Introduction and Research Question

As economies become increasingly cashless, the use of credit cards has increased substantially. Credit card defaults, if left unmanaged, can impose significant credit losses on banks, making accurate forecasts essential. However, predictive models often rely on detailed financial histories that may not always be available. This project investigates whether using only basic background information can provide meaningful insights into default outcomes. Thus, we investigate the effectiveness of various machine learning models in predicting the outcome of a binary response variable from a set of continuous and categorical predictors. Using a 70/30 train/test split (4900 training and 2100 testing observations), K-Nearest Neighbours (KNN), Naive Bayes (NB) and Decision Tree models are compared against a logistic regression model which serves as a benchmark. NB model performed best (AUC 0.76) while KNN performed worst (AUC 0.7).

Visualisation of Key Variables

The bar charts in Figure 1 illustrate how customers' past payment behaviour relates to their likelihood of default. Default proportions (in turquoise) rose significantly among customers who had payment delays of two months and above (codes 2 to 9), while it remained low among customers who had zero balance, paid duly or used credit revolving facility (codes -2 to 0). This suggests that Repayment Status history is a strong indicator of default risk.

Figure 1: Bar Plot of Proportion of Default by Repayment Status from Apr to Sep

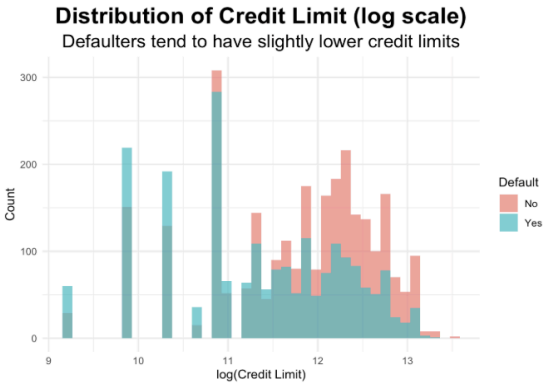

Figure 2: Histogram of credit limit (log_X1)

In Figure 2, the histogram of credit limits compares the distribution of credit extended by the bank to each customer between defaulters and non-defaulters, captured by the variable log_X1 (log-transformed credit limit). It reveals that defaulters generally hold lower credit limits than non-defaulters, suggesting that higher-risk or lower-income borrowers with smaller credit lines are more prone to default.

Figure 3 shows the proportion of defaulters across categories of each categorical variable. Differences between genders (X2) and marital statuses (X4) were minimal, but there was a clearer trend observed across education levels (X3). The proportion of

defaulters decreased as education level rose from High School to Graduate School, possibly due to the positive correlation between years of formal education and level of financial literacy (Ansar et al., 2023). Higher financial literacy would facilitate better financial decision-making and hence reduce likelihood of default. Thus, subsequent models should include repayment status, credit limit and education level as key predictors, as they demonstrate strong associations with default risk.

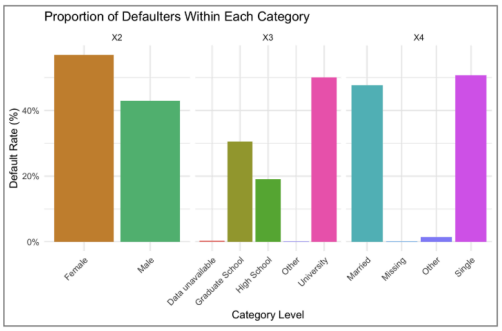

Figure 3: Bar Plot of Proportion of Defaulters within Each Category

Processing of Key Continuous Variables

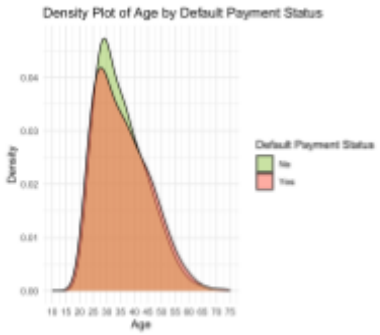

Figure 4.1 Density plot of X5

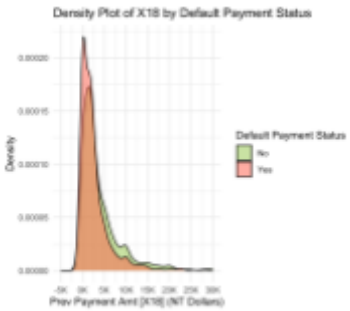

Figure 4.2 Density Plot of X18

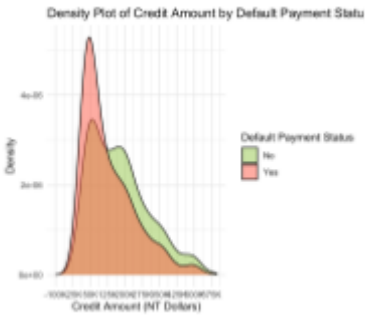

Figure 4.3 Density plot of X1

Figures 4.1-4.3 show that age (X5), and payment amounts (represented by X18) and credit limit (X1) are all right-skewed due to a small number of customers with very high values. To address this skewness,these continuous variables were log-transformed to reduce the influence of extreme observations and produce distributions that are more suitable for modelling techniques like logistic regression.

