Final Project

Understanding Credit Default:
An Analytical Approach to Predicting Payment Behaviour

Jiaxuan (Jasmine) Tian
1007935415
University of Toronto
Department of Statistical Sciences
STA303H1S Methods of Data Analysis II
Mohammad Kaviul Anam Khan
April 7, 2024

I. Introduction

In Canada, where credit delinquencies have increased, it is crucial to comprehend the factors that contribute to credit card defaults. Delinquency rates increased from 1 in 31 to 1 in 25 Canadians between Q3 2022 and Q3 2023, with significant increases in British Columbia and Ontario (Equifax, 2023).

To better understand the relationship between credit card default risk and financial behavior, this study examines the relationship between financial behavior and demographic characteristics. This study builds on the logistic regression approach that was supported by Çāgsar & Unal (2019) and Sharma & Mehra (2018), who identified past payment behavior and demographic factors as important predictors. Merikoski et al. (2018) suggests adding balance limits and credit utilization to the predictive model.

Despite adopting the logistic regression approach because of its reliability and track record of success in related studies, this study stands out for its attempt to obtain a more complex picture of default determinants by enlarging the variable pool. Our study expands the field of inquiry by including variables such as balance limits in an effort to improve default risk forecasts in this dynamic financial environment.

II. Methods

2.1 Choice of Method

A Generalized Linear Model (GLM) was used to simulate the likelihood of credit card default since it can handle binary outcomes and produce easily understood coefficients. The logistic regression, a GLM with a logit link function, is widely used in risk assessment studies and is effective in handling dichotomous dependent variables.

The model is provided by:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = X^{'}\beta$$

With assumptions of :

1. The response default (Y) is a binary variable.
2. All the predictors are independent with each other.
3. Y= $\log\left(\frac{\pi}{1-\pi}\right)$ has linear relationship with each $x_i\beta_i$

2.2 Variable Selection

The selection of initial variables was based on the literature's indication of their significance. By reclassifying some variables, among other preprocessing techniques, this selection was improved. By keeping variables with significant explanatory power, the iterative stepwise method integrated both the Akaike Information Criterion (AIC) and

Bayesian Information Criterion (BIC) to optimize the model and maintain a balance between model simplicity and fit.

2.3 Model diagnostics and validation

To evaluate whether the chosen model was appropriate, model diagnostics were carried out. The residuals of deviation were compared to the fitted values to find any outliers or significant cases that might cause the model to be distorted. To verify that the expected probabilities and the observed results line up, the calibration plots were also looked at, which added to the model's dependability. Receiver Operating Characteristic (ROC) curves were used to assess the consistency of the model and make sure it could discriminatively distinguish between default and non-default events. A model's performance was measured by its Area Under the Curve (AUC) - a value near 1 indicates a better model performance.

III. Results

3.1 Description of Data

The dataset includes 29601 credit card user observations that record financial behaviour and demographic data. With a right-skewed distribution suggesting higher credit given to fewer customers, summary statistics show that balance limits vary amongst individuals. The average credit limit is 167550.5, which is different from the median of 140000.0. This suggests that there are differences in the availability or use of credit.

Statistic	Min	1st Qu.	Median	3rd Qu.	Max	Mean
Value	10000.0	50000.0	140000.0	240000.0	1000000.0	167550.5

Table 1: Summary Statistics for Balance Limit

A composite visual analysis of continuous and categorical variables relevant to credit card usage and repayment behaviours. The visual data analysis shows that the sampled credit card users exhibit a wide range of financial behaviours, with a tendency toward lower expenses and payments while acknowledging the existence of high-spending outliers. This variability indicates the complexity of forecasting default risk and how important it is to approach credit risk assessment.

3.2 Analysis Process and Model Selection

Using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as indicators of model fit and complexity, three logistic regression models were compared during the model selection stage (Table 2). AIC and BIC values for the first set of predictors in Model 1 were 27884.47 and 27959.13, respectively. Model 2 showed somewhat higher AIC and BIC values after a few predictors were eliminated, indicating a less frugal fit. With the elimination of less significant predictors, Model 3 has the lowest BIC scores of 27951.43 and the second lowest AIC of 27893.36. Overfitting is a

significant risk that increases with sample size; BIC helps manage this risk by raising the penalty for the number of parameters.

So, in large datasets with nearly 30,000 samples, BIC tends to be a better model selection criterion due to the higher penalties it imposes on model complexity, and it was chosen as the final model for further evaluation.

