Direction Detection of Select Stocks with Machine Learning

Anand Mohan [1][0000-0002-9434-0465] Rashmi Agarwal[2] [0000-0003-1778-7519]

J.B Simha [3][0000-0002-6733-7535]

[1, 2]REVA Academy for Corporate Excellence (RACE), REVA University, Bangalore, India
[3]CTO, Abiba Systems, Bangalore, India
anand.ba06@reva.edu.in

Abstract. Several research initiatives have been taken to predict stock market returns using historical data. Investors can find plenty of algorithms that detect the exact closing price of any stock but will not tell the direction of the closing price. During this proposed work, twenty-two years' price of the stock's daily close price is being utilized for direction detection. The objective of this paper is to get the right stock, perform exploratory data analysis for data preparation and then build the right models by using multiple modelling techniques to predict whether the price will move up or move down. Closing prices are being utilized as six different feature variables for building the classification model. The difference between the seventh and eighth day closing price is determined. The 0.7%, 1%, and 1.5% differences are different classes of direction to determine either positive, negative or no change. A similar process is again repeated for the feature variable increased to ten days and fourteen days respectively. Then momentum, trend, volatility, and volume indicators are utilized as feature variables and different classification models are built to determine upward direction detection. Random forest modelling has given the highest efficiency in direction detection. Logistic regression modelling done for percentage change in close price as 0.5% has given the highest efficiency for volume and momentum indicators whereas the extreme gradient boost classifier provided the best prediction performance for trend and volatility indicators. Therefore, various classification modelling techniques had been remarkably useful in direction detection for the stock under consideration.

Keywords: Direction Detection, Stock Market, Technical Indicators, Classification Models, HDFC, KOTAK, SBI

1   Introduction

The stock market encourages the free economy concept. It is one of the significant financial tools in the hands of the corporate and enterprises to raise their funds through investments done by the common man. In return for investors putting their stake in company stocks, it is expected that they earn profits through dividends and

upward stock movements, which would also enhance their economic status apart from the growth of the participant company whose stocks are at stake in the public domain.

Live validations are still becoming a grim prospect, because of several things like value variations, quiet news, and existing noise[1]. Several machine learning associated techniques are developed which have created the potential to predict the market to an extent[2]. For the transaction of shares via a broker, there is mostly a fee paid to the broker for each buy and sale which will almost eat up the gains[3]. The requirement is to overcome the ambiguities of fundamental and technical evaluation, and advanced development in the modelling strategies has pushed several researchers to check new strategies for stock value forecasting[4].

In the next section, some of the available literature is scanned which throws light on various related aspects of machine learning methods and other methodologies, and also study and research other related issues which help assist better in direction detection in the stock market.

2 Literature Review

Literature review initially scans through technical and fundamental analysis of stocks. Further, it discusses as to how algorithmic trading based on fundamentals and technical indicators helps investors in their decision making. Further, it emphasizes the merits of machine learning and artificial intelligence over algorithmic trading. It discusses unsupervised and various supervised classification techniques used in this paper. Later it reviews the literature on confusion matrix discussing various metrics for evaluation of the modelling techniques used for this proposed work.

2.1 Fundamental and Technical Analysis

Rajkar et al. in their paper comprehensively talk about the numerous parameters impacting value movements in varied sizes and layers in the stock market [5]. Therefore, different analysis namely technical and fundamental analysis is being done to invest in stock markets.

Elbialy in his paper worked on fundamental analysis and suggest that it helps to identify and implement short positions by selling the shares of companies showing downtrends and then covering these positions by buying back the shares of these companies when they start showing upward trends [6]. Fundamental analysis helps to identify stock quality and therefore, stock technical analysis done later performs better on the strong fundamental stock.

Thanekar and Shaikh in their paper conclude their study that technical analysis can demarcate and recognize commerce openings in the stock market by examining identifiable patterns similar to volume and price action movements [7]. Kimbonguila et al. in their paper used many technical indicators like Moving Average Convergence Divergence (MACD), moving average, etc on the past costs to identify better stocks for trading purposes [8].

2.2 Algorithmic Trading

Taking the discussion further, Hansen in his paper mentions algorithmic trading which is a systematic method of trading without subjective assessment through a manual trader using computer programs. Hansen further observes that fast algorithms improve traders' ability to scan the market and seize opportunities most appropriately [9]. Mukerji et al. in the paper caution that though algorithmic trading gives better results than manual trading, regulators have restrained algorithmic trading following accusations of market manipulations [10].

2.3 Supervised and Unsupervised Learning

This paper introduces a new stock market prediction model that includes three major phases namely feature engineering, non correlated feature selections, and finally direction detection.

