DSE Project: Card Default

1. Introduction

In the ever-changing landscape of financial services, the ability to accurately predict and manage credit card defaults is of paramount importance for banks worldwide. The negative impacts of defaults are far-reaching, affecting both financial institutions and their customers. To address this challenge, this project will focus on whether basic background information alone can serve as effective indicators for forecasting credit card defaults.

2. Data Cleaning

This project will leverage a dataset from Yeh et.al. (2009), containing information on credit card defaults in Taiwan in 2005. In the dataset, incongruous data for Education Level (X3) and Marital Status (X4) were manually removed, reducing the sample size from 7000 to 6932. For categorical variables, X6-X11, the amount of data across the various categories was notably uneven. Some categories exhibited merely single-digit occurrences, thereby introducing the potential for imprecise fitted probabilities of 0 or 1. As a mitigating strategy, categories spanning from 1 to 9 on the repayment status scale were merged into a singular category denoted as "payment delay for one month and above." Bill statement variables (X12-X17) were aggregated into "Total Bill Statements from April to September," and X18-X23 into "Total Amount of Past Payments from April to September." The dataset was split into a 70% training set (4853 observations) and a 30% test set (2079 observations).

3. Key Descriptive Statistics

- Examination of variable X1 reveals notable distinctions in both median values and overall trends between default and non-default categories (Figure 1). Specifically, the median given credit for non-default is observed to be 160,000 TWD, whereas the corresponding median for default is 90,000 TWD. Additionally, the 25th and 75th percentiles for non-default cases surpass those for defaults, suggesting that the variable X1, denoting the given credit, may serve as an indicative factor for default propensity.
- For variable *Total Bill Statements from April to September* (I(X12+X13+X14+X15+X16+X17)), it is not obvious to see big differences in the median, 25th, and 75th percentile of the total amount of bill statements from April to September (Figure 2). However, for some defaults, the billing statement is negative, which is strange in the sense that this may indicate the default cardholder overpays the bill.

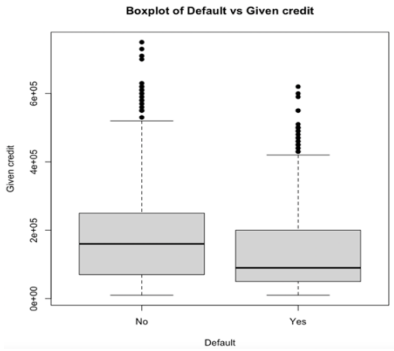

Figure 1

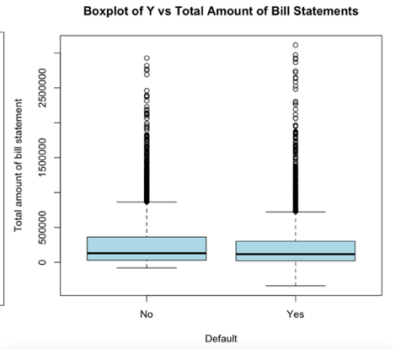

Figure 2

- For variables from X6 to X11, the number of payment delays for any month for the default is greater than that for the non-default. The example of the bar plot for X6 (Figure 3) can visualize the "strange" trend. Given that the total number of non-defaults is much greater than the default, this observation suggests a plausible association between payment delays and the likelihood of default.

