Stock trend prediction using Ensemble learning techniques in python

P.Rajesh,N.Srinivas, K.Vamshikrishna Reddy, G.VamsiPriya, Vakula Dwija.M, D.Himaja .

***Abstract*: Stock trends are generated in huge volume and it changes every second. Stock market is a complex and volatile system where people will either gain money or lose their entire life savings. This project is about taking quantifiable data from finance API about the top 500 companies in S&P stock exchange and predicting its future stock trend with ensemble learning. To achieve it we have considered mainly two prediction methods, Heat Map and Ensemble Learning, which based on the percentage change in the stock price data will classify the stock into buy, sell or hold categories. Heat map is generated based on the correlation coefficient of the quantifiable data to further classify the stock as one of the three above mentioned categories. On the other hand, we used the ensemble learning model to classify the stock into a majority vote-based system that considers 3 main classification models. Observations shows that Random Forest, SVM and K-neighbors classifiers show the most prominent results of all other possible combinations. The accuracy of the prediction model is more than 51% whereas in comparison with prediction models with a single classifier labelling with 30% accuracy the model has increased the accuracy by 23%.**

***Index Terms*: Stock trends, Machine Learning, Ensemble Learning, Heat map, K-Neighbors, Random Forest, SVM.**

I. INTRODUCTION

A. Stock Market

Stock market prediction is the act of trying to determine the future value of a company stock or other financialinstrument traded on an exchange. It is a known fact that the stock market is one of the driving factors that run the economy of a country or a person depending on the amount of time he/she spends on learning its patterns. [1]Stock market has always been volatile and subjective where a successful prediction is often considered as a lucky guess and would have a low possibility of happening ever again. To most of the people a stable prediction model for stocks has been a dream and something that companies and individuals are investing a whole lot of money in order to achieve it. [2] The successful prediction of a stock's future price could yield significant profit. The efficient-markethypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. [3]

Manuscript published on 30 March 2019.
*Correspondence Author(s)

Dr.P.Rajesh,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.
Dr.N.Srinivasu,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.
K.VamshiKrishna Reddy,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram,India.
G.Vamsi Priya,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.
Vakula Dwija.M,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.
D.Himaja,Computer Science Department,Koneru Lakshmaiah Education Foundation, Vaddeswaram, India.

B. Data Sources for Market Prediction

There are wide ranges of sources for an effective market data but the main ingredient that make it all unique for everyone is the type of processing that one wants to do using the data and can vary from one researcher to another. [4] In this project we have considered the Yahoo finance Aliyah finance API is an entity that provides free stock data for a certain period that the user specifies. It has the universal structure where the data provides 7 columns in its stock data

1. DATE
2. OPEN
3. CLOSE
4. HIGH
5. LOW
6. VOLUME
7. ADJACENT CLOSE

C. Correlation

Correlation is any statistical association, though in common usage it most often refers to how close two variables are to having a linearrelationship with each other

D. HeatMap

It is one of the prediction models that uses the data from the correlation coefficient generated using the cr() function in python and generate a correlation table for all the data sets available. This can be further represented using the color coding. Negative gives Red, Positive gives Green, Neutral gives Yellow. Thus, by using this technique it will be easy to understand the trend

E. Machine Learning Classifiers

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). [6]
There are two types of learners in classification as lazy learners and eager learners.

Lazy learners: Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting. [7]

Ex. k-nearest neighbor, Case-based reasoning

Eager learners: Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict. [8]

Ex. Decision Tree, Naive Bayes, Artificial Neural Networks

F. Ensemble Learning

It is a concept where instead of using a single classifier on the data set it considers 2 or more classifiers and performs a majority vote-based classification. [9] In this type of classification, the data that is classified by the majority as true and will be considered true thus increasing its efficiency. It is a rather new learning method and can be used on small scale to medium scale data sets.

G. Python Libraries

Python is and has been the best finance analytics tool that has wide range of libraries that help in gathering the data, performing analytics and representing the data. PANDAS provide a wide range of data assimilation tools making it easier to carry the data across the algorithm. MATPLOTLIB provides wide range of data representation tools and methods the help in making the results visually attractive to people. BEAUTIFUL SOUP is a part of natural language processing that helps in gathering the data from the web-servers. [10]
The rest of paper is structured as follows. Section II describes about the literature survey of various stock trend predictions using data mining techniques. Section III Brief explanation of proposed Methodology of ensemble stock trend prediction and system architecture, this module predicts whether the stock to be bought or sold or to hold. Evaluation and comparative based analytical results of Stock Trend Prediction Using Ensemble Learning Techniques in Python were presented in section IV. Section V consists of conclusion and future scope.

