Fraud İşlem Tespit Modeli Raporu

Hazırlayan: Ali Kamiloğlu

20201101019

Giriş

Bu rapor, kredi kartı işlemlerindeki sahtekarlıkları tespit etmek amacıyla yapılan veri analizi ve makine öğrenimi modelleme çalışmalarını içermektedir. Veri seti, farklı işlem bilgilerinden ve işlemlerin sahtekarlık olup olmadığını belirten etiketlerden oluşmaktadır. Bu raporda, veri setinin detaylı analizi, eksik değerlerin doldurulması, veri temizleme işlemleri ve çeşitli makine öğrenimi modellerinin uygulanması ele alınmıştır. Ayrıca, sonuçlar çeşitli görselleştirmelerle desteklenmiştir.

Veri Seti Kaynağı

Bu çalışmada kullanılan veri seti, Kaggle'dan alınmıştır.

Veri Seti Kaynağı: [Kaggle - Fraud Detection Dataset]

Veri Seti Tanımı

Veri seti, kredi kartı işlemleriyle ilgili çeşitli bilgileri içermektedir. İşlem verileri ve sahtekarlık olup olmadığını belirten etiketler, veri setinde yer almaktadır. Aşağıda veri setindeki başlıca sütunlar ve açıklamaları verilmiştir:

- **Unnamed: 0**: Benzersiz işlem ID'si
- **trans_date_trans_time**: İşlem tarihi ve saati
- **cc_num**: Kredi kartı numarası
- **merchant**: İşlemin yapıldığı mağaza
- **category**: İşlemin kategorisi (gıda, eğlence, vb.)
- **amt**: İşlem tutarı
- **first**: Kart sahibinin adı
- **last**: Kart sahibinin soyadı
- **gender**: Kart sahibinin cinsiyeti
- **street**: Kart sahibinin adresi
- **city**: Kart sahibinin yaşadığı şehir
- **state**: Kart sahibinin yaşadığı eyalet

- **zip**: Kart sahibinin posta kodu
- **lat**: Kart sahibinin yaşadığı yerin enlemi
- **long**: Kart sahibinin yaşadığı yerin boylamı
- **city_pop**: Kart sahibinin yaşadığı şehrin nüfusu
- **job**: Kart sahibinin mesleği
- **dob**: Kart sahibinin doğum tarihi
- **trans_num**: İşlem numarası
- **unix_time**: İşlem zamanı (Unix formatında)
- **merch_lat**: Mağazanın enlemi
- **merch_long**: Mağazanın boylamı
- **is_fraud**: İşlemin sahtekarlık olup olmadığını belirten etiket (0: Sahtekarlık değil, 1: Sahtekarlık)

Veri Setinin İlk İncelemesi

Veri setinin incelenmesi işlemi sırasında aşağıdaki adımlar uygulanmıştır:

Veri Setinin İlk Satırlarının Gösterimi

Veri setinin ilk birkaç satırı şu şekildedir:

	Unnamed: 0	trans_date_trans_time	cc_num	...	merch_lat	merch_long	is_fraud
0	0	6/21/20 12:14	2291163933867244	...	33.986391	-81.200714	0
1	1	6/21/20 12:14	3573030041201292	...	39.450498	-109.960431	0
2	2	6/21/20 12:14	3598215285024754	...	40.495810	-74.196111	0
3	3	6/21/20 12:15	3591919803438423	...	28.812398	-80.883061	0
4	4	6/21/20 12:15	3526826139003047	...	44.959148	-85.884734	0

Eksik Değerlerin İncelenmesi

Veri setindeki eksik değerlerin analizi sonucu eksik değer bulunmamıştır:

Unnamed: 0 0

trans_date_trans_time 0

cc_num 0

merchant 0

category 0

amt 0

first 0

last 0

gender 0

street 0

city 0

state 0

zip 0

lat 0

long 0

city_pop 0

job 0

dob 0

trans_num 0

unix_time 0

merch_lat 0

merch_long 0

is_fraud 0

Yapay Eksik Değerlerin Eklenmesi ve Doldurulması

Orijinal veri setinde eksik değer bulunmadığından, eksik değerlerin nasıl ele alınacağını göstermek amacıyla bazı eksik değerler yapay olarak eklenmiştir:

Unnamed: 0 0

trans_date_trans_time 0

cc_num 0

merchant 0

category 0

amt 6

first 0

last 0

gender 6

street 0

city 6

state 0

zip 0

lat 0

long 0

city_pop 0

job 0

dob 0

trans_num 0

unix_time 0

merch_lat 0

merch_long 0

is_fraud 0

Bu eksik değerler, `ffill` (forward fill) yöntemi kullanılarak ortanca değerleriyle doldurulmuştur:

Unnamed: 0	0
trans_date_trans_time	0
cc_num	0
merchant	0
category	0
amt	0
first	0
last	0
gender	0
street	0
city	0
state	0
zip	0
lat	0
long	0
city_pop	0
job	0
dob	0
trans_num	0
unix_time	0
merch_lat	0
merch_long	0
is_fraud	0

