Model #101: Credit Card Default Model

Model Development Guide

Mimi Trinh

Section 1: Introduction

The credit card default model in this report is based upon the research aimed at the case of customer default payments in Taiwan. The original study concentrates on comparing the predictive accuracy of probability of default among six data mining methods. However, this study only focuses on developing the most accurate predictive model to forecast default outcome using four predictive models: random forest, gradient boosting, logistic regression with variable selection, and support vector machine (SVM). The measurement metrics used in this study to compare the effectiveness of four models are true positive rate (TPR), false positive rate (FPR), and accuracy.

This study is conducted in multiple phases to approach the problem to develop the most accurate predictive model to forecast default outcome 1) data understanding 2) feature engineering 3) exploratory data analysis (EDA) using traditional and model-based decision tree EDA 4) predictive model development 5) predictive model result comparison. Two major highlights of the study are the facts that model-based decision tree EDA doesn't produce accurate predictive models and the fact that logistic regression with variable selection is the most accurate method to predict default outcome, using both train and test dataset.

Section 2: The Data

The dataset comes in the R data format, which is loaded into R studio to analyze, process, and develop models. The data.group variable is constructed to partition the dataset in a single dimension to divide the dataset into three parts: train dataset to develop the models, test dataset to test the model accuracy, and validate dataset to monitor the chosen models. The figure below shows that there are 15,180 observations in the train dataset of 30,000 observations, which accounts for 50.6% of the total dataset. The test dataset includes 7323 observations, which represent 24.4% of the total dataset. Finally, the validate dataset has 7497 observations, which account for 25% of the total dataset. Sections 2-4 of this paper only focuses on the train dataset.

Figure 1

train	test	validate	Sum
15180	7323	7497	30000

There are 30 variables included in the dataset. Below is the data dictionary table of the default variables.

Figure 2

Variable	Name	Descript
1	ID	Observation identification number
2	LIMIT_BAL	Amount of the given credit (NT dollar in Taiwan), including both the individual consumer credit and his/her family (supplementary) credit
3	SEX	Gender (1 = male, 2 = female)
4	EDUCATION	Education (1 = graduate school, 2 = university, 3 = high school, 4 = others)
5	MARRIAGE	Marital status (1 = married, 2 = single, 3 = others)
6	AGE	Age (year)
7	PAY_0	Repayment status in September 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
8	PAY_2	Repayment status in August 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
9	PAY_3	Repayment status in July 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
10	PAY_4	Repayment status in June 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
11	PAY_5	Repayment status in May 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
12	PAY_6	Repayment status in April 2015 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, …, 9 = payment delay for nine months and above)
13	BILL_AMT1	Amount of bill statement in September 2015

14	BILL_AMT2	Amount of bill statement in August 2015
15	BILL_AMT3	Amount of bill statement in July 2015
16	BILL_AMT4	Amount of bill statement in June 2015
17	BILL_AMT5	Amount of bill statement in May 2015
18	BILL_AMT6	Amount of bill statement in April 2015
19	PAY_AMT1	Amount paid in September 2015
20	PAY_AMT2	Amount paid in August 2015
21	PAY_AMT3	Amount paid in July 2015
22	PAY_AMT4	Amount paid in June 2015
23	PAY_AMT5	Amount paid in May 2015
24	PAY_AMT6	Amount paid in April 2015
25	DEFAULT	Binary variable, default payment (1 = yes, 0 = no) as the response variable
26	u	Added variable to partition the dataset
27	train	Flag variable to indicate observations in train dataset (1 = train dataset, 0 = not in train dataset)
28	test	Flag variable to indicate observations in test dataset (1 = test dataset, 0 = not in test dataset)
29	validate	Flag variable to indicate observations in validate dataset (1 = validate dataset, 0 = not in validate dataset)
30	data.group	Added variable to partition the dataset in a single dimension to divide the dataset into three parts: train, test, validate

To conduct a data quality check, it's necessary to use the data summary for each variable.

Figure 4

```
> summary(train)
       ID           LIMIT_BAL          SEX          EDUCATION        MARRIAGE          AGE       
 Min.   :    1   Min.   : 10000   Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :21.00  
 1st Qu.: 7509   1st Qu.: 50000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:28.00  
 Median :14958   Median :140000   Median :2.000   Median :2.000   Median :2.000   Median :34.00  
 Mean   :14994   Mean   :168065   Mean   :1.603   Mean   :1.847   Mean   :1.551   Mean   :35.48  
 3rd Qu.:22472   3rd Qu.:240000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:41.00  
 Max.   :29999   Max.   :800000   Max.   :2.000   Max.   :6.000   Max.   :3.000   Max.   :75.00  
     PAY_0                PAY_2             PAY_3              PAY_4              PAY_5              PAY_6        
 Min.   :-2.00000   Min.   :-2.000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000  
 1st Qu.:-1.00000   1st Qu.:-1.000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000  
 Median : 0.00000   Median : 0.000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000  
 Mean   :-0.02009   Mean   :-0.134   Mean   :-0.1632   Mean   :-0.2165   Mean   :-0.2611   Mean   :-0.2868  
 3rd Qu.: 0.00000   3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000  
 Max.   : 8.00000   Max.   : 8.000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000  
   BILL_AMT1         BILL_AMT2        BILL_AMT3        BILL_AMT4         BILL_AMT5        BILL_AMT6      
 Min.   :-165580   Min.   :-69777   Min.   :-61506   Min.   :-170000   Min.   :-61372   Min.   :-339603  
 1st Qu.:   3528   1st Qu.:  3010   1st Qu.:  2772   1st Qu.:   2418   1st Qu.:  1754   1st Qu.:   1320  
 Median :  22576   Median : 21492   Median : 20089   Median :  19106   Median : 18138   Median :  17112  
 Mean   :  51218   Mean   : 49297   Mean   : 47021   Mean   :  43197   Mean   : 40154   Mean   :  38722  
 3rd Qu.:  66608   3rd Qu.: 63659   3rd Qu.: 59596   3rd Qu.:  53914   3rd Qu.: 49858   3rd Qu.:  48932  
 Max.   : 626648   Max.   :624475   Max.   :632041   Max.   : 628699   Max.   :823540   Max.   : 699944  
    PAY_AMT1         PAY_AMT2          PAY_AMT3         PAY_AMT4         PAY_AMT5           PAY_AMT6     
 Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0  
 1st Qu.:  1000   1st Qu.:    836   1st Qu.:   396   1st Qu.:   281   1st Qu.:   273.5   1st Qu.:   138  
 Median :  2129   Median :   2011   Median :  1800   Median :  1500   Median :  1506.5   Median :  1500  
 Mean   :  5658   Mean   :   5872   Mean   :  5149   Mean   :  4806   Mean   :  4712.8   Mean   :  5334  
 3rd Qu.:  5002   3rd Qu.:   5000   3rd Qu.:  4500   3rd Qu.:  4000   3rd Qu.:  4016.8   3rd Qu.:  4005  
 Max.   :873552   Max.   :1215471   Max.   :889043   Max.   :621000   Max.   :417990.0   Max.   :443001  
    DEFAULT             u                 train        test      validate   data.group
 Min.   :0.0000   Min.   :0.0000251   Min.   :1   Min.   :0   Min.   :0   Min.   :1  
 1st Qu.:0.0000   1st Qu.:0.1263411   1st Qu.:1   1st Qu.:0   1st Qu.:0   1st Qu.:1  
 Median :0.0000   Median :0.2524615   Median :1   Median :0   Median :0   Median :1  
 Mean   :0.2255   Mean   :0.2514970   Mean   :1   Mean   :0   Mean   :0   Mean   :1  
 3rd Qu.:0.0000   3rd Qu.:0.3777524   3rd Qu.:1   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1  
 Max.   :1.0000   Max.   :0.4999846   Max.   :1   Max.   :0   Max.   :0   Max.   :1  
```

From the data summaries above, the data quality check shows that there is no missing value in the dataset. However, there are data integrity issues in the dataset. In other words, the dataset is dirty, especially in variables EDUCATION, MARRIAGE, PAY 0-6, and BILL_AMT 1-6. Therefore, it’s necessary to do a brief analysis and check each variable before conducting any in-depth exploratory data analysis (EDA) or building any predictive model.

First, it’s necessary to convert the DEFAULT variable from numeric values to factor values since this is a nominal variable. The table below shows that among 15,180 observations in the train dataset, only 3423 observations or 22.55% actually default on their payment. We can conclude that majority of the observations do not default on their payment. The fact that the two classes within the response variable are not equally distributed is an important note to keep in mind as we build predictive models in section 5 of this report.

Figure 5

no default	default on payment	Sum
11757	3423	15180

Second, it's necessary to convert the numeric values in the SEX column into factor values since this is a nominal variable. The table below shows that 40% of the observations are male and 60% of them are female, which is close to the 50-50 equal split between the two classes.

Figure 6

1	2
6020	9160

Third, it's necessary to convert the numeric values to factor values in the EDUCATION column because it's a nominal variable. Below is the breakdown of all levels within this variable. However, this variable has dirty data because the dataset has seven classes from 0 to 6 whereas the provided data dictionary only has four levels from 1 to 4.

Figure 7

0	1	2	3	4	5	6
7	5389	7115	2443	64	139	23

Fourthly and similarly, we have to convert the numeric values to factor values in MARRIAGE variable. Below is the breakdown of all levels in the MARRIAGE column. Again, there's a data discrepancy here in which the dataset has four levels from 0 to 3 whereas the data dictionary only has three levels from 0 to 3.

Figure 8

0	1	2	3
29	6939	8037	175

Fifthly, it's necessary to convert numeric values to factor values in all fix variables PAY 0-6 because they represent nominal variables. There are data discrepancies in these columns as well since the data dictionary only has -1 and positive values whereas the dataset has -1, -2, 0, and positive values.

In order to clean the dataset and fix these data discrepancy issues, it's necessary to run a brief EDA to determine how we should handle the dirty observations. It's bad practice to simply delete these dirty observations. Rather we should conduct a brief EDA to understand their patterns to determine how to best remap them.

Starting with the EDUCATION variable, below is the breakdown of default rate for each class within the variable.

Figure 9

Class	0	1	2	3	4	5	6
Default rate	0%	20%	24%	26%	8%	9%	9%

The dataset has seven classes from 0 to 6 whereas the data dictionary only has four classes from 1 to 4 with 4 being the unknown category. Since categories 5 and 6 have the default rate very close to category 4, it's safe to remap them into category 4 of others. Since there are only seven observations in category 0, it's a very small sample size, so we will remap category 0 into category 4 of others. Thus, overall, observations in category 0, 5, 6 will be remapped into category 4 of others.

Next regarding the MARRIAGE variable, below is the breakdown of default rate for each class within the variable.

Figure 10

Category	0	1	2	3
Default rate	3%	24%	21%	25%

Since category 0% has a default rate that's very different from the remaining categories, we can conclude that this category behaves very differently from the other three categories. Therefore, it's bad practice to map observations in category 0 into category 3 of others. Therefore, we will leave this variable the way it is and consult the industry expert to understand what this category means since it's not reported in the data dictionary.

Then, regarding the PAY 0-6 variables, first it's necessary to convert the variable PAY_0 to PAY_1 to keep the naming consistent among the variables. Then we need to determine how to handle observations with values of -2 and 0 since the data dictionary doesn't have these values. Since all PAY 1-6 variables have a similar pattern in which the majority of the observations fall in the categories of -2, -1, 0, and 1, it's only necessary to conduct a brief EDA on the default rate of observations with values of -2, -1, 0, 1, and 2 in the first PAY variable. Below is the breakdown.

Figure 11

Class	-2	-1	0	1	2
Default rate	13%	14%	17%	34%	70%

Since the default rates for values -2 and 0 are very close to the default rate for 1 category, we can conclude that these three categories behave similarly. Thus, we can remap observations with categories of -2 and 0 to category -1, which means customers paying duly.

Finally, it's interesting to see negative values in the variables BILL_AMT 1-6. However, after conducting research in consumer credit, it's possible to have a negative bill statement if the customer overpays the bill and/or if the customer has a credit back for certain items on their accounts. Therefore, it's reasonable and permissible to have negative account balance.

