Evaluating future stock value asset using machine learning

R. Soujanya [a,*], P. Akshith Goud [b], Abhishek Bhandwalkar [b], G. Anil Kumar [a]

[a] Computer Science and Engineering, GRIET, India
[b] Department of CSE, GRIET, India

ARTICLE INFO

ABSTRACT

Stock Trading is one of the intuitive ways through which people make money. It operates just like an auction bridge where purchaser and trader negotiate prices and make trades. Stock trading involves putting our calculative minds and applying for some crunch numbers on the graphs to get the desired profits. But with the advancement of technology stock trading has reached another level. With algorithmic trading forex bots and predictions coming into the picture, people are trying to make much more money than the usual ways which involves going to the market with their gut feeling. This paper will be telling about stock prediction using machine learning as a tool to determine the future value of a stock. There is a lot of quantitative and technical analysis that goes into the stock prediction when done by the stockbrokers which are accurate up to a great extent or sometimes isn't. But there is a better scope to improve our prediction by using certain machine learning models. In this paper, the model will be using the linear regression model to predict stock prices for capitalizations in different markets employing assets with daily and up to date minute frequencies.

1. Introduction

The stock market is pretty volatile and its indices keep on fluctuating which creates a sense of disbelief in the trader while investing. Two types of stock analysis are taking into consideration. The first one is the fundamental analysis where the brokers look at the performance of the industry, political factors, economy, etc [1]. The second one is the technical analysis which involves applying statistical models to the market activity over some time. Stockbrokers use this information to trade on platforms. Stock prediction has come into the picture these days and people are leaning towards it in higher numbers. In recent years the dominance of machine learning has led many industries and firms to apply machine learning models to the field which has led to many promising results.

The environment of the share market is highly dynamic and susceptible mainly because of its underlying factors like the performance of a particular company, rumours or fundamental reasons like the P/E ratio, Moving Average, etc. So, based on these factors the basic prediction models consist of just forecasting the market or prediction restricting to a particular attribute. This paper will develop a prediction model that will help a person to trade with less uncertainty of losing his/her money over gut feeling. The model makes use of historical data which is a dataset containing stock prices. The dataset is split into a train set and testing set later fed to the machine learning model of this data [2]. The main perspective of our predictive model would be finding the safest share to invest in and maximize a user's investment profits.

2. Machine learning in stock prediction

Machine Learning: Machine learning is an subset of Artificial Intelligence application which helps the system by providing ability to grasp automatically and boost itself based on experience without being programmed explicitly. Stock: Stock (also capital stock) of an enterprise, is all of the shares into which ownership of the enterprise is divided. Stock Prediction: Stock prediction is the process of predicting the future value of the stock by using means of prediction models that apply technical and statistical analysis by means of mathematical logic. Linear Regression: It is a machine learning model that is used for the prediction of known parameters that are correlated with the outputs [3]. Frequent techniques can be changed in the industry at any time by tweaking its

* Corresponding author.
E-mail address: soujanya.manik24@gmail.com (R. Soujanya).

functionality and trying to come up with different ways to approach a particular problem. The use of machine learning in any field comes with results that make people ecstatic. When it comes to stock prediction, people always tend to hesitate to move towards the technical approach involving machine learning models because of the stock markets' volatile nature and susceptibility towards quick changes. But the advancements of machine learning models have invigorated people's choice to switch to algorithmic trading [4]. The use of predictive models in the stock market gives a detailed analysis of the data and various factors like the moving average, P/E ratio or the earnings per share which affect a stock's value. Many machine learning models ranging from the simple linear regression and SVM to the advanced Auto ARIMA and LSTM are used for prediction in the financial industry [13]. In this paper, we use the linear regression algorithm to develop a model that is used to predict the average price of a particular stock. Linear Regression is used in a plethora of applications driven by the financial industry.