Omta et al. in their paper used machine learning and artificial intelligence for the analysis of image-based cellular screens. It is suggested in the paper that exploratory data analysis should be performed as an initial step to gain a better data understanding before executing machine learning algorithms. Machine learning can again be further categorized into supervised and unsupervised learning [11].

Alhomadi findings in his paper were that some literature has used both supervised and unsupervised machine learning techniques for securities market predictive modelling and located that both kinds of models will create predictions with satisfactory accuracy (Acc) [12]. Dar researched further on unsupervised machine learning techniques by deeply studying principal component analysis and suggested that the central plan of principal component analysis is to spot correlations and patterns in a dataset with high dimensionality and scale back it to a considerably lower dimension without losing any important info [13].

further, various supervised classification machine learning techniques have been used in this paper namely logistic regression, decision tree, random forest, k nearest neighbours, and extreme gradient boosting.

2.4 Classification Machine Learning Techniques

Al-Bairmani and Ismael worked further on exploring logistic regression and infers that logistic regression is used instead of linear regression in situations where the target variable is not numeric, but a nominal or an ordinal variable [14]. Jena and Dehuri suggest that the simple linear modelling algorithms become more complex as the size of the datasets increases which is being handled using more advanced algorithms in the decision tree for classification and regression problems [15]. Schonlau and Zou infer that random forest modelling is quite flexible to non-linearity in the dataset and is the most appropriate ensemble learning algorithm for medium-sized to very large-sized datasets [16]. Wang studied k nearest neighbours and informs that it is the most popular statistical technique utilized in pattern identification over the last four decades [17]. Zhang et al. researched on extreme gradient boost which according to him is extensively recognized as an extremely useful ensemble learning algorithm. However, its performance needs more improvements ideally in scenarios where the dataset is imbalanced [18].

2.5 Confusion Matrix for Classification Models

Various classification algorithms as discussed have to be built for the data. Subsequently, all these algorithms have to be tested. The confusion matrix for classification models is a step in that direction. Markoulidakis et al. in their paper evaluate numerous performance metrics which include Acc, precision (Pr), and recall (Rcl) [19].

3 Methodology

Initially fundamental and technical analysis of the stocks under consideration is performed to demonstrate why a particular stock dataset has been used for this proposed work. The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework has been used in this paper. In data understanding, the different feature variables used for the proposed work are being studied and their univariate analysis is performed. Based on the data understanding phase, various steps are being taken in the data preparation phase namely handling missing values, features addition, and data scaling using the Minmax scaler. Once the data has been prepared, different modelling algorithms are implemented on them namely logistic regression, decision tree, random forest, k nearest neighbour, and extreme gradient boost classifiers. The data evaluation phase further examines the results of different modelling techniques which are used in the data modelling phase. Deployment speaks about developing a front end Application Programming Interface (API) for the deployment dashboard.

3.1 Data Collection

Various classification algorithms as discussed have to be built for the data. Subsequently, all these algorithms have to be tested. The confusion matrix for classification models is a step in that direction. Markoulidakis et al. in their paper evaluate numerous performance metrics which include Acc, Pr, and Rcl [19].

Daily trading data of HDFC, KOTAK, and SBI banks from the year 2000 to 2022 are being used for this study. This study uses National Stock Exchange (NSE) data.

The *symbol* column tells the corporate symbol mentioned for the stock. The *opening price* is the first trade worth that is recorded throughout the day's trading. The *high and low* is the highest and lowest value respectively at that a stock is listed during a period.

The *previous closing* is going to be a consecutive session's opening price. The *last price* is the one at which the foremost recent transaction happens. The close is the last value recorded once the market is closed on the day. The *Volume Weighted Average Price* (VWAP) is a trading benchmark based on both volume and worth. The trading *volume* shows the number of shares listed for the day, listed in lots of hundreds of quantities of shares.

Table 1 discusses details for every column used in the HDFC, KOTAK, and SBI datasets.

