Stock Market Prediction Using Machine Learning

P.Poovizhi[1], S.L.Gayathri[2], M.Indhuja[3], R.K.NaveenKumar[4]
[1]Assistant professor, Dept of Information Technology
[2, 3, 4]Dept of Information Technology
[1, 2, 3, 4] SNS College of Engineering Coimbatore India.

Abstract*- Stock market prediction is mainly used for investors as well as researchers for many years. It is used based on its complex, volatile and making it difficult to make reliable predictions, regularly changing in nature. This Proposed stock prediction system uses machine learning models like Random Forest model and Support Vector Machine to provide an approach towards prediction of market trends. Support vector machine is a machine learning model for classification and analysis. However, this model is mostly used for classification. The Random Forest model is an ensemble learning method that has been an exceedingly successful model for classification and regression. 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, Data Pre-processing, Data Mining, Dataset, Stock, Stock Market.

I. INTRODUCTION

A stock also know as an share more commonly in general represents ownership claims on business by a particular individual or a group of people. The stock market is basically an aggregation of various buyers and sellers of stock. The prediction process is expected to be robust, accurate and efficient. The system must work according to the real-life scenarios and should be well suited to real-world settings. The system is also expected to take all the variables that might affect the stock's value and performance of the stock. In 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. The trend in stock market predictions is not a new thing yet this issue is kept being discussed by the various organizations. Being able to predict accurately the longer term financial outcome is like earning pile. Proposed work analysis the problem in an academic way which provides a different ways to predict on the market trend. By using proposed system for develop a financial data predictor program there will be dataset storing all historical stock prices and data will be treated as training sets for the program. To predict the NSE of stock prediction, this Paper aims at developing a program, which serves best solution for accurate predicted stock result. Random Forest model 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. It is also called as an ensemble learning method. Random Forest Model involves a 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. 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 SURVEY

Using support vector machines for time series prediction

Thissen, U et all Financial market prediction, electric utility load forecasting, weather and environmental state prediction, and reliability forecasting uses the technique called Time series prediction method. The system models and time series data generates processes are generally complicated for these applications and the models for these systems are usually not known a priori. exact and impartial estimation of the time series data given by these systems datasets cannot be achieved using linear techniques which is an well known format, and thus in an advanced time series prediction algorithms the estimation process required. This paper provides information about the usage of novel machine learning approach: support vector machines (SVM). SVM have the ability of finding the accurate forecast time series data when the underlying system processes are typically nonlinear, non-stationary and not defined a-priori. To perform better than other non-linear techniques including neural-network based non-linear prediction techniques such as multi-layer perceptrons therefore SVMs techniques can be used since it is proved method. The ultimate goal is to provide the reader with deep understanding into the applications using SVM for time series prediction, to give a brief tutorial on SVMs for time series prediction, to outline some of the advantages and challenges in using SVMs for time series prediction, and to provide a origin for the reader to locate books, technical journals, and other online SVM research resources.

[1] Thissen, U., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). C hemometrics and intelligent laboratory systems, 6 9(1), 35-49.

Support vector machines

Hearst, M et all The issue for collecting the essays should help us to make familiar our readers with this interesting new training skill in the Machine Learning stable. Bernhard Scholkopf in an initial overview, notice that the particular advantage of support vector machine (SVMs) over other learning algorithms is that it can be analyzed theoretically using concepts from computational learning theory, and at an equivalent time are able to do good performance when applied to world problems.

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

Building decision tree classifier on private data

Du, W., & Zhan, Z et all Alice and bob defines there idea about how to build a decision tree classifier for the following scenario : a database is vertically partitioned into two parts one part proposed by Alice and the other part proposed by Bob. Alice and Bob build a decision tree classifier based on dividing the database, but due to some privacy concern, they wants to reveal their private pieces to the other party and present a protocol that allows a classifier building without having to compromise their privacy concern. The proposed system use third-party server and built over a useful building block, the scalar product protocol. This protocol gives a solution to the scalar product protocol is more efficient than any existing solutions and how to classify on the vertically partitioned data. They also developed a method that allows them to build a decision tree classifier based on their joint data.

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

A training algorithm for optimal margin classifiers

Boser, B et all A training algorithm that maximizes the margins between the training patterns and the decision boundary is presented. This technology is applicable to a wide variety of the classification functions that including perceptrons, polynomials, and Radial Basis Functions. The successful number of parameters is adjusted automatically to match the complexity of the problem. The solution is shows a linear combination of supporting patterns. The subset of the training patterns are closest to the decision boundary. Bounds on the generalization performance based on the leave-one-out cross validation method and the VC-dimension are given. Tested results on the optical character recognition problems demonstrate the good generalization obtained when compared with other learning algorithms.

