Investigation of factors for Credit Card Default Prediction

Di Wu

Data Analytics Department, Denison University

1. Abstract

In this project, the likelihood of credit default is predicted using a dataset containing various client features, such as demographics, credit limit, and payment history. Three predictive models are built and compared: decision tree, k-nearest neighbors (kNN), and logistic regression. The analysis reveals key factors influencing credit default risk, and the performance and limitations of the models are assessed.

2. Introduction

Credit default risk is a critical factor for financial institutions when assessing creditworthiness and making decisions about credit approval or monitoring. Accurately predicting default risk can help minimize losses and optimize lending practices. This study aims to build and compare predictive models for credit default risk using a dataset containing client information such as gender, education, marital status, age, payment history, and bill statements.

The dataset used in this study contains information on credit default risk for a sample of clients. It contains 30000 observations on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are 25 variables, including both categorical and numerical features, and a total of N observations, where N represents the number of clients in the dataset. The variables are as follows:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_x: Repayment status for last 6 months (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
- BILL_AMTx: Amount of bill statement for last 6 months (NT dollar)
- PAY_AMTx: Amount of previous payment for last 6 months (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

The dataset contains a mix of demographic, financial, and behavioral variables, which can be used to build predictive models for credit default risk. The target variable, "default.payment.next.month," is a binary categorical variable indicating whether a client is expected to default on their payment in the next month.

3. Research hypothesis

The research hypothesis for this study is that there is a relationship between the input features (e.g., LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_x, BILL_AMTx, PAY_AMTx) and the target variable (default.payment.next.month). The hypothesis aims to investigate whether the selected input features can be used to predict the likelihood of a client defaulting on their payment in the next month.

Null Hypothesis (H0): The null hypothesis states that there is no relationship between the input features and the target variable. In other words, the selected input features do not have any predictive power in determining the likelihood of a client defaulting on their payment in the next month.

Alternative Hypothesis (H1): The alternative hypothesis states that there is a relationship between the input features and the target variable. This means that the selected input features can be used to predict the likelihood of a client defaulting on their payment in the next month.

The null hypothesis (H0) and alternative hypothesis (H1) will be tested using the decision tree, k-nearest neighbors (kNN), and logistic regression models. If the models demonstrate statistically significant predictive power, it will provide evidence in favor of the alternative hypothesis (H1), suggesting that the input features can be used to predict credit default risk. If the models do not show statistically significant predictive power, it would fail to reject the null hypothesis, indicating that there is no relationship between the input features and the target variable.

4. Methodology

In this study, three different predictive modeling techniques are employed to assess the relationship between the input features and the target variable (default.payment.next.month): decision tree, k-nearest neighbors (kNN), and logistic regression. These methods were chosen for their ability to handle categorical and numerical input features and their widespread use in classification tasks.

1) Decision Tree: A decision tree is a hierarchical, tree-like structure that recursively splits the data into subsets based on the values of the input features. The tree is built by selecting the feature that best splits the data at each node, maximizing the separation of the target variable's classes. The decision tree model is easily interpretable and can handle both categorical and numerical features. In this study, the ‘rpart’ package in R is used to build the decision tree model, and the complexity parameter (cp) is set to control the tree's size and prevent overfitting.

2) k-Nearest Neighbors (kNN): kNN is a non-parametric classification method that assigns a new observation to the majority class of its k-nearest neighbors in the feature space. The kNN model does not make any assumptions about the data's distribution and can be used with both categorical and numerical features. In this study, the ‘knn’ function from the ‘class’ package in R is used to implement the kNN model. The optimal value of k (the number of neighbors) is determined using cross-validation on the training dataset.
3) Logistic Regression: Logistic regression is a parametric method used for binary classification tasks. It models the probability of the target variable belonging to a particular class as a function of the input features using a logistic function. Logistic regression can handle both categorical and numerical features but requires preprocessing steps, such as encoding categorical variables and scaling numerical variables. In this study, the ‘glm’ function from the ‘stats’ package in R is used to build the logistic regression model, and a threshold for classifying default probabilities is determined using cross-validation.

