Credit Card Default Model

Name: Wenying Wu, Declan Stockdale
Date: 16/05/2021

INTRODUCTION

The report details the process we undertook in competing in a Kaggle competition as part of the machine learning algorithms and applications subject. The objective of the competition was to predict which customers of a bank would default on their credit card within the next month.

This report has been prepared following the CRISP-DM process which is broken into the following sections: business understanding, data understanding, data preparation, modelling and evaluation and conclusion.

Both members contributed equally to this report. Each member pursued the modelling process independently and results were discussed over email and WhatsApp. Sections of the report were initially written by each member individually and as the process came to its conclusion, further revisions were made by both members. This was done over a shared online document. A more detailed breakdown is available in Appendix 1.

BUSINESS UNDERSTANDING

The competition aims to gain insight into the important factors that determine whether a customer will default on their credit card payment during the next month. This is quite important to banks as it determines if the bank will suffer a financial loss for a particular client. It may also be useful to identify if a customer needs additional financial support which may lead to positive outcomes for both the bank and the customer.

Various possible predictor variables have been collected and computed for both defaulters and non-defaulters. We want to determine which of these variables are useful for prediction. The main aims are to find out whether there is a relationship between the risk of default and other factors and generate a model to predict the default risk of a customer. The analysis and model building will be undertaken in the statistical language 'R'. The chosen language is widely used for this purpose and no additional resources are necessary.

There are numerous ethical concerns associated with the application of machine learning to predict credit cards default. Various state of the art algorithms focus purely on predictive power and not on interpretability. If such an algorithm is implemented, then the transparency of the decision-making process is significantly harder to identify and explain. Financial advisors employed at a bank may struggle to communicate to a customer why they designated as a defaulter and this may lead to frustration and the loss of the customer. This may change as more tools are built to improve interpretability but, in the meantime, it would raise concerns.

Most algorithms improve with access to more data which may involve the collection of additional personal data to improve model performance. This may lead to discrimination due to age, gender and education for our supplied data set but could extend to race, area of residence and ethnicity with additional data collection which would be sought after to potentially improve performance.

In terms of creating value for the bank and the shareholders, concerns arise when non-defaulters are classified as defaulters and defaulters are classified as non-defaulters. Misclassification of the first type would lead to unfair treatment towards the customer and their access to further credit may be restricted. Misclassification of the second kind would lead to a loss of revenue. Both misclassifications need to be considered when assessing the performance of the algorithms.

2. DATA UNDERSTANDING

The data has been supplied as part of the Kaggle competition "[UTS MLAA] Credit Card Default Model Default payments of credit card clients". It's based on a previous Kaggle competition using similar data collected from a Taiwanese bank. The supplied data has been separated into a train set containing data from 23101 customers with a default predictor included. A test set containing information from a remaining 6899 customer with no default predictor has also been included and will be used to assess the predictive power of our model.

The data dictionary for the train and test set are detailed below: (Source from MLAA AT2 Brief (Parts A and B))

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_X: Repayment status for the past X months. -1 = paid on time; 1 = payment delay for one month; 2 = payment delay for two months etc
- BILL_AMTX: Bill amount for past X months
- PAY_AMTX: Payment amount for last X months
- default: Default payment next month (1=yes, 0=no)

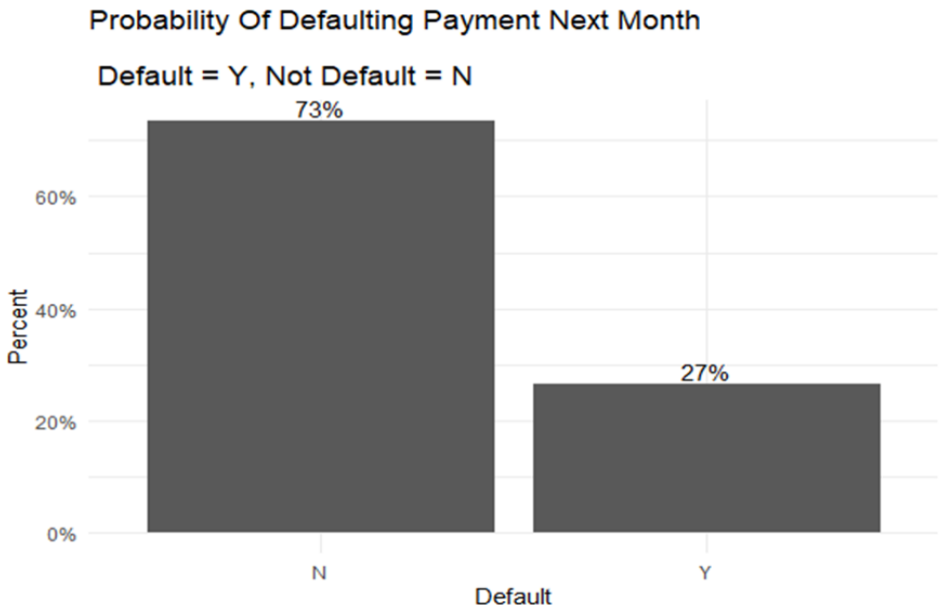
Exploring the data

Because the train and test sets are already supplied, and the test set doesn't include the default variable (dependent variable to be predicted). The data exploration and analysis are performed solely on the train set.

The first step taken was to check for any missing values. From the output below we found that there are no missing values.

```
      ID LIMIT_BAL       SEX EDUCATION  MARRIAGE       AGE     PAY_0     PAY_2     PAY_3     PAY_4
       0         0         0         0         0         0         0         0         0         0
   PAY_5     PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6  PAY_AMT1  PAY_AMT2
       0         0         0         0         0         0         0         0         0         0
PAY_AMT3  PAY_AMT4  PAY_AMT5  PAY_AMT6   default      AGE1
       0         0         0         0         0         0
```

The train set consists of 73% non-defaulters and only 27% defaulters. This indicates the dataset is highly imbalanced, however, it is common in these datasets because most people make credit card repayments on time.

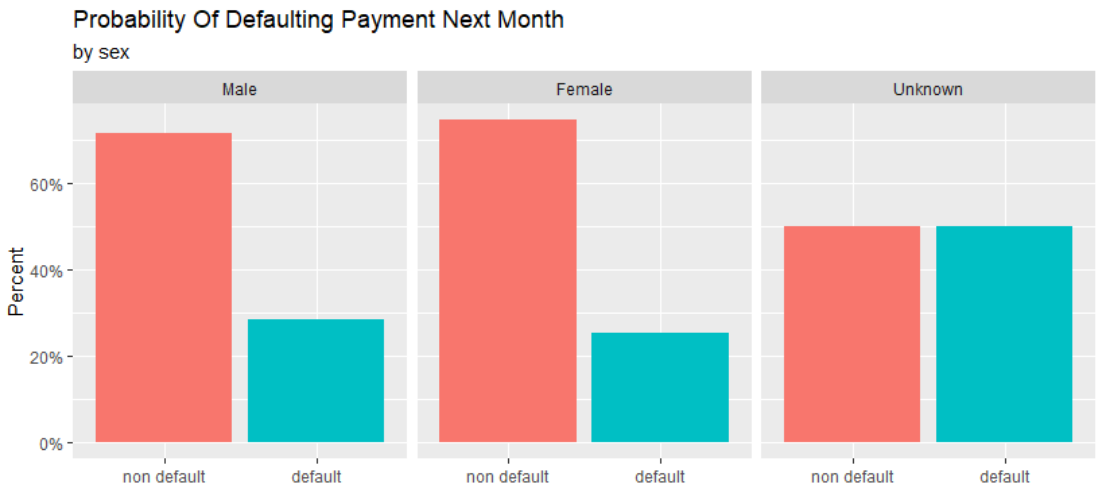


Data exploration detail for each independent variable are in Appendix 2 and the summary is in the below table.

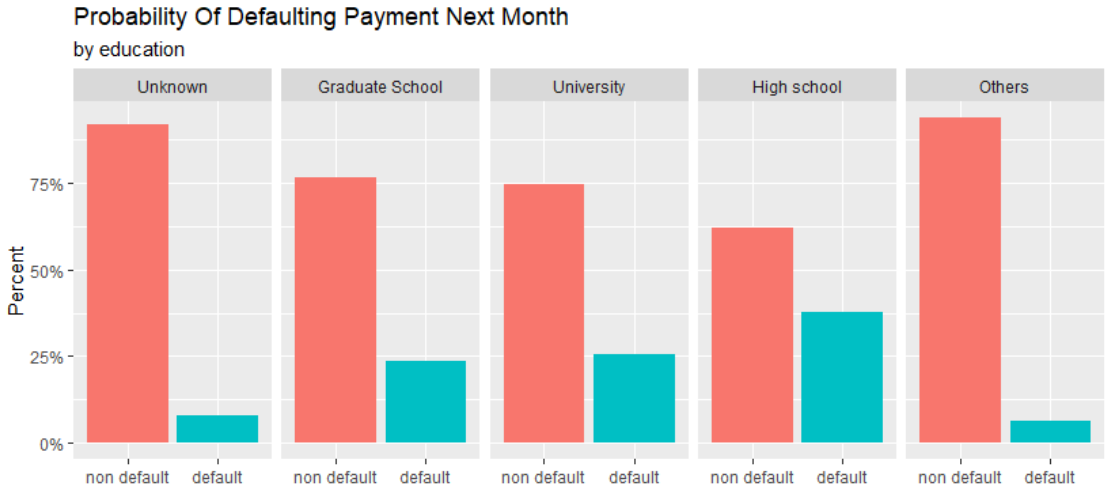
No.	Data Exploration Finding
1	There are more female than male in this data set.
2	The majority of the clients are with education level of University or Graduate School.
3	The number of single clients is higher than the one of married clients.
4	Most of the clients are age from 20 to 60.
5	Most credit limits are in the range of 0-$500,000.
6	There is a large amount of undocumented data (-2 and 0) input in the PAY_X variable.
7	Invalid input (-99) exists in BILL_AMTX; Most of BILL_AMTX are in the range of 0-$360,000; Default clients generally have less bill amount.

EDA – Relationship between Default and other variables

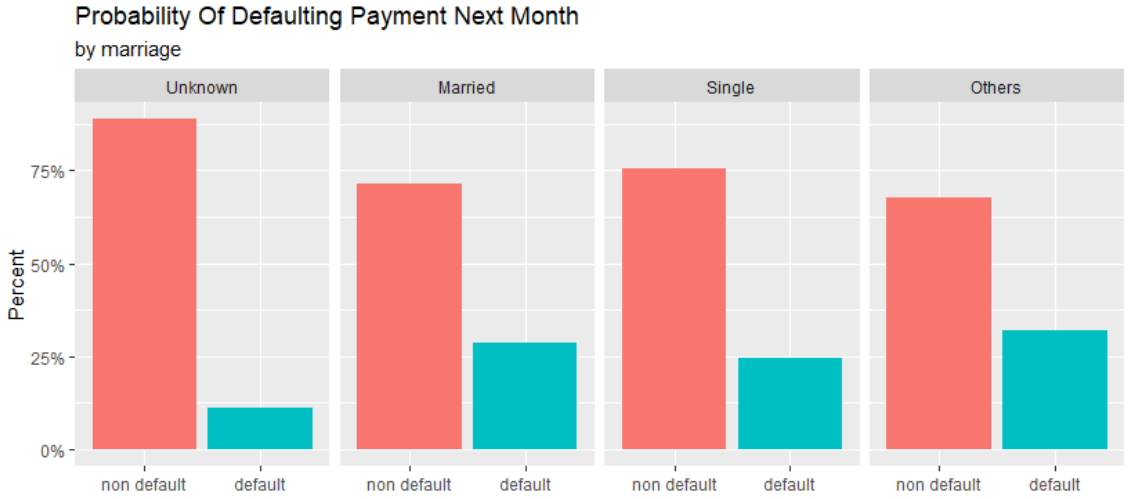
Female clients have a lower chance to default in general.



