Application of Business Analysis in Stock Market Forecasting - Machine Learning in Stock Market Practice

Xinyi Zhang[1, *]

[1]Beijing Normal University - Hong Kong Baptist University United International College, China

*Corresponding author: q030026207@mail.uic.edu.hk

Abstract. With the development of global economic integration, the stock market occupies an important position in the global economy. Accurately predict the stock market is of important social and economic value, the stock market has huge amounts of data sources, such data features to capture the hidden rule of the stock market, and associated accurately predict proposed the new challenge, with the vigorous development of the data mining technology and the data sample is unceasingly rich, The value of data is more fully recognized and more widely concerned, and business data analysis has been gradually applied to stock market forecasting, For example: Machine learning, data mining. This paper studies the business analysis in the era of big data, so that the stock market can get higher economic benefits. For the stock market, machine learning, statistical reasoning and other methods can be used as theoretical research, but the practical application needs to be prepared and improved according to the real market environment.

Keywords: Business analysis; Stock market forecast; data processing and application; machine learning; Data visualization.

1. Introduction

The current stock market can be analyzed from three aspects. The first is from the economic perspective, as the Federal Reserve has raised interest rates and shrunk its balance sheet, capital in the market has shrunk. The market will further sell off stocks with high valuations and high debt and some investments that do not generate cash flow and value, so there is more focus on quality stocks and companies that can provide good cash flow. [1] The second is the industry perspective, which depends on several key points. The first is the relationship between supply and demand. When demand outstrips supply asp goes up and the company makes more money. Of course, the opposite company will earn a lot less, so we have to pay attention to supply and demand. The third is the perspective of individual stocks, which is technical analysis and fundamental analysis. A company's business model, cash flow, financial position and valuation all require sound business analysis.

At the present stage, business analysis mainly refers to the economic benefit analysis of marketing programs, and judge whether they meet the development goals of enterprises at the present stage at the financial level. When the product concept meets this requirement, it will enter the product research and development stage. The steps of business analysis include the first: verifying the reasonableness and accuracy of data sources. Second, verify the practicality and validity of the survey data. Third, analyze the consumption characteristics of target customers. Fourthly, clarify the company's business scope and products. Fifth, determine reasonable business model, business scope, business strategy and other information.

From the analysis of the business of the era of big data, and a big data technology for high hadoop, it is an open source distributed data, vast amounts of data can be stored in a piece of the server, thereby improving the efficiency and reduces the cost, improve the speed of data processing at the same time, improve the precision of data processing. [2] And an algorithm called machine learning, in machine learning methods, learning supervision is the most impressive achievements so far, want to undertake supervised learning, we need to start from a set of sample data, label each sample with a computer can learn, machine through the input variable and output, autonomous learning algorithm. [3] Computer programs can calculate data during machine learning. In supervised learning, a computer can learn how to calculate an output from an input by itself, given sample data of the program's inputs and outputs. Its ability to extract features from a large set of raw data without relying on prior

knowledge of predictors makes machine learning potentially attractive for stock market prediction at high frequencies. [4] Therefore, stock market can be predicted based on machine learning algorithms.

With the rapid development of financial economics, the efficient market hypothesis in not much support at the same time, has been questioned by some, such as some exist in the stock market "calendar effect", "size effect", the "reversal effect" and so on the abnormal phenomenon, and with the rise and development of cognitive psychology, to understand the price behavior, will study the decision making process, And then we have behavioral finance theory, the efficient market hypothesis which says that markets are efficient and invincible, and behavioral finance theory which says that markets are irrational and invincible. [5] Behavioral finance theory says that if the stock market fails again and again, then the market is not that efficient. The reason why markets are often ineffective is that people are irrational. They tend to choose the one that everyone else thinks is best, and they tend not to choose the one that they think is best. The decision-making process includes data collection, data preprocessing, modeling, model testing, model optimization, predictive validation to predict future stock market prices, as well as analyzing the behavior of investors and shareholders to make predictions based on their psychology. [6] There are a few differences with the so-called efficient market hypothesis. In this study, we use the Knowledge framework of Python financial quantification and follow the steps to build the framework system of Python quantitative investment. Behavioral finance theory tells us that people's psychology is influenced by desire and greed because most people are inclined to speculate or gamble. Therefore, the stock market is irrational in most of the time and it is easy to have extreme stock price fluctuations. Time series is the most common data type in financial quantitative analysis, which records the value of a variable or feature along the time axis. [7] For example, the daily closing price of a stock from 2006 to 2016. One of the important links of quantitative analysis is to analyze and mine based on historical data, trying to explore the change rule trend of something from the dimension of historical data for prediction. Time series analysis theory has been mature, including general statistical analysis, such as stationarity, autocorrelation, statistical modeling and judgment, and time series prediction, including machine learning, deep learning, such as LSTM model. [8] This research is based on the behavioral finance theory in the background of business analysis using Python machine learning through modeling to predict the short-term future of the stock market. [9]

The introduction part of this paper first introduces the background status quo of today's stock market, and then lists the results of others' research, such as the business analysis of efficient market theory. After that, it describes the deficiencies of others' research, the advantages of my own research and specific research ideas. In the Method part of the second part, the model used in this study is firstly introduced, followed by the data sources used in this study and selected characteristic values. The third part is the Result part, which describes the results and analysis of the research. In the last part, I first write the research questions, the results and significance, the shortcomings of my own research and the direction of future research.

