DATA SCİENCE BOOTCAMP BİTİRME PROJESİ

HİSSE SENEDİ TAHMİNİ

ŞEYDA TAŞDAN

1. GİRİŞ

Finansal piyasalarda hisse senedi değerlerini tahmin etmek, yatırımcılar ve finansal kurumlar için hayati öneme sahiptir. Bu projede, Nifty 500 endeksindeki hisse senetlerine ait veriler kullanılarak bir tahmin modeli oluşturulmuş ve bu modelin başarısını değerlendirmek için sentetik veri üretilmiştir.

Bu rapor, projenin adım adım ilerleyişini açıklamaktadır. Projede kullanılan veri seti, ön işleme adımları, özellik seçimi, model seçimi, sentetik veri üretimi ve sonuç değerlendirmesi ayrıntılı olarak ele alınacaktır.

Bu projenin amacı, hisse senedi değerlerini tahmin etmek için bir model geliştirmektir.

2. VERİ SETİ ve ÖN İŞLEME

Projede kullanılan veri seti, NIFTY 500 endeksinde yer alan hisse senetlerinin günlük ticaret verilerini içerir. Bu veri seti, hisse senetlerinin açılış, yüksek, düşük, kapanış fiyatları gibi günlük ticaret verileriyle birlikte sektör bilgisi gibi önemli özellikleri içerir. Veri seti Kaggle.com'dan alınmıştır.

Veri setindeki temel sütunlar şunlardır:

Company Name (Firma Adı): Hisse senedinin şirketinin adı.

Symbol (Sembol): Hisse senedi sembolü, bir menkul kıymete alım satım amacıyla atanan benzersiz bir harf dizisidir.

Industry (Sektör): Hisse senedinin ait olduğu sektörün adı.

Series (Seri): **EQ** , Eşitlik anlamına gelir. Bu seride teslimata ek olarak gün içi ticaret de mümkündür ve **BE,** Kitap Girişi anlamına gelir. Ticaretten Ticarete veya T segmentine giren hisseler bu seride işlem görür ve gün içi işlem yapılmasına izin verilmez.

Open (Açılış Fiyatı): Piyasada işlem başladığında finansal menkul kıymetin açıldığı fiyattır.

High (En Yüksek Fiyat): Hisse senedinin işlem günü boyunca işlem gördüğü en yüksek fiyat.

Low (En Düşük Fiyat): Hisse senedinin işlem günü boyunca işlem gördüğü en düşük fiyat.

Previous Close (Önceki Kapanış): Bir önceki günün kapanış fiyatı.

Last Traded Price (Son İşlem Fiyatı)**:** Günün son işlem fiyatı, gerçek son işlem fiyatıdır.

Change (Değişim): Bir hisse senedi veya tahvil kotasyonu için değişim, mevcut fiyat ile önceki günün son işlemi arasındaki farktır.

Percentage Change (Yüzde Değişim): Günlük yüzde değişim oranı.

Share Volume (Hisse Hacmi): Belirli bir zaman diliminde alınan veya satılan hisselerin toplam sayısı.

Value (Değer, Hindistan Rupisi) : Piyasa değeri (piyasa değeri olarak da bilinir) bir şirketin tedavüldeki hisselerinin mevcut piyasa fiyatıyla çarpılmasıyla hesaplanır.

52 Week High (52 Haftanın En Yüksek Değeri) : 52 haftanın en yüksek değeri, bir hisse senedinin geçen yıl içinde işlem gördüğü en yüksek hisse fiyatıdır.

52 Week Low (52 Haftanın En Düşük Değeri) : 52 haftanın en düşük seviyesi, bir hisse senedinin geçen yıl boyunca işlem gördüğü en düşük hisse fiyatıdır.

365 Day Percentage Change (365 Günlük Yüzde Değişim) : 365 günlük yüzde değişim.

30 Day Percentage Change (30 Günlük Yüzde Değişim) : 30 günlük yüzde değişim.

Eksik Veri Kontrolü:

Eksik veriler, modeli yanıltabilir ve tahminlerin doğruluğunu düşürebilir. Bu nedenle eksik verileri belirlemek ve uygun şekilde ele almak önemlidir.

