Kanada’daki Evlerin Regresyon Modelleri Kullanılarak Tahmin Edilmesi

Son Teslim : 11 Şubat 23:59

Regresyon

Regresyon, istatistik ve makine öğrenimi alanlarında sıklıkla kullanılan bir analiz tekniğidir. Temel olarak, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modelleme amacını taşır. Bu ilişki, bağımsız değişkenlerin değerleri kullanılarak bağımlı değişkenin tahmin edilmesi veya açıklanması olarak ifade edilir.Genellikle sürekli sayısal verilerle çalışan regresyon analizi, değişkenler arasındaki matematiksel ilişkiyi temsil eden bir model oluşturur. Bu model, veri setindeki değişkenlikleri anlamak ve gelecekteki değerleri tahmin etmek için kullanılabilir.

Regresyon modellerinin kullanım alanları oldukça geniştir. Örneğin, ekonomi, finans, biyoloji, mühendislik, sosyal bilimler ve makine öğrenimi gibi birçok alanda kullanılırlar. Bir şirketin gelirini tahmin etmek, bir ilacın etkisini ölçmek veya hava durumu tahminleri yapmak gibi birçok uygulama örneği bulunmaktadır.Örneğin, bir regresyon modeli kullanarak bir şirketin gelirini tahmin etmeyi düşünelim. Bu modelde bağımlı değişken gelir olacak ve bağımsız değişkenler arasında reklam harcamaları, personel maaşları, pazarlama stratejileri gibi faktörler bulunabilir. Model, bu faktörleri kullanarak gelecek dönemdeki geliri tahmin edebilir.

Assignment 1

Elimizde Kanada'nın 45 farklı şehrinden toplanmış 29 Kasım 2023 yılında toplanmış bir veri seti bulunmaktadır. Veri seti 9 farklı değişkenden ve 35768 veriden oluşmaktadır. Sizden isteğimiz verilen veri setini kendi içerisinde karmadan verinin ilk yüzde 80'lik kısmında yer alan verileri kullanarak verileri eğitmeniz ve son yüzde 20'lik kısmı kullanarak "Price" değeri hakkında tahminde bullunmanızdır. Ödev için izlemeniz gereken adımlar aşağıda yer almaktadır:

1-) Ön İşleme

1. Her bir sütunun dağılımını histogram kullanarak gösterilmesi
2. Modelde sonucu olumsuz yönde etkileyceği düşülen sütunların çıkarılması (Olumsuz etkileme sebebinin yorumlanması)
3. Box Plot veya seçtiğiniz farklı bir yöntem kullanarak aykırı değerlerin (outlier) sayısı hakkında yorum yapılması ve isteğe bağlı olarak aykırı değerlerin elimine edilmesi
4. Z Score normalizasyonunun yapılması
5. Veri setini inceleyerek bireysel olarak ön işleme işlemlerinin eklenmesi ve bunların eklenme sebeplerinin açıklanması

2-) Farklı Modellerin Denenmesi

Ödevimizin ikinci aşamasında aşağıdaki regresyon modellerini sırasıyla eğitmeniz gerekmektedir.

1. Multiple Linear Regression
2. kNN Regression (bu ödev için k = 20 alabilirsiniz veya kendiniz optimal bir değer de hesaplayabilirsiniz)
3. Random Forest Regression
4. Support Vector Regression
5. Neural Network Regression (Python'daki MLPRegressor kütüphanesini kullanabilirsiniz)
6. Gradient Boosting Regression

3-) Modellerin Kıyaslanması

Model	Mean Absolute Error (MAE)	R-Squared	Mean Squared Error (MSE)
Multiple Linear Regression			
kNN Regression			
Random Forest Regression			
Support Vector Regression			
Neural Network Regression			
Gradient Regression			

Tablo 1

Bu aşamada sizden 2. aşamada eğittiğiniz modelleri test ederek , MAE, R-Squared ve MSE metrikleri ile modelin ne kadar iyi olduğunu ölçmeniz , tabloyu doldurmanız ve en düşük hataya sahip modelin neden “Price” özelliğini daha iyi tahmin edebildiğini ve en yüksek hataya sahip modelin neden “Price” özelliği için daha kötü bir tahmin modeli olduğunu açıklamanız beklenmektedir.

Veri Seti

Veri setine şu adresten ulaşabilirsiniz : https://www.kaggle.com/datasets/jeremylarcher/canadian-house-prices-for-top-cities/data

Rapor ve Puanlama

Raporun içeriği sırasıyla

1. Giriş Kısmı: Regresyon modelinin ne olduğu ve hangi verilerde kullanılabiliceğinin tanımlanması
2. Modellerin Açıklanması: Ödev içerisinde kullanılacak olan modellerin tanımları yapılarak hangi modelin hangi tür veriler üzerinde daha iyi çalıştığının yorumlanması
3. Ön İşleme: Verinin ön işleme kısmında neler yapıldığının anlatılması
4. Modellerin Kıyaslanması: Tablo 1 doldurularak modellerin bu veri seti üzerindeki performanslarının incelenmesi
5. Sonuç: Tablo 1'e göre en iyi ve en kötü tahmin değerlerine sahip modellerin neden en iyi ve en kötü olduğunun açıklanması

Puanlama kısmı aşağıdaki gibi olacaktır:

1. Rapor
    a. Giriş : 5 Puan
    b. Modellerin Açıklanması : 15 Puan
    c. Ön İşleme Kısmının Açıklanması : 20 Puan
    d. Modellerin Kıyaslanması ve Sonuç: 50 Puan
2. Kod
    a. Clean Code : 10 Puan

Not : Kodunuzu Python Programlama dili yazmanız istenmektedir. Kodunuzu hazır kütüphaneler kullanarak yazabilirsiniz ancak kendi oluşturduğunuz her bir model için ek puan alacaksınız.

Not : Raporları değerlendirirken ne kadar öğrenmiş olduğunuz dikkate alınmaya çalışılacaktır. Bu sürece katılma sebebinizin ARGE biriminde rol alabilme şansının yanı sıra kendinizi geliştirmek için de olduğunu lütfen unutmayın. Sizden temennimiz amacınızın bir an önce ödevi teslim etmek değil, ödev içerisinde kullanmanızı istediğimiz modelleri güzel bir şekilde öğrenmiş olmanızdır. Unutmayın yapmış olduğunuz ödev sadece öğrendiğinizi değerlendirebilmemiz için bir araç, önemli olan bu süreç boyunca kendiniz geliştirmiş olmanız ve semester boyunca harcayacağınız vaktinizin olumlu ve verimli bir sonuç vermesidir. Sorularınızı whatsapp grubu üzerinden sorabilirsiniz. Şimdiden başarılar.

Ödev Teslim: Kodunuzu ipynb veya py uzantılı dosya olarak ve raporunuzu pdf formatı olarak **argeaiclub@gmail.com** mailine 11 Şubat Pazar günü 23:59'a kadar gönderebilirsiniz.