1-) Projenin Amacı

Bu proje, belirli bir ürünün bir mağazada gelecekteki aylarda ne kadar satılacağını tahmin etmeyi amaçlamaktadır. Proje kapsamında, geçmiş satış verileri analiz edilerek gelecekteki satışları tahmin etmek için bir makine öğrenimi modeli geliştirilmiştir. Bu süreçte, tahmin modelinin doğruluğunu artırmak için veri işleme, model eğitimi ve performans değerlendirmesi gibi adımlar gerçekleştirilmiştir.

2-) Kullanılan Veri Seti

Bu projede, Kaggle'ın "Predict Future Sales" yarışmasından alınan veri seti kullanılmıştır. Veri seti, mağaza ve ürün kombinasyonlarının geçmiş satış bilgilerini içerir. Dosyaların açıklamaları şu şekildedir:

2.1-) sales_train.csv

- **İçerik**: Mağaza ve ürün kombinasyonlarının geçmiş satışlarını içeren ana veri setidir.
- **Sütunlar**:
    - **date**: Satışın gerçekleştiği tarih.
    - **date_block_num**: Her ayı temsil eden sayısal bir değer (örneğin, 0: Ocak 2013, 1: Şubat 2013).
    - **shop_id**: Mağaza kimliği.
    - **item_id**: Ürün kimliği.
    - **item_price**: Ürünün birim fiyatı.
    - **item_cnt_day**: Belirli bir günde satılan ürün sayısı.

2.2-) items.csv

- Ürünlere dair ek bilgileri içerir.
    - **item_id**: Ürün kimliği.
    - **item_name**: Ürün adı.
    - **item_category_id**: Ürünün kategori kimliği.

2.3-) item_categories.csv

- Ürün kategorilerinin detaylarını içerir.
    - **item_category_id**: Kategori kimliği.
    - **item_category_name**: Kategori adı.

2.4-) shops.csv

- Mağazalarla ilgili bilgileri içerir.
    - **shop_id**: Mağaza kimliği.
    - **shop_name**: Mağaza adı.

2.5-) test.csv

- Test veri setidir ve gelecekteki satışların tahmin edilmesi gereken mağaza ve ürün kombinasyonlarını içerir.

3-) Veri Hazırlığı

Veri seti, modelin performansını artırmak amacıyla bir dizi ön işleme adımından geçirilmiştir:

3.1-) Tarih Sütununun Formatlanması

- **date** sütunu, yıl, ay ve gün olarak ayrıştırılmıştır. Bu işlem, satış verilerinin zamana bağlı analiz edilmesini kolaylaştırmıştır.

3.2-) Eksik ve Hatalı Verilerin Düzeltilmesi

- **item_price** ve **item_cnt_day** sütunlarında hatalı değerler tespit edilmiştir:
  - Negatif veya sıfır fiyatlar, ilgili ürünlerin medyan fiyatlarıyla değiştirilmiştir.
  - Negatif satış değerleri 0 olarak düzeltilmiştir.

3.3-) Verilerin Aylık Düzeye Getirilmesi

- Günlük satış verileri, **date_block_num**, **shop_id** ve **item_id** kombinasyonlarına göre gruplanarak aylık satış toplamlarına dönüştürülmüştür.
  - **item_cnt_day** sütunundaki değerler toplanarak **item_cnt_month** adıyla yeni bir sütun oluşturulmuştur.
  - **item_price** sütunundaki değerler, aynı dönemdeki ürünlerin ortalama fiyatını hesaplamak için kullanılmıştır.

3.4-) Gecikme Özelliklerinin Eklenmesi

- Geçmiş satış bilgilerini modelin öğrenebilmesi için gecikme özellikleri (**lag features**) eklenmiştir:
  - Örneğin, bir ürünün 1 ay önceki satışları **item_cnt_month_lag_1** sütununda yer almaktadır.

3.5-) Eksik Değerlerin Doldurulması

- Gecikme sütunlarında oluşan eksik değerler (NaN), 0 ile doldurulmuştur.

3.6-) Eğitim ve Test Setlerine Ayırma

- Veri seti, %80 eğitim ve %20 test olmak üzere ikiye ayrılmıştır.

4-) Modelin Eğitimi

4.1-) Model Seçimi

- **Random Forest Regressor** kullanılmıştır:
  - Birden fazla karar ağacı kullanarak çalışan güçlü bir regresyon algoritmasıdır.
  - Verilerdeki doğrusal olmayan ilişkileri öğrenmede başarılıdır.
  - Aykırı değerlere dayanıklıdır.

