International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

QUANTITATIVE ANALYSIS OF DIFFERENT ALGORITHMS FOR STOCK PRICE PREDICTION

Sarvadnya Navti[1], Dr. Nilesh Fal Dessai[2]

[1]Student, Department of Information and Technology, Goa College of Engineering, India
[2]Associate Professor and HOD, Department of Information and Technology, Goa College of Engineering, India

---

Abstract – *With the advent of social involvements over the internet, company-specific advertisements, and market influencers, stock investments came into a trend across the investors and other communal platforms. Direct online investment through third-party applications and daily market observation facilities has reduced the use of brokers thus reaching a wider audience. Machine Learning (ML) and Artificial Neural Networks (ANN) eases the goal of achieving a system that can replicate a market to identify when the market and particular stock in a market is suitable for investing. Many forecasting algorithms are available in ML that helps in making human error-free decisions at a faster rate. To extrapolate and forecast any sort of quantity, multiple algorithms can be used to abridge complexity and increase optimality. Widely used algorithms: Linear Regression (LR), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) that are reviewed in different kinds of literature and are used for prediction compared for accuracy by calculating their Root Minimum Squared Error Rates (RMSE). The dataset constituted of time series data and news strings. The news assessment system is scored using VADER (Valence Aware Dictionary and Sentiments Reasoner) tool that is used for semantic computation. Accuracy for the combined result is checked for all algorithms. This taxonomical study is projected to minimize the error rate and increase the accuracy of a prediction model. Market investments are considered the fastest mode of earning a profit if invested appropriately and wisely. Although, multiple factors termed as Market Risks can affect the market and price of the stock. For that, specific data or entity that reflects these factors needs to be traced and scored. Tally of these scores converges to the current situation of the market and observations can be noted from that.*

Key Words: **ML, RF, XGBoost, LR, DT, LightGBM, RMSE, VADER**

1.INTRODUCTION

There are many challenges to be addressed in this topic, targeting the crucial components of this sector helping the equity traders expand the revenues by precise prediction. The stock market prediction is classified into fundamental and technical analysis. The fundamentals are as carried out in past using formulas and daily market hypotheses. These techniques are also used in the present time and have lots of drawbacks. Simple inference using such methods cannot be reticulated in making big financial decisions.
To identify and define a track between these drawbacks technical seek to determine the future price of a stock based solely on the trends of the past price. The data is time-related so a series analysis needs to be carried out using series of analyses in real-time.

There are elementary assumptions considered in the technical analysis part of the stock market. Primary assumption relates to the significant information about the performance of a company that is already priced into the stock. Other is about what trend does it follows. Lastly that how often historical prices tend to repeat, generally due to market psychology.

Fundamental analyses are often used in long-term approaches as the results obtained here are more suitable. In contrast to technical analysis, the technique infers good results for short-term strategies. Merging these two techniques using some quantitative ML techniques will benefit the investors to track the market in both situations. The users can vary their mindsets based on the external situation of the market whether to invest for long or marginalize the profit of short investments. The ideal system is expected to provide accuracy that is expected from the end-user. ML offers a variety of algorithms for forecasting the time series datasets of the stock market. Each algorithm can form a pattern that is used for prediction. The features learn from the pattern and real-world interfaces to generalize the explicit problem statement. It also has an advantage with the coding interface since the market prediction is more often rule-based. Some factors and drawbacks of some algorithms make the output noisy. Also, external factors might affect the market and are exposed by a simple price predictor. An opinion mining or news assessment using a sentiments analysis system provides a better response when blended with the price forecasting model. Thus gauging criteria attain external factors score enclosed for a better forecast.

1.1 Stock Investments

Getting future insights about the impulsive financial sector, where hard-earned money can be invested and frequency of profit can be earned still attracts lots of researchers towards the topic of stock prediction. Stock is a very crucial sector for a company. Buying a stock as a customer of a company, helps

the company to grow its economy. The part of this economy is the float for the stock owner. The stock's rate for a corporation is not fixed and proportional to the condition of the market. It keeps changing frequently from minute to minute depending upon the buying selling sales percentage of the market. Displacement of the stock prices is less or negligible for the giant companies in the market, but still, they are also vulnerable to market risks.

