Group 75 – Final Project Report

Team Members

- Ngoc Tu Pham; npham72
- Kaixuan Ma, kma304
- Ari Jones; aridjones
- Anirudh Kolapalli, akolapalli3
- Qaiyim Harris; qharris7

Background Information and Problem Overview

The COVID-19 pandemic has led to increased volatility in the stock market, as evidenced by the fluctuations in the S&P 500 index. This volatility poses challenges for investors seeking to manage risk and protect their investments. The SPDR S&P 500 Trust ETF (SPY), which closely tracks the S&P 500 index, has become a popular investment vehicle for those looking to gain exposure to the broader market.

Accurate prediction of SPY prices can provide investors with valuable insights to make informed decisions. By anticipating market trends, identifying optimal entry and exit points, and managing risk effectively, investors can potentially enhance their returns and mitigate losses during periods of market uncertainty.

However, predicting stock prices accurately is a complex task that requires analyzing various factors, such as economic indicators, company fundamentals, and market sentiment. Traditional forecasting methods may not always capture the intricate relationships and patterns within the data, leading to suboptimal predictions.

This project aims to address this challenge by leveraging machine learning techniques and relevant data to develop robust models for predicting SPY ETF prices accurately. Our goal is to empower investors with actionable insights to make well-informed decisions and optimize their investment strategies in the face of market uncertainties.

Analytics Approach

Our approach involves several key steps:

1. Data Collection and Preparation: We compile a comprehensive dataset including historical prices of the SPY ETF, its trading volumes, and different economic indicators (e.g., FED interest rates, CPI/inflation rates, GDP growth, and unemployment rates). This dataset undergoes rigorous data cleaning and preprocessing to ensure quality and consistency.
2. Exploratory Data Analysis: We explore characteristics of the variables in the dataset as well as their potential relationships. Data visualization is used to assist this process. We may also do feature engineering to explore new features from the existing dataset that can provide additional insights.
3. Model Exploration: We start with multi-variate linear regression and then explore other analytics models such as decision tree, random forest, gradient boosting, neural network, and time series analysis.
4. Dimensionality Reduction: We deal with potential multicollinearity by keeping most relevant predictors and exploring PCA to enhance model performance.
5. Model Evaluation and Refinement: We employ varied metrics such as R2, RMSE, and MAE and validation techniques to assess and refine our models. This iterative process will help us identify the most effective models for predicting SPY ETF prices.

Initial Hypotheses

- Efficient Market Hypothesis

With information technology advancement and increasingly powerful algorithmic trading, the SPY ETF price is supposed to be very efficient. However, there can be a possibility that the price does not consistently behave as predicted due to randomness or other influential predictor variables not yet considered.

- Behavioral Finance

Investor behavior and psychological biases can influence stock price movements, leading to deviations from rational expectations. There can be oversold or overextended price movements due to fear of missing out (FOMO). This may be relevant to explaining erratic price movements.

- Fundamental Analysis

Fundamental analysis (e.g., P/E ratio) can be proved ineffective to predict stock prices. Empirical evidence shows that it is relative and that while it can drive stock price movements over the long term, it can fail to explain short term fluctuations.

- Technical Analysis

Technical indicators such as volume data, historical price (via support and resistance) should play a role in influencing stock prices, including the SPY ETF. Technical analysis can help explain short term deviations from the long-term trend.

Overview of Data

Variable	Overview / Description	Data Source
SPY ETF Price	Adjusted daily closing price of the SPY ETF that tracks the performance of the S&P 500 index.	R library tidyquant
SPY Volume	Average daily trading volume of the SPY ETF	
GLD Price	Adjusted daily closing price of SPDR Gold Trust that invests in physical gold.	
CPI	Consumer price index readings reflect the level of inflation in the US.	FRED St. Louis FED https://fred.stlouisfed.org/
GDP	US's gross domestic product index	
FED Fund Rate	Federal Reserve fund rate as a tool to support monetary policies	
Jobs Opening	Number of job openings in the US	
Population	The population of the US	
Unemployment Rate	US's unemployment rate overtime	
PCE	Personal consumption expenditures, also known as consumer spending, is a measure of the spending on goods and services by Americans.	

Data Cleaning and Preprocessing

- Data joining: The data available to download on the FRED website need to be joined together by month. Most of these economic data are reported monthly.

- Data aggregation: The adjusted closing prices and trading volumes of SPY and GLD are daily. Thus, we calculated the monthly averages of these variables to merge with the economic data.

