Çevrimiçi Ödemelerde Sahtekarlığın Anlaşılması

Makine Öğrenmesi

Hakan SÖNMEZ
502231006

Problem

Tanımı

Günümüzde son kullanıcılar için çevrimiçi alışveriş git gide artan öneme sahip olmakla beraber bu satışı yapan firmalar için ödemelerde sahtekarlığın anlaşılması ve ödemenin o anda reddedilmesi firma için büyük önem arz etmektedir. Çünkü anlaşılmayan sahtekarlık firmanın iade ve iptal süreçleriyle uğraşması ve postalanan ürünün geri alınamaması ya da son kullanıcının zarar etmesi sonuçlarını doğurmaktadır. Bunların engellenmesi çevrimiçi ödeme platformları için büyük bir sorundur.

2020 yılında e-ticaret siteleri çevrimiçi ödemelerdeki sahtekarlıklar yüzünden 20 milyar dolardan fazla zarar etti. Bu sayı 2022 yılında 41 milyar dolar oldu ve 2023 yılında ise 48 milyar doları aşması beklenmektedir.
E-ticaret sitelerin bu sahtekarlıklarla klasik yordamlarla ya da insan gücüyle başa çıkması imkansız bir görevdir.

Çözümü

Günümüz dünyasında milyonlarca insanın artık çevrimiçi alışveriş yaptığı düşünülürse ve bu sayının hiç bir zaman azalmayıp her zaman artacağı da eklenirse bu problemin çözümü ancak yapay zeka tarafından yapılabileceği analşılır. Problemin çözümü için kaggle üzerinden bulunan veri setinde KNN, Naive Bayes, Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, Gradient Arttırma ve Karar Destek Sistemleri algoritmalarıyla modeller denenmiş ve sonuçları karşılaştırmalı olarak gösterilmiştir.

Literatür Taraması

Ulusal Tez Merkezi'nde "fraud" kelimesiyle yapılan aramada 106 adet Bilgisayar Mühendisliği ait tez bulunmaktadır. Bu veri setini doğrudan kullanan bir makale ya da tez bulunamamıştır. Ama PCA ile dönüştürülmüş creditcard.csv üzerine 2 adet tez bulunup incelenmiştir.

Four Classification Methods - LAYTH RAFEA HAZIM

Tezinde kredi kartında sahtekarlığın tespitini 4 adet sınıflandırma algoritmasıyla incelemiştir. Bulduğu sonuçlar ise NB 97.46%, SVM 95.04%, KNN 97.55% and RF 97.7%'dir.

Kredi Kartı Sahte İşlem Tespiti - Kazım SOYLU

Tezinde aynı veri setini kullanmış olup Derin öğrenme, Rastgele orman ve Yığınlar üzerinde çalışmıştır. Bu algoritmaların sonuçlarını grafiksel olarak paylaşmıştır. Kesinlik değerinde Rastgele orman daha iyi performans vermiş olup, sahte işlem tespitinde oranında ise Yığın daha başarılı olmuştur.

Not: Sınıflandırıcı yığınında temel öğrenici olarak derin öğrenme ve rastgele orman modelleri kullanılmış, meta öğrenici olarak da yine rastgele orman algoritması kullanılmıştır.

Veri Seti

Kaynağı

Kaggle üzerinden 3 adet veri seti bulunmuştur.

1. Online payments fraud detection
2. Credit card fraud detection PCA ile veriler gizlenmiştir.
3. Credit card fraud detection

Bu çalışmada 1. veri seti tercih edilmiştir. Kaggle'da bu veri setiyle 10 adet çalışma bulunmaktadır.

Sütunların Açıklaması

step: 1 adımın 1 saate eşit olduğu bir zaman birimini temsil eder.
type: İşlemin tipi
amount: İşlemin miktarı
nameOrig: Müşternin işleme başlaması
oldbalanceOrg: İşlemden önce bakiye
newbalanceOrig: İşlemden sonra bakiye
nameDest: İşlemin alıcısı
oldbalanceDest: İşlemden önce alıcının ilk bakiyesi
newbalanceDest: İşlemden sonra alıcının ilk bakiyesi
isFraud: Sahtecilik mi değil mi

