Table of Contents

Introduction

Financial institutions need a precise credit risk evaluation to achieve profitability goals and maintain regulatory requirements. The report analyzes actual Taiwanese credit card user data to predict payment defaults through analysis of customer actions and personal characteristics. The analysis process includes data preparation work followed by feature development and model evaluation. The scorecard package in R enables the creation of a transparent logistic regression-based scorecard which generates an interpretable credit score for default probability estimation. Moreover, the analysis evaluates whether another advanced machine learning model delivers better predictive results than basic methods through a comparison study.

Dataset Understanding

1. Dataset Description

The dataset contains financial information from 30,000 Taiwanese credit card users in New Taiwan (NT) dollars starting from April to September 2005. Although it is not a full year dataset, this six-month span provides sufficient data to analyze repayment behavior and detect default trends. The dataset includes 25 variables which combine personal characteristics with payment patterns according to the following table:

Variable	Description	Data Type
ID	Client ID	Integer
LIMIT_BAL	Total credit limit in NT dollars that a person can use.	Integer
SEX, EDUCATION, MARRIAGE, AGE	Demographic attributes about customer's sex, education level, marriage status and their age.	Integer
PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6	Financial variables PAY_0 to PAY_6 show each client's repayment status from September 2005 to April 2005, indicating whether they paid on time or were late	Integer
BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6	Financial variables BILL_AMT1 to BILL_AMT6 show the amount of each client's bill to be paid from September 2005 to April 2005.	Numeric
PAY_AMT1, PAY _AMT2, PAY _AMT3, PAY _AMT4, PAY _AMT5, PAY _AMT6	Financial variables PAY_AMT1 to PAY_AMT6 show the amount each client paid in the previous month from September 2005 to April 2005.	Numeric
default.payment	Default status of customers	Integer

Table 1. Dataset variable description.

2. Data Exploration and Preparation

Some records show no late payments but are marked as default (738 records), while others show delays yet are non-default (398 rows). This may occur because the dataset spans April to September 2005, while the default flag may represent outcomes for the entire year. Since there is no detailed description confirming whether the "default.payment" reflects the six months or the full year, these records are kept maintaining dataset integrity and reflect realistic repayment behavior beyond the observed months.

The dataset is not normally distributed. For example, descriptive statistics shows that "LIMIT_BAL" has a mean of 167,523 NT dollars and a median of 140,000 NT, ranging from 10,000 to 1,000,000 NT showing a right skewness. But this does not affect the analysis since logistic regression of credit scoring does not require normality. Noticeably, the dataset contains no missing values which makes it rare for real-world data sets.

Moreover, "PAY_0" was renamed to "PAY_1" to maintain consistency with the repayment status sequence (e.g., "PAY_2", "PAY_3", "PAY_4", and so on). 68 records with undefined values (0) in "EDUCATION" or "MARRIAGE" were removed and left 29,932 records. "Unknown" values (5 and 6) in "EDUCATION" were combined with "Others" (4) to represent similar unclear categories, and values like –2 and –1 in PAY_0 to PAY_6 were recoded to 0 to indicate no payment delay.

New variables were also created to measure each client's credit utilization ratio (e.g., "util_1", "util_2" and so on), showing how much of their credit limit is used each month. Negative ratios caused by overpayments were converted to 0 to represent no credit usage. An average - "avg_util_6month" was then calculated to summarize usage over six months. This new variable captures the average delay value across six months, where higher values indicate greater credit usage and higher risk. Additionally, the maximum delay - "max_delay_6m" experienced by each client and the number of months with late payments - "num_late_months" are also calculated as both reflect the severity and frequency of repayment issues.

