IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Stock Price Prediction Using Random Forest Method and Twitter Sentiment Analysis

Mr D.B Khadse[1], Ashutosh Ambule[2], Tuba Khan [3], Abhishek Shende[4], Vaibhav Mundhe[5]
Assistant Professor, Department of Computer Science & Engineering[1]
UG Students, Department of Computer Science & Engineering[2,3,4,5]
Priyadarshini Bhagwati College of Engineering, Nagpur, Maharashtra, India

Abstract: *Stock price forecasting could be a vital and thriving topic in financial engineering especially since new techniques and approaches on this matter are gaining ground constantly. Within the contemporary era, the ceaseless use of social media has reached unprecedented levels, which has led to the belief that the expressed public sentiment could be correlated with the behaviour of stock prices. The concept is to acknowledge patterns which confirm this correlation and use them to predict the future behaviour of the assorted stock prices. With little doubt, though uninteresting individually, tweets can provide a satisfactory reflection of public sentiment when taken in aggregate. We develop a system which collects past tweets, processes them further, and examines the effectiveness of varied machine learning techniques like Naive Bayes Bernoulli classification and Support Vector Machine (SVM), for providing a positive or negative sentiment on the tweet corpus. Subsequently, we employ the identical machine learning algorithms to research how tweets correlate with exchange price behaviour. Finally, we examine our prediction's error by comparing our algorithm's outcome with next day's actual close price. Overall, the final word goal of this project is to forecast how the market will behave within the future via sentiment analysis on a collection of tweets over the past few days, also on examine if the idea of contrarian investing is applicable. the ultimate results seem to be promising as we found correlation between sentiment of tweets and stock prices.*

Keywords: Stock Market Prediction, Sentiment Analysis, Twitter, Machine Learning

I. INTRODUCTION

Modern data processing techniques have led to the event of sentiment analysis, [1] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of exchange Prediction Using Machine Learning Approach", ICECA 2017. an algorithmic approach for detecting the predominant sentiment a few product or company using social media data. A positive field for the employment of sentiment analysis has been stock exchange forecasting, a theme undeniably undergoing intense studies nowadays, an excellent volume of knowledge, which contains information about numerous topics, is being transmitted online through various social media, a wonderful example is Twitter, where over 400 million tweets are sent daily. Though each tweet might not be significant as a unit, an oversized collection of them can provide data with valuable insight about the common opinion on a specific subject. Gauging the public's sentiment by retrieving online information from Twitter, will be valuable in forming trading strategies. The proper prediction about the fluctuation of stock prices depends on many factors, and public sentiment is arguably included. Exchange price prediction for brief time windows appears to be a random process. The stock price movement over an extended period of your time usually develops a linear curve. People tend to shop for those stocks whose prices are expected to rise within the near future. The uncertainty within the stock exchange refrain people from investing in stocks. Thus, there's a desire to accurately predict the exchange which may be utilized in a real-life scenario. [2]

The methods wont to predict the stock exchange includes a statistic forecasting together with technical analysis, machine learning positive and predicting the variable exchange. The datasets of the stock exchange prediction model include details just like the terms opening price, the info and various other variables that are needed to predict the article variable which is that the price in a very given day. The previous model used traditional methods of prediction like statistical method with a prediction statistic model. Exchange prediction outperforms when it's treated as a regression problem but performs well when treated as a classification.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

The aim is to style a model that gains from the market information utilizing machine learning strategies and gauge the long run patterns available value development. The Support Vector Machine (SVM) may be used for both classification and regression. It's been observed that SVMs are more employed in classification based issue like ours. The SVM technique, we plot every single data component as some extent in n-dimensional space (where n is that the number of features of the dataset available) with the worth of feature being the worth of a specific coordinate and, hence classification is performed by finding the hyper-plane that differentiates. Predictive methods like Random forest technique are used for the identical. The random forest algorithm follows an ensemble learning strategy for classification and regression. The random forest takes the common of the varied subsamples of the dataset, this increases the predictive accuracy and reduces the over-fitting of the dataset.