Section 3: Feature Engineering

In this project, we need to engineer additional features from the variables provided. First, the AGE variable by default has the age measured in years. The min AGE in the dataset is 21 and max is 75, so we bin the variable into four categories 1) young adults 25 years old and less 2) adults 26-40 years old 3) middle age 41-64 years old 4) the elder 65 and over.

Second, we add the following additional variables to the dataset.

- AVG_BILL_AMT: average bill amount by averaging the monthly bill amount (expenditure) over the six months
- AVG_PAY_AMT: average payment amount can be used as a prxy for income or ability to pay
- PAY_RATIO 1-5: how much of each bill does the customer pay each month? Do they pay in full or less than the full amount? Since there's a time delay here, PAY_RATIO1 = PAY_AMT1 / BILL_AMT2, and so on. For N/A values in these variables with 0 payment / 0 bill, we define them as 100
- AVG_PAY_RATIO: average of the five payment ratio variables above
- AVG_UTIL 1-6: utilization of how much of the credit line the customer is using. Utilization = current balance / credit limit
- AVG_UTIL: average utilization of the six utilization variables above
- BILL_GROWTH 2-6: the balance growth of each month. BILL_GROWTH2 = BILL_AMT2 – BILL_AMT1
- UTIL_GROWTH 2-6: the utilization growth of each month. UTIL_GROWTH2 = UTIL2 – UTIL1
- MAX_BILL_AMT: the max billed amount over the six months

- MAX_PAY_AMT: the max payment amount over six months
- DLQ 1-5: delinquency of each month. Negative values mean customer owe money whereas positive values mean customer overpay. DLQ1 = PAY_AMT1 – BILL_AMT2
- MAX_DLQ: maximum of delinquency amount over months

Since we engineer features, we don't use the raw variables in the predictor pool anymore. Instead we use the engineer features to replace the raw variables to build the predictive models. For example, instead of using the six raw variables of payment amount, we use the average and max payment amount.

Section 4: Exploratory Data Analysis (EDA)

Section 4a: Traditional EDA

Below are the data summaries to act as a data quality check for the engineered features. From the results below, there's no issue with the engineered features.

Figure 12

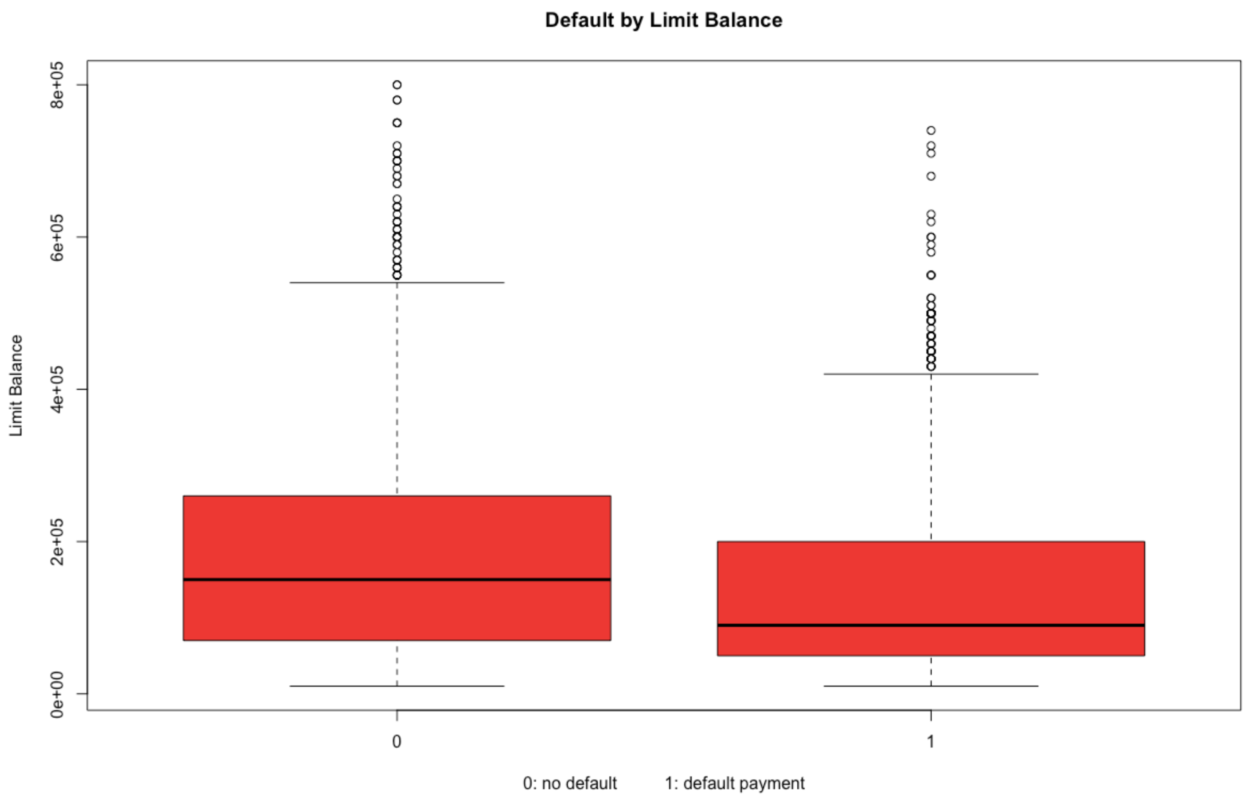
```
       ID           LIMIT_BAL        SEX      EDUCATION MARRIAGE AGE          PAY_1          PAY_2
 Min.   :    1   Min.   : 10000   1:6020   1:5389    0:  29   1:1964   -1      :11742   -1      :12944
 1st Qu.: 7509   1st Qu.: 50000   2:9160   2:7115    1:6939   2:9041   1       : 1874   2       : 1981
 Median :14958   Median :140000            3:2443    2:8037   3:4117   2       : 1340   3       :  159
 Mean   :14994   Mean   :168065            4: 233    3: 175   4:  58   3       :  153   4       :   54
 3rd Qu.:22472   3rd Qu.:240000                                        4       :   39   1       :   15
 Max.   :29999   Max.   :800000                                        5       :   13   5       :   13
                                                                       (Other):   19   (Other):   14
     PAY_3           PAY_4           PAY_5           PAY_6         BILL_AMT1         BILL_AMT2
 -1      :13041   -1      :13382   -1      :13651   -1      :13617   Min.   :-165580   Min.   :-69777
 2       : 1935   2       : 1622   2       : 1365   2       : 1406   1st Qu.:   3528   1st Qu.:  3010
 3       :  129   3       :   88   3       :   89   3       :   91   Median :  22576   Median : 21492
 4       :   39   4       :   38   4       :   36   7       :   26   Mean   :  51218   Mean   : 49297
 7       :   14   7       :   30   7       :   30   4       :   23   3rd Qu.:  66608   3rd Qu.: 63659
 5       :   10   5       :   15   5       :    7   6       :   10   Max.   : 626648   Max.   :624475
 (Other):   12   (Other):    5   (Other):    2   (Other):    7
   BILL_AMT3         BILL_AMT4          BILL_AMT5          BILL_AMT6           PAY_AMT1          PAY_AMT2
 Min.   :-61506   Min.   :-170000   Min.   :-61372   Min.   :-339603   Min.   :     0   Min.   :      0
 1st Qu.:  2772   1st Qu.:   2418   1st Qu.:  1754   1st Qu.:   1320   1st Qu.:  1000   1st Qu.:    836
 Median : 20089   Median :  19106   Median : 18138   Median :  17112   Median :  2129   Median :   2011
 Mean   : 47021   Mean   :  43197   Mean   : 40154   Mean   :  38722   Mean   :  5658   Mean   :   5872
 3rd Qu.: 59596   3rd Qu.:  53914   3rd Qu.: 49858   3rd Qu.:  48932   3rd Qu.:  5002   3rd Qu.:   5000
 Max.   :632041   Max.   : 628699   Max.   :823540   Max.   : 699944   Max.   :873552   Max.   :1215471

    PAY_AMT3          PAY_AMT4          PAY_AMT5            PAY_AMT6        DEFAULT          u
 Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0   0:11757   Min.   :0.0000251
 1st Qu.:   396   1st Qu.:   281   1st Qu.:   273.5   1st Qu.:   138   1: 3423   1st Qu.:0.1263411
 Median :  1800   Median :  1500   Median :  1506.5   Median :  1500             Median :0.2524615
 Mean   :  5149   Mean   :  4806   Mean   :  4712.8   Mean   :  5334             Mean   :0.2514970
 3rd Qu.:  4500   3rd Qu.:  4000   3rd Qu.:  4016.8   3rd Qu.:  4005             3rd Qu.:0.3777524
 Max.   :889043   Max.   :621000   Max.   :417990.0   Max.   :443001             Max.   :0.4999846

     train        test       validate   data.group  AVG_BILL_AMT      AVG_PAY_AMT        PAY_RATIO1
 Min.   :1   Min.   :0   Min.   :0   Min.   :1   Min.   :-56043   Min.   :     0   Min.   :-497.8000
 1st Qu.:1   1st Qu.:0   1st Qu.:0   1st Qu.:1   1st Qu.:  4789   1st Qu.:  1112   1st Qu.:   0.0432
 Median :1   Median :0   Median :0   Median :1   Median : 21198   Median :  2389   Median :   0.0908
 Mean   :1   Mean   :0   Mean   :0   Mean   :1   Mean   : 44935   Mean   :  5255   Mean   :      Inf
 3rd Qu.:1   3rd Qu.:0   3rd Qu.:0   3rd Qu.:1   3rd Qu.: 56880   3rd Qu.:  5554   3rd Qu.:   1.0000
 Max.   :1   Max.   :0   Max.   :0   Max.   :1   Max.   :592432   Max.   :627344   Max.   :      Inf
```

```
  PAY_RATIO2          PAY_RATIO3          PAY_RATIO4          PAY_RATIO5          AVG_PAY_RATIO
Min.   :-40.75000   Min.   :-500.0000   Min.   :-3.03e+03   Min.   :-31.02329   Min.   :-605.4581
1st Qu.:  0.04285   1st Qu.:    0.0368   1st Qu.: 3.60e-02   1st Qu.:  0.03735   1st Qu.:    0.0470
Median :  0.09091   Median :    0.0762   Median : 6.77e-02   Median :  0.07791   Median :    0.1584
Mean   :      Inf   Mean   :       Inf   Mean   :      Inf   Mean   :      Inf   Mean   :       Inf
3rd Qu.:  1.00000   3rd Qu.:    1.0000   3rd Qu.: 1.00e+00   3rd Qu.:  1.00000   3rd Qu.:    1.0004
Max.   :      Inf   Max.   :       Inf   Max.   :      Inf   Max.   :      Inf   Max.   :       Inf

    UTIL1              UTIL2              UTIL3              UTIL4              UTIL5
Min.   :-0.61989   Min.   :-1.39554   Min.   :-1.0251   Min.   :-1.04330   Min.   :-0.87674
1st Qu.: 0.02191   1st Qu.: 0.01863   1st Qu.: 0.0162   1st Qu.: 0.01539   1st Qu.: 0.01113
Median : 0.31646   Median : 0.29684   Median : 0.2752   Median : 0.24000   Median : 0.21104
Mean   : 0.42435   Mean   : 0.41252   Mean   : 0.3920   Mean   : 0.35926   Mean   : 0.33203
3rd Qu.: 0.83020   3rd Qu.: 0.81098   3rd Qu.: 0.7532   3rd Qu.: 0.66542   3rd Qu.: 0.60255
Max.   : 6.45530   Max.   : 6.38050   Max.   : 5.3914   Max.   : 5.14685   Max.   : 4.92625

    UTIL6               AVG_UTIL          BILL_GROWTH2       BILL_GROWTH3       BILL_GROWTH4
Min.   :-1.212868   Min.   :-0.23259   Min.   :-384675   Min.   :-512650   Min.   :-418926
1st Qu.: 0.008023   1st Qu.: 0.03114   1st Qu.:  -2145   1st Qu.:  -2596   1st Qu.:  -3434
Median : 0.184772   Median : 0.28614   Median :      0   Median :      0   Median :      0
Mean   : 0.318171   Mean   : 0.37306   Mean   :  -1920   Mean   :  -2276   Mean   :  -3823
3rd Qu.: 0.582484   3rd Qu.: 0.68764   3rd Qu.:   1572   3rd Qu.:   1389   3rd Qu.:   1022
Max.   : 3.885550   Max.   : 5.36431   Max.   : 489972   Max.   : 391348   Max.   : 429981

 BILL_GROWTH5       BILL_GROWTH6       UTIL_GROWTH2       UTIL_GROWTH3       UTIL_GROWTH4
Min.   :-432730   Min.   :-400000   Min.   :-2.63080   Min.   :-4.91340   Min.   :-3.02322
1st Qu.:  -2705   1st Qu.:  -1625   1st Qu.:-0.01986   1st Qu.:-0.02349   1st Qu.:-0.02800
Median :      0   Median :      0   Median : 0.00000   Median : 0.00000   Median : 0.00000
Mean   :  -3043   Mean   :  -1432   Mean   :-0.01182   Mean   :-0.02049   Mean   :-0.03277
3rd Qu.:    996   3rd Qu.:   1184   3rd Qu.: 0.01731   3rd Qu.: 0.01496   3rd Qu.: 0.01078
Max.   : 341696   Max.   : 381629   Max.   : 1.63324   Max.   : 1.98697   Max.   : 1.68050
```

```
     UTIL1              UTIL2              UTIL3             UTIL4              UTIL5
 Min.   :-0.61989   Min.   :-1.39554   Min.   :-1.0251   Min.   :-1.04330   Min.   :-0.87674
 1st Qu.: 0.02191   1st Qu.: 0.01863   1st Qu.: 0.0162   1st Qu.: 0.01539   1st Qu.: 0.01113
 Median : 0.31646   Median : 0.29684   Median : 0.2752   Median : 0.24000   Median : 0.21104
 Mean   : 0.42435   Mean   : 0.41252   Mean   : 0.3920   Mean   : 0.35926   Mean   : 0.33203
 3rd Qu.: 0.83020   3rd Qu.: 0.81098   3rd Qu.: 0.7532   3rd Qu.: 0.66542   3rd Qu.: 0.60255
 Max.   : 6.45530   Max.   : 6.38050   Max.   : 5.3914   Max.   : 5.14685   Max.   : 4.92625

     UTIL6               AVG_UTIL          BILL_GROWTH2      BILL_GROWTH3      BILL_GROWTH4
 Min.   :-1.212868   Min.   :-0.23259   Min.   :-384675   Min.   :-512650   Min.   :-418926
 1st Qu.: 0.008023   1st Qu.: 0.03114   1st Qu.:  -2145   1st Qu.:  -2596   1st Qu.:  -3434
 Median : 0.184772   Median : 0.28614   Median :      0   Median :      0   Median :      0
 Mean   : 0.318171   Mean   : 0.37306   Mean   :  -1920   Mean   :  -2276   Mean   :  -3823
 3rd Qu.: 0.582484   3rd Qu.: 0.68764   3rd Qu.:   1572   3rd Qu.:   1389   3rd Qu.:   1022
 Max.   : 3.885550   Max.   : 5.36431   Max.   : 489972   Max.   : 391348   Max.   : 429981

  BILL_GROWTH5      BILL_GROWTH6      UTIL_GROWTH2       UTIL_GROWTH3       UTIL_GROWTH4
 Min.   :-432730   Min.   :-400000   Min.   :-2.63080   Min.   :-4.91340   Min.   :-3.02322
 1st Qu.:  -2705   1st Qu.:  -1625   1st Qu.:-0.01986   1st Qu.:-0.02349   1st Qu.:-0.02800
 Median :      0   Median :      0   Median : 0.00000   Median : 0.00000   Median : 0.00000
 Mean   :  -3043   Mean   :  -1432   Mean   :-0.01182   Mean   :-0.02049   Mean   :-0.03277
 3rd Qu.:    996   3rd Qu.:   1184   3rd Qu.: 0.01731   3rd Qu.: 0.01496   3rd Qu.: 0.01078
 Max.   : 341696   Max.   : 381629   Max.   : 1.63324   Max.   : 1.98697   Max.   : 1.68050

  UTIL_GROWTH5       UTIL_GROWTH6       MAX_BILL_AMT     MAX_PAY_AMT           DLQ1              DLQ2
 Min.   :-1.99750   Min.   :-2.01725   Min.   : -2900   Min.   :      0   Min.   :-597607   Min.   :-609666
 1st Qu.:-0.02220   1st Qu.:-0.01460   1st Qu.: 10051   1st Qu.:   2196   1st Qu.: -56954   1st Qu.: -53412
 Median : 0.00000   Median : 0.00000   Median : 31588   Median :   5000   Median : -16804   Median : -16070
 Mean   :-0.02723   Mean   :-0.01386   Mean   : 60426   Mean   :  15621   Mean   : -43640   Mean   : -41149
 3rd Qu.: 0.01021   3rd Qu.: 0.01180   3rd Qu.: 79120   3rd Qu.:  12201   3rd Qu.:      0   3rd Qu.:      0
 Max.   : 2.02000   Max.   : 2.00897   Max.   :823540   Max.   :1215471   Max.   : 696809   Max.   :1181069

      DLQ3              DLQ4              DLQ5             MAX_DLQ
 Min.   :-622699   Min.   :-823540   Min.   :-685789   Min.   :-823540.0
 1st Qu.: -48270   1st Qu.: -45625   1st Qu.: -44183   1st Qu.: -66104.5
 Median : -15220   Median : -13723   Median : -11790   Median : -23012.0
 Mean   : -38048   Mean   : -35348   Mean   : -34010   Mean   : -49998.7
 3rd Qu.:      0   3rd Qu.:      0   3rd Qu.:      0   3rd Qu.:   -767.2
 Max.   : 683112   Max.   : 355569   Max.   : 339603   Max.   :   4341.0
```

