Veri Madenciliği
Ev Kirası Fiyat Tahmin

Kocaeli Üniversitesi
Teknoloji Fakültesi
Bilişim Sistemleri Mühendisliği
Serhat Kaçmaz - 181307036
serhatkacmaz3@gmail.com

I. Giriş

Günümüzde yapay zeka kavramı her geçen gün hayatımıza daha çok girmektedir. İlerleyen teknoloji ile yapay zekadan beklentiler de giderek artmıştır. Veri madenciliği ise büyük ölçekli veriler arasından faydalı olan bilgiye ulaşmaya çalışmaktır. Büyük veri ambarları veya yığınlarından gelecek için anlamlı ifadeler olan bilgilerin çıkarımların yapılması için kullanılır.

II. Proje

A. Konu

Proje konusunu detaylı anlatmadan önce genel olarak bakacak olursak. Bir nesnesin özelliklerine göre piyasadaki fiyatının tahmini veren bir uygulama geliştirilmesi. Burada fiyatının tahmin edilmesinin isteyeceğim nesne ise 'Ev'. İstanbul içerisinde bulunan evlerin özelliklerine göre fiyatının tahmini gerçekleştirme. Evin özellikleri olarak kaç yıllık olduğu kaçıncı kat, hangi semt, kaç odalı, eşyalı mı eşyasız mı, brüt veya net değeri ne gibi özelliklerini analiz ederek istenilen kriterdeki evin kirası tahmin edecektir.

B. Amac

Bazı ülkelerin nüfüsündan bile daha büyük olan İstanbul gibi bir büyük şehrin içerisinde milyonlarca insan olduğu kadar evler de bulunur. Her insan barınma ihtiyacını giderirken de barındığı veya barınacağı ortamın bazı özelliklere sahip olmasını ve kendisi için uygun bir fiyatda olmasını talep eder. Bundan dolayı bu proje eşliğinde kullanıcı talep ettiği ev için piyasadaki kira fiyatını görebilecektir.

III. Kullanilan Teknolijiler

A. Python

Veri setini oluşturmak için internetten bana gerekli olan bilgileri indirmem ve bilgisayarımda depolamam gerekiyor. Python programlama dili bu konuda en kolay ve verimli olarak benim işimi görücektir. Python, birçok amaç için kullanılabilen, dinamik bir söz dizime sahip, yaygın olarak kullanılan, nesne yönelimli ve üst

B. Selenium

Selenium test komut dosyaları Java, Python, C# ve daha pek çok farklı programlama dillerinde yazılabilir. Bu test komut dosyaları Chrome, Safari, Firefox, Opera gibi çeşitli tarayıcılarda çalışabilir. İlgili sayfalardan istediğim verileri kazımam da yardımcı olacaktır.

C. Jupyter Notebook

Birden çok programlama dilinde etkileşimli bilgi işlem için açık kaynaklı yazılım, açık standartlar ve hizmetler geliştirme hedefleri olan bir projedir.

Bir web tarayıcısı üzerinden notebook belgesi formatındaki kodları düzenlemeyi ve çalıştırmayı sağlayan bir sunucu-istemci uygulamasıdır. İlk çıktığında isim olarak IPython Notebook diye biliniyordu. Başlangıçta sadece Python'ı desteklesede zamanla gelişerek Julia, Octave, R, Haskel, Ruby gibi dilleri de desteklemeye başladı geliştirme hedefleri olan bir projedir.

IV. Veri Setinin Oluşturulması

Web kazıması için kendime belirli siteleri belirmem gerek, verileri kazıyabileceğim bazı web siteleri;

- sahibinden.com
- hepsiemlak.com
- emlakjet.com

Verileri çekerken site tarafından bot gibi algılanabiliriz bu sorunu gidermek için ise python da olan Time Modülünü kullanabiliriz. Projeyi geliştirirken sahibinden.com sitesi tarafından veri çekilmedi ancak veri çekilecek şekilde python üzerinden bot yazıldı. Hepsiemlak.com ve emlakjet.com sitelerinden veriler

çekildi ve cekilen veriler tek bir dataset içerisinde birleştirildi.

V. VERI KAZIMA AŞAMALARI

Verileri kazımaya başlamadan önce toplayacağım verileri hangi sitelerden toplayacağımı karar verdim. Bunun için siteleri araştırmaya başladım. Veri kazıma yaparken siteyi incelemek gerekiyor. Bunun içinde sitenin html kodlarını inceleyerek verileri düzenli çekebilmek için bir örüntü yakalamaya çalıştım.