II. LITERATURE SURVEY

Survey of securities market Prediction Using Machine Learning Approach

The exchange prediction has become an increasingly important issue within the nowadays. one in every of the methods employed is technical analysis, but such methods don't always yield accurate results.[1]

Impact of monetary Ratios and Technical Analysis on Stock Price Prediction Using Random Forests

The use of machine learning and computing techniques to predict the costs of the stock is an increasing trend.[2]

Stock Market Prediction via Multi-Source

Multiple Instance Learning Accurately predicting the exchange could be a challenging task, but the fashionable web has proved to be a really useful gizmo in making this task easier.[3]

Stock Market Prediction: Using Historical Data Analysis

The stock exchange prediction process is stuffed with uncertainty and might be influenced by multiple factors. Therefore, the exchange plays a very important role in business and finance.[4]

A Survey on securities market Prediction Using SVM

The recent studies provide a well-grounded proof that the majority of the predictive regression models are inefficient in out of sample predictability test.[5]

Predicting Stock Price Direction Using Support Vector Machines

Financial organizations and merchants have made different exclusive models to try and beat the marketplace for themselves or their customers, yet once in a very while has anybody accomplished reliably higher-than-normal degrees of profitability.[6]

Prediction Method supported Support Vector Machines (SVM) and Independent Component Analysis (ICA)

The statistic prediction problem was researched within the work centres within the various financial organisation. The prediction model, which is predicated on SVM and independent analysis, combined called SVM-ICA, is proposed for exchange prediction.[7]

Machine Learning Approach Available Market Prediction

The overwhelming majority of the stockbrokers while making the prediction utilized the specialized, fundamental or the statistical analysis. Overall, these techniques couldn't be trusted completely, so there emerged the necessity to administer a powerful strategy to financial exchange prediction [8]

III. PROBLEM STATEMENT

The aim of this project is to facilitate users with the past, current and future market trends and behavior using Twitter Sentiment Analysis techniques, which involves training of past data to analyze the data patterns. The main objective of this project is to find the best model to predict the value of the stock market. During the process of considering various techniques

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

and variables that must be taken into account, we found out that techniques like random forest, support vector machine were not exploited fully. In, these papers we are going to present and review a more feasible method to predict the stock movement with higher accuracy. This project also presents a machine-learning model to predict the longevity of stock in a competitive market. The successful prediction of the stock will be a great asset for the stock market institutions and will provide real-life. The goal of the project is to use historical stock data in conjunction with sentiment analysis of news headlines and Twitter posts, to predict the future price of a stock of interest. The headlines were obtained by scraping the website, FinViz, while tweets were taken using Tweepy. Both were analyzed using the Vader Sentiment Analyzer.

IV. METHODOLOGY

4.1 Random Forest Algorithm

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithm, it gives good accuracy in the prediction. This is usually used in the classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we are using random forest classifier which has the same hyper parameters as of a decision tree.

4.2 Support Vector Machine Algorithm

The main task of the support machine algorithm is to spot an N dimensional space that distinguishably categorizes the information points. Here, N stands for variety of features. Between two classes of knowledge points, there is multiple possible hyper-planes which will be chosen, the target of this algorithm is to seek out a plane that has maximum margin. Maximizing margin refers to the space between data points of both classes. The benefit related to maximizing the margin is that it provides is that it provides some reinforcement so future data points may be more easily classified. Decision boundaries that help classify data points are called hyper-planes.[9] supported the position of the information points relative to the hyper-plane they're attributed to different classes. The dimension of the hyper-plane relies on the amount of attributes, if the quantity of attributes is 2 then the hyper-plane is simply a line, if the quantity of attributes is three then the hyper-plane is 2 dimensional.

4.3 Naïve Bayes Classifier:

Naïve Bayes classifier is used for sentiment analysis. This algorithm is structured to provide either of the three classes: positive, negative and neutral from the news headlines and twitter tweets. In sentiment analysis we figure out, if text express negative or positive feeling. The basic idea of Naïve Bayes technique is to find the probabilities of classes assigned to texts by rising the joint probabilities of words and classes. To avoid underflow, log probabilities can be issued.

