Analysis of Overdue Loans from Customers and Factors Affecting Banks' Recovery Strategies

Xuening Bai[1,a,*]

[1]*Department of Art and Science, University of Toronto, Toronto, M5S 1A1, Canada*
a.shirley000510@gmail.com
**corresponding author*

Abstract: Banks still want to recover part of overdue loans, and choose a corresponding debt collection strategy for each customer based on some factors. The article uses cross-sectional analysis and nonparametric tests to determine the factors that influence banks' recovery of loans. From the analysis, recovery strategies and sex are independent of each other. Loan recovery ratios are related to the cost of debt collection strategies. A binary logistic regression model concludes that gender, age and expected recovery amount did not influence the choice of debt collection strategy.

Keywords: non-performing loan, recovery strategy, cross-table, comparative analysis, Kolmogorov-Smirnov test, non-parametric test, binary logistic regression

1. Introduction

Banks play an important role in people's lives and have an interdependent relationship. People can earn profits by depositing money they do not need on a daily basis, or they can obtain funds from banks by taking out loans to relieve financial pressure. On the other hand, banks can lend money to those who need it. Not only do they earn interest income, but they also help banks diversify their risk. However, all investments carry risks. Loans that cannot be collected, often referred to as non-performing assets, can significantly impact banks. At the end of 2015, banks in the Eastern Caribbean Currency Union (ECCU) had 17% of their total loans as non-performing loans, well above the region's 5% limit line, resulting in a significant deterioration in their profitability or even insolvency [1]. Since the loans released to borrowers are savings that others have structured within the bank, if they cannot be collected back home, the damage is to the depositors' money and over time the bank's operations can become problematic. According to the Basel Accord, commercial banks should make different provisions for different bad loans to cope with the risk [2]. In this process, banks need to take up a lot of monetary capital, making less money circulating in society. When a debt is legally declared "uncollectible" by the bank [3], the account is considered "written off", but this does not mean that the bank is free from the debt. Banks still want to collect the money owed by some customers. They score the account to assess the expected recovery amount, which is the amount the bank is likely to receive from the customer in the future. Different strategies are used depending on the threshold [4].

Contrastive analysis and logistic regression models are widely used in the quantitative analysis of bank-to-bank strategies. Bingxin Wang used contrastive analysis to conduct separate longitudinal and cross-sectional analyses comparing the consumption of the same comparable subject with different characteristics in different situations over time and comparative analysis of different consumption conditions and characteristics over the same period [5].

This article analyzes the effectiveness of banks' debt collection strategies for different borrowers and the factors that influence different debt collection strategies. A Comparative Analysis is used to conclude whether different thresholds used by banks to adopt different strategies are effective. A Binary Logistic Regression model is used to predict the factors that influence the banks' use of different debt collection strategies. We find that the proportion of loans recovered by banks (actual amount recovered divided by the expected amount recovered) is distributed differently across different recovery strategies, indicating that the choice of strategy affects the actual amount recovered. In addition, the proportion of recovered loans is less pronounced across gender, responding that the actual amount recovered is not related to the gender of the customers. Age, gender and expected recovery amount influence less on the choice of recovery strategies.

The paper is organized as follows: Chapter 2 preprocess the original data and created new variables with research significance; Chapter 3 introduces the preliminaries of the Comparative Analysis such as the Cross-table, Kolmogorov-Smirnov test to check normality, and Nonparametric test; Chapter 4 further study factors influence the chosen of recovery strategies using Binary Logistic Regression; Chapter 5 concludes the paper and provides suggestions.

