TÜRKİYE CUMHURİYETİ
YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

YAPAY ZEKAYA GİRİŞ
2.ÖDEV

Öğrenci Adı: Ali Albayrak
Öğrenci NO: 20011910

Öğrenci Adı: Mahmut Alomeyr
Öğrenci NO: 19011920

Ders Yürütücüsü
Prof.Dr. Mehmet Fatih AMASYALI

MAYIS, 2023

İçindekiler

● Ödev Anlatımı

Bizim ödev bir Websiteden alınan veri kümesi ile bir apartman fıyat Tahmin (Regression) modeli oluşturmaya yönelik bir ödevdir.

● Çözüm Genel Anlatımı

kullanıdığmız programlama dili: Python

çözümümüz 3 aşamalıydı

1- verileri siteden çekme

seçtiğimiz site: emlakjet.com

kullanılan kütüphaneler: rquests, BeautifulSoup

bu aşama sonunda 2 tane dosya bulunur biri verileri çekilmiş dairelerin linkleri diğeri ise her daireye ait özellikleri ham şekliyle içeren dosya

2- verileri önişleme ve temizleme

burda ilgisiz özellikleri silindi ('İlan Numarası', 'Görüntülü Gezilebilir mi?', …vb)

sonra kalan özellikleri tek tek temizlendi (sonraki kısımda anlatılacak)

3- modeleri oluşturup eğitme

burda 2 tane taminleyici 4'er farklı işleme derecesine sahip veri kümeleri kullanarak eğitildi

modeller: Linear Regression, Random Forest Regressor

işleme dereceleri: normal temizlenen, outlier değerlerin çıkarılan, normalizasyon işlemi gören, PCA işlemi gören veriler

● Veri Kümesini Nasıl Oluşturuldu

veri kümesini oluşturma sırasında 3 aşamadan geçildi.

ilk aşama

veri ilk önce emlakjet bahçelievlerdeki kiralık dairelerin linkleri çeklip bir dosyaya yazıldı (apt_links.csv)

sonra o linklere tek tek girip dairenin tüm özellikleri çekilip ayrı bir dosayaya yazıldı (apt_raw_data.csv)

bu aşma sonunda örnek sayısı 240 tanedir

bu aşamada scraping kısmı bitmiş olup veri kullanabilmek için hazırlama aşamasına geçilir

ikinci aşama

burda kullanılmaycak sütünler silindi (toplam 12 özellik kalıyor)

sonra her sütündeki veriler aynı şekilde göstermek için işlemler yapıldı

yani NaN diğerleri defult bir değer olarak atanır ve yazı içeren özellikler sadece sayıya çevirildi

burda yeni bir veri dosyamız oluşmuş oluyor (apt_pre_data.csv)

bu aşma sonunda örnek sayısı 240 tanedir

üçüncü aşama

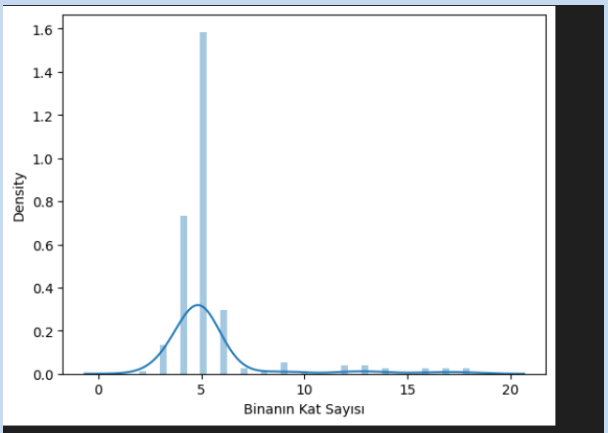
en son verinin oulier değerleri silip ayrı ve son veri dosyamızı oluşturyoruz (apt_cleaned_data.csv)

