Research on Estate Mortgage Risk Control Based on \textit{RF-XGboost Model}.

Ziyan Zhang, Nan Liu\textsuperscript{*} and Xinxin Xu
School of Economics and Management, Chongqing Jiaotong University, Chongqing, 400074, China
\textsuperscript{*}Corresponding author’s e-mail: liunan@cqjtu.edu.cn

Abstract. With the expansion of the real estate mortgage market, commercial banks urgently need efficient and convenient means to deal with the high bad debt rate of real estate mortgage loans. Meanwhile, government departments also have desire to establish a scientific and effective system to supervise the real estate mortgage market. Therefore, the concept of machine batch evaluation in this paper establishes an intelligent real estate evaluation model based on \textit{RF-XGboost model}, which can not only complete the market value evaluation subject of the designed real estate by means of machine batch evaluation with high efficiency and low cost, but also help to promote the standardization process of the evaluation industry, make the mortgage market transparent, and finally realize the effective control of real estate mortgage risks.

1. Project Background

1.1. \textit{There are great risks in the real estate mortgage loan business of commercial banks}
With the acceleration of urbanization and the explosion of the urban resident population, people's housing demand continues to rise. By 2017, the total scale of real estate mortgage loans of commercial banks in China has exceeded 322.5 billion, accounting for more than 70\% of the total amount of commercial bank loans. According to the practice, the bank's loan amount is generally not more than 70\% of the evaluation results, in order to objectively determine the security of the real estate mortgage loan business, but in reality, it often backfires.

Due to the limitations of traditional evaluation methods, the imperfection of bank lending platforms and the incompleteness of lending supervision, the bad debt rate of lending business of commercial banks and other financial institutions has remained high [1]. Also in 2017, the overall non-performing loan ratio of China's commercial banks reached 1.74\% and showed a continuous upward trend. Therefore, it is difficult to assess and confirm the value of the collateral to offset the principal and interest of the loan, causing commercial banks as lenders to still suffer huge economic losses [2].

1.2. \textit{The CBRC continues to strengthen supervision of the mortgage lending market}
Under the background of economic globalization and accumulation of real estate bubble, the government financial supervision department has always adhered to the principle and bottom line of "no systematic financial risks", paid attention to enhancing macro-prudential supervision of the real estate market, established an effective supervision system, adhered to legal supervision, implemented penetration supervision, strengthened comprehensive supervision, and highlighted functional supervision and behavioral supervision[3] . Since the beginning of 2017, the Management Guidelines and Interim Measures issued by the CBRC have not only forced banks to increase the intensity and investment of
post-lending re-evaluation, but also clearly classified the risk of bank assets and tightened the scope of identification of non-performing assets. In this way, it is enough to promote commercial banks to improve their own level of bearing market risks, to achieve the effect of standardizing the market, and to avoid the incentives of market risks to a certain extent.

As a result, the non-performing loan ratio has been more truly reflected and the market transparency has been further improved. As a result, customers' awareness of non-performing assets control has been further enhanced. At the same time, it is imperative for the government financial regulatory agencies to achieve the goal of market standardization.

2. Research Status

2.1. Current Situation of Risk Control of Real Estate Collateral in the Era of Big Data

As a lender, the market risks of commercial banks and other financial institutions are induced by the rise in bad debt rate caused by the fluctuation of market value after mortgage loan. Therefore, for the main body of the bank, realizing the value monitoring of collateral is the only way to control its risks. As a necessary key work, the market value assessment of real estate collateral is often completed by commercial banks entrusting professional assessment agencies and appraisers.

However, there are three main shortcomings in traditional evaluation methods [4]:

- High cost of batch operations: when the appraisal company collects the appraisal fees, it generally accrues the service fees according to the market appraisal of the appraisal object, and manual batch appraisal of real estate collateral requires high cost, which may even lead to the chaos that the cost of risk control is higher than the risk loss.

