CS-1390/PHY-1390-1 Final Project Report: Stock Market Predictor (SMP): Predicting a closing price of Google.inc stock across a given period of time using deep learning

Bertrand Kwibuka
Computer Science and Economics Department
Ashoka University
Sonepat, India
January, 22nd, 2022

I. Introduction

The stock market is highly volatile, which is the extent to which a trading price series fluctuates over time. In that regard, investors need some information in order to make well-informed and rational decisions. Thus, a variety of mathematical and statistical techniques can be used to try and predict the stock market prices. The task of predicting stock prices is very challenging because of multiple (macro and micro) factors, such as politics, global economic conditions, unexpected events, a company's financial performance, etc.

Various techniques were used to predict the future value of the stock market as a result of the advent of machine learning. Even if It is difficult to forecast how the stock market will perform, through machine learning, you can discover the future value of company stocks and other assets traded on exchanges with the hopes that these predictions generate profits for investors.

In addition to physical and psychological factors, rational and irrational behavior are also considered in the prediction process. The combination of these factors causes share prices to fluctuate. Due to this, stock price predictions are less accurate than they could be. However, these factors also make it possible to detect patterns in a large amount of data. Financial analysts, researchers, and data scientists explore various analytical methods to detect stock market trends. As a result, algorithmic trading emerged and uses automated and pre-programmed strategies to execute orders.

II. Problem statement and its description

The idea of stock prediction is growing and gaining significant interest from various researchers around the world. Stock market investors make predictions about the future. It is important to seek methods and tools that will maximize profits while minimizing risks when predicting stock market prices. Stock market business involves complex and challenging predictions, which are fundamental to success.

Stock market analysts and indices generate a huge amount of data in time series format everyday. The problem here is founded in the idea on how to use this data to be able to predict future stock prices. Many strategies are already in use, but how to leverage machine learning techniques gives a high accuracy and performance on this problem. Thus, machine learning helps in discovering the future value of a company and other financial assets traded on a stock market exchange.

Traditionally, trying to predict the value of a stock on the market takes a gruesome effort from numerous people including financial analysts and those who use fundamental methods in finance. Stock analysis attempts to determine if a security's value is correct within the broader market. A fundamental analysis identifies securities that are not priced correctly by the market by taking a macro-to-micro perspective. In the process of arriving at a fair market value for a stock, analysts examine first the state of the economy and then the strength of the industry before concentrating on individual company performance.

To evaluate the value of a stock or any other type of security, fundamental analysis uses public data. Performing a fundamental analysis on the value of a bond, for instance, involves looking at a bond's value in light of the overall economy and interest rates, then taking into account information about the bond issuer, such as potential changes in its credit rating. As a measure of a stock's underlying value and potential for growth, fundamental analysis examines its revenues, earnings, future growth, return on equity, profit margins, and other factors. This data can all be found in a company's financial statements.

In this project, I leverage the use of machine learning to shorten this process and see if the predicted prices match the actual prices of the stocks on the market.

III. Literature Reviewed

Research on the stock market is one of the most important issues in recent years because of its nonlinear nature. The reliability of the prediction may not be guaranteed by traditional methods like fundamental and technical analysis. The majority of predictions are made using regression analysis. [8] experiments with well-known efficient regression methods to predict the stock market price from stock market data and conclude that multiple regression approaches can be improved by adding more variables. [8] mentions polynomial regression, radial basis function regression, and linear regression as possible means of forecasting stock market prices. Their main conclusion is, given the nature of the stock market data, that this regression analysis helps understand how the typical value of the dependent variable changes when any one independent variable is changed, while the other independent variables are held constant. In regression analysis, the underlying assumption is that the dependent variable will have the expected value given the independent variables, that is, the average value of the dependent variable, as long as the independent variables are constant.

[11] predict the market performance of Karachi Stock Exchange (KSE) on day closing using different machine learning techniques. In the model, different attributes are used as inputs and the market is predicted as positive or negative. They viewed attributes such as oil price, gold price, silver price, interest rate, forex rate, news feed, and social media rate. Additionally, they referred to the old statistical techniques, including Simple Moving Average (SMA) and Autoregressive Integrated Moving Average (ARIMA). The researchers examined machine learning techniques such as Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM). They found that the algorithm MLP performed best when compared to other algorithms, and that the oil rate attribute was most relevant to market performance. Based on their findings, the KSE-100 index can be predicted using machine learning techniques.

