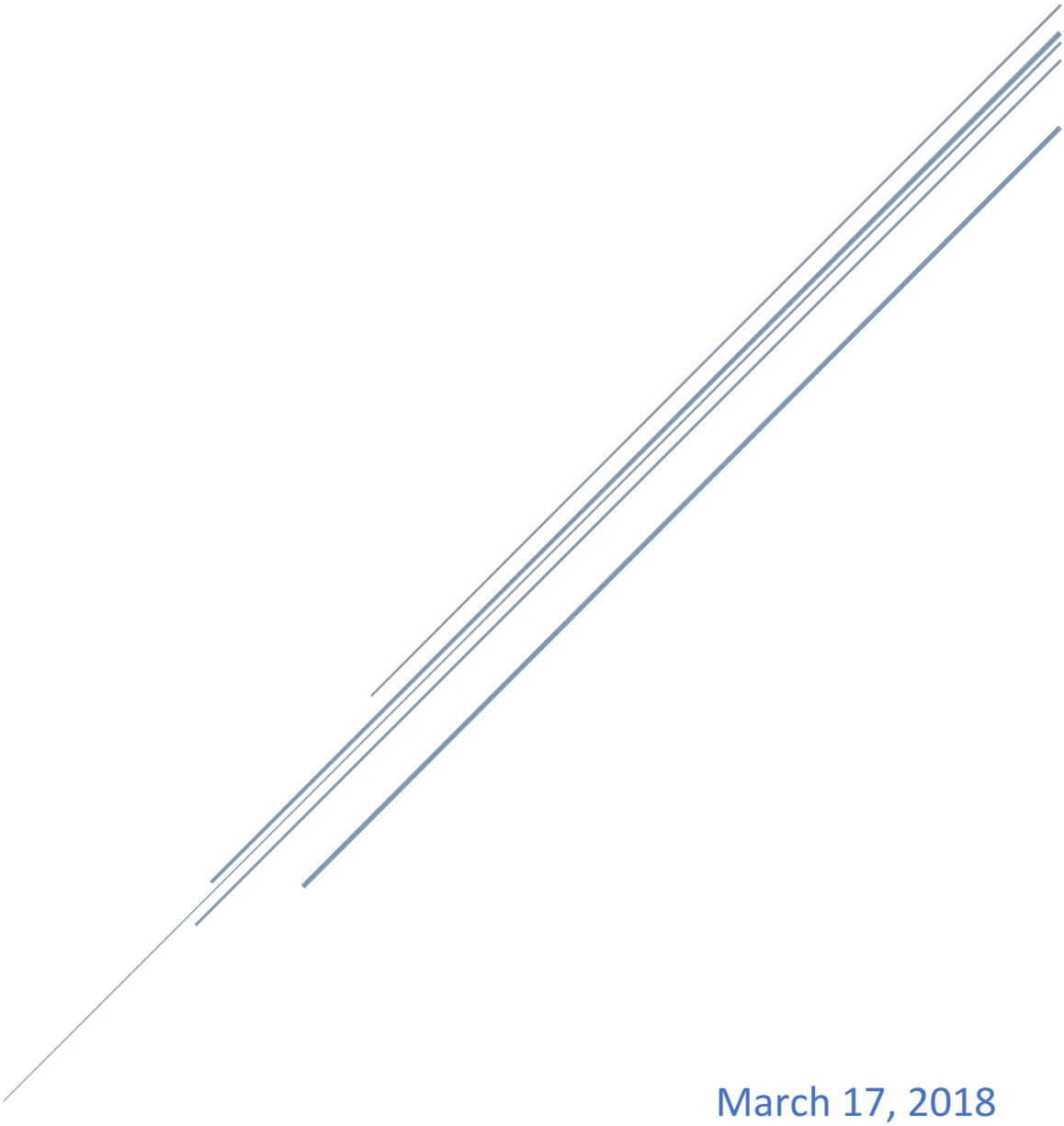
DATA ANALYSIS REPORT FOR CREDIT CARD DATASET

Diner's Club

Yutao Lu, Fangyu Yan, Wenyi Zhang, Yu Liu



March 17, 2018

1. Introduction

The dataset "Credit card data for participants.csv" is a dataset of monthly credit card data collection from January to July. The main goal of our data analysis is to construct a model to predict the kind of card holder who has high possibility of being defaulted and also the type of card holder who has relatively low possibility of default in the future. The response variable in the model is July_Payment_Status and others are predictors. The ideal model can give us a good prediction of which kind of person is most likely to default and which is not. Based on the results, the bank (in different period) can be better off.

2. Data Cleansing

The original dataset has 30602 observations and 24 variables including rows with missing attributes, duplicated rows, corrupted values and so on. Those values are meaningless and therefore need to be "cleaned" to some extent.

2.1 Cleaning Procedure

* Remove 17 duplicated rows.
* Delete all rows including meaningless data in column July_Payment_Status.
* Replace decimal values by integer values.
* Replace all meaningless (out of range) values by NA.