1.2 Forecasting Algorithms

The impulsive nature of the stock market is very difficult to predict. The hikes in prices sometimes range infrequent times. Marginalizing errors in such situations become a very difficult task. The theme of this study is identifying the algorithm with better accuracy, least error, execution time, and fewer complexities among the different algorithms. Different algorithms considered are:

i. Linear regression

It belongs to the class of statistics and uses a simple linear function for prediction. The linearly separable line separates the sample points and using that equation error value is minimized for a better fit of the line. [7] It has two variables: one is explanatory that is independent of other factors and the second is responsive that depends on the explanatory variable. The best-fit decision boundary is very much suited for prediction models like stock regression. This method is rule-based and provides explicit coding ease. This method of ML is considered to be the easiest and best forecasting algorithm as its variants are superior to the rest of ML algorithms in most categories.

Output generated by linear trend is not accurate and has a high error rate. Note LR works best when fine-tuned with Stored Vector Machines and confusion matrix. Also, multiple regression techniques can be used for better predictability in iterative form.

ii. Light gradient boosting machine

LightGBM ML library provides an efficient and effective implementation of the forecasting algorithms for multiple cases using gradient boosting algorithms. [6]The algorithm extends itself by adding automatic feature selection by centering the boosting examples with larger gradients. This results in fast-track training and ameliorated predictive performance. As such, LightGBM has developed a genuine algorithm for ML competitors when working with tabular time series data predictive modeling tasks.

iii. Extreme gradient boosting machine

XGBoost algorithm is considered for most gradient boosting for classification and regression problems. Because of its fast and efficient nature and range of predictive modeling tasks it has gained lots of popularity over the years. In this review, XGBoost is used for time series data forecasting. Again we have a drawback of this algorithm to address and that is it requires a time series dataset to be transformed into a supervised learning problem in primary stages. [6]It also requires walk-forward validation for model evaluation, as evaluating the model might result in biased results.

iv. Random Forest

Random Forest is an ML algorithm that creates different decision trees by dividing the problem into multiple sub-problems. [9]A random forest classifier identifies fits per each decision tree and control overfitting. The entire dataset is grouped according to best-suited criteria and the best converging result is grouped for the optimistic result. Complexity depends upon the size of the dataset. The bigger the decision trees and the greater the number of trees, the feasibility of storage and execution time also demerits the algorithm for prediction analysis.

v. Decision trees

DT belongs to the non-parametric supervised learning method that is used for problems related to groups of regression as well as classification. [8]DT has the objective to generate a pattern that predicts the value of a decision variable by adapting simple decision rules inferred from the data features. A tree is formed of leafs each contributing constant approximation. DT learns from data to approximate a sigmoidal graph with a set of if-then-else decision rules. The disadvantage of this model is: the depth is directly proportional to the complexity of the model and is subject to over-fitting.

1.3 News mining

The market risks are not considered by time series prediction algorithms. Algorithms do not work for strings unless modeled to some parametric values. If news assessment scores are considered alongside the price prediction model, the output we get will have sentiments associated. In case of a stock is doing well historically and a sudden external factor like political issue, pandemic, company threats, natural disaster, market competitors, etc. affects the market. This will lead to a loss for the investors. To track this, a news mining system is also required with valid scores associated in terms of positives, negatives, and neutral.

2. RELATED WORK

Stock market forecasting is considered as a stimulating research area. It thought-provokes a lots of effort for achieving accuracy in price predictions models. Objective is investment with profit connotation maximization. Research Thesis [14] Predicting the Movement Direction of OMXS30 Stock Index Using XGBoost and Sentiment Analysis, Author P. Elena pointed that tree-based ensemble ML methods (XGBoost) have proved positive in studies. The research also stated that the trend to incorporate multiple data sources such as historical price datasets and textual datasets, in prediction models is leaning, aiming to achieve superior forecasting performance.

