Credit Card Default Prediction

Liqing Jing, Juan Zhang, Qinghua Lin

Overview

Banks can have serious impact from credit defaults. Based on the dataset, the total billed amounts from April 2005 to September 2005 is 8,095,850,136 (NT dollar), and the payment amount is 949,541,777 (NT dollar), only 11.73% of the total billed amount. Being able to identify which individual is more likely to default on their credit card bills can help banks better manage default risks and balance reserves. How can we prevent the risk of default payments and predict the probability of default? That would be the research we will focus on.

Data

Data source: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

The dataset is called Default of credit card clients provided by UCI machine learning. We did our research based on the variables we have from the dataset. Are default payments related to 23 variables (Amount of given credit, Gender, Education, Marital Status, Age and etc?

Default payment next month in column Y will be considered dependent variable. Limit balance, sex, education, marriage, age, history of past payment, amount of bill statement and amount of previous payment will be considered independent variables. We are adding months delayed in payment which is the count of history of past payment and maximum months delayed which is the maximum months of past payment history. We are splitting age category into three binary variables consisting of young age(21-40), middle age(41-60) and senior age(61-79) and previous payment into binary variable of delay in each month. We hypothesize months delayed in payment and maximum month delayed being the most important predicting variables.

The original raw data has 30000 rows of data and it includes 25 columns. There are 24 variables plus an 'ID' column. The definition of the variables in the dataset are as followings:

Variable Name	Definition
Y: default payment next month	(1 = default; 0 = not default)
X1: Amount of the given credit (NT dollar)	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender	Gender (1 = male; 2 = female).
X3: Education	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status	Marital status (1 = married; 2 = single; 3 = others).
X5: Age	Age (year).
X6 - X11: History of past payment	History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar)	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005;. . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar)	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

The sample data is shown as below:

	...1	...2	...3	...4	...5	...6	...7	...8	...9	...10	...11	...12	...13	...14	...15	...16	...17	...18	...19	...20	...21	...22	...23	...24	...25
1	NA	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
2	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment
3	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
4	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
5	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
6	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
7	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
8	6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
9	7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
10	8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0

Data Cleaning and Feature Engineering

A few steps were taken to clean and manipulate the data including:

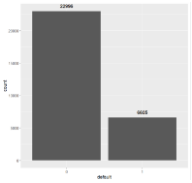
1) Cleaned the data, removed useless rows and columns, renamed variables
2) Added new variables and transformed variables
3) Removed columns and variables that are not required for the model. 13 independent variables left and one dependent variable.

```
tibble [29,601 × 14] (S3: tbl_df/tbl/data.frame)
 $ default          : chr [1:29601] "1" "1" "0" "0" ...
 $ LIMIT_BAL        : chr [1:29601] "20000" "120000" "90000" "50000" ...
 $ SEX              : chr [1:29601] "2" "2" "2" "2" ...
 $ AGE              : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 1 1 1 1 1 ...
 $ EDUCATION        : chr [1:29601] "2" "2" "2" "2" ...
 $ MARRIAGE         : chr [1:29601] "1" "2" "2" "1" ...
 $ num_months_delay: num [1:29601] 2 2 0 0 0 0 0 0 1 0 ...
 $ max_delay_months: chr [1:29601] "2" "2" "0" "0" ...
 $ delay_month1     : num [1:29601] 1 0 0 0 0 0 0 0 0 0 ...
 $ delay_month2     : num [1:29601] 1 1 0 0 0 0 0 0 0 0 ...
 $ delay_month3     : num [1:29601] 0 0 0 0 0 0 0 0 1 0 ...
 $ delay_month4     : num [1:29601] 0 0 0 0 0 0 0 0 0 0 ...
 $ delay_month5     : num [1:29601] 0 0 0 0 0 0 0 0 0 0 ...
 $ delay_month6     : num [1:29601] 0 1 0 0 0 0 0 0 0 0 ...
```

4) Converted variables to factor

Imbalance Data Resolution

An issue of imbalanced data is shown in the plot where 78% are nondefault and 22% are default.



