Journal of Computer Science

Original Research Paper

Ultimate Prediction of Stock Market Price Movement

[1]Rebwar M. Nabi, [1]Soran Ab. M. Saeed, [2]Habibolah Bin Harron and [3]Hamido Fujita

[1]Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq
[2]University of Technology Malaysia, Johor, Malaysia
[3]Iwate Prefectural University, Takizawa, Japan

Article history
Received: 20-10-2019
Revised: 04-12-2019
Accepted: 21-12-2019

Corresponding Author:
Rebwar M. Nabi
Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq
Email: rebwar.nabi@spu.edu.iq

Abstract: Investment in the stock market is currently very popular due to its economic gain. Numerous researchers' and academicians' work is focused on financial time series prediction due to its data availability and profitability. Therefore, this study presents the design and implementation of a novel binary classification framework to predict stock market trends. The framework is composed of data preprocessing, feature engineering, feature selection and classification algorithms. The model is built on multiple sector stock market companies' data collected from NASDAQ over a period of ten years. Various feature selection algorithms are applied in combination with several machine learning algorithms. Furthermore, as the new contribution, we have constructed two new features which have been found to be promising in terms of improving overall performance. Ultimately, a collaboration of feature selection and classification techniques is employed. The application of Principal Component Analysis (PCA) with Multilayer Perceptron and Support Vector Machine (SVM) to added featured datasets shows 100% accuracy on the majority of datasets. In summary, an intensive comparison is presented among the various feature selection and classification algorithms.

Keywords: Stock Market Forecasting, Feature Engineering, Feature Selection, Machine Learning Mechanism, Predictive Analysis, Predictable Movement, Java and WEKA

Introduction

Every country's economic growth depends upon stock market performance. The stock market is highly volatile and unpredictable by nature. Therefore, investors continually take risks in hopes of making a profit. People want to invest in the stock market and expect profit from their investments. Many factors influence stock prices. Examples include supply and demand, market trends, the global economy, corporate results, historical price, public sentiments, sensitive financial information and popularity (such as good or bad news related to a company), all of which may result in an increase or decrease in buyer's strength etc. Although one may analyze many factors, better stock market performance and future price prediction remain challenging. The forecasting of rapidly changing stock prices is a very challenging task (Fama *et al*., 1969). Real-life news impacts the stock market. Over the last few years, there have been many ups and downs in the stock market, as there are n factors that can affect a share market. Thus, due to its dynamic nature, it is highly challenging to predict a stock price. To address this issue, there should be some system that can both detect the pattern in stock prices when influenced by the political, economic and natural environment and take into account people's sentiment about a particular company.

Although one may analyze several factors, the achievement of better performance in estimating future movement remains thought provoking. Markets are efficient or, at a minimum, movement for a particular duration is. Public information is reflected in stock prices and the pricing mechanism rapidly and efficiently processes new information sets. Attempting to gain an edge is nearly impossible, especially when one tries to process widely accessible public information. Investors are, therefore, better off holding a well-diversified portfolio of stocks.

There are two primary schools of thought in analyzing the financial markets. The first approach is known as fundamental analysis. The methodology used in fundamental analysis evaluates a stock by measuring

its intrinsic value through qualitative and quantitative analysis. This approach examines a company's financial reports, management, industry, micro and macro-economic factors (Graham *et al*., 2015; Idrees *et al*., 2019; Chen *et al*., 2019). The second approach is technical analysis; The methodology used in technical analysis employs the learning of historical company data. The stock analysis uses a variety of charts to anticipate what is likely to happen. The stock charts include most of the type charts, Open-High-Low-Close (OHLC) chart and mountain charts. The charts are viewable in different time frames with price and volume. Many types of indicators are used in the charts, including resistance, support, breakout, trending and momentum (Kirkpatrick and Dahlquist, 2010).

Analyzing financial data insecurities has been an important and challenging issue in the investment community. Dozens of elements influence the stock market (King, 1966; Chen *et al*., 1986). Experiments for generating new features that do not exist in the dataset and may be necessary to a better predictive accuracy rate have been performed (Long *et al*., 2019). In high dimensional data, not all features are relevant and influence the outputs. Improved feature representation, based on stock market prediction, is evaluated to investigate the statistical metrics used in feature selection that extract the most relevant features to diminish the attributes list of datasets (Zhou *et al*., 2017). There are many machine learning techniques. The determination of which techniques are superlative in the prediction of stock movement is of major concern in financial data.

This research project, based on NASDAQ financial stock market data, aims to solve the following problem:

- Interval of stock data prediction: Decide stock movement prediction on a yearly, quarterly, monthly, weekly, or daily basis.
- Feature Engineering: Which features to consider to increase accuracy.
- Feature Selection: Finding the best feature selection algorithm for stock movement.
- Machine Learning: Explore the best-supervised machine learning technique based on training and testing classification.

Related Work

Stock Market Forecast

Market prediction using different analysis techniques is regularly practiced in modern marketing systems by collecting and analyzing market information (Subha and Nambi, 2012). Traders in any part of the world are interested in a market that is profitable and uses multiple technical indicators, macroeconomic factors and stock market indexes to study the market. This diverse market drivers' information reflects existing market price characteristics and facilitates prediction of future market price characteristics (Caley, 2013; Yang *et al*., 2019). As a result, we can prevent anticipated negative changes in the market due to new information about the market. However, market analysis is not a common practice and is often carried out using traditional tools and manual practices making the processing time consuming and prone to errors (Caley, 2013). Although other countries have carried out various studies on market prediction, a direct implementation of their findings is not practical. Each study follows different approaches based on the country's economic and market situation. Moreover, the market features that have an impact in one country may not have a similar impact in another country. As a result, we need to take a closer look at the target market to form or improve market strategies.

Machine Learning

Machine Learning is manipulated to analyze datasets to generalize and observe the patterns of that data or information. To predict future value or behavior from those observations or patterns it will then iteratively learn from data, unlike typical computer programs. The purpose of machine learning is to program computers to use sample data as an experience or model and use the patterns of this data to predict the future based on that data (Nayak *et al*., 2016; Feng *et al*., 2019). Machine Learning not only deals with database problems but is also an application of artificial intelligence (AI). It helps solve several problems in face recognition, biometrics authentication, medical diagnoses, agriculture, economics, computer networks and robotics (Alpaydin, 2014; Mohri *et al*., 2018). Machine Learning involves training a computer model with data or historical information (Lison, 2015) to potentially predict behavior of the system in the future. Machine Learning can be divided into three main subsets: (1) Supervised Learning; (2) Unsupervised Learning; (3) Reinforcement Learning.

