Proceedings of ICFTBA 2024 Workshop: Finance in the Age of Environmental Risks and Sustainability
DOI: 10.54254/2754-1169/94/2024OX0151

The Comparison of Feature Selection Methods and Feature Combinations Based on the Stock Prediction

Xiaowan Shen[1,a,*]

[1]*Guanghua Cambridge International School, Shanghai, 201319, China*
a. wilda@usf.edu
**corresponding author*

Abstract: Predicting short-term stock prices is a significant and challenging research area due to market volatility. Machine learning (ML) uses algorithms to learn patterns from data, improving prediction accuracy over time. Feature selection (FS) methods enhance model accuracy and efficiency. Evaluating and selecting the best FS methods and feature combinations is essential for improving prediction performance. This paper evaluates three feature selection (FS) methods by scoring technical indicators and using three models with 30 different indicator combinations to predict outcomes. Error rates are used to measure accuracy. The analysis reveals that incorporating all features yields the lowest average error rate in this paper, and Williams R is attached to the greatest importance in this paper. Among the FS methods, Mutual Information (MI) and Random Forest (RF) outperform the correlation coefficient. Future work can focus on two main areas: considering more indicators and combinations, and exploring additional feature selection (FS) methods to identify the best one.

Keywords: Stock Prediction, Machine Learning, Feature Selection, Error Rates.

1. Introduction

Predicting stock price in the short term is a research subject that has been widely and fully discussed. It is challenging due to the market's volatile nature. This difficulty attracts researchers and academics to improve prediction models. Accurately interpreting crucial stock information early can lead to profitable trading.

Machine learning (ML) involves the use of algorithms that learn from data and build a standard that minimizes error rates. These algorithms analyze large datasets to identify correlations, trends, and relationships that might not be evident through traditional analytical methods. By iteratively improving their performance based on feedback from the data, machine learning models can make increasingly accurate predictions and decisions.

Machine learning plays a significant role in the field of financial markets, particularly in the predicting stock price movements. In this context, machine learning methods are used to analyze vast amounts of historical price data, trading volumes, market sentiment, and other relevant financial indicators. By identifying patterns and trends within this data, machine learning models can forecast future stock prices or market trends with a higher degree of accuracy than traditional statistical methods.

DOI: 10.54254/2754-1169/94/2024OX0151

Technical indicators as features with machine learning methods can be used to predict stock price. In particular, the feature selection (FS) methods can improve accuracy and computation time by selecting proper features [1, 2]. There are various FS methods including filter methods and Wrapper methods with different criteria [3]. Choosing one proper FS method is necessary in predicting the stock price. Hence, there is need to evaluate these FS methods by comparing their FS scores with their performance in models. In addition, it is also essential to select the correct combination of features after evaluating these features. Therefore, the comparison of prediction of stock with different feature combinations is the other objective of the paper.

2. Literature Review

The past literature demonstrates that feature selection can enhance prediction performance, scalability, and the classifier's generalization ability. In the realm of knowledge discovery, feature selection is crucial for minimizing computational complexity, storage requirements, and costs [1]. Chen & Hao's paper, and Ram ırez-Gallego et al.'s paper has used the information-theory based FS method in their models [2, 3]. Naik & Mohan's study carried out Boruta feature selection technique with a number of technical indicators [4].

Apart from FS methods mentioned above, Venkatesh & Anuradha's study and Htun et al.'s study both reviewed three types of common FS methods [5, 6]. Among these methods, the Pearson's correlation method and the mutual information method (MI method) stand out for their convenience and straightforward interpretation. Random Forest (RF) has gained popularity as a feature selection method due to its numerous beneficial attributes, including internal error estimates, correlation measurements, and feature importance scores. These three FS methods are used in this paper.

Singh et al.'s paper reviews the machine learning algorithms that are appropriate for stock prediction and discusses the current tools and techniques [7]. Logistic regression, commonly used for classification problems, calculates the probability of an event being a success or failure. Support Vector Machine (SVM) classifies data by finding the hyperplane that maximizes the margin between two classes. Decision Tree builds classification models in a tree structure. These three models are among the most basic and widely used for classification tasks. This paper employs them to predict stock prices.

The comparison of feature selection methods has been made on many types of datasets such as biomedical informatics. and text classification with different ways of testing [8-10]. This paper compares three basic FS methods with technical indicators and stock datasets. The novelty point is that instead of complex algorithms, the performance of models with the different indicator combinations as features is considered as the measure of FS methods. on top of that, 30 combinations of indicators are tested in this paper to give a relatively more accurate and reliable comparison among FS methods. This has not been considered in the paper before.

