Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2451

Robotic Process Automation for Stock Selection Process and Price Prediction Model using Machine Learning Techniques

Saumitra Chattopadhyay
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number:1609 - 1618
Publication Issue:
Vol 70 No. 2 (2021)

Abstract

The usage of AI in the financial industry has increased the reliance on stochastic models for forecasting market behaviour. Quantitative analysts are constantly working to increase the precision with which machine learning models predict stock returns. Regression models using Support Vector Machine (SVM) and Random Forest are well renowned for their ability to predict closing prices with high accuracy. This study suggests a method for analysing and forecasting stock prices using an ensemble of these algorithms.Using basic market price data from India's National Stock Exchange (NSE), datasets are used that have been preprocessed to add common technical indicators as features. The size of the training dataset is decreased by using feature selection techniques to rank the features according to their impact on the final closing price. Sentiment analysis is also used in the study to examine the effect of investor sentiment on stock prices. Twitter postings with a specific corporate hashtag are rated as good or bad using a trained Word2Vec model. The ensemble model is then trained on a fresh dataset made up of counts of both positive and negative tweets across time as well as technical indicator data.This paper makes a contribution to the area by offering an ensemble model for stock price prediction that blends SVM and Random Forest regression models. The study illustrates the significance of feature selection in lowering dataset size as well as the limited influence of aggregated sentiment analysis from Twitter data on the performance of the model as a whole. For researchers and quantitative analysts looking to improve the precision of stock price prediction models in the financial industry, these findings offer invaluable insights.

Article History
Article Received*: 05 September 2021*
Revised*: 09 October 2021*
Accepted*: 22 November 2021*
Publication*: 26 December 2021*

Index Terms—Machine Learning (ML), Predication, Classification, Ensemble Regression, Sentiment Analysis

I. Introduction

Quantitative analysts have struggled to forecast stock market movements, but thanks to the success of machine learning algorithms, the financial technology sector has embraced AI-driven models in order to increase earnings. Despite the fact that the Efficient Market Hypothesis makes it difficult to predict the market with absolute certainty, traders continue to make an effort to maximise their investment returns.

According to EMH, there are more variables that affect market behaviour than just basic price data. Social media user feedback regarding a particular company has been shown to affect the stock's closing price the day after. These societal elements should be taken into consideration while

developing a model that can anticipate stock trends effectively. Sentiment analysis is a method for determining the degree of a sentence's positive or negative connotation. With the help of this technique, tens of thousands of messages relevant to a particular company's stock can be analysed via microblogging sites (like Twitter [5]), and the insights gained can be utilised to train the model, which should then function more effectively.

The major goal of this research is to determine experimentally whether two models that are each proficient in forecasting stock prices may function more effectively together. We take into account Support Vector and Extremely Randomised (ExtRa) tree-based regressors trained on datasets of technical indicator obtained from companies listed on the Indian National Stock Exchange.

This study goes beyond training on technical indicator data and explores the impact of ncorporating public opinion from Twitter on stock price prediction. By utilizing sentiment analysis, we investigate how the sentiment expressed in tweets can affect the performance of the prediction model. Our contributions can be summarized as follows:

- Feature Extractiuon: Extra trees are used to accomplish feature selection on a dataset based on technical indicators. This method aids in decreasing the quantity of the dataset, increasing computational effectiveness, and concentrating on the most crucial features for stock price prediction.

- Stacked Regressor: Using a Stacked Regressor, which we train, we may combine the knowledge gained through Support Vector and ExtRa tree regressors. By utilising the advantages of these models, we hope to improve the precision and stability of our price prediction.

- Using Twitter sentiment analysis: Using sentiment analysis of Twitter data, we examine how public opinion affects stock prices. We attempt to quantify the impact of public sentiment on changes in stock prices by averaging positive and negative tweet counts over time.

II. Related work

A. *Predictive Machine Learning for Stocks*

Technical indicators are computations based on a security or contract's price, volume, or open interest. Technical analysts frequently use these indicators to examine previous data and forecast future market changes. As features for our suggested work, we will use a variety of well-known indicators in this research. Although there are many different sorts of indicators, some have a stronger impact on stock price than others. These indicators' precise formulas, which will be used as features in our analysis, are provided in the study's appendix.