4.4 Xgboost Algorithm

XGBoost algorithm is employed for stock price prediction. After the sentiment analysis process, we combine it to most up-to-date and in trend algorithm to process with stock data to predict stock price. XGBoost could be a most powerful machine learning algorithm today. XGBoost stands for gradient boosted trees which means it's an enormous machine learning algorithm with plenty of parts remember boosting is an ensemble method. XGBoost can automatically handle the missing values. Regularized boosting or prevents overfitting. multiprocessing, tree pruning a number of the features of XGBoost algorithm.

V. DESIGN AND IMPLEMENTATION

5.1 Classification

Classification is an instance of supervised learning where a group is positive and categorized supported a standard attribute. From the values or the information are given, classification draws some conclusion from the observed value. If quite one input is given then classification will attempt to predict one or more outcomes for the identical. some classifiers that are used here for the exchange prediction includes the random forest classifier, SVM classifier.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

A. Random Forest Classifier

Random forest classifier may be a form of ensemble classifier and also a supervised algorithm. It basically creates a group of decision trees, that yields some result, the fundamental approach of random class classifier is to require the decision aggregate of random subset decision tress and yield a final class result supported the votes of the random subset of decision trees.[10]

Parameters:

The parameters included within the random forest classifier are n_estimators which is total number of decision trees, and other hyper parameters like oob-score to work out the generalization accuracy of the random forest, max_features which incorporates the amount of features for best-split. Min_weight fraction_leaf is that the minimum weighted fraction of the accumulation of weights of all the input samples required to be at a leaf node. Samples have equal weight when sample weight isn't provided.

B. SVM Classifier

SVM classifier could be a form of discriminative classifier. The SVM uses supervised learning i.e. a positive training data. The output are hyper-planes which categorizes the new dataset. They're supervised learning models that uses associated learning algorithm for learning models that uses associated learning algorithm for classification and likewise as regression.

Parameters

The tuning parameters of SVM classifier are kernel parameter, gamma parameter and regularization parameter. Kernels are often categorized as linear and polynomial kernels calculates the prediction line. In linear kernels prediction for a replacement input is calculated by the real number between the input and also the support vector. Parameter is thought because the regularization parameter; it determines whether the accuracy of model is increases or decreases. Gamma parameter measures the influence of one training on the model. Low values signifies aloof from the plausible margin and high values signifies closeness from the plausible margin.

VI. Module Identification

The various modules of the project would be divided into the segments as described.

6.1 Data Collection

Data collection may be a very basic module and also the initial step towards the project. It generally deals with the gathering of the proper dataset. The dataset that's to be employed in the market prediction must be wont to be filtered supported various aspects. Data collection also complements to boost the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we are going to be [14] ositive the Kaggle dataset and consistent with the accuracy, we'll be using the model with the information to ositiv the predictions accurately.

6.2 Pre Processing

Data pre-processing may be a a part of data processing, which involves transforming data into a more coherent format, data is typically, inconsistent or incomplete and typically contains many errors. the information pre-processing involves trying out for missing values, trying to find categorical values, splitting the data-set into training and test set and eventually do a feature scaling to limit the range of variables in order that they'll be compared on common environs.

6.3 Training the Machine

Training the machine is analogous to feeding the information to the algorithm to the touch up the test data. The raining sets are wont to tune and fit the models. The test sets are untouched, as a model shouldn't be judged supported 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 hyper parameters just like the number of trees during a random forest. We perform the whole cross-validation loop on each set of hyper parameter values. Finally, we are going to calculate a cross-validated score, for individual sets of hyper parameters. Then, we select the simplest hyper

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

parameters. the thought behind the training of the model is that we some initial values with the dataset so optimize the parameters which we wish to within the model. this can be kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it's divided within the ratio of 80:20 where 80% is for the training set and therefore the rest 20% for a testing set of the information. [5]

6.4 Data Scoring

The method of applying a predictive model to a collection of information is observed as scoring the information. The technique wont to process the dataset is that the Random Forest Algorithm. Random forest involves an ensemble method, which is typically used, for classification and additionally as regression. supported the educational models, we achieve interesting results. The last module thus describes how the results of the model can help to predict the probability of a stock to rise and sink supported certain parameters. It also shows the vulnerabilities of a selected stock or entity.

