Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017)
IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9

Stock Market Prediction using Data Mining Techniques

Sahaj Singh Maini[1]
SCOPE, VIT, Vellore, India
sahajsingh.maini2014@vit.ac.in

Govinda.K[2]
SCOPE, VIT, Vellore, India
kgovinda@vit.ac.in

***Abstract*—Stock market prediction has been an area of interest for investors as well as researchers for many years due to its volatile, complex and regularly changing in nature, making it difficult to make reliable predictions This paper proposes an approach towards prediction of stock market trends using machine learning models like Random Forest model and Support Vector Machine. The Random Forest model is an ensemble learning method that has been an exceedingly successful model for classification and regression. Support vector machine is a machine learning model for classification. However, this model is mostly used for classification. These techniques are used to forecast whether the price of a stock in the future will be higher than its price on a given day, based on historical data while providing an in-depth understanding of the models being used.**

Keywords—Machine Learning, Random Forest Model, Stock Market, Support Vector Machine.

I. Introduction

This paper proposes the use of machine learning techniques for evaluating data concerned to the stock market using a novel method for prediction of stock prices to minimize the risk of investment in a stock market. Natural Language Processing is used on content derived from news articles and other relevant sources along with an Ensemble learning model called Random Forest model and Support Vector Machine. The forecasting problem of stock price is treated as a classification problem to make better decisions.

Stock market prediction has gained monumental prevalence in financial markets globally. The ability to forecast the direction of a stock price or an index is very important for various purposes. A few of them being a potential reduction in the risk of investment for the investors and supporting in identifying opportunities for speculators seeking to make profits by investing in stock indexes.

In statistics and machine learning the problem that involves deciding which category a new observation belongs to among a set of categories using a dataset for training which contains observations whose category membership is mentioned. In machine learning, classification is regarded as an example of supervised learning in which learning takes place using a training dataset containing observations with correctly identified classes provided in advance.

Natural Language Processing is a way for the computer program to interpret all human derived languages. It helps in analyzing the text and making sense of what is presented to the program in the form of human language. This paper proposes the use of sentiment analysis of text from the content derived from news sources for understanding the natural language and inferring whether the message it provides is positive, negative or neutral towards its influence on the stock price.

Random Forest model is an ensemble learning method that creates multiple decision trees based on random subsets of data during the training and outputs the mode of classes that have resulted from these decision trees using an input for classification. Random Forest Model involves a general technique called Bootstrap Aggregation which leads to better model performance by decreasing the variance.

Support Vector Machine is a machine learning algorithm used for classification and regression. Support Vector Machine is predominantly used for classification problems, it is a maximum margin classification algorithm. In Support Vector Machine each data item is plotted on an n-dimensional space, where n is equal to the number of features and then an attempt is made to classify these points by finding the most suitable plane that differentiates these points better than all the other planes.

II. Literature Review

In recent times, sentiment Analysis has been used in multiple areas such as blogging websites, review websites, online retail etc.. Sentiment analysis has been a very important social media analytics tool and has been effectively used by

E-commerce websites like Snapdeal and Flipkart to filter irrelevant reviews. Sentiment analysis at present plays a vital role in customer service, management of brand reputation and business intelligence. It has also been influential in politics by helping political strategists determine the public opinion on the internet. It played a crucial role in Barack Obama's campaign in 2011 where sentiment analysis was used to predict the responses to campaign messages.

The Efficient Market Hypothesis is a basic rule in finance that contradicts the ability of prediction algorithms to determine the future trends in the stock market. According to the Efficient Market Hypothesis, the prices of stocks in the market majorly depend upon information which is new and follows a random pattern. It indicates that if someone identifies a method that can analyze the historical data to predict the prices in the future, the whole market would eventually know about it which would lead to the prices of stocks being corrected. Although this hypothesis is widely accepted as a central paradigm guiding the markets, There are numerous researchers who have rejected this hypothesis and have attempted to draw out patterns of market's behavior with respect to the external stimuli.

There are various algorithms that have been used in stock market prediction research like linear regression, logistic regression, neural networks and naive bayes classifier. The conditions considered to make a prediction being quotes related to commodity prices. There have also been considerable attempts made for reliable prediction based upon results from textual analysis of Twitter feeds.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X.[17] proposed a technique for stock market prediction on the basis of sentiments of Twitter feeds which was experimented on S&P 100 index. A continuous Dirichlet Process Mixture model was used to learn the daily topic set. Stock index and Twitter sentiment time-series were then regressed to make a prediction.

