Gerçek Zamanlı Stok Analizi ve Tahmin Aracı Raporu

1. Projenin Amacı

Bu proje, kullanıcının belirli bir hisse senedine (S&P500 şirketlerinden biri) bugün yapacağı yatırımın geçmiş ve gelecek performansını analiz etmek amacıyla geliştirilmiştir. Kullanıcılar, seçtikleri hisse senedine yapacakları yatırımın:

- **Geçmiş Performansı:** 1 hafta, 1 ay, 3 ay, 6 ay ve 1 yıl önce yapmış olması durumunda bugünkü değerini ve
- **Gelecek Tahmini:** 1 hafta, 1 ay, 3 ay, 6 ay ve 1 yıl sonrası olası değerinin tahminini elde ederler.

Uygulama, gerçek zamanlı hisse senedi verileri ve makine öğrenimi modellerini kullanarak kullanıcıya kapsamlı bir yatırım analizi sunmayı amaçlar.

2. Veri Toplama ve İşleme Süreci (Data Collection & Preprocessing)

2.1 Veri Kaynakları ve Yapısı:

Projede aşağıdaki veri kaynakları kullanılmaktadır:

- **Hisse Senedi Verileri:** Yahoo Finance API kullanılarak seçilen hisse senedinin geçmiş fiyat verileri elde edilmiştir.
- **Döviz Kuru Verileri:** Open Exchange Rates API aracılığıyla USD/TL döviz kuru verileri elde edilmiştir.

Elde edilen veriler; tarih, açılış fiyatı, kapanış fiyatı, en yüksek, en düşük, hacim gibi temel teknik finansal göstergeleri içermektedir.

2.2 Veri Ön İşleme Adımları:

Veri ön işleme sürecinde aşağıdaki adımlar uygulanmıştır:

- **Tarih Formatlama:** Veri setindeki tarih sütunları, analiz ve modelleme için uygun bir biçime dönüştürülmüştür.
- **Eksik Değerlerin İşlenmesi:** Eksik veya hatalı veriler tespit edilerek uygun yöntemlerle (‘inf’ değerlerin lineer interpolasyon ile doldurulması) doldurulmuştur.

- **Özellik Mühendisliği (Feature Engineering):** Modelin performansını arıtmak için aşağıdaki teknik göstergeler hesaplanmıştır:
    - **Hareketli Ortalamar (Moving Average):** 5, 21, 63, 126 ve 252 günlük periyotlar için (tatil günleri dikkate alınmıştır) basit hareketli ortalamalar hesaplanmıştır.
    - **Üstel Hareketli Ortalamalar (Exponential Moving Average):** 5, 21, 63, 126 ve 252 günlük periyotlar için (tatil günleri dikkate alınmıştır) üstel hareketli ortalamalar hesaplanmıştır.
    - **Göreceli Güç Endeksi (Relative Strength Index - RSI):** 5, 21, 63, 126 ve 252 günlük periyotlar için (tatil günleri dikkate alınmıştır) RSI değerleri hesaplanmıştır.
    - **Volatilite Ölçümleri:** Fiyatların standart sapmaları hesaplanarak volatilite bilgileri hesaplanmıştır.
    - **Günlük Yüzdelik Değişim:** Kapanış fiyatlarının günlük yüzdelik değişimi hesaplanmıştır.
    - **Günlük İşlem Hacmi (Volume) Yüzdelik Değişimi:** Günlük işlem hacminin (Volume) yüzdelik değişimi hesaplanmıştır.

3. Modelleme Süreci (Modelling Process)

3.1 Denenen Modeller ve Performans Karşılaştırmaları

Modelleme aşamasında her şirket için ayrı ayrı 12'şer makine öğrenmesi modeli eğitilmiş ve her şirket için en iyi sonucu veren makine öğrenmesi modeli seçilmiştir. Denenen modeller aşağıdaki gibidir:

- **Doğrusal Regresyon (Linear Regression):** Temel bir model olarak kullanılmıştır. En yüksek ortalama test skoruna sahip modellerden biri olup (0.9879), başarılı tahminler yapmıştır. **FI (Fiserv, Inc.)** kodlu şirkette en iyi model olarak seçilmiş ve 0.9986 test skoruna ulaşmıştır.
- **Ridge ve Lasso Regresyon:** Aşırı uyumu (overfitting) önlemek için kullanılmıştır. Ridge Regresyon, en yüksek test skoruna ulaşan model olup (0.9989), **NI (NiSource Inc.)** kodlu şirkette en iyi model olarak belirlenmiştir. Lasso Regresyon ise **TYL (Tyler Technologies, Inc.)** kodlu şirkette 0.9942 test skoru ile en iyi model olmuştur.
- **Destek Vektör Regresyonu (Support Vector Regression - SVR):** Lineer, Polinomal ve Radyal Bazlı Fonksiyon (RBF) çekirdekleri ile denenmiştir. Linear SVR, **FOX (Fox Corporation)** kodlu şirkette en iyi model olarak belirlenmiştir ve 0.9984 test skoruna ulaşmıştır. RBF SVR, **O (Realty Income Corporation)** kodlu şirkette en iyi model olarak 0.9717 test skoruna ulaşmıştır.