The LIMIT_BAL variable has the mean of $179,039 under the no default group and of $130,371 under the default group. The 25th, 50th, and 75th limit credit for the no default group are $70,000, $150,000, and $260,000 whereas for the default group are $50,000, $90,000, and $200,000. Therefore, the customers in default tend to have a lower credit limit, perhaps because they have a bad credit history, so they can’t obtain a higher credit limit.

Figure 13

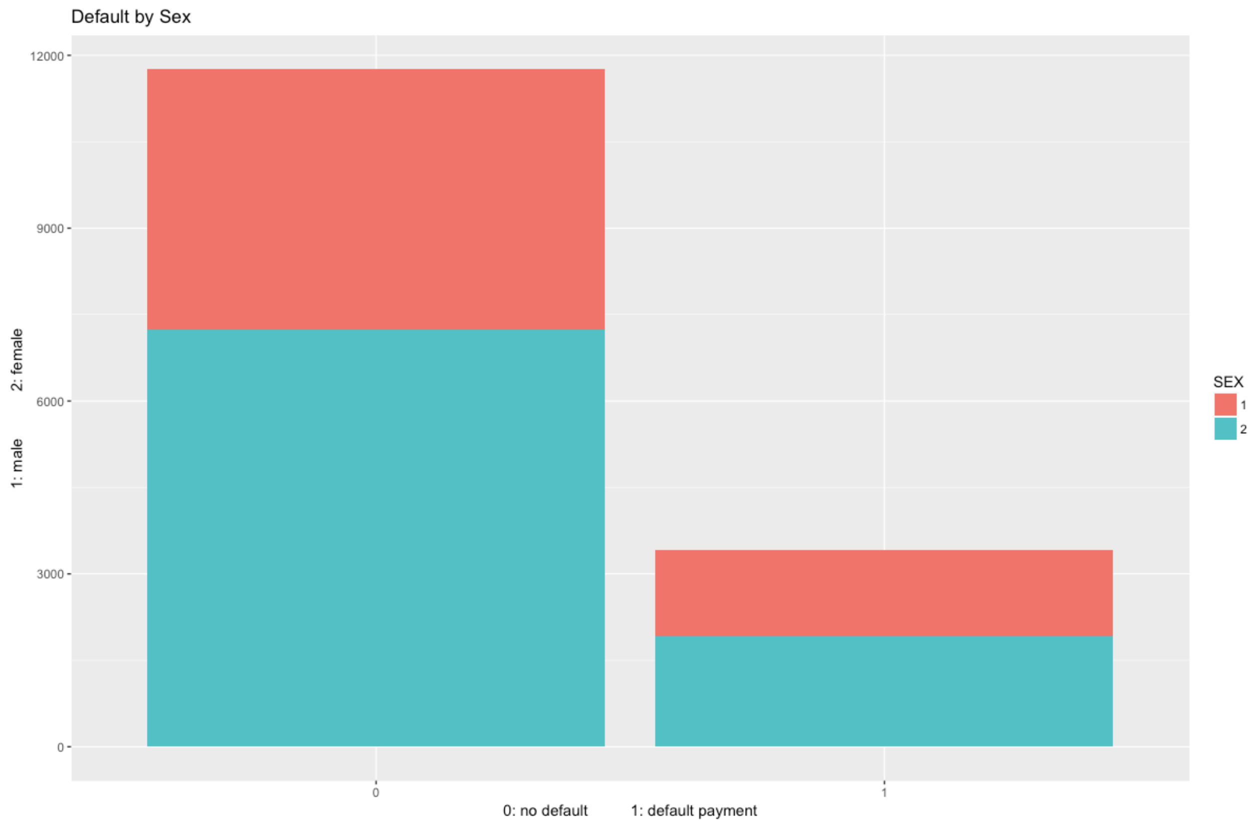


Among those in no default, 38% are male. Among those in default, 44% are male. Among both groups, 40% are male. Thus, the relationship between gender and payment default is not clear.

Figure 14

	male	female	Sum
no default	4523	7234	11757
default payment	1497	1926	3423
Sum	6020	9160	15180

Figure 15

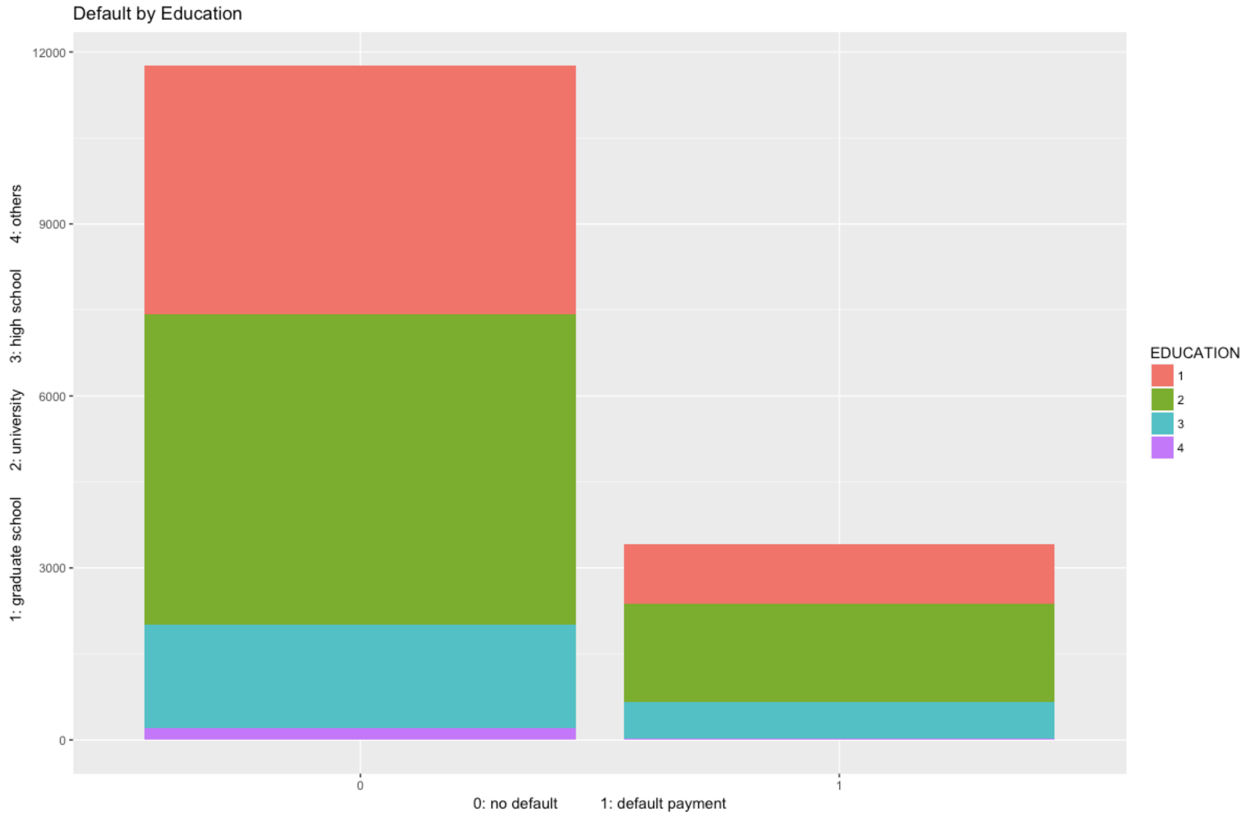


Among both default and no default groups, majority observations belong to the graduate school and university classes. The relationship between education and default payment thus is unclear.

