The Study of Using Machine Learning Algorithms to Construct Portfolio Formation

Zefeng Chen[1,*]

[1] *Jinan University - University of Birmingham Joint Institute, Jinan University, Guangzhou 511400, China*
[*] *Corresponding author. Email:chenzefeng2019051719@stu2019.jnu.edu.cn*

ABSTRACT
The earning of the investors in the stock market is related to their investment choice and the behavior of the stocks. The key to investors making a more intelligent investment decision is to predict the price and volatility of the stocks accurately. In addition, building a portfolio formation is also an effective way to reduce investment risk. However, the stock price is beyond the predictive ability of any individual, and the portfolio formations are hard to build due to the stock market volatiles incessantly. In such a condition, this project attempts to propose a methodology for designing portfolio formations by using machine learning methods. The ultimate aim of this project is to predict the profit and volatility of stocks in the American stock market and use these predictions to construct portfolio formations. In this paper, the Support Vector Machine (SVM), Light Gradient Boosting Machine (LGBM), and Long Short-Term Memory model (LSTM) are used to predict the tendency of stocks. The experiment result shows that the LGBM model has the best performance on stocks price prediction. Based on the prediction of price, investing profit is easy to predict. The financial model called exponential weighted moving average (EWMA) is used to calculate the volatility of the stocks. After getting the prediction of profit and volatility, using the spectral clustering algorithm can get the portfolio formation. The result shows that the portfolio formation given by the proposed model has high profit and low volatility.

Keywords: *Stock Prediction, EWMA Model, Clustering Algorithms, Machine Learning, Portfolio Formation*

1. INTRODUCTION

The stocks variate every day, and the return that the investors can get is not stable. If the investors only buy one stock, they may lose most of their money when it is volatile. Therefore, constructing a suitable portfolio formation can help investors avoid the risk of losing money. A good portfolio formation should contain stocks with high profit and low volatility. It ensures that the yield will not fluctuate much and can minimize the risk of losing money. However, the traditional methods to construct a portfolio formation need much experience, and its component may be affected by the constructor's preferences. The purpose of this project is to find a method to generate an appropriate portfolio formation automatically.

Machine learning algorithms have been used by several researchers on the stock market. Das et al. used backpropagation neural networks and SVM to construct the models to predict future prices in the Indian stock market [1]. Ji et al. proposed a new prediction method based on deep learning technology [2]. K. Hiba Sadia et al. reviewed the use of random forest, SVM on the stock dataset [3]. Lin et al. put forward a new SVM-based approach for predicting the stock trend. The proposed model contains the feature selection part and prediction model part. In the feature selection part, the SVM filter is used to select a subset of financial indexes. This selected subset contains the features with a high correlation with the tendency of stock. After selecting the features, a SVM implemented by the piecewise linear principle is applied in the prediction model part. The optimal separating hyperplane of the SVM regressor is constructed based on the weight of features in the selected subset [4]. F.Ye et al. proposed a stock prediction model based on LGBM. This model can give suggestions about the decision in the stock market. The criterion of evaluating the model performance is to calculate the cumulative returns under different machine learning models [5]. K.Chen et al. used the LSTM model to forecast the earning of stock in china stock market. Instead of using the daily profit rate, this approach aims to predict the 3-day earning rate. The

result showed the different stock datasets would affect the performance of the prediction [6].

EWMA model is a model used to predict the volatility of stock based on historical data. Axel had found the optimal decay parameter for this model by comparing the result of different parameter values [7]. Clustering algorithms have been applied to construct portfolio formation by researchers. Victoria et al. used three standard clustering algorithms: K-means, K-medoids, and hierarchical clustering on a large financial dataset to determine the portfolio formation [8].

In this study, SVM, LGBM, and LSTM algorithms are used to predict stock price, respectively. The comparison of the performance of these three algorithms is shown in this paper, and the result shows that LGBM has the best performance in this project. Based on the prediction of stock price, stock profit can be calculated. Then the EWMA model is used to predict the volatility of each stock in this project. After predicting the profit and volatility, all stocks are mapped on the coordinate system. Then, using the clustering algorithm to get serval portfolio formations. After analyzing the result, the optimal portfolio formation is the output. This project aims to help any person deal with the stock market.

