2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Prediction of Stock Prices using Machine Learning (Regression,Classification) Algorithms

Srinath Ravikumar
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
srinath.ravikumar16@vit.edu

Prasad Saraf
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
prasad.saraf16@vit.edu

***Abstract* - The stock market is an interesting industry to study. There are various variations present in it. Many experts have been studying and researching on the various trends that the stock market goes through. One of the major studies has been the attempt to predict the stock prices of various companies based on historical data. Prediction of stock prices will greatly help people to understand where and how to invest so that the risk of losing money is minimized. This application can also be used by companies during their Initial Public Offering (IPO) to know what value to target for and how many shares they should release. So far there have been significant developments in this field. Many researchers are looking at machine learning and deep learning as possible ways to predict stock prices. The proposed system works in two methods – Regression and Classification. In regression, the system predicts the closing price of stock of a company, and in classification, the system predicts whether the closing price of stock will increase or decrease the next day.**

Keywords—Stock prices, stock market, machine learning, regression, classification, confusion matrix

I. INTRODUCTION

Stock exchanges are the financial institutions which allow exchange of different types of goods between stock broker components. Stock market prediction is the method of determining the future value of a stock or other financial instrument traded on an exchange. A misconception is also associated with people that buying and selling of the stocks/shares in the market is an act of gambling. This misconception can be changed and bringing awareness among people for this.

Over the past few years, 90 percent of the data in the world has been created as a result of the creation of 2.5 quintillion bytes of data on a daily basis. A very large amount of data is generated by financial market. It's very difficult for a trader to recognize a pattern and then devise an optimal strategy for making decisions. Predicting how the stock market will perform is one of the most difficult things to do. There are so many factors involved in the prediction – physical factors vs physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.[1]

Machine Learning can be used as a game changer in predicting the values of stock prices. Machine learning techniques have the potential to unearth patterns and insights we didn't see before, and these can be used to make unerringly accurate predictions. The machine learning is growing at a phenomenal pace in today's world.[2]

II. LITERATURE SURVEY

This survey involved a lot of study and analysis of fields like Deep Learning, Stock markets, Factors affecting stock prices, etc various studies have shown prediction of stock prices considering various factors.

A study which gave a theoretical approch to predict stock prices with the help of basic regresion models was developed by Ashish Sharma and his colleagues[7].The study explained the regression models and how their application can be useful in price prediction. This study gave us an idea about how chosing appropriate factors affecting stock price as variables can give some predictions.

This work needed an appropriate set of features to predict most accurate value. A study by Amit Kumar Sirohi gave a detailed explanation of an 2-tier model[8] in which first tier gave information about feature selection like opening price, closing price of a stock, etc and the second tier build various kernels on the extracted features. This gave us a fair idea about apply suitable features to models.

Later a study by Pushkar Khanal and Shree Raj Shakya[5] stated that Support Vector Machine algorithm gave best results for prediction with an effective accuracy as compared to most other machine learning algorithms and traditional technical methods. This study effectively explained how classification can be useful in prediction.

A study by Yaojun Wang and Yaoqing Wang[6] explained the effect of the stock comments information from social media with help of social media mining on the stock price. Bring this factor in consideration along with other important factors led to more accurate results. This study applied SVM with emotion index effectively.

Later it was realized that Deep learning models could improve the accuracy significantly. Rohit Verma with his colleagues proposed a theoretical approach for using Artificial Neural Network to predict stock prices[4].Here the results were obtained from Nifty stock index dataset on the basis of values from the past days. An accuracy of 96% was obtained from this study.

III. METHODOLOGY

The functioning of the system mentioned in this paper is given below:

