STOCK MARKET TREND PREDICTION USING K-NEAREST NEIGHBOR (KNN) ALGORITHM

[1]KOMMERA MADHAVA REDDY, [2]D.SAI KRISHNA
[1]MCA Student, [2] Assistant Professor
Department Of MCA
Sree Chaitanya College of Engineering, Karimnagar

Abstract This paper examines a hybrid model which combines a K-Nearest Neighbors (KNN) approach with a probabilistic method for the prediction of stock price trends. One of the main problems of KNN classification is the assumptions implied by distance functions. The assumptions focus on the nearest neighbors which are at the centroid of data points for test instances. This approach excludes the non-centric data points which can be statistically significant in the problem of predicting the stock price trends. For this it is necessary to construct an enhanced model that integrates KNN with a probabilistic method which utilizes both centric and non-centric data points in the computations of probabilities for the target instances. The embedded probabilistic method is derived from Bayes' theorem. The prediction outcome is based on a joint probability where the likelihood of the event of the nearest neighbors and the event of prior probability occurring together and at the same point in time where they are calculated. The proposed hybrid KNN Probabilistic model was compared with the standard classifiers that include KNN, Naive Bayes, One Rule (OneR) and Zero Rule (ZeroR). The test results showed that the proposed model outperformed the standard classifiers which were used for the comparisons.

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

I Introduction

Analyzing financial data in securities has been an important and challenging issue in the investment community. Stock price efficiency for public listed firms is difficult to achieve due to the opposing effects of information competition among major investors and the adverse selection costs imposed by their information advantage.

There are two main schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis. The methodology used in fundamental analysis evaluates a stock by measuring its intrinsic value through qualitative and quantitative analysis. This approach examines a company's financial reports, management, industry, micro and macro-economic factors.

The second approach is known as technical analysis. The methodology used in technical analysis for forecasting the direction of prices is through the study of historical market data. Technical analysis uses a variety of charts to anticipate what are likely to happen. The stock charts include candlestick charts, line charts, bar charts, point and figure charts, OHLC (open-high-low-close) charts and mountain charts. The charts are viewable in different time frames with price and volume. There are many types of indicators used in the charts, including resistance, support, breakout, trending and momentum.

Several alternatives to approach this type of problem have been proposed, which range from traditional statistical modelling to methods based on computational intelligence and machine learning. Vanstone and Tan surveyed the works in the domain of applying

soft computing to financial trading and investment. They categorized the papers reviewed in the following areas: time series, optimization, hybrid methods, pattern recognition andclassification. Within the context of financial trading discipline, the survey showed that most of the research was being conducted in the field of technical analysis. An integrated fundamental and technical analysis model was examined to evaluate the stock price trends by focusing on macro-economic analysis. It also analyzed the company behaviour and the associated industry in relation to the economy which in turn provide more information for investors in their investment decisions.

A nearest neighbor search (NNS) method produced an intended result by the use of KNN technique with technical analysis. This model applied technical analysis on stock market data which include historical price and trading volume. It applied technical indicators made up of stop loss, stop gain and RSI filters. The KNN algorithm part applied the distance function on the collected data. This model was compared with the buy-and-hold strategy by using the fundamental analysis approach.

Fast Library for Approximate Nearest Neighbors (FLANN) is used to perform the searches for choosing the best algorithm found to work best among a collection of algorithms in its library. Majhi etal. examined the FLANN model to predict the S&P 500 indices, and the FLANN model was established by performing fast approximate nearest neighbor searches in high dimensional spaces.

Artificial neural networks (ANN) exhibit high generalization power as compared to conventional statistical tools. ANN is able to infer from historical data to identify the characteristics of performing stocks. The information is reflected in technical and financial variables. As a result, ANN is used as a statistical tool to explore the intricate relationships between the related financial and technical variables and the performance of stocks.

Neural network modelling can decode nonlinear regularities in asset price movements. Statistical inference and modifications to standard learning techniques prove useful in dealing with the salient features of economic data.

Some research has been carried out through the use of both qualitative and quantitative analysis. Shynkevich et al. studied how the performance of a financial forecasting model was improved by the use of a concurrent, and appropriately weighted newsarticles, having different degrees of relevance to the target stock. The financial model supported the decision-making process of investors and traders. Textual pre-processing techniques were utilized for the predictive system. A multiple kernel learning technique was applied to predict the stock price movements. The technique integrated information extracted from multiple news categories while separate kernels were used to analyze each category. The news articles were partitioned according to some factors from the industries and their relevance to the target stock. The experiments were performed on stocks from the health care sector. The results showed that the financial forecasting model had achieved better performance when data sources contain increased categories of the number of relevant news. An enhanced model for this study incorporated additional data source using historical prices and made predictions based on both textual and time series data. Additional kernels can be employed for different data sources. The use of new categorical features was to improve the forecasting performance.