3. Methodology

The whole experiment consists of three main steps. First, technical indicators are calculated with datasets. Then, the FS score is given to each indicator. Subsequently, the paper uses machine learning methods with 30 indicator combinations to predict the stock and calculate the error rates. Finally, the error rates and FS scores are compared to see whether they show similar tendency so that it can evaluate FS methods and research on the performance of different indicator combinations.

3.1. Data Collection

In this paper, stock data are obtained from https://uk.finance.yahoo.com/ for three companies including Tesla, BYD and Xiaomi (see Table 1). The time ranges of the data for Tesla and BYD are

DOI: 10.54254/2754-1169/94/2024OX0151

from 2014.5.12 to 2024.5.10. The time range of the data for Xiaomi is from 2018.8 to 2024.5.12. In this scenario, the decision is whether the stock will increase or decrease.

Table 1: Partial dataset of Tesla.

Date	Open	High	Low	Close	Adj Close	Volume
2014-05-12	12.258000	12.479333	11.992000	12.311333	12.311333	105034500
2014-05-13	12.250667	12.756000	12.200000	12.677333	12.677333	106458000
2014-05-14	12.596667	12.898667	12.473333	12.708000	12.708000	81100500
2014-05-15	12.665333	12.844000	12.353333	12.572667	12.572667	90606000
2014-05-16	12.596667	12.802667	12.514667	12.770667	12.770667	67315500

3.2. Technical Indicators

Technical indicators used include Simple Moving Average(SMA), Exponential Weighted Moving Average (EWMA), Momentum Indicator (MOM), Relative Strength Index (RSI) and Williams R (R). Table 2 is listed to show calculation of indicators. Table 3 is the meaning of the notation in Table 2.

Table 2: Technical indicators and relevant formulas.

Technical indicators	Calculation	Number of days used (n)
Simple Moving Average (SMA)	$(C_t + C_t{-}_1 + ...C_t{-}_{n+1})/n$	5,10
Exponential Weighted Moving Average (EWM)	$(C_t - SMA_{t-1}) * (2/n + 1) + SMA_{t-1}$	none
Momentum Indicator (MOM)	$Ct - Ct{-}n - 9$	5,10
Relative Strength Index (RSI)	$100 - (100/(1 + Avg_n(Gain)/Avg_n(Loss)))$	14
Williams R (WR)	$((H_n - C_t)/(H_n - L_n)) * 100$	14

Table 3: Notation definition.

Notation	Meaning
n	Total number of days.
H_n	the highest day stock price in period of n days .
L_n	the lowest day stock price in period of n days .
C_t	day close stock price at time t.
$Avg_n(Gain)$	average day gain in period of n days.
$Avg_n(loss)$	average day loss in period of n days.

3.3. Feature Selection Methods

3.3.1. Pearson's Correlation

The Pearson Correlation (PC) can be used to detect the linear relationship between two variables. The following equation is used to calculate the PC (ρ) between two variable, x and y [5].

$$\rho(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

DOI: 10.54254/2754-1169/94/2024OX0151

3.3.2. Mutual Information (MI)

Mutual Information (MI) is information-theory-based method used in feature selection (FS). It measures the mutual dependence between two variables (X and Y). MI evaluates the "amount of information" about one random variable carried by the other random variable. The following equation is used to calculate MI between two discrete random variables, x and y.

$$I(X,Y) = \sum_{x\in B}\sum_{y\in A} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

Where p(x, y) is the joint probability function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively [5].

3.3.3. Random Forest (RF)

Random Forest (RF) is a learning method used for both classification and regression problems. It employs bootstrapped aggregation (bagging) and random feature selection techniques. Random forest can also be used in feature selection. Random Forest (RF) can produce two types of importance scores: mean decrease accuracy (MDA) and mean decrease impurity (MDI) [6]. This paper uses the Gini index as a measure of MDI.

3.4. Machine Learning Methods

3.4.1. Logistic Regression

Logistic Regression assumes the probability of one class p(x) to be:

$$p(x) = \frac{e^{\beta_1 X_1+\beta_2 X_2+\cdots}}{1+e^{\beta_1 X_1+\beta_2 X_2\cdots}} \tag{3}$$

X_n is the predictor; β_n is the coefficient awaiting for being estimated. The equation can be transformed into linear form. Then, β_n can be estimated by the maximum likelihood method.