The dataset is split into training and testing sets, with the training set used to build the models and the testing set used to evaluate their performance. Accuracy, the proportion of correctly classified instances, is used as the main evaluation metric. The models are compared based on their accuracy scores, and the significance of the input features is assessed using variable importance measures and coefficient estimates.

5. Data exploration and visualization

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, leading to unreliable and unstable estimates of the regression coefficients. In the given dataset, the heatmap shown in *figure 1* of the 25 variables reveals multicollinearity between the repayment status variables (PAY_x) and the amount of bill statement variables (BILL_AMTx).

1) Repayment Status (PAY_x) Variables: The repayment status variables (PAY_0 to PAY_6) represent the client's payment history for the last six months. A high correlation between these variables, such as the 0.6 correlation between PAY_0 and PAY_2, suggests that a client's repayment status in one month is strongly associated with their repayment status in other months. This multicollinearity may cause issues in estimating the regression coefficients and determining the individual importance of each repayment status variable in the prediction models.
2) Amount of Bill Statement (BILL_AMTx) Variables: The amount of bill statement variables (BILL_AMT1 to BILL_AMT6) represents the client's bill statement amount for the last six months. The heatmap shows a high correlation between these variables, with BILL_AMT1 having a 0.95

correlation with BILL_AMT2, indicating a strong association between the bill amounts in consecutive months. This multicollinearity can lead to unstable estimates of the regression coefficients and make it difficult to assess the individual contribution of each bill amount variable in the predictive models.

To address the issue of multicollinearity, several approaches can be considered:

1) Variable Selection: Removing one of the highly correlated variables from the model can help reduce multicollinearity. For example, if PAY_0 and PAY_2 are highly correlated, only one of them can be included in the model. The choice of which variable to retain can be based on domain knowledge, the correlation with the target variable, or other feature selection methods.
2) Creating Composite Variables: Instead of using the original variables, composite variables that capture the common information can be created. For example, the average repayment status (avg_pay_time) and the average bill amount (avg_crd_bill) across the six months can be calculated and used as input features in the models. These composite variables may help mitigate the effects of multicollinearity while still incorporating the relevant information from the original variables.

