Machine learning models for mutual funds assessment in fund selection

Hanlei Huang

27th, 1st lane, Chengzhong Road, Huaiji County, Zhaoqingcity, Guangdongprovince, China

3284672482@qq.com

Keywords: mutual funds, machine learning, select, return prediction, model comparisons

Abstract. Mutual funds represent a considerable portfolio of financial securities and provide investors an alternative way in investment. It is profound and helpful because of a lower risk than equity and derivatives and a highly qualified management. The optimal mutual fund is chosen in terms of the maximum return with a certain risk or the minimum risk with a certain return for risk-aversion investors. However, return prediction of models like dividend discount model and Residual income model is complicated due to different obstacles such as cashflow uncertainty or financial report manipulation. To solve it problem, this study will introduce some basic machine learning models like linear regression, logistic regression and decision tree. The implementation of these three algorithms involves 'train' and 'test' the data used and measures an expected return and select a suitable mutual fund for different types of investors. In addition, some model comparisons in these related models will also exhibit their performances, accuracy and complexity.

1 Introduction

With an increasing of economy and household income, a considerable proportion of incomes of families flow to financial markets, where can provide a larger number of financial tools to invest. In all financial securities provided in financial markets, investing mutual funds is considered to be a valid and suggested choice because they are managed by institute investors who have a larger number of high qualified fund managers and more information sources. Mutual funds are portfolios of assets (like stocks and bonds etc.), with lower volatility than equity but higher volatility than bonds investment, which are considered to have a balanced return and diversification. With the advantages of cheapness, specialized management and less time consumption, individual investors with less expertise are prone to invest mutual funds for passive investment and risk management.

Return prediction, together with risk, is crucial in selecting mutual funds. For traditional models like dividend discount model, Residual income model, Free cash flow valuation and Price Multiples in some cases are not effective because of their own assumptions. For example, dividend discount model is invalid when the company does not have a clear dividend policy. Residual income model is invalid when the financial report in a company is manipulated. Free cash flow valuation is invalid when the Free cash flow for equity (FCFE) computed is less than zero. Price Multiplesis invalid because of the trailing P/E ratio.

Machine learning approaches are able to generate an artificial and intelligent relationship between dependent variables and independent variables. By training and testing some useful financial data of mutual funds in training set and validation set, an expected return or decision will come out and a suitable mutual fund will be selected to meet the demand of different types of investors.

2 Model variables and Data sources

2.1 Input variables

A set of financial indicators and the proportion of invested sectors are needed to be considered as independent variables. For financial indicator, some related ratios like sharp ratio and P/E ratio and related indicators such as net asset and fund yield should be considered.

2.2 output variables

Output variables must obey the properties of linear regression, logistics regression and decision tree respectively. Therefore, for linear regression, three-years mutual fund return needs to be considered due to the assumption that investors pay more attention on the return within three years. This phenomenon is obvious because automatic investment plan (AIP) is recommended by financial analysts to invest a fixed amount of money regularly (such as monthly or weekly). Usually AIP is approximated three-year plan so that investors are more likely to withdraw fund within three years. Besides this, for logistics regression and decision tree, an indicator variable (need_invested) is considered. It is aclassifier that judges whether the 3-years mutual fund return is larger than 9.6 (%) *. If the 3-years mutual fund return of a certain mutual fund is larger than 9.6, need_invested will return 1 unless it will return as 0.

In this case, 9.6(%) is regarded as the yield of 10-year Treasures bond in October 2018. It is often regarded as a risk-free rate that investors will not suffer from risk when they invest. And investing 10-year Treasures bond in United States is safe because no investors will believe governments of United Sates will default in the future.In the website of economy of United States, the yields of 10-year Treasures bond in October and November 2018 are 3.085 and 3.134(%) respectively. Therefore, the average of them can be used to compute the approximated risk-free rate:

$$r_f = \frac{(r_{f10}+r_{f11})}{2} = 3.1095(\%) \tag{1}$$

Furthermore, computing the three-year return is necessary due to match the term of 3-years mutual fund return.

$$r_f^* = (1 + r_f)^3 - 1 = 9.621576(\%) \approx 9.6(\%) \tag{2}$$

2.3 Data sources

There are 25,308 mutual funds with general aspects in 2018 from Kaggle. However, in order to eliminate the negative effects brought by N.A. values and some missing values, the objects of mutual funds with these kinds of values should be removed in the process of data cleaning, causing 3,845mutual funds left. Besides this, standardization in input variables is essential before fitting the model.

$$x_{standardization} = \frac{x_i - x_{mimimum}}{x_{maximum} - x_{mimimum}} \tag{3}$$

In the process of constructing efficient frontier, the time series data of close prices in Dow Jones Industrial Average and Nasdaq Composite are collected.

