Prediction Model for the Stock Market Using Machine Learning Techniques

Ria Arora (arorari6) ria.arora@mail.utoronto.ca

Cong Liu (liucon16) congcrystal.liu@mail.utoronto.ca

Yiqi Wu (wuyiqi1) yiqi.wu@mail.utoronto.ca

ECO481: Macroeconomic Finance (with machine learning applications)

University of Toronto

Department of Economics

Final Project Report

15 December 2022

Introduction

The stock market has always been difficult to predict due to the complex behaviour and volatility within the market. In the past, investors have tended to use their subjective judgments to buy and sell stocks, but some have come to a loss. For example, American investor Bill Hwang lost 20 billion dollars in the value of stock returns. In the age of the technological revolution, embracing artificial intelligence in the financial market is the next frontier for investors, firms, and stockholders. Combining the financial and technical domains will allow firms to exploit specific trading opportunities through calculated decisions on expected values of earnings and losses.

These ideas fascinated our group and established our interest in determining how we could blend both domains to base these solutions. The COVID-19 pandemic hit the world unpredictably. Under this shock, we are interested in comparing different machine learning models and evaluating whether their prediction performance is still robust. The research question we plan to investigate is: **which machine learning model is the best at predicting stock market direction during the COVID-19 pandemic**?

There have been multiple papers published on this topic of interest to predict the nature of the stock market and provide further insights for successful investment using a variety of methods. Insights from statistical classifiers showed that the prediction of stock price movements can be more efficient by capturing the movement of daily high prices to observe its volatility right before the market closes (Novak, et al, 2015). An early paper did a thorough comparison between older statistical techniques and machine learning algorithms for their predictive performance of the Karachi Stock Exchange and concluded that an ML technique performs the best (Usmani et al, 2016). A recent paper established an efficient prediction model for companies

on the New York Stock Exchange using the random forest algorithm with a high prediction precision of 93.23% (Madeeh and Abdullah, 2021).

However, those works were based on early-year repositories of stock price data. The findings of past studies are no longer applicable to the stock market during the COVID-19 pandemic. Our study focuses on comparing the predictive performance of machine learning models during the COVID-19 pandemic. Considering that the random forest is highly efficient in predicting the stock market, we hypothesize that the tree-based methods will perform better than the other classifiers. Our group hopes that our paper can contribute to the lack of research for this recent pandemic shock and that our forecasting model of the stock returns can help investors make more informed decisions.

Data and Methodology

Since our project is interested in comparing model performances during the COVID-19 period, we focus on the time horizon from November 2019 to November 2022. We extracted the historical stock price data of the International Business Machines Corporation (IBM) from Yahoo finance. To enhance the performance of the machine learning algorithms, we collected stock price data from firms that are comparable to IBM: Oracle and Microsoft (Gartner, n.d.). Including data from these two companies minimizes the variation among the three companies while increasing the sample size, as we do not want to study the impact of companies on stock price movements. The combined dataset has daily records of the open, high, and low, closing prices and trading volume of the three firms' stock.