2.1. Analysing large data in finance

Analysing data in stock analysis is one of the big and complex issues faced by developers. People often tend to rely on their imagination and fail to put the image in the form of code or an algorithm that is plausible in the real world [5]. The stock market is driven by many attributes like the previous close price, the open price, high price, low price, average price, turnover, etc. It comes to the developer to choose which attributes he wants to inculcate in developing the model. Most of the time, people use the open price, close price and try to predict the adjusted close price. The data is fetched from many sources like yahoo finance or NSE's website. Extraction of data can be simply done by web scrapping in python or individually selecting the number of years of data we want and downloading the .xlsx file. The paper discusses the development of the predicting model using the python programming language. The model hits the NSE website to web scrap the data and store it in an excel sheet producing a dataset which is then processed to store the necessary attributes in the SQL database [6].

2.2. The most effective method to use machine leaning in stock prediction

A prediction model is said to be effective when it ticks most of the factors like accuracy, precision, speed, etc. But the most important point to use machine learning in stock analysis is to increase the accuracy of prediction so that the trader can have a robust view and confidence about buying or selling a particular share. Mainly the prediction model must take into consideration the plethora of factors that are responsible for the moment of the stock price of a company. The model must be designed in such a way that the trader can make a maximum amount of profit from trading his securities on the platform. It should be easy to use and deployable so that there may be no confusion for the person using it. The prediction model should be able the answer certain question like what has happened, what could happen? and what should we do? [7]. It should give a descriptive and diagnostic analysis of the data collected and using the optimized algorithms give a plausible outcome i.e. prediction. In the end, any prediction model must overcome the existing stereotypes and provide better accuracy than the existing machine learning models as shown in (Fig. 1).
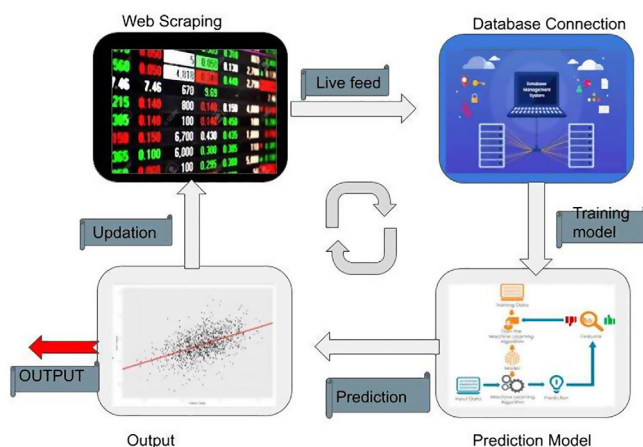

Fig. 1. Working architecture.

3. Tools used in developing the model

3.1. Anaconda navigator: – anaconda navigator is an IDE (integrated development environment).

It is a free and open-source software for both the Python and R programming languages used for scientific computing like machine learning applications, large-scale data processing, etc, which aims to simplify package management and deployment. In our project, the code is written using the jupyter notebook among various other built-in options provided by anaconda navigator.

3.2. MySQL

MySQL is a free open-sourced Relational Database Management System (RDBMS) which uses Structured Query Language (SQL). In our project, we used the MySQL database to store the various datasets of companies' stock prices such as previous close, open price, etc and kept on updating our database using web scrapping (Beautiful Soup). Our program only runs 6 days a week excluding Sundays as the stock market is closed on Saturday and Sunday and only updates values during stock market working hours. Our data on the current stock values keep on updating every 2 min and by the end of the day, the company's dataset is updated with the current day's data in the database as shown in (Fig. 2).

4. Model architecture

The architecture of the predictive model consists of three modules:

- Web Scrapping Module: In this method, the model tries to extract historical data from the NSE website of the stocks which are being targeted and save it to an excel document. The web scraping is done by using the built-in python library called beautiful soup. Initially, around 10 years of raw stock values from the NSE website without changing the attributes is gathered and stored in the excel file. The best part of web scraping data from the NSE website is that we are not required to perform any pre-processing of the data. After the use of the data set in the training and testing of the prediction model, once the model is up and running it tries to hit the yahoo finance website every 2 min to update the dataset with the current stock values and again perform processing [14].
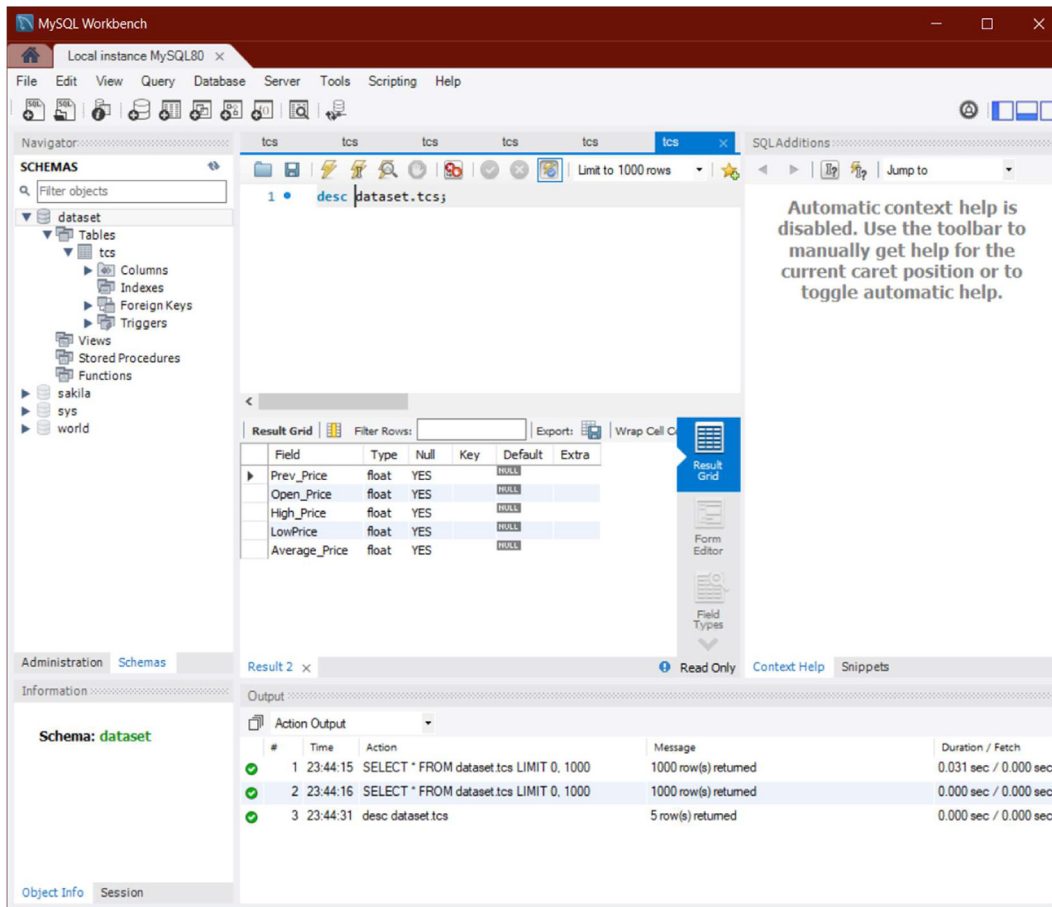
Fig. 2. Database connection.

- Prediction Module: This module involves predicting the stock using the predefined data sets that we have collected from the national stock exchanges website. Using the linear regression algorithm, we try to divide the dataset into a train set and test set. Linear regression is used to develop a relationship between the dependent variable and independent and an independent variable. It follows the equation y = a + bx. The goal of the model is to find the best fitting line with the least sum of squared errors with the actual average value of our stock to the predicted value. The model once deployed works for 5 days a week and will be off during the weekends, as the stock market is closed on Saturdays and Sundays.
- Database Connection Module: The data from the database is taken as input for the prediction model which in turn gives the ultimate output. Initially, we store the complete dataset into the MySQL database with desired attributes and use them in the python code. These values are used for the first phase of the training of the model and then once the h model is trained and the values are predicted, these predicted values are updated into the database for future use [12]. This process is repeated for every 2 min.