Supervised Learning

Supervised learning involves the use of historical predictors and outcomes with the intent that the model will provide useful predictions of new values given new combinations of predictors (Lison, 2015). This learning is "supervised" in that the outcomes of particular sets of predictors are already known and can be used to monitor the accuracy of the predictions that the model produces.

Supervised learning algorithms come in many forms with specific strengths, weaknesses and purposes (Hastie *et al*., 2001). Specific models that are suitable for the research in this thesis include Linear Regression and Random Forests. Each of these models can provide insight into outcomes in a manner that permits the prediction of new values given novel feature inputs

DOI: 10.3844/jcssp.2019.1795.1808

(Kuhn and Johnson, 2013). Linear regression models can be used for performing standard regression for a series of outcomes given a particular set of inputs, or features. This is to say that the real mean of the outcome varies linearly with the features (Rawlings *et al*., 2001). Linear regression models fail when the relationship is non-linear as well as when too many features are used to fit the model. In the former case, either the data must be transformed into the linear domain, or a different model must be used. In the latter case, adding more features to a model may decrease the error on the training dataset. However, the testing dataset may then exhibit increased error. This is known as an over fitted model (Hastie *et al*., 2001).

Unsupervised Learning

Unsupervised learning is used when the order of outcomes of a dataset is unknown and the user is looking for a pattern to analyze (Lison, 2015). This is especially useful in circumstances in which the distribution of the data is unknown and the researcher is looking for additional information about the behavior of the dataspace. No expectations of results are fed into the system by the analyst. Instead, unsupervised learning models are used to find the patterns and behavior and help derive expectations of the data for further analysis and understanding (Hinton and Sejnowski, 1999).

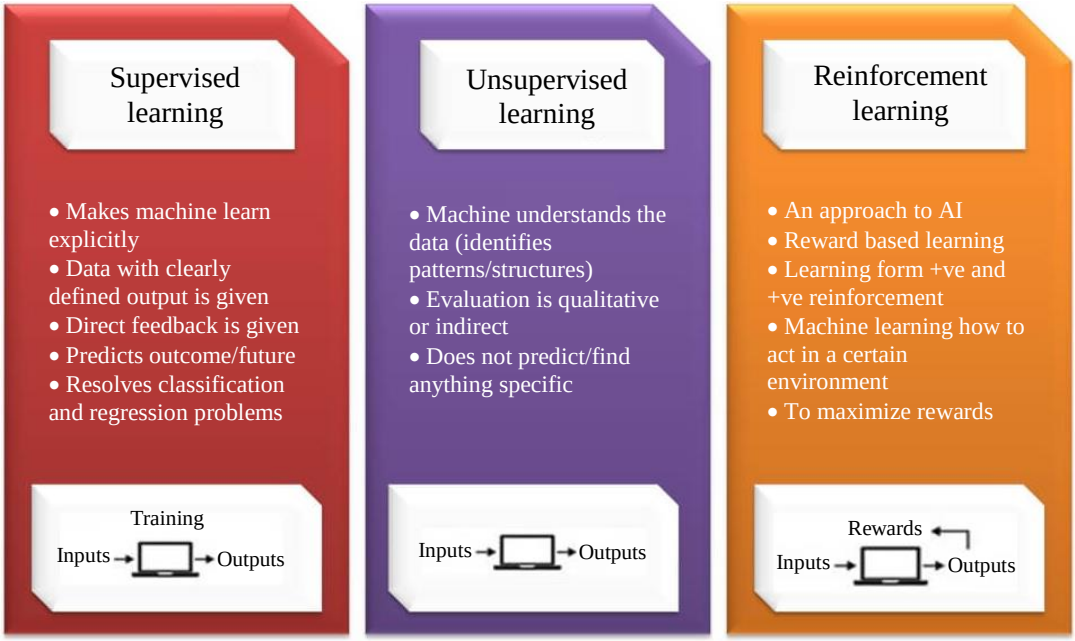

Fig. 1: Types of machine learning

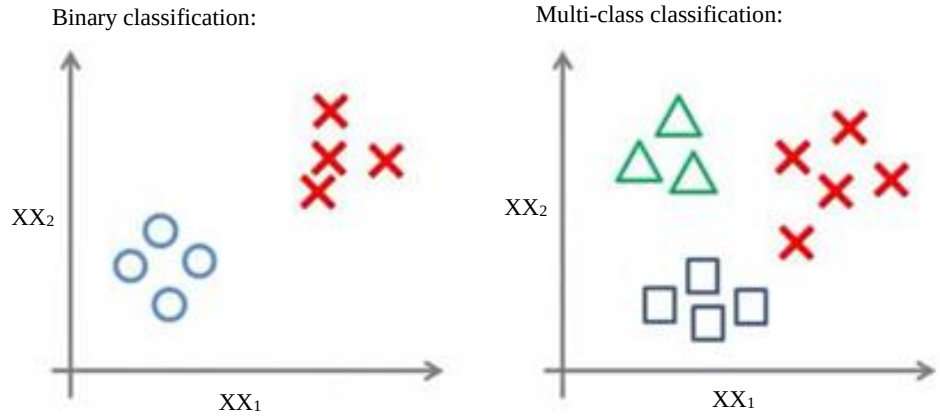

Fig. 2: Binary-label and multi-class classification

Binary Vs. Multiclass Classification

Binary classification is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) based on a classification rule. Contexts require a decision as to whether or not an item has some qualitative property. Figure 2 explain the difference between binary and multi-class classification.

Machine learning, on the other hand, employs multiclass or multinomial classification. This method classifies instances into one of three or more classes. (Classifying instances into one of two classes is called binary classification.) (Zhou *et al*., 2016).

Feature Engineering

Feature engineering is a vast topic and more methods are being invented every day, particularly in the area of automatic feature learning. The basic concepts of Machine Learning are data and model. Data looks at stock market data including daily stock prices, announcements of earnings by individual companies and even opinion articles from pundits. Each piece of data provides a small window into a limited aspect of reality. The aggregate of all these observations gives us a picture of the whole.

Nevertheless, the picture is messy because it is composed of a thousand little pieces. Measurement noise and missing pieces add to the confusion. Trying to understand the world through data is like trying to piece together reality using a noisy, incomplete jigsaw puzzle with a bunch of extra pieces (Zemke, 2003). This is where mathematical modeling-in particular, statistical modeling-comes in. The language of statistics contains concepts for common characteristics of data, such as wrong, redundant, or missing. Wrong data is the result of a mistake in measurement. Redundant data contains multiple aspects that convey precisely the same information.