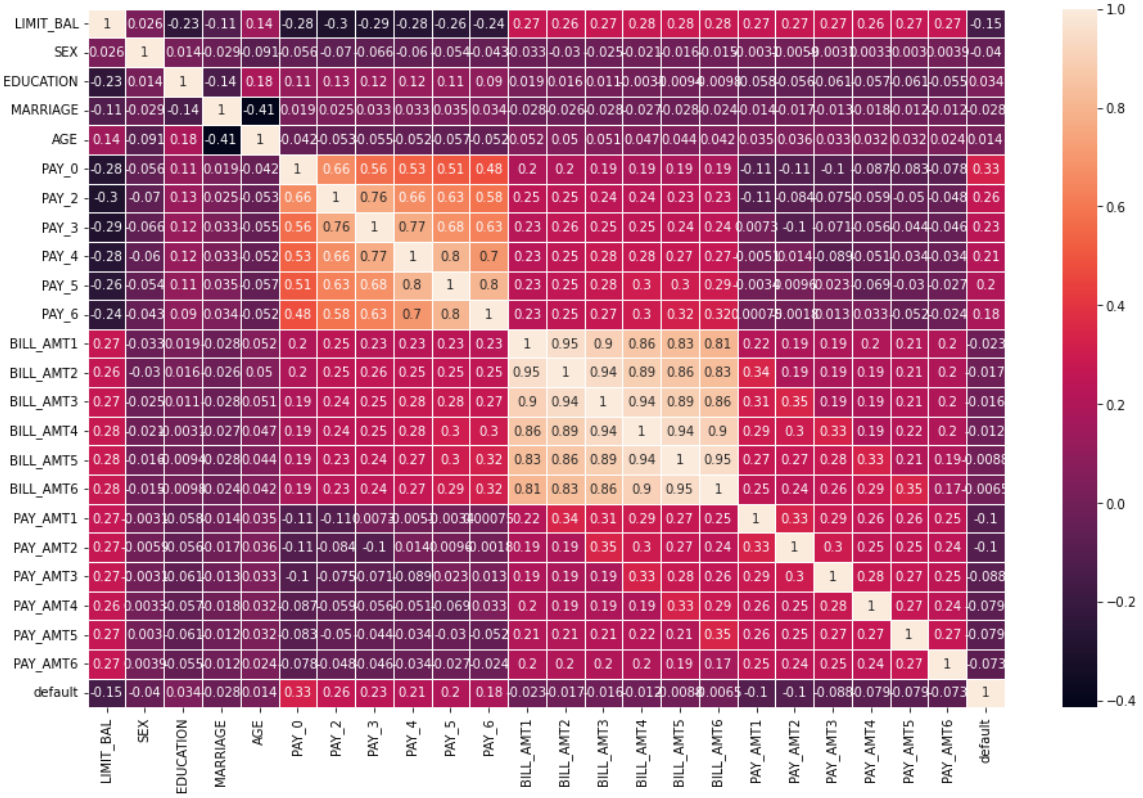

Figure 1-heatmap for all features

The bar plot of the proportion of education level by marriage and gender combination provides insights into the default behavior across different demographic groups. The plot reveals that for high school, university, and graduate school education levels, the default pattern remains consistent across all combinations of marriage and gender. In these groups, more than 75% of the clients do not default on their payments (default = 0). This observation suggests that clients with high school, university, and graduate school education levels demonstrate similar credit behavior, regardless of their marital status and gender.

On the other hand, clients with "others" as their education level exhibit a different trend, with over 90% of them not defaulting on their payments. This group, which represents clients with alternative educational backgrounds or those who have not provided details about their education, displays a lower default rate compared to the other education levels. This finding might suggest that clients in this category have distinct credit management practices or financial circumstances that result in lower default rates.