II. LITERATURE SURVEY

Title	Year	Methodology	Disadvantages
Improved Stock Market Prediction by Combining Support Vector Machine and Empirical Mode Decomposition	2012	This paper gives a brief sketch about a two-stage neural network architecture which is constructed by combining Support Vector Machine (SVM) and Empirical Mode Decomposition (EMD) used to propose the stock market prices.	The main drawback of this model is the financial data will be divided into many regions on the bases EMDS, then these regions will be using different SVMs which are having different kernel function and various parameters for prediction of financial data. This will be a time-consuming process because of rigorous splitting and execution of data.
Stock Volatility Prediction using Multi-Kernel Learning based Extreme Learning Machine	2014	This paper mainly concentrates how to design a methodology which is going to improve prediction accuracy through implementing a multi-kernel learning based on the Extreme learning machine using the HKEx 2001 stock market datasets. These two methodologies will help in improving the prediction accuracy and prediction speed	There is a small drawback present here is, if the input data become more complicated then MKL-ELM will be more time-consuming .
Prediction of Stock Market by Principal Component Analysis	2017	this paper had used a method called the principal component analysis (PCA) with linear regression to reduce the problems faced in stock market price prediction like high dimensionality (means reducing redundancy in the data).	. The main drawback in this model is we need to careful select the principal components, otherwise the model will not show accurate results.
Stock Market Prediction using Optimum Threshold based Relevance Vector Machines	2016	this paper mainly concentrates on prediction of prices in the stock market domain using optimum threshold-based Relevance Vector Machines (RVM),	The drawback present in this system is if there is a negligible degradation for datasets used then there a reduction in number of RVs (Relevance Vector) is observed
Stock Market Prediction Using Hidden Markov Model	2014	This paper has used a Hidden Markov models (HMM) to predict the change in values of stock prices.	The performance of this model decreases with increase in training data
Survey of Stock Market Prediction Using Machine Learning Approach	2017	This paper predicts the stock price values using various regression model to get the better output values. The regression model used are polynomial regression which is a linear regression in which nth degree polynomial is formed by considering the relationship between the independent variable x and the dependent variable y	The least squares approach can be used to fit models that are not linear models.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Stock Market Prediction Using Machine Learning Techniques	2016	The main theme of the paper is to predict stock prices performance using Karachi Stock Exchange (KSE). The input attributes are compared to predict the results as either positive or negative.	We need to be very careful while choosing the attributes otherwise this model does not give any good results.
Stocks Market Prediction Using Support Vector Machine	2013	This research has put the concentration only on four companies and their factors affecting the stock trend, are chosen for stock multivariate analysis. The approach used is Support Vector Machines strategy which can be used for both classification and regression models for analyzing the relationship among all the factors and predict the stock market performance.	The drawback is if we feed the model with n number of parameters it may not give accurate results, and if there are any minor fluctuations in the training datasets then major impact will be shown in prediction ultimately predictive results will decrease.
Forward Forecast of Stock Price Using Sliding-window Metaheuristic-optimized Machine Learning Regression	2017	this paper mainly concentrates on analyzing the time series and modeling of finance time series based on many guidelines taken from the investors.	This prediction techniques is only for the non-linear time series, it will not work for the traditional models.

III. METHODOLOGY

A. System Architecture

Following architecture is considered for the project to predict the stock market trend using python.

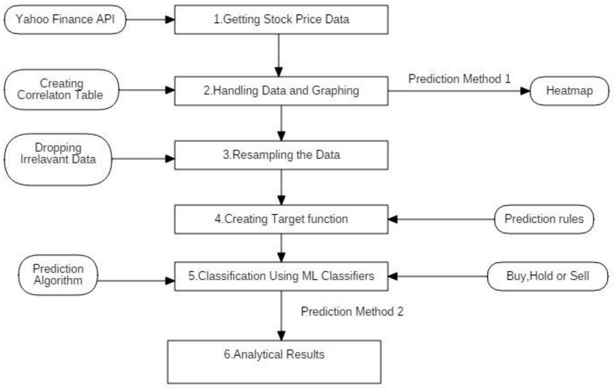

Figure 3.1: System Architecture

S&P 500 Data Collection: This project collects data from yahoo finance API based on the S&P 500 companies i.e. top 500 companies of the S&P stock exchange. Using beautiful soup, the tickers of top 500 companies are collected and stored locally. Pandas python library analytics is used to collect stored sticker stock data from 2013 to 2018 for further preprocessing. This collected data is stored into csv files with the following columns date, open, high, close, volume, adjacent close.