5. Methodology and implementation

In this paper we discuss the Stock dataset, in the beginning, later we use various tools to analyse the data to predict the stock prices. Then later on test how efficiently data got processed. Firstly, we start by web scraping the data and then filling up the .xlsx file. Then comes the process of writing the code for

the model. We use the python programming language to write the code. Libraries like pandas and NumPy are used for computational and data reading purposes whereas libraries like matplotlib and cufflinks are used for data visualization purposes [8]. After doing some data visualization we found out that there was no data cleaning required for the extracted dataset. As a result, it could be directly used for processing. Now comes the part of fitting the model. The linear regression model is imported from the sklearn library which is present in python. Using this we try to divide the dataset into training and test set. Afterward, the model is trained and fitted using the train test and the test set. Now to tell why we choose linear regression over other algorithms we try to compare it with multiple regression [9]. This comparison is done to tell the accuracy, precision, and correctness of one model over the other. So, we choose a single parameter to predict the average price of a stock in a linear regression whereas we choose many attributes like opening price, closing price, etc in multiple regression. We split 67% of the dataset into train set and 33% of it into test set. The trained model is then used to predict the future average price of the stock. The predicted value is then updated into the database and then these predicted values will be also considered in future predictions as input. MySQL is used as database input in the code. MySQL is an open source software used as a database service. It supports Sequential Query Language (SQL).SQL queries are used to grab data from the database table's and do analysis on the data and also update the data values in the database. Here data is sent to the database using web scraping tool in python as shown in (Fig. 3).
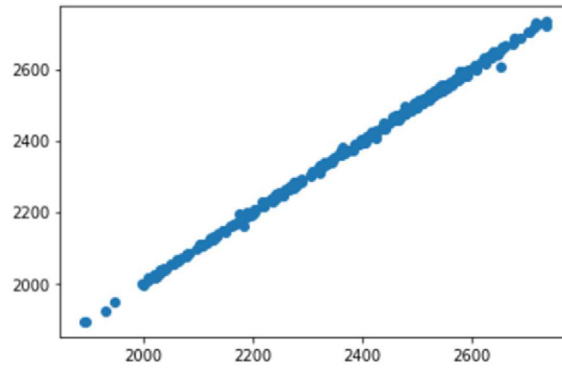
```
In [24]: sns.scatterplot(y_test,pre,)
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x238415cd588>
```
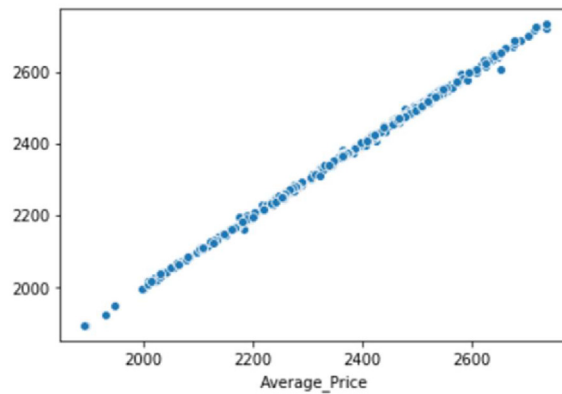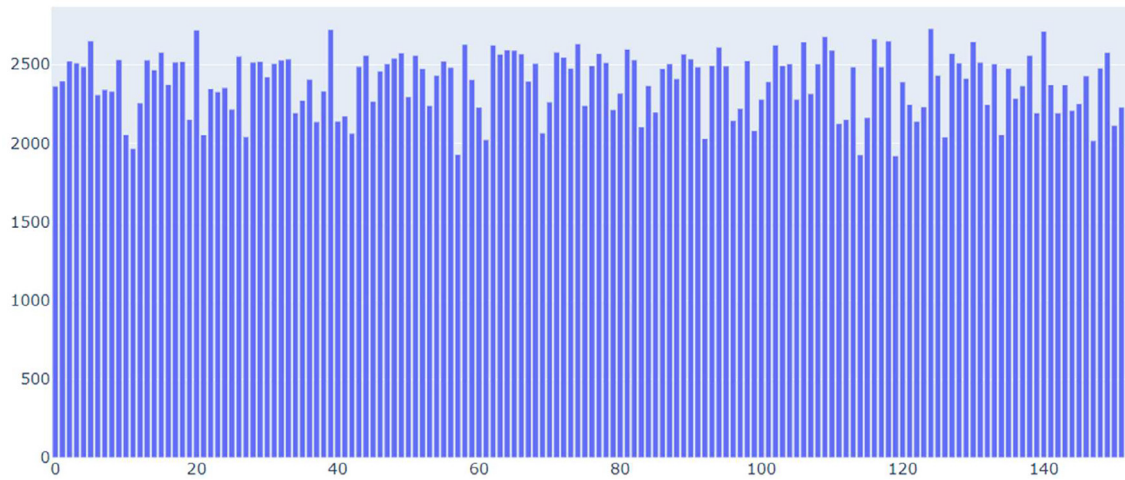