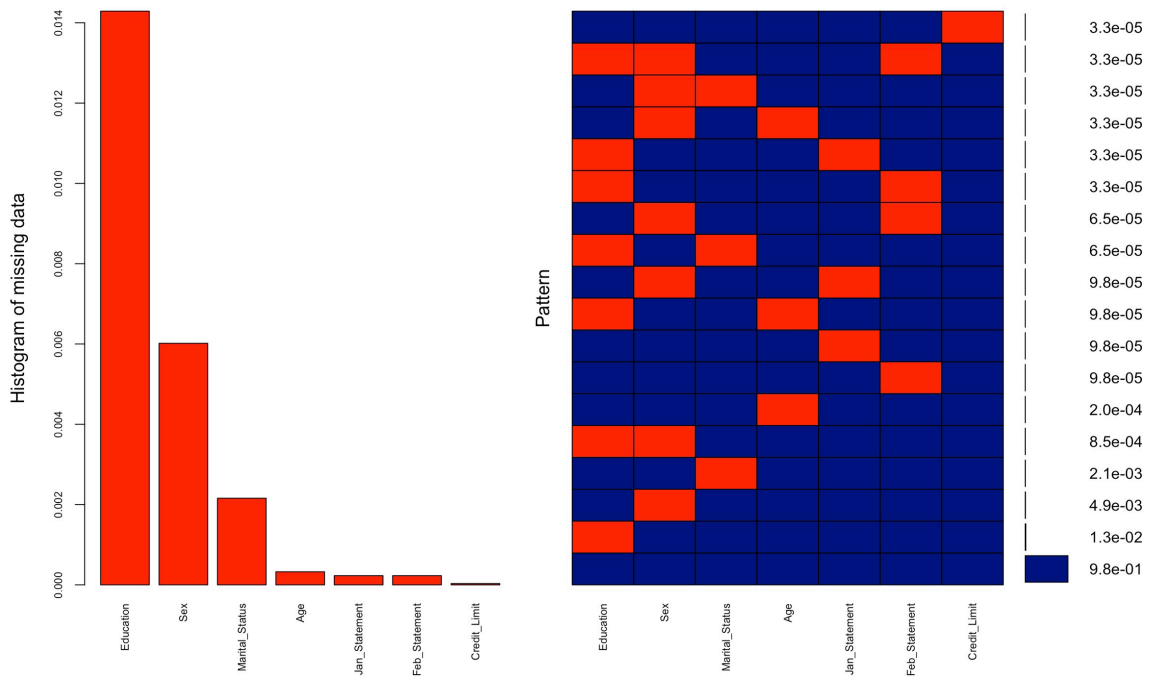
3. Data Analysis

3.1Data Adjustment

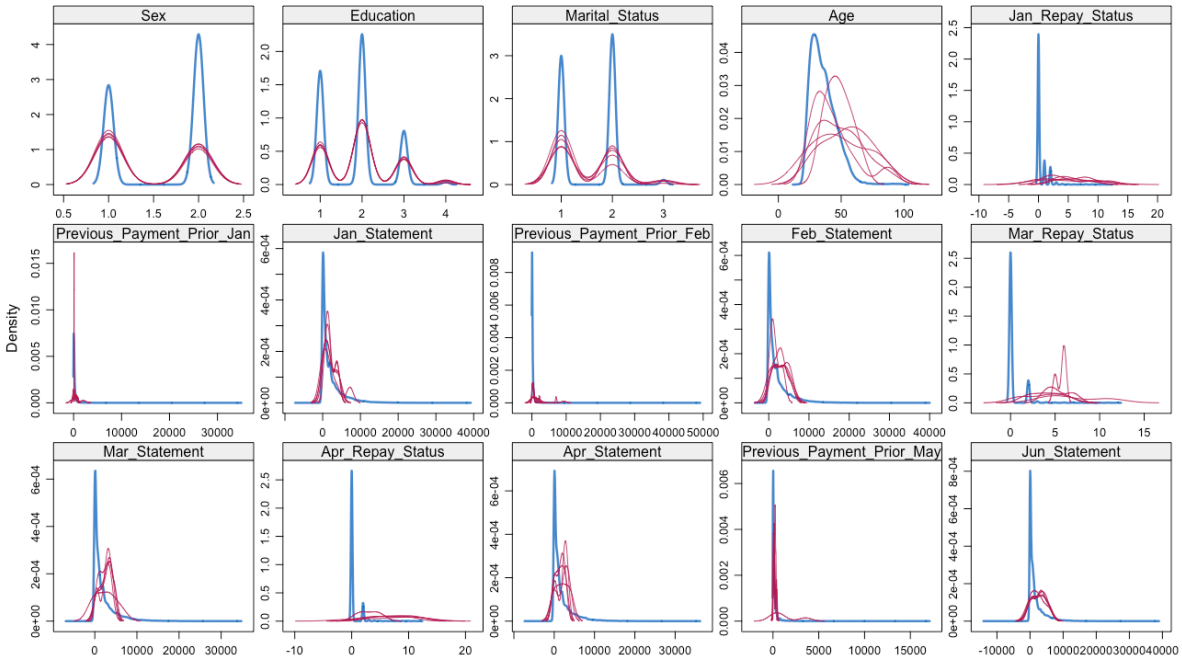
The dataset after cleansing procedure is incomplete and inconvenient to process, therefore, we made further adjustment on the dataset.