3. Demographic background understanding

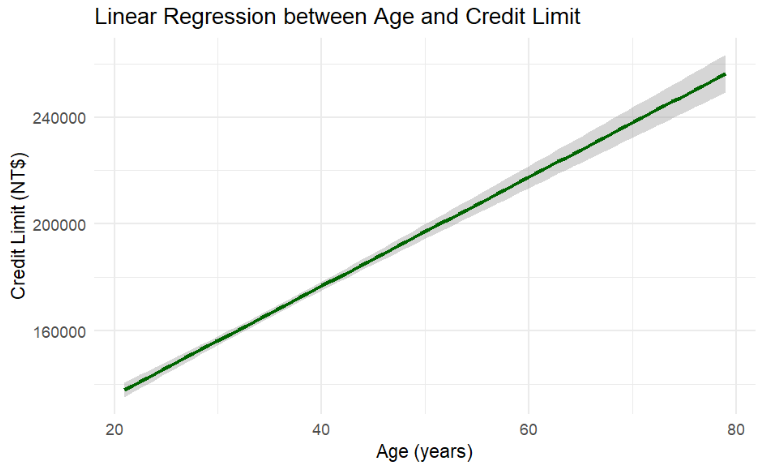


Figure 1. Relationship between age and limit credit balance.

Figure 1 shows that older clients generally receive higher credit limits, which makes sense because they are likely to have more stable incomes and longer credit histories. For each additional year of age, the credit limit increases by about 2,039 NT dollars (Appendix A).

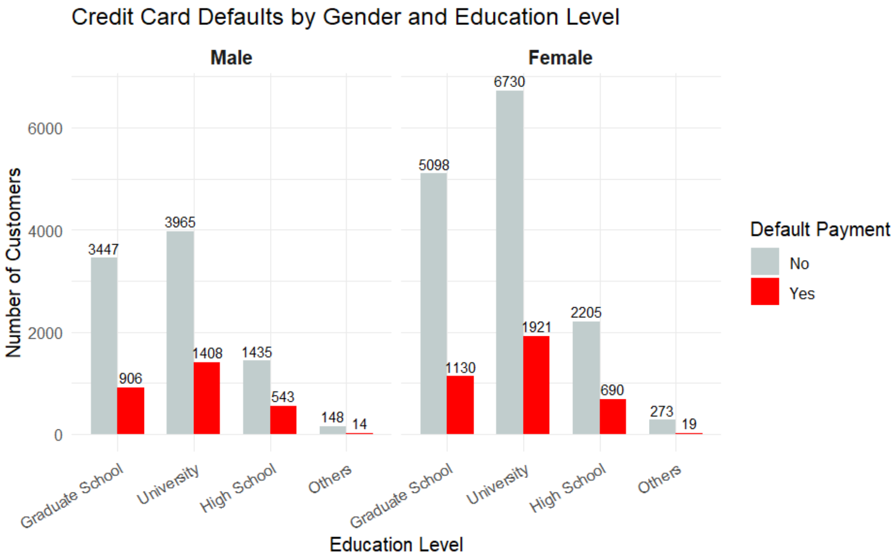


Figure 2. Credit card defaults by gender and education level.

The data in Figure 2 indicates that university and graduate students make up most clients while female clients exceed male clients at each educational level (e.g., 6,730 females vs. 3,965 males at the university level). The default rates among university-educated clients show that 1,921 females and 1,408 males failed to meet their payment obligations despite their higher level of education.