To resolve the imbalanced data, the dataset was split into cross validation dataset(train + validation)(70%) and test dataset(30%). Then the oversampling approach was applied to the cross validation dataset(train + validation) dataset. Below is size of each set after oversampling:

	Nondefault	Default	Total
Train + validation dataset	16167	16167	32334
Test dataset	6829	2009	8838
Total			41172

Data exploration and analysis

1. Bar plots of each variable allow us the first claims on the data.

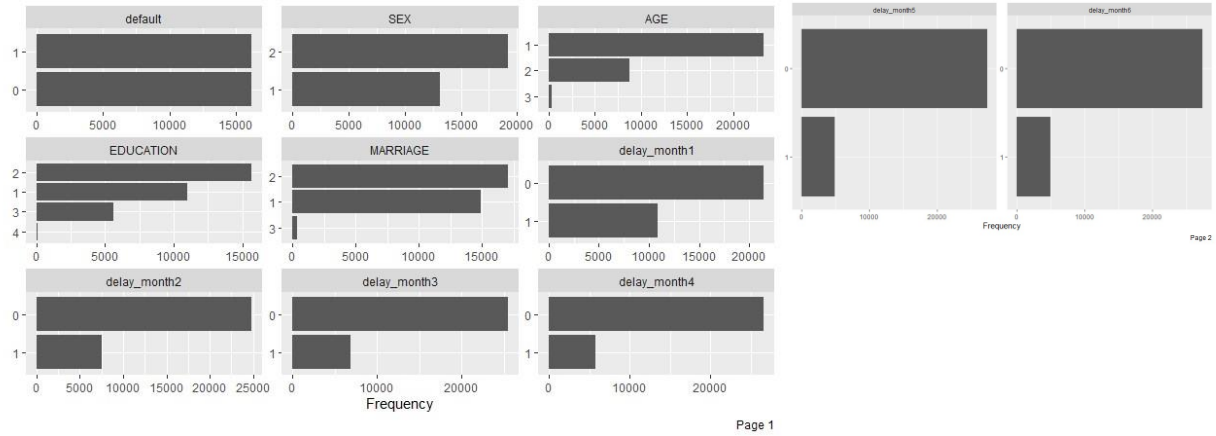


Figure 1

2. Explore distribution and density plots. We can see that Given Credit(limit_balance), max_delay_months and num_months_delay are not normally distributed and more likely to be positively skewed distribution.

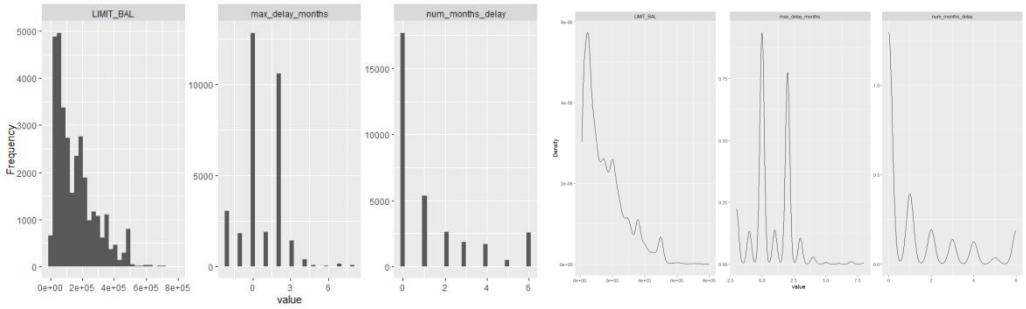
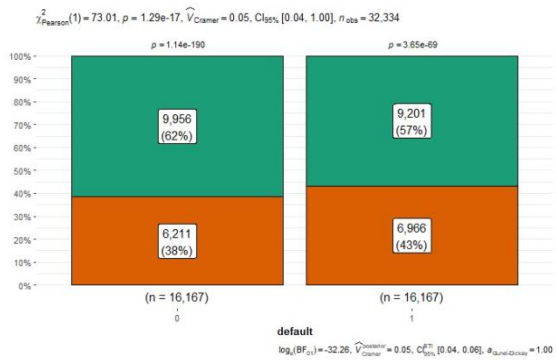