2. METHODOLOGY

2.1. Data Acquisition

This project studies three datasets, S&P 500 stock data, SP500 Stock Market Index, and CBOE Volatility Index [9, 10, 11]. All of them come from Kaggle. A new dataset is created from these three datasets. The new dataset content includes serval features, including stock daily data, stock market index data, CBOE volatility index value, etc.

2.2. Exploit the Relationship between Different Features and Stock Price

In this experiment, covariance and correlation coefficient are used to analyze the dependency of the stock price on different features. Both these two criteria can reflect the linear correlation between the two variables X and Y.

The covariance is represented as Cov and it can be calculated as follows:

$$Cov(X,Y) = E[X * Y] - E[X] * E[Y] \quad (1)$$

where *E[X]*, *E[Y]* represents the expectation of X and Y, respectively.

The correlation coefficient is recorded as *r* and the equation of calculating the correlation coefficient is as follows:

$$r(X,Y) = \frac{cov(X,Y)}{\sqrt{Var[X] * Var[Y]}} \quad (2)$$

where *Var[X]*, *Var[Y]* represents the variance of X and Y, respectively.

2.3. Stock Price Prediction

In this experiment, three models are used to predict the future price of each studied stock. These three models are SVM, LGBM, and LSTM. The daily profit is the difference between the close price of adjacent two days. The equation for calculating daily profit is as follows:

$$profit = Today\ Price - Previous\ Day\ Price \quad (3)$$

SVM algorithm is a supervised learning method based on the statistical learning theory, which can deal with regression and classification tasks. In classification tasks, the data points are plotted on the n-dimensional space. Then the algorithm will find a hyperplane that lets the distance between the nearest data points (called support vectors) and the decision boundary maximum in the model training process. In regression tasks, the aim is to let as many instances as possible lie on the region, the width of the region can be artificially set. SVM not only can deal with the linear classification tasks and regression tasks but also can solve non-linear tasks. The kernel method can be used in non-linear tasks to avoid consuming too many computational resources. The kernel method studied in this project is called the Gaussian RBF kernel.

LGBM is a gradient boosting framework that uses tree-based learning algorithms. The main idea of boosting is training weak models iteratively to get the optimal model. Different from XGBoost, LGBM uses the histogram algorithm to decrease memory usage. When it handles the largescale data, it uses the gradient-based one-side sampling method to do the sampling. In addition, it uses the exclusive feature bundling algorithm to deal with sparse data. The application of these methods leads LGBM to have the advantages of faster training speed, lower memory consumption, and better accuracy.

LSTM is a kind of recurrent neural network (RNN), which introduces the memory cell, a unit of computation. LSTM is constructed by a single input layer, multiple LSTM layers, a dense layer, and a single output layer. By the unique structure, LSTM is suitable to deal with important events with a long delay in the time series. It can effectively associate memories and input remote in time by the memory cells. Therefore, it performs well on the time series data.

In this project, the performance of different models is measured by the criterion called root mean squared error (RMSE). RMSE gives an idea of how much error the system typically makes in its predictions. The smaller the value of MSE is, the closer is the observed value to the predicted value.

The equation of calculating RMSE is as follows:

$$RMSE(X,h) = \sqrt{\frac{1}{m} * \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2} \quad (4)$$

In Equation (4), m is the number of instances in the dataset. $x^{(i)}$ is a vector of the feature values of the i^{th} instance in the dataset, and $y^{(i)}$ is the corresponding label. X is a matrix containing all the feature values of all instances and h is the prediction function.

2.4. Stock Volatility Prediction

The volatility of a stock represents that whether it can create a consistent return for the investors or not. Therefore, volatility is an important indicator to determine a stock is worthy of investment or not.