.2 Literature survey

Sneh Kalra et al. in 2019, in this paper authors, did research on the fluctuation of stock market prices with respect to the relevant new articles of

a company. They used classifier Naïve Bayes to separate negative or positive statements for prediction purposes based on daily news variance the social media data, blogs data may be considered for future work

Aditya Menon et al. in 2019, this paper is focused on a review of neural model for forecast the stock tread after reviewing on a neural model they think that The long short term memory algorithm for predicting the economic information in confluence into the trendy era, this would be prioritized algorithm for forecasting

Ashish Sharma et al. in 2017, they found that regression analysis is mostly used for stock market trend prediction they survey of regression technique for stock prediction using stock market data. In the future result could improve by using more numbers variables.

Andrea Picasso et al. in 2019, in this research, authors worked which will alliance the economic and elemental analysis for market trend prediction through the various kind of application and automation methods neural network is machine learning technique the problem of trend stock and those are charts with forecasting data. As an input data sentiment of a news article is exploited. According to their research the problem in the most problematic accomplishment among the use of information about news astral one-off. To overcome this problem in the future the proper feature fusion technique will be suitable for the future.

Gangadhar Shobha et al. in 2018, this paper provided a full overview of machine learning techniques which will help to reader for use of equations and concept the author discussed about three type of all machine learning technique and also various kind of metrics like accuracy, confusion matrix, recall, RMSE, precision and quintile of errors. The author thinks that this review can help those people who are new to machine learning because most of the people confuse to use most of the machine learning techniques for prediction or others.

Suryoday Basak et al. in 2018, the author developed an experimental framework for predicting stock prices whether the price goes up or down in this experiment author uses the two algorithms name as a random forest classifier and Gradient boosted decision' n trees, and they got more accuracy in comparison to others research papers where others experiments got 50% to 67% results on the other hand according to the author of this paper they got 78% result accuracy for long term window. In the future, they could use the build boosted tree model for short term data window.

Arash Negahdari kia et al. in 2018, as the stock prediction so many experiments and models, have been developed for prediction purpose on historical data like as in this paper the author present HyS3 graph-based semi-supervised model and through a network views Kruskal based graph algorithm called ConKruG. In the future they think social media data, Twitter data could be used for the prediction of stock for better results using these algorithms.

Bruno Miranda et al. in 2019, in this paper the survey of bibliographic techniques that focus on text area for research the author works on the prediction of financial market values by using the machine learning models support vector machine (SVM) and neural network with data set from North American market new models may have opportunities for north American market data for prediction purpose in future.

K. Hiba Sadia et al. in 2019, author aim for this paper us to preprocess the raw data firstly then they are doing a comparison between random forest and SVM algorithm the main aim of the

author is to find out the best algorithm for stock trend prediction in the last they have given the best-fit algorithm for future stock forecasting which is random forest algorithm for future work they think that for getting more accuracy in result the adding of more parameters can be good.

Akash et al. in 2019, the author introduced two more algorithm name as “LS SVM” which is least square support vector machine and another one is “PSO” (particle swarm optimization) the work “PSO” basically select the be stun bounded parameter with the “LS SVM” to reduce the overfitting and some technical indicators which will basically enhance the result accuracy. On the other hand at the same time, the proposed algorithm is being compared with artificial neural network model.

Aparna Nayak et al.in 2016, in this paper authors, woked to predict the stock market trend by using the supervised learning methods, here authors predicted the data based on daily live data which is directly calling by the program using yahoo financial website and also predicting the monthly Mu yen chen et al. in 2019, authors did research to calculate the impact of news articles on the stock prices using deep learning approach LSTM (long short-term memory) and they think this study can predict the stock market trend. based prediction where in this paper they got a better result for daily live prediction instead of monthly prediction further future work they think if we consider more sentiments to the monthly [prediction that would also generate the best result .

Nuno Oliveira et al. in 2016, in this paper the author purposed a methodology by which they can access the value of stock prediction and microblogging data they used, for stock prices and return indices and some more like a portfolio. For this experiment, they have used huge data of Twitter, for all this experimental work they use Kalman filter to merge the microblogging data and some external sources and as a result, they found twitter data and blogging data were relevant for the purpose of forecasting these datasets were very useful. This result can be improved by using some more and different data such as social media datasets and others.