Mittal, A., & Goel, A.[18] applied sentiment analysis on Twitter feeds to discover the interrelationship among "public sentiment" and the "market sentiment". Data retrieved from Twitter is used to predict the public mood. A Self Organizing Fuzzy Neural Network is used on predicted mood from the Twitter feeds and Dow Jones Industrial Average values from the previous day to predict the movement of the stock market.

Gidofalvi, G., & Elkan, C.[19] use the data from financial news articles to predict short-term movement of stock price. The movement of the stock price is classified into three different classes representing three different directions, namely "up", "down", and "unchanged". A naive Bayesian text classifier is used to predict the direction of the movement of stock price by deriving a set of indicators from the textual data retrieved from various financial news articles.

III. Methodology

The dataset consists of data from the timeline 2000 to 2016 which is used for the prediction of Dow Jones Industrial Average Index. The data in this dataset has been retrieved from three different sources-

1. News data consisting of historical news from Reddit World News Channel.Only the top twenty-five headlines of each day are considered.
2. The stock data for Dow Jones Industrial Average (DJIA) is considered for the timeline of 2008 to 2016.Yahoo finance is used for compilation of the stock data.
3. Data from the Guardian's restful news API from the year 2000 to 2008.

The top twenty-five headline for a day are arranged in a row along with a 'label' column holding '0' if these headlines lead to the decline or lack of change in the stock price the next day and holding '1' if these headlines lead to increase in the stock price the next day. The data retrieved from the top 25 headlines and the label is compared to get an accurate prediction.

The text in the training dataset is converted into numeric values using 'Bag of Words' model. This model basically counts the number of occurrences of each word in the provided data. The values of number occurrences of the words in the data are used to form feature vectors derived from the model.

Only using the basic 'Bag of Words' model leads to the formation of feature vectors only based upon the frequency of words without considering the order in which they occur. Therefore, skip-gram model or an n-gram model is used which stores the words in the order in which they occur in the data. The n-gram refers to n words in a given order. This paper uses a unigram(n=2) and a bigram(n=2) model which stores the count of sets of two words in order. The vectors derived from these methods are used to train the machine learning models.

Random Forest model is a supervised classification algorithm. It creates multiple decision trees based on random subsamples from the data, each capable of producing a result when presented with values for prediction. The higher the number of trees, the higher is the accuracy and less is the risk of overfitting compared to other models.

The Random Forest model produces 'n' number of trees as weak classifiers and merges all the trees into a forest. When the Random Forest model is used for regression the mean of resulting values from all the decision trees is the resulting prediction value and when it is used for classification, the resulting class is the mode of the resulting classes from the decision tree. To classify a new object each tree gives a classification that can be described as a vote and the class with the highest votes is chosen as the class of the new object.

Assume a training set -

$A=\{(x_1, y_1),(x_2, y_2), .. (x_n, y_n)\}$

The building of a combination of decision trees takes place using the given dataset A , The elements of which, individually are weak classifiers.

$h=\{h_1(x), h_2(x), . . ., h_n(x)\}$, where $h_i(x)$ represents the i th decision tree. Together these form a random forest.

$\Theta_i=(\theta_{i1}, \theta_{i2}, . . ., \theta_{ip})$, these are the parameters which influence the creation of decision trees such as the variables which are to be split in a node, structure of the tree to be constructed, etc.

Each decision tree classifier can be represented as

$h_i(x) = h(x/\Theta_i)$.

The general technique incorporated in Random Forest model to increase its stability and accuracy is called Bootstrap Aggregation which is also known as bagging. In this technique, when presented with training input along with the responses, random subsamples are chosen continuously from the data presented to us with replacement to build decision trees.

Given training set- $X= x1, x2, x3,. . . , xn$

Given Responses relative to the above training set-

$Y=y1, y2, y3,. . ., yn$

for i in 1 to n

1. Choose Subsamples from the training set with replacement from both X and Y calling these, Xi, Yi.
2. Create a decision tree Fi for regression or classification using Xi and Yi.

After the creation of multiple decision trees the prediction for unseen sample of data, x' can be made by calculation of the average of prediction values from all the trees.

$$F'=(1 \div B) \sum_{i=1}^{B} Fi(x')$$

And in case of classification, by taking the mode of predicted classes from all the trees.

