IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 4, Ser. III (Jul. – Aug. 2020), PP 08-17
www.iosrjournals.org

Stock Market Prediction Using Machine Learning

Shamik Palit. Chandrima Sinha Roy

Abstract
Being the exchange where the issuing and trading of equities or stocks of publicly held companies take place, the stock market is one of the most vital components of a market's economy. Stock market prediction is the process of attempting to predict the future values of a company based on its previous data to enhance the probability of a successful trade for an investor. In a financially volatile stock market, it is important to have a very precise prediction of future trends. Stock prediction involves the prediction in advance on whether the future market will close higher or lower compared to its opening levels. The stock market data is highly noisy, irregular and chaotic in nature. Hence proven to be a daunting task for market researchers and investors to make buy or sell decisions. A number of techniques as well as combinations of algorithms have been proposed over time to try and make a reliable and stable prediction. This paper aims at outlining the research work for Stock Market Prediction with special focus to daily, monthly and yearly stock predictions based on the technical approaches that have been proposed or implemented with varying levels of success rates. The algorithms being studied are implemented on a dataset and their accuracies are compared. Rules are proposed at the end of the implementation process to help a developer make predictions on their computers.

Date of Submission: 11-08-2020 Date of Acceptance: 27-08-2020

I. Introduction

1.1 GENERAL

The stock market is the collection of markets and exchanges where the shares of public-listed companies, bonds and other classes of securities are issued and traded. It provides companies with access to capital in exchange for a piece of ownership to the investor. Stock market prediction is a process of forecasting future price movements based on the past price movements within stock charts. It helps the investor make a more financially sound investment decision.A secure prediction of the values of the stocks is mandatory to score profits. The setting of the right price is the key to success in stock market investment. The focus of prediction varies in three ways: 1) the targeting price change can be daily, monthly and yearly, 2) the set of stocks used can be less than 10 particular stocks, to stocks in a particular industry, to generally all stocks; 3) the predictors used can range from global news and economy trend, to particular characteristics of the company, to purely time series data of stock price. After fixing the above criteria, the learned model can be used to make future predictions about stock values by evaluating the history of stock prices as well as other features. The procedure involves feature extraction from the dataset to be used, followed by the training and testing of the model using various ML algorithms before the result is evaluated.

II. Problem Definition And Objectives

2.1 DEFINITION

The ability to predict stock prices will provide significantly profitable to both the investors and the companies. Unfortunately, no techniques have proven to be completely accurate. This paper focuses on exploring some of the common algorithms used during the prediction process followed by an implementation of these algorithms on WEKA. Finally, this paper tries to set some guidelines which developers can adapt to make their own predictions.

2.1 OBJECTIVES

- To study and discuss selected algorithms used in daily, monthly and yearly stock prediction.
- To analyze the algorithms and make a comparative study based on their accuracies.
- To implement the algorithms on some datasets in WEKA and analyze the results achieved.
- To suggest stock prediction policies for developers.

III. Background I

3.1 STOCK MARKET

The ownership of a company is divided within the shares of that company. A single share of a company represents a fractional ownership of that company in proportion to all the shares possessed by that company. A stock is a share in a company. A stock is possessed by paying money. This money goes to the company who possessed that stock. This money is then used for building up the company. Companies, thus, benefit from the stock market.The stockholders can later sell their stocks at a higher price than the buying price. In this way, the stockholders also make profits. This buying and selling of the stocks takes place on a global network called the stock market and this whole process is called stock marketing.

3.2 WHY DO WE NEED STOCK MARKET PREDICTION?

The value of a stock is a function of the amount in the form of dividends that is expected to be paid by investors in the future. [1]These expectations of investors form the current price of stocks. Whenever the expectations for a company turn out to be true, the price of a stock remains the same. But in reality, the opinions of investors keep changing as new information comes into the picture. This information either worsens the opinion of the company or improves it.Thus, the value of a stock is also constantly fluctuating which is why a need for prediction of future stock prices arises.

