The Erdős Institute Data Science Bootcamp Fall 2023

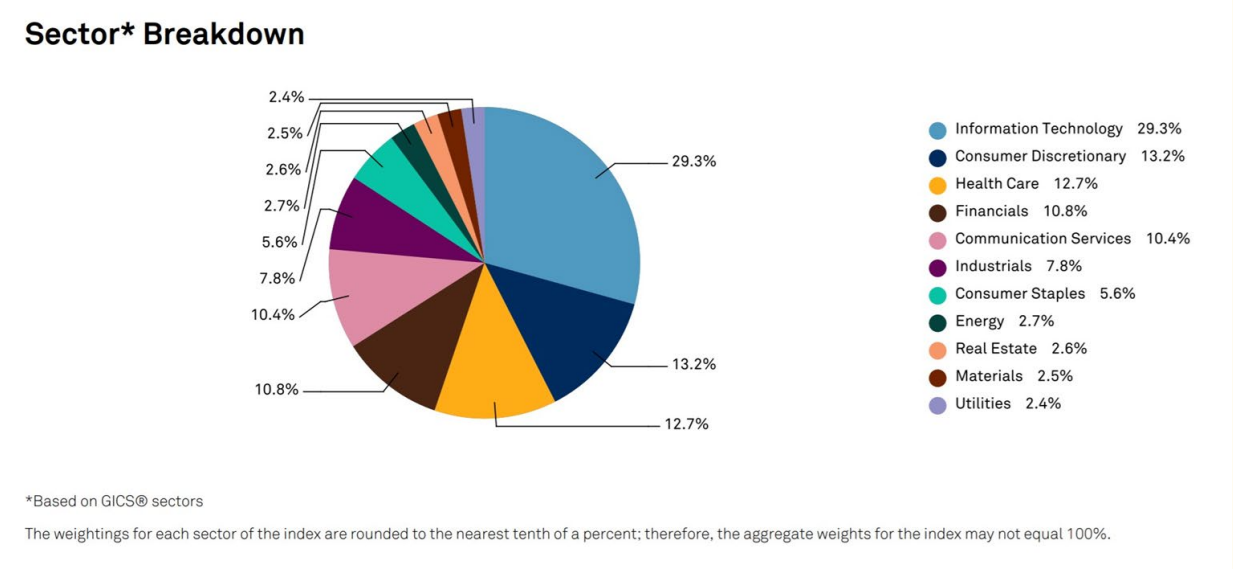
Forecasting Financial Markets: Predictive Modeling Using S&P 500 Data

By Guess-timate Gang
(Hitesh Gakhar, Michail Paparizos, Kriti Sehgal, Limin Wang)

Group Mentor: Soheyl Anbouhi

Stock Market and S&P index

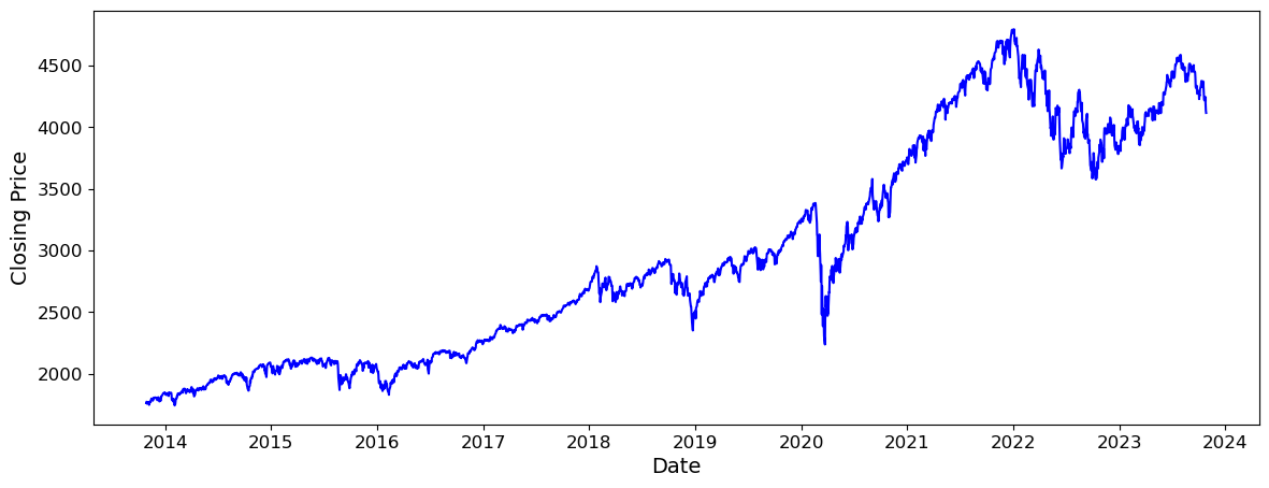
The S&P (standard and poor) 500 is a stock market index composed of ~500 of the leading publicly-traded companies.