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

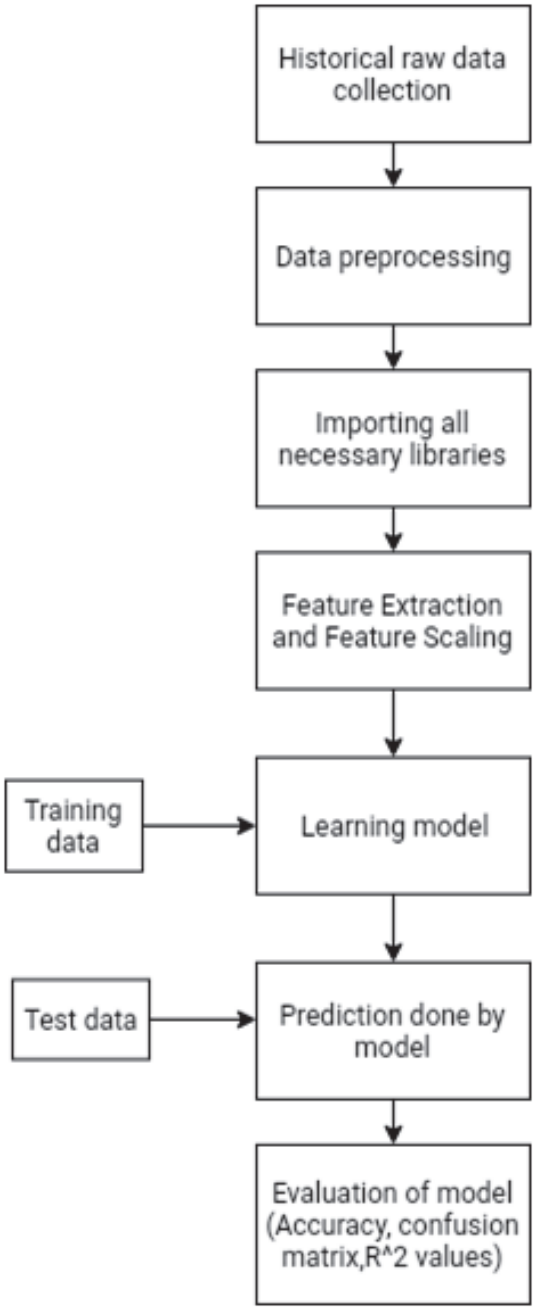

Fig. 1. Flowchart of application

1. The raw data used is from Yahoo finance[3].
2. The attributes used for feature extraction are ‘date’ and ‘closing price’ of a stock.
3. Features used to predict the momentum of stock price of particular company are ‘stock momentum’, ‘index volatility’, ‘sector momentum’. These features are scaled.
4. The dataset is split into training and test data set.
5. The training dataset is used for model training and test dataset is used for prediction .The significance of a feature is determined using the R^2 values
6. The values of the test data are predicted and the results are evaluated. The result is given on the basis of accuracy, confusion matrix and time required for the model used.

IV. Technical Terms

Before proceeding to the machine learning model, it is necessary to understand the technical terms related to finance and stock market in general.

A. Stock Market Index

Sometimes referred as index, it gives a measure of the relative value of the group of stocks within the index in numerical terms. If the stocks within an index change their values, the value of the index gets affected.

B. Outstanding Shares

They refer to a company's stock currently held by all its shareholders, including share blocks held by institutional investors and restricted shares owned by the company’s officers and insiders.

C. Market Capitalization

It refers to the total dollar market value of a company's outstanding shares. Usually referred to as ‘market cap’, it is calculated by multiplying a company's outstanding shares by the current market price of one share.

D. S&P500

It is an American stock market index with market capitalizations of 500 large companies that have common stock listed in NASDAQ, NYSE, etc.

V. Dataset Selection and Preprocessing

Yahoo finance provides an easy way to fetch any historical stock values of a company with the help of the ticker-name programmatically using in-built API's. It provides a feature to get prices with initial-date and final-date provided.

The companies which have been considered here are belonging to the S&P500 index. The index is used to reflect the price fluctuation and performance of the large 500 companies stocks traded in NYSE and NASDAQ. But, why choose S&P500? The U.S economy affects the stock market of all other countries. From a global perspective, the U.S accounts for more than half of the global stock market. S&P500 accounts for 80-85% of market-cap in U.S stock market and the current value is 23.77 trillion dollars. In most of the markets, S&P500 based products are among the most traded, liquid and most invested among all the other available alternatives. Also, S&P500 represent three-quarters of a stock-market in terms of market capitalization. And, the S&P500 is market cap weighted index, meaning that the companies with larger stock value have got more influence on the index. Thus, S&P500 is referred to as a benchmark to determine the state of the overall economy. And that's why, many investors and fund-managers compare their performance against S&P500.

The list of S&P500 companies are obtained by scraping the wikipedia page: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

The companies with ticker names and the sector they belong to are fetched. Using the ticker values of a company the historical stock data of a company is acquired from Yahoo finance.

The data is present in a raw format and is not feasible for analysis. The data contains highest value, lowest value, opening value, closing value and volume of traded stocks for a given particular date. Out of these, the date and closing value of the stock are of utmost importance to us. Using the closing value of a stock we calculate two more parameters – 'Momentum' and 'Volatility' for each company, sector and index. For each company in the dataset, the corresponding stock momentum, sector momentum and index momentum are considered. The volatility of the company, sector and index are also considered. This is done for each and every company.

