WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Stock Price Trend Forecasting Using Machine Learning

Rahul Kumar Deo

M. Tech
Department of Computer Science and Engineering
Jharkhand University of Technology, Ranchi.

ABSTRACT

This paper seeks to analyze a broad range of prediction procedures to predict future stock returns based on past performance and mathematical news indicators. To construct an ensemble of different stocks to magnify the risk, Machine Learning strategies are employed to interpret the seemingly chaotic market data. By selecting a set of parameters with a significant impact on the offer value of an association, a connection between the selected factors and the offer value is established, which can be used to predict the correct outcomes. Although the offer market is difficult to predict due to its immaterial nature, this paper focuses on applying Machine Learning techniques and stock marker ideas to forecast stock expenses. Additionally, the paper examined the effect of the new coronavirus outbreak on the Nifty 50 index. The final product is a web-based interactive platform to enable stock market predictions.

Keywords: Machine learning, Stock returns, Foreseen, Statistical analysis, Stock indicator.

I. INTRODUCTION

The Stock Exchange Forecast and Examination is a demonstration of the effort to identify the longer-term value of partnership stocks. The stock exchange is a fundamental element of the nation's economy and plays a major role in the development of the business that, over time, affects the nation's economy. Financial experts and industry professionals are concerned about the stock exchange and must be aware of whether a stock can increase or decrease in value over a predetermined period of time. If the interest in an organization's stock is higher, the corporate offer value will increase, and if the interest in the organization's stock is low, the corporate offer worth would decrease. We have collated the actual securities exchange data of 500 organizations in the SP 500 market from the company's history. We have used data from American airlines and normalized it using six different regression models for fore-casting and comparison. The coronavirus pandemic has had a significant impact on the financial exchanges of over 200 nations in a short period of time. This has had a considerable impact on the stock market in India and the reactions to the pandemic. This analysis has also examined the movements of the NIFTY -50 and the underlying factors that have contributed to them. Furthermore, the financial exchanges around the world have seen a surge in volatility due to the speculative behavior of speculators.

II. DESIGN DETAILS AND IMPLEMENTATION

2.1 System Overview

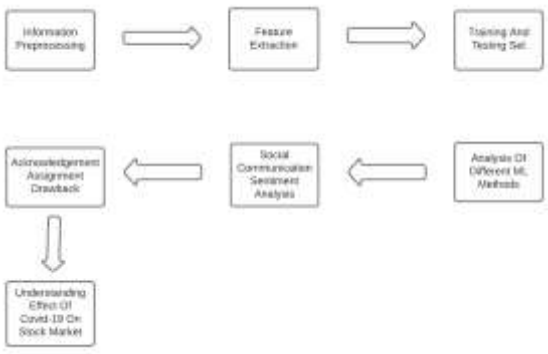


Fig 1. System Overview

2.1.1 Information Preprocessing

- Information discretization: - A small amount of data that has a particular meaning, especially when it comes to math.
- Information transformation: - Basically, it's about standardizing a set of data.
- Data cleaning: - Filling Missing Values Using Machine Learning Techniques.
- Data Integration: - In short, it is a collection of data files.

2.1.2 Feature Selection

In essence, we developed new features from the baseline features to gain a better comprehension of the dataset, such as 50-day operating mean, previous day contrast and so on. To reduce the number of features that are rarely used, we select features based on the k highest values in Feature Selection, using a direct model to test the effect of a single regressor, and then for some regressors. Generally, we use a select K-based algorithm, using f regression as the score for analysis. Additionally, we attempted to incorporate Twitter's Daily Sentiment Score into the analysis, as this element for each organization is dependent on the client's tweets about that organization and the tweets on its website.

2.1.3 Training and Testing set

Once the informational index has been restored to a clean informational index, the information index is divided into preparing and testing sets for evaluation. Here, train test outcomes are considered to be later qualities, and testing set is around 10.

2.1.4 Analysis of different Machine learning methods

- Grouping Methods: This section covers administered order techniques such as Support Vector Machines (SVM), Neural Networks (NNNs), Naive Tom Bayes (Naive Tom Bayes), Adaboost classifiers (Random Forest Classifiers), and other related techniques.
- Regression Strategies: These models will be used to get the average mathematical price of interesting stocks. This part will use regulated relapses methods like Linear Regression, Support Vector Regression, Kernel methodologies, etc.