2. Method

The pandas library is used to process date data, and the resample function in pandas is used to convert the date samples, such as high and low frequency data. Time series sample conversion is divided into low frequency data to high frequency data, from high frequency data to low frequency data. Market trading is generally high frequency, fundamentals are generally monthly, quarterly, annual and other data, quantitative analysis, often combine fundamental data and market trading data for statistical review analysis, so the conversion of sample data frequency includes: df.resample().

The main purpose of time series analysis is to use the history and current situation of the characteristic variables of things to predict the possible situation in the future. The stationarity of time series means that the basic characteristics remain unchanged, which requires the stock data obtained by us to continue along the current state in the future for a period of time. So in Python we introduce statsmodels and scipy.stats to draw QQ and PP plots. The two plots are very similar, except that PP

plots use accumulative ratios of distributions, while QQ plots use quantiles of distributions to check. The data points are basically on the diagonals. After that, numpy was used to simply simulate the white noise process, and it was concluded that the process was random and fluctuated near 0. In addition, the random walk is the model of time series XT: XT = XT-1 +wt, wt is a discrete white noise sequence, and the random walk is not stable, so the covariance is related to time. If the time series we model is a random walk, then it is unpredictable. Then it can be concluded that the sequence is unstable, and the random walk model can be obtained by transposition: XT-XT-1 =wt. Therefore, the first-order difference of the random walk is equal to the white noise, and the NP. diff function can be used for the time series to see if it is true. Viewing trends in data using python's drawing tools is intuitive, but it's also subjective, and different people may come to different conclusions about the same graph. Therefore, in the previous stage of data extraction and analysis in this study, a more objective statistical method is needed to test the stationarity of time series

In our study, the data of the previous n days were used as the prediction, and the difference between the n+2 and n+1 days was used as the prediction target for training. The samples were randomly divided into two groups, 60% training set and 40% test set. Generally, for supervised learning, model training needs to be based on labeled data, while for stock trend classification, it is actually a little cumbersome, so this study divides into four types of labels. Classify the increase into 0-5%, and label it as category 1 and category 2 respectively if it is greater than 5%. The training set is used to train the classifier, and then the test set is used to verify the model. The four classifiers include k-nearest neighbor, naive Bayes, decision tree and support vector machine. Classify declines as 0-5% and greater than 5% and label them as category 3 and Category 4 respectively. Let's take n to be 60 days. Firstly, data preprocessing is carried out. We need to take the data of a stock in the first 60 days as the input feature, so we need to group and sort the stock names, and then do time series processing according to 60 days. Then, the GBDT model was used as an example for model training, model prediction, and output model performance indicators to see the accuracy of classification results.

GBDT algorithm is a combination of the gradient boosting algorithm and the regression tree algorithm. The idea of gradient boosting algorithm is originally proposed by Friedman which is an integration of multiple weak predictors and a branch of boosting algorithm, that constantly adding new predictors to make up for the shortcomings of known predictors. [10] GBDT-Stacking model uses the GBDT to automatically extract and transform features suitable for prediction and uses Stacking model to predict CTR of user, which improves the performance of baseline effectively.[11]

The purpose of our study is to predict whether it can constitute the above patterns and predict the probability of generating this pattern according to the pattern over a period of time, such as the past period to the present. Since this is not a traditional classification, but a probability predicted by a model, the practical usefulness of this is relatively high. The prediction part of the model also uses the stochastic forest model, which is suitable for algorithms such as decision trees or logistic regression in the field of quantitative trading that requires high interpretability. Random forest is a machine learning algorithm proposed by LEO BREIMAN in 2001 by combining Bagging ensemble learning theory with random subspace method. On the basis of constructing Bagging ensemble based on decision tree learning, random forest further introduces the characteristics of human random attribute in the training process of decision tree. Firstly, x samples are extracted from the original training set by autonomous sampling method, and the sample size of each sample is the same as that of the original training set. Secondly, X decision tree models are established for x samples to obtain X regression results. Finally, x decision tree results are combined by taking average values

In addition, our model can also be optimized, the above model is called using the function of SkLearn, the parameters of the model are default, we can try to use different parameters to improve the performance of the model, we can use the function GridSearchCV provided by SkLearn to find the parameters of the grid, therefore, we can find a predictive model. Finally, we can use this formula to evaluate the effectiveness of the prediction model: goodness of fit R^2

$$R^2 = 1-\frac{\sum_{m=1}^{M}(y_m-\widehat{y_m})^2}{\sum_{m=1}^{M}(y_m-\overline{y_m})^2}$$

Where $\overline{y_m}$ is the mean value of the true value being predict. R^2 is an effective index to evaluate the linear regression model, of which the range is [0,1]. The closer the value of R^2 is to 1, the better the prediction effect is.