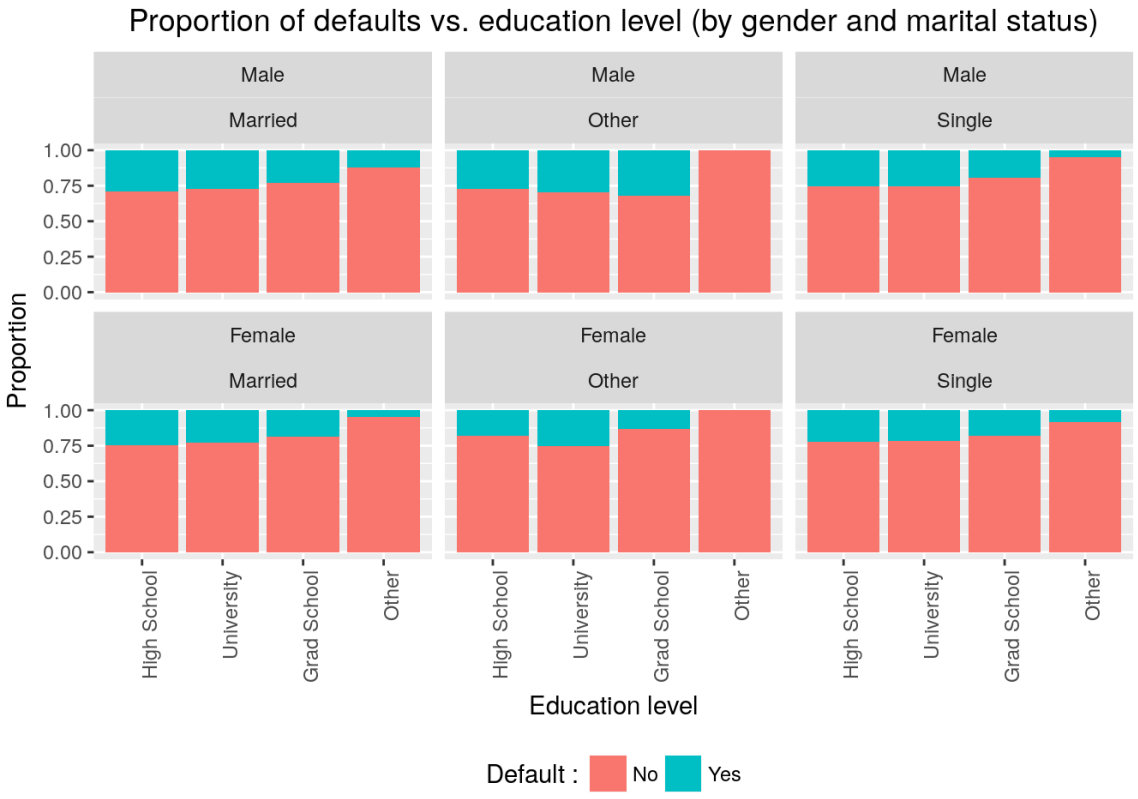

Figure 2- Proportion of education level by gender and marital status

The bar plot of limit balance by default payment next month provides insights into the distribution of credit limits and their relationship with default behavior. The plot reveals that most clients do not default on their payments (default = 0), and among those who do not default, the majority have a limit balance of less than 250,000. Furthermore, a significant portion of these clients, approximately 6,000 individuals, have a limit balance of 100,000. This observation suggests that clients with lower credit limits are more likely to manage their credit responsibly and avoid defaulting on their payments.

The high number of clients with a limit balance of 100,000 could be due to various factors, such as financial institutions offering this credit limit as a common starting point for new credit card holders or clients with limited credit history. Additionally, clients with lower credit limits might be more cautious about their credit usage and repayment, resulting in a lower likelihood of default.

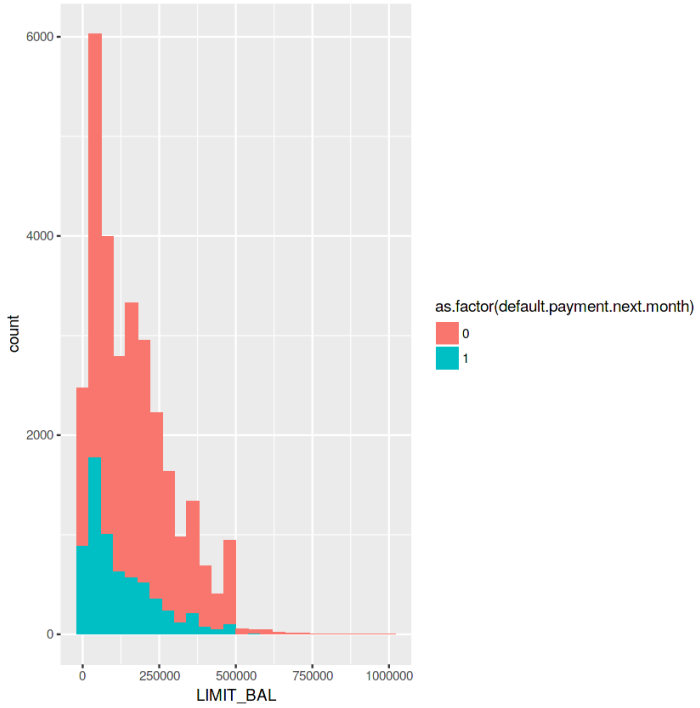

Figure 3-limit balance by default payment next month

6. Statistical analysis

The decision tree model was built using a complexity parameter (cp) of 0.00119, which controls the size of the tree and helps prevent overfitting. The dataset was split into a 70% training set and a 30% test set for cross-validation purposes. This process ensures that the model's performance is assessed on previously unseen data, providing a more reliable estimate of its accuracy.

