BÜYÜK VERİ FİNAL PROJESİ

Kısım-2 Kaggle Predict Future Sales

Onur Kaplan

*Özet—Bu çalışmada amaç Kaggle üzerindeki Predict Future Sales adlı yarışmada en büyük Rus yazılım firmalarından biri olan 1C Şirketi tarafından sağlanan günlük satış verilerinden oluşan oluşan zorlu bir zaman serisi verisi üzerinden istenilen aylık satış miktarını bir takım algoritmalar kullanılarak tahminin sağlanmasıdır.*

Anahtar Kelimeler—XGBoost, Kaggle, CatBoost, LightGBM, RandomForestRegressor

# GİRİŞ

Büyük bir Rus yazılım firması olan 1C Şirketi tarafından sağlanan günlük satış verilerinden oluşan zorlu bir zaman serisi veri kümesi üzerinden firmanın gelecek ay için her ürün ve mağaza için toplam satışların tahmin edilmeye çalışılmasıdır.

# VERİSETİ

## Veri Kümeleri

sales\_train.csv : Eğitim seti Ocak 2013-Ekim 2015 arasındaki geçmiş verileri içerir.

test.csv : Test seti Kasım 2015’deki geçmiş verileri içerir.

sample\_submission.csv : Gönderilecek tahmin dosyası için örnek bir taslak içerir.

items.csv : Ürünler hakkında ismi, kategori numarası gibi ek bilgiler içerir.

item\_categories.csv : Kategori ismi gibi ek bilgiler içerir.

shops.csv : Mağazalar hakkında mağaza ismi gibi ek bilgiler içerir.

## Veri Açıklamaları

ID : Test veri seti içindeki (mağaza,ürün) demetinin temsil bilgisini içerir.

shop\_id : Mağazanın benzersiz tanımlayıcısını içerir.

item\_id : Ürünün benzersiz tanımlayıcısını içerir.

item\_category\_id : Ürünün kategorisinin benzersiz tanımlayıcısını içerir.

item\_cnt\_day : Satılan ürün sayısının bilgisini içerir.

item\_price : Ürünün fiyat bilgisini içerir.

date : Tarih bilgisini içerir.

date\_block\_num : Date bilgisinin kaçıncı ayda olduğunu içerir.

item\_name : Ürünün isim bilgisini içerir.

shop\_name : Mağazanın isim bilgisini içerir.

item\_category\_name : Ürünün kategorisinin isim bilgisini içerir.

# ALGORİTMALAR

## XGBoost

XGBoost, yüksek verimli, esnek ve taşınabilir olacak şekilde tasarlanmış, optimize edilmiş dağıtılmış bir gradient boosting kütüphanesidir. Gradient Boosting çerçevesi altında makine öğrenme algoritmaları uygular. XGBoost, birçok veri bilimi problemini hızlı ve doğru bir şekilde çözen paralel bir ağaç güçlendirmesi (GBDT, GBM olarak da bilinir) sağlar. Aynı kod, büyük dağıtılmış ortamda (Hadoop, SGE, MPI) çalışır ve milyarlarca örneğin ötesinde sorunları çözebilir.[1]

## CatBoost

CatBoost, karar ağaçlarında gradient boosting için bir algoritmadır. Yandex araştırmacıları ve mühendisleri tarafından geliştirilen, şirket içinde görevleri sıralamak, tahmin etmek ve önerilerde bulunmak için yaygın olarak kullanılan MatrixNet algoritmasının halefidir. Evrenseldir ve çok çeşitli alanlarda ve çeşitli problemlere uygulanabilir.[2]