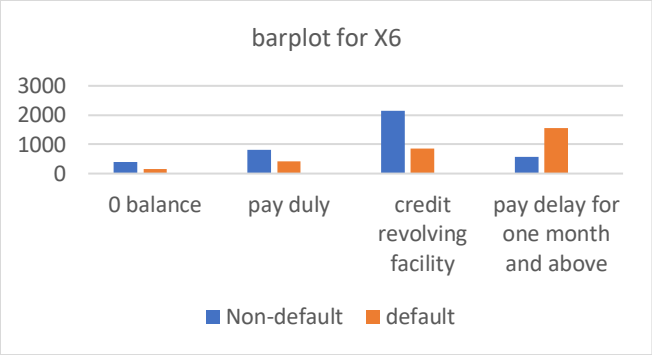

Figure 3

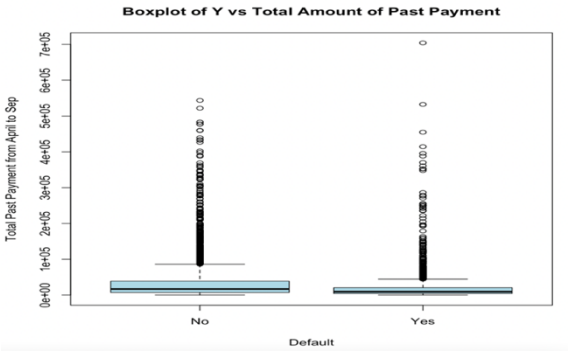

Figure 4

- The variable denoted as the *Total Amount of Past Payments from April to September* (I(X18+X19+X20+X21+X22+X23)), exhibits a discernible discrepancy in the range between default and non-default instances (Figure 4). Notably, the range for defaults appears considerably narrower than that observed for non-defaults, barring outliers. This observation suggests that the cumulative amount of payments made over the specified period may serve as an indicator for predicting default occurrences. Additionally, it is significant to note the presence of numerous outliers which may affect the robustness and efficacy of predictive models.

4. Result

4.1 Logistic model

For analyzing factors that will predict Card Default, this project uses a logistic model with variables (X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, I(X12+X13+X14+X15+X16+X17), I(X18+X19+X20+X21+X22+X23)). Considering the purchasing power of TWD is relatively low compared with other currencies, and the range of variables in terms of TWD is large (X1, I(X12+X13+X14+X15+X16+X17), I(X18+X19+X20+X21+X22+X23)), these variables will be scaled down by 10000. The significance level chosen for this logistic model is 0.05.

For this logistic model, according to the confusion matrix (refer to Table 1 and Table 2), the training error is approximately 27.69%, while the test set manifests a slightly higher error of 28.63%. It is discerned from these metrics that the model's accuracy is not perfect. Nevertheless, the discerned patterns in relation to card default can offer some valuable insights considering over 70% of cases are correctly classified in the model.

In the analysis conducted (refer to Table 3), it can be deduced from the statistical significance tests that both the given credit variable (X1) and the total past payment variable (I(X18+X19+X20+X21+X22+X23)) exhibit significance. In the case of variables from X6 to X11, a mixed pattern emerges, with some variables demonstrating statistical significance while others do not. Notably, variables I(X6=credit revolving facility in Sep),

I(X6=payment delay for one month and above in Sep), I(X7=credit revolving facility in August), and I(X7=payment delay for one month and above in August) display exceptionally small p-values in comparison to other variables, including those that are statistically significant. This observation suggests that the repayment status of the credit revolving facility and the occurrence of payment delays for one month and above in the most recent months are pivotal in determining default status.

	No	Yes
False	2385	955
True	389	1124

Table 1: Confusion Matrix for Training Set

	No	Yes
False	981	409
True	186	502

Table 2: Confusion Matrix for Test Set

Regressor	Pr(>|z|)	Significance
I(X1/10000)	1.323474e-07	Significant
I(X6=credit revolving facility in Sep)	1.145111e-06	Significant
I(X6=payment delay for one month and above in Sep)	2.176160e-06	Significant
I(X7=pay duly in August)	2.401516e-02	Significant
I(X7=credit revolving facility in August)	4.904182e-04	Significant
I(X7=payment delay for one month and above in August)	3.819917e-03	Significant
I(X8=payment delay for one month and above in July)	3.174640e-02	Significant
I(X9=payment delay for one month and above in June)	4.083790e-02	Significant
I(X10=pay duly in May)	1.197404e-02	Significant
I(X10=credit revolving facility in May)	8.452944e-02	Significant
I((X18 + X19 + X20 + X21 + X22 + X23)/10000)	7.462258e-07	Significant

Table 3: Statistically Significant Variables

4.2 Naïve Bayes

The application of Naïve Bayes in this project yielded slightly inferior performance with AUC=0.74 (Figure 6) compared to the logistic model with AUC=0.75 (Figure 5). Notably, when employing variables X12 to X17 (The amount of bill statement) and X18 to X23 (the amount of past payment) separately and in combination with Total Bill Statement from April to September and Total Past Payment from April to September, differential behaviors were observed. Specifically, the latter configuration exhibited a slightly superior performance with an AUC of 0.74 (Figure 6), while the AUC for utilizing variables separately was 0.72 (Figure 7). These findings suggest dependencies in the predictive efficacy of the Naïve Bayes model based on the treatment of bill statements and past payment variables. This distinction underscores the potential significance of considering the amount of the billing statement and the past payment over an extended duration when forecasting default likelihood.