In constructing a new feature, it is desirable for the result to be interpretable. Moreover, interpretable features and models are more natural and lead to the most accurate model. In addition, adding complexity improves the accuracy of classification. The goal of feature engineering, however, is not so much to limit the number of feature dimensions as much as possible but to arrive at the right features for the task (Long *et al*., 2019).

Stock market data as numeric data is already in a format that's easily ingestible by mathematical models. A mathematical model of data describing the relationships that predict stock prices might be a formula that maps a company's earning history, past stock prices and industry to the predicted stock price. Useful features should not only represent salient aspects of the data but also conform to the assumptions of the model. Hence, ransformations are often necessary. Numeric feature engineering techniques are fundamental. Distribution summarizes the probability of taking on a particular value. The distribution of input features matters to some models more than others. In stock market data, additional features can be added to improve the classification result. Contrarily, in the literature it is challenging to find proper studies that try to construct new features. Therefore, in this study, we intend to investigate and propose new features using mathematical operations.

Methodology

An executable Jar Project was developed with a Graphical User Interface (GUI) using Java Programming Language. Waikato Environment for Knowledge Analysis (WEKA) was used as a machine learning platform for using the algorithms. However, several other JREs were added separately into the project since feature selection and even a few classifiers were not included in the WEKA platform. The overall look of the main windows can be seen in the Fig. 3.

As can be seen in the above figure, the system first allows us to choose the dataset and subsequently the feature selection algorithm that is to be tested. Next, the system prompts to select the Machine Learning classifier. Finally, the attributes to be included in the experiment are added. The original attributes are either accepted as downloaded or features that have been added using feature Engineering techniques are chosen. Lastly, the classify button will be pressed to see the result. An example of the result is shown in the Fig. 4.

The project mainly consists of six steps and it is explained in the Fig. 5.

Dataset Collection

Datasets were downloaded from NASDAQ for the following companies: CMCSA, CSCO, FOX, FOXA, LRCH, MCHP, MSFT, NTAP, QCOM and SWSK. Each dataset contains 2520 daily records. The total of 119 rows from ten years as months. 90 rows used as a training and the rest 29 used for testing for all classifiers. The attributes are composed of the following:

1. Date
2. Close Price
3. Volume (total transactions)
4. Open Price
5. High price
6. Low Price

The dataset files in CSV file format. Sample dataset of CSCO is shown in the following Table 1.

DOI: 10.3844/jcssp.2019.1795.1808

The companies' details are shown in Table 2 which consist of ten companies stock data.

Going forward this paper will employ company symbols, rather than names.