## LightGBM

Verilerin boyutu gün geçtikçe artmaktadır ve geleneksel veri bilimi algoritmalarının daha hızlı sonuç vermesi zorlaşmaktadır. LightGBM, yüksek hızı nedeniyle "Light" olarak eklenir. LightGBM büyük boyutlu verileri işleyebilir ve çalışması için daha az bellek alır. LightGBM'nin popüler olmasının bir başka nedeni, sonuçların doğruluğuna odaklanmasıdır. LGBM ayrıca GPU öğrenimini de destekler ve bu nedenle veri bilimcileri veri bilimi uygulama geliştirme için LGBM'yi yaygın olarak kullanırlar.[3]

## RandomForestRegressor

Rastgele Orman, hiperparametre ayarı yapmadan bile, çoğu zaman büyük bir sonuç üreten, esnek, kullanımı kolay bir makine öğrenmesi algoritmasıdır. Aynı zamanda en çok kullanılan algoritmalardan biridir, çünkü hem basitlik hem de sınıflandırma ve regresyon görevleri için kullanılabilir.[4]

# GELİŞTİRİLEN ÇÖZÜM

sales\_train ve test verilerinin ön işlemeleri yapıldıktan sonra sales\_train algoritmalara gönderilmek için eğitim ve test olarak bölünmüştir. Bölünen veri XGBoost, CatBoost, LightGBM, RandomForestRegressor algoritmalarında çalıştırıldıktan sonra en başarılı sonucu veren RandomForestRegressor gerçek test verisini tahmin etmek için kullanılmıştır.

## Yalancı Kod

Kütüphaneleri ekle

Verileri oku

Veri kümelerini incelemek için yazdır

Eksik verileri kontrol et

sales\_train verilerini incelemek için görselleştir

Görselleştirilen veriden outlier verileri kontrol et

Outlier verileri ve tekrar eden verileri temizle

sales\_train verilerini aylık olarak grupla

sales\_train’de gruplanmış verilerin fiyat ortalamasını hesapla

test verisini tahmin için hazırla

sales\_train verilerini X,Y şeklinde ayır

sales\_train verisini eğitim ve test olarak böl

Bölünen değerleri modeller üzerinde dene

XGBoost dene

CatBoost dene

LightGBM dene

RandomForestRegressor dene

Denenen modellerin başarılarını kıyasla

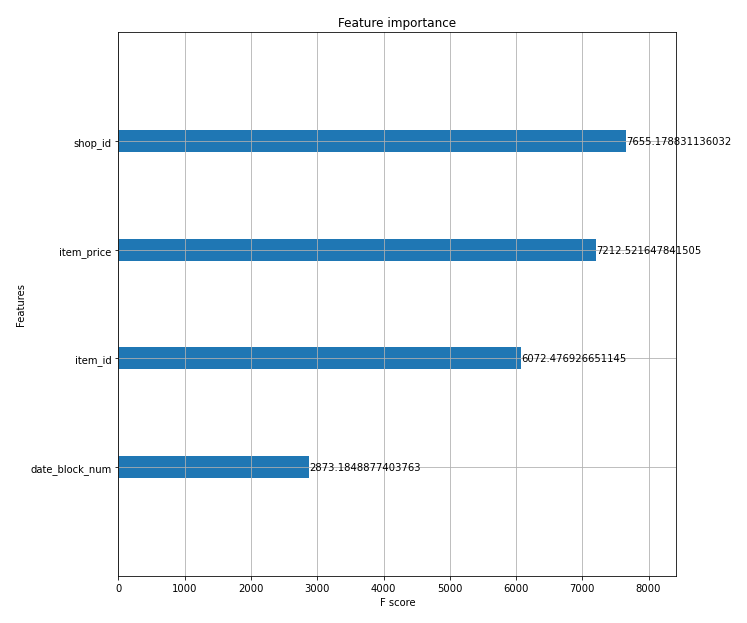
En başarılı model ile test verilerini tahmin et

