Executive Summary

Introduction: The S&P 500 is a stock market index that measures the performance of 500 of the largest publicly traded companies in the United States. It is considered a key indicator of the overall health and performance of the U.S. stock market. This index is widely used by investors and financial professionals as a benchmark for the stock market's performance.

Objectives:

- Regression: Forecast the S&P 500 index price at the time of closing

Data: The dataset we consider is the S&P index closing prices over the last 10 years and was obtained from www.nasdaq.com

Methods:

1. Data Preprocessing:
    - The data is cleaned to handle missing values and outliers
    - Relevant information is extracted and new features are engineered
2. Model Training:
    - A wide variety of forecasting techniques are utilized such as ARIMA, double exponential smoothing, decision trees and various ensemble methods. These allowed us to train the data to capture the underlying patterns and temporal dependencies.
    - A wide variety of classification techniques are utilized such as k-nearest neighbors, support vector machines, decision trees, and random forests.
3. Validation and Evaluation:
    - The models are rigorously validated using time-series cross-validation techniques to ensure robust performance.
    - Evaluation metrics such as root mean square error (RMSE), Max Absolute Error Percentage (Max AEP), and Mean Absolute Error Percentage (Mean AEP) are employed to quantify the accuracy of the forecast.
4. Parameter Selection**:**
    - The hyperparameters were tuned by a standard optimization routine.

Results: For the regression based forecast task, the best performing model was Gradient Boost. For optimal parameters, the RMSE of the forecast and the test data is 31.34 (scale ~4400) and the MAEP is ~2.02%. This says that the through-out the testing, the forecast was never more than 2.02% away from the actual data that day.

Expected Impact and Next Steps: This project aims to provide accurate and reliable predictions of S&P 500 index prices, aiding investors, financial analysts, and policymakers in making informed decisions. There are a few future directions we could take, like further exploration of ensemble methods, perhaps by borrowing tools from applied topology or network science, and the development of a pipeline that executes a large variety of methods to pick the best strategy for any given financial time series.

Note: A pipeline for classification of price momentum was set up as well, however, we were unable to get favorable results.