IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

Ensemble Learning for Stock Market Prediction: Leveraging Uncertainty in Decision Tree-Based Methods

Dr. Sachin S Agrawal[1], Dr. Pravin R. Satav[2] and Mr. Bhushan Talekar[3]
Assistant Professor, Department of Computer Science and Engineering[1]
Lecturer, Department of Computer Engineering[2]
Research Scholar, Department of Computer Science and Engineering[3]
College of Engineering and Technology, Akola, Maharashtra, India[1,3]
Government Polytechnic, Amravati, Maharashtra, India[2]
sachin.s.agrawal@gmail.com and prsatav@gmail.com

Abstract: *This papers delves into the critical parameters of decision tree-based ensemble models, such as the number of trees (T), which necessitates careful user selection. We investigate the impact of T on model performance, aiming to determine whether setting it to the largest computationally manageable value is optimal or if a smaller T might yield better results with proper tuning. Furthermore, we extend traditional decision tree classifiers to accommodate uncertain data, a common occurrence in real-world scenarios due to factors like measurement errors, data staleness, and multiple measurements. Our research addresses the uncertainty problem associated with decision tree-based methods, providing valuable insights into its consequences and implications. The research context revolves around stock market prediction, a challenging and dynamic domain crucial for financial decision-making. Stock market behavior is influenced by various macroeconomic factors, making precise predictions complex. This study categorizes stock price prediction into trend classification and price forecasting and emphasizes the significance of incorporating fundamental analysis methods to forecast stock prices accurately. Additionally, the role of news in stock market prediction is explored, highlighting the challenges posed by unstructured news data. In summary, this work contributes to the understanding of ensemble methods, decision tree-based algorithms, and their application in stock market prediction, while also addressing the novel challenge of uncertainty in predictive modeling.*

Keywords: Rotation Forest, Random Forest Algorithm

I. INTRODUCTION

Combining several base learners generally has better results than using only one learner. The main reason is that combination or averaging reduces the variance of the model. There is increasing popularity of ensemble because of their superior performance. In several experimental works, they have been named as the most accurate algorithms. Ensemble performance depends ontwo main properties: 1) the individual success of the base learners and 2) the independence of thebase learners' results from each other (low error, high diversity). It is possible to build diverse base learners by using the same or different type of base learners. When the same type of base learners are used, the diversity is created by using different training data set for each base learner in the ensemble. There are several methods for creating different training data sets such as bagging, boosting, random subspaces, random forests, rotation forest, and extremely randomized trees. The existing ensemble methods create different training data sets by deleting (bagging) or weighting samples or deleting (random subspace, random forest, extremely randomized trees) or rotating (rotation forest, random projections) features. If an ensemble algorithm generates training sets those are very similar to the original training set, the average individual accuracy of base learnerswould be high, but their decisions would be similar (low diversity). Based on these observations, adding new and hopefully, more accurate features to the original data set are proposed to obtain.

Since its introduction, Random forests have enjoyed much success as one of the most widely used decision tree based methods in machine learning. But despite their popularity and apparent simplicity, random forests have proven to be very difficult to analyze. Indeed, many of the basic mathematical properties of the algorithm are

still not completely well understood, and theoretical investigations have often had to rely on either making simplifying assumptions or considering variations of the standard framework in order to make the analysis more tractable. One advantage of decision tree based methods like random forests is their ability to natively handle categorical predictors without having to first transform them (e.g., by using feature engineering techniques). However, we would like to show how this capability can lead to an uncertainty problem for decision tree based methods that have, to the best of our knowledge, never been thoroughly discussed, and whose consequences have never been carefully explored.

The random forest algorithm for classification and regression, which is based on the aggregation of a large number T of decision trees. T is one of several important parameters whichhave to be carefully chosen by the user. Some of these parameters are tuning parameters in the sense that both too high and too low parameter values yield sub-optimal performances for an earlystudy on the effect of such parameters. It is unclear, however, whether the number of trees T shouldsimply be set to the largest computationally manageable value or whether a smaller T may be sufficient or in some cases even better, in which case T should ideally be tuned carefully.

Traditional decision tree classifiers work with data whose values are known and precise. We would like to extend such classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Example sources of uncertainty include measurement/quantization errors, data staleness, and multiple repeated measurements. With uncertainty, the value of a data item is often represented not by one single value, but by multiple values which creates uncertainty.