The EWMA model is a kind of volatility estimator, whose main advantage is computation simplicity. Using this model can predict the future volatility of a stock based on its historical yield rate. First, the continuously compounded log-return of stock is as follows:

$$\mu_n = ln\left(\frac{S_i - S_{i-1}}{S_{i-1}}\right) \quad (5)$$

where S_i is the close value of this stock on day i. The equation of the EWMA model is as follows:

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 - (1 - \lambda)\mu_{n-1}^2 \quad (6)$$

where $0 < \lambda < 1$. This model uses all historical data to predict the volatility. If λ moves away from 1, this model assigns high wights to the recent observation than the past data.

2.5. Construct Portfolio Formations

After getting the prediction of price, the profit of the stock can be predicted. Based on the prediction of profit and volatility of each stock, the stock can be mapped on the coordinate whose x-axis is profit and the y-axis is volatility. By using the clustering algorithms, the stocks are divided into serval groups. The optimal portfolio formation is the one that has high profit and low volatility.

The spectral clustering algorithm is a technique based on graph theory. The main idea is to treat all data as points in space. The weight of an edge between every two points is related to the distance between them. The weight of the edge will be higher if the two points are closer while the value will be lower if the two points are further. Based on the weight of edges, the algorithm clusters the data points into serval groups. The objective of the algorithm is to let the sum of the weight of the edge between different clusters be as low as possible, and the sum of the weight of the edge within the cluster be as high as possible.

2.6. Proposed Model

The idea implemented in this paper is to use the machine learning algorithm to predict the price and use the EWMA model to predict the volatility of each stock. Based on the price prediction, the profit can be predicted. After prediction, the stocks are mapped on the coordinate system. Finally, using a clustering algorithm to construct the portfolio formation. The flow chart of the proposed model is shown in Figure 1.

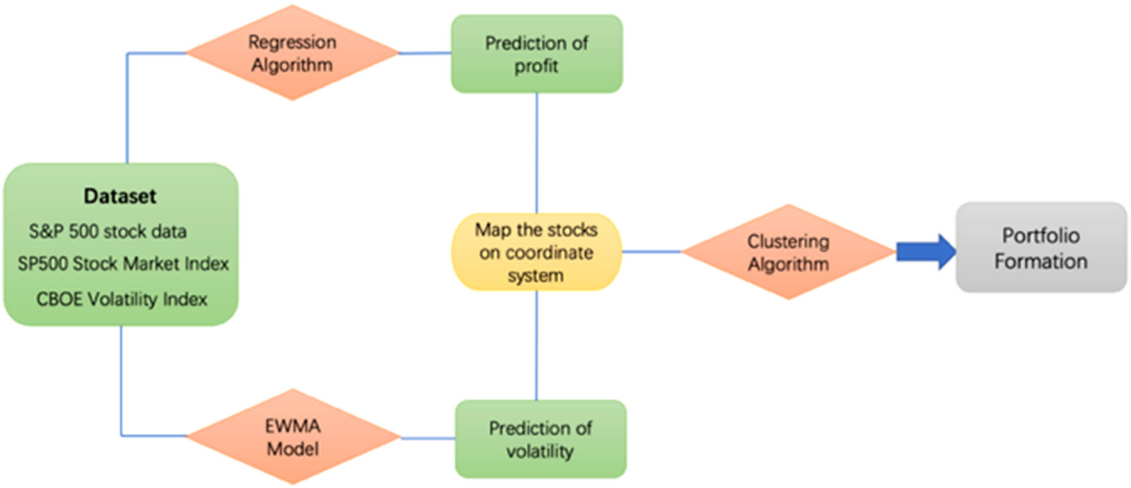

Figure 1 Flow Chart of Proposed Model

3. RESULTS AND DISCUSSION

3.1. Prediction of Stock Price

The stock named "JEC" is used as an example in this subsection to show the prediction of stock price and volatility.

3.1.1. Dependency of Stock Price on the features

Table 1 shows the result of the correlation coefficient and covariance of stock profit and other features.