Kumaretal.[1]To predict stock values, a thorough investigation was done on a number of classifiers, including SVM, Random Forest, KNN, Naive Bayes, and Softmax. SVM and Random Forest classifiers demonstrated superior f-measures, which are regarded as better performance metrics, but Naive Bayes classifiers displayed higher accuracy. To prevent potentially harmful trade recommendations for inexperienced traders, the Volatility Index was not taken into consideration as a feature. Additionally, because the Stochastic oscillator and Williams%R indicator offer identical interpretations, they were both left off of the feature list to prevent duplication and shorten computation time. Additionally, support for the features of Moving Average Convergence

Divergence (MACD) and Exponential Moving Averages (EMA) was introduced to take advantage of their importance as leading indicators.

Li and Liao [3] The study contrasted shallow machine learning methods (Naive Bayes, Support Vector Machine, Decision Tree) with deep learning methods (Multi-Layer Perceptron, Recurrent Neural Network, Long-Short Term Memory Network). Decision tree was a viable candidate for the suggested ensemble technique because it had the highest f-measure but the lowest accuracy. Deep learning models were computationally expensive and did not significantly outperform faster shallow models in terms of insights. Furthermore, it was discovered that basic technical indicators like Moving Average and trailing indicators were less effective at forecasting future stock performance.

Wendong et al. [4The study suggested a support vector machine prediction method and feature weighting method using genetic algorithms. The characteristics were determined from the valuation index's components, which also included General Capital, Total Market Value, Price Earnings, Price to Book, Price to Sales, and Price Cash Flow. Important technical indicators were found via feature selection utilising a genetic algorithm based on closing prices over numerous trading periods. As a result, the dataset's dimensionality was decreased, making it possible for a machine learning classifier to forecast stock movement more accurately and for less money. However, when utilised separately, the genetic algorithm model for feature selection proved computationally sluggish.

Yujun et al. [6] The study's finding that SVMs perform better in developed markets than in developing markets suggests that they are useful for price prediction in the established Indian stock market. However, the research's broad conclusion is constrained because it only takes into account two indexes (the S&P 500 USA and the HSI China) as indicators of market success. For traders utilising different technical indicators, more information on SVMs' accuracy in predicting the prices of stocks outside the index would be beneficial. As opposed to just depending on the NIFTY50 index, the suggested approach emphasises training with datasets from specific Indian companies. Additionally, nothing is known about the effects of hyper-parameter adjustment, which is important in SVM models.

Jaiwang and Jeatrakul [2] In order to choose the most crucial technical and fundamental indicators for each stock, the researchers used a large voting system. They used SVM to forecast stock prices and assessed several kernel operations inside the SVM framework. The dot function outperformed other kernel functions including rbf, sigmoid, and polynomial, according to experimental findings. They also emphasised the difficulty of managing a large number of features, which can take up a lot of storage space and computing resources, thereby lessening the influence of key technical indications on the estimated price in the end.

ManojlovicandStajduhar[7]The study concentrated solely on utilising an ensemble model called Random Forest to forecast stock prices. They produced promising accuracies and utilised a wide variety of technical indicators. The effect of hyper-parameter adjustment on the model's accuracy and speed, specifically changing the size of the forest's trees, has not yet been fully investigated.

B. Sentiment Evaluation

Sentiment analysis was used by Bhardwaj et al. [12] to analyse news articles from online newspaper

websites in order to anticipate stock prices and understand how the SENSEX and NIFTY50 indexes behave. The accuracy of stock trend forecasting using SVM was enhanced by Lima et al.'s work [13] by factoring in a general public mood variable. Positive public sentiment towards the stock is indicated by a day where the number of positive tweets outweighed the number of negative tweet

Mittal and Goel [14] The researchers developed a custom sentiment analysis framework based on the Profile of Mood States (POMS) questionnaire to capture public mood. They combined the sentiment analysis results with the previous day's Dow Jones Industrial Average (DJIA) Index values to train a Self Organizing Fuzzy Neural Network (SOFNN) for market movement prediction.Pagoluetal.[15]The researchers found a substantial association between public opinion and the swings in DJIA prices through their use of N-gram and Word2Vec representations of tweets.

