Stock Market Prediction Using Machine Learning

*Note: Prediction of price of future in different stock

Gehna Sachdeva
University Institute of Computing
Chandigarh University
Mohali, India

Babalu Kushwaha
University Institute of Computing
Chandigarh University
Mohali, India

Abstract- The share market is a complex and risky way to conduct business. It can be difficult to predict the future price due to its volatility and complexity. But the machine is very efficient to calculation so that we are going to calculate the stock market price by calculating past records and its employee and its shared holder trust, market valuation and how much company are old and many more, via this data we are made an algorithm and then predict the stock feature price it is called "Stock market prediction model." This paper proposes an algorithm that learns to predict the future value of stocks using machine learning based on various factors. Index Terms—Stock Market, Data Analysis, Social Media Mining, Machine Learning

Keywords – —Stock Market, Data Analysis, style, Social Media Mining, Machine Learning

I. INTRODUCTION

Stock market is a platform where a individual or firm invest on company's stocks or shares to earn profit. If this investment is done wisely, it can give huge profits to the investor. But the ups and downs in stock prices are unpredictable. Here, we have machine learning that can help us to predict the future ups and down in stock prices. The basic idea behind the stocks market is that businesses list their share as tiny commodities know as stock. This is done to raise money for company or to have financial aid form outsidefirms. The term IPO refers to the price at which a corporation offers to sell its stock. After buying stock at IPO, customer can sell them at any cost to buyers. After every profitable transaction, the prices for a particular share increase simultaneously. The market price drops and traders lose money if more stocks are released at a lower initial public offering. That's why investing in stock is a big fear for investors.[1] The prediction of stock prices can be enhanced by the application of gadget learning. Device learning techniques can reveal patterns and insights that were previously hidden from view, and these can be utilized to produce predictions that are incredibly precise. In the cutting-edge, global device research space, advancements are occurring at a rate never seen before.[2] One choice made in the stock market can have a big effect on someone's life. One choice made in the stock market can have a big effect on someone's life. This study present an algorithm that uses machine learning to learn how to forecast the future value of stock based on multiple inputs. The model's purpose is to forcast the KSE-100 index's performance. It uses various factors such as commodity prices, foreign exchange, interest rate, and public sentiment to predict the market. Different variants of Artificial Neural networks (ANN) are used for prediction. Some of these include Deep Belief Network, single layer Perception, Radial Basis Function, and Support Vector Machine. The results indicated that Multi-layer perception performed well and predicted the market with an accuracy of 77In COVID, we are not predicting how stock price works but as you know price going to drop and suddenly going to the highest ever stock price it is unbelievable, but when we are given more data algorithms, we give you a more accurate result.

II. PREDICTION MODEL

A. *Data Analysis*

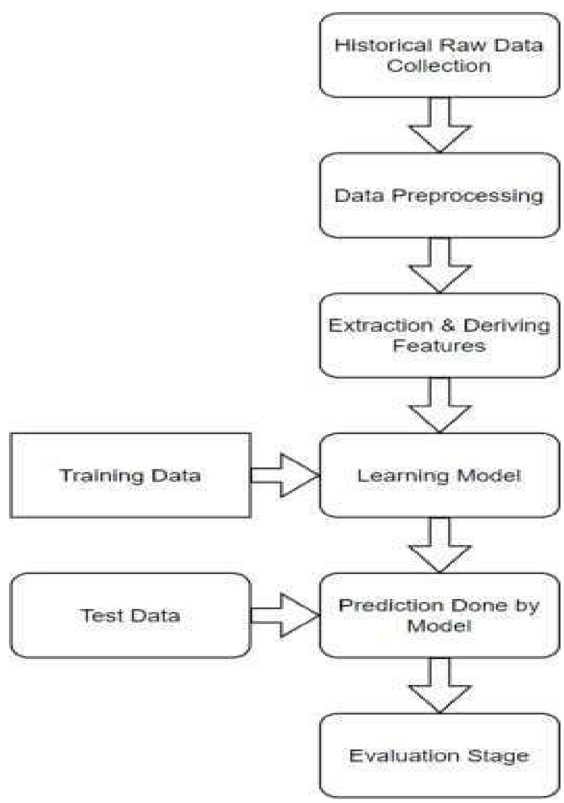
On this stage, we will examine the raw records to be had to use and have a look at it in -order to become aware of appropriate attributes for the prediction of our decided on label. The data -set's characteristics consist of

1. Open (stock's opening price
2. High (Maximum amount attainable at a given moment)
3. Low (Lowest feasible price at a given moment)
4. Shut (Stock closing price)
5. Volume (Total number of trades in a given day)
6. Dividend percentage

Adjusted values of Attributes:

1. Adj. Open: This is opening price of stocks
2. Adj. High: This is highest stock price.
3. Adj. Low: This is lowest stock price.
4. Adj. Close: This is closing price and decides the opening price of next day.
5. Adj. Volume: Volume traded directly affects the opening price of next day, so it is most important feature.