Temel İstatistikler ve Veri Analizi

Cinsiyet Dağılımı

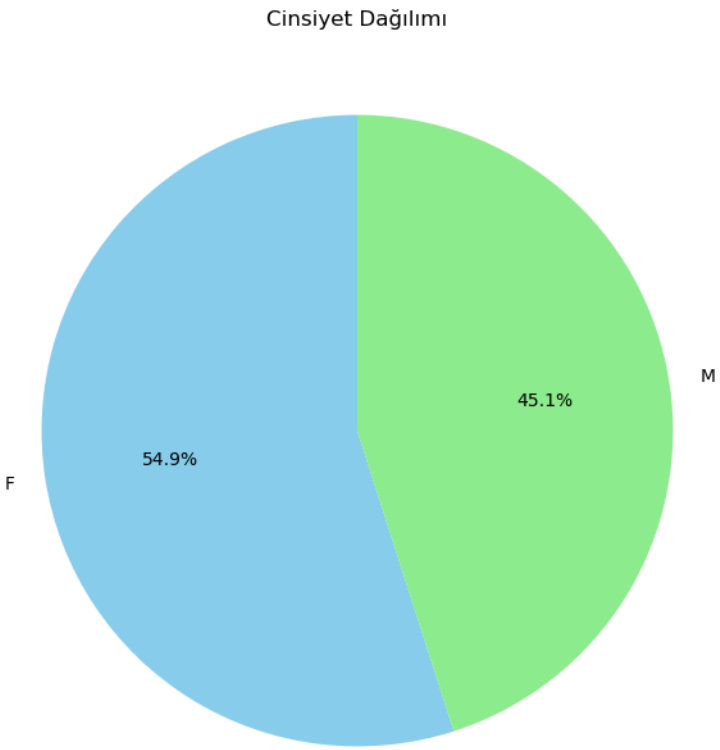
Veri setindeki cinsiyet dağılımı şu şekildedir:

F 36008

M 29527

Name: gender, dtype: int64

Cinsiyet dağılımı, kadınların (F) ve erkeklerin (M) oranlarını göstermektedir. Kadınların sayısı erkeklerden fazladır.



İşlem Tutarı İstatistikleri

İşlem tutarlarının temel istatistikleri aşağıdaki gibidir:

count 65535.000000

mean 69.167788

std 144.185586

min 1.000000

25% 9.670000

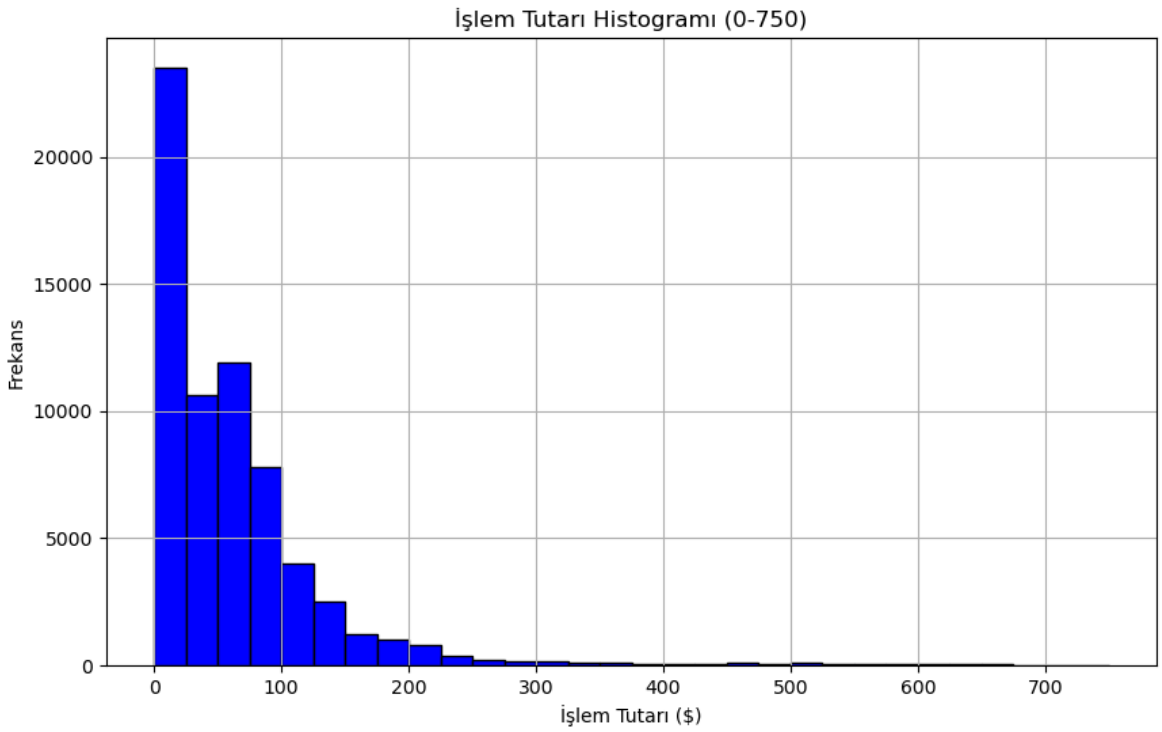
50% 47.150000

75% 82.850000

max 13149.150000

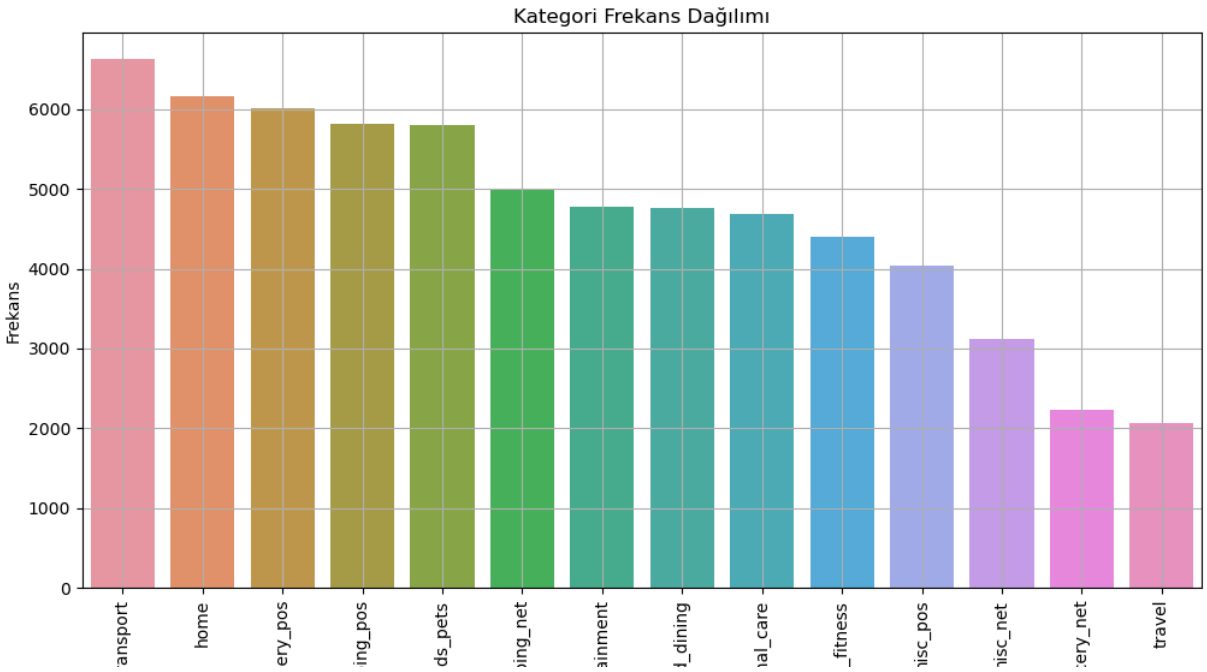
Name: amt, dtype: float64

İstatistikler, işlemlerin çoğunun küçük tutarlarda olduğunu, ancak bazı işlemlerin oldukça yüksek tutarlarda gerçekleştiğini göstermektedir. Bu dağılımı daha iyi anlamak için aşağıdaki histogramı oluşturdum:



Kategori Dağılımı

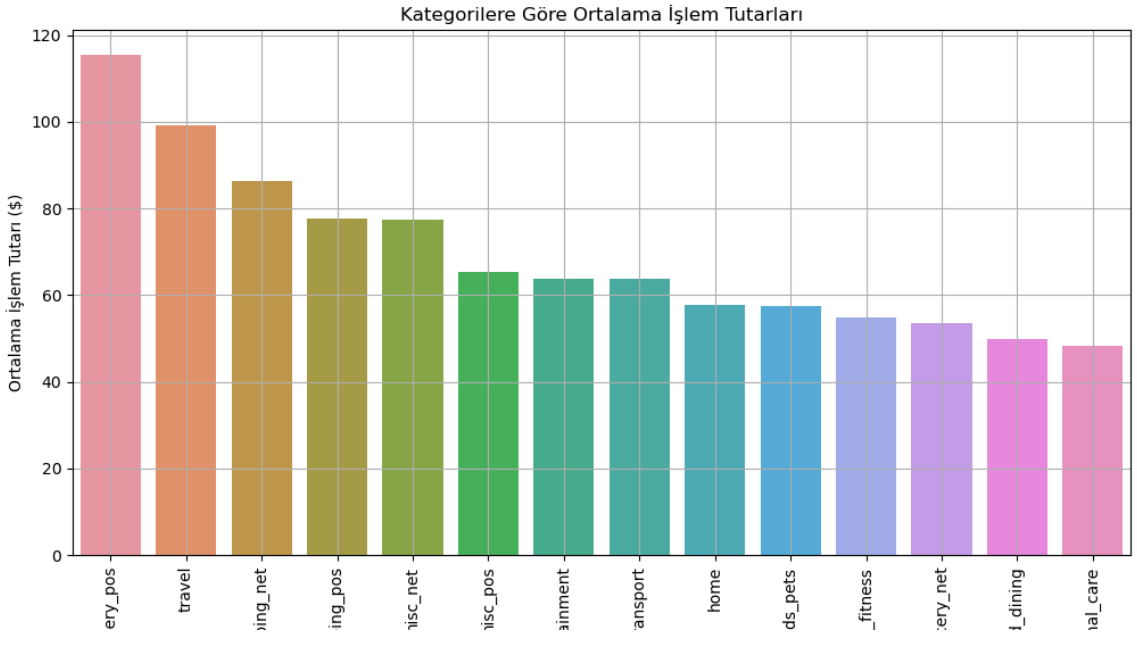
İşlem kategorilerinin dağılımı şu şekildedir:



İşlemler en çok "gas_transport" kategorisinde gerçekleşmiştir. Diğer yaygın kategoriler arasında "home", "grocery_pos" ve "shopping_pos" yer almaktadır.

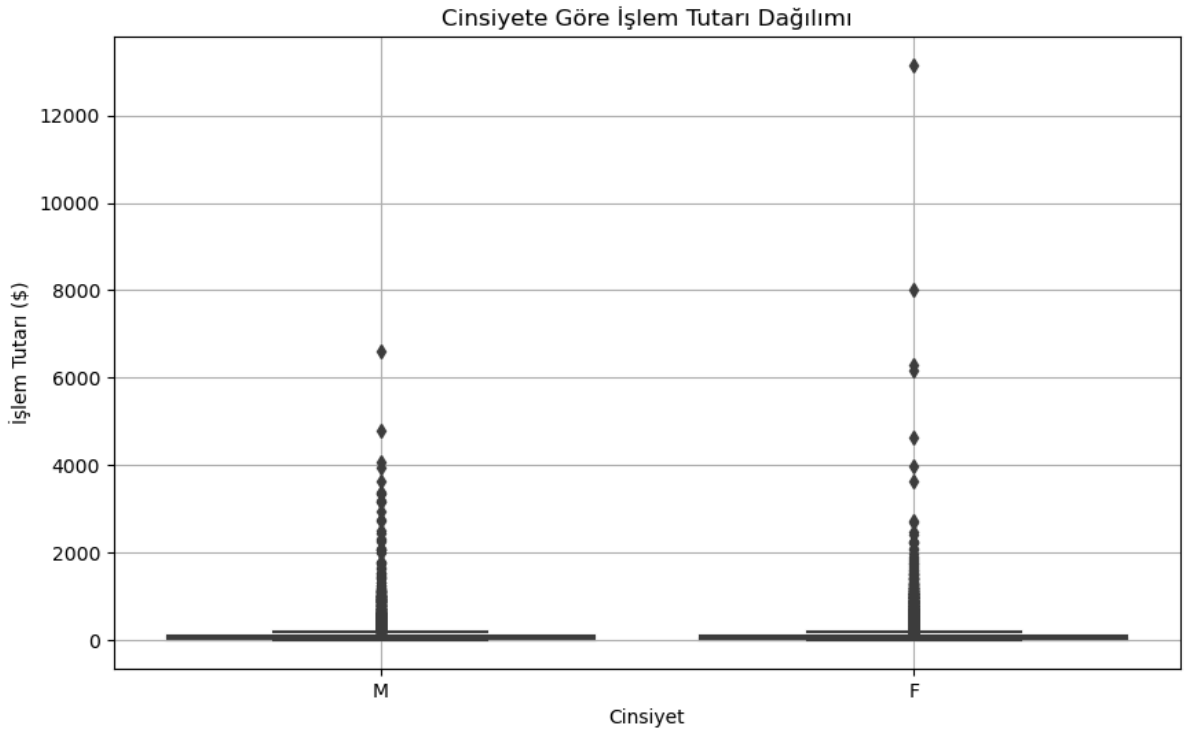
Kategorilere Göre Ortalama İşlem Tutarı

Kategorilere göre işlem tutarlarının ortalaması aşağıdaki şekilde görselleştirilmiştir:



Bu grafik, her kategori için ortalama işlem tutarını göstermektedir. Bazı kategorilerdeki işlemlerin diğerlerine göre daha yüksek tutarda olduğu gözlemlenmiştir.

