Prediction and Mitigation of Defaulting Credit Card Customers

Team 18

Raghu Bishnoi, Kenneth Greenwood, Boyuan Li, Michelle Li Li, Praphulla Nadimpalli

Business Understanding

As overseers of our digitized marketplaces, credit card companies have a bird's eye view of what we buy. To know what consumers all over the world are interested in, there's perhaps no better way that to examine their purchase histories, so it's no wonder that credit card companies, are at the forefront of big data mining. While one of the major use cases of data mining in this industry is fraud detection, we want to focus on another aspect: ***predicting which customers are likely to default and taking preventive measures.***

According to Business Insider, the delinquency rate on credit-card loan balances at commercial banks in the US - spiked to 2.5% in Q1 of 2018. In Taiwan, our target market, the average ratio of delinquency credit (past-due over three months) against account receivables (including non-accrual amounts) was 0.23% in the same time period. This translates to NT$ 4.8 Billion in losses every year. If we can first predict the customers who are likely to default and then take follow up action on these customers to minimise the default amount, we can expect to reduce these losses across the industry significantly.

We seek to evaluate risk in two ways. First, by identifying common characteristics of the customer base in our data set, we can group customers to better understand their demographic and credit behaviors in order to implement market segmentation. Secondly, by evaluating likelihood of default based on financial history (ie. drawing up balances, and payment behavior) we can better identify risk, possible determine thresholds for risk tolerance, and proactively work with customers to mitigate default events, in order to implement risk control. Combining both, profitable customer life-cycle management and risk control is the ultimate goal.

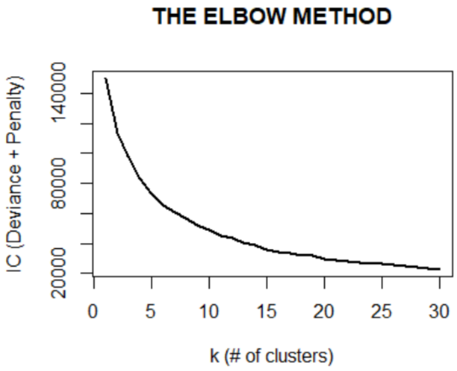
Data Understanding / Preparation

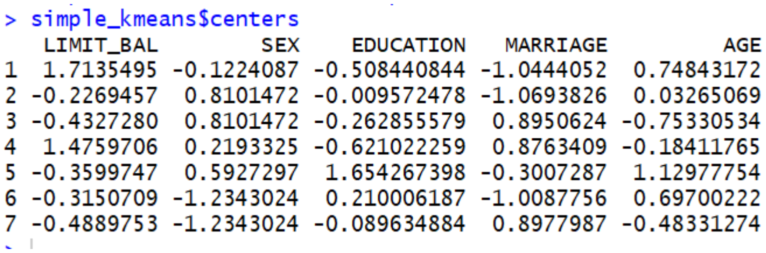
The UCI Credit dataset is a comprehensive dataset featuring credit card client information for the Taiwan market, from the period April to September 2005. It contains 25 variables of client information. 4 categorical variables are related to demographics (marital status, age, education and gender), and the rest provides financial information (credit limits, bill payment per month, bill statement amount, and previous payment status for the past month). Specifically, Bill payment status for each month ranges from values between -2 and 8, with -2 and -1 being the full amount paid duly, 1 being payment delayed for 1 month, 2 being payment delayed for 2 months and so on.

The target variable is a factor, indicating whether or not customers defaulted in the following month. The demographic data can be used to segment customers, while numeric financial data is mainly used in default prediction and classification model.

Regularization and Clustering

We used k-means clustering to identify groups of customers with similar demographic(sex, education, marriage, age, limit_balance) attributes. Seven optimized clusters are chosen in terms of the elbow method.