There are various theories associated with market forecasts. Two of the theories are explained in [1] Quantitative

Analysis of Stock Market Prediction for Accurate Investment Decisions in Future, are about Efficient Market Hypothesis (EMH) and Random Walk Theory (RWT). Both theories expresses that it is not possible to outperform the market since the prices of stock reflects all accessible data about the resources. For this, the market hypothesis are categorized into: Weak, Semi-strong and strong EMH. In Weak EMH class, only past data is considered. All information is utilized in semi strong EMH class and strong EMH utilizes all publicly and privately available information. Whereas in RWT, hypothesis assumes that it is difficult to forecast price of stock as the price does not depend on historical time series data related to stock.

Paper [11] Review Of Stock Prediction Using Machine Learning Techniques discusses about various ML approaches such as Natural Language Processing (NLP), LR, K Nearest Neighbors (KNN), Stored Vector Machines(SVM), Long Short Term Memory(LSTM) chains, ANN ,etc. for achieving the goal of building the model that help investors and brokers to invest in the market.

3. PROPOSED APPROACH

The diagram visualizes proposed forecasting approach of stock performance. The share price predictor module constitutes of data pertaining to one company's stock price with respect to time in the market. That data runs through the algorithm selected and predicts the price score.

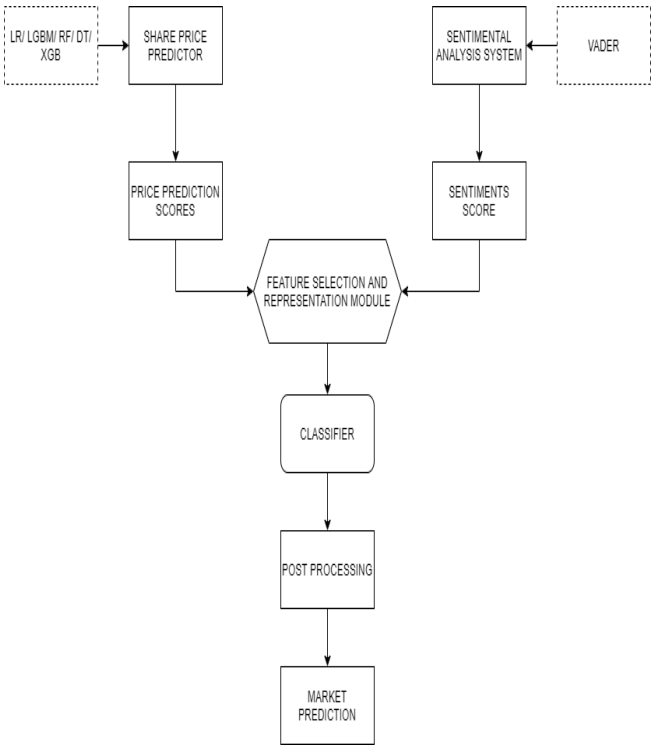

Fig 3.1: Proposed approach

Similarly, on the other hand the news data is fed to the sentimental analysis unit. [15]VADER is a lexicon and rule-based tool that analyses sentiments that is specifically attuned to sentiments expressed in media. VADER assigns sentiment scores based on polarity of the news. This tool is precisely defined for each news type and scores accordingly. This scores are used to identify market performance. The data obtained from both units is merged and represented. Classification algorithm makes the buying-selling decisions and notifies end user accordingly. It represents market in binary form in terms of 0: Market condition is bad or 1: Market is good for selling. Bad condition of market sometimes is suitable for buying but stock must me chosen prudently.