Now that the dataset is clean, it can be fed to a machine learning model.

VI. Algorithmic Approach

The first step in building a machine learning model is to obtain an optimal dataset. The open soured data which is available on the internet consists of many discrepancies like having missing data, having repeated rows of the same data, data being unstructured etc. Before feeding the data to the machine learning model, the data needs to be modified of preprocessed so that the model is able to deliver the results which are as accurate as possible.

The main attributes that are found in financial datasets (historical data about stock prices of a particular company) are as follows:

1. Date of that particular stock price
2. Opening stock price
3. High stock price (highest value of that stock during that day)
4. Low stock price (lowest value of that stock price during that day)
5. Closing stock price
6. Volume of stocks traded

Of all these above parameters, the closing price is predominantly used as an attribute to feed the model. Using this single value, the future stock price of a company can be predicted using various regression models available in machine learning. In regression, according to the input given, a curve is plotted in a graph. The curve represents the variations in the stock prices over the years. Here, the X-axis will contain the date of the stock and the Y-axis will contain the closing price of a stock.

The regression models that the dataset will be applied on are as follows:

1. Simple Linear Regression
2. Polynomial Regression
3. Support Vector Regression (SVR)
4. Decision Tree Regression
5. Random Forest Regression

However, the closing price of a stock does not give the user a lot of information of the future prices of that particular stock. So, the following parameters are added to the dataset. They are as follows:

A. Momentum

If the current day closing price is greater than the previous day's closing price, current day's momentum is 1, else it is 0

B. Volatility

It is calculated as follows

$$\text{Volatility}[i] = \frac{\text{close}[i-1] - \text{close}[i]}{\text{close}[i-1]}$$

close[i]: current day stock closing price

close[i-1]: previous day stock closing price

C. Index Momentum

It is an average of previous 5 days index momentum. NASDAQ is the index used in this application.

D. Index Volatility

It is calculated as an average of Volatility of index for previous 5 days.

E. Sector Momentum

It is calculated as an average of momentum value for all the companies belonging to a particular sector.

F. Stock Momentum

It is an average of previous 5 days momentum of a company.

G. Stock Price Volatility

It is an average of previous 5 days volatility of a company.

Considering the new attributes, a more accurate and precise model can be created using classification models. Classification models provide the result whether a particular input belongs to which of the classes mentioned. Binary classification is done, where the result is of Yes/No type. In this application, the result will be based on momentum. The model will provide with the result whether the stock will go high or low the next day depending on all the past data.

The various classification algorithms applied are as follows:

1. Support Vector Machine (SVM)
2. K – Nearest Neighbors (KNN)
3. Logistic Regression
4. Naïve Bayes
5. Decision Tree Classification
6. Random Forest Classification

In the case of SVM, there are various kernels that are implemented for the dataset.

The kernels are as follows:

1. linear
2. poly
3. rbf
4. sigmoid

VII. Implementation

The above models are implemented through Spyder. The specifications of the laptop on which the model was implemented are mentioned below:

- Company – HP
- Processor – Intel Core i5-7200U CPU @ 2.50GHz 2.71GHz
- OS – Microsoft Windows 10
- RAM – 8 GB DDR4
- Graphics – 4 GB NVIDIA GeForce 940MX
- Storage – 1 TB SATA HDD

VIII. Results and Experiment Evaluation

The results of the experiment are mentioned below.