3.1.1 Adjusting Dataset Procedure

* Replace negative values (-1, -2, -3) in columns: Jan_Repay_Status, Feb_Repay_Status, Mar_Repay_Status, Apr_Repay_Status, May_Repay_Status, Jun_Repay_Status by 0. The following plot shows some of the variables with highest number of NA's. The y-axis of histogram shows the percentage out of the total number of observations of that variable. Since all percentage values are less than 5%, we can safely use mice function to replace NA's with imputed values. The pattern plot shows the patterns of NA's across all variables.



* Use mice to replace NA values. The following graph is the visualization result of mice transformation. The densities of the imputed data are shown in red, and the densities of the original data are shown in blue. From the plot, we can tell that they have similar density, which means the imputed values are reasonable and doesn't mess around the original data.

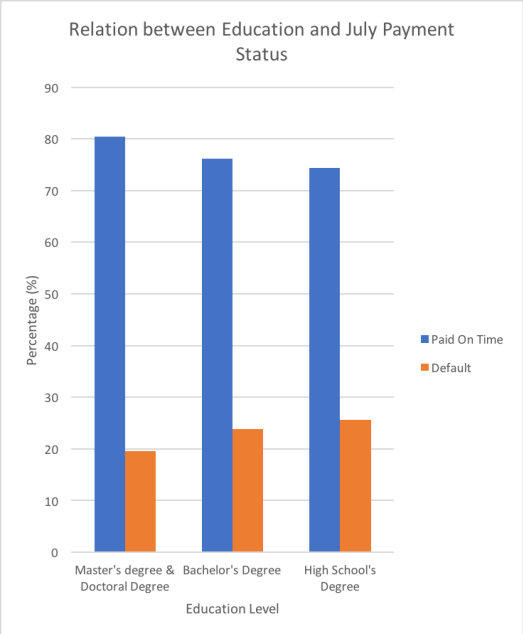


3.2 Excel Table Results

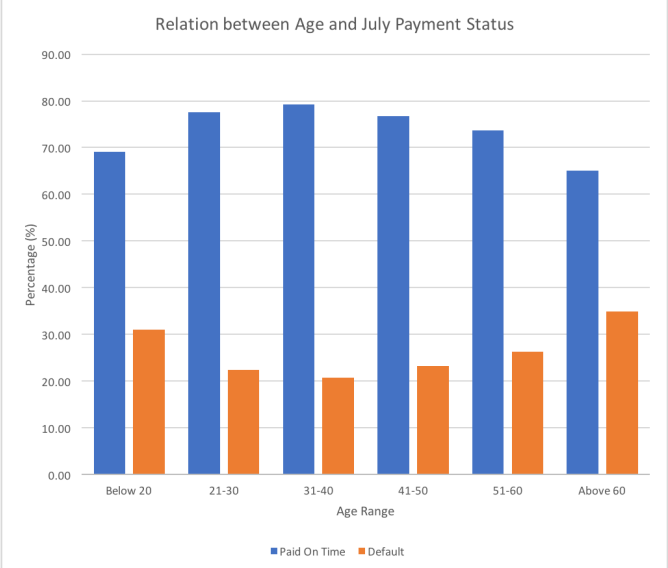
According to the adjusted dataset, we can derive following graphs/tables.

	Frequencies (n)		Percentage (%)	
July Payment Status	0 (Paid on time)	1 (Default)	0 (Paid on time)	1 (Default)
Sex				
Male	9135	2967	75.48	24.52
Female	14466	3829	79.07	20.93
Age				
Below 20	107	48	69.03	30.97
21-30	8359	2420	77.55	22.45
31-40	8715	2275	79.30	20.70
41-50	4630	1402	76.76	23.24
51-60	1563	559	73.66	26.34
Above 60	321	172	65.11	34.89
Marital Status				
Single	12746	3422	78.83	21.17
Married	10569	3275	76.34	23.66
Education				
Master's degree & Doctoral D	8615	2088	80.49	19.51
Bachelor's Degree	10782	3372	76.18	23.82
High School's Degree	3760	1298	74.34	25.66
Others	178	51	77.73	22.27

