International Journal of Advanced Science and Technology
Vol. 29, No. 12s, (2020), pp. 2486-2496

Feature Engineering for Stock Price Prediction

*Rebwar M. Nabi[1], Soran AB. Saeed[2], Abdulrahman M. W. Abdi[3]

[1] *Sulaimani Polytechnic University, rebwar.nabi@spu.edu.iq*

[2]*Sulaimani Polytechnic University, soran.saeed@spu.edu.iq*

[3] *Kurdistan Technical Institute, ebdulrahman.ebdi@gmail.com*

Abstract

It is generally accepted that stock market is viewed as a multifaceted non-linear dynamic system, in which it is normally affected by numerous aspects. Previously, it had been stated that traditional analysis and forecasting methods are not performing well to accurately expose the intrinsic pattern for stock market. Consequently, big differences between expected and observed results. Nevertheless, recently many researchers have implemented several machine learning approaches to forecast the future stock market prices more accurately and precisely

Based on the literature, feature engineering has not been the interested of researchers for stock price prediction, Therefore, the aim of this study is to proposes a novel feature engineering approach to predict the stock prices based on historical data using both binary and multiclass classification. We have used 11 datasets from Nasdaq and S&P 500 Index to evaluate the accuracy of the proposed approach. More importantly, two new features were engineering and added to the original dataset. The feature engineering has improved the classification accuracy significantly. The overall prediction results improved by 25.64% compared to applying the same procedure without feature engineering. Last but not least, this study can be considered as the unique to use feature engineering for stock prediction.

Keywords*: Stock Market Price Prediction, Feature Engineering, Feature Selection, Machine Learning, WEKA, Gradient Boosting Machine, Multiclass Classification*

1. Introduction

The stock market prediction has attracted much attention from both academia and business. The question remains: "To what extent can the history of a common stock's price be used to make meaningful predictions concerning the future price of the stock [1]." Early research on stock market prediction was based on the Efficient Market Hypothesis (EMH) and the random walk theory [1], [2]. These early models suggested that stock prices cannot be predicted since they are driven by new information (news) rather than present/past prices. Therefore, stock market prices will follow a random walk and their prediction accuracy cannot exceed 50% [3].

Conversely, an increasing number of studies [4]–[14] provide evidence contrary to what is suggested by the EMH and random walk hypothesis. These studies show that the stock market can be predicted to some degree and therefore question the EMH's underlying assumptions. Many within the business community also view Warren Buffet's ability to consistently beat the S&P Index as a useful indicator that the market can be predicted.

Predicting stock prices is an essential objective in the financial world [7], [8], [12] since a reasonably accurate prediction can yield significant financial benefits and hedge against market risks. However predictable, it remains difficult to forecast the stock price movement, mainly as the financial market is a complex, evolutionary, and nonlinear dynamic system which interacts with political events, general economic conditions, and traders' expectations [12]. However, realizing accurate forecast of stock prices in the short term (1 day, 5 days ahead), medium-term (10 days, 15 days ahead), (20 days, 30 days ahead), and long term (Quarterly) is one of the most attractive and meaningful research subjects in the investment field and its applications. The benefits involved in inaccurate predictions have been motivating enthused researchers to

develop newer and more advanced tools and methods. In general, there are two common methods to predict the stock market prices, which are fundamental analysis and technical analysis. The former utilizes economic factors to estimate the intrinsic values of securities, whereas the latter relies on historical data on stock prices.

Feature engineering is a vast topic and more methods are being invented every day, particularly in the area of automatic feature learning. The basic concept of ML is data and model. Data could be stock market data containing 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.

When constructing new features [15], it is advantageous when the result is interpretable. Interpretable features and models are more accessible and lead to the most accurate model; it is a good idea to add complexity to improve the accuracy of classification. The goal of feature engineering, however, is not to make the feature dimensions as low as possible but to arrive at the right features for the task. Stock market data as numeric data are already in a format that is easily ingestible by mathematical models. A mathematical model of data that describes the relationships which predict stock prices might be a formula that maps a company's earning history, past stock prices, and industry to the predicted stock price. Researchers in [16] applied a deep learning approach for stock prediction and found that the results were promising. Besides, a feature engineering approach for educational data mining competition has been implemented using ensemble classifiers.