2.1.5 Social Communication Sentiment Analysis

Analyzing the current market conditions from the most up-to-date news and social media channels such as Twitter to gain insights into the longer term of stock expense.

2.1.6 Acknowledgement Assignment Drawback

In this step, material weightage is passed to optional ways used for information collection.

2.1.7 Understanding Effect of Covid-19 on Stock Market

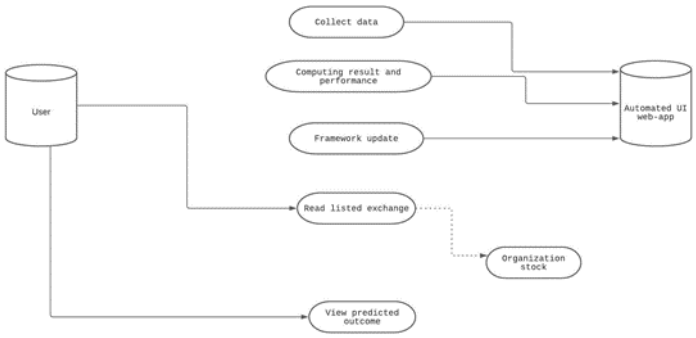
For the analysis, we first collected data of stocks of Nifty index in the last few months from January to June. After the dataset collection, we imported the necessary libraries in python and normalized some features. Basically, we analyzed the cumulative returns of NIFTY index in the last couple of months. Then, we calculated the drawdown of maximum downside risk. After that, we started analyzing google search trend results of India during the COVID-19 pandemic. The important conclusion is that it is public sentiment that has the most impact on the market price. We applied the same for Dow Jones and then analyzed the stock return by industry.

2.2 System Architecture

The architecture employed in sock price trend forecasting's are as under:

- Collect Data: - The necessary stock knowledge of the NSE shall be provided on the NSE website. A machine-driven machine learning program in the Backend shall be prepared to acquire the knowledge required for the system. Knowledge on American airline stocks and of five hundred completely different sp firms shall be acquired from various sources and through human tweets.
- Computing Result and Performance: - The prediction results are processed and created by an automated Machine learning program within Backend. It is the system and its corresponding web-applications that are intended to analyze the results of the prediction and system performance.

- Framework Update: - As the market changes and the industry innovates, it's important to keep the framework up-to-date. The usual trade effects and their real value will automatically be updated when needed by the AI software, and any changes will be made on the backend on a regular basis.



2. System Market Analysis and Prediction System.

- Read listed exchange: - The company selling that's regulated at the NSE in India is usually seen by users and the predictions are seen on websites together.
- Organization Stock: - It is an integral part of monitoring recorded trade. It includes the stock price of specific organizations.

III. METHODOLOGY & EFFECTIVENESS ANALYSIS

3.1 METHODOLOGY

3.1.1 Information Collection and Normalization: -

For the past five years, we have compiled daily trade data for five hundred organizations of SP and five hundred markets since the organization's inception. The primary characteristics of this data that have been consistently collected include Open, High and Low, Close and Volume, Adjusted Shut for each company and SP Index. To date, we have also collected information on peachy stocks and the Dow Jones industrial average throughout the pandemic period from January to June, as well as a normalization control of certain factors. To standardize our data, we have employed the Z-Score method. The mathematical formula for calculating new data from the old data when using this method is as follows: $Z = (X - \mu) \div \sigma$

The Z-score method was chosen as the preferred method due to its ability to address the issue of outliers, which is not present with the Min-Max standardization method. Additionally, it allows for the data to be normalized to conform to the normal distribution, which is more straightforward to interpret. Furthermore, the graph of normalized data allows for the calculation of the number of standard deviations from the mean. Unlike the Min-Max method, which only returns values within a range of 0 to 1, the Z-score method offers flexibility in this regard. It returns values with a zero-mean value, with values greater than or equal to the mean converted to positive values, values equal to the mean set to 0, and values lower than the mean set to negative values.