Table 1.Top rows of HDFC, KOTAK, and SBI stock dataset

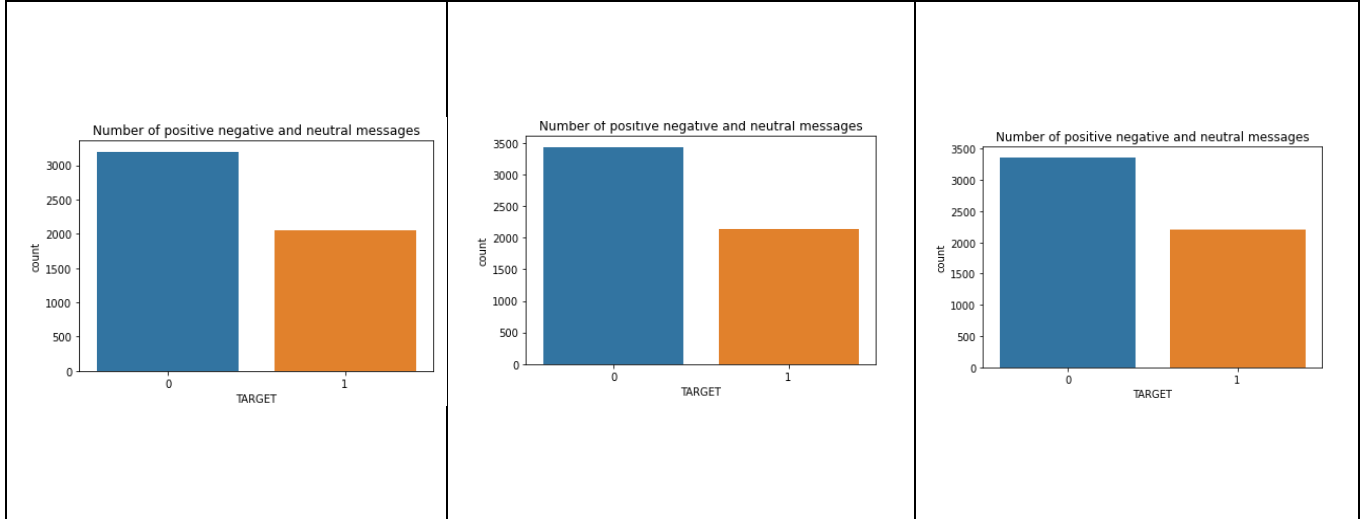| Date	Symbol	Prev Close	Open	High	Low	Last	Close
1/3/2000	HDFC	272	294	294	294	294	294
5/30/2022	HDFC	2330	2368	2388	2362	2367	2367
1/3/2000	KOTAK	212	220	229	220	229	229
5/30/2022	KOTAK	1946	1945	1952	1896	1907	1903
1/3/2000	SBI	226	236	244	234	244	244
5/30/2022	SBI	469	473	477	471	475	475

6

3.2   Data Exploration

Fig.1. Class Distribution for HDFC, KOTAK, and SBI stock

As shown in Fig. 1, HDFC STOCK is moving 2140 times in an upward direction whereas 3435 times, it is not moving in an upward direction. KOTAK STOCK is 2055 times suitable for long trading whereas 3199 times, it is not moving in an upward direction. SBI STOCK is 2211 times suitable for long trading whereas 3364 times, it is not moving in an upward direction.
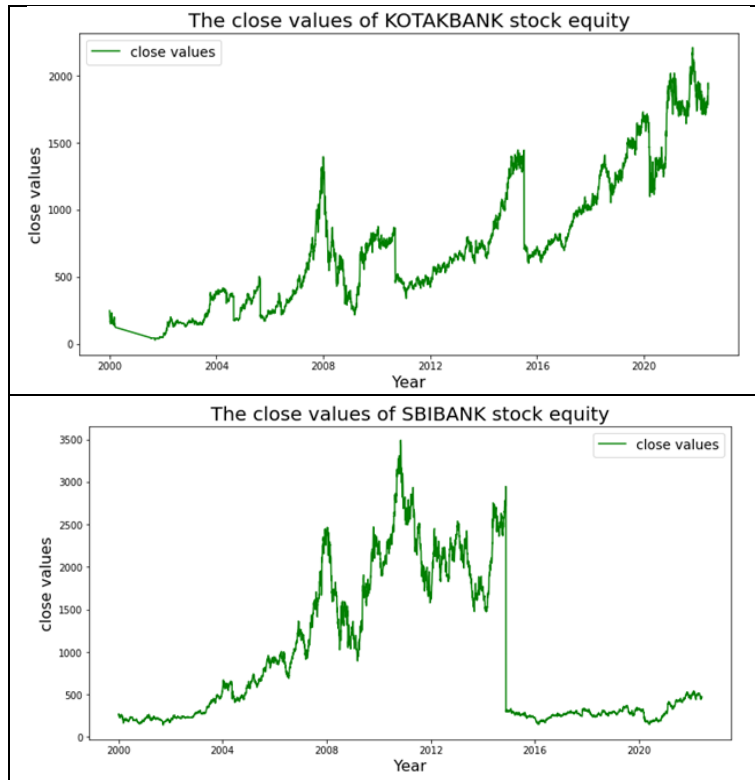
Fig.2. Close values of HDFC, KOTAK, and SBI stock from 2000 to 2022
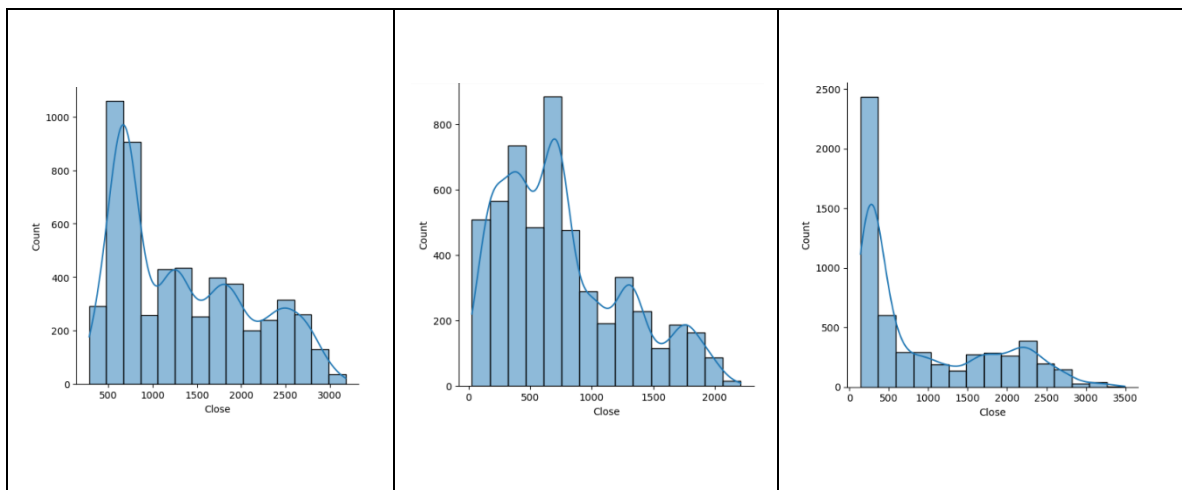