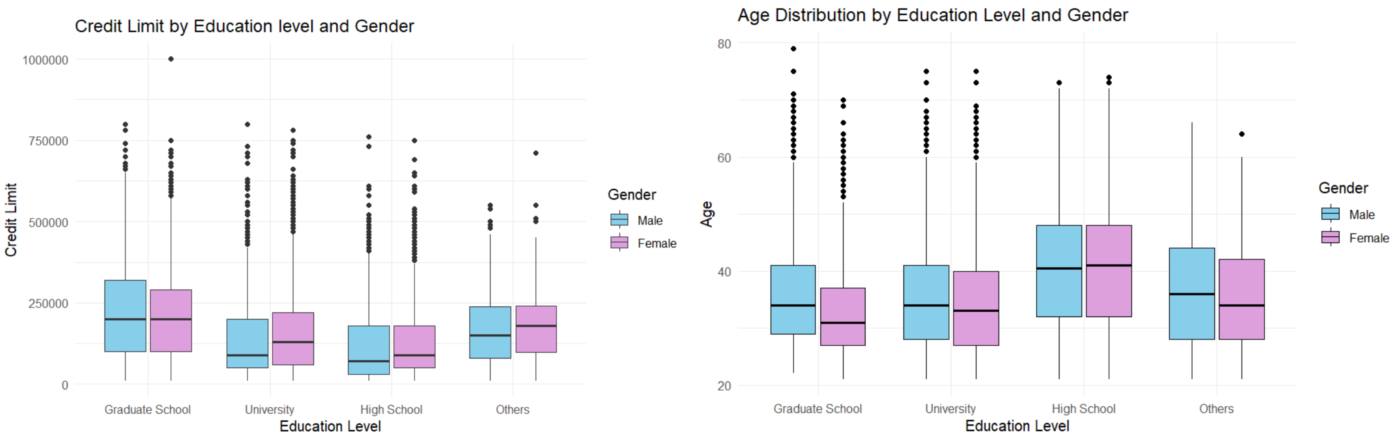


Figure 3. Credit limit and age distribution by education level and gender.

Figure 3 shows two boxplots analyzing credit limits and age distribution by different educational backgrounds and between male and female clients. The medians of credit limits for the clients with graduate-level education are below 250,000 NT dollars, whereas those with university or high school education have lower limits near or below 125,000 NT dollars. The credit limit ranges for both genders are similar across all education levels, and the age tends to increase with education level as high school and Others include older clients on average (37-40 years).

The initial overview of this dataset shows that most clients are female university graduates, suggesting a young and financially capable customer base. Credit limits rise with age and education, indicating older and more educated clients receive higher credit lines. However, defaults across these groups show that demographics alone cannot fully predict credit risk and it needs to further analyze using credit scoring.

Credit Scoring Model Development

1. **Data Partitioning and Binning**

Before building the credit scoring model, the dataset was split into 70% training and 30% validation subsets. Binning was then applied to the training set using the woebin function to convert continuous variables into discrete groups.

2. **Variable selection**

A credit scoring model was built using the "Scorecard" package to estimate default probability and convert it into an interpretable score. Information Value (IV) was first calculated for all variables to assess their predictive strength. Only variables with IV ≥ 0.1 which are strong predictors were kept for further modelling. However, AGE was kept for interpretability as it had the highest IV (0.02) among demographic factors (SEX-0.009, MARRIAGE-0.005, EDUCATION-0.01). The remaining variables were “avg_delay_6month”, “num_late_months”, “max_delay_6m”, “PAY_1” to “PAY_6”, “LIMIT_BAL”, “PAY_AMT1” to “PAY_AMT3,” and “AGE”. Since “num_late_months”, “avg_delay_6month” and “max_delay_6m” were derived from “PAY_1” to PAY_6”, “PAY_1” to PAY_6” and “PAY_AMT1” to ”PAY_AMT3” were removed, leaving “num_late_months”, “avg_delay_6month”, “max_delay_6m”, “LIMIT_BAL” and “AGE”.

3. **Fine Tuning**

The Weight of Evidence (WoE) plots in Appendix B show smooth patterns without irregular jumps, so no regrouping is required. For “avg_delay_6month”, clients with the lowest delay (<0.167) have an 88.4% positive probability, which drops steadily to 38% for those with average value above 1. A similar decline appears in “num_late_months”, where the probability falls from 88.4% to 40% once three or more late months occur. “max_delay_6m” follows the same trend, decreasing from 88.4% to 54% for longer delays. In contrast, “LIMIT_BAL” rises with positive probability, reaching 88% among high-limit clients, suggesting stronger financial reliability. “AGE” remains stable across groups showing no major fluctuations. Timely repayments and higher credit limits are strong indicators of creditworthiness, while age appears to have minimal influence on default risk.

4. **Logistic Regression**

Variable	Estimate	p-value
Intercept	-1.24	< 2e-16
avg_delay_6month_woe	-0.22	0.00012
num_late_months_woe	-0.69	< 2e-16
max_delay_6m_woe	-0.06	0.259
LIMIT_BAL_woe	-0.39	4.02e-16
AGE_woe	-0.44	0.000351

Table 2. Logistic regression summary model 1.