- NA filling: Unlike the other economic data, the GDP data is measured and reported quarterly. For the months that are not reported, we take the average of the right after and right before quarterly GDP figures.

- Data normalization: For multi-variate linear regression, we normalize variables in ranges [0; 1]

- Data scaling for PCA: Independent variables are scaled around the mean and standard deviation.

Insights from Exploratory Data Analysis

- Correlation

	Avg. SPY Volume	GDP	CPI	FED Rate	GLD Price	Jobs Opening	Population	PCE	Unemployment
Avg. SPY Price	-0.32	0.93	0.93	0.50	0.91	0.90	0.91	0.93	-0.21

Please also refer to Figure 1: Scatter Plots – Average SPY Price and Normalized Variables and Figure 2: Normalized SPY Prices and Normalized Variables by Month in the Visualizations folder in our GitHub. We combine our economics knowledge and understanding of the stock market for the past 10 years to draw useful insights below.

- SPY price vs. SPY volume

  The SPY volume tends to spike in periods when the SPY price experiences sharp declines. On the other hand, in periods when the SPY price increases steadily, the SPY volume fluctuates mildly. Behavioral finance can explain this. When stock prices decrease sharply, investors become fearful and want to sell stocks before other investors do to preserve their wealth. Stop-loss orders set up as part of risk management are also triggered, accelerating the price decline and trading volumes.

- SPY price vs. GDP, CPI, PCE

  The patterns of GDP, CPI, and PCE to SPY price are very similar. Their movements to SPY become more erratic from 2020 before converging in early 2024. This can be explained by the impact of FED fund rate changes on the S&P 500 since 2020 as a response to COVID-19.

- SPY price and FED fund rate

  In March 2020, both the SPY price and FED fund rate plunged due to the rapidly expanding COVID-19 pandemic. As the risk of lockdown and economic activity stoppage became elevated, investors sold their equity shares, causing the decline in the SPY price. To support the economy, the FED lowered interest rates to 0%/0.25% and kept this rate low until early 2022.

  Under that low interest rate environment that makes costs of borrowings and equity premium lower and thanks to the successful discovery and fast manufacturing of COVID-19 vaccines later in 2020, economic

activities quickly recovered, the SPY price rose sharply. Then as inflation (measured by CPI) ran out of control, in 2022 the FED began rapidly increasing interest rates to combat it.

Increases in interest rates contract economic activities and lower equity multiples. As a result, the SPY price fell until the breakthrough in Artificial Intelligence with OpenAI ChatGPT fueling optimism in future productivity increases and economic growth. The slowdown of inflation also helped boost SPY price targets as when inflation is under control, the FED will lower interest rates to foster economic activities.

- SPY price and GLD price

  GLD tends to outperform SPY during periods of economic uncertainty (for example, during the COVID-19 spread in 2020) as investors look for safe assets. Although there can be contradictory movements between them, in the long term they both increase in value and positively correlate with each other.

- SPY price and number of jobs opening

  Larger numbers of jobs opening signal economic booms, and during economic expansion, SPY price also rises as companies experience profitability increases. However, there is a divergence between them starting 2023. As jobs opening continued to decrease due to tightening monetary policies (FED rate hikes since 2022) in 2023, SPY recovered its 2022 selloff and climbed higher thanks to economic optimism from AI breakthrough (e.g., large language models like OpenAI ChatGPT) and continued GDP growth.

- SPY price and population

  A higher population means more labor in the workforce fosters economic activities, increasing stock prices.

- SPY price and unemployment rate

  The few months in 2020 were extreme with high unemployment rates due to COVID-19 lockdowns. This also supports the general negative correlation between unemployment and SPY price. High unemployment rates mean the economy is in contraction, lowering business profitability and SPY price.

Analytics Modeling Results

- Multi-variate linear regression

First, using normalized data, we modelled SPY Price on all the nine available predictor variables. Due to multicollinearity (e.g., VIF value of 4,123 for CPI, and VIF value of 5,314 for PCE), the resulting coefficients are too extreme (e.g., -1,419.41 for CPI and 1815.62 for PCE). In addition, the negative coefficient signs of CPI and GDP are unreasonable compared to the positive correlation between them and SPY price.