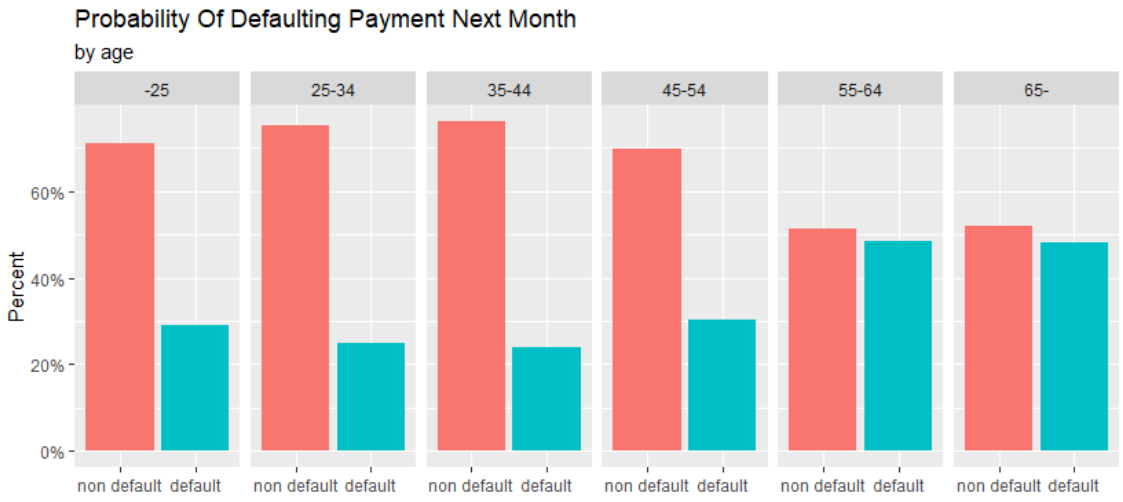
Higher education level generally has a lower chance to default.



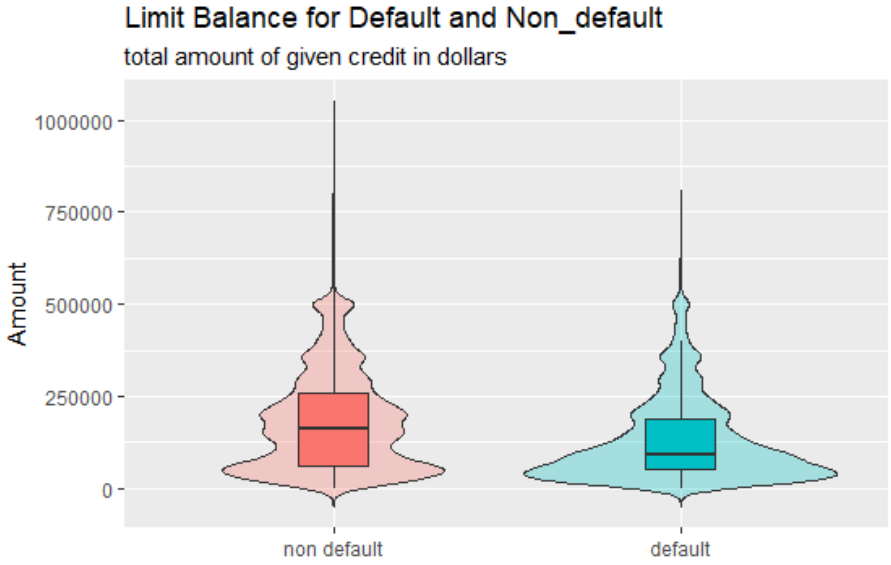
Married and Other clients are more likely to default in general.



Looking at the age distribution, the age groups that have the biggest chance to default are 55-64, more than 65 and under 25 years old.



Clients with lower credit limit are more likely to default. More analysis of limit balance is in Appendix 2 Part 3.



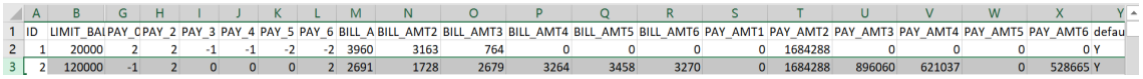
3. DATA PREPARATION

This section details the various procedures and rationale behind features that were engineered from the supplied data.

Variable Selection

PAY_AMTX

As abovementioned, PAY_AMTX is either 0 or another number across PAY_AMT1 to PAY_AMT6, and the numbers are far exceeding BILL_AMTX and even the LIMIT_BAL. This seems illogical for all clients in this dataset to have such a large amount of repayment on each month if it is not 0. Below is an example from the original train set.



	A	B	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	ID	LIMIT_BAL	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_A	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	defau
2	1	20000	2	2	-1	-1	-2	-2	3960	3163	764	0	0	0	0	1684288	0	0	0	0	Y
3	2	120000	-1	2	0	0	0	2	2691	1728	2679	3264	3458	3270	0	1684288	896060	621037	0	528665	Y

For clearer presentation, data of Client 2 (ID) with a credit limit of $120,000 is summarised in below table.

Client 2	1	2	3	4	5	6	SUM
PAY_X	-1	2	0	0	0	2	
BILL_AMTX	$2,691	$1,728	$2,679	$3,264	$3,458	$3,270	$17,090
PAY_AMTX	0	$1,684,288	$896,060	$621,037	0	$528,665	$3,730,050

The PAY_AMT's are abnormally high value, and the value for each row (client) is the same across the dataset, which means every client is paying, for example, $1,684,288 for the last 2 month. This seems unreasonable for real-world scenarios. Therefore, to eliminate any effect of these invalid data, we decided to remove all PAY_AMTX in both the train and test set in the modelling section. A model including PAY_AMTX variables will also be built to test the impact of PAY_AMTX.

0 and -2 in PAY_X

There are 74,011 values of 0 and 18,630 values of -2 (67% of total input together, see Appendix 2 Part4 b.) in the train set and these 2 values are not included in the data dictionary. Strictly speaking, these are invalid data input. There are a few standard methods to treat these data points. The first method would be to remove observations with these invalid inputs, but the loss is too big (3117 remaining from 23101). Second one is to delete variables, but we decide to explore the impact of these inputs before removing them. The third one is to replace invalid input by major class, but in this case, invalid input is already a major class. Hence, we decide to leave these input 'as is' which seems to be the most appropriate method.

Other undocumented data

There are some undocumented inputs in SEX variable (listed below) that are invalid, we decided to remove observations as there are only 4 in total.

SEX	integer [6] (S3: table)	9088 14009 1 1 1 1
1	integer [1]	9088
2	integer [1]	14009
2113	integer [1]	1
martian	integer [1]	1
orthodontist	integer [1]	1
RYDE	integer [1]	1

There are also 12 inputs of 0 in EDUCATION and 45 inputs of 0 in MARRIAGE. We decided to leave them 'as is' because the number is relatively larger than those in SEX

EDUCATION	integer [7] (S3: table)	12 8180 10761 3784 98 225 ...
0	integer [1]	12
MARRIAGE	integer [4] (S3: table)	45 10510 12291 255
0	integer [1]	45

Outliers Cut

According to the "rule of thumb" (Han, J., & Kamber, M., 2001), There are also some rows deleted because containing invalid input, namely -99 in LIMIT_BAL, and age larger than 122 (the longest human lifespan in Wikipedia). After removing outliers and invalid inputs, 22824 rows (98.8% of original 23101) remaining.

Evaluation Metric

In Kaggle competition, ROC-AUC is the chosen metric of model evaluation. However, we choose Recall (TP/(TP+FN)) as the evaluation metric to maximise the proportion of actual positive (default) identified correctly. The reason is that any bank will not want any defaults to not be identified to avoid any monetary loss.

4. MODELLING

Various models were built and refined for default prediction. Similar work has been performed in the previously mentioned Kaggle competition and the successful methods serve as a guide for our purpose. The list of models chosen for this work and their relevant metrics are listed below:

Model summary table

	Model name	AUC	Recall
1	Linear Logistic Model	0.7210	0.2533
2	Random Forest	0.7966	0.4488
3	XGBoost	0.8141	0.5046
4	GBM	0.8004	0.4260
5	Lasso	0.6551	0.0013

Model Evaluation

Model summary including AUC, Recall, confusion matrix (from the 'train set' we split from the train set), and feature importance plot are in Appendix 3. We can see from the above summary table that the XGBoost model has the highest AUC score as the XGBoost package in R has an embedded function to maximise AUC. In addition, XGBoost model performs the best in Recall (the evaluation metric we choose). Therefore, we recommend implementing the XGBoost model over the alternative models to ensure minimal financial losses to the bank.

Another interesting finding is that the AUC and Recall of XGBoost models with and without PAY_AMTX (invalid input as abovementioned) are identical, see below comparison. Though the reason is unknown, we still recommend not to include PAY_AMTX in the model.

5. EVALUATION AND CONCLUSION

Various machine learning algorithms have been applied to predict credit card default as part of a Kaggle competition. The performance metric used in the competition was the area under the receiver operator curve (AUC) where an XGBoost model scored the highest AUC of 0.795 outperforming a random forest model, generalized linear model, and a linear logistic model. Furthermore, due to the implications of classifying non-defaulters as defaulters (false positives) in a banking scenario, we have decided that recall is the most appropriate metric and would recommend a XGBoost model.

REFERENCE:

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.

APPENDIX

Appendix 1

Contributions by member

Declan Stockdale

- Worked through various algorithms including XGBoost, Random Forest and SVM
- Worked on a neural network that wasn't included due to poor performance and replicability
- Generated various graphs used in this report
- Wrote and revised major sections of the report namely the introduction, business understanding, data understanding and the conclusion

Wenying Wu

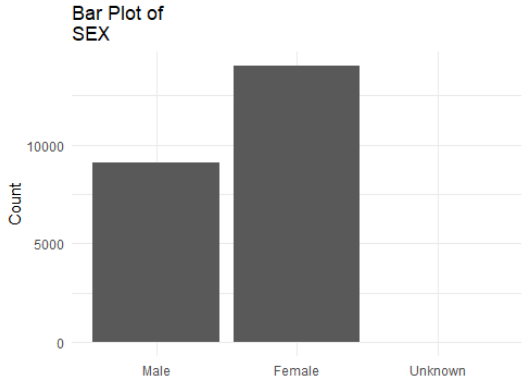
- Worked through various algorithms including XGBoost, GLM, GBM, Random Forest, Lasso and linearized logistic model
- Generated and improved the quality of various graphs used in this report
- Wrote and revised major sections of the report namely the introduction, business understanding, data understanding and model evaluation
- Responsible for most submissions for Kaggle challenge

Appendix 2

Part 1

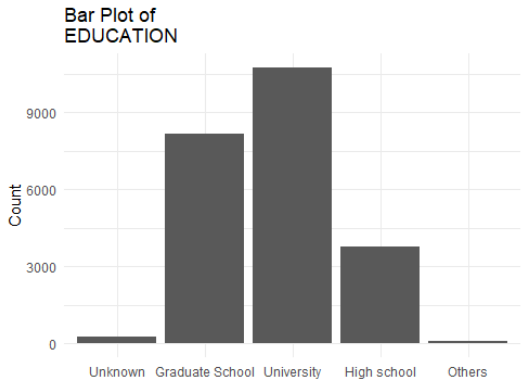
a. SEX

There are more female than male in this data set.