Figure 2

3. Visualization with statistical details(especially plot each variable with the proportions of different groups)



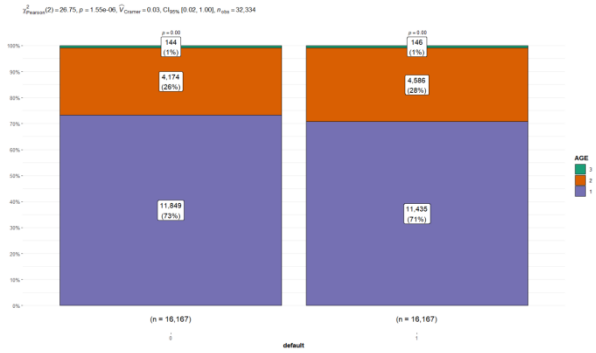
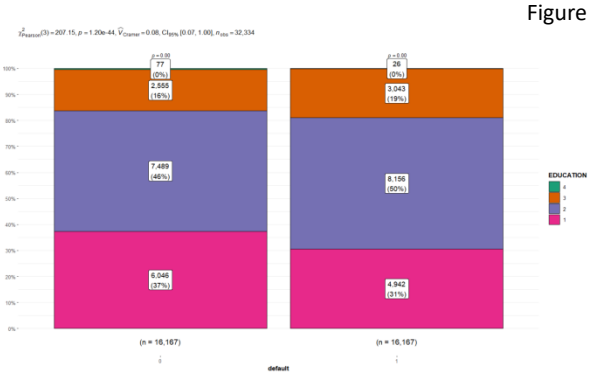


Figure 6 (left) & Figure 7(right)



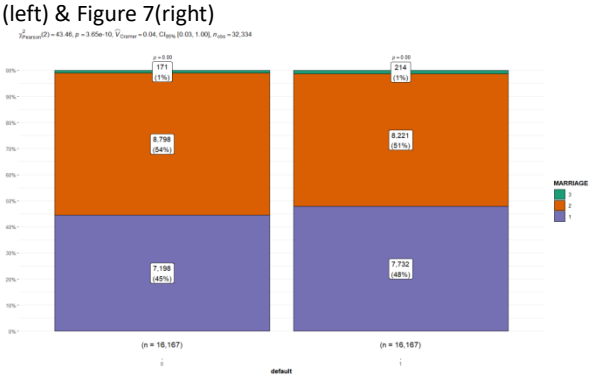
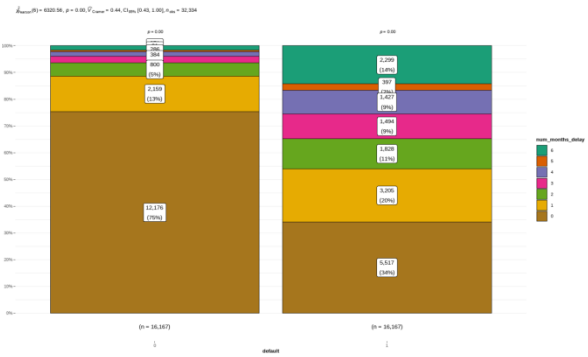


Figure 8 (left) & Figure 9(right)



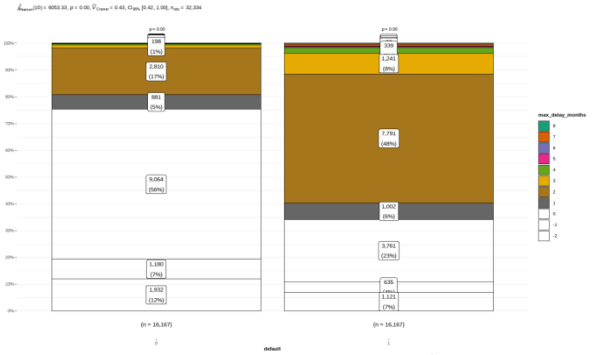
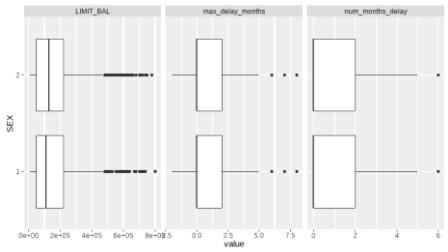


Figure 10(left) & Figure 11(right)

According to the plots, we count and calculate percentages for every category and visualizes frequency table in the form of stacked bars as well as provides numerous details, which allows us to conclude that:

- The gender could be associated with the default, the males may more likely to be default. About 53%(6966/13177) males would be default while 48% (9201/19157) in females(Figure 6)
- The middle-age groups are more likely to be default compared with the young group and senior. 52.35% of middle age groups would be default while 49.11% in young age group and 50.34% in senior age group.(Figure7) ;
- The education level is strongly associated with the default, namely the more educated they get, the more likely they would not be default. 44.98% of graduate school group would be default while 52.13% in university group and 54.36% in high school group.(Figure8);
- The married group are more likely to be default, they may have more pressure from family. 51.79% of married group would be default while 48.30% in single group.(Figure9)
- From the plot, it is obviously that the more number of months had delays in prepayment, the more likely they would be default(Figure 10 & Figure 11). So we could hypothesis that the past delay history correlate with default in next month. People who have bad history may be more likely to be default;