[1]Fig.1



[2]Fig.2

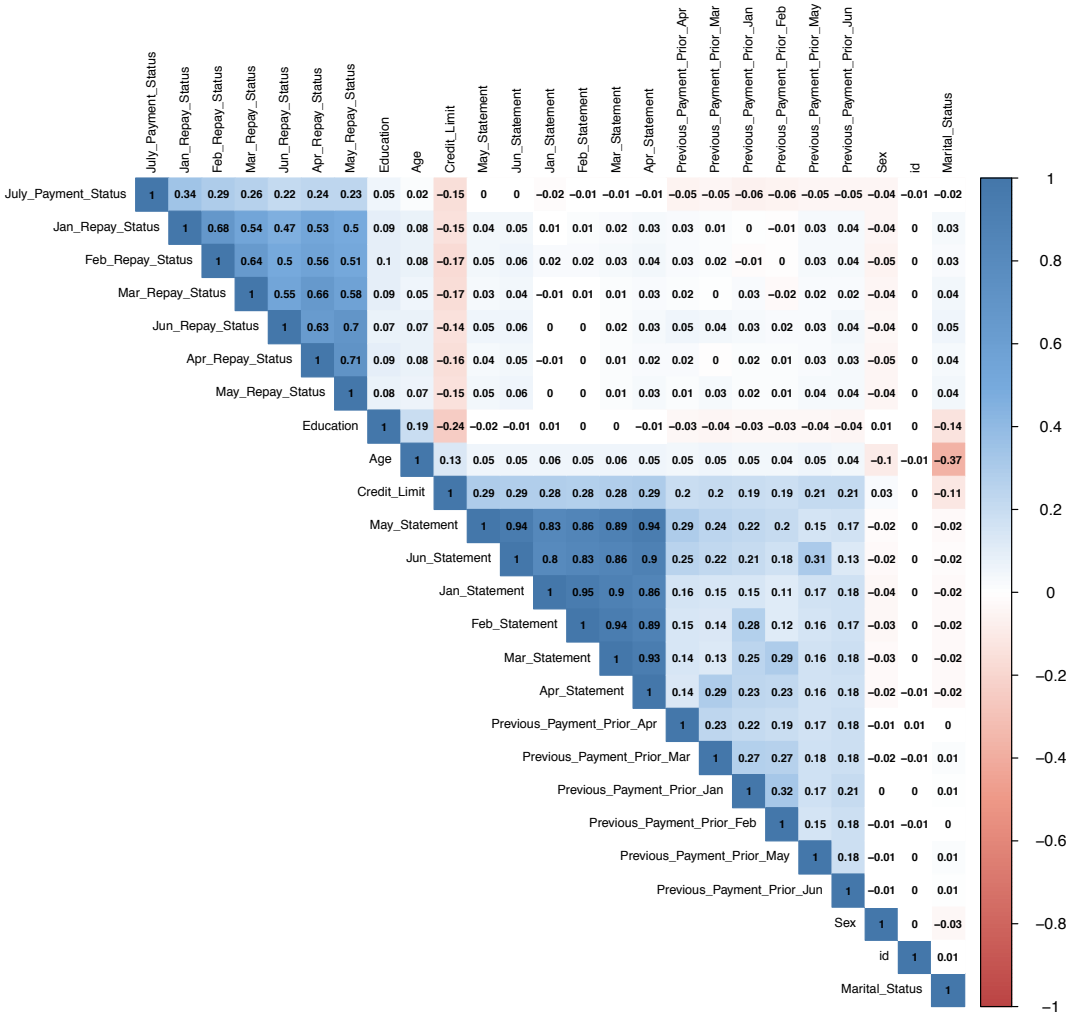


[3]Fig.3

[1] Table from excel.
[2] Relation between Education and July_Payment_Status.
[3] Relation between Education and July_Payment_Status.

3.2 Pairwise Correlation Plot

There are 24 variables in total. To observe the correlations between each variable we plotted a pairwise correlation graph using R. The blue color in the graph represent the high correlation coefficient between two variables and the white and orange color represent the relatively low correlation coefficient.

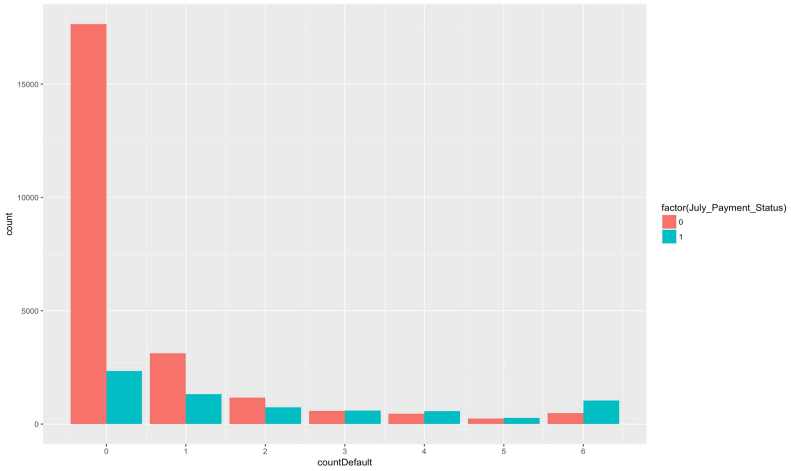


[4]Fig.4

[4] Pairwise Correlation plot.

3.3 Histogram

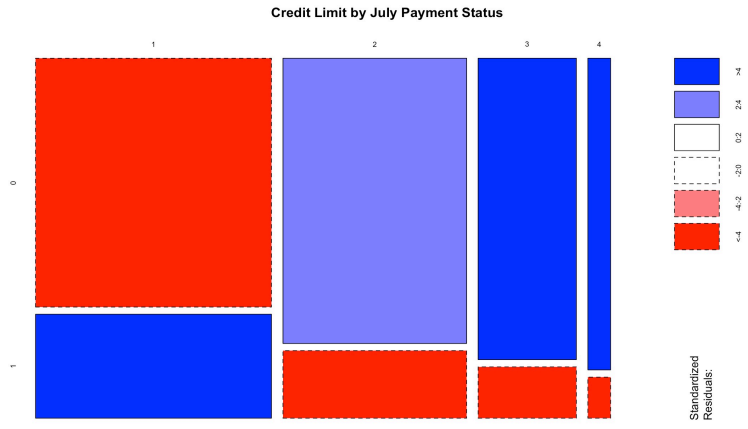
The histogram below shows the number of people who has never defaulted, has been defaulted once, twice, and up to six times during the past 6 months. While the red bar represents people whose July Payment Status is paid on time and the green bar indicates people who defaulted in July Payment Status. For example, 3 represent people who has 3 default records in the past 6 months.



