Stock Price Prediction Using Machine Learning

Nalawade Sakshi[1], Mule Sarita[2], Adsare Monika[3]

[1-3]*Dept. of Computer Engineering, JOCE, Maharashtra, India*

Abstract - *The stock market or stock market is one of the most complicated and sophisticated ways of doing business. Small assets, brokerage firms, the banking sector, all depend on this same body for the distribution of income and risk; a very complicated model However, this document proposes to use a machine learning algorithm to predict the future price of shared resources for exchange using open source libraries and pre-existing algorithms to help make this business format from unpredictable to predictable. We will see how this simple application will bring acceptable results. The result is based entirely on numbers and assumes a large number of axioms that may or may not follow in the real world as the time of prediction.*

***Key Words*: basic concepts, data analysis, fundamental, implementation, linear regression, stock market, supervised machine learning**

1. INTRODUCTION

The stock market is one of the oldest methods by which a normal person trades stocks, makes investments and earns some money from companies that sell a part of themselves on this platform. This system turns out to be a potential investment scheme if done wisely. However, the price and liquidity of this platform are highly unpredictable and this is where we bring the technology to help us. Machine learning is one of those tools that helps us get what we want. The following 3 paragraphs will briefly explain the key components of this document:

The stock market, as we know, is a very important trading platform that affects everyone individually and nationally [2]. The basic principle is quite simple, companies will list their shares in companies as small commodities called shares. They do this to raise money for the company. A company lists its shares at a price called an IPO or initial public offering. This is the offer price at which the company sells the shares and raises money. After that these shares are owned by the owner and you can sell them at any price to a buyer on a stock exchange such as the BSE or the Bombay Stock Exchange. Traders and buyers continue to sell these shares at their own price, but the company can only retain the money earned during the IPO. The continued hope of the hare aside aside from realizing more profits, it translates into a particular increase in the share price after each profitable transaction. However, if the company issues more shares at a lower initial public offering, the market price for the exchange falls and traders suffer a loss. This exact phenomenon is the reason for the fear that people have when investing in stock markets and the reason for the fall and rise in stock prices in a nutshell.

Now, if we try to plot a chart of the stock market price over the time period (say 6 months), is it really difficult to predict the next outcome on the chart?

A human brain is very capable of stretching the graph by some coordinates simply by looking at it for a few minutes.

[1] And if we create a group of random people who try to extend the chart for a fixed period of time (say a week), we will get a very reasonable and approximate answer to a real-life chart .

Because many brains will try to interpret the scheme and make a hypothesis and this activity of this type has proved to be much more effective in practice than it appears in theory.

[5] That said, the best estimate of the real value of the stock is made using the crowd calculation method.

But since it is highly avoidable that crowd computing is a very slow activity, we tried to use a computer here to simulate such an example with a more scientific and mathematical approach.

In statistics, there is a way in which we look at the values and attributes of a problem in a graph and identify dependent and independent variables and try to establish or identify an existing relationship between them [3 and 4]. This technique is known as linear regression in statistics and is very commonly used due to its very simple and effective approach. In machine learning we have adapted the same algorithm in which we use the characteristics to train the classifier which then predicts the value of the tag with a certain accuracy that can be verified during the training and testing of the classifier. In order for a classifier to be accurate, you must select the correct characteristics and have enough data to train the classifier. The accuracy of your classifier is directly proportional to the amount of data provided to the classifier and the selected attributes. So with a basic knowledge of the stock market, charts and data analysis along with machine learning; now we are ready for the program device.

2. PREDICTION MODEL

A. Data Analysis Stage

In this phase, we will analyze the available raw data and study it to identify the appropriate attributes for the

prediction of our selected tag.

Now the data we will be using for our program is taken from www.quandl.com, a world-class dataset delivery platform.

At this stage, we will analyze the available raw data and study it to identify the appropriate attributes for the prediction of our selected tag.

Now the data we will be using for our program is taken from www.quandl.com, a world-class dataset delivery platform.

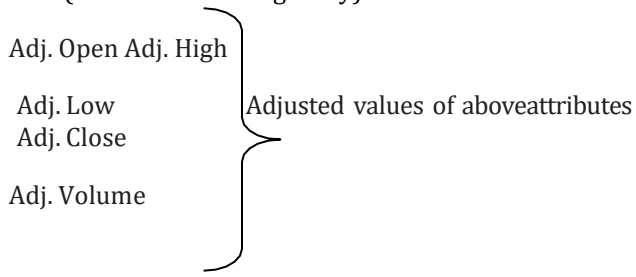
The dataset taken is for GOOGL from WIKI and can be extracted from quandl using the "WIKI/GOOGL" token. We extracted and used about 14 years of data.

