Hacettepe Yapay Zeka Topluluğu

Kanada'daki Evlerin Regresyon Modelleri Kullanılarak Tahmin Edilmesi

Regresyon Ödevi

ECEM KAYA
11.02.2024

İçindekiler

1.GİRİŞ

Regresyonun Tanımlanması

Regresyon analizi, makine öğrenimindeki en önemli ve en yaygın alanlardan biridir. Doğrusal regresyon, belirli bir bağımsız değişken seti ile bağımlı değişken arasındaki ilişkileri modellemek için kullanılan istatistiksel bir makine öğrenmesi yöntemidir.

Genel olarak regresyon analizinde, bazı hedefler dikkate alınır ve çok sayıda gözlem yapılır. Her gözlemin bir veya daha fazla özelliği vardır. Özelliklerden (en az) birinin diğerlerine bağlı olduğu varsayımından sonra, aralarında bir ilişki kurulmaya çalışılır. Başka bir deyişle, bazı özellikleri veya değişkenleri diğerleriyle yeterince iyi eşleştiren bir işlev bulunması gerekir. Bağımlı özelliklere bağımlı değişkenler, çıktılar (output) veya yanıtlar (responses) denir. Bağımsız özeliklere ise bağımsız değişkenler, girdiler (intup) veya tahmin ediciler (predictors) denir. Regresyon problemlerinde genellikle bağımlı değişken sürekli verilerdir. Ancak girdiler sürekli, ayrık hatta cinsiyet, ırk, marka vb. gibi kategorik veriler olabilir.

Regresyon tabanlı makine öğrenmesi algoritmalarına ise tipik olarak, bir değişkenin diğerini etkileyip etkilemediğini, nasıl etkilediğini veya birkaç değişkenin nasıl etkili olduğunu cevaplamak için ihtiyaç duyulur.

2.MODELLERİN AÇIKLANMASI

1. Multiple Linear Regression (Çoklu Doğrusal Regresyon)

Çoklu doğrusal regresyon, bir bağımlı değişkenin birden fazla bağımsız değişkenle ilişkisini modellemek için kullanılan bir istatistiksel yöntemdir. Bu model, bağımsız değişkenlerin kombinasyonunu kullanarak bağımlı değişkenin varyansını açıklamaya çalışır. Temel olarak, regresyon denklemindeki katsayılar, bağımsız değişkenlerin etkilerini belirtir ve en küçük kareler yöntemiyle tahmin edilir. Bu yöntem, karmaşık ilişkileri anlamak ve tahminler yapmak için kullanılır, ancak bazı istatistiksel varsayımları doğru bir şekilde karşılamak gerekir.

2. kNN Regression (k- En Yakın Komşu Regresyonu)

k-En Yakın Komşu (kNN) Regresyonu, tahmin yapmak için kullanılan bir makine öğrenimi yöntemidir. Bir noktanın tahmin edilen değerini belirlemek için, ona en yakın k komşunun değerlerini kullanır. kNN, örnekler arasındaki uzaklık ölçüsünü kullanarak tahmin yapar. Küçük k değerleri değişken tahminlere, büyük k değerleri ise pürüzsüz tahminlere yol açar. k-En Yakın Komşu (kNN) Regresyonu, genellikle niceliksel (sayısal) verilerle kullanılır. Bağımlı değişken sürekli ise, kNN regresyonu tercih edilir. Ancak, bazı durumlarda kategorik verilerle de kullanılabilir.

3.Random Forest Regression (Rastgele Orman Regresyonu)

Random Forest regresyonu, bir makine öğrenimi yöntemi olup, birden fazla karar ağacının bir araya getirilmesiyle tahmin yapar. Her bir ağaç, rastgele seçilen örneklem verileri ve özellikler üzerinde eğitilir. Sonuç olarak, tüm ağaçların tahminlerinin ortalaması alınarak tahmin yapılır. Bu yöntem, karmaşık ilişkileri modellemek ve istikrarlı tahminler elde etmek için kullanılır. Random Forest regresyonu, her türlü sayısal ve kategorik veri türüyle uyumlu olarak kullanılabilir.

4.Support Vector Regression (Destek Vektör Regresyonu)

Support Vector Regression, regresyon problemlerinde kullanılan bir makine öğrenimi algoritmasıdır. Veri noktalarını bir hiper düzlem üzerinde böler ve bu hiper düzlemi en iyi şekilde uyarlamak için bir hata tolerans bandı kullanır. Küçük veri setleri ve gürültülü verilerle etkili çalışır, ancak parametre ayarı önemlidir.

