Team Member: Nihar Muniraju, Sushma Nagula, Xuyang Ji

Project Scope

- Ticker = ["AAPL", "MSFT","GOOG","INTC","NVDA"]
- The training and validation set contains data
  - start_date = "2019-01-01" and end_date = "2023-12-31" from Yahoo Finance
  - Training 70%, Validation 30%, with random seed: 961
- The test set contains data from 2024-01-01 to 2024-04-01

Progress Update:

Machine Learning Models

We are developing an advanced stock price prediction model by leveraging a wide array of technical indicators and machine learning algorithms. We are trying to capture various market trends and volatility.

By implementing a diverse set of regression models such as ridge, SVR, Random forest, AdaBoost, etc., to evaluate and compare their performance using RMSE to identify the best-performing model. Additionally we are using integrated exponential smoothing methods to remove noise and to produce smooth future predictions, enhancing the accuracy of our forecasts. Finally we have visualised both historical and future stock predictions to provide clear insights into market dynamics and model predictions.

Next Step:

Next we are going to validate it using various optimization techniques and use different feature selectors to see

how well we can fit into the model to achieve a good accuracy and precision. We are planning our own strategies to check how well the stock with different volatility can be observed and then do some back-testing strategy to have a good performing model.

SARIMA

Our team has been diligently working on fine-tuning the SARIMA model for forecasting. The data underwent preprocessing steps, including transformation and scaling to ensure its suitability for modeling. Prior to the model development, we conducted extensive data exploration to gain insights into the underlying trends of each stock in the portfolio with Dash and Matliplot. Rolling statistics and decomposition were employed to uncover the structure in the data.

To identify the period with the most consistent residual value across all stocks, we employed seasonal decomposition analysis, with a focus on analyzing the stability of residual values over different periods. We then identified the time period with the lowest variability in the residual values, indicating the most consistent period across the portfolio. With insights from seasonal decomposition, we proceeded to apply auto_sarima to determine the optimal seasonal parameters on the training and validation set and evaluate its performance on test data. Performance metrics such as MAE, MSE were calculated to assess the accuracy of the model prediction.

Next Step:
Moving forward, we plan to incorporate the technical indicators in the ML regression models to SARIMA and compare models' predictive capability. In addition, sensitivity analysis will be conducted to assess the robustness of the SARIMA model under different market conditions.

Sentiment Analysis

Our project at the convergence of finance and NLP is progressing through the model building stage, aiming to enhance stock price forecasting with sentiment analysis. Currently, leveraging the threading concepts successfully we implemented preprocessing pipelines for news headlines and integrated a pre-trained DistilBERT model for sentiment classification. However, we face challenges with scarcity of positive sentiment data and data standardization across sources.

Next Step:
Moving forward, we aim to address bias in sentiment analysis, refine model integration, and evaluate the impact on forecasting accuracy, ultimately advancing the predictive capabilities of our pipeline.