Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 3, No. 3, 2022

A Review of Risks Associated with Machine Learning in Application to Quantitative Investment

Siyin Shen*

Shanghai Australia International School, Shanghai 200000, China
*Corresponding Author's Email: xjf9699@126.com

Abstract: Based on inspirations and ideas from relevant literatures, this paper evaluated the risks associated with using random forest, XG Boost and logistic regression for quantitative investment from the perspective of its accuracy, adaptability, efficiency, simplicity and interpretability. Overall, the random forest and the XG Boost contains better accuracy and have higher adaptability than the logistic regression as they are susceptible to different data types. The XG Boost have the fastest processing speed which gives it higher efficiency over the other two, however it is also the most difficult to implement as it is written in C++. All three algorithms are relatively clear and easy to understand. This work hopes to assist investors in their decision making on which model to use.

Keywords: Logistic regression, Random forest, XG Boost, Risk evaluation.

1. Introduction

Quantitative investment is associated with using financial transactions through quantifying methods and computer programs. Through mining large amounts of data, machine learning methods can obtain technical indicators that are often ignored by traditional methods.

Through literature research method, this paper was able to establish an evaluation on the risks associated with different machine learning algorithms in application to quantitative investment. The key algorithms this paper analysed include the random forest, XG Boost, and logistic regression. Although many past researches conducted a comparison between the different algorithms in the fields of quantitative investment, this paper offers a new approach as it breaks down the factors that would affect the risk in using the model.

The paper would evaluate each of the following components; accuracy, adaptability, efficiency, simplicity and interpretability which would all play influence on the model failure risks and strategy application risks of quantitative investment. Model failure risk refers to how changes in the market could cause the strategies applied to become invalid. Strategy application risk refers to operational errors or usage of unsuitable parameters which can lead to risks in the transaction.

Increased accuracy would result in lower model failure risks. The adaptability of the model is important to face the changes in the operational environment and would also lower model failure risks. Greater complexity, and weaker interpretability would increase the difficulties for investors to become familiar with the characteristics of the model. This, therefore, would increase the strategic application risks. The efficiency of the model could allow investors to make decisions more quickly.

2. Random Forest

2.1. Introduction to random forest

Random forest is a combinatorial classifier algorithm which combines Bragging and Randomization. It is composed of many single classification regression trees, and the generation of a single tree depends on an independent and identically distributed random vector. The generalization error of the whole tree depends on the classification efficiency of individual trees and the degree of correlation among the trees.

The upper bound of generalization error PE^* of random forest can be obtained as:

$$PE^* \leq \bar{p}(1 - s^2)/s^2$$

Where s is the overall classification efficiency of the combinatorial classifier

2.2. Risk evaluation on random forest

2.2.1. Accuracy

Assuming s > 0, when there are enough classification trees in the random forest, the generalization error of converges to a finite value everywhere. Therefore, as the number of classification trees increases, risk of overfitting is reduced. However, when it comes across classification and regression problems with large noise, the random forest algorithm can easily become overfit.

For data with imbalances, the random forest algorithm can effectively balance these errors. For data with missing values, the random forest algorithm has a relatively large processing capacity for missing problems. For out-of-set data, an unbiased estimate of the true error can be obtained during model generation without loss of training data.

On the other hand, Guan Runjing's paper showed that when the market is poor, the random forest multi-factor stock selection model cannot create stable positive interests in the stock market, so it cannot be applied to actual transactions and needs improvement [1].

2.2.2. Adaptability

Due to the combination of trees, the random forest can process nonlinear data and can handle both discrete and continuous data without normalization. Liu Wei and others found that the nonlinear characteristic of their forecast model based on random forest would reduce model mismatches as its accuracy is not susceptible to model definition errors [2]. In

addition, it is suitable for the many characteristic factors of the fund's heavy stock holding and can be applied to large-scale datasets, which is suitable to the complexity of the financial market.