1.1 Background

Stock market prediction is a significant task for the financial decision-making process and investment. Even though stock price prediction is a key problem in the financial world, it contributes to the growth of efficient methods for stock exchange transactions. Generally, stock markets are in the form of non-stationary, non-linear and uncertain even so financial experts recognized it is complex to produce precise predictions. Stock market prediction is a challenging job due to its high dynamic and unstable. Stock market prediction plans to compute the future value of a company stock trade on exchange as well as consistent prediction of future stock prices obtains high profits to investors. Various researches applied numerical data and news for the prediction of the stock market. Commonly, based on the number of information sources, the stock market prediction technique is experimented on selecting the numerical data by analysing the news data (Liu and Wang, 2018). In basic, forecasting behaviours are separated into three levels, such as short, medium and long. Furthermore, stock market movements are influenced by various macroeconomical aspects, like bank exchange rate, commodity price index, investors' expectations, bank rate, general economic conditions, investor's psychology, firms' policies, institutional investors' choices, political events and so on (Patel et al., 2015). Additionally, stock value indices are computed using higher market capitalization stocks, whereas several technical parameters are also employed to obtain statistical information about stock price values (Patel et al., 2015). In the stock market, there are two assumptions for predicting stock price value. The first one is EMH stating at any time, stock price completely confines all identified information about stock where all identified information's are utilized through market participants and also random price variations obtains new random information's. Therefore, stock prices execute a random walk, that is every future price does not follow any patterns or trends. This assumption deduces fluctuations, so incomplete or delayed information controls the stock market prices. In addition, an exterior incident influences successive stock market prices, although the precise prediction of a stock price is complex. From the prediction perception, it can be categorized into two types, namely stock price trend and stock price forecast. The stock price trend is also named as classification, and stock price forecast is also termed as regression (Ballings et al., 2015; Yuan et al., 2020). Basically, the time duration for stock price trend prediction is highly related with previously selected features (Yuan et al., 2020). The prediction of stock market future price is very significant for investors, because of the identification of suitable movement of stock price decreases the risk of future trend calculation. The industry, economy and other correlated features are considered to compute the intrinsic value of a company, which helps to forecast stock prices from fundamental analysis method. Stock market decision-making technique is a very complex and significant job because of unstable and complex nature of the stock market. It is necessary to discover a huge quantity of valuable information created through the stock market. In addition, every investor has an imminent requirement for identifying future behaviours of stock prices. Although, it helps the investors to achieve the best profit by identifying the best moment to sell or buy stocks. Normally, trading in stock market can be performed electronically or physically. The investor becomes the owner or partnership of a particular company, while an

investor obtains a particular company share. Furthermore, financial data of the stock market is very complex in nature, so for predicting stock market behaviour is also complex. The stock market prediction helps the investors to take investment decisions by offering strong insights regarding stock market behaviour for reducing investment risks. Consequently, news plays a high authority on stock price behaviour (Zhou et al., 2018). Stock market prediction using news mining is a recent attractive research area and it has various challenges due to its unstructured form of news. Conventionally, stock market prediction highly depends on historical stock data. The variety of technical indicators is extracted from historical data for predicting stock market trend. Moreover, financial news articles are the most significant element of market information for predicting the stock market. Normally, the prediction outcomes are evaluated using two elements, like Root Mean Square Error (RMSE) or Root Mean Square Relative Error (RMSRE). The developments in performance prediction lead to very much gainful for investors. On the other hand, economists established another assumption based on the advancement of behavioural, intelligence and computational finance, termed Inefficient market Hypothesis (IMH) (Zhou et al., 2018). Accurate prediction of variable stock price refers to sustainable realization and acquires good amount of profitable shares for shre traders. The pragmatic analysis performed on United States of America (USA) stock market has displayed that statistical importance in return rate is identified by several data analysis. The above statement is offered as verification of stock market prediction prospect (Kim and Park., 2009). The present researchers proved that the probable for predicting stock price with high probability based on current deep learning method is better than artificial intelligence technology. The stock market prediction has two types in economics, namely fundamental and technical analysis (Kim, 2012). Recently, various novel approaches are developed to predict and model the stock market together with nonlinear and linear approaches. Moreover, rational precise prediction of stock market movement has high possibility to obtain the best financial gains (Booth et al., 2014; Nayak et al., 2019).

1.2 Motivation

Due to the lack of timing, the technical experts defect the current prediction and transaction. Hence it is very difficult to consider the prediction time, generalization, accuracy, and confidence of the signal prediction. Some of the major issues in dealing with the time-series data are classification, clustering, dimensionality reduction, rule discovery, rule summarization, and pattern discovery. Also, the prediction is affected due to the global trends, local trends, and noise which is the detrending methods required cleaning the data for further applications

1.3 Problem Statement

Till date many classifiers are proposed which work with data whose values are known and precise, but they fail to handle data with uncertain information. The uncertainty arises in many applications during the data collection process, including measurement errors/quantization errors, data staleness, and multiple repeated measurements and this uncertainty leads to poor performanceof the predictor.

The diversity is generally created by random selection of features. But this procedure leads to less accurate base learners. Many researchers have proposed the strategies which were focusing on improving the performance of Random Forest. Obviously, the latency could not be improved while processing huge data generated by ubiquitous sensing node in the era without new technology help.

Random Forest is been mainly been used for binary classification and regression but not implemented for multi-categorical variables.

The major problem of the Random Forest algorithm is that redundant nodes not only hurt the performance of the building phase but also reduce the accuracy of results. In this, we will study onTuning of parameters (number of trees, maximum features, and minimum sample leaf for considering split point) and try to optimize the performance of ensemble algorithm.

Random Forest is a hybrid model comprises of many base learners which come together for improving the prediction accuracy as a whole, but due to not considering the errors, missing data and redundant data during the training of base learners the resultant ensemble predictor results in poor performance. In the bootstrap process, the data set provided in the base learners' results in the error propagation in an exponential manner, which dramatically reduces the prediction performance of the final model.

1.4 Objectives

This research work will try to achieve some or all of the following objectives:

- To explore Uncertainty present in the dataset and to trained base learners by considering the uncertainty present in the data sets.
- To prove tuning of the parameters mathematically.
- To create the diversity for each base learner using different transformation techniques for improving the prediction performance.
- To check the performance of the proposed technique with respect to a different application domain by applying to a large number of data sets from the public database open ML.
- To provide a theoretical prove in favor of setting T for a computationally feasible large number as long as learning of the model is improved.