- Kullanacağın bir tarayıcı için tarayacı özelinde bir web driver kurulumu.
- Yazdılan kod içersinde webdriver'ın yolunu verme.
- Yazılacak kod içersinde kullanılcak kütüphanelerin pip install ile kurulup import edilme işlemi.
- Toplanan verileri csv içersinde kaydetme.
- İki farklı siteden topladığım verileri tek csv haline merge etme.
- Merge edildikten sonra aynı olan kayıtlar duplicate edildi.

A. HepsiEmlak

HepsiEmlak sitesinde veri çekerken izlediğim adımlar.

- Webdriver siteyi açar.
- Döngü ile tüm ilanların olduğu sayfalar dolaşılır ve div içerisindeki linkler bir listede toplanır.
- Liste içindeki linklere teker teker webdriver ile gidip içerden veri alınır.
- Veri alındıktan sonra sonraki linke gidilir.
- Link içerisindeki veri toplanırken fiyat, ilçe ve diğer özellikleri ayrı şekilde topladım site tasarımından dolayı. Teknik ozellikler table içinde iken, fiyat ve ilce farkli bileşenler içindeydi.

B. EmlakJet

EmlakJet sitesinde veri çekerken izlediğim adımlar.

- Webdriver siteyi açar.
- Sayfa içersindeki ilk ilana tıklar.
- İlan içeriğindeki bilgileri toplar.
- Geri gelir onceki sayfaya.
- Sonrasında birsonraki ilana tıklar.
- Sayfa içerinde tüm ilanlar toplandıktan sonra diğer sayfalara geçiş yapar.
- Link içerisindeki veri toplanırken fiyat, ilçe ve diğer özellikleri ayrı şekilde topladım site tasarımından dolayı. Teknik ozellikler table içinde iken, fiyat ve ilce farkli bileşenler içindeydi.

VI. VERİ ÇEKERKEN ZORLUK VEREN ADIMLAR VE ÇÖZÜMLERİ

A. Sorunlar

- Verilerin tutarlı gelmemesi.
- Bazı kayıtları çekmeye çalışırken patlaması.
- Veri Çekerken bilgisayar ekranı kapanırsa veri çekmede sorun oluşması.
- Web siteleri tarafından bot olarak algılanıp engel yedikten sonra uzun süre erişim engeli yemek.
- Sayfa giderken reklam çıkması ve ondan dolayı patlaması
- Div aralarında reklam olması

B. Çözüm

Kod içerisinde js kodu çalıştırıp ekranda birden belircek olan reklamı engleyebiliriz. Div aralaırnda reklamlarda belirili örüntü yakalanıp çözülebilir. Bot olarak algılanmadan kurtulmak için belirli farklı sürelerde istek yapma ve bekleme ile giderilebilir. Veri çekilmeyen linkler exception ile yakalanılıp tekrar indirilir.

VII. ÖN İŞLEME VE TEMİZLEME

- İki farklı siteden topladığım verileri merge ettikten sonra aynı olan verilerden dolayı duplicate ettim.
- Her bir özelliğimde bulunan verilere baktım neler var, en çok ne var şeklinde analiz ettim.
- Her bir özelliği teker teker inceledim çok aşarı saçma tutarsız olan verileri replace ettirdim.
- Çok fazla tutarsız özelliği olan kayıtları uçurdum.
- Mesala Oda sayısı 2+1 varken, bir de 2 oda 1 salon şeklinde olan kayıtlar bulunuyordu bu sadece tek bir örneği bunları sabit ortak bir değerle güncelledim.
- Veriler içersindeki boşlukarı trim mantığında uçurdum.
- M2 m^2 gibi gereksiz alanları kaldırdım.
- Brüt değeri net değerinden kücük olan kayıtlar vardı bunları düzenledim.

Tüm sütünlarım için yukarda ki maddeleri uyguladım.

VIII. NUMERİC YAPMA

Preprocessing kütüphanesini import edip topladığım verileri numeric hale getirmek için bu kütüphaneyi kullandım. labelEncoder üzerinden transform sayesinde içersine verdiğim sütünü numeric hale getirip bana geri veriyordu. Fiyat dışındaki tüm özelliklerimi numeric hale getirdim. Daha sonrasında csv olarak son halini kaydettim.

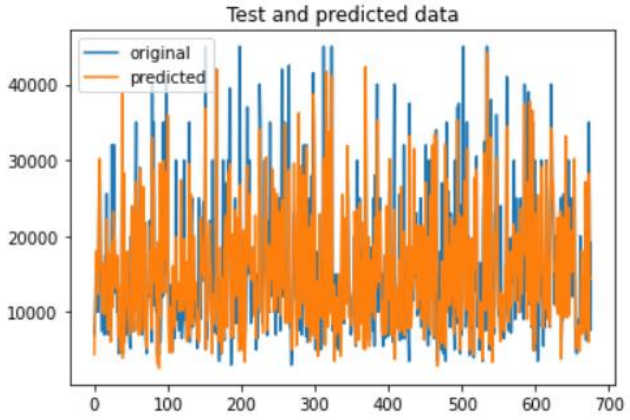
IX. MODEL EĞİTME

Ev kira fiyat tahmini projesi bağımlı bağımsız değişkenleri içerisinde barındırdığı için kullanıcağım modelleri regrasyon olanlardan tercih ettim. Model eğitimlerini yaparken kullandığım modeller XGB regression, KNN regression, Gradient Boosting regression, Random Forest Regressor, LGBM regression, DecisionTree modellerini kullandım.