2. Preprocess Data

The dataset about bank lending includes more than 1800 data and 6 variables which are customer ID, expected recovery in dollars, actual recovery in dollars, different levels of the recovery strategy, age of agent, and gender. Since the bank has different expected and actual recovery amounts for each customer, it is very curious whether the final recovery amount matches the implemented strategy. In other words, whether the amount the bank wants to recover is worth using the corresponding strategy. Therefore, two new variables were created: repaid ratio and adjusted ratio. The repaid ratio (actual recovery divided by expected recovery) represents the actual recovery amount under different strategies. Since different recovery strategies have different costs, e.g. Level 0 recovery costs $0, Level 1 recovery costs $50, Level 2 recovery costs $100, etc., when the costs of the different recovery strategies are subtracted from the actual recovery amount, the new actual recovery amount divided by the expected recovery is the adjusted ratio. In addition, based on the original Level of recovery strategies, they are regrouped into two categories: the first category includes Level 0 to Level 2 recovery strategies, and the other category includes Level 3 and Level 4 recovery strategies. The minimum, maximum, and mean values of numeric type variables and the number of each category of categorical variables are presented in table 1.

Table 1: Description of variables.

Numerical Variables in the Dataset			
Variables	Minimum	Maximum	Mean
Expected Recovery in Dollars	194	9964	2795.97
Actual Recovery in Dollars	200.43	34398.48	4000.97
Age	18	84	39.65
Repaid ratio	0.07	6.25	1.18

Table 1: (continued)

Numerical Variables in the Dataset			
Variables	Minimum	Maximum	Mean
Adjusted Ratio	0.05	6.22	1.15
Categorical Variables in the Dataset			
Variables	Count		
Sex	Female: 909, Male: 973		
Recovery Strategy	Level 0: 247, Level 1: 670, Level 2: 670, Level 3: 368, Level 4: 364		

3. Comparative Analysis

For every overdue loan, even if full recovery is unlikely, the bank still wants to get some of it back to minimize losses. Therefore, the bank will estimate the expected recovery amount based on the borrower's past economic situation and characteristics and choose the corresponding collection policy based on that. In this way, the cross-tabulation is selected to explore whether gender, as an influencing factor, is related to the choice of collection policy.

3.1. Testing for Discrete Variables

The cross-tabulation shows the joint frequency of data values based on two or more categorical variables [6]. The Pearson Chi-square test is combined to determine whether the results of the cross-tabulation are statistically significant [7]. In other words, whether the two categorical variables are independent of each other. If the p-value is equal to or less than the statistical threshold (usually 0.05), the result is significant. That is, we reject the null hypothesis that the two variables are independent of each other.

Specifically, the null hypothesis is that there was no association between gender and the level of the recovery strategy. If the difference between observed and expected counts within each group is greater, the higher the chi-square score, the more likely it is to be significant, which makes it more likely that we will reject the null hypothesis and conclude that the variables are correlated [8]. As seen from the cross-tabulation, under the null hypothesis (variables are independent) being true, the number of women or men in the first group of recovery strategy is very close to the expected number, and the other group has the same trend. The p-value of the Pearson Chi-square is 0.294 (see table 2), which is larger than the statistic threshold (0.05). In this way, we would accept the null hypothesis that asserts the two variables are independent of each other. Meaning that banks do not differentiate in their choice of strategy based on the gender of the customer. The choice of strategy is a cost to the bank, but there may be subtle differences in the methods used because there are still differences in the thinking of men and women.

Table 2: The result of Chi-square Test.

Chi-square Test	
	Significance
Pearson Chi-square	0.294

3.2. Testing for Continuous Variables

The repaid ratio most directly reflects the amount the bank recovered versus the amount it expected after using the recovery strategy. A repaid ratio greater than 1 indicates that the amount recovered from the customer is greater than the bank expected, and less than 1 indicates that the amount actually recovered by the bank is less than expected. When subtracting the bank's cost of the recovery strategy, an adjusted ratio greater than 1 indicates that the amount recovered from the customer is greater than the bank's expectation and that the recovery strategy is effective, while a ratio less than 1 indicates that the effectiveness of the recovery strategy is questionable. We use the repaid ratio and the adjusted ratio to compare the normality of the distributions in the low-cost strategy and the high-cost strategy.

For continuous variables, we use the Kolmogorov-Smirnov test [9] to check the normality of the two groups and focus on the magnitude of the p-value to determine whether the non-parametric test [10] analysis is needed to assist in the determination.

