ISSN 2959-6157

Research on the Influencing Factors of Home Loan Approvals

Chen Chen

School of Statistics, Southwestern University of Finance and Economics, Chengdu, 610000, China
Corresponding author: 42256080@smail.swufe.edu.cn

Abstract:

. The purpose of this paper is to comprehensively discuss the factors that may impact the approval of housing loans using Binary Logit Regression and Random Forest. As housing loans have become more common, the increase in non-performing rates of housing loans in the last two years has led to stricter loan approvals and greater uncertainty for loan applicants. This study examines data from home loan lenders to create a model for predicting an applicant's ability to obtain a home loan and to assist applicants in planning for the future. The dataset comprises 480 loan records and 12 variables. The model based on Binary Logit Regression passes the Likelihood Ratio Test with a final prediction accuracy of 81.64%, which is considered acceptable. The results indicate that the applicant's region and credit status significantly affect loan approval. Through the Random Forest, it is found that in addition to credit history, the weights for monthly applicant income and loan term are also high. Overall, applicants can predict loan approval based on the degree of influence of these factors.

Keywords: Home loan; binary logit regression; random forest.

1. Introduction

The real estate industry is a crucial sector in modern society and holds a significant position in the economies of countries worldwide. As the real estate industry expands and home prices rise, home loans are becoming increasingly common. However, loan applicants often lack knowledge of the necessary conditions for a successful housing loan application. This paper aims to examine the impact of various factors on loan approval and assist applicants in determining whether their housing loan applications can be approved.

Real estate not only fulfills people's fundamental housing needs but also serves as a vital investment and financial tool. Increasing real estate investment can promote national economic growth, drive the development of related industries such as construction, building materials, and manufacturing, create jobs, and improve the appearance of the city [1]. The real estate industry is gaining momentum in the world's economic development due to population growth and urbanization. The scale of the real estate market is increasing in countries around the world. The global real estate market reached a total value of $326.5 trillion in 2020, which represents a 5% increase compared to the previous year. The residential sector, which is the largest segment of the real estate market, experienced an 8% year-on-year increase in value, reaching approximately $258.5 trillion in 2020. This segment accounts for around 79% of global real estate value.

During the development of the real estate market in developed countries, three distinct types of development emerged: a steady development type, represented by Germany; a fluctuating development type, represented by South Korea; and a rapid development type, represented by Australia [2]. Housing system reform was initiated in China after the country's reform and opening-up period. This led to two stages of real estate system reform: the first stage from 1978 to 1997 and the second stage of comprehensive marketization of real estate from 1998 to the present [3]. The real estate industry has developed differently across various countries. However, one commonality is that its rapid growth has led to economic expansion and a surge in property prices, particularly in residential areas. According to Knoll's research, global housing prices have experienced a rapid increase since the 1960s, surpassing historical fluctuation ranges [4]. The increase in housing prices has resulted in significant pressure for the public to purchase homes. However, few individuals can afford to pay the full amount upfront, leading many to apply for a home loan.

Regarding the long-term equilibrium relationship, the elasticity coefficient between real estate prices and bank credit is 1.092. This indicates that if real estate prices increase by 1% in the long run, bank loans will also increase by 1.092% in the same direction [5]. The rise in real estate prices has led to a substantial increase in personal

mortgage business. While housing credit can boost the purchasing power of residents and increase their wealth benefits, it also raises the risk of over-indebtedness [6]. Real estate price fluctuations have an isotropic effect on the credit risk of commercial banks. However, instability in real estate prices increases the likelihood of mortgage defaults and banks' exposure to credit risk [7]. The U.S. subprime crisis was caused by the decline of its real estate market and the subsequent rise in interest rates. This led to a significant number of subprime borrowers being unable to repay their mortgages, resulting in a sudden surge in foreclosures. The subprime lending chain was impacted from the point of fracture, and the risk eventually spread throughout the country and even globally [8]. In 2022, the balance of non-performing real estate loans for 14 listed banks in China totaled 255.919 billion yuan, which represents a 72% increase. Additionally, the balance of non-performing personal housing loans increased by over 50% year-on-year [9]. In the post-epidemic era, non-performing housing loans have increased worldwide, and the personal housing loan industry is once again facing a similar challenge.