A. *XGB Regression*

XGBoost , fonksiyon uzayında gradyan iniş olarak çalışan gradyan artırmanın aksine, fonksiyon uzayında Newton-Raphson olarak çalışır , Newton Raphson yöntemiyle bağlantı kurmak için kayıp fonksiyonunda ikinci dereceden bir Taylor yaklaşımı kullanılır.

```
Training score:  0.893228466423388
MSE: 24330347.30
RMSE: 4932.58
R2:  0.7170652744832102
MAE:  3408.503613714636
MAPE:  3408.503613714636
```

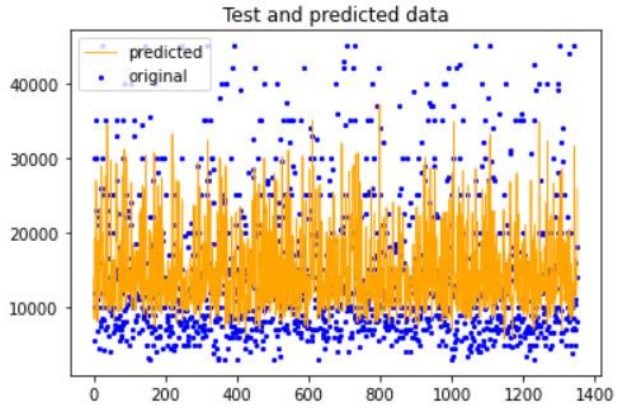


Şekil 1 XGB Regression

B. *KNN Regressor*

En iyi k seçimi verilere bağlıdır; genel olarak, daha büyük k değerleri , gürültünün sınıflandırma üzerindeki etkisini azaltır [8] , ancak sınıflar arasındaki sınırları daha az belirgin hale getirir. İyi bir k , çeşitli buluşsal tekniklerle seçilebilir.

```
Training score:  0.4055968674593031
MSE: 59493789.27
RMSE: 7713.22
R2:  0.31470813392561003
MAE:  5822.969181459566
MAPE:  5822.969181459566
```

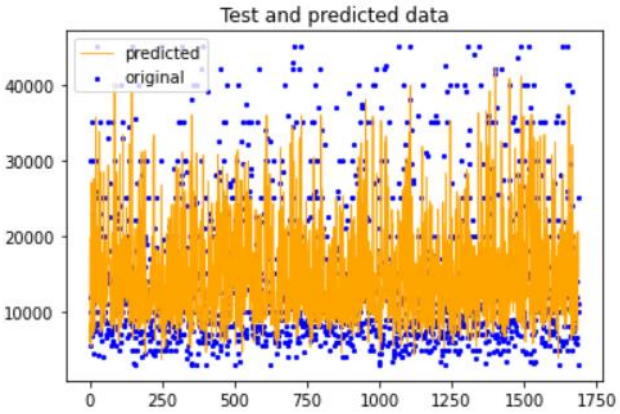


Şekil 2 KNN Regression

C. *Gradient Boosting Regressor*

Gradient boosting regressor , diğerleri arasında regresyon ve sınıflandırma görevlerinde kullanılan bir makine öğrenimi tekniğidir . Tipik olarak karar ağaçları olan zayıf tahmin modelleri topluluğu şeklinde bir tahmin modeli verir.

```
Training score:  0.7159646611950126
MSE: 33340149.21
RMSE: 5774.09
R2:  0.6329507291365148
MAE:  4131.199215972948
MAPE:  4131.199215972948
```