In [7], a Machine Learning approach was proposed, which was trained using stock data and gained intelligence. They applied the gathered knowledge for an accurate prediction. As part of the study, the authors used a machine learning technique called Support Vector Machine (SVM) to predict stock prices for large and small capitalizations as well as for three different markets, using prices on both a daily and up-to-the-minute basis. This SVM algorithm is used to process data from a large dataset, which is collected from different global financial markets and does not lead to overfitting. According to their findings, this method is more efficient and produces higher profits than the selected benchmark.

[6]] focuses on the use of Regression and LSTM based Machine learning to predict stock values by considering some factors of the data as open, close, low, high and volume. They highlighted that The vital part of machine learning is the dataset used and this dataset should be as concrete as possible because a little change in the data can perpetuate massive changes in the outcome. In this project, they used supervised machine learning on a dataset obtained from Yahoo Finance. They use regression and LSTM models with regression involved in minimizing errors and LSTM contributing to remembering the data and results in the long run. The most important result is that they found the LSTM Model to offer more accuracy than the regression based Model.

IV. Project description and goals

In this project, I use the LTSM (long short term memory) model and publicly available data from Alpha Vantage. I used Alpha Vantage because it uses an API which allows me to pull the data directly to my code without downloading it from public records such as Google Finance or Yahoo Finance. I specifically predict the stock price of Google.inc listed on S&P 500.

I used many common libraries in machine learning such as tensorflow, keras, folium and others to perform different functions. The LSTM I used was forked from the Neptune framework. Neptune provides specific frameworks of models online and a platform to directly visualize all logs, and performance metrics without spending much time doing it on a computer locally.

I arrived at the intended results by following a few critical steps: setting up the environment, downloading the data, pre-processing data, model development and hyperturning, and then finally predicting the stock closing price. My goal was to achieve a considerable high accuracy measured using RMSE(Root Mean Square Error) and MAPE (Mean absolute percentage error) as my model uses time-series data.

V. Dataset specification

I used publicly available S&P 500 index data on a company named Google with a ticker symbol GOOGL on the index. Google is a technology conglomerate company which started publicly selling its data in 2004. To download this data, I needed to find a way to incorporate it directly in my code without going through other databases such as Yahoo Finance and Google Finance for easily coding it. It is in this way that I used a platform called Apha Vantage which provides public API keys that pull data directly from the online finance databases and download it with a file format in an CSV format. The data used was a historical daily time series data dating for a period of time since Google.inc started selling its shares publicly till 2022. The data had attributes such as date, daily open, daily high, daily low, daily close, daily volume. I only took care of the variable 'daily close' because I needed to predict the daily close price of a stock. Daily close refers to the last price of a stock on a given trading day. I used this because it is the last daily price and the least volatile on a daily basis.

Data preprocessing process in this project consists of a way to aggregate the sequence of information because the data is time series data with a sequential nature consisting of observations taken at successive points in time (daily

basis). This was achieved by using the MA (Moving Average) technique to smooth out unnecessary short time fluctuations in the data. This MA calculates the average of a range of stock (closing) prices over a specific number of periods in that range.

Another critical point in preprocessing is using the 'StandardScaler' function to normalize the data. There are different ranges in this dataset and normalization helps in changing the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. This is also because the model doesn't know exactly the distribution of the data.

The last step in data preparation is to transform the data already in raw format into a predefined format that can be used in an LTSM model. This is achieved by using the function 'extractseqXoutcomeY'. This is useful because in order to predict n price, this function inputs vectors of $n-1$ data points prior and uses the n price as the outcome value [10].

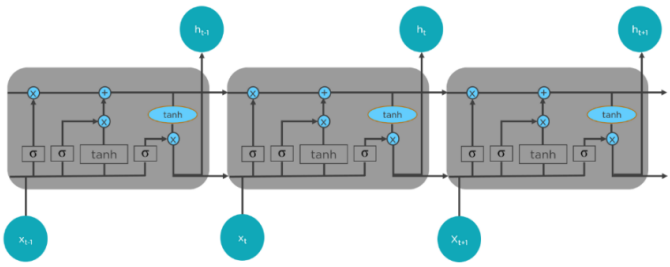
VI. Methods, Experiments and Explanation, Schematics

A. *Implementation details*

In this project, I use a Long Short Term Memory (LSTM) network for building my model to predict the stock prices of Google. LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning which uses feedback connections instead of feedforward neural networks. It took me some considerable personal time to study this model because it was not covered in class, but my associations with the field of economics pushed me to study it and understand it. LSTM is commonly used for processing and predicting time-series data.