The table below shows the characteristics of the data set used in the project.

Open	Opening price of stocks on a particular date
High	Highest price recorded on a particular date
Low	Lowest price recorded on a particular date
Close	Closing price of stocks on a particular date
Shares Traded	Number of shares bought or sold on a particular date
Turnover	Total currency exchange for that stock on a particular date

Table 1: Features used in dataset

Date	**Open**	**High**	**Low**	**Close**	**Shares Traded**	**Turnover (Rs. Cr)**
01-Jan-2020	12202.15	12222.20	12165.30	12182.50	304078039	10445.68
02-Jan-2020	12198.55	12289.90	12195.25	12282.20	407697594	15256.55
03-Jan-2020	12261.10	12265.60	12191.35	12226.65	428770054	16827.27
06-Jan-2020	12170.60	12179.10	11974.20	11993.05	396501419	16869.22
07-Jan-2020	12079.10	12152.15	12005.35	12052.95	447818617	17797.68

Table 2: Top 5 rows in dataset

3.1.2 Preprocessing and cleaning: -

The movement includes the inclusion or recovery of missing data and the deletion of redundant data. Similarly, the movement includes the framing of additional supporting highlights from existing highlights.

3.1.3 Extraction of Features: -

This movement focuses on searching for a space of potential element subsets. Subsets that are ideal or close to ideal are then selected to reduce the amount of work that should be limited. To reduce the number of features that are rarely used, a linear model is used to determine features based on the k highest values. A single regressor is tested consecutively for multiple regressors. To avoid overfitting or underfitting of the dataset, a Select K-based algorithm is employed, with f regression as the scorecard. This process is divided into three steps: 1. Start with a consistent machine learning model (M0). 2. Test all models (M1) consisting of only one feature and select the best based on the F statistic. 3. Try all models (M2) including M1 and select the best.

3.1.4 Data Normalization: -

The information should be standardized for accuracy by making sure that all highlights aren't given too much or too little weightage. It also ensures that the optimization algorithm converges faster.

3.2 Effective Analysis

The proficiency of the framework has been evaluated by taking into account two variables: the Root mean square error and the R-square worth.

3.2.1 Root Mean Square Error: -

In insights, the Mean squared inaccuracy (MSE), or mean squared deviation (MSE), of an assessor, quantifies the normal squares of the errors, i.e., the normal squared difference between the assessed quality-effectiveness and the true worth. The main difficulty with mean square error lies in the fact that the requirement for inaccuracy is greater than the requirement for information. As my information has a requirement for 1, and the misfortune work has a requirement for 2, it is impossible to accurately compare information with the error. To address this, we take the root of the average square error, which is the root mean square error. The root serves as the basis for the average square error. RMSE is highly normal, making it an ideal mistake metric for mathematical-emotional expectations. On the other hand, the Mean Absolute Error is more extreme and significantly reduces large inaccuracies.

3.2.2 R-Square worth: -

R-squared is a measure of the degree of variation within a reaction variable, which can vary between zero and one. The direct relapse model clarifies this higher value, which is undoubtedly worth the additional accuracy of the relapse model due to the additional changeability it provides. The R square cost is a pro-particular measure of the variety within the reaction variable, which is clarified by independent factors. It is likely that R-square may be mathematical in nature, but it is to be closed to the information unit to be fitted to the curve. It is to be referred to as the steady-state of assurance or the consistent-of-multiple-assurance for factual technique.

IV. RESULTS AND DISCUSSION

4.1 Actual and predicted values of different bagging and boosting regression models:

Following are the results of our Machine Learning code, which shows the real and forecast values of various bagging and boost regression models based on RMSE and r-squared value.

Algorithm	**RMSE Worth**	**R-Squared worth**
Random Regressor	1.43254343-07	0.956669
Bagging Regressor	1.329966e-07	0.959771
Adaboost Regressor	2.988297e-07	0.909611
K-Neighbors Regressor	0.00039015	-117.01176
Gradient Boosting	1.274547e-07	0.961448

4.1.1. Random Regressor:

Random Forest is a type of partner degree group method that uses multiple call trees to perform expressions for every relapse and characterization errand. It also uses a method called Bootstrap Aggregation and is commonly referred to as sacking. What is sacking, you'll ask? Well, sacking in the Random Forest methodology involves coaching each call tree on a specific information sample wherever sampling is done with replacement. The idea behind this is that it's often important to mix different call trees together to get the best result, instead of relying on individual call trees.