Fig.3. Distribution Plot for the HDFC, KOTAK, and SBI Stock

As shown in Fig. 2 –Fig. 3, the data has a positively skewed distribution which is observed in all 3 stocks namely HDFC, KOTAK, and SBI bank stock. SBIBANK stock is looking as the least volatile stock followed by HDFC and then KOTAK.
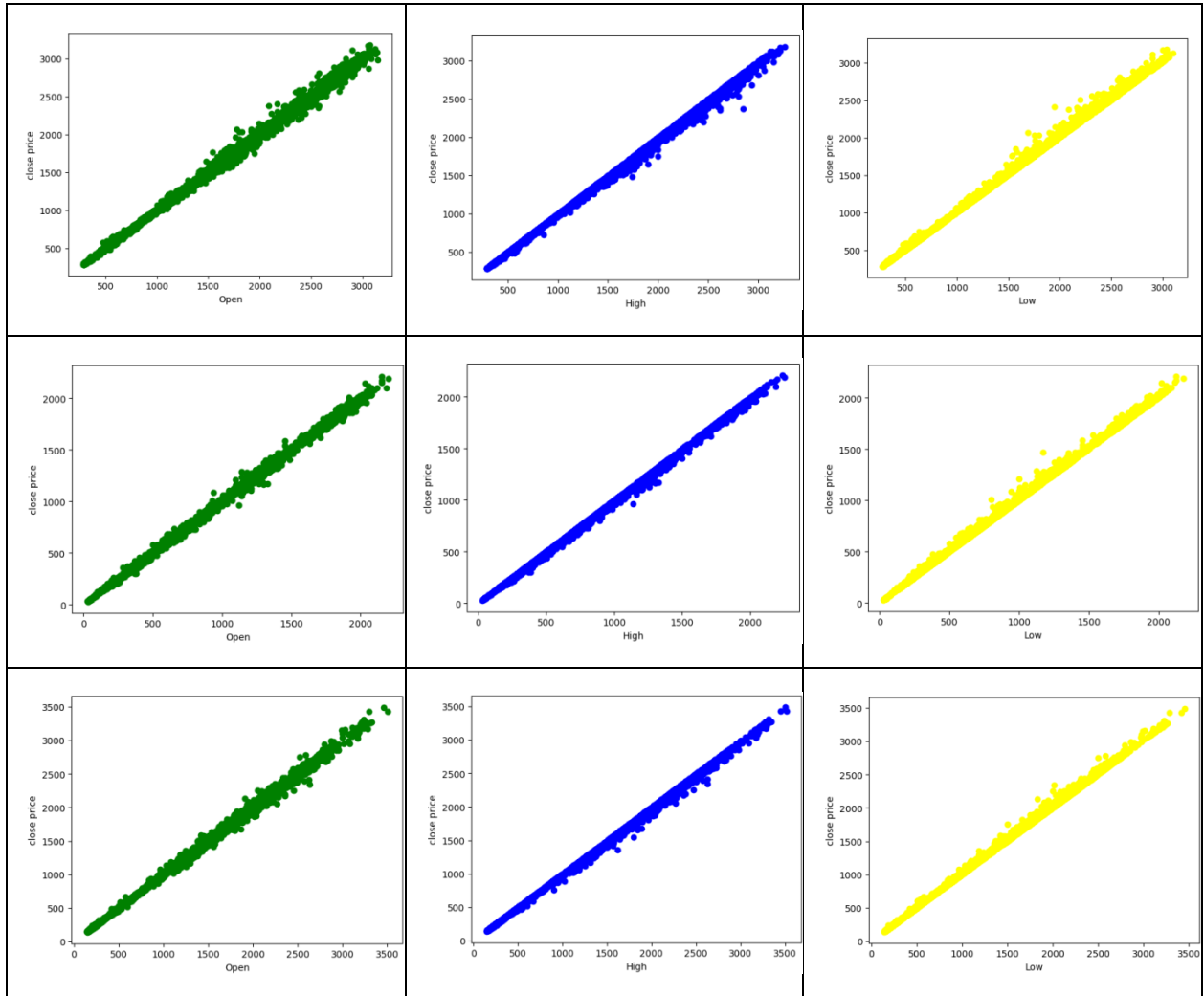