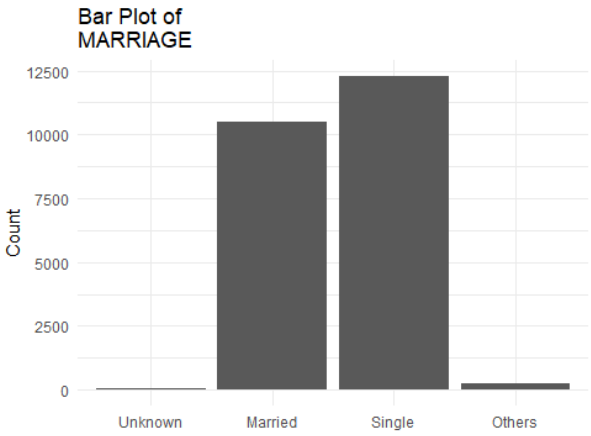
b. EDUCATION

The majority of customers/clients are with education level of University and Graduate School.



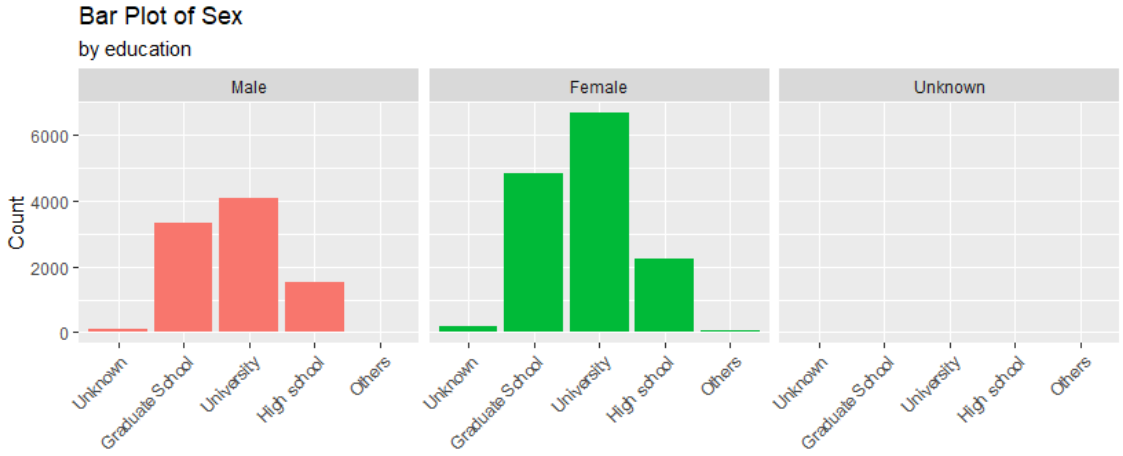
c. MARRIAGE

Single and Marriage status are the major groups, while the proportion of single clients is higher than married clients.

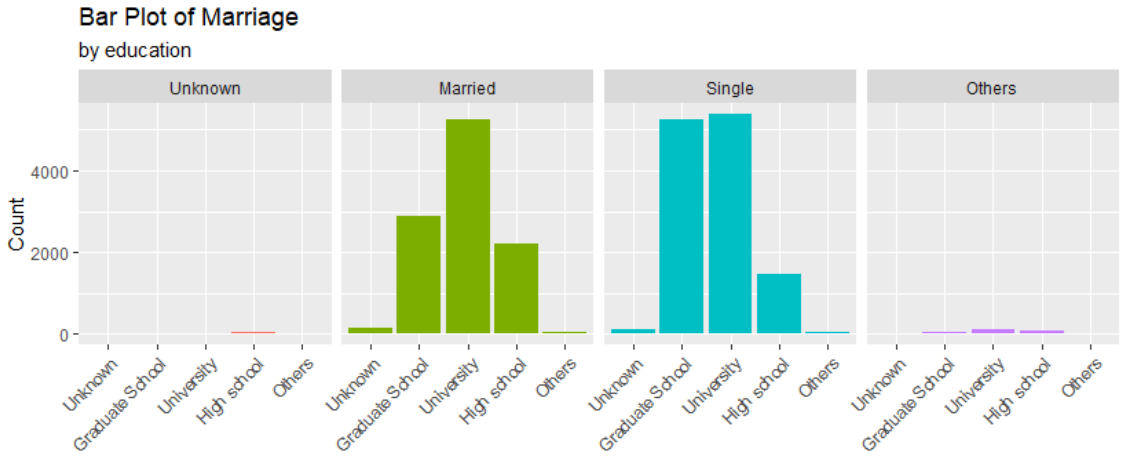


d. Combination Analysis

The distribution of Education for both male and female are similar. The highest proportion is the university, followed by graduate school, high school, unknown and others.

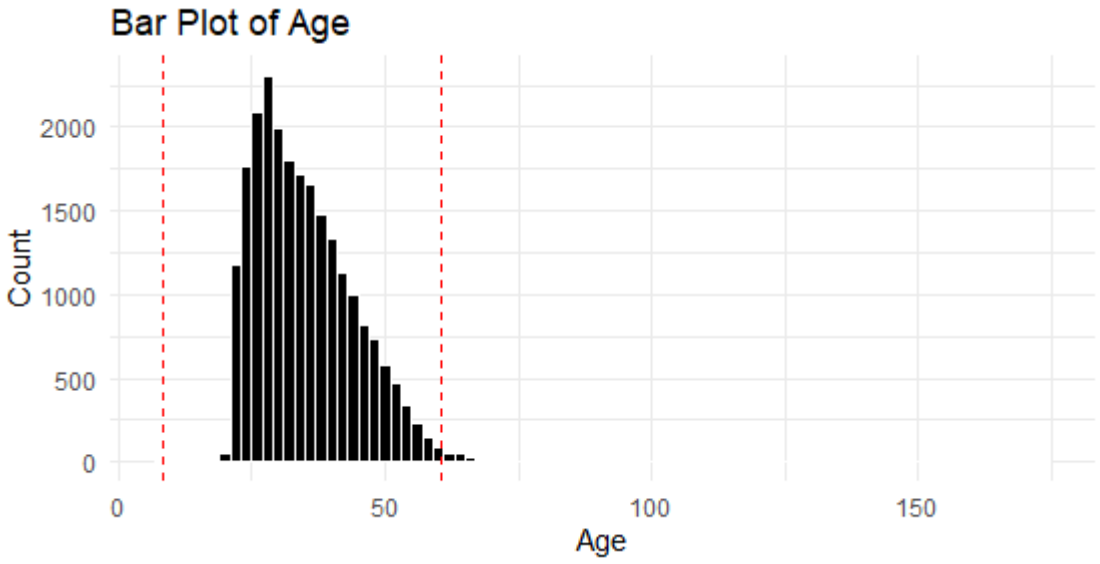


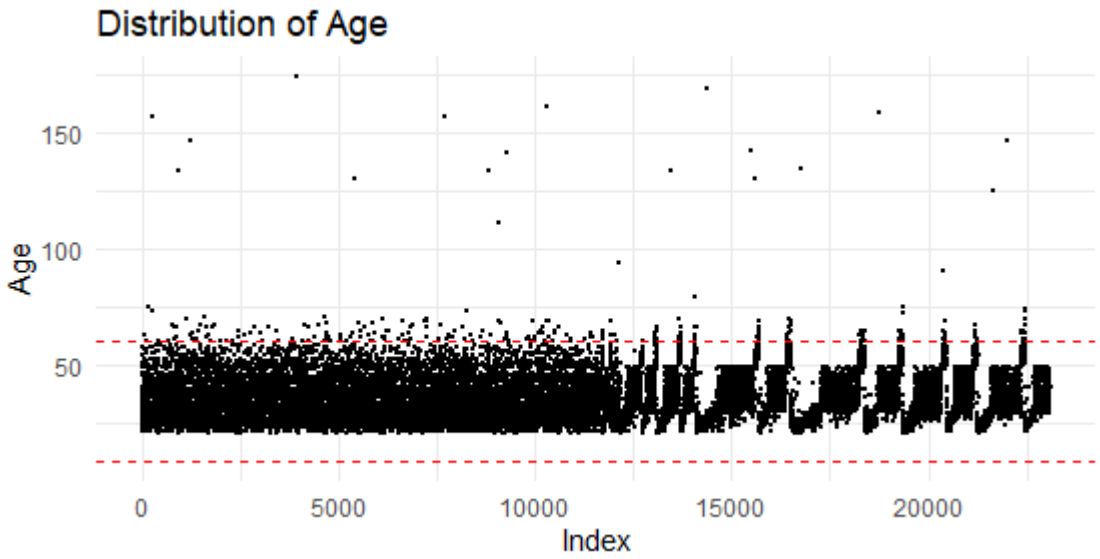
Education level distribution across all marriage status is similar, same as the distribution on genders (male and female).