4. ARCHITECTURE

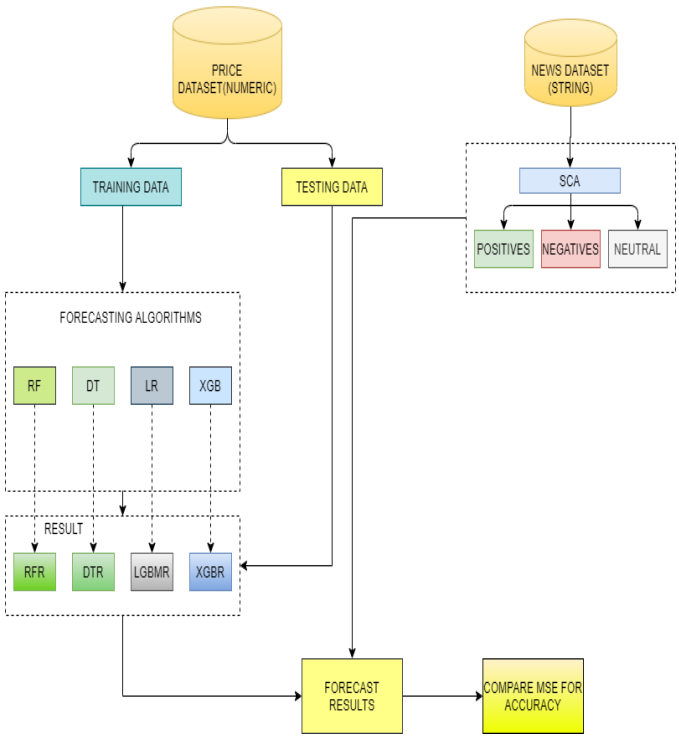

Fig 4.1: Architecture for comparative analysis of different algorithms

The above architecture has two units. One is stock performance evaluation, and other is news mining and news sentiment scoring unit. The stock performance evaluation unit is where the different algorithms are compared. The time series stock price data is distributed into training and testing data. The training data is fed to different ML algorithms to predict the future insights with respect to price. Then later, predicted output is compared with the Testing data. Using that error is calculated. Sentiments scores are merged with the forecasted result and ultimately the RMSE and accuracies are gauged to find algorithm with best outcome.

5. IMPLEMENTATION

Different algorithms were implemented independently to calculate RMSE. Data collected was cleaned and normalized. Then necessary features were extracted from the same. The processed data was used as input for LR, RF, DT, LightGBM and XGBoost algorithms. The figure below shows trained datasets with predicted versus testing datasets for linear regression algorithm.

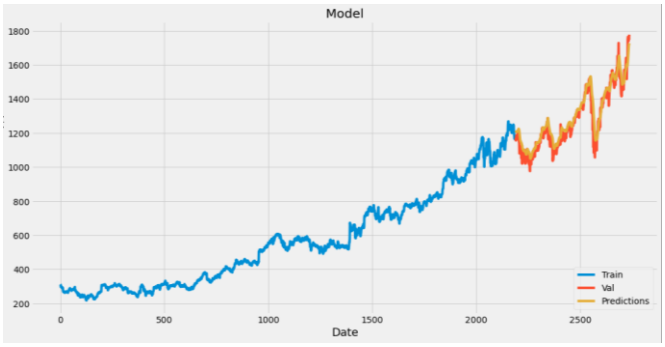

Fig 5.1: Comparing predicted result with testing

Similarly, sentiment analysis using news assessment system was implemented using VADER tool. The data from news headlines was accessed from a web source.