3.4.2. Support Vector Machine

The support vector classifier classifies into two sides of a hyperplane, as Figure 1 shown. The hyperplane is selected to correctly separate most of the training observations. Only the observations that lie near the margin change the position of the hyperplane. The movement of other observations does not alter the classifier, as long as they do not cross the hyperplane.

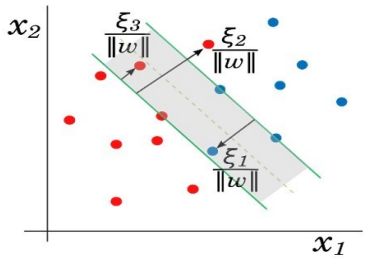

Figure 1: Support vector machine [7].

DOI: 10.54254/2754-1169/94/2024OX0151

3.4.3. Gradient Boosting Decision Trees

Boosting is a learning technique that enables decision trees to perform better. Gradient boosting works as following: Models are trained sequentially, with each new model fitting the residual errors of the combined ensemble of all previous models. The predictions of all models are combined through a weighted sum.

4. Results

There is huge amount of data. To be straightforward and clear, this paper only shows the key points and list the data.

4.1. Feature Selection Score

The FS scores of indicators in all three cases give a similar tendency. With respect to correlation coefficient, all indicators have a relatively lower value. WR has a much higher score in both RF score and MI score than all other indicators. RSI and MOM show their relatively higher importance in RF score. A heat map of the scores is shown as Figure 2.

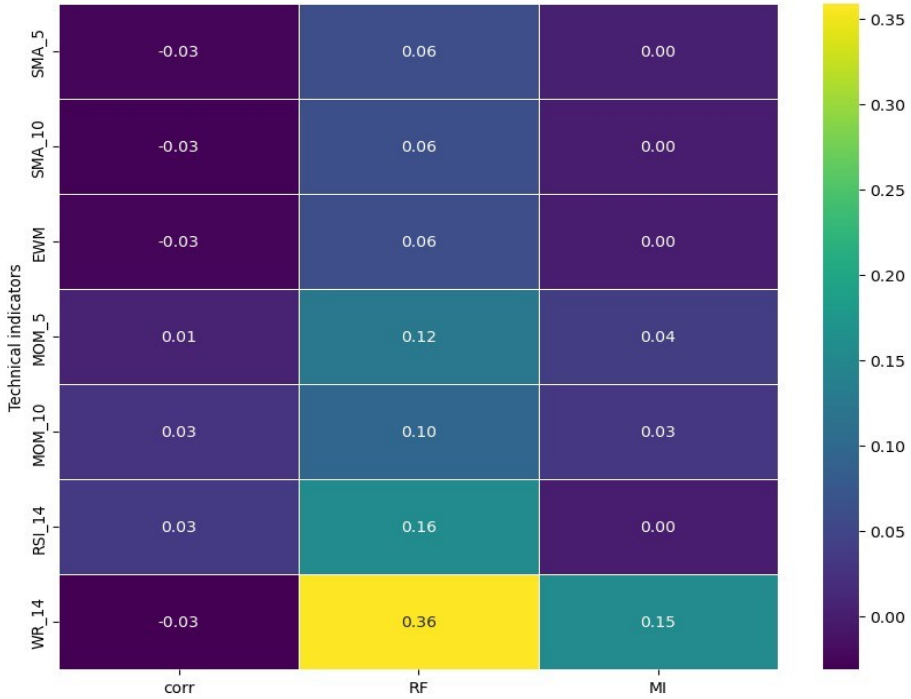

Figure 2: The heat map of the FS score in the case of Tesla.

To sum up, correlation criteria tells all indicators share equal importance, while RF and MI lay their emphasis on RSI and WR.

4.2. Error Rates of Prediction

The prediction is repeated with these three companies, and the three results are average. The following mainly discusses the average results. First of all, the prediction with WR shows relatively lower average error rates. Figure 3 for average error rates is shown below. The red bars are error rates of the prediction with WR and the blue bars are the ones without WR.