If done right, stock marketing could yield significant profits to the investor. Forecasting future stock prices will also help companies prepare and tackle future situations and even come on top. They could allocate their funds in a way wherein they earn the most at any given time.

3.3 TYPES OF STOCK PREDICTION

Many attempts have been made to forecast the stock prices. The focus of each prediction project varies significantly in three respects.

1. Based on Time Period

Changes in the target price may be

- near-term (less than a minute),
- short-term/ daily (tomorrow to a few days later),
- long-term/monthly (months in the future), [2]
- yearly (years in the future)

2. Based on the Stocks

The stockscan be limited to

- less than 10 specific stocks,
- stocks in a particular industry to all general stocks. [2]

3. Based on the Type of Predictors used

The predictors used can

- vary from global news and economic trends,
- to company's specific characteristics,
- to purely time-series stock price data. [2]

IV. Background II

4.1 MACHINE LEARNING

ML is a sub-field of artificial intelligence that allow systems to be coded which will give them the ability to make predictions with minimal human intervention. These systems are capable of automatically learning from examples of data or direct experienceand identifying patterns to make accurate predictions. As the model is fed more data, it updates its output accordingly and makes predictions.

4.2 WHY MACHINE LEARNING?

ML has revolutionized various areas of life. The finance industry has used ML for stock market prediction since the 1970s. They are helpful tools for people navigating the investment and risk assessment decision-making process.[3] Human emotions is mostly the greatest hinderance to the prediction process. Machines with the correct algorithms can overcome this difficulty as well as execute trades faster and provide better predictions. They provide effective strategies for trading by automating the selection process. This allows investors to monitor multiple markets and respond accordingly. More markets imply increased opportunities. ML algorithms thus move a step further by not only finding associationsbased on previous data but also adapt to future trends with new data. In recent years, the use of ML algorithms has proven to be highly profitable.

4.3 TYPES OF LEARNING

ML can be broadly divided into the following categories-

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

4.3.1 Supervised Learning

Supervised learning algorithms focuses on making future predictions by using past examples which are fed to the model in the form of labeled examples. In this learning method, the correct output is known for each input also concluding that a relationship exists between the two. The algorithm first infers a mapping function from the known dataset which is used to make output predictions for future input data. After a certain amount of training, it can correctly make predictions about the output values for new inputs. Iterative predictions are made on the training data and the output is investigated and modified to equate to the correct output. Supervised ML algorithms are further categorized into classification and regression problems.Some common examplesof supervised algorithms include Linear Regression, SVM and Decision Trees.

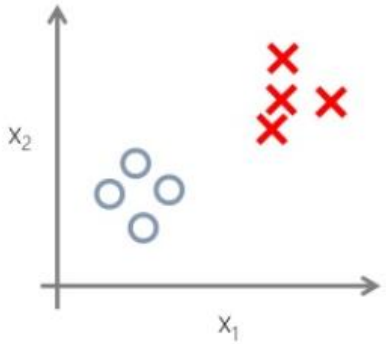
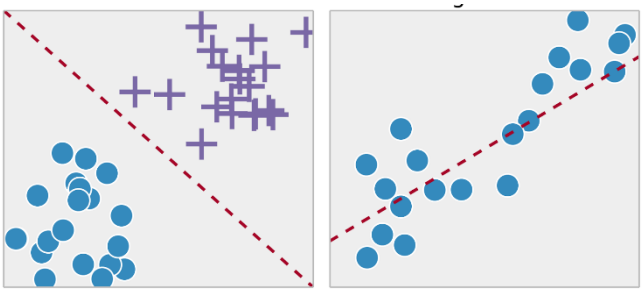


Fig. 3.1 Example of Supervised Learning

Classification is the process of identifying the particular category to which an input belongs. The training data is divided into different categories and the output prediction will basically classify the input to a pre-defined category.



a. b.

Fig 3.2 Examples of Classification and Regression- a. represents classification and b. represents regression

Regression is related to prediction. Output predictions are aimed to take place within a continuous output, that is, the input is mapped to a continuous function. Again, the prediction process takes place considering historical data.