- The efficiency of the traditional operation method is relatively low: when using the three major asset evaluation methods (market method, income method and cost method), a large amount of information needs to be collected. Meanwhile, the integrity and authenticity of data and cases need to be ensured, and the data processing also needs to take a long time, thus greatly reducing the efficiency of the evaluation operation.

- The evaluation results are highly subjective: in the process of evaluation, real estate appraisers need to artificially correct the influencing factors. This work is not only difficult, but also highly subjective, requiring appraisers to have strong professional skills and good moral qualities.

This shows that even if professional appraisal institutions and real estate appraisers have the ability to undertake appraisal business, the disadvantages of traditional operation methods are still exposed. Therefore, the demand for scientific and efficient collateral intelligent appraisal systems by commercial banks continues to rise [5]. At present, in many countries with developed real estate markets, machine batch evaluation technology has been widely used in the assessment of specific purposes and has been widely used in the tax base assessment of real estate.

2.2. Definition of Machine Batch Evaluation

Standardization of domestic evaluation industry. The level of intelligence is relatively low. At present, it is mainly based on manual operation. The batch evaluation of machines only stays in the field of scientific research and has not been put into practice. As an imported product, China has given different interpretations to the relevant definitions given by the International Assessment Standards, the Standards of the International Association of Tax Assessors and the American Association for the Promotion of Assessment. However, the core of the three definitions is based on the three basic methods of market method, cost method and income method, which are combined with mathematical statistics for assessment. The definition of machine batch evaluation in this paper refers to the relevant provisions of International Evaluation Standards: machine batch evaluation refers to the evaluation of various types of property within a given time by using systematic, unified and statistical inspection and analysis evaluation techniques and methods taking into account the results [6].
2.3. Review of Research on Batch Evaluation of Machines at Home and Abroad
As early as the second half of the 19th century, the theoretical system of modern asset evaluation in western capitalist countries took shape. In contrast, the real estate appraisal industry, which started in the 1980s in China, is inferior to the appraisal industry in developed capitalist countries in terms of its technical level and operational efficiency due to traditional appraisal methods, high cost of batch operations and subjectivity of appraisers.

O'Connor [7] analyzed and compared three regression models of residential buildings, namely addition, multiplication and linearity, and provided effective theoretical basis for evaluation work by comparing the advantages and disadvantages of the three methods. Kettani and Oral [8] design and implement a real estate evaluation system based on Quebec, Canada, and use analogy regression method as the technical barrier of the evaluation system. In addition, factors such as personal evaluation grade and real estate type of collateral holders are also comprehensively considered in the evaluation process. Mccluskey and Deddis [9] combine GIS technology with batch evaluation to realize large data collection and data distribution statistics, and to carry out data storage, processing analysis and static prediction. Mccluskey and Mccord [10] used 2694 sets of residential data to compare the two batch evaluation methods of regression model and artificial neural network. The results show that artificial neural network has more advantages in estimation accuracy, and regression model is easier to be understood and approved by people. At present, the international real estate evaluation industry is showing a trend of big data and intelligence, which can provide reference for commercial banks' mortgage lending. In this paper, the high precision and high efficiency evaluation of real estate collateral will be realized through the construction of collateral machine batch evaluation system and the RF-XGboost joint algorithm, and feedback will be given to commercial banks based on the evaluation results to realize real-time monitoring of collateral market value and timely stop-loss and effective control of lending risks.

3. Design of real estate collateral evaluation index system
Through the analysis of the three major asset appraisal principles, it is not difficult to find that in the actual appraisal business process, professional appraisers often make qualitative analysis of some indicators and influencing factors according to their own appraisal work experience and experience. Because the evaluation process is often mixed with appraisers' personal understanding and cognition, this operation method is easily interfered by subjective factors [11]. Due to the lack of standardized and standardized process management, the real estate collateral value assessment business is always difficult to ensure its accuracy and objectivity.