3 Process of mutual funds selection

The process showed below demonstrates how can financial analysts help investors select a measurable and suitable mutual fund.

3.1 Significant process

The process consists of five steps. At the beginning of the process, the data frame is divided into two parts: training set and validation set respectively. This means that only the data in training set can be ‘trained’ and the data in validation set are used to ‘test’ the model to observe the accuracy. After training the data by the models showed above, expected predictions are computed by using the input data in the validation set. More precisely, while a linear regression can be used to predict an exactly expected returns of mutual funds, logistics regression and decision tree only return an approximated classification of mutual funds which return 1. Furthermore, in the step of ‘testdata’, it is worth that implementing sensitivity analysis in linear regression, in which financial analysts or mutual funds managers are able to observe how sensitive for each input variables to expected return, encouraging to make new strategies in asset allocation and securities selection. Compared withimplementing sensitivity analysis in linear regression, the probability of accuracy in logistics

regression and decision tree are computed by matching how many values of 1 between predictions with validation set. The most crucial and significant step is mutual fund selection. Since the numerical results are obtained in the former steps, for each types of investors (estimated by IPS: Investment Policy Statement), different methods are used. For risk-aversion investors, Markowitz efficient frontier1are implemented. The optimal mutual fund is the one which is the closest to the tangent point on efficient frontier. For risk-neural investors, the largest return of prediction of mutual fund is chosen and for risk-seeking investors, the largest risk of prediction (estimated by standard deviation) of mutual fund is selected.

$$distance = \sqrt{(\sigma_i - \sigma_{tangent})^2 + (r_i - r_{tangent})^2} \quad (4)$$

$$optimal\ distance = \min(distance) \quad (5)$$

3.2 key procedure in efficient frontier construction

One of the most import concept is Markowitz efficient frontier. Itis the set of optimal portfolios that offer the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. In the process of constructing Markowitz efficient frontier, obtaining the time series returns of Dow Jones Industrial Average and Nasdaq Composite is the key steps in order to implement the package of 'IntroCompFinR' in R.

The return of time series:

$$\log\left(\frac{X_{t+1}}{X_t}\right) = \log(X_{t+1}) - \log(X_t)2 \quad (6)$$

Where X is the time series data of close prices in Dow Jones Industrial Average and Nasdaq Composite

This is because discretely compounding returns is approximately equal to Continuously compounding returns when the size of data frame is large.

$$R_t = \frac{X_t - X_{t-1}}{X_{t-1}} = \frac{X_t}{X_{t-1}} - 1 \quad (7)$$

$$r_t = \ln\left(\frac{X_t}{X_{t-1}}\right) = \ln(X_t) - \ln(X_{t-1}) = p_t - p_{t-1} \quad (8)$$

$$X_t = X_{t-1}(1 + R_t) = X_0(1 + R_1) \dots (1 + R_{t-1})(1 + R_t) \quad (9)$$

$$= X_{t-1}e^{r_t} = X_0 e^{r_t + r_{t-1} + \cdots + r_1} = X_0 e^{\sum_{i=1}^{t} r_i} \quad (10)$$

Since $\ln(1 + x) \approx x\ for\ x\ small, then$:

$$r_t = \ln\left(\frac{X_t}{X_{t-1}}\right) = \ln(1 + R_t) \approx R_t \quad (11)$$

4 Results of the mutual fund selection

4.1 Linear regression

[1]Dr. Graeme West.An introduction to Modern Portfolio Theory: Markowitz, CAP-M, APT and Black-Litterman. CAM Dept, University of the Witwatersrand. August 14, 2005: 1-15

[2]Ruey S. Tsay: An Introduction to Analysis of Financial Data with R. John Wiley & Sons, Incorporated 2012-10-29 :2-8

