Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/23/20230393

Credit Card Default Prediction Analysis: Based on Default Data of Taiwanese Customers from April to September 2005

Hanlu Sun[1,a,*]

[1]*Department of finance, Nanhang Jincheng College, Nanjing, 211156, China*
a. 1914801028@qq.com
**corresponding author*

Abstract: Credit cards are widely used due to their overdraft function, which establishes a loan relationship between customers and financial institutions. However, defaulting on credit card payments can result in negative consequences, such as bad credit records for cardholders and economic losses for financial institutions. This research paper analyzes credit card default data in Taiwan from April to September 2005, with the aim of providing support for financial institutions to effectively monitor credit card risks.

Keywords: credit card, default data, factor analysis

1. Introduction

In recent years, our country's credit card market has demonstrated a significant upward trend in terms of growth and development. In response to this, various financial institutions have introduced a range of attractive policies and marketing strategies to enlarge their share of the credit card market [1-2]. Unfortunately, some of these institutions have not exercised sufficient scrutiny in assessing loan applicants, thereby contributing to a rise in the delinquency rate of credit cards. This, in turn, poses a significant danger to the health and sustainability of the credit card industry. By the close of 2022, the total value of credit card debt overdue by six months or more stood at a staggering 188.296 billion yuan, which accounted for 1.86 percent of the total outstanding balance. Therefore, it is important to conduct an in-depth analysis of credit card risks in Taiwan, as such an investigation can help identify the sources and underlying causes of these risks, thereby allowing financial institutions to better manage credit risk and secure the future of the credit card market [3].

2. Problem Raising

The dataset used in this study comprises 30,000 observations with 8 dimensions of data, including LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY, BILL_AMT (with BILL_AMT1 denoting the September 2005 bill amount (NT)), and PAY_AMT. These data can be categorized into demographic information, behavioral information, and product information. Demographic information primarily includes SEX, AGE, MARRIAGE, EDUCATION, and other data with relatively minor variations, while behavioral information involves dynamic information generated by consumer spending and payment behavior.

Based on personal experience and intuition, initial assumptions were made about which individuals are more susceptible to credit card default. Subsequently, these assumptions were tested

DOI: 10.54254/2754-1169/23/20230393

and confirmed or refuted through data analysis. The study further conducts an importance analysis of factors affecting default, providing data-driven insights for enhancing credit card risk management, setting credit ratings for cardholders, and informing policy optimization decisions for personnel involved in credit granting and risk management [4].

3. Data Analysis

3.1. Analysis of Default Factors of Credit Card Users

3.1.1. Brief Analysis

According to the dataset spanning April to September 2005, the number of customers who defaulted on their credit card payments was 6,636, representing 22.12% of the total sample. The average credit limit for customers was 167,484 yuan, while the mean age was 35.48 years old. Postgraduates and undergraduates constituted 35.28% and 46.77% of the sample, respectively. Additionally, male customers accounted for 40% of the total population.

3.1.2. Demographic Information Analysis

(1) Analysis from gender Dimension

The hypothesis put forward in this study suggests that there are significant differences between males and females regarding employment, consumption habits, performance awareness, risk preferences, and other factors that may affect credit card default rates. Specifically, it is hypothesized that women are generally more cautious and risk-averse, while men are more inclined to engage in advanced consumption [5].

As shown in Table 1, the default rate for female customers is 20.77%, which is about 86% of the default rate for male customers. This finding is consistent with the hypothesis that females are more cautious, while males are more likely to engage in advanced consumption.

Table 1: Number of credit card households and expected default situation by gender.

Row tag	Summation term	default.payment.next.mont
1	11888	2873
2	18112	3763

Opinion: Male users are more likely to be late than female users.

(2) Analysis from age dimension

Another hypothesis put forth in this study posits that borrower age is positively correlated with social experience, career development, working ability, and mental stability, and that borrowers who are either too young or too old are at increased risk of default.