Fig.4. Scatter Plot against close price for HDFC KOTAK, and SBI Stock from 2000 to 2022

As shown in Fig. 4, a linear relationship exists between Independent variables and the target variable except for fewer outliers which is quite negligible.

3.3   Data Pre-Processing

The HDFC, KOTAK, and SBI data which are taken from NSE come with a lot of limitations that have to be processed.

3.3.1   Handling Missing Values

Three of the features' trades, 'deliverable volume', and'% deliverable are dropped as they are having several missing values.

3.3.2   Features Addition

Computed variables added to the dataset are simple and exponential moving averages for rolling periods of seven, thirteen, twenty, hundred, and two hundred days. The one day's previous lag values of volume are also added as features. Six, ten, fourteen, and thirty days' consecutive closing prices are tabulated week on week for the entire dataset and utilized as different feature variables. Momentum, trend, volatility, and volume indicators are also used as feature variables.

3.3.3   Data Scaling

Minmax Scaler is the data scaling approach that is being used. MinMax Scaler shrinks the data inside the given range, from zero to one.

3.4   Data Modeling

Based on direction detection Acc, it can be suggested to the prospective investor whether to invest or not invest in stock. Direction prediction Acc is further determined using momentum, trend, volatility, and volume indicators as feature variables and building different classification models on them. Table 2 explains the modelling strategies and model evaluation Rule used for this paper.

Table 2. Modelling Strategies and Model Evaluation Rule

Modelling Strategies	Model Evaluation Rule
Direction detection by six, ten, and fourteen days consecutive closing prices split week on the week.	percentage change on closing price>0.7% =>Positive Trend percentage change on closing price<-0.7% =>Negative Trend percentage change on closing price between 0.7 and 0.7% =>Neutral
Go long direction prediction is performed separately using momentum, trend, volatility, and volume indicators.	percentage change on closing price>0.5% =>Positive Trend percentage change on closing price<=0.5% =>Not Positive Trend

4   Implementation and Results

The data evaluation phase is the result of the data modelling phase and discusses the metrics utilized to determine the extent of the success achieved by the different modelling algorithms employed on the target variable.

4.1   Model Evaluation using Logistic Regression Classifier for Go Long Direction Prediction

Various classification models are utilized to predict the direction of the close value of HDFC, KOTAK, and SBI stock and estimate using different error metrics. All the results derived from the various models are examined below.