bu aşma sonunda örnek sayısı 218 tanedir

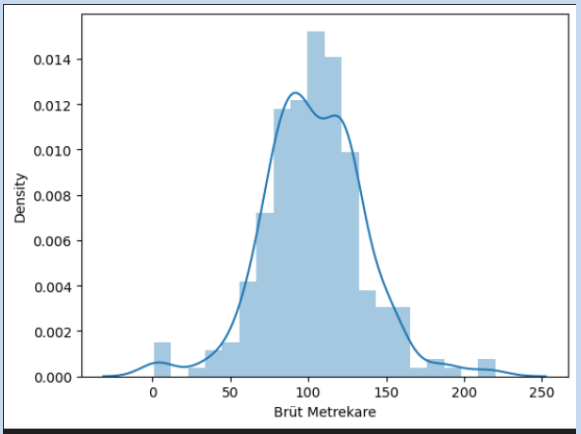
özellik sayısı: 12 özelliktir

- ## Tablo ve Grafikler

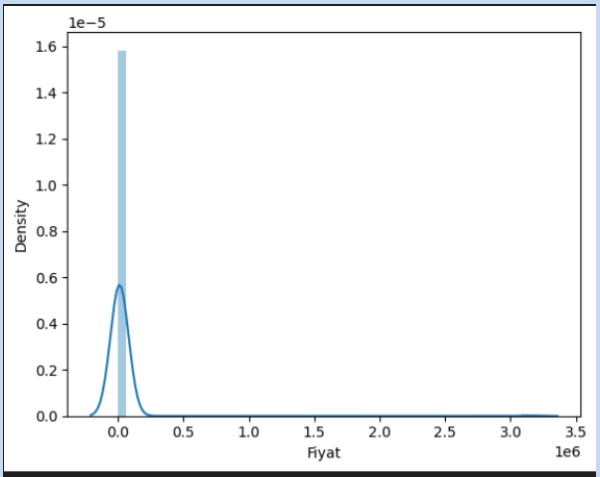
outlier değerleri kaldırmakta dağlımı gösteren grafikler



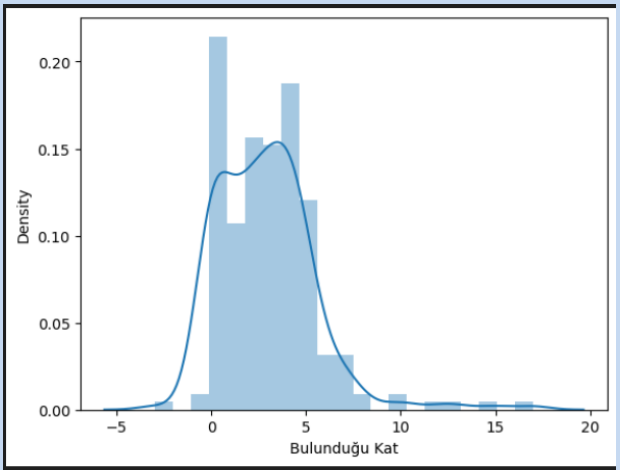

```
Binanın Kat Sayısı => lower:  -2.3920397269790623 => upper:  13.347791939368443
prev count:  226
new count:  218
```



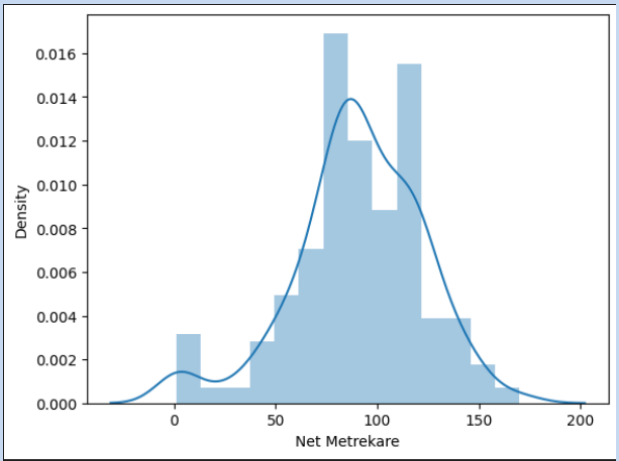

```
Brüt Metrekare => lower:  6.828350549986666 => upper:  200.58914945001334
prev count:  240
new count:  235
```