Part 2 - AGE

The majority of the age is between 20 and 60. (Red lines are Q1-1.25IQR & Q3+1.25IQR in all graphs)

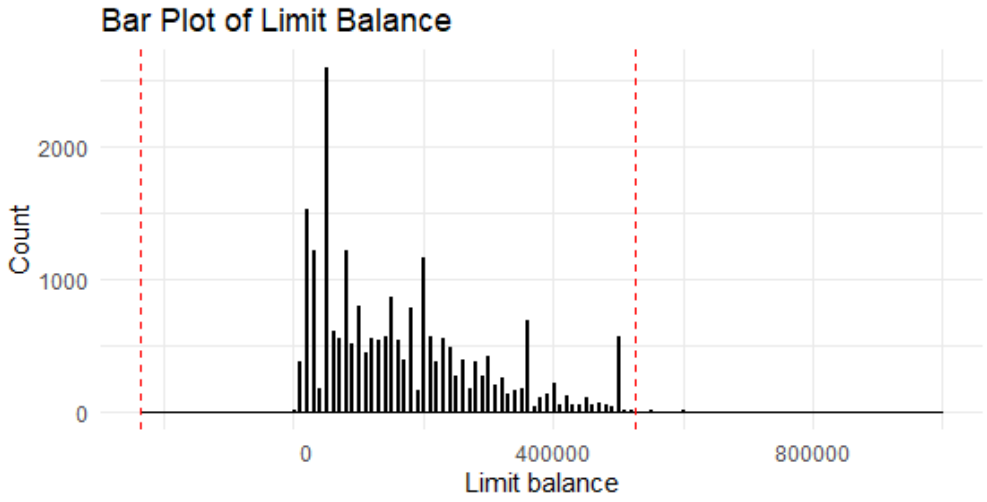


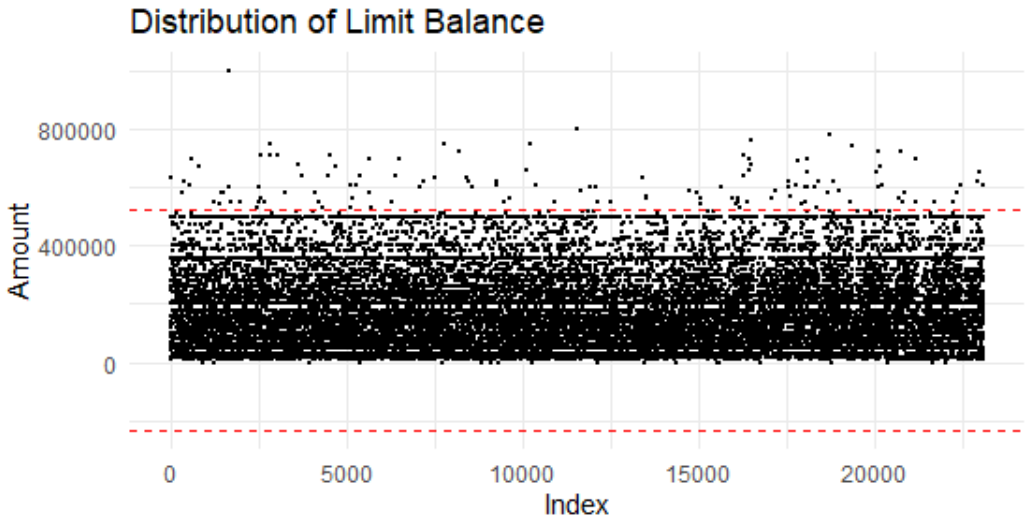


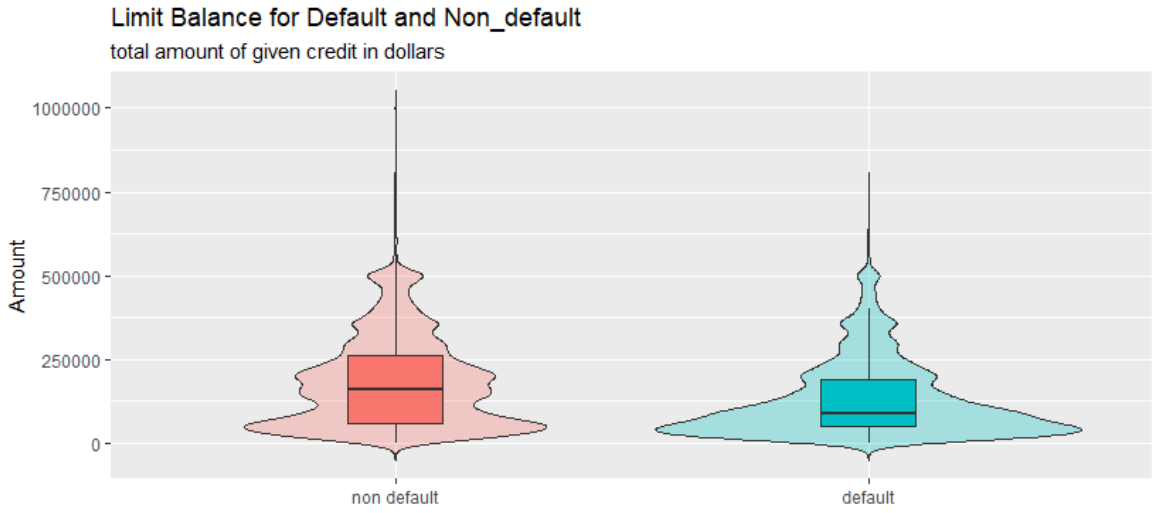
Part 3 - Limit Balance

Below graphs are Limit_bal analysis. They suggest that most of LIMIT_BAL fall between 0 – $500,000, across gender, educational level marriage status and age group.

a. Total

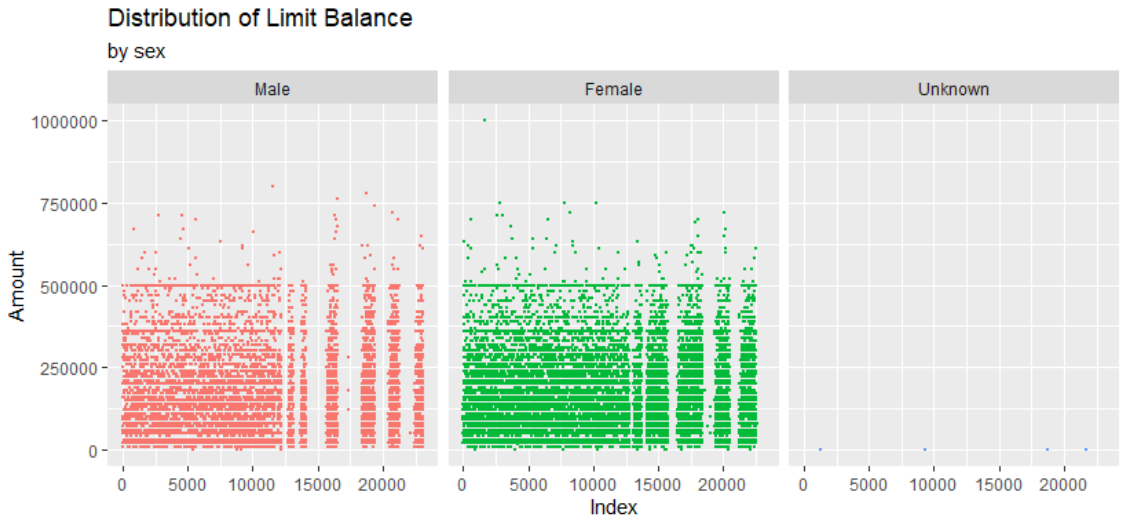


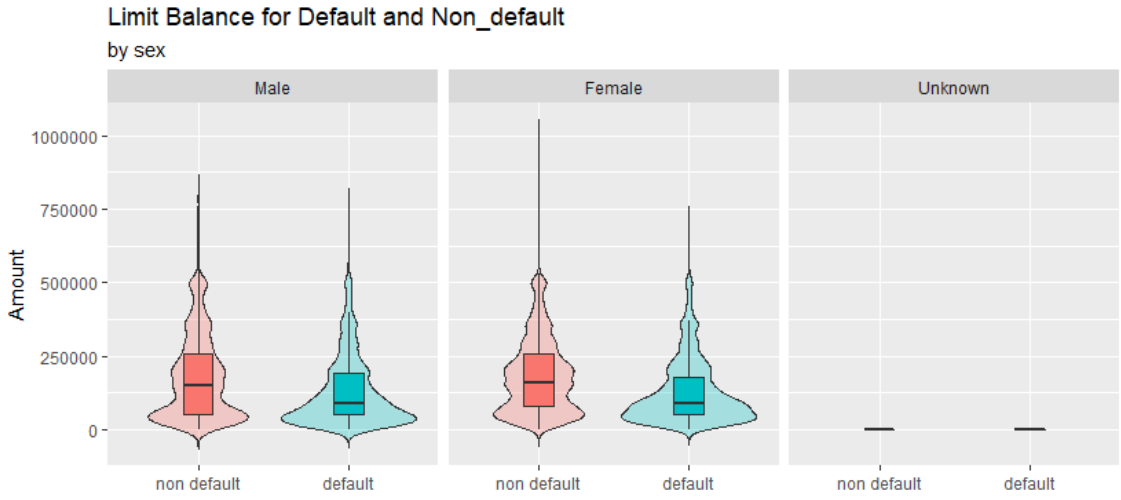




b. By age

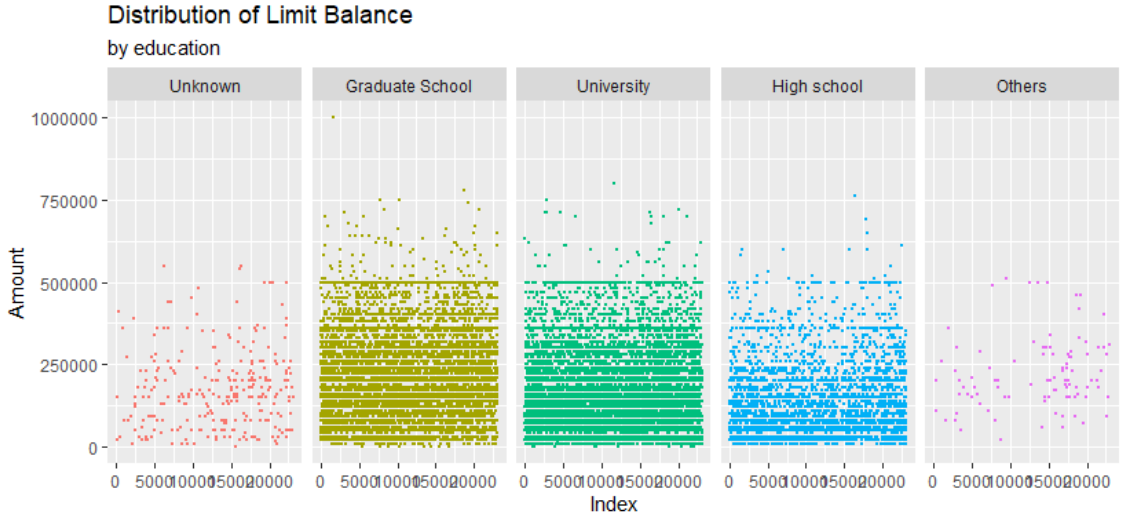
Female clients generally have a higher limit balance.