4. Boxplot of categorical variables and numeric variables.



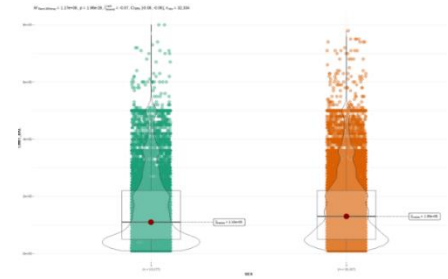
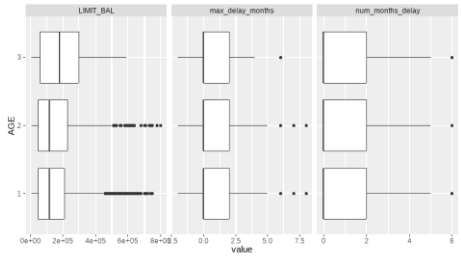


Figure 12



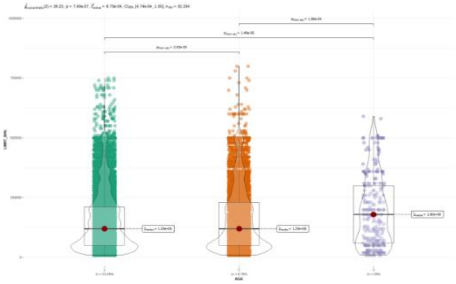
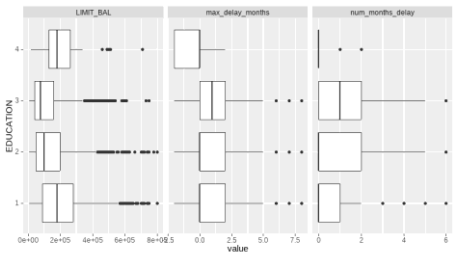


Figure 13



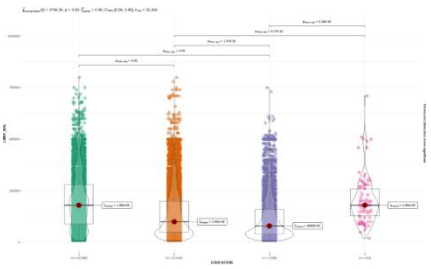
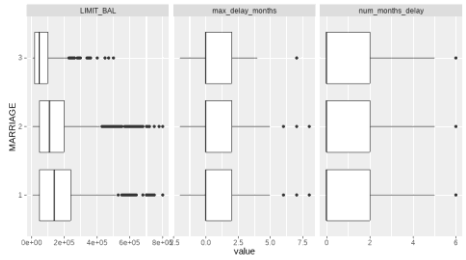


Figure 14



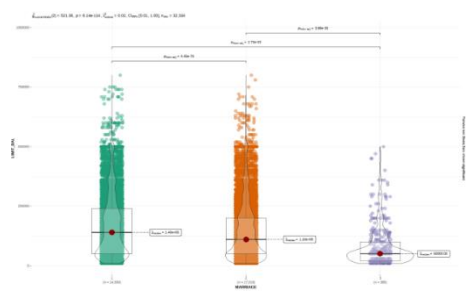


Figure 15

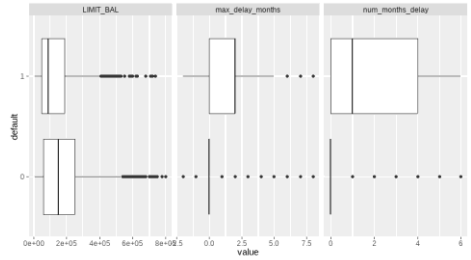
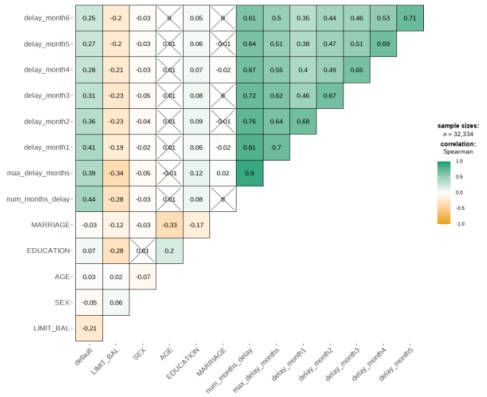


