Hacettepe Üniversitesi
YAPAY ZEKA
Topluluğu

Kanada'daki Evlerin Regresyon Modelleri Kullanılarak Tahmin Edilmesi

Proje Raporu

Hazırlayan

Nurşah Satılmış

İçindekiler

1.Giriş

Regresyon analizi, istatistiksel bir yöntemdir ve bağımlı bir değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılır. Bu ilişkinin doğasını anlamak ve gelecekteki değerleri tahmin etmek için kullanılan güçlü bir araçtır. Regresyon modelleri, ekonomi, finans, mühendislik, sağlık bilimleri ve sosyal bilimler gibi çeşitli disiplinlerde sıkça rastlanan analiz araçlarıdır.

Regresyon analizi, bağımlı değişkenin sürekli bir sayısal değer olduğu durumlarda kullanılır. Örneğin, bir şirketin geliri, hisse senedi fiyatları, ev fiyatları veya hastaların tedavi süresi gibi değişkenler regresyon analizi ile modelleme yapılabilir.

Bu modeller genellikle veri analizi ve tahmin süreçlerinde kullanılır. Geçmiş verilere dayanarak gelecekteki olayları veya değerleri tahmin etmek için kullanılabilecekleri gibi, değişkenler arasındaki ilişkiyi anlamak ve faktörlerin bağımlı değişken üzerindeki etkilerini değerlendirmek için de kullanılırlar. Dolayısıyla, regresyon analizi, analitik ve kestirimci bir araç olmanın ötesinde, birçok alanda karar verme süreçlerini destekleyen temel bir yöntemdir.

Bu çalışmada, regresyon modelinin ne olduğunu ve hangi verilerde kullanılabileceğini inceleyeceğiz. Ayrıca, Kanada'daki evlerin özelliklerini içeren veri seti üzerinde regresyon modellerini uygulayarak aldığımız sonuçları değerlendireceğiz.

2.Kullanılan Regresyon Modellerinin Tanımları ve Performansının Karşılaştırılması

2.1 Multilinear Regression (Çoklu Doğrusal Regresyon)

Multilinear regression, birden fazla bağımsız değişkenin kullanıldığı ve bağımlı değişkenin sürekli bir değer aldığı durumlarda kullanılır. Lineer ilişkileri modellemek için kullanılan temel bir regresyon yöntemidir. Veri setinde bağımsız değişkenlerin bağımlı değişken üzerinde lineer bir etkisi olduğu varsayımına dayanır. Eğer bağımsız değişkenler arasında lineer bir ilişki varsa ve bu ilişki iyi bir şekilde modelleyebilirse, multilinear regression iyi sonuçlar verebilir.

2.2 K-Nearest Neighbors (KNN) Regression

KNN regression, bir tahmin yaparken benzerlik esasına dayanan bir yöntemdir. Tahmin edilen noktanın çevresindeki en yakın k komşulardan alınan değerlerin bir ortalama veya ağırlıklı ortalaması kullanılarak tahmin yapılır. Bu model, veri setindeki lokal yapıyı yakalamak için etkili olabilir. Özellikle veri setindeki ilişkilerin lineer olmadığı veya karmaşık yapılar içerdiği durumlarda iyi çalışabilir.

2.3 Random Forest Regression

Random forest regression, birçok karar ağacının bir araya gelerek bir tahmin yapmak için kullanıldığı bir ensemble yöntemidir. Her bir ağaç, rastgele seçilen alt veri setleri üzerinde eğitilir ve sonuçlarını bir araya getirerek tahmin yapılır. Bu model, non-lineer ilişkileri ve etkileşimleri modellemek için güçlü bir araç olabilir. Ayrıca, veri setindeki gürültüyü azaltmak ve overfitting'i engellemek için etkili bir yöntemdir.

2.4 Support Vector Regression (SVR)

SVR, özellikle küçük veri setleri ve yüksek boyutlu özellik uzaylarında etkili olan bir regresyon yöntemidir. SVR, veri setinin dağılımını yakalayan bir düzlem veya hiperdüzlemi bulmaya çalışır. Bu model, lineer olmayan ilişkileri modellemek için de kullanılabilir ve aynı zamanda outliers'a karşı dirençlidir.

2.5 Neural Network Regression (Yapay Sinir Ağı Regresyonu)

Neural network regression, karmaşık ve non-lineer ilişkileri modellemek için kullanılan derin öğrenme yöntemlerinden biridir. Yapay sinir ağları, veri setindeki karmaşık yapıları yakalayabilir ve özellikle büyük veri setleri üzerinde iyi çalışabilir. Ancak, aşırı öğrenme riski bulunmaktadır, bu nedenle iyi bir şekilde düzenlenmelidir.