III. Proposed Methodology

The proposed research intends to analyse the Indian National Stock Exchange using a combination of sentiment analysis and machine learning methods.

MachineLearning

1) Data Acquisition and Preprocessing:

In the proposed work, price information of individual stocks listed in the NSE was obtained using Quandl's API. The data was then processed using Pandas to create a new dataset, which included additional features based on selected technical indicators (excluding Simple Moving Average and Exponential Moving Average). These indicators were chosen for their effectiveness as leading indicators in predicting future trends, while lagging indicators were deemed less useful for machine learning techniques. The leading indicators were derived for each trading period through simple calculations involving parameters such as opening/closing/highest/lowest price, traded volume, and turnover, readily available from Quandl.

Table I: Feature for QUANDL

FeatureIndex	Details
0	open
1	high
2	low
3	close
4	totaltradequantity
5	turnover(lacs)

*2) Feature Selection:*After preprocessing the dataset, Ex- tRA trees feature selection will be used to narrow down the training dataset size by identifying the features that have the greatest impact on the stock's closing price. ExtRa tree, a Random Forest variant, takes the whole data sample at each split and chooses decision boundaries at random, producing a less compact but computationally less expensive model. Extra trees may perform less accurately on high-dimensional

datasets with noisy features, but in practical applications, they are comparable to a conventional random forest. A predictor offered by Scikit-Learn makes use of many randomised decision trees, increasing predicting accuracy by averaging and reducing overfitting. The number of estimators affects the algorithm's runtime more than any other factor and does not produce noticeable or substantial changes in the selected features.

3) *Stacked Regressor:*Stacking regression is an ensemble learning technique that combines multiple regression models using a meta-regressor. The individual regression models are trained on the full training set, and the meta-regressor is then fitted using the outputs (meta-features) from the individual models in the ensemble [21].

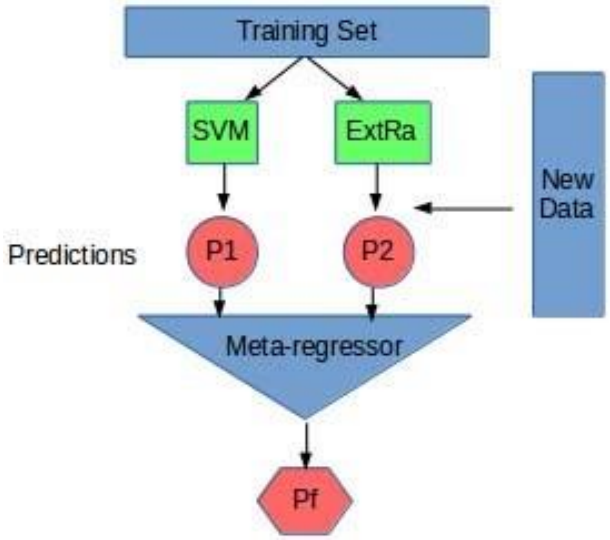

Figure1.Architecture of Regressor as Stacked

IV. Experimental Setup

a. Technical Indicator Data Training:

The proposed model prioritizes making predictions while learning from multiple datasets of varying nature, leading to the exclusion of hyper-parameter tuning. For SVM, a low gamma value is chosen to avoid repetitive price predictions, while a linear kernel is selected due to better performance compared to other kernels. For the Extra tree regressor, randomness in decision tree construction can yield different results, but fixing the random state attribute ensures consistency in retrieving previous results. To evaluate the superiority of the stacked regressor model, 10 independent train-test executions are performed with different random states. However, during feature selection, a constant random state is set to maintain consistent feature selection across executions.

b. Training on Technical indicator data with positive/negative tweet counts:

The separate twitter sentiment dataset for Infosys includes the total number of favourable and negative tweets as of February 2017. Based on the technical indicator dataset size (3-6 months) going backwards from January 31st 2019, the dataset is appropriately split. The generated tuples are then added to the collection of technical indicators, and model training is then performed.