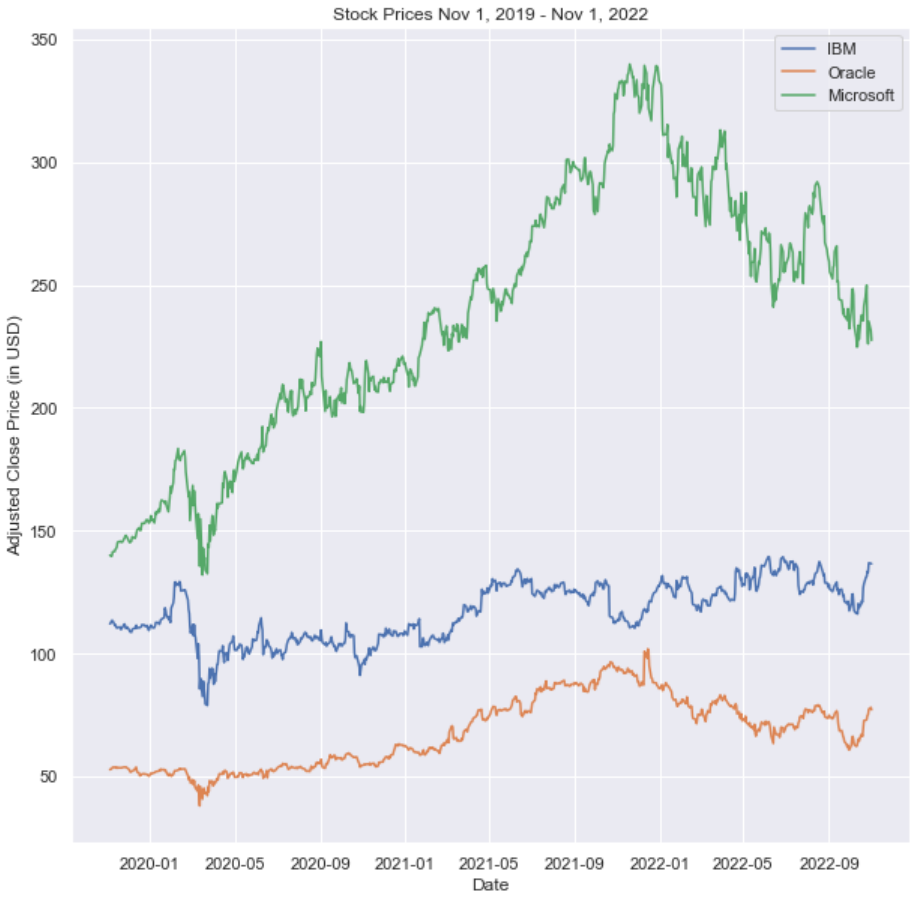

Figure 1: Stock price from 2019-11-01 to 2022-11-01 of Microsoft, IBM and Oracle

The stock price movements of the three companies are very similar as shown in figure 1. All three firms' stock prices hit rock bottom in the first quarter of 2020. Over the past three years, the stock prices of all three firms show an increasing trend while having fluctuations. At the beginning of 2022, the stock prices of all three companies reach a point higher than the initial adjusted close path that each firm started with at the beginning of 2020.

For our prediction models to incorporate the impacts of COVID on the U.S. stock market, we decided to pull daily search popularity data in the U.S. from Google trends. When people want to know more about something they are interested in, they often search for relevant keywords. Thus, the Google search popularity of COVID keywords can measure how much the

public is paying attention to COVID-related news. We decide to build a daily COVID search index to represent the level of public interest. We define the daily COVID search index as the average Google search popularity of the following keywords: coronavirus, covid-19, lockdown, mask, pandemic, quarantine, and vaccine.