After removing highly correlated variables with extreme VIF values (i.e., PCE, CPI) and FED rate which is too statistically insignificant, we come up with a linear regression model with six normalized predictor variables.

model <- lm(AVG_PRICE~AVG_VOLUME+GDP+JO+POP+UNRATE+GLD_PRICE, data=data_normalized)

The model has R2 0.9665, indicating good quality. VIF values are also sharply lower than those of the original l model with nine variables. The coefficients also are reasonable with the actual correlation between SPY price and the six variables.

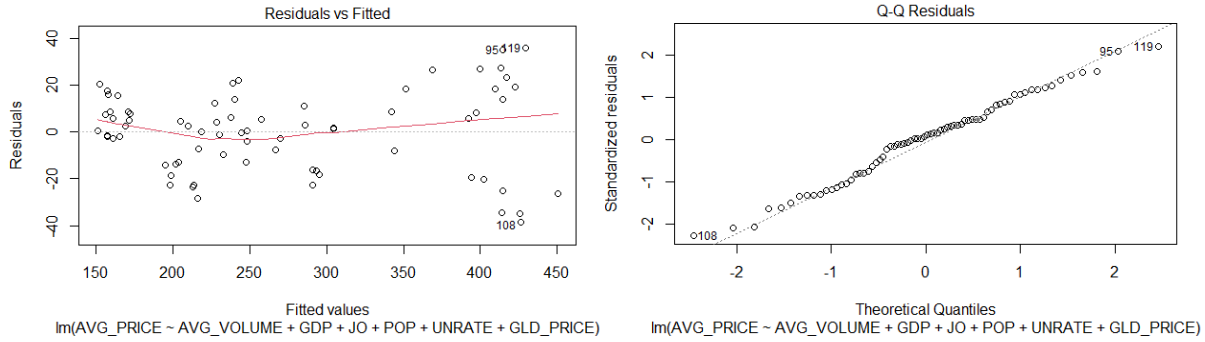
Thus, we decided to finalize a linear regression model with these six predictor variables, performing a train-validation-test split. We set seed in R as 68 for reproductivity and randomly allocate 60%, 20%, and 20% respectively for the train, validation, and test data sets.