SVR, problemde kullanılan veri dağılımı bilinmediğinde, yüksek boyutlu ya da az örnek içeren veri kümelerinde başarılı sonuçlar elde edebilmektedir.

5. Neural Network Regression (Sinir Ağı Regresyonu)

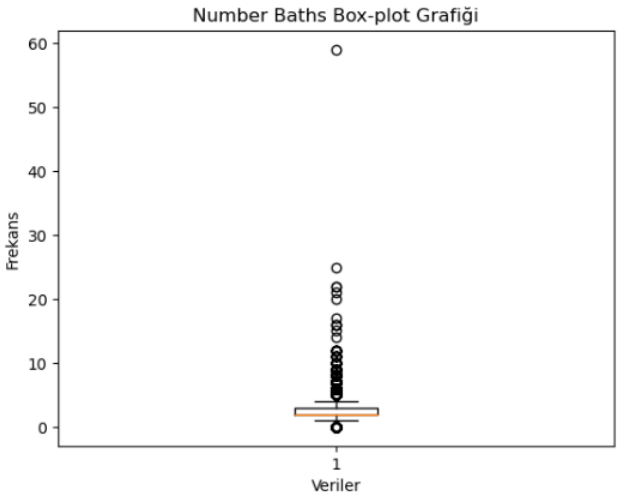
Neural Network Regression, karmaşık ilişkileri modellemek için kullanılan bir makine öğrenimi yöntemidir. Yapay sinir ağları kullanılarak girdi verilerinden sürekli bir çıktı tahmini yapılır. Sinir ağı regresyonunda, girdi verileri sinir ağının katmanlarına aktarılır ve ardından bu veriler üzerinde çeşitli matematiksel işlemler uygulanarak çıktı üretilir. Bu işlemler arasında ağırlıklandırma, aktivasyon fonksiyonları ve geri yayılım algoritmaları gibi teknikler bulunur. Ağ, birçok gizli katman içerebilir ve her katman birbirine bağlı nöronlardan oluşur. Her nöron, önceki katmandaki nöronların çıktılarına dayalı olarak bir ağırlıkla çarpılır, ardından bir aktivasyon fonksiyonuna sokulur ve sonuçlar bir sonraki katmana iletilir. Eğitim sürecinde, ağın hata oranı minimize edilmeye çalışılır ve doğru tahminler yapabilmesi için ağın ağırlıkları ayarlanır. Sinir ağı regresyonu genellikle hem niceliksel (sayısal) hem de kategorik (sınıflandırma) veri türleriyle kullanılabilir. Başka bir deyişle, hem sürekli bir çıktı değişkeni olan regresyon problemleri hem de kategorik sınıflandırma problemleri için kullanılabilir.

6. Gradient Boosting Regression (Gradyan Artırma Regresyonu)

Gradient Boosting Regresyonu, bir dizi zayıf tahminleyici modeli bir araya getirerek güçlü bir tahminleyici oluşturan bir makine öğrenimi tekniğidir. Her bir model, önceki modelin hatalarını düzeltmeye odaklanır ve böylece toplam hatayı azaltır. Bu yöntem, genellikle yüksek doğruluk gerektiren tahminleme problemlerinde kullanılır. Gradient Boosting Regresyonu, genellikle hem niceliksel (sayısal) verilerle hem de kategorik (sınıflandırma) verilerle kullanılabilir.