DieRefich proved through experiences, the combination of Bragging method and Randomization can effectively reduce the influence of noise, and hence random forest can effectively process the data containing noise [3]. Random forest is also suitable for feature selection of high-dimensional input space and have strong adaptability to changing data.

2.2.3. Efficiency

Random forest obtains the efficiency of single classification tree and have a relatively fast training speed. Zhang Xiao constructs a trend-tracking model based on RF-GB algorithm. He found that random forest reduced the number of tuning parameters needs of GBDT algorithm and improve the training speed [4]. However, the efficiency of random forest algorithm is compromised when dealing with unbalanced data or continuous data. Li Xiang concluded that the XG Boost algorithm have stability over random forest [5], while Zhu Yangbao concluded that though random forest have good stock selection capabilities, XG Boost have better stock selection capabilities and are faster than the Random Forest when applied to quantify stock selection [6].

2.2.4. Interpretability

Random forest is easy to interpret. After training the model using the random forest algorithm, the algorithm can determine which features are important and the random forest algorithm can be used to make feature selection.

2.2.5. Simplicity

Due to the large number of stock data features and high noise, random forest is considered to be introduced into the feature selection of the model. Random forest is simple in its application. It can observe the change of model accuracy by randomly adding noise interference to each feature, and measure the importance of feature by the extent of accuracy reduction. If the model accuracy increases after noise reduction for a feature, it indicates that the feature is of high importance.

3. XG Boost

3.1. Introduction to XG Boost

Boosting algorithm is a machine learning method that integrates many weak classifiers together to generate a strong classifier by reducing the bias in supervised learning. XG Boost is an algorithm based on many boosting algorithms such as AdaBoost and GDBT, whose principle is to optimize the objective function and minimize it. By optimizing the objective function, the error and complexity are optimized.

The XG Boost algorithm uses a large number of base classifiers, which need a more general algorithm to achieve gradient descent. This can be achieved by using the Taylor second-order expansion is used:

$$Obj_m \cong \sum_{i=1}^{n} [l(y_i, y_i^{m-1}) + g_i f_m(x_i) + \frac{1}{2} h_i f''(x) \Delta x^2] + \emptyset(f_m)$$

where n represents the number of samples used, the number of current iterations of the m table, and f(m) represents the current iteration error.

3.2. Risk evaluation on XG Boost

3.2.1. Accuracy

XG Boost adds the regularization term into the objective function to ensure that every iteration hedged the complexity of the model and effectively reduces the possibility of overfitting. Since the numerical value of each feature is only used for size comparison, the XG Boost model have good tolerance for outliers. For the treatment of missing values, XG Boost can automatically learn the splitting direction of missing values.

Tian Hao analysed the XG Boost algorithm and compared it with other mainstream algorithms. Tian applied a data pre-processing method in which the original data can be displayed in a smaller specification. This method can achieve the purpose of dimensionality reduction and retain the information of the original data more completely, so it reduces the time required for modelling, while maintaining the accuracy of the model [7].

3.2.2. Adaptability

The weak classifier that makes up XG Boost is a decision tree, and its appropriate data type is categorical data. Therefore, it is necessary to convert the data input into classed data. The XG Boost algorithm can be used for both nonlinear classification and linear classification.

Xiang Li showed based on the classification of data from XG Boost, quantify stock selection objectives and a yield that outperform the broader market can be achieved [5]. For the first time, the XG Boost algorithm was applied to quantitative stock selection, which made it possible to handle high-dimensional. As multi-factor selection stock method became possible, it presents broader development ideas about new multifactor models.

Runjing Guan constructed a quantitative multi factor stock selection model based on XG Boost. The model can identify the downtrend of the stock well, thus it reduces volatility. However, the model fails to identify the rise in the stocks, making it difficult for higher returns to be achieved [1].