Model Comparison

Model	AIC	BIC
Model 1	27884.47	27959.13
Model 2	27903.85	27961.92
Model 3	27893.36	27951.43

Table 2: Model Comparison using AIC and BIC

The dataset was cleaned during the finalization of our predictive model, with the predictors being reclassified and the missing values being pruned to improve data integrity. The Generalized Linear Model (GLM) was utilized, and Table 3 presents the important coefficients and their significance. The most significant predictors of credit default risk. These findings offer insightful information that both supports theoretical predictions and adds empirical data from earlier credit risk assessment studies.

Coefficient	**Estimate Value**	**Coefficient**	**Estimate Value**
Intercept	-0.4893	sex	-0.1365
bal_lim	-0.000001342	marriage	-0.2244
bill_sep	-0.000000129	status_sep	0.6989
paid_sep	-0.00001375		

Table 3: Coefficients of the final GLM

3.3 Model Validation and Results

In validating the performance of Model 3, several diagnostic plots were analyzed.

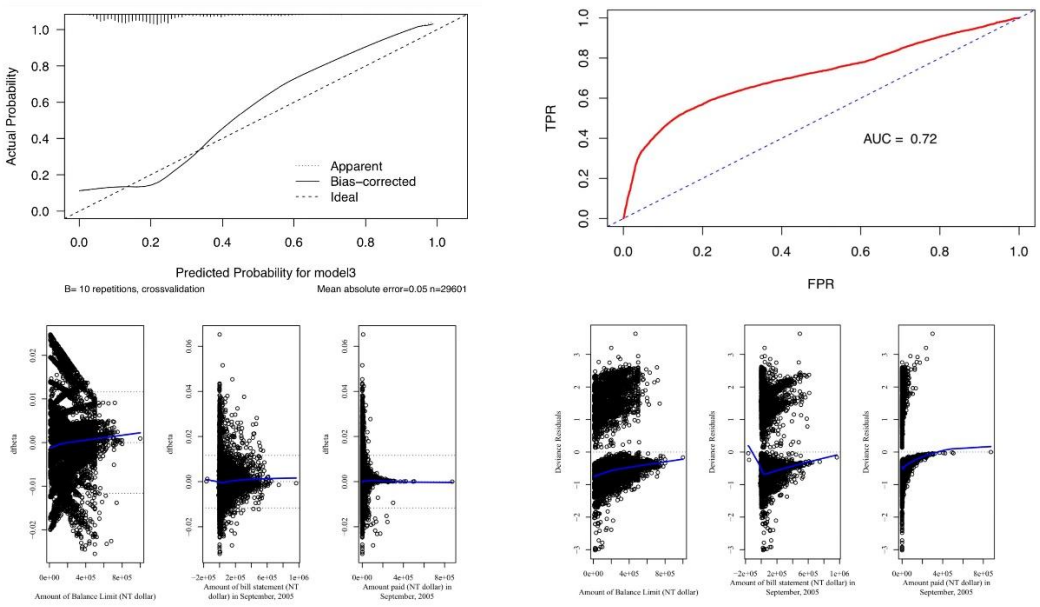

Figure 1: Model Validation Metrics for Credit Card Default Prediction

The calibration curve is shown in the top-left plot and shows how the model was calibrated for various probabilities. The bias-corrected curve shows how overfitting was taken into account.

The ROC curve, shown on the top right, has an AUC of 0.72, indicating a fair discriminative ability for predicting defaults.

The amount of balance limit (left) and amount paid (middle) dfbetas plots in the bottom row show how different data points affect the regression coefficients. Robustness was ensured because no single data point was found to have an excessive impact on the model. The deviance residuals for the balance limit and amount paid, which evaluate the model fit, are shown in the bottom-right plot. The residuals' comparatively random distribution around the 0 line indicates a good fit devoid of obvious systematic outliers.

bal_lim	bill_sep	paid_sep	sex	marriage	status_sep
1.210930	1.199862	1.102257	1.004228	1.013576	1.082343

Table 4: Variance Inflation Factor (VIF) value of the predictors of the final model

When considering how well these predictors explain the variability in the response variable, it appears that all of them have VIF values considerably less than 5. This is good news for the model's coefficients' reliability because it implies that the estimates of the various coefficients are not unduly impacted by the model's other predictors.

Together, these plots support Model 3's validity and dependability in predicting credit card payment defaults in the context of the study.

IV. Discussion

4.1 Final Model Interpretation and Importance

The coefficients of the final model provide important information about the variables affecting credit card defaults.

'Amount of balance limit' had a coefficient of -0.000001342, indicating that for each unit increase in the balance limit, the probability of a credit card default is slightly reduced, though the effect is minor. It can be inferred from this that clients with higher balance limits are less likely to default, possibly due to their more stable financial circumstances.