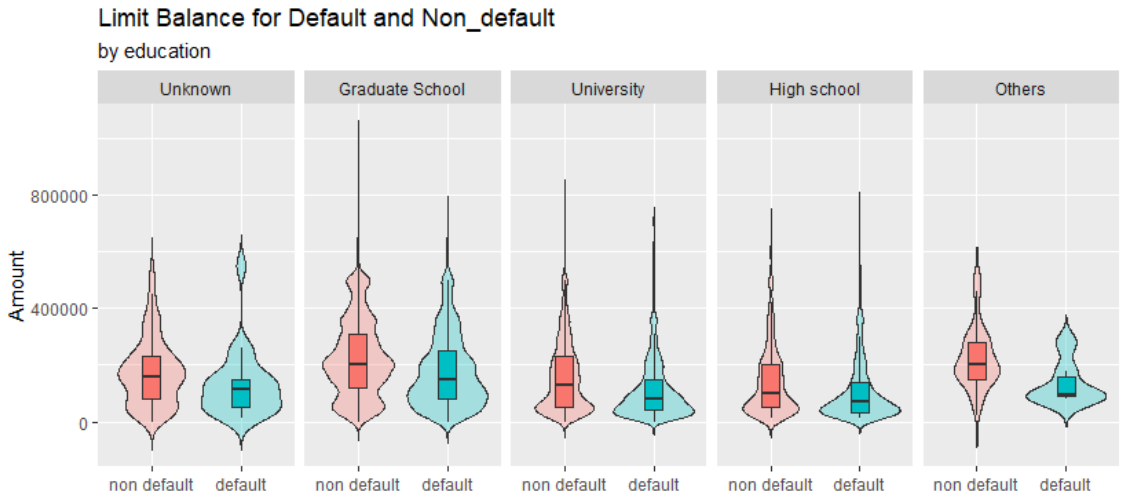




c. By education

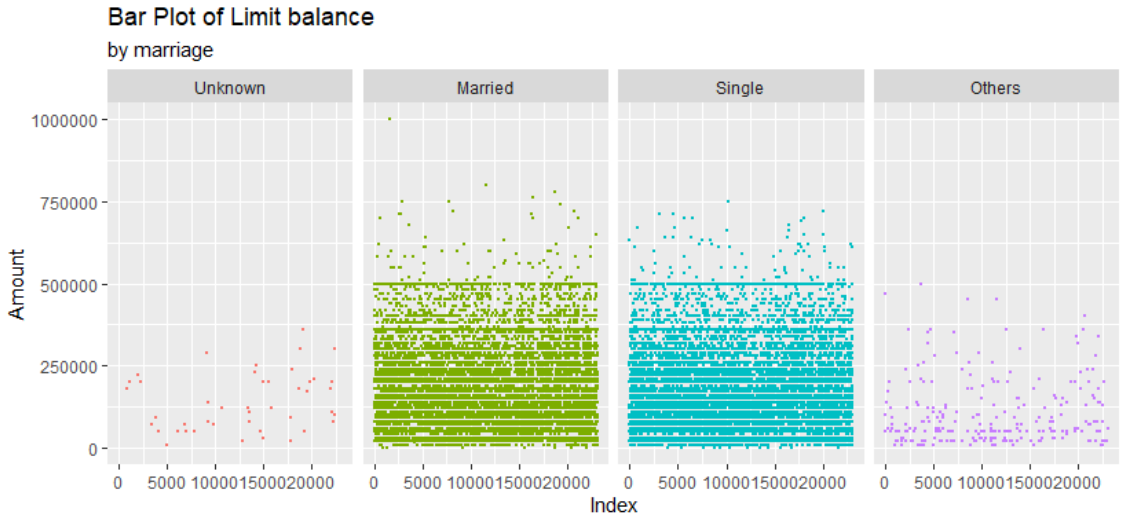
Higher education generally has a higher limit balance.

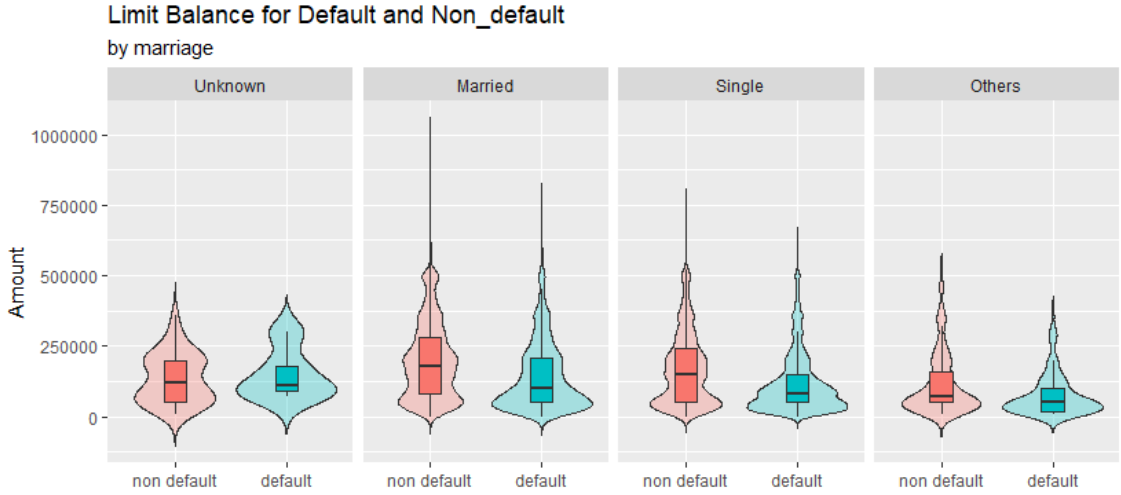




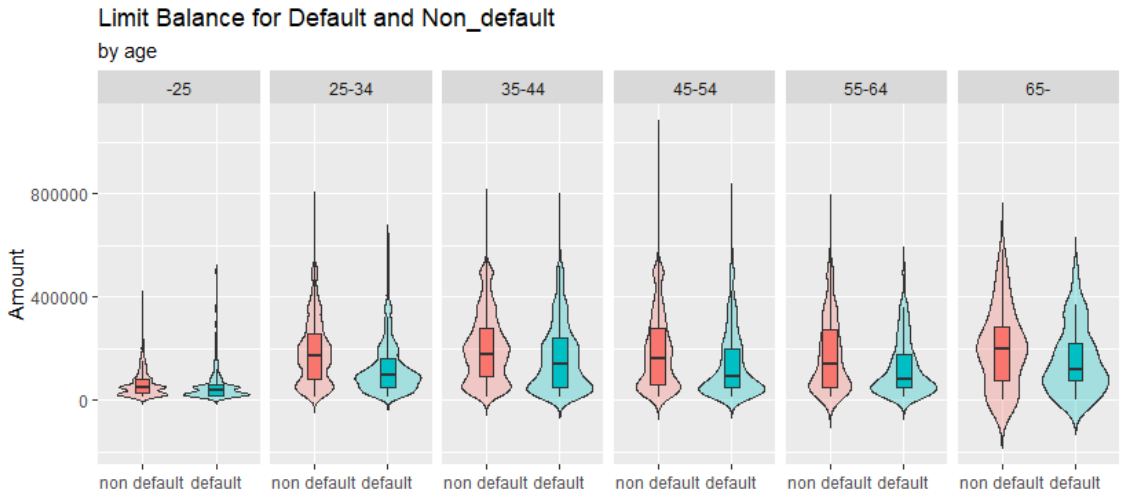
d. By marriage

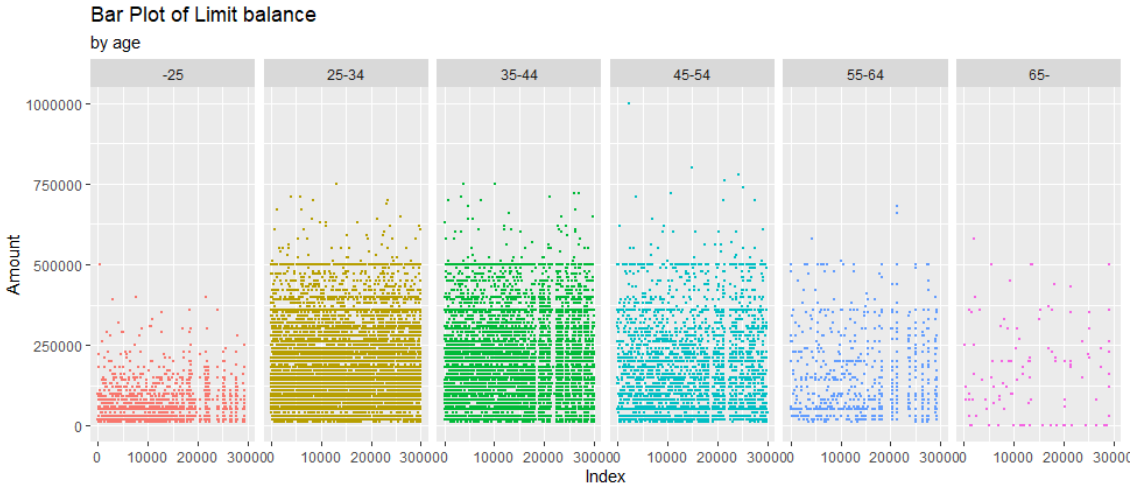
Limit Balance for Married and single clients are almost the same.





e. By Age

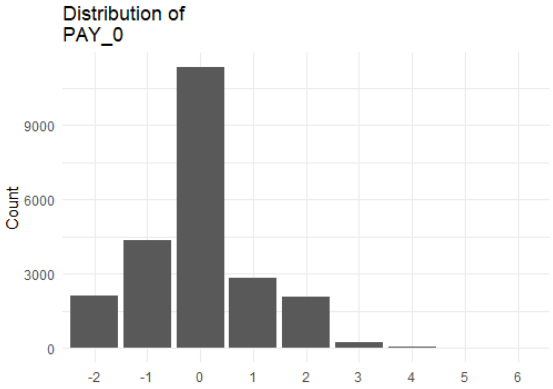

Limit Balance for Default and Non_default
by age
-25
25-34
35-44
45-54
55-64
65-
800000
400000
0
Amount
non default default

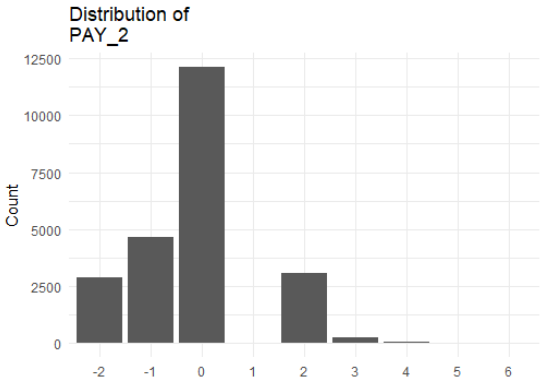

Bar Plot of Limit balance
by age
-25
25-34
35-44
45-54
55-64
65-
1000000
750000
500000
250000
0
Amount
0
10000
20000
300000
Index

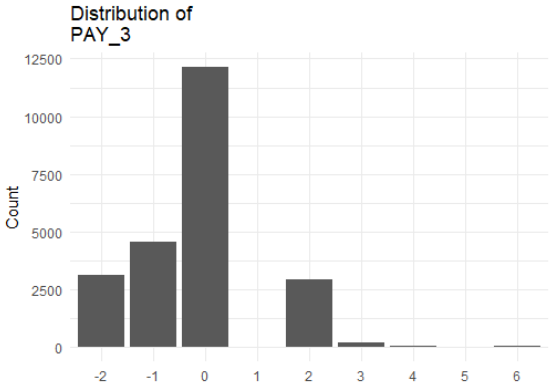
Part 4 - PayX

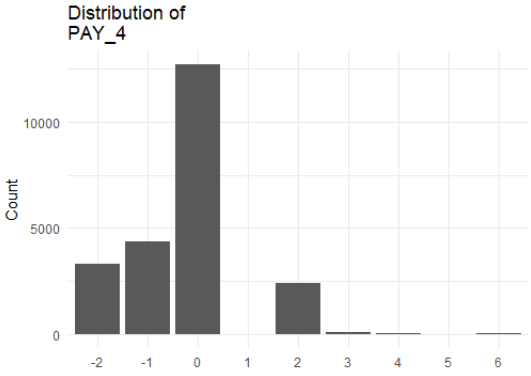
Below graphs are Pay_X distributions, we can see the biggest proportion is captured by 0, followed by -1 then -2. It is noted that both 0 and -2 are not documented in the data dictionary. Due to the amount of 0 and -2 captured, we decided to leave them untreated, otherwise, the data will be undersampled.