4.3.2 Unsupervised Learning

Unsupervised learning algorithms can be used in situations where our information about the dataset is limited, meaning little or no idea is known about what the results are supposed to look like. The dataset is neither labeled nor classified and the system operates without any supervision and guidance. The algorithm devises its own interpretation about the structure of the data and tries to effectively depict the data in a compressed form. Another important task ofthese algorithmsis reducing dimensionality, that is, minimizing the number of features taken into consideration or the selection of only the main contributors for the prediction

process. Unsupervised algorithms are further classified into association and clustering problems. Unsupervised machines are capable of performing more intensive processing tasks as opposed to supervised learning machines. Some common examples of these algorithms include k-means clustering, Apriori algorithm.

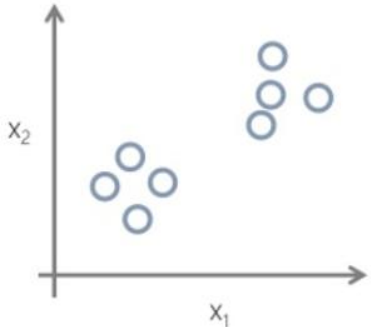


Fig. 3.3 Example of Unsupervised Learning

Clustering is the process of grouping the unlabeled input data into groups based on their similarities. This technique is used for grouping data and finding hidden patterns in the data.
Association rules are if-then rules that describe the probability of relationships between two or more data items in the dataset. It involves setting a minimum support threshold to filter out the frequent item sets followed by the application of a minimum confidence constraint to achieve useful rules.

4.3.3 Semi-Supervised Learning

Semi-supervised learning method majorly consists of unlabeled data and some labeled data. Therefore, it is a mix of the supervised and unsupervised learning methods.The increased number of unlabeled data during training has proven to increase the accuracy. The system begins to identify groups by processing the small amount of labeled data. Next, the system trains and learns from the large amount of unlabeled data by using the knowledge it acquired from the labeled data. This algorithm is even capable of identifying new groups which were not seen in the labeled data.