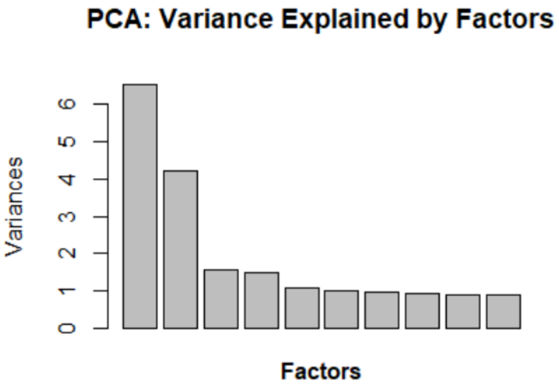


```
> simple_kmeans$centers
   LIMIT_BAL        SEX    EDUCATION   MARRIAGE         AGE
1  1.7135495 -0.1224087 -0.508440844 -1.0444052  0.74843172
2 -0.2269457  0.8101472 -0.009572478 -1.0693826  0.03265069
3 -0.4327280  0.8101472 -0.262855579  0.8950624 -0.75330534
4  1.4759706  0.2193325 -0.621022259  0.8763409 -0.18411765
5 -0.3599747  0.5927297  1.654267398 -0.3007287  1.12977754
6 -0.3150709 -1.2343024  0.210006187 -1.0087756  0.69700222
7 -0.4889753 -1.2343024 -0.089634884  0.8977987 -0.48331274
```

The first cluster is comprised of highly educated older people with high credit limits. The second is comprised of university graduated, middle aged, married females. The third is comprised of young, unmarried females. The fourth cluster is comprised of highly educated, high credit limit young, unmarried females. The fifth cluster is comprised of less educated older married females. The sixth cluster is comprised of married less educated men with lower credit limits. The final cluster is that of unmarried university graduate young men with lower credit limits. Within cluster classification can also shed light on the different average default rates across clusters.

First Factor -

BILL_AMT4	BILL_AMT5	BILL_AMT3	BILL_AMT2	BILL_AMT6	BILL_AMT1	PAY_5
0.3522465	0.3501291	0.3481048	0.3441620	0.3429270	0.3327495	0.2135893



Second Factor -

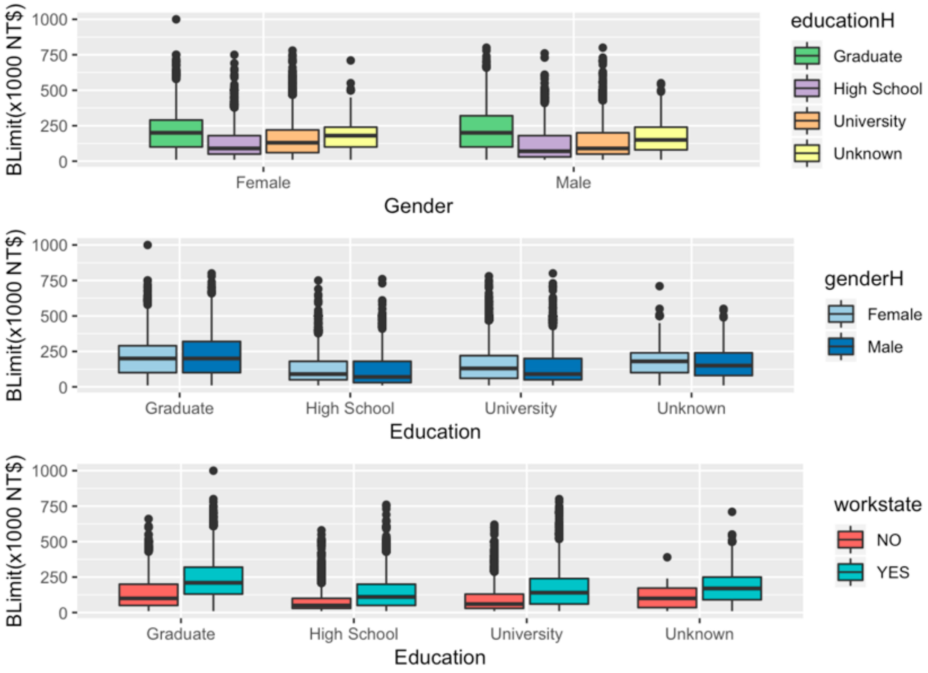
PAY_3	PAY_4	PAY_2	PAY_5	LIMIT_BAL	PAY_0	PAY_6	default.	PAY_AMT3
-0.3349	-0.33438	-0.3279	-0.32131	0.31151	-0.296	-0.2959	-0.1763	0.150