[5]Fig.5

3.4 Mosaic Plot

We merge credits limits into 4 groups with 1 representing low credit limits, 4 representing highest group of credit limits. And the following mosaic plot shows higher the credit limit, higher proportion will default on July statement. The color of the rectangle represents the standard deviation of each type and the size of the rectangle for each column represents the proportion of that group of credit limits.



[6]Fig.6

[5] Histogram.

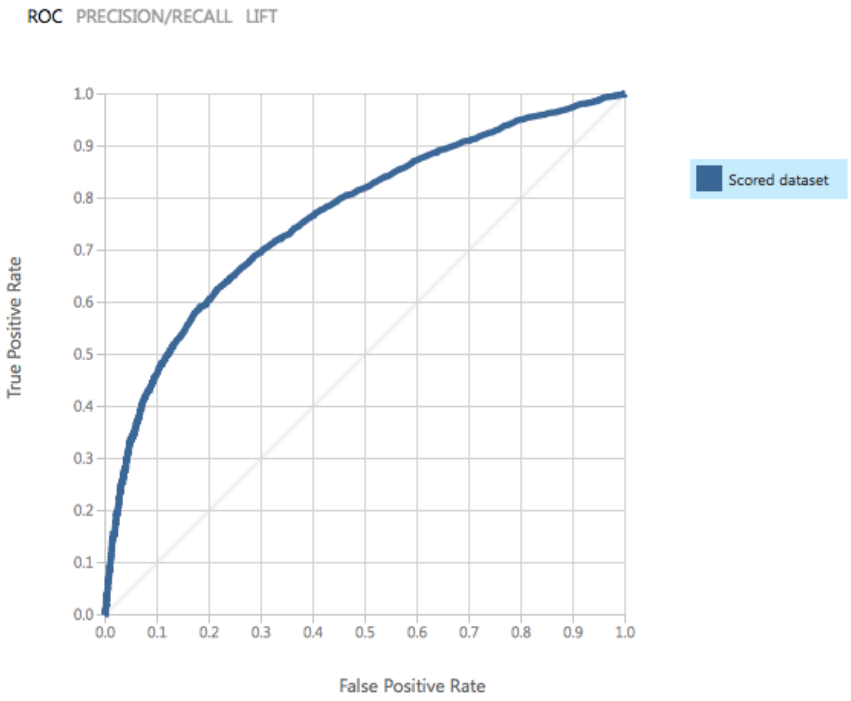
[6] Pairwise Correlation plot.

4. Prediction

After data analysis, we use several models to predict the probability of clients who may pay on time or default in July. In general, we use 4 models to predict the probability of default: logistic regression model, random forest model, two-class neural network model, PCA-anomaly detection model. The PCA-anomaly detection model has a better performance in predicting default while the other three has a good performance in predicting people who paid on time.

4.1 Logistic Regression Model

We use logistic regression model to predict the response variable July Payment Status. Multiple logistic regression suggested that Sex, Marital Status, Credit Limit, Previous Payment Prior Jan, Previous Payment Prior Feb, Payment Prior Apr, Previous Payment Prior May, Previous Payment Prior Jun, Jan Repay Status, Feb Repay Status, Mar Repay Status, Jun Repay Status, Jan Statement, Mar Statement had the biggest influence on the probability of paid on time, and the logistic regression equation could be used to predict the probability of paid on time. This model is good for predicting people who paid on time according to the azure output, and the accuracy is 80.7% with 0.5 threshold and AUC (area under the curve) 0.765.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
622	1139	0.807	0.650	0.5	0.765

False Positive	True Negative	Recall	F1 Score
335	5549	0.353	0.458

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	66	20	0.011	0.776	0.071	0.767	0.037	0.776	0.997	0.000
(0.800,0.900]	168	56	0.041	0.790	0.226	0.755	0.133	0.792	0.987	0.001
(0.700,0.800]	141	76	0.069	0.799	0.328	0.712	0.213	0.805	0.974	0.003
(0.600,0.700]	126	80	0.096	0.805	0.402	0.683	0.284	0.818	0.961	0.007
(0.500,0.600]	121	103	0.125	0.807	0.458	0.650	0.353	0.830	0.943	0.012
(0.400,0.500]	136	157	0.164	0.804	0.503	0.606	0.430	0.843	0.916	0.023
(0.300,0.400]	179	350	0.233	0.782	0.529	0.527	0.532	0.860	0.857	0.052
(0.200,0.300]	251	765	0.366	0.715	0.522	0.425	0.675	0.882	0.727	0.131
(0.100,0.200]	366	2046	0.681	0.495	0.446	0.298	0.882	0.915	0.379	0.406
(0.000,0.100]	207	2231	1.000	0.230	0.374	0.230	1.000	1.000	0.000	0.765