Motivation

- S&P 500 is a key indicator of the health and direction of the U.S. stock market.
- S&P 500's influence extends beyond the United States.
- Many investors and fund managers,e.g., hedge fund, use the S&P 500 as a benchmark to evaluate the performance of their investments.
- Many financial products, such as index funds and exchange-traded funds (ETFs), are designed to track the performance of the S&P 500.

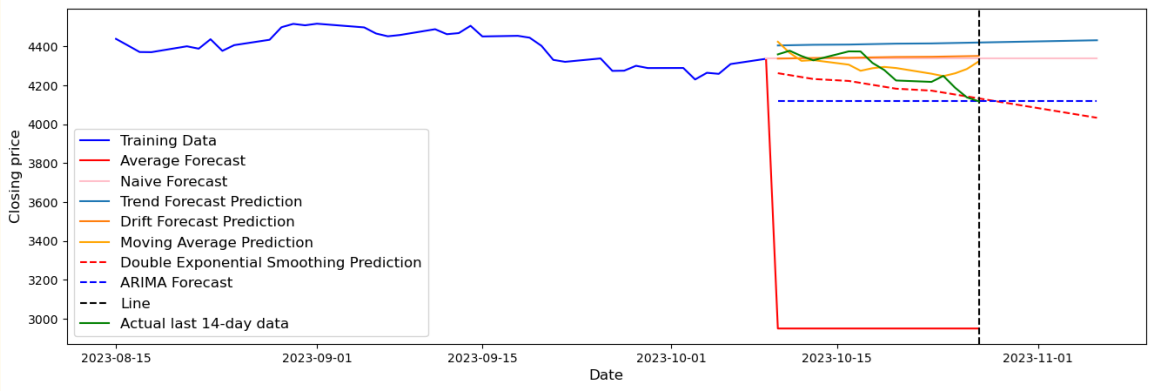
The Data: First Look and Preprocessing



- The S&P 500 dataset for the last 10 years was sourced from www.nasdaq.com
- The dataset was well-organized but minor cleanup and reorganization was performed
- Other features and quantities, like rolling averages, differences, fractional changes were created and explored to understand the data better
- The closing price has an upward trend, however, fluctuates heavily on a more microscale

Exploratory Data Analysis

To forecast the closing price of S&P 500 index, we trained and tested a large variety of models. A subsample is shown on the right.



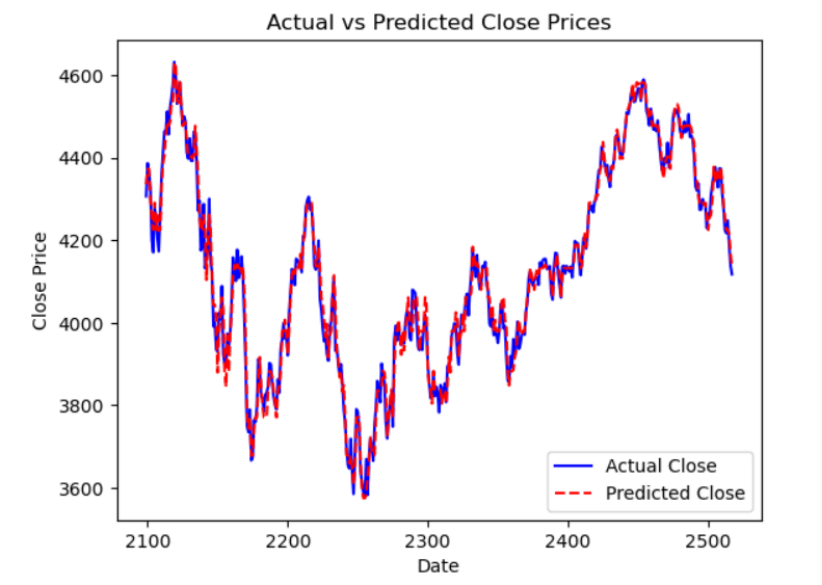
The models are rigorously validated using time-series cross-validation techniques to ensure robust performance.

