STOCK MARKET PREDICTION USING MACHINE LEARNING ALGORITHMS

A.SARANYA, Dr.R.ANANDAN

Abstract: ***The purpose of this work is to predict the stock price fluctuation and find the best usable algorithm for predicting stock price by comparing the outcomes of various algorithms of machine learning considering various factors. The algorithms we are proposing to use are linear regression, logical regression, LSTM, random forest algorithm, SVM and naive Bayes' algorithm. We are aiming to apply various algorithms and predict the one with best outcome. The factor which we have used as an attribute in our model is news that we assume will affect the price of the stock market. We have taken the top 25 news of the day and each news will be evaluated as if it's positive or negative and then the positive new is assigned with +1 value and so as negative with -1. For the particular day the sum of the news values are taken if the sum is positive then the predicted price will increase as compare to previous price but if the sum is negative the value should decrease base on these facts evaluation is done between predicted and occurred value and so the algorithms are used to generate the prediction and hence used to calculate the accuracy provided by the algorithm .Using news as the factor may help us in the more chance of increase in the detecting the fluctuation in the values as the news is one of the greatest factor effecting the change in stock prize as news contain the brief every possible event happened in the previous day and also contain about the company that is their release of product , status , bonds , funds , investments.***

Index Terms**: *Machine learning, Random forest, Linear Regression, Logical Regression, LSTM, Naïve Bayes' algorithm, SVM.***

I. INTRODUCTION

Machine learning is to anticipate the future from past information. It is a kind of AI that gives computer the capacity to upgrade being expressly customized. In machine learning the system learn from the experience by implementation of different machine learning algorithm using python. It uses a specialized algorithm for training the data sets and based on the training data it gives a prediction on new data sets. Machine learning has been divided into categories that are learning that are supervised, unsupervised and reinforcement. Learning which is supervised is a algorithm which the input data is provided and corresponding the data is trained with correct answer labeled by human being. Unsupervised Learning has no labels. It provides the learning algorithm. This algorithm has to find the pattern of the input data. Finally, Reinforcement learning dynamically interacts with the environment and it's receive, positive or negative feedback to improve its performance.

In recent years, Machine learning has been widely used in detection and achieved favorable performance. Stock market prediction is one of the most popular process in world and also it has wide spectators all around the world. As predictive analyzing strategy would be very useful for investors, brokers, business men and common people. Investors and brokers can change their strategy or companies can advertise their product based on outcome of a prediction which is being predicted by using machine learning. From historical data we created feature set that includes the dataset. The aim is to investigate machine learning based techniques for forecasting by prediction result in best accuracy.

Share market prediction is used to generate the upcoming values of the stock of a company. It is successful there will be more profits. If stock price is predicted successfully it will increase profit of investor. The purpose of the paper is to propose the appropriate machine learning algorithm to predict stock market price. In this project we use various modules to get the best outcome as in previous paper such as use of k-mean algorithm has a drawback that it can't provide us the best output with highly varying inputs. Basically the thought is of initially taking a static data set on which the various machine learning algorithms are been applied to get the best algorithm by comparing their results and then finally we are aiming to adding on various features or attributes and then by applying the algorithms to get the best results and learn how these features affect the stock market.

II. RELATED WORK

We were taken various papers as reference which are mentioned in the above table.

The paper [1] is about using the K mean algorithm to discover the effective technical trading pattern of the stock market but it is with the restriction in the amount of input data which can be used. The paper 2 discusses about using various algorithm to find the best algorithm for predicting the share price of KSE market but it is not using any specific feature. The paper 3 is about using linear regression model for stock market prediction but the problem is that it can't be used for the prediction where high fluctuation is faced in share market. The paper 4 is about using adaptive SVM for high frequency stock price forecasting where problem is faced for predicting the value when there is low frequency in the share price.

Revised Manuscript Received on July 05, 2019.