In terms of bill amount and payment amounts, it is not obvious at first glance what these latent features mean. However, we have discovered that the first latent feature involves the customers who have higher than average bill amounts but do not pay them back fully, aka those who use revolving credit. The second latent feature comprises of those customers who have pay the bill amounts in full, aka who do not use revolving credit. These two components explain 43% of the point variability.

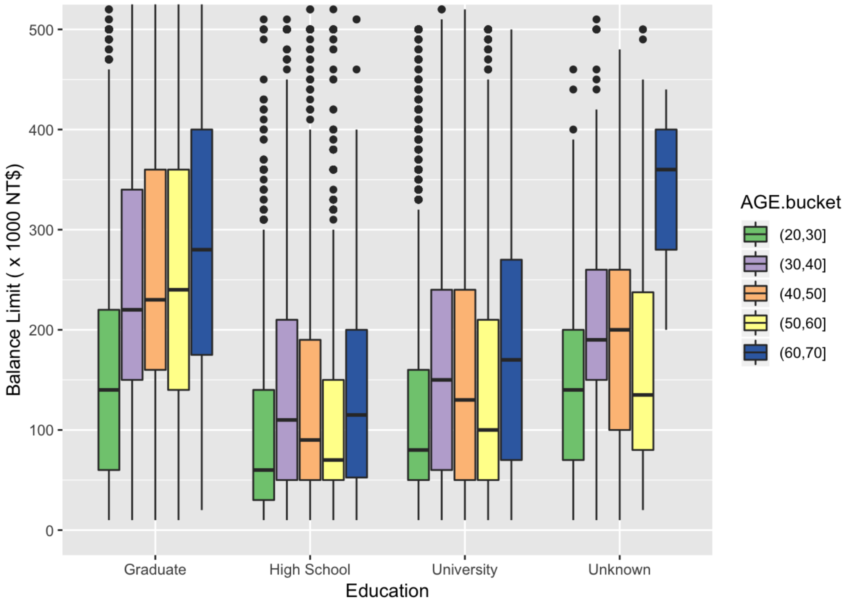
Data transformation and Visualization

Data transformations, including centering and scaling, are implemented to better fit the regression predictive model. Data resampling (SMOTE and over-sampling) is also pre-proceeded to handle imbalanced target variable issue (only 22% default rate). Age and Repayment status variables are binned to create new variables columns.

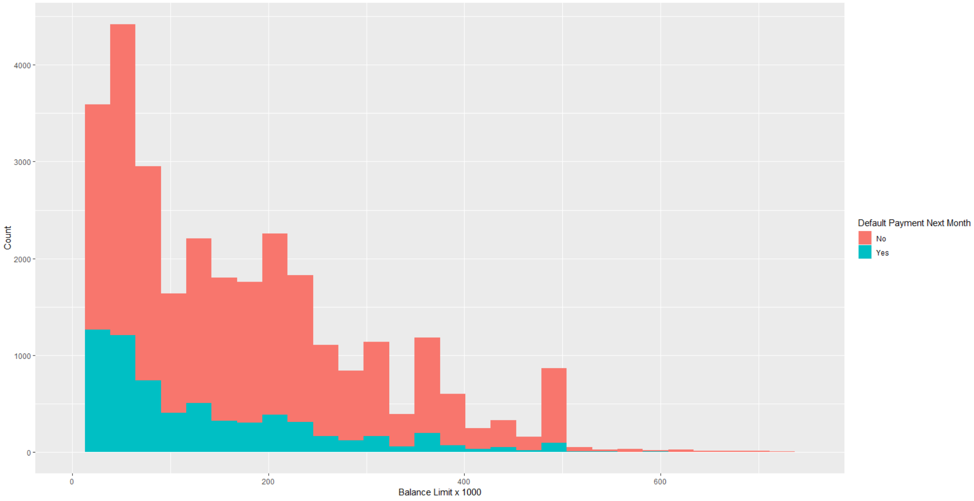
Following is the exploratory data analysis.