A. Regression Models

1) Simple Linear Regression

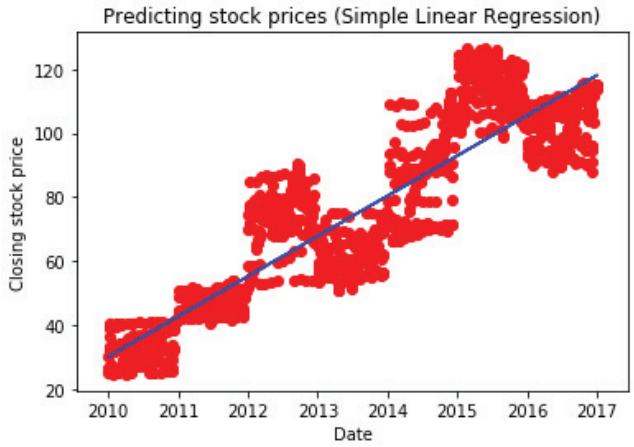

Fig. 2. Simple Linear Regression

2) Polynomial Regression

Degree = 10

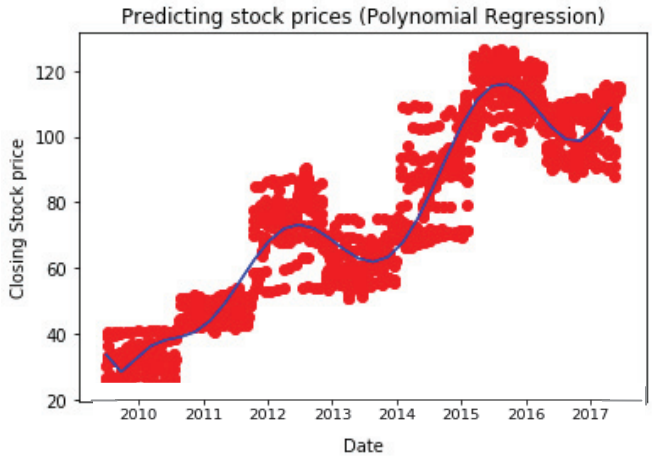

Fig. 3. Polynomial Regression

3) Support Vector Regression (SVR)

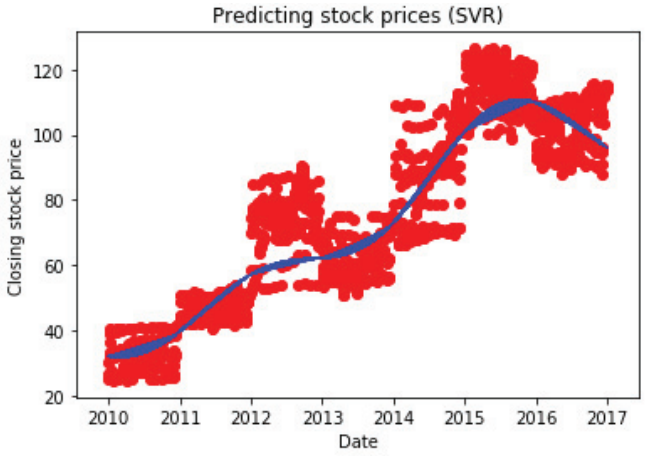

Fig. 4. Support Vector Regression

4) Decision Tree Regression

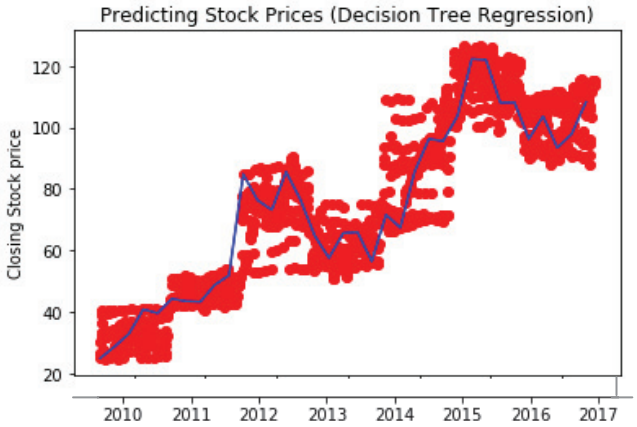

Fig. 5. Decision Tree Regression

5) Random Forest Regression

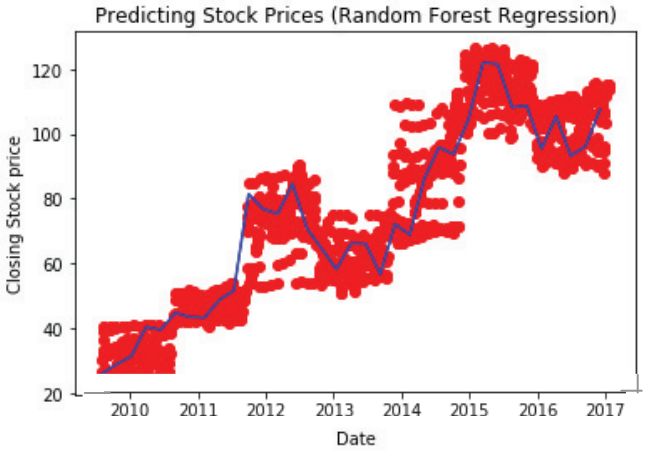

Fig. 6. Random Forest Regression

TABLE I. Result Analysis of Regression Models

Model	Accuracy	Time(in seconds)
Simple Linear Regression	81.52	0.77
Polynomial Regression	91.45	0.98
Support Vector Regression (SVR)	87.41	1.16
Decision Tree Regression	98.09	0.79
Random Forest Regression	99.57	1.06

Note that the accuracies mentioned above are not a direct indicator of the power and precision of that particular algorithm or model. It largely depends on the input dataset fed to the model. For example, in the case of stocks of Apple company, all the regression models gave high accuracy values, but it may not be so for other datasets.