Figure 16

	graduate school	university	high school	others	Sum
no default	4335	5403	1806	213	11757
default payment	1054	1712	637	20	3423
Sum	5389	7115	2443	233	15180

Figure 17

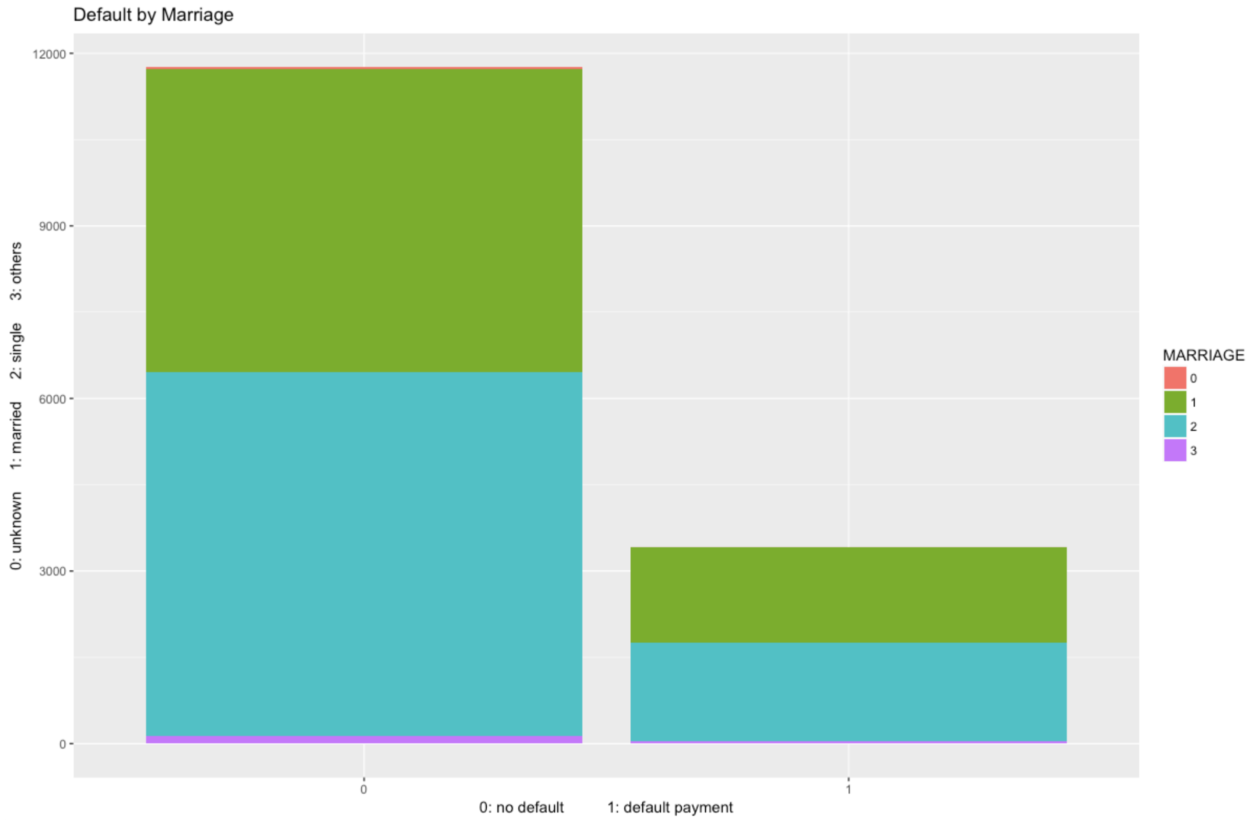


Similarly, we only have a few others and unknown observations in the MARRIAGE variable. There are more single observations than married, but the proportion is similar between default and no default group. Thus, the relationship between marriage and default payment is unclear.

Figure 18

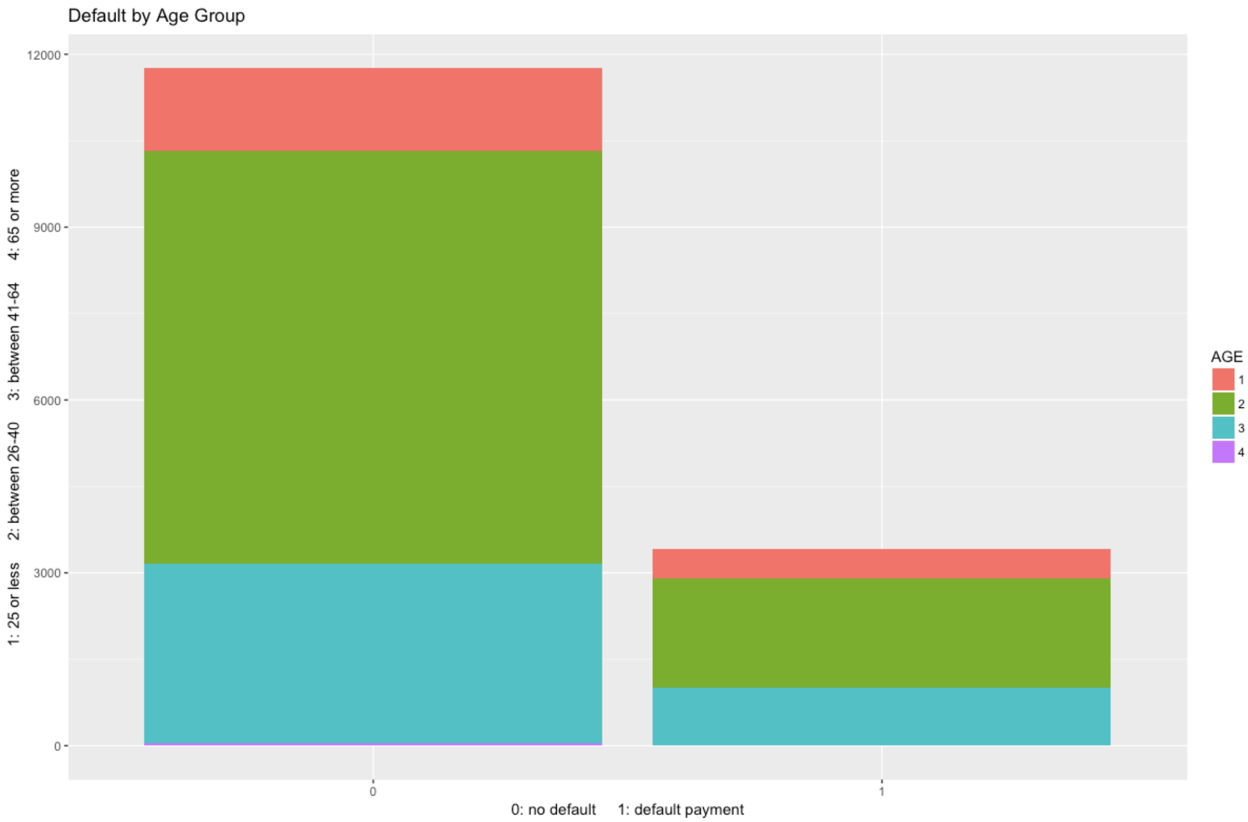
	unknown	married	single	others	Sum
no default	28	5277	6321	131	11757
default payment	1	1662	1716	44	3423
Sum	29	6939	8037	175	15180

Figure 19



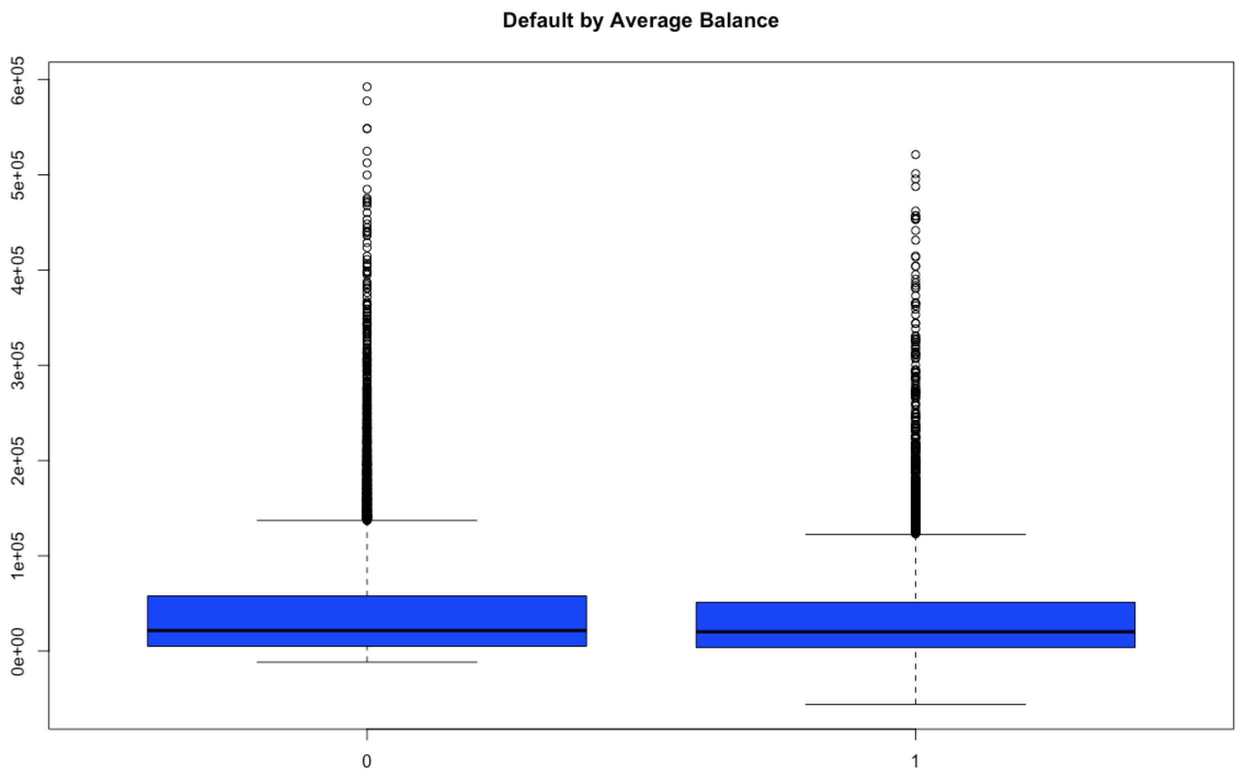
Similarly, using the figure below, both the default and no default groups have the same distribution of age. Thus, there's no clear relationship between age and default payment.