An estimate of uncertainty in the predictions can be made by calculating the standard deviation of the prediction values obtained from each decision tree.

$$\sigma = \sqrt{(\sum_{i=1}^{B} (Fi(x') - F'))^2 \div (B-1)}$$

This technique is widely used to decrease the variance of the model without increasing the bias.

Decision trees use various methods to decide the node on which the split occurs. Decision trees commonly decide which node to split based on the calculation of gini index, chi-square, Information gain and also use reduction in variance algorithm for regression problems. This paper uses information gain as a factor to make the decision of choosing the node where the split occurs.

The degree of disorganization of a system is called entropy of the system.The value of entropy on calculating entropy for a homogeneous system is always zero. If the presented sample is divided equally then the entropy for the system is supposed to be one.

The formula required to calculate the entropy is-

$Entropy= -p \log_2(p) - q \log_2(q)$

'p' and 'q' refer to the probability of success or failure respectively in that node. The split with the lowest entropy is chosen compared to the parent node and other splits to make a split.

Steps on building and using a Random Forest Model,

1. Given there are n cases in the training dataset. From these n cases, sub-samples are chosen at random with replacement. These random sub-samples chosen from the training dataset are used to build individual trees.
2. Assuming there are k variables for input, a number m is chosen such that m < k. m variables are selected randomly out of k variables at each node. The split which is the best of these m variables is chosen to split the node. The value of m is kept unchanged while the forest is grown.
3. Each tree is grown as large as possible without pruning.
4. The class of the new object is predicted based upon the majority of votes received from the combination of all the decision trees.

Another model being used in this paper is Support Vector Machine. Support Vector Machine is a supervised machine learning algorithm which after plotting the data items in an n-dimensional plane performs classification by dividing the data points into two classes using a plane. The plane which divides the data points better than all the other planes is chosen. This plane is called a hyperplane.

Support Vector Machine is a maximum margin classification algorithm. In a maximum margin classification, a hyperplane is defined as a plane that divides the input variable space into two separate classes. Here, the margin refers to the region between the closest data points and the plane. This length of the margin from the plane is calculated as the perpendicular distance from the closest data point to the plane. The plane with the largest margin is chosen to be the hyperplane or the plane that differentiates the two classes at best. The two closest points to the plane are called support vectors, these vectors define a hyperplane.

When provided with 'n' points say,

$(\bar{x}_1, y_1), (\bar{x}_2, y_2), \ldots, (\bar{x}_n, y_n)$, where y_i is either 1 or -1 and $\bar{x}_i$ is a k-dimensional real vector.

D=$\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$

The hyperplane that divides the above-mentioned points into two different classes can be represented as

$\overline{w} \cdot \bar{x} + b = 0$, where $\overline{w}$ is a normal vector to the plane that divides the points into two classes.

Hard margin Support Vector Machine results in a model that allows zero errors. Such a model performs well in linearly separable problems where it is possible to successfully classify the data points into two classes. In linearly separable problems, Two hyperplanes are selected which are parallel and farthest from each other. The region between these hyperplanes can be identified as the margin and the plane that lies in the middle of the two hyperplanes is chosen as the optimal plane that divides the points into two classes. This plane is known as maximum margin hyperplane.

The equations for the two hyperplanes can be given by,

$\overline{w} \cdot \bar{x} + b = 1$ and $\overline{w} \cdot \bar{x} + b = -1$

The expression for maximum margin in linearly separable Support Vector Machine is given by,

$$\text{margin} = arg_{x \epsilon D} min(|x.w + b| / \sqrt{\sum_{i=1}^{n} w_i^2})$$

The distance between the two parallel hyperplanes chosen to identify the maximum margin hyperplane is given by,

$2 / ||\overline{w}||$

The goal of implementing Support Vector Machine is to correctly classify the data.Therefore, The given data points are inhibited from falling in the margin.

$$y_i = +1 \; when \; \overline{w} \cdot \bar{x}_i + b \geq +1$$

$$y_i = -1 \; when \; \overline{w} \cdot \bar{x}_i + b \leq -1$$

Hence from the above two inequalities,

$$y_i(\overline{w} \cdot \bar{x}_i + b) - 1 \geq 0$$

A Quadratic Programming formulation of Support Vector machine having hard margin can be given by,

$$min_{w,b} (||w|| /2) \text{ , such that } y_i(\overline{w} \cdot \bar{x}_i + b) \geq 1$$
$$\forall i \epsilon \{1, 2, 3, \ldots n\}$$