This initial logistic regression model in Table 2 shows that “avg_delay_6month_woe”, “num_late_months_woe “, “LIMIT_BAL_woe” and AGE_woe” are statistically significant with all p-values <0.05 since statistical significance is assessed at the 95% confidence level. The most influential variable is “num_late_months_woe”, followed by “LIMIT_BAL_woe”, with estimates of –0.69 and –0.39 respectively. These negative values mean more late payments and lower credit limits are strongly associated with higher credit risk. This will reduce the credit score for riskier customers and increase for safer ones. Although “max_delay_6m_woe” was included, its p-value of 0.259 suggests it is not statistically significant and could be removed in the final model.

Variable	p-value	VIF
avg_delay_6month_woe	0.00012	9.53
num_late_months_woe	< 2e-16	13.23
max_delay_6m_woe	0.259	6.58
LIMIT_BAL_woe	4.02e-16	1.16
AGE_woe	0.000351	1.07

Table 3. Logistic regression summary model 1 VIF check.

VIF was then applied to check multicollinearity in Table 3, values are greater than 10 considered confusing the model. Despite its high VIF, “num_late_months_woe” (13.23) is highly statistically significant ($p < 0.05$). In contrast, “max_delay_6m_woe” had the lowest predictive strength ($p = 0.259$). Stepwise selection was then used to keep only the most relevant predictors which are “avg_delay_6month_woe”, “num_late_months_woe “, “LIMIT_BAL_woe” and AGE_woe”.

Variable	Estimate	p-value
Intercept	-1.24	< 2e-16
avg_delay_6month_woe	-0.23	6.02e-05
num_late_months_woe	-0.73	< 2e-16
LIMIT_BAL_woe	-0.39	3.06e-16
AGE_woe	-0.45	0.000311

Table 4. Logistic regression summary model 2.

The refined logistic regression model 2 in Table 4 shows all variables remain statistically significant with p-values below 0.05 which indicates strong predictive power. In addition, “num_late_months_woe” shows the largest estimate (-0.73), reaffirming its dominant role in assessing credit risk, followed by AGE_woe (-0.45) and LIMIT_BAL_woe (-0.39). These values highlight that frequent late payments, older age, and lower credit limits are closely linked to higher risk of default.