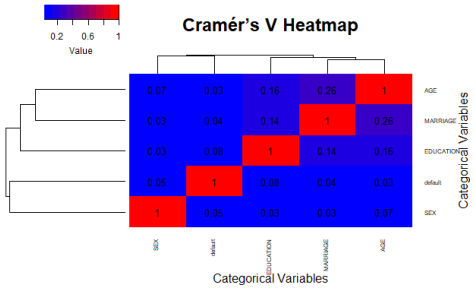
Figure 16

- Averagely, the females tend to have more given credit than the males, and the p-value of pairwise test is significantly less than 0.001.(Figure 12)
- It looks like the senior age groups would have more given credit than the other two while the median of middle age and young age almost the same. the p-value of pairwise test is significantly less than 0.001(Figure 13)
- The education level seems to be associated with the given credit, the more educated they get, the more likely they have the credit. The median of graduate school is significantly higher than the university and high school groups. (Figure 14)
- The married groups has higher given credit than single group, the p-value of pairwise test is significantly less than 0.001(Figure 15)
- The less given credit they would have, the more likely they may be default. Usually, the less likely they could pay duly, the less given credit would be. So that makes sense.(Figure16)

5. Explore Correlation



Non-parametric spearman is appropriate for not normally or not very linearly distribution. The plot displays correlation coefficients and shows the strength of the correlation while the color shows the direction where green is positive and orange is negative correlation. And if it is not significant correlations, it is simply crossed out.



Cramér's V is a measure of association between categorical variables, and it is an extension of the chi-squared test. The above is the Cramér's V heatmap for the categorical variables of "SEX", "EDUCATION", "MARRIAGE", "AGE" and "default". It ranges from 0 to 1, where 0 indicates no association, and 1 indicates a perfect association.

Overview of Modeling

We implemented the following four models in this project:

1. GLM

We try this traditional classification GLM model and used 10-fold cross validation and calculated the accuracy for each fold. Then we explore the most significant factors on default and find the model goodness of fit and predictive power.

2. DT(Decision Tree)

For the Decision Tree model, we employed a recursive partitioning approach to create a tree structure that recursively splits the dataset based on the most significant features. Utilizing 10-fold cross-validation, we assessed the model's performance across different folds and examined key metrics such as accuracy. Additionally, we investigated the interpretability of the resulting tree and evaluated its ability to capture the underlying patterns in the data.

3. XGB

Similar to random forests, XGBoost uses additive methods to build trees one at a time with gradient boosting to learn the optimal discriminative model for prediction. We use 10-folds cross validation to get the average cross validation accuracy.

4. KNN

We chose this model because it is easy to interpret, understand, and implement. We used a loop with 10-fold cross validation to find the optimal K with the highest model accuracy.

Model performance and results

1. GLM

- We notice that the NA in the variables “delay_month61”. There may be the multicollinearity problems in this model, which means two or more independent variables in GLM model are highly correlated. Then we use alias function, 1 and -1 show that “num_months_delay” and several “delay_month” variables are linearly dependent on “delay_month6”. This means that they are highly correlated. So we decide to remove the variable “delay_month6” and re-run the GLM model. We also notice that the VIF values of num_months_delay is obviously above 10, but this variable is significant . So as what Ogreens said, if a regression coefficient is statistically significant even when there is a large amount of multi-collinearity, it is statistically significant in the ‘face of that collinearity’. So we decide to keep this variable.

- The accuracy of classification is 0.7989 and AUC is 0.7378. The delta value 0.1975 when k equal to be 10 in the cross validation is low, which suggests that the model is not overfitting to the training data and the performance of training data is consistent with one on new data.

- The deviance residuals in the summary would show how well our model is fitting the data. We want the 1Q and 3Q to be similar in absolute value 0.91. We want our median to be close to 0. And we want the minimum and maximum to be similar to each other and also under three, which means that it is not deviating from a normal distribution.

- Conclusion:
  - The negative value shows that there is a correlation between credit limit and defaulting on credit card payment. It shows that the higher the given credit balance they have, the less likely they would be default.

- The age group (41-60) is more likely to default on credit card payment.
- The single person is more likely to default on credit card payment.
- There is a correlation between delayed payment history and defaulting on credit card payment. The people who constantly delay on payment are more likely to default on credit card payment.
- There is a higher probability of defaulting if the individual delayed in payment in the previous month.
- Those results would be valuable from the perspective of risk management. It would be effective for the bank to decide to whether they should approve the credit card and how much the consumer credit should be. It could manage default risks and balance reserves. If the bank could accurately estimate the real probability of default, they could lower the risk of default payment and save a lot.

```
Model :
default ~ LIMIT_BAL + SEX + AGE + EDUCATION + MARRIAGE + num_months_delay +
    max_delay_months + delay_month1 + delay_month2 + delay_month3 +
    delay_month4 + delay_month5 + delay_month6

Complete :
             (Intercept) LIMIT_BAL SEX2 AGE2 AGE3 EDUCATION2 EDUCATION3 EDUCATION4 MARRIAGE2 MARRIAGE3
delay_month61           0         0    0    0    0          0          0          0         0         0
             num_months_delay max_delay_months delay_month11 delay_month21 delay_month31 delay_month41
delay_month61                1                0            -1            -1            -1            -1
             delay_month51
delay_month61            -1
```