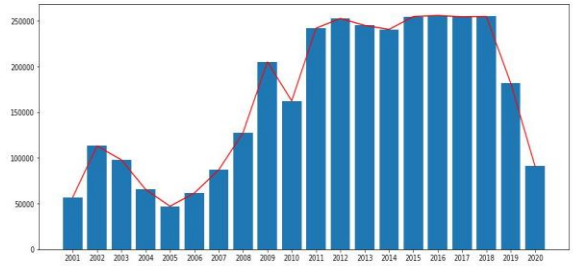

Fig5.2: Year-wise breakdown of news datasets related to stock news

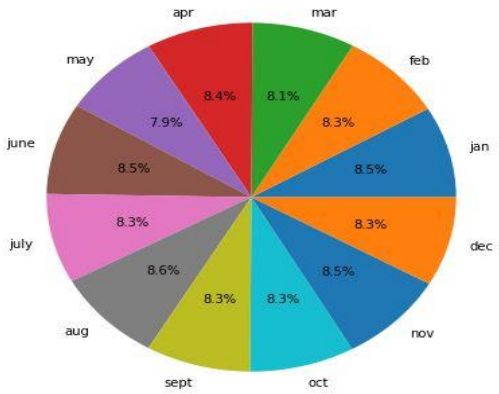

Fig5.3:Month-wise involvement of news regarding market

6. SAMPLE WORKING

The proposed system takes external input that depends upon the scenarios of market. The information is inclusive of stock price in market up's and down's and news published in economic sector relating to the stock market. The news is scored if the news contains optimistic metaphors like rise/growth in production, increasing trade. Else it is assigned a value of 0. If it contains negative words like decrease, loss then it is assigned a negative score. Based on total mean of this score the stock is assigned a sentiment alongside component price predictor. The review is specifically done to compare different algorithms for forecasting. So different algorithms were used and their RMSE is then calculated and compared for accuracy using testing data. After compiling both results, the condition of market was represented in binary form. 1 represents positives for investment and 0 represents bad market condition. In case if the live data can be used as an input to system, better accuracy can be achieved.

7. RESULT AND SOLUTIONS

Sr. no.	Algorithm	Function	RMSE
1	Linear Regression	algo = LinearRegression() algo.fit(x_train, y_train) predictions= algo.predict(x_test)	0.05692347045438241
2	LightGBM	gbm = lightgbm.LGBMRegressor() gbm.fit(x_train, y_train) predictions = gbm.predict(x_test)	0.0583079056070462
3	XGBoost	xgb = xgboost.XGBRegressor() xgb.fit(x_train, y_train) predictions = xgb.predict(x_test)	0.05968830860645 93
4	Random Forest	rf = RandomForestRegressor() rf.fit(x_train, y_train) prediction=rf.predict(x_test) print(predictions[:10]) print(y_test[:10])	0.05257968397499098
5	Decision Tree	dtr = DecisionTreeRegressor() dtr.fit(x_train, y_train) predictions = dtr.predict(x_test) print(predictions[:10]) print(y_test[:10])	0.10831900809236311

Fig 7.1: Comparative analysis of different algorithms for RMSE

Even after decades of study by the brightest minds in finance, no provable solution could be obtained on this research. A good conclusion that can be drawn is that there may be some momentum effects observed in all algorithms that are used for forecasting in ML. ML offers diversified solutions in terms of its algorithm so choosing best algorithm sometimes becomes difficult with the type of data

that is handled. For a specific data one algorithm proves to be efficient but same algorithm sometimes lacks for other datasets. This study concerned in identifying best suit algorithm for two separate datasets. The RMSE for all algorithms was calculated as displayed in tabular form. The RMSE was almost similar in all cases excluding DTs. DT is a part that is used in RF where RF shows least RMSE. LightGBM and XGBoost works closely. Linear regression was the easiest and fast executed multifaceted algorithm from the list.

Different stock market related challenges that are addressed in this study:

(1). Dynamic nature of Stock Prediction happens due to the fluctuations in stock market which changes every day. To overcome this challenge real time data was obtained periodically using web scrapping to know the prediction of every change.

(2) Training and retraining the model frequently to predict the stock market change. There is an ambiguity that after how much period model should be trained to get optimum prediction.

(3) Selecting the appropriate machine learning model for the respective prediction by computing quantitative analysis of different models. Suppose if the training data is much larger than no. of feature in such case, linear regression is better than XGBoost but if no. of features is much larger than training data, then RF outperforms all compared algorithms.