Moreover, several studies have been conducted that implemented feature engineering; however, none are related to stock prediction. Researchers in [17] used feature engineering fault diagnosis of induction motors. A semantic feature model in concurrent engineering was conducted by researchers in [18]. Another study used feature engineering for energy theft detection using gradient boosting and found useful combinations from the origin features [19]. The authors of [20] presented a feature engineering framework for short-term earthquake prediction based on AETA data. Feature engineering for search advertising recognition was investigated by researchers in [21].

Based on the literature it can observed that there is lack of studies which applied feature engineering for stock price prediction. Therefore, this study aims to propose a novel feature engineering method to predict the stock prices for monthly basis. It is worth mentioning that our study can be considered a pioneer in studying and implementing feature engineering for stock prediction utilizing ensemble methods. To support and prove this, we have searched and investigated international databases such as Science Direct, Elsevier, Scopus, IEEE Digital Library, Springer, and ACM. Furthermore, several other platforms and databases were investigated, for example, Google Scholar, EBSCO Information Services, and DOAJ.

We will study various ensemble methods as well as feature selection algorithms to showcase the importance of feature engineering.

The rest of the paper is structured as follows. In section two, the research methodology is explained. The results and discussion can be found in the following section. The conclusion of the paper is presented in section four.

2. Research Framework

Our research framework is composed of six main stages such as dataset collection, Data pre-processing, feature Engineering, applying feature selection and ensemble classifier, evaluation analysis. Figure one illustrates the overall research framework in this study.

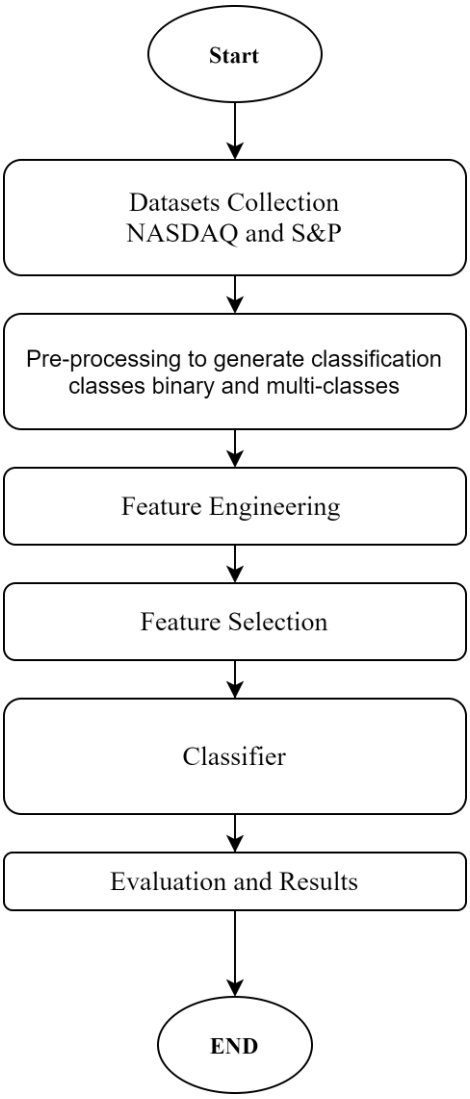

Figure 1: A research framework.

2.1. Dataset Collection

The NASDAQ and S&P 500 index datasets were downloaded. On the whole, 25 years of historical data were downloaded for the CMCSA, CSCO, AAPL, SBUX, LRCX, MCHP, MSFT, NTAP, QCOM, SWSK and in S&P 500 stocks. The duration of the data was Jan 1995 to Jan 2020. Altogether, we have mined 3270 months records as the monthly based prediction. The original downloaded data is daily bases and generally, each dataset has around 6294 records of the historical data. Generally, each dataset has six attributes:

1. Date: The current date of the stock movement.
2. A close price: The closing price of the stock.
3. Volume: Volume is commonly reported as the number of shares that changed hands during a given day.
4. open price: Open price of a stock.
5. high price: The highest price during a given day.
6. low price: The lowest price during a given day.

2.2. Pre-processing

It is widely known that pre-processing is considered as a crucial step in ML and data mining. Therefore, in our study, we propose a new approach to pre-process the data to generate binary (BC) and multiclass (MC) classes for monthly prediction. The stock movement to compare the predicted and real percentage change every month assigns the class to monthly data. To find the monthly movement here, stock movement is the difference between the monthly close and open price:

Difference = close price (last date of the month) – open price (first date of the month)

The stock price movement in terms of percentage (%) is calculated as follows:

Percentage_Difference = Difference / open price (first date of the month)

For assigning the classification class in a multiclass classification case:

If Percentage_Difference >1, then the class is positive;

If Percentage_Difference <-1, then the class is negative;

Otherwise, the class is neutral.