- Step 1: Loading the most class which is answerable for training and prediction.
- Step 2: processing of live twitter data fetched via respective tweets through API.
- Step 3: Pre-processing of the fetched data which is finished to get rid of special characters, stop 4words and perform tokenization.
- Step 4: Perform sentiment analysis of the obtained new tweet data using naïve bayes Classifier.
- Step 5: Respective company stock data (open, close, adj close) factors taken into consideration together with sentiment analysis details fed to XGBoost algorithm.
- Step 6: Stock price of respective company displaying in between time window of half-hour. Thus, the expected value are going to be obtained.

VII. SOFTWARE REQUIREMENT

- Python
- IDE(Jupyter / Colab)
- Numpy
- Scikit-learn
- Pandas
- Matplotlib
- Pyplot
- Tweepy
- NLTK

VIII. RESULT

	Date	Tweets	Prices	Comp	Negative	Neutral	Positive
0	2022-03-05	Delay unitedAIRLINES flight to Houston at Lib...	40	0.9796	0.038	0.849	0.9796
1	2022-03-04	Never flown before ever in my Life since 88 ...	36	0.9975	0.061	0.803	0.9975
2	2022-03-03	Every time I fly United Airlines theres a pro...	40	0.9969	0.042	0.836	0.9969
3	2022-03-02	CHARITY MEANS LOVE Part 55httpstcoLtVou02xgG...	42	0.996	0.048	0.833	0.996
4	2022-03-01	Change your trip due to Covid is a big scam b...	41	-0.9002	0.12	0.776	-0.9002
5	2022-02-28	UnitedAirlines doesnt refund their Fully Refu...	44	0.9946	0.057	0.789	0.9946
6	2022-02-27	Vintage United Airlines Playing Cards In Case...	40	0.6934	0.137	0.689	0.6934
7	2022-02-26	RT BlueQueenBree United Airlines unitedAIRLIN...	40	0.9658	0.1	0.762	0.9658

Figure 1: Fetching live tweets

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Date	adj_close_price	Date	Comp	Negative	Neutral	Positive
2007-01-01	12469	2007-01-01	-0.9814	0.159	0.749	-0.9814
2007-01-02	12472	2007-01-02	-0.8179	0.114	0.787	-0.8179
2007-01-03	12474	2007-01-03	-0.9993	0.198	0.737	-0.9993
2007-01-04	12480	2007-01-04	-0.9982	0.131	0.806	-0.9982
2007-01-05	12398	2007-01-05	-0.9901	0.124	0.794	-0.9901
...	...	...	...	...	...	...
2016-12-27	19945	2016-12-27	-0.9898	0.178	0.719	-0.9898
2016-12-28	19833	2016-12-28	0.2869	0.128	0.763	0.2869
2016-12-29	19819	2016-12-29	-0.9789	0.138	0.764	-0.9789
2016-12-30	19762	2016-12-30	-0.995	0.168	0.734	-0.995
2016-12-31	19762	2016-12-31	-0.2869	0.173	0.665	-0.2869

