



Analysis of Trends of Purchasers of Motorcycles in Latin America

Rintaro Tanabe^{1(✉)} and Yumi Asahi²

¹ School of Information and Telecommunication Engineering, Course of Information Telecommunication Engineering, Tokai University, Tokyo, Japan

5bjml2l4@mail.u-tokai.ac.jp

² School of Information and Telecommunication Engineering, Department of Management System Engineering, Tokai University, Tokyo, Japan

asahi@tsc.u-tokai.ac.jp

Abstract. The Latin American economy experienced the currency crisis and the associated confusion from the early 1990s through the early 2000s. Since 2003, rapid economic growth has been achieved. As a result, in Latin America “A” country, the impact of external demand led to the expansion of the consumer finance market. Furthermore, financial services expanded due to income disparity correction policy implemented from 2003 to 2010. By these, purchasers due to loans increased of motorcycle and automobile, but on the other hand rate of loans outstanding increased. In this research, we look for factors of loans outstanding from customer data. The data used in this study is customer data of anonymized motorcycles in Latin America “A” country from September 2010 to June 2012. From the usage data it turns out that the proportion of loans standing is high. Therefore, it is necessary to extract variables that are factors of loans outstanding. From there, it is necessary to grasp the characteristics of loans outstanding. The analysis flow is data cleaning, basic aggregation, grouping of data, variable extraction, binomial logistic regression analysis. Data is organized by data cleaning. Data was grouped by income amount by grouping of data. Basic aggregation allows to determine the characteristics of the data. Next, we extract the variables that cause the factor of loans outstanding by AUC. Finally, binomial logistic regression analysis finds out how the variables extracted by AUC affect loans outstanding. In addition, analysis results and Beforehand studies have considered that specific variables greatly affect loans outstanding. Therefore, this studies deeply dig up that variable. Based on the results of the analysis, we explore the tendency of loans outstanding.

Keywords: Motorcycle · Customer data · Loans outstanding within 18 months

1 Introduction

The Latin American economy experienced the currency crisis and its associated confusion from the early 1990s through the early 2000s, so the GDP growth rate in 1990–2002 was sluggish [1]. However, vigorous capital investment is expanded because of the global economic expansion and the rise in primary commodity prices, expansion of exports and inflow of investment funds since the 2000s. In addition, it has achieved

rapid economic growth due to the expansion of personal consumption since 2003. Therefore, In the Latin American “A” country, the expansion of exports and the influence of external demand due to the rise in primary commodity prices led to the expansion of the consumer finance market. In addition, the financial service expanded to the people who cannot take out a loan due to raising the poor to the middle class by income disparity correction policy implemented from 2003 to 2010 [2].

From the above, although the purchases increased by loans of motorcycles and cars [3], there were many customers who did not understand the contract contents of loans [4], and the rate of bad debts due to excessive debt consumption has raised [2].

In this research, we look for factors of bad debt from customer data.

2 The Data Overview

The data used in this study is anonymized customer data of motorcycles in Latin America “A” country from September 2010 to June 2012. The data is composed of the score of the credit agency A, the score of the credit agency B, the history of the tax, payment, the working year, the married/unmarried, the sex, the working state, the resident state, the main income, the side income, the dealer assessment, the division number, the down payment, the loan, interest amount, occupation, academic background, house type, region, product type, displacement, size, 6 months Bad, 12 months Bad, 18 months Bad, and so on.

(6 months Bad, 12-month Bad, and 18-month Bad mean that if it doesn’t reach the price that the customer has to pay until the limits after purchasing, they will be checked. Therefore, a person who is eligible for Bad for 6 months Bad will be checked for 12-month Bad, and 18-month Bad. It has never returned).

The number of customers was 14,304 in Latin America “A” country.