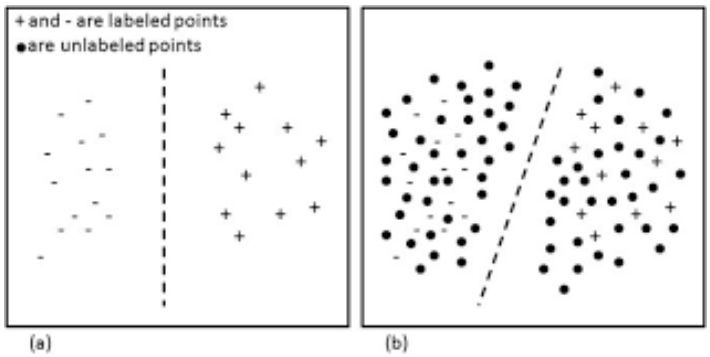


Fig. 3.4 Example of Semi-Supervised Learning

4.3.4 Reinforcement Learning

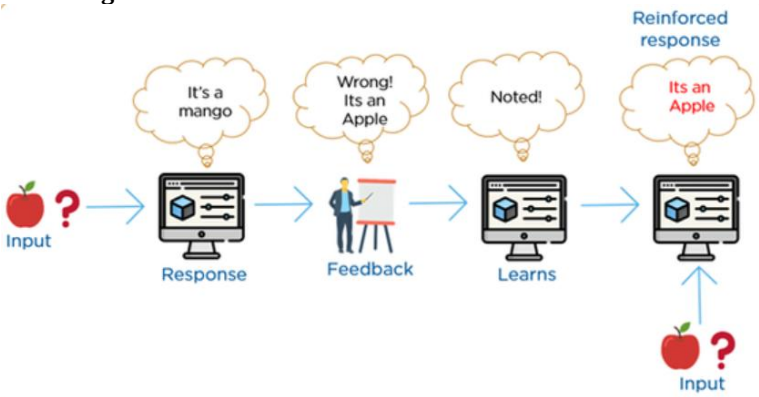


Fig. 3.5 Example of Reinforcement Learning

In reinforcement learning, the system interacts with its environment using a reward-system to make a decision. It neither requires labeled data to come to a conclusion nor does it require a training set, in the absence of which it just learn from experience. When the system first processes an input, it performs an action. If this action is correct, the system receives a reward from the environment. If the action is incorrect, it receives feedback from the external environment and the system learns the best behavior in that state. Based on this trial and error approach, it finally concludes to a decision. Reinforcement learning focuses on maximizing the rewards obtained.Thus, it tries to find the best action to be taken in a particular context that would maximize the performance.

4.4 APPLICATIONS OF MACHINE LEARNING

The scope of ML in today's world is so vast that it is being applied in almost every field. A few of the most common applications are-

- Health Care: On exploring ML techniques, it was found that doctors could be assisted in making a quicker and maybe even, a more accurate diagnosis. Further, these algorithms could tell when the patients were deteriorating so that an earlier treatment could be done.
- Intelligent Conversation: Businesses have harnessed this technology and applied it to their websites so that their customers receive a highly personalized experience with minimal human intervention. This can be seen in the implementation of chatbots which answer frequently asked questions and tailor to the needs of the user.
- Hiring Process: Recruiters often find it difficult to shortlist the best candidates for a job. Narrowing the best of the lot is often a tedious task given the vast number of candidates. ML software is used to quicken this process by identifying the required characteristics and shortlisting the best candidates.
- Fraud Detection: Banks and insurance companies have also started using ML models to detect fraudulent exchanges and protect their customers' accounts. The algorithms filter data and detect suspicious patterns.
- Recommendation Systems: These systems include targeted advertisements which is displayed to a user on their social media accounts like Instagram, Facebook and other sites such as Netflix as well as on e-commerce sites. The main intention of these systems is to recommend the right product to their customers whether it to buy products on Amazon or watch shows on Netflix. These systems have proven to be an effective marketing and advertising strategy.
- Transportation: Transportation firms use ML to make decisions by using the travel history and identifying patterns across various routes. These solutions are then presented to the customer to provide the route with the least potential problems when they travel.

4.5 WEKA TOOLKIT

WEKA (Waikato Environment for Knowledge Acquisition) is a software developed by the University of Waikato, which is used as a graphical user interface to explore ML. Datasets from existing sources can be uploaded to the software. Once uploaded,itsupplies the user witha number of tools to provide functionalities such as data pre-processing, data visualization and allows users to apply ML algorithms such as regression, classification, association rules mining and clustering. These algorithms can be directly applied to the required dataset or a separate Java code can be inserted.

V. Overview of Algorithms

5.1 LOGISTIC REGRESSION

LR is a regression-technique which is used for predictive analysis of a dataset having one or more independent variables. The resultant value is a dichotomous variable, that is, a binary value. The core function used is the sigmoid or logistic function. It is an S-shaped curve that varies between 0 and 1.

$$\text{Logistic Function, } \sigma(z) = \frac{1}{(1+e^{-z})}$$

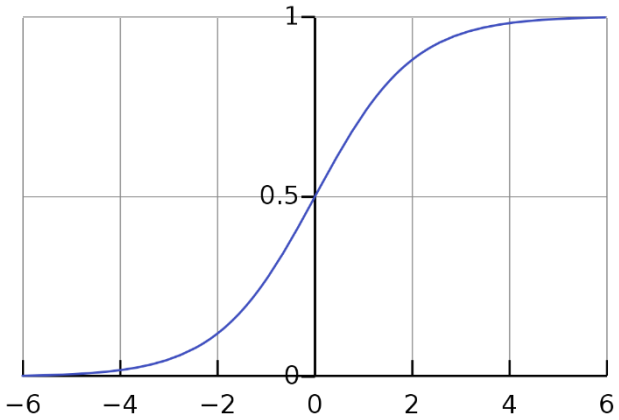


Fig. 4.1 Curve of Logistic Function

At the end of the process, a model is described that gives the relationship between the dependent dichotomous output variable and a set of independent predictors (numerical and categorical).