Veri setindeki eksik verileri kontrol etmek için isnull() fonksiyonunu kullanıldı.. Eksik verilere sahip sütunları belirlemek için nan_columns adında bir liste oluşturuldu. Bu sütunları daha sonra doldurma amacıyla bu liste oluşturuldu. Ayrıca veri setinde eksik değerlerin yerine ‘-’ karakteri kullanıldığını farkedildiği için bu karakterleri ‘np.nan’ ile değiştirildi.

```
# Tüm veri setindeki eksik değerleri NaN ile değiştirme
df.replace('-', np.nan, inplace=True)
```

+ Code + Markdown

```
# Eksik verilere sahip sütunları kontrol etme
nan_columns = df.columns[df.isnull().any()].tolist()
print("Eksik Verilere Sahip Sütunlar:")
print(nan_columns)
```

Eksik Veri Doldurma:

Eksik verilere sahip sütunları doldurmak için medyan değerleri kullanıldı. Veri setindeki diğer değerlerden büyük ölçüde sapmış olabilecek aykırı değerlerin etkisini azaltacağı için bu yöntem kullanıldı. Örneğin, 'Change', 'Percentage Change', '365 Day Percentage Change' ve '30 Day Percentage Change' sütunlarındaki eksik değerleri medyan değerleriyle doldurulduu.

Kategorik Sütun Analizi:

'Industry', 'Company Name' ve 'Series' sütunlarındaki benzersiz değerleri ve frekanslarını analiz edildi. Bu analiz, kategorik değişkenlerin veri setindeki dağılımını anlaşılmasına yardımcı oldu.

One-Hot Encoding:

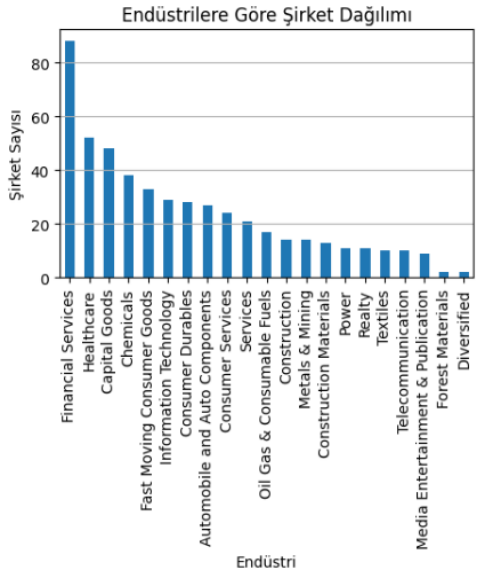
Veri setindeki kategorik değişkenleri sayısal değerlere dönüştürmek için one-hot encoding yöntemi kullanıldı. Bu yöntem, 'Industry' ve 'Symbol' sütunlarındaki kategorik değerleri binary vektörlere dönüştürür. 'Industry' sütunundaki endüstri isimlerini temsil eden yeni binary sütunlar oluşturuldu. Benzer şekilde, 'Symbol' sütunundaki hisse senedi sembollerini temsil eden binary sütunlar eklendi.

Ayrıca, 'Series' sütununu iki ayrı binary sütuna dönüştürüldü: 'Series_EQ' ve 'Series_BE'. Bu sütunlar, hisse senedi serisinin (EQ veya BE) varlığını ifade eder.

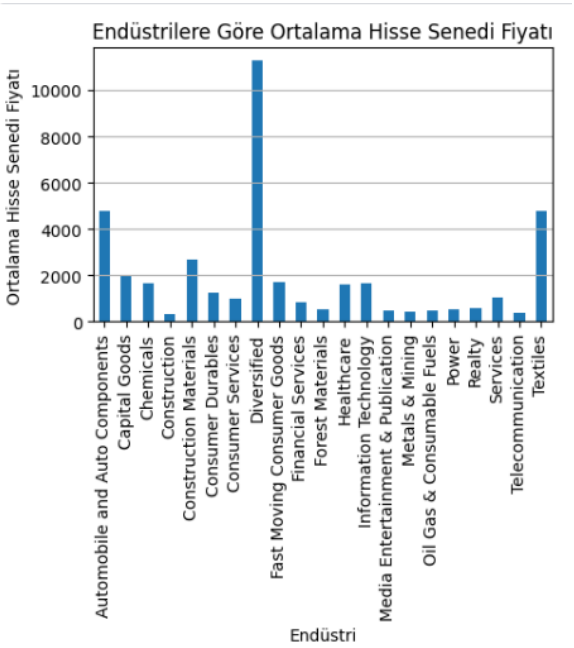
'Company Name' sütununu, hisse senedi fiyatlarının ortalamaları ile eşleştirildi. Böylece, şirketlerin isimleri yerine, hisse senedi fiyatlarını temsil edecek şekilde kullanabilir hale getirildi.