2.3. Feature Engineering (FE)

As mentioned earlier, the study aimed to explore and engineer few features to advance the classification accuracy. Two new features were added to the dataset. The first new feature is called High_Low_Difference (HL_Diff), which is the difference between the high and low price of the month. The mean value of close open difference as daily bases was also made. The following mathematical equations were used to harvest new features:

$$\mathrm{HL}_{\mathrm{Diff}} = High_{\mathrm{Max}} - Low_{Min} \qquad \text{eq (1)}$$

Where:
$High_{\mathrm{Max}}$ = Maximum high price of month
Low_{Min} = Minimum low price of month

$$Mean = \frac{\sum f\ (close - open)}{total\ days} \qquad \text{eq (2)}$$

Where:
$\sum f = Sum\ of\ the$ difference of close and open price
And:
$total\ days = number\ of\ days\ in\ the\ duration$

As can be seen in equation one, HL_Diff is generated by the difference between the whole month high and low price. It displays the whole month's maximum movement in the price. The mean is calculated based on the mean values of all differences between close and open prices. That shows the average movement in price.

2.4 Feature Selection (FS)

In this study, multiple feature selection algorithms were used to find the best feature selection algorithm for classification approach. It is worth mentioning that the WEKA's default configuration was implemented for all algorithms, which means no parameter configuration was considered since it was not within the scope of this study. Generally, the following algorithm was considered:

- Sequential Feature Selection (Best First) Search and CFS Subset Evaluation (SEQ)
- Genetic Search and CFS Subset (GEN)

- Ranker Search and Chi-Squared (CHI)
- Ranker Search and Recursive Feature Elimination (REF)
- Ranker Search and Correlation Coefficient (CC)
- Ranker Search and Info Gain Evaluation (IG)
- Ranker Search and ReliefF and it is Variant Evaluation (RV)
- Ranker Search and Principle Components Analysis Evaluation (PCA)

2.5 Implementation of Ensemble Classifier Techniques

As the classifier, we implemented eight ensemble learning algorithms, all of which are used as a default configuration in the WEKA application programming interface library, which technically means we did not play with the parameters, base learners, and other parameters. The following algorithms were chosen for this study:

- Bagging Classifier (BAG)
- Stacking Classifier (SC)
- Voting Ensemble Classifier (VE)
- AdaBoost Classifier (ADA)
- Gradient Boosting Machine (GBM)
- Multi-Boosting Classifier (MB)
- Random Forest (Random Subspace) (RF)

2.6 Evaluation Method

For evaluation, we have followed and used the WEKA library default evaluation methods [22]. We have used CA as the evaluation metrics. For reliable testing and results, we have divided our data into training and testing. Approximately, 65% used for training purposes, and 35% are used for testing. Since we are using the default WEKA evaluation, we are not going to provide the equations and mathematical details behind them as it would lead to repetition. Details of all the evaluation methods can be found in [22], [23].

3. Experimental Results and Discussion

In this section, the results will be demonstrated for both binary and multiclass classifications. Additionally, the experiments which have been steered to accomplish the research aim will also be established. As we have stated earlier, the results will be presented based on the WEKA evaluation methods. We will be mostly using Classification Accuracy (CA) as it is extensively acknowledged as the most spontaneous evaluation metric.

We steered numerous trials on 11 companies' dataset in NASDAQ and S&P 500 index composite stocks. The duration of the data is Jan 1995 to Jan 2020. for the monthly prediction we have extracted total of 3270 months records. We have used 65% for training and 35% for testing.