3.2.3. Efficiency

The XG Boost algorithm supports CPU parallel operations and column sampling which can assist in reducing the amount of computation. Column sampling would also reduce overfitting. Though XG Boost have many parameters, but in actual tuning, there are very few parameters that can significantly improve the predictive ability of the model. If the amount of data is very large, the modelling and parameter optimization process will waste more time for the time it takes. The distributed method can be used to improve the efficiency when computing as due to the gradient descent algorithm used, there is no correlation between the multiple classifiers it uses.

As mentioned before, Yangbao Zhu found that the XG Boost are better than the random forest in terms of running speed or stock selection ability [6]. Speed can be improved further by organizing the original data to be displayed in a smaller specification like Tian Hao did in his research [7].

3.2.4. Simplicity

Both tree building process and boosting process of XG Boost are based on the objective function, and all operations

are evaluated by minimizing objective function. The formula for it is relatively clear and easy to understand. However, its initial code implementation is written in C++, hence when XG Boost is used in such as programs such as Python and R, the installation can be cumbersome.

3.2.5. Interpretability

Due to Taylor's second-order expansion, the process of building trees and boosting relies only on the first derivative and second derivative of the loss function, and the formula is relatively clear and easy to understand.

4. Logistic Regression

4.1. Introduction to logistic regression

Logistic regression is a probabilistic model in which the probability of occurrence of an event is the dependent variable and the influencing factor is the independent variable. It is widely used in many fields such as medical health, disaster prediction, and risk rating. It assumes that P(y = 1) = p, hence P(y = 0) = 1 − p. From this, $Ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$, and $Ln\left(\frac{p}{1-p}\right)$ is denoted as Logit P.

$$Logit(P) = Ln\left(\frac{p}{1-p}\right)$$

From which, β_0 is constant, β_1, β_2, …, β_n represents the size of impact of independent variable x_i on the dependent variable Y, ε_i is a random perturbation term. Assuming that the x_i value remained unchanged, then the regression coefficient β_i and the probability P shares a positive relationship. $\frac{p}{1-p}$ represents the odds ratio, which is the ratio of the probability of an event occurring and not occurring. If β_i is positive the odds ratio would increase. Conversely, if β_i is negative, then the odds ratio would decrease. From these, it can be established:

$$P = \frac{1}{1 + exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)\}}$$

4.2. Risk evaluation on logistic regression

4.2.1. Accuracy

As a linear model, logistic models have high accuracy, stability. Logistic regression requires more data to fit a better model; yet a large number of multiclass features or variables are not handled well as logistic regression has limited learning ability. Logistic regression is sensitive to outliers and is easy to overfit and underfit. Therefore, its classification accuracy is not high.

For variable filtering or dimensionality reduction, the model has high requirements of financial indicators and data quality. If the number of metrics is large and the processing of it is inappropriate, it reduces the accuracy of the model's predictions. Therefore, before the model predicts the classification, it needs to add a combination of feature indicators to the original indicator and apply them to logistic regression model predictions.

To compensate for the logistic model's high data requirements and insufficient interpretation of the target, Logistic regression is often combined with the GDBT model. Facebook proposed in 2014 that GBDT is used to solve the problem of logistic model feature combination, which successfully improves the accuracy of model prediction confirmatory rate [8].

4.2.2. Adaptability

Unlike multilinear regression, logistic regression model does not require both independent and dependent variables to be continuous variables. However, the dependent variable can only take the values 0 and 1. Therefore, the model is more suitable for binary classification problems

Ohlson was the first to apply logistic regression to financial risk warning, which confirmed to have a good classification effect [9]. The model was able to determine whether financial risks have occurred and can objectively assess the probability of financial risk in a business. At the same time, the model overcomes the shortcomings of multivariate discriminant analysis, which requires the data to satisfy the normal distribution.

4.2.3. Efficiency

The amount of calculation involved is small, hence it is fast and does not require much storage. However, there is a trade-off between feature space and logistic regression performance.