The decision tree model achieved an accuracy of 0.8034226, indicating that it correctly classified approximately 80.34% of the instances in the test set. This performance level suggests that the decision tree model can effectively predict the default payment behavior of clients based on the selected input features.

The variable importance result reveals the relative contribution of each input feature in the decision-making process of the tree model. In this case, the average payment time (avg_pay_time) is the most important feature, with a value of 1,220.804. This implies that the payment time has the most significant impact on determining the likelihood of a client defaulting on their payment.

The other variables, in descending order of importance, are average credit bill amount (avg_crd_bill) with a value of 126.6485, average payment amount (avg_pay_amt) with a value of 95.43764, limit balance (LIMIT_BAL) with a value of 24.13638, age (AGE) with a value of 11.45028, marital status (MARRIAGE) with a value of 0.1751429, and education (EDUCATION) with a value of 0.03986686. These values indicate that the average credit bill and payment amounts are the next most critical factors influencing the default payment behavior, followed by the client's limit balance and age. Marital status and education have the least impact on default payment behavior in this decision tree model.

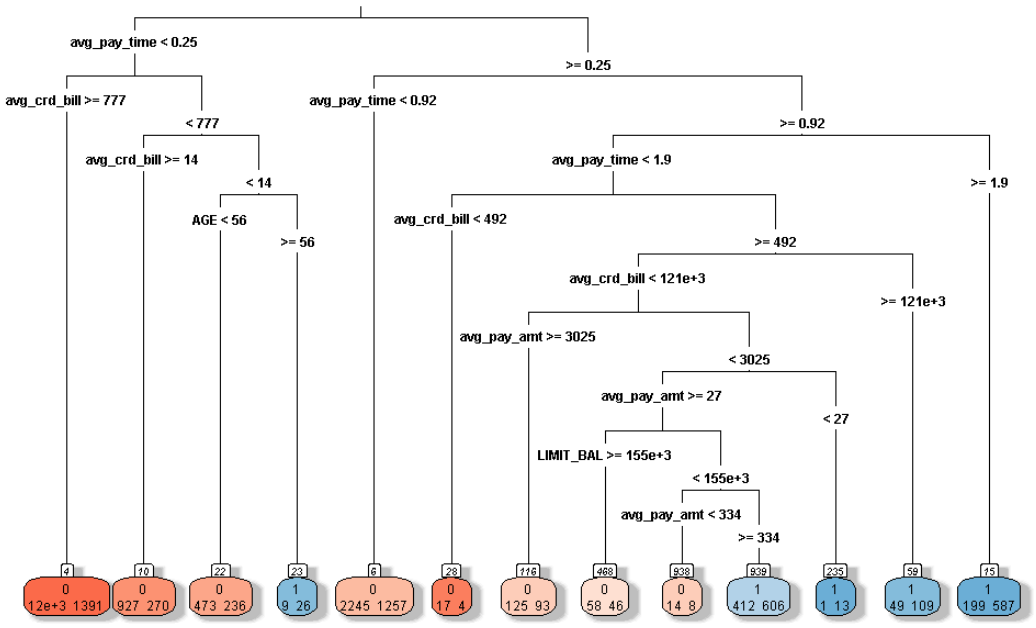

Figure 4- Decision tree plot (cp=0.00119)

The k-Nearest Neighbors (kNN) model, when applied with k=69, yields the highest accuracy of 0.7796422. This indicates that the kNN model with 69 nearest neighbors can correctly classify approximately 77.96% of the instances in the test set. However, it is important to consider that the kNN algorithm can be challenging to implement when dealing with a mix of categorical and numeric variables, as is the case with the given dataset.

To address this issue, only numeric variables were used in the kNN model, specifically limit balance, age, average credit bill amount (avg_crd_bill), average payment amount (avg_pay_amt), and average payment time (avg_pay_time). By excluding categorical variables such as sex, education, and marital status, the model's accuracy may have been affected, as these variables might also have an impact on the default payment behavior.