3 The Research Purpose

From the definition in the usage data summary, 18 months Bad customers are the most number in 6, 12, and 18 months Bad. From Fig. 1, it can be assumed the serious situation because about 20% of customers of all data are with Bad for 18 months. Due to the usage data contains many data items, it is necessary to extract variables that become a core of Bad for 18 months. In this study, we extract the variables which influence precisely 18 months bad AUC and analyze the influence of data extracted by logistic regression analysis. Based on the results, we grasp the characteristics of 18 months Bad customers. AUC is defined as the region below the ROC curve and it is an index for measuring the accuracy of the model. Therefore, AUC can be judged that the larger the numerical value of the region, the higher the accuracy.

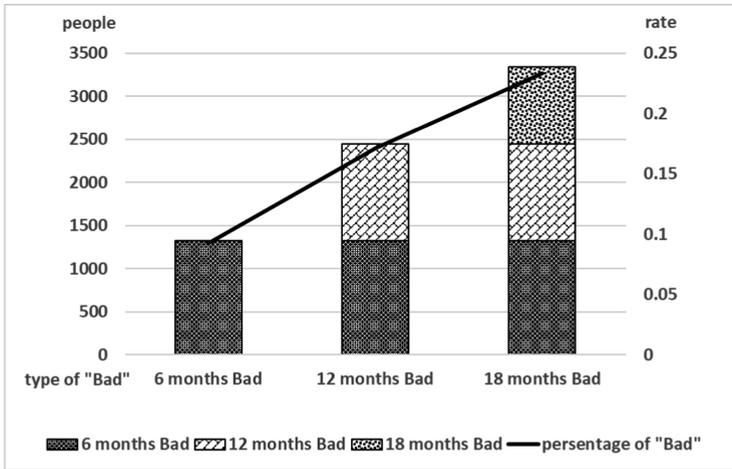


Fig. 1. Per 6,12,18 months “Bad” transition of proportion and total

4 The Analysis

The flow of analysis is performed by following procedures; data cleaning, basic tabulation, grouping, AUC, and logistic regression analysis.

At first, we supplement missing data and remove with data cleaning. Next, the data trend is grasped by basic aggregation. Then we divide into some group by the amount of main income. In AUC, the variable affecting 18 months Bad is extracted by the area under the ROC curve. Finally, we examine how the variable extracted by AUC has an impact on 18 months Bad.

4.1 The Data Cleaning

At first, we complement the missing data. As there was no blank data on the score of credit agency B, we complemented the score of credit agency A and its main income by using the score of credit agency B. In addition, we removed the interest, the down payment, the borrowing money, the age, the blank data because of the variable after complementing the credit agency A and its income, and the customer data which is impossible to calculate. As the result, the number of customer data is 13,217.

We will explain the calculation method about the score of credit agency A. We use the score of credit agency B of the customer who has a blank in the score of credit agency A and calculate the average the score of both credit agency A and B. After that, we input the calculated the score of credit agency A in the blank data. Then we supplement the main income with the score of credit agency A as well.

4.2 The Basic Aggregate

According to Fig. 2, the percentage of Bad in the Midwest, Northeast, and Northern is high. In addition, the average of main income is lower in the regions with the higher

rate of Bad. Next, we look at the trends in educational backgrounds and types of occupations that are relevant to “Bad” customer’s main income by region. View from Figs. 3 and 4. As a result, the same tendency was seen in all areas. The proportion of “Salary earners” in the classification of occupation and The proportion of “Graduated from Educational background 3” in the academic record was found to be large.

