SUPERVISED MACHINE LEARNING TECHNIQUES FOR STOCK MARKET ANALYSIS

K. Praveen Kumar[1] , Dr. Subhash Pamera[2]

[1] Assistant Professor, Department of Business Management, Aurora's PG College (MBA),
Uppal, Hyderabad
Email: praveen.kumar9065@gmail.com

[2] Assoc. Professor, Department of Business Management, Aurora's PG College (MBA),
Uppal, Hyderabad
Email: pamerasubhashdr@gmail.com

ABSTRACT

This research investigates a hybrid model for stock price trend prediction that combines a probabilistic strategy with a K-Nearest Neighbors (KNN) approach. The assumptions made by distance functions pose a major challenge to KNN classification. The closest neighbors, or the centroid of the test cases' data points, are the subject of the assumptions. When predicting stock price trends, this method leaves out non-centric data items that may be statistically significant. In order to do this, an improved model that combines KNN with a probabilistic technique that computes probabilities for the target instances using both centric and non-centric data points must be built. The Bayes theorem is the foundation of the embedded probabilistic approach. The joint probability, which is derived from the probability of the event of the nearest neighbors and the event of prior probability happening simultaneously and at the same time that they are calculated, serves as the basis for the prediction outcome. One Rule (OneR), Zero Rule (ZeroR), Naive Bayes, and KNN are examples of traditional classifiers that were evaluated with the suggested hybrid KNN probabilistic model. According to the test findings, the suggested model performed better than the benchmark classifiers that were employed in the comparisons.

Index Terms: - Probabilistic Method, Bayes' Theorem, Naive Bayes, K-Nearest Neighbors, Stock Price Prediction

I INTRODUCTION

Within the investing world, analyzing financial data in securities has proven to be a significant and difficult problem. Due to the competing effects of information rivalry among significant investors and the adverse selection costs imposed by their knowledge advantage, achieving stock price efficiency for publicly traded companies is challenging. When it comes to financial market analysis, there are two primary schools of thinking. The initial strategy is referred to as fundamental analysis. Fundamental analysis employs an approach that measures a stock's intrinsic value using both qualitative and quantitative examination. This method looks at the management, industry, micro and macroeconomic aspects, and financial reporting of a corporation. Technical analysis is the name for the second strategy. Technical analysis employs historical market data analysis as an approach for predicting price direction. Technical analysis makes use of a range of charts to predict future events. The OHLC (open-high-low-close) charts, mountain charts, point-and-figure charts, line charts, bar charts, and candlestick charts are among the stock charts. You may view the price and volume charts in various time intervals. The charts employ a variety of indicators, such as momentum, trending, breakout, resistance, and support.

Different approaches to solving this kind of problem have been put forth; these include approaches based on computational intelligence and machine learning as well as conventional statistical modeling. Vanstone and Tan conducted an overview of the literature about the use of soft computing in financial trading and investment. The studies were divided according to the following categories: classification, hybrid approaches, time series, optimization, and pattern recognition. According to the report, the majority of research being done in the area of financial trading discipline is focused on technical analysis. By concentrating on macro-economic study, an integrated fundamental and technical analysis model was looked at to assess the stock price patterns. In relation to the economy, it also examined the company's behavior and the related industry, which gives investors additional information to consider when making investment decisions.

When the KNN approach was combined with technical analysis, the closest neighbor search (NNS) method yielded the desired outcome. Technical analysis was used by this model to analyze historical price and trading volume data from the stock market. It used RSI filters, stop loss, and gain as technical indicators. The distance function was applied to the gathered data by the KNN algorithm portion. Using the fundamental analysis method, this model was contrasted with the buy-and-hold technique.

The Fast Library for Approximate Nearest Neighbors (FLANN) is utilized to conduct the searches necessary to select the optimal approach from a library of algorithms that has been found to work well. Majhi and others. investigated the FLANN model to forecast the S&P 500 indices. Quick approximation nearest neighbor searches were used to create the FLANN model in high-dimensional regions.

The generalization capacity of artificial neural networks (ANN) is higher than that of traditional statistical methods. An artificial neural network (ANN) may deduce the features of performing stocks from past data. The data is represented by financial and technical variables. ANN is therefore employed as a statistical tool to investigate the complex correlations between the relevant technical and financial indicators and stock performance. Asset price fluctuations exhibit nonlinear regularities that can be decoded by neural network modeling. When handling the prominent characteristics of economic data, statistical inference and adjustments to conventional learning methods come in handy.