It is worth noting that the kNN model's accuracy of 0.7796422 is lower than the decision tree model's accuracy of 0.8034226. This difference could be attributed to the exclusion of the categorical variables in the kNN model, which might have provided valuable information for predicting default payment behavior. Furthermore, the kNN model's accuracy could also be influenced by the choice of k, the distance metric, and the scaling of the numeric variables.

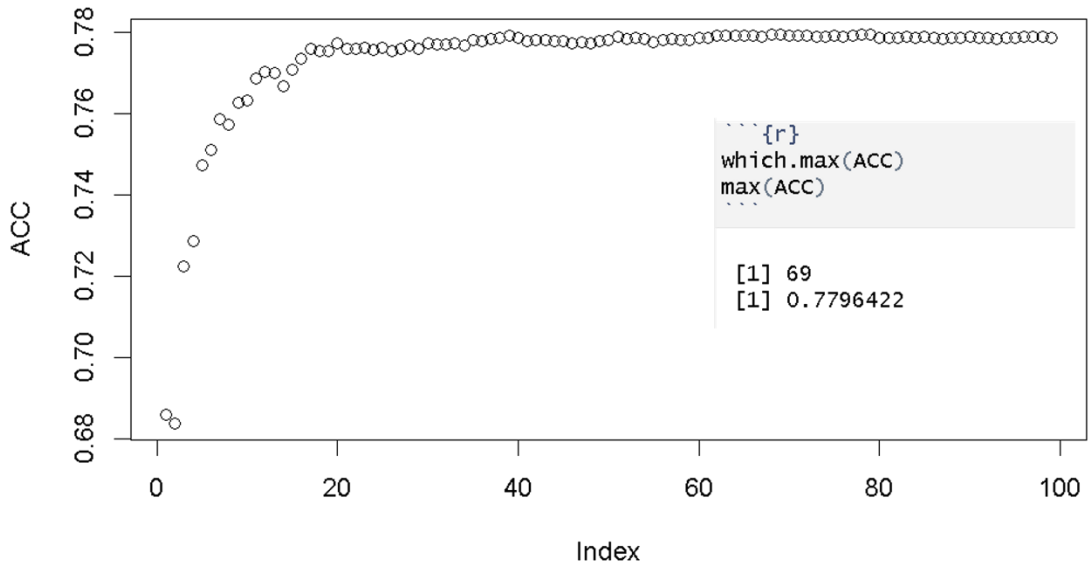

Figure 5- KNN accuracy vs number of k plot

The logistic regression model achieved an accuracy of 0.8048672 when the regularization parameter (q) was set to 0.357, indicating that it correctly classified approximately 80.49% of the instances in the test set. The logistic regression model's accuracy is slightly higher than the decision tree model's accuracy of 0.8034226 and notably higher than the kNN model's accuracy of 0.7796422 (with k=69 and only numeric variables). This improvement in accuracy could be attributed to the logistic regression model's ability to incorporate both categorical and numeric variables, which allows for a more comprehensive understanding of the relationships between the input features and the target variable (default.payment.next.month).

Moreover, the logistic regression model provides the added advantage of estimating the probability of default for each client, which can be useful for risk assessment and decision-making processes in credit management. However, it is essential to note that the choice of the regularization parameter (q) can impact the model's performance and should be carefully selected, usually through techniques like cross-validation, to prevent overfitting or underfitting.

This table shows the estimated coefficients (or weights) for each input feature, their standard errors, z values, and the associated p-values (Pr(>|z|)).