Wang developed a theoretical model to assess the risk of personal housing loan assets by analyzing the 'rent to own' approach. The rules for determining whether the risk of housing loan assets is increasing were deduced [10]. Yu utilized the decision tree technique in data mining to analyze the pre-processed personal housing loan mining dataset. The goal was to identify implicit patterns hidden within a large amount of data and ultimately develop a personal housing loan risk assessment model [11]. Qiang conducted a detailed analysis of consumer personal credit risk and established a personal credit score model for Bank X using logistic regression. This model was then used to assess the risk [12]. Due to the rising non-performing rate of housing-related loans, banks must scrutinize loan qualifications more strictly. For borrowers, the conditions for applying for housing loans and the approval of their housing applications are major concerns that affect their basic quality of life. This paper aims to construct a model using real-life data on housing loan approvals through Binary Logit Regression and further analyze the variables of interest using Random Forest. The model will analyze the impact of 11 factors, including gender, applicant income, and credit history, on the approval of housing loans.

2. Methods

2.1 Data Source

The dataset used in this paper was obtained from the website Kaggle and collected by Rushikesh Konapure at Dream Housing Finance company. It comprised 614 data sets that contain 11 factors that may impact whether an applicant receives a home loan or not. The original data was saved in CSV format.

2.2 Variable Selection

The data used in this paper contains 614 samples and 11 variables with 134 missing values. The algorithm created uses all available loan applicant information, including Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, and Property_Area, to predict loan approval (Table 1).

Table 1. List of Variables

Term	Type	Meaning
Gender	Categorical	Male/ Female
Married	Categorical	Applicant married (Y/N)
Dependents	Numeric	Number of dependents
Education	Categorical	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Categorical	Self-employed (Y/N)
ApplicantIncome	Numeric	Applicant's income
CoapplicantIncome	Numeric	Co-applicant income
Loan_Amount	Numeric	Loan amounts in thousands
Loan_Amount_Term	Numeric	Term of the loan in months
Credit_History	Categorical	Credit history meets guidelines
Property_Area	Categorical	Urban/ Semi Urban/ Rural
Loan_Status	Categorical	Loan approved (Y/N)

2.3 Method Introduction

The method used in this paper is Binary Logit Regression and Random Forest. The logit regression model can be represented as:

$$P = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m)} \quad (1)$$

Where β_0 is a constant term, $\beta_1, \beta_2, \cdots, \beta_m$ are partial regression coefficients. Perform a logit transformation on $f(x) = \frac{1}{1+e^{-x}}$, then $L(p) = ln\frac{p}{1-p}$. The logit regression model can be expressed in the following linear form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (2)$$

2.4 Data Preprocessing

The analysis includes 614 samples, but 134 missing data points are eliminated. To enhance the study, the data are pre-processed. After removing missing values, the sample size is reduced to 480. Dummy variables are used to represent the qualitative variables, and symbols are assigned to variables (Table 2).

Table 2. List of Variables after Data Preprocessing

Term	Symbol	Range
Gender	x_1	0-Male, 1-Female
Married	x_2	0-NO, 1-Yes
Dependents	x_3	0 to 3 persons
Education	x_4	0-Under Graduate, 1-Graduate
Self_Employed	x_5	0-NO, 1-Yes
ApplicantIncome	x_6	150 to 81,000 dollars per month
CoapplicantIncome	x_7	0 to 33,837 dollars per month
Loan_Amount	x_8	9 to 600 months
Loan_Amount_Term	x_9	36 to 480 thousand dollars
Credit_History	x_{10}	0-No, 1-Yes
Property_Area	x_{11}	0-Semiurban, 1-Urban, 2-Rural
Loan_Status	Y	0-NO, 1-Yes

3. Results and Discussion

3.1 Basic Information

This paper uses Loan_Status as the dependent variable and the other variables as independent variables. The table above indicates the participation of 480 samples in the analysis, revealing the absence of any missing data (Table 3).