B. *Classification*

TABLE II. RESULT ANALYSIS OF CLASSIFICATION MODELS

Model	Acc.	Time(s)
Support Vector Machine (linear)	68.41	158.48
Support Vector Machine (poly)	64.80	195.38
Support Vector Machine (rbf)	67.86	201.15
Support Vector Machine (sigmoid)	58.65	160.81
K – Nearest Neighbors	61.50	19.02
Logistic Regression	68.27	10.51
Naïve Bayes	67.10	10.14
Decision Tree Classification	57.99	198.57
Random Forest Classification	63.33	202.54

Confusion matrix –

A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known.

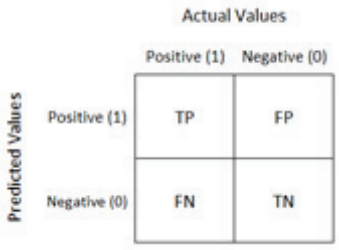

Fig. 7. Confusion Matrix

The following convention is followed in the table given below.

- TP – True Positive
- FP – False Positive
- FN – False Negative
- TN – True Negative

TABLE III. CONFUSION MATRIX VALUES OF CLASSIFICATION MODELS

Model	TP	FP	FN	TN
SVM (Linear)	178	79	90	228
SVM (Poly)	136	121	63	255
SVM (rbf)	170	87	93	225
SVM (sigmoid)	153	104	138	180
K – Nearest Neighbors	153	104	109	209
Logistic Regression	174	83	88	230
Naïve Bayes	177	80	102	216
Decision Tree Classification	141	116	126	192
Random Forest Classification	165	92	122	196

IX. CONCLUSION

The input dataset obtained from Yahoo Finance was efficiently preprocessed and various significant attributes were added to the dataset like momentum, volatility and sector details[3].

The stock price and momentum of Apple stock dataset was predicted and the accuracies of the various machine learning models were compared and analyzed. Learning and understanding the various terminologies and techniques present in the stock market was very helpful in preprocessing the dataset in order to achieve best possible results. The Logistic Regression Model gave maximum mean accuracy of 68.622%.

X. FUTURE WORK

Various neural network techniques can be applied to the processed data to make the machine more powerful. Advanced neural network techniques can be applied to the processed dataset. The neural network techniques make use of time series to accurately predict the stock values. Various other features like asset value, equity ratio, etc. can be taken into consideration to further improve the accuracy. Related tweets from twitter can also be considered while predicting the future stock prices.

The methods mentioned in the paper can be applied to real time data to get real time predictions of the stock value. Thus it can be deployed in real world application providing a more reliable way of understanding stock price changes. Accurate graphs can be plotted for a particular company to understand the patterns of stock values and thus making it easy to understand specific patterns.

REFERENCES

[1] TrevirNath "How Big Data Has changed finance" https://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp.

[2] Saheli Roy Choudhury."Machines will soon will be able to learnwithout being programmed" https://www.cnbc.com/2018/04/17/machine-learning-investing-in-ai-next-big-thing.html

[3] Yahoo finance data -https://in.finance.yahoo.com/

[4] Rohit Verma Astral institute Technical and Research, Indore (M.P.) , PROF. Pkumar Choure (CSE) Astral institute Technical and Research, Indore (M.P.)"Neural Networks through Stock Market Data Prediction " International Conference on Electronics,Communication and AerospaceTechnology ICECA 2017

[5] Pushkar Khanal ,Shree Raj Shakya "*Analysis and Prediction of Stock Prices of Nepal using different Machine Learning Algorithms*" Department of Mechanical Engineering, Pulchowk campus, Institute of Engineering, Tribhuvan University, Nepal

[6] Yaojun Wang, Yaoqing Wang "Using Social Media Mining Technology to Assist in Price Prediction of Stock Market" .

[7] Ashish sharma , dinesh bhuriya , upendra singh "Survey of Stock Market Prediction Using Machine Learning Approach "International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.

[8] Amit Kumar Sirohi, Pradeep Kumar Mahato, Dr. Vahida Attar "Multiple Kernel Learning for Stock Price Direction Prediction " IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014),August 01-02, 2014, Dr. Virendra Swarup Group of Institutions, Unnao, India

[9] Book-Elements of Artificial Neural Network,Kishan Mehrotra,Vhilukuri K K Mohan,Sanjay Ranka.