Machine Learning Project Report: Stock Price Prediction

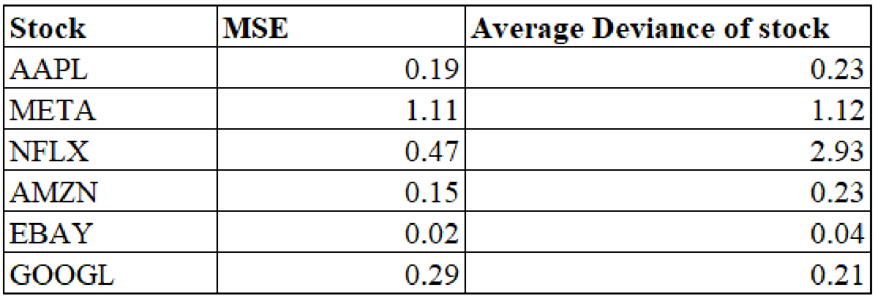
By,
Vamsee Narahari,
Aishwarya,
Vindhya Mandekar

Table of Contents

Executive Summary

This project aims at creating a machine learning model that predicts the price of a stock for every 15 minute time intervals. Stock prices of Google, Apple, Meta, Netflix, Amazon, Ebay have been collected from Yahoo finance. Variables high, low, close, open and volume have been collected for every 15 minute intervals from December 14th and March 13th. This complete data has been used for training, hyper-tuning and validation. Finally, the selected model has been tested on the data from March 14th to March 23rd.

The programming language chosen for this project is Python. Some of the libraries used are sklearn, statsmodel, and tensorflow . We tried different approaches like linear regression, lasso regression, random forest and neural network to build a stock price prediction model. Based on our analysis we found that different stocks move differently in the market. We also learned that each stock's close depends on a different set of features. Some of the more volatile stocks depend on more lag values as well as volume information whereas the less volatile stocks depend mostly on the previous lag variables. The main observation was that we saw a strong linear correlation between lag values and close price and hence the final model we chose was a lasso linear regression model. Below are the statistics for this model.



Stock	MSE	Average Deviance of stock
AAPL	0.19	0.23
META	1.11	1.12
NFLX	0.47	2.93
AMZN	0.15	0.23
EBAY	0.02	0.04
GOOGL	0.29	0.21

Table-1 Final Model Results & Comparison of Average Deviance

Background

Stock market trading can be an exciting and potentially lucrative endeavour, but it can also be a challenging one for those who are new to it. Entering the stock market trading requires knowledge, skill, patience, and a significant amount of dedication. Personal day trading is a type of trading where individuals buy and sell financial instruments, such as stocks or currencies, within a single trading day. One of the most significant difficulties individuals may face when starting personal day trading is the level of risk involved. Personal day trading involves making quick decisions based on market fluctuations, which can result in significant losses if the market moves against the trader. It also requires knowledge of trading platforms and software, as well as the ability to analyse market data quickly. Individuals who are new to personal day trading may find it difficult to acquire the necessary knowledge and skills, which can result in significant losses. Our project attempts to make this feat easier for people just starting their trading journey. Traditionally, stock trading analysis is seen as a time series forecasting model, but we are looking at it in a different light. We are using machine learning prediction techniques to identify the important features that need to be considered before trading stocks. We also want to throw light on how the stock market behaves at every 15 minute intervals and what factors are important and need to be considered before trading. Our analysis also helps provide insight on stock trends and show patterns of how the stock moves during the day. But before we can get started on explaining how we do this, here are some terms that you need to be familiar with.

Term	Definition
Close	The last price at which a stock trades during a trading interval
Open	The price of the stock at the start of the trading interval
High	The highest value of the stock during the trading interval
Low	The lowest value of the stock during the trading interval

Volume	The number of shares traded in a particular stock during the trading interval
Volatility	The rate at which the price of a stock increases or decreases over a particular period

How it is traditionally done

The stock market is known for being volatile and dynamic. Accurate stock price prediction is extremely challenging because of multiple factors, such as politics, global economic conditions, unexpected events, a company's financial performance, and so on. But financial analysts, quantitative researchers and data scientists continue to explore the data and use analytical techniques to detect stock market trends. This has given rise to Algorithmic trading that uses advanced mathematical models for making transaction decisions in the financial markets.

The analysis of stock data involves examining successive observations taken over time, which can be treated as a time series. Time series forecasting is a suitable method for predicting future values of stocks based on their historical values. Moving Average technique is a popular method used to help smooth out price data by creating a constantly updated average price. By taking the average of a specific number of past observations, a moving average can reveal patterns and trends in the stock prices that may be difficult to identify otherwise. A rising moving average indicates that the security is in an uptrend, while a declining moving average indicates a downtrend.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is another technique particularly well-suited for stock price prediction due to its ability to capture and remember long-term dependencies and patterns in time series data. LSTMs can model complex, non-linear relationships between the input features and the output variable, making it suitable for predicting stock prices, which are affected by a variety of factors that can be difficult to model using traditional statistical techniques. It also uses memory cells to remember past information and selectively forget irrelevant information. This allows the model to capture long-term dependencies and patterns in stock price data.