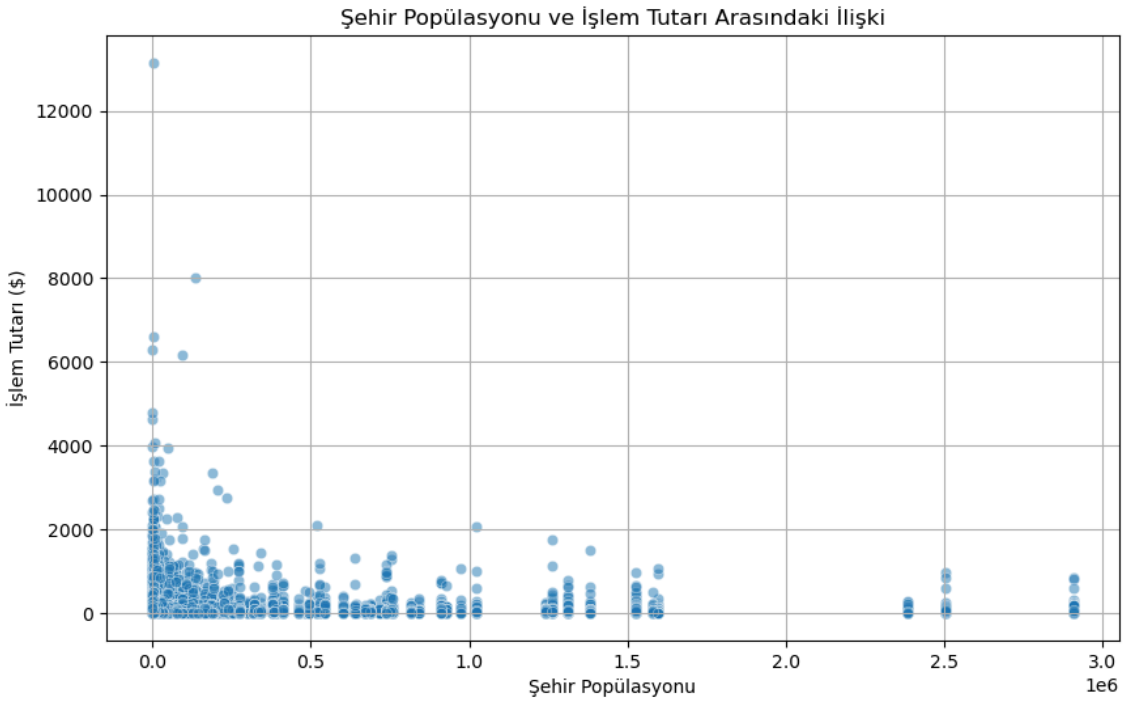
Cinsiyet ve İşlem Tutarı İlişkisi



Boxplot grafiği, kadınların ve erkeklerin işlem tutarlarını karşılaştırmaktadır. Her iki cinsiyet için de geniş bir işlem tutarı aralığı gözlemlenmektedir.

Şehir Popülasyonuna Göre Sahtekarlık Dağılımı

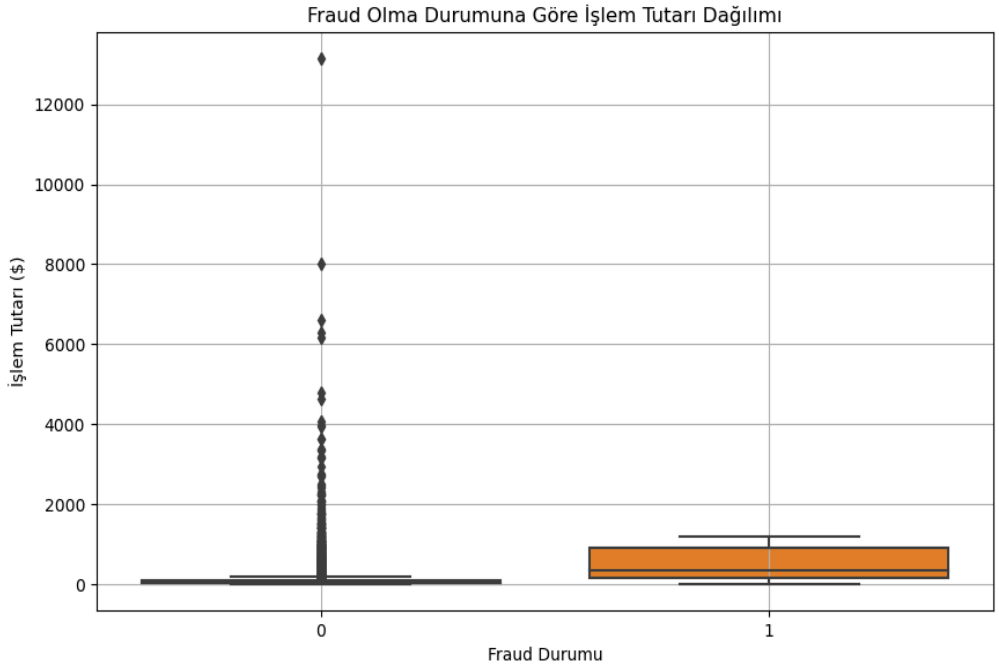
Şehir popülasyonuna göre sahtekarlık işlemlerinin dağılımı aşağıdaki gibidir:



Bu grafik, şehir popülasyonu ile sahtekarlık arasında bir ilişki olup olmadığını incelemek için kullanılmıştır.