In addition to single data points, LSTM can also process entire sequences of data (for example, audio and video). An example would be the recognition of unsegmented, connected handwriting, and speech recognition.

Time series analysis is made easier with LSTM. LSTM is capable of capturing historical trends and predicting future values more accurately. I will now give a brief overview of the LSTM model.



Schematic of LSTM method. Source: Google Images

As you can see in the image above, LSTMs are structured like chains. A general RNN has one layer of neural connections. In contrast, LSTMs have four interacting layers that communicate extensively.

LSTMs work in the following three-step process:

1) In LSTM, the first step is to decide which information should be omitted from the cell in that particular step. This decision is made using a sigmoid function. It computes the function based on the previous state (h_{t-1}) and the current input x_t
2) In the second layer, there are two functions. In the first case, it is the 'sigmoid function', and the second, it is the '$tanh$' function. By using the sigmoid function, we can decide which values to let through $(0 \text{or} 1)$. By using the $tanh$ function, the values passed are weighted according to their importance, ranging from -1 to 1.
3) In the third step, we determine what the final output will be. To begin, you must run a sigmoid layer that determines which parts of the cell state make it to the output. This means that the cell state must be put through the $tanh$ function so that the values are pushed between -1 and 1 and multiplied by the output of the sigmoid gate.

With this basic description of the LSTM used, the following parts of this report are standard machine learning processes that can be easily understood.

After the model development, I used the oldest 80% of the data for training, and saved the most recent 20% as the hold-out testing set. After this step, I trained the model, tested it and then calculated the performance.

VII. Observations

The performance of this LSTM based model is done using statistical methods RMSE(Root Mean Square Error) and MAPE (Mean absolute percentage error) because the data is in a time series format. Since this is not a regression method, these evaluations are the most adequate current model evaluation metrics.

RMSE indicates the difference between predicted and true values, while MAPE (%) indicates the difference relative to the true values. As an example, a MAPE value of 20% indicates that the expected stock price and the actual price differ by 20%. The following are the mathematical formulae of RMSE and MAPE.

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{At-Ft}{At}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(At-Ft)^2}$$

where N = the number of time points, At = the actual / true stock price, Ft = the predicted / forecast value [10].

Once the training was completed, I tested the model against the hold-out test set. I visualized all my graphs using the Neptune dashboard and the standard machine learning libraries.

My LSTM model achieved an RMSE of $114.38\$$ and MAPE = 3.35% ; which represents a relatively high accuracy compared to standard stock prediction methods such as simple moving averages and other regression based methods. Please refer to the references and literature review for any benchmark analysis on this.

The RMSE =114.38$ represent an error difference of 114.38$ between the actual and predicted prices. This is a pretty low number compared to how Google stock changed in the last 18 years with its high daily volatility considered.

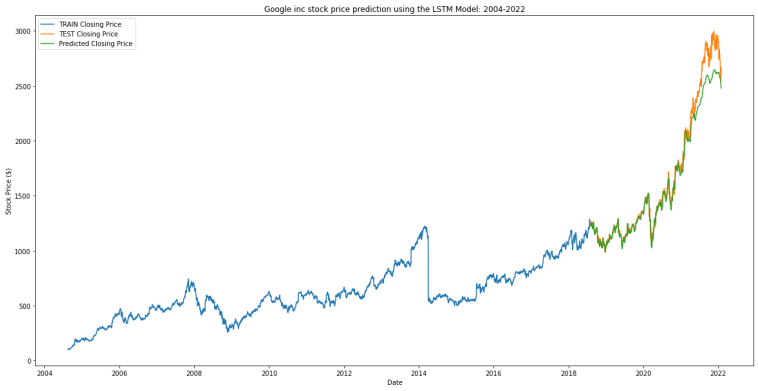
The MAPE of 3.35% a good performance of this LSTM: an error of 3% difference between the actual price vs the predicted price. However this can be slightly improved by tuning the parameters and playing around with other forms of data such as quartely and monthly returns which are less volatile than daily returns.

For this paper ergonomics, please find all graphs and performance metrics in the index.