To test this hypothesis, the distribution of credit card holders by age was analyzed, and the results indicate a U-shaped distribution, as shown in Figure 1, with 72.51% of borrowers under the age of 40. Moreover, the probability of overdue payments increases with borrower age, with users aged between 50 and 60 exhibiting the highest default rate at 26.17%. These findings support the hypothesis that borrowers who are older or younger may be at greater risk of default, potentially due to factors such as declining mental acuity or unstable employment prospects.

DOI: 10.54254/2754-1169/23/20230393

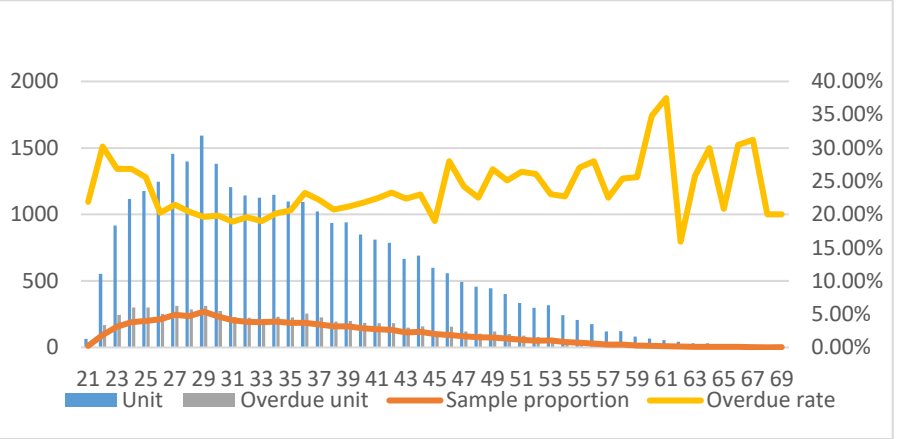

Figure 1: Age distribution of overdue users.

Based on the analysis conducted in this study, it can be concluded that the results are broadly consistent with the expected hypotheses put forth. Specifically, the finding that female users are less likely to default on their credit card payments compared to male users aligns with the hypothesis that women are more cautious and risk-averse. Similarly, the finding that borrowers who are either too young or too old are at higher risk of default supports the hypothesis that borrower age is positively correlated with social experience, career development, working ability, and mental stability. Overall, the study provides valuable insights into credit card risk management and policy optimization, and highlights the importance of demographic and behavioral factors in predicting credit card default rates.

(3) Analysis from Marriage Dimension

Another hypothesis explored in this study suggests that married individuals are more likely to default on their credit card payments compared to unmarried individuals, and may exhibit a stronger tendency towards passive repayment behavior. However, it is also posited that married individuals may have better repayment ability overall, and exhibit lower rates of delinquency due to a greater sense of responsibility.

To test this hypothesis, the data was analyzed to compare the default rates of married and unmarried credit card users. Figure 2 shows the number of married users who defaulted on their payments was 3,192, with an overdue probability of 23.68%, while the number of single users who defaulted was 3,329, with an overdue rate of 21.06%.

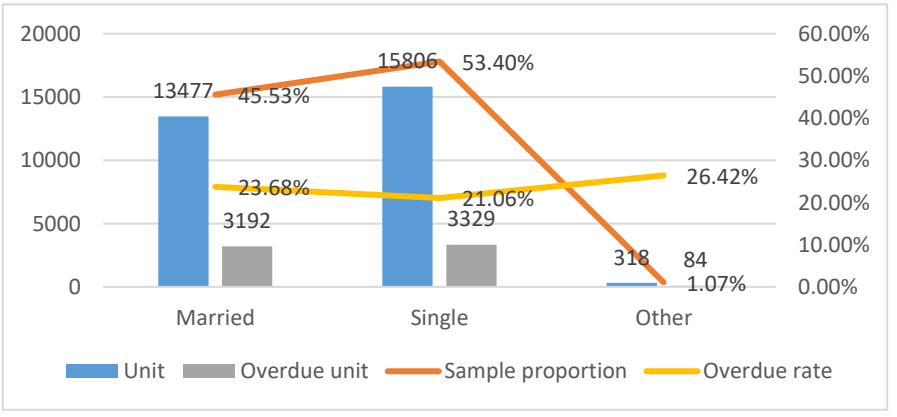

Figure 2: Marital status and overdue distribution of users.