Based on random forest model and XGBoost model, this paper establishes a rapid assessment model of real estate collateral. In this process, the sklearn toolkit needs to be imported to realize random forest classification, and then the evaluation results are calculated and output through the boosted tree toolkit XGBoost. As far as Python is concerned, we must quantify the indicators in the process of programming, and even comprehensively reflect the first-level indicators through the second-level indicators. This paper selects 19 characteristic variables including floor area, decoration grade, floor number, room number, building age, building type, property management fee, floor area ratio, greening rate, parking space ratio, park presence, proximity to universities, proximity to key schools, educational facilities, living facilities, average residential price, distance from the nearest light rail station or subway station, distance from the nearest business circle and number of bus stations to comprehensively reflect the market value of real estate collateral to be evaluated. The above 19 indicators can be divided into three categories: construction factor (A), neighbor hood factor (B) and location factor (C).

- **A**: Building factors refer to the influencing factors related to the building engineering structures where the real estate collateral is located, which match the needs of consumers, thus directly affecting the market value of the real estate collateral.
- **B**: Each neighborhood unit must be equipped with corresponding public facilities, such as educational institutions, social service facilities, medical facilities and entertainment facilities.
- **C**: Location factors refer to the incentives and conditions that make geographic location function or change.

![Real estate collateral assessment indicator system](image)

**Figure 1. Evaluation Index System of Real Estate Mortgage.**

4. **Establish a real estate collateral rapid evaluation model based on RF-XGboost.**

4.1. **Basis for Selecting Random Forest Regression Analysis**

This paper selects three first-class indexes of construction factor, neighborhood factor and location factor, and uses 19 basic factors to comprehensively reflect the above three first-class indexes, and then calculates the market value of the real estate collateral to be evaluated. In addition, for the real estate in some special areas, it is necessary to add secondary indicators, even tertiary indicators. Therefore, due to the characteristics of large index dimension and large sample data volume, the collateral rapid assessment model cannot meet the requirements of data processing accuracy and stability such as least square regression method, gray correlation method and other traditional regression analysis methods. Therefore, this paper intends to use machine learning algorithm to build a rapid evaluation model of real estate collateral, and then realize accurate evaluation of the market value of real estate collateral. Rich Caruana [12] finally concluded that the comprehensive performance of the random forest model is due to BP neural network and support vector machine through comparison and selection of three machine learning algorithms. In addition, considering that when part of the sample data in the training sample set is true, the random forest model can guarantee the accuracy of the evaluation results of the market value of collateral, and can measure the importance of index features by quantitative means after the training is finished.

Therefore, this paper uses random forest to build a rapid assessment model of real estate collateral, and uses Python's scikit-learn toolkit to conduct random forest regression analysis on the above influencing factors. as shown in Figure 2, it is a visual random forest regression principle.

![Fig 2. Schematic diagram of random forest regression model](image)
4.2. Model Improvement

Although random forest is superior to training decision tree in precision by virtue of Bootstrap Aggregating Integration Principle, it is prone to over-fitting. In addition, with the number of training sets further increasing and the size of trees further expanding, how to improve the efficiency of program operation has become a top priority. Researchers need to prune trees appropriately to reduce the time needed for regression analysis and to improve random forest ensemble by expanding memory. Therefore, in order to solve the above problems, this paper introduces XGBoost algorithm to improve the model.

With the help of the features of the random forest that can measure the importance of the feature indexes by quantitative means, and taking advantage of the combined classification, the combination of the two machine learning algorithms, i.e. classifying the feature indexes by means of the random forest model and sorting them according to the quantitative importance degree, obtains the features with higher importance and generates new sample data. Then, the XGBoost model is used to train and test the sample data. Finally, the accuracy of evaluation results and the efficiency of program operation are improved. It can not only solve the regression analysis problem under more complicated conditions and complete the market value evaluation of real estate collateral under more complicated conditions, but also provide reference and basis for complicated evaluation problems. The improved collateral rapid evaluation model is shown in Figure 3.

