OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Simplified approach for analyzing Stock Market by making use of machine learning

Pavuluri Swarupa[1], Siddu Pujitha[2],Bheemaraju Geetha[3],Indla Sushma[4],Savalam Lakshmi[5]

[1,2,3,4,5]B.Tech,*Deapartment of Computer Science and Engineering, Mekapati Rajamohan Reddy Institute of Technology & Science,Udayagiri,SPSR Nellore,AP,India.*

ABSTRACT: *Stock market is probably the most seasoned technique where an ordinary individual would exchange stocks, make ventures and gain some cash out of organizations that sell a piece of themselves on this stage. This framework ends up being a potential venture plans whenever done admirably. The primary target of this paper is to locate the best model to anticipate the estimation of the financial exchange. During the way toward considering different methods and factors that must be considered, we discovered that strategies like arbitrary backwoods, bolster vector machine were not misused completely. In, this paper we are going to present and survey a progressively attainable technique to anticipate the stock development with higher precision. The main thing we have considered is the dataset of the securities exchange costs from earlier year. The dataset was pre-handled and adjusted for genuine examination. Subsequently, our paper will likewise concentrate on information preprocessing of the crude dataset. Furthermore, after pre-handling the information, we will survey the utilization of irregular woodland, bolster vector machine on the dataset and the results it produces the fruitful expectation of the stock will be an incredible resource for the financial exchange establishments and will give genuine answers for the issues that stock speculators face.*

KEY WARDS: *Machine Learning, Data Pre-processing, Data Mining, Dataset, Stock, Stock Market.*

Date of Submission: 18-05-2020 Date of acceptance: 03-06-2020

I. INTRODUCTION

Stock Market as we probably am aware is a significant exchanging stage which influences everybody at an individual and national level. The essential guideline is very straightforward; Companies will list their offers in the organizations as little wares called Stocks. They do as such so as to fund-raise for the firm. An organization records its stock at a cost called the IPO or first sale of stock. This is the offer cost at which the organization sells the stock and fund-raises. After which these stock are the property of the proprietor and he may offer them at any cost to a purchaser at an Exchange, for example, BSE or Bombay Stock Exchange. Dealers and purchasers keep selling these offers at their own cost yet the organization just gets the opportunity to keep the cash made during the IPO. The keep trusting of rabbits starting with one gathering then onto the next so as to make more benefits, brings about an expansion of cost of the specific offer after each productive exchange. Be that as it may, in the event that the organization gives more stocks at a lower IPO, at that point the market cost for trade goes down and merchants endure a misfortune. This careful wonder is the explanation behind the dread individuals have in putting resources into securities exchanges and the purpose behind the fall and ascent of stock costs basically. Presently in the event that we attempt to diagram the stock trade cost over the timeframe (state a half year), is it extremely difficult to foresee the following result on the chart? A human cerebrum is entirely fit for expanding the chart a couple of directions by simply straightforward taking a gander at it for a couple of moments. Furthermore, in the event that we swarm figure for example make a gathering of irregular individuals attempt to expand the diagram by a fixed measure of time (say seven days), we will get an entirely sensible and surmised answer to a genuine chart.

II. PROBLEM DEFINITION

The issue with evaluating the stock cost will stay an issue if a superior securities exchange forecast calculation isn't proposed. Anticipating how the securities exchange will perform is very troublesome. The development in the financial exchange is normally dictated by the suppositions of thousands of speculators. Securities exchange expectation, requires a capacity to anticipate the impact of ongoing occasions on the speculators. These occasions can be political occasions like an announcement by a political pioneer, a bit of news on trick and so forth. It can likewise be a worldwide occasion like sharp developments in monetary forms and item and so on. Every one of these occasions influences the corporate profit, which thus influences the feeling of financial specialists. It is past the extent of practically all financial specialists to effectively and reliably foresee these hyper parameters. Every one of these elements makes stock value expectation troublesome. When the correct information is gathered, it at that point can be utilized to prepare a machine and to produce a prescient outcome. Financial exchange forecast is fundamentally characterized as attempting to decide the stock worth and offer a powerful thought for the individuals to know and anticipate the market and the stock costs. It is for the most part introduced utilizing the quarterly money related proportion utilizing the dataset. In this way, depending on a solitary dataset may not be adequate for the expectation and can give an outcome which is wrong. Henceforth, we are considering towards the investigation of AI with different datasets combination to foresee the market and the stock patterns.

III. LITERATURE SURVEY

During a literature survey, we focused on some of the information about Stock market prediction mechanisms currently being used.