A little amount of research has been conducted utilizing both qualitative and quantitative analysis. Shynkevich et al. investigated how using concurrent, suitably weighted news pieces with varying degrees of relevance to the target stock enhanced the effectiveness of a financial forecasting algorithm. The financial model assisted traders and investors in making decisions. For the prediction system, textual pre-processing methods were used. Utilizing a multiple kernel learning approach, the stock price fluctuations were forecasted. The method combined data taken from several news categories, each of which was analyzed using a different kernel. The news stories were divided based on industry-specific characteristics and their applicability to the target stock. The health care industry's equities were used for the studies. The findings demonstrated that when data sources include more categories with a greater number of pertinent news stories, the financial forecasting model performed better. For this study, an improved model used historical pricing as well as textual and time series data to make predictions. It also included additional data sources. Different data sources can be used with additional kernels. The purpose of including new category variables was to enhance the accuracy of forecasting.

2. SURVEYS OF LITERATURE

The authors of this work, Sneh Kalra et al., conducted research in 2019 on how stock market values fluctuate in relation to a company's important new articles. In order to differentiate positive and negative utterances for prediction purposes based on daily news variance, they utilized the classifier Naïve Bayes. Future work on social media and blog data may take this technique into consideration. The 2019 paper by Aditya Menon et al. focuses on a review of a neural model for stock market forecasting. After reviewing the neural model, the authors believe that the long short-term memory algorithm for predicting economic information in conjunction with the trendy era should be given priority when it comes to forecasting. Regression analysis is mostly utilized for stock market trend prediction, according to a 2017 study by Ashish Sharma et al. that examined regression techniques for stock prediction using stock market data. Additional numerical variables could be used in the future to improve the outcome. According to Andrea Picasso et al. (2019), this study's authors combined economic and fundamental analysis with a variety of automation techniques and applications to predict market trends. Neural networks are a machine learning technique that addresses the issue of trend stocks, and the charts in question contain forecasted data. The sentiment of a news story is used as input data. Their investigation indicates that the most troublesome achievement in using information about astral one-off news is this one. The appropriate feature fusion technique will be appropriate in the future to solve this issue.

In 2018, Gangadhar Shobha et al. published a paper that gave readers a comprehensive overview of machine learning techniques. The author covered three different types of machine learning techniques as well as a variety of metrics, including accuracy, precision, RMSE, quintile of errors, recall, and confusion matrix. Since most individuals are confused about whether to utilize machine learning techniques for prediction or other purposes, the author believes that this overview can be helpful to those who are new to machine learning. In Suryoday Basak et al.'s 2018 paper, the authors developed an experimental framework for predicting stock prices, regardless of price movements. The authors used two algorithms, known as a random forest classifier and gradient boosted decision trees, and they obtained higher accuracy than other research papers, where the authors reported that their long-term window results for the experiment ranged from 50% to 67%. They might utilize the built-in boosted tree model for short-term data windows in the future.

According to Arash Negahdari Kia et al. (2018), numerous experiments and models have been developed for the purpose of stock prediction using historical data. For example, the author of this paper presents a HyS3 graph-based semi-supervised model and uses a network views Kruskal based graph algorithm called ConKruG. They believe that in the future, social media and Twitter data may be utilized to predict stocks with more accurate outcomes when these algorithms are applied.

In 2019, Bruno Miranda and colleagues conducted a survey of bibliographic techniques pertaining to text area research. The author's focus is on financial market value prediction, using machine learning models such as support vector machines (SVM) and neural networks, with data sets sourced from the North American market. Future research may offer

opportunities for these new models to leverage North American market data for predictive purposes.
K. The goal of Hiba Sadia et al.'s 2019 paper is to preprocess raw data before comparing the random forest and SVM algorithms. The primary goal of the authors is to determine which algorithm is best for predicting stock trends. In the end, they have identified the random forest algorithm as the best-fit algorithm for future stock forecasting. However, they believe that adding more parameters to future work will increase accuracy.