Fig. 3: Initial GUI of project

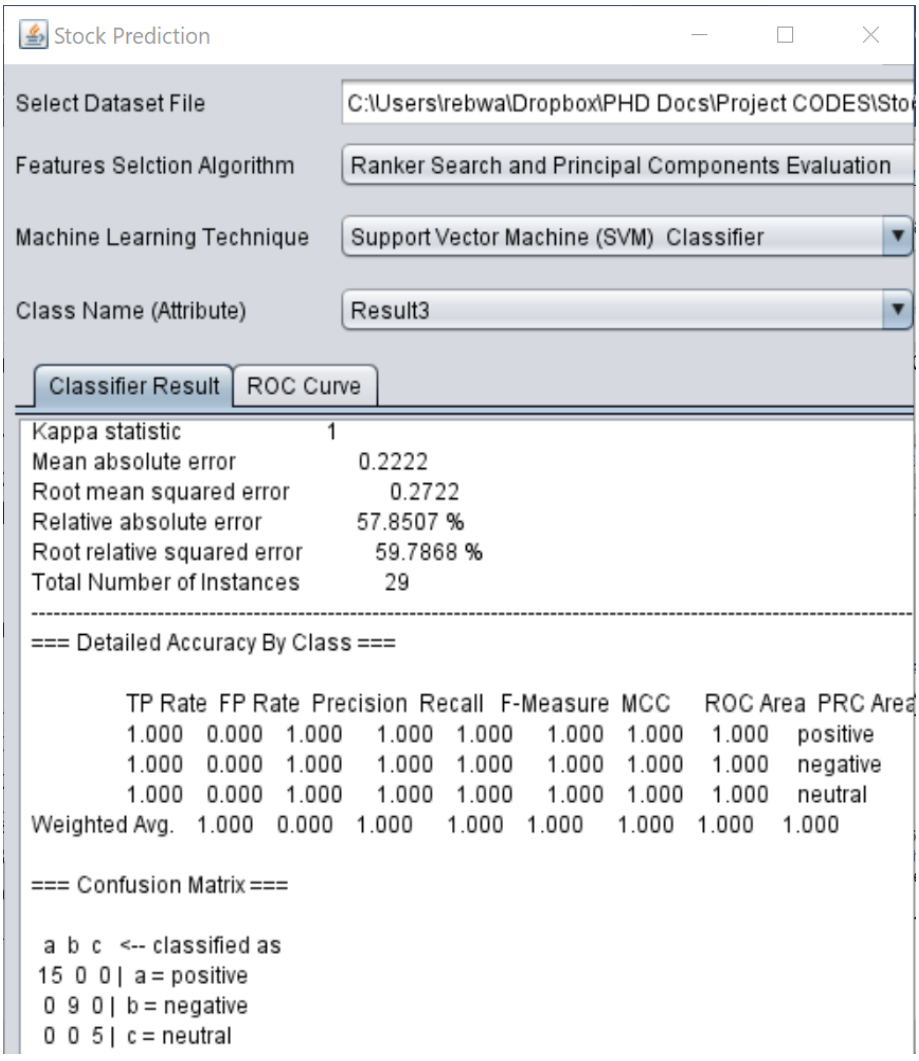

Fig. 4: Developed system demo

DOI: 10.3844/jcssp.2019.1795.1808

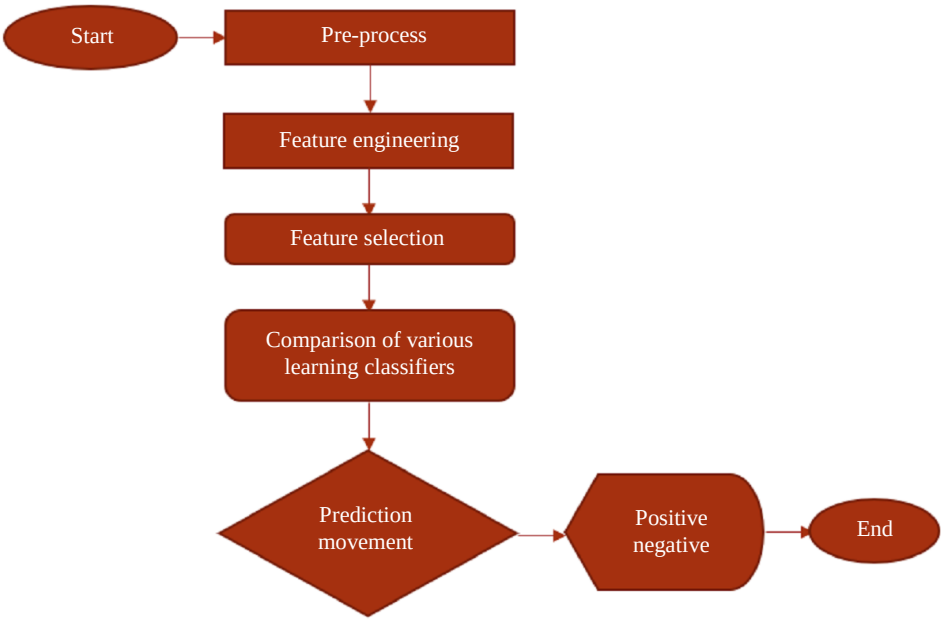

Fig. 5: Proposed project flow-chart

Table 1: Downloaded data from Nasdaq - Sample

Date	Close	Volume	Open	High	Low
2/22/2019	38.61	20376890	38.48	38.730	38.40
2/21/2019	38.47	22071940	37.61	38.470	37.44
2/20/2019	37.79	14939590	37.70	37.925	37.53
2/19/2019	37.55	12853460	37.75	37.940	37.48
2/15/2019	37.77	15872050	37.60	37.780	37.15
2/24/2009	6.94	57832580	6.325	6.965	6.31
2/23/2009	6.345	51067300	6.435	6.70	6.32

Table 2: Company's bio details

No	Symbol	Company Name
1	CMCSA	Comcast Corporation
2	CSCO	Cisco Systems, Inc.
3	FOX	Fox Corporation
4	FOXA	Fox Corporation
5	LRCH	Lam Research Corporation
6	MCHP	Microchip Technology Incorporated
7	MSFT	Microsoft Corporation
8	NTAP	NetApp, Inc.
9	QCOM	QUALCOMM Incorporated
10	SWSK	Skyworks Solutions, Inc.

Pre-Processing

Pre-processing involved the collection of historical data for the last ten years from the NASDAQ, pre-processing that data with methods of generating the monthly based interval and application of the class label as a movement. For binary classification, it is (a). Positive and (b). Negative. An example of the pre-processed data is shown in Table 3.

Feature Engineering

As previously mentioned, one of the goals of this study was to add new features to improve the accuracy of classification. For this purpose, two features were

studies and added. The added feature is calculated by finding the difference between the high and low price of the month. Whereas the second feature is the mean value of close open difference as daily bases:

- **HLdiff:** Difference in whole month high and low price. Mostly, it shows the whole month maximum movement in the price
- **Mean:** Mean values of all Difference between close and open price. That would show the price average moment

Feature Selection

The best features to forecast stock were selected, including selected monthly data features, based on that data. Different techniques which can forecast learned data, such as filter methods, wrapper methods and embedded systems, were applied, including:

- None (No Feature Selection Algorithm)
- Sequential feature selection (Best First) Search and CFS Subset Evaluation (SEQ)
- Genetic Search and CFS Subset Evaluation (GEN)
- Ranker Search and Chi-Squared Evaluation (CHI)
- Ranker Search and Recursive Feature Elimination Evaluation
- Ranker Search and Correlation Coefficient Evaluation
- Ranker Search and Info Gain Evaluation (IG)
- Ranker Search and ReliefF and it is Variant Evaluation (REF)
- Ranker Search and Principle Components Analysis Evaluation (PCA)
- The Best Feature Selection (Finding the best from above all 9)

Comparison of Various Learning Classifiers

When developing a classifier using various functions from different classifiers, it is essential to compare the performances of the classifiers. Simulation results can provide us with direct comparison results for the classifiers with a statistical analysis of the objective functions. In this study, fifteen known classifiers are tested and compared; the algorithms are:

- Naive Bayesian Classifier (NB)
- Bagging Classifier (BAG)
- Stacking Classifier (STA)
- Voting Ensemble Classifier (VOT)
- Support Vector Machine (SVM)
- AdaBoost Classifier (Ada)
- Gradient Boosting (GBM) Classifier
- Multi Boosting Classifier (MB)
- Decision Tree Classifier (DT)
- Random Forest Classifier (RF)
- Logistic Model Tree (LMT) Classifier
- Logistic Classifier (LOG)
- Simple Logistic Classifier (SL)
- Multilayer Perceptron Classifier (MP)
- Multiclass Classifier (MC)
- The Best Classifier (Finding the best from all 15 above)

Prediction

Calculate performance, including accuracy, confusion matrix, ROC Curve, accurate time-wise prediction, binary classification optimization and compare with all feature selection and machine learning mechanisms. After comparison, if the actual value is predicted, the algorithm goes on to different and targeted value prediction then evaluates the result of the prediction. This will reveal the ultimate collaboration of feature selection approaches along with machine learning techniques.

ROC Curve of Binary-Label Classification

ROC Curve of False Positive Rate (X-Axis) vs True Positive Rate (Y-Axis) for Binary-Label Classification of Dataset (CMCSA added featured Dataset) and Machine Learning Technique as Bayesian Classifier. The ROC curve is embedded in a box having unit-length sides. It begins at the origin defined by a sensitivity of 0.0 and a specificity of 1.0 and ends at a sensitivity of 1.0 and a specificity of 0.0. The ROC Curve shown in Fig. 1 is shown for POSITIVE Label. The ROC Curve show in Fig. 6 shows NEGATIVE Label.