5.2 NAÏVE BAYES

Naïve Bayes is a probabilistic classification technique that classifies the input into predefined classes based on the concept of Bayes theorem. The Bayes theorem in probability describes the probability of occurrence of an event by considering the prior knowledge of possible causes for the event.

According to Bayes theorem, $P(A|B) = \frac{P(B|A)\times P(A)}{P(B)}$

Where P(A|B) = posterior probability of A given B

P(A) = prior probability of A
P(B|A) = likelihood of B if A is true
P(B) = prior probability of B

It assumes that the features are not correlated. The value of A can be replaced with a class and B with the set of features. The Naïve Bayes classifier finds the probability of every feature and then selects the outcome with the highest probability. It is suited for those datasets which have a high dimensionality of inputs.

5.3 SUPPORT VECTOR MACHINES

This algorithm can be applied to solve classification and regression problems. It distinctly classifies the data by producing an output separating hyperplane. The model is first trained with a training set where each record corresponds to a particular category. Using these inputs, a model is formed and future data is classified. The data points that lie closest to the hyperplane are called support vectors.

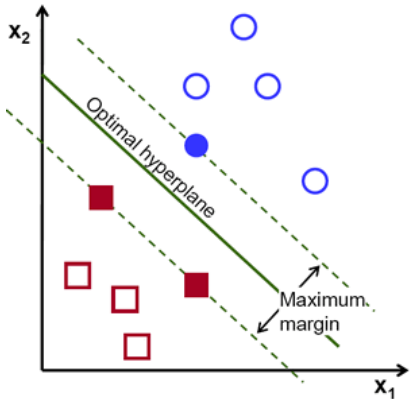


Fig. 4.2 SVM with optimal hyperplane

Many hyperplanes are formed by the model. Support vectors affect the orientation and position of the hyperplane. So, these points are used to maximize the margin between the various classifications. Larger is the distance between the hyperplanes lower is the generalization error of the classifier.

5.4 RANDOM FORESTS

This is an ensemble supervised learning method which is based on decision trees. It can be applied to classification and regression problems. The random forest model builds anumber of decision trees and combines them to achieve a higher accuracy. During the model building process, the best feature is selected among a group of random features and node splitting is performed based ion that feature. This increases the randomness and diversity in the model and helps in giving a better result. During the training phase, the model learns to map the historical data to the output. The model identifies relationships between the data during this period. After enough training is done with good quality data, the random forest will be able to make very good predictions.

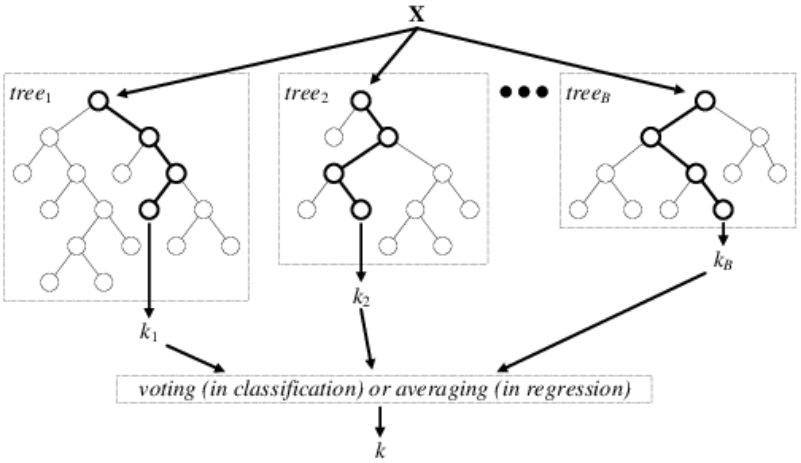


Fig 4.3 Random Forest Structure

VI. STUDY

6.1 STUDY OF ALGORITHMS

This section provides an overview of some selected algorithms after certain study.