Sonuçları submission.csv’ye kaydet

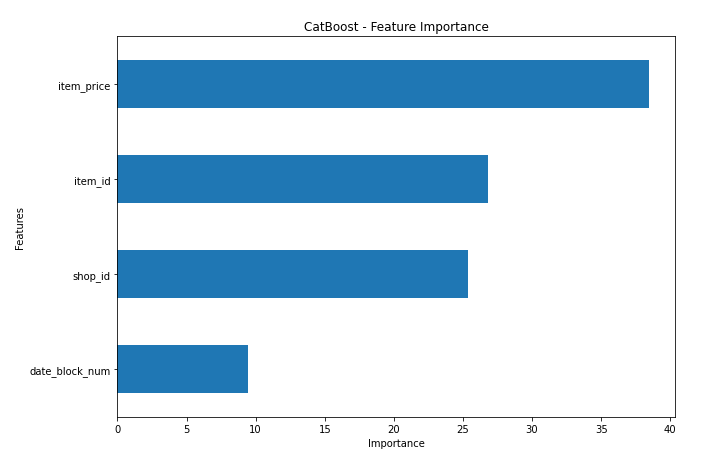
# SONUÇLAR

|  |  |  |
| --- | --- | --- |
|  | **RMSE** | **R2** |
| **XGBoost** | 5.35415 | 0.5644 |
| **CatBoost** | 5.24958 | 0.5813 |
| **LightGBM** | 0.56908 | 5.3257 |
| **RandomForestRegressor** | 0.70616 | 4.3977 |

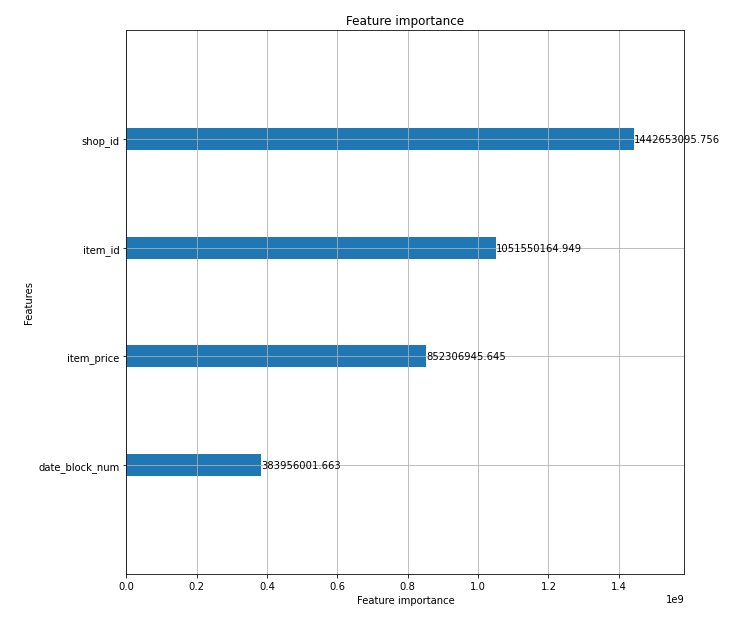
## XGBoost Feature Importance



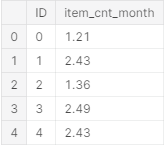
## CatBoost Feature Importance



## LightGBM Feature Importance



Yukarıdaki tabloda gözüktüğü gibi en başarılı R2 ve RMSE skorunu verdiği için RandomForestRegressor algoritması seçilmiştir. RandomForestRegressor kullanılarak tahmin edilen test çıktısıda aşağıda verilmiştir.



KAGGLE KERNEL LİNKİ : <https://www.kaggle.com/eddam123/kou-b-y-k-veri-final-160202061>

# KAYNAKÇA

1. Shubham Malik(2020), XGBoost: A Deep Dive Into Boosting*)*
2. <https://towardsdatascience.com/https-medium-com-talperetz24-mastering-the-new-generation-of-gradient-boosting-db04062a7ea2>,
3. <https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc>,
4. <https://devhunteryz.wordpress.com/2018/09/20/rastgele-ormanrandom-forest-algoritmasi/>