DOI: 10.54254/2754-1169/23/20230393

Based on the data analysis, the initial hypothesis that married individuals generally have a higher default cost than unmarried individuals, and a stronger passive repayment willingness, appears to be unsupported. The data indicates that the probability of overdue payment for married users is 23.68%, while the overdue probability for single users is 21.06%. This suggests that the marital status of credit card users may not be a significant factor in predicting the likelihood of default.

However, further investigation is needed to determine the relationship between marital status and credit limit, as the average limit for married users is higher than that for single users. Additionally, it is important to consider potential confounding factors, such as consumer loans or child support, that may affect the financial situation of married users. Therefore, additional data and analysis are necessary to fully understand the relationship between marital status and credit card default risk.

(4) Analysis from the dimension of education level

Hypothesis: It is posited that higher education levels are positively associated with social positioning, personal quality, wider economic sources, greater understanding of the importance of personal reputation, increased attention to reputation, and stronger willingness and ability to repay debts, thereby resulting in lower late payment rates.

Test: As shown in Table 2, a significant association between education level and default payments ($x^2(3) = 116.86$, $p < 0.001$). Specifically, individuals with a university education were found to be more likely to default on their payments than those with a graduate school education (OR = 1.31, 95% CI = 1.23-1.39), while individuals with a high school education were less likely to default on their payments (OR = 1.41, 95% CI = 1.30-1.52). The association between education level and default payments was not statistically significant for individuals with other types of education (OR = 0.25, 95% CI = 0.11-0.50).

Table 2: Logistic Regression Results for Impact of Education on Detault Payments.

	Estimate	Std.Error	z-value	P-value
(Intercept)	-1.43483	0.02466	-58.184	< 2e-16 ***
(EDUCATION)Graduate School	0	-	-	-
(EDUCATION)University	0.26756	0.03165	8.453	< 2e-16 ***
(EDUCATION)High School	0.34460	0.04109	8.387	< 2e-16 ***
(EDUCATION)Other	-1.37285	0.38998	-3.520	0.00043 ***

Opinion: Individuals with higher levels of education exhibit better risk awareness, greater attention to personal credit information, and a lower probability of late payments [6]. This finding is consistent with the hypothesis put forth.

3.1.3. Product Information Analysis

Analysis from the Quota Amount Dimension:

Hypothesis: The historical credit information of a user serves as a critical determinant for the bank in determining their credit amount. Users with good credit information are likely to receive higher credit amounts, and as a result, have a lower probability of late payments.

Inspection: Figure 3 shows, it was observed that 84.73% of households had a limited credit amount of less than 300,000 yuan, with users in this category exhibiting a significantly higher probability of late payments compared to those with credit amounts exceeding 300,000 yuan. For users with a credit limit ranging between 10,000 and 50,000 yuan, the probability of late payments is even higher, exceeding 32%.

DOI: 10.54254/2754-1169/23/20230393

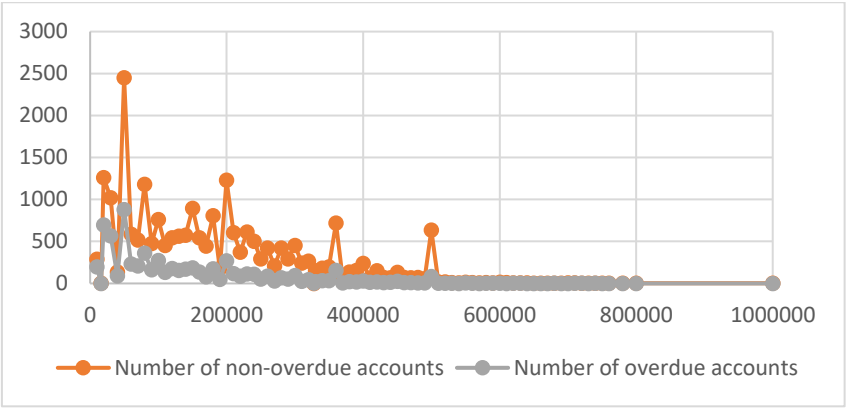

Figure 3: Overdue distribution of the limit amount.