Firstly, we have showed experiments on all datasets using the classifiers and FS algorithms applied in this study. For each dataset a separate graph is produced. To compare and evaluate feature engineering, each graph displays two dataset results. Firstly, the results with feature engineering (FE) are represented by a solid-line. However, the results without feature engineering are represented by a dotted line. An example of the generated graph is shown in figure two The example of a performance comparison

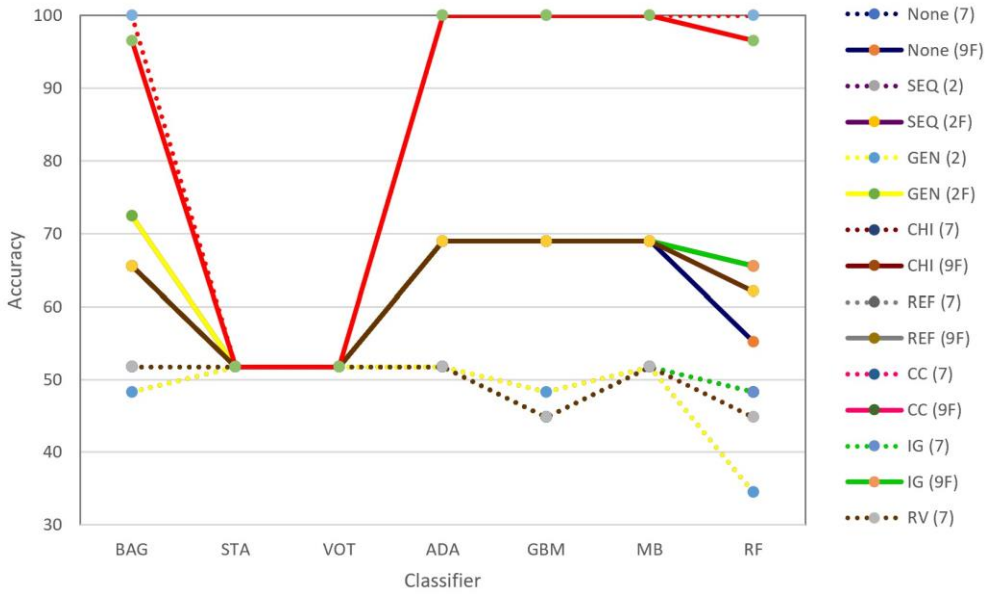

Figure 2: CA result graph sample.

graph shown in figure two contains the following:

- The X-Axis indicates the classifier techniques
- Y-Axis denotes CA in percentage
- Various lines and columns, in which the (F) mean with feature engineering, otherwise without feature engineering
- Colors of the line with a different color for each feature selection algorithm

Figure three illustrates the overall MC classification prediction for LRCX dataset.

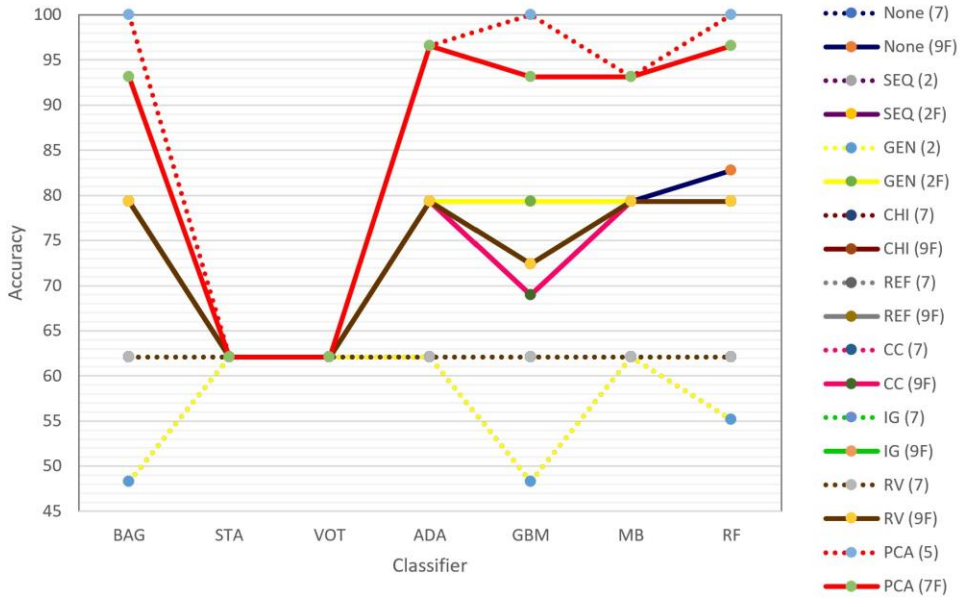

Figure 3:Multiclass Classification result for LRCX company dataset.