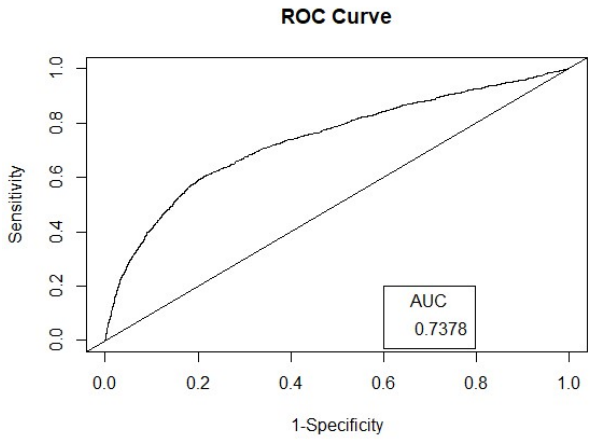
```
Call:
glm(formula = default ~ LIMIT_BAL + SEX + AGE + EDUCATION + MARRIAGE +
    num_months_delay + max_delay_months + delay_month1 + delay_month2 +
    delay_month3 + delay_month4 + delay_month5, family = "binomial",
    data = new_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.39774  -0.91925  -0.05247   0.94438   1.97830

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.511e-01  4.227e-02  -5.942 2.82e-09 ***
LIMIT_BAL       -1.796e-06  1.100e-07 -16.329  < 2e-16 ***
SEX2            -1.557e-01  2.559e-02  -6.086 1.16e-09 ***
AGE2             7.690e-02  3.083e-02   2.494 0.012627 *
AGE3            -5.138e-02  1.339e-01  -0.384 0.701220
EDUCATION2       2.787e-02  2.917e-02   0.956 0.339322
EDUCATION3       1.339e-03  3.948e-02   0.034 0.972950
EDUCATION4      -5.237e-01  2.367e-01  -2.212 0.026939 *
MARRIAGE2       -1.762e-01  2.756e-02  -6.393 1.63e-10 ***
MARRIAGE3       -7.269e-03  1.146e-01  -0.063 0.949444
num_months_delay  3.209e-01  5.343e-02   6.006 1.91e-09 ***
max_delay_months  4.687e-02  1.398e-02   3.352 0.000803 ***
delay_month11     9.592e-01  6.375e-02  15.047  < 2e-16 ***
delay_month21    -4.559e-02  7.330e-02  -0.622 0.533954
delay_month31     9.387e-02  7.272e-02   1.291 0.196756
delay_month41     5.439e-03  7.427e-02   0.073 0.941612
delay_month51    -1.092e-01  9.752e-02  -1.120 0.262901
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44824  on 32333  degrees of freedom
Residual deviance: 37529  on 32317  degrees of freedom
AIC: 37563
```



2. DT(Decision Tree)

- The model needs to be reliable in the future for predicting default rates. This is measured by maximizing the loglikelihood of the test set. the decision tree was reduced in size (pruned) by choosing the complexity parameter that minimizes the out-of-sample error validation, or x error. The results between cp and the out-of-sample error are shown below:

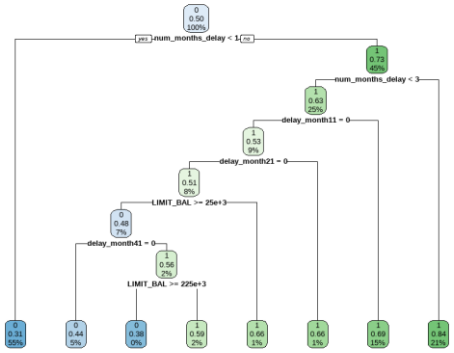
```
        CP nsplit rel error
1 0.41188841      0   1.00000
2 0.00154636      1   0.58811
3 0.00088363      7   0.57549
```

  The optimized complexity parameter value `0.00088363`.

- The max accuracy value of classification is 0.7953 and ROC value is 0.7151. The hyper parameter is obtained by the model with 10 folder cross validation on training dataset. The accuracy value and ROC value on the test dataset suggests that the model is not over fitting to the training data and the performance of training data is consistent with the test data.