DOI: 10.54254/2754-1169/94/2024OX0151

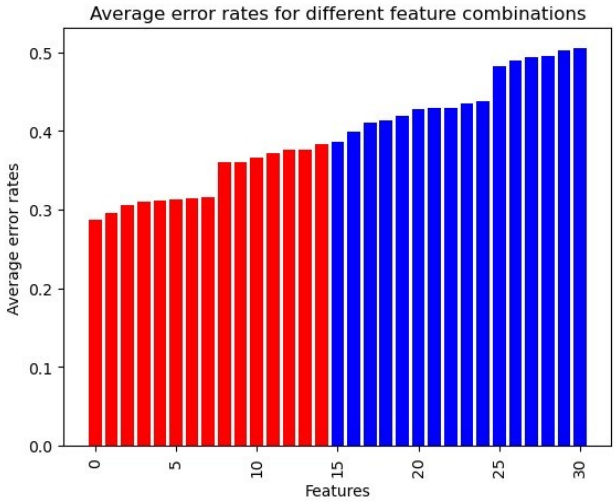

Figure 3: Average error rates of prediction with different feature combinations.

Additionally, considering all the features in a model gives the lowest average error rate. This might because the accuracy of model increases as the information gained by the model increases. Error rates table of 5 feature combinations with the lowest error rates is shown in Table 4 below.

Table 4: Error rates of 5 feature combinations with the lowest error rates.

Feature Combination	LR	SVM	BDT	Average
['SMA_5', 'SMA_10', 'EWM', 'MOM_5', 'MOM_10', 'RSI_14', 'WR_14']	0.350333	0.260000	0.250000	0.286778
['MOM_10', 'RSI_14', 'WR_14']	0.359333	0.257667	0.271000	0.295667
['SMA_10', 'MOM_10', 'EWM', 'WR_14', 'RSI_14']	0.383333	0.233333	0.295000	0.303222
['EWM', 'MOM_10', 'WR_14', 'RSI_14']	0.392333	0.268333	0.283333	0.314000
['RSI_14', 'WR_14']	0.395000	0.268333	0.275000	0.312111

MI and RF score give a consistent conclusion with performances of the error rates of prediction. It is that WR is attached to the greatest importance. In contrast, correlation criteria do not give a strong and consistent selection standard.

5. Conclusion

This paper uses three FS methods to give scores to technical indicators and use three models with 30 different indicator combinations to predict. The error rates is used as the measure of accuracy. Finally, the FS scores and error rates are analyzed to research on which feature combination can improve the performance of models better in this case and which FS methods perform better.

In conclusion, incorporating all the features in a model results in the lowest average error rate. This could be because the model's accuracy improves with an increase in the information it receives. Furthermore, MI and RF do perform better than correlation coefficient. The reason might be the linear nature of correlation criteria. In fact, much literature on models with correlation criteria as FS method usually has another FS method to help select features.

Future work can be focused on mainly two parts. First, more indicators and combinations can be considered. The generalization is very limited in this paper because only five technical indicators are considered. Besides, more FS methods might be explored so that the range of the choices can be larger and a relatively better FS method can be selected out.

DOI: 10.54254/2754-1169/94/2024OX0151

References

[1] *Khan, N. M., Madhav C, N., Negi, A., & Thaseen, I. S. (2019). Analysis on Improving the Performance of Machine Learning Models Using Feature Selection Technique. Advances in Intelligent Systems and Computing, 69–77.*

[2] *Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications, 80, 340–355.*

[3] *Ram´ırez-Gallego, S., Mourin˜o-Tal´ın, H., Mart´ınez-Rego, D., Bolo´n-Canedo, V., Ben´ıtez, J. M., Alonso-Betanzos, A., & Herrera, F. (2018). An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(9), 1441–1453.*

[4] *Naik, N., & Mohan, B. R. (2019). Optimal Feature Selection of Technical Indicator and Stock Prediction Using Machine Learning Technique. Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics, 261–268.*

[5] *Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies, 19(1), 3–26.*

[6] *Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. Financial Innovation, 9(1).*

[7] *Singh, N., Khalfay, N., Soni, V., & Vora, D. (2017). Stock Prediction using Machine Learning a Review Paper. International Journal of Computer Applications, 163(5), 36–43.*

[8] *Phyu, T. Z., & Oo, N. N. (2016). Performance Comparison of Feature Selection Methods. MATEC Web of Conferences, 42, 06002.*

[9] *Drota´r, P., Gazda, J. & Sm´ekal, Z. (2015). An experimental comparison of feature selection methods on two-class biomedical datasets. Computers in Biology and Medicine, 66, 1–10.*

[10] *Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., & Mahoney, M. W. (2007). Feature selection methods for text classification. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’07.*