In 2019, Akash et al. published two more algorithms: "LS SVM," or least square support vector machine, and "PSO," or particle swarm optimization. The latter algorithm selects the best unbounded parameter in conjunction with "LS SVM" to minimize overfitting and some technical indicators, thereby improving the accuracy of the results. However, the suggested approach is also being contrasted with an artificial neural network model at the same time.
In a 2016 article by Aparna Nayak et al., the authors used supervised learning techniques to forecast the trajectory of the stock market. The authors used daily live data that was obtained directly from the Yahoo Finance website by the algorithm, as well as monthly forecasts.
In 2019, Mu yen chen et al. used a deep learning approach called LSTM (long short-term memory) to calculate the effect of news stories on stock prices. The authors of this study believe that their research can be used to anticipate the trend of the stock market. based prediction, while this paper's results for daily live prediction were better than those for monthly prediction. Going forward, they believe that if we take more sentiments into account for the monthly prediction, that will likewise produce the best results.
In a 2016 research, Nuno Oliveira et al. described a system for determining the value of stock prediction and microblogging data for stock prices, return indices, and more metrics similar to a portfolio. They used a lot of Twitter data for this experiment, and they combined it with data from external sources and microblogging using the Kalman filter. As a result, they discovered that the data from blogging and Twitter was relevant for forecasting, and these datasets were very helpful. By utilizing additional and diverse data, such as social media datasets and others, this outcome can be enhanced.

Regression analysis was used by Han lock Siew et al. (2017) to determine the accuracy of the stock trend forecasting. For this experiment, the authors used WEKA software, which is used for data mining and machine learning algorithms to execute them. The dataset they used included heterogeneous values and was intended to be used for handling currency values and functional ratios. For the purpose of forecasting, Bursa Malaysia provides the dataset used to calculate the movement of stocks. The more consistent ordinal format of data could potentially improve forecasting utilizing the regression method for future extensions, according to the authors' thoughts.

In 2019, Smruti Rekha Das et al. used the firefly method to forecast stock prices. The authors collected the dataset from four different websites, NSE-India, BSE, S&P 500, and FTSE. The collected dataset was then properly transformed using mathematical formulas, backpropagation, neural networks, and additional two methods for prediction, forecasting according to the time horizon of alternate days, such as one day, three days, five days, and so forth. Adding more parameters to the implemented algorithms could potentially lead to more accurate findings in future study.

Authors Dattatray P. Gandhimal et al. published a review of stock prediction techniques in 2019. The writers of this work examined approximately fifty research papers that were published in accordance with the publication years, and they recommended the most effective prediction method. According to the review, KNN and fuzzy-based techniques—which include SVM, SVR, and many more—are the best; however, these two methods can work better when employing historical data, as the authors recommended. They plan to analyze further papers in the future in order to determine the best-fit prediction algorithm. Companies that provide financial services are creating solutions that support future forecasting. Stock prediction, also known as stock market mining, is one of the many financial information sources available online that might be a useful field for study. The ability to anticipate stocks becomes more and more crucial, particularly if a variety of guidelines are developed to assist in improving investing choices across various stock markets. Shin et al. (2005) had adopted the genetic algorithm; the Korea Stock Price Index 200 (KOSPI 200) generated a large number of trading rules. Hellestrom and Homlstrom (1998) in Sweden employed a statistical analysis based on a modified kNN to ascertain where correlated areas fall in the input space in order to improve the prediction performance for the 1987–1996 period. The Weightless Neural Network (WNN) model and the Single Exponential Smoothing (SES) model were two of the methods offered by the Zimbabwe Stock Exchange to forecast stock values (Mpofu, 2004). Gavrilov et al. (2004) offered a clustering stocks strategy to group 500 Standard & Poor companies. A total of 252 numbers, including the starting stock price, were included in the data. Cao (1977) introduced a fuzzy evolutionary method to find pair relationships in stock data according to user preferences. The study created possible guidelines for mining stock-trading rules, market pairs, and stocks; it also demonstrated the usefulness of such a method for actual trading. Furthermore, kNN prediction algorithms were used in other research by Subha et al. (2012), Liao et al. (2010), Tsai and Hsiao (2010), and Qian and Rasheed (2007).

3. IMPLEMENTATION

By using KNN, a closest neighbor search (NNS) approach yielded the desired outcome. method combined with technical evaluation. Technical analysis was used in this model to analyze stock market data. which contain trade volume and price history. Technical indicators consisting of RSI, stop loss, and stop gain filters. The distance function was applied to the gathered data in the KNN algorithm portion. The buy-and-hold strategy was contrasted with this model utilizing the fundamental analysis method.

Negative aspects:

Restricted examination

Utilizing past price and trading volume data

The buy-and-hold approach

.