We employ "Adj. Open, Adj. High, Adj. Close, Adj. Low, and Adj. Volume" to ectract the features that would improve our ability to predict the result. We choose the attribute "Close" to be our label (the variable which we shall be predictiong.) Adjusted valuesat place of raw values are used as those are processed to getrid of common data errors. HL PCT: It is percentage changeused to reduce no. of features, but retain the data gathered. PCT Change: Same treatment is done adj. open and adj. closeprices to reduce features and used derived one.



B. Training Model

After converting data frames to NumPy arrays we use visualization libraries of python to make plots of the data processed.[1] We can use various regression models for prediction of stock prices:

a. First, Basic Linear Regression
b. The Polynomial Regression
c. Regression using Support Vectors
d. Regression using Decision Trees
e. Forest Regression at Random

C. Equations

$$HLPCT = ((Adj.high - Adj.low) * 100)/Adj.close)$$

$$PCTChange = ((Adj.Close - Adj.Open) * 100)/Adj.Open$$

D. Some Common Mistakes

The data is scaled such that for any value

X

The data should split into test data and train data respective to its type i.e. label and feature.

III. STOCK MARKET INDEX

The stock market index is a widely used statistical measure that shows changes in the prices of stocks. It is computed using the values of the underlying stocks. A stock market index is like a combination of many companies' stocks that have at the top of the list like in India we are use NIFTY50 that is a combination of top 50 Companies in India that valuation is very high and also there are some funds like Sen-sex it is a top 100 Company stock that takes and makes them it is all are called index funds below are some index funds that you can search about it. – Direct Placment of the UTI Nifty Index Fund

- Prudential Nifty Index Fund by ICIC
- Fund for SBI Nifty Index -Direct Plan
- Direct Plan for HDFC Index Fund
- LIC MF Nifty Plan Index Fund
- HDFC Index Fund - Direct Plan
- LIC MF Index Fund - Sen-sex Plan
- and many more

IV. REMARKABLE SHRE

They refer to the shares of a company that are currently held by all of its investors, including restricted shares held by ins. It is called which share that is held by all stockholders, institutional investors, and also held by its own company owner and other people are related to the company all are called outstanding shares.

V. MARKET CAPITALIZATION

It depends on another competitor that is held all by like outstanding share price are have more than another company share then it is called these company market cap is very high. also, it depends on how much have 1 stock price has and how much is held by company board members and owners.

VI. METHODOLOGY

The Learning method employed in this study paintings is arbitray woodland. After Obtaining and smoothing the time-series data, the technical indicators are retrieved. Technical signs are parameters that offer insights into the expected inventory fee conducted in destiny. those technical signs are then employed in the random forest's training In this segment, the info of each step is discussed.

A. Pre-Processing

The historical inventory data collected over time is exponentially smoothed as part of the pre -processing of the records. In this, the most recent observations are given more weight, and the as the observations get older, the weighted keeps getting lower. The exponential smoothing for time series X can be completed iteratively as follows:

$$[3] S0 = X0$$

$$[4] fort?`0, \; St = sf * Xt + (1 - sf) * Xt - 1$$

Where sf is the smoothing factor, whose fee varies from zero to This preprocessing makes the historical statistics suitable for determining the fee fashion in the stock values by eliminating noise outliers, and missing facts,.The technical indicators are derived from the smoothed time-series data, which is predicated on the forecast of the target charge cost, or TPi of the ith day as:

$$[5] TPi = Sign(CPi + d - CPi)$$

Where d is the large range of days that the forecast is to be carried out. The fee shift is determined by the TPI signal, which, however lovey, indicated a fantastic shift in the inventory charges after d days, and vice versa

B. Features

In order to predict the movement of the inventory rate, feature extraction is carried out based on the technical indicators that are computed from smoothed time series data. Analysts use those in particular to determine the directions of stock rate movements. The indicators employed in the CNX case The starting rate, high price, low fee stocks exchange, and turn over (in rupees) are all considered Nifty information. The remaining price is the structured/expected variable. Similar to SP BSE Sensex figures, opening, exorbitant, and sporadic expenses are considered as indications.