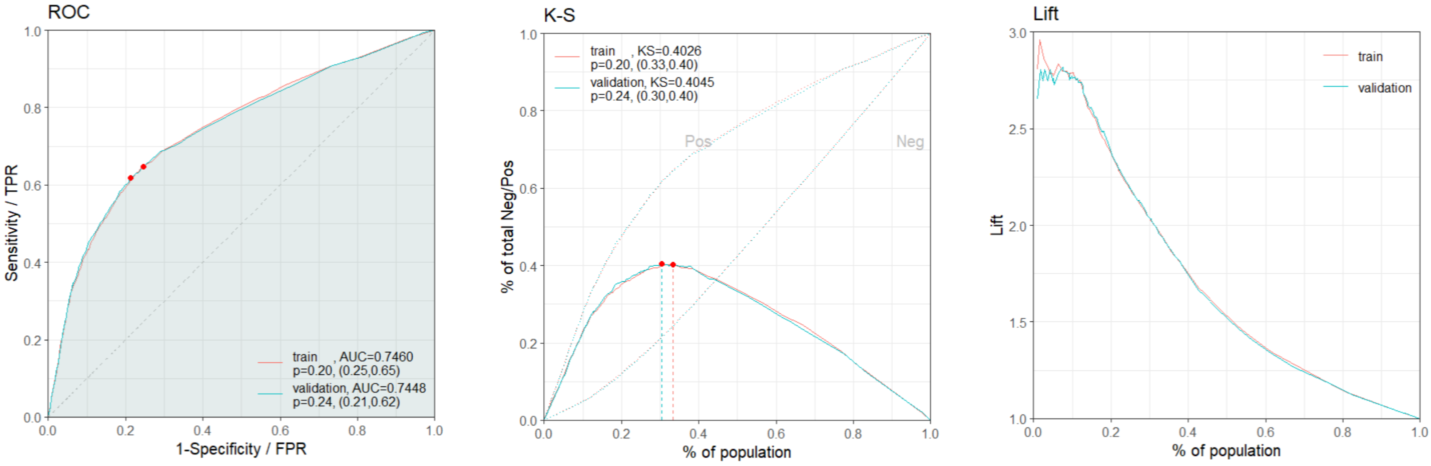


Figure 4. Logistic regression model 2 evaluation.

The model shows stable and reliable performance with AUC scores of train (0.746) and validation (0.744), it can effectively distinguish between high and low-risk customers. A K-S value above 0.4 further supports its strong predictive power, while the lift chart shows the model successfully prioritizes risky customers early making it valuable for targeting interventions or credit decisions.

5. Scorecard

Characteristics	Attribute	Scorecard Points
AGE	< 26	129
	26 <= age < 29	133
	29 <= age < 36	134
	36 <= age < 46	132
	age >= 46	130
Average Delay Days (across 6 months)	days < 5	145
	5 <= days < 15	137
	15 <= days <= 30	135
	days > 30	129
Number of late payment months	0 late months	145
	1 month late	120
	2 months late	112
	>= 3 months late	94
Credit Limit (NTD)	< 40000	129
	40000 <= limit < 140000	134
	140000 <= limit < 380000	140
	Limit >= 380000	144

Table 5. Final scorecard for business to validate customer credit risk.

The scorecard function used the selected bins and logistic model. A baseline score of 600, odds of 1 to 30, and PDO of 20 reflected the three percent default rate, meaning one default per thirty good customers and every twenty points doubled repayment odds. The result returns to a basepoint of 538 and scores for bin ranges.

Some bins have negative scores, which can be confusing for business users. To make the scorecard easier to interpret, all bin scores were converted to positive. For each variable, the lowest bin score was identified, and its absolute value was added to all bins of that variable. The updated basepoint (517) was calculated by subtracting the total increase from the original basepoint to ensure that final credit scores remained consistent across customers.

The scorecard was then restructured into a final scorecard table for business users in Table 5. The updated basepoint of 517 was evenly divided across the four variables, giving 129.25 points each to maintain a balanced score distribution across all variables. Final bin scores were then rounded to integers and displayed clearly in terms of variable name, attribute and points. Finally, the average scores for default (525) and non-default (550) customers were calculated. The final cut-off point was selected as the midpoint between the two groups, which is 538 and provides a decision boundary. It means if a customer achieves their point equal or greater than 538, they are eligible for credit borrowing.

For example, a 35-year-old customer with an average delay of 10 days over the past six months, one late payment month, and a credit limit of 160,000 NTD would receive the following score: 134 points for age, 137 for delay days, 120 for late payments, and 140 for credit limit. As a result, they have 531 points. As this score falls below the cutoff point of 538, the customer would not qualify for credit approval based on the scorecard.

Build and Compare Another Prediction Model

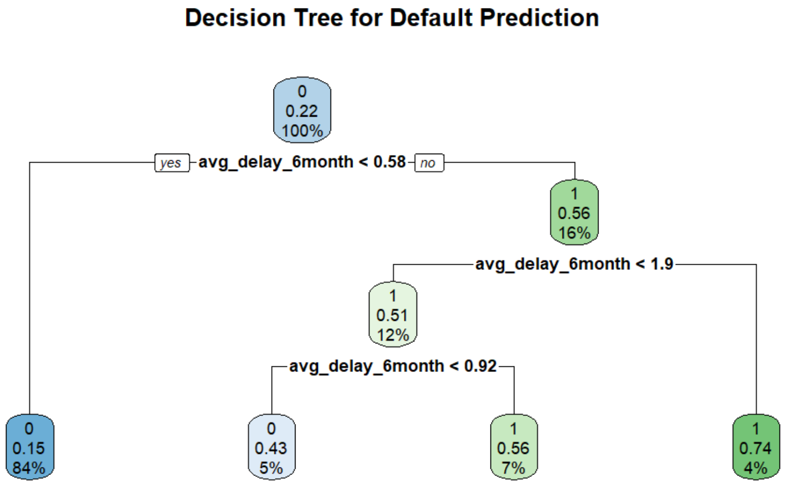


Figure 5. Decision tree model.