3.1 Survey of Stock Market Prediction Using Machine Learning Approach

The stock market expectation has become an inexorably significant issue in right now. One of the techniques utilized is specialized examination, yet such strategies don't generally yield precise outcomes. So it is essential to create techniques for an increasingly precise forecast. By and large, speculations are made utilizing expectations that are gotten from the stock cost in the wake of considering all the elements that may influence it.

3.2 Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests

The utilization of AI and man-made consciousness methods to foresee the costs of the stock is an expanding pattern. An ever increasing number of analysts put their time each day in thinking of approaches to show up at procedures that can additionally improve the exactness of the stock expectation model. Because of the tremendous number of alternatives accessible, there can be n number of ways on the most proficient method to anticipate the cost of the stock, however all strategies don’t work a similar way

3.3 Stock Market Prediction through Multi-Source Multiple Instance Learning

Precisely anticipating the stock market is a difficult undertaking, yet the advanced web has end up being a helpful instrument in making this errand simpler. Because of the interconnected configuration of information, it is anything but difficult to separate certain suppositions in this manner making it simpler to set up connections between different variable and generally scope out an example of venture.

3.4 Stock Market Prediction: Using Historical Data Analysis

The stock market forecast process is loaded up with vulnerability and can be impacted by different components. In this manner, the stock market assumes a significant job in business and fund. The specialized and principal investigation is finished by wistful examination process. Online networking information has a high effect because of its expanded use, and it very well may be useful in anticipating the pattern of the stock market. Specialized examination is finished utilizing by applying AI calculations on authentic information of stock costs

3.5 A Survey on Stock Market Prediction Using SVM

The ongoing examinations give a very much grounded confirmation that the vast majority of the prescient relapse models are wasteful in out of test consistency test. The purpose behind this wastefulness was parameter shakiness and model vulnerability. The examinations likewise finished up the conventional systems that guarantee to take care of this issue. It is utilized to learn polynomial spiral premise work and the multi-layer perception classifier. It is a preparation calculation for arrangement and relapse, which chips away at a bigger dataset. There are numerous calculations in the market however SVM furnishes with better productivity and exactness. The connection examination among SVM and stock market shows solid interconnection between the stock costs and the market file.

IV. LIMITATIONS OF EXISTING SYSTEM

The existing system fails when there are rare outcomes or predictors, as the algorithm is based on bootstrap sampling. The previous results indicate that the stock price is unpredictable when the traditional classifier is used. The existence system reported highly predictive values, by selecting an appropriate time period for their experiment to obtain highly predictive scores. The existing system does not perform well when there is a change in the operating environment. It doesn't focus on external events in the environment, like news events or social media. It exploits only one data source, thus highly biased

V. PROPOSED SYSTEM

In our model we can predict the future sales based on past predictions of year by year and month by month. Now if we try to graph the stock exchange price over the time period (say 6 months), is it really hard to predict the next outcome on the graph but by using our proposed system we can easily plot a graph and see the analysis. We are using machine learning algorithm to predict the future stock price for exchange of stock and find the best fits. We can see the last few year's data and day to day changes so we can easily estimating the future sales. Kaggle is an online community for data analysis and predictive modeling. It also contains dataset of different fields, which is contributed by data miners. Various data scientist competes to create the best models for predicting and depicting the information. The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for.

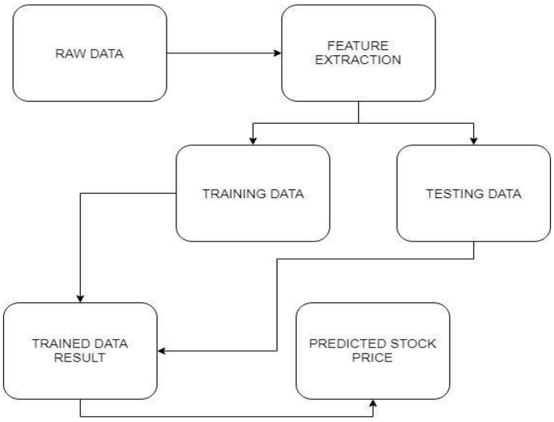


Fig 5.1.System Architecture

The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset.