The suggested approach

The act of attempting to forecast the future value of a stock or other financial instrument traded on a financial exchange is known as stock market prediction. Python is the programming language used to apply machine learning to predict the stock market. In this work, we present a Machine Learning (ML) method that will learn from the stock data that is currently accessible, gain intelligence, and use that information to make precise predictions. This study employs prices with both daily and up-to-the-minute frequencies, and a machine learning technique called K-Nearest Neighbor to predict stock prices for the large and small capitalizations in the three separate marketplaces.

Benefits:

Assess the potential worth of a stock.

Accessible stock information and get intelligence analysis

Precise forecasting

both big and small capitalizations, as well as across the three marketplaces

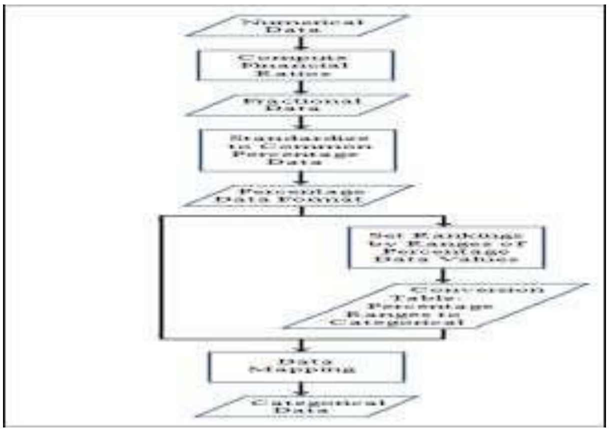
Numerical Data
Compute Financial Ratios
Fractional Data
Standardise to Common Percentage Data
Percentage Data Format
Set Rankings by Ranges of Percentage Data Values
Conversion Table: Percentage Ranges to Categorical
Data Mapping
Categorical Data

Fig 1:- proposed model

3.2.Alogirhtam

The steps adopted for classification by KNN are illustrated as follows:

Steps:

- Classification: KNN
- Initialization of k value on nearest neighbors
- Compute the distance between the X query instance and all the training samples.
- Sort the distance values
- Determine the nearest neighbors to the query instance based on the k value
- Calculate the number of Profit instances of the nearest neighbors in the vicinity of Xquery instance
- Calculate the number of Loss instances of the nearest neighbors in the vicinity of X query instance

The steps adopted for classification by the probabilistic method is illustrated as follows:Steps:

- Classification: Probabilistic method
- Calculate the prior probabilities of Profit class and Loss class from the data set
- Calculate the KNN's probabilities of Profit class and Loss class based on the numberof Profit nearest neighbors and the number of Loss nearest neighbors.
- Calculate the joint probabilities from the prior probabilities and KNN's probabilitieson Profit class and Loss class
- Compare the joint probabilities of Profit class and Loss class
- Select the predictive value from the class values with the highest joint probability

4. METHODOLOGY

Data Collection

Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we will be analyzing the live dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

Pre Processing

Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

Training the Machine

Training the machine is similar to feeding the data to the algorithm to touch up the test data. The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyperparameters like the number of nearest neighbours. We perform the entire cross-validation loop on each set of hyperparameter values. Finally, we will calculate a cross-validated score, for individual sets of hyperparameters.

Data Scoring

The process of applying a predictive model to a
set of data is referred to as scoring the data. The technique used to process the dataset is the KNN Algorithm. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