Table 6.1 Study of Algorithms used for Analysis

ML Technique	Type of Problems	Advantages	Disadvantages
LR	Classification and regression problems	It does not require the input features to be scaled and is simple and very efficient.	It requires large datasets to achieve a stable and good accuracy. [4]
		The output is more informative and explains the contributing factors.	It cannot be applied to problems that cannot be linearly separated.
SVM	Classification and regression problems	It produces good results when there the dimensionality in the data is high.	Performs poorly when the dataset is noisy.
		The risk of overfitting is less as they provide good generalization of the data. [5]	It lacks transparency in its output which is caused due to the high number of dimensions. [5]

Table 6.1 (contd.)

Naïve Bayes	Classification problems	Can produce good estimation of parameters in small datasets. [6]	It does not produce accurate results when the dataset contains high correlations. [6]
		The training process is simple and fast.	It is not suitable to use for large datasets.
Random Forests	Classification and regression problems	They are simple and do not require much input preparation.	The size of the model is the main issue.
		They have high computational efficiency during the training period and yield high results. [7]	It performs poorly on imbalanced data, that is, rare outputs.

6.2 COMPARATIVE ANALYSIS OF ALGORITHMS

The algorithms have been implemented on various datasets and a comparative analysis of the algorithms is performed based on their output accuracies.

Table 6.2 Performance Comparison for Daily Stock Prediction

Problem	Algorithm	Dataset Used	Percentage of Accuracy (%)	Author(s)
Daily Stock Prediction	Naïve Bayes	S&P 500	60.38	[8]
		KSE-Banks	86.09	[9]
	LR	S&P 500	60.62	[8]
		Istanbul Stock Exchange National (ISEN)	56.47	[10]

Table 6.3 Performance Comparison for Monthly Stock Prediction

Problem	Algorithm	Dataset Used	Percentage of Accuracy (%)	Author(s)
Monthly Stock Prediction	Random Forests	AAPL	94.53	[11]
		Samsung Electronics Co. Ltd	93.96	
	LR	NSE-Bank	65.4	[12]
		NSE-Mining	61	

Table 6.4 Performance Comparison for Yearly Stock Prediction

Problem	Algorithm	Dataset Used	Percentage of Accuracy (%)	Author(s)
Yearly Stock Prediction	Random Forest	Amazon	72.36	[13]
		Bata	66.28%	
	SVM	Amazon	67.16	[13]
		Bata	62.35	

VII. Implementation

In this implementation, the SVM will be used on tstock.arff which is a dataset consisting of 950 instances and 10 attributes. This dataset contains the daily stock prices of 10 companies over a span of nearly 4 years from January 1988 to October 1991. The attributes are all of numeric datatype.