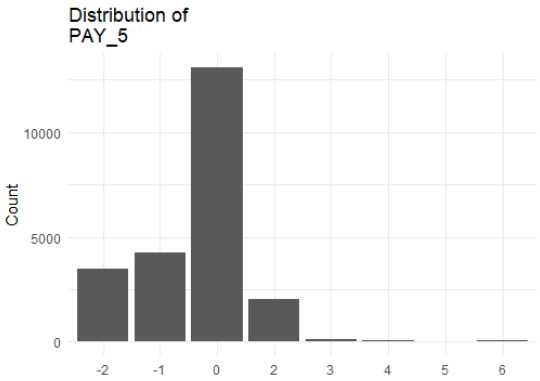
a. Distribution

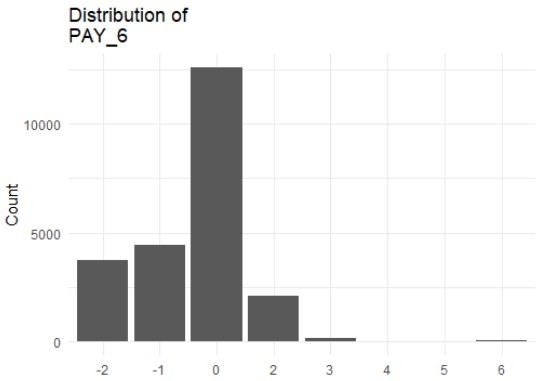




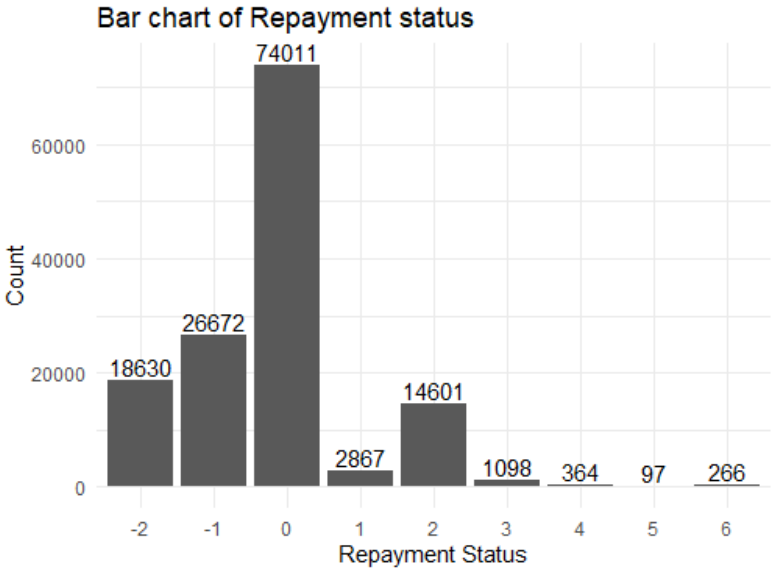








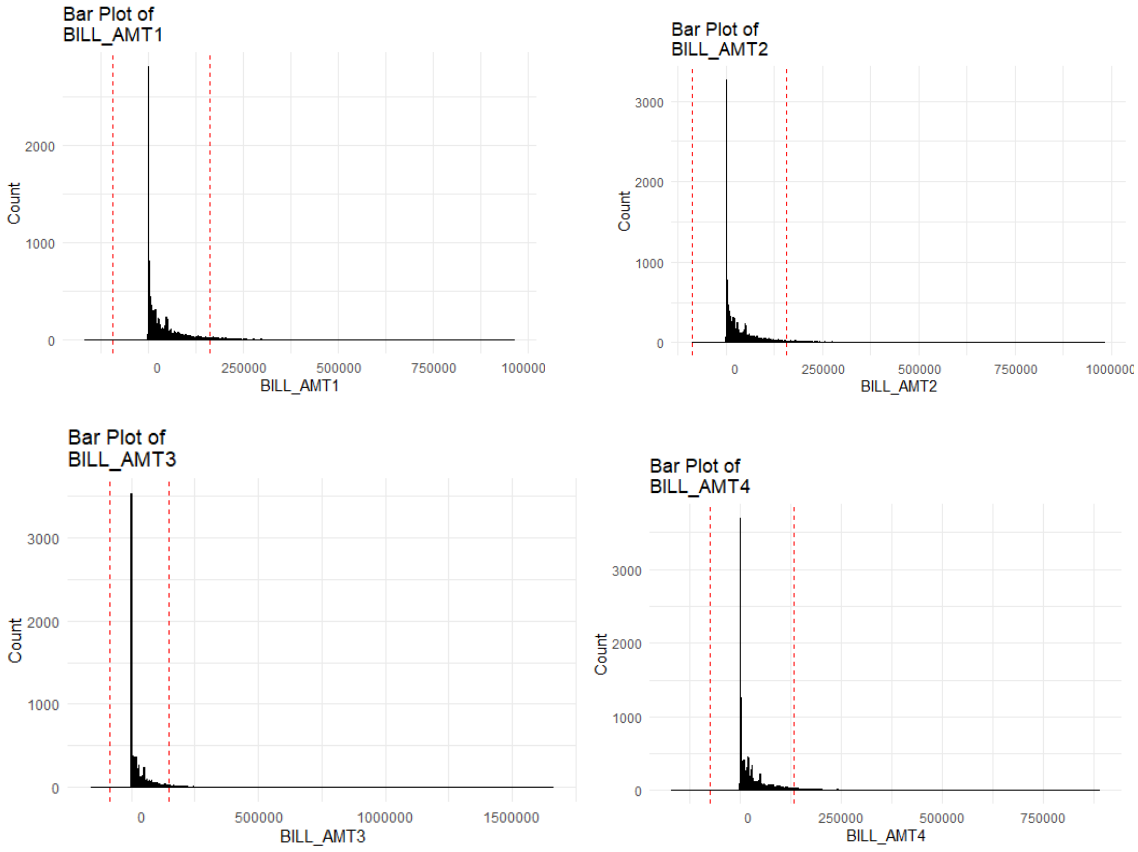
b. Total

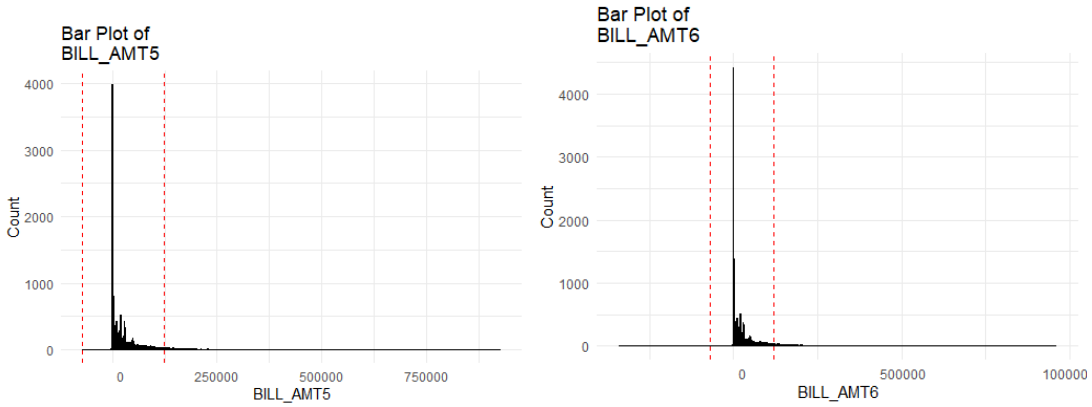


Part 5 - Bill_AMT X

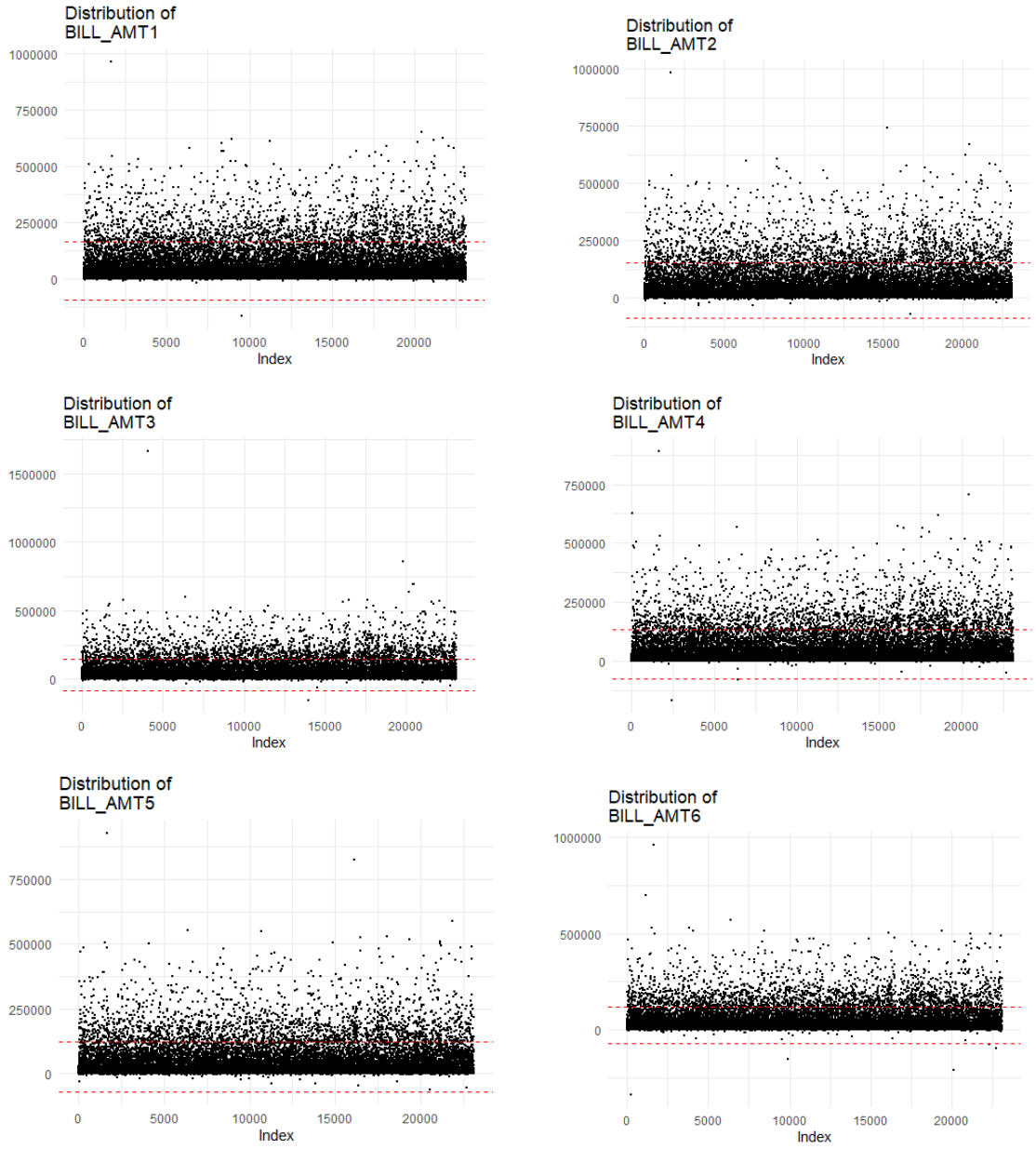
a. Bar Plot

Below are the description of variables Bill_amtX. There are many records falls above the Q3 line, so we decided to cut outliers above Q3+22000 to include more data point.