In this project we attempt to apply different techniques and models to help understand market trends as well as predict stock price.

Exploratory Data Analysis (EDA):

The EDA was performed on the training dataset to identify anomalies or outliers, understanding the data and determine the relationships between the variables. The stock prices is a time-series data, so it is checked for any abnormal changes in the prices and verified the accuracy of the data through internet. One of the insights into the movement of the stocks is that the coefficient of variation in the META and NFLX are relatively higher. This volatility is due to the recent news & performance reports of those companies. Every stock's close price is strongly and linearly related to the previous lag terms of the closing and opening prices as illustrated in Figure-1.

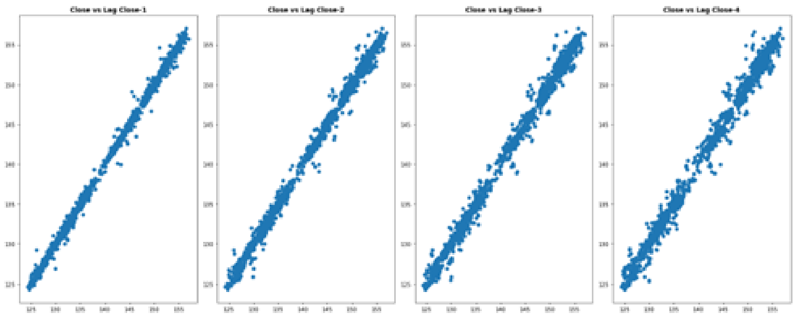


Figure-1: Sample Scatter Plot Close Prices vs Lag terms

The magnitude of difference in the closing price from one interval to another interval is slightly dependent on the volume of the stock which has been traded in the lag intervals. While the volume traded in the previous intervals increased, the magnitude difference in the intervals also increased.

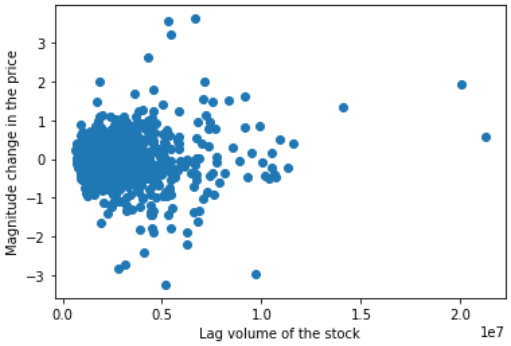


Figure-2: Sample Scatter Plot Lag-1 Volume vs Close Price magnitude change

Moreover, there is an existing pattern between the time of the day and the stock movement as well. In the early hours of the day, the volume and magnitude difference is higher than the rest of the day. Also, there is high correlation between the different lag terms.

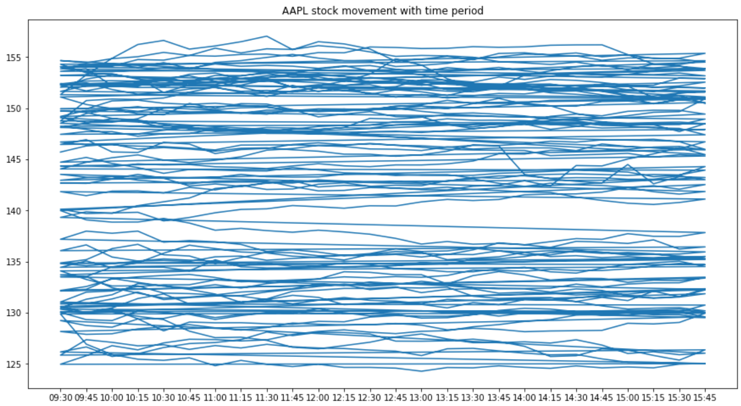


Figure 3: Sample Stock Movement with time of the day

If n-number of lag terms are used there will be multicollinearity hence, principal component analysis has been performed to visualise the data using only 3 lag terms. As we can see in the figure-4, the principal component 1 was able to capture the trend of the google stock. The first two principal components capture 91% of the closing price variance.