Table 3: Pre-processing the data

M	Close	Vol	Open	High	Low	Res.
2	38.6	2.68E+	36.7	38.7	36.2	pos
1	36.5	4.96E+	33.4	37.4	33.4	pos
12	34.0	4.32E+	39.0	39.2	32.6	neg
11	39.0	4.08E+	38.1	39.6	36.6	pos
10	38.1	6.15E+	35.3	38.6	33.5	pos
2	38.6	2.68E+	36.7	38.7	36.2	pos

DOI: 10.3844/jcssp.2019.1795.1808

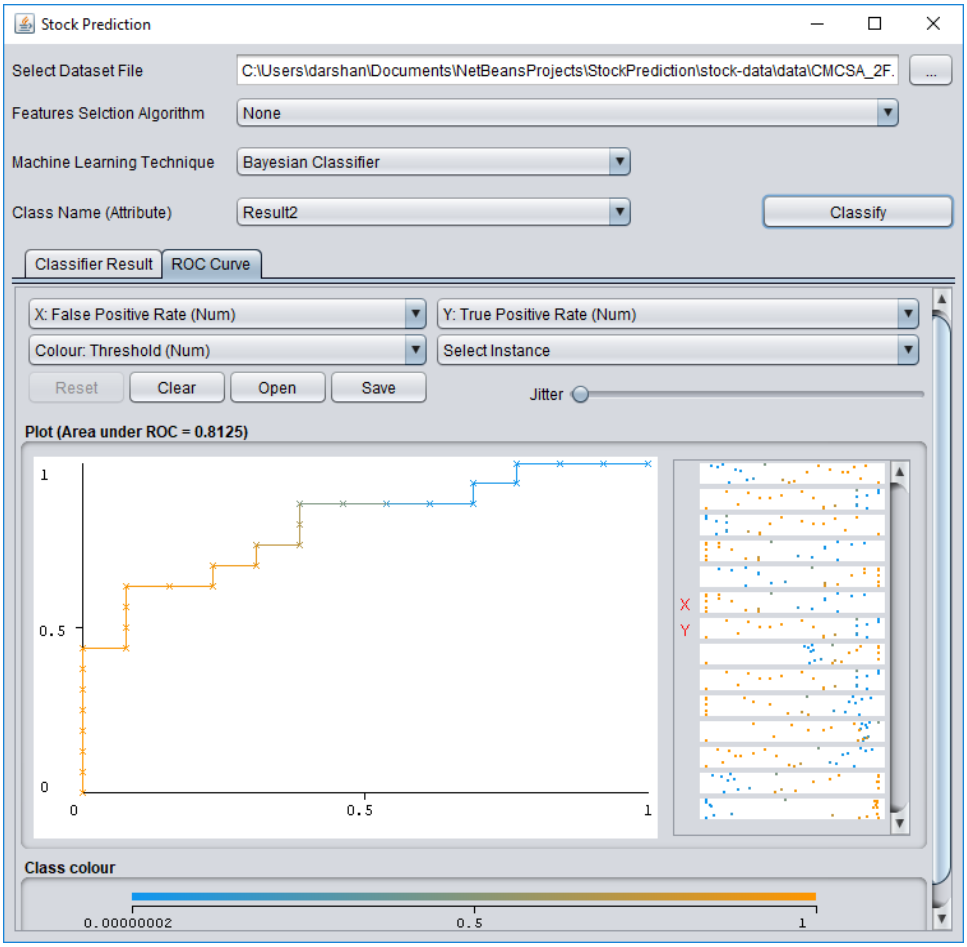

Fig. 6: ROC curve binary classification sample

Result and Discussion

Preliminary Prediction (Feature Engineering and Feature Selection)

First, the experiments were conducted on the entire 40 datasets of ten companies. Furthermore, all the classifier techniques and feature selection algorithms on all 40 datasets as supervised (training and testing based) classifiers and separate graphs have been generated. To demonstrate the importance of feature engineering, two different types of lines are used. Solid lines are used to represent the experiments with added features (Feature Engineering).

Dotted lines are used to demonstrate the experiments on the datasets without added features, employing only the original features as downloaded from Nasdaq. An individual graph is generated for each company's data. The below figure demonstrates an example of the overall prediction result.

Figure 7 demonstrates the overall look of the graph produced. Horizontally, the algorithm is listed and vertically the feature selection algorithm is listed. The performance graph contains:

- X-Axis [Classifier techniques Symbol-all 15 classifiers]
- Y-Axis [Prediction Accuracy in percentage]
- 18 lines [solid line-with added feature (9F) and dotted line-without the added feature (7)]
- 9 line colors [Different color for each feature selection algorithm]
- PCA (6F) [6 Features selected and F for with feature] As previously mentioned, the data of ten companies was used to evaluate the accuracy of the algorithms. Figure 9 demonstrates the overall performance and accuracy of the CMCSA company.

Figure 8 illustrates the prediction results over various feature selection and classification algorithms, demonstrating that the results are different. The worst results occur using random forest with ranker search and correlation coefficient evaluation. However, PCA performs

DOI: 10.3844/jcssp.2019.1795.1808

as the best feature selection algorithm when it is used with Support Vector Machine, achieving 100% accuracy.

QCOM company data is also chosen. The comparison result is shown in Fig. 9.