Table 3. Overview of Analysis

Name	Options	Frequency	Percentage
Loan_Status	0	148	30.83%
	1	332	69.17%
	Total	480	100.0%
Summary	Hiatus	480	100.00%
	Total	0	0.00%
	Hiatus	480	100.0%

3.2 Descriptive Analysis

As the histogram for Loan_Amount is right-skewed, the logarithm of the value is taken and the corresponding histogram is drawn. The maximum frequency for Loan_Amount falls between 2 and 2.25. Additionally, it is observed that more people opt for long-term loans than short-term ones (Figure 1).

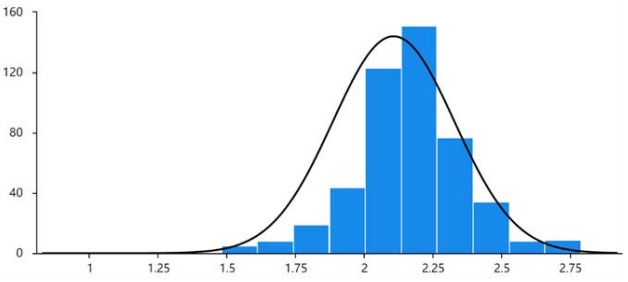

Fig. 1 The Histogram of Loan_Amount

A box-and-whisker plot is created to display ApplicantIncome categorized by approved and not-approved loans. The plots show no statistically significant difference, suggesting that ApplicantIncome may not be a significant factor in home loan approval (Figure 2).

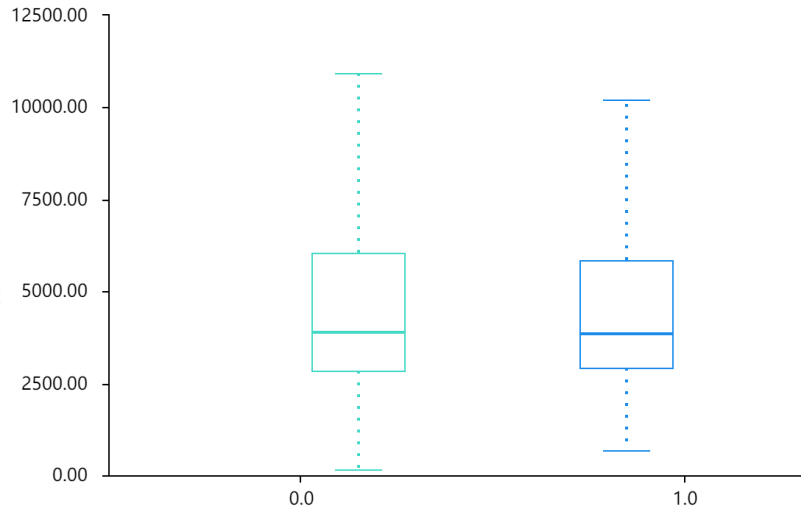

Fig. 2 Box-and-whisker Plot of ApplicantIncome

3.3 Binary Logit Regression Results

From Table 4, it can be seen that the formula for the model is:

$$ln\left(\frac{p}{1-p}\right) = -1.822 - 0.344x_1 + 0.510x_2 + \cdots + 3.163x_{10} - 0.468x_{11} \quad (3)$$

The table shows that $x_{10}\,(z = 8.451, p = 0.000 < 0.05)$, $x_{11}(z = -3.199, p = 0.001 < 0.05)$ presents a significance level of 0.05, which means that they have a significant effect on Y. The value of the regression coefficient for x_{10} is 3.613, and its OR value is 37.059, which means that when x_1 increases by one unit, the change (increase) in Loan_Status is 37.059 times. The value of the regression coefficient for x_{11} is -0.468, and its OR value is 0.626, which means that when x_2 increases by one unit, the change (decrease) in Loan_Status is 0.626 times (Table 4).