- **Karar Ağaçları (Decision Trees):** Hızlı ve yorumlanabilir bir model olarak kullanılmıştır. **HES (Hess Corporation)** kodlu şirkette en iyi model olarak belirlenmiş olup test skoru 0.7999 olarak hesaplanmıştır. Ancak değişkenliğin yüksek olduğu şirketlerde düşük performans göstermiştir.
- **Rastgele Ormanlar (Random Forests):** Birden fazla karar ağacının birleşimi ile daha güçlü bir model oluşturulmak amacı ile kullanılmıştır. Ortalama test skoru 0.9804 olup, **EQR (Equity Residential)** şirketinde en iyi model olarak seçilmiştir.
- **Gradient Boosting ve AdaBoost:** Zayıf öğrenicilerin birleştirilmesiyle güçlü tahmiciler elde edilmesi amacı ile kullanılmıştır. Gradient Boosting, **CHRW (C.H. Robinson Worldwide, Inc.)** kodlu şirkette 0.9858 test skoruyla en iyi model olmuştur. AdaBoost, KEY şirketinde 0.9541 test skoru ile başarılı olmuştur.
- **K-En Yakın Komşu (K-Nearest Neighbors - KNN):** Veri noktalarının yakınlıklarına dayalı tahminler yapılması amacı ile kullanılmıştır. **AVB (AvalonBay Communities, Inc.)** kodlu şirkette en iyi model olarak belirlenmiş ve 0.960 test skoruna ulaşmıştır. Ancak yüksek volatiliteye sahip şirketlerde doğruluk oranı düşük kalmıştır.

Her model, eğitim ve test verileri üzerinde değerlendirilmiş ve performans metrikleri (R^2, MSE, RMSE, MAE, MAPE ve her periyot için Walk-Forward Validation) hesaplanmıştır.

3.2 Neden Sektör Bazlı Model Seçimi Yapılmadı?

Model seçiminde sektör bazlı optimizasyon yapılmamıştır çünkü şirketlerin finansal hareketliliği, sektör içindeki diğer firmalara kıyasla farklılık göstermektedir. Örneğin, aynı sektörde bulunan iki firma arasında fiyat dinamikleri ve volatilite büyük değişkenlik gösterebilmektedir. Bu yüzden her şirket için bireysel bazda en iyi model belirlenmiş ve sektör bazlı genellemelerden kaçınılmıştır.

4. Model Performans Analizi (Evaluation & Performance)

4.1 Walk-Forward Validation ile Model Seçimi

Doğru model değerlendirmesi yapabilmek için Walk-Forward Validation (WFV) yöntemi kullanılmıştır. WFV, özellikle zaman serisi verilerinde modelin geçmiş verilere dayalı olarak gerçekçi tahminler yapmasını sağlamak amacı ile tercih edilmiştir. Geleneksel çapraz doğrulama yöntemleri, bağımlı değişkenin zaman bağımlılığını ihmal edebileceğinden dolayı, WFV ile ilerleyen zaman aralıklarında modelin performansı adım adım test edilerek değerlendirilmiştir. Bu sayede her şirketin yatırım performansı geçmiş veriler üzerinden tutarlı bir şekilde analiz edilmiştir:

- Her şirket için kurulan 12'şer model için farklı zaman aralıklarında (1 hafta, 1 ay, 3 ay, 6 ay ve 1 yıl) tahmin yapılmıştır ve bu tahminlerin hata oranları karşılaştırılarak en iyi model seçilmiştir.
- **Uzun vadeli tahminlerde** en düşük hata oranına sahip modeller Random Forests (WF_RMSE_1yıl = 2.32) ve GradientBoosting (WF_RMSE_1yıl = 3.79) olmuştur.
- **Orta vadeli tahminlerde (3 ay – 6 ay),** en iyi performansı gösteren modeller Random Forest (WF_RMSE_3ay = 2.32) ve Gradient Boosting (WF_RMSE_3ay = 2.89) olmuştur.
- **Kısa vadeli tahminlerde (1 hafta - 1 ay),** en düşük hata oranına sahip modeller Linear Regression (WF_RMSE_1ay = 1.25) ve Lasso Regression (WF_RMSE_1ay = 1.49) olmuştur.
- KNN, uzun vadeli tahminlerde en düşük hata oranına (WF_RMSE_1yıl = 1.94) sahip olmasına rağmen bazı şirketlerde çok zayıf performans göstermiştir.
- Bazı şirketlerde ise ML modelleri başarısız olmuş ve alternatif olarak LSTM veya hibrit modeller denenmesi için daha sonraya bırakılmıştır.