II. LITERATURE REVIEW

Jin Young Choi, Rhee Man Kil and Chong-Ho Choi , introduces the piecewise linear regression network, a novel method for function approximation based on piecewise linear regression technology (PLRN). The PLRN is intended to accomplish three things: 1) reduce the challenge brought on by the high dimensional settings of the provided data; 2) do away with the requirement of creating ordered topological maps as in the case of traditional piecewise linear approximation techniques; and 3) achieve fast learning without being constrained to local minima of an error surface. The PLRN is used to forecast Mackey-Glass chaotic time series and is compared to alternative ways to demonstrate the efficacy of our method.

T. Jinyu and Z. Xin, This article uses multiple linear regression to create prediction models for the audit opinion and then performs an effectiveness test. It chose 30 businesses as samples from the Shanghai and Shenzhen stock markets. The findings demonstrate that the analytical model has a greater accuracy rate, excellent interoperability, and offers a fresh perspective on how to forecast the audit opinion.

Y. E. Cakra and B. Distiawan Trisedya, The demand for a stock heavily influences its price, and no one factor can reliably predict that demand each day. This makes stock price prediction a challenging endeavour. Efficient Market Hypothesis (EMH), however, asserted that stock price also greatly relied on new information. The opinions of users on social media are one of many information sources. The reputation of a company may be influenced by the public's perception of its goods, which may then influence the public's decision to purchase the firm's shares. Making an appropriate analysis of opinion is crucial when using it as primary data. Sentiment analysis is one well-known instance of using opinion as data. Sentiment analysis is a method for identifying the feelings or emotions people have about something, in this case, the products of some businesses. Research has been done on the use of sentiment analysis to forecast stock prices. According to Bollen's research, Twitter users' opinions may accurately anticipate DJIA value with an accuracy rate of 87.6%. This demonstrates that stock prices and sentiment analysis are related. In this study, we aim to forecast the Indonesian stock market using straightforward sentiment analysis. To categorise tweets and determine the sentiment against a corporation, the Naive Bayes and Random Forest algorithms are utilised. The outcomes of sentiment analysis are applied to forecast the price of a company's shares. To create the prediction model, we employ the linear regression technique. Our study demonstrates that hybrid features and prior stock prices used in prediction models provide the best predictions, with coefficients of determination of 0.9989 and 0.9983, respectively.

Kavitha S, Varuna S and Ramya R, Consumer interest, behaviour, and product revenues are the business insights needed to forecast the future of the industry using recent or historical data. With the help of statistical approaches, these insights can be produced for forecasting purposes. Depending on the needs of the data, the statistical methods can be assessed for the prediction model. Time series data are frequently used in forecasting and prediction. For greater accuracy, most applications including weather forecasting, finance, and stock market mix historical data with the most recent streaming data. Regression models are used to examine the time series data, though. In order to choose the best model for improved prediction and accuracy, the training data set is used in this study to evaluate the linear regression and support vector regression models.

A.Izzah, Y. A. Sari, R. Widyastuti and T. A. Cinderatama, The development of stock prediction comes from the fields of data mining and economics. Due to their significance in developing a more effective and efficient strategy, stock projections have received special attention. In this work, a mobile application based on the Android platform was constructed to predict stock price using Improved Multiple Linear Regression (IMLR). IMLR is a combination of multiple linear regression and moving average. The requirements analysis, system design, implementation, and testing phases of the app's development are broken down into several parts. Data were automatically gathered by utilising the Yahoo Finance API from the finance.yahoo.com website with the

category "Jakarta Composite Index (A JKSE)". Users of this software could view real-time stock price predictions in addition to daily stock history. The MSE, RMSE, and MAPE values for the mobile app accuracy prediction are 15087.465 in MSE, 122.831 in RMSE, and 3.255 in MAPE, which is a superior result than the conventional algorithm.

F. Mar'i, U. Pratiwi, I. Oktanisa and F. Utaminingrum, One of the few difficult problems to resolve is the forecast of stock indexes across several nations. The presence of stock indices provides information about the state of the market in a nation, including Indonesia. Predicting the company's future worth is vital due to a country's significant stock index. In this study, the stock index is predicted using the Jakarta Islamic Index (JII). We presented a Multiple Linear Regression as a method with coefficient determination using many numerical approaches, such as the Gauss-Jordan method, Gaussian elimination, and Cramer's rule, to predict the stock indeks value. The system of linear equations involving multiple linear regression is solved using a variety of numerical techniques with the goal of determining the most effective numerical technique. When comparing the three methods for solving a system of linear equations, the Mean Absolute Percentage Error (MAPE) is employed. The less the MAPE value, the better the performance of the prediction models. According to the test results, Gauss Elimination and Cramer's rule approach yield the lowest MAPE error values, respectively, of 0.43% and 0.44%, whereas Gauss-Jordan yields MAPE of 0.83%.

B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, Since its inception, the stock market has demonstrated the effects of both high and low prices. It is the pinnacle of all financial activity and trade. When the Dow Jones Industrial Average dropped 777.68% in 2008, the stock market meltdown revealed to the world that business had reached its lowest point. These stock prices can be forecasted using a number of machine learning techniques, and these algorithms may be applied using the supervised learning method. We have test data for supervised learning, and we train the models using this data. Despite the possibility that the outcomes from training the model may be different from the actual, it has been noted that accuracy is frequently acceptable. The initial aim in this article is to get datasets from stock data using web scraping. The data is then plotted on a graph, from which we may determine if stock prices are rising or falling. After that, we will use SVM and linear regression to predict stock prices, with linear regression being superior to SVM in stock market analysis.