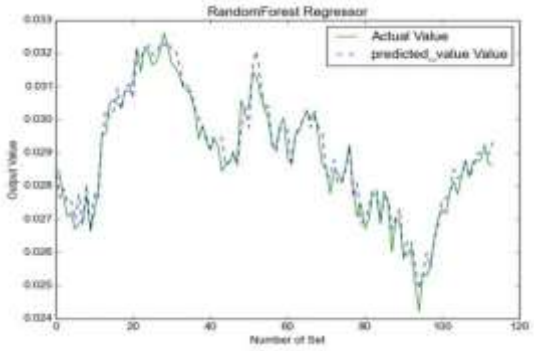


Fig: RMSE worth = 1.4325434e-07, R-squared worth = 0.956

RMSE -: 10^{-07}

R-squared worth: - 0.956

4.1.2 Adaboost Regressor:

AdaBoost is the first extremely well-made boosting algorithmic program designed for binary classification. AdaBoost is short for Adaptational Boosting. It may be a very trendy boosting technique that combines several 'weak classifiers' into one 'strong classifier'.

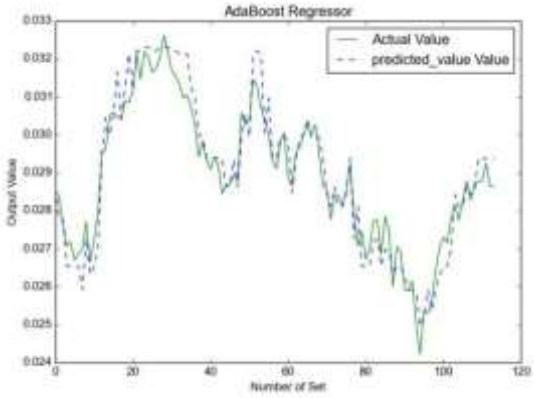


Fig: RMSE worth = 2.9882972e-0, R-squared worth = 0.909

Green Line Shows the Actual value and Blue dotted line depicts predicted value.

RMSE -: 2.988 e-07 (pretty low)

R-squared worth: - 0.909 (pretty high)

4.1.3 Bagging Regressor:

Bootstrap summation, also known as bagging, is a widely applicable method for reducing the variation of a real-world learning approach. Generally, averaging a large number of perceptions reduces disparities. A typical approach to reduce the variation and henceforth increase the accuracy of a measurable learning approach is to collect multiple preparing sets from the population and construct a different expectation model using each preparation set. After that, the subsequent expectations are normalized. Additionally, rehashed tests can be taken from the (unused) preparing informational collection.

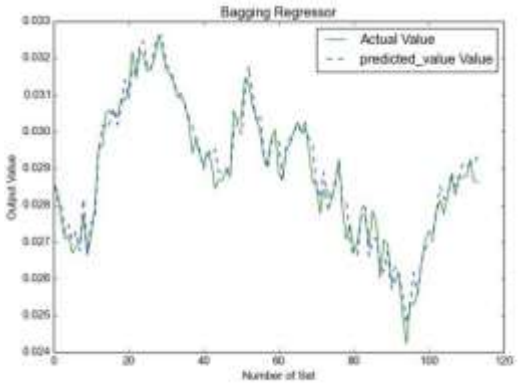


Fig: RMSE worth= 1.329966e-07, R-squared worth=0.961

Similar results are found for bagging and boosting regressor. Bagging regressors have slightly lower R-Square values and

RMSE values as compared to gradient boosting regressors, and RMSE values are slightly higher compared to gradient boosting.