Specifically, the null hypothesis is that the repaid ratio and adjusted ratio are normally distributed in the two types of recovery strategies, respectively. In the normality test, the Kolmogorov-Smirnov test shows that the p-values of repaid ratio and adjusted ratio in two groups of recovery strategies are less than 0.05, which means that the null hypothesis is rejected. In other words, the repaid and adjusted ratios are not normally distributed in the two groups of recovery strategies, respectively. Based on this condition, a Non-parametric test is used. In this case, the null hypothesis is that the distribution of the repaid ratio and the adjusted ratio is consistent in the two groups of the recovery strategy, respectively. By independent-samples Mann-Whitney U test, we found that both p-values are less than 0.05 implying rejection of the null hypothesis. Therefore, the distribution of repaid ratio or adjusted ratio is different across two general levels of the recovery strategy.

Looking at the descriptives of the normality test (see Table 3), the mean value of the repaid ratio is less than 1 for the low-cost recovery strategy, indicating that the actual recovery amount is less than the expected recovery. However, the mean of the repaid ratio is greater than 1 for hight cost recovery strategies, indicating that actual recoveries are greater than expected recoveries. From the results, it is easy to see that the bank will get a higher return for spending more money to collect the loan. If the bank's collection costs are removed from the actual recovery amount, which is the bank's net recovery amount, the adjusted ratio has the same trend as the repaid ratio in both categories of recovery strategies. Banks will use higher-rated and more costly collection strategies for customers who are more delinquent on their loans. For delinquencies under $1,000, the bank will not spend extra to collect and may choose to remind the customer by email, text message, phone call, etc. If it is over $1,000, the bank will spend an additional $50 for each additional $1,000 as a watershed to urge the customer to pay the debt. The highest collection strategy will be used when the loan is over $5000, which means the bank will spend $200. The bank will use the additional collection cost to hire a third-party collection team to assist in getting the loan back. This is more in line with the results of our experiment, where the higher the cost, the more the bank will recover the loan.

Table 3: The describe of ratios in different strategies.

Descriptive			
	Recovery strategies	Mean	Median
Repaid Ratio	Low-cost	0.925	0.869
	High-cost	1.688	1.579
Adjusted Ratio	Low-cost	0.895	0.839
	High-cost	1.652	1.545

Repeating the methods used above (Kolmogorov-Smirnov test and Non-parametric test), we wanted to investigate whether the repaid ratio and adjusted ratio were normally distributed across genders. In this way, the null hypothesis is that the repaid ratio and adjusted ratio were normally distributed across genders. The Kolmogorov-Smirnov test showed that the p-values of repaid ratio and adjusted ratio were less than 0.05 across genders, which means that the null hypothesis was rejected. Therefore, the Non-parametric test was used to continue the comparison for both variables across gender. The null hypothesis at this point was that the distribution of repaid ratio or adjusted ratio was consistent across genders. The p-value of both sets of variables in the Hypothesis test summary is greater than 0.05, indicating that the results do not reject the null hypothesis. In other words, the distribution of repaid ratio and adjusted ratio is consistent across genders. Recalling the Descriptives table 4 of the normality test, the mean and median of the repaid ratio are very close across genders, while the adjusted ratio also has a similar trend. This also aids in verifying the accuracy of the findings of the nonparametric test. Therefore, gender does not have a significant effect on the bank's repossession of loans.

Table 4: Hypothesis Test on ratios across sex.

Hypothesis Test Summary		
No.	Null Hypothesis	Significance
1	The distribution of repaid ratio is the same across categories of sex	0.095
2	The distribution of adjusted ratio is the same across categories of sex	0.094

4. Binary Logistic Regression

The collection strategy used by the bank for each customer can be determined by considering various aspects such as gender, age, amount owed, number of days past due, etc. By analyzing their level of influence on the collection strategy, the most appropriate and effective method is matched. Not only does this increase the success rate, but it also saves costs. Specifically, the variables in our dataset are based on the study of which affect the choice of debt collection strategy. Since debt collection strategy is a two-class variable (low-cost strategy and high-cost strategy), a binary logistic regression model is most suitable for analyzing this problem [11]. T hus the dependent variable (y) is debt collection strategy and the independent variables (x) are gender, age, and late recovery of loans.