When we compare the balance limit with gender, education and work status, we saw that gender has no effect on balance limit determination process of bank, while the education level has a positive effect on this process. Moreover, work status is a very important factor at balance limit determination (work state means that there is no repayment delay in the previous 6 month).



Moreover, age also has an impact on the balance limit that a customer receives. We would expect that the customers with lower balances are more likely to default because of the higher amount of risk in pre-screening would lead to a lower balance. We explored that in the following graph.



There are differences in different demographics and the likelihood of default. The causal relationship between them and the dependent variable needs to be further explored.

Causal Modelling

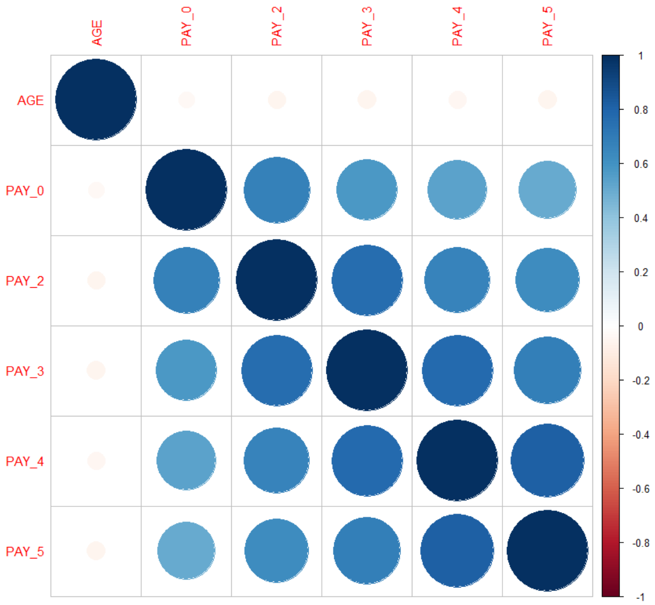
In order to understand which demographic variables impact the probability of default in a casual manner, we looked at the four demographic variables available to us: age, sex, marital status and education. We ran these through the Rubin model in order to estimate the causal relationship. Here were our findings:

1. There is a significant causal relationship between sex and log odds of default. Male customers increase the log odds of default by 0.18.
2. There is an insignificant causal relationship between education and default.
3. There is a very small but significant causal relationship between age and log odds of default. An increase in the age of the customer by a year increases the log odds of default by 0.0042.
4. There is a significant causal relationship between marital status and log odds of default. Being married increases the log odds of default by -0.079.

Modeling

Predictive classification model is the goal of this model building section. Highly interpretive models such as Logistic Regression, Logistic with Interaction, Classification tree, Lasso and, Null models are designed to predict the probability of default. Low predictive bias complex models such as Random Forest and XGBoost are also included to compare the prediction accuracy. Lasso is basically in regularization and feature selection. Within the Lasso method, we tried the the 1SE and Min methods of Lambda selection. 1SE model that resulted in the selection of 14 variables is chosen.

One major issue in the modelling was computational costs. Since the data was full of factor variables, running logistic interaction became really difficult. In order to reduce the variables, we decided to only use the payment status as of the current month as a instrument for the payment status in previous months since we found these to be highly correlated.