4.2.4. Simplicity

Logistic regression holds the advantage of having fewer parameters. It is a generalized linear regression analytical model, and is simple to implement. It is an excellent classification model for linear separable problems.

4.2.5. Interpretability

The explanatory ability of logistic regression is relatively strong and convenient to use, and most scorecards are based on logic regression build. The output of the prediction result is probability, between 0 and 1. The weighting of each variable is also evident and the calculated coefficients are interpretable, so the result is easy to understand. However, if the number of metrics is large and the processing of it is inappropriate, the interpretation of the metrics on the objective is quite poor.

5. Conclusion

It is impossible to anticipate the changes in exogenous factors, hence it is particularly important for quantitative investment institutions to enhance their capabilities and reserve differentiated investment strategies to cope with market shocks. Asset managers should be equipped with the ability to understand the inner workings of a model as they have a fiduciary responsibility to understand and communicate the risks of their clients' portfolios. This places particular emphasis on the interpretability of their models. In addition, institutions should set appropriate investment target and select the strategic model that can undertake risks under the target frame.

Through evaluations of past researches, this paper first highlighted the characteristics of random forest, XG Boost, and logistic regression. Then this paper evaluates the accuracy, validity and the adaptability of the model which would affect the model failure risks, and the efficiency, simplicity and the interpretability of the results which would affect strategic application risk.

Many interesting potential avenues of research to draw more meaningful and institutive conclusions from financial machine learning models. This paper aimed to construct a theoretical comparison between different algorithms of machine learning that is used in quantitative investment. It

hopes to provide investors undertaking quantitative investment with a guideline, from a risk point of view, that can be considered when deciding upon which model they should use.

References

[1] Guan Runjing (Southwestern University of Finance and Economics), 2020. Research on Quantitative Investment Stock Selection Based on Machine Learning. http://jour.ucdrs.superlib.net/views/specific/2929/thesisDetail.jsp?dxNumber=390108795987&d=24E1BB8447C99E46036E3917F65046A2&sw

[2] Liu Wei. Luo Linkai. Wang Huazhen. (2008) A forecast of bulk – holding stock based on random forest. Journal of Fuzhou University (Natural Science), 36: 134-139.

[3] Dietterich T.G., 1998. An experimental comparison of three methods fottp://dx.doi.org/10.1023r constructing ensembles of decision tree: bagging，boosting and randomization. Machine Learning. http://dx.doi.org/10.1023/A;1007607513941

[4] Zhang Xiao (Guangxi University), 2018. Quantitative Investment Model Based on Improved GBDT. http://jour.ucdrs.superlib.net/views/specific/2929/thesisDetail.jsp?dxNumber=390106882309&d=61E8F81F5E085DC2DD7DD12E879976AD&sw

[5] Li Xiang (China Academic Journal Electronic Publishing House), 2017. Multi-Factor Quantitative Stock Option Planning Based on XGBoost Algorithm. http://www.cnki.net

[6] Zhu Yangbao (Nanjing University), 2017. Multi-Factor Stock Selection Scheme Design Based on XGBoost and LightGBM Algorithm. http://jour.ucdrs.superlib.net/views/specific/2929/thesisDetail.jsp?dxNumber=390108102120&d=26F9D47EA1EC93275B365AEB779DA30D&sw

[7] Tian Hao (Shanghai Normal University), 2018. Analysis on Shanghai and Shenzhen 300 Quantitative Investment Based on XGBoost Algorithm. http://jour.ucdrs.superlib.net/views/specific/2929/thesisDetail.jsp?dxNumber=390106715466&d=8A3A3E77129F17DBDF536E9D07308645&sw

[8] He X, Pan J, Jin O, et al (ACM), 2014. Practical Lessons from Predicting Clicks on Ads at Facebook. https://dl.acm.org/doi/abs/10.1145/2648584.2648589

[9] Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), 109–131. https://doi.org/10.2307/2490395