M. Lutfi, S. P. Agustin and I. Nurma Yulita, Stock predictions have been made by researchers in a variety of ways. One investment option that is appealing to investors is stock. However, investors and researchers (academics) are quite curious about how investors might forecast stock prices in the future because of the possibility of relatively substantial stock price fluctuations. There are two different categories of factors that influence stock price changes. Both internal and external variables exist. In this study, linear regression, SMO regression, and random forest algorithms are used to forecast stock prices. Then contrast the three models to determine which one is the best at forecasting stock prices. The findings demonstrate that SMOReg, which has the minimum error value among other models and a metric with an MAE number of 7.4758, MSE 157.0944, RMSE 12.5337, and MAPE 0.524, is the most effective model in this study. This indicates that the model is trustworthy enough to be applied to stock price forecasting. SMOReg still has issues with examining sequential data and overfitting, though.

C. Ebenesh and K. Anitha, The suggested framework's main goal is to predict new stock price movements and anticipate trends in changes to the stock market price index for a limited number of three equities. This approach aims to create two classification groups based on linear regression (LR) and long short-term memory (LSTM) (LS TM). In terms of each model's capacity to forecast the movement of the Bombay Stock Exchange (BSE) index, their performances are contrasted. With an anticipated sample size of 30, the proposed framework has been evaluated for the prediction of three stocks (AAPL, MSFT, and AMZN). The average performance of the LR model (98.2%) is determined to be considerably better than that of the LS TM model (72.3%) with (p0.05), according to experimental data. In comparison to the LS TM model, the LR model showed promising performance, according to the evaluation of stock index prediction parameters.

2.1 Logistic Regression

P. V. Sairam and L. K, In this work, an unique long short term memory algorithm (LSTM), which is compared to the Logistic Regression technique, is used to provide a comparative study of enhanced F1 score in stock market values. Materials and Procedures To increase F1 score for forecasted stock market values, novel long short term memory (N=10) and logistic regression technique (N=10) were iterated. To optimise pH, two strategies are simulated by changing the NLSTM and logistic regression parameters. For two groups, the sample

size is estimated using Gpower 80%, and 20 samples were employed in this study. Results and Discussion: LSTM's accuracy percentage (68.24%) is significantly higher than that of logistic regression (53.71%) with a 0.407 ($p>0.05$). In order to increase F1 score, long short term memory algorithms aid in automatic stock market price prediction.

B. Kumar Jha and S. Pande, In a variety of problem domains, including sales, finance, healthcare, the stock market, etc., forecasting techniques are applied. The time-series dataset contains information about time that is helpful for statistical analysis and forecasting. The supermarket sales forecast aids in increasing sales in a professional setting. The method aids in problem-domain decision-making. For predicting, a variety of methods are available, including the regression model and the logistic exponential model. The most recent instrument to demonstrate enhanced effectiveness in terms of forecast accuracy is the Facebook (FB) Prophet. For the purpose of predicting sales based on data from supermarkets, this research has suggested a technology called FB Prophet. A few forecasting models, including the additive model, the Autoregressive Integrated Moving Average (ARIMA) model, and the FB Prophet model, have been explored in the suggested research effort. FB Prophet is a better prediction model in terms of low error, better prediction, and better fitting, according to the proposed research work.

H. R. Putri and A. Dhini , Due to its strong correlation to profitability, the difficulties and rivalry in the investment industry have recently attracted a lot of attention. It is generally acknowledged that a significant decline in a company's profitability is closely related to its financial hardship, which results in difficulty in meeting its financial obligations. Previous study used traditional statistical approaches to construct a financial distress prediction model that has drawbacks because it is heavily dependent on several constrictive assumptions. This study looked at both financial and non-financial characteristics to predict financial distress, and it used data mining techniques because of their superiority with less restrictive assumptions. Focusing on listed companies in the financial and infrastructure sectors on the Indonesia Stock Exchange (IDX) over a 4-year period, ensemble classifiers, logistic regression, and C4.5 decision trees have all been constructed and evaluated. Given that boosting outperformed other methods in prior research, the decision tree with boosting model did the best. It had the lowest error rates and maximum accuracy, and its overall accuracies, sensitivity, and specificity were 94.61%, 94.6%, and 94.5%, respectively. The findings of this study also provide a number of inferences, such as the fact that return on assets is the most significant or crucial predictive factor in identifying financially distressed enterprises.

U. Ananthakumar and R. Sarkar , Predicting the future value of a company's shares or another financial instrument traded on an exchange is referred to as stock market prediction. This study uses a data mining technique to try to solve this unpredictability even though accurately predicting a stock's future price is a difficult occurrence that would still result in a sizable benefit. The firms listed on the BSE SENSEX are selected for this study as a representative group of those that are traded the most. In order to determine which ratios are crucial and how they affect stock prices, logistic regression is employed on a variety of essential financial ratios of these companies as well as a few macro financial variables. When compared to a related study in the literature, the proposed model yields greater classification accuracy.

Z. Jiang and Y. Chen Investors find it challenging to make stock market predictions. The processing of massive data is now possible thanks to advancements in computer science and technology. Data mining is one example that has seen extensive use. The most crucial tool for gathering market data is a search engine in the interim. As a result, stock trading could potentially have a search behaviour component. Therefore, in this study, we attempt to forecast stock volatility using the Baidu Index (BDI). This study developed a logistic regression model to investigate the link between searched terms and stock volatility using the strong R language and data mining methods. We suggested a straightforward dictionary for this use, which is the paper's only tiny contribution.