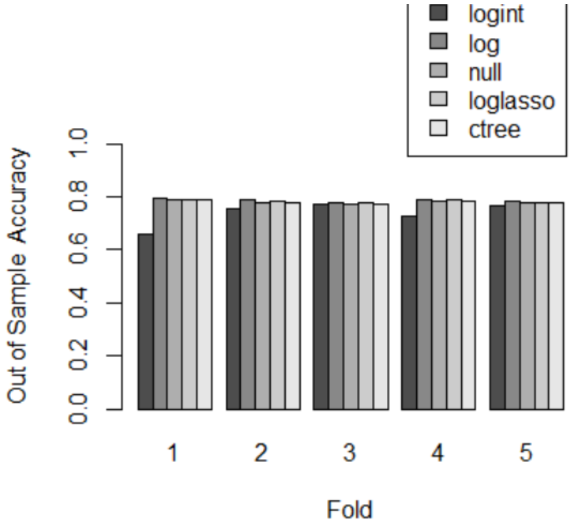
To assess the insample effectiveness of the models, we looked at the in sample R squared and the accuracy. We took a high threshold for our logistic regression of 0.75 because of the business justification. We wanted our false positive rate to be low as most of our customers are highly educated high income individuals and approaching them with suggestions of avoiding default might alienate the customer base.

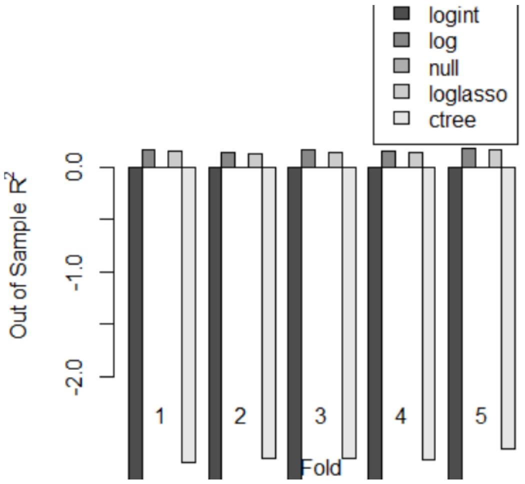
Model	Comments	In sample R^2	In sample Accuracy with threshold of 0.75
Logistic	All variables included	0.164	0.785
Logistic with Interaction	All variables with interaction	-15.84	0.753
Lasso	The lasso selected 14 variables using the 1SE method.	0.146	0.783

Classification Tree	All variables included	0.142	0.820
XGBoost	All variables (100 iterations)	-	0.751
Null	For comparison purpose only	~0	0.778

Evaluation

To understand the model performance, we did a 5 fold cross validation and observed how each model performs in terms of out-of-sample R squared and accuracy. Logistic Interaction, Logistic, Logistic Lasso(1SE lambda value - 14 variables), Classification Tree and Null models performed in the below manner:

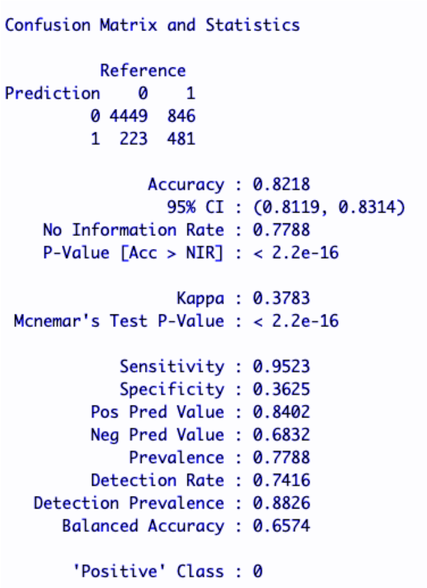




The OOS R squared graph indicates that logistic regression and classification tree models have a negative R squared value, implying that they are overfitting the model and performing worse than the null model. In terms of accuracy, logistic interaction seems to be performing bad when compared to all the other models. The logistic regression model with all variables has a better accuracy than the null model and an R squared value of 0.16. The logistic lasso model with 14 variables has accuracy and OOS R squared values pretty close to that of the logistic regression model.

From this analysis, we identified Logistic regression with all variables to be an effective model in terms of R squared value and Accuracy. From our logistic regression, we interpreted that being a male, married, older, and having a lower limit balance are all positively correlated with defaulting.