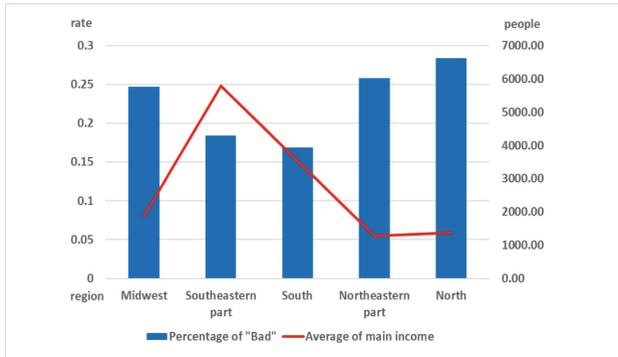


Fig. 2. By region percentage of main income and “Bad” ratio

In addition, in the prior research, the score of credit agency A and the score of credit agency B was grouping done. The ratio of 18 months Bad was calculated using them, and as a result, Fig. 5 was listed. According to the score of the credit agency A, the lower the score value, the clearly the higher the proportion of 18 months Bad. Looking at the score of credit agency B, the lower the percentage of the score, the higher the proportion of 18 months Bad is, but the groups 0, 1, 2 are not cleaning.

4.3 The Grouping

It can be seen that there is a clear difference between the ratio of Bad in the north and the south. Moreover, it is considered that the bad factor is influenced by the main income, so we categorize groups by the main income for each revenue amount. The grouping criterion is classified by income which is published by the Ministry of Economy, Trade and Industry.

The A/B stratum get 7,475 or more a per month, the C stratum gets 1,734 to less than 7,435 a per month, D stratum gets less than 1,085 to 1,734 a per month, and E stratum gets less than 1,085 a per month.

Therefore, the A/B stratum is an affluent class, the C stratum is an intermediate class, and the D/E stratum is a poor class. We classified based on this criterion. From Table 3, it can be seen that the proportion of “Bad” in poor D and E stratum is high.

There is also a reason for summarizing the A/B stratum. There is also a reason for summarizing the A/B stratum. According to Fig. 6 the A/B stratum is a small part of the whole, and even this data is small number value [3] (Table 1).

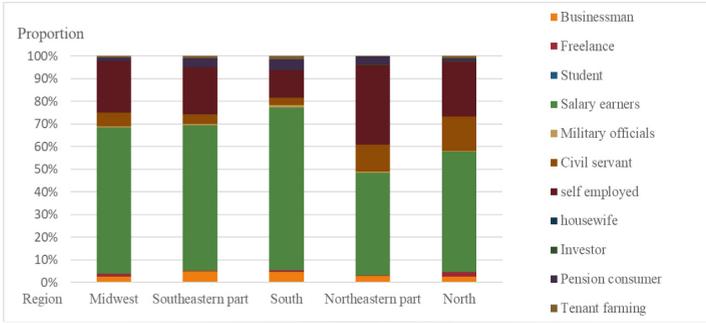


Fig. 3. Percentage of type of occupation by region

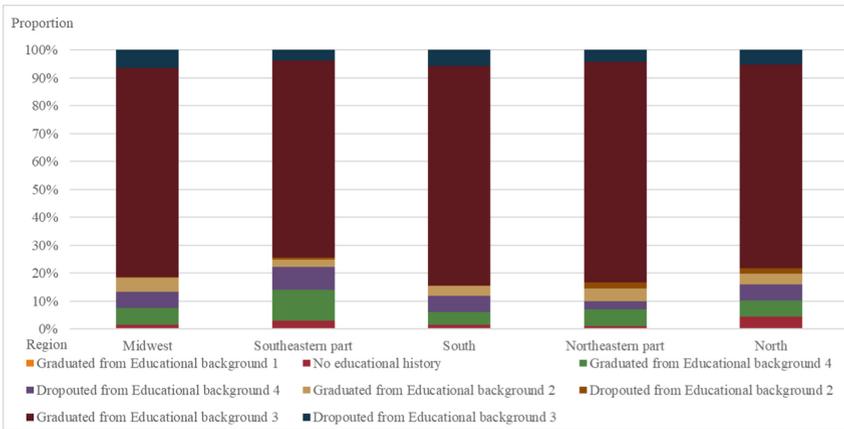


Fig. 4. Percentage of type of educational background by region variable.