Figure2.Time Series Visualization validation =5

c. Data Validation

Traditional validation techniques like K-folds cannot be employed since the time-series data used in the proposed model are sequential. The model's knowledge of the sequential dependencies between data points is disrupted by randomly dividing the data. Instead, a time-series split is carried out, where the model is tested on the following sequential segment after it has been trained on a portion of the data. The amount of the test data is then decreased by gradually adding a new chunk of the testing data to the training set. Ten time-series splits are used in this study. The fully trained model's performance when put to the test on the entire dataset is represented by the reported R2 and RMSE values.

V. Results & Analysis

Overall, compared to the SVM regression model, the ensemble model outperformed it significantly, but only slightly better than the ExtRa tree-based regressor. The percentage difference between the ensemble model's R2 similarity and RMSE and the SVM regressor is shown in the following graphs.

A. *Feature Selection Method*

Table II presents the features chosen by the ExtRa tree feature selection algorithm. Among the nine technical indicators, the Stochastic Oscillator (% D), Price Volume Trend, and Typical Price exhibit the strongest influence on the closing price across different dataset sizes. Specifically, in the case of a three-month dataset, which is relatively small in terms of the number of data points, the ExtRa tree algorithm identifies a greater number of features that impact the closing price.

Table II: Selected Features By Time-Series Duration

No.ofMonths	SelectedIndices	Indicatorname
3 Months	1,2,3,6, 7	CCISOKIndicatorSODPVTTurnover
4Months	3,6,7, 8	SODIndicatorPVTTurnoverTypical
5Months	3,6,8,	SODIndicatorPVTTypical
6Months	3,6,8,	SODIndicatorPVTTypical

