Financial Economics Insights

Article

Application of Data Analysis in Bank Mortgage Loan Risk Assessment

Chuhan Wang [1,*]

[1] Carey Business School, Johns Hopkins University, Washington, DC, 20001, USA
* Correspondence: Chuhan Wang, Carey Business School, Johns Hopkins University, Washington, DC, 20001, USA

Abstract: Under the background of the continuous growth of bank mortgage loan business, how to accurately assess loan risk has gradually become a key issue in bank risk management. The traditional mortgage risk assessment method has shown its deficiency in efficiency and accuracy, and the application of data analysis technology has brought new possibilities to solve this problem. This paper discusses the application of data analysis technology in bank mortgage risk assessment, including the integration of machine learning algorithm, big data analysis, risk assessment model construction and data visualization, by analyzing data collection and pre-processing and common data analysis techniques.

Keywords: data analysis; bank mortgage loans; risk assessment

Published: 27 May 2025

1. Introduction

In the risk management of bank mortgage loan, how to accurately evaluate the repayment ability of borrowers and the actual value of mortgage assets is always the key problem that banks must solve. With the evolution of financial markets and the development of information technology, traditional evaluation methods have become inadequate to meet the new challenges. The use of data analytics, especially machine learning and big data analytics, can extract key variables from a wide range of information to help banks more accurately assess the degree of risk in loans. With the help of data analysis technology, banks not only improve the efficiency of risk identification, but also provide strong support for loan approval and risk control. Therefore, the research on the practical application of data analysis in bank mortgage loan risk assessment has far-reaching application prospects and academic value.

2. Overview of Data Analysis Methods

2.1. Data Collection and Preprocessing

In the assessment of mortgage loan risk, data collection and pre-processing are very important to ensure the accuracy and quality of data analysis. Initially, banks had to collect data on loan applicants through multiple channels, including personal finances, credit history, valuation of collateral, and market interest rates. This information may come from the institution's internal systems, external credit rating agencies, or publicly available government data sources. Once collected, the data must be initially processed to maintain its integrity, consistency, and accuracy. The preliminary processing involves data purification, filling blank values, abnormal data detection, data normalization and so on. Data purification helps to eliminate duplicate or useless information, and processing of blank values and outliers prevents these data from interfering with subsequent analysis [1]. After completing this pre-processing, the bank will be able to build a high-quality data set, which will lay a solid foundation for the construction of credit risk assessment models.

2.2. Common Data Analysis Techniques

The commonly used data analysis techniques include regression analysis, decision tree, cluster analysis and neural network. In terms of the relationship between the predictor variables, regression analysis is to build a mathematical model to explore the internal relationship between the borrower's financial status and credit risk. Decision trees classify information hierarchically to help banks evaluate the credit risk of loan applicants. Cluster analysis aims to group borrowers according to risk levels, so as to implement differentiated management strategies for customer groups. The neural network mimics the structure of human brain neural network, mining potential regularity in large-scale data sets, and thus improving the accuracy of prediction. Each technology has its own advantages and uses cases, usually depending on the specific situation to choose the appropriate analysis method [2].

3. Traditional Methods of Bank Mortgage Loan Risk Assessment

3.1. Credit Scoring Model

Banks generally use credit scoring models as a traditional method for risk assessment of loan applicants. The model estimates the likelihood of future default based on an applicant's credit history, ability to repay loans, and financial status. Among many scoring models, logistic regression model, discriminant analysis model and support vector machine model are common. The logistic regression model is commonly used to build credit score models. The core idea of logistic regression is to predict the probability of default from a set of independent variables (such as income, debt ratio, loan history, etc.). Specifically, the basic form of a logistic regression model is as follows:

$$P = \frac{1}{1+\exp\left(-\left(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n\right)\right)} \tag{1}$$

Among them, P is the probability that the borrower defaults, b_0, b_1, $\cdots$, b_n are the regression coefficient, X_1, X_2, $\cdots$, X_n are independent variables that affect credit scores (such as borrower's age, income level, debt ratio, etc.). The model uses the maximum likelihood estimation method to determine the value of each regression coefficient to generate a credit score. The advantage of this model is that it can process the influence of multiple factors and quantify the results as the probability of default, helping banks to assess the risk of loans. By setting a threshold for a credit score, banks can classify borrowers into "low risk" and "high risk" and thus develop different lending strategies [3].