Fig. 5. Percentage of “18 months Bad” by group of credit agency

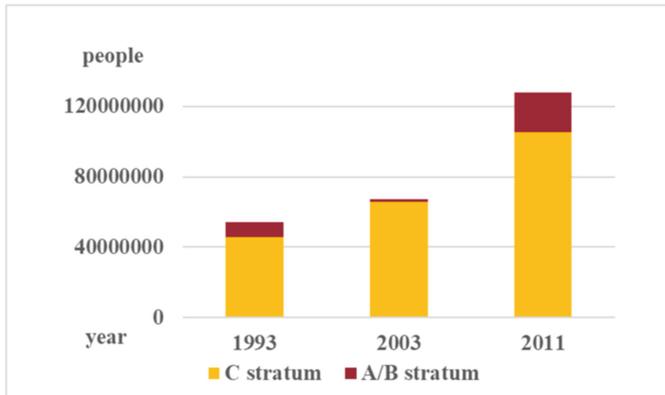


Fig. 6. Population transition by stratum

Table 1. Number of data by income stratum and percentage of 18 months Bad

	A/B stratum	C stratum	D stratum	E stratum
The number of data	147	2425	5516	5216
18 months “Bad”	13%	15%	27%	22%

4.4 The Extracting Variables

As the data of 18 months after cleaning is an objective variable and other variables are explanatory variables, we analyze the explanatory variables by logistic regression analysis and obtain the predicted probability. After that, we obtain the area below the ROC curve based on the explanatory variables. As a result, the presence/absence of negative information in the customer list, the number of inquiries to the customer list, the age, the value of real estate, the rate of list price, the borrowing money, the interest amount, the score of credit agency A, the score of credit agency B, product A, product B, D stratum were adopted. Also, from the results of AUC, the score of credit agency A is 0.688, and the score of credit agency B is 0.594. From here the credit agency A’s score is strong relationship to 18 months Bad, indicating higher credibility (Table 2).

4.5 Logistic Regression Analysis

We analyze the logistic regression analysis by using the adapted variables in 4.4. The logistics regression analysis is a way to predict occurrence probability. Based on the analysis, we judge the occurrence probability of Bad customers. According to Fig. 2, the variable whose Exp (B) value is 1 or more is the number of inquiries to the customer list, the ratio of the list price, the product B, and D stratum. The customers who apply to these variables tend to be “Bad” easily. And also, the variable whose Exp (B) value is less than 1 is the presence or absence of negative information of the customer list, and product A. The customers who apply to these variables might not tend to be “Bad” easily. In addition, the higher the score of the data, the easier to be

Table 2. Analysis result of ROC curve

Variables adopted	Area	Significance probability
With or without of negative information on customer list	.554	.000
Number of Inquiries to customer lists	.553	.000
Age	.566	.000
Real estate value	.572	.000
Ratio to list price	.567	.000
Debt	.578	.000
Interest amount	.570	.000
Credit agency A score	.688	.000
Credit agency B score	.594	.000
Product A	.564	.000
Product B	.578	.000
D stratum	.552	.000

Table 3. Analysis result of logistic regression analysis

Variables	Exp(B)	Significance probability
With or without of negative information on customer list	.831	.009
Number of inquiries to customer lists	1.089	.000
Ratio to list price	1.012	.000
Debt	1.000	.000
Interest amount	.999	.001
Credit agency A score	.996	.000
Credit agency B score	.999	.000
Product A	.760	.008
Product B	1.399	.000
D stratum	1.188	.002

“Bad” because the debt is numerical data, and the value of Exp (B) is more than one point. The interest amount, the score of the credit institution A, and the score of the credit institution B are numerical data as well. The lower the score of the data, the easier to be “Bad” (B) is less than one point.