4 RESULTS AND EVOLUTION METRICS

Fig 1:. Home Screen

Fig 2: Download Dataset

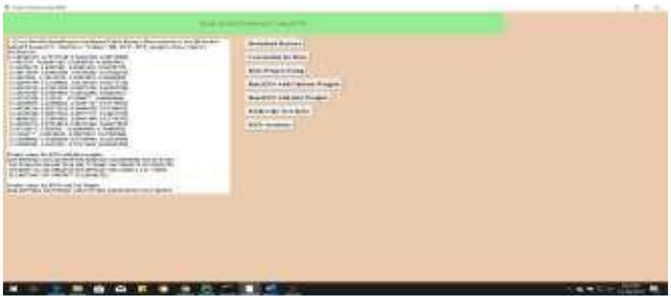

Fig 4:- Data Pre-processing

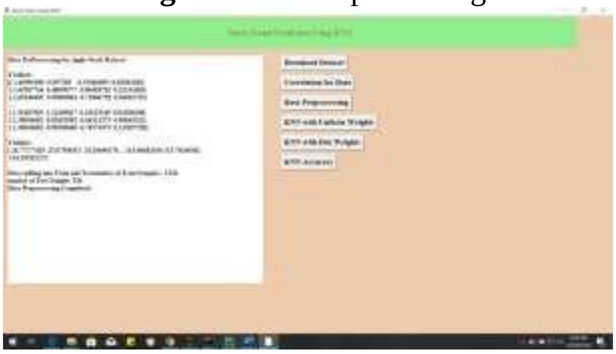

Fig 5 KNN Accuracy

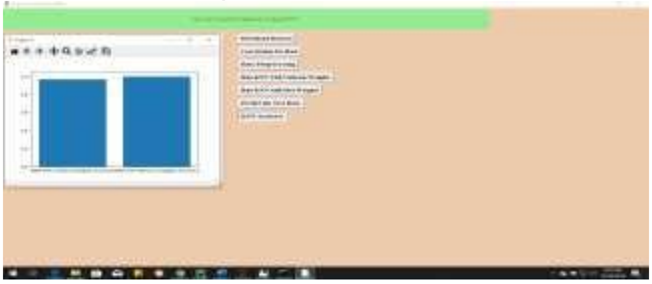

Fig :- Accuracy graph

5 CONCLUSION

The limitation of the proposed model is that it applies a binary classification technique. The actual output of this binary classification model is a prediction score in two- class. The score indicates the model's certainty that the given observation belongs to eitherthe Profit class or Loss class. For future work, the knowledge component is to transform the binary classification into multiclass classification. The multiclass classification involves observation and analysis of more than the existing two statistical class values. Additional research will include the application of the probabilistic model to multiclass data in order to provide more specific information of each class value. The newly formed multiclass classification will contain five class labels named “Sell”, “Underperform”, “Hold”, “Outperform”, and “Buy”. In numerical values for mapping purpose, we will convert “Sell” to -2 which implies strongly unfavorable; “Underperform” to -1 which implies moderately unfavorable; “Hold” to 0 which implies neutral; “Outperform” to 1 which implies moderately favorable; and “Buy” to 2 which implies strongly favorable.

References

[1] 1 *S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash* Benjamin Graham, Jason Zweig, and Warren E. Buffett, The Intelligent Investor, Publisher: Harper Collins Publishers Inc, 2003.

[2] Charles D. Kirkpatrick II and Julie R. Dahlquist, Technical Analysis: The Complete Resource for Financial Market Technicians (3rd Edition), Pearson Education, Inc., 2015.

[3] Bruce Vanstone and Clarence Tan, A Survey of the Application of Soft Computing to Investment and Financial Trading, Proceedings of the Australian and New Zealand IntelligentInformation Systems Conference,

[4] Vol. 1, Issue1, http://epublications.bond.edu.au/infotech_pubs/
13/,2003, pp. 211–216.

[5] Monica Tirea and Viorel Negru, Intelligent Stock Market Analysis System - A Fundamental and Macro-economical Analysis Approach, IEEE, 2014.

[6] Kian-Ping Lim, Chee-Wooi Hooy, Kwok-Boon Chang, and Robert Brooks, Foreign investors and stock price efficiency: Thresholds, underlying channels and investor heterogeneity, The North American Journal of Economics and Finance. Vol. 36, http://linkinghub.elsevier.com/retrieve/pii/S106 2940815001230, 2016, pp. 1–28.

[7] Lamartine Almeida Teixeira and Adriano Lorena Inácio de Oliveira, A method for automatic stock trading combining technical analysis and nearest neighbor classification, Expert Systems with Applications, http://linkinghub.elsevier.com/retrieve/pii/S095 7417410002149, 2010, pp. 6885–6890.

[8] Banshidhar Majhi, Hasan Shalabi, and Mowafak Fathi, FLANN Based Forecasting of S&amp;P 500 Index. Information Technology Journal, Vol. 4, Issue 3, http://www.scialert.net/abstract/?doi=itj.2005.2 89.292, 2005, pp. 289–292.

[9] Ritanjali Majhi, G. Panda, and G. Sahoo, Development and performance evaluation of FLANN based model for forecasting of stock markets, Expert Systems with Applications, Vol. 36, Issue 3, http://linkinghub.elsevier.com/retrieve/pii/S095 7417408005526, 2009, pp. 6800–6808. *http://dx.doi.org/10.2139/ssrn. 2646618*