3.2. Asset Value Assessment

In the process of risk assessment of bank mortgage loans, an accurate assessment of the value of assets occupies the core position. The main goal of this step is to clarify the actual value of the collateral provided by the borrower in the market, so as to ensure that the loan limit and the value of the collateral assets are kept in a reasonable range. Conventional asset valuation methods include market comparison method, cost calculation method and income forecasting method. Among them, the market comparison method judges the value of assets by referring to the transaction prices of similar assets in the market. This method is especially applicable when there are a large number of transactions of similar assets in the market, and is often used in the valuation of real estate. Evaluators will refer to the market transaction price of similar properties, and take into account the location of the property, construction time, size and other conditions, in order to calculate the market value of the target asset. Costing, which values an asset based on its replacement cost, is often used for special assets, such as non-residential properties or commercial projects, where it is difficult to find a transaction in the market. The valuation expert will calculate the current net value of the asset based on factors such as construction costs, depreciation and market demand. The income forecasting method is more applicable to the valuation of assets that generate stable cash flows, such as rental properties or commercial buildings [4].

3.3. Analysis of Borrower's Income and Repayment Ability

When the bank evaluates the risk of mortgage loan, the financial return of the applicant and its solvency are the key factors in the evaluation. Conventional income audits rely heavily on financial reports, tax documents and proof of income submitted by applicants to determine whether they have sufficient funds to repay debt. First, the bank will verify the income received by the applicant on a monthly or annual basis, which usually covers salary income, bonuses, investment returns, leasing income and other diversified income categories. In-depth analysis of these returns allows banks to estimate the amount of money an applicant has at his disposal and compare it with the monthly payments to determine whether there is a likelihood of difficulty servicing the debt. In addition, the bank will also assess the stability of the applicant's income, including the diversity of income sources and the potential risk of income fluctuations. If the applicant's income comes mainly from a single source, especially if it is closely linked to a specific industry or enterprise, the bank will exercise greater caution in assessing solvency. In addition, the applicant's debt status is also an important basis to measure its solvency. Traditionally, the applicant's Debt-to-Income Ratio (DTI) is calculated. DTI is a measure of an applicant's total monthly debt repayments as a proportion of their monthly income. If the DTI ratio is too large, it means that the applicant is under greater debt pressure, and the bank may consider raising the risk premium, reducing the loan amount, or even directly refusing the loan application [5].

3.4. Loan Default Prediction Model

The loan default prediction model is an important part of bank mortgage risk assessment. Traditional default prediction methods mainly rely on statistical models, usually by analyzing historical data to predict the lender's repayment ability. Common default prediction models include logistic regression model, discriminant analysis, decision tree, support vector machine (SVM) and so on. Logistic regression model is the most common default prediction model. Its basic principle is to estimate the default probability by regression analysis of multiple risk characteristics of the lender. The specific formula is as follows:

$$P = \frac{1}{1+\exp(-z)} \tag{2}$$

Among them, P is the probability that the lender will default, $\exp(-z)$ is an exponential function, z is a linear combination of predictor variables, in the form:

$$z = \beta_0 + \beta_1 X_1 \beta_2 X_2 + \cdots + \beta_n X_n \tag{3}$$

Among them, β_0 is the intercept term, β_1, β_2, $\cdots$, β_n is the regression coefficient, X_1, X_2, $\cdots$, X_n is a variety of characteristic variables of the lender (such as credit score, income, debt ratio, etc.). By training the data set, the bank can estimate the regression coefficient and calculate the default probability for each lender. When the default probability exceeds a certain threshold, the bank can judge the loan as high risk and take corresponding risk control measures.

4. Application of Data Analysis Technology in Mortgage Loan Risk Assessment

4.1. Application of Machine Learning Algorithm

Machine learning algorithms have become important tools in the risk assessment of bank mortgage loans. Traditional credit evaluation methods rely on manual experience and rules. They are easily influenced by subjective factors, and their evaluation efficiency is low. Machine learning algorithms can automatically identify potential risk factors through in-depth analysis and modeling of historical data, thereby improving the accuracy and efficiency of risk assessment. Machine learning algorithms mainly build models through data training, and commonly used algorithms include decision trees, random forests, support vector machines (SVM) and neural networks. Through these algorithms, banks can analyze customers' credit history, income level, debt situation, loan amount and

other factors to assess their repayment ability and default risk. Table 1 below shows examples of several machine learning algorithms used in mortgage risk assessment.