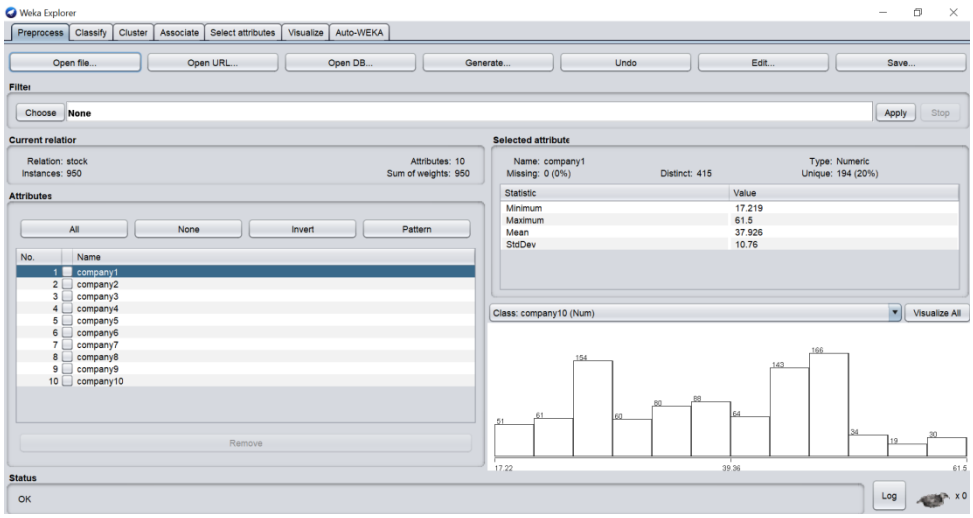


Fig. 7.1 Uploading an existing dataset on WEKA

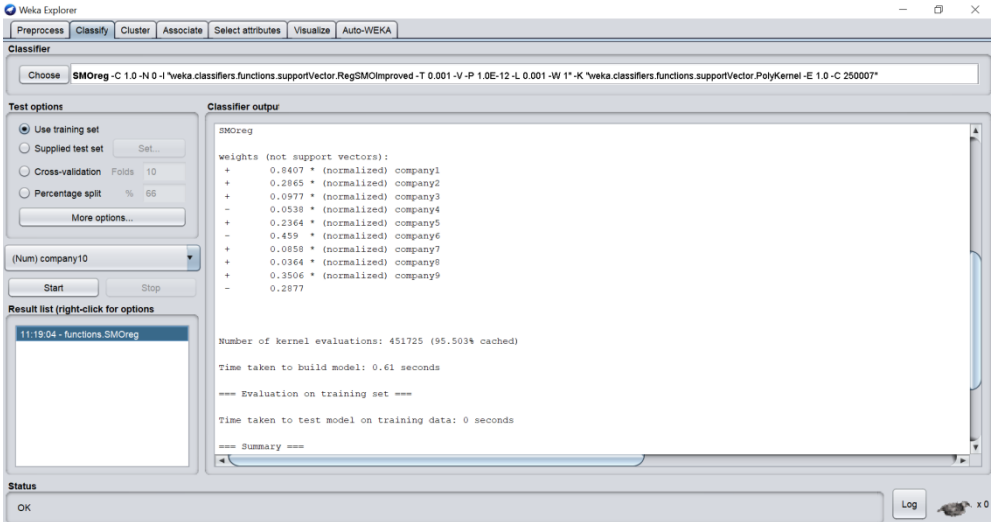


Fig. 7.2 Selecting the SVM classifier and running the algorithm

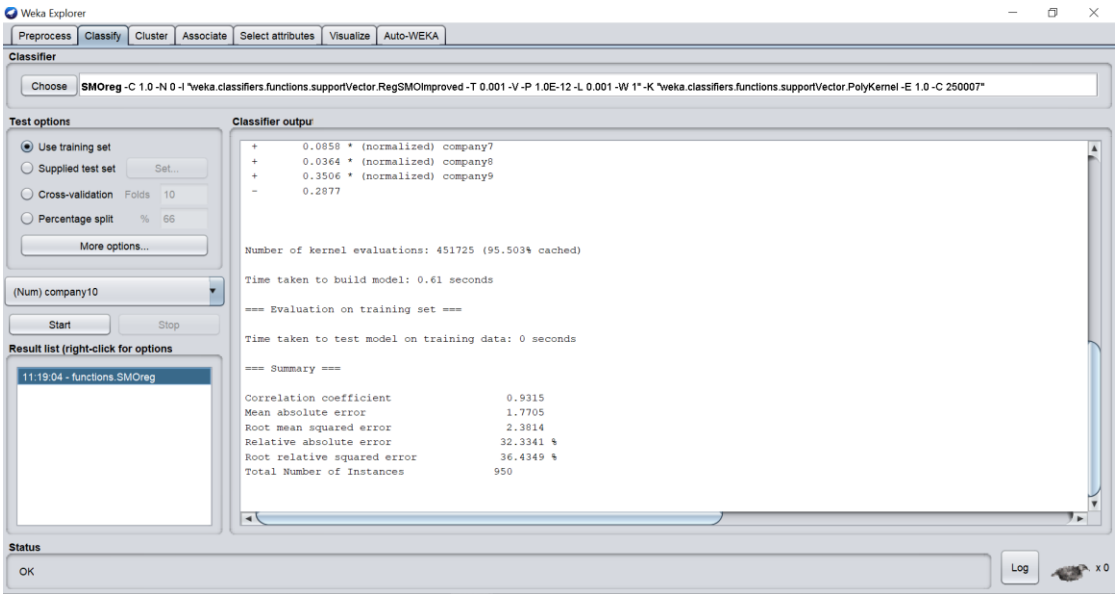


Fig. 7.3 Results obtained after running the algorithm

Steps involved to make predictions-

1. Download an existing dataset from any of the multiple online data sources like Kaggle, Yahoo Finance in .csv format.
2. Run WEKA and proceed to the Explorer tab.
3. Load the file using the **Open File…** option in Weka. Direct to the file directory containing the dataset. Select the file.
4. Select the **Save…** option and save the file as .arff format. Make sure the **File of Type…** is selected to **All Files**.
5. Reload the new .arff file using the **Open File…** option.
6. In the **Cluster** tab, choose the SVM classifier by selecting **Choose > functions >SMOreg.**
7. To run the algorithm, select **Start.**

VIII. Conclusion

In this paper, the concepts necessary for basic understanding of stock market prediction using machine learning is discussed. Various algorithms arestudied and a comparative analysis based on their accuracies is given. Finally, the SVM algorithm is implemented on a dataset and a set of rules are provided to guide naïve users make predictions.

References

[1]. M. Piper, "obliviousinvestor.com," 2009. .
[2]. A. Zheng and J. Jin, "Using AI to Make Predictions on Stock Market," pp. 1–6.
[3]. "sigmoidal.io," LLC, Sigmoidal. .
[4]. N. Singh Pahwa and N. Khalfay, "Stock Prediction using Machine Learning a Review Paper," Int. J. Comput. Appl., vol. 163, no. 5, pp. 975–8887, 2017.