2.6 Gradient Boosting Regression

Gradient boosting regression, zayıf öğrenicileri bir araya getirerek güçlü bir tahminci oluşturan bir ensemble yöntemidir. Her bir ağaç, önceki ağaçların hatalarını düzeltmeye odaklanarak eğitilir. Bu model, karmaşık yapıları yakalamak ve overfitting'i azaltmak için etkili olabilir

3. Veri Ön İşleme

3.1. Keşifçi Veri Analizi

Adres değişkeni şehir ve eyalet bilgilerini de içerdiğinden , hedef değişkeni tahmin etmek için adres değişkeni yeterli olup eyalet ve şehir bilgisi gerekli değildir, “Province” ve “City” sütunları veri setinden silinmiştir.

Değişkenlerin veri türleri incelendi , eksik veri olup olmadığı kontrol edildi. Eksik veriye rastlanmadı. Tekrarlı satır olup olmadığı incelendi, 2516 satırın tekrarlı satır olarak tespit edilmesine rağmen satırları incelediğimizde birkaç farklı değişkenin olduğu ve buna bağlı olarak hedef değişkenin farklı olduğu saptandığı için verinin yapısını değiştirmemek adına bu satırlar için işlem yapılmadı.

Histogram grafikleri ile verinin dağılımı incelendi.

3.2. Aykırı Değer Analizi

Veri seti büyük olduğu için istatiksel testler yerine veri görselleştirme ile (Q-Q grafik ile) verinin normal dağılımlı olup olmadığına bakıldı ve normal dağılımlı olmadığına karar verildi. Bu işlem z-score aykırı değer tespiti yapılıp yapılamayacağını öğrenmek amacıyla yapıldı. Veri normal dağılıma sahip olmadığı için z-score ile değil IQR ile aykırı değer tespiti yapıldı.

IQR ile saptanan aykırı değerler minimum veya maximum değerlere eşitlendi (Winsorizasyon). Kutu grafiği ile her değişken için aykırı değerlerin giderilmeden önceki ve sonraki halleri görselleştirildi.

3.3. Encoding

Kategorik değişkenler, bir veri setinde belirli kategorileri veya grupları temsil eden değişkenlerdir. Bu tür değişkenler genellikle metin veya sembolik değerlerle temsil edilirler. Ancak, birçok makine öğrenimi algoritması ve istatistiksel teknik, sayısal verilerle çalışır. Bu nedenle, kategorik değişkenlerin sayısal değerlere dönüştürülmesi gereklidir. Bu işleme "encoding" denir.

Encoding, kategorik değişkenlerin belirli bir sıralama veya hiyerarşi içinde temsil edilmesini sağlar, böylece algoritmalar bu değişkenlerle çalışabilir. Bu dönüşüm, genellikle one-hot encoding veya label encoding gibi teknikler kullanılarak gerçekleştirilir. One-hot encoding, her kategoriye ayrı bir sütun atayarak kategorik değişkenleri ikili (0 veya 1) sayısal değerlere dönüştürür. Label encoding ise kategorik değişkenlerin her bir kategorisine benzersiz bir sayı atayarak onları sayısal değerlere dönüştürür. Bu yöntemin avantajlarından biri, veri kümesinin boyutunu azaltmasıdır, çünkü her kategori için sadece bir tane ek sütun gerektirir. Ayrıca, bazı makine öğrenimi modelleri, özellikle ağaç tabanlı modeller, label encoded verilerle daha iyi

çalışabilir çünkü bu modeller, kategorik değişkenleri doğrudan işleyebilir ve kategoriler arasındaki ilişkileri öğrenebilir. Ancak, label encoding'in dezavantajlarından biri, sayısal değerlerin sıralı olduğunun algılanabileceği ve modelin yanlış sonuçlara yol açabileceği durumlar olabilir.

Veri setinin boyutunu büyütmek istenmediği için kategorik değişkenler label encoding ile kodlandı.

Kodlama işlemi sonrası heatmap ile koralasyonlar incelendi. “Province” ve “Population” değişkenleri ile hedef değişken arasında görece daha az koralasyon olduğu gözlemdi

.

3.4. Normalizasyon

Z-Score normalizasyonu, verilerin ortalama değerini 0'a ve standart sapmasıyla da ölçekler. Bu, regresyon modelinin, değişkenler arasındaki farklı ortalama ve yayılımların neden olduğu yanlılıkları gidererek katsayıların daha doğru yorumlanmasını sağlar. Ayrıca, Z-Score normalizasyonu, veri dağılımını değiştirmez, sadece ölçekler, böylece regresyon modeli, verilerin orijinal dağılımına dayalı olarak daha doğru tahminler yapabilir. Diğer normalizasyon yöntemleri veri dağılımını değiştirirken, Z-Score normalizasyonu bu özelliği korur. Bu avantajlardan dolayı, Z-Score normalizasyonu uygulandı.