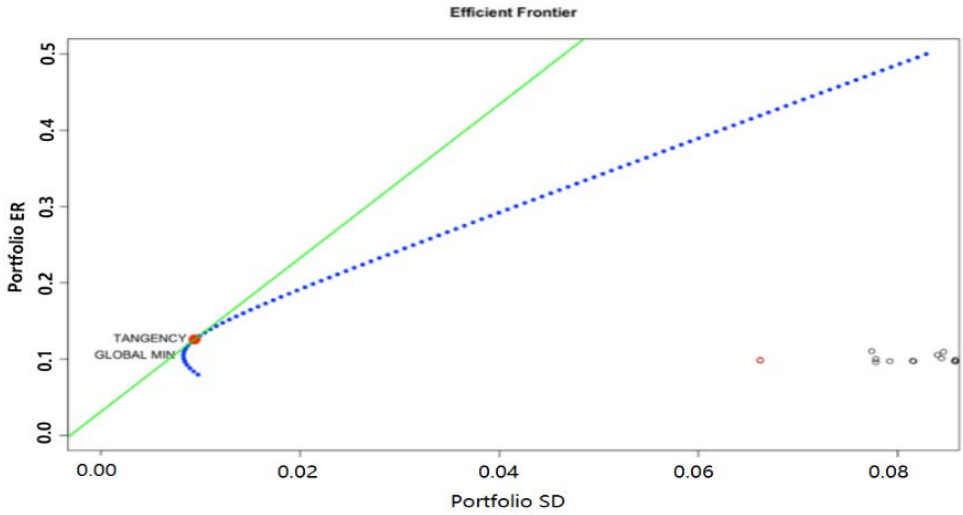

Fig 1. Results of the mutual fund selection

Where the red point is the mutual fund chosen.

4.2 forrisk-aversion investors:

Tab 1. Forrisk-aversion investors

Mutual fund name	VWINX
Fund return 3 year	9.870595(%)
Fund standard deviation	6.62(%)
distance	0.06292521

4.2.1 forrisk-neural investors:

Tab 2. Forrisk-neural investors

Mutual fund name	VALIX
Fund return 3 year	13.49593(%)
Fund standard deviation	12.64(%)
distance	0.1173571

4.2.2 forrisk-seeking investors:

Tab 3. Forrisk-seeking investors

Mutual fund name	PSCSX
Fund return 3 year	10.04535(%)
Fund standard deviation	16.15(%)
distance	0.1541941

4.3 Logistics regression

4.3.1 forrisk-aversion investors:

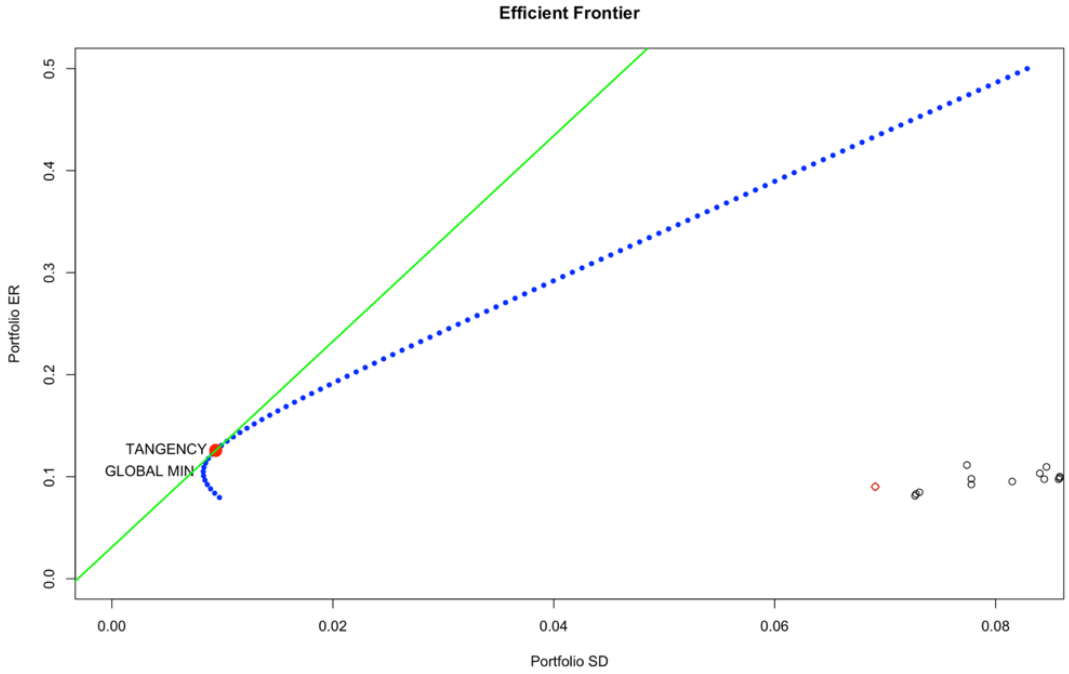

Fig 2. Logistics regression

Tab 4. Forrisk-aversion investors

Mutual fund name	PTOAX
Fund return 3 year	9.02(%) small than risk-free rate
Fund standard deviation	6.91(%)
distance	0.106950195