Table 1. Comparison of Application of Machine Learning Algorithms in Risk Assessment.

Algorithm class	Application scenario	Advantage	Shortcoming
Decision tree	It is used to classify loan customers and assess risk levels	Easy to understand and explain; Missing values can be handled	Easy to overfit, poor generalization ability
Random forest	Integrated learning method based on decision tree enhances prediction accuracy	The prediction accuracy is improved and overfitting is avoided	The model has high complexity and large computation
Support Vector Machine (SVM)	For classification problems of high dimensional data	It performs well in high dimensional space and has strong robustness	Sensitive to parameter selection and long training time
Neural network	Deal with complex nonlinear relationship and predict the probability of loan default	Powerful fitting capability, suitable for complex pattern recognition	Sensitive to parameter selection and long training time; It requires a lot of data and computing resources

By training on historical loan data, the machine learning model can identify potential defaulting customers and set loan interest rates or repayment strategies based on their risk level. In addition, as the amount of loan data collected by banks increases, the machine learning model will continue to optimize, and the accuracy and real-time nature of the assessment will continue to improve.

4.2. Big Data Analysis Technology

Nowadays, banks use big data analysis to assess mortgage risk, which has become particularly critical. By analyzing numerous customer information, market dynamics, and historical lending behavior, banks can rely on big data technology to make more accurate risk assessments, effectively reducing the occurrence of defaults, and optimizing the accuracy and rationality of loan approvals. The application range of this technology is wide, covering data collection, storage, processing, analysis and other links, among which the most critical is data mining and pattern recognition. First, banks use big data technology to build a customer credit scoring model through comprehensive analysis of customers' social behavior, consumption pattern, income level, credit history and other data. For example, banks use machine learning algorithms to build more accurate risk assessment models based on non-traditional data such as customers' bank transactions, social media information, and purchase behavior. These models can automatically identify potential default risks and help banks to take timely risk control measures.

Furthermore, by using big data technology, banks can monitor and forecast external factors such as supply and demand changes in the real estate market, regional economic development trends, and government policies in real time to help banks assess the future market value of collateral. Changes in these factors have a direct impact on loan risk. Through big data analysis, banks can more accurately judge the appreciation potential or depreciation risk of collateral, so as to optimize the approval and quota setting of loans. In addition, through big data analysis, banks can continuously monitor various risk indicators of borrowers and their collateral. Once market conditions fluctuate, banks are able to quickly adjust lending conditions or implement necessary risk management measures. The following Table 2 shows examples of big data applications in mortgage credit risk assessment.

Table 2. Application of Big Data Analysis in Mortgage Loan Risk Assessment.

Application field	Technical method	Specific application
Customer credit evaluation	Machine learning, deep learning algorithms	Evaluate the credit rating based on the customer's historical behavioral data, social data, etc.
Appraisal of the value of the collateral	Big data analysis, market prediction model	Assess changes in the real estate market and predict fluctuations in collateral value
Risk monitoring and management	Data mining, real-time analysis model	Dynamic monitoring of loan default and mortgage depreciation

In general, big data analysis technology provides a new idea and method for bank mortgage risk assessment. Through comprehensive data mining and intelligent analysis, the accuracy and timeliness of risk assessment can be greatly improved, and the credit risk of banks can be effectively reduced.

4.3. Construction of Risk Assessment Model

A typical mortgage risk assessment model can be divided into several key steps, such as data collection, feature selection, model construction, model training and evaluation (see Figure 1). First, banks need to collect data from a variety of data sources, such as the borrower's personal credit history, income level, repayment ability, and the market value of the collateral. Then, these data are cleaned and pre-processed to remove outliers and missing data to ensure the quality of the data. Next, the most critical variables for predicting loan risk are selected through feature engineering, such as borrower credit score, debt-to-income ratio, collateral location and valuation. These characteristics can be selected through statistical analysis or based on domain knowledge. Once the appropriate algorithm has been selected, the next step is to train the model. In the training process, the training set and the test set are usually divided for the training and evaluation of the model. The training set is used for model learning, while the test set is used to verify the model's predictive power in real-world scenarios. The model can be further optimized by cross-validation and other techniques. Finally, the performance of the model is verified by evaluating indicators such as accuracy, recall rate, F1 score, etc. The evaluation results can help banks decide whether to use the model for actual loan risk assessment.