As demonstrated in figure three in the popular cases, the accuracy is expressively enhanced. For instance, when GEN feature selection is considered, the feature engineering has improved the classification accuracy significantly, in which for the BAG algorithm, the accuracy of approximately 50% is achieved, whereas, without feature engineering, less than 50% is achieved.

Furthermore, the overall CA result for SWKS company dataset is exposed in figure four Alike to the LRCX company, feature engineering has subsidized meaningfully to improving the classification accuracy. For example, the RF with GEN achieves approximately 67% accuracy when tested with FE, while around 50% accuracy is attained without FE. To sum up, on the SWSK company dataset, PCA is also considered as the best feature selection algorithm, and the Genetic Algorithm comes second. Lastly, FE contributed astonishingly to enahnce the overall CA.

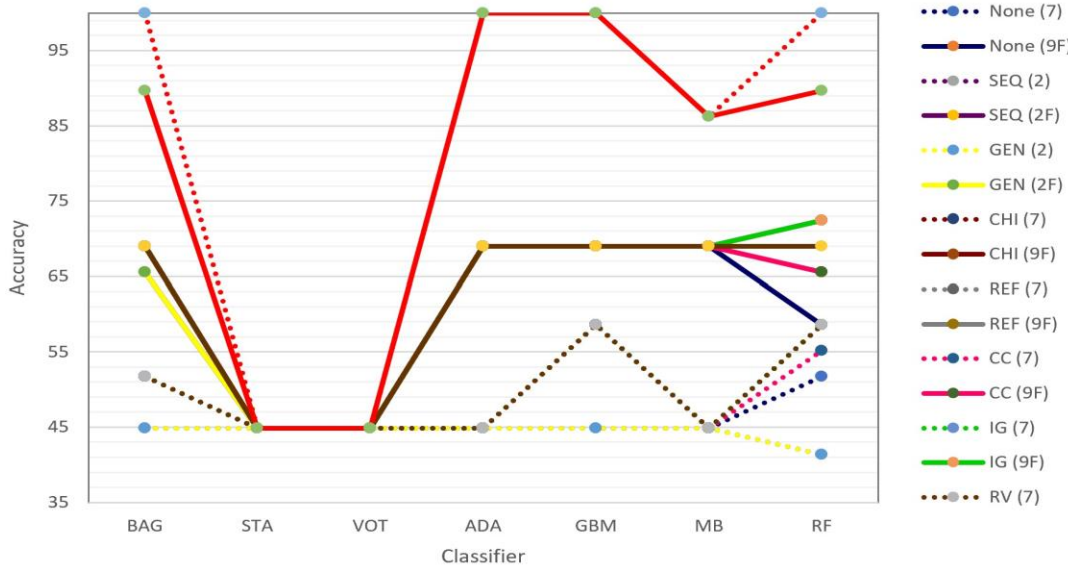

Figure 4: MC Classification result for SWSK dataset.

Similarly, the MC classification, in this section, the experiment results for binary classification is offered. Figure five exemplifies the BC results over several feature selection and classification algorithms on the CMCSA dataset. It can be recognized that FE has contributed massively, in which the CA is climbed to 75% approximately as it is represented with solid-yellow line for FE.