RMSE: - 1.329 e-07

R-Squared worth: - 0.959

4.1.4 Gradient boosting Regressor:

The gradient boosting algorithm is a common algorithmic rule in which each predictor corrects the error of its predecessor. Unlike in the Adaboost algorithm, the weights for the coaching instances are not adjusted, but instead, each predictor is taught to treat the remaining errors of the previous predictor as labels. This algorithm is referred to as Gradient Boosted Trees, the base learner for which is the CART algorithm (Classification and Regression trees). A key parameter is used in this system, which is called 'Shrinkage'. This parameter refers to the fact that the predictions of each tree in the ensemble are shrunk when the training rate is increased by a factor between zero and one.

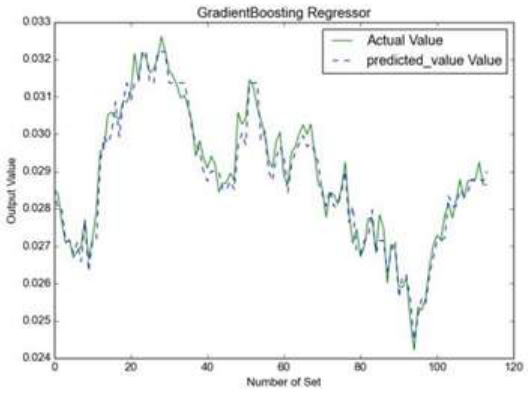


Fig: RMSE worth= 1.274547e-07, R-Squared worth=0.959

The actual and predicted value is pretty much in unison along with

RMSE: - 1.275e-07

R-Squared worth: - 0.961

But this is performing a little better than the Adaboost regressor.

4.1.5 K-Neighbors Regressor:

K-nearest neighbors is one of the most widely used basic algorithms in Machine Learning, and has a prominent place in the regulated learning space. It is particularly useful in design recognition, information processing, and interruption detection. It is widely applicable in all circumstances, as it is non-parametric, meaning that it does not make any assumptions regarding the distribution of data (as opposed to elective calculations, such as Gausian mixture models, which accept a Gaussian representation of the information).

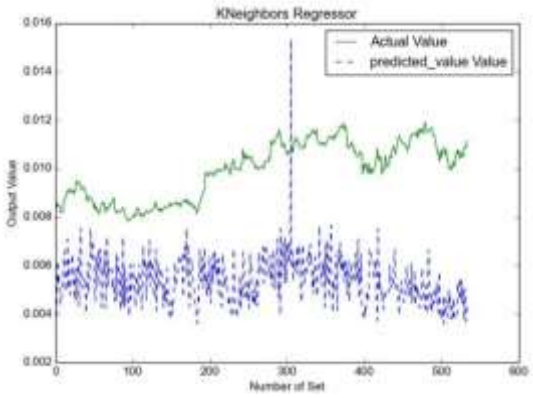


Fig: RMSE Worth= 0.00039015, R-Squared worth= Too High

It is evident that the RMSE value of K-Neighbor is very high and the R-squared is also sufficiently negative. This is a very weak regressor that can be observed from the plot of this particular regressor graph.

RMSE: - 0.00039015

R-Squared worth: -117.01176

V. CONCLUSION AND FUTURE SCOPE

5.1 American Airline Industry Conclusion:

In the analysis of the results obtained for American airlines, it has been determined that the gradient boosting regressor is the most efficient. This is followed by the bagging regressor, random forest regressor, Adaboost regressor and K-Neighbour regressor. The bagging regressor has been found to be more effective than bagging (Bootstrap sampling) due to the fact that the combination of multiple freelance base learners significantly reduces the error. Therefore, we propose to provide several freelance base learners as potential. Each base learner should be generated by randomly sampling the first set of knowledge with replacement. Furthermore, it is evident from the results that additional hidden layers improve the model scores. Random forest is an additional variant of bagging, where the primary difference is the inclusion of irregular features.