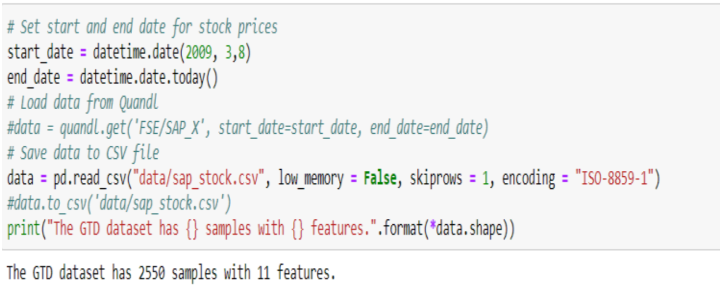
VI. MODULES IDENTIFICATION

6.1.Data Collection: Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we will be analyzing the Kaggle dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

6.2.Pre Processing: Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

6.3.Training the Machine: Training the machine is similar to feeding the data to the algorithm to touch up the test data. The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyper parameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyper parameter values. Finally, we will calculate a cross-validated score, for individual sets of hyper parameters. Then, we select the best hyper parameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

VII. EXPERIMENTAL RESULTS

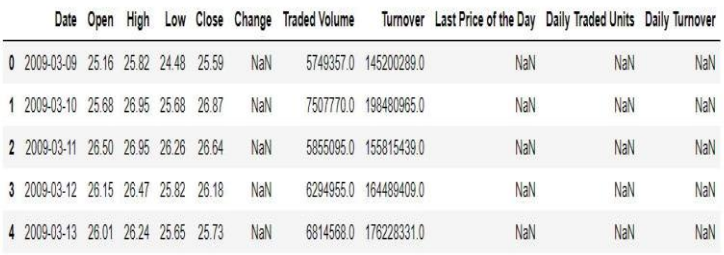


```
# Set start and end date for stock prices
start_date = datetime.date(2009, 3,8)
end_date = datetime.date.today()
# Load data from Quandl
#data = quandl.get('FSE/SAP_X', start_date=start_date, end_date=end_date)
# Save data to CSV file
data = pd.read_csv("data/sap_stock.csv", low_memory = False, skiprows = 1, encoding = "ISO-8859-1")
#data.to_csv('data/sap_stock.csv')
print("The GTD dataset has {} samples with {} features.".format(*data.shape))

The GTD dataset has 2550 samples with 11 features.
```

Fig 7.1:Store the data set and see the number of samples with features.

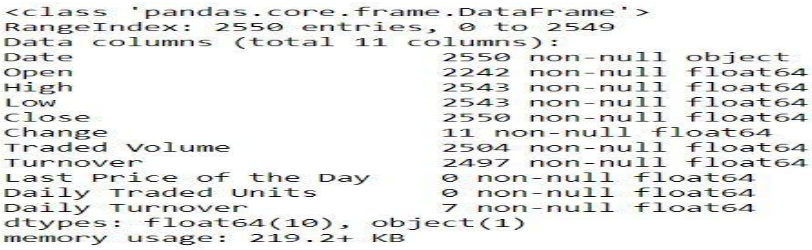
Data loading the attributes of the data include: open,high, low,close,volume,split ratio adj.open high,adj close.



	Date	Open	High	Low	Close	Change	Traded Volume	Turnover	Last Price of the Day	Daily Traded Units	Daily Turnover
0	2009-03-09	25.16	25.82	24.48	25.59	NaN	5749357.0	145200289.0	NaN	NaN	NaN
1	2009-03-10	25.68	26.95	25.68	26.87	NaN	7507770.0	198480965.0	NaN	NaN	NaN
2	2009-03-11	26.50	26.95	26.26	26.64	NaN	5855095.0	155815439.0	NaN	NaN	NaN
3	2009-03-12	26.15	26.47	25.82	26.18	NaN	6294955.0	164489409.0	NaN	NaN	NaN
4	2009-03-13	26.01	26.24	25.65	25.73	NaN	6814568.0	176228331.0	NaN	NaN	NaN

Fig 7.2: Display the dataset 1st 5 rows.

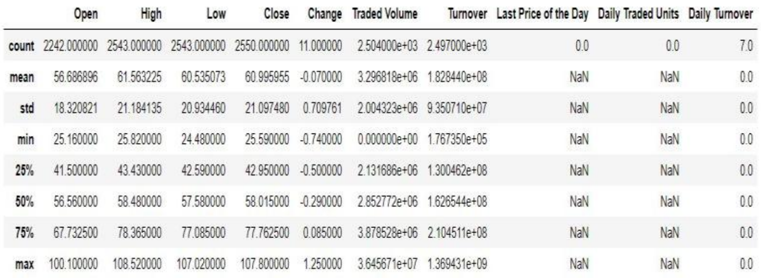
We select the attribute close to be our label and with the help of head keyword we can see the first five rows data and use open, high,adj high,adj open ,adj low,adj volume to extract the features that will help us predict the outcome better