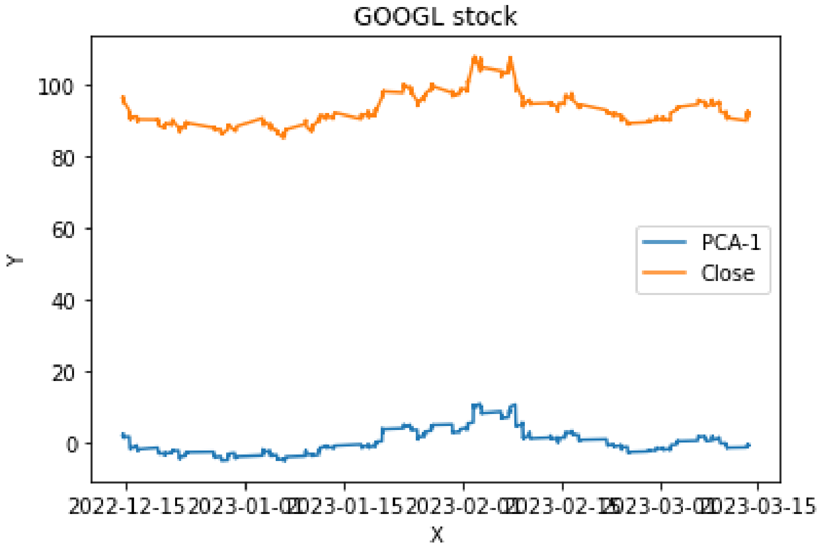
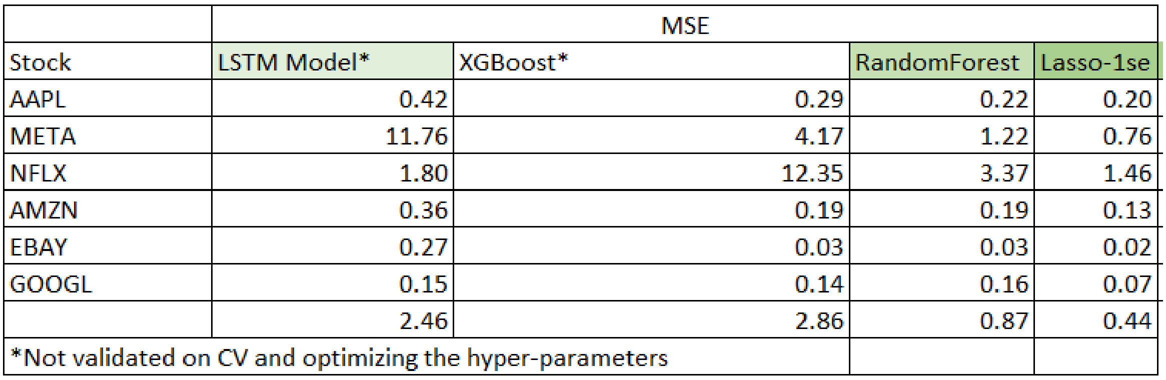


Figure-4: Sample PCA-1 and Closing Price Trend

Apart from using only 15-min interval data, we incorporated the day based interval as well. We observed correlation between the previous day closing price and the trading day prices.

Model Development & Results:



	MSE			
Stock	LSTM Model*	XGBoost*	RandomForest	Lasso-1se
AAPL	0.42	0.29	0.22	0.20
META	11.76	4.17	1.22	0.76
NFLX	1.80	12.35	3.37	1.46
AMZN	0.36	0.19	0.19	0.13
EBAY	0.27	0.03	0.03	0.02
GOOGL	0.15	0.14	0.16	0.07
	2.46	2.86	0.87	0.44
*Not validated on CV and optimizing the hyper-parameters				

Table-2 All Models Performance

Looking at our initial EDA, we saw a strong linear relationship between the explanatory and response variables. But we also saw multicollinearity between the independent variables. Hence we started our analysis by running Linear Regression. The model for each stock was trained by regressing Close price at time t using Close, Open, High, Low and Volume at t-1, t-2 and t-3 as well as the daily prices for these 5 attributes. For the linear regression models, the R square obtained was close to 1 which shows that the model has overfitted the training data. We saw that the close of lag 1 and time period variable were statistically significant in all the models. Time period was also significant in suggesting that there is a time trend. We noticed the mean square error for Netflix and Meta stocks were higher compared to the others. This indicates that these stocks are more volatile compared to the others. For all the individual stock models the condition number is large which suggests that there is multicollinearity among the independent variables.

We then ran a Lasso regression to see if we have a common set of features that could contribute to explaining the variation in the close price but we saw that different stock models have different sets of features that are important to the prediction. So, the initial assumption of one stock's volatility and movements impacting other stocks was violated. We moved to developing models for each individual stock.

Also, there are multiple regression models developed and tested based on the stock-market fundamentals, experimenting with different sets of features. We found that while increasing the number of lag-terms, the MSE decreased. AIC also decreased parallelly suggesting that the multi-component time-series analysis can be applied.

