**Big Data Analytics**

**Credit Default Prediction**

**PROJECT REPORT**

***Submitted by***

Narendra Bandi (002607680)

Pratik Chaudhari (002601825)

Santhosh Babu Rajamanickam Natarajan (002612609)

Sai Srujith Guduri (002602121)

**Table of Contents**

1. **Introduction**……………………………………………………………………………….....3
   1. Dataset………………………………………………………………………………………......3
   2. Dataset attributes………………………………………………………………………………..3
      1. Demographic variables……………………… ………………………………………….....3
      2. Credit Payment variables………………...………………………………………………....3
   3. Glimpse of data………………………………………………………………………………………………4
   4. Problem statement…………………….…………………………………………………………4
2. **Data Cleaning**……………………………………………………………………………..…4
   1. Missing Value…………………………………………………………………………..……….4
   2. Unbalance data……………………………………………………………………….....……….4
   3. Data summary……………………….……………………………………………………..……5
3. **Data Visualization**…………………………………………………………………..……....5
   1. Education level Vs Default………………………………………………...……….……......….5
   2. Gender Vs Default……………………………..……………………………………………..…5
   3. Plotting graph for Pay\_X to see the variation in values…………………………...……………6
   4. Credit balance distribution with default payment……………………..………………………..6
   5. Frequency of credit limits……………………...……………………………………………….7
      1. Correlation test………………………...……………………………………………………7
   6. Bill statements...…………………………………………………………………………..…….8
      1. Correlation test………………………………………………...……………………………8
   7. History of Past Payment……………………………………………..………………………….9
      1. Correlation test..……………………….……………………………………………………9
4. **Model Fitting** ……………………………………….......…………………………..……....10
   1. Logistics Regression………………………………..........………………………………….….10
      1. Lasso Regression....................................................................................................................11
   2. Classification Tree…………….....................................………………..…………………….…11
      1. Random Forest Ensemble Method.........................................................................................12
   3. Naïve Bayes Classification Method.…………………………...………..............................…...13
   4. KNN Classification ……………………..……..............................................………………….13
   5. Majority Voting Ensemble ……………………...……………………………………………...14
   6. Neural Network............................................................................................................................14
5. **Dimension reduction using PCA** ……………………………………………………..……..........15
6. **Results and discussion** …………………………………………………………………..…….......16
7. **Learnings** …………………………………………………………………..……............................17
8. **References** …………………………………………………………………..……...........................17

**1.INTRODUCTION**

**1.1 Dataset**

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from **April 2005**to **September 2005**.

**1.2 Dataset Attributes**

**1.2.1 Demographic variables**

These variables contain the data of customer’s personal information.

* 1. **ID**: ID of each client
  2. **SEX**: Gender (1=male, 2=female)
  3. **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
  4. **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
  5. **AGE**: Age in years

**1.2.2 Credit Payment variables**

These variables mostly contain the details related to the payments and the credits

* 1. **PAY\_0**: Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
  2. **PAY\_2**: Repayment status in August 2005 (scale same as above)
  3. **PAY\_3**: Repayment status in July 2005 (scale same as above)
  4. **PAY\_4**: Repayment status in June 2005 (scale same as above)
  5. **PAY\_5**: Repayment status in May 2005 (scale same as above)
  6. **PAY\_6**: Repayment status in April 2005 (scale same as above)
  7. **LIMIT\_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
  8. **BILL\_AMT1**: Amount of bill statement in September 2005 (NT dollar)
  9. **BILL\_AMT2**: Amount of bill statement in August 2005 (NT dollar)

10. **BILL\_AMT3**: Amount of bill statement in July 2005 (NT dollar)

11. **BILL\_AMT4**: Amount of bill statement in June 2005 (NT dollar)

12.**BILL\_AMT5**: Amount of bill statement in May 2005 (NT dollar)

13. **BILL\_AMT6**: Amount of bill statement in April 2005 (NT dollar)

14.**PAY\_AMT1**: Amount of previous payment in September 2005 (NT dollar)

15. **PAY\_AMT2**: Amount of previous payment in August 2005 (NT dollar)

16. **PAY\_AMT3**: Amount of previous payment in July 2005 (NT dollar)

17. **PAY\_AMT4**: Amount of previous payment in June 2005 (NT dollar