Figure 20



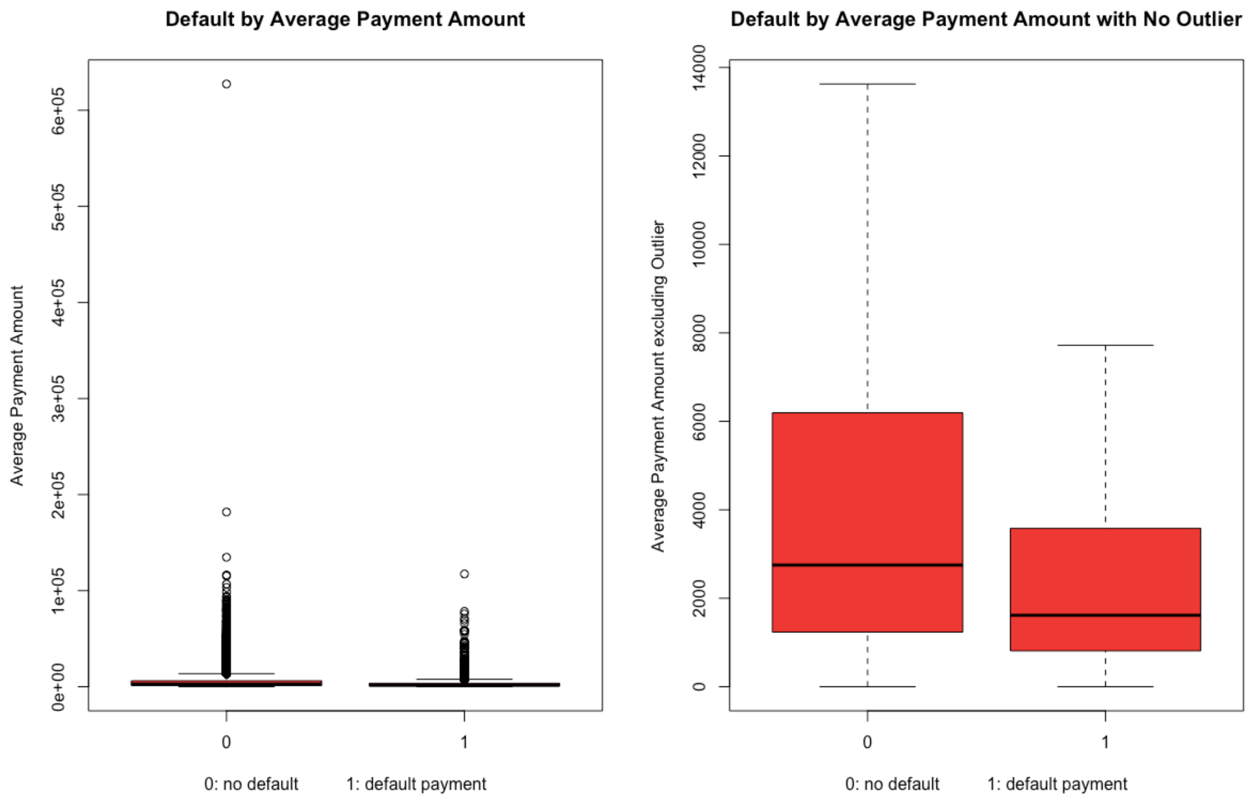
Under the no default group, the average bill amount is $ 45,119.49 whereas the mean for the default group is $44,300.94. The 25th, 50th, 75th percentile of average bill amount for the no default group are $5,040.66, $21,558.16, and $57,864.66 whereas for the default group are $3,700.66, $20,261.83, and $51,184.16. The min value for the default group is lower than that of the no default group, and the no default group has a larger outlier range than the default group. However, there's still no clear relationship between the bill amount and default payment.

Figure 20



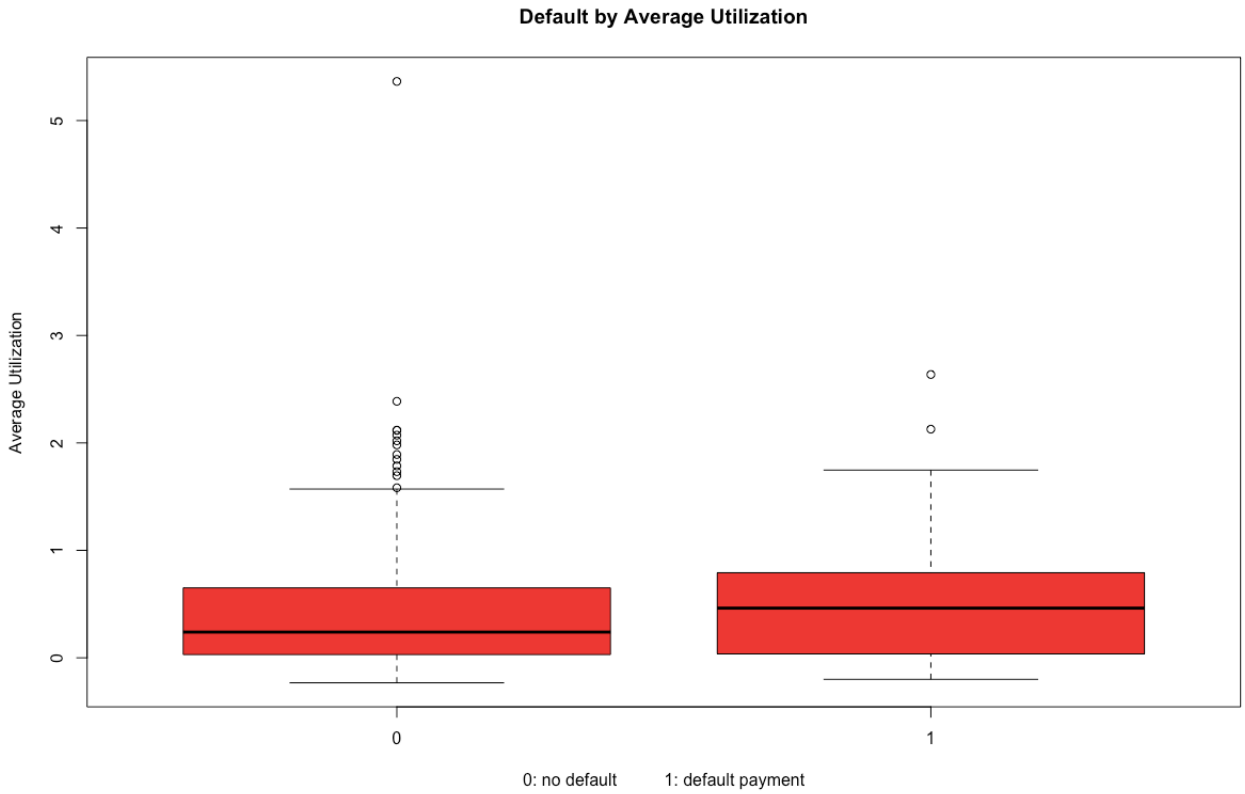
The average payment amount under no default group is $5,808.24 and under the default group is $3,355.24. The 25th, 50th, and 75th percentile of payment amount under the no default group are $1,236.16, $2,750.00, $6,194.33 and under the default group are $811.66, $1,614.50, $3,578.33. Under the no default group, there's a significant upper outlier that skews the dataset. If we remove this outlier, the payment of the no default group is higher than that of the default group. Therefore, perhaps customers who pay less amount tends to pay default customers.

Figure 21



The average utilization among no default group is 35% and among the default group is 45%, which is a 10% difference. The 25th, 50th, 75th utilization percentile in the no default group are 3%, 24%, 65% and in the default group are 4%, 46%, 79%. Perhaps people who utilize less of their credit limit are more likely to default on payment.

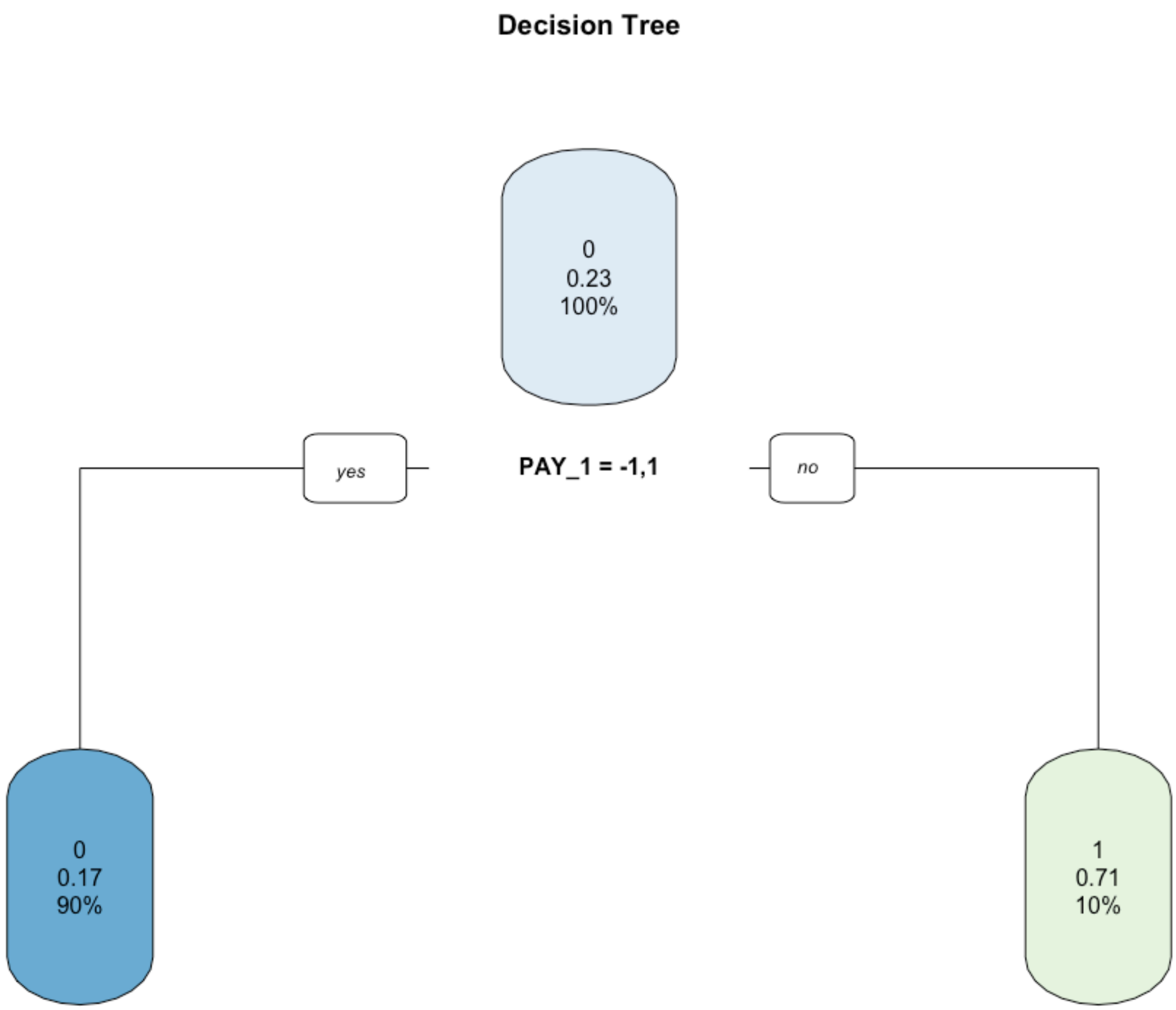
Figure 22



Section 4b: Model Based EDA

When we fit a decision tree using rpart and plot the tree dendogram in R, we have the following result.

Figure 23

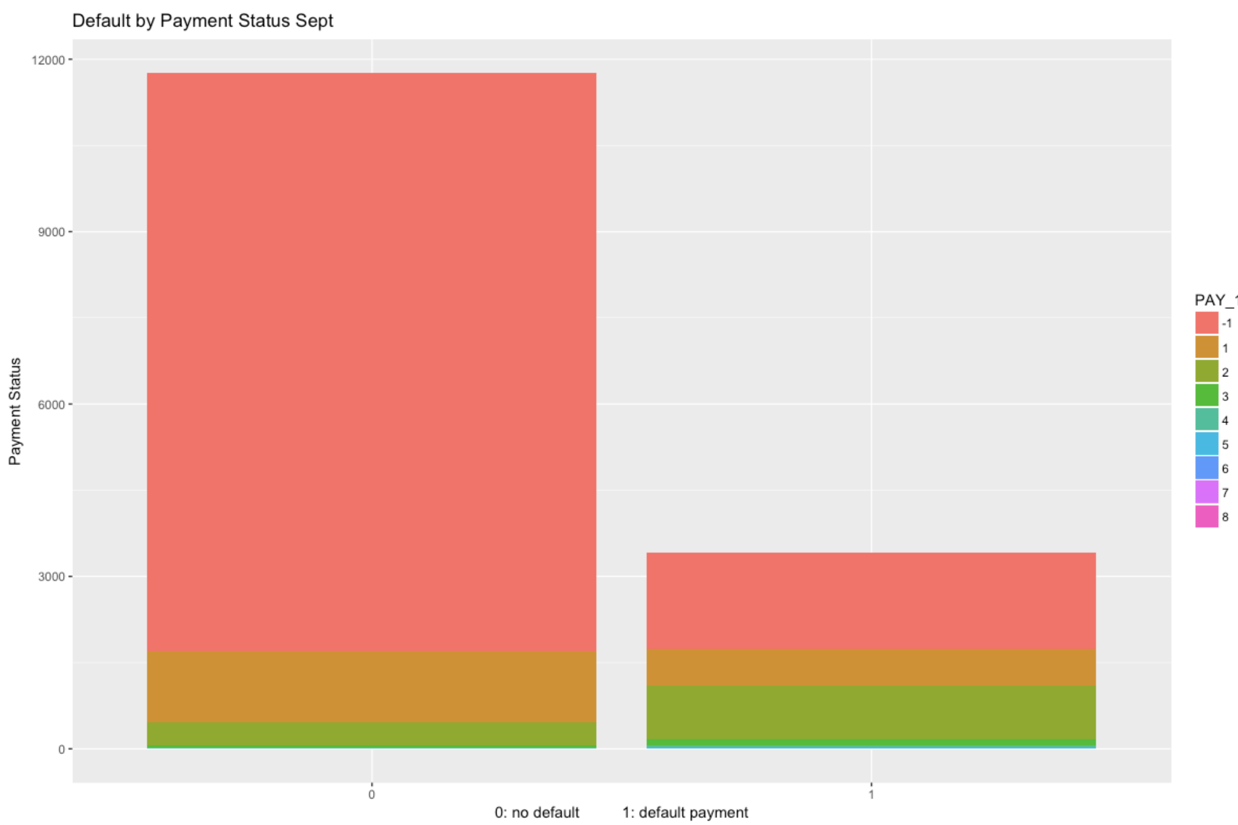


The decision tree result above keys on the variable PAY_1 or the payment status of the month of September. So we go back to the dataset to plot this variable to see a separation between two classes of the variable.