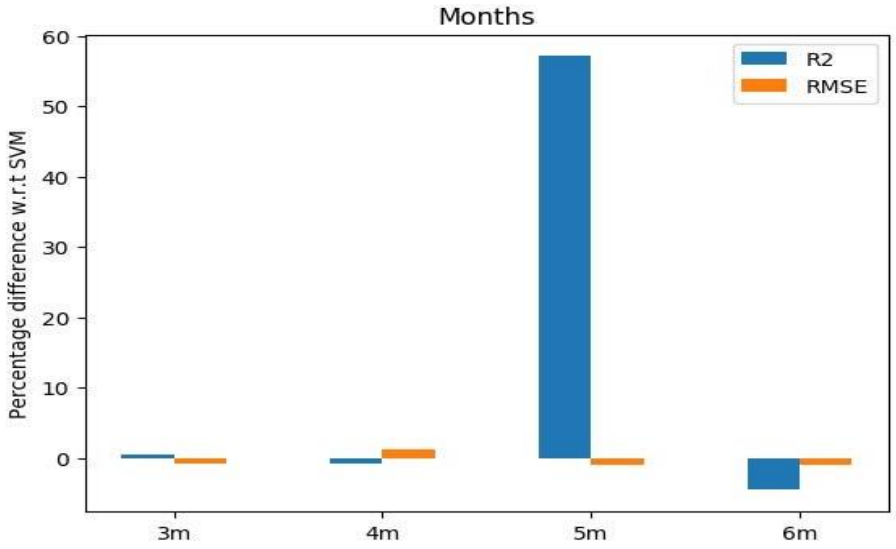

Figure.3.Percentageof SVMmeta-regressor

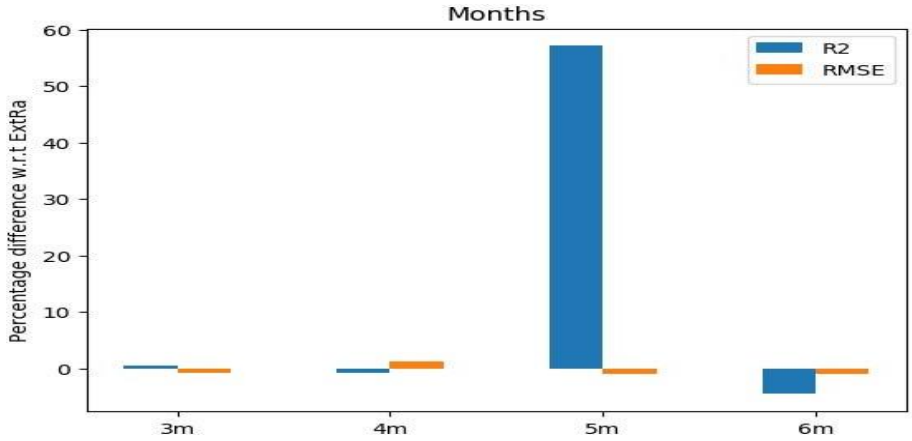

Figure 4.Ensemble model Percentage Improvement

B. Information on technical indicator counts for both positive and negative tweets

When trained using both the technical indicator dataset and the positive/negative tweet counts for each date, the performance of the ensemble model is compared to independent SVM and ExtRa tree models. It's interesting to note that using SVM or ExtRa trees as the ensemble's meta-regressor

improves performance, particularly on the 5-month sample. Figures 3 and 4 show that while the RMSE difference is less favourable for the 6-month dataset, the R2 difference is noticeably improved for both the 5-month and 6-month datasets.

VI. Conclusion

Inthiswork, the application of Robotic Process Automation (RPA) in the stock selection process and the utilization of machine learning techniques for price prediction have shown promising results. By automating repetitive tasks and data retrieval using RPA, the stock selection process becomes more efficient and less prone to human error. Incorporating machine learning algorithms, such as SVM, ExtRa trees, and ensemble models, has improved the accuracy of price prediction models. The inclusion of sentiment analysis from social media data has also provided valuable insights into the relationship between public opinion and stock price movements. However, further research is needed to explore the impact of hyper-parameter tuning and to evaluate the models on a wider range of datasets and markets. Overall, this approach offers potential benefits for investors and traders in making informed decisions and maximizing their returns in the stock market.

References

[1] I.Kumar,K.Dogra,C.Utreja,andP.Yadav,"Acomparativestudyofsu-pervised machine learning algorithms for stock market trend prediction,"in *2018 Second International Conference on Inventive CommunicationandComputationalTechnologies(ICICCT)*,042018,pp.1003–1007.

[2] G. Jaiwang and P. Jeatrakul, "A forecast model for stock trading usingsupport vector machine," in *2016 International Computer Science andEngineeringConference(ICSEC)*,Dec2016,pp.1–6.

[3] W. Li and J. Liao, "A comparative study on trend forecasting approachfor stock price time series," in *11th IEEE International Conference onAnti-counterfeiting,Security,andIdentification(ASID)*,102017,pp.74–78.

[4] Y.Wendong,L.Zhengzheng,andJ.Bo,"Amulti-factoranalysismodel of quantitative investment based on ga and svm," in *2017 2ndInternational Conference on Image, Vision and Computing (ICIVC)*, June2017,pp.1152–1155.

[5] "Twitterapidocumentation,"https://developer.twitter.com/en/docs.html.

[6] Y. Yujun, Y. Yimei, and L. Jianping, "Research on financial time seriesforecasting based on svm," in *2016 13th International Computer Confer-ence on Wavelet Active Media Technology and Information Processing(ICCWAMTIP)*,Dec2016,pp.346–349.

[7] T.ManojloviandI.tajduhar,"Predictingstockmarkettrendsusingrandom forests: A sample of the zagreb stock exchange," in *2015 38thInternational Convention on Information and Communication Technol-ogy, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1189–1193.

[8] T. Ye, "Stock forecasting method based on wavelet analysis and arima-svrmodel,"in*20173rdInternationalConferenceonInformationManagement(ICIM)*,April2017,pp. 102–106.