F. Yang, H. Yang and M. Yang Stock market price manipulation practises have significantly skewed stock prices and jeopardised the interests of small and medium-sized investors. The discrimination against and prevention of such practises is crucial for the stock markets' future development. By using the Shanghai and Shenzhen stock markets as examples and significant indicators of stock price manipulation based on Primary Component Analysis, this paper establishes a logistic regression model for the discrimination of stock price manipulations after analysing the characteristics of stock price manipulations. In comparison to earlier models, this one has a higher level of relevance and a higher rate of prediction success.

Q. Li and W. Shi Since the Shanghai and Shenzhen stock exchanges were created thirty years ago, our capital market has flourished and offers the listed company solid assistance. However, as the market economy's

competitive environment heats up, more and more publicly traded enterprises are being exposed to significant financial risk. This circumstance not only interferes with the capital market's normal development, but it also causes significant financial losses for all parties involved, including creditors and investors. In this article, fifteen common financial indices are chosen and the listed firms in the biochemical industry are categorised. In this work, the factor analysis is first used to analyse the sample, after which the scores of the five principal components are computed. Finally, logistic regression is used to examine the factor scores. To study financial crisis prediction, a logistic regression model is built. This model's accuracy, which is roughly 74 34 percentage points, indicates that it might be utilised to study the financial crisis in this sector.

III. EXISTING METHODOLOGY

Stock market analysis or prediction is an interesting research topic in the area of investment. The effective prediction of the stock market provided more profit besides reducing the risk rate. Hence, it is important to analyse the future direction of stock market with good accuracy. This section discussed the literature review of the existing stock market analysis and prediction methods. The stock market prediction methods are streamlined such as Statistical method, Machine learning method, Pattern recognition method, Hybrid method and Sentiment analysis methods. Also in stock market analysis work research gaps and the issues encountered are discussed.

3.1 Prediction:

Dataset :

Context: Stock market data is widely analyzed for educational, business and personal interests.

Content: The data is the price history and trading volumes of the fifty stocks in the index NIFTY 50 from NSE (National Stock Exchange) India. All datasets are at a day-level with pricing and trading values split across .cvs files for each stock along with a metadata file with some macro-information about the stocks itself. The data spans from 1st January, 2000 to 30th April, 2021.

Feature extraction: Appropriate features are extracted following the recording of the data that represents the problem. It is a vital step as it permits to measure or compute features that might contain information concerning the process status. Briefly, a feature extraction scheme calculates various metrics reflecting exclusive features in the collected data. Obtaining descriptors that better illustrate the issue is the major objective. The feature extraction process provides output as a structured table generated by feature columns. Every row is a pattern, with an extra random column representing each sample's current position (usually called a label or class).The patterns are not labeled when the status is not known.

Feature selection and reduction: This step makes use of either feature selection or feature reduction schemes to treat resultant attributes to obtain less space or a set of new features. This is a voluntary process that allows to select or reduce the number of features extracted. Feature reduction is for creating new features using the original features, whereas feature selection is for finding a reduced set of attributes that better defines a procedure. These steps are intended to reduce issues, e.g., time expenditure and the obscenity of size and so on. These methods are usually classified into Filters, Wrappers and Embedded Schemes, which in turn can be devised by machine learning algorithms.

Prediction: A novel dataset is generated from the original dataset on the basis of selected attributes. The offline run makes the utilization of the new dataset for developing build models using which classification and regression tasks can be performed among other things. The Algorithm Selection block includes procedures and techniques for selecting the most adequate ML (machine learning) model. This approach is extensively executed for discovering various solutions with the implementation of several ML models. For a variety of ML methods, it is essential to discover the best model for classifying the prediction.

3.2 Machine Learning based Prediction Approaches

Prediction based on machine learning algorithms has attracted research interest in view of its expected accuracy and efficiency. AI and Machine learning algorithms involve various steps when adopting supervised learning. First, prediction features that represent the attributes of the flows (e.g., packet length) are identified. Second, the machine learning model is constructed. Third, the classifier is trained to associate specific features with known prediction classes. Finally, the model is applied to classify data prediction, predicting the classes in prediction flow.

4.3.1 Logistic Regression

Logistic regression is a classification technique borrowed by machine learning from the field of statistics. Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The intention behind using logistic regression is to find the best fitting model to describe the relationship between the dependent and the independent variable.

IV. PROPOSED METHODOLOGY

Stock market analysis or prediction is an interesting research topic in the area of investment. The effective prediction of the stock market provided more profit besides reducing the risk rate. Hence, it is important to analyze the future direction of stock market with good accuracy. This section discussed the literature review of the existing stock market analysis and prediction methods. The stock market prediction methods are streamlined such as Statistical method, Machine learning method, Pattern recognition method, Hybrid method and Sentiment analysis methods. Also in stock market analysis work research gaps and the issues encountered are discussed.

The prediction of the stock market is a challenging task in day-to-day lives. One of the major problems associated with the technical techniques is “self-destructing”. The opportunity will go away from the traders if one understands the profitable trading strategy and the traders choose the same buy or sell action that avoids copying some successful strategy (Qian and Rasheed, 2007). Simple voting and stacking ensemble methods failed to operate as a result of greater correlations of predictions between classifiers (Enke et al., 2011). The drawback of Generative Topographic Maps (GTM) is that the number of categories employed for discretization was restricted and the number of the categories changes based on the nature of the individual features. Also, Deep learning algorithms yielded poor performance in case of the higher dimensional data and large window sizes (Singh and Srivastava, 2017). There exist several aspects in the prediction that involves physiological factors, physical factors, rational and 39 irrational behaviour and so on. All these factors are combined for making share prices volatile and complex to predict with high accuracy and less error.