Basic stock data can be obtained from the API data interface provided by Tushare, Sina Finance and Flush, and the data interval is daily trading data from 2020 to 2021 including open price, close price, high price, low price and other data. We can save historical data and build data sets.

3. Result

For the stock market with high noise and high unpredictability, various exploration methods including machine learning, statistical reasoning and other technologies can be used as theoretical research, but the practical application needs to be more fully prepared and improved according to the real market environment. The result of this study is that we can eventually find a model with certain predictive performance. We can input the basic data of all stocks in the last 60 days to get a model and predict the return category for stock screening. We can see that a major factor for the performance of the model is data. We should try our best to make the data large enough and have enough data features, so as to improve the accuracy of prediction, and then divide the data into prediction set and training set, and the data of prediction set and training set should not overlap. It can be seen from the results of this study that machine learning can predict the stock market, but further decisions need to be made according to changes in people's psychology, policies and market supply and demand relations. Due to the complexity of the financial market, the construction and fitting of the model need to combine the actual data, market environment and other factors, and also need to take into account the subjective emotions of investors and other influences

Because the theoretical innovation involved in this study is not very much, but its engineering implementation is relatively sufficient, and the system can also be combined with various machine learning models for prediction, so it is very extensible. In addition, it is also valuable to identify more trend patterns in combination with other methods of feature selection to improve prediction accuracy. Market information is changeable, often reflected in the relationship between volume and price. Therefore, using quantitative analysis of technical analysis can improve the success rate of stock selection. But all of the technical analysis based on the application of historical induction and large data, but history is always not completely similar, the trend of future development is different also, so technical analysis has certain limitations in the, with high probability choose stocks can determine stock prices form rise in the future, but not necessarily way as we wish to rise, may be in the middle of the instability makes us out

4. Conclusion

This paper examines the problem of stock market forecasting based on machine learning, and by building a model in Python, it is possible to predict the decline and increase of stock prices in the short term in the future. However,the classification attributes selected in this study only come from traditional technical analysis indicators. Although they are highly representative, they are relatively single and only reflect part of the stock information. There is room for improvement in the selection and combination of attributes. The trainer in this study added the characteristics of stock decline and rise, and screened the database records before the training data, but the prediction effect was not obtained in the sample test. In actual investment, the model needs to be further modified. All kinds of speculators can combine their own investment needs and characteristics to design trainers to meet their own requirements. The selected attributes can be macroeconomic indicators, industry development trends, and financial indicators in addition to technical indicators. Our research can prove that machine learning has a certain role in predicting the stock market, and can also provide

reference suggestions for people who buy stocks, but the amount of data used in our research is not large enough, and the amount of data can be increased later, this study uses day-level data, we can also find hour-level data with shorter time levels, minute-level data.

References

[1] Y. Fan and M. Stevenson, "A review of supply chain risk management: definition, theory, and research agenda," International Journal of Physical Distribution & Logs Management, vol. 48, no. 3, pp. 205–230, 2018.

[2] L. Gao and J. Xiao, "Big data credit report in credit risk management of consumer finance," Wireless Communications and Mobile Computing, vol. 2021, no. 4, Article ID 4811086, 7 pages, 2021.

[3] L. BREIMAN. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[4] Eunsuk Chong and Chulwoo Han and Frank C. Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies[J]. Expert Systems with Applications, 2017, 83: 187-205.

[5] A. Titan The efficient market hypothesis: Review of specialized literature and empirical research Procedia Economics and Finance, 32 (2015), pp. 442-449

[6] B. Malkiel the efficient market hypothesis and its critics. The Journal of Economic Perspectives, 17 (1) (2003), pp. 59-82

[7] P. D. Trung, "Research risk factors and management competence of Vietnam commercial banks from 2006-2020," Environmental Science and Engineering: B, vol. 5, no. 8, p. 5, 2016

[8] T.B. Trafalis, H. Ince. Support vector machine for regression and application to financial Forecasting [C], Neural Networks, 2000. IJCMM 2000, Proceedings of the Ieeeinns-enns International Joint Conferenece on IEEE, 2000.

[9] L.J, Cao, F, Tay. Support vector machine with adaptive parameters in financial time series forecasting [J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1506-1518.

[10] J. Friedman Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 2001. 29(5), pp. 1189-1232

[11] He X, Pan W, Cheng H. Research on hit rate prediction model based on ensemble learning method. Computer engineering and Science. 2019. 41(12). pp. 2278-2284.