The 'amount paid in September' coefficient of -0.00001375 indicates a significant decrease in the probability of default with increased payments, emphasizing the importance of payment magnitude on creditworthiness.

'Marriage', the categorical predictor, has a coefficient of -0.2244, indicating that married clients are less likely to default than single clients. This could be because of shared financial responsibilities or other socio-economic factors related to married status.

Better risk management and consumer credit education are guided by the model's confirmation that a number of factors, including a cardholder's background and payment history, significantly influence credit default risk.

4.2 Limitations of the Analysis

Despite its strength, our model has certain drawbacks that may limit its applicability. The omission of interaction terms between predictors, which might offer more information about the dynamics between various variables, is one possible problem. By looking into these interactions, the model could be enhanced to capture more intricate relationships.

Based on historical data, the model predicts that credit behavior in the future will resemble that of the past. Economic conditions and consumer behavior patterns are dynamic in reality, which could eventually make the model less predictive. To keep the model's predictive power intact, it must be validated using current data and updated continuously.

Appendix:

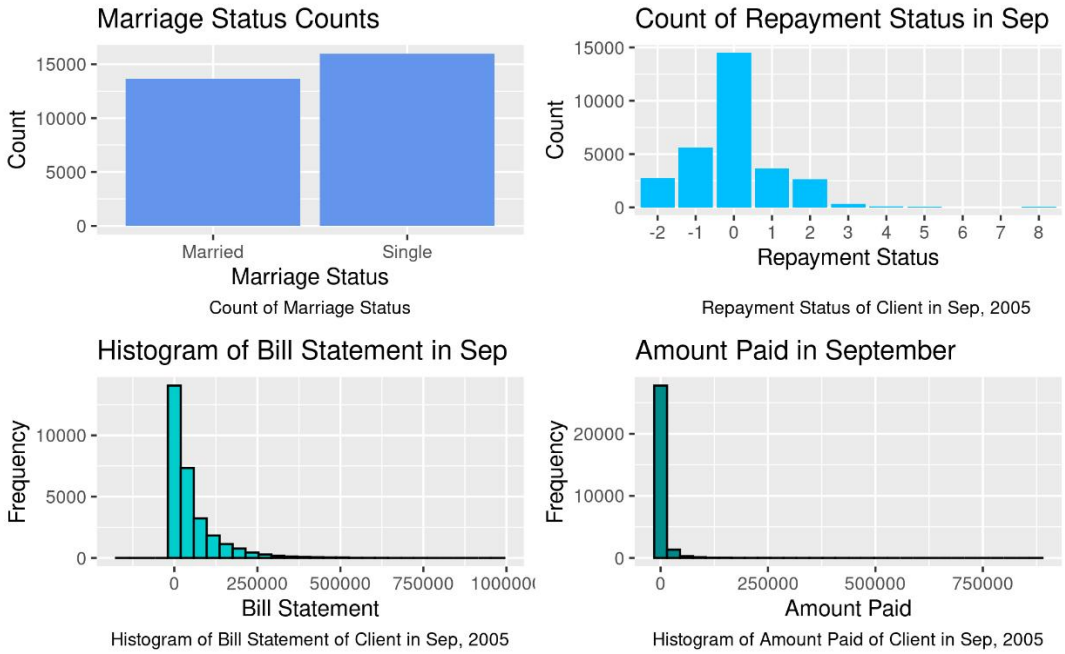

Figure 1: Exploratory Data Analysis (EDA) plots of the predictors

Reference:

Equifax. (2023, December 5). *Missed Payments and Financial Strain Are Increasing for Many Canadians, Credit Card Debt Continues To Rise*. Equifax® Canada Market Pulse - Consumer Quarterly Credit Trends Report. https://www.consumer.equifax.ca/about-equifax/press-releases/-/blog/missed-payments-and-financial-strain-are-increasing-for-many-canadians-credit-card-debt-continues-to-rise/

Merikoski, M., Viitala, A., & Shafik, N. (2018, May 18). *Predicting and preventing credit card default.* https://sal.aalto.fi/files/teaching/ms-e2177/2018/McKinseyFinal.pdf

Sharma, S., & Mehra, V. (2018, July). *Default payment analysis of credit card clients.* https://www.researchgate.net/publication/326171439_Default_Payment_Analysis_of_Credit_Card_Clients

Yeh, I.-C., & Lien, C. (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.* Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Çığşar, B., & Ünal, D. (2019). *Comparison of data mining classification algorithms determining the default risk.* Scientific Programming, 2019, 1–8. https://doi.org/10.1155/2019/8706505