4.3.2 forrisk-neural investors:

Tab 5. Forrisk-neural investors

Mutual fund name	PSLDX
Fund return 3 year	16.95(%)
Fund standard deviation	13.27(%)
distance	0.1308169

4.3.3 forrisk-seeking investors:

Tab 6. Forrisk-seeking investors

Mutual fund name	PSCSX
Fund return 3 year	15.2(%) same company but different return due to the model difference
Fund standard deviation	16.15(%)
distance	0.1541941

4.4 Logistics regression

4.4.1 forrisk-aversion investors:

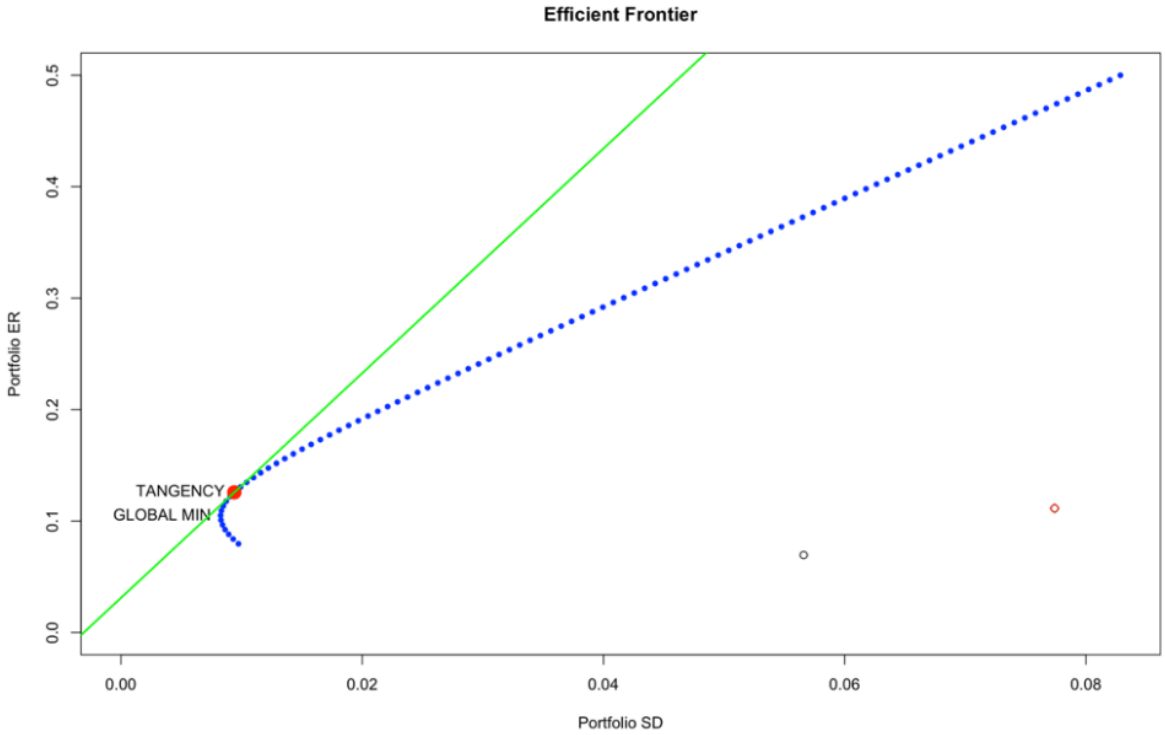

Fig 3. Logistics regression

Tab 7. Forrisk-aversion investors

Mutual fund name	VALIX
Fund return 3 year	11.14(%)
Fund standard deviation	7.74(%)
distance	0.06950424

4.4.2 forrisk-neural investors:

Tab 8. Forrisk-neural investors

Mutual fund name	PSPRX
Fund return 3 year	16.95(%)
Fund standard deviation	13.27(%)
distance	0.1308169

4.4.3 forrisk-seeking investors:

Tab 9. Forrisk-seeking investors

Mutual fund name	PSCSX
Fund return 3 year	15.2(%) same company but different return due to the model difference
Fund standard deviation	16.15(%)
distance	0.1541941

It is obvious that different mutual funds will be chosen when using different models. For risk-aversion investors, it is not recommended to use logistics regression due to the lower expected return than risk-free rate. However, the expected return chosen by linear regression is smaller than that of decision tree but has lower risk. For risk-aversion investors, it is suggested to use logistics regression and decision tree model because they predict a higher expected return. For risk-seeking investors, the same mutual fund is chosen, though different expected returns are predicted.