The decision tree model was built using four key predictors: “avg_delay_6month”, “num_late_months”, “LIMIT_BAL”, and “AGE”. As shown in Figure 5, the tree splits customers based on “avg_delay_6month” indicating it provides the most useful splits to distinguish between default and non-default customers . The first major split occurs at 0.58, effectively separating low-risk customers which is left branch with default probability 0.15 from higher-risk groups. Further splits refine risk levels, with customers having "avg_delay_6month" above 1.9 showing the highest default likelihood at 0.74. This tree provides a simple, interpretable structure for identifying risk segments based on repayment patterns.

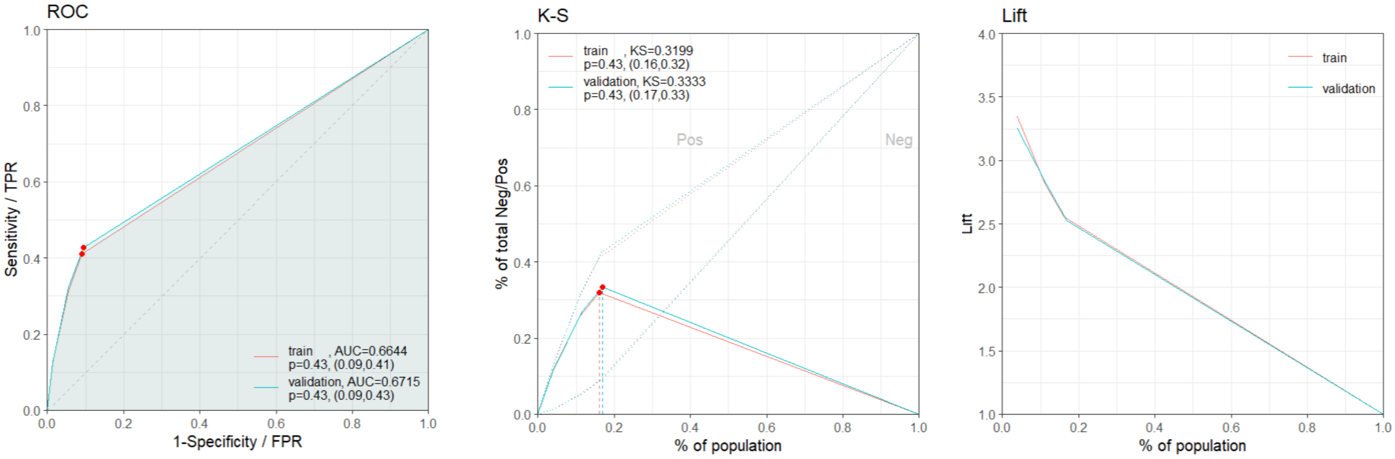


Figure 6. Decision tree model evaluation.

The tree model also uses splitting with 70% training and 30% validation. Figure 6 demonstrates moderate predictive ability, with AUC scores of 0.664 and 0.671 and K–S values around 0.32 showing weaker separation between defaulters and non-defaulters and a rapid decline in lift beyond the top deciles. In contrast, the logistic regression model in Figure 4 achieves higher AUC values (0.746 train, 0.744 validation) and stronger K–S statistics above 0.4, producing a more consistent discrimination across risk segments.