A.Saranya , Research scholar, Department of computer science and Engineering, VELS Institute of Science and Technology,

Assistant professor (O.G),Department of Software Engineering, SRM Institute of Science and Technology, Chennai 603203,India (aksaranya@gmail.com)

Dr.ANANDAN , Professor, Department of computer science& Engineering, VELS Institute of Science and Technology, Chennai 603203 India.(anandan.se@velsuniv.ac.in)

III. EXISTING SYSTEM

Data Mining and Machine learning in stock exchange, could be a bran-new analysis in engineering fields with a heap of challenges. For creating a result system that forecasts for a fluctuation in stock price, particularly for any top BSE market whereas the stock price is ongoing different machine learning algorithm and approach that are statistical were taken to estimate the best possible outcomes. A math approach which is multiple linear regression can be framed for comparison. The model is useful in predictive modeling. Currently, in Stock market various factor responsible for large fluctuation in opening and closing price. Furthermore, closing value in stock market depends upon the trading done in whole day and to predict the opening value of a particular market using Multiple Variable Linear regression combined with Logistic regression and ultimately the buying the stock will be benefit or not using Random Forest algorithm.

Classification problems are widely solved through this. Logical regression don't need linear relationship btw variables and conditional. It can be used by multiple types of interconnections as if bid non-linear log for prediction of odd ratio. Eliminate over fitting and under fitting, it will take all important variables. A marvelous approach to make sure the way is to use a stepwise technique to calculate the LR algorithm. It needs massive sample sizes than normal least square. However, the recourse for involve effects that are interaction of variables which are categorical within the calculation and for the model. If we have ordinal dependent variable, then it is called as ordinal logistic regression. If the problem is having multi- class dependent variable, then we can call is as multinomial logistic regression

To develop a system for predicting the result of and stock price whereas the share market trading is developing. It uses the previous stock value data of a particular market in for designing our skeleton and Multiple Variable Linear Regression is been used for designing. Estimation was also carried out successfully in our work. MLR (multiple liner regression) has been very useful for predicting in regular intervals and the winner is generated by the final value of the stock price.

IV. DATASET DESCRIPTION

Our data set is about the features of the US based stock market which includes the feature like opening price, closing price, date, highest, lowest.
In our upgraded dataset we are including how the news and tweets are having some impact on the stock market price

V. PROBLEM ANALYSIS

A. Naive Bayes' algorithm: The naïve Bayes' algorithm uses the Bayes' theorem where the data set is converted to the frequency table from where we need to find out the probabilities of various outcomes. The probability tables is been updated throughout the training period and then finally using this theorem we can get out final outcome. For introducing a new survey, we need to look for the probabilities of class in the table of probability which is based on the features.

Why naïve Bayes' algorithm:-
1. Quick and through for predicting the class of data.
2. Does well in multiclass prediction.
3. Useful for categorical data rather than numerical
4. Used for independent data attributes

B. Random forest algorithm: The random forest algorithm is used to create many decision trees and then finally combine them together to get the more faultless and stable prediction.
1. Used for both classification and regression types of problem.
2. Handle the missing value and maintain accuracy for missing data.
3. Will not over fit the system.
4. Estimate larger datasets with better dimensionality.

C.SVM: The aim of the SVM algorithm is an N-dimensional space that distinctly classifies the data points. In SVM algorithm, we used the n-dimensional spaces where we use n for representing the number of traits, also with the value of every feature is the value of a coordinate which is particular to plot each data item. Kernel functions are used. They return the dot product of 2 vectors so as to map the points in the higher dimensional space. The data point are referred as vectors. Kernels are basically the distance between 2 observations.
Reason for SVM algorithm:-
1. Can be useful for nonlinear dataset
2. Robust against over fitting
3. Effective in high dimensional space.
4. Combination of kernel functions can be used for better results.

D. Logistic regression: Grouping tasks can be performed by this. Logistic regression uses logistic function also known as sigmoid function. The binary variable are dependent variable in logistic regression which have data code as 1 or 0. The structures are linear, due to which they work good when you use linearly separable classes. Logistic regression can also be generalizing by completing coefficients with a tunable penalty strength. We can use more complex decision boundaries by fitting complex parameters using high order polynomials. The coefficients of the logistic regression algorithm must be estimated from your training data. Logistic regression will not perform good when there are many and non-linear decision boundaries. They will never be ample plaint to capture naturally more relationships that are complex.
Reason for logistic regression:-
1. Often run faster
2. Can numerically approximate gradient of your values
3. Avoid over fitting
4. Don't need pick learning rate