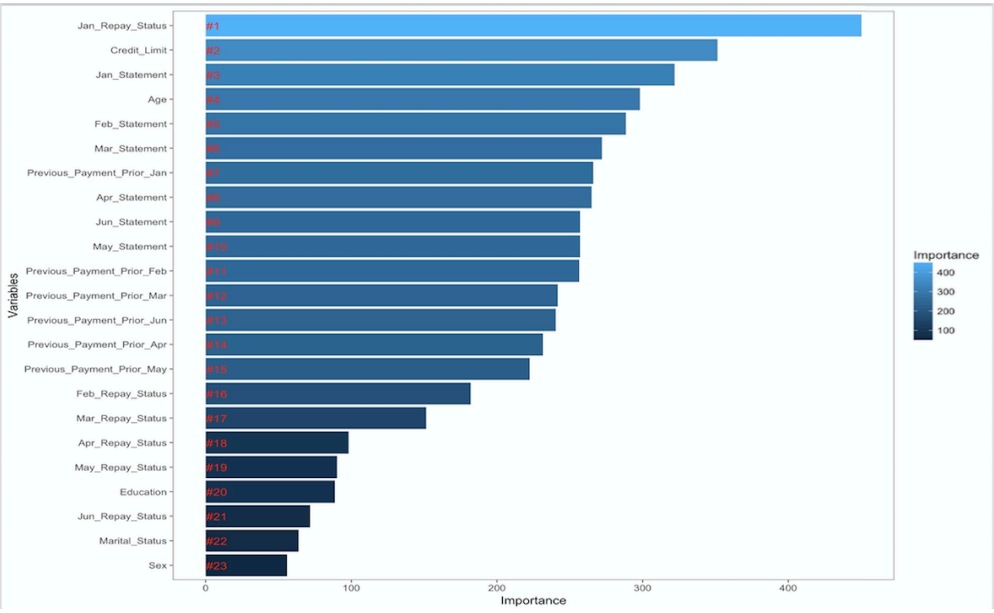
```
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -7.832e-01  1.366e-01  -5.734 9.82e-09 ***
## Sex                          -1.334e-01  3.575e-02  -3.732 0.000190 ***
## Marital_Status               -1.889e-01  3.701e-02  -5.106 3.30e-07 ***
## Age                           1.402e-03  2.018e-03   0.695 0.487264
## Education                    -5.377e-02  2.640e-02  -2.036 0.041709 *
## Credit_Limit                 -3.297e-05  4.382e-06  -7.523 5.34e-14 ***
## Previous_Payment_Prior_Jan   -4.639e-04  7.110e-05  -6.525 6.80e-11 ***
## Previous_Payment_Prior_Feb   -2.832e-04  5.817e-05  -4.869 1.12e-06 ***
## Previous_Payment_Prior_Mar   -1.112e-04  5.367e-05  -2.073 0.038213 *
## Previous_Payment_Prior_Apr   -3.131e-04  6.235e-05  -5.021 5.13e-07 ***
## Previous_Payment_Prior_May   -1.962e-04  5.992e-05  -3.274 0.001061 **
## Previous_Payment_Prior_Jun   -1.191e-04  4.121e-05  -2.891 0.003842 **
## Jan_Repay_Status              7.104e-01  2.656e-02  26.750  < 2e-16 ***
## Feb_Repay_Status              8.025e-02  2.642e-02   3.037 0.002388 **
## Mar_Repay_Status              1.280e-01  2.793e-02   4.582 4.60e-06 ***
## Apr_Repay_Status              3.628e-03  3.019e-02   0.120 0.904351
## May_Repay_Status              6.331e-02  3.297e-02   1.920 0.054874 .
## Jun_Repay_Status              1.006e-01  2.915e-02   3.452 0.000556 ***
## Jan_Statement                -1.023e-04  3.195e-05  -3.201 0.001372 **
## Feb_Statement                 4.350e-05  4.251e-05   1.023 0.306199
## Mar_Statement                 1.130e-04  3.790e-05   2.982 0.002862 **
## Apr_Statement                -4.122e-05  3.768e-05  -1.094 0.273979
## May_Statement                 7.669e-06  4.239e-05   0.181 0.856434
## Jun_Statement                 3.385e-05  3.518e-05   0.962 0.335971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23987  on 22409  degrees of freedom
## Residual deviance: 20667  on 22386  degrees of freedom
##   (525 observations deleted due to missingness)
## AIC: 20715
##
## Number of Fisher Scoring iterations: 5
```

However, this model is not good enough for predicting default and the AIC is high to some extent. We reduced number of variables in the model and generated several similar results in AIC and true positive rate, therefore, we move to the other model seeking for better solutions.