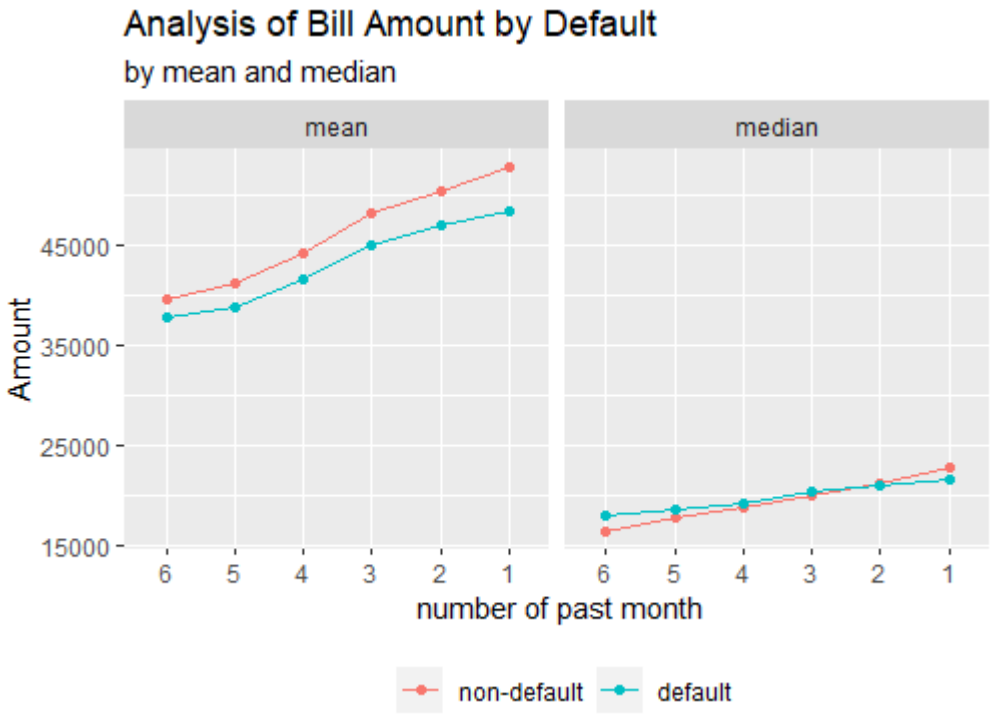


b. Distribution



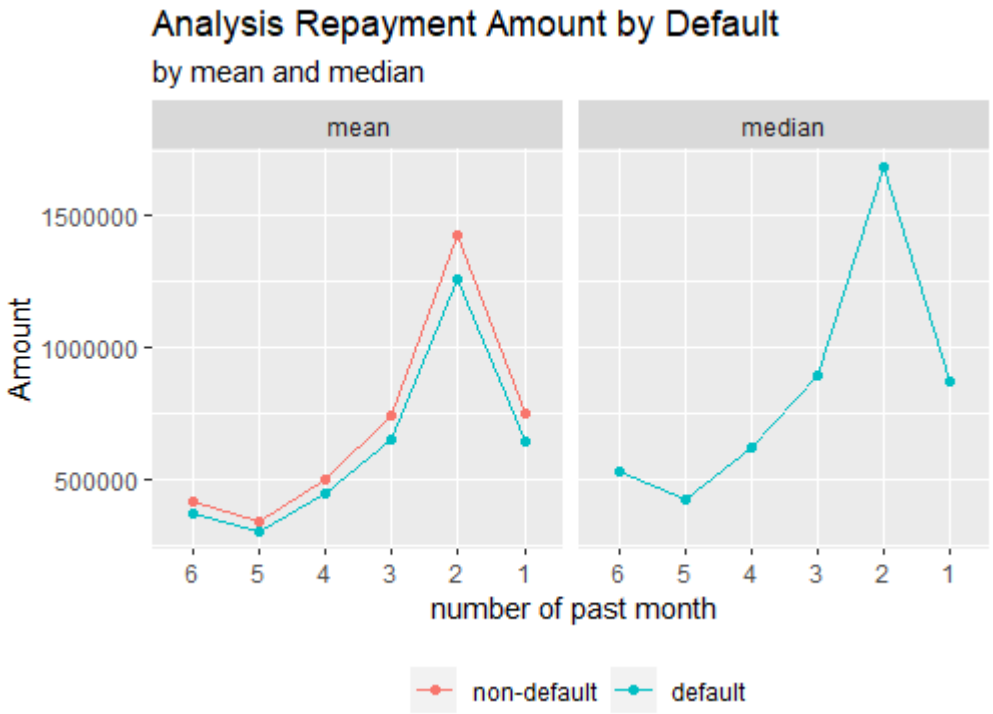
c. By default

We can see an upward trend in both mean and median of Bill Amount, and default client generally has less bill amount.



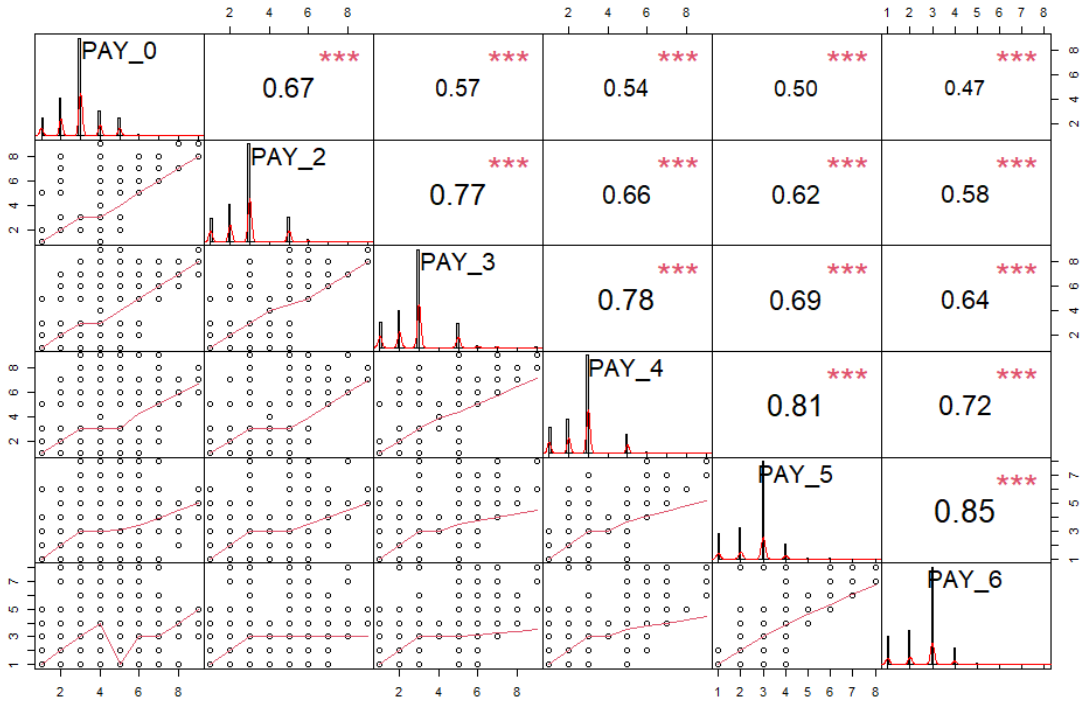
Part 6 - PAY_AMTX

Though this is treated as invalid data, it shows some feature that makes logical sense: default clients generally make less repayment than non-default clients.

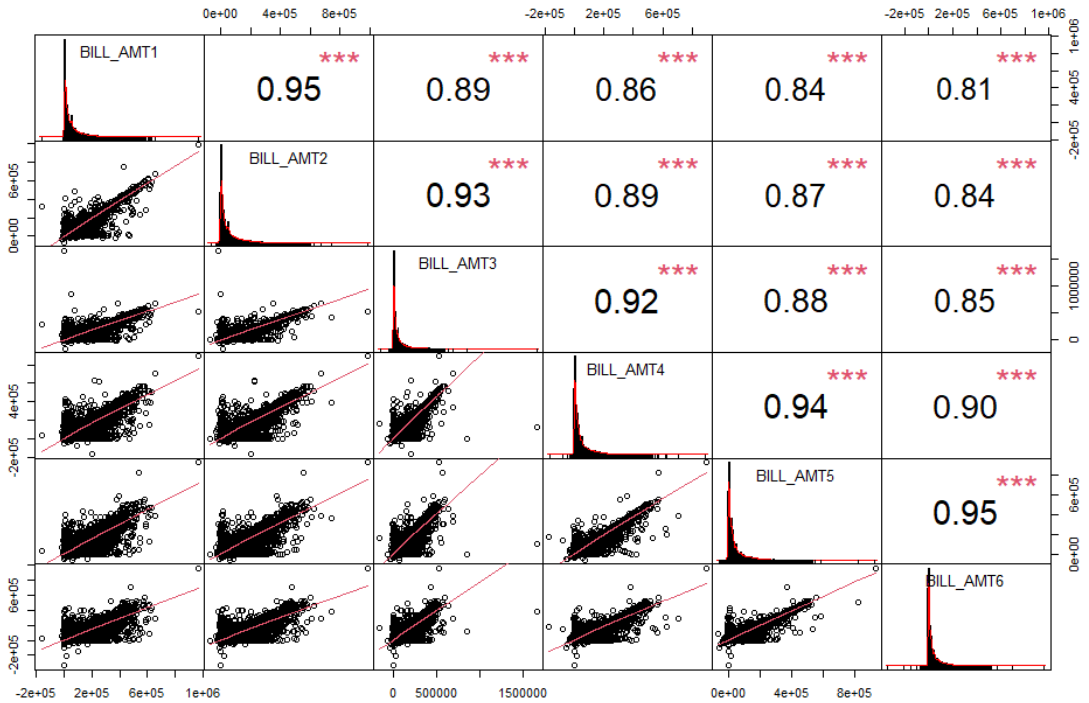


Part 7 - Correlation graph

a. PAY_X



b. BILL_X



Appendix 3

Part 1 - Linear Logistic Model

a. Model

```
Call:
glm(formula = default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 +
    PAY_3 + PAY_5 + PAY_AMT1 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5,
    family = "binomial", data = trainlm)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8047  -0.7530  -0.5980   0.8249   2.6986

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.342e-02  1.337e-01  -0.250 0.802596
LIMIT_BAL   -2.185e-06  1.750e-07 -12.483  < 2e-16 ***
SEX2        -6.795e-02  3.821e-02  -1.778 0.075373 .
MARRIAGE    -1.419e-01  3.892e-02  -3.646 0.000266 ***
AGE          1.468e-02  2.161e-03   6.793 1.10e-11 ***
PAY_0        4.202e-01  2.104e-02  19.974  < 2e-16 ***
PAY_3        1.615e-01  2.212e-02   7.303 2.82e-13 ***
PAY_5        1.006e-01  2.316e-02   4.343 1.40e-05 ***
PAY_AMT1    -4.500e-07  6.195e-08  -7.263 3.78e-13 ***
PAY_AMT3    -4.519e-07  6.232e-08  -7.250 4.16e-13 ***
PAY_AMT4    -3.403e-07  8.240e-08  -4.130 3.62e-05 ***
PAY_AMT5    -4.424e-07  1.206e-07  -3.669 0.000243 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19802  on 17111  degrees of freedom
Residual deviance: 17582  on 17100  degrees of freedom
AIC: 17606

Number of Fisher Scoring iterations: 4
```