Fig. 3. Scatterplot of the prediction [10].

Fig. 4. Output of Multiple Regression.

5.1. Implementation

Stock price Dataset: The dataset consists of different parameters which later on used in the analysis of data The dataset attributes are:

1. Prev_close: previous data closing price
2. Open_Price: present days opening price
3. High_Price: highest price of the stock that day.
4. Low_Price: Lowest price of the stock that day

5. Average_Price: Average price of the stock that day.

Step 1: collect dataset values for a particular company.
Step 2: convert the dataset values into .csv format and send it into the database.
Step 3: create a table of the data using the sql query.
Query: create table dataset.tcs (Prev_Price float, Open_Price float, High Price float, LowPrice float, Average_Price float);
Step 4: insert data into the database.
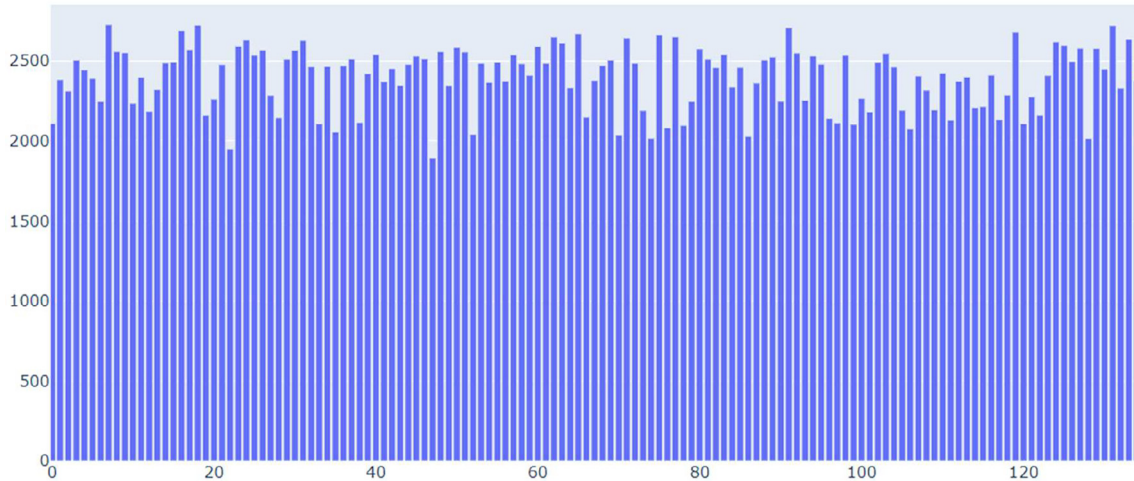Step 5: use the data from the tables and train a regression model.