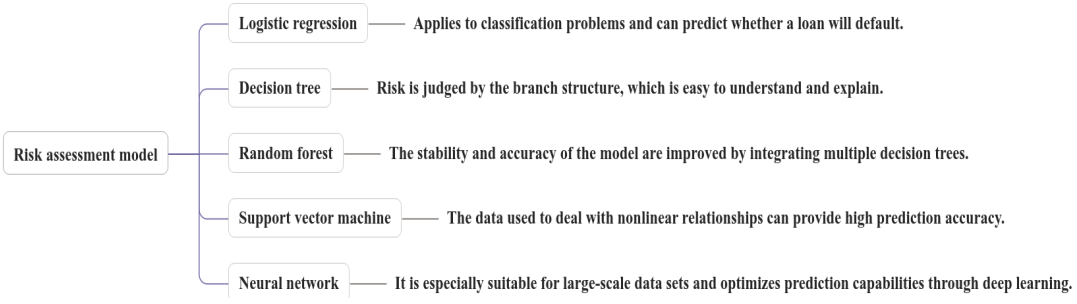

Figure 1. Risk Assessment Model Framework.

4.4. Data Visualization and Decision Support

In the process of mortgage loan evaluation, visual tools such as heat map, line chart and bar chart can be used to clearly show the relationship between borrower's credit score, loan amount, repayment cycle and default risk. Data visualization can also reveal the correlation between different factors, so that decision makers are not limited to a single number and table when faced with a large amount of information, but can analyze and make decisions from a global perspective. With the increasing complexity of banking business, the limitations of traditional mortgage risk assessment methods are gradually exposed.

Decision support systems (DSS) can integrate data visualization technology to automatically generate loan risk assessment reports and assist loan examiners to make scientific decisions quickly. Table 3 below illustrates the impact of different risk factors on the probability of loan default and visualizes the changing trends of default risk under these factors.

Table 3. Influences of Different Risk Factors on Loan Default Probability.

Risk factor	Default probability (%)
Credit score below 600	25
The loan amount exceeds the property valuation	15
The repayment period is more than 20 years	10
The borrower's income is insufficient to cover the payments	20

The combination of data visualization and decision support technology enables banks to achieve more accurate risk assessment and management during loan approval, thereby optimizing the loan approval process and improving the risk control ability of banks.

5. Conclusion

With the increasing complexity of banking business, the limitations of traditional mortgage risk assessment methods are gradually exposed. The introduction of data analysis technology has brought significant improvement to mortgage loan risk management. The application of technologies such as machine learning and big data analysis not only improves the accuracy of risk prediction, but also optimizes the efficiency and transparency of the decision-making process. By building a more scientific risk assessment model, banks can better identify and manage potential risks, thus ensuring the stability of the financial system.

References

1. T. Dombrowski, R. K. Pace, and J. Wang, "Imputing borrower heterogeneity and dynamics in mortgage default models," *J. Real Estate Finance Econ.*, vol. 68, no. 3, pp. 462–487, 2024, doi: 10.1007/s11146-022-09934-9.
2. Y. Ge, H. Song, and B. Li, "Bank loan strategy based on evaluation and decision model," in *J. Phys.: Conf. Ser.*, vol. 1865, no. 4, p. 042018, Apr. 2021, IOP Publishing, doi: 10.1088/1742-6596/1865/4/042018.
3. J. Duanmu, Y. Li, M. Lin, S. Tahsin, et al., "Natural disaster risk and residential mortgage lending standards," *J. Real Estate Res.*, vol. 44, no. 1, pp. 106–130, 2022, doi: 10.1080/08965803.2021.2013613.
4. B. Wiggins, *Calculating Race: Racial Discrimination in Risk Assessment*. Oxford University Press, 2020. ISBN: 9780190068722.
5. M. Iosifidi, E. Panopoulou, and C. Tsoumas, "Mortgage loan demand and banks' operational efficiency," *J. Financ. Stability*, vol. 53, p. 100851, 2021, doi: 10.1016/j.jfs.2021.100851.