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2550 entries, 0 to 2549
Data columns (total 11 columns):
Date                          2550 non-null object
Open                          2242 non-null float64
High                          2543 non-null float64
Low                           2543 non-null float64
Close                         2550 non-null float64
Change                        11 non-null float64
Traded Volume                 2504 non-null float64
Turnover                      2497 non-null float64
Last Price of the Day         0 non-null float64
Daily Traded Units            0 non-null float64
Daily Turnover                7 non-null float64
dtypes: float64(10), object(1)
memory usage: 219.2+ KB
```

Fig 7.3. Information features in a dataset

Check the data type in each columns and also gives the range index and memory range and total number of columns



	Open	High	Low	Close	Change	Traded Volume	Turnover	Last Price of the Day	Daily Traded Units	Daily Turnover
count	2242.000000	2543.000000	2543.000000	2550.000000	11.000000	2.504000e+03	2.497000e+03	0.0	0.0	7.0
mean	56.686896	61.563225	60.535073	60.995955	-0.070000	3.296818e+06	1.828440e+08	NaN	NaN	0.0
std	18.320821	21.184135	20.934460	21.097480	0.709761	2.004323e+06	9.350710e+07	NaN	NaN	0.0
min	25.160000	25.820000	24.480000	25.590000	-0.740000	0.000000e+00	1.767350e+05	NaN	NaN	0.0
25%	41.500000	43.430000	42.590000	42.950000	-0.500000	2.131686e+06	1.300462e+08	NaN	NaN	0.0
50%	56.560000	58.480000	57.580000	58.015000	-0.290000	2.852772e+06	1.626544e+08	NaN	NaN	0.0
75%	67.732500	78.365000	77.085000	77.762500	0.085000	3.878528e+06	2.104511e+08	NaN	NaN	0.0
max	100.100000	108.520000	107.020000	107.800000	1.250000	3.645671e+07	1.369431e+09	NaN	NaN	0.0

Fig7.4.Descriptive statistics summary of data set:

It give the descriptive statistics that means mathematical calculations like mean,count,min value ,max value 25% value 50% and 75% value ,standard deviation are obtained with the help of description keyword.

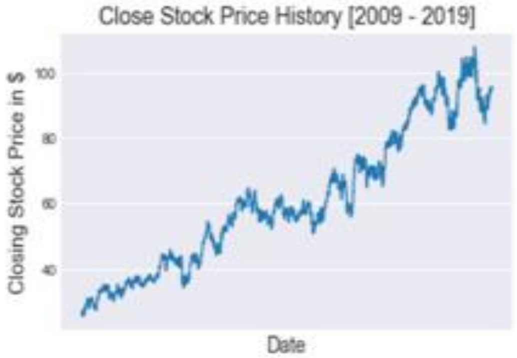


Fig 7.5.Create subplots to plot graph and control axes:

Here we can take a look at the price movement over time by simply plotting the closing price vs time we can already see that the price continuously increase our time and we can also estimate that trend could be linear.

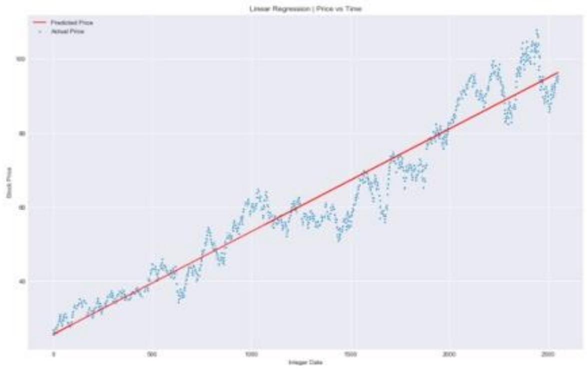


Fig 7.6.Linear Regression: Using linear regression plot the data:

In linear regression model using the train test split method. In large data set of training data contributes to a stronger and more accurate classifier with ultimately increases the overall accuracy. Here the red line indicates the prediction line and the blue curves indicate the previous data

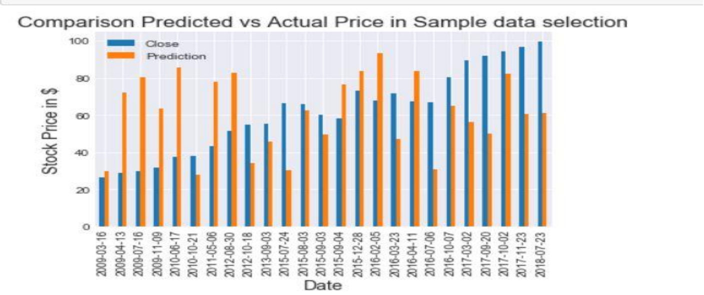


Fig7.7.Comparison Predicted vs Actual Price in Sample data selection:

Here we can take the predicted values compare with the actual values on random sample from our dataset. We can see the large variations between predicted and actual values in the random sample.

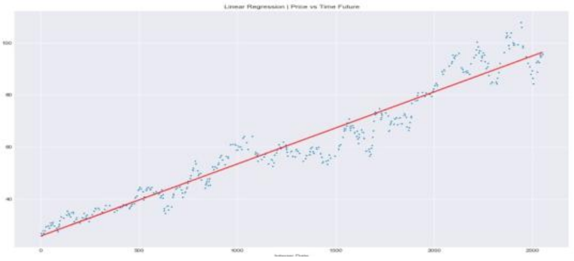


Fig 7.8.Price vs. Time Future:

The graph shows that comparison between price and time the prediction line shows when the time is increases the price also increases here we can see how the time is predicted with price.

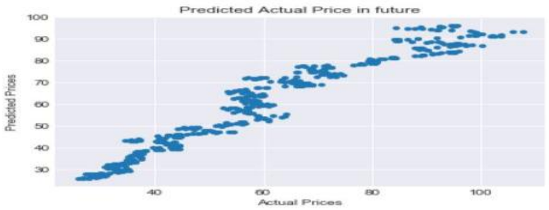


Fig 7.9.Predicted Actual Price in future:

We can see the future predicted price here the data points are mostly close to a diagonal which indicates that the predicted values are close to the actual values and the model performance is largely quite good.

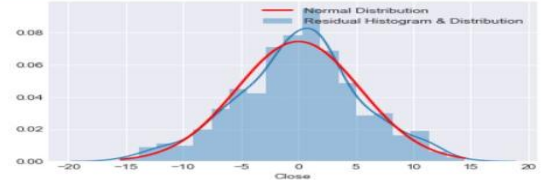
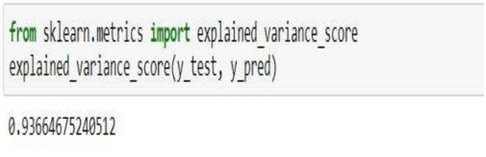


Fig.7.10.Residual Histogram & Distribution:

In the above histogram there are two distributions the red line or red curve indicates the normal distribution and blue curve indicates the residual histogram and distribution so the residual histogram is related to the normal distribution curve.