Table 1. Dependency of Stock Price on the features

Feature Name	Correlation	Covariance
3 Days Price Average	0.997	7.860
Previous Day Price	0.995	7.844
Index Close	0.166	1.310
Gap between Open & Close	0.072	0.565
Stock Amplitude	-0.222	-1.750
Index Volume	-0.291	-2.298
Stock Volume	-0.296	-2.333
Index Amplitude	-0.375	-2.960
VIX Value	-0.528	-4.160

3.1.2. Performance of Different Algorithms

The above-mentioned three machine learning algorithms (SVM, LGBM, LSTM) are trained on this dataset, respectively. Cross-validation is used on all models to find the fit parameters to raise the performance of algorithms. The error of all machine learning algorithms on the training set and test set are compared in Table 2.

Table 2 Evaluation of Different Algorithm

Algorithm	Training Error	Test Error
SVM	1.803	1.923
LGBM	0.799	0.904
LSTM	1.320	1.514

It is evident from Table 2 that the LGBM has the lowest error on the test set. This result represents that the LGBM performs best among these algorithms on this dataset. Figure 2 shows the true price and the price predicted by the LGBM of the stock.

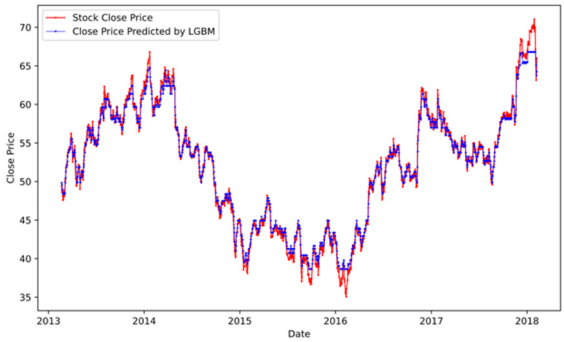

Figure 2 True Price and Predicted Price of Stock

3.2. Prediction of Stock Volatility

Apply the EWMA model to predict the stock volatility of each day. In this project, the decay parameter λ is set as 0.94. Figure 3 shows the volatility of the stock.

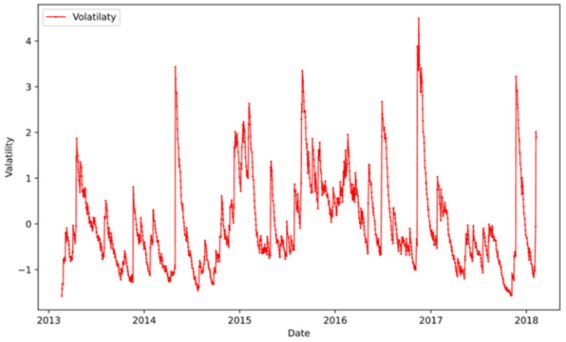

Figure 3 Volatility of Stock

3.3. Construct Portfolio Formation

In this project, the optimal portfolio formation is constructed based on the stocks named "AAPL", "AXP", "BDX", "OMC", "WMT", "ECL", "GGP", "ZTS" and "HOLX". According to the prediction of price, the prediction of profit can be calculated. After getting the prediction of profit and volatility, the stocks are mapped on the coordinate system. Then, using the spectral clustering algorithm to group the similar instance to the same cluster. Figure 4 shows the mapping result of stocks and different portfolio formations.

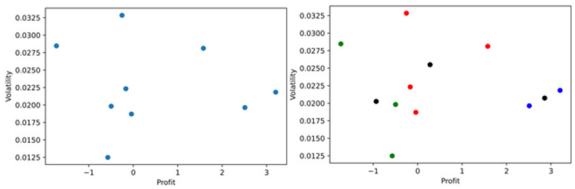

Figure 4 Stocks Mapping and Portfolio Formation

In this project, the number of clusters is chosen as 3. It means that all the stocks are divided into three groups. After applying the spectral clustering algorithm, the stocks are divided into the blue group, green group, and red group. The black points are the center points of each cluster, which represent the characteristics of the corresponding group. Each group represents a kind of portfolio formation.

3.4 Result Analysis

Table 3 shows the coordinate of the center point and the stocks component of each portfolio formation.

