Yapay Zekâ Topluluğu

Arge semester Assignment 1

Kanada’daki Evlerin Regresyon Modelleri Kullanılarak Tahmin Edilmesi

Hazırlayan:

Onur Günenç

Son teslim tarihi:

11/02/2024

I. GİRİŞ KISMI:

Regresyon bağımsız ve bağımlı değişkenler arasındaki ilişkiyi matematiksel olarak ifade etmemize yarayan bir istatiksel ölçümdür.

Bu verileri analiz ettikten sonra, bir evin fiyatının tahmini için bir model oluşturmanız gerekir. Dolayısıyla, **bağımlı değişkenin davranışını büyük bağımsız değişkenlere dayanarak tahmin etmek için regresyon analizinin kullanıldığını** söyleyebiliriz. Burada bağımsız değişkenler yatak sayıları , banyo sayıları , hangi ülkede bulunduğu ülkenin nüfusu ve aile geliri medyanın göre oluşturduk.

II. Modellerin Açıklanması:

1. Multiple lineer regression:

Birden fazla bağımsız değişkenin farklı oranlarda etki etmesine bağlı olan ve bağımlı değişkenle doğrusal bir oranın hesaplanmasını sağlayan yöntemdir.

2. kNN Regression:

KNN algoritması yeni bir verinin tüm yakın komşularını arar ve bu komşuları arasındaki mesafeye göre yeni verinin hangi kategoride olacağına karar verir**.**

3. Random Forest Regression:

Temeli birden çok karar ağacının ürettiği tahminlerin bir araya getirilerek değerlendirmesine dayanır. Nihai tahmin için ağaçlardan tahmin değerleri talep edilirken her bir ağacın daha önce hesaplanan hata oranları göz önüne alınarak ağaçlara ağırlık verilir.

4. Support Vector Regression:

Kullanılma Amacı bir marjin aralığına max noktayı en küçük hata ile olabilecek şekilde doğru ya da eğriyi belirlemektir.

5. Neural Network Regression:

Yapay sinir ağları da aynı bunun gibi kendi belleğinde depolanan; öğrendiği bilgileri kullanarak söylenilen duruma karar verir. Diğer algoritmalar kesin olarak karar verir, verilen emirleri belirli kurallara göre işleme koyar ama ANN için bu geçerli değildir. ANN deneyimlerle hareket eder. Bu durumdan dolayı diğer algoritmalara göre yavaş olabilir.

6. Gradient Regression:

Gradient Boosting, zayıf tahmin edicilerin (genellikle karar ağaçları) bir araya getirilerek güçlü bir tahmin edici oluşturulmasını amaçlayan bir ensamble yöntemdir. *Temel fikir, önceki tahmin*

edicilerin hatalarını düzeltmeye çalışarak yeni tahmin ediciler eklemektir. Bu, karmaşık veri yapısını ve gürültüyü ele alabilen son derece etkili bir model oluşturmanın bir yolunu sunar.

III. Ön İşleme:

Modelde sonucu olumsuz etkileyeceğini düşündüğüm satırların çıkarılması için veride aykırı değerlerin sayılarına bakıyoruz.

```
Değişkenlerdeki aykırı değer sayısı:
{'Price': 2470, 'Number_Beds': 445, 'Number_Baths': 2670, 'Population': 5489, 'Median_Family_Income': 1336}
```

Sayılar bakıldığında çok fazla aykırı değer olduğunu o yüzden çıkarmamaya karar verdiğimi sadece yatak sayısı değişkeninde