[4] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July), In P roceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM

Stock Market Prediction using Data Mining Techniques

Sahaj Singh Maini et all Stock market prediction it is mainly used for investors as well as researchers for many years. It is used based on its complex, volatile and making it difficult to make reliable predictions, regularly changing in nature. This Proposed stock prediction system uses machine learning models like Random Forest model and Support Vector Machine to provide an approach towards prediction of market trends. Support vector machine is a machine learning model for classification and analysis. However, this model is mostly used for classification. The Random Forest model is an ensemble learning method that has been an exceedingly successful model for classification and regression. 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.

[5] Sahaj Singh Maini, SCOPE, VIT, Vellore, India; Govinda.K, SCOPE, VIT, Vellore, India; Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part
Number : CFP17M19-ART , ISBN : 978-1-5386-1959-9

III. EXISTING SYSTEM

Sentiment Analysis has been used in multiple areas such as blogging websites, review websites, online retail, E-Commerce websites to filter relevant data etc. 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.

A. Drawbacks

- However, sentiment analysis cannot be applied in stock prediction, in needs strong fundamental evaluation on reading the price behavior and predicting the price values.
- 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.

IV. PROPOSED SYSTEM

Proposed system studies stock market prediction using machine learning techniques for evaluating the historical data using a novel method. Proposed work uses machine learning algorithm namely Random Forest model and Support Vector Machine.

A. RANDOM FOREST MODEL

- Random Forest model it 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. It is called as ensemble learning method.
- 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.
- 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.
- Each tree is grown as large as possible without pruning.

B. SUPPORT VECTOR MACHINE

- SVM is a machine learning algorithm mainly used for classification and regression.
- NSE stock is considered for stock prediction, the dataset collected from 2000 to till date is used.
- The forecasting problem of stock price is treated as a classification problem to make better decisions.
- 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 hyper-plane. Support Vector Machine is a maximum margin classification algorithm. In a maximum margin classification a hyper-plane 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 hyper-plane 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 hyper-plane.

C. PRICE FORECAST

Price forecast is done for 5 days using machine learning techniques such as Decision tree and random forest. The result is compared with the score value to identify the accuracy value and plotted.

- df['label']=df[forecast_col].shift(forecast_out)
- X=np.array(df.drop(['label'],1))
- X = preprocessing.scale(X)
- X_lately = X[-forecast_out:]
- X = X[:-forecast_out]

The X lately variable contains the most recent features, which are going to predict against. As you see, defining a classifier, training, and testing was all extremely simple. The forecast set is an array, showing that not only could you just seek out a single prediction, but you can seek out many at once.

D. Advantages of Proposed System

- To achieve good accuracy
- Used for better decisions for investors.

V. SYSTEM IMPLEMENTATION

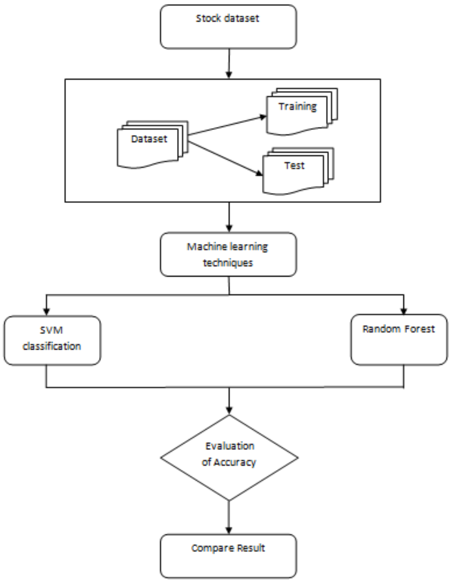


Fig 1 System Architecture

Fig 1 explains that the NSE is an online community used for predictive modeling and predictive modeling. It contains dataset of different fields, which is contributed by data miners. To create the best models for predicting and depicting the information various data scientist is used. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges.

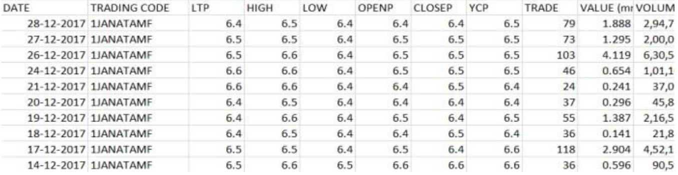
The first step is to convert raw data into processed data. This can be done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step process is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. The feature extracted from an initial state of measured data and builds derived values or features. These features can be informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction may be a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for simple management, while still precisely and totally depicting the first informational collection.