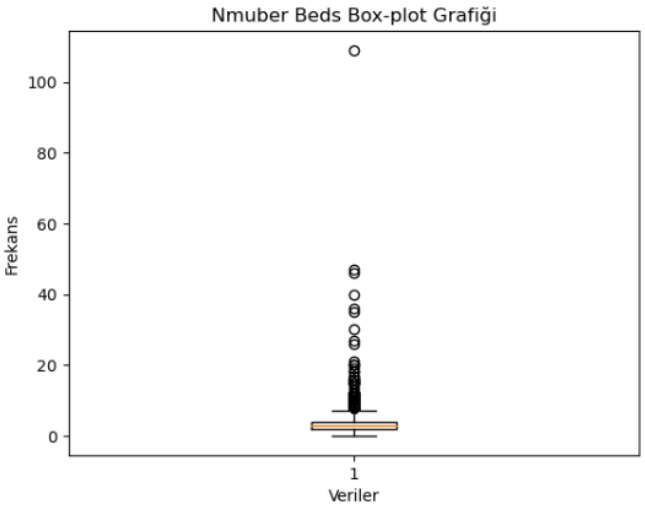
3.VERİ ÖN İŞLEME

Boxplot Grafiklerinin Yorumu

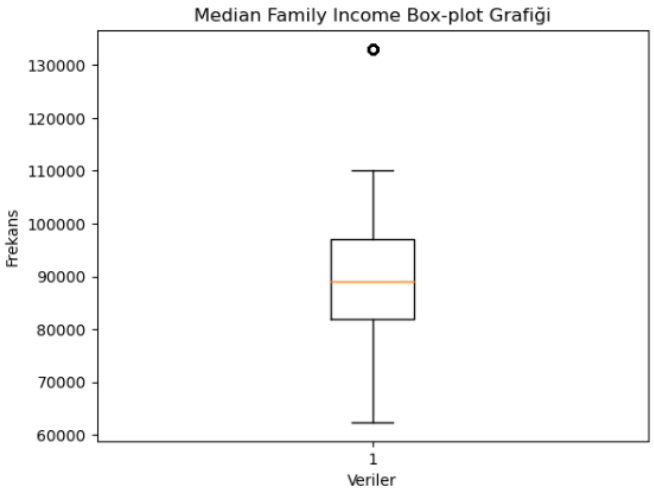


Boxplot, veri çeyreklerini (veya yüzdelikleri) ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğin görsel olarak dağılımını göstermek için kullanılır.
Boxplot bir veri kümesinin beş özelliğini gösterir: minimum değer, ilk (%25) çeyrek, medyan, üçüncü (%75) çeyrek ve maksimum değer.

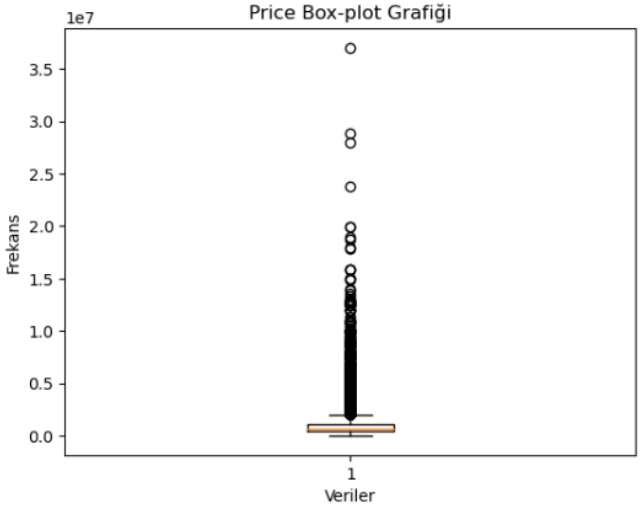
Yandaki grafikte Number Baths değişkenin boxplot grafiği görülmektedir. Bu grafikte görüldüğü üzere Number Baths değişkeninde aykırı değerler sıklıkla görülmektedir. Bu da modeller için olumsuzluk yaratacağı düşünüldüğü için aykırı değer temizlemesi yapılmaya karar verilmiştir.



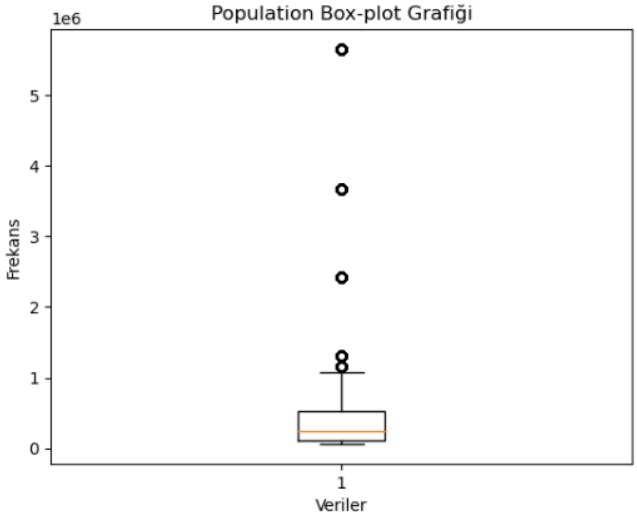
Yandaki grafikte Number Beds değişkenin boxplot grafiği görülmektedir. Bu grafikte görüldüğü üzere Number Beds değişkeninde aykırı değerler sıklıkla görülmektedir. Bu da modeller için olumsuzluk yaratacağı düşünüldüğü için aykırı değer temizlemesi yapılmaya karar verilmiştir.



Yandaki grafikte Median Family Income değişkenin boxplot grafiği görülmektedir. Bu grafikten yola çıkarak Median Family Income değişkeninin modelleri olumsuz etkileyecek aykırı değerleri olmadığı görülmektedir. Sonuç olarak bu değişkende aykırı değer temizlemesi yapılmamaya karar verilmiştir.