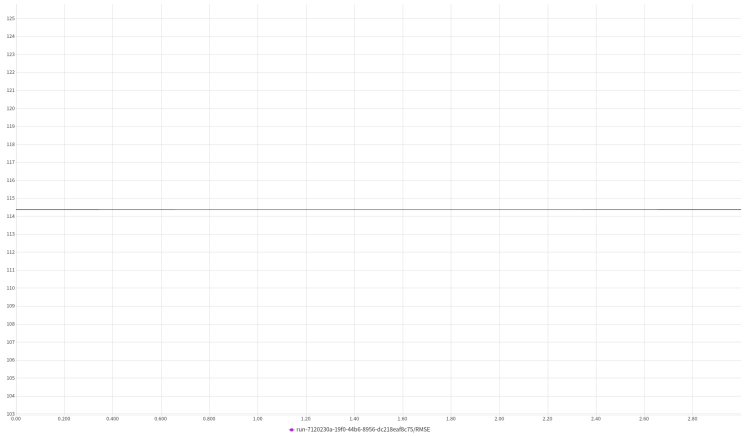
VIII. Conclusion and future directions

We have seen that the problem of predicting most accurate future stock stock prices is very founded and concerning for investors. Our daily lives are profoundly affected by the stock market. The stock market contributes significantly to a country's economy. The purpose of this project was to learn the basics of the stock market and how to make stock price predictions using the machine learning method LSTM. There are other alternative methods which are being improved which can be worked and compare their accuracies with LSTM. These include some regression methods, advanced forecasting economic algorithms, Echo State Networks (ESN), and others. The reality is that we are getting closer to eliminating the cumbersome lengthy process of playing a guess game with regards to a security price by taking advantage of the availability of large data sets and math tools which together make machine learning algorithms able to perform the work easily.

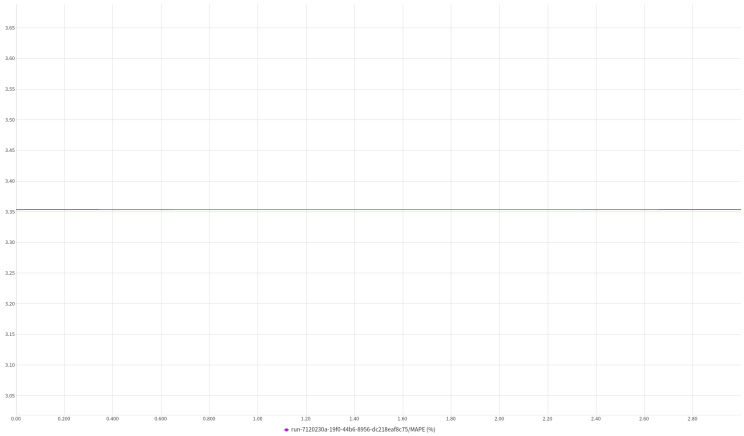
IX. INDEX



Plot of Stock Predictions with LSTM



Plot of RMSE. RMSE =114.38$



Plot of MAPE. MAPE= 3.35%

REFERENCES

[1] Arthur Charpentier, Romuald Élie, and Carl Remlinger. "Reinforcement learning in economics and finance". In: *Computational Economics* (2021). DOI: 10.1007/s10614-021-10119-4.

[2] Colah. *Understanding LSTM networks*. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[3] Scott Counts, Justin Cranshaw, and Stevie Chancellor. *Measuring Employment Demand with Search Data*. URL: https://www.msrinteractivescience.com/employment.

[4] Rajat Dhyani. *Stock-price-predictor: This project seeks to utilize deep learning models, long-short term memory (LSTM) neural network algorithm, to predict stock prices.* URL: https://github.com/Rajat-dhyani/Stock-Price-Predictor.

[5] *How to get the data you need from google's search analytics API*. URL: https://moz.com/blog/how-to-get-the-data-you-need-from-googles-search-analytics-api.

[6] Ishita Parmar et al. "Stock Market Prediction Using Machine Learning". In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. 2018, pp. 574–576. DOI: 10.1109/ICSCCC.2018.8703332.

[7] V Kranthi Sai Reddy. "Stock Market Prediction Using Machine Learning". In: *International Research Journal of Engineering and Technology* 5.10 (Oct. 2018), pp. 1032–1035.

[8] Ashish Sharma, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach". In: *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. Vol. 2. 2017, pp. 506–509. DOI: 10.1109/ICECA.2017.8212715.

[9] *Stock Market Predictor (SMP): Predicting a closing price of a stock across a given period of time using deep learning (LSTM)*. URL: https://github.com/GaelKBertrand/SMP-predictor-Google-stock.

[10] *Time Series: Predicting Stock Prices Using Machine Learning*. URL: https://neptune.ai/blog/predicting-stock-prices-using-machine-learning.

[11] Mehak Usmani et al. "Stock market prediction using machine learning techniques". In: *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. 2016, pp. 322–327. DOI: 10.1109/ICCOINS.2016.7783235.