We use the root mean square error (RMSE) to quantify the accuracy of these forecasts in the regression tasks

Gradient Boosting for Regression

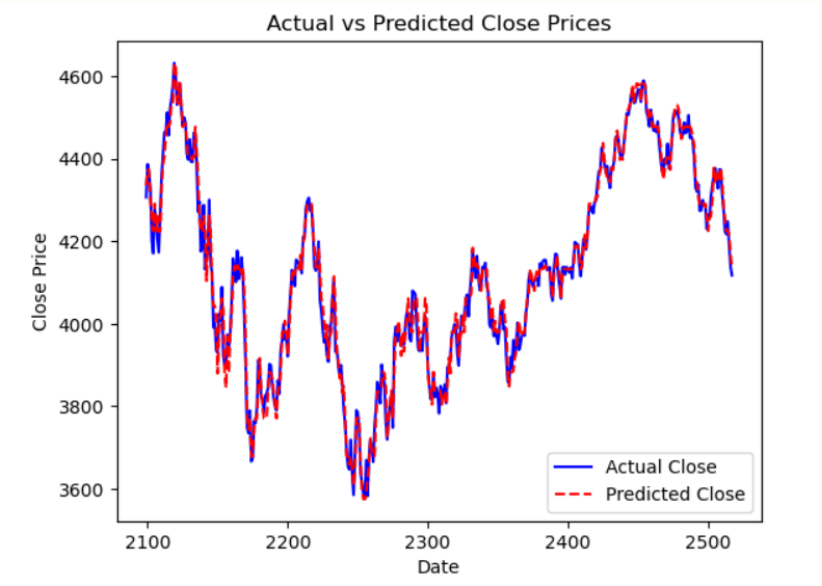
We get best performance with the Gradient Boosting Regressor

- This method produces a strong forecast from an ensemble of weak learners
- The features used to train and predict the closing price were the previous 3 days' value.
- We tune for the parameters: number of weak learners and learning rate to find the most accurate forecast
- We use cross validation by splitting the data into 25 splits
- On the left, we see a forecast with 120 learners and learning rate = 0.07
- On the testing set, this has a RMSE of 31.34 and a Max Absolute Error Percentage of 2.01% and Mean Absolute Error Percentage of 0.57%. This means on any given day, the prediction was no worse than 2% of the actual value and the average prediction is 0.57% away from the actual value.



What's the point?

- S&P data gives insight into the overall market trend.
- Investors can use this to make an informed decision to buy, sell, or hold stocks or indexes that mirror the S&P 500 index.
- This index is often seen as an indicator of the broader economy's health, so it may help:
  - companies align their goals with market expectations and adjust their strategies.
  - policymakers make informed decisions about fiscal policies.
  - day traders capitalize on short term predictions to make quick decisions.



Remarks and Possible Future Work:

- The key challenge in financial forecasting stems from a multitude of factors affecting the price fluctuations. It would be beneficial to bootstrap these methods with more qualitative ideas, like sentiment analysis.
- While we have experimented with some combination and ensemble methods, further exploration holds promise, by perhaps borrowing tools from network science and topological data analysis [1].
- An ambitious future work could involve developing a streamlined pipeline that executes a large variety of methods, thereby picking out the best forecasting strategy tailored to the stock or index based on its behavior.
- We also attempted to study the data via a classification problem: one that answers whether the price goes up or down. We were unable to get favorable results, so one direction we are interested in is exploring feature engineering to enhance accuracy.

[1] How topology is applied in financial market analysis, Medium (Link)