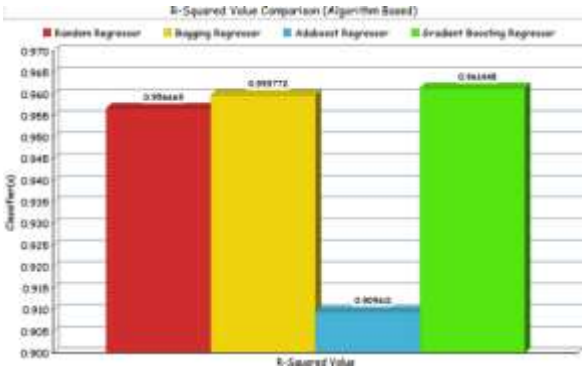


Fig: 5.1 RMSE Value Comparison

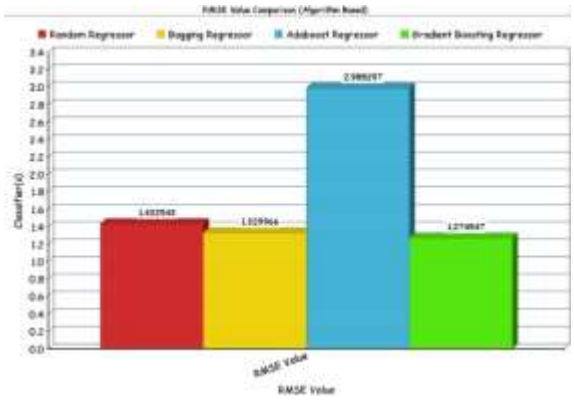


Fig. 5.2 R-Squared value comparison

5.2 Covid effect on Nifty-50 Conclusion:

Based on the results we've collected; we can say that absolute affirmed cases and day-to-day cases in India and around the world have a correlation with Nifty50 closing costs. Also, Nifty50 shutdown costs are a lot worse with completely affirmed and everyday cases. The prognostic analysis shows that Random Forest Regression is outperforming every other model. We'll be looking at how COVID-19 will affect the cost of Nifty50 records. We'll be using the Pearson Correlation and looking at how it'll affect the cost of the list. We'll also be using different AI relapse models to predict the cost. The results show that the unpredictability of the market is directly related to the number of cases. The irregular Forest Regression model usually has higher RMSEs and R-Squares.

5.3 Future Scope:

The prediction of stocks can be performed on more intricate datasets, which can be used to forecast the general and more complex stocks. In the near future, the availability of high-performance computing systems will be available, which could facilitate the training of gradient boosting models on large datasets of training data to achieve higher accuracy and faster predictions. A method of making predictions using gradient boosts and other boosting regression can be beneficial in the future. Additionally, for the analysis of model efficiency, other fac-tors can be considered in addition to RMSE and R-Squared Worth.

VI. REFERENCES

[1] Yuning Zhang,Yuqing Dai,"Machine Learning in Stock Price Trend Forecasting"

[2] Shunrong Shen, Haomiao Jiang,"Stock Market Forecasting Using Machine Learning Algorithms"

[3] Wei Huangb, Yoshiteru Nakamoria, Shou-Yang Wangb,"Forecasting stock market movement direction with support vector machine"

[4] Xin Guo,"How can machine learning help stock investment?"

[5] Mark T. Leung , Hazem Daouk , An-Sing Chen,"International Journal of Forecasting (2000) 173–190/ ijforecast Forecasting stock indices: a comparison of classification and level estimation models"

[6] Joumana Ghosn Yoshua Bengio,Multi-Task Learning for Stock Selection: "* Dept. Informatique et Recherche Opérationnelle Université de Montréal Mon- tréal, Qc H3C-3J7

[7] Tsai, C.-F., Lin, Y.-C., Yen, D. C., and Chen, Y.-M,Predicting stock returns by classifier ensembles." Applied Soft Computing 11 (2011), 2452–2459

[8] Wenzhi Ding Ross Levine Chen Lin Wensi Xie, CORPORATE IMMUNITY TO THE COVID-19 PANDEMIC; Levin, A. U,Stock selection via nonlinear multi-factor models.

[9] Breiman, L.,Random forests. Machine Learning 45 (2001)

[10] Yasean Tahat, Stock Market Returns, liquidity, and COVID-19 Outbreak: Evi- dence from the UK

[11] Ayo CK, Stock price prediction using the ARIMA model,In: 2014 UKSim-AMSS

[12] 16th international conference on computer modelling and simulation.

[13] Atsalakis GS, Valavanis KP, forecasting stock market short-term trends using a neuro-fuzzy based methodology