Algorithm: LS-RF

Input: Dataset D (set of predictor variables and regression/dependent variable)

for $t = 1$ to k

d ← D

T$_t$ ⭘ CART (d) where CART (d) = agmin ($\sum^{|d|}_{i=1}(y_i - r)^2$)

end for

return T

$$[6] mi = jjxij$$

Whereas j are the values determined by the evidence. Let ri represent the based variable. The Tree ensemble Model is a rigid version of CART trees, where the totalof several timbers; forecasts is considered as:

$$[7] yi = ki = 1mixij, j, mkM$$

Where M is the designated are and ok is the total number of bushed in the random woods. The goal function, which is derived from the regularization time period() and training loss L(), determines the overall performance of the model

C. Proposed Approach

Based on least rectangular improvement Random Forest (LS-RF) Carts are extensively utilized in numerous assessments and predictions applications. But the trees that are designed to learn incredibly strange styles tend to outlive the educational units. The tree may also develop in an entirely unusual fashion due to noisy data or an outlier because the decision timber is relatively basic and predictive methods occasionally exhibit bias and have large variance. for this reason, this hassle is conquered by means of Random Forest through schooling numerous Carts on a couple of characteristic areasat the cost of slightly elevated bias. This suggest that not every desirable tree in the chosen forest gets all of the educational information. Partitions are created recursively from the education statistics. The division has implemented the utilization of mean square errors as indicators or impurity. The predictions made by each selection tree are pooled as soon as they are shaped to provide a final predictions. A technique called LSboost is used to combine the outputs of several CART novices in order to achieve more faborable overall performance. Moreover, it is employed to reduce the decision tree's variation and over becoming. The tree ensemble version used in our suggested method is a linear model, which can be expressed as follows: the collections of predictor variable xj is used to calculate the regression/structured variable mi

$$[8] Obj() = ki = 1L()i + ki = 1()i$$

The version's prediction accuracy is measured by the educations Loss L(), which uses the logistic loss for logistic regression.

$$[9] L() = I[yiln(1 + -yi)] + [(1 - yi)ln(1 + eyi)]$$

A. Experimental outcomes

This study proposes a Random Woodland regression model for inventory market rate predictions that is mostly based on LSboost data. This section provides details about the dataset that was used, the experimental setup, and an analysis of the results obtained by applying the suggested approach (LS-RF) to the specified dataset additionally,

The suggested method's performance is contrasted with that of the standard Vector Regression using the same data set and experimental configuration.

Table-1 Error Measures

Prediction-Model	Mape	Name	RMSE	
LS-RF	0.2573	0.0025	0.0025	420.1312
SVR	2.0085	0.0201	0.0201	25581.515
2 days ahead of time	0.3978	0.0039	0.0042	1089.2167
LS-RF				
SVR	1.18	0.0118	0.0144	13165.9144
3 days ahead of time	0.4096	0.0041	0.0043	1105.0762
LS-RF				
SVR	1.1891	0.0119	0.0137	11720.7024
4 days ahead of time				
LS-RF	0.4738	0.004 7	0.0049	1494.080 3
SVR	1.1357	0.011 4	0.0128	10216.89 4
5 days ahead of time				4
LS-RF	0.5709	0.005 7	0.0062	2250.428 4

Table-2 Error Measures

Prediction-Model	RMSE	MAE			
LS-RF		0.0002	0.0025	0.0025	420.1312
SVR		0.0454	0.0201	0.0201	25581.515
2 days ahead of time LS-RF		0.0034	0.0039	0.0042	1089.2167
SVR		0.0388	0.0118	0.0144	13165.914 4
3 days ahead of time LS-RF		0.0027	0.0041	0.0043	1105.0762
SVR		0.0387	0.0119	0.0137	11720.702 4
4 days ahead of time					
LS-RF	0.1980	0.0020	4932.4472	0.0049	1494.0803
SVR	3.7748	0.0377	963306.050	0.0128	10216.8944
5 days ahead of timeLS-RF	0.2173	0.0022	5016.9331	0.006	2250.4284

SVR	0.9969 9	0.009	0.0116	8396.699 5
6 days ahead of time				
LS-RF	0.639	0.0064	0.0069	2799.232 2
SVR	1.0341 3	0.010	0.0117	8432.273 4
7 days ahead of time				
LS-RF	0.5629 6	0.005	0.0064	2408.576 7
SVR	1.1199 2	0.011	0.0125	9412.041
8 days ahead of time				
LS-RF	0.5557 6	0.005	0.0062	2287.578 8
SVR	1.0663 7	0.010	0.0119	8572.082 9
9 days ahead of time				
LS-RF	0.7181 1	0.007	0.0089	4618.789 8
SVR	1.1203 2	0.011	0.0124	9149.674 7

				2	
SVR	3.6304	0.0363	885782.60 2	0.011	8396.6995
6 days ahead of timeLS-RF	0.2654	0.0027	3	6	2799.2322
			6830.4108	0.006 9	
SVR	3.6537	0.0365	885410.31 7	0.011	8432.2734
7 days ahead of timeLS-RF	0.2497	0.0025	8	7	2408.5767
			6067.8055	0.006 4	
SVR	3.7087	0.0371	902542.88 1	0.012	9412.041
8 days ahead of timeLS-RF	0.3105	0.0031	4	5	2287.5788
			9433.4056	0.006 2	
SVR	3.6813	0.0368	882470.28 5	0.011	8572.0829
9 days ahead of timeLS-RF	0.3962	0.0040	6	9	4618.7898
			16420.262 7	0.008 9	
SVR	3.7595	0.0376	916365.38 1	0.012	9149.6747

VII. DATASE TS.

This examines the use of then years's worth of historical data, form January 2006 to December 2015, for two inventory market indices-the SP BSE Sensex and the CNX Nifty-which could be incredibly lage. all the facts are acquired from NSE and BSE websites.

