The Efficacy of Different ML Algorithms in Predicting Stock Prices

Presented by

WeiChih Lee

Alvee Chowdhury

Summary

Purpose

The ability to predict stock prices would yield great rewards. If we could predict stock prices with great accuracy, we would be able to currently identify which stocks to invest in. Moreover, we would also know which stocks to avoid or sell. Therefore, predicting the stock prices is clearly profitable for the financial traders of our company. This report attempts to use various machine learning algorithms to try find an efficient and accurate way of predicting the closing stock price of different companies. This report is suitable for all the financial traders in our comapany and advocates the importance of using predictive analytics as a tool to make investment decisions.

Main Finding

We used four ML algorithms to predict the stock prices - Support Vector Machine, Extreme Gradient Boosting, ARIMA and hybrid model consisting of ARIMA and Extreme Gradient Boosting. We used Root mean square as our evaluation metric. They all performed similar to each other. However, Support Vector Machine performed slightly better than the rest, so we should use SVM to predict future stock prices.

Contents

Introduction

Predictive analysis is defined as forecasting future events by identifying the trend and patterns of data. It is used in almost every industry like management, insurance, or the stock market. It uses data mining, data modeling, artificial intelligence, statistical learning, and machine learning algorithms to predict future outcomes by analyzing various threats, possible future demand, etc. It helps in anticipating uncertain events.

One of the fields where predictive analytics plays a vital part is the stock market. The ability to predict stock prices would yield huge rewards. Being able to predict which stock prices are going to rise or fall can be really important in making business decisions. We could use that knowledge to buy the shares that would most likely rise in the near future while avoid or sell the shares that are predicted to fall soon. So, clearly stock price prediction is a very useful tool in forming investment strategies and development of risk management models.

However, the stock market is a complex structure, and the stock price depends on various factors like demand and supply, the company's profit, etc. It is almost impossible to predict future stock prices with 100% accuracy due to the uncertain nature of the market. Predictive models help to analyze the past stock prices to determine the future stock prices with great accuracy. It assists in recognizing the financial pattern of the companies' stock prices. Predicting stock prices can be challenging since the data is volatile, non-parametric, and non-linear. Over the years, development in Machine Learning algorithms have greatly improved prediction models and thus, are a great asset for companies to calculate and analyze the stock market price, financial services, and the global capital market.

Introduction

In this report, we evaluate the performance of four ML algorithms: Support Vector Machine, Extreme Gradient Boosting (XGBoost), ARIMA, and a hybrid ARIMA + XGBoost model. Our goal was to find the best model that has the highest accuracy in predicting the stock price.

This report is broken down into few sections. We first talk a bit about the data we used and how we processed it. Then we give a summary of each model. We then summarise the performance of the models. Finally, we end by summarising our findings and what we could do for our future work.

Data Preprocessing

We aim to predict the closing price by training our models using historic data. We are collecting data from Yahoo finance. We intend to predict the stock price for Toyota, Google, Facebook, Nike, Adidas, Well Fargo, American Express, Bank of Canada, Walmart, Costco and JPMorgan.

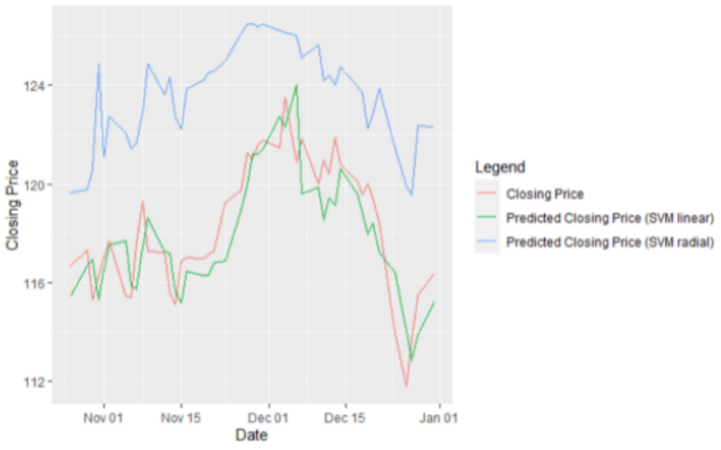
The data contains 6 variables: - “Open”, “Close”, “High”, “Low”, “Adj. close”, “Volume”. The Open price is the first trading price for the trading day. The High price is the highest value during the trading day. The Low price is the lowest value during the trading day. The Close price is the last trading price of the day. The Adj. Close price is the closing price after the accounting actions like a dividend. Volume is the total number of shares traded for a stock on the trading day.