Table 4. Results of Binary Logit Regression Analysis

Sum	Coefficient	Error	z	Wald $\chi 2$	p	OR	ORprice95% CI
x_1	-0.344	0.327	-1.051	1.106	0.293	0.709	0.374 ~ 1.346
x_2	0.510	0.284	1.798	3.233	0.072	1.666	0.955 ~ 2.905
x_3	0.081	0.133	0.608	0.369	0.543	1.084	0.835 ~ 1.408
x_4	0.381	0.300	1.271	1.616	0.204	1.464	0.813 ~ 2.637
x_5	-0.132	0.349	-0.378	0.143	0.706	0.877	0.442 ~ 1.737
x_6	0.000	0.000	0.147	0.022	0.883	1.000	1.000 ~ 1.000
x_7	-0.000	0.000	-1.180	1.392	0.238	1.000	1.000 ~ 1.000
x_8	-0.003	0.002	-1.432	2.052	0.152	0.997	0.994 ~ 1.001
x_9	-0.000	0.002	-0.217	0.047	0.828	1.000	0.996 ~ 1.003
x_{10}	3.613	0.427	8.451	71.419	0.000	37.059	16.034 ~ 85.655
x_{11}	-0.468	0.146	-3.199	10.236	0.001	0.626	0.470 ~ 0.834
Intercept	-1.822	0.850	-2.144	4.597	0.032	0.162	0.031 ~ 0.855

Based on the model's prediction accuracy, it is evident from the table above that the overall accuracy of the research model is 81.46%, indicating an acceptable model fit. The accuracy of predicting a loan not being approved is 44.59%, while the accuracy of predicting a loan being approved is 97.89% (Table 5).

Table 5. Prediction Accuracy with Binary Logit Regression

0		Predicted Value		Forecast Accuracy
		1		
True Value	0	66	82	44.59%
	1	8	324	97.59%
Gather				81.25%

Linear regression calculations are performed and combined with VIF values to determine the presence of covariance. If the VIF value exceeds 10 or the tolerance value is less than 0.1, covariance is present. The table above indicates that there is no covariance issue in this model (Table 6).

The original hypothesis of the model test is that the model quality is the same in both cases, whether or not the independent variables. The p-value is less than 0.05, indicating that the original hypothesis is rejected. This model construction is meaningful because the independent variables used have validity (Table 7).

Table 6. Results of Covariance Diagnostics

Sum	VIF	Tolerance
x_1	1.195	0.837
x_2	1.331	0.751
x_3	1.222	0.818
x_4	1.064	0.94
x_5	1.036	0.966
x_6	1.455	0.687
x_7	1.135	0.881
x_8	1.526	0.655
x_9	1.037	0.964
x_{10}	1.013	0.987
x_{11}	1.023	0.978

Table 7. Results of Likelihood Ratio Test

Model	-2-fold Log-likelihood	chi-square Value	df	p	AIC	BIC
Intercept Only	593.050	-	-	-	-	-
Final Model	440.052	152.999	11	0.000	464.052	514.137

3.4 Random Forest Results

The analysis is repeated using random forests as there are fewer variables with significant effects in the model based on binary logit regression. For random forest modeling, the dependent variable used is Loan_Status, the training set ratio is set to 0.8.

The table 8 shows that x_{10} has the highest weight at 26.2% which plays a key role in model construction. Additionally, x_6 accounts for 19.4%, and x_8 accounts for 18.9%. Together, these three features account for 64.5% of the weight (Table 8).