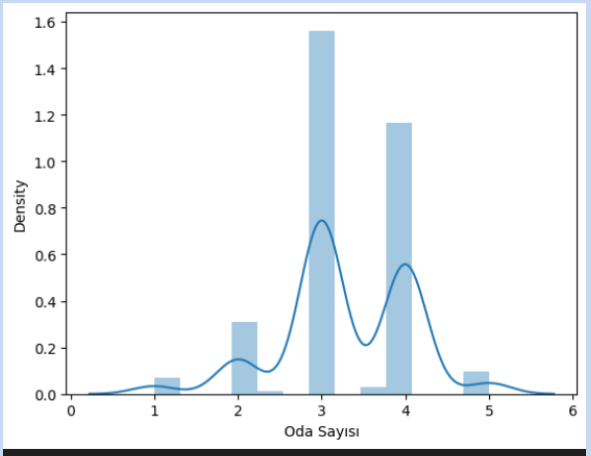

```
Fiyat => lower:  -181812.56796501382 => upper:  232197.76796501383
prev count:  230
new count:  229
```



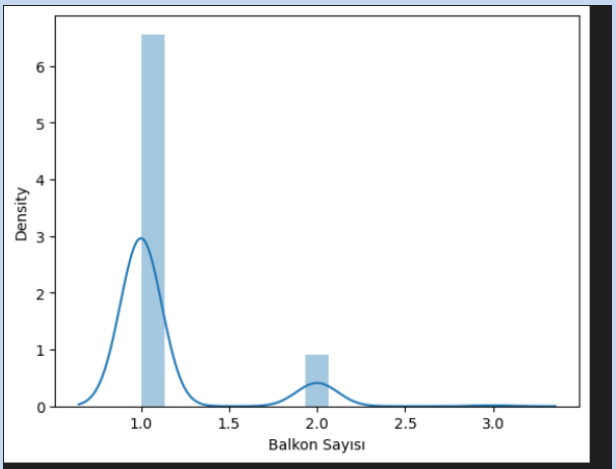

```
Bulunduğu Kat => lower:  -5.013038171137051 => upper:  10.783250937094499
prev count:  235
new count:  231
```



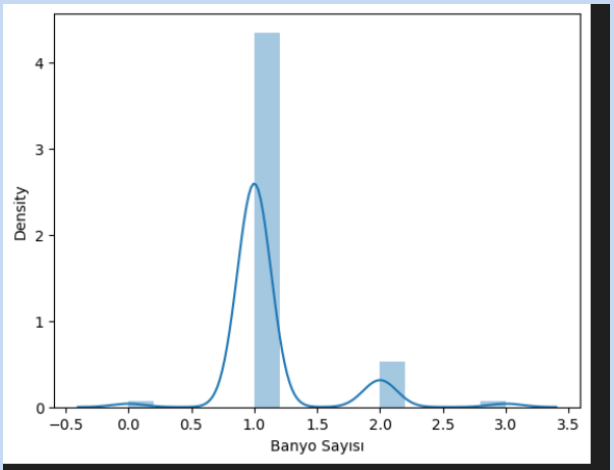

```
Net Metrekare => lower:  -4.949714764309178 => upper:  186.35397008345814
prev count:  235
new count:  235
```




```
Oda Sayısı => lower:  0.9836467585644835 => upper:  5.583452808535084
prev count:  231
new count:  231
```




```
Balkon Sayısı => lower:  0.08119588415284484 => upper:  2.178544375587415
prev count:  231
new count:  230
```




```
Banyo Sayısı => lower:  -0.07113472213009997 => upper:  2.3069425823921086
prev count:  229
new count:  226
```

tahminleyicilerin ortalma performasını gösteren puanlar:

```
score avgs:  [-154.75971450257563, 0.04107929934086539, 0.04107929934086506, 0.041079299340866, -353.3767515985369, -0.06532672069693952, -0.02997648238082822, -0.131303420871914]
```