1) LIMIT_BAL: The negative coefficient (-1.480202e-06) indicates that as the LIMIT_BAL increases, the probability of default payment next month decreases. The p-value (7.559519e-16) is very small, suggesting that the effect is statistically significant.
2) SEXfemale: The negative coefficient (-1.190449e-01) indicates that being female is associated with a lower probability of default payment next month compared to being male. The p-value (1.411249e-03) shows that the effect is statistically significant.
3) EDUCATION: The positive coefficients for graduate_school (9.019487e-01), university (8.796401e-01), and high_school (8.428833e-01) suggest that clients with these education levels have a higher probability of default payment next month compared to the reference category (others). The p-values (1.876351e-04, 2.559436e-04, and 5.295031e-04, respectively) indicate that these effects are statistically significant.
4) MARRIAGE: The coefficients for married (7.702171e-02) and single (-1.053541e-01) are not statistically significant (p-values of 6.335698e-01 and 5.186955e-01, respectively), indicating that the marital status has no significant impact on the probability of default payment next month.
5) AGE: The positive coefficient (5.044925e-03) indicates that as the client's age increases, so does the probability of default payment next month. The p-value (2.587609e-02) suggests that the effect is statistically significant.
6) avg_crd_bill: The positive coefficient (9.865618e-07) suggests that as the average credit bill amount increases, the probability of default payment next

month also increases. The p-value (5.313849e-03) indicates that the effect is statistically significant.

7) avg_pay_amt: The negative coefficient (-2.895872e-05) implies that as the average payment amount increases, the probability of default payment next month decreases. The p-value (6.176025e-14) shows that the effect is statistically significant.
8) avg_pay_time: The positive coefficient (1.333837e+00) indicates that as the average payment time increases, the probability of default payment next month also increases. The p-value (0.000000e+00) suggests that the effect is highly statistically significant.