5 Model comparison

5.1 Linear regression

Linear regression reveals that most of financial variables contribute well to 3-yearsfund return except for the days from the issue date and the type of mutual funds. It makes sense because the economy of United States in the period was booming, meaning that investors pay more attention to those with high-quality financial statements and better rating by financial analysis, no matter when the mutual funds were issued and which types of the mutual funds were.

In addition, the sectors of utilities, energy and communication services has stronger negative relationship to 3-years fund return, while the more percentage of fund invested in the sector of healthcare, the more return will earn, showing a positive relationship between healthcare and 3-years fund return.

Moreover, the data predicted is approximately similar with the data of validation, meaning that once investors obtain the values of independent variables, it is a proper way to input all the related value into the model and return an approximated three-year return of the mutual fund they want to invest.

Last but not least, standard deviation has the strongest negative relationship, followed by fundyield and beta value, while Morningstarrating and sharpratiohave the strongest and the second strongest positive relationship, which can provide individual investors an effective way to select fund mutual.

5.2 Logistics regression

In this model, the trend of accuracy increases at first and peaks at 97.27%, though it will decrease in the end. If investors believe that the mutual funds is regarded as a ‘need invested’ in reality when the probability of need_invested (y=1) is less than 50%, the accuracy of this model is about 90%, while the model will have approximate 95% accuracy when investors belief they can regard mutual funds chosen as needed invested securities when the probability of need_invested (y=1) is in the interval of [90%,100%].

5.3 Decision tree

In this model, most of mutual fund with a smaller YTM (the yield to maturity is less than 0.49) will return to 0 (not recommended mutual funds), while those whose have larger YTM, fund yield, net assets and market capitalization will return 1(recommended mutual funds).

In addition, for those with larger proportion position in basic materials and shorter time to now also return 1, though these two variables are not necessary in linear regression and the variable of time to now is not necessary in both of linear regression and logistics regression.

Lastly, compared with logistics regression, the graph of decision tree has a more stable trend of accuracy and higher approximated accuracy which is within 90%.

6 Conclusion

This research has provided alternative methods of machine learning for investors to select a suitable mutual fund. It has also avoided the limitations of other traditional models when predicting expected returns.

The main obstacle in predicting expected returns is risk uncertainty. To deal with this problem, machine learning models has proposed because they pay more attention to the trends of expected returns. The input variables and financial index are usually available from the balance sheets of mutual funds companies or websites.

From IPS (Investment Policy Statement), the types of investors will be assessed. Therefore, Different expected results predicted by modelswill be implemented according to the types of investors. For most of investors (risk-aversion investors), linear regression model and decision tree are recommended because of predicting a higher return and decision tree has a more stable trend of accuracy.

Acknowledgments

I would like to express the deepest appreciation to the professor Allen, who has introduced an interesting concept: machine learning to me. He continually opens my mind in implementing machine learning methods in financial areas, which inspires me a lot. Without his guidance, this research would not have been possible. In the future, I would like to continue this area and implement more machine learning methodsin finance.

Reference

[1] Alpaydin, Ethem. Introduction to Machine Learning. Third edition. Cambridge, MA: The MIT Press. 2014.1-26

[2] R. Duda, P. Hart and D. Stork. Pattern Classification, Second edition. Wiley, 2001. 1-5

[3] Carmona, R. (2014, second edition) Statistical Analysis of Financial Data in R. Springer. Pages 199-276

[4] Ivanovski, Zoran, NadicaIvanovska, and Zoran Narasanov. 2015. Application of Dividend Discount Model Valuation at Macedonian Stock-Exchange. UTMS Journal of Economics 6 (1): 147–154.

[5] Stephen H.L. Yip. Dividend discount, discounted cash flow and residual income model mispricing ratios, and stock return prediction: value line evidence. The University of Queensland: 21-31

[6] Marsland, S. Machine Learning: An Algorithmic Perspective (2nd edition). 2015: 249-267

[7] Dr. Graeme West.An introduction to Modern Portfolio Theory: Markowitz, CAP-M, APT and Black-Litterman. CAM Dept, University of the Witwatersrand. August 14, 2005: 1-15

[8] Ruey S. Tsay: An Introduction to Analysis of Financial Data with R. John Wiley & Sons, Incorporated2012-10-29 :2-8

[9] CHAO-YING JOANNE PENGKUK LIDA LEEGARY M. INGERSOLL: An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research 96(1):3-14 · September 2002: 3-10

[10] Eric Zivot: Computing Efficient Portfolios in R. November 11, 2008