Figure 24

	no default	default payment	Sum
pay in full	10053	1689	11742
1 month late	1246	628	1874
2 month late	403	937	1340
3 month late	35	118	153
4 month late	8	31	39
5 month late	7	6	13
6 month late	3	4	7
7 month late	0	6	6
8 month late	2	4	6
Sum	11757	3423	15180

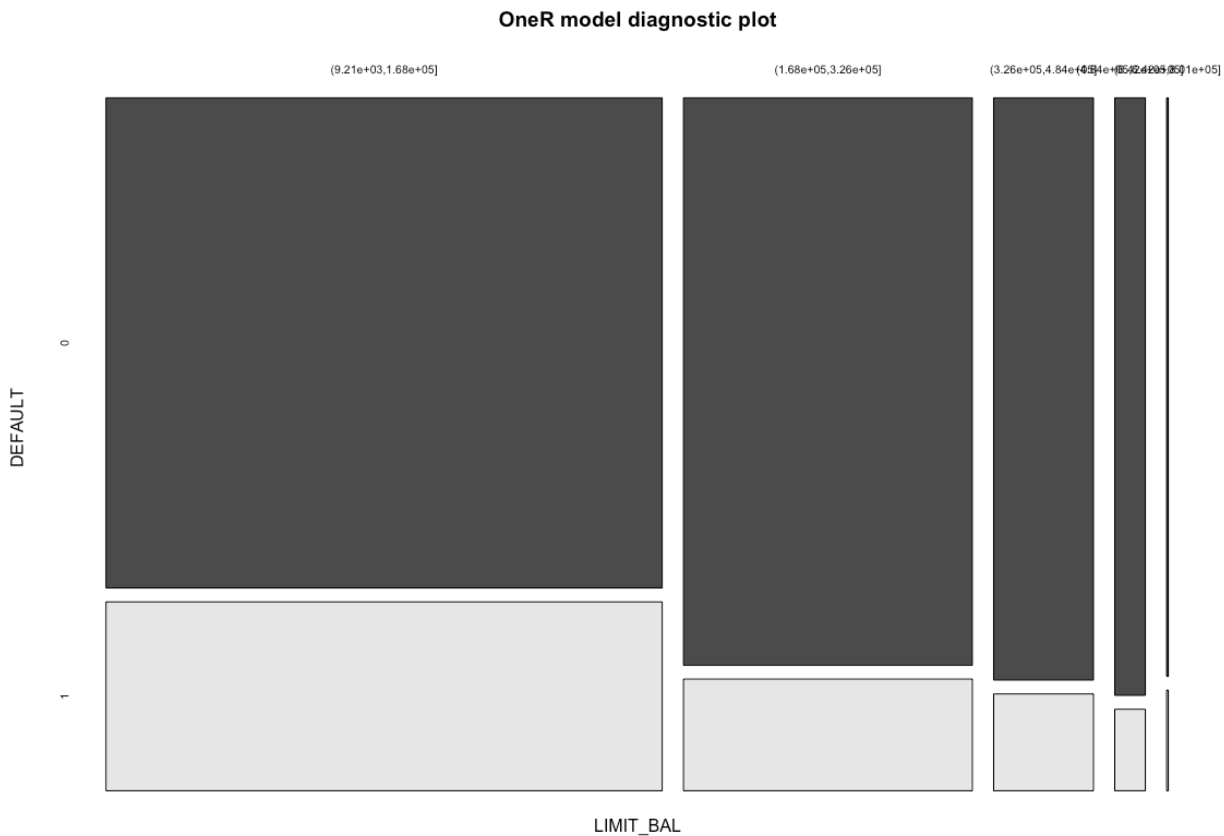
Figure 25



Using the results above, among those with no default, 86% paid in full in September. Among those with default payment, only 49% paid in full in September. Thus, perhaps customers who fail their payment right at the beginning are more likely to default on their payment later on.

```
Call:
OneR.formula(formula = DEFAULT ~ LIMIT_BAL + SEX + EDUCATION +
    MARRIAGE + AGE + AVG_BILL_AMT + AVG_PAY_AMT + AVG_UTIL +
    MAX_BILL_AMT + MAX_PAY_AMT + MAX_DLQ, data = model_train,
    verbose = TRUE)

Rules:
If LIMIT_BAL = (9.21e+03,1.68e+05] then DEFAULT = 0
If LIMIT_BAL = (1.68e+05,3.26e+05] then DEFAULT = 0
If LIMIT_BAL = (3.26e+05,4.84e+05] then DEFAULT = 0
If LIMIT_BAL = (4.84e+05,6.42e+05] then DEFAULT = 0
If LIMIT_BAL = (6.42e+05,8.01e+05] then DEFAULT = 0

Accuracy:
11757 of 15180 instances classified correctly (77.45%)

Contingency table:
       LIMIT_BAL
DEFAULT (9.21e+03,1.68e+05] (1.68e+05,3.26e+05] (3.26e+05,4.84e+05] (4.84e+05,6.42e+05] (6.42e+05,8.01e+05]
    0                * 6236              * 3750              * 1330               * 418                * 23
    1                  2403                 738                 221                  57                   4
    Sum                8639                4488                1551                 475                  27
       LIMIT_BAL
DEFAULT   Sum
    0   11757
    1    3423
    Sum 15180
---
Maximum in each column: '*'

Pearson's Chi-squared test:
X-squared = 325.36, df = 4, p-value < 2.2e-16
```

Using OneR to build model-based decision tree EDA, the result above shows that the model is statistically significant with p-value less than 0.05 alpha. The model has a high accuracy because its accuracy is 77.45%.



Furthermore, to analyze the model deeper, by examining the OneR plot above, the most significant predictor to forecast DEFAULT is the LIMIT_BAL variable, which means that the limit balance of an individual is a good indicator to determine whether the individual will default on their credit card payments.

```
Confusion matrix (absolute):
          Actual
Prediction     0     1   Sum
       0   11757  3423 15180
       1       0     0     0
       Sum 11757  3423 15180

Confusion matrix (relative):
          Actual
Prediction    0    1  Sum
       0   0.77 0.23 1.00
       1   0.00 0.00 0.00
       Sum 0.77 0.23 1.00

Accuracy:
0.7745 (11757/15180)

Error rate:
0.2255 (3423/15180)

Error rate reduction (vs. base rate):
0 (p-value = 0.5046)
```

Thus far, the OneR decision tree model is proven solid. However, when the confusion matrix above is examined, the model doesn't perform as well as expected. Specifically, the model has a high accuracy metric because it predicts all cases to be under no default, which is the majority of the actual observations. Therefore, the OneR decision tree model is actually useless because it doesn't predict anything at all. It simply mirrors after the actual breakdown of default vs. no default in the response variable.

In conclusion, the decision tree results above are interesting but should not be used alone as the predictive model for this project. Therefore, it's necessary to build other more sophisticated predictive models such as random forest, gradient boosting, logistic regression, and neural networks, which are presented in section 5 of this report.

Section 5: Predictive Modeling – Methods and Results

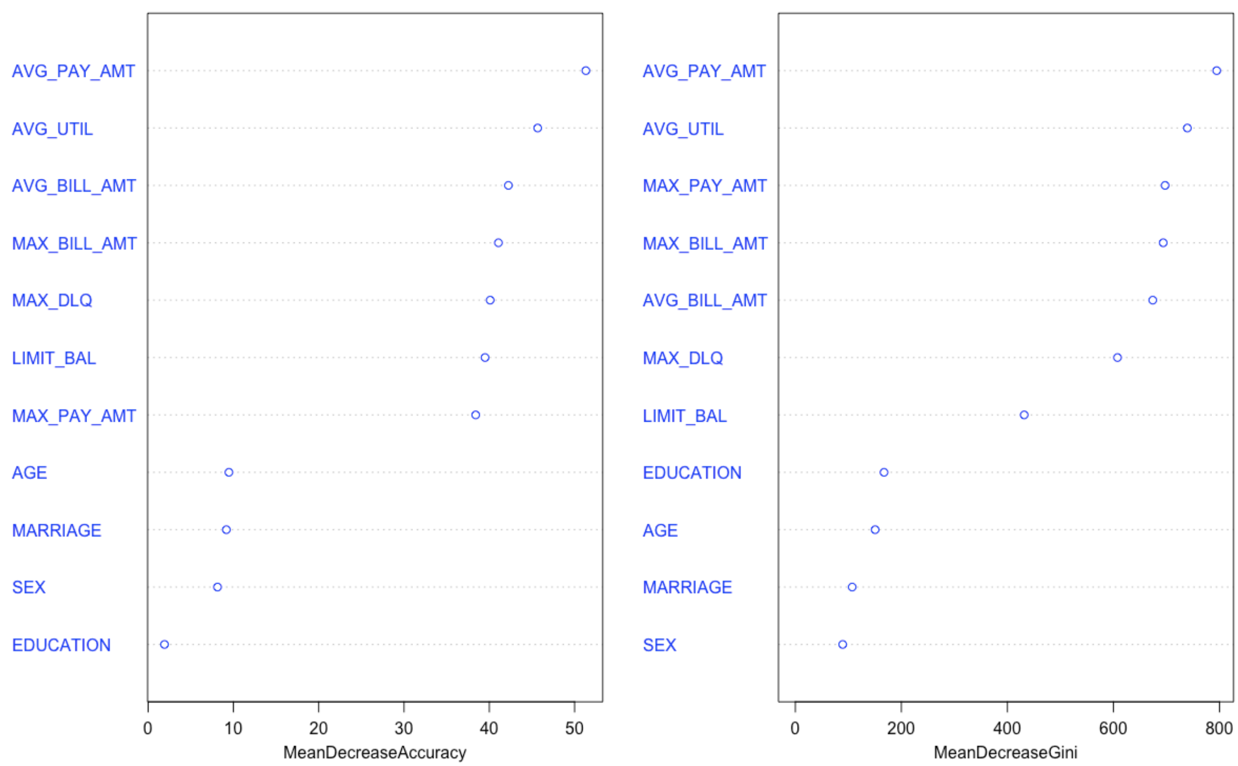
In this section, four predictive models are developed 1) random forest 2) gradient boosting 3) logistic regression with variable selection 4) support vector machine. For each model, relevant and useful mode output, in-sample model performance results on train dataset, and out-of-sample performance results on test dataset are presented. Three universal metrics are utilized to measure and compare four models' performance, including 1) true positive rate (TPR) 2) false positive rate (FPR) 3) accuracy.

Section 5a: Model #1 – Random Forest

A random forest is developed using the train dataset with 300 trees.