Predicting stock prices can be a challenging task since the stock price is affected by numerous factors. We need all the parameters that affect the market volatility. But since a lot of these factors are unknown, it can be difficult to build efficient models. Mehar, Deeksha, Tikkiwal, and Kumar (2019) addressed this by creating six new features. We took a similar approach. We created 8 new features: Stock high – low price, opening – Closing price, stock prices’ 7,14- and 30-day moving averages respectively. We also defined some technical indicators like: The Relative Strength Index (RSI) calculates a ratio of the recent upward price movements to the absolute price movement, The Parabolic Stop-and-Reverse which calculates a trailing stop, the Directional Movement Index, and the trend.

SVM

We chose SVM because It helps to analyze classification and regression analysis for the data. SVMs seek to minimize generalization error. Here, we decided to train our model with both linear and Gaussian Kernal. We needed to specify the Kernel, which takes data as input and transforms it into the required form.

You can learn more about the performance of this model and all the other models in the "Performance Summary" section. Here we included a graph of the original closing price of the Toyota company and also the predicted closing prices produced by the two models we trained.

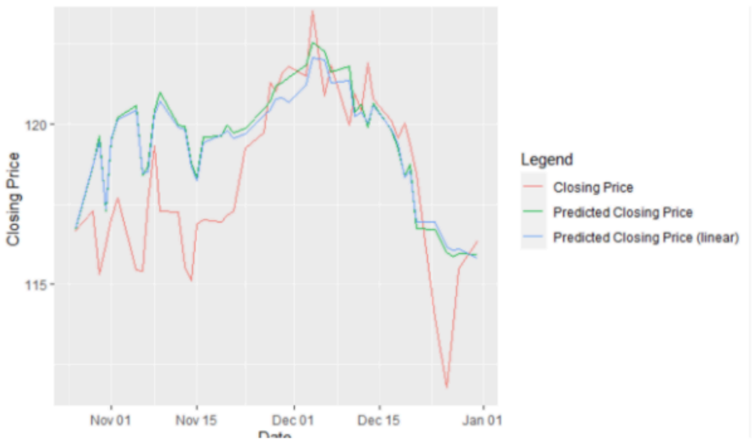


As you can see above, the model with the linear kernel performs really well, while the Gaussian one under-performed. We saw similar results for all the companies we tested on.

XGBOOST

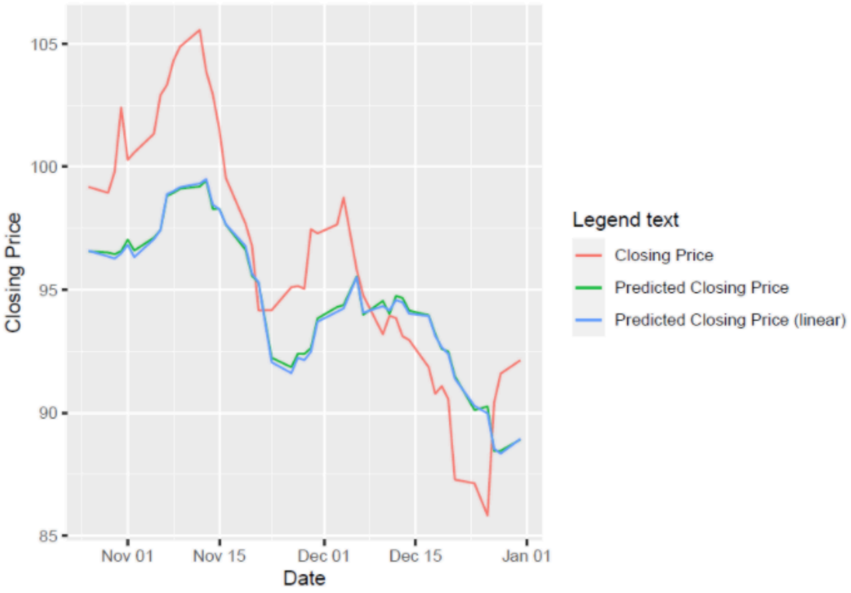
We chose XGBoost because of its efficiency and fast yet fairly accurate output. It uses either decision trees or generalized linear models as its base model to build a more complex model. Decision trees and generalized linear models are just some simple ML algorithms. So they have their limitations. But now imagine adding them one by one as building blocks to build this complex model where the new model created predicts the errors of the prior model and improves from it. Clearly, we will get a better model! It uses "Boosting" to train the base models successively and uses gradient descent to minimize the errors.