Han lock Siew et al. in 2017, the author in this paper used regression technique for finding the accuracy in the forecasting of a stock trend for all this experiment they used WEKA software which used for data mining and machine learning algorithms to execute them, the dataset they used which contains heterogeneous values and which is used for handling of currency values and functional ratios. The dataset for calculating the stock movement is collected from Bursa Malaysia for forecasting purposes. For the future extension, the authors thought that the forecasting using the regression method can be improved by using the more standardized ordinal format of data.

Smruti rekha das et al. in 2019, in this paper authors, used firefly method for forecasting the stock prices as an input dataset author collect from four different websites name as NSE-India, BSE, S&P 500 and FTSE, and all collected dataset is well transformed by using proper mathematical formulas by using the backpropagation, neural network and more two methods used for prediction, forecasting according to the time horizon of alternate days 1 day, 3 days, 5 days and so on. For future work, there may be some chance to get more accurate results by giving more parameters to the implemented algorithms .

Dattatray P.Gandhmal et al. in 2019, authors had written the paper on the review of stock prediction techniques. In this paper, the authors reviewed about 50 published research papers

according to the publication years, and the authors suggested the best technique for prediction. KNN and fuzzybased techniques as the authors suggested are the best techniques according to the review such as KNN, SVM, SVR, and much more but these two techniques can be more effective for the purpose of using historical data. In the future, they will review more papers to get the best-fit algorithm for prediction .

Financial services companies are developing their products to serve future prediction. There are a large amount of financial information sources in the world that can be valuable research areas, one of these areas is stock prediction and also called stock market mining. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets. The genetic algorithm had been adopted by Shin et al. (2005); the number of trading rules was generated for Korea Stock Price Index 200 (KOSPI 200), in Sweden Hellestrom and Homlstrom (1998) used a statistical analysis based on a modified kNN to determine where correlated areas fall in the input space to improve the performance of prediction for the period 1987-1996. Both models mentioned were provided in the Zimbabwe stock exchange to predict the stock prices which included Weightless Neural Network (WNN) model and single exponential smoothing (SES) model Mpofu (2004). Clustering stocks approach was provided by Gavrilov et al. (2004) to group 500 stocks from the Standard & Poor. The data represented a series of 252 numbers including the opening stock price. A fuzzy genetic algorithm was presented by Cao (1977) to discover pair relationship in stock data based on user preferences. The study developed potential guidelines to mine pairs of stocks, stock-trading rules, and markets; it also showed that such approach is useful for real trading. Moreover, other studies adopted kNN as prediction techniques such as (Subha et al., 2012; Liao et al. 2010; Tsai and Hsiao 2010; Qian and Rasheed, 2007)

3 Implementation Study

A nearest neighbor search (NNS) method produced an intended result by the use of KNN technique with technical analysis. This model applied technical analysis on stock market data which include historical price and trading volume. It applied technical indicators made up of stop loss, stop gain and RSI filters. The KNN algorithm part applied the distance function the collected data. This model was compared with the buy-and-hold strategy by using fundamental analysis approach.

Disadvantagas:-

- Limited analysis
- Applied on historical price and trading volume
- buy-and-hold strategy

.3.1proposed methodology

Stock market prediction is an act of trying to determine the future value of a stock other financial instrument traded on a financial exchange. The programming language is used to predict the stock market using machine learning is Python. In this paper we propose a Machine Learning (ML) approach that will be trained from the available stocks data and gain intelligence and then uses the acquired knowledge for an accurate prediction. In this context this study uses a machine learning technique called K-Nearest Neighbor to predict stock prices for the large and small capitalizations and in the three different markets, employing prices with both daily and up-to-the-minute frequencies.

Advantages:-

- determine the future value of a stock
- available stocks data and gain intelligence analysis

- accurate prediction
- large and small capitalizations and in the three different markets

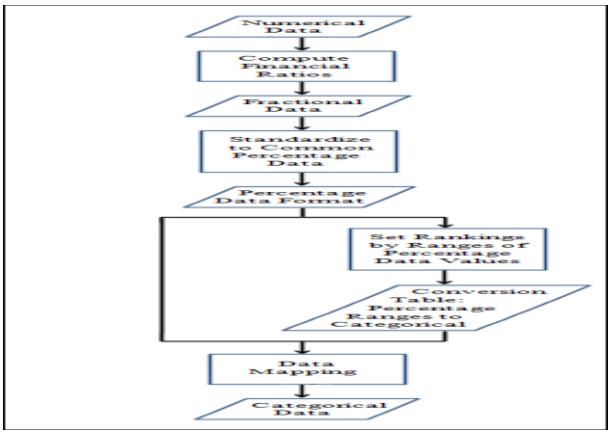

Fig 1:- proposed model

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 Xquery 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 number of 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

4.1 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.