Pre-processing: In this module the collected data is preprocessed in order to support the requirements of our two prediction models heat map and ensemble learning. During the process of data collection from stock API even though pandas does a little preprocessing of the data is not enough. For example, the dates present will be in a random order for the specified year this column will be parsed in our and a index column will be added to the dataset. This gives the data a certain edge to the already present dataset making it robust and ready for further processing.

Heat map generation: Heat map is one the prediction models that we implement in this project. Heat map works with correlation coefficient derived from the data sets collected and preprocessed as per last step. It is achieved using predefined correlation function from the python libraries. This gives an output of 500 companies correlation values with one other creating a one 500 * 500 matrix in a dataset. Now that we have the correlation matrix we can build a heat map using it, in order to represent the entire matrix and to differentiate it color coding is given to the matrix. This occurs as follows positive correlation gives green color, negative correlation value gives red and finally neutral correlation value gives orange color. This color coding gives it a sense of visual acceptance by the end user.If a company heat map gives a major green coding it would be the company that we refer to a user for buying the stock and vice versa.

Resampling Data: Although heat map is a great prediction method it has its shortcomings, when faced with an enormous dataset it will be quite difficult for the user to search for the required correlation match out of the all available combinations.

Thus, in order to work towards a clear prediction model with concise results we have implemented machine learning more precisely ensemble learning methods for our prediction model. For every machine learning application, we know that there are 3 key ingredients, target function, training datasets and classifiers.

Creating target function: For any machine learning algorithm target function is an important element that defines its efficiency in achieving the given task. We have considered our data sets and performed necessary data resampling to aid the target function we devised.The basic principle of our target function takes the user given percentage change in the stock trend regardless of it being a loss or gain and works with a counter. This ultimate counter values after the analysis of all the data present in the dataset is passed to the ensemble learning classifiers in order to generate our prediction.

Using machine learning classifiers: Using the target function we would get a basic idea on how our prediction model works, using this knowledge we modified our datasets and passed it to the classifiers.We have observed all the previous projects that worked on the stock market prediction used mostly a single classifier for its training. We in this project considers 3 classifiers, k neighbors, SVM and Random forest, using ensemble learning we have combined its efficiency and merged into one new voting classifier. This improved the efficiency of the training data exponentially.

Predicting the trend: Thus, ensemble learning method for the voting classifier works such that if the majority of the 3 key classifiers i.e. one or more classifiers agree to the result it will be given as positive, negative or neutral based on the conditions specified. This module predicts whether the stock to be bought or sold or neither. (fig 4.5)

Analytical results: This module depicts the results that we collected in the previous modules. The prediction of the trend is attained by using the counter that which ensemble learning provides making it the ultimate prediction for the given stock. This can be coupled with already provided data to represent in a visual format given in figure 4.6 Making it easily understandable by the end user.This can be further classified for more than one companies using multiple plots where each one relates to a certain company.

IV. IMPLEMENTATION RESULTS

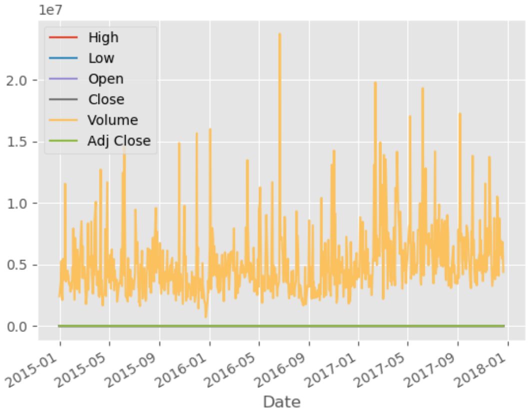

Fig IV.1: Plot of sample Data set that has been imported from YAHOO finance API. Since the volume has a greater number than others it dominated the entire plot

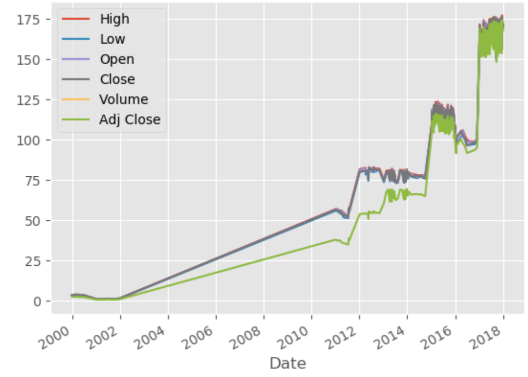

Fig IV.2: Removing the volume from the dataset we can see the other datasets plots. But even they are adjacent to each other and are not visible clearly.