Table 3. Model evaluation using logistic regression classifier

Modelling Strategies	HDFC	KOTAK	SBI
Direction detection by six, ten, and fourteen days consecutive closing prices split week on week	Pr-0.35 Rcl-0.60 Acc-0.35	Pr-0.37 Rcl-0.74 Acc-0.36	Pr-0.36 Rcl-1.00 Acc-0.36
Go long direction prediction using volume indicators	**Pr-0.98 Rcl-0.83 Acc-0.92**	**Pr-0.99 Rcl-0.93 Acc-0.97**	**Pr-0.92 Rcl-0.80 Acc-0.90**
Go long direction prediction using momentum indicators	Pr-0.71 Rcl-0.63 Acc-0.76	Pr-0.73 Rcl-0.61 Acc-0.75	Pr-0.69 Rcl-0.62 Acc-0.74
Go long direction prediction using trend indicators	Pr-0.83 Rcl-0.59 Acc-0.80	Pr-0.76 Rcl-0.48 Acc-0.72	Pr-0.78 Rcl-0.49 Acc-0.74
Go long direction prediction using volatility indicators	Pr-0.93 Rcl-0.47 Acc-0.77	Pr-0.90 Rcl-0.40 Acc-0.74	Pr-0.81 Rcl-0.30 Acc-0.70

From Table 3, it is observed that go long direction prediction using volume indicators has given considerable Pr, Rcl, and Acc in direction prediction.

4.2 Model Evaluation using Random Forest Classifier for Go Long Direction Prediction:

Table 4. Model evaluation using random forest classifier

Modelling Strategies	HDFC	KOTAK	SBI
Direction detection by six, ten, and fourteen days consecutive closing prices split week on the week	**Pr-0.85** **Rcl-0.89** **Acc-0.87**	**Pr-0.71** **Rcl-0.79** **Acc-0.74**	**Pr-0.83** **Rcl-0.88** **Acc-0.85**
Go long direction prediction using volume Indicators	**Pr-0.91** **Rcl-0.82** **Acc-0.90**	**Pr-0.92** **Rcl-0.79** **Acc-0.89**	**Pr-0.90** **Rcl-0.73** **Acc-0.86**
Go long direction prediction using momentum indicators	Pr-0.76 Rcl-0.51 Acc-0.75	Pr-0.79 Rcl-0.46 Acc-0.74	Pr-0.72 Rcl-0.55 Acc-0.74
Go long direction prediction using trend indicators	Pr-0.87 Rcl-0.56 Acc-0.80	Pr-0.87 Rcl-0.55 Acc-0.79	Pr-0.83 Rcl-0.57 Acc-0.78
Go long direction prediction using volatility indicators	Pr-0.89 Rcl-0.50 Acc-0.77	Pr-0.89 Rcl-0.50 Acc-0.78	Pr-0.83 Rcl-0.61 Acc-0.80

From Table 4, it is observed that direction detection has given the highest Pr, Acc, and Rcl in prediction. Also, go long direction prediction using volume indicators has given considerable Pr and Acc in direction prediction but Rcl can still be improved.

4.3 Model Evaluation using Extreme Gradient Boost Classifier for Go Long Direction Prediction

Table 5. Model Evaluation using Random Forest Classifier

Modelling strategies	HDFC	KOTAK	SBI
Direction detection by six, ten, and fourteen days consecutive closing prices split week on the week	Pr-0.35 Rcl-0.42 Acc-0.40	Pr-0.38 Rcl-0.41 Acc-0.40	Pr-0.38 Rcl-0.47 Acc-0.37
Go long direction prediction using volume indicators	**Pr-0.90** **Rcl-0.73** **Acc-0.86**	**Pr-0.92** **Rcl-0.90** **Acc-0.93**	**Pr-0.88** **Rcl-0.82** **Acc-0.89**
Go long direction prediction using momentum indicators	Pr-0.70 Rcl-0.61 Acc-0.75	Pr-0.75 Rcl-0.62 Acc-0.77	Pr-0.70 Rcl-0.59 Acc-0.74
Go long direction prediction using trend indicators	Pr-0.85 Rcl-0.74 Acc-0.85	Pr-0.82 Rcl-0.61 Acc-0.79	Pr-0.83 Rcl-0.67 Acc-0.81
Go long direction prediction using volatility indicators	Pr-0.86 Rcl-0.75 Acc-0.85	Pr-0.81 Rcl-0.63 Acc-0.79	Pr-0.80 Rcl-0.67 Acc-0.81

From Table 5, it is observed that go long direction prediction using volume indicators has given considerable Pr, Rcl, and Acc in direction prediction.