Opinion: If a user's credit limit is less than 50,000 yuan, it is reasonable to assume that the bank may offer a smaller credit amount due to the user's suboptimal credit status. This assumption can be further substantiated by verifying the user's historical credit information. Thus, the available information supports the aforementioned hypothesis.

3.1.4. Behavior Information Analysis

(1) Analysis from the consumption amount dimension

Hypothesis: It is hypothesized that there is a positive association between a user's bill amount and their repayment pressure, leading to a higher probability of late payments.

Test: The probability density curve of overdue user bill amount remains consistent when the latest issue of BILL_AMT1 is taken as the sample. Figure 4 illustrates that the distribution of BILL_AMT1 is skewed to the right, with a prolonged tail towards higher values. This indicates that while there are many users with smaller bill amounts, there are also a considerable number of users with substantially high bill amounts. These findings suggest that users with larger bill amounts may experience greater repayment pressure, which could potentially lead to a higher probability of late payments.

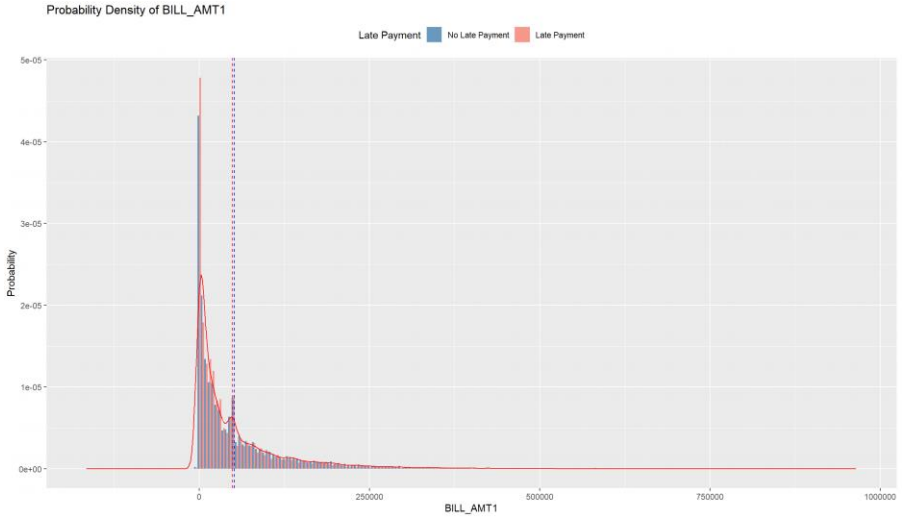

Figure 4: BILL_AMT1 distribution.

In the probability density graph for BILL_AMT1, two lines can be observed, represented by different colors. The blue line illustrates the mean value of the entire BILL_AMT1 variable, while

DOI: 10.54254/2754-1169/23/20230393

the red line represents the mean value of the BILL_AMT1 variable for those who made a late payment on their credit card bill in the following month.

The peak position of the probability density map for late payment users moves towards the left relative to the overall distribution, indicating that the bill amount of late payment users is more concentrated in a lower range. It is important to note that while the red line shifts towards the right, it does not imply that all high bill payers will be late payers.

Moreover, by comparing the two charts, it can be observed that the bill amount of late payers is more concentrated in the lower range. While individuals with higher bill amounts may be more likely to pay late, this represents only a trend. In reality, individuals who actually pay late tend to have lower bill amounts relative to the overall distribution.

Opinion: It is important to note that the test results do not necessarily prove that the hypothesis is false. Rather, they indicate that the relationship between bill amount and overdue probability is more complex than initially hypothesized. While the hypothesis suggests that higher bill amounts lead to greater repayment pressure and higher overdue probabilities, the test results show that this relationship is not as straightforward as previously thought.