However, it is crucially important to notice that even though the accuracy rate is as high as around 0.8, yet the specificity rate (true negative rate) is far lower than the sensitivity rate (true positive rate). For instance in the rpart decision tree confusion matrix below. Further improvement in the true positive rate is needed since wrongly predicted default records cost more than other cases.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4449  846
         1  223  481

               Accuracy : 0.8218
                 95% CI : (0.8119, 0.8314)
    No Information Rate : 0.7788
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3783
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9523
            Specificity : 0.3625
         Pos Pred Value : 0.8402
         Neg Pred Value : 0.6832
             Prevalence : 0.7788
         Detection Rate : 0.7416
   Detection Prevalence : 0.8826
      Balanced Accuracy : 0.6574

       'Positive' Class : 0
```

Deployment

Today, many credit card companies actively exercise their contractual right to change the terms, or terminate use of their cards for certain users. If an individual is deemed to be a risk, whether due to exceeding credit limits, consistent delinquency, etc. card companies can seek to limit their losses by enacting certain restrictions. That said, we believe our model has the potential to be incredibly useful in predicting the individuals who present risk. Earlier detection could help both the companies and consumers. Companies will be able to identify risk much earlier, helping to reduce the losses associated with delinquent customers. Additionally, before a critical point is reached, card companies can work with consumers to adjust credit terms, like credit limits or interest rates, without having to cut someone's card off completely. Our model also has the potential to be used in the screening process before cards are awarded. Assuming, card companies will have access to a person's credit history, and perhaps spending behavior with their other cards, the card companies will be able to forecast the credit risk an individual presents if they were to issue an individual one of their cards.

We believe that by increasing the accuracy of the decision making process by even 1% and saving the costs by that amount, the total savings would come down to NT$ 50 million every year. That makes this a very valuable analysis indeed.

Strategy Based on the Model:

1. We want to target customers who have a likelihood of default of between 60 to 75% by making regular phone calls to check on their repayment status and by giving them small incentives to pay on time.
2. To the customers whose likelihood of default which falls above 85%, we would like to immediately talk to them about cancelling their card service in order to avoid future losses.