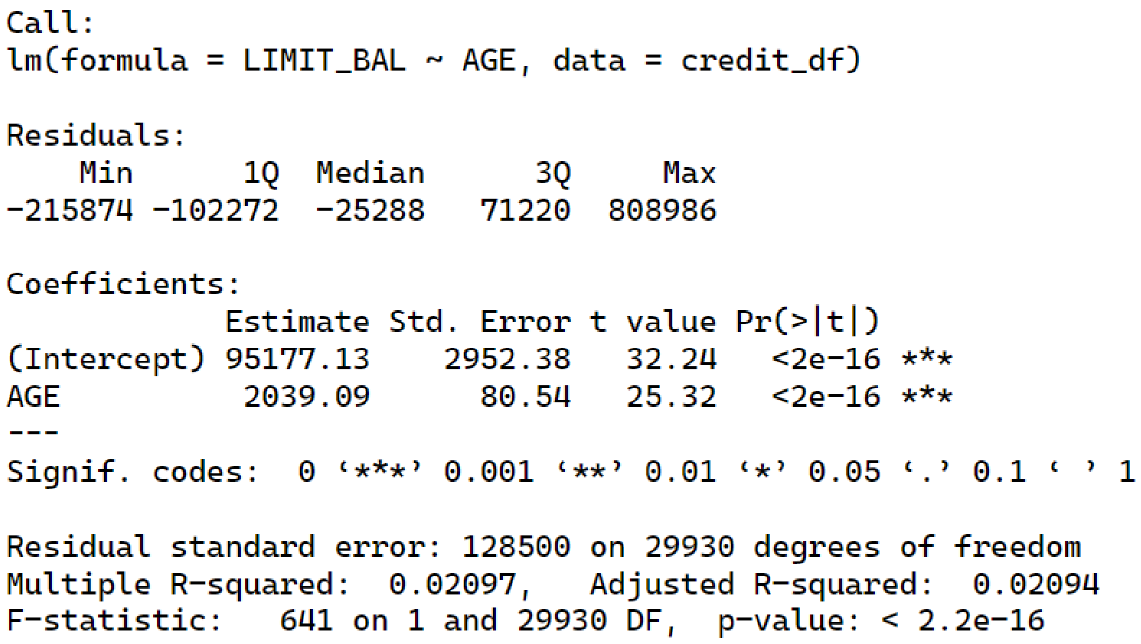
If these models were used for real customers, the tree model would make more mistakes because it splits people into fixed groups and ignores small differences near the cut-off points. For example, a customer earning 49,000 NT and another earning 51,000 NT might fall into different groups even though they are very similar, which can lead to wrong predictions. The scorecard model, however, considers these small differences and gives each customer a continuous score that better reflects their true risk, helping the business identify potential defaulters earlier and make fairer credit decisions.

Conclusion

Above all, this credit scoring analysis chose average repayment delay, number of late months, credit limit and age as key variables through Information Value analysis and multicollinearity tests. The scorecard also received business-friendly point values and a specific cutoff point at 538 which enabled risk-based decision support. The total score remained unchanged when the model applied negative score adjustments for better understanding.

The logistic regression demonstrated better performance to the decision tree model through AUC, KS and lift metric evaluations in both training and validation datasets. The tree model identified average delay as the primary risk indicator, but its basic structure restricted its ability to make accurate predictions. The scorecard functions as a useful data-based system which helps users evaluate customer credit risk through simple methods.

Appendix A: Linear Regression Summary



```
Call:
lm(formula = LIMIT_BAL ~ AGE, data = credit_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-215874 -102272  -25288   71220  808986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 95177.13    2952.38   32.24   <2e-16 ***
AGE          2039.09      80.54   25.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 128500 on 29930 degrees of freedom
Multiple R-squared:  0.02097,	Adjusted R-squared:  0.02094 
F-statistic:   641 on 1 and 29930 DF,  p-value: < 2.2e-16
```

Figure A1: Linear Regression Summary.

The regression results mean for each additional year of age, the credit limit increases by about 2,039 NT dollars, and this relationship is statistically significant (p < 0.001). However, the R-squared value of 0.02 indicates that age alone explains only a small portion of the variation in credit limits and suggests a weak relationship between the two variables. This weak relationship hints that banks should rely more on other factors, such as income, repayment history, and spending behavior, when determining appropriate credit limits.