4.2 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.

4.3 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.

4.4 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

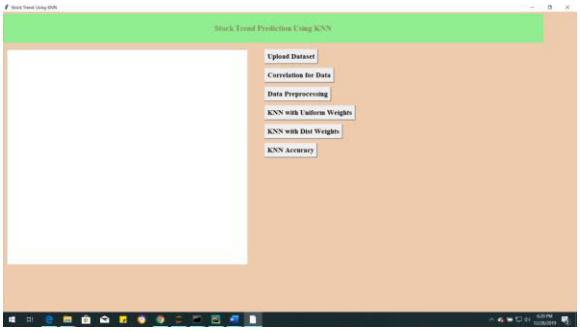

Fig 1:. Home Screen

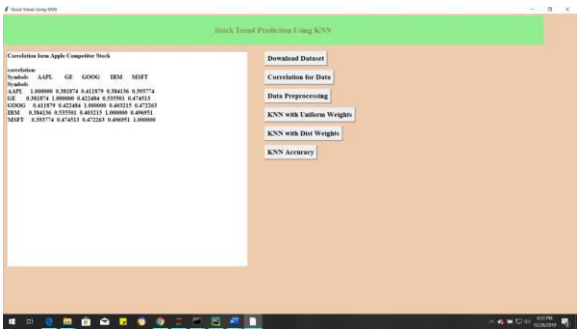

Fig 2: Download Dataset

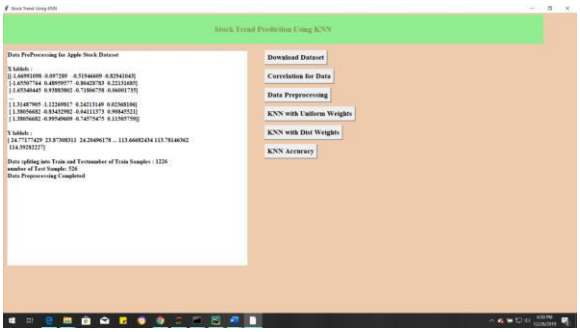

Fig 3:- Correlations for Data

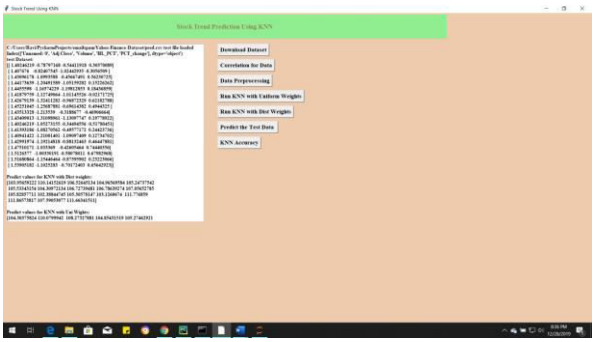

Fig 4:- Data Pre-processing

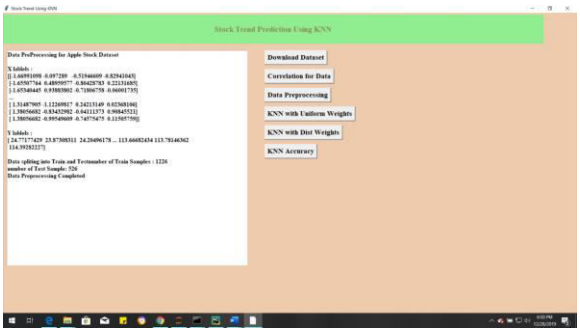

Fig 5 KNN Accuracy

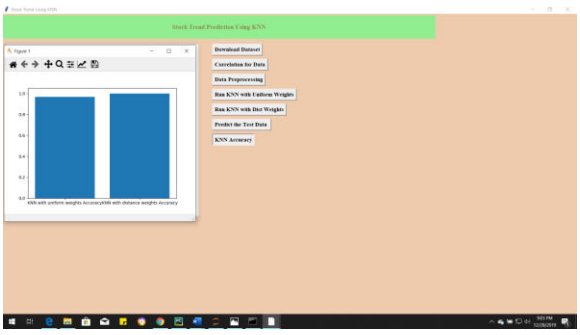

Fig :- Accuracy graph

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.

6 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 Intelligent Information Systems Conference,

[4] Vol. 1, Issue 1

[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/S1 06 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/S0 95 7417410002149, 2010, pp. 6885–6890.

[8] Banshidhar Majhi, Hasan Shalabi, and Mowafak Fathi, FLANN Based Forecasting of S&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.

[1] http://dx.doi.org/10.2139/ssrn. 2646618