We trained our model using: Decision trees and generalized linear models as the base models. We included the graph depicting its performance on the data from the Toyota company.



ARIMA

The Autoregression Integrated Moving Average (ARIMA) model uses the predictors as lags to forecast time-series data.
It learns from its past values. It is a simple yet powerful forecasting tool for time series data. We decided to include this since we implemented a hybrid model using this as our base model. So we wanted to include this as a reference. However, it performs quite well on its own. The only drawback was that we needed to make sure the data was stationary, i.e it had no underlying trend.

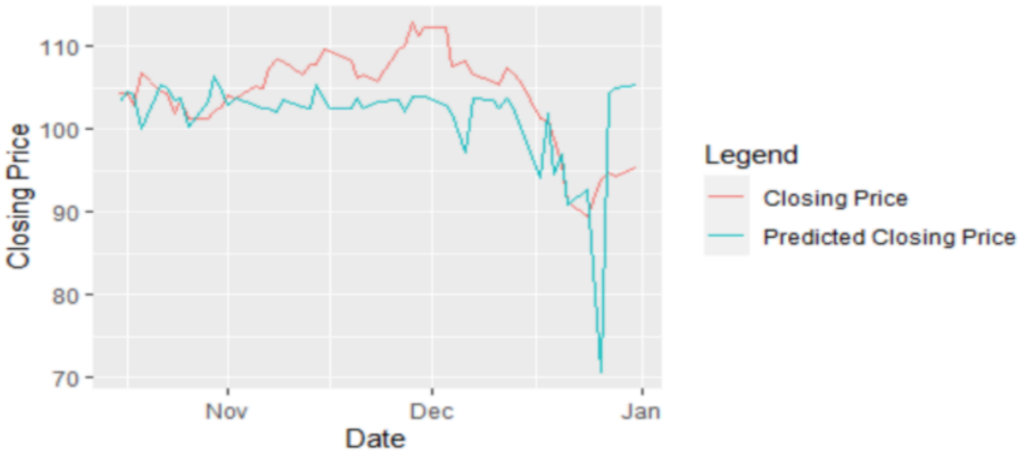


Hybrid Model

As mentioned earlier, predicting stock prices can be challenging since the data is volatile, non-parametric, and non-linear. From the ARIMA model we saw that if we were to just use the linear part of the data (i.e., make the data stationary) we can get a good prediction. However, we lose the nonlinear part of the data. So here, we use an ARIMA model to account for the linear component while the XGBoost is used to predict the nonlinear component of the data.

We first use a suitable ARIMA model to predict the model at time t. Then we find the residuals of this model. Once we find the residuals, we will feed them into a well-tuned XGBoost model. This will allow us to predict the non-linear nature of the prices. We will then add these two up to get our final result.

The plot below is the actual closing price and the closing price

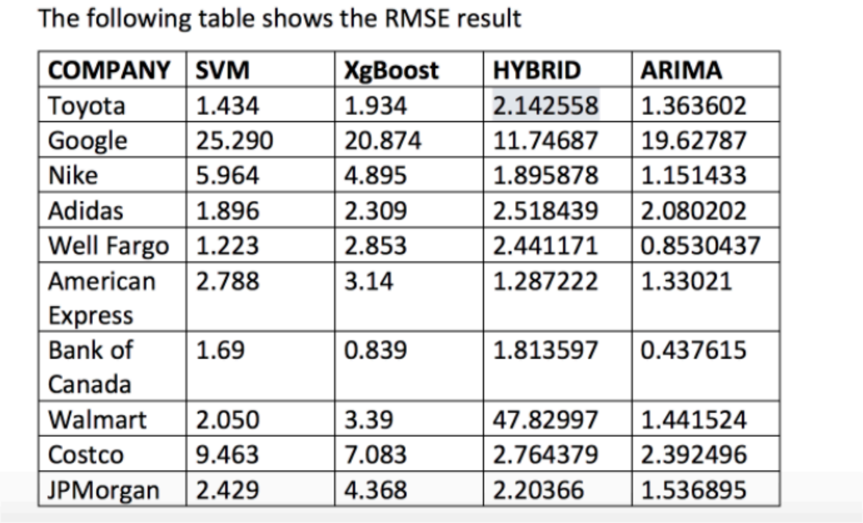


Summary of the Performance

For evaluation of model fitting and prediction, the following parameter was calculated for each model. The smaller the Root Mean Square Error value, the better the model performs.