Fraud Durumuna Göre İşlem Tutarı Dağılımı

Aşağıdaki grafik, `is_fraud` etiketiyle belirlenen sahtekarlık durumu (1: Fraud, 0: Non-Fraud) ile işlem tutarı arasındaki dağılımı göstermektedir.



Bu grafikte, sahtekarlık işlemlerinin ve sahtekarlık olmayan işlemlerin nasıl bir tutar aralığında yer aldığını ve bunların dağılımını gözlemleyebiliriz.

Makine Öğrenimi Modelleri ile Sahtekarlık Tespiti

Veri seti üzerinde çeşitli makine öğrenimi modelleri uygulanarak sahtekarlık tespiti yapılmıştır. Bu bölümde, uygulanan modellerin sonuçları detaylı olarak ele alınmıştır.

Model Hazırlığı ve Eğitim

Bağımlı ve Bağımsız Değişkenler

Veri setindeki bazı bağımsız değişkenler (özellikler) ve bağımlı değişken (hedef) seçilerek modelleme için hazırlanmıştır. Aşağıda kullanılan bağımsız değişkenler ve bağımlı değişken verilmiştir:

- **Bağımsız Değişkenler (X)**: `amt` (işlem tutarı), `gender` (cinsiyet), `city_pop` (şehrin nüfusu), `category` (işlem kategorisi)
- **Bağımlı Değişken (y)**: `is_fraud` (işlemin sahtekarlık olup olmadığını belirten etiket)

Niteliksel bağımsız değişkenler (`gender` ve `category`) sayısal hale getirilerek (encoding) modellemeye uygun hale getirilmiştir.

Veri Setinin Eğitim ve Test Setlerine Bölünmesi

Veri seti, model eğitimi ve değerlendirmesi için eğitim ve test setlerine ayrılmıştır. Eğitim seti, modelin öğrenmesi için kullanılırken, test seti modelin performansını değerlendirmek için kullanılmıştır.

Modelleme ve Değerlendirme

Aşağıda, veri seti üzerinde uygulanan çeşitli makine öğrenimi modelleri ve sonuçları verilmiştir.

Lojistik Regresyon Modeli

Lojistik regresyon modeli, sahtekarlık tespiti için kullanılan temel yöntemlerden biridir. Modelin sonuçları aşağıdaki gibidir:

Lojistik Regresyon Sonuçları:

Confusion Matrix:

[[19548 17]

[96 0]]

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19565
1	0.00	0.00	0.00	96
accuracy			0.99	19661
macro avg	0.50	0.50	0.50	19661
weighted avg	0.99	0.99	0.99	19661

ROC AUC Skoru: 0.7890

Lojistik regresyon modeli, doğruluk oranı yüksek olmasına rağmen sahtekarlık işlemleri (etiket 1) tespit etmede yetersiz kalmıştır.

Naive Bayes Modeli

Naive Bayes modeli, özellikle metin sınıflandırma ve sahtekarlık tespiti gibi problemlerde yaygın olarak kullanılan bir modeldir. Naive Bayes modelinin sonuçları:

Naive Bayes Sonuçları:

Confusion Matrix:

[[19403 162]

[57 39]]

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	19565
1	0.19	0.41	0.26	96
accuracy			0.99	19661
macro avg	0.60	0.70	0.63	19661
weighted avg	0.99	0.99	0.99	19661

ROC AUC Skoru: 0.7676

Naive Bayes modeli, sahtekarlık işlemlerini daha iyi tespit edebilmiştir ancak hala çok düşük bir hassasiyet (precision) ve f1 skoru sergilemektedir.

Karar Ağacı Modeli

Karar ağaçları, karar kurallarını ve veri örneklerini dallandırarak sınıflandırma yapar. Karar ağacı modelinin sonuçları:

Karar Ağacı Sonuçları (max_depth=5):

Confusion Matrix:

[[19553 12]