The heatmap in Figure 5 shows a distinct block of strong correlations among X12-X17, the monthly bill statement amounts in six consecutive months. This indicates substantial redundancy as customers with high bills in one month typically have high bills in subsequent months. Including all six variables would add little predictive value and may introduce multicollinearity. To address this, these variables were consolidated into a single median measure that summarises overall bill burden while being more robust to extreme values. A log transformation was then applied to form log_med_bill to reduce skewness.

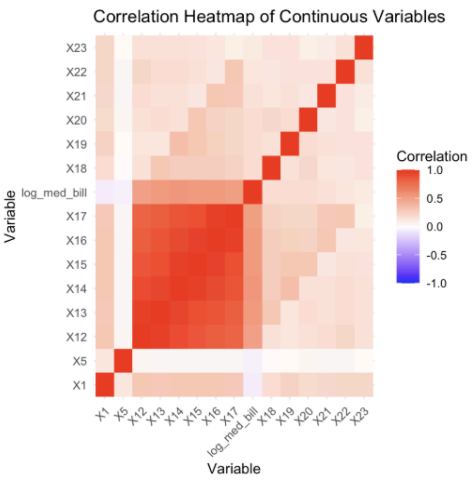

Figure 5: Correlation Heatmap of Continuous Variables (with log_med_bill)

Logistic Regression Model

The logistic regression model was chosen as the benchmark model for this analysis, providing a baseline against which other models can be compared. Predictor selection was performed using the forward selection process, which sequentially adds variables to improve model fit while penalising overfitting. This procedure identified 11 key predictors: log_X1 (credit limit), X3 (education), log_X5 (age), X6, X9, X11 (repayment status), log_X18-21 (payment amounts) and log_med_bill, as the most significant contributors to predicting default.

K-Nearest Neighbours (KNN)

As KNN measures the Euclidean distance between values, categorical predictors would not yield meaningful results and hence only continuous variables log_X1, log_X5, log_X18-21, and log_med_bill were used. After variable selection, the variables were standardised to ensure that no single predictor would dominate distance calculations. LOOCV was used to find the k-value that minimised misclassification. With the selected variables, scaled data and k = 24, the KNN model was used to predict test data.

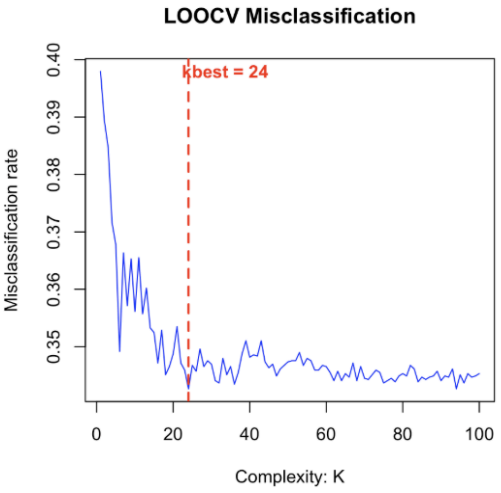

Figure 6: LOOCV Misclassification graph

Naive Bayes (NB)

The Naive Bayes model was trained using all log-transformed continuous variables and categorical predictors. Due to the assumption that predictors are conditionally independent given the class, the Naive Bayes model does not require formal variable selection.

Decision Tree

Since decision trees only select useful variables during splitting, no explicit variable selection was required. Although some continuous variables (namely X12-X17) were strongly correlated, they were not combined into a single variable, since decision trees inherently handle multicollinearity by selecting only the most informative variables at each split. Log transformations were not applied, as decision trees are invariant to monotonic transformations and interpretability of split points is clearer on the original scale (e.g. X20 < 175 is easier to interpret than log_X20 < 2.24).

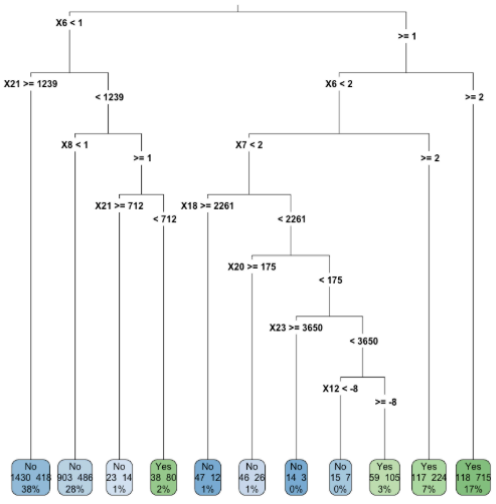

Figure 7: Decision Tree

To ensure that the model could explore all possible patterns, the initial tree was grown with minimal restrictions (cp = 0.001 and maxdepth = 30), resulting in a tree with 158 leaves. This tree was then pruned using the best cp value (determined by LOOCV). This pruning process produced the decision tree seen in Figure 7, which has 11 leaves.

Evaluation and Comparison of Model Performance

ML Model	Logistic Regression	K-Nearest Neighbours	Naive Bayes	Decision Tree
AUC	0.75	0.7	0.76	0.74

Table 1: Comparison of AUC values for Logistic Regression, KNN, NB and Decision Tree models.

The AUC values of the models ranged between 0.70 and 0.76 (Table 1), indicating moderate discriminative ability in predicting credit card default. While the models perform better than random guessing, they are not yet strong enough for high-stakes applications such as credit card approval, where false negatives (predicting no default when a default occurs) carry substantial cost for banks.

Logistic regression already performs strongly (AUC = 0.75), likely because the key predictors in the dataset repayment history variables (X6, X9, X11), outstanding bill burden (log_median_bill), and repayment amounts (log_X18-21) are intuitively closely linked to repayment behaviour. These predictors already separate high-risk and low-risk borrowers well, so a model more flexible than logistic regression was unlikely to yield a significantly larger improvement.

The KNN model performed the worst (AUC = 0.7), despite using the optimal k value that minimised misclassification. The poorer performance for the model was primarily due to the differences in the number and type of predictors used. The KNN model used only 7 continuous variables, compared to 11 in logistic regression, reducing the information available for class separation.

The Naive Bayes model performed the best (AUC = 0.76) and was the only model that performed better than logistic regression. One likely reason is that NB assumes independence between predictors. Before preprocessing, the six monthly bill variables (X12–X17) were highly correlated, which could cause NB to overweight them. Combining them into a single log_med_bill variable reduced this redundancy. All models used this improved feature set, but NB benefited more because the predictors better aligned with its independence assumption, improving its accuracy and performance.

The decision tree model achieved an AUC of 0.74, which is only slightly lower than logistic regression. This indicates that the tree was able to capture most of the useful signal in the data, particularly the strong relationship between recent repayment status (X6 -X11) and default risk, which appears repeatedly in the main splits of the pruned tree. However, after capturing the major

distinctions, the tree added splits for very small groups, reflecting training-set noise rather than consistent customer behaviour, which reduced generalisation. Because the strongest predictors in the dataset (such as repayment status) follow fairly direct relationships with default risk, logistic regression was able to model them effectively with a simpler structure, while the decision tree's additional complexity did not translate into better performance.

Limitations

A key limitation of this study lies in the context of the dataset, which is drawn from credit card customers in Taiwan in 2005, reflecting the regulatory environment, economic conditions and credit market structure of that period in. Changes in these factors since then would reduce the various ML models' applicability to modern or non-Taiwan contexts. In addition, the data provide only a static six-month snapshot of customer behaviour, rather than a longer-term panel, so the ML models cannot capture longer-run dynamics such as gradual changes in income, credit limits or repayment behaviour.

Another limitation lies in the predictors available. The dataset only includes basic demographics and credit card activity, but excludes key financial information such as income, employment stability, and household wealth. The data set also lacks economic variables that are known to influence borrower default behaviour, such as economic downturns, rising unemployment or tightening credit conditions (Podpiera & Ötker, 2010). Without these variables, the models can only detect risk once problematic credit card activity has already occurred, limiting the models' accuracy. Incorporating richer personal and macroeconomic variables would likely improve accuracy and capture earlier signs of default risk.

Concluding remarks

Returning to the research question, the results show that basic demographic and limited financial attributes do provide meaningful predictive insight into default risk, but only to a moderate degree. While the models performed better than chance, the absence of richer predictors limited overall predictive strength. Future work should incorporate additional temporal, micro-financial and macroeconomic data to enhance generalisability, and consider ensemble methods like Random Forests or boosting algorithms to further strengthen performance while maintaining interpretability.

An interesting point to note is that the Logistic Regression performed on par with or better than more complex models like Decision Tree and Naive Bayes. Despite its simplicity, it could capture key relationships effectively, likely because default risk is driven by mostly linear factors such as credit limit and repayment history. This shows that with good variable selection and preprocessing, a simple model can match the performance of more sophisticated methods.

Appendix

References

1. Ansar, S., Klapper, L., & Singer, D. (2023). *The Importance of Financial Education for the Effective Use of Formal Financial Services*. https://documents1.worldbank.org/curated/en/099346003072335535/pdf/IDU06db1d7f504f4e04bdc086250f256d40d4253.pdf
2. Ötker-Robe, İ., & Podpiera, J. (2010). The Fundamental Determinants of Credit Default Risk for European Large Complex Financial Institutions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1750682