The portfolio formation with negative profit is not the one that the investors want to invest in. Therefore, the green cluster is not an ideal portfolio formation. Comparing the red cluster with the blue cluster, the blue one has a higher profit and lower volatility. Thus, the blue cluster which contains "WMT" and "ZTS" is the optimal one among all portfolio formations and it is also the output of the proposed model.

Table 3 Profit, volatility and component

Cluster	Profit	Volatility	Component
Green	-0.9332	0.0203	'AXP' 'OMC' 'ECL'
Red	0.2787	0.0255	'AAPL' 'BDX' 'GGP' 'HOLX'
Blue	2.8580	0.0207	'WMT' 'ZTS'

4. CONCLUSION

This project aims to use machine learning algorithms and the EWMA model to construct portfolio formation based on historical data. Firstly, three different machine learning algorithms called SVM, LGBM, and LSTM are used to predict the tendency of the stocks, respectively. The result shows that the LGBM algorithm outperforms the SVM algorithm and the LSTM algorithm. The LGBM algorithm achieves 0.904 of RMSE, which indicates a quite accurate performance. Based on the prediction of stock price, the profit of stocks can be predicted. Secondly, the EWMA model is used to predict the volatility of the stocks. Thirdly, the stocks are mapped on the coordinate system according to the prediction of profit and volatility. Then, the clustering algorithm called spectral clustering is applied to the mapped data points to cluster the stocks, each cluster represents a kind of portfolio formation. Finally, the optimal portfolio formation is chosen to be the output of the proposed model.

The result shows that LGBM has an excellent performance on the stock market data. Different from most other Gradient Boosting Decision Tree (GBDT) models, the decision trees in LGBM use the leaf-wise growth strategy to split nodes. The algorithm will search all the current leaf nodes to find the one with the greatest splitting gain and then split it in the model training process. Compared with using the level-wise growth strategy, in the case of the same times of splitting nodes, using the leaf-wise growth strategy can reduce more errors and get better accuracy.

The above result concludes that the machine learning algorithms can be used to construct portfolio formation in the stock market. Based on the prediction of profit and volatility, the spectral clustering algorithm can cluster the stocks to give a good portfolio formation. Investing in a suitable portfolio formation can decrease the volatility of the yield of the investment. This method can help the investor who has not enough experience do the correct decision in the stock market to avoid the risk of investment. This project can be further extended to include more clustering algorithms to get more different portfolio formations and compare their results. In addition, the project can be also further extended to study how to apply machine learning algorithms on other types of financial data.

REFERENCES

[1] Das, S.P., & Padhy, S. (2012). Support Vector Machines for Prediction of Futures Prices in Indian Stock Market. *International Journal of Computer Applications, 41*, 22-26.

[2] Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. *International Journal in Computer Simulation*, *5*(1), 55–72.

[3] K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, SarmisthaPadhi, Saurav Sanyal "Stock Market Prediction Using Machine Learning Algorithms" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-4, April 2019.

[4] Lin, Y., Guo, H., & Hu, J. (2013). An SVM-based approach for stock market trend prediction. Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE.

[5] Ye, F., Wang, J., Li, Z., Jihan, Z., & Yang, C. (2021). Jane Street Stock prediction model based on LightGBM. In 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP).

[6] Chen, K., Zhou, Y., & F Dai. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE International Conference on Big Data (pp.2823-2824). IEEE.

[7] Axel A. Araneda (2021)," Asset volatility forecasting: The optimal decay parameter in the EWMA model" arXiv:2105.14382.

[8] Victoria Lemieux, Payam S. Rahmdel, Rick Walker, B. L. William Wong, and Mark Flood. 2014. Clustering Techniques And their Effect on Portfolio Formation and Risk Analysis. In Proceedings of the International Workshop on Data Science for Macro-Modeling (DSMM'14).

[9] S&P 500 stock data https://www.kaggle.com/camnugent/sandp500

[10] SP500 Stock Market Index

https://www.kaggle.com/elvinagammed/sp500-stock-market-index

[11] CBOE Volatility Index

https://www.kaggle.com/jonathanbesomi/cboe-volatility-index-vix