Figure 26

Variable Importance Plot - Random Forest Model



The variable importance plot is an expected output for random forest modeling. There are two approaches or two types of measurement utilized in a variable importance plot: accuracy and gini. Accuracy tests to see how worse the model performs without each variable. Gini goes deeper into decision tree to measure how pure the nodes are at the end of each tree. Both measurements indicate that the higher the score, the more significant the variable. In the variable importance plot above, we can conclude the following about the predictors.

- The most significant variable is AVG_PAY_AMT or the average payment amount
- The payment-related predictors are also significant
- The demographic variables, however, are not significant, including AGE, MARRIAGE, SEX, EDUCATION

Figure 27

predict1train	No Default Actual	Default Actual	Sum
No Default Predicted	11748	91	11839
Default Predicted	9	3332	3341
Sum	11757	3423	15180

Using the classification above, we can calculate the following performance metrics for the train dataset.

- TPR = 3332 / 3423 = 97.34%
- FPR = 9 / 11,757 = 0.0766%
- Accuracy = (11,748 + 3332) / 15,180 = 99.34%

Figure 28

predict1test	No Default Actual	Default Actual	Sum
No Default Predicted	5549	1348	6897
Default Predicted	217	209	426
Sum	5766	1557	7323

Using the classification above, we can calculate the following performance metrics for the test dataset.

- TPR = 209 / 1557 = 13.42%
- FPR = 217 / 5766 = 3.7634%
- Accuracy = (5549 + 209) / 7323 = 78.63%

Section 5b: Model #2 – Gradient Boosting

In this model, a cutoff of 0.4633751 is utilized to determine the classification of the predicted values with 0 as no default and 1 as default payment.

Figure 29

predict2train	No Default Actual	Default Actual	Sum
No Default Predicted	11570	3155	14725
Default Predicted	187	268	455
Sum	11757	3423	15180

The classification table is used to calculate the following performance metrics for the train dataset using gradient boosting model.

- TPR = 268 / 3423 = 7.8294%
- FPR = 187 / 11,757 = 1.5905%
- Accuracy = (11,570 + 268) / 15,180 = 77.98%

Figure 30

predict2test	No Default Actual	Default Actual	Sum
No Default Predicted	5666	1433	7099
Default Predicted	100	124	224
Sum	5766	1557	7323

The classification table above is used to calculate the following performance metrics for the test dataset.

- TPR = 124 / 1557 = 7.964%
- FPR = 100 / 5766 = 1.7343%
- Accuracy = (5666 + 124) / 7323 = 79.07%

Section 5c: Mode #3 – Logistic Regression with Variable Section

Using the results of the first two models of random forest and gradient boosting, we identify a pool of interesting predictors to use in the logistic regression model. Specifically, we remove the four insignificant demographic variables MARRIAGE, AGE, SEX, EDUCATION and leave the seven payment-related predictors remain in the predictor pool to develop a logistic regression model. Then among these seven variables, we use the stepwise automatic variable selection method to arrive at the optimal logistic regression model.

Figure 31

```
Call:
glm(formula = DEFAULT ~ LIMIT_BAL + AVG_BILL_AMT + AVG_PAY_AMT +
    AVG_UTIL + MAX_BILL_AMT + MAX_PAY_AMT + MAX_DLQ, family = binomial(),
    data = model_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4782  -0.7867  -0.6500  -0.2307   4.9009

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.342e-01  5.612e-02 -14.865  < 2e-16 ***
LIMIT_BAL    -1.960e-06  2.622e-07  -7.473 7.82e-14 ***
AVG_BILL_AMT  9.209e-06  1.826e-06   5.043 4.59e-07 ***
AVG_PAY_AMT  -1.787e-04  1.638e-05 -10.905  < 2e-16 ***
AVG_UTIL      2.838e-01  8.895e-02   3.190  0.00142 **
MAX_BILL_AMT -8.600e-06  1.719e-06  -5.003 5.65e-07 ***
MAX_PAY_AMT   2.970e-05  3.125e-06   9.504  < 2e-16 ***
MAX_DLQ      -4.876e-06  2.102e-06  -2.319  0.02037 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16205  on 15179  degrees of freedom
Residual deviance: 15464  on 15172  degrees of freedom
AIC: 15480

Number of Fisher Scoring iterations: 6
```

The result above shows that all seven predictors are statistically significant at 95% confidence level, each with p-value less than 0.05 alpha. Thus, the stepwise automatic variable selection algorithm indicate that all variables in the model are significant. Among these seven predictors, LIMIT_BAL, AVG_PAY_AMT, MAX_BILL_AMT, MAX_DLQ have negative coefficients, which meant that they have a negative correlation with the dependent variable. In other words, the lower the limit balance and the lower the average payment amount and the lower of the maximum bill and the lower the maximum delinquency value, the higher the probability of default on payment. The other three predictors AVG_BILL_AMT, AVG_UTIL, MAX_PAY_AMT have positive coefficient, which mean that these variables have a positive correlation with the response variable. In other words, the higher the average billing amount and the higher the utilization rate and the higher the maximum payment amount, the higher the chance of default on payment.

Figure 32

	No Default Actual	Default Actual	Sum
No Default Predicted	7443	1332	8775
Default Predicted	4314	2091	6405
Sum	11757	3423	15180

The classification table above is used to calculate the following performance metrics for the train dataset.

- TPR = 2091 / 3423 = 61.09%
- FPR = 4314 / 11,757 = 36.69%
- Accuracy = (7443 + 2091) / 15,180 = 62.81%

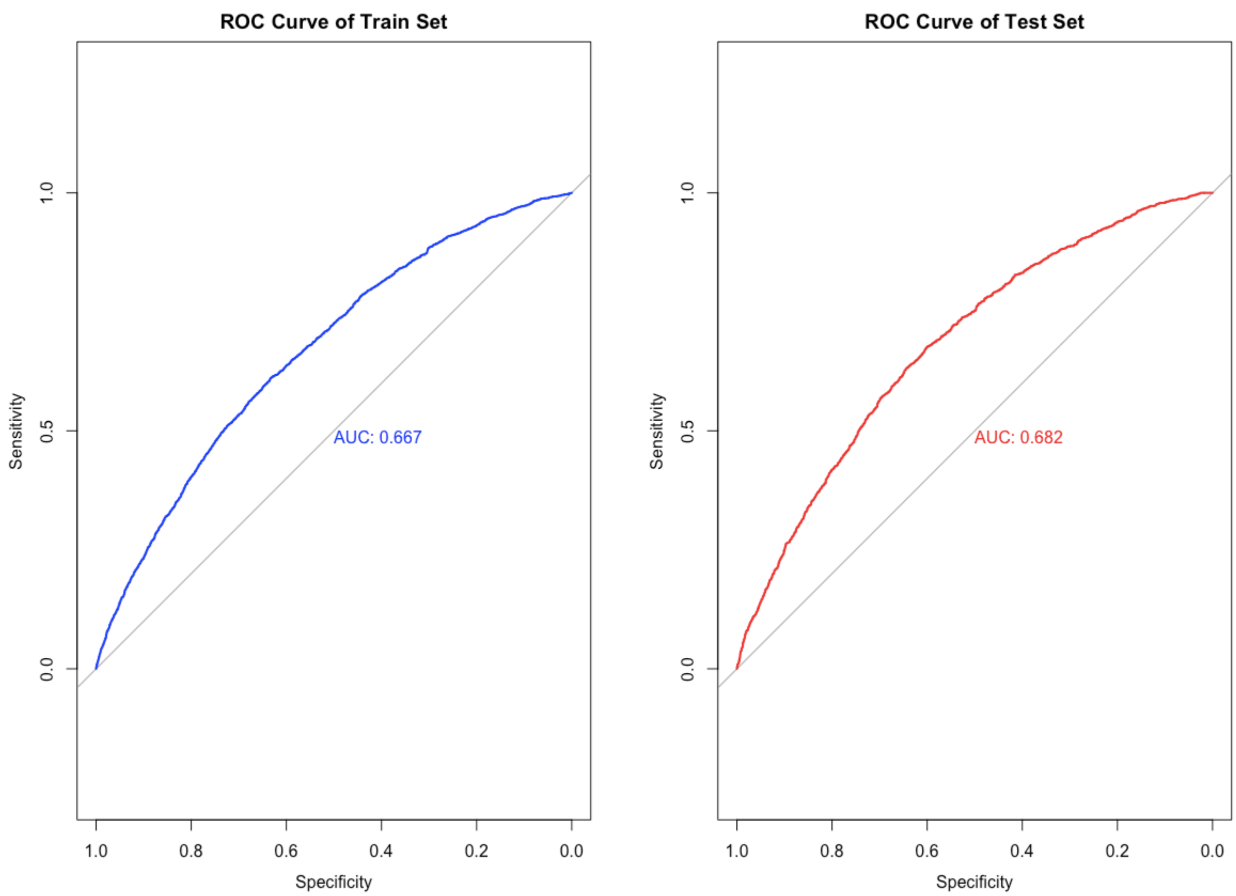
Figure 33

	No Default Actual	Default Actual	Sum
No Default Predicted	3470	506	3976
Default Predicted	2296	1051	3347
Sum	5766	1557	7323

The classification matrix above is used to calculate the following performance metrics for the test dataset.

- TPR = 1051 / 1557 = 67.5%
- FPR = 2296 / 5766 = 39.82%
- Accuracy = (3470 + 1051) / 7323 = 61.74%

Figure 34



The two ROC curves and AUC above for train and test sets are very similar to one another. Thus, we can conclude that there's no overfitting issue in the logistic regression model.

Section 5d: Model #4 – Support Vector Machine (SVM)

Using linear kernel to build the SVM model, we generate the following results.

Figure 35

predict4train	No Default Actual	Default Actual	Sum
No Default Predicted	11757	3423	15180
Default Predicted	0	0	0
Sum	11757	3423	15180

The classification table above is used to calculate the following performance metrics for the train dataset.

- TPR = 0 / 3423 = 0%
- FPR = 0 / 11,757 = 0%
- Accuracy = (11,757 + 0) / 15,180 = 77.45%

From the results above, SVM is the worst model since it predicts all observations to belong to the no default group. In other words, the model predicts nothing and is useless.

Figure 36

predict4test	No Default Actual	Default Actual	Sum
No Default Predicted	5766	1557	7323
Default Predicted	0	0	0
Sum	5766	1557	7323

The situation is similar with the test result, as indicated in the classification table above.

- TPR = 0 / 1557 = 0%
- FPR = 0 / 5766 = 0%
- Accuracy = (5766 + 0) / 7323 = 78.74%

Thus, we can conclude that SVM is the worst model since it doesn't predict anything in both train and test datasets.