4.1 Proposed Methodology : Random Forest

It is one of the powerful supervised learning algorithm, which can perform both regression and classification problems. This is a combination of multiple decision tree algorithms and higher the number of trees, higher the accuracy. It works as same as the decision tree, which based on information gain. In classification, each decision tree will classify the same problem and the overall decision will be calculated by considering the majority vote of the results. The most important advantage of this model is that it can handle missing values and able to handle large datasets.

V. RESULTS AND DISCUSSION

The number of K trees to make a random forest and the number of F randomly chosen features to build a decision tree are the algorithm's two important factors. A large K and F should be used for huge and high dimensional data. Induce an accurate generalising function from a set of labelled training cases is the aim of supervised machine learning. The examples in a data set are often treated identically since, in most situations, everything that is initially known about a task is contained in the set of training instances. To induce a model of the data, some cases are more helpful than others. For instance, outliers or incorrectly categorised instances are frequently less advantageous than boundary instances. Additionally, even if other cases are appropriately identified and do not constitute outliers, they may still be harmful for developing a model of the data (Michael R. Smith and Tony Martinez, 2014).

The creation of a random forest involves noisy and outlier data because of the intricacy of the data distribution in a high-dimensional future space. As a result, each tree in the forest will act less predictably. In order to improve the prediction performance by lowering the error rate in the Random Forest, this research aims to optimise sample selection by probability proportional to size sampling (weighted sampling), in which the noisy and outlier observations are down weighted

5.1 Data Visualization

ISSN PRINT 2319 1775 Online 2320 7876
  UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

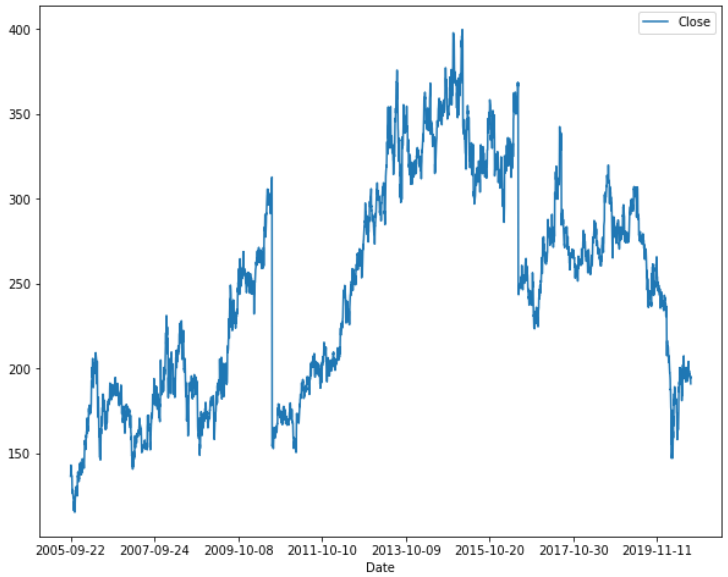

Figure 5.1. Data Visualization

5.2 Model Training and Testing

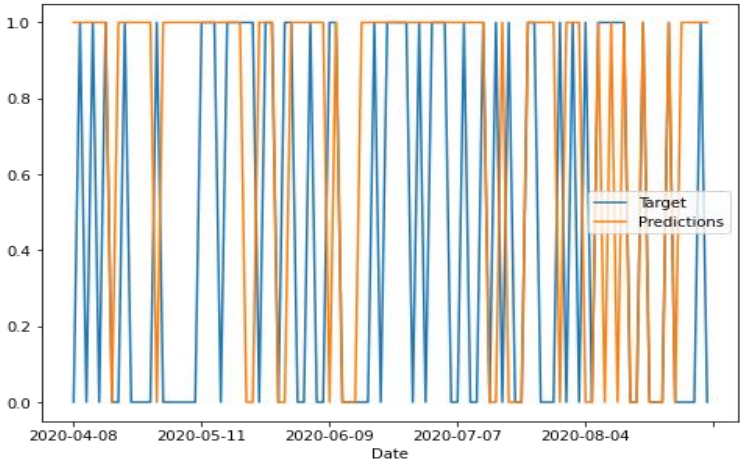

Figure 5.2 Model Training

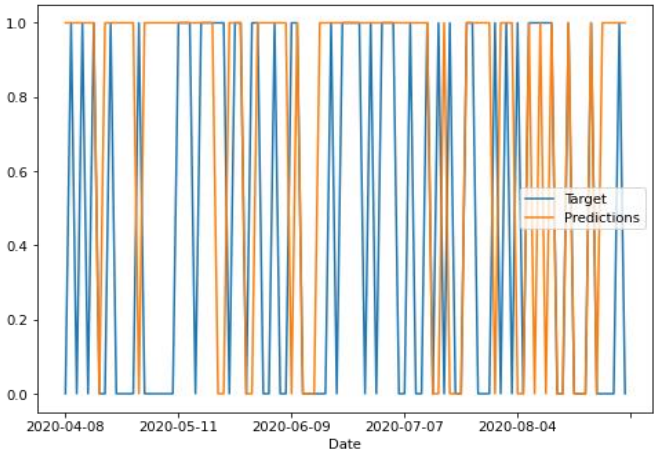

Figure 5.3 Model Testing

ISSN PRINT 2319 1775 Online 2320 7876
  UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, 2022