Yandaki grafikte Price değişkenin boxplot grafiği görülmektedir. Bu grafikte görüldüğü üzere Price değişkeninde aykırı değerler sıklıkla görülmektedir. Bu da modeller için olumsuzluk yaratacağı düşünüldüğü için aykırı değer temizlemesi yapılmaya karar verilmiştir.



Yandaki grafikte Population değişkenin boxplot grafiği görülmektedir. Bu grafikten yola çıkarak Population değişkeninin modelleri olumsuz etkileyecek aykırı değerleri olmadığı görülmektedir. Sonuç olarak bu değişkende aykırı değer temizlemesi yapılmamaya karar verilmiştir.

Eksik Veri Tespiti

```
: print(data_df.isnull().sum())
```

```
City                  0
Price                 0
Address               0
Number_Beds           0
Number_Baths          0
Province              0
Population            0
Latitude              0
Longitude             0
Median_Family_Income  0
dtype: int64
```

Yandaki grafikte veri setimizdeki değişkenler üzerinde eksik veri olup olmadığı üzerine analiz yapılmıştır. Ve görüldüğü üzere değişkenlerde eksik veri bulunamamıştır.

Olumsuz Etkileyen Değişken

Address değişkeninde her bir gözlem değeri farklı olduğundan dolayı Address değişkeninin modeli olumsuz etkileyeceği düşünülmektedir. Bunun sonucunda Address değişkeninin çıkartılmasına karar verilmiştir.

4.MODELLERİN KIYASLANMASI

MODEL	Mean Absolute Error (MAE)	R-Squared	Mean Squared Error (MSE)
Multiple Linear Regression	0.41	0.56	0.36
kNN Regression	0.52	0.50	0.58
Random Forest Regression	0.44	0.63	0.43
Support Vector Regression	0.55	0.43	0.67
Neural Network Regression	0.53	0.51	0.56
Gradient Boosting Regression	0.48	0.57	0.50

Tablo.1

5.SONUÇ

Multiple Linear Regression modelinde, , modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Ekstra olarak modeli kurarken Price değişkeniyle diğer değişkenlerin korelasyonlarına bakılıp korelasyon değeri 0.2'den büyük olanlar alınıp bir X değişkeni oluşturulmuştur. Bu işlem R^2 değeri negatif bulunduğu için yapılmıştır. Bu modelde min max normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.56 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %56'sını açıkladığını görülmektedir.

kNN regression modelinde, modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Bu modelde Z-Score normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.50 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %50'sini açıkladığını görülmektedir.

Random Forest Regression modelinde, modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Modele en uygun ağaç sayısını belirlemek için GridSearchCV kullanılarak çapraz doğrulama (cross validation) yapılmıştır. Bu modelde Z-Score normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.63 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %63'ünü açıkladığını görülmektedir.

Support Vector Regression modelinde, modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Bu modelde en uygun parametreyi bulabilmek için elle ayarlama yapılmıştır çünkü GridSearchCV kullanmak parametreler fazla olduğundan dolayı üzün sürmüştür. Bu modelde Z-Score normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.43 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %43'ünü açıkladığını görülmektedir.

Neural Network Regression modelinde, modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Bu modelde Z-Score normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.51 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %51'ini açıkladığını görülmektedir.

Gradient Boosting Regression modelinde, modeli kurarken diğer modellerde de yapıldığı gibi Price, Number Beds ve Number Baths değişkenlerinde aykırı değer temizlemesi yapılarak model kurulmuştur. Bu modelde Z-Score normalizasyonu yapılmıştır. Bunun sonucunda MSE ve MAE değerleri daha çok yorumlanabilir hale getirilmiştir. R^2 değeri 0.57 elde edilmiştir. Bunun sonucunda kullanılan bağımsız değişkenlerin birlikte bağımlı değişkendeki varyansın %57'sini açıkladığını görülmektedir.

Yapılan modeller sonucunda elde edilen Tablo1'de görüldüğü üzere Random Forest Regression modelinin veri seti üzerinden Price değişkenini tahmin etme işleminde daha iyi olduğu görülmektedir. Bunun nedenlerinden birisi Random Forest Regression' da birden fazla karar ağacının bir araya getirilmesiyle tahmin yapılması ve bu tahminlerin ortalamaları alınıp sonuç tahminin oluşturulmasıdır. Bu yöntem, karmaşık ilişkileri modellemek ve istikrarlı tahminler elde etmek için kullanılır. Price değişkenini tahmin etme işleminde diğer modellere kıyasla en kötü olan model Support Vector Regression olarak görülmektedir.