The summary of the finalized linear regression model on the train data set is as follows.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.649      8.074  15.191  < 2e-16 ***
AVG_VOLUME   -13.209     18.085  -0.730    0.468
GDP           14.244     22.908   0.622    0.536
JO           121.850     17.991   6.773 4.37e-09 ***
POP           86.788     16.777   5.173 2.40e-06 ***
UNRATE       -14.872     18.675  -0.796    0.429
GLD_PRICE    138.248     19.568   7.065 1.33e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.88 on 65 degrees of freedom
Multiple R-squared:  0.9684,    Adjusted R-squared:  0.9655
F-statistic: 332.4 on 6 and 65 DF,  p-value: < 2.2e-16
```

R2 is 0.9684, together with the F-statistics, indicates this model is a good fit. The Residual vs Fitted graph and the Q-Q residuals also show no heteroscedasticity and residuals are normally distributed.



Using this model on the validation and test data sets, we got the R2 results as 0.9635 and 0.9500 respectively, indicating that the model performs well in predicting SPY price.

- Multi-variate linear regression with PCA

We also explored PCA for linear regression as another method to deal with multicollinearity. Out of nine principal components, the first four take up 96.71% of the variance. The linear regression model based on these four principal components gives an R2 of 0.963. Thus, using PCA does not result in a better model.

Overall, for linear regression, our selected model is the non-PCA and more straightforward model with the six predictor variables: Average SPY volume, GDP, Jobs opening, Population, Unemployment rate, and GLD price. For more details and coding, please refer to “Linear Regression.R” in the Code folder in our GitHub.

- Decision Tree

To build a decision tree regression model, our approach involves feature selection and engineering, hyperparameter tuning, model optimization, and training and evaluation of the decision tree model.

We select relevant features (CPI, GDP, FEDRATE, JO, PCE, POP, UNRATE, AVG_VOLUME) and generate polynomial features of degree 2 to capture potential non-linear relationships. Then we split the data into training and testing sets, reserving 20% of the data for testing. For hyperparameter tuning, we use GridSearchCV with a reduced range of hyperparameters and a 5-fold cross-validation strategy.

Then, we train the decision tree model using the best hyperparameters found during tuning before evaluating the model's performance on the test set using various metrics such as mean squared error, R^2 score, mean absolute deviation, mean absolute percentage error, and Theil's inequality coefficient. Finally, we analyze feature importances to identify the most influential features in predicting the average price of the SPY ETF.

Our decision tree model achieved the following performance metrics on the test set:

- Mean Squared Error: 235.69
- R^2 Score: 0.96
- Mean Absolute Deviation: 11.13
- Mean Absolute Percentage Error: 3.86%
- Theil's Inequality Coefficient (Theil's U): 0.03

These results indicate that the decision tree model provides a good fit to the data and can predict the average price of the SPY ETF with reasonable accuracy. The high R^2 score suggests that the model explains a significant portion of the variance in the target variable. The low mean absolute percentage error and Theil's inequality coefficient further confirm the model's predictive performance.

- Random Forest

The approach is similar to approach to decision tree regression; however, we train the random forest model with 100 trees using the best hyperparameters found during tuning.

Our random forest model achieved the following performance metrics on the test set:

- Mean Squared Error: 141.30
- R^2 Score: 0.98
- Mean Absolute Deviation: 7.47
- Mean Absolute Percentage Error: 2.72%
- Theil's Inequality Coefficient (Theil's U): 0.02

These results show that the random forest model outperforms the decision tree model in predictive accuracy. The higher R^2 score and lower error metrics indicate that the random forest model provides a better fit to the data and can predict the average price of the SPY ETF more accurately. The ensemble nature of random forests, which combines multiple decision trees, contributes to its superior performance compared to the individual decision tree model.

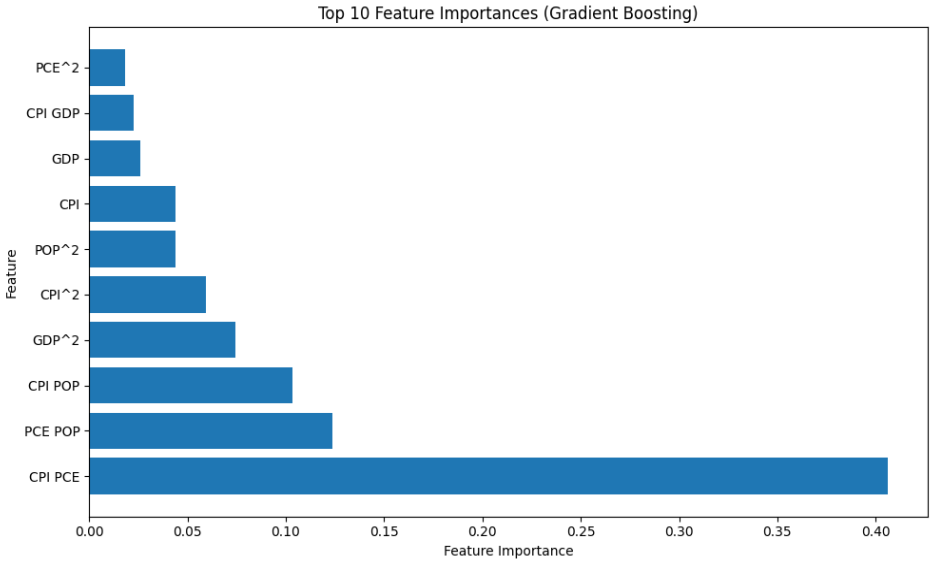
- Gradient Boosting

Gradient boosting is another ensemble learning method that sequentially builds decision trees, with each new tree attempting to correct the errors made by the previous trees. By iteratively improving the model, gradient boosting can capture complex relationships in the data and provide high predictive performance.

The approach is like that of our random forest model. This time, we train the gradient boosting model with 100 boosting stages using the best hyperparameters found during tuning. The gradient boosting model achieved the following performance metrics on the test set:

- Mean Squared Error: 149.18
- R^2 Score: 0.98
- Mean Absolute Deviation: 7.62
- Mean Absolute Percentage Error: 2.74%
- Theil's Inequality Coefficient (Theil's U): 0.02

These results are comparable to those of the random forest model, indicating that both the ensemble methods provide similar predictive performance for this task. The gradient boosting model's ability to iteratively improve the model by focusing on the samples that are harder to predict contributes to its strong performance. The high R^2 score and low error metrics suggest that the gradient boosting model captures the underlying patterns in the data and can accurately predict the average price of the SPY ETF.



- LSTM Neural Network

LSTM neural networks are a type of recurrent neural network (RNN) well-suited for time series forecasting tasks. They can learn long-term dependencies and patterns in data to predict stock price movements.

With this modeling method, we preprocess historical SPY ETF prices, addressing trend and seasonality. We split the data into training and testing sets, with the final 25 days allocated for testing. Then we design an LSTM neural network architecture with a single hidden layer of 100 units before training the model using the training set without scaling the input values.

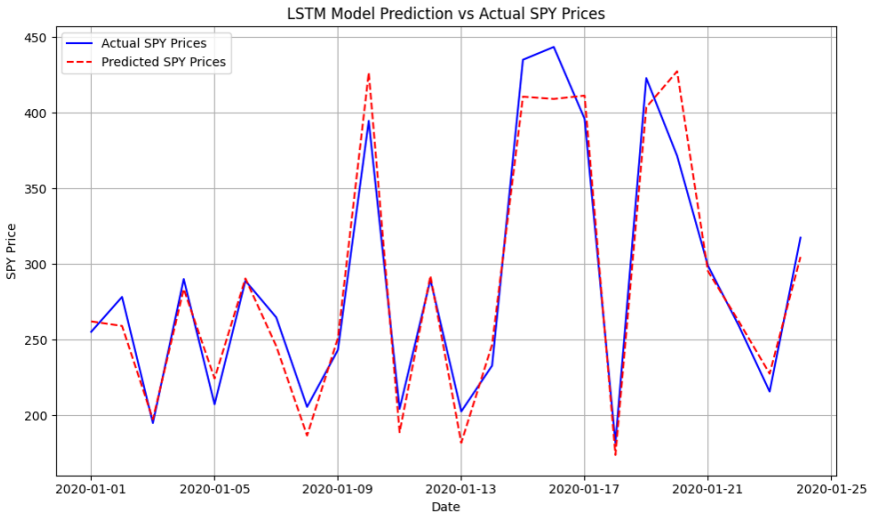
Then, we evaluate the model's performance on the testing set using various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Theil's Inequality Coefficient (Theil's U). Finally, we compare the performance of the LSTM model with traditional time series models.

The LSTM neural network achieved the following performance metrics on the testing set:

- Mean Absolute Error (MAE): 15.5151
- Mean Squared Error (MSE): 19.7093
- Mean Absolute Percentage Error (MAPE): 0.0530
- Theil's Inequality Coefficient (Theil's U): 1.0119

The visualization provides valuable insights into the performance and effectiveness of the LSTM model in predicting SPY ETF prices. The LSTM model closely follows the actual SPY prices, capturing the overall trend and fluctuations in the time series. The model performs well in predicting the prices during the focused forecast period, as evident from the close alignment between the predicted and actual prices. The visualization helps to assess the model's ability to learn and forecast the complex patterns in the S&P 500 index price data. The accuracy of the predictions during the testing period demonstrates the effectiveness of the LSTM neural network in capturing long-term dependencies and making reliable forecasts.

The analysis was underpinned by a comprehensive approach involving preprocessing steps such as calculating moving averages and input scaling, followed by the LSTM model training and performance evaluation using several metrics. The LSTM model's ability to learn long-term dependencies in the time series data contributes to its strong performance in forecasting the SPY ETF price.



- Other Time Series Analysis

We explore naive forecasting techniques, ARIMA (Autoregressive Integrated Moving Average), Exponential Smoothing, and Prophet, an open-source tool released by Facebook’s Core Data Science team for forecasting time series data. We divide data into initiation and evaluation sets, use the first data set to develop a set of forecast models, and use the second data set to evaluate the models. Then we compare the MADs, MFEs and tracking signals of each model.

Time Series Model Performance Metrics:

	mad	rmse	mape	TS
NAIVE	442.951296	536.445107	9.486809	1.211070
DRIFT	317.734621	396.053269	6.786437	1.246491
ARIMA	309.957954	397.557841	6.614257	1.282619
ETL	296.973982	359.243067	6.369862	1.209679
PROPHET Basic	595.286904	691.393296	12.825960	1.161446
PROPHET Linear	599.261053	695.482027	12.913791	1.160566
NAIVE 6 Year	442.951296	536.445107	9.486809	1.211070
DRIFT 6 Year	300.775328	377.140660	6.421700	1.253895
ARIMA 6 Year	355.077558	448.046798	7.582112	1.261828
ETL 6 Year	384.688047	458.616955	8.262592	1.192179
PROPHET 6 Year	1065.239757	1159.458588	23.220695	1.088448
ARIMA 1 Qtr	233.847782	278.009328	4.630239	1.188847
Exponential Smoothing 1 Qtr	171.562416	206.079049	3.396598	1.201190
PROPHET 1 Qtr	760.513098	782.721228	15.241438	1.029202

Through decomposition, we could detect the trend component present within the series, however, there was little evidence of apparent seasonal or cyclical components to the data set. We leveraged naive forecasting methods to establish baseline performance metrics to evaluate more complex approaches. We also evaluated model performance with a smaller historical data set and smaller forecasting window.