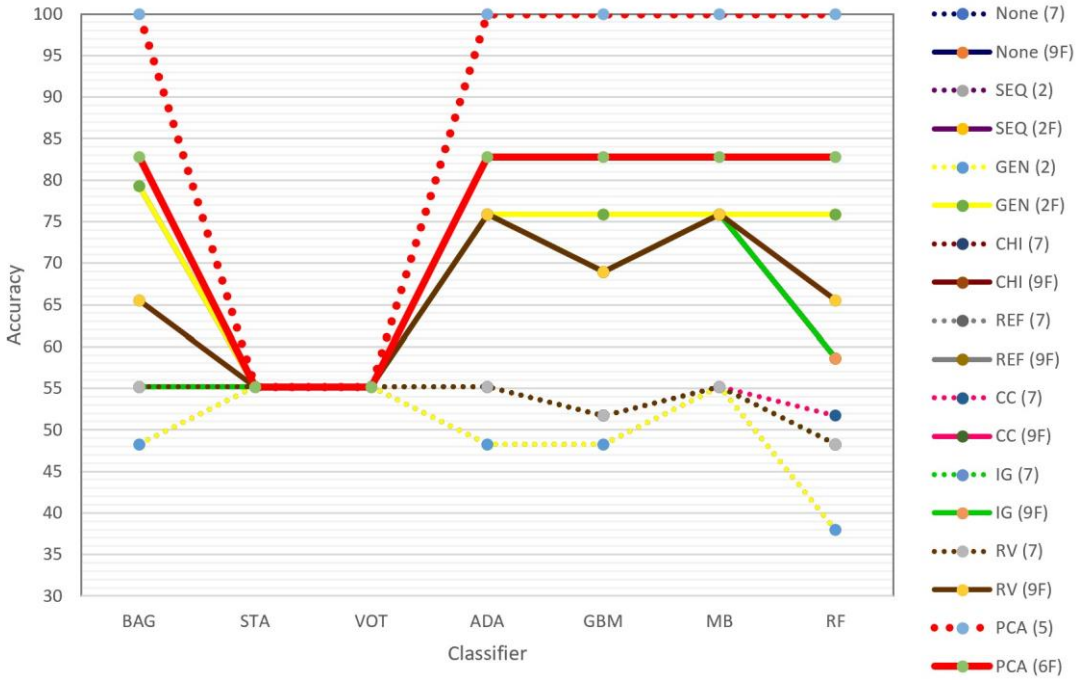

Figure 5: BC

classification result for CMCSA dataset.

the QCOM company dataset is also selected to display case an experiment. The forecast result is revealed in figure six.

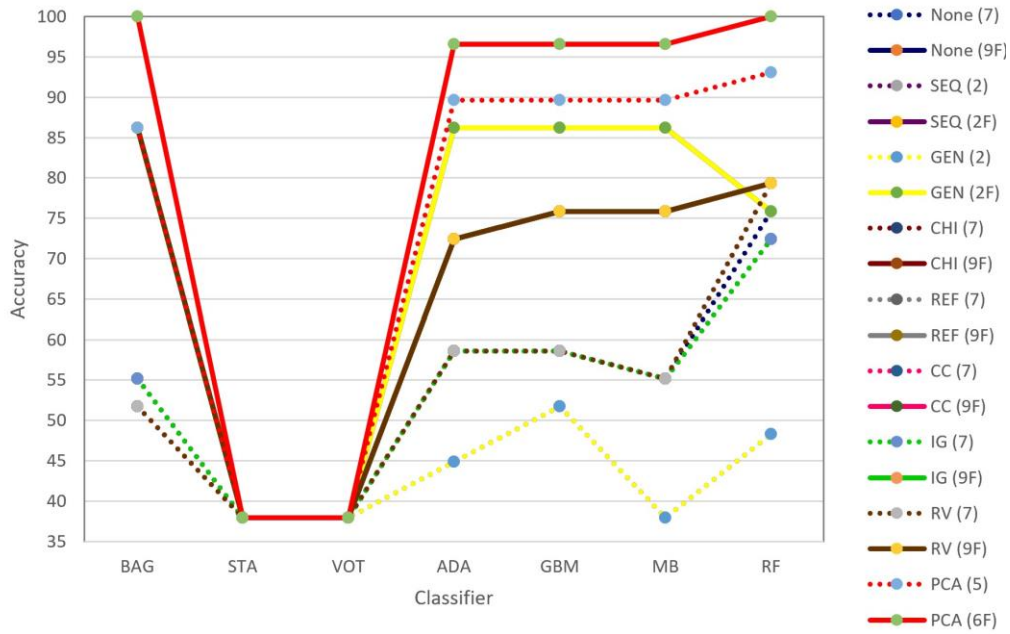

Figure 6: BC classification result for QCOM dataset.

For the QCOM dataset, in the majority cases, the predictions were auspicious. In contrast to earlier findings with CMCSA, the Correlation Coefficient with FE was found to be outperforming, where, in some cases, an accuracy of 85% was achieved when it is tested with ADA, GBM and MB Algorithms. More importantly, the FE has meaningfully upgraded CA when we use of GEN with the BAG Classifier. The Solid yellow line denotes the forecast with the FE by triumph the CA of 80%. Though, the dotted yellow line displays the forecast without FE, in which roughly 47% is reached.

3.1 Feature Engineering Benchmark With WEKA

To discover the influence and the implication of the projected FE approach, we steered intensive experiments on several datasets. Tables one and two reveal the overall forecast of the suggested method with FE and without FE using GBM as the classifier. The motive behind the benchmark with WEKA is that we were not find any study or paper, which painstaking FE for stock prediction. Table one shows the CA result with feature Engineering.