Our model helps us in understanding the likelihood of default of every individual customer and we can see what strategy to use on which customer in order to maximise the revenue by encouraging them to pay back the money and by minimising the write offs by cancelling or renegotiating the terms of their credit service.

APPENDIX

Contributions

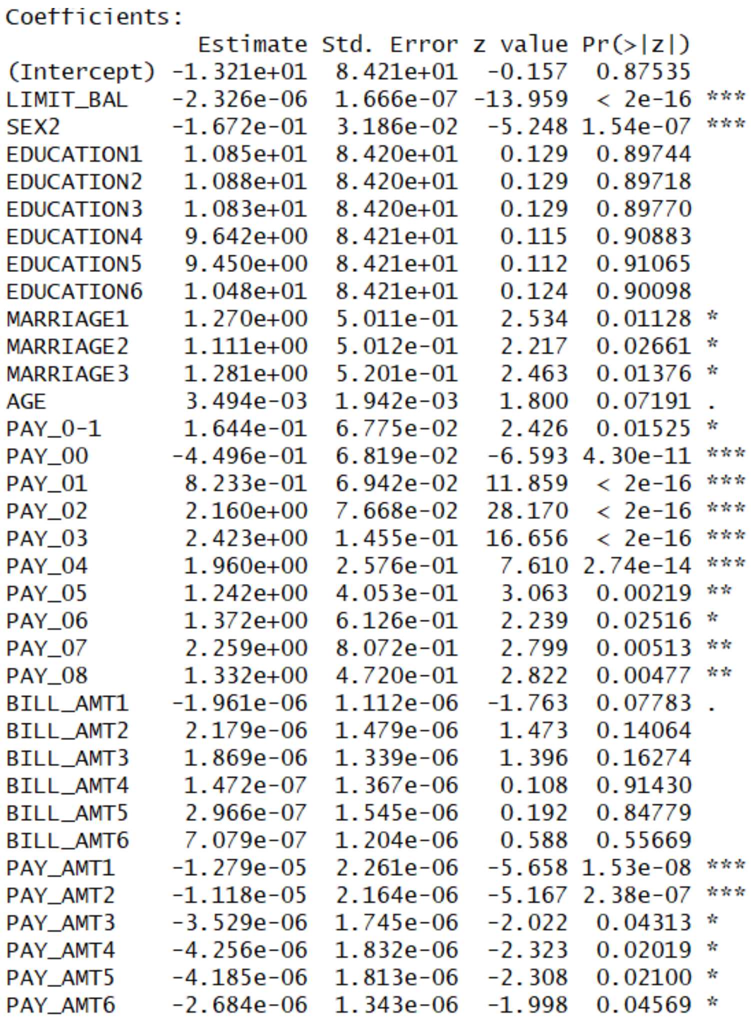
All team members are committed to the project and have made great contributions. Raghu contributed to the clustering and EDA part. Boyuan and Praphulla build the models and the measurements. Kenneth and Michelle contributed to the business definition and deployment.

Data Dictionary:

- 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_0: Repayment status in September, 2005 (-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)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

The final logistic regression model is as follows:



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.321e+01  8.421e+01  -0.157  0.87535    
LIMIT_BAL   -2.326e-06  1.666e-07 -13.959  < 2e-16 ***
SEX2        -1.672e-01  3.186e-02  -5.248 1.54e-07 ***
EDUCATION1   1.085e+01  8.420e+01   0.129  0.89744    
EDUCATION2   1.088e+01  8.420e+01   0.129  0.89718    
EDUCATION3   1.083e+01  8.420e+01   0.129  0.89770    
EDUCATION4   9.642e+00  8.421e+01   0.115  0.90883    
EDUCATION5   9.450e+00  8.421e+01   0.112  0.91065    
EDUCATION6   1.048e+01  8.421e+01   0.124  0.90098    
MARRIAGE1    1.270e+00  5.011e-01   2.534  0.01128 *  
MARRIAGE2    1.111e+00  5.012e-01   2.217  0.02661 *  
MARRIAGE3    1.281e+00  5.201e-01   2.463  0.01376 *  
AGE          3.494e-03  1.942e-03   1.800  0.07191 .  
PAY_0-1      1.644e-01  6.775e-02   2.426  0.01525 *  
PAY_00      -4.496e-01  6.819e-02  -6.593 4.30e-11 ***
PAY_01       8.233e-01  6.942e-02  11.859  < 2e-16 ***
PAY_02       2.160e+00  7.668e-02  28.170  < 2e-16 ***
PAY_03       2.423e+00  1.455e-01  16.656  < 2e-16 ***
PAY_04       1.960e+00  2.576e-01   7.610 2.74e-14 ***
PAY_05       1.242e+00  4.053e-01   3.063  0.00219 ** 
PAY_06       1.372e+00  6.126e-01   2.239  0.02516 *  
PAY_07       2.259e+00  8.072e-01   2.799  0.00513 ** 
PAY_08       1.332e+00  4.720e-01   2.822  0.00477 ** 
BILL_AMT1   -1.961e-06  1.112e-06  -1.763  0.07783 .  
BILL_AMT2    2.179e-06  1.479e-06   1.473  0.14064    
BILL_AMT3    1.869e-06  1.339e-06   1.396  0.16274    
BILL_AMT4    1.472e-07  1.367e-06   0.108  0.91430    
BILL_AMT5    2.966e-07  1.545e-06   0.192  0.84779    
BILL_AMT6    7.079e-07  1.204e-06   0.588  0.55669    
PAY_AMT1    -1.279e-05  2.261e-06  -5.658 1.53e-08 ***
PAY_AMT2    -1.118e-05  2.164e-06  -5.167 2.38e-07 ***
PAY_AMT3    -3.529e-06  1.745e-06  -2.022  0.04313 *  
PAY_AMT4    -4.256e-06  1.832e-06  -2.323  0.02019 *  
PAY_AMT5    -4.185e-06  1.813e-06  -2.308  0.02100 *  
PAY_AMT6    -2.684e-06  1.343e-06  -1.998  0.04569 *  
```