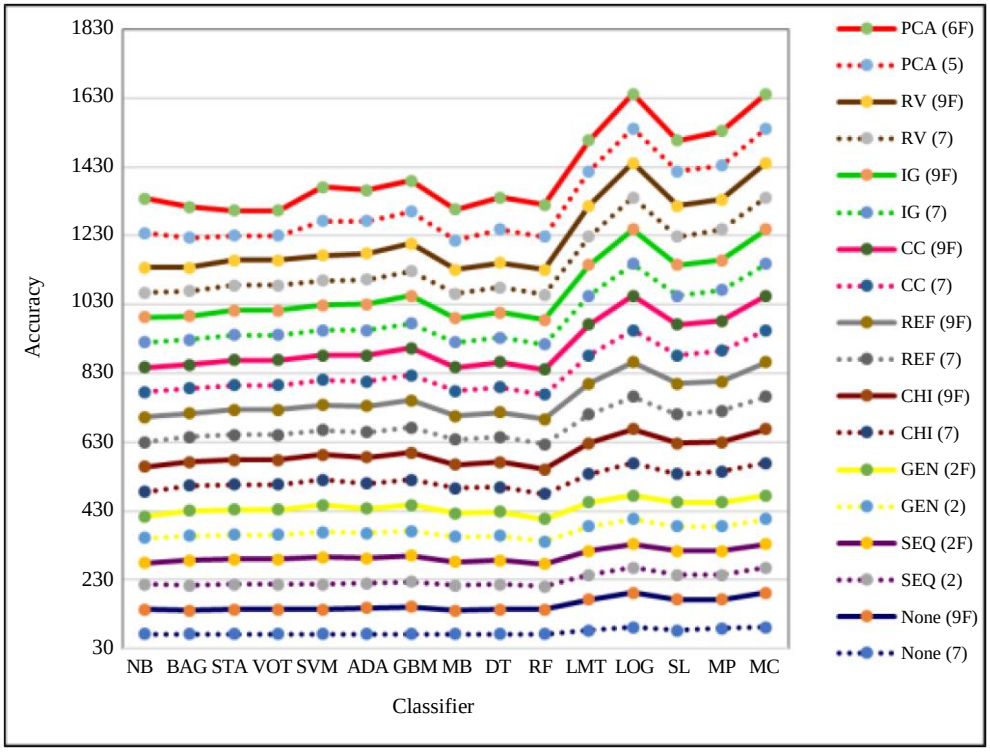

Fig. 7: Graph: Performance comparison graph

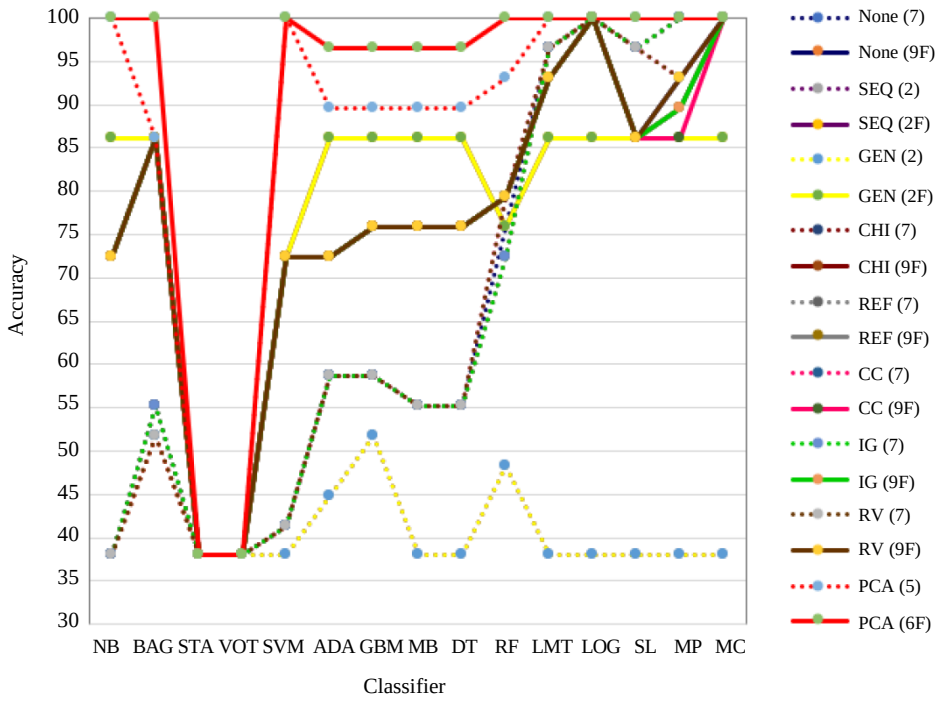

Fig. 8: CMCSA prediction result

DOI: 10.3844/jcssp.2019.1795.1808

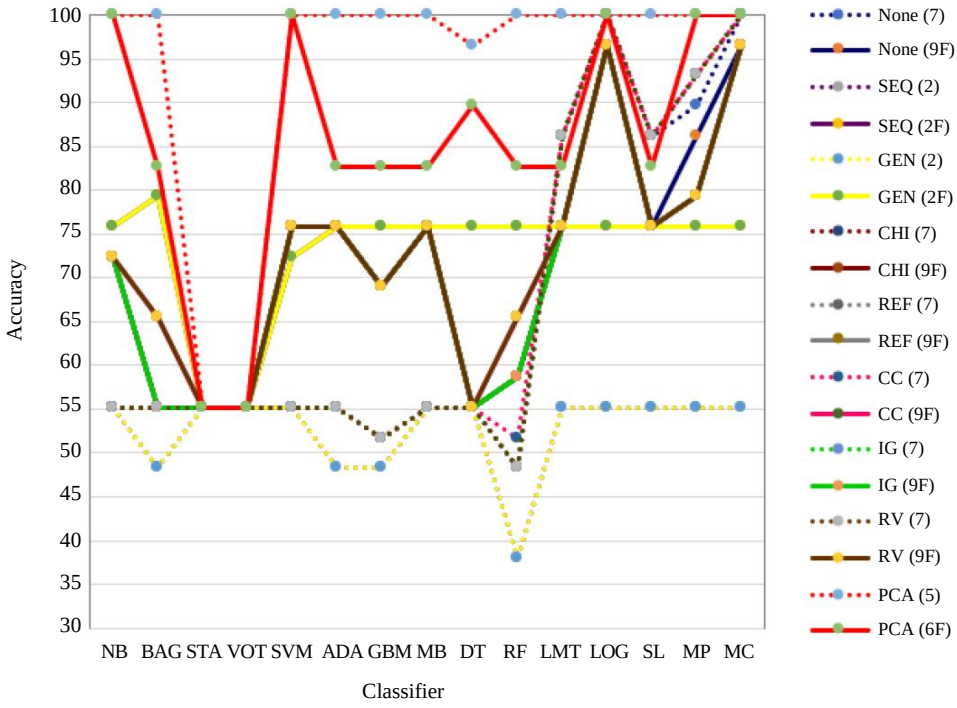

Fig. 9: QCOM Prediction result

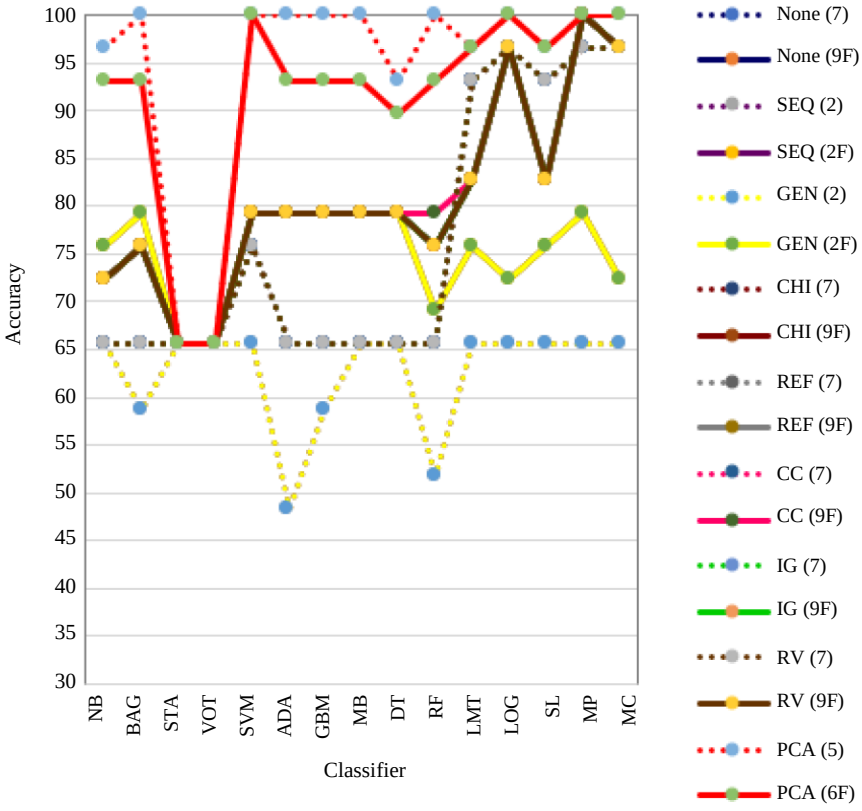

Fig. 10: CSCO company prediction result

For the QCOM company, in the majority cases, the predictions were promising. In contrast to earlier findings with CMCSA, the Correlation Coefficient using added features was found to be outperforming, where, in some cases, an accuracy of 100% was achieved. Likewise, for the CMCSA company, the support vector machine without using any feature selection algorithm, was the top performer. It is also worth mentioning that adding features significantly improved accuracy. An example of this is the use of Genetic Search and CFS Subset Evaluation (GEN) with Bagging Classifier. The Solid yellow line represents the prediction with the added feature and an accuracy of 80% is attained. However, the dotted yellow shows the prediction when the original features are included in which approximately 37% is achieved. Figure 10 illustrate the overall prediction result for CSCO company. As it can be noticed MP and SVM with PCA outperform all other algorithm and combinations.

Preliminary Investigation

The produced graphs of all ten companies' binary-label classification primarily demonstrated the difference between the solid and dotted line graph and expanded the accuracy through the addition of new features to the dataset as part of feature engineering. Based on the experiments in feature selection, Principal Component Analysis (PCA), which is shown with red color in all Graphs, was found to be a very efficient algorithm compared to others.

The Best Classifier

Another aim of this study was to find the best classifier technique. As previously explained, the testing was done on all ten companies' datasets with added new features and PCA as a feature selection algorithm. The binary-label method was tested with all 15 classifier techniques used in this study. Table 4 to 6 show binary label classification results. The tables are separated into three Table Sections (TS) to match the two-column format. The full table is attached as an appendix for a better and easier understanding.

The results show the accuracy of each classifier with each company's dataset. The average accuracy of each classification technique for all companies was taken into account. Based on that, in binary-label classification, two classifiers give the best results on all companies' datasets. The First ranked is Support Vector Machine (SVM) and the second-best algorithm is Multilayer Perceptron (MP). These show the best performance among all 15 classifiers.

Study Performance Achieved Benchmark

It can be shown that this study has achieved a new benchmark by constructing new features that add to the original datasets. Based on the results, a remarkable improvement was noticed. More importantly, we have studied the best combination of feature selection and classifier techniques.