4.2 LSTM Neden Başarısız Oldu?

Başlangıçta LSTM ile tahmin denemeleri yapılmıştır ancak aşağıdaki sebeplerle beklenen performans elde edilememiştir. Bu durumun sebepleri aşağıdaki gibidir:

- **Düşük Veri Miktarı:** S&P500'deki her şirket için 7 yıllık (özellik mühendisliği sonrası 5 yıllık) günlük veri ile yeterli olsa da zaman serisi tahminlerinde LSTM modellerinin genellikle daha fazla veriye ihtiyaç duyması.
- **Yüksek Varyans ve Overfitting:** Bazı şirketlerde LSTM modellerinin train skoru 0.99 ancak test skoru 0.6 olarak hesaplanmıştır.
- **Modelin Finansal Piyasalardaki Kaotik Yapıyı Yakalayamama İhtimali:** ML modelleri, LSTM'ye kıyasla daha genelleştirilebilir çıktılar vermiştir.
- **Eğitim Süresi: BAX (Baxter International Inc.)** kodlu şirkette yapılan eğitim yaklaşık olarak 3 saat sürdüğü için diğer şirketler üzerinde eğitim yapılması yerine özellik mühendisliğine odaklanılmıştır.

5. Boyut İndirgeme ve Özellik Seçimi

Başlangıçta, **Temel Bileşen Analizi (Principal Component Analysis - PCA)** kullanılarak boyut indirgeme denemeleri yapılmıştır ancak PCA'nın bağımsız değişkenler ile hedef değişken arasındaki ilişkiyi tam olarak koruyamadığı ve zaman serilerindeki ardışıklık ilişkisini göz ardı ettiği gözlemlenmiştir. Bu nedenle, **Kısmi En Küçük Kareler Regresyonu (Partial Least Squares - PLS)** yöntemi uygulanarak modelin hedef değişkenle en iyi ilişkili öznitelikleri koruması sağlanmıştır. Bunun sonucunda:

- Yapılan testlerde, PCA uygulandığında modelin performansı ortalama $R^2 = 0.65$ seviyesinde kalırken, PLS uygulandığında bu değer $R^2 = 0.92$ seviyesine yükselmiştir. Bu nedenle PLS kullanımı tercih edilmiştir. Böylece model doğruluğu artırılmış, overfitting azaltılmış ve öznitelikler azaltılarak modelin daha hızlı çalışması, gereksiz hesaplama yapılmasının önüne geçilmiştir.

6. Uygulama ve Kullanılabilirlik (Deployment & Usability)

6.1 Flask ile Web Uygulaması

- Kullanıcılar web arayüzü üzerinden; yatırım miktarını, para birimini (USD veya TL) ve hisse kodunu girerek hem geçmişe yönelik analiz hem de geleceğe yönelik tahminleri elde edebiliyor.
- Hisse senedi verileri; kullanıcı girişi yaptığı anda, anlık olarak çekiliyor ve kodu girilen şirket için seçilmiş olan model, güncel veri ile eğitiliyor.

6.2 Pipeline Yapısı ve Model Seçimi

- Pipeline ile her şirket için en uygun model otomatik olarak seçiliyor.
- En iyi modeller bir **.pkl** dosyasına kaydediliyor ve her çalıştırıldığında en uygun model yüklenerek tahmin yapılıyor. Yeni model eğitmek yerine, her seferinde güncel verilerle model eğitiliyor.

7. Geliştirlebilecek Noktalar

- MLOps süreçleri eklenebilir (CI/CD, model monitoring, model güncelleme).
- Elle model seçimi yerine AutoML ile her şirket için en uygun model otomatik ve dinamik olarak seçilebilir.
- Dışsal faktörler (ekonomik göstergeler, haber duyarlılığı) modele entegre edilebilir.
- Veri görselleştirmeleri eklenerek analizler daha anlaşılır hale getirilebilir.
- LSTM ve Transformer tabanlı modeller daha geniş veriyle test edilebilir.