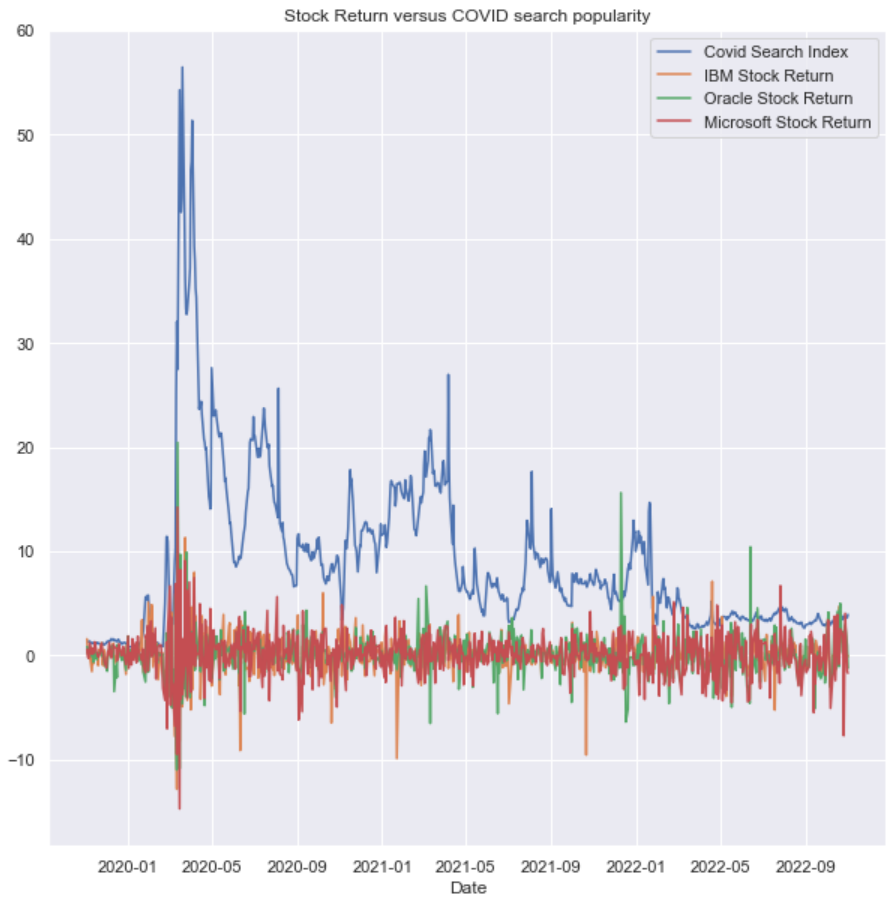

Figure 2: COVID search index versus stock return of Microsoft, IBM and Oracle from 2019-11-01 to 2022-11-01

Looking at the overlapping return series, the similarity in stock prices of these three companies is even more pronounced. The returns in early 2020 and 2022 were notably volatile, which is in line with observations from the price movement. With the onset of the COVID pandemic at the beginning of 2020, we observe a spike in the COVID search index while the

stock prices fluctuated sharply. The impact of COVID on the U.S. stock market is confirmed by the fact that both stock prices and Google search popularity experienced dramatic changes in the same time frame. We use these data to build prediction models for stock price movements. The models are trained based on the following feature inputs: the daily opening price, high and low stock prices, the trading volume of the stock, the firm name and the daily COVID search index. Our response variable of interest is an indicator of whether the daily return is positive or not.

We split 80% of the data for model training, which consists of 1812 entries of all three firms' stock prices from 2019-11-04 to 2022-03-28. Since all features are indexed by time, we did not do a random shuffle before the train/test split so that the training data preserves the time ordering. We performed two additional preprocessing steps to the training data so that the machine learning algorithms can have an ideal behaviour. We encoded the firm labels with integers values between 0 and 2 and standardized all features to have zero mean and unit variance.

For our project, we decide to train and compare three machine learning algorithms: L1-penalized logistic regression, decision tree and random forest. Our project focuses on predicting whether the daily stock return is positive or not, it is suitable to use logistic regression as it is commonly used for binary classification. The L1 penalty term (also known as LASSO) can help identify the features that are most powerful when predicting the direction of stock prices. The LASSO penalty shrinks coefficient estimates for irrelevant predictors down to zero when the tuning parameter is sufficiently large. In this way, we can focus on the most important features and interpret how they affect the stock price direction. As the stock market exhibits highly complex behaviours, we anticipate that the probability of earning a positive return may not be linearly related to our data features. The tree-based methods are known for capturing potential

non-linearities in the relationship, given that the tree is properly pruned. The tree algorithms can also provide us with an intuitive visualization of the "rules of thumb" for predicting the difficulties of stock prices.

All machine learning algorithms mentioned above require hyperparameter tuning. The penalized logistic regression needs to tune a hyperparameter for the L1 penalty strength. To prune the trees, we decide to restrict the maximum tree depth so that we do not have an overgrown tree. We select the hyperparameter value based on results from 10-fold cross-validation. After a grid search of candidate hyperparameters, we set all three models with the hyperparameter value that maximizes the average cross-validation accuracy. To answer our research question, we then compare each model's predictive accuracy and declare the optimal prediction model for stock price direction.

Results

After conducting 10-fold cross-validation for all three machine learning algorithms, we set the L1 penalty strength for the logistic regression at the default level, while the maximum depth of the tree-based methods is set to one.