5.3 Prediction

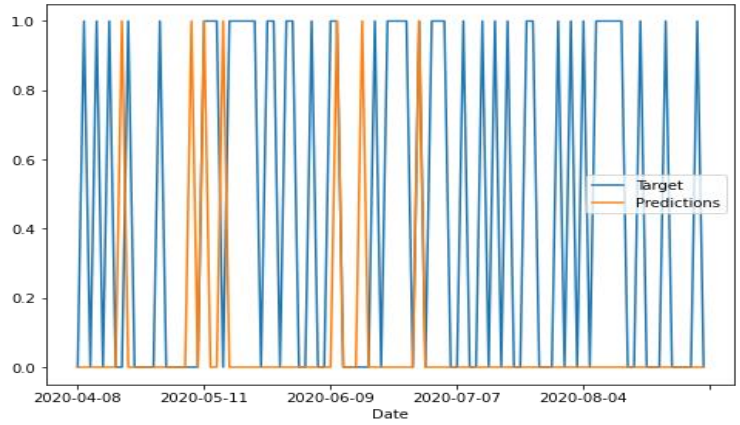

Figure 5.4 . Prediction Result

5.4 Confusion Matrix

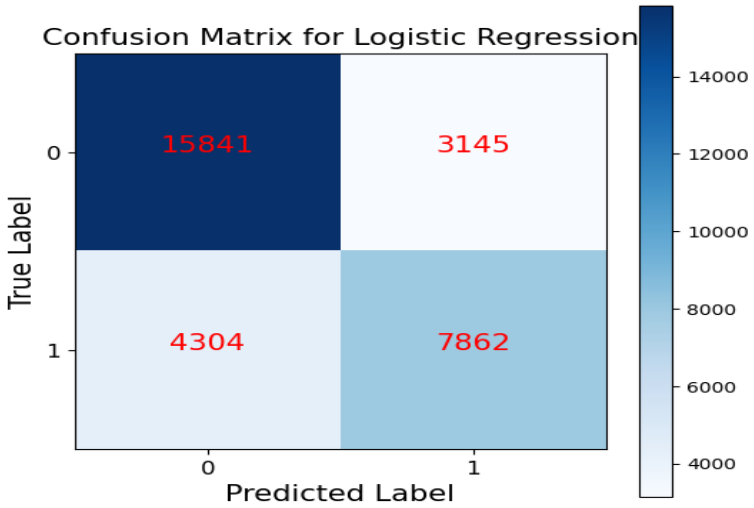

Figure 5.5. Confusion Matrix for Logistic Regression

	precision	recall	f1-score	support
0	0.79	0.83	0.81	18986
1	0.71	0.65	0.68	12166
accuracy			0.76	31152
macro avg	0.75	0.74	0.74	31152
weighted avg	0.76	0.76	0.76	31152

Table 5.6. Result Parameters for Logistic Regression

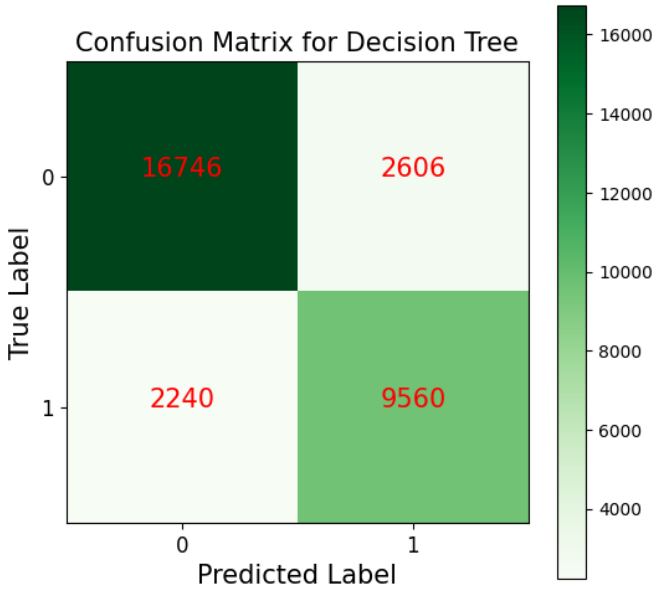

Figure 5.7 . Confusion Matrix for Decision Tree

	precision	recall	f1-score	support
0	0.87	0.88	0.87	18986
1	0.81	0.79	0.80	12166
accuracy			0.84	31152
macro avg	0.84	0.83	0.84	31152
weighted avg	0.84	0.84	0.84	31152

Table 5.8 . Result Parameters for Decision Tree

VI. CONCLUSION

Up until recently, stock market forecasting was a mystery to academics. Theoretically, there are a number of ways to predict future prices with acceptable error terms, but in practise, none of these approaches are workable due to future uncertainty and some elements that cannot be accounted for by mathematical models. Thus, a mystery is created. As a result, financial forecasting is constantly difficult for researchers who value creativity and innovation. This study attempted to build a financial forecasting model utilising artificial intelligence technologies like ANN, GA, etc. with a constructive attitude toward this problem. To determine whether ANN is appropriate for stock market prediction in general and in the Indian context specifically, the model's performance data is examined.