Next, we experimented with the non-linear models such as Random Forest and Boosting. We tuned the hyper-parameters with the time-series split cross validation and determined the optimal parameters for the models. As mentioned earlier, we tested on LSTM and the respective lag values, to see how the model performs, on the higher level with decent number of the layers, Lasso model performed better than LSTM, so instead of increasing the complexity of solution we stuck with the Linear regression model with the features selected through L1 norm.

The evaluation criteria for the stock market would be critical since, at the 15 minutes interval the less volatile stock wouldn't change much hence R-squared can't be used as a metric.

Comparing MSE with the average deviation of the stock between the intervals is a good indicator of prediction performance.

Through our evaluation and performance measures, we concluded that the linear regression performs better. Even the exploratory data analysis showcases the same as there is strong linear relationship between the closing price and its lag terms. The performance measures of all the different models have been highlighted in the Table-2, showcasing lasso selected features with linear regression performs better than the other models. Below figure-5 displays the model prediction and actual closing price for the test data.

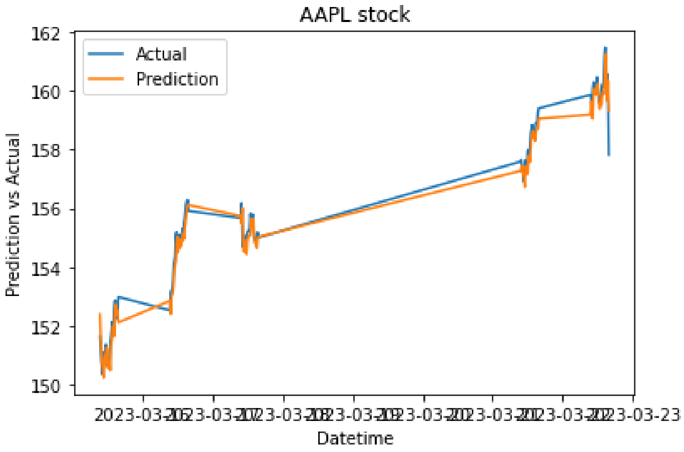


Figure-5 Final Model Actual vs Prediction

Recommendations and Business Value

To buy or not to buy? : Using our statistical models you will be able to predict the price of the stock in the next time window. This will tell you how you should trade in the next time period - if you should buy, sell or do nothing. For example, if the predicted close price of the stock is significantly higher than the price of the previous interval or price at which they bought the stock, traders may see this as an opportunity to sell the stock. On the other hand, if the predicted close price of the stock is shown to have a price drop, traders may see this as an opportunity to buy the stock. The prediction using the linear model we built has a lower mean squared error than the average deviance. Hence, the model can be used to make buy or sell decisions.

What's trending : Our analysis can also help understand the trends in the individual stocks. It tells you how the stock behaves at different times during the day. The trend of a stock during the day can provide insights into intraday market sentiment. If a stock is trending upwards during the day, it may indicate positive intraday market sentiment and investor confidence in the stock. Conversely, if a stock is trending downwards during the day, it may indicate negative intraday market sentiment and investor uncertainty. This model provides insights into the trend making it easier for traders to follow along and keep track of these stock prices.

Uncalculated Risk, No Reward : Understanding the trend of the stocks can also help traders manage their risk by identifying potential support and resistance levels. For example, if a stock has been trending downwards during the day and has established a resistance level, traders can use this level as a stop loss to limit their potential losses if the stock starts to rise. Since, the mode provides a 15 min interval prediction, it will capture a good amount of stock volatility. Any sudden increase or decrease will be highlighted and captured in the prediction by this model.

Move to a different beat : Our analysis also proves that different stocks have different patterns and move differently in the market. Hence it is important to understand not one model works for all and each stock needs to be analysed differently.

Having a good price prediction, understanding the trend of the stock during the day and knowing that each stock warrants individual analysis, our models and analysis gives a new trader the confidence they need to invest in stocks without going in blind and taking a huge risk.

The current model is unable to capture the extreme volatility in stock closing prices. This volatility could be due to external factors such as friction between countries, a major financial breakdown etc. The model could be fine tuned by feeding such information in the form of sentiments so the extremities could also be predicted up to an extent.

Summary and Conclusions

As we have seen, there are multiple techniques that can be used to predict stock price and trends. The linear model is giving the best prediction out of all the models explored. The models explored are linear regression, lasso regression, random forest and neural network. As a result, Netflix and Meta stocks are found to be more volatile than others. Another finding was that the volume and magnitude of the trade is higher in the early hours. This model prediction can be used to make investment decisions, track stock trends, and understand market volatility. Some of these techniques can be used in conjunction with the others to build superior models with higher predictive power. As a next step, the analysis so far can be used with other modelling techniques to understand the combined effect of some of these stock prices.