(4) Out of vast the real time amount of Normalization, dataset sorting and filtering, pre-processing, features extraction was a challenging task to estimate.

One of problem causing in most scenarios is that the prediction is way lower than the price. Mostly in the tree-based algorithms (like DT) cannot extrapolate data. If you take a look at the plot with data split, imagine that all what an algorithm sees when training is train set obviously. For the algorithm, train set bounds are total bounds for the rest of data. If the validation or test set reach higher than training set maximum value, tree-based classes cannot predict them correctly. They will predict the value as the maximum they are aware of.

8. CONCLUSION

The RMSE varies from algorithm to algorithm for the dataset. Accuracy and performance of the algorithms depend on the type of dataset collected. This study holds positive for linear regression, but LightGBM, XGBoost and Random forest also suited better. The advantages of using linear regression is its non-complex implementation nature and multi-functional nature. Use of iterative regressions converges the solution to the best fit decision parameter. In this study the market mimicry is performed using different algorithms and tested accordingly. Stock market prediction is the act of mitigation of risk through the spreading of investments across multiple entities, which is achieved by the pooling of a number of small investments into a large bucket. Stock Market is the most suitable investment for the common man as it offers an opportunity to invest in a diversied, professionally managed portfolio at a relatively low cost. The analysis of the stock market cycles helps to gain how market acts in reference period. Identification of the gains and losses period in simulation also helps in obtaining market acumens. The results of our analysis also show that the stock market cycles have dampened in the recent past.

REFERENCES

[1] Sharma, Surbhi & Kaushik, Baij. (2018). Quantitative Analysis of Stock Market Prediction for Accurate Investment Decisions in Future. Journal of Artificial Intelligence. 11. 48-54. 10.3923/jai.2018.48.54. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Chellaswamy, Karthigai Prakasam & Nm, Natchimuthu & Faniband, Muhammadriyaj. (2021). Stock Market Reforms and Stock Market Performance. International Journal of Financial Research. 12. 202. 10.5430/ijfr.v12n2p202. K.

[3] https://www.investopedia.com/articles/07/mean_reversion_martingale.asp

[4] Huang, W., Nakamori, Y., Wang, S., "Forecasting Stock Market Movement Direction with Support Vector Machine," Computer & Operations Research, Vol. 32 (10), pp. 2513–2522, 2005.

[5] Kandananond, Karin. (2012). A Comparison of Various Forecasting Methods for Autocorrelated Time Series. International Journal of Engineering Business Management. 4. 1. 10.5772/51088.

[6] https://machinelearningmastery.com

[7] https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

[8] https://scikit-learn.org/stable/modules/tree.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[10] Mohammed, Walid. (2020). Challenges of Stock Prediction. 10.4018/978-1-7998-1086-5.ch013.

[11] Patel, Ramkrishna & Choudhary, Vikas & Saxena, Deepika & Singh, Ashutosh. (2021). Review Of Stock Prediction Using Machine Learning Techniques. 10.1109/ICOEI51242.2021.9453099.

[12] Li, Bozhao & Chen, Na & Wen, Jing & Jin, Xue-bo & Shi, Yan. (2015). Text Categorization System for Stock Prediction. International Journal of u- and e-Service, Science and Technology. 8. 35-44. 10.14257/ijunesst.2015.8.2.04.

[13] Shah, Dev & Isah, Haruna & Zulkernine, Farhana. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. International Journal of Financial Studies. 7. 26. 10.3390/ijfs7020026.

[14] P. Elena, 'Predicting the Movement Direction of OMXS30 Stock Index Using XGBoost and Sentiment Analysis', Dissertation, 2021.

[15] https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader

[16] Polamuri, Subba & Srinivas, Kudipudi & Mohan, A.. (2019). Stock Market Prices Prediction using Random Forest and Extra Tree Regression. International Journal of Recent Technology and Engineering. 8. 1224 - 1228. 10.35940/ijrte.C4314.098319.