Fig. 5. Output of Linear Regression.


```
In [31]: print(metrics.mean_squared_error(y_test,pre))
         34.03511330895994

In [32]: print(metrics.mean_absolute_error(y_test,pre))
         4.071035282816069

In [33]: print(metrics.mean_squared_log_error(y_test,pre))
         5.786958624784098e-06
```

Fig. 6. Performance metrics of multiple regression.

```
In [50]: print(metrics.mean_squared_log_error(y_test,pre))
         0.0001454683799080659

In [51]: print(metrics.mean_absolute_error(y_test,pre))
         21.699852590270808

In [52]: print(metrics.mean_squared_log_error(y_test,pre))
         0.0001454683799080659
```

Fig. 7. Performance metrics of linear regression.

Step 6: update the dataset every 2 min using a web scraping tool on yahoo finance known as beautifulSoup.

Step 7: Re-train the regression model with new data values for more accurate predictions.

6. Result and analysis

Multiple Regression: In this algorithm we train our model using open price, high price and low price to predict the nearly accurate average price. The graph represents the predicted average price on the y-Axis and the numbers of days the average price is predicted for in the near future on the X-Axis as shown in (Fig. 4).

Linear Regression: In this algorithm we use previous days price to train our model and then predict the following1 day's average price. The graph represents the predicted average price on the y-Axis and the numbers of days the average price is predicted for in the near future on the X-Axis as shown in (Fig. 5).

6.1. Analysis

Mean square error comparision between linear and multiple regression.

6.1.1. Multiple regression
See Fig. 6.

6.1.2. Linear regression
See Fig. 7.

7. Conclusion

The algorithm is going to be an exemplary resource for brokers and investors for investing within the share market since it's trained on an enormous collection of historical data and has been selected after being tested on a sample data. This project demonstrates models to predict the stock value with more accuracy as compared to previously implemented machine learning models. In future we would like to store the dataset using the cloud architecture using tools like AWS or IBM Cloud and probably implement the prediction model using Artificial Neural Networks [11].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Nirbhey Singh Pahwa, Neeha Khalfay, Vidhi Soni, Deepali Vora, Paper: Stock Prediction using Machine Learning a Review Paper.
[2] Hongduo Cao, Tiantian Lin, Ying Li, Hanyu Zhang, Paper: Stock Price Pattern Prediction Based on Complex Network and Machine Learning.
[3] K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, Sarmistha Padhi, Saurav Sanyal, Paper: Stock Market Prediction Using Machine Learning Algorithms.
[4] Tejas Mankar, Tushar Hotchandani, Manish Madhwani, Akshay Chidrawar, Lifna C.S., Paper: Stock Market Prediction based on Social Sentiments using Machine Learning.
[5] Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan. Paper: Stock Market Prediction Using Machine Learning.
[6] S.P. Pimpalkar, Jenish Karia, Muskaan Khan, Satyam Anand, Tushar Mukherjee, Paper: Stock Market Prediction using Machine Learning.
[7] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. Paper: Survey of Stock Market Prediction Using Machine Learning Approach.
[8] V. Kranthi Sai Reddy, Paper: Stock Market Prediction Using Machine Learning.
[9] Mariam Moukalled, Wassim El-Hajj, Mohamad Jaber, Paper: Automated Stock Price Prediction Using Machine Learning.
[10] Yibig Ng Paper: https://towardsdatascience.com/machinelearning-techniques-applied-to-stock-price-prediction6c1994da8001.
[11] Aishwarya Singh, Paper: Stock Prices Prediction Using Machine Learning and Deep Learning Techniques (with Python codes).
[12] P. Kora, ECG based Myocardial Infarction detection using Hybrid Firefly Algorithm, in: Computer Methods and Programs in Biomedicine, 2017, 152, pp. 141–148. 2-s2.0-85030103123.
[13] N. Krishna Chythanya, L. Rajamani, Neural Network Approach for Reusable Component Handling, in: Proceedings – 7th IEEE International Advanced Computing Conference, IACC 2017, art. no. 7976764 2017, pp. 75–79.
[14] P. Gayathri, S. Umar, G. Sridevi, N. Bashwanth, R. Srikanth, Hybrid cryptography for random-key generation based on ECC algorithm, Int. J. Electr. Comp. Eng. 7 (3) (2017) 1293–1298.