The feature extraction process is followed by a classification process wherein the info that was obtained after feature extraction is split into two different and distinct segments. Classification is the main issue of recognizing to which set of categories a new observation belongs. The training data set is specifically used to train the model whereas the test data is used to predict the accuracy of the model. The splitting process can be done in a way that training data maintain a higher proportion than the test data. The collection of random decision trees to analyze the data is done using random forest model. In layman terms specific attributes in the data are been looked from the total number of decision trees in the forest and in a cluster of the decision trees. This is known as data splitting. In this case, since the end goal of our proposed system is to predict the price of the stock by analyzing its historical data.

EXPERIMENTAL RESULTS

The various modules of the Paper stock market prediction would be divided into the segments as described.

Data Collection



DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mr	VOLUM
28-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.5	79	1.888	2,94,7
27-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.5	6.5	73	1.295	2,00,0
26-12-2017	1JANATAMF	6.5	6.6	6.4	6.5	6.5	6.5	103	4.119	6,30,5
24-12-2017	1JANATAMF	6.6	6.6	6.4	6.5	6.5	6.5	46	0.654	1,01,1
21-12-2017	1JANATAMF	6.6	6.6	6.4	6.4	6.5	6.4	24	0.241	37,0
20-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.4	37	0.296	45,8
19-12-2017	1JANATAMF	6.4	6.6	6.4	6.5	6.4	6.5	55	1.387	2,16,5
18-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.5	6.4	36	0.141	21,8
17-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.4	6.6	118	2.904	4,52,1
14-12-2017	1JANATAMF	6.5	6.6	6.5	6.6	6.6	6.6	36	0.596	90,5

The data collection process involves the selection of quality data for analysis. Data collection model use NSE INDIA stock dataset taken from nse.com for machine learning implementation. The job of a data analyst is to find out the ways for sources of collecting relevant and comprehensive datasets, interpreting it, and analyzing results with the help of statistical techniques.

Data Visualization

A large amount of information represented in graphic form is easier to understand and analyze the training dataset. Some companies specify that a knowledge analyst must skills to make slides, diagrams, charts, and templates the training dataset. In our approach, stock rate from 200 to till dare is shown as data visualization part.

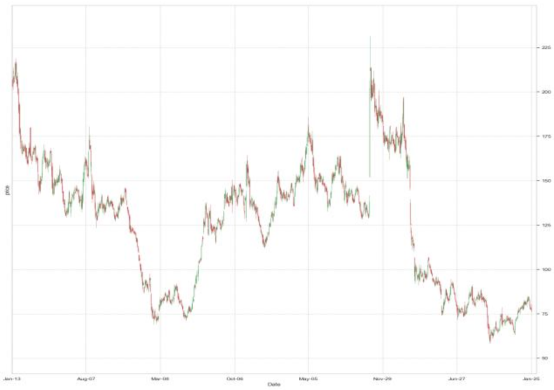


Fig 2: Data Visualization of Intrusion detection rate

Data Preprocessing

The purpose of preprocessing is to convert data into a form that matches machine learning. Structured and clean data allows a knowledge scientist to urge more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Model Training

After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entered the "feeding" the algorithm with training dataset. An algorithm will process data and input, output a model that is able to find a target value (attribute) in new data an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Model Evaluation And Testing

The goal of this step is to develop the simplest model able to formulate a target value faster. A data scientist can achieve this goal through dataset model tuning. That's the optimization of model parameters to achieve an algorithm's best performance and best outcomes.

VI. CONCLUSION AND FUTURE ANALYSIS

This Paper mainly aims to predict the direction of stock market trends in the future. In this Paper, study the performance on predictive analysis on NSE Index stock, of dataset from 2000 to January 2019 is considered. This is decisive to the traders as such analysis can influence the decision making with regard to buying or selling an instrument in a positive manner and discussed about two statistical machine learning models, namely Random Forest Model and Support Vector Machine which are used to provide a reliability prediction of stock market trends based on historical data. On the basis of the results obtained, and say that both the models exhibited notable performance in predicting the direction of the stock index.

REFERENCES

[1] Thissen, U., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). C hemometrics and intelligent laboratory systems, 6 9 (1), 35-49.

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

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

[4] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July), In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM

[5] Sahaj Singh Maini, SCOPE, VIT, Vellore, India; Govinda.K, SCOPE, VIT, Vellore, India; Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) IEEE Xplore Compliant - Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9