puanların sırayla gösterdiği değerler

linear sadece veri

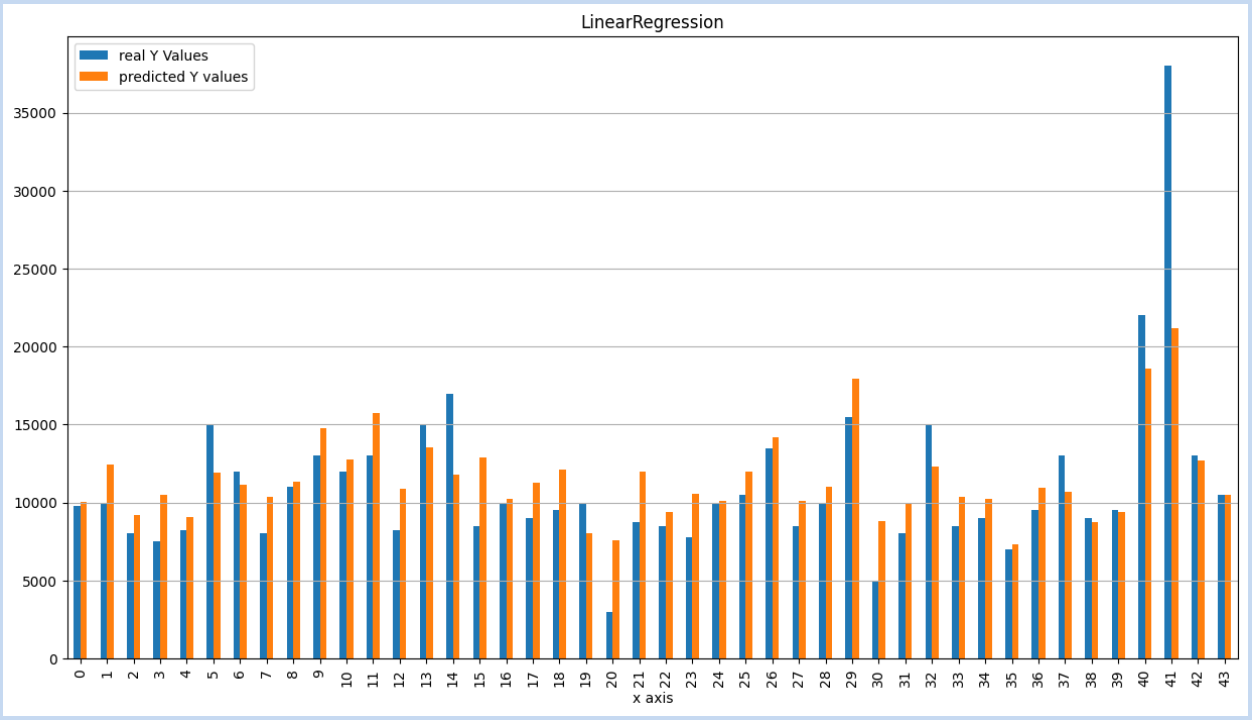
linear outlierleri silnmiş veri

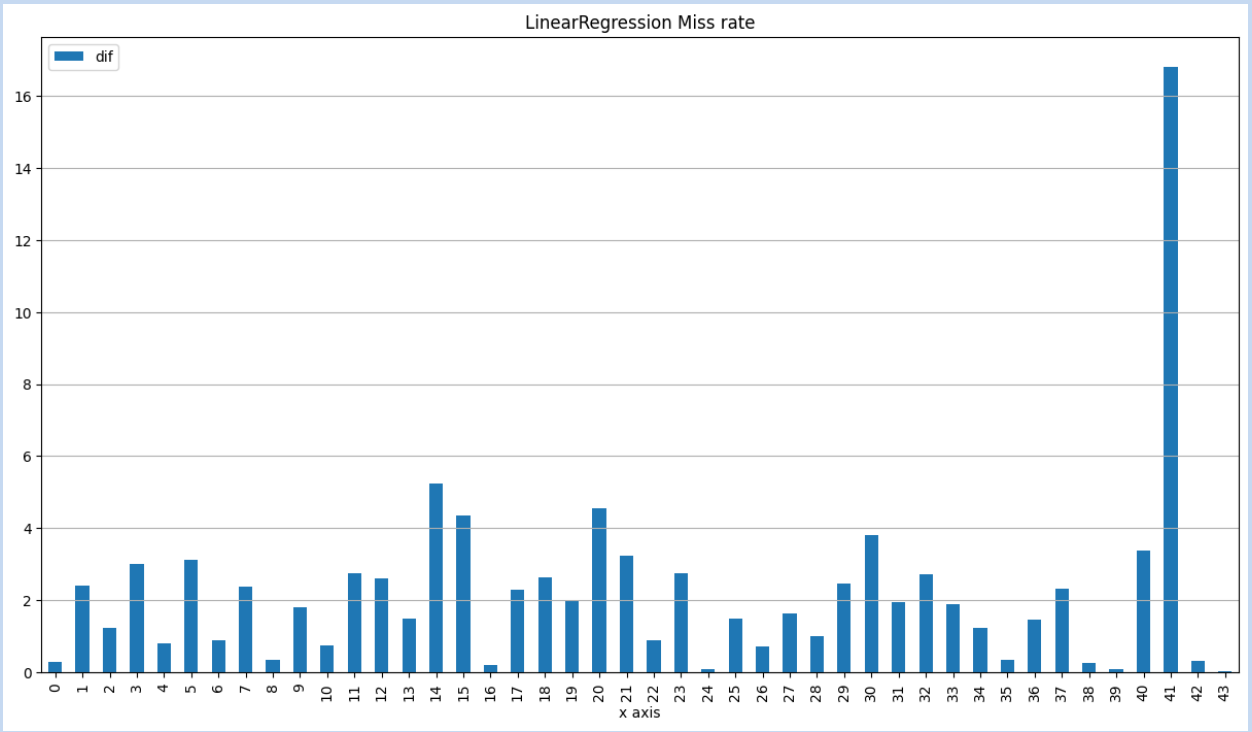
linear outlierleri silnmiş + normalizasyon uygulanan veri