Appendix B: Fine classing plots

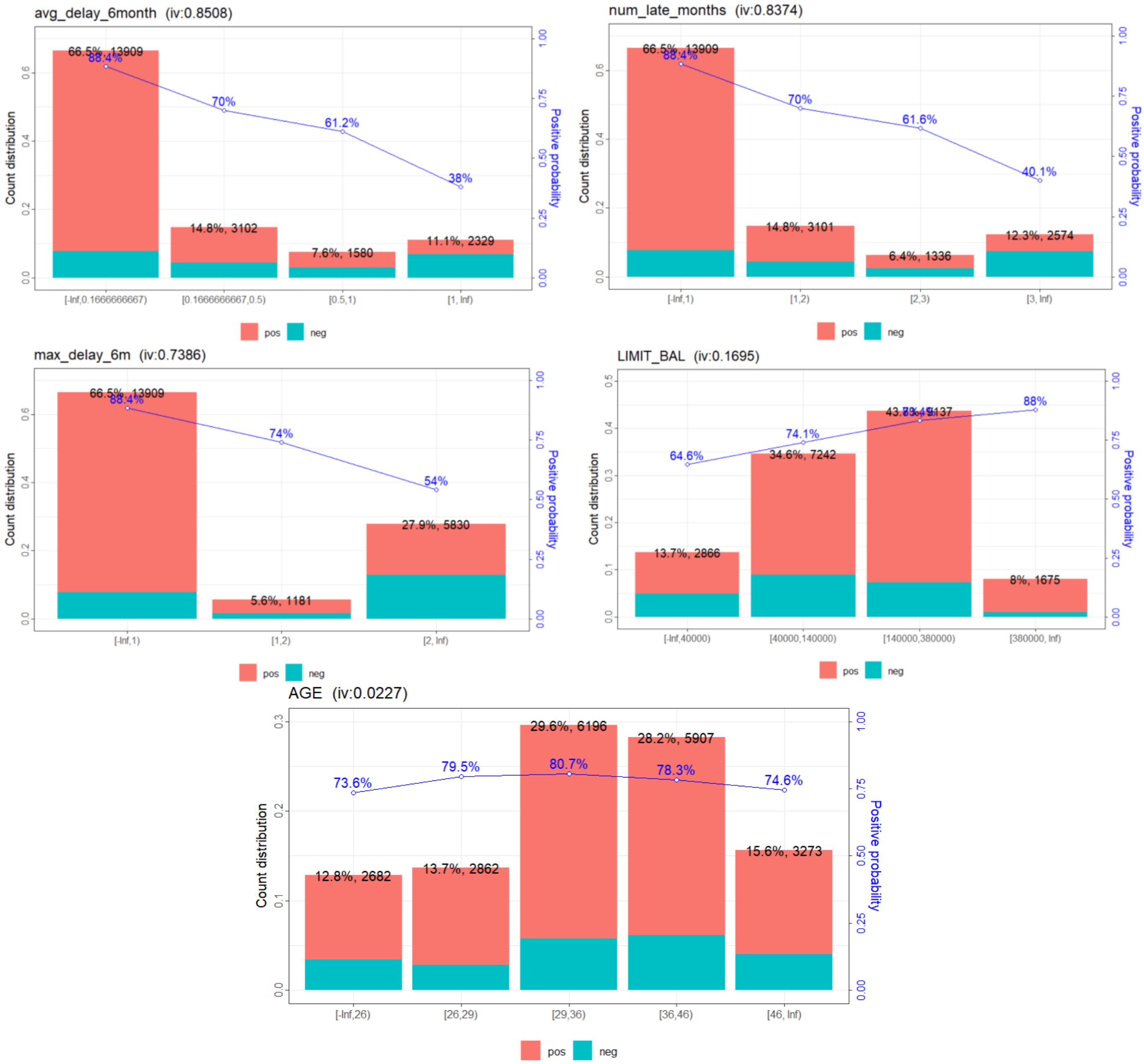


Figure B1: Fine classing plots.