Figure 37

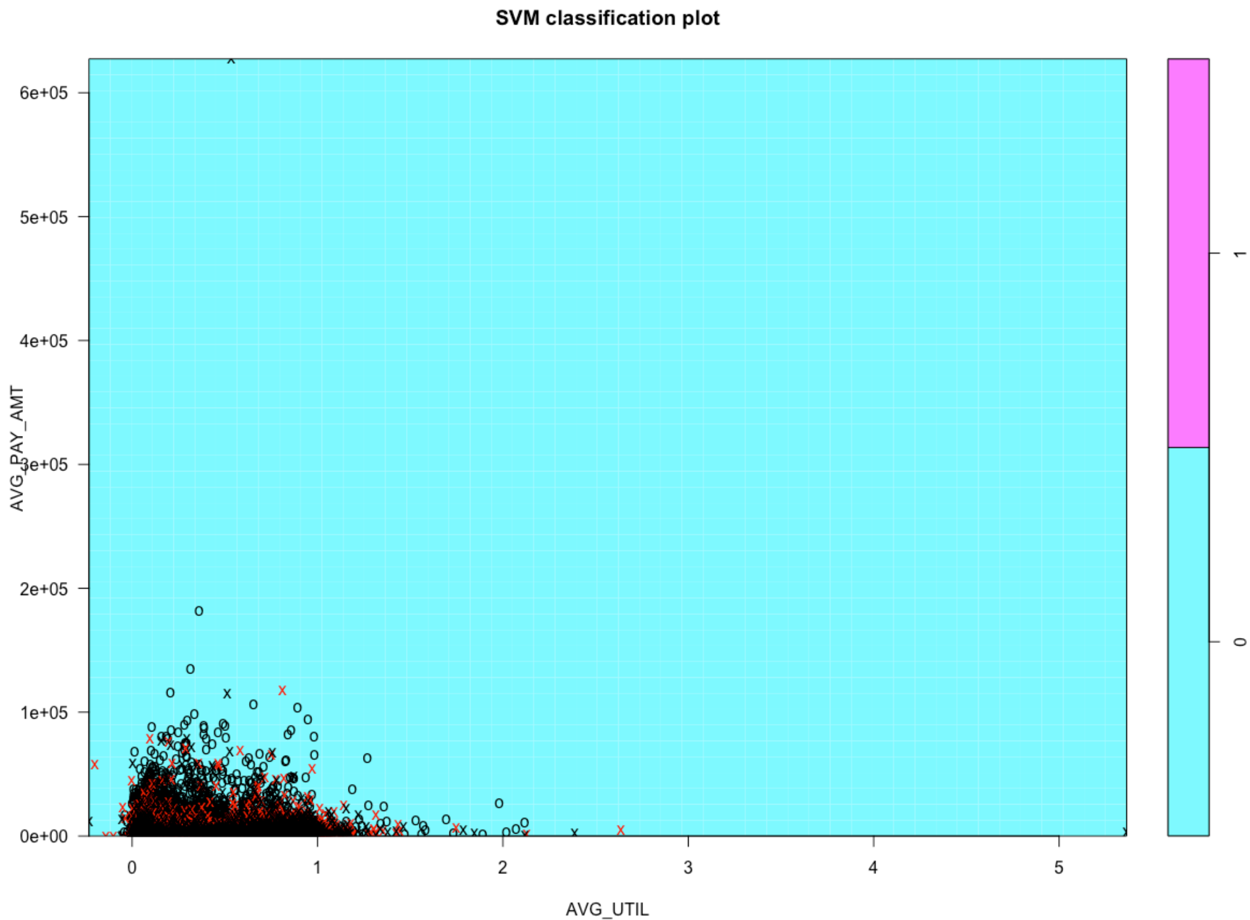

SVM classification plot
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
AVG_PAY_AMT
0
1
2
3
4
5
AVG_UTIL
1
0

Figure 38

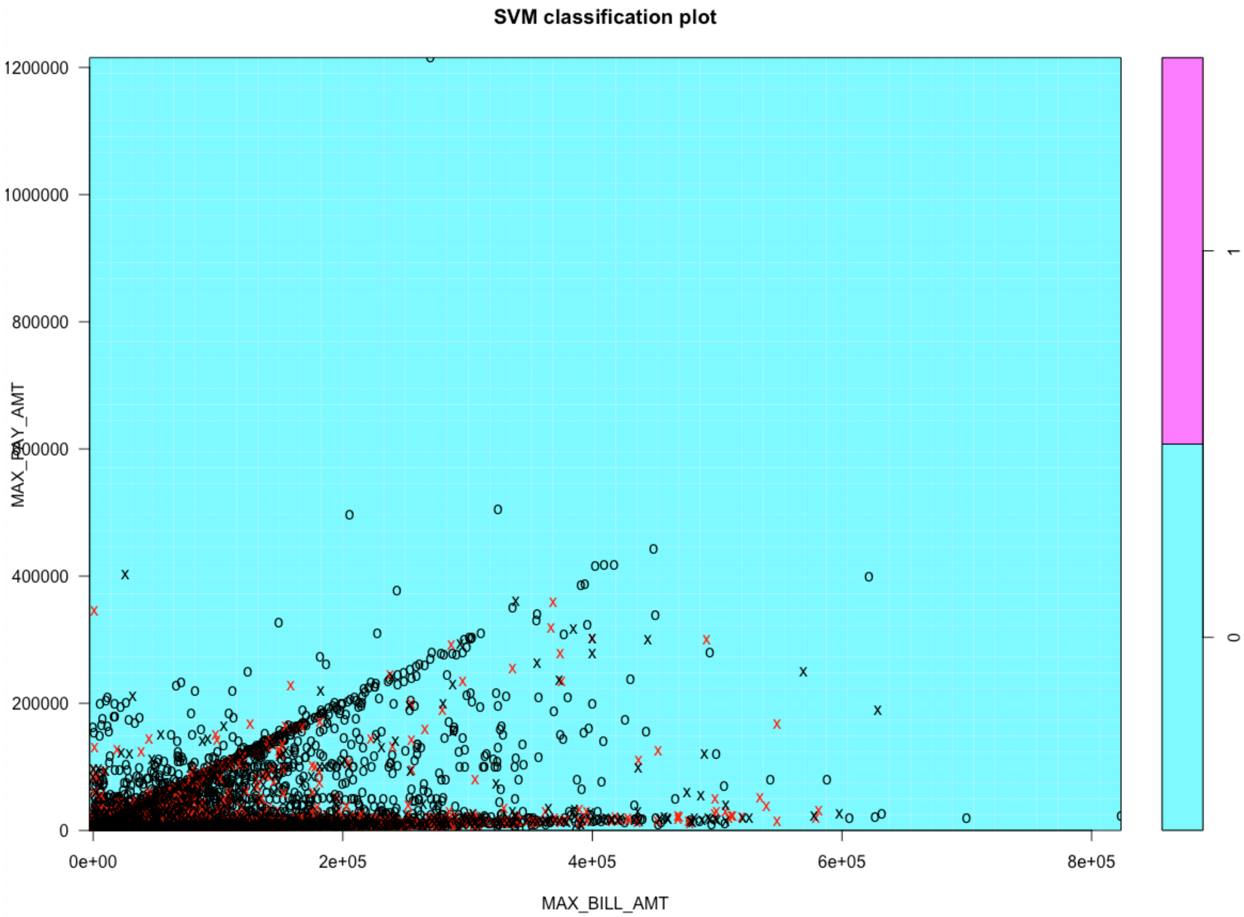
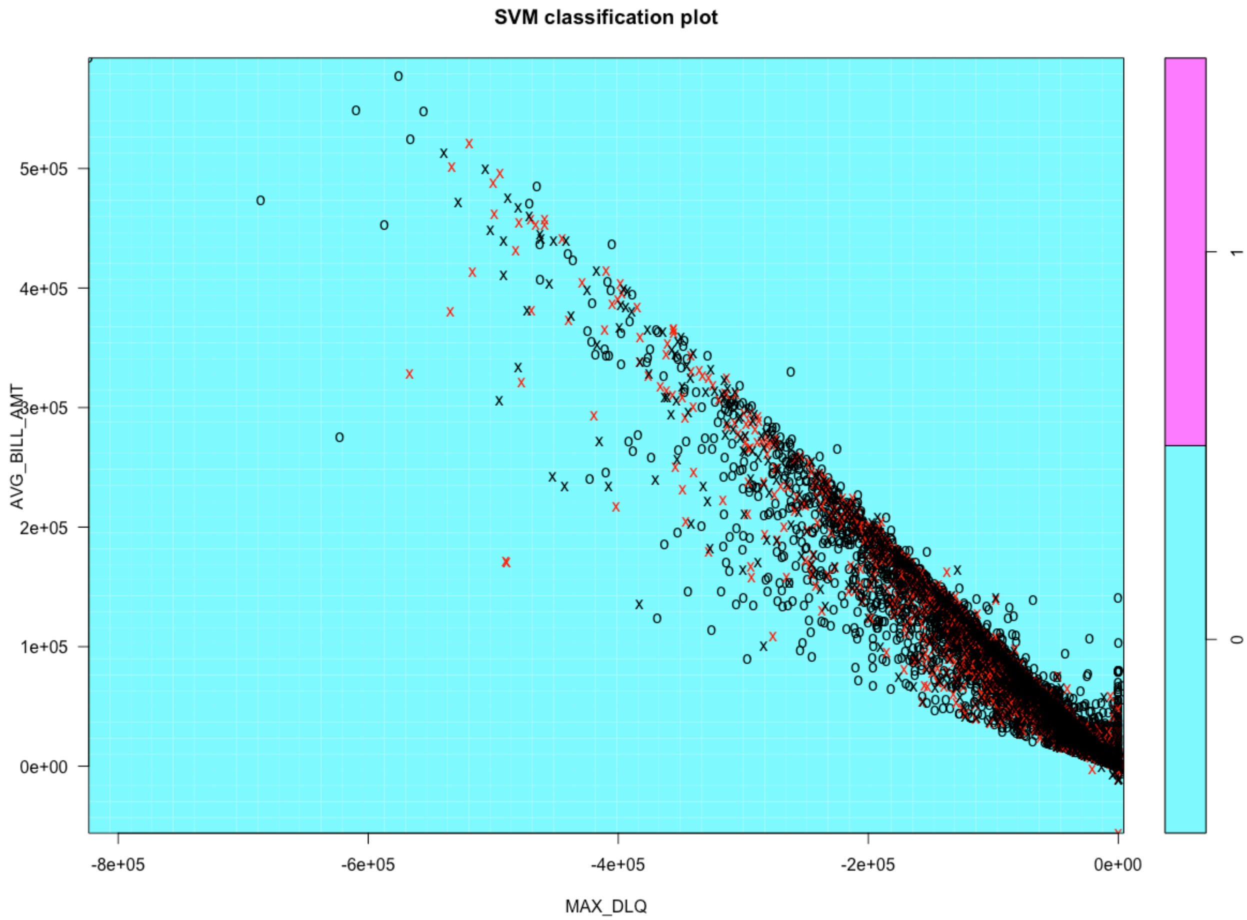

SVM classification plot
MAX_PAY_AMT
MAX_BILL_AMT
1200000
1000000
800000
600000
400000
200000
0
0e+00
2e+05
4e+05
6e+05
8e+05
1
0

Figure 39



The three classification plots or margin plots above show that there's no clear distinction between the default and no default observation. As a result, the plots above confirm the conclusion that the SVM model is useless since it doesn't predict anything.

Section 6: Comparison of Results

Using the results in section 5, below is the summary table that includes the performance metrics of all four models using with train and test datasets.

Figure 40

	Model #1	Model #2	Model #3	Model #4
Train Set				
TPR	97.34%	7.83%	61.09%	61.74%
FPR	0.08%	1.59%	36.69%	0.00%
Accuracy	99.34%	77.98%	62.81%	77.45%
Test Set				
TPR	13.42%	7.96%	67.50%	0.00%
FPR	3.76%	1.73%	39.82%	0.00%
Accuracy	78.63%	79.07%	61.74%	78.74%
Ranking	3	2	1	4

From the summary table above, model #4 SVM is the worst since it predicts all observations to be in no default category, which means that this model is useless and doesn't predict anything. The biggest problem with model #1 random forest is overfitting. In other words, the model trains the dataset very well to build a strong model, but it fails to apply to the test dataset. Model #2 gradient boosting has low TPR and FPR, so it's not as reliable. Model #3 logistic regression is the best model because there's no overfitting issue, and both TPR and FPR are reasonable. Because the DEFAULT variable has an unequal proportion of classes with the majority of observations falling into the no default category, accuracy is not a reliable performance metric. Thus, TPR and FPR carry more weight and indicate the model performance more accurately than the accuracy metric. As a result, using the summary table above, below is the model ranking based on the performance metrics using both train and test dataset.

1. Model #3: logistic regression with variable selection
2. Model #2: gradient boosting
3. Model #3: random forest
4. Model #4: SVM

Section 7: Conclusion

In conclusion, the credit card default project uses data from a research in Taiwan, aiming to study the customer default payments. The first half of the project is dedicated to get the data ready via understanding of the dataset, feature engineering, and EDA. The second half of the project is dedicated to building and comparing four predictive models: random forest, gradient boosting, logistic regression, and SVM. Using three measurement metrics TPR, FPR, and accuracy on both train and test data, the logistic regression model produces the best predictive outcomes. It doesn't overfit the dataset and has a balanced performance among all three metrics on both train and test dataset.

Future researchers are encouraged to approach the problem with the following recommendations. First, the data scientists should consider more relevant predictors. From the results of this project, demographics variables such as sex, marriage, age don't have significant impact on the response variable. Thus, data scientists should consider additional payment-related predictors such as FICO score, payment method, etc. Second, future modelers are encouraged to try different modeling techniques such as neural networks. Though the logistic regression model performs well, it can still be improved. Third, future researchers should consider options such as zero-based Poisson and zero-based negative binomial approaches along with the logistic regression model to address the imbalance of default vs. no default in response variable.