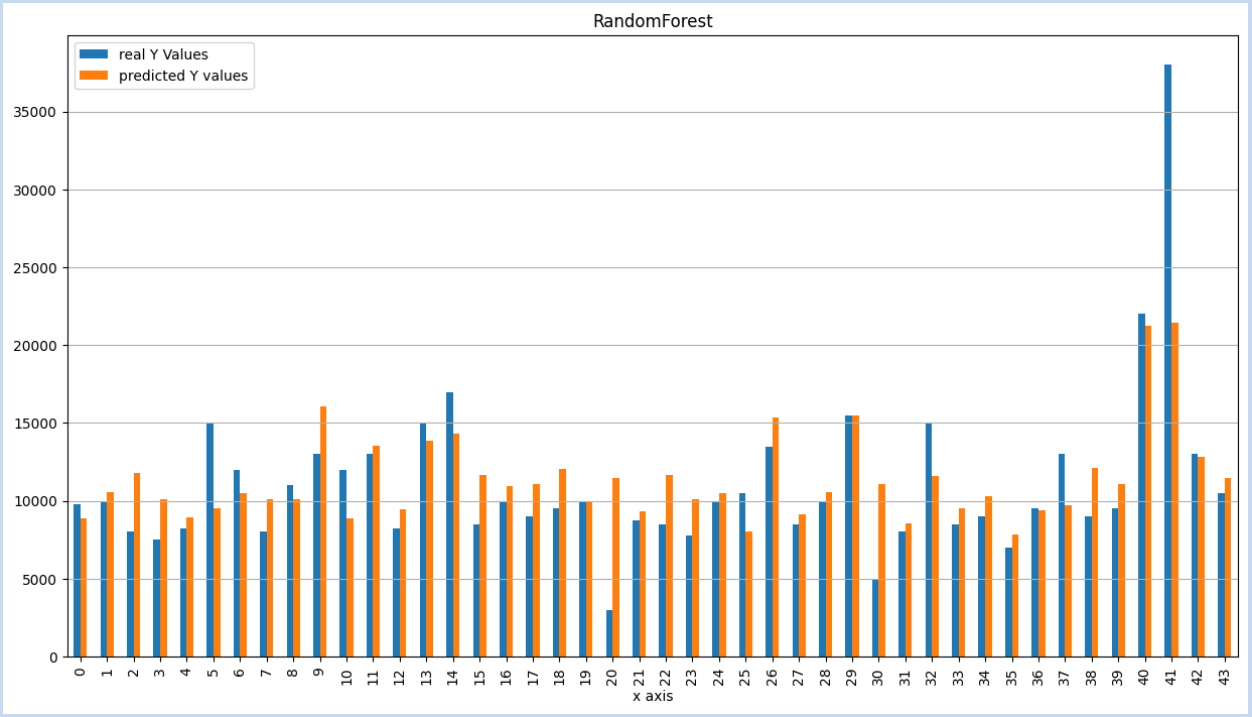
linear outlierleri silnmiş + normalizasyon ve PCA uygulanan veri