In soft margin Support Vector Machine, the model allows for some data points belonging to one class to appear on the other side of the boundary. The soft margin is used when the data is not linearly separable. In this case, slack variables are introduced for each $\underline{x}_i$. The Quadratic Programming formulation for soft margin Support Vector Machine can be given by,

$$min_{w,b} (||w|| /2) + C \sum_{i=1}^{n} \xi_i \text{ such that } y_i(\overline{w} \cdot \bar{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

$$\forall i \epsilon \{1, 2, 3, \ldots n\}$$

Here, the value of C influences the number of data points that are allowed to be placed within the margin. As during the process of identifying the hyperplane all the data points that lie within the margin called support vectors affect the position of the hyperplane, the value of C has an influence on the number of support vectors used for identifying the hyperplane thereby influencing the number of data points allowed within the margin.

The sensitivity of the algorithm towards the data being used for training increases with the decrease in the value of C which leads to increasing variance and decreasing bias and the sensitivity of the algorithm decreases with increase in the value of C which leads to decreasing variance and increasing bias.

In the case of problems where the data is linearly inseparable, a kernel is used to nonlinearly map the data received as input into higher dimensions. The resulting mapping after using a kernel is then linearly separable. Kernels are functions that transform a low dimensional input space into a higher dimensional space. The Transformation of data into a feature space provides a possibility to define a measure of similarity based on the dot product.

The Quadratic formulation of Support Vector Machine where a kernel is used is given by,

$$min_{w,b}\ (\ ||w||/2\ )\ +\ C\sum_{i=1}^{n}\xi_i\ ,$$

such that $y_i(\ \overline{w}\ \varphi(\overline{x}_i)+b)\geq 1\ -\xi_i$ and

$\xi_i \geq 0 \quad \forall i \epsilon \{1,2,3,\ \ldots\ n\}$.

Here ϕ refers to the kernel function that maps the input data to higher dimensional space.

The mapping by a kernel is defined as,

$$k(\overline{x}_i\ ,\ \overline{x}_j)\ =\ \varphi(\overline{x}_i)\ .\varphi(\overline{x}_j)$$

$\overline{w}$ in the transformed space can be defined as,

$$\overline{w} = \Sigma_i \alpha_i\ y_i\ \ \varphi(\overline{x}_i)$$

Few common kernels have been mentioned below-

Polynomial - The polynomial kernel provides the possibility of lines which are curved in the input space. For two samples represented as feature vectors in a input space,vectors $\overline{x}_i$ and $\overline{x}_j$, the polynomial kernel can be defined as

$$k(\overline{x}_i\ ,\ \overline{x}_j)\ =\ (\overline{x}_i\ .\ \overline{x}_j)^d$$

$$k(\overline{x}_i\ ,\ \overline{x}_j)\ =\ (\overline{x}_i\ .\ \overline{x}_j\ +\ 1)^d$$

Gaussian Radial Basis Function - Radial Function is local and has the capability to create complex regions within a feature space.For two samples represented as feature vectors in a input space,vectors $\overline{x}_i$ and $\overline{x}_j$, the gaussian radial function can be defined as

$$k(\overline{x}_i\ ,\ \overline{x}_j)\ =\ exp(-\ \left|\left|x_i - x_j\right|\right|^2 \div 2\sigma^2\ )$$

Exponential Radial Basis Function - The exponential radial basis function is function whose value is depended upon the distance from the origin. Generally Euclidian distance function is used for calculating the distance.For two samples represented as feature vectors in a input space, vectors $\overline{x}_i$ and $\overline{x}_j$, the polynomial kernel can be defined as

$$k(\overline{x}_i\ ,\ \overline{x}_j)\ =\ exp(-\ \left|\left|x_i - x_j\right|\right|\ \div 2\sigma^2\ )$$

IV. Results

The evaluation of the results from the models trained using the given data can be presented using a confusion matrix, precision ,recall, F1 score and accuracy classification score.A confusion matrix is generally used in describing the performance of the model for classification on test data for which the true values are familiar. Precision can be defined as the fraction of number of instances which are relevant over the number of all the instances which have been retrieved. Recall is the fraction of a number of instances that are relevant and have been retrieved over the total number of instances that are relevant.