3653 rows × 6 columns

Figure 2: Fetching past data

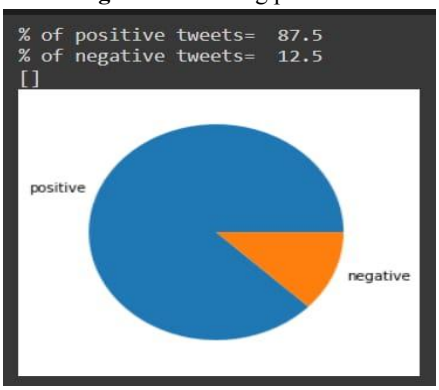

Figure 3: Percentage of positive and negative tweets

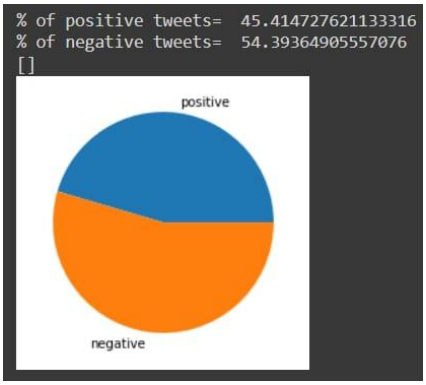

Figure 4: Percentage of Positive and Negative past data

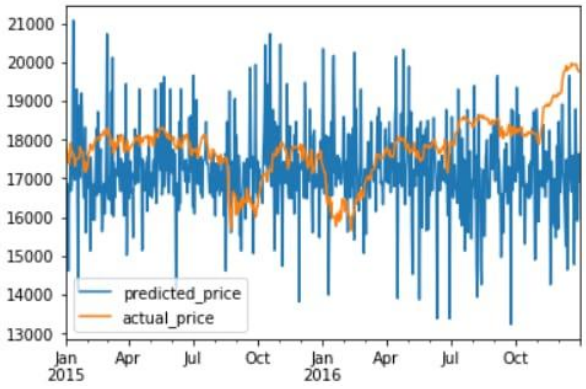

Figure 5: Plotting predicted and actual data of Stocks

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

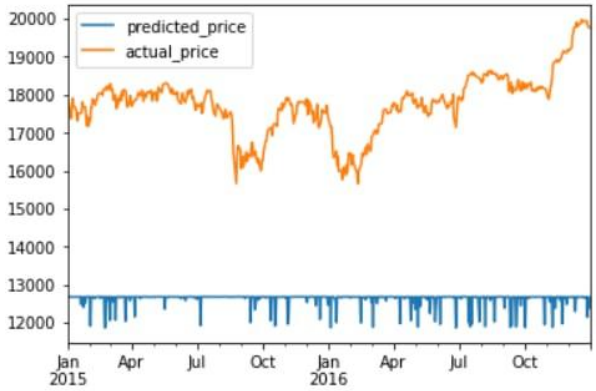

Figure 6: Plotting Predicted and actual data of Stocks

Figure 7: Plotting Predicted and Actual data of Stocks

IX. CONCLUSION

By measuring the accuracy of the various algorithms, we found that the foremost suitable algorithm for predicting the value of a stock supported various data points from the historical data is that the random forest algorithm. The algorithm are going to be a good asset for brokers and investors for investing money within the securities market since it's trained on a large collection of historical data and has been chosen after being tested on a sample data. In this paper, we investigated whether public sentiment, as measured from tweets, is correlated or even predictive of stock values and specifically for 16 of the foremost popular tech companies to keep with Yahoo! Finance. Our results show that changes within the general public sentiment can affect the exchange, which implies that we are going to indeed predict the stock exchange with high chances.

REFERENCES

[1] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of exchange Prediction Using Machine Learning Approach",ICECA,2017.
[2] Loke. K.S. “Impact of monetary Ratios And Technical Analysis On Stock Price Prediction Using Random Forests”, IEEE,2017.
[3] Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market . Prediction via Multi-Source MultipleInstanceLearning.” IEEE 2018.
[4] VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.
[5] Jabaseeli, A. Nisha, and E. Kirubakaran. “A Survey on Sentiment Analysis of (Product) Reviews.” International

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Journal of Computer Applications 47.11, 201

[6] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Languages in Social Media. Association for linguistics, 2011

[7] a. Mittal and a. Goel. "Stock Prediction Using Twitter Sentiment Analysis." Tomx.Inf. Elte.Hu, (June), 2012.

[8] a. Mittal and a. Goel. Stock Prediction Using Twitter Sentiment Analysis. Tomx.Inf.Elte.Hu, (June), 2012. ˚

[9] F. . Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors.

[10] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Lrec, pages 1320–1326, 2010.