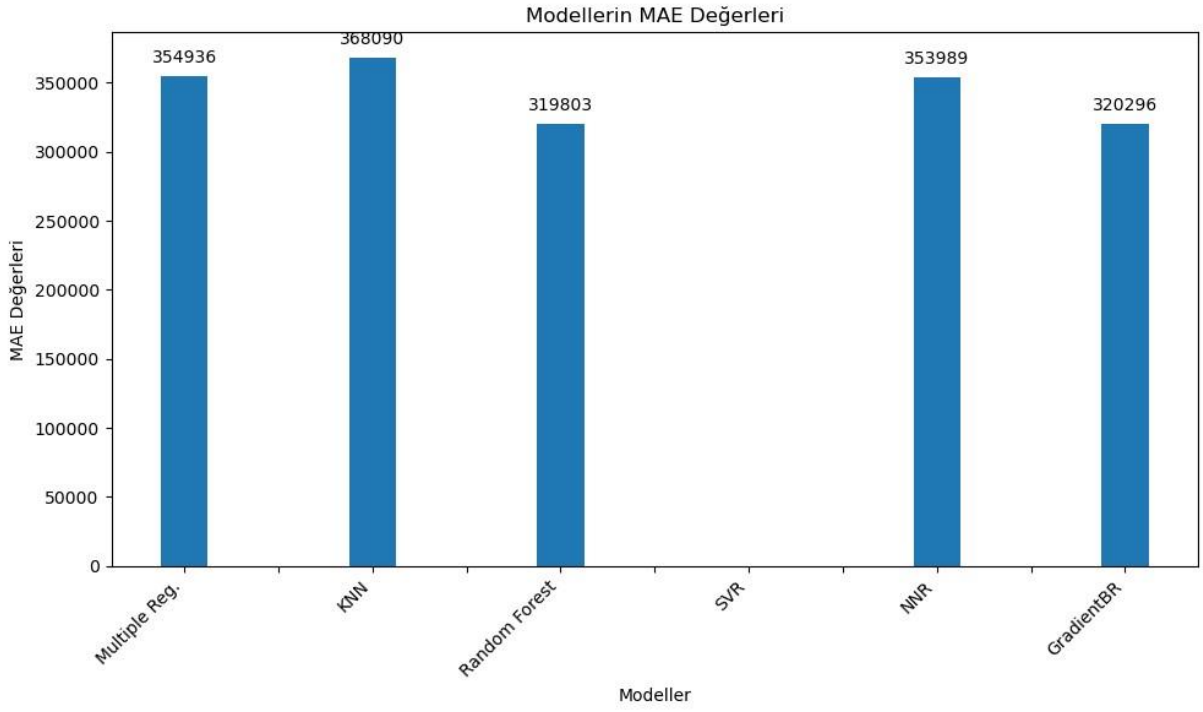
```
[407]: 3      10174
       4       7831
       2       7724
       5       4004
       1       2729
       6       1581
       0        769
       7        511
       8        253
       9         96
       10        35
       11        17
       12        13
       15         6
       16         5
       20         3
       17         2
       18         2
       21         1
       13         1
       26         1
       27         1
       47         1
       40         1
       35         1
       109        1
       46         1
       14         1
       36         1
       30         1
       19         1
```

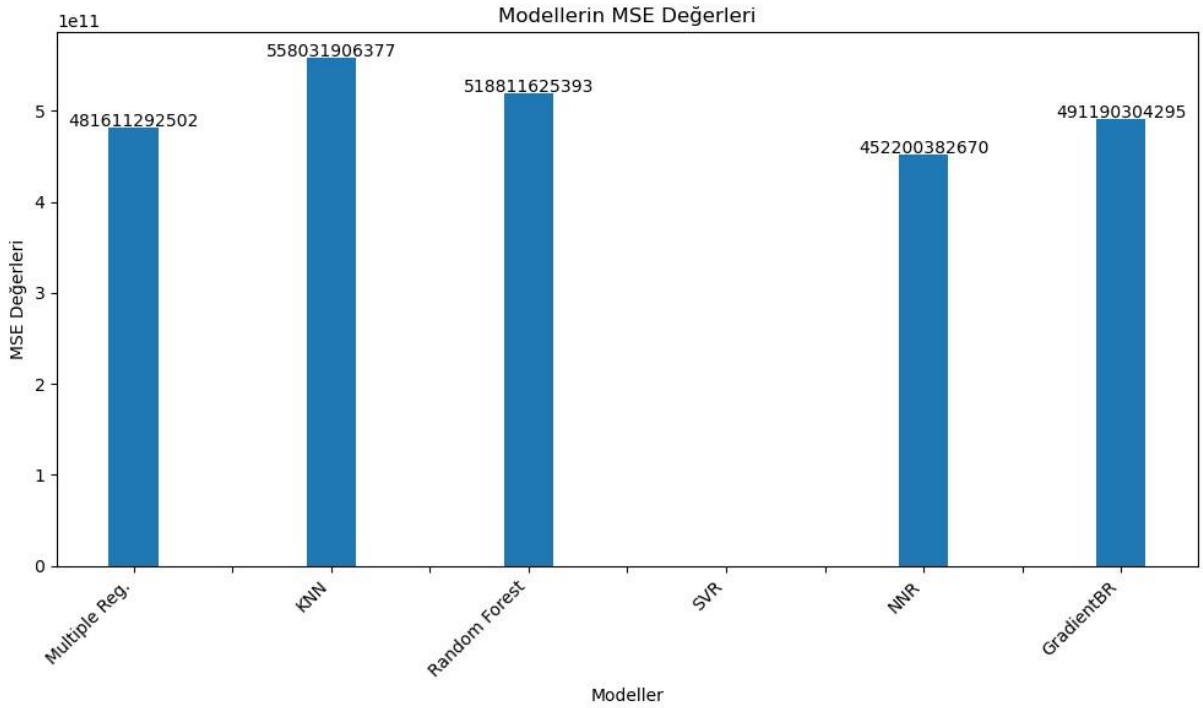
Sayılara bakıldığında 109 yatak sayısı olan gözlemin diğerlerinden çok daha fazla olduğunu veriyi diğerlerine göre çok daha kötü etkileyebileceği için sadece onu veri setimizden siliyoruz.