3.5. PCA

Modele verilecek inputun boyutunu azaltmak adına değişkenleri elemek yerine , bunun model performansları üzerinde negatif etkisi olabileceği düşünüldü, PCA ile verinin yapısını değiştirmeden iyi bir şekilde ifade edecek bileşenler yaratıldı. Kaç bileşenle maximum kümülatif varyansa ulaşalabileneceği grafik ile gözlemlendi . Grafik yorumlanarak 5 bileşenin veriyi en iyi şekilde açıklayabileceğine karar verildi.

4. Modellerin Karşılaştırılması

Support Vector Regresyon için hiperparametre araması çok uzun zaman aldığı(2 saat yaklaşık) için arama yaptığım değerleri aralık olarak vermedim.Aynı şekilde ,Neural

Network ve Gradient Boosting regresyon modelleri için de hiperparametre ayarları çok uzun sürdü(4-5 saat).Bu nedenle bu modeller için hiperparametre aralıklarını belirli değerler arasında arama yaparak buldum. Belki donanımsal belki farklı bir sebepten ötürü işlemin bu denli uzun sürmesi , modeller için en iyi parametreleri bulmamda işleri zorlaştırdı.

Bunların sonucunda elde edilen veriler aşağıda verilmiştir.

(En iyi alınan değerler)

Model	Mean absolute error (MAE)	R-Squared	Mean Squared Error (MSE)
Multiple linear regression	0,550	0,593	0,463
kNN Regression	0,385	0,749	0,285
Random Forest Regression	0,389	0,458	0,339
Support Vector Regression	0,389	0,298	0,461
Neural Network Regression	0,549	0,238	0,556
Gradient Regression	0,389	0,298	0,461

Multiple Linear Regression (Çoklu Doğrusal Regresyon):

- Bu model, bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkiyi modellemek için kullanılır.
- Katsayılar sayesinde her bir bağımsız değişkenin etkisi açıkça anlaşılabilir.
- Ancak, bu model doğrusal olmayan ilişkileri yakalayamaz ve aşırı uyum sorunuyla karşılaşabilir.
- Veri belirgin bir normal dağılıma sahip olmadığı için diğer regresyon modellerine göre daha başarısız oldu.

K-Nearest Neighbors Regression (KNN) (K-En Yakın Komşu Regresyonu):

- KNN, tahmin yapmak için en yakın komşuların etrafındaki verileri kullanır.
- Esnek bir yapıya sahiptir ve doğrusal olmayan ilişkileri modellemekte etkilidir.
- Ancak, tahmin yapmak için tüm eğitim verisinin saklanması gerektiği için büyük veri setlerinde hesaplama yoğunluğu artabilir.

- Modele verilen veri seti büyük olmadığından hesaplama maaliyeti çok olmadı , pca uyguladığımız input ile 0,285 mse değeriyle en az hataya ve 0,749 R-kare değeri ile veriyi en iyi şekilde açıklamasıyla en iyi performası gösteren model oldu.

Support Vector Regression (SVR) (Destek Vektör Regresyonu):

- SVR, veri noktaları arasındaki optimum hiper düzlemi bulmaya çalışarak doğrusal ve non-dogrusal regresyon problemlerinde etkilidir.
- Farklı çekirdek fonksiyonları kullanarak esneklik sağlayabilir.
- Ancak, hiperparametrelerin ayarlanması uzun sürdüverilen değerler arasında ne n

Multi-layer Perceptron (Neural Network) Regression (Çok Katmanlı Algılayıcı Regresyonu):

- Bu model, gizli katmanlar aracılığıyla karmaşık ilişkileri modellemek için kullanılır.
- Yapay sinir ağları, veri içindeki karmaşıklığı yakalayabilir ve esnek bir yapıya sahiptir.
- Ancak, eğitim süresi oldukça uzundu ,model için en iyi parametreleri bulmak çok uzun sürdü ve yeterli deneme yapılamadığı için istenilen sonuç alınamadı .
- Verilen parametreler arasında en iyi parametre ile eğitilen model 0,556 mse ile başarılı olamadı , ve r-kare değeri 0,238 geldiği için veriyi açıklamada iyi bir performans sergileyemedi.

Gradient Boosting Regression (Gradyan Arttırma Regresyonu):

- Bu model, zayıf öğrenicilerin bir araya gelmesiyle güçlü bir tahmin modeli oluşturur.
- Adım adım artan bir yaklaşım kullanarak, her adımda hata azalır.Aşırı uyumu doğal olarak azaltır ve büyük veri setlerinde etkilidir.
- Tuhaf bir şekilde svr ile aynı performans metriklerine sahip olan veri , metriklere göre başarılı olamadı. Yapılan bir hatadan kaynaklı olabilir.

5. Sonuç

KNN regresyon modeli düşük ortalama kare hatası ile en iyi sonuca ulaştı ve ayrıca yaklaşık %80 lik R-kare oranıyla veriyi en iyi açıklayan model oldu.