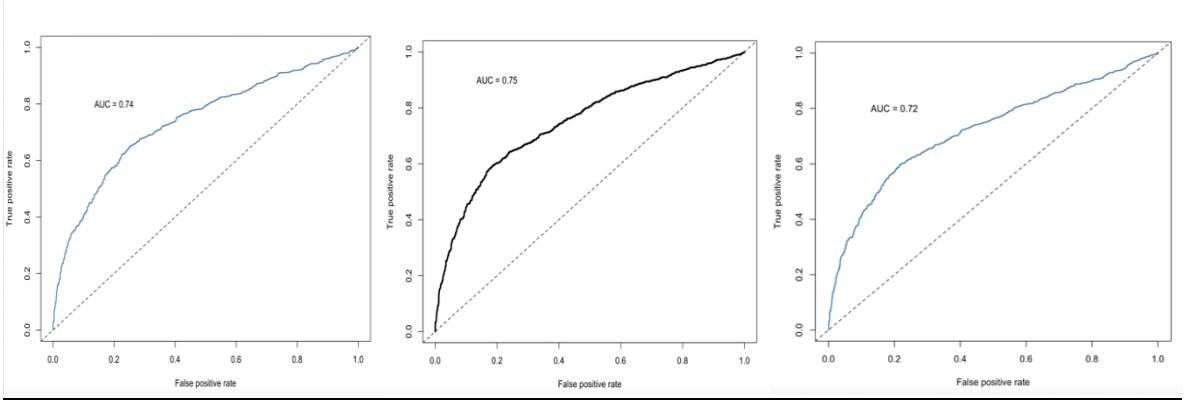

Figure 5: ROC for Naïve Bayes with *Total Bill Statement* and *Total Past Payment* (combined variables)

Figure 6: ROC for logistic model with *Total Bill Statement* and *Total Past Payment* (combined variables)

Figure 7: ROC for Naïve Bayes with X12-X17 and X18-X23 (separated variables)

4.3 Decision Tree

In congruence with the logistic model, the significance of repayment statuses in August (X7) and September (X6) is notably pronounced, evident by their positioning in the upper echelons of the decision tree (Figure 8). Furthermore, consonant with the graphical representation elucidated in the boxplot (Figure 2), the aggregate sum of past payments emerges as a pivotal determinant in the prediction of default given it appears twice in the decision tree (Figure 8). The decision tree analysis reveals a discernible trend wherein individuals with elevated Total Past Payment values from April to September tend to be categorized as non-default; conversely, for variables pertaining to the History of Past Payment (X6, X7, X9), individuals characterized by a measurement scale denoting payment delay are predisposed to be predicted as default.

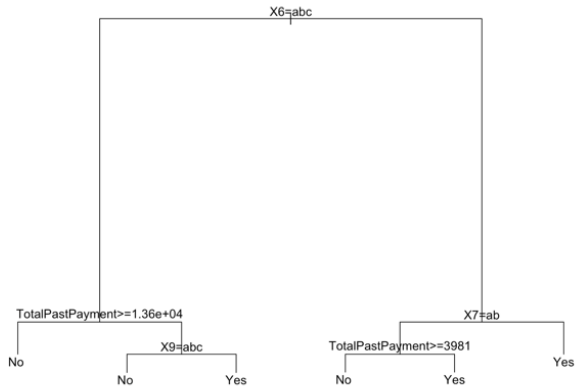

Figure 8: the decision tree

(a=0 balance; b=pay duly; c=credit revolving facility; d=pay delay for one month and above)

5. Limitation and Discussion

The negative billing statement in key statistics still lacks explanation after modeling. While merging variables (Total Past Payment & Total Bill Statement) alleviates huge variance, outliers persist, hindering robust model formation. Exploring alternatives, like assigning weights to variables based on temporal proximity, could be a potential avenue for further refinement. Combining repayment status categories from 1-9 as payment delay of one month and above may oversimplify; detailed categorizations reflecting time intervals might improve predictions, but additional data is needed for validation.

6. Conclusion

The study employs logistic model, Naïve Bayes, and decision tree models to predict credit card defaults. Results highlight the crucial role of financial variables, specifically given credit, total past payments, and payment status, in forecasting defaults. Surprisingly, personal information like gender, education, and marital status holds limited significance. The project urges financial institutions to prioritize billing and payment data over personal details in predictive models.