4.2-) Modelin Eğitimi

- Model, eğitim setindeki özellikler (**X_train**) ve hedef değişkenler (**y_train**) üzerinde eğitilmiştir.

```
▾ RandomForestRegressor
RandomForestRegressor(random_state=42)
```

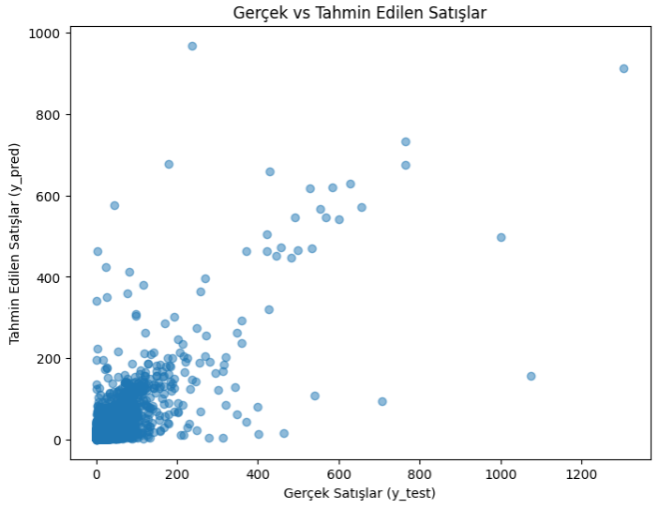
4.3-) Tahminler

- Model, test setindeki verilere (**X_test**) dayanarak satış tahminleri yapmıştır.

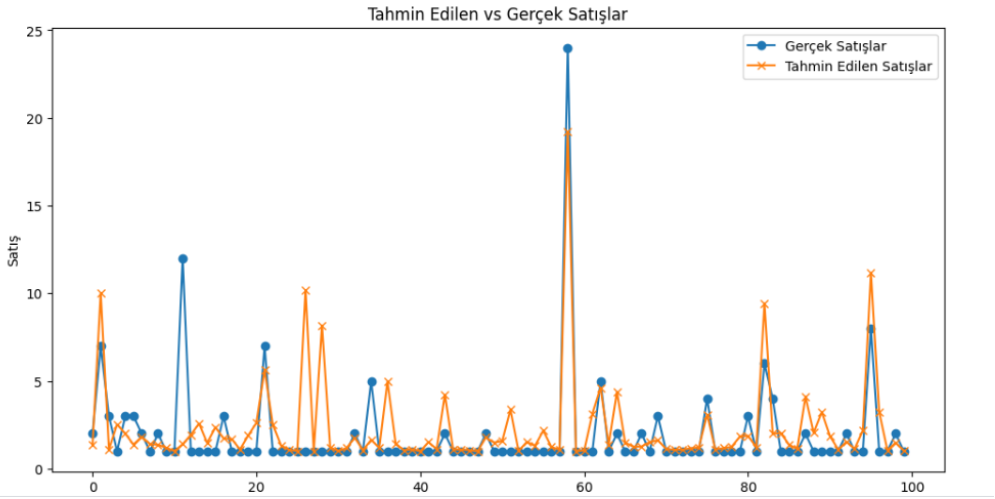
5-) Performans Analizi

Modelin performansı, çeşitli görselleştirmeler ve metriklerle değerlendirilmiştir:

- **Gerçek vs Tahmin Edilen Dağılım Grafiği**: Tahmin edilen satışların gerçek değerlerle ne kadar uyumlu olduğunu gösterir. Çoğu tahmin düşük değerler için başarılı olmuştur.



- **Tahmin Edilen ve Gerçek Satışların Çizgi Grafiği**: Modelin belirli bir veri alt kümesindeki performansını görselleştirir. Genel trendlerin takip edildiği, ancak uç değerlerde hatalar olduğu gözlemlenmiştir.



- **Performans Metrikleri:**
  - **MAE (Mean Absolute Error)**: Ortalama hata.
  - **MSE (Mean Squared Error)**: Hataların karesi ortalaması.
  - **R^2 Score**: Modelin veri varyansını ne kadar iyi açıkladığını ölçer.

6-) Projenin Sonuçları

- **Güçlü Yönler**:
  - Model, düşük satış değerlerinde oldukça başarılı sonuçlar verdi.
  - Genel trendleri yakalamada etkili oldu.
- **Zayıf Yönler**:

- Yüksek satış değerlerinde tahmin hataları gözlemlendi.
- Modelin, uç (outlier) değerler üzerinde geliştirilmesi gerekiyor.

Beyza BAL

202213171816