Instead, the test results suggest that there may be other factors at play, such as financial planning and management skills, that influence the relationship between bill amount and overdue probability. Users who overbill may have a better understanding of their financial situation and may be more likely to pay their bills on time, regardless of the amount. On the other hand, users with lower bills may be more susceptible to temporary changes in their financial situation, leading to more late payments.

Overall, the test results suggest that the relationship between bill amount and overdue probability is more complex than initially hypothesized, and that additional research may be needed to fully understand the factors that influence late payments.

(2) Analyzing Quota Usage Dimensions

A hypothesis is proposed in this study that posits that customers with low consumption rates have a lower default rate, while customers with high consumption rates experience an increasing default rate as their consumption amounts increase. To test this hypothesis, the sample used for analysis is the PAY_AMT1 of the most recent period. As shown in Figure 5, the probability density curve reveals that there are more customers who are overdue with low payment amounts. When the user's payment amount ranges from 0 to 5000, the probability of default is the highest, surpassing 25%.

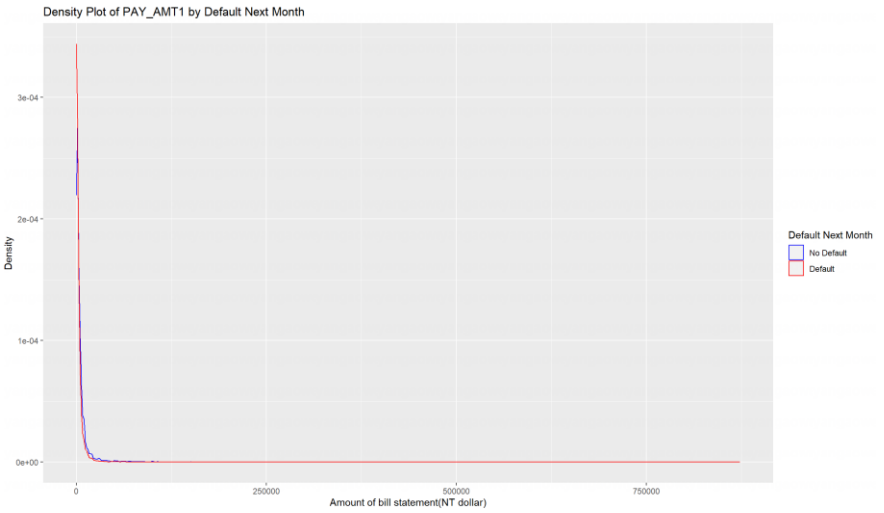

Figure 5: PAY_AMT1 distribution.

DOI: 10.54254/2754-1169/23/20230393

Opinion: The test results are inconsistent with the hypothesis proposed.

(3) Analysis from Overdue History Dimension

This study proposes a hypothesis that suggests the longer a user's payment is overdue, the higher the likelihood that their repayment ability has declined, resulting in an increased probability of overdue payments. To test this hypothesis, the sample analyzed was the PAY_0 of the most recent issue. If the values of -2 and -1 exist in PAY_0, they should be combined with 0, representing that the user has pre-paid at that time. Furthermore, due to the small number of households that are more than four months overdue (139), a separate analysis would affect the results. Therefore, households that are four to eight months overdue are consolidated into three. Figure 6 shows the longer a user's payment is overdue, the higher the probability of overdue payments. When the user's payment is overdue for more than three months, the probability of the user being overdue is the highest.