b. Confusion matrix

```
       Confusion Matrix and Statistics

                 Reference
       Prediction    N    Y
                N 4062 1141
                Y  181  387

                     Accuracy : 0.7709
                       95% CI : (0.7599, 0.7817)
          No Information Rate : 0.7352
          P-Value [Acc > NIR] : 2.43e-10

                        Kappa : 0.2636

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

                   Sensitivity : 0.25327
                   Specificity : 0.95734
                Pos Pred Value : 0.68134
                Neg Pred Value : 0.78070
                    Prevalence : 0.26477
                Detection Rate : 0.06706
          Detection Prevalence : 0.09842
             Balanced Accuracy : 0.60531

              'Positive' Class : Y
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.25327225	0.95734150	0.68133803	0.78070344
Precision	**Recall**	**F1**	**Prevalence**
0.68133803	0.25327225	0.36927481	0.26477214
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.06705944	0.09842315	0.60530688	

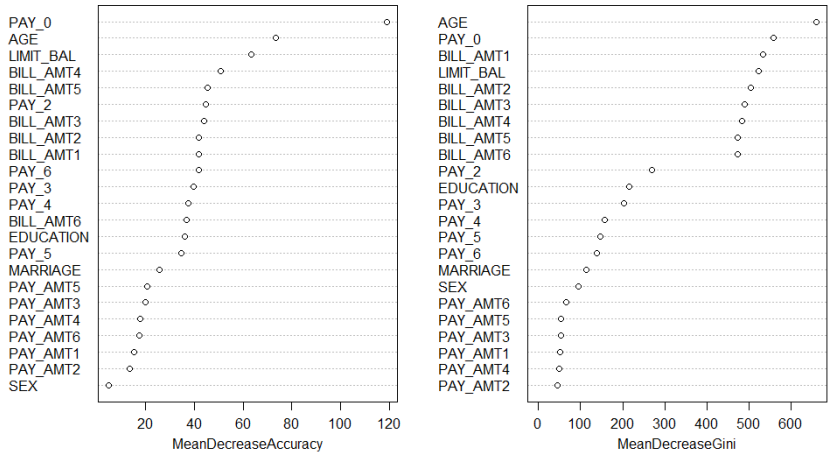
Part 2 - Random Forest

a. Confusion matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 3939  806
         Y  304  722

               Accuracy : 0.8077
                 95% CI : (0.7972, 0.8178)
    No Information Rate : 0.7352
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.448

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

            Sensitivity : 0.4725
            Specificity : 0.9284
         Pos Pred Value : 0.7037
         Neg Pred Value : 0.8301
             Prevalence : 0.2648
         Detection Rate : 0.1251
   Detection Prevalence : 0.1778
      Balanced Accuracy : 0.7004

       'Positive' Class : Y
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.4725131	0.9283526	0.7037037	0.8301370
Precision	**Recall**	**F1**	**Prevalence**
0.7037037	0.4725131	0.5653876	0.2647721
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.1251083	0.1777855	0.7004328	

b. Feature importance graph



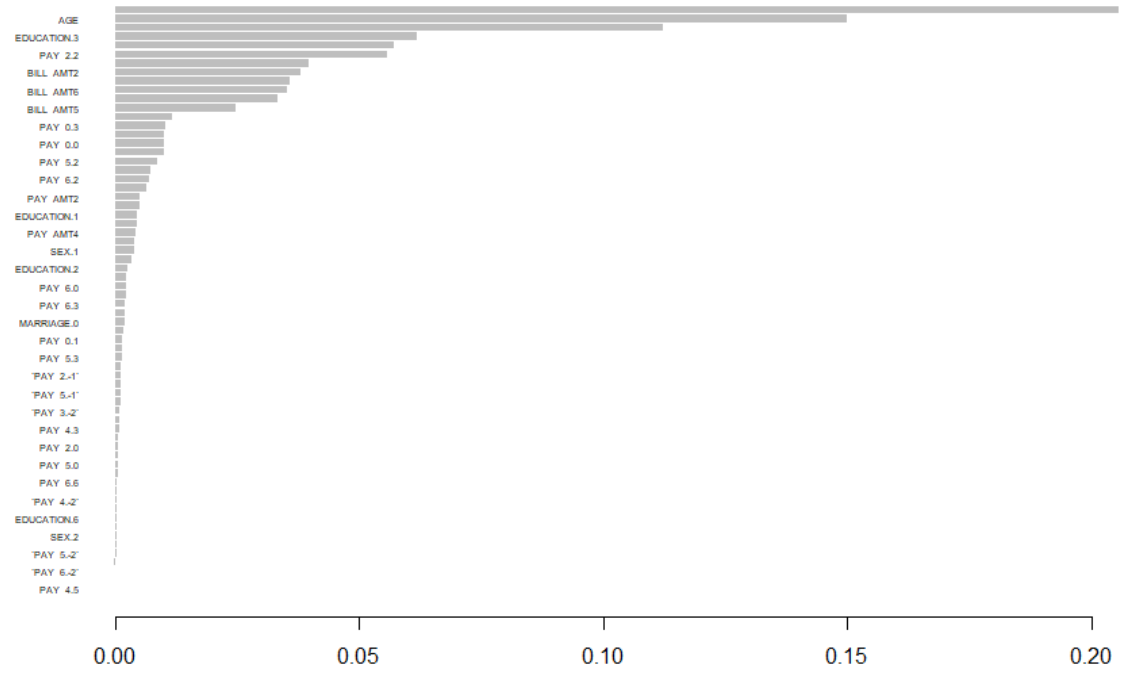
Part 3 - XGBoost

a. Confusion matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3964  757
         1  279  771

               Accuracy : 0.8205
                 95% CI : (0.8103, 0.8303)
    No Information Rate : 0.7352
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4876

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

            Sensitivity : 0.5046
            Specificity : 0.9342
         Pos Pred Value : 0.7343
         Neg Pred Value : 0.8397
             Prevalence : 0.2648
         Detection Rate : 0.1336
   Detection Prevalence : 0.1819
      Balanced Accuracy : 0.7194

       'Positive' Class : 1
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.5045812	0.9342446	0.7342857	0.8396526
Precision	**Recall**	**F1**	**Prevalence**
0.7342857	0.5045812	0.5981381	0.2647721
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.1335990	0.1819442	0.7194129	

b. Variable Importance



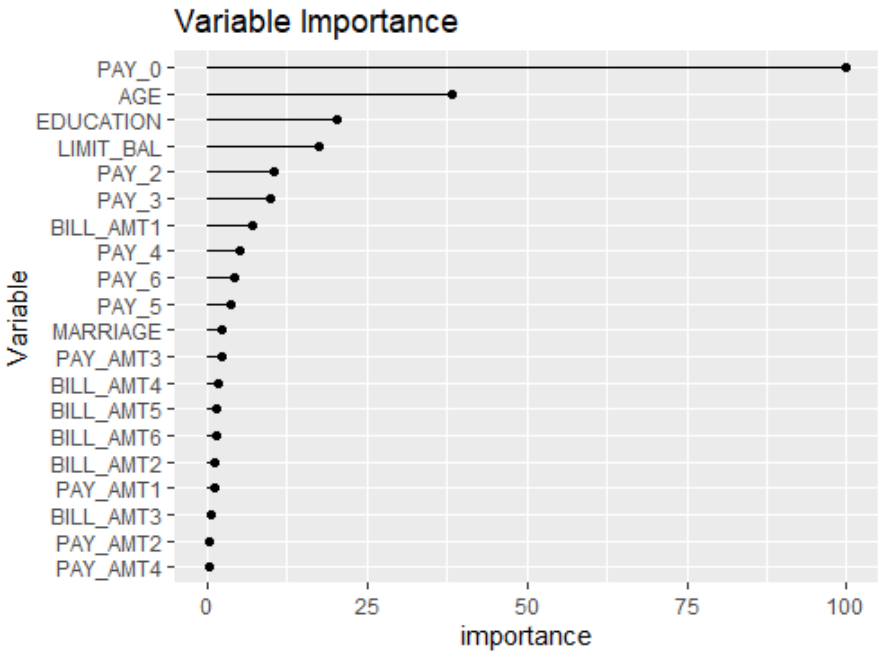
Part 4 - GBM

a. Confusion Matrix

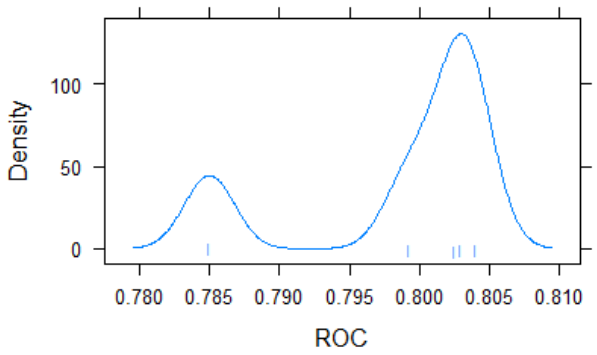
```
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N 3985  877
         Y  258  651

               Accuracy : 0.8033
                 95% CI : (0.7928, 0.8135)
    No Information Rate : 0.7352
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4196

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

            Sensitivity : 0.4260
            Specificity : 0.9392
         Pos Pred Value : 0.7162
         Neg Pred Value : 0.8196
             Prevalence : 0.2648
         Detection Rate : 0.1128
   Detection Prevalence : 0.1575
      Balanced Accuracy : 0.6826

       'Positive' Class : Y
```

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
0.4260471	0.9391940	0.7161716	0.8196216
Precision	**Recall**	**F1**	**Prevalence**
0.7161716	0.4260471	0.5342634	0.2647721
Detection Rate	**Detection Prevalence**	**Balanced Accuracy**	
0.1128054	0.1575117	0.6826205	

b. Variable Importance



c. Plot



Part 5 - Lasso

a. Confusion matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4233 1526
         1   10    2

               Accuracy : 0.7338
                 95% CI : (0.7222, 0.7452)
    No Information Rate : 0.7352
    P-Value [Acc > NIR] : 0.6009

                  Kappa : -0.0015

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0013089
            Specificity : 0.9976432
         Pos Pred Value : 0.1666667
         Neg Pred Value : 0.7350234
             Prevalence : 0.2647721
         Detection Rate : 0.0003466
   Detection Prevalence : 0.0020794
      Balanced Accuracy : 0.4994760

       'Positive' Class : 1
```

```
    Sensitivity         Specificity      Pos Pred Value      Neg Pred Value
   0.0013089005        0.9976431770        0.1666666667        0.7350234416
      Precision              Recall                  F1         Prevalence
   0.1666666667        0.0013089005        0.0025974026        0.2647721365
Detection Rate Detection Prevalence    Balanced Accuracy
   0.0003465604        0.0020793623        0.4994760388
```

b. Plot