[9] J. Yang, R. Rao, P. Hong, and P. Ding, "Ensemble model for stock pricemovement trend prediction on different investing periods," in *2016 12thInternationalConferenceonComputationalIntelligenceandSecurity(CIS)*,Dec2016,pp.358–

361.

[10] B. Labiad, A. Berrado, and L. Benabbou, "Machine learning techniquesforshorttermstockmovementsclassificationformoroccanstockexchange,"in*11thInternationalConferenceonIntelligentSystems:TheoriesandApplications(SITA)*,102016,pp.1–6.

[11] M.Ceci,G.Pio,V.Kuzmanovski,andS.Deroski,"Semi-supervised multi-view learning for gene network reconstruction," *PLOSONE*,vol.10,no.12,pp.1–27,122015.[Online].Available:https://doi.org/10.1371/journal.pone.0144031

[12] A. Bhardwaj, Y. Narayan, Vanraj, Pawan, and M. Dutta, "Sentimentanalysisforindianstockmarket prediction using sensex andnifty," *Procedia Computer Science*, vol. 70, pp. 85 – 91,2015,proceedingsofthe4thInternationalConferenceonEco-friendly Computing and Communication Systems. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S187705091503207X

[13] M.L.Lima,T.P.Nascimento,S.Labidi,N.S.Timbo,M.V.L.Batista,G. N. Neto, E. A. M. Costa, and S. R. S. Sousa, "Using sentiment anal-ysis for stock exchange prediction," *International Journal of ArtificialIntelligence&Applications*,vol.7,pp.59–67,012016.

[14] A.MittalandA.Goel,"Stockpredictionusingtwittersentimentanalysis,"2011.

[15] V.S.Pagolu,K.N.Reddy,G.Panda,andB.Majhi,"Sentimentanalysis of twitter data for predicting stock market movements," in *2016International Conference on Signal Processing, Communication, PowerandEmbeddedSystem(SCOPES)*,Oct2016,pp.1345–1350.

[16] S.BharathiandA.Geetha,"Sentimentanalysisforeffectivestockmarket prediction," *International Journal of Intelligent Engineering andSystems*,vol.10,pp.146–154,062017.

[17] J. Kordonis, S. Symeonidis, and A. Arampatzis, "Stock price forecastingviasentimentanalysisontwitter,"in*The20thPanhellenicConferenceonInformatics(PCI'16)*,112016.

[18] "Nationalstockexchangedata-quandlapi,"https://www.quandl.com/data/NSE-National-Stock-Exchange-of-India.

[19] "pandas:PythonDataAnalysisLibrary,"Online,2012.[Online].Available:http://pandas.pydata.org/

[20] F.Pedregosa,G.Varoquaux,A.Gramfort,V.Michel,B.Thirion,O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander-plas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch-esnay, "Scikit-learn: Machine learning in Python," *Journal of MachineLearningResearch*,vol.12,pp.2825–2830,2011.

[21] L. Breiman, "Stacked regressions," *Machine Learning*,vol.24,no.1,pp.49–64,Jul1996.[Online].Available:https://doi.org/10.1023/A:1018046112532

[22] S.Raschka,"Mlxtend:Providingmachinelearninganddatascienceutilitiesandextensionstopythonsscientificcomputingstack,"*TheJournalofOpenSourceSoftware*,vol.3,no.24,Apr.2018.[Online].Available:http://joss.theoj.org/papers/10.21105/joss.00638

[23] B.Pang,L.Lee,andS.Vaithyanathan,"Thumbsup?:Sentimentclassificationusingmachinelearningtechniques,"in*ProceedingsoftheACL-02ConferenceonEmpiricalMethodsinNaturalLanguageProcessing-Volume10*,ser.EMNLP'02.Stroudsburg,PA,USA:AssociationforComputationalLinguistics,2002,pp.79–86.[Online]. Available:https://doi.org/10.3115/1118693.1118704

[24] “Tweepy: An easy-to-use python library for accessing the twitter api.”https://tweepy.readthedocs.io/en/v3.5.0/api.html.

[25] “tweet-preprocessor,”https://pypi.org/project/tweet-preprocessor/.

[26] R.Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.