```
from sklearn.metrics import explained_variance_score
explained_variance_score(y_test, y_pred)
```

```
0.93664675240512
```

Fig.7.11.The value of R2 shows that are model accounts for nearly 94% of the differences between the actual stock prices and the predicted prices:

So finally the r2 values shows that 94% of the difference between the actual stock price and the predicted stock price and it give the good prediction now we can easily analyze weather we want to invest in the stock market or not with the above prediction.

VIII. CONCLUSION

By estimating the precision of the various calculations, we found that the most reasonable calculation for foreseeing the market cost of a stock dependent on different information focuses from the chronicled information is the irregular woods calculation. The calculation will be an incredible resource for representatives and speculators for investing cash in the stock market since it is prepared on a colossal assortment of verifiable information and has been picked in the wake of being tried on an example data. The venture exhibits the AI model to foresee the stock an incentive with more precision when contrasted with recently executed AI models.

REFERENCES

[1]. Kunal Pahwa, Neha Agarwal,” Stock Market Analysis using Supervised Machine Learning”,IEEE-2019

[2]. K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, SarmisthaPadhi, Saurav Sanyal. ” Stock Market Prediction Using Machine Learning Algorithms”, IJEAT-April-2019.

[3]. Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.

[4]. Loke.K.S. “Impact Of Financial Ratios And Technical Analysis On Stock Price Prediction Using Random Forests”, IEEE, 2017.

[5]. Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market Prediction via Multi-Source Multiple Instance Learning.” IEEE 2018.

[6]. VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.

[7]. SachinSampatPatil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET 2016.

[8]. https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde. pdf

Pavuluri Swarupa, et. al." A Simplified approach for analyzing Stock Market by making use of machine learning." *International Journal of Modern Engineering Research (IJMER),* vol. 10(03), 2020, pp 44-50.