modelin başarısını ölçmek için en kritik grafiklerden biridir. Çizgi sol üst köşeye ne kadar yakınsa model o kadar iyidir.



Feature Importance: XGBoost modelinin kararı verirken en çok hangi özelliklere (Örn: V14, V10, Amount vb.) dikkat ettiğini çubuk grafik olarak gösterir.



5. Sonuç

Uygulanan modellerin çoğu, genel doğruluk açısından çok yüksek (%99.96) başarı göstermiştir. Ancak, asıl odaklanılması gereken metrik olan **Recall (Geri Çağırma)** metriği açısından, **XGBoost ve Random Forest** en iyi performansı sergilemiştir (Recall: 0.82).

Model karşılaştırma sonuçları şöyledir:

- **Genel Performans:** Üç model de genel doğrulukta (Accuracy) yaklaşık %99.96 ile çok başarılıdır.
- **Recall (Geri Çağırma):**
    - **XGBoost** ve **Random Forest** Recall ile en iyi performansı göstermiştir. Bu, dolandırıcılık işlemlerinin %82'sini doğru bir şekilde tespit ettikleri anlamına gelir.
    - **ANN** biraz daha düşüktür ().
- **F1-Score:**
    - **XGBoost**, ile en yüksek F1-Score'a sahiptir. F1-Score, Precision ve Recall'ın dengeli bir ortalaması olduğu için en iyi genel dengeye sahip modeldir.
- **Hata Analizi:**
    - **Yanlış Negatifler (FN):** Her üç model de 25-26 adet dolandırıcılık işlemini gözden kaçırmıştır. İş hedefi tespit olduğu için, bu **FN'ler kritik hatadır**.
    - **Yanlış Pozitifler (FP): XGBoost** (FP=6) en az sayıda normal işlemi hatalı bir şekilde dolandırıcılık olarak sınıflandırmıştır, bu da en yüksek hassasiyetin (Precision) nedenidir.