3. VERİ GÖRSELLEŞTİRME

Veri setindeki değişkenler arasındaki ilişkileri ve dağılımları anlamak amacıyla farklı görselleştirme tekniklerini kullanıldı.

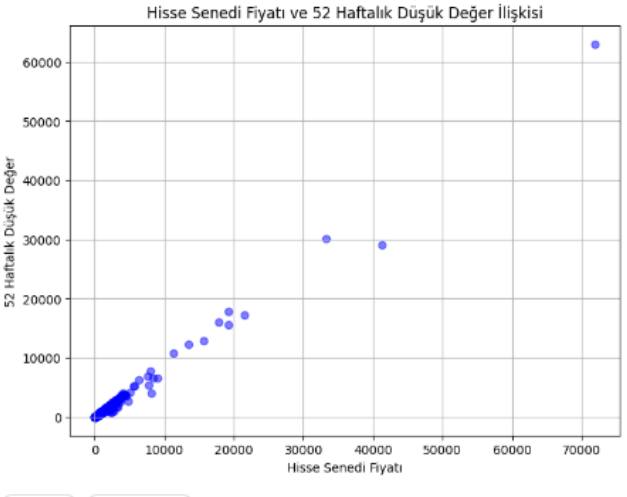


Bu grafik, her endüstrinin içinde kaç şirket olduğunu gösterir. Bu, veri setindeki endüstrilerin dengesiz olup olmadığını değerlendirilmesine yardımcı olur. Bu grafiğe göre Financial Services endüstrisi tahmini 80-100 arasında şirket sahibi olup veri setindeki en fazla şirkete sahip endüstridir.

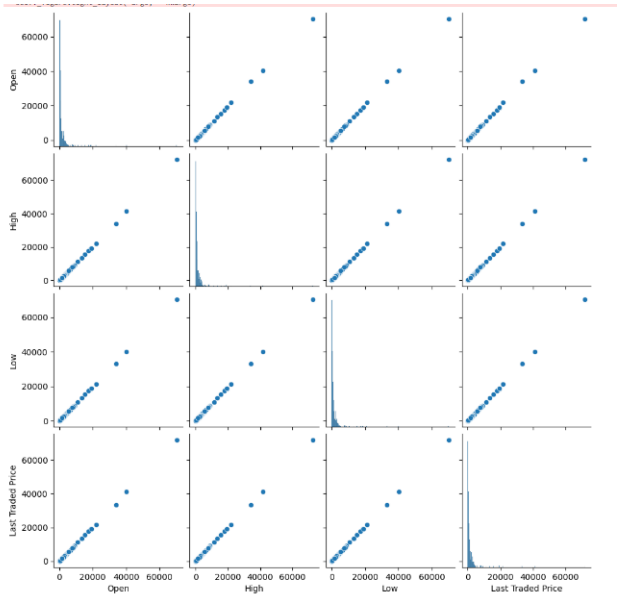


Bu grafik, her endüstrinin içindeki şirketlerin hisse senedi fiyatlarının ortalamasını gösterir. Bu, farklı endüstrilerin hisse senedi fiyatlarının genel seviyesini karşılaştırılmasına yardımcı olur.