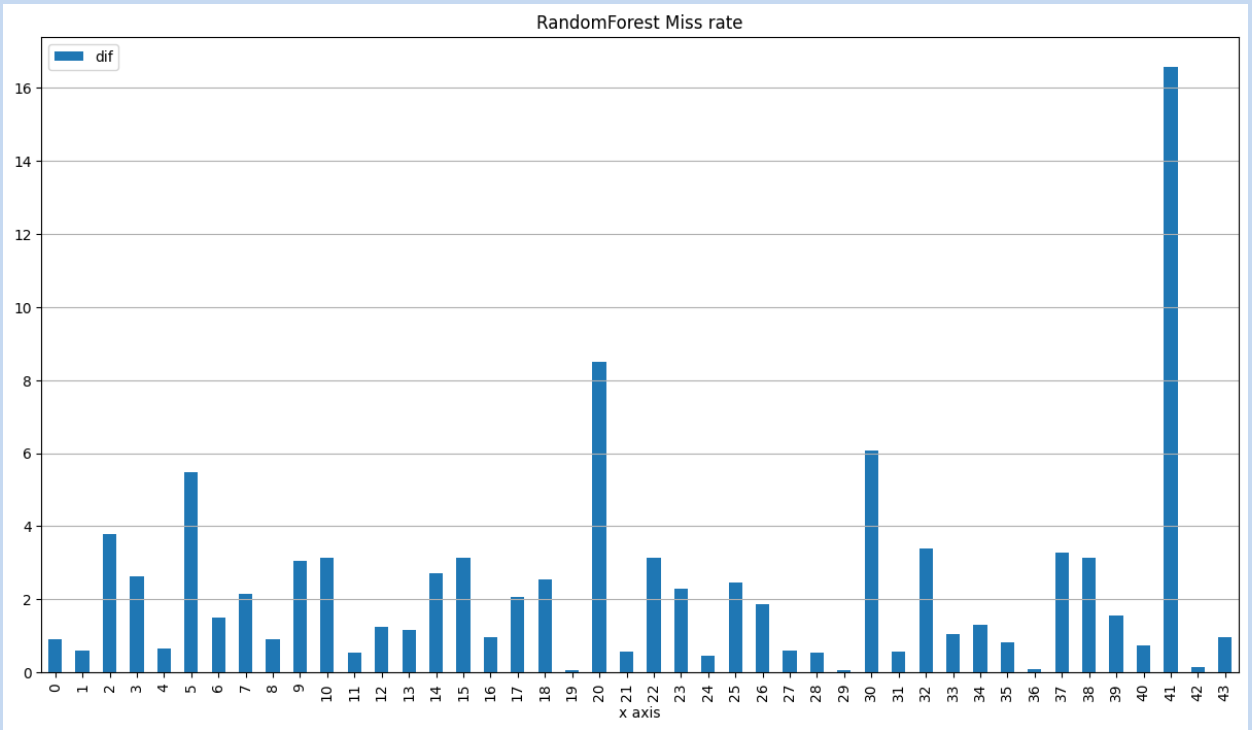
sonraki 4 diğerde aynı şey ifade eder faka Random Forest için

tahminleyicilerin bir örneğini gösteren grafikler:






RandomForest
real Y Values
predicted Y values
35000
30000
25000
20000
15000
10000
5000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
x axis


RandomForest Miss rate
dif
16
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
x axis

● Bulgular ve Yorumlar

bu ödevin sonunda bir ML modeli oluştururken bir çok etkileyici eleman olabileceğini gördük

bunlardan birisi ve bilirgin olanı outlier değerleri kaldırma

puanlama kısımında belli olduğu gibi veri kümesinde yanliş değerlere sahip özellik (veya etiket) varsa tahminleyicinin doğruluk oranına çok etkili olabilir. özellikle bu ödevde olduğu gibi veri kümesi küçük olduğu durumlarda.

ikinci olarak normalizasyon ve PCA uygulamak bizim bu veri kümemize pek etkilemedi. bunun sebebi özelliklerin çoğu bir category şeklinde olması olablilir oyüzden normalize edildiğinde pek değişklik olmuyor

son olarak seçilen modeller ile ilgili bir yorum. o da Random Forset modeli genel olarak burdaki veri kümesine çok uygun olmadığı görüldü. fakat onun şöyle güzel yanı var yakın tahmin ettiği değerler çok yakın olanları (linear'e göre) fazla miktarda var.

Not 1: t-test kullanılmadı çünkü regrasyon yöntemi kullandık