Our research is mostly focused on predicting financial time series. We have conducted analysis utilising stock price data. Using only the past price value of the same stock, we tried to predict the future price of the stock. Unquestionably, such an effort is mostly intended to test a poor version of market efficiency. However, in order to assess the predictive power of forecasting models, this study primarily focused on error minimization elements of those models. We thought that while analysis might not be able to totally eliminate risk, it might be able to lessen it for investors.

The ability to predict stock prices is essential for making investments and financial decisions. However, because the stock market is so unpredictable, investing there carries a significant risk. Numerous studies have been conducted to forecast the market in order to generate revenue using a variety of methodologies, including fundamental analysis, statistical analysis, and technical analysis, with varying degrees of success. However, these methods do not offer the thorough analysis that is necessary, making it ineffective for predicting stock values.

Predicting stock market values has always been challenging. It is believed that the company's stock price is not largely influenced by the nation's economic situation and is not directly linked to the economic growth of the

nation overall or of a particular region. As a result, stock price forecasting is now more difficult than ever. Politics, business news, natural disasters, and other factors can all have an impact on the stock value on a given day. The stock prices have changed quickly as a result of the advanced technology's quick data processing of the events. As a result, huge investors, stock brokers, and financial institutions must acquire and sell equities as quickly as possible. By capturing its nonlinear behaviour, the most recent developments in data mining techniques offer useful tools for forecasting chaotic situations like the stock market.

The ability to predict stock prices is essential for making investments and financial decisions. However, because the stock market is so unpredictable, investing there carries a significant risk. Numerous studies have been conducted to forecast the market in order to generate revenue using a variety of methodologies, including fundamental analysis, statistical analysis, and technical analysis, with varying degrees of success. However, these methods do not offer the thorough analysis that is necessary, making it ineffective for predicting stock values. Predicting stock market values has always been challenging. It is believed that the company's stock price is not largely influenced by the nation's economic situation and is not directly linked to the economic growth of the nation overall or of a particular region. As a result, stock price forecasting is now more difficult than ever. Politics, business news, natural disasters, and other factors can all have an impact on the stock value on a given day. The stock prices have changed quickly as a result of the advanced technology's quick data processing of the events. As a result, huge investors, stock brokers, and financial institutions must acquire and sell equities as quickly as possible. By capturing its nonlinear behaviour, the most recent developments in data mining techniques offer useful tools for forecasting chaotic situations like the stock market.

As a result, the main contribution of this study effort is on modelling the sampling mechanism, and choosing the training and testing datasets was a challenging task in stock market forecasting. The random forest machine learning method, which performs well at stock market forecasting, served as the foundation for the development of the stock market prediction model.

REFERENCES

[1]. Jin Young Choi, Rhee Man Kil and Chong-Ho Choi, "Piecewise linear regression networks and its application to time series prediction," Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), 1993, pp. 1349-1352 vol.2, doi: 10.1109/IJCNN.1993.716793.

[2]. T. Jinyu and Z. Xin, "Apply multiple linear regression model to predict the audit opinion," 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, 2009, pp. 303-306, doi: 10.1109/CCCM.2009.5267661.

[3]. Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015, pp. 147-154, doi: 10.1109/ICACSIS.2015.7415179.

[4]. Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear regression and support vector regression," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.

[5]. A.Izzah, Y. A. Sari, R. Widyastuti and T. A. Cinderatama, "Mobile app for stock prediction using Improved Multiple Linear Regression," 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017, pp. 150-154, doi:10.1109/SIET.2017.8304126.

[6]. F. Mari, U. Pratiwi, I. Oktanisa and F. Utaminingrum, "Comparative Study of Numerical Methods in Multiple Linear Regression For Stock Prediction Jakarta Islamic Index (JII)," 2019 International Conference on Sustainable Information Engineering and Technology (SIET), 2019, pp. 110-115, doi: 10.1109/SIET48054.2019.8985999.

[7]. B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 629-631, doi: 10.1109/ICACITE51222.2021.9404733.

[8]. M. Lutfi, S. P. Agustin and I. Nurma Yulita, "LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, And Random Forest," 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 1-5, doi:10.1109/ICAIBDA53487.2021.9689749.

[9]. C. Ebenesh and K. Anitha, "A Novel Approach to Minimize the Mean Square Error in Predicting Stock Price Index using Linear Regression in Comparison with LSTM Model," 2022 International

Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 1365-1370, doi: 10.1109/ICSCDS53736.2022.9760764.

[10]. P. V. Sairam and L. K, "Automatic Stock Market Prediction using Novel Long Short Term Memory Algorithm compared with Logistic Regression for improved F1 score," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), 2022, pp. 578-582, doi: 10.1109/ICIPTM54933.2022.9754116.

[11]. B. Kumar Jha and S. Pande, "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 547-554, doi: 10.1109/ICCMC51019.2021.9418033.

[12]. H. R. Putri and A. Dhini, "Prediction of Financial Distress: Analyzing the Industry Performance in Stock Exchange Market using Data Mining," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), 2019, pp. 1-5, doi: 10.1109/ICSSSM.2019.8887824.

[13]. U. Ananthakumar and R. Sarkar, "Application of Logistic Regression in Assessing Stock Performances," 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2017, pp. 1242-1247, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.199.

[14]. Z. Jiang and Y. Chen, "BDI based stock prediction," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), 2016, pp. 119-122, doi:10.1109/ICOACS.2016.7563061.

[15]. F. Yang, H. Yang and M. Yang, "Discrimination of China's stock price manipulation based on primary component analysis," 2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014), 2014, pp. 1-5, doi:10.1109/BESC.2014.7059519.