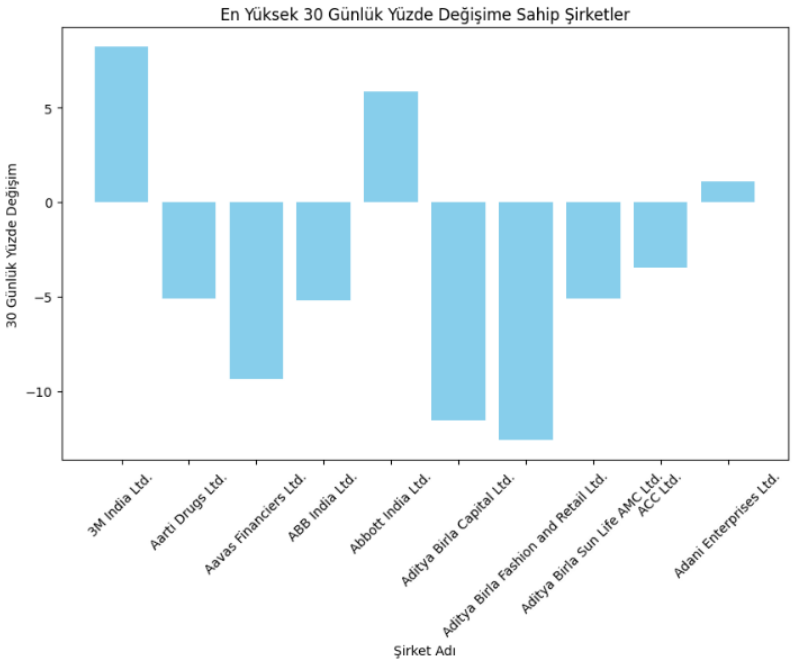
Bu grafiğe bakıldığında bakıldığında Diversified endüstrisinin ortalama hisse senedi fiyatının en yüksek değerde olduğu görünüyor.



Bu scatter plot, hisse senedi fiyatlarının ve 52 haftalık düşük değerlerin ilişkisini gösterir. Bu tür grafikler, iki değişken arasındaki ilişkiyi değerlendirmek için kullanılır. Eğer hisse senedi fiyatı ile 52 haftalık düşük değer arasında pozitif bir korelasyon varsa bu bilgi yatırımcılar için değerlidir çünkü düşük değerlerle karşılaştırıldığında mevcut fiyatlar uygun olabilir. Buna göre hisse senedi fiyatı 0-10000 arasında iken pozitif bir korelasyon vardır.



Bu grafik, hisse senetleriyle ilgili sayısal değişkenler arasındaki ilişkileri gösterir. Örneğin, açılış fiyatı ile kapanış fiyatı arasındaki ilişkiyi veya en düşük fiyat ile en yüksek fiyat arasındaki ilişki değerlendirebilir. Bu, değişkenler arasındaki olası korelasyonları veya örüntülerin görünmesine yardımcı olur.



Bu grafik, şirketler arasında en yüksek 30 günlük yüzde değişime sahip olan ilk 10 şirketi göstermektedir. Şirketlerin son 30 gün içindeki değer değişimleri, yatırımcılar için önemli bir gösterge olabilir.

4. MODEL EĞİTME

Bu projede, hisse senedi değerlerini tahmin etmek için iki farklı regresyon algoritması kullanıldı: Random Forest Regressor ve Gradient Boosting Regressor.

Bu projede, hisse senedi değerlerini tahmin etmek için iki farklı regresyon algoritması kullanıldı: Random Forest Regressor ve Gradient Boosting Regressor.

Random Forest Regressor:

Random Forest, çok sayıda karar ağacını bir araya getirerek tahminlerde bulunan ve aynı zamanda veri setindeki önemli özellikleri belirleyen bir algoritmadır. Yüksek doğruluk, düşük overfitting oranı ve açıklanabilirlik gibi avantajlarıyla bilinir. Bu özellikler, hisse senedi değerlerini tahmin etmek için kullanılması için uygun kılar.

Gradient Boosting Regressor:

Gradient Boosting, zayıf öğrenciler (genellikle karar ağaçları) kullanarak tahmin performansını artırmayı amaçlayan bir algoritmadır. Gradient Boosting, önceki ağaçların hatalarını düzelterek yeni ağaçlar ekler ve bu şekilde tahmin doğruluğunu artırır. Overfitting eğiliminde olmasına rağmen, uygun şekilde ayarlandığında oldukça güçlü tahminler yapabilir.

Her iki algoritmanın sentetik veri üzerindeki performansını değerlendirmek için 10 farklı sentetik veri noktası oluşturuldu. Bu sentetik veriler oluşturulurken rand() fonksiyonunu kullanarak 0 ile 1 arasında rastgele sayılar üreterek sentetik veri oluşturuldu. Bu sayılar, gerçek veri setindeki sayısal değişkenlerin aralıklarına benzer değerler içeriyor. Bu şekilde, hisse senetlerinin günlük ticaret verilerine benzer sentetik veri örnekleri elde edildi.