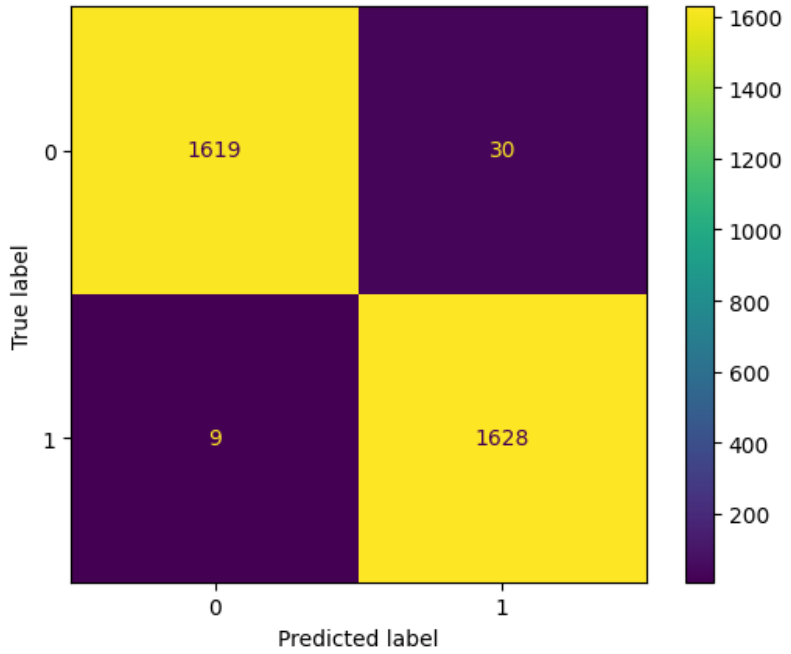
Modeller

Tüm modeller için hiperparametre optimizasyonları yapılmış ve bulunan en iyi değerler ile eğitim ve test süreci gerçekleştirilmiştir. Bunun için GridSearchCV metodu kullanılmıştır ve cross validation için 5 değeri belirlenmiştir.

KNN (K-Nearest Neighbors)

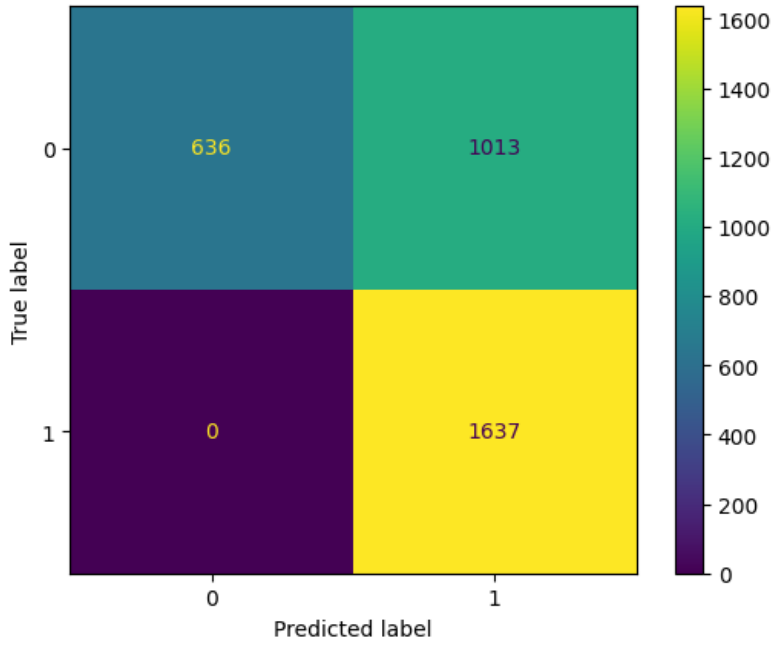
KNN en basit anlamı ile içerisinde tahmin edilecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfını tahmin etmeye dayanır.

En iyi parametreler n_neighbors: 1, p: 2’dir.
Training Time: 0.02
Test Time : 0.25
Accuracy 0.99
Recall 0.99
Precision 0.98
F1-Score 0.99



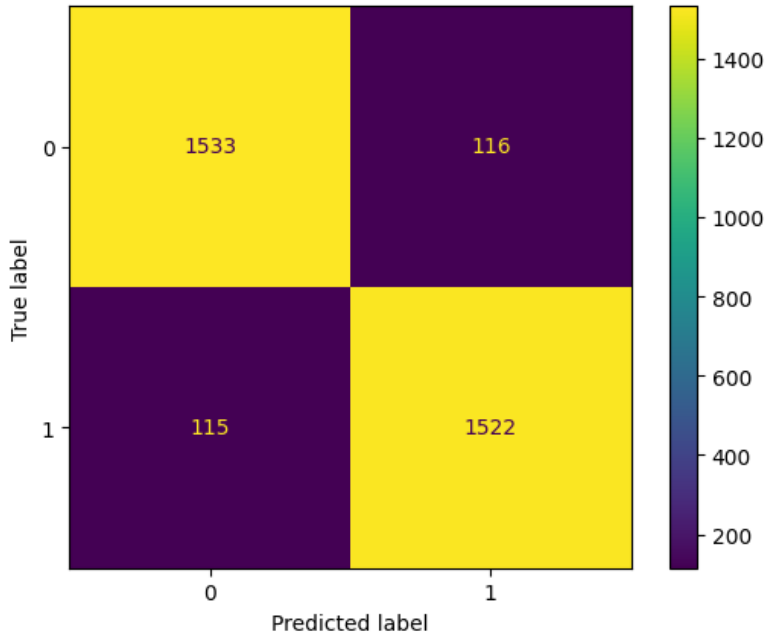
Naive Bayes (Gaussian Naive Bayes)