Feature Engineering Contribution

The project compared the results of both the original dataset and feature added dataset as shown in section 4. The graphs present lines that represent feature selection algorithm and classifier techniques, accuracy of training and are testing based. There are two types of lines: Dotted lines present the original dataset and solid lines represent the added new constructed features dataset. All the graphs show that added feature performance outperforms the original dataset in most cases. Therefore, the project aimed at expanding feature engineering is completed impeccably.

Table 4: Overall Performance result TS1

Alg\C	CMCSA%	CSCO%	FOXA%	FOX%
NB	100.00	93.10	100.00	100.00
BAG	82.76	93.10	96.55	100.00
STA	55.17	65.52	55.17	51.72
VOT	55.17	65.52	55.17	51.72
SVM	100.00	100.00	100.00	100.00
ADA	82.76	93.10	100.00	100.00
GBM	82.76	93.10	100.00	100.00
MB	82.76	93.10	100.00	100.00
DT	89.66	89.66	100.00	100.00
RF	82.76	93.10	100.00	100.00
LMT	82.76	96.55	100.00	100.00
LOG	100.00	100.00	96.55	96.55
SL	82.76	96.55	100.00	100.00
MP	100.00	100.00	100.00	100.00
MC	100.00	100.00	96.55	96.55

Table 5: Overall Performance result TS2

Alg\C	LRCX%	MCHP%	MSFT%	NTAP
NB	100.00	100.00	100.00	96.55
BAG	96.55	93.10	89.66	96.55
STA	68.97	62.07	72.41	58.62
VOT	68.97	62.07	72.41	58.62
SVM	100.00	100.00	100.00	100.00
ADA	96.55	100.00	89.66	96.55
GBM	96.55	100.00	89.66	96.55
MB	96.55	93.10	89.66	96.55
DT	93.10	86.21	93.10	96.55
RF	96.55	100.00	93.10	96.55
LMT	100.00	100.00	89.66	96.55
LOG	96.55	93.10	100.00	96.55
SL	100.00	100.00	89.66	96.55
MP	100.00	100.00	100.00	100.00
MC	96.55	93.10	100.00	96.55

Table 6: Overall Performance result TS3 Final

Alg\C	QCOM%	SWKS%	Avg of all companies
NB	100.00	100.00	98.97
BAG	100.00	96.55	94.48
STA	37.93	48.28	57.59
VOT	37.93	48.28	57.59
SVM	100.00	100.00	100.00
ADA	96.55	96.55	95.17
GBM	96.55	96.55	95.17
MB	96.55	96.55	94.48
DT	96.55	96.55	94.14
RF	100.00	96.55	95.86
LMT	100.00	100.00	96.55
LOG	100.00	100.00	97.93
SL	100.00	100.00	96.55
MP	100.00	100.00	100.00
MC	100.00	100.00	97.93

The Ultimate Collaboration of Feature Selection Along with Classification Techniques

As discussed in section two, the majority of the latest and most innovative market research has been studied. Additionally, the latest feature selection approaches and classifier techniques available have been implemented in the project. The current market studies show the best results of prediction as 80 to 95% accuracy in most cases. This project's performance in some cases, achieved 100% accuracy of the classifier. To sum up, the Principal Component Analysis (PCA) with Multilayer Perceptron and Support Vector Machine (SVM) on added feature datasets shows 100% accuracy. Thus, the ultimate collaboration of feature selection, along with classifier techniques, achieved the highest results at 100%.

Compared with current studies in the literature, it can be demonstrated that our proposed prediction model outperforms the majority of studies. Mehta *et al.* (2019) achieved 99.2% accuracy employing support vector regression, whereas we achieved 100% accuracy.