Dataset attributes include:

Open (opening price of shares)

High (highest possible price in a time instance) Low (lowest possible price in a time instance) Close (closing price of shares)

Volume (total trades during a day)



We select the "Close" attribute as our label (The variable we are going to predict) and use "Adj. Open, Adj. High, Adj. Close, Adj. Low and Adj. Volume" to extract the characteristics that will help us to better predict the result.

It should be noted that we use values mounted on raw, since these values are already processed and free of common data collection errors.

We now know that charts made for stock analysis use the above attributes to track them. This charts are called OHLCV

charts [11] and are very informative about the state of stocks. Now we use the similar graphical parameters to decide the characteristics of the classifier.

We define the set of features that we will use:

• Closing: It is an important source of information, since it determines the market opening price for the next day and the expected volume for the day.

• HL_PCT: this is a derived characteristic defined by:

$$HL\_PCT = \frac{\text{Adj. High} - \text{Adj. Low}}{Adj.\,Close} \times 100$$

We use percentage change, as this helps us to reduce the number of functions, but to retain the net information involved. High-Low is a relevant feature because it helps us formulate the shape of the OHLCV chart.

• PCT_change: this is also a derived characteristic,

defined by:

$$PCT\_Change = \frac{Adj.Close - Adj.Open}{Adj.Open} \times 100$$

We do the same treatment with Open and Close as High and Low, since both are very relevant in our forecasting model and help us to reduce the number of redundant functions as well.

• Volume: This is a very important decision-making parameter since the volume traded has the most direct impact on the future stock price than any other feature. Then we will use it as in our case.

We have successfully analyzed the data and extracted the useful information that we will need for the classifier. This is a very crucial step and should be treated with extreme care.

A lack of information or a small error in the derivation of useful information will lead to a fault prediction model and a very inefficient classifier.

In addition, the extracted features are very specific to the theme used and will certainly vary from theme to theme. Generalization is possible if, and only if, the data of the other subject are collected with the same consistency as the previous subject.

B. Training and testing stage

Training and testing phase In this phase, we will use what we have extracted from our data and implemented in our machine learning model.

We will impliment the SciPy, Scikit-learn and Matplolib libraries in python to program our model, train them with the features and labels we have extracted and then test them with the same data.

First we will process the data to make the data that includes:

Moved values of the label attribute of the percentage that you want to predict.

• The Dataframe format is converted to the Numpy array format.

• All NaN data values removed before being sent to the classifier.

• The data is scaled in such a way that for any X value,

$$X \in [-1,1]$$

• The data is divided into test data and training data according to the type, i.e. label and characteristic.

Now the data is ready to be entered into a classifier. We will use the simplest classifier, namely the linear regression, defined in the Sklearn library of the Scikit-learn package.We chose this classifier because of its simplicity and because it serves our purpose in the right way. Linear regression is a widely used technique for data analysis and prediction. It essentially uses key features to predict the relationships between variables based on their dependencies on other features. [9] This form of prediction is known as supervised machine learning. Supervised learning is a method in which we enter tagged data, that is, the characteristics are matched to their tags. Here we train the classifier in such a way that it learns the patterns of which combination of characteristics results in which label.

Here, in our case, the classifier sees the features and simply looks at their label and remembers it. Remember the combination of features and their respective label which in our case is the stock price a few days later. Then go ahead and find out which model is followed by the features to produce the respective label. This is how supervised machine learning works [10]. For tests in supervised machine learning, we put a combination of features into the trained classifier and check the output of the classifier with the actual tag. This helps us determine the accuracy of our classifier. Which is very crucial for our model. A classifier with an accuracy of less than 95% is practically useless.

Accuracy is a very crucial factor in a machine learning model.

It is necessary to understand what accuracy means and how to increase its accuracy in the next subtopic.

C. Results

Once the model is ready, we use the template to get the desired results in any way we want. In our case, we will plot a graph of our results (fig. 1) according to our requirements that we discussed earlier in this document.

Fig. 1. Graph showing stock price of GOOGL from year 2005 till July 2018. Red is the line representing given data and blue is representing the forecasted or the predicted value of stock.

It must be according to our needs, and as mentioned above, a model with an accuracy of less than 95% is practically useless. There are some standard methods for calculating accuracy in machine learning, some are as follows:

• R2 value of the model.

• Adjusted R2 value

• RMSE value

• Confusion matrix for classification problems.

• And many more

Fig. 2. Graph showing the exact amounts of predicted values.

Accuracy is the component that every machine learning developer is always committed to contributing to. After the model has been developed, there is an endless effort to optimize the model for more and more accurate results. There are some very common and simple ways to increase the efficiency of the model and they have been discussed

above.