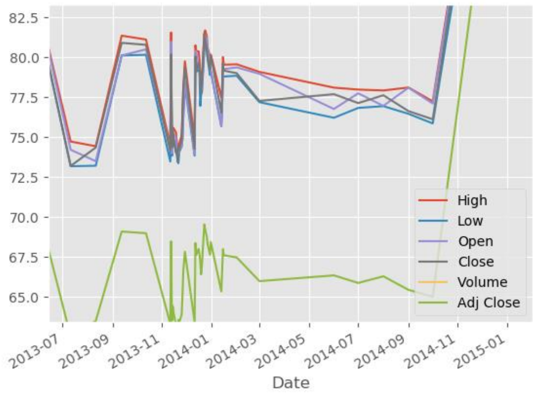

Fig: IV.3 This further deeper plot gives the different data parameters and changes it takes place that are not visible in Fig 5.1 and Fig 5.

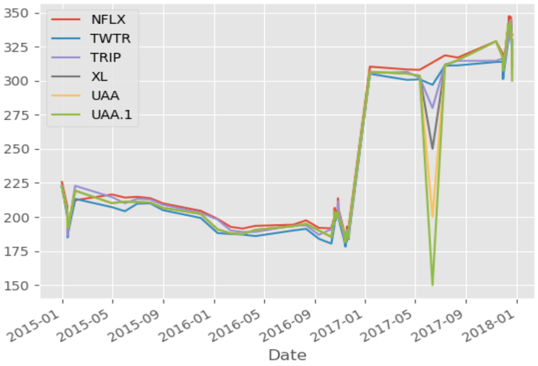

Fig:IV.4 This graph shows the top 6 performers of the S&P stock Exchange during the fiscal year 2018.

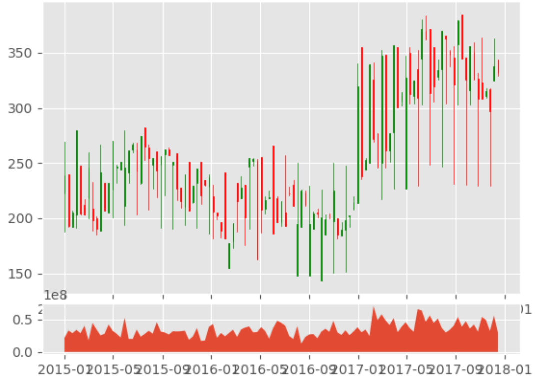

Fig:IV.5 It shows the candle stick representation of the data set where red means losing stock and green gives rising stock market shares during the respective years given on x axis.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

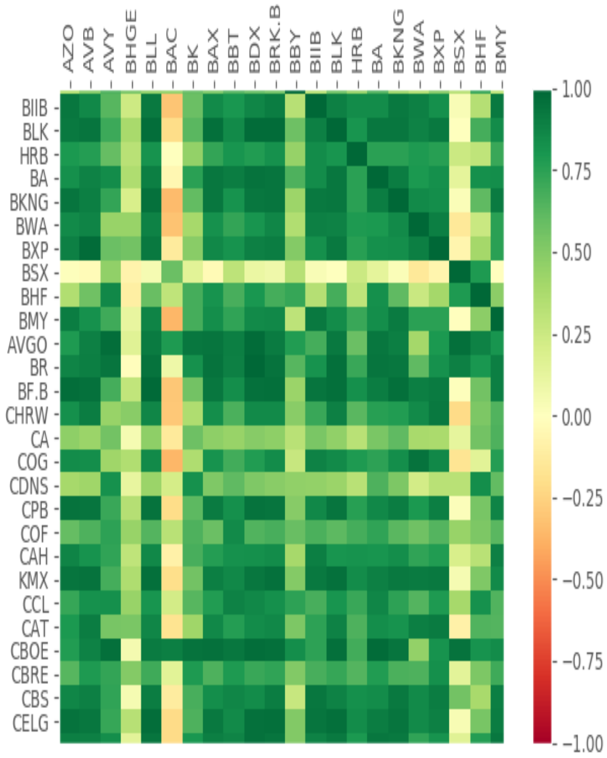

Fig:IV.6 This shows the heat map where all the 500 companies are represented in a correlation table and has color coding that suggest green-buy, red-sell and yellow-hold making it easier to compare with respect to other companies for a better understanding of the stock.