Additionally, the thorough comparison with similar benchmark models (with other forecasting models) verifies the superiority of the proposed novel model. In particular, it allows the ultimate collaboration of feature selection along with classification techniques, which can rarely be found in the literature.

Conclusion

This study aimed to investigate and discover the best feature selection algorithm. PCA was found to be the best compared to others. Another objective was to test and explore the new features using feature engineering techniques. In this study, we have constructed two new features using mathematical procedures. The contribution of the added features has been found to be promising in the majority of cases.

Finally, we have also attempted to find the best classifier using 15 different machine learning algorithms. The ultimate collaboration with feature selection and feature engineering was also studied. The Support Vector Machine and Multilayer

Perceptron with PCA were found to outperform other algorithms and an accuracy of 100% was achieved in most cases. We used NASDAQ Stock Data, employing the data of 10 different companies to validate the proposed methodology.

To conclude, most studies in the literature have not paid attention to the use of binary classification concepts; instead, they more used numeric prediction. Likewise, it can be said this study is one of the most comprehensive studies in this area, using nine feature selection algorithms, feature engineering and 15 machine learning algorithms.

A few future works emerge from this paper. First, experiment with daily and weekly movements instead of using monthly movements. Second, it is also recommended that multi-label classification be considered along with the comparison of results with binary- label classification. Finally, the addition of numeric prediction to one project followed by a comparison between binary and multi-label classification could be promising.

Acknowledgment

I would like to show my massive appreciation for all supervisors of my Ph.D. Without them the work would not be possible. Their contribution was the key for publishing this paper.

Author's Contributions

Rebwar Nabi: This work has been accomplished as one of the taken papers from the Ph.D. Dissertation. He is the core contributor of this work.

Soran Saeed: This author was core supervisor of the thesis and work extensively towards preparing the article.

Habibollah Harron: This author was second supervisor of the thesis and work extensively towards preparing the article.

Hamido Fujita: This author was second supervisor of the thesis and work extensively towards preparing the article.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Alpaydin, E., 2014. Introduction to Machine Learning. 1st Edn., MIT Press, Cambridge, ISBN-13: 0262028182, pp: 613.

Caley, J.A., 2013. A survey of systems for predicting stock market movements, combining market indicators and machine learning classifiers. Dissertations and Theses.

Chen, N.F., R. Roll and S.A. Ross, 1986. Economic forces and the stock market. J. Bus., 383-403. DOI: 10.1086/296344

Chen, S.B., Y.M. Zhang, C.H.Q. Ding, J. Zhang and B. Luo, 2019. Extended adaptive lasso for multi-class and multi-label feature selection. Knowledge-Based Syst., 173: 28-36. DOI: 10.1016/j.knosys.2019.02.021

Fama, E.F., L. Fisher, M.C. Jensen and R. Roll, 1969. The adjustment of stock prices to new information. Int. Economic Rev., 10: 1-21. DOI: 10.2307/2525569

Feng, F., X. He, X. Wang, C. Luo and Y. Liu *et al*., 2019. Temporal relational ranking for stock prediction. ACM Trans. Inform. Syst. DOI: 10.1145/3309547

Graham, B., W.E. Buffett and J. Zweig, 2015. The Intelligent Investor: The Definitive Book on Value Investing. Harper Collins Publishers.

Hastie, T., R. Tibshirani and J. Friedman, 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 1st Edn., Springer, New York, ISBN-10: 0387952845, pp: 533.

Hinton, G.E. and T.J. Sejnowski, 1999. Unsupervised Learning: Foundations of Neural Computation. 1st Edn., MIT Press, Cambridge, ISBN-10: 026258168X, pp: 398.

Idrees, S.M., M.A. Alam and P. Agarwal, 2019. A prediction approach for stock market volatility based on time series data. IEEE Access, 7: 17287-17298. DOI: 10.1109/ACCESS.2019.2895252

King, B.F., 1966. Market and industry factors in stock price behavior. J. Bus., 39: 139-90. DOI: 10.1086/294847

Kirkpatrick, I.I. and J.A. Dahlquist, 2010. Technical Analysis: The Complete Resource for Financial Market Technicians. 2nd Edn., FT Press, ISBN-10: 0132599627, pp: 704.

Kuhn, M. and K. Johnson, 2013. Applied Predictive Modeling. 1st Edn., Springer, New York, ISBN-10: 1461468493, pp: 600.

Leottau, D.L., J. Ruiz-del-Solar and R. Babuška, 2018. Decentralized reinforcement learning of robot behaviors. Artificial Intell., 256: 130-59. DOI: 10.1016/j.artint.2017.12.001

Lison, P., 2015. An Introduction to Machine Learning. 1st Edn., Springer, Berlin, Germany.

Long, W., Z. Lu and L. Cui, 2019. Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Syst., 164: 163-73. DOI: 10.1016/j.knosys.2018.10.034

Mehta, S., P. Rana, S. Singh, A. Sharma and P. Agarwal, 2019. Ensemble learning approach for enhanced stock prediction. Proceedings of the 12th International Conference on Contemporary Computing, Aug. 8-10, IEEE Xplore Press, Noida, India. DOI: 10.1109/IC3.2019.8844891

Mohri, M., A. Rostamizadeh and A. Talwalkar, 2018. Foundations of Machine Learning. 1st Edn., MIT Press, ISBN-10: 0262039400, pp: 504.

Nayak, A., M.M. Manohara Pai and R.M. Pai, 2016. Prediction models for Indian stock market. Proc. Comput. Sci., 89: 441-49. DOI: 10.1016/j.procs.2016.06.096

Rawlings, J.O., S.G. Pantula and D.A. Dickey, 2001. Applied Regression Analysis: A Research Tool. 2nd Edn., Springer Science and Business Media, ISBN-10: 0387984542, pp: 660.

Subha, M.V. and S. Thirupparkadal Nambi, 2012. Classification of stock index movement using k-Nearest Neighbours (k-NN) algorithm. WSEAS Trans. Inform. Sci. Applic., 9: 261-70.

Yang, F., Z. Chen, J. Li and L. Tang, 2019. A novel hybrid stock selection method with stock prediction. Applied Soft Comput. DOI: 10.1016/J.ASOC.2019.03.028

Zemke, S., 2003. Data mining for prediction. Financial Series Case.

Zhou, L., K.P. Tam and H. Fujita, 2016. Predicting the listing status of Chinese listed companies with multi-class classification models. Inform. Sci., 328: 222-236. DOI: 10.1016/j.ins.2015.08.036