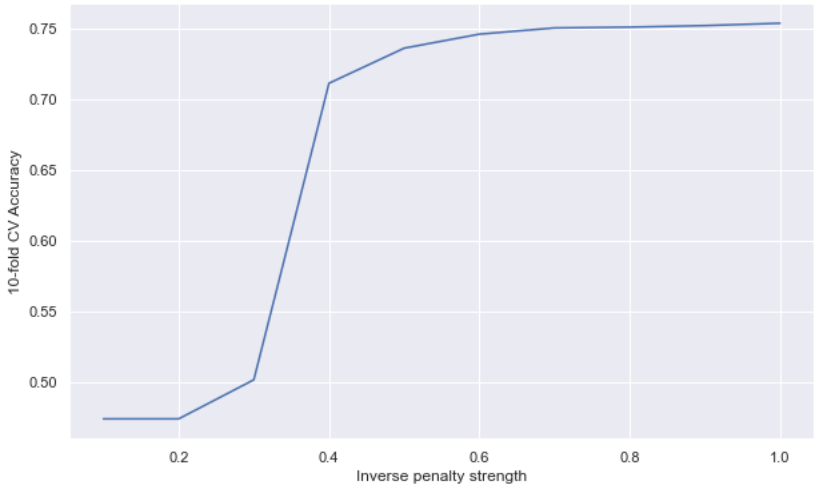

Figure 3: Inverse L1-penalty strength versus 10-fold cross-validation accuracy for the penalized logistic regression

The L1-penalized logistic regression achieves predictive accuracy of 77.70%, which is the highest among the three models. We calculated the feature importance as the Euler number to the power of coefficients (Serengil, 2021).

Feature	Coefficient	Importance
Low	38.2261	3.9936×10^{16}
High	19.4228	2.7239×10^{8}
Volume	0.1422	1.1528
firm	-0.0022	0.9978
covid_search	-0.0523	0.9491
Open	-57.7504	8.3048×10^{-26}

Table 1: Coefficient estimates from the L1-penalized logistic regression, sorted by feature importance

The top two key features from the penalized logistic regression are the daily low and high prices of the stock, which is not surprising since the stock return is proportional to the price differences. Both features are positively correlated with the probability of earning a positive stock return. The direction of the coefficients is consistent with our intuition. If a stock has a higher daily low price, this may indicate an upward momentum and therefore a higher chance of a positive stock return. Selling at a high price is also more likely to earn a positive return. The trading volume of the stock is the third most important feature and is also positively associated with the probability of earning a positive stock return. Higher trading volume suggests that more investors are interested in buying or selling the stocks, thus affecting the stock return. Although the stocks of the technology companies kept rising during the pandemic, the pandemic still has an overall negative effect on the stock market. Greater public attention in COVID-related news is

correlated with a lower probability of having a positive stock return. People paying more attention to the COVID updates is likely because economic activity was more severely impacted by the pandemic, which also affects the stock market.

Conclusion

Determining the behaviour of the stock market has always been a puzzle to many investors. However, literature published in the past allows researchers to analyze the effects of applying the techniques of machine learning to calculate the trajectory of stock returns. In addition to this, research on this topic provides further proof that innovative machine learning methods can build better prediction models. This paper proposes three different machine learning models: L1-penalized logistic regression, decision tree and random forest. All three models have their hyperparameter tuned based on 10-fold cross-validation results. To select the **optimal model** for predicting the stock return direction of U.S. technology companies (IBM, Oracle and Microsoft) **during the COVID-19 pandemic**, we compared each model's predictive accuracy. To conclude, the logistic regression with an L1-penalty is the optimal model by its superior predictive accuracy of 77.70%. In addition to the higher accuracy, the coefficient estimates from the penalized logistic regression are also in line with stock market intuitions. Results from the L1-penalized logistic regression are accurate and sensible, and thus can indeed help investors make more informed decisions in the stock market.

References

Gartner, Inc. (n.d.). *Top IBM Competitors & Alternatives 2022: Gartner Peer Insights - Metadata Management Solutions*. Gartner Peer Insights. Retrieved December 12, 2022, from https://www.gartner.com/reviews/market/metadata-management-solutions/vendor/ibm/alternatives

Gorenc Novak, M., & Velušček, D. (2016). Prediction of stock price movement based on daily high prices. *Quantitative Finance*, *16*(5), 793-826.

Madeeh, O. D., & Abdullah, H. S. (2021, February). An efficient prediction model based on machine learning techniques for prediction of the stock market. In *Journal of Physics: Conference Series* (Vol. 1804, No. 1, p. 012008). IOP Publishing.

Serengil, S. I. (2021, January 6). *Feature importance in logistic regression for machine learning interpretability*. Sefik Ilkin Serengil. Retrieved December 12, 2022, from https://sefiks.com/2021/01/06/feature-importance-in-logistic-regression/

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016, August). Stock market prediction using machine learning techniques. In *2016 3rd international conference on computer and information sciences (ICCOINS)* (pp. 322-327). IEEE.