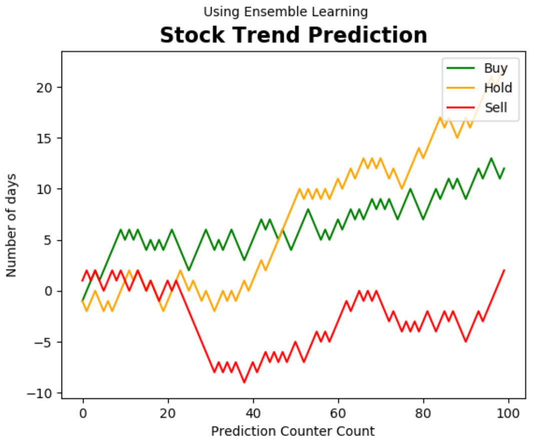

Fig:IV.7 This plot gives the final representation of the ensemble learning model that gives the prediction counter for a stock given by the user. The highest counter suggestion will be considered as the ultimate decision for the stock in this case hold has higher counter making it the ideal decision.

V. CONCLUSION

This project concentrates mainly on the static data that was obtained from yahoo finance API for the given time period. This can be further optimized for real time stock market prediction by implementing other sources of stocks data such as news journal policy etc. The goal is to obtain maximum efficiency with minimum classification complications of the stock trends.

ACKNOWLEDGMENT

Authors would like to express our gratitude to our guide, professors, family and friends who supported in the fulfillment of our minor project. Thankful to all who helped during this project.

REFERENCES

1. Stock Market Prediction Using Hidden Markov Model, PoonamSomani, ShreyasTalele and SurajSawant.
2. Survey of Stock Market Prediction Using Machine Learning Approach, ASHISH SHARMA, Dinesh Bhuriya and Upendra Singh.
3. Stock Market Prediction Using Machine Learning Techniques, MehakUsmani, Syed Hasan Adil, KamranRaza and Syed Saad Azhar Ali.
4. Stocks Market Prediction Using Support Vector Machine, Zhen Hu, Jie Zhu, and Ken Tse
5. Forward Forecast of Stock Price Using Sliding-window Metaheuristic-optimized Machine Learning Regression,Jui-Sheng Chou and Thi-Kha Nguyen.
6. Improved Stock Market Prediction by Combining Support Vector Machine and Empirical Mode Decomposition, HonghaiYu and Haifei Liu-2012.
7. Stock Volatility Prediction using Multi-Kernel Learning based Extreme Learning Machine, Zhiyong Zhao, Xiaodong Li, FeiYu and Hao Zhang
8. Prediction of Stock Market by Principal Component Analysis, Muhammad Waqar, Hassan Dawood, Muhammad Bilal Shahnawaz, Mustansar Ali Ghazanfar and Ping Guo
9. Stock Market Prediction using Optimum Threshold based Relevance Vector Machines, Karthik HS, Nishanth VA and Manikandan
10. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm, Lei Zhao ; Lin Wang

AUTHORS PROFILE

Dr.P.Rajesh, Professor at Koneru Lakshmaiah Educational Foundation, has done his **M.Tech,Ph.D** of Computer Science Engineering discipline from JNTU,Hyderabad. He is a reviewer for the journal **DMKD**,Wiley publications,Canada , Elsevier Journal of computational Statistics and Data Analysis and for the conference of international conference on innovative methods in engineering applications, applied sciences, **anveshana**. He has published various international journals and conferences at **IEEE ,Springer, ACM, ICMLDA** etc. He is a Technical program committee member for international conferences such as **FSDM 2018**,Bangkok,Thailand , **SCML2019**, wuhan, China, and the second international workshop on **Data science Engineering and its Applications,**Spain.

Dr.N.Srinivasu, Dr. N. Srinivasu obtained Phd Computer Science and Engineering from Nagarjuna University in 2012.AndhraPradesh ,India. He is currently working as Professor in Koneru Lakshmaiah Education Foundation. research interests are Cloud Computing, Big data analytics, Soft Computing. He has more than 35 Publications in various International Journals and Conferences.

K.Vamshikrishna Reddy is a student of Computer Science Department at Koneru Lakshmaiah Educational Foundation,Vaddeswaram,Andhra Pradesh.He is doing his research work in Knowledge Engineering.

G.Vamsi Priya is a student of Computer Science Department at Koneru Lakshmaiah Educational Foundation,Vaddeswaram,Andhra Pradesh.She is doing her research work in Knowledge Engineering.

Vakula Dwija.M is a student of Computer Science Department at Koneru Lakshmaiah Educational Foundation,Vaddeswaram,Andhra Pradesh.She is doing her research work in Knowledge Engineering.

D.Himaja is a student of Computer Science Department at Koneru Lakshmaiah Educational Foundation,Vaddeswaram,Andhra Pradesh.She is doing her research work in Knowledge Engineering..