F1 score can be defined as,

$$F1 = 2\times(precision \times recall\ )/(precision\ +\ recall)$$

Here support refers to the number of samples that lie in that particular class.

Random Forest Model using 1-gram-

For Random Forest Model using 1-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	145	41
1	18	174

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	0.89	0.78	0.83	186
1	0.81	0.91	0.86	192
Avg / Total	0.85	0.84	0.84	378

The accuracy classification score is equal to 0.843915343915 .

Random Forest Model using 2-gram-

For Random Forest Model using 2-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	140	46
1	6	186

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	0.96	0.75	0.84	186
1	0.80	0.97	0.88	192
Avg / Total	0.88	0.86	0.86	378

The accuracy classification score is equal to 0.862433862434 .

Linear Support Vector Machine Model using 1-gram-

For Linear Support Vector Machine using 1-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	151	35
1	32	160

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	0.83	0.81	0.82	186
1	0.82	0.83	0.83	192
Avg / Total	0.82	0.82	0.82	378

The accuracy classification score is equal to 0.822751322751 .

Linear Support Vector Machine Model using 2-gram-

For Linear Support Vector Machine using 2-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	160	26
1	32	160

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	0.83	0.86	0.85	186
1	0.86	0.83	0.85	192
Avg / Total	0.85	0.85	0.85	378

The accuracy classification score is equal to 0.846560846561 .

Nonlinear Support Vector Machine Model using Gaussian Radial Basis Function and 1-gram-

For Nonlinear Support Vector Machine using 1-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	130	56
1	0	192

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	1.00	0.70	0.82	186
1	0.77	1.00	0.87	192
Avg / Total	0.89	0.85	0.85	378

The accuracy classification score is equal to 0.851851851852 .

Nonlinear Support Vector Machine Model using Gaussian Radial Basis Function and 2-gram-

For Nonlinear Support Vector Machine using 2-gram model for analysis of text, the confusion matrix is given by

Predicted- Actual-	0	1
0	130	56
1	0	192

The precision, recall and F1 score are

	Precision	Recall	F1-score	Support
0	1.00	0.70	0.82	186
1	0.77	1.00	0.87	192
Avg / Total	0.89	0.85	0.85	378

The accuracy classification score is equal to 0.851851851852 .

V. Conclusion

This study aims to predict the direction of stock market trends in the future. In this study, we have performed predictive analysis on Dow Jones Industrial Average Index, which consists of thirty companies and is majorly owned by S&P Global. This is crucial to the traders as such analysis can influence the decision making with regard to buying or selling an instrument in a positive manner. We have discussed two statistical machine learning models, namely Random Forest Model and Support Vector Machine which are used to provide a reliable prediction of stock market trends based on historical data. On the basis of the results obtained, we can say that both the models exhibited notable performance in predicting the direction of the stock index. The Random Forest model using a 1-gram model for text analysis produced an accuracy of 84.3% and on using a 2-gram model produced an accuracy of 86.2%. The linear Support Vector Machine using 1-gram model and 2-gram model for text analysis produced predictions with an accuracy of 82.2% and 84.6%, while the nonlinear Support Vector Machine produced predictions with an accuracy of 85.1% for both 1-gram and 2-gram models. We have observed that the Random Forest Model outperforms the Support Vector Machine while using the given dataset. However, Support Vector Machine is a favorable substitute for financial forecasting when the data is in time-series format. There can be other factors that can influence the prediction performance of the models being used like the selection of optimum parameters that remain to be compelling topics for further research.

References

[1] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[2] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

[3] Thissen, U., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). Using support vector machines for time series prediction. *Chemometrics and intelligent laboratory systems*, *69*(1), 35-49.

[4] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

[5] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

[6] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

[7] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

[8] Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[9] Du, W., & Zhan, Z. (2002, December). Building decision tree classifier on private data. In *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14* (pp. 1-8). Australian Computer Society, Inc..

[10] Weston, J., & Watkins, C. (1999, April). Support vector machines for multi-class pattern recognition. In *ESANN* (Vol. 99, pp. 219-224).

[11] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth*

annual workshop on Computational learning theory (pp. 144-152). ACM.

[12] Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Machine learning: ECML 2004*, 39-50.

[13] Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP*, *12*, 231-238.

[14] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. *ACL (2)*, *2013*, 24-29.

[15] Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Stanzford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentiment Analysis. pdf)*, *15*.

[16] Gidofalvi, G., & Elkan, C. (2001). Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*.