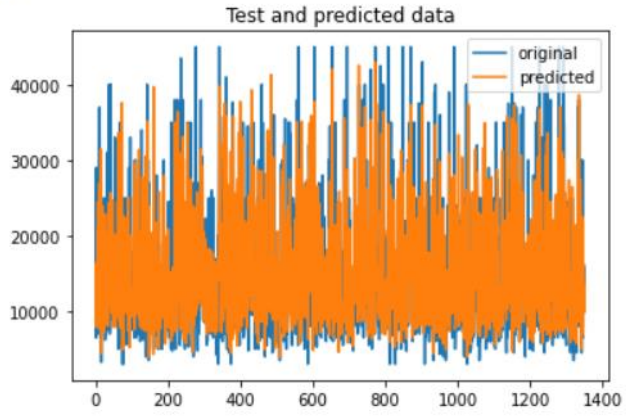


Şekil 3 Gradient Boosting Regression

D. *Random Forest Regresser*

Random forest regresyon birden fazla karar ağacını kullanarak daha uyumlu modeller üreterek isabetli tahminlerde bulunmaya yarayan bir regresyon modelidir. Karar ağaçlarını kullandığı için kesiklidir. Yani belli bir aralıkta istenen tahminler için aynı sonuçları üretir.

```
Training score:  0.9424146528746122
MSE: 31013025.35
RMSE: 5568.93
R2:  0.6419245895142396
MAE:  3930.6095150465894
MAPE:  3930.6095150465894
```

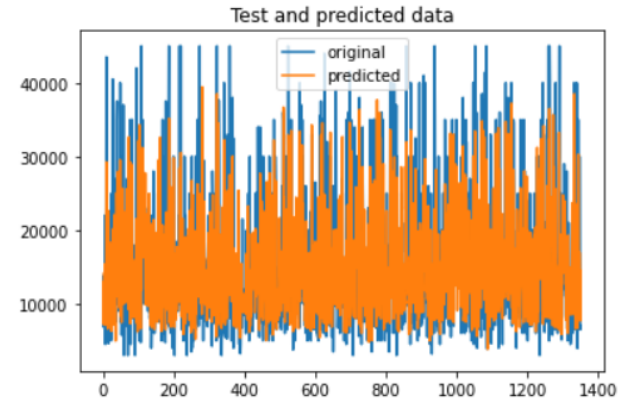


Şekil 4 KNN Random Forest Regression

E. LGBM Regression

Light GBM çerçevesi, GBT, GBDT , GBRT , GBM, MART ve RF gibi farklı algoritmaları destekler . Light GBM, XGBt'un seyrek optimizasyon, paralel eğitim, çoklu kayıp fonksiyonları, düzenli hale getirme, paketleme ve erken durdurma gibi birçok avantajına sahiptir. İkisi arasındaki en büyük fark, ağaçların yapımında yatmaktadır. LightGBM, diğer çoğu uygulamanın yaptığı gibi, ağaç düzeyinde - satır satır - büyümez. Onun yerine yaprak döken ağaçlar yetişir. Kayıpta en fazla azalmayı sağlayacağına inandığı yaprağı seçer.

```
MSE: 33895750.89
RMSE: 5822.01
R2:  0.6195064149065161
MAE:  4150.669672389584
MAPE:  4150.669672389584
```



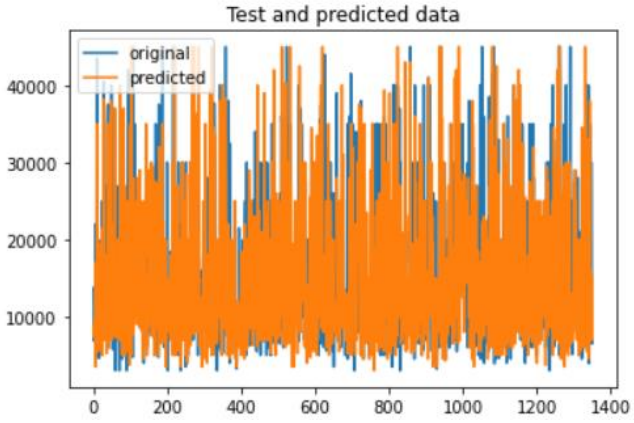
Şekil 5 LGBM Regression

F. DecisionTree

Karar ağacı öğrenimi , istatistik , veri madenciliği ve makine öğreniminde kullanılan denetimli bir öğrenme yaklaşımıdır . Bu biçimcilikte, bir dizi gözlem hakkında sonuçlar çıkarmak için bir tahmin modeli olarak bir sınıflandırma veya regresyon karar ağacı kullanılır.

Hedef değişkenin ayrı bir değerler kümesi alabildiği ağaç modellerine sınıflandırma ağaçları denir ; bu ağaç yapılarında, yapraklar sınıf etiketlerini temsil eder.

```
Training score:  0.9989837924535885
MSE: 64656805.47
RMSE: 8040.95
R2:  0.27420107047717956
MAE:  5225.670857988166
MAPE:  5225.670857988166
```



Şekil 6 DecisionTree

X. REFERANSLAR

- https://www.selenium.dev/
- https://www.youtube.com/watch?v=A1nJ66Vmkaw
- https://www.youtube.com/watch?v=tbXuI9HzNsQ&t=536s
- https://www.youtube.com/watch?v=CpeoWX_jyxw
- https://www.datatechnotes.com/