Based on the table above, the final model achieved an accuracy of 0.72 on the test set, with a combined precision of 0.71, a combined recall of 0.72, and a combined f1-score of 0.69. It appears that the model is relatively ineffective. In general, the binary logit regression model performs better (Table 9).

Table 8. Values of Feature Weight

Term	Symbol	Weight
Gender	x_1	0.017
Married	x_2	0.032
Dependents	x_3	0.047
Education	x_4	0.025
Self_Employed	x_5	0.019
ApplicantIncome	x_6	0.194
CoapplicantIncome	x_7	0.123
LoanAmount	x_8	0.189
Loan_Amount_Term	x_9	0.044
Credit_History	x_{10}	0.262
Property_Area	x_{11}	0.049

Table 9. Evaluation Results of the Test Set Model

Term	Accuracy	Recall Rate	f1-score	Sample Size
0	0.69	0.33	0.45	33
1	0.72	0.92	0.81	63
Accuracy	-	-	0.72	96
Average	0.71	0.63	0.63	96
Average (combined)	0.71	0.72	0.69	96

4. Conclusion

In this paper, 480 valid samples are selected with Loan_ Status as the dependent variable and the other variables as independent variables. This study analyzes the factors affecting loan approval using binary logit regression and random forest. After analyzing the sample data, this paper obtains a series of statistical results. Based on binary logit regression, the findings indicate that applicants from semi-urban areas are more likely to obtain housing loans compared to those from urban and rural areas. Additionally, meeting the credit history requirements has a significant impact on the success of loan applications. Based on Random Forest, loan approval outcomes are significantly influenced by credit history, applicant monthly income, and loan tenure. This study aims to assist loan applicants in evaluating their eligibility for a housing loan and planning for their future. However, it is important to note that the study has limitations due to a small sample size and a lack of independent variables. Future studies should address these limitations by collecting relevant data on an ongoing basis and exploring additional factors that may impact loan approval.

References

[1] Li Qiming. On the Relationship between China's Real Estate Industry and the National Economy. China Real Estate, 2002, (6): 4.

[2] Lv Tingui, Wang Yawen. Characteristics of real estate market development history in developed countries and its revelation - A comparative study based on Australia, South Korea, and Germany. China Real Estate, 2019, 12: 20-25.

[3] Cao Y. China's real estate development history. Enterprise technology development, 2014, 33(11): 110-111.

[4] Guo Kesha, Huang Yanyan. Problems and way out of China's real estate market development from international comparison. Finance and trade economy, 2018, 39(01): 5-22.

[5] Duan Zhongdong, Zeng Linghua, Huang Zexian. An empirical study of real estate price volatility and bank credit growth. Financial Forum, 2007, 2: 40-45.

[6] Xie Jialin. Analysis of the impact of housing credit on household consumption behavior. Journal of Economic Research, 2023, 19: 59-61.

[7] Zhao Yan, HAN Ning. Empirical analysis of the impact of real estate price fluctuation on the credit risk of China's commercial banks. Times Economy and Trade, 2017, 31: 6-8.

[8] Du Houwen, Chu Chunli. The U.S. Subprime Crisis: Roots, Trends, and Impact. Journal of Renmin University of China, 2008, 1: 49-57.

[9] Xu Qian. Banks' non-performing rate of housing-related loans rises due to real estate downturn. China Real Estate News, 2023.

[10] Wang Jianfeng. Lenders' Adverse Selection Behavior and Commercial Banks' Credit Risks: A Theoretical Model and Empirical Analysis of the Housing Credit Asset Risks Affected by "Rent to Own". Financial Research, 2003, 11: 21-27.

[11] Yu Zhuo. Application of decision tree to construct a risk assessment model for personal housing loans. Northeast University of Finance and Economics, 2008.

[12] Gan Q. Identification and early warning of personal consumption credit risk. Zhejiang University, 2019.