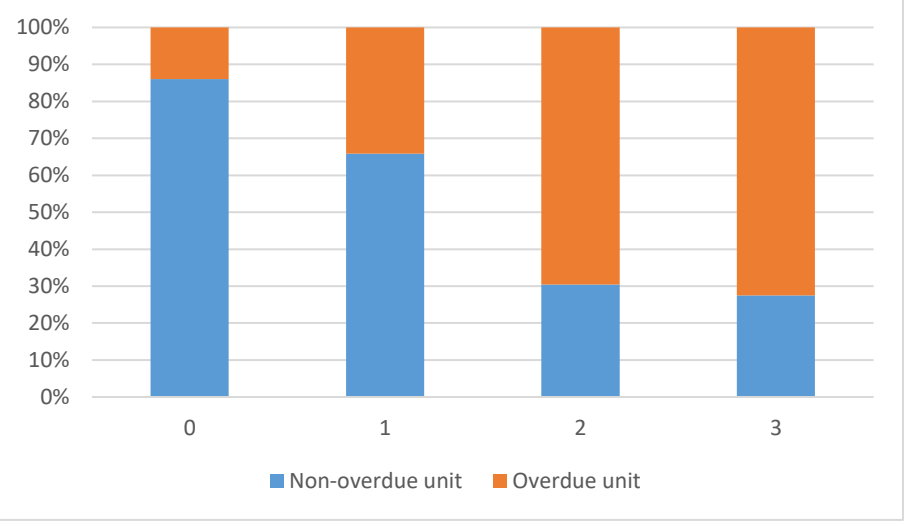

Figure 6: Customer overdue history distribution.

In conclusion, this study finds that there is a positive correlation between the length of time a payment is overdue and the probability of overdue payments. The test results are consistent with the proposed hypothesis, suggesting that users with a longer overdue payment history have a higher likelihood of being overdue in the future. These findings have significant implications for credit management and risk assessment in the financial industry. Future studies could further explore the factors that contribute to overdue payments and how they can be mitigated to improve credit management practices.

3.2. Significance Analysis of Default Factors

To better understand the problem of credit card default and improve risk management strategies, this study utilized a random forest model to analyze the relationship between several predictors and the probability of credit card default. The goal was to identify the most important predictors of default and gain insight into the factors that contribute to the problem. Figure 7 shows the degree of significance of each factor.

The random forest model is a powerful tool that can handle large datasets with numerous variables, making it an ideal approach for this analysis. By identifying the most important predictors of default, this study provides valuable insights that can inform risk management strategies and improve credit management practices.

DOI: 10.54254/2754-1169/23/20230393

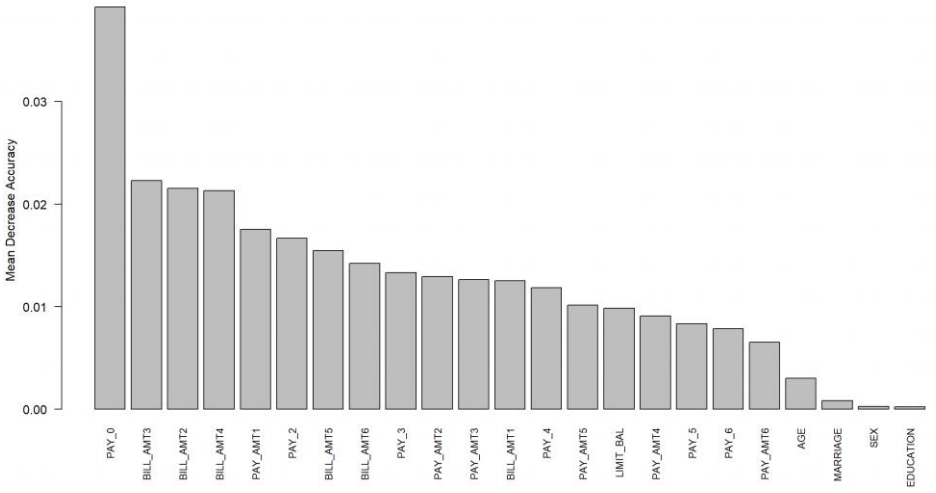

Figure 7: Ranking of significance factors.

Table 3: Significance of default factors.