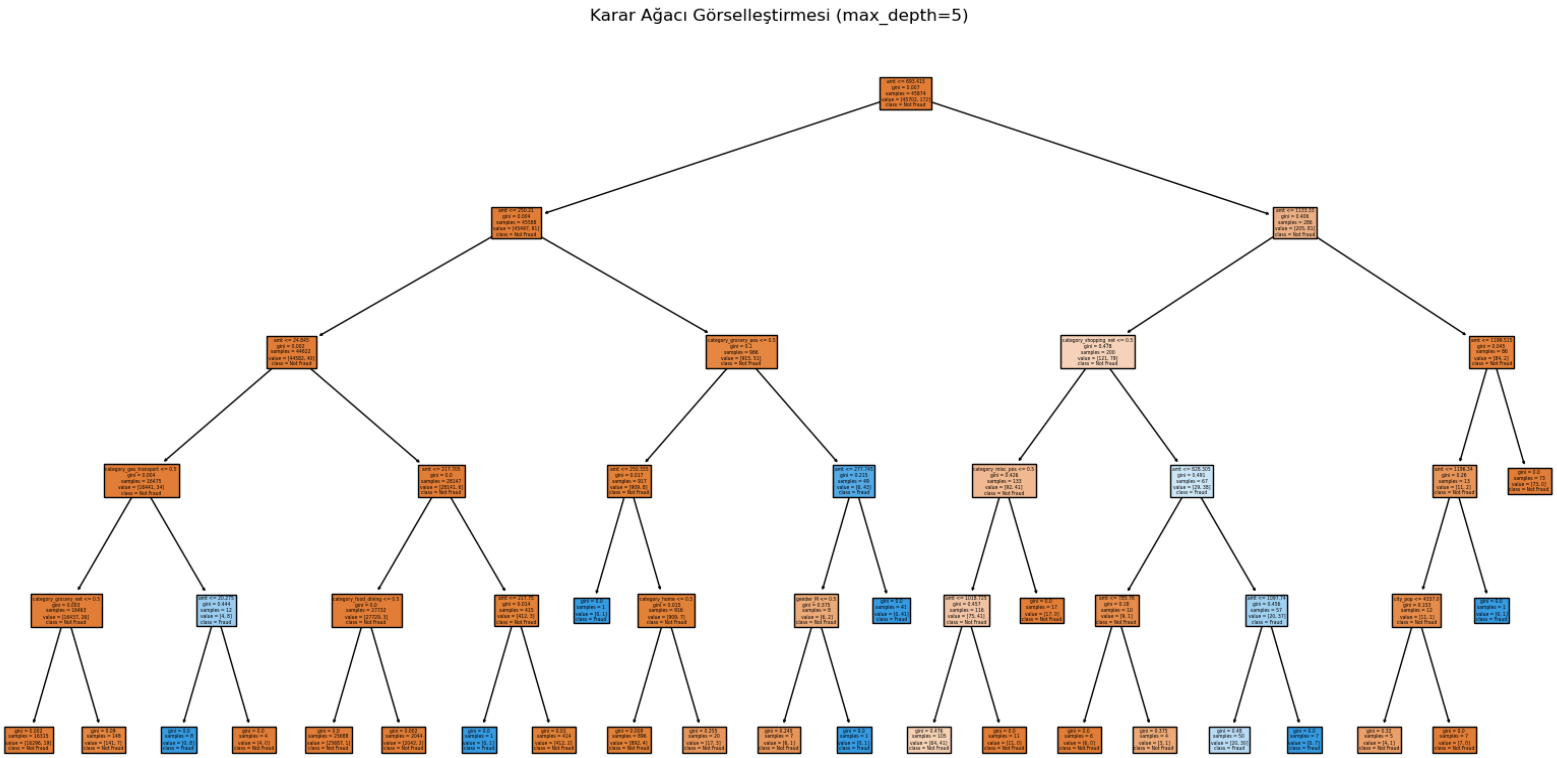
[59 37]]

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19565
1	0.76	0.39	0.51	96
accuracy			1.00	19661
macro avg	0.88	0.69	0.75	19661
weighted avg	1.00	1.00	1.00	19661

ROC AUC Skoru: 0.8267

Karar ağacı modeli, sahtekarlık işlemlerini tespit etmede lojistik regresyon ve Naive Bayes modellerine göre daha iyi performans göstermiştir. ROC AUC skoru, modelin genellikle iyi bir ayırıcı performansa sahip olduğunu göstermektedir.



Yapay Sinir Ağı Modeli (MLPClassifier)

Yapay sinir ağları, karmaşık veri ilişkilerini öğrenebilen güçlü modellerdir. Ancak, bu modelin performansı, diğer modellere göre daha düşük kalmıştır:

Yapay Sinir Ağı Sonuçları:

Confusion Matrix:

[[19565 0]

[96 0]]

Classification Report:

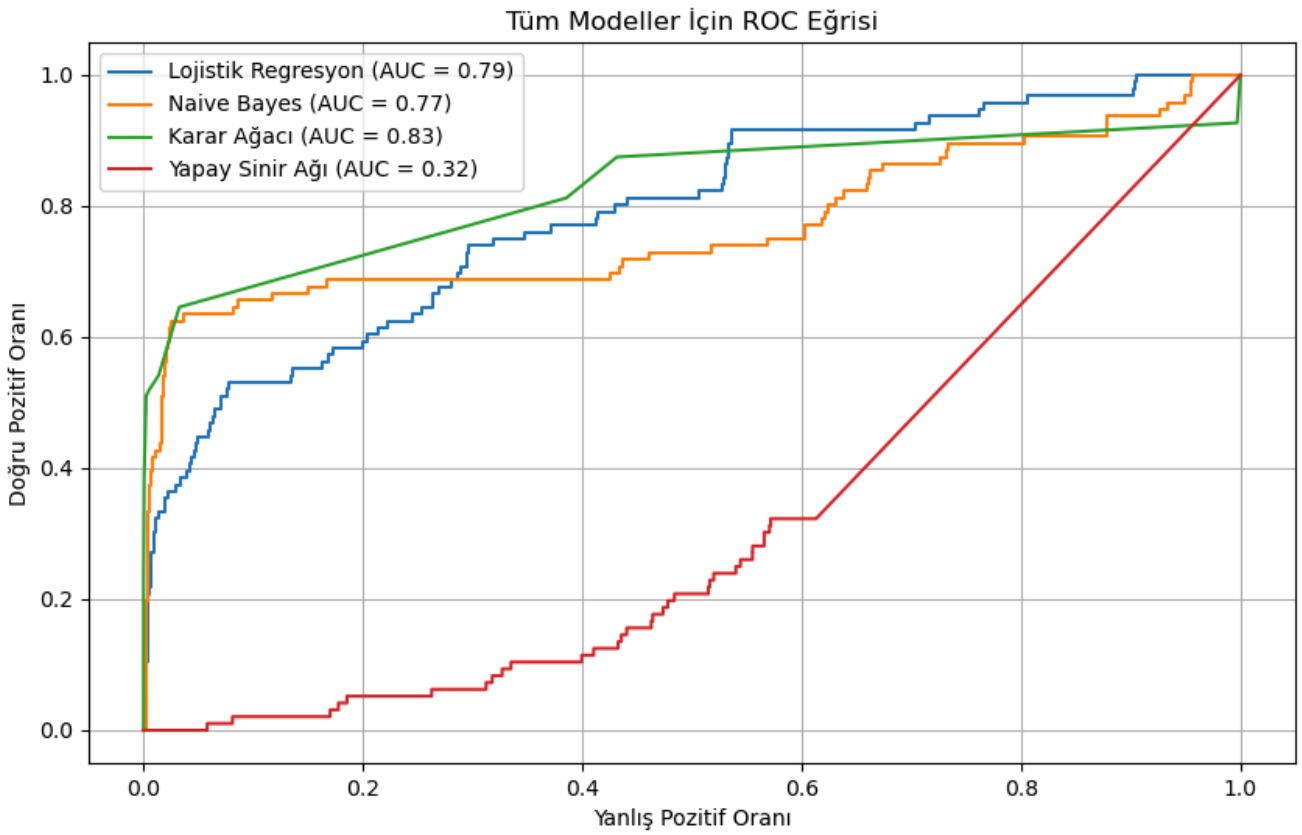
	precision	recall	f1-score	support
0	1.00	1.00	1.00	19565
1	0.00	0.00	0.00	96
accuracy			1.00	19661
macro avg	0.50	0.50	0.50	19661
weighted avg	0.99	1.00	0.99	19661

ROC AUC Skoru: 0.3213

Yapay sinir ağı modeli, sahtekarlık işlemlerini tespit etmede başarısız olmuştur. Bu, modelin öğrenme sürecinde veya veri dengesizliğinde bir problem olduğunu gösterebilir.

ROC Eğrileri

Tüm modeller için ROC eğrileri çizilmiş ve performansları karşılaştırılmıştır.



Sonuçlar

Verilerin İncelenmesi ve Temizlenmesi

Veri setinin incelenmesi ve temizlenmesi işlemleri, eksik değerlerin doldurulması ve gereksiz değişkenlerin çıkarılmasını içeriyordu. Bu adımlar, veri setinin modelleme için uygun hale getirilmesini sağladı.

Model Performansı

Çeşitli makine öğrenimi modelleri kullanılarak sahtekarlık tespiti yapılmıştır. Karar ağacı modeli, diğer modellere kıyasla sahtekarlık işlemlerini tespit etmede daha iyi performans göstermiştir. Lojistik regresyon ve Naive Bayes modelleri de iyi sonuçlar vermiş ancak sahtekarlık işlemlerini tespit etmede yeterince etkili olamamıştır. Yapay sinir ağı modeli, sahtekarlık işlemlerini tespit etmede başarısız olmuştur, bu da modelin veri dengesizliği veya eğitim sürecinde bir problem olabileceğini göstermektedir.

Kaynakça

1. Kaggle. (n.d.). *Veri bilimi ve makine öğrenimi yarışmaları, veri setleri ve topluluk kaynakları*. Retrieved from Kaggle
2. Towards Data Science. (n.d.). *Veri bilimi, makine öğrenimi ve yapay zeka üzerine makaleler*. Retrieved from Towards Data Science
3. Scikit-Learn Documentation. (n.d.). *Python'da makine öğrenimi için Scikit-Learn kütüphanesinin resmi belgeleri*. Retrieved from Scikit-Learn Documentation
4. Matplotlib Documentation. (n.d.). *Python'da veri görselleştirme için Matplotlib kütüphanesinin resmi belgeleri*. Retrieved from Matplotlib Documentation