Table 1: CA result with FE.

Alg/DS	CA% CSCO	CA% AAPL	CA% SBUX	CA% LRCX	CA Avg.%
SEQ	69.23	75	66.34	79.8	72.59
GEN	69.23	75	66.34	79.8	72.59
CHI	70.19	72.11	67.3	72.11	70.43
RFE	70.19	72.11	67.3	72.11	70.43
CC	70.19	72.11	67.3	72.11	70.43
IG	70.19	72.11	67.3	72.11	70.43
RV	70.19	72.11	67.3	72.11	70.43
PCA	100	100	100	100	100.00
Average					74.67

Instead, Table two shows the CA result without feature Engineering.

Table 2: CA result without FE.

Alg/DS	CA% CSCO	CA% AAPL	CA% SBUX	CA% LRCX	Avg.%
SEQ	44.55	46.15	45.19	51.92	46.95
GEN	44.55	46.15	45.19	51.92	46.95
CHI	46.53	52.88	42.3	57.69	49.85
RFE	46.53	52.88	42.3	57.69	49.85
CC	46.53	54.8	42.3	57.69	50.33
IG	46.53	52.88	42.3	57.69	49.85
RV	46.53	50	42.3	57.69	49.13
PCA	100	100	93.26	99.03	98.07
Average					55.12

Tables one and two define the contribution of the projected FE model. The average CA result on all datasets with all feature selection algorithms is exceptionally enhanced. The average CA for SEQ and GEN is remarkably increased with FE by reaching 69.23%, while the average CA is only 44.55% without FE. Similarly, the average CA is enlarged by 25.64% when the CHI, RFE, CC, IG, and RV are used. Furthermore, the average CA is also improved with PCA with FE by 1.93%. Finally, on average, the CA with FE is 55.12% while it is increased to 74.67% with feature engineering.

Based on the result of the above experiments, it can be supposed that FE advances the CA extremely. There is the conceivable reason behind this enhancement. Initially, new features can make the results interpretable, in which the interpretable features and models are more accessible and lead to the most accurate model [15]. More importantly, sometimes it is a good idea to add complexity to improve the accuracy of classification. However, is not to make the feature dimensions as low as possible but to arrive at the right features for the task. Stock market data as numeric data are already in a format that is easily ingestible by mathematical models. Lastly, engineers in WEKA emphatically stated that feature engineering regarded as one of the best ways to improve the overall machine learning algorithm.

To the best of our knowledge and based on literature our study is considered to be extremely rare to investigate and implement feature engineering for stock prediction, especially with ensemble methods. To support and prove this claim we have searched and investigated the international databases such as Science Direct[1], Elsevier[2], IEEE[3], Springer[4] and ACM[5]. Furthermore, few other platforms and databases were investigated, for example scholar[6], EBSCO Information Services and DOAJ. As a result, only one paper has been found, which the states that they have applied feature engineering in their study [24].

They proposed a multi-filters neural network using deep learning methodologies for feature engineering called (MFNN) for regression domain. The MFNN integrates the convolutional and recurrent neurons to build the multi-filters structure in order to the information from different

[1] https://www.sciencedirect.com/

[2] https://www.elsevier.com/en-xm

[3] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=2188

[4] https://www.springer.com/gp

[5] https://www.acm.org/

[6] https://scholar.google.com/

feature spaces and market views can be obtained. Since, they adopted a different type of classification, they have used different metrics to evaluate their model. Therefore, it was challenging to find an appropriate benchmark. However, based on their conclusion they stated that their model was at the best level, improved by 7.19% compare with traditional machine learning models as well as algorithms. But our approach has approved the classification accuracy by 25.64% on majority cases and sometimes higher, which means our approach beats their model.

On the other hand, it is evident that there are huge differences between feature engineering and feature extraction. In their study, they did not mention what features they constructed rather they only mention the feature extraction. Based on the literature, feature extraction is considered either as dimensional reduction or changing the feature into the desired form[25].