	OOB Estimate	95% CI
Error rate	0.180	(0.176,0.184)
Precision	0.843	(0.839,0.846)
Recall	0.946	(0.943,0.948)
Specificity	0.384	(0.378,0.389)
Kappa	0.388	(0.383,0.393)

Of the 23 predictor variables included in the model, PAY_0, AGE, and BILL_AMT1 are the three most important variables for predicting default payment, as shown in Table 3, with relative importance values of 0.04, 0.016, and 0.013, respectively. The high importance of PAY_0 suggests that a customer's most recent payment status is the strongest predictor of whether they will default on their next payment. It is important to note that the remaining predictor variables have relatively low importance values, with the majority having values below 0.04. The model's overall accuracy is 82.0%, indicating that the model is performing relatively well in predicting whether a customer will default on their next payment. However, it is crucial to consider additional measures such as precision, recall, and AUC when evaluating the model's performance. Overall, this Random Forest model provides valuable insights into which predictor variables are most important for predicting default payment, which can help credit card companies identify and prioritize customers who are at a higher risk of defaulting.

4. Conclusion

After analyzing the data in the 8 dimensions mentioned above, several conclusions can be drawn. Firstly, male users have a higher probability of defaulting compared to female users. Secondly, the age range of 25 to 40 years old has the lowest probability of default. Thirdly, education level is negatively correlated with the probability of default, indicating that users with a higher education level are less likely to default. Fourthly, the amount of credit extended to the user is inversely proportional to the probability of default, meaning that the higher the credit amount, the lower the probability of default. Fifthly, the probability of default is positively correlated with the bill amount, indicating that the higher the bill amount, the higher the probability of default. Sixthly, the

DOI: 10.54254/2754-1169/23/20230393

probability of default is positively correlated with the duration of overdue payment. Finally, the predictors PAY_0, AGE, and BILL_AMT1 have significant impacts on the probability of default.

The analysis of multiple groups of data, such as the combination of gender and education level, can provide more insights and help in the assessment of credit card risk. The results of this analysis can also aid financial institutions in identifying customers with higher risks of default and develop appropriate internal control measures. Financial practitioners should also focus on managing customers with smaller consumption bills, instead of solely focusing on customers with larger consumption bills.

Through this practice, it is evident that big data analysis plays a crucial role in the future of financial careers. The core competency of big data risk control lies in proposing hypotheses, testing them, forming opinions, and making judgments and predictions based on them. Risk control personnel should have the ability to propose hypotheses, form opinions after verification, and make judgments and predictions based on the opinions under the condition of big data risk control.

Acknowledgments

We would like to express our sincere gratitude to all the individuals and organizations who supported and contributed to this research project.

Firstly, we would like to thank the financial institution that provided us with the credit card default data used in this study. Without their support and cooperation, this research would not have been possible.

We would also like to thank our supervisors and colleagues for their guidance, advice, and encouragement throughout the project. Their insights and feedback were invaluable in shaping our analysis and conclusions.

Thank you all for your invaluable contributions to this research paper.

References

[1] *Xingliang Zhu,Jia Wang,and Jiaoju. Ge.Empirical Analysis of Factors Influencing Credit Card Repayment based on Probit Model [J]. Consumer economy,2013(4):1557-1580*

[2] *Kuangnan Fang,Guijun Zhang,andHuiying Zhang. Personal Credit Risk warning Method Based on Lasso-Logistic Model [J]. Quantitative and technical economic research,2014(2):125-136*

[3] *Ruiting Mei,Yang Xu,and Guochang Wang. Analysis of credit card default prediction model and its influencing factors [J]. Statistics and Application, 2016, 5(3): 263-275.*

[4] *Xie,Xuanli,Yan Shen,Haoxing Zhang,and Feng Guo.2018. Can Digital Finance Promote Entrepreneurship? – Evidence from China. Economic Quarterly [Chinese, Jingjixue Jikan],17(4):1557–1580*

[5] *Manning, R. D. (2000). Credit card nation: the consequences of America's addiction to credit. New York: Basic Books.*

[6] *Jagtiani, Julapa, and Catharine Lemieux. 2019. The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform. Financial Management 48(4):1009–1029.*