According to the logistic regression model, the Monthly_Repay_Status is positively related to the July Payment default rate, i.e. a client who has a higher Monthly_Repay_Status rate also has larger probability of defaulting and vice versa. Meanwhile, the Previous_Payment_Prior_Month term is negatively related to the July default prediction, i.e. as the Previous_Payment_Prior_Month increases, the client has a higher possibility of not defaulting. Therefore, the bank can focus on the previous monthly record on repay status and previous payment, especially on the values and sign of both records such that the bank can select the "preferred client" to put forward preferential policy and gain more interest. On the other hand, the bank can select the "not so preferred client" and put some specific punishment policies to reduce the loss of revenue due to default.

4.2 Random Forest Model

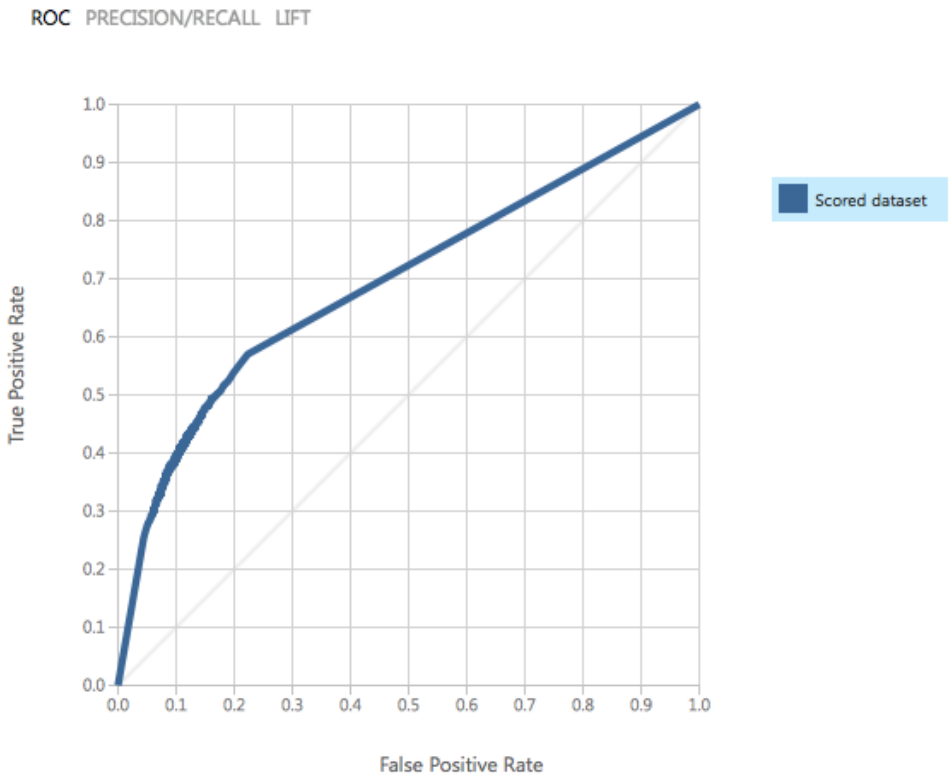
In the random forest model, the following plot represents the significance level of each variables. In general, we focus on the variables that have the level of importance greater than 200. The important variables are Monthly statements, January repay status, monthly previous payment prior, credit limit and age. Therefore, in the random forest model, we expand the random tree specified on the important variables that were mentioned above. The evaluation of this model is good. The accuracy of the model is 81.57%, the model has a good prediction on people who paid on time though a relatively inferior result on predicting default.



```
> cMatrix = confusionMatrix(prediction, testData$payment_Status)
> cMatrix
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5634 1085
         1  324  603

               Accuracy : 0.8157
                 95% CI : (0.8068, 0.8244)
    No Information Rate : 0.7792
    P-Value [Acc > NIR] : 2.23e-15

                  Kappa : 0.3612
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9456
            Specificity : 0.3572
         Pos Pred Value : 0.8385
         Neg Pred Value : 0.6505
             Prevalence : 0.7792
         Detection Rate : 0.7369
   Detection Prevalence : 0.8788
      Balanced Accuracy : 0.6514

       'Positive' Class : 0
```

4.3 Neural Network Model

The neural network model is a convenient model for regression in machine learning. The following pictures shows the results of the model. The model's accuracy is 78.3% and with a good prediction on people who paid on time (90.04%). AUC of this model is 0.701, compared to logistic regression model's 0.765, this model is slightly worse than the previous one. Furthermore, this model does not have a higher true positive rate means this model is still not good for predicting default. Therefore, we move to the other model seeking for a better solution of default prediction.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
690	1071	0.783	0.541	0.5	0.701