Bu değişkenleri oluştururken, gerçek veri setindeki dağılımlar düşünülerek veri setinde de benzer bir dağılım hedeflendi.

Bu sentetik veriler, projede kullanılan gerçek veri setiyle aynı özelliklere sahip. Her iki algoritmanın sentetik veri üzerinde yaptığı tahminleri değerlendirmek için Mean Squared Error (Ortalama Kare Hata), R-squared (R-kare) ve Explained Variance

Score (Açıklanan Varyans Skoru) gibi metrikleri kullanıldı. Bu metrikler, her algoritmanın tahminlerinin gerçek değerlere ne kadar yakın olduğunu ölçer.

5. SONUÇLAR

Random Forest algoritması ile oluşturulan sentetik veri tahminleri şu şekildedir:

```
Sentetik Veri Tahminleri:
[3605.5585 3605.4095 3605.426    10.699    10.9885 3605.2025 3605.291
   10.8215 3605.3375 3605.4455]

Mean Squared Error Değeri: 441248.9494392386
R-squared (R-kare) Değeri: 0.978473635180587
Explained Variance Score: 0.9785438847534977

True/False Tablosu:
     Gerçek Değer  Tahmin Edilen Değer  True/False
338        131.40             134.9575       False
498         56.00              55.6120        True
212          8.50               9.4260       False
214         32.40              32.1310       False
383       2406.00            2439.8175       False
..            ...                 ...         ...
349       4111.00            4162.1715       False
401       2651.00            2659.5280       False
89        4111.25            4096.6365       False
203      33244.05           33244.9640       False
275       9068.00            8924.1750       False

[501 rows x 3 columns]
```

Gradient Boosting algoritması ile oluşturulan sentetik veri tahminleri şu şekildedir:

```
Gradient Boosting Sentetik Veri Tahminleri:
   Tahmin Edilen Değer
0            41.388602
1            74.798599
2            53.910320
3            52.894218
4            68.000433
5            37.014486
6             6.524674
7            54.794697
8            42.315969
9            28.476334
Gradient Boosting Gerçek Veri Test Sonuçları:
Mean Squared Error Değeri: 7662808.777866686
R-squared (R-kare) Değeri: 0.7924762695983428
Explained Variance Score: 0.7935861224785223

Gradient Boosting True/False Tablosu:
     Gerçek Değer  Tahmin Edilen Değer  Doğru/Tuş
433        298.50           288.418052      False
186       2205.30          2161.521762      False
99        1458.00          1510.270641      False
86          16.80            27.233721      False
98         948.00           896.491121      False
..            ...                 ...        ...
443        413.70           391.752081      False
22         469.50           476.289617      False
320        649.90           634.965146      False
453       2036.25          2100.650957      False
432          6.80            25.836037      False

[151 rows x 3 columns]
```

Random Forest algoritmasıyla oluşturulan sentetik verilerle yapılan tahminler oldukça başarılı oldu. Sentetik verilere uygulanan modelle elde edilen R-squared değeri 0.978, Mean Squared Error değeri ise 441248.95 olarak hesaplandı. Ancak, gerçek veriler üzerinde yapılan tahminlerde bazı hatalar ortaya çıkmış gibi görünüyor. Bu, modelin gerçek verilere uyum sağlama konusundaki sınırlamalarını gösteriyor.

Gradient Boosting algoritmasıyla oluşturulan sentetik verilerle yapılan tahminlerin performansı Random Forest'a göre biraz daha düşük oldu. Sentetik verilere uygulanan Gradient Boosting modeliyle elde edilen R-squared değeri 0.792, Mean Squared Error değeri ise 7662808.77 olarak hesaplandı. Bu sonuçlar, Gradient Boosting algoritmasının veriye daha az uyum sağlayabildiğini gösteriyor.

Her iki algoritmanın da gerçek verilere uygulanmasıyla elde edilen tahminlerde, özellikle düşük değerli hisse senetlerinde doğruluk oranının düştüğünü gözlemlendi. Sentetik verilerle yapılan tahminlerdeki başarı, gerçek dünya verileriyle uygulandığında azalmaktadır. Bu durum, modelin gerçek dünya verilerine uyum sağlama yeteneğinin sınırlı olduğunu göstermektedir.