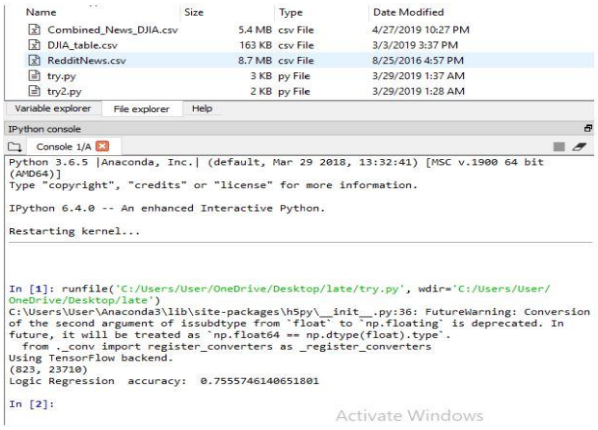
E.LSTM: It is an artificial RNN structure which is used in the field of DL. LSTMs are designed for avoiding the problems which are long-term dependency i.e. be able to join old data to the ongoing work. The core idea is of cell state. The cell state is a type of conveyer belt. It go through the full chain, with only some less important linear interactions. It is easy for unchanged information to flow
Reason for LSTM:-
1. Get deep learning in the project.
2. Control vanishing gradient/context problem

VI. PROPOSED MODEL

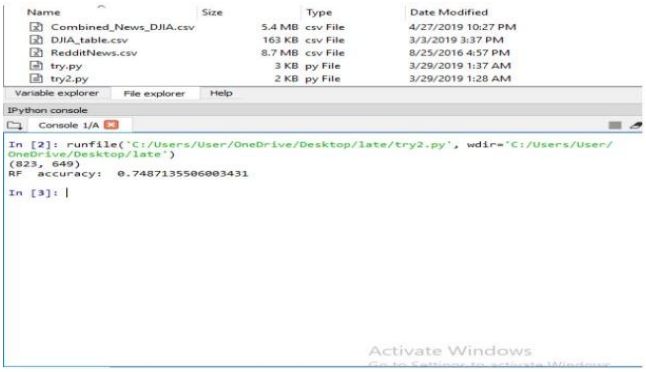
So considering the various factors and problems we are proposing a model where we are trying to overcome some of the mentioned problems and for that we are using various machine learning algorithms. And also we are here with a new feature where we are considering the news for our attribute to predict the fluctuation the stock market price.

VII. RESULTS

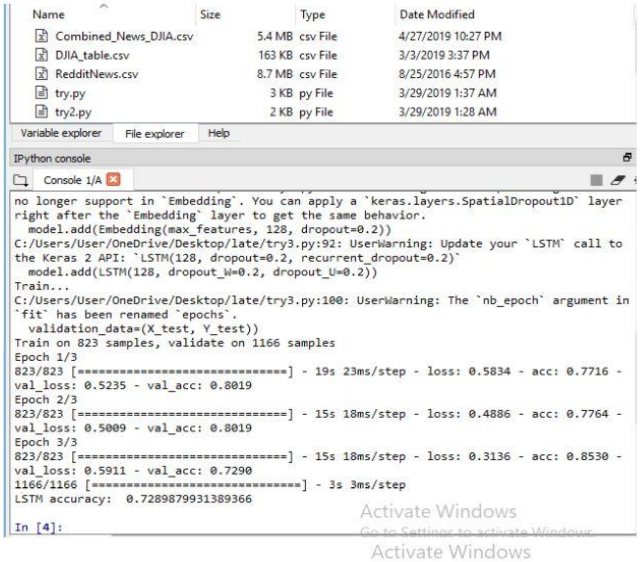
A. LOGISTIC REGRESSION OUTPUT:



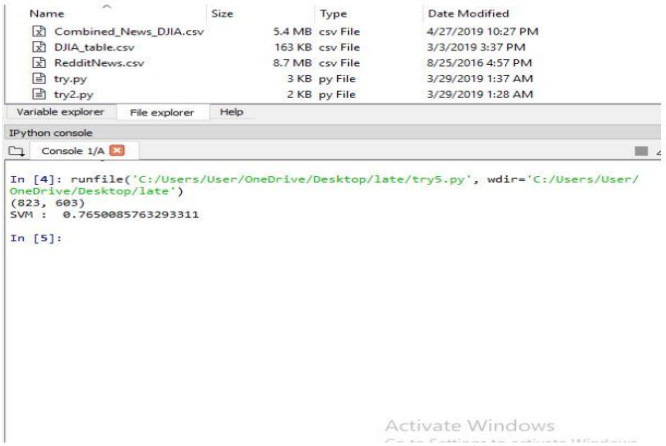
B. RANDOM FOREST OUTPUT:



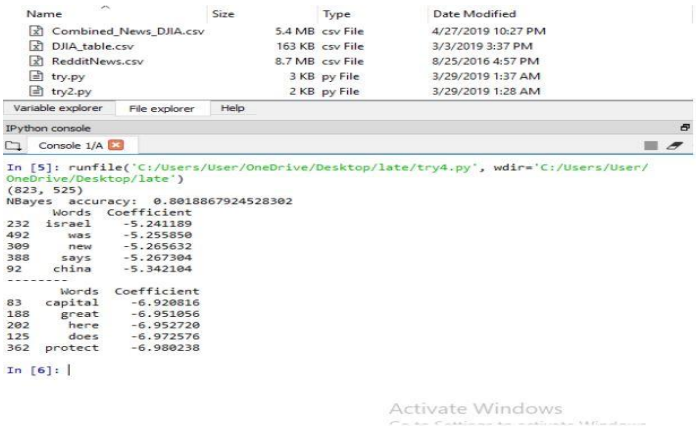
C. LSTM OUTPUT:



D. SVM OUTPUT:



E. NAÏVE BAYES' OUTPUT:



VIII. RESULTS ANALYSIS

From the complete project we can conclude when the features such as news and tweets are taken in consideration then we get a better result with more accurate output. Also among the algorithms used naïve baye's algorithm is the one with the best accuracy. We got the best accuracy for naïve baye's algorithm which was 80.100% when we used news as an attribute. Naïve baye's algorithm also helps us to find out the particular words that are used most in the dataset set and hence having the most impact on the calculation of the accuracy. And same for the words having minimum impact on the calculation of the accuracy. Hence we can conclude that the naïve baye's algorithm is the best possible algorithm for the stock market prediction of our case. The least accuracy is coming for LSTM that is 72.8% and eventually the accuracy of SVM algorithm, Logistic regression and Random forest algorithm is 76.5%, 75.5% and 74.87%.And also using the news as an attribute increased the accuracy of the result.

Sr. No.	Algorithm	Result
1	Random forest algorithm	**74.871%**
2	SVM	**76.501%**

3	LSTM	**72.898%**
4	Naïve Bayes'	**80.188%**
5	Logistic Regression	**75.557%**

IX. CONCLUSION AND FUTURE WORK

This result can be used along with the live news and tweets for creating a dynamic dataset which would help us create a better prediction of the share market prices. Using news as the factor may help us in the more chance of increase in the detecting the fluctuation in the values as the news is one of the greatest factor effecting the change in stock prize as news contain the brief every possible event happened in the previous day and also contain about the company that is their release of product, status, bonds, funds, investments, etc. In this we can use tweets and company reviews so to increase more effective predictions as it can provide more accuracy as the more number factors the higher can we achieve our accuracy.

REFERENCES

1. Stock Price Prediction Using the ARIMA Model (1 Ayodele A. Adebiyi, 2 Aderemi O. Adewumi, 3 Charles K. Ayo),2014 IEEE DOI 10.1109.
2. Detecting Stock Market Manipulation using Supervised Learning Algorithm (Koosha Golmohammadi, Osmar R. Zaiane) march 2015 Research gate.
3. Stock Market Prediction using Machine Learning Techniques(Mehak Usmani, Kamran Raza, Syed Hasan Adil, Syed Saad Azhar Ali).
4. Learning Linear Regression algorithm for stock market data :- http://beancoder.com/linear-regression-stock-prediction/
5. https://www.geeksforgeeks.org/confusion-matrix-machine-learning
6. http://rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier
7. Stock Price Prediction Using the ARIMA Model (1 Ayodele A. Adebiyi., 2 Aderemi O. Adewumi, 3 Charles K. Ayo)
8. Detecting Stock Market Manipulation using Supervised Learning Algorithms (KooshaGolmohammadi, Osmar R. Zaiane)
9. Stock Market Prediction using Machine Learning Techniques (MehakUsmani, Kamran Raza, Syed Hasan Adil, Syed SaadAzhar Ali).
10. Forward Forecast of Stock Price Using Sliding-Window Metaheuristic-Optimized Machine-Learning Regression (Jui-Sheng Chou and Thi-Kha Nguyen).
11. A Decision Support Approach for Online Stock Forum Sentiment Analysis (Desheng Dash Wu, Lijuan Zheng, and David L. Olson).