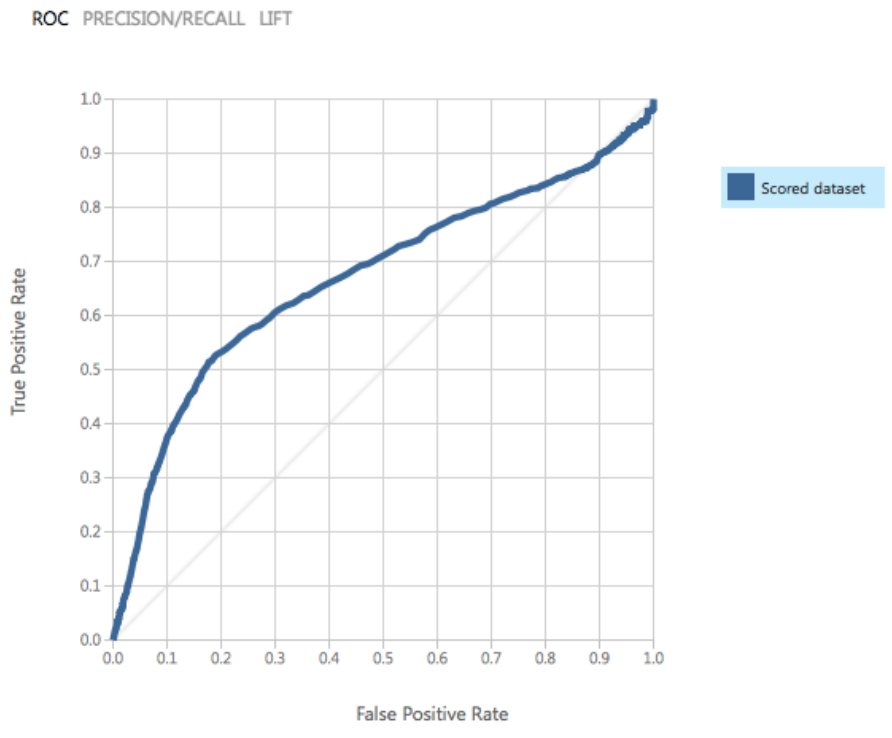
False Positive	True Negative	Recall	F1 Score
586	5298	0.392	0.454

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	593	438	0.135	0.790	0.425	0.575	0.337	0.823	0.926	0.011
(0.800,0.900]	39	47	0.146	0.789	0.439	0.566	0.359	0.827	0.918	0.014
(0.700,0.800]	24	36	0.154	0.787	0.447	0.557	0.373	0.829	0.911	0.016
(0.600,0.700]	16	36	0.161	0.785	0.449	0.547	0.382	0.830	0.905	0.018
(0.500,0.600]	18	29	0.167	0.783	0.454	0.541	0.392	0.832	0.900	0.020
(0.400,0.500]	20	36	0.174	0.781	0.459	0.533	0.403	0.834	0.894	0.023
(0.300,0.400]	19	36	0.181	0.779	0.463	0.526	0.414	0.835	0.888	0.025
(0.200,0.300]	25	45	0.191	0.776	0.469	0.518	0.428	0.837	0.881	0.028
(0.100,0.200]	28	65	0.203	0.771	0.472	0.505	0.444	0.839	0.869	0.033
(0.000,0.100]	979	5116	1.000	0.230	0.374	0.230	1.000	1.000	0.000	0.701

4.4 PCA-Based Anomaly Detection Model

PCA-Based Anomaly Detection method is a common tool used in initializing model, especially in detecting anomaly. The following results shows that the accuracy is only 22.1% with a low AUC of 0.669. It seems like a bad model at the beginning, however, if we specify the true positive rate 99.41%, it is the highest among all the models above. This model is not good for predicting people who paid on time but it is a good model for predicting default.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1686	10	0.221	0.221	0.5	0.669

False Positive	True Negative	Recall	F1 Score
5949	0	0.994	0.361

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	496	435	0.122	0.786	0.378	0.533	0.292	0.821	0.927	0.011
(0.800,0.900]	833	3403	0.676	0.450	0.387	0.257	0.784	0.852	0.355	0.363
(0.700,0.800]	237	1734	0.934	0.254	0.355	0.219	0.923	0.744	0.063	0.609
(0.600,0.700]	46	189	0.964	0.235	0.355	0.219	0.950	0.691	0.032	0.639
(0.500,0.600]	74	188	0.999	0.221	0.361	0.221	0.994	0.000	0.000	0.669
(0.400,0.500]	10	0	1.000	0.222	0.363	0.222	1.000	1.000	0.000	0.669
(0.300,0.400]	0	0	1.000	0.222	0.363	0.222	1.000	1.000	0.000	0.669
(0.200,0.300]	0	0	1.000	0.222	0.363	0.222	1.000	1.000	0.000	0.669
(0.100,0.200]	0	0	1.000	0.222	0.363	0.222	1.000	1.000	0.000	0.669
(0.000,0.100]	0	0	1.000	0.222	0.363	0.222	1.000	1.000	0.000	0.669

5. Conclusion

Based on our analysis of the dataset, credit limit and whether the clients paid their statements on time in the previous six consecutive months are the most prominent factors that bring about the default behaviour, which must be taken into consideration in the new decision-making mechanism. After fitting four different models to the dataset, we found PCA-Based Anomaly Detection Model is good at predicting default behaviour. In contrast, logistic regression, neural network, and random decision forest models are relatively good at classifying people who will

pay their statements on time.

Model comparison: in the above 4 models, the logistic regression model, neural network model and random forest model have similar accuracy on prediction and they can predict the “bank preferred client” well though the prediction on default is inferior to the PCA-Based anomaly detection model. Among the three models, the logistic regression model and random forest model have a slightly better accuracy and true negative rate to the neural network model. The PCA-Based anomaly detection model has only 22.1% accuracy but with a relatively high true positive rate which is convenient for predicting default but not a good model in total due to the lack of accuracy on true negative response prediction.