Naive Bayes sınıflandırıcısının temeli Bayes teoremine dayanır. Tembel bir öğrenme algoritmasıdır aynı zamanda dengesiz veri kümelerinde de çalışabilir. Algoritmanın çalışma şekli bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.
En iyi parametreler
var_smoothing: 3.5111917342151275e-06dir.
Training Time: 0.006
Test Time : 0.001
Accuracy 0.69
Recall 1.00
Precision 0.62
F1-Score 0.76



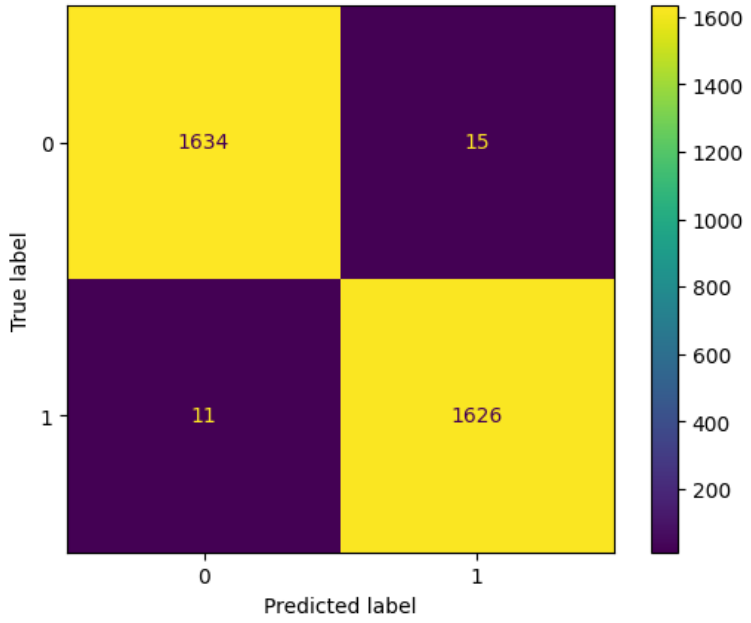
Logistic Regression

Lojistik regresyon analizi sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir. Ayırma (Diskriminant) analizi verilerin sınıflandırılması ve belirli olasılıklara göre belirli sınıflara atanmasını sağlayan bir yöntemdir.
En iyi parametreler
C: 100.0, penalty: l2, solver: liblinear
Training Time: 0.095
Test Time : 0.002
Accuracy 0.93
Recall 0.93
Precision 0.93
F1-Score 0.93



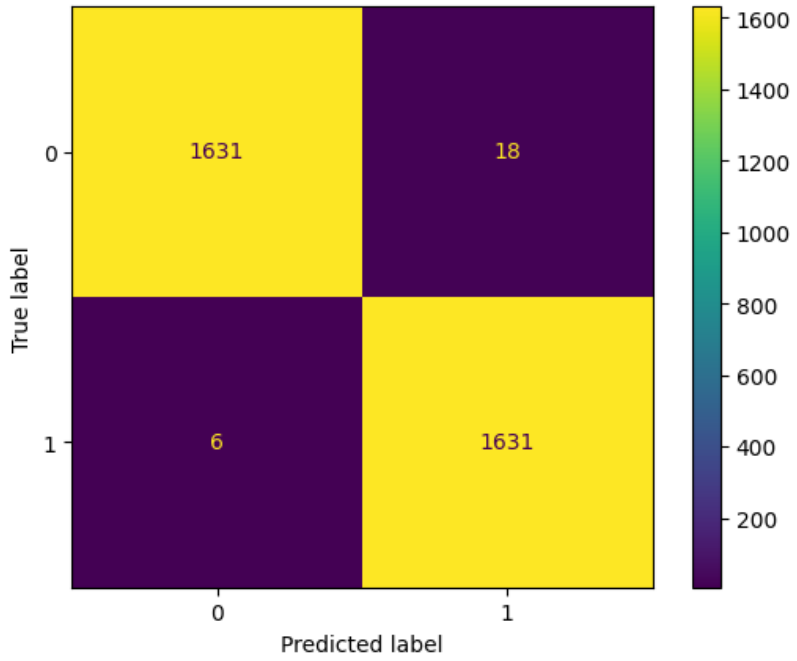
Decision Tree

Bir karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bir dizi karar kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır. Yani basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına bölerek kullanılan bir yapıdır.
En iyi parametreler
criterion: gini, max_depth: 10, min_samples_leaf: 5
Training Time: 0.084
Test Time : 0.005
Accuracy 0.99
Recall 0.99
Precision 0.99
F1-Score 0.99