5 Consideration

Based on this analysis result, we can mainly consider two points. One is comparison between product A and product B. Since purchasers of product B are likely to become 18 months Bad customers, it is considered that it is necessary to review customer data of purchasers of product B. Buyer of product A is difficult to become 18 months Bad

customer. Therefore, it can be considered to expand the range of purchasers of product A.

Next, it is about factors that make it easier for D stratum customers to become 18 months Bad customers. It is considered to be due to the expansion of financial services as well. Also, the lowest E stratum is hard to become 18 months Bad customer because it can be thought that a loan was not originally constructed.

6 The Future Tasks

In this analysis, we identified the variables which have impact on 18 months Bad. However, the payment collection rate of the customers who are 6 months Bad and 12 months Bad is lower than the 18 months Bad because of the number of times of payment. Therefore, we analyze same things in 6 months Bad and 12 months Bad and extract the influenced variables. Moreover, we reanalyze from the adapted variables respectively. For example, using the decision tree analysis with this analysis method, we search for the customers who will be “Bad”, especially which variables are particularly applicable. And also, when we conduct the analysis, we analyze both cases that with the score of the credit institution “A” and “B” and not using them respectively. From the result, we can also judge whether a new credit risk model is necessary. We propose new evaluation method of credit risk model by these results.

In addition, the scores of the credit agency A are the most effective evaluation. The factors include basic aggregation, AUC results, and prior research. The content constituting the score of the credit organization A is unclear. Therefore, it is necessary to grasp the characteristics of the score of the credit agency A.

Appendix

We will explain AUC. AUC is the name of the lower area of the ROC curve as shown in Fig. 7. AUC is an index for evaluating the strength of the relationship between the analysis target and other variables. As an evaluation criterion, if AUC is 1, there is a certain relationship. If AUC is 0.5, there is no relation. The closer the AUC is to 1.0, the more the relationship is strong. As practical examples, there are the following two. “Performance evaluation of tests for diagnosing healthy group/disease group”, “Examination of predictive factors of early prognosis after organ transplantation”.

Next, the binomial logistic regression analysis will be explained. Introduce variables to be analyzed as categories, not numerical values you want to predict. It found whether it will be 18 months Bad or not. Factors can be found, you can predict the outstanding probability and the payment completion probability. Since the odds ratio is 1.0 or more, it can be said that it affects the analyzed variables. Examples of use include creation of credit score models and medical sites.

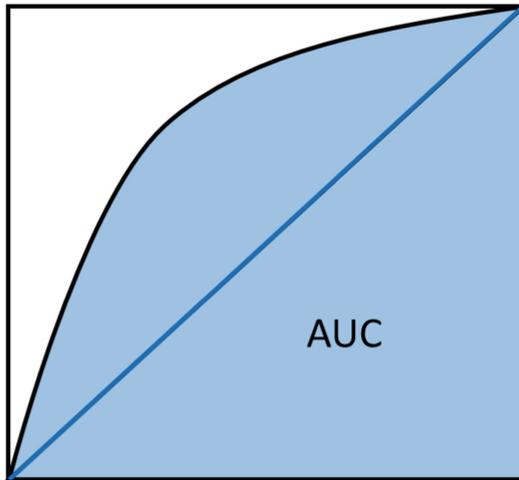


Fig. 7. The ROC curve

References

1. Cabinet Office Policy Division (Economic and Fiscal Analysis): “Trend of the global economy 2009 I”, Emerging economies the impact of the financial crisis and future prospects, vol. 2, no. 3 (2009)
2. Latin America, “A” country’s property and casualty insurance market, Sampo Japan Research Report, pp. 59–60. www.sjnk-ri.co.jp/issue/quarterly/data/qt64_3.pdf. Accessed 18 Dec 2017
3. METI: Commerce white paper 2012, Trends in the global economy, vol. 1, no. 6 (2012)
4. The emerging middle class of A country to push forward with loan consumption. Nikkei Business business.nikkeibp.co.jp/article/world/20110526/220227/. Accessed 20 Dec 2017