Tsai et al. have modeled the prediction of consumer loan delinquency for the analysis of unsecured consumer loan customers. The model uses borrowers' demographic variables (such as gender, age, education, monthly income, etc.) and money attitude as discriminatory information, and the money attitude increases the predictive power of the model. Therefore, it is reasonable to suspect that the variables that may influence the choice of collection strategy after a loan is overdue include gender and age. The overdue recovered loan is the amount predicted by the bank based on the customer's probability of payment, total debt and willingness to pay, in other words, it includes the customer's current economic status and money attitude, so it is possible to try it as the third variable. From Table 5, we can see that the p-value of each variable is less than the statistical bound of 0.05, which means that they have little influence on debt collection strategies, which means that the variables that influence debt collection strategies are not put into the model. Banks can consider other variables other than gender, age and late collection of loans as references when choosing debt collection strategies for delinquent customers.

Table 5: The variables in the model.

Variables in the Model	
Variables	Significance
Age	0.982
Sex	0.909
Expected recovery amount	0.611

5. Conclusion

The bank still wants to recover part of the loan to reduce the loss of the bank, so the choice of debt collection strategy is very important because the effect of using different cost strategies will be different. The results of the cross-tabulation found that gender and debt collection strategies were independent of each other, indicating that gender had no direct impact on debt collection strategies. Using the Non-parametric test, the distribution of repaid ratio or adjusted ratio is different across high-cost and low-cost recovery strategies, and the higher the cost the bank spends, the more likely it is to recover more loans. However, the distribution of repaid ratio or adjusted ratio is the same among different genders, which means that gender does not affect the loan recovery amount. Finally, through binary logistic regression model, it is concluded that gender, age and expected loan recovery have little effect on the choice of debt collection strategy. In general, banks use debt collection strategies to recover unpaid loans from customers more effectively, and the cost of debt collection invested by banks is proportional to the effect. However, factors that influence the choice of strategy are other than age, gender and expected loan recovery, such as monthly salary, other loans, credit history, etc. If we want to study the choice of debt collection strategy more accurately, we need to know and mobile more customer information and economic situation to make the best choice.

References

[1] *Beaton, Ms Kimberly, Ms Alla Myrvoda, Shernnel Thompson. 2016 Non-performing loans in the ECCU: Determinants and macroeconomic impact. International Monetary Fund.*

[2] *BCBS. 2017 Basel III: International Regulatory Framework for Banks. BIS, 7 Dec.www.bis.org/bcbs/basel3.htm.*

[3] *Chang R D, Shen W H, Fang C J. 2008 Discretionary loan loss provisions and earnings management for the banking industry[J] International Business & Economics Research Journal (IBER) 7 (3).*

[4] *Migwi, James M. 2013 Credit Monitoring and recovery strategies adopted by Commecial Banks in Kenya. Diss. University of Nairobi.*

[5] *Wu B X. 2011 Consumption and management: New discovery and applications. Elsevier.*

[6] *Kamakura, Wagner A., Michel W. 1997 Statistical data fusion for cross-tabulation[J] Journal of Marketing Research 34 (4) 485-498.*

[7] *McHugh, Mary L. 2013 The chi-square test of independence Biochemia medica 23 (2) 143-149.*

[8] *Charles P. 1967 A Framework for the Comparative Analysis of Organizations[J] American Sociological Review 32 (2) 194–208.*

[9] *Massey Jr, Frank J. 1951 The Kolmogorov-Smirnov test for goodness of fit[J] Journal of the American statistical Association 46 (253) 68-78.*

[10] *Hoeffding, Wassily. 1994 A non-parametric test of independence. The Collected Works of Wassily Hoeffding. Springer, New York, NY 214-226.*

[11] *Harrell, Frank E. 2015 Binary logistic regression. Regression modeling strategies. Springer, Cham 219-274.*