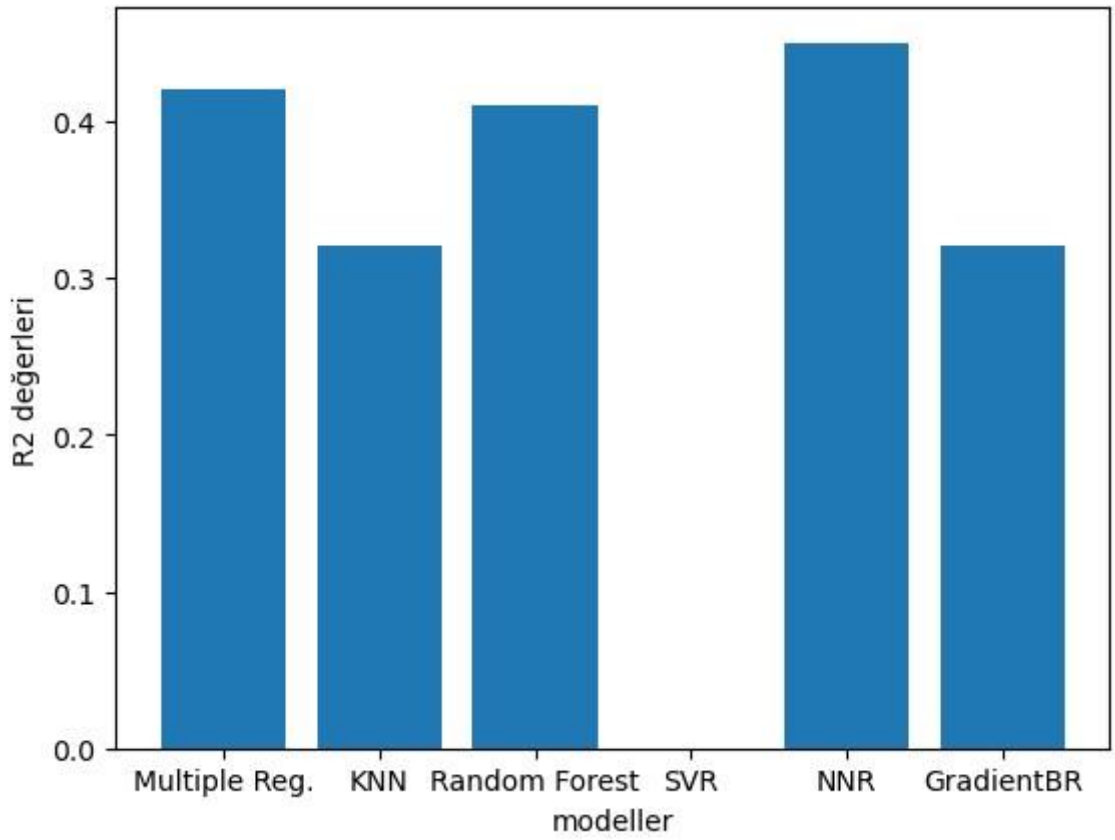
Model	Mean Absolute Error (MAE)	R-Squared	Mean Squared Error (MSE)
Multiple Linear Regression	354936	0.42	481611292502
kNN Regression	368090	0.32	558031906377
Random Forest Regression	319803	0.41	518811625393
Support Vector Regression	0	0	0
Neural Network Regression	353989	0.45	452200382670
Gradient Regression	320296	0.32	491190304295

IV. **Modellerin Kıyaslanması**:

Normalleştire yapmadım çünkü bazı modelde normalleştirme yapıp bazılarına uygulanamayacağından karşılaştırma imkânımız kalmıyor o yüzden böyle büyük sayılar elde ettik ve gene gözlemlenmesi zorlaştı o yüzden aşağıdaki grafikleri ekliyorum.







Sonuçlarını elde ettik.

SVR modelinin yapamadım çünkü bilgisayarımda sebebini bilmediğim bir sorundan kod saatlerce çalışmadı ve öyle kaldı.

v. **Değerlendirme**:

R2 değerlerine bakınca en iyi sonucu veren modelin NNR olduğunu görüyoruz. Sebebinin geriye yayılımlı olmasından kaynaklı olduğunu düşünüyorum .

KNN`in ise en kötü model olduğunu sebebinin modelde çok fazla aykırı değer olduğundan ve KNN modelinin aykırı değerlere çok hassas olmasından kaynaklı en kötü sonuçlar getiren model olmuştur.