![Diagram of real estate collateral rapid evaluation model based on RF-XGboost](image)

5. Conclusion and prospect

5.1. Significance and Advantages of Machine Batch Evaluation

Today, with the level of urbanization and informatization rising continuously, on the one hand, facing the dilemma of high bad debt rate of mortgage loan business, commercial banks urgently need a scientific and efficient risk control method to reduce bad debt losses. On the other hand, the strength of the government's regulation of the real estate market is evident, and a financial supervision system led by the government is imperative. This paper introduces the concept of machine batch assessment. The rapid assessment model of real estate collateral based on random forest -XGBoost joint algorithm has the following five advantages:

- Effectively improving the efficiency of the evaluation operation and saving the time of the evaluation operation;
• Greatly reducing the time cost and labor cost of the assessment operation, and promoting the standardization and standardization of the mortgage loan market;
• With the help of big data algorithm and mathematical statistics theory, the accuracy of the evaluation results is improved to a certain extent, the accuracy of the bank loan amount is improved, and then the bad debt rate of the loan business is reduced;
• It can realize the dynamic monitoring of real estate collateral and meet the requirements of domestic and foreign norms and methods for regular post-loan evaluation;
• The rapid assessment model based on *random forest* and *XGBoost* can be applied to more assessment scenarios and can be used for efficient and accurate assessment under more flexible assessment conditions. In addition, it is a great innovation in the field of machine evaluation to quote big data algorithms with higher quantization and intelligence.

5.2. *Existing Deficiencies and Prospects*

In this paper, random forest and XGBoost are combined to expand the application scope of the model and improve the efficiency of program operation to a certain extent. However, both of them are algorithms developed based on the theory of training decision tree, which are prone to over-fitting, reduce the operation efficiency and affect the overall operation progress. In the future, the above algorithm will be optimized and upgraded to further reduce the impact caused by the insurmountable defects of the algorithm itself, or an algorithm program with wider application scope and higher operation efficiency will be developed to provide more time-honored means for the evaluation of the market value of collateral and the research in the field of loan risk control of commercial banks.

References

[1] Zhang M.J. (2018) Risk and Prevention of Commercial Bank Mortgage. *Economic and Trade Practice*, 18:155-157.
[2] Chai Y.P, Li Y.Q. (2015) A Preliminary Study on the Risk Management of Agricultural Bank Mortgage. *Rural Finance Research*, 05: 56-58.
[3] Tang N. (2018) Analysis of the Impact of “Strong Supervision” on Commercial Banks. *Times Finance*, 20:81-83.
[4] Pang F. (2017) Application of random forest model in batch evaluation of second-hand housing. *Chongqing Jiaotong University*, Chongqing.
[5] Xuan D. (2018) Research on real estate loan risk control based on machine batch evaluation. *Chongqing Jiaotong University*, Chongqing.
[6] (2018) International Assessment Criteria - Defining Value Types. *China Assets Appraisal*, 01: 7+11+16+39.
[7] O’Connor, P. M. (2002). "Comparison of Three Residential Regression models: Additive, Multiplicative, and Nonlinear." *Assessment Journal*.
[8] Ossama Kettani, Muhittin Oral. Designing and implementing a real estate appraisal system: The case of Québec Province, Canada[J]. *Socio-Economic Planning Sciences*, 2015.49.
[9] Mccluskey, W., et al. (1997). "Interactive application of computer assisted mass appraisal and geographic information systems." *Journal of Property Valuation Investment* 15(5): 448-465.
[10] Mccluskey, W. J., Mccord, M., Davis, P. T., Haran, M., & Mcilhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, 30(4), 239-265.
[11] Zhang Z.Y. (2019) IGFSS evades confirmation bias in construction management. In: Proceedings of 2019 5th International Conference on Economics, Management and Humanities Science (ECOMHS 2019). Bangkok. pp.1127-1132.
[12] Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *International Conference on Machine Learning*. 