Random Forest

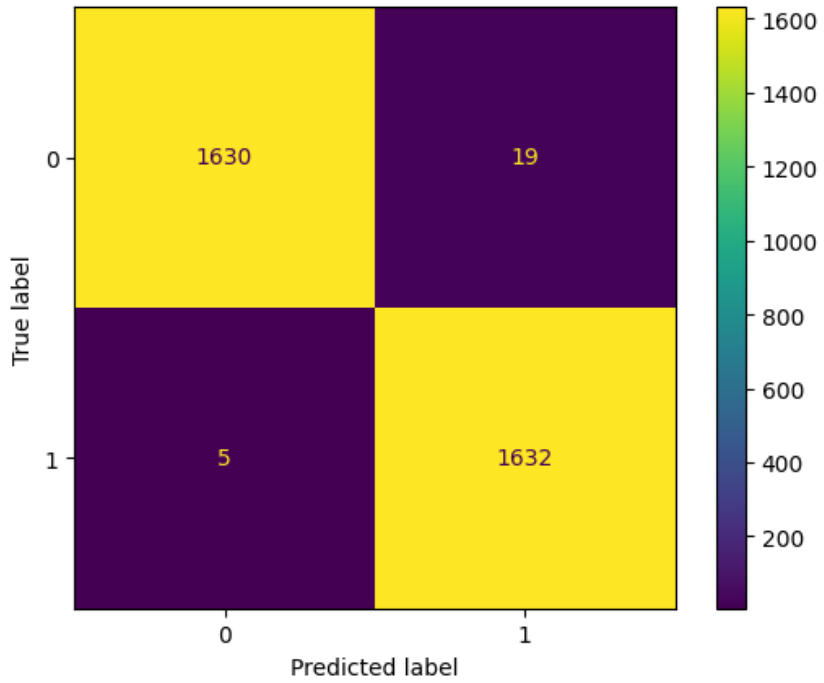
Karar ağaçlarının en büyük problemlerinden biri aşırı öğrenme-veriyi ezberlemedir (overfitting). Rassal orman modeli bu problemi çözmek için hem veri setinden hem de öznitelik setinden rassal olarak onlarca yüzlerce farklı alt-setler seçiyor ve bunları eğitiyor. Bu yöntemle yüzlerce karar ağacı oluşturuluyor ve her bir karar ağacı bireysel olarak tahminde bulunuyor. En çok oy alanı sonuç olarak veriyor.
En iyi parametreler
max_depth: 20, min_samples_leaf: 5, n_estimators: 200
Training Time: 7.278
Test Time : 0.097
Accuracy 0.99
Recall 1.00
Precision 0.99
F1-Score 0.99



Gradient Boosting

Boosting, zayıf öğrenicileri(weak learner) güçlü öğreniciye(strong learner) dönüştürme yöntemidir. Bunu iterasyonlar ile aşamalı olarak yapar. Boosting algoritmaları arasındaki fark genellikle zayıf öğrenicilerin eksikliğini nasıl tanımladıklarıdır.
Gradient Boosting'de öncelikli olarak ilk yaprak(initial leaf) oluşturulur. Sonrasında tahmin hataları göz önüne alınarak yeni ağaçlar oluşturulur. Bu durum karar verilen ağaç sayısına ya da modelden daha fazla gelişme kaydedilemeyinceye kadar devam eder.

En iyi parametreler
max_depth: 8, min_samples_leaf: 4, n_estimators: 100
Training Time: 4.5119
Test Time : 0.018
Accuracy 0.99
Recall 1.00
Precision 0.99
F1-Score 0.99



Sonuçların Karşılaştırılması

	Training Time	Test Time	Accuracy	Recall	Precision	F1-Score
GB	4.51e+00	1.80e-02	0.99	1.00	0.99	0.99
RF	7.28e+00	9.72e-02	0.99	1.00	0.99	0.99
DT	8.44e-02	5.83e-03	0.99	0.99	0.99	0.99
kNN	2.07e-02	2.52e-01	0.99	0.99	0.98	0.99
LR	9.59e-02	2.01e-03	0.93	0.93	0.93	0.93
NB	6.98e-03	1.10e-03	0.69	1.00	0.62	0.76

Kullanılan tüm algoritmaların en iyi optimize hallerinin karşılaştırılmasında Logistic Regression ve Naive Bayes hariç geri kalanlarının tamamının f1 Skorunun 0.99 olduğu görülmüştür. Bunların içerisinde en hızlı eğitim süresi tabi ki kNN'de daha sonra Decision Tree'de ve test süresi yine en hızlı Decision Tree'dir.