4. Conclusion

The study aimed to propose a novel feature engineering approach for stock prediction. This study collected Nasdaq and S&P 500 index listed stocks for the last 25 years as the dataset. The dataset included data of various companies, such as CMCSA, CSCO, AAPL, SBUX, LRCX, MCHP, MSFT, NTAP, QCOM, and SWKS. Monthly stock movement is predicted for each month. We have implemented feature engineering to add two features to the dataset as 1. Mean value of Open and Close price difference and 2. The high low difference, which is part of feature engineering that improved the performance, shows in the results as increasing accuracy. The technology uses the interface of Java and WEKA to judge varied styles of feature selection and classifier over the given dataset. For the feature selection part, various algorithms were applied, which are CFS Subset, Chi-Squared, Recursive Feature Elimination, Correlation Coefficient, Info Gain, Relief and its Variant, PCA, Sequential Feature Selection (Best First), Genetic Search, and Ranker Search, for ML techniques applied different classifiers on datasets, such as Stacking, AdaBoost, GBM, Multi-Boosting, and Random Forest. We tested all the techniques using classification on stock market movement as positive, negative, and neutral.

In general, the CA accuracy enhanced by our approach has approved the classification accuracy by 25.64% on majority cases and sometimes higher.

Acknowledgments

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. Furthermore, many thanks to my university (SPU) for providing me with such amazing opportunity to study PhD.

References

1. E. F. Fama, “The Behavior of Stock-Market Prices,” *J. Bus.*, vol. 38, no. 1, pp. 34–105, 1965.
2. E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll, “The Adjustment of Stock Prices to New Information,” *Int. Econ. Rev. (Philadelphia).*, vol. 10, no. 1, pp. 1–21, 1969.
3. J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market.”
4. M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, “Evaluating multiple classifiers for stock price direction prediction,” *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, 2015.
5. Y. Chen and Y. Hao, “A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction,” *Expert Syst. Appl.*, vol. 80, pp. 340–355, Sep. 2017.
6. T. A., “Improvement on Classification Models of Multiple Classes through Effectual Processes,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, 2015.
7. E. Chong, C. Han, and F. C. Park, “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,” *Expert Syst. Appl.*, vol. 83, pp. 187–205, Oct. 2017.

8. R. T. Farias Nazário, J. L. e Silva, V. A. Sobreiro, and H. Kimura, "A literature review of technical analysis on stock markets," *Q. Rev. Econ. Financ.*, vol. 66, pp. 115–126, 2017.
9. G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
10. L. Wang, Z. Wang, S. Zhao, and S. Tan, "Stock market trend prediction using dynamical Bayesian factor graph," *Expert Syst. Appl.*, vol. 42, no. 15, pp. 6267–6275, 2015.
11. A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," *J. Econ. Financ. Adm. Sci.*, vol. 21, no. 41, pp. 89–93, 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. B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Syst. Appl.*, vol. 79, pp. 153–163, Aug. 2017.
14. Y. Zhao, J. Li, and L. Yu, "A deep learning ensemble approach for crude oil price forecasting," *Energy Econ.*, vol. 66, pp. 9–16, 2017.
15. U. Khurana, D. Turaga, H. Samulowitz, and S. Parthasrathy, "Cognito: Automated feature engineering for supervised learning," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 1304–1307.
16. W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," *Knowledge-Based Syst.*, vol. 164, pp. 163–173, 2019.
17. P. S. Panigrahy, D. Santra, and P. Chattopadhyay, "Feature engineering in fault diagnosis of induction motor," in *2017 3rd International Conference on Condition Assessment Techniques in Electrical Systems, CATCON 2017 - Proceedings*, 2018, vol. 2018-Janua, pp. 306–310.
18. Y. J. Liu, K. L. Lai, G. Dai, and M. M. F. Yuen, "A semantic feature model in concurrent engineering," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 3, pp. 659–665, Jul. 2010.
19. R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019.
20. J. Huang, X. Wang, S. Yong, and Y. Feng, "A feature enginering framework for short-term earthquake prediction based on AETA data," in *Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, ITAIC 2019*, 2019, pp. 563–566.
21. Y. Sun and G. Yang, "Feature engineering for search advertising recognition," in *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 2019, pp. 1859–1864.
22. E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."* 2016.
23. M. Swamynathan, *Mastering Machine Learning with Python in Six Steps - review and good into in ML and NN approaches and basics + Python samples --Each topic has two parts: the first part will cover the theoretical concepts and the second part will cover practical impleme*, vol. 19, no. 2. 2017.
24. W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," *Knowledge-Based Syst.*, vol. 164, pp. 163–173, Jan. 2019.
25. S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.