However, let's look at some of the standard ways to optimize a machine learning algorithm:

• Unrestricted optimization

o Decent gradient

o Newton's method

o Batch learning

o Decent stochastic gradient

• Limited optimization

o Lagrange duality

o SVM in primary and double forms

o Limited Methods

Most machine learning problems are, in the end, optimization problems, where we minimize a function subject to some restrictions.

3. USEFUL TIPS

A. Requirements and specifications

You need to thoroughly know the exact requirements of the problem and the specifications of the machine and performance as a first step. Do not rush this step, as this step is very crucial in deciding on the overall plan for the development of the program. Carefully study the case, do a little background check, gather extensive knowledge of the topic in question and identify what you really want and set as a goal.

B. Careful analysis of the function

You need to be very careful when deriving data characteristics, as they play a direct role in the forecasting model. Everyone should have direct meaning along with labels. Minimizing functions subject to requirements restrictions, as far as possible, is also highly recommended.

C. Application

You must select the appropriate model in which you will implement mathematics to produce results.The selected or designed model must be in combination with the input data. An incorrect template designed or selected for inappropriate data, or vice versa, will result in a completely useless junk model. You must refer to supported SVMs or other available methods to process data. Testing several models at once to see which one works most effectively is also a good practice.

In addition, implementation is the easiest step of all and should take the least amount of time to save us some time from the total cost of time that could be used on some other important steps.

D. Training and testing

Training a model is very easy. You just need to make sure that the data is consistent, consistent, and available in great abundance. A large training dataset contributes to a stronger and more accurate classifier that ultimately increases overall accuracy.

Testing is also a very simple process. Make sure your test data is at least 20% the size of your training data. It is important to understand that tests are the test of the accuracy of classifiers, and sometimes it is observed that they are inversely proportional to the score of the classifiers. However, the accuracy of the classifier has no dependence or correlation with the evidence. Sometimes it seems so, but the tests have no relation to the classifier.

E. Optimization

It is almost impossible to create a versatile classifier at once, so we must always continue to optimize. There is always some chance for improvement. When optimizing, consider standard methods and basic requirements.

Switch to SVM, test and test different models, look for new and improved features, edit the entire data model to fully fit the model, etc. are some very basic ways to optimize your classifier.

4. SOME COMMON MISTAKES

We mention some of the common mistakes made by professionals in this field, which you should avoid[12]:

• Poor annotation of training and test datasets

• Poor understanding of algorithm assumptions

• Poor understanding of algorithm parameters

• Lack of understanding of the goal

• Do not understand the data

• Prevent leaks (features, information)

• Insufficient data to train the classifier

• Use machine learning where you don't need it

5. CONCLUSIONS

Machine learning as we have seen so far, is a very powerful tool and therefore avoidable, it has a great application. So far we have seen that machine learning relies heavily on data. Therefore, it is important to understand that data is quite valuable, and as simple as it may seem, analyzing the data is

not an easy task.

Machine learning has found an amazing application and has evolved further into deep learning and neural networks, but the basic idea is pretty much the same for all of them.

This document provides a fluid view of how to implement machine learning. There are various ways, methods, and techniques available to manage and solve various problems, in different imaginable situations. This document is limited to supervised machine learning only and seeks to explain only the fundamentals of this complex process.

6. Acknowledgment

This article is written on the basis of the author's summer project during his degree course in 2018 after the valuable guidance of Neha Agarwal (Asst. Prof.) AUUP, Noida.

REFERENCES

[1] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore "A Machine learning approach to Building domain-specific Search engine", IJCAI, 1999 - Citeseer

[2] Yadav, Sameer. (2017). STOCK MARKET VOLATILITY - A STUDY OF INDIAN STOCK MARKET. Global Journal for Research Analysis. 6. 629-632.

[3] Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[4] Draper, N.R.; Smith, H. (1998). Applied Regression Analysis (3rd ed.). John Wiley. ISBN 0-471-17082-8.

[5] Robert S. Pindyck and Daniel L. Rubinfeld (1998, 4h ed.). Econometric Models and Economic Forecasts

"Linear Regression", 1997-1998, Yale University http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[6] Agarwal (July 14, 2017). "Introduction to the Stock Market". Intelligent Economist. Retrieved December 18, 2017.

[7] Jason Brownlee, March 2016, "Linear Regression for machine learning", Machine learning mastery, viewed on December 2018, https://machinelearningmastery.com/linear-regression-for-machine- learning/

[8] Google Developers, Oct 2018, "Decending into ML: Linear Regression", Google LLC, https://developers.google.com/machine-learning/crash-course/descending-into-ml/linear-regression

[9] Fiess, N.M. and MacDonald, R., 2002. Towards the fundamentals of technical analysis: analysing the information content of High, Low and Close prices. Economic Modelling, 19(3), pp.353-374.

[10] Hurwitz, E. and Marwala, T., 2012. Common mistakes when applying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429.