A. Experimental Setup

Every experiment is conducted on an Intel Core i7 PC running Windows 10 with eight gigabytes of RAM and a clock speed of three GHz. We use Matlab 2016 for our investigations. There are 100 trees in the ensemble in LS-RF

B. Assessment Measures

metrics, specifically, mean Absolute percent errors (MAPE), suggest Absolute errors (MAE), relative Root suggests Squared mistakes (rRMSE), and recommend Squared error (MSE) are employed to evaluate to evaluate the regression models' performance, These assessment measure' mathematical notations are demonstrated in E

$$[10](|A-P|/|A|)a100$$

$$[13]MSE = 1/nn((A - P)/A)2)$$

in where AI and PI represent the ith day's actual and predicated values, respectively. The total number of days for which a prediction is made is denoted by n.

C. *Consequences and Discussion*

The results obtained for the suggested LS-RF method and SVR approach, across four overall performance metrics, are displayed in Desk I and Table II for Nifty CNX FACTS AND SP BSE Sensex records, respectively, an i.e. rRmse, MSE, MAPE, and MAE. The performace for forecast made 1- 10, 15-30, and 40 days ahead of time is shown in both tables.

The consequences received for the proposed LS-RF ap- proach and SVR technique are shown in table I and table II for Nifty CNX information and SP BSE Sensex information re- spectively, over the course of four total performance metrics: MSE, Rrmse, MAE, AND MAPE. Performance for forecastes made 1-10, 15-30, and 40 days ahed of time is displayed in both tables.

VII. CONCLUSION

This paper supplies a clean perception of how to implement gadget getting to know. there are numerous methods, tech- Niqueand strategiesavailable to handle and solve numerous issues, in unique conditions possible. This paper is restricted to simplest supervised gadget gaining knowledge of, and tries to give an explanation for simplest the fundamentals of this complicated process. Mastering and information the various terminologies and strategies gift in the stock market was very helpful in preprocessing the dataset that allows you to achieve high- quality possible outcomes. The Logistic Regression version gave maximum suggest accuracy of 68.622.

REFERENCES

[1] S.. Ravikumar and P. Saraf,” Prediction of Stock Prices using Machine Learning (Regression, Classification) Al- gorithms,” 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5, doi: 10.1109/IN- CET49848.2020.915406
[2] Fama and F. Eugene, ”Random walks in stock market prices,” Financial analysts journal 51,vol. 1, pp. 75-80, 1995.
[3] Miao, Kai, C. Fang and . Z. G. Zhao., ”Stock Price Forecast Based on Bacterial Colony RBF NeuralNetwork [J],” 2007.
[4] Bollerslev and Tim, ”Generalized autoregressive conditional heteroskedasticity,” Journal of econometrics 31, vol. 3, pp. 307-327, 1986.
[5] Hsieh and A. David, ”Chaos and nonlinear dynamics: appli- cation to financial markets.,” The journal of finance 46, vol. 5, pp. 1839-1877, 1991.
[6] Rao, T. Subba and M. M. Gabr, An introduction to bispectral analysis and bilinear time series models, vol. 24, Springer Science Business Media, 2012
[7] Hadavandi, Esmaeil, S. Hassan and G. Arash , ”Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting,” Knowledge-Based Systems 23, vol. 8, pp. 800-808, 2010.
[8] Lee, Yi-Shian and Lee-Ing Tong, ”Forecasting time series using a methodology based on autoregressive integrated mov- ing average and genetic programming,” Knowledge-Based Systems 24, vol. 1, pp. 66-72, 2011.
[9] M. H. Zarandi, H. Esmaeil and I. B. Turks, ”A hybrid fuzzy intelligent agent ˇGbased system for stock price prediction,” International Journal of Intelligent Systems 27, vol. 11, pp. 947-969, 2012.
[10] Welling and Max, ”Support vector regression,” 2004. J. Patel, S. Shah, P. Thakkar and K. Kotecha, ”Predicting stock market index using fusion of machine learning techniques,” Expert Systems with Applications 42, vol. 4, pp. 2162-2172, 2015
[11] Breiman and Leo, ”Random forests,” in Machine learning 45, 2001.
[12] [12]2017 2nd International Conference for Convergence in Technology(I2CT) 978-1-5090- 4307-1/17