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - F_i)^2}{n}}$$



The following table shows the RMSE result

COMPANY	SVM	XgBoost	HYBRID	ARIMA
Toyota	1.434	1.934	2.142558	1.363602
Google	25.290	20.874	11.74687	19.62787
Nike	5.964	4.895	1.895878	1.151433
Adidas	1.896	2.309	2.518439	2.080202
Well Fargo	1.223	2.853	2.441171	0.8530437
American Express	2.788	3.14	1.287222	1.33021
Bank of Canada	1.69	0.839	1.813597	0.437615
Walmart	2.050	3.39	47.82997	1.441524
Costco	9.463	7.083	2.764379	2.392496
JPMorgan	2.429	4.368	2.20366	1.536895

The table above shows us that there is no one model that works for all 10 companies. SVM and ARIMA performed well and so did our hybrid. Most of these models failed to predict sudden changes. The Hybrid Model did not necessarily perform better than expected. When trying to compute hybrid models, we need to be careful. Ping-Feng Pai and Chih-Sheng (2004) mentioned how hybridizing two models that are different from each other are likely to minimize forecasting errors but also warned that just combing two models would not necessarily improve prediction accuracy. This is consistent with our findings.

Conclusion

Our goal was to find the best model that had the highest accuracy in predicting the stock price. After conducting the research, we found SVM to perform the best out of all the models. So we suggest using SVM to predict future closing prices. In the research paper, "Stock Closing Price Prediction using Machine Learning Techniques" by Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar mentioned, "For future work, deep learning models could be developed which consider financial news articles along with financial parameters such as a closing price, traded volume, profit and loss statements, etc., for possibly better results". We will take this into account for future research, and we would like to develop deep learning models which consider financial news articles along with financial parameters such as a closing price, traded volume, profit, and loss statements, etc., for possibly better results. Our main problem was that the features we included were not enough to properly characterize the uncertain nature of the market. So, for future work, we will have to focus more on feature engineering to find more important variables. Having said that, we still believe including any one of the above models will be a great way for us to predict the stock prices and help identify which companies to invest in.

References

- Predictive Analytics (CLAY HALTON, 2019) from Predictive Analytics Definition (investopedia.com)
- Prerana, Pratheeksha, Tahmin, Anusha, Madhu (2020): STOCK MARKET PREDICTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES
- Ping-Feng Pai and Chih-Sheng (2004): A hybrid ARIMA and support vector machines model in stock price forecasting: The International Journal of Management Science
- Mehar, Deeksha, Tikkiwal, and Kumar (2019) “Stock Closing PricePrediction using Machine Learning Techniques”: International Conference on Computational Intelligence and Data Science
- Aparna, Manohara and M. Pai (2016) “Prediction Models for Indian Stock Market”: Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)
- Prabhakaran, S. (2021, March 22). ARIMA model - complete guide to time SERIES forecasting in Python. Retrieved April 15, 2021, from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
- Joseph, E., Mishra, A., &amp; Rabiu, I. (2019, February 14). Forecast on close stock market Prediction usingSupport vector Machine (SVM). Retrieved April 15, 2021, from https://www.ijert.org/forecast-on-close-stock-market-prediction-using-support-vector-machine-svm
- Sangarshanan. (2019, April 07). Time series forecasting - ARIMA models. Retrieved April 15, 2021, from https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06

References

- Brownlee, J. (2020, August 14). Boosting and AdaBoost for machine learning. Retrieved April 15, 2021, from https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/
- Meltzer, R. (2020, October 21). What is random forest? Retrieved April 15, 2021, from https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/
- Dwivedi, R. (n.d.). Xgboost algorithm for classification and regression in machine Learning: Analytics steps. Retrieved April 15, 2021, from https://www.analyticssteps.com/blogs/introduction-xgboost-algorithm-classification-and-regression
- Directional movement Index (DMI) - OVERVIEW, Calculation, Trading. (2020, September 12). Retrieved April 15, 2021, from https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/directional-movement-index-dmi/#:~:text=The%20Directional%20Movement%20Index%20%28DM