4.4 Direction Detection and Go Long Direction Prediction using the Best Classifier Model

Table 6. Leader board comparison of metrics for direction detection and go long direction prediction using the best classifier model

Modelling Strategies	HDFC	KOTAK	SBI
Direction detection by six, ten, and fourteen days consecutive closing prices split week on the week(random forest classifier)	**Pr-0.85** **Rcl-0.89** **Acc-0.87**	**Pr-0.71** **Rcl-0.79** **Acc-0.74**	**Pr-0.83** **Rcl-0.88** **Acc-0.85**
Go long direction prediction using volume indicators(logistic regression classifier)	**Pr-0.98** **Rcl-0.83** **Acc-0.92**	**Pr-0.99** **Rcl-0.93** **Acc-0.97**	**Pr-0.92** **Rcl-0.80** **Acc-0.90**
Go long direction prediction using momentum indicators(logistic regression classifier)	Pr-0.71 Rcl-0.63 Acc-0.76	Pr-0.73 Rcl-0.61 Acc-0.75	Pr-0.69 Rcl-0.62 Acc-0.74
Go long direction prediction using trend indicators (extreme gradient boost Classifier)	Pr-0.85 Rcl-0.74 Acc-0.85	Pr-0.82 Rcl-0.61 Acc-0.79	Pr-0.83 Rcl-0.67 Acc-0.81
Go long direction prediction using volatility indicators (extreme gradient boost classifier)	Pr-0.86 Rcl-0.75 Acc-0.85	Pr-0.81 Rcl-0.63 Acc-0.79	Pr-0.80 Rcl-0.67 Acc-0.81

From Table 6, it is observed that random forest classifier modelling has given the highest efficiency in direction detection among all modelling techniques namely logistic regression, decision tree, random forest, k nearest neighbour, and extreme gradient boost modelling. This has been tested and proven with six, ten, and fourteen day consecutive closing prices split week on week as six, ten, and fourteen feature variables. Also, logistic regression classifier modelling has provided the best Pr, Rcl, and Acc for go long direction prediction using volume indicators.

4.5 Utility from the Business Perspectives

For a stop loss of 2.0 reward-risk ratio for approximately 0.8 Pr would be 2*.8/2*.2=4:1 if a 0.5% difference in consecutive day close price for any stock is only 2.0.for higher percentage difference reward to risk ratio would be higher.

Here, modelling algorithms provide the close price of HDFC BANK, KOTAK BANK, and SBI BANK Stock over twenty years with the train test split of 70%:30%. If we invest Rs.10000 for six years and roughly calculate profit with 0.5% change on close price with the highest Pr in detecting true positives then the following results are possible as per the formulae given in Eq. (1):

$$n = number\,of\,days\,it\,is\,true\,positive$$
$$m = number\,of\,days\,it\,is\,false\,positive$$
$$pr = precision$$
$$p = percentage\,change$$
$$c = capital\,invest$$
$$netreturns = (p*c*n*pr - p*c*m*pr)/100 \qquad (1)$$

Using trend indicators with the highest Pr of 0.85 for HDFC BANK stock, the confusion matrix provides information as shown in Fig. 5.

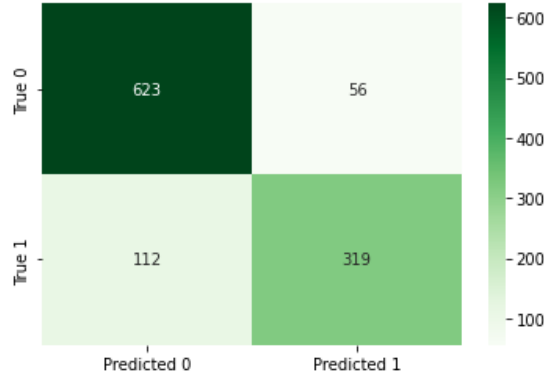**Fig.5.** confusion matrix For HDFCBANK stock using trend indicators as feature variables

Therefore, Net Returns are:

$$0.5*10000*319*0.85/100 - 0.5*10000*56*0.85/100$$
=Rs.11177.5 profit which would be
$$11177.5/(10000*6)*100$$
=18.63% returns

4.6   Risk Adjusted Returns

The real data dump is imported for HDFC, KOTAK, and SBI stock between 2000 till 2022. Then the return, variance, and volatility of these stocks are calculated following which the annualized return to risk ratio and finally, the Sharpe ratios are calculated. The Sharpe ratio for HDFC, KOTAK, and SBI Stock is calculated as

0.173818, 0.149589, and 0.005306 respectively. Therefore, from the results obtained it becomes evident that HFDC shows a better Return vs. Risk performance over the specified period compared to KOTAK stock followed by the SBI stock which shows the least Return vs. Risk performance.

5   Conclusion and Future Scope

This paper solely focuses on predicting the direction of the close price of the HDFC stock using classification algorithms techniques. Later similar process is applied for predicting the direction of the close price of other stocks in the banking sector namely SBI and KOTAK stocks. In the future, there is a deployment dashboard proposed. As per the proposal for future assignments, the dashboard takes API as an input derived from the machine learning algorithms and can be utilized in predicting the direction of the close price for any stock in the banking sector. Any stock on the stock market can utilize the same procedure to forecast buy or not to buy choices, which is helpful.