While all the calculated tracking signals fall within an acceptable range, none of the models do a great job modeling the reality of the magnitude of movements and randomness contained within the SPY ETF price data set during the forecasted period. Exponential Smoothing with trend and ARIMA forecasts leveraging the full training data set provided the best measured performance considering all factors, including number of forecasted predictions. These specific forecasts were superior to the other time series models but not particularly close to the actual values. However, the range of potential probable values reported by the ETL and ARIMA forecasts did appear to approximate the potential magnitude of movements present within the actual data.

This insight could give investors a probable price range the SPY price is forecasted to be trading within for a future period. This is potentially very valuable information. Given updated data, current SPY ETF prices, recent price movements information, and other contextual details, investment decisions could be planned accordingly.

Overall Conclusion and Key Takeaways

A linear regression model with six predictor variables (Average SPY volume, GDP, Jobs opening, Population, Unemployment rate, and GLD price) was found to perform well in predicting SPY price. This model achieved an R2 of 0.9684 on the training set, 0.9635 on the validation set, and 0.9500 on the test set, indicating a good fit and strong predictive performance. Using PCA did not result in a better performing linear regression model.

A decision tree regression model was trained to forecast the average price of the SPT ETF price. After feature selection, hyperparameter tuning, and model evaluation, the decision tree achieved a mean squared error of 235.69, R2 score of 0.96, mean absolute deviation of 11.13, mean absolute percentage error of 3.86%, and Theil's inequality coefficient of 0.03 on the test set. These results indicate the decision tree model provides a good fit and reasonable predictive accuracy for the SPY price.

A random forest regression model with 100 trees outperformed the decision tree, achieving a mean squared error of 141.30, R2 score of 0.98, mean absolute deviation of 7.47, mean absolute percentage error of 2.72%, and Theil's inequality coefficient of 0.02 on the test set. The ensemble nature of random forests, combining predictions from multiple trees, contributed to its superior performance to the individual decision tree.

A gradient boosting regression model with 100 boosting stages was trained and tuned. It achieved comparable performance to the random forest, with a mean squared error of 149.18, R2 score of 0.98, mean absolute deviation of 7.62, mean absolute percentage error of 2.74%, and Theil's inequality coefficient of 0.02 on the test set. The gradient boosting model's iterative improvement focusing on harder to predict samples enables it to capture complex patterns and accurately forecast the S&P 500 index price.

A LSTM neural network we developed achieved a mean absolute error of 15.5151, mean squared error of 19.7093, mean absolute percentage error of 0.0530, and Theil's inequality coefficient of 1.0119 on the test set. Visualizations show the LSTM model closely followed actual prices, capturing overall trends and fluctuations. The model's ability to learn long-term dependencies contributed to its strong forecasting performance.

Time Series with ARIMA detected a trend component in the SPY price series, but little evidence of seasonal or cyclical components. Naive forecasting methods established baseline metrics. Exponential smoothing with trend and ARIMA models using the full training set provided the best performance. While no time series models well captured the magnitude and randomness of actual price movements, the range of probable values from ETL and ARIMA forecasts approximated the potential magnitude of changes. This insight could provide investors with a probable future price range for planning decisions.

In summary, we used different machine learning and time series analysis methods to predict the price of the SPY ETF that tracks the S&P 500 index. Linear regression, decision tree, random forest, gradient boosting, and LSTM neural networks all achieved strong predictive performance, with random forests and gradient boosting showing the highest accuracy. Time series methods like exponential smoothing and ARIMA also provided valuable forecasts and additional insights into probable price ranges. The combination of these techniques can empower investors with actionable information to make well-informed decisions in the face of market uncertainties.

Code Files (in GitHub)

For codes used to generate the project and accompanying figures, please go to the Code folder in our GitHub: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-75

This Code folder also has a ***Requirements.txt*** file for Python libraries required for our Python scripts and iPython notebooks. Regarding R libraries, we think the R libraries used in the “Linear Regression.R” are within the content of the course.

The parental Team 75 folder in the Github repository has a ***Readme.docx*** file with installation and running instructions as well as an overview of our directory structure.

For data sets, please browse the Data folder in our GitHub.