Forecasting Prices Using Stock Market Index Data

Student Number: 690065435

December 2022

1 Introduction

Stock markets from across the world are tracked using indices that measure a section of the stock market, such as the Nasdaq Composite Index. Price forecasting is a very important task in the financial industry, as it can be used to guide strategies.

Historically, financial institutions have used discretionary methods to make investment decisions – they rely on fundamentals and the judgement of analysts [1]. However, with the rise of big data and computational power, systematic methods have become increasingly popular – institutions use rules-based strategies that are implemented by a computer and involve little to no human intervention [1]. Systematic methods enable decisions to be made quickly, which leading market makers and high-frequency trading firms such as Jane Street Capital and Hudson River Trading use to exploit arbitrage opportunities and maximise profits by trading at high volumes [2]. However, these firms do not publicly disclose their strategies, which makes it difficult to understand how they make decisions.

In this project, we will investigate the following question: can we use regression models on stock market index data to forecast prices effectively? My initial hypothesis is that this is not possible, as the stock market is a complex system that is difficult to predict.

Previous studies have investigated the prediction of stock market trends with regression on moving averages [3], but they have focused on individual stocks whereas I am focusing on a stock market index. This is a potential snag, as the index is not a direct representation of the stocks it tracks. The efficient market hypothesis also suggests that this will be difficult, as it states that asset prices reflect all available information, yet we would only be predicting according to a subset of information [4].

2 Methodology and Dataset

2.1 About Stock Exchange Dataset

We are using a stock exchange dataset on Kaggle that was collected from Yahoo Finance, containing the daily price data for stock market indices across the world. Each record in the CSV file contains the following information for each trading day:

- Index
- Date
- Opening price
- Highest price
- Lowest price
- Closing price adjusted for splits
- Adjusted closing price adjusted for both dividends and splits
- Volume of shares traded

2.2 Data Cleaning

I loaded the CSV file into a pandas dataframe and converted the date to a datetime64 object to make it easier to work with. It originally consisted of 112457 rows. After looking at the information of the data, I noticed that there were 2204 incomplete rows, so I removed them to ensure that the data was consistent. There were also rows with trading volume recorded as 0 in earlier dates where this information was not available, so I removed them to prevent them from skewing the distribution of volume data – this further reduced the dataset from 110253 rows to 68160 rows.

2.3 Data Exploration

I calculated the distribution of features to understand patterns in the data. The distributions of open, high, low, close, and adjusted close are extremely similar, as demonstrated in the histograms – this makes sense as open, close, and adjusted close are between the low and high for each day.

Meanwhile, the graph of daily volume traded over time shows that market activity has increased over time, which can be attributed to the rise of electronic trading making the stock market more accessible to everyone, including high-frequency trading firms.

2.4 Feature Engineering

2.5 Data Filtering

2.6 Regression Models

2.7 Models Evaluation

3 Results

3.1 Ridge Regression

3.2 LASSO Regression

3.3 Polynomial Regression

4 Discussion

4.1 Limitations of the Study

The efficient market hypothesis states that asset prices reflect all available information, which makes it difficult to get arbitrage opportunities. However, Renaissance Technologies, a hedge fund that uses systematic methods, serves as a perfect counter-example – their Medallion Fund has achieved 66.07% annualised returns since 1988 [5]. This suggests that it may be possible to achieve better results given access to big data and computational power, whereas in this study we relied on limited information.

4.2 Future Work

References

[1] C. R. Harvey, S. Rattray, A. Sinclair, and O. Van Hemert, Man vs. machine: Comparing discretionary and systematic hedge fund performance *The Journal of Portfolio Management*, vol. 43, no. 4, pp. 55–69, 2017.

[2] I. Aldridge, *High-frequency trading: a practical guide to algorithmic strategies and trading systems*, vol. 604. John Wiley & Sons, 2013.

[3] S. Dinesh, N. Rao, S. Anusha, and R. Samhitha, Prediction of Trends in Stock Market using Moving Averages and Machine Learning in *2021 6th International Conference for Convergence in Technology (I2CT)*, pp. 1–5, IEEE, 2021.

[4] E. F. Fama, Efficient capital markets: A review of theory and empirical work *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[5] B. Cornell, Medallion Fund: The Ultimate Counterexample? *The Journal of Portfolio Management*, vol. 46, no. 4, pp. 156–159, 2020.