This paper has not discussed how to address one major drawback of stock prediction, namely that over different periods the stock returns can change drastically. In future research work, it can be shown how to define bullish and bearish regimes using modern machine learning techniques. The sentiment analysis approach may also need to be explored using text analytics for predicting stock market returns. An intelligent automated system for options trading would be also the next step forward.

References

[1]   D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," *Int. J. Financ. Stud.*, vol. 7, no. 2, 2019, doi: 10.3390/ijfs7020026.

[2]   P. Sonkiya, V. Bajpai, and A. Bansal, "Stock price prediction using BERT and GAN," 2021, [Online]. Available: http://arxiv.org/abs/2107.09055

[3]   Y. Huang, L. F. Capretz, and D. Ho, "Machine Learning for Stock Prediction Based on Fundamental Analysis," *2021 IEEE Symp. Ser. Comput. Intell. SSCI 2021 - Proc.*, 2021, doi: 10.1109/SSCI50451.2021.9660134.

[4]   N. Rouf *et al.*, "Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions," *Electron.*, vol. 10, no. 21, 2021, doi: 10.3390/electronics10212717.

[5]   A. Rajkar, A. Kumaria, A. Raut, and N. Kulkarni, "Stock Market Price Prediction and Analysis," *Int. J. Eng. Res. Technol.*, vol. 10, no. 06, pp. 115–119, 2021.

[6]   B. A. Elbialy, "The Effect of Using Technical and Fundamental Analysis on the Effectiveness of Investment Decisions of Traders on the Egyptian Stock Exchange," *Int. J. Appl. Eng. Res.*, vol. 14, no. 24, pp. 4492–4501, 2019,

[Online]. Available: http://www.ripublication.com

[7]   G. S. Thanekar and Z. S. Shaikh, "Analysis and Evaluation of Technical Indicators for Prediction of Stock Market," *Int. J. Eng. Res. Technol.*, vol. 10, no. May, pp. 341–344, 2021.

[8]   A. Kimbonguila, L. Matos, J. Petit, J. Scher, and J.-M. Nzikou, "Effect of Physical Treatment on the Physicochemical, Rheological and Functional Properties of Yam Meal of the Cultivar 'Ngumvu' From Dioscorea Alata L. of Congo," *Int. J. Recent Sci. Res.*, vol. 10, pp. 30693–30695, 2019, doi: 10.24327/IJRSR.

[9]   K. B. Hansen, "The virtue of simplicity: On machine learning models in algorithmic trading," *Big Data Soc.*, vol. 7, no. 1, 2020, doi: 10.1177/2053951720926558.

[10]   P. Mukerji, C. Chung, T. Walsh, and B. Xiong, "The Impact of Algorithmic Trading in a Simulated Asset Market," *J. Risk Financ. Manag.*, vol. 12, no. 2, p. 68, 2019, doi: 10.3390/jrfm12020068.

[11]   W. A. Omta *et al.*, "Combining Supervised and Unsupervised Machine Learning Methods for Phenotypic Functional Genomics Screening," *SLAS Discov.*, vol. 25, no. 6, pp. 655–664, 2020, doi: 10.1177/2472555220919345.

[12]   A. Alhomadi, "Forecasting stock market prices : A machine learning approach," *Digit. Commons*, vol. 11, no. 2, pp. 16–36, 2021.

[13]   A. N. Dar, "PRINCIPAL COMPONENT ANALYSIS (PCA) (Using Eigen Decomposition)," *Gsj*, vol. 9, no. 7, pp. 240–252, 2021, [Online]. Available: www.globalscientificjournal.com

[14]   Z. A. A. Al-Bairmani and A. A. Ismael, "Using Logistic Regression Model to Study the Most Important Factors Which Affects Diabetes for the Elderly in the City of Hilla / 2019," *J. Phys. Conf. Ser.*, vol. 1818, no. 1, 2021, doi: 10.1088/1742-6596/1818/1/012016.

[15]   M. Jena and S. Dehuri, "Decision tree for classification and regression: A state-of-the art review," *Inform.*, vol. 44, no. 4, pp. 405–420, 2020, doi: 10.31449/INF.V44I4.3023.

[16]   M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[17]   L. Wang, "Research and Implementation of Machine Learning Classifier Based on KNN," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 677, no. 5, pp. 0–5, 2019, doi: 10.1088/1757-899X/677/5/052038.

[18]   P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.

[19]   I. Markoulidakis, G. Kopsiaftis, I. Rallis, and I. Georgoulas, "Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem," *ACM Int. Conf. Proceeding Ser.*, pp. 412–419, 2021, doi: 10.1145/3453892.3461323.