- Conclusion:
  - With the optimized complexity parameter value 0.00088363, we get the following decision tree as shown below: this model has seven splits, starting with num_months_delay, then further splitting the largest remaining bucket by delay_month11, delay_month21, LIMIT_BAL,

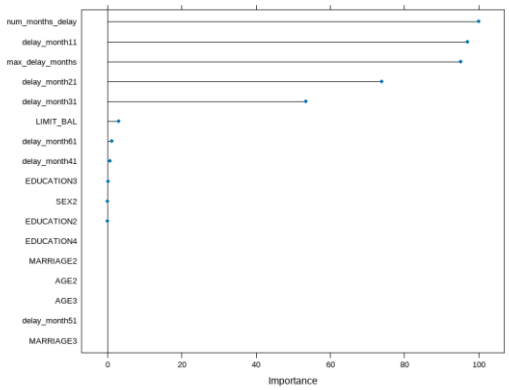
delay_month41 and so on. Based on this following obtained decision tree, the key feature is num_months_delay, and then the other features.

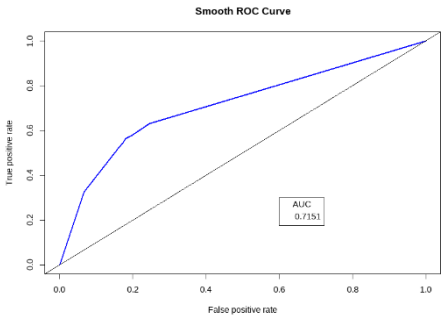


- We also can get the importance of features to predict the credit defaults of the model. As illustrated in the following figure, the relative importance of num_months_delay is 100%, making it the most influential factor in predicting credit defaults. and the importance of delay_month11 is 97%, highlighting its substantial contribution to the predictive power of the model. On the other hand, the `MARRIAGE3',`delay_month51', `EDUCATION4', `AGE2', `AGE3', and `MARRIAGE2', have negligible importance in predicting credit defaults.

rpart variable importance

	Overall
num_months_delay	100.00000
delay_month11	97.00599
max_delay_months	95.19587
delay_month21	73.92143
delay_month31	53.54259
LIMIT_BAL	2.93776
delay_month61	1.13714
delay_month41	0.66052
EDUCATION3	0.14494
SEX2	0.05201
EDUCATION2	0.03443
AGE3	0.00000
AGE2	0.00000
EDUCATION4	0.00000
MARRIAGE3	0.00000
delay_month51	0.00000
MARRIAGE2	0.00000





3. XGB

- To obtain the optimized hyperparameters for an XGB model, we used grid search to conduct a systematic hyperparameter tuning process with 10 folder cross validation on training dataset, iterating through different combinations and evaluating performance metrics to identify the set that maximizes model effectiveness.. The optimized hyperparameters are as follows:

nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
300	5	0.4	0	0.8	1	0.8

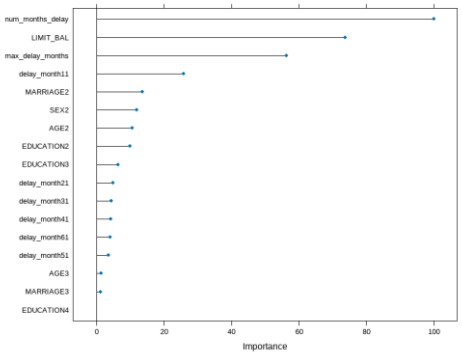
- Using the optimized hyperparameters, the max accuracy value of classification is 0.7946 and ROC value is 0.7056. The accuracy value and ROC value on the test dataset suggests that the model is not over fitting to the training data and the performance of training data is consistent with the test data.

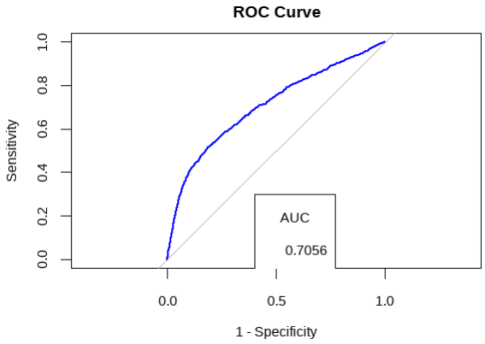
- Conclusion:

  - We also can get the importance of features to predict the credit defaults of the model. As illustrated in the following figure, the relative importance of num_months_delay is 100%, indicating it is the most key factor in predicting credit defaults. and the importance of LIMIT_BAL is 73%, showing that the LIMIT_BAL has considerable predictive power of the model. On the contrary, the `EDUCATION4`, have negligible importance in predicting credit defaults.

xgbTree variable importance

	Overall
num_months_delay	100.000
LIMIT_BAL	73.226
max_delay_months	64.066
delay_month11	22.848
MARRIAGE2	13.330
SEX2	11.730
AGE2	11.272
EDUCATION2	9.523
EDUCATION3	7.961
delay_month21	5.347
delay_month31	5.046
delay_month41	4.780
delay_month61	4.741
delay_month51	2.921
AGE3	1.918
MARRIAGE3	1.782
EDUCATION4	0.000



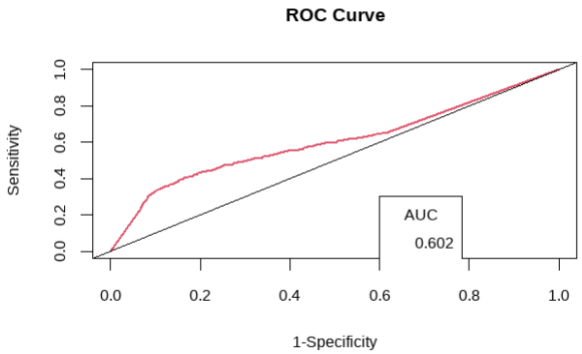


4. KNN

We tested different values of K (1 to 5), result showed the optimal model with accuracy 0.77 is when K=1. Cross validation was set to 10-fold. After training and fitting the data into KNN model, the AUC is 0.602.

k	Accuracy	Kappa
1	0.7773858	0.5547716
2	0.7540357	0.5080717
3	0.7443249	0.4886500
4	0.7371498	0.4743002
5	0.7336858	0.4673722

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 1.



Conclusion

This project focused on predicting credit card defaults to help banks manage risks and balance reserves effectively. The dataset, comprising 30,000 rows and 25 columns, was sourced from UCI machine learning and explored using 23 variables. Key independent variables included credit limit, gender, education, marital status, age, payment history, bill amounts, and previous payments, with default payment next month as the dependent variable.

Data cleaning and feature engineering involved adding variables like num_months_delay and max_delay_months, transforming age categories, and creating dummy variables for delayed payments. Imbalanced data was addressed by oversampling the minority class.

Exploratory data analysis revealed insights, such as gender, age group, education level, and marital status influencing credit defaults. Boxplots highlighted variations in given credit across categories, emphasizing the correlation between credit limits and default probability.

The correlation analysis, using Spearman and Cramér's V, unveiled relationships between variables, providing a foundation for modeling.

Generalized Linear Model (GLM): The GLM analysis revealed several key insights into credit default prediction. Notably, the correlation between credit limit and defaulting on credit card payments was negative, indicating that higher credit balances were associated with a lower likelihood of default. Age, marital status, and a history of delayed payments also are significant predictors. The model exhibited a classification accuracy of 79.89% and ROC value of 0.7378.

Decision Tree (DT): After pruning the decision tree to optimize its reliability in predicting future default rates.The model achieved a classification accuracy of 79.53% and a ROC value of 71.51%. Feature importance analysis highlighted the pivotal role of num_months_delay and delay_month11 in predicting credit defaults.

XGBoost (XGB): Utilizing grid search for hyperparameter tuning, the XGB model's optimized configuration included 300 rounds, a max depth of 5, and specific values for parameters such as eta, gamma, colsample_bytree, min_child_weight, and subsample. The resulting model demonstrated a classification accuracy of 79.46% and a ROC value of 70.56%. Feature importance analysis underscored the significance of num_months_delay and LIMIT_BAL in predicting credit defaults.

k-Nearest Neighbors (KNN): For KNN, different values of K were tested, with K=1 yielding the optimal accuracy of 77%. The model, although simple, provided insights into credit default prediction. However, its AUC value of 60.2% suggested moderate discriminative power compared to other models.

In conclusion, each model exhibited unique strengths and limitations. GLM provided interpretability and highlighted demographic factors, DT emphasized the importance of historical payment behavior, XGB showed the power of ensemble learning, and KNN offered simplicity. The choice of the best model depends on the specific goals and priorities of the analysis, with potential applications in risk management and credit assessment. In addition, our analysis may also impact the risk ratings when new clients applying for new credit products. Clients will be classified as higher default risk if they meet those risk criteria we conclude from our analysis, this will help with the decision making whether or not to approve the application and prevent future loss.