[5]. L. Auria and R. a Moro, "Support Vector Machines (SVM) as a Technique for Solvency Analysis," Discuss. Pap. Dtsch. Inst. Wirtschaftsforsch., no. August, 2008.
[6]. N. Udomsak, "How do the naive Bayes classifier and the Support Vector Machine compare in their ability to forecast the Stock Exchange of Thailand ?," 2015.
[7]. A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischo, "On-line Random Forests," 2009 IEEE 12th Int. Conf. Comput. Vis. Work. ICCV Work., pp. 1393–1400, 2009.
[8]. C. Liu, J. Wang, D. Xiao, and Q. Liang, "Forecasting S & P 500 Stock Index Using Statistical Learning Models," pp. 1067–1075, 2016.
[9]. M. A. Ghazanfar, S. A. Alahmari, asmeen F. Aldhafiri, A. Mustaqeem, M. Maqsood, and M. A. Azam, "Using Machine Learning Classifiers to Predict Stock Exchange Index," Int. J. Mach. Learn. Comput., vol. 7, no. 2, pp. 24–29, 2017.
[10]. P. Tüfekci, "Classification-based prediction models for stock price index movement," Intell. Data Anal., vol. 20, no. 2, pp. 357–376, 2016.
[11]. L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," vol. 00, no. 00, pp. 1–20, 2016.
[12]. A. Nayak, M. M. M. Pai, and R. M. Pai, "Prediction Models for Indian Stock Market," Procedia Comput. Sci., vol. 89, pp. 441–449, 2016.
[13]. I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A comparative study of supervised machine learning algorithms for stock market trend prediction," Second Int. Conf. Inven. Commun. Comput. Technol., no. Icicct, pp. 1003–1007, 2018.

Shamik Palit, et. al. "Stock Market Prediction Using Machine Learning." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(4), 2020, pp. 08-17.