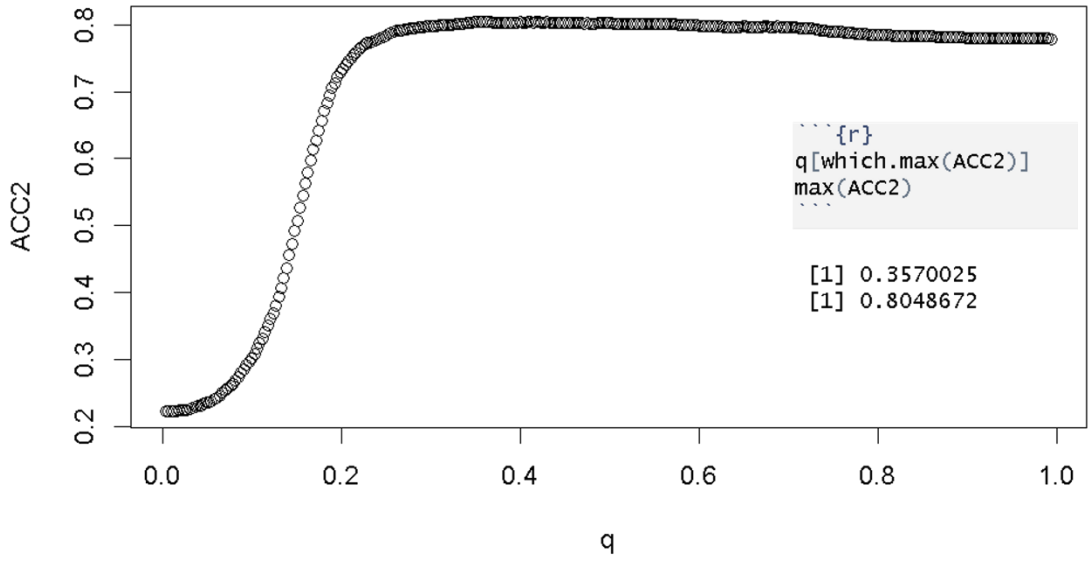

Figure 6- Logistic Regression accuracy vs q

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.373760e+00	3.047737e-01	-7.7885977	6.775699e-15
LIMIT_BAL	-1.480202e-06	1.836222e-07	-8.0611255	7.559519e-16
SEXfemale	-1.190449e-01	3.729080e-02	-3.1923404	1.411249e-03
EDUCATIONgraduate_school	9.019487e-01	2.414787e-01	3.7351075	1.876351e-04
EDUCATIONuniversity	8.796401e-01	2.405861e-01	3.6562381	2.559436e-04
EDUCATIONhigh_school	8.428833e-01	2.432304e-01	3.4653697	5.295031e-04
EDUCATIONothers	5.070071e-02	4.871372e-01	0.1040789	9.171067e-01
MARRIAGEmarried	7.702171e-02	1.615699e-01	0.4767083	6.335698e-01
MARRIAGEsingle	-1.053541e-01	1.632492e-01	-0.6453575	5.186955e-01
AGE	5.044925e-03	2.264262e-03	2.2280664	2.587609e-02
avg_crd_bill	9.865618e-07	3.539406e-07	2.7873656	5.313849e-03
avg_pay_amt	-2.895872e-05	3.858952e-06	-7.5042946	6.176025e-14
avg_pay_time	1.333837e+00	3.280143e-02	40.6639717	0.000000e+00

Table 7- Logistic regression coefficient

7. Conclusion

It is important to note that while the decision tree model's accuracy is relatively high, further analyses and comparisons with other classification algorithms may be necessary to determine the most suitable model for predicting default payment behavior. Additionally, the variable importance values provide insights into the relative importance of each input feature but do not imply causation, and further investigations are needed to better understand the relationships between these variables and default behavior.

The kNN model with k=69 yields a relatively high accuracy of 0.7796422 but might be less effective in predicting default payment behavior compared to the decision tree model due to the exclusion of categorical variables. Further investigations and comparisons with other classification algorithms, such as logistic regression or ensemble methods, might be necessary to determine the most suitable model for predicting default payment behavior while considering all relevant variables.

The logistic regression model demonstrates a relatively high accuracy of 0.8048672 in predicting default payment behavior while considering all relevant input features, making it a strong candidate for further analysis and comparison with other classification algorithms. Its ability to handle a mix of categorical and numeric variables and estimate the probability of default provides valuable insights for credit management and risk assessment in the financial industry.

In conclusion, the logistic regression model demonstrated the best performance with an accuracy of 0.8048672 when the threshold (q) was set at 0.357. The data exploration phase revealed multicollinearity issues among some of the variables, particularly within PAY_x and BILL_AMTx variables. The decision tree and logistic regression models were able to handle mixed variable types and take into account the relationships between the target variable and input features. The logistic regression model provided further insights into the relative importance of the input features and their impact on the probability of default payment.

Despite the limitations related to multicollinearity and mixed variable types, the study demonstrated that machine learning techniques, such as decision tree and logistic regression, can effectively predict the probability of default payment next month. These models can be useful for financial institutions and credit providers in assessing the creditworthiness of their clients, allowing for better risk management and decision-making processes.

Future research could explore additional methods to address multicollinearity and mixed variable types or include more advanced techniques such as ensemble methods or deep learning for improved predictive performance. Furthermore, exploring additional features or external factors that may influence the probability of default payment could provide even more valuable insights for credit risk assessment.

8. Ethical considerations

Ethical considerations play a significant role in any data-driven project, particularly when dealing with sensitive financial information. Some ethical aspects and potential stakeholders to be considered in this project include:

Data ownership and privacy: The dataset used in this project contains personal financial information of clients, which may raise privacy concerns. It is crucial to ensure that the data is anonymized and all personally identifiable information is removed or encrypted. Obtaining informed consent from the clients, as well as respecting data ownership rights, are essential in maintaining ethical standards in data handling.

Human ethics and fairness: Machine learning models, including the ones used in this project, may inadvertently perpetuate biases present in the data. It is essential to consider the fairness of the models to ensure that they do not unfairly discriminate against certain demographic groups (e.g., based on gender or education level). By conducting bias and fairness audits of the models, one can identify and mitigate potential issues.

9. Reference

I-Cheng Yeh, Che-hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients" (2009)

https://www.sciencedirect.com/science/article/pii/S0957417407006719

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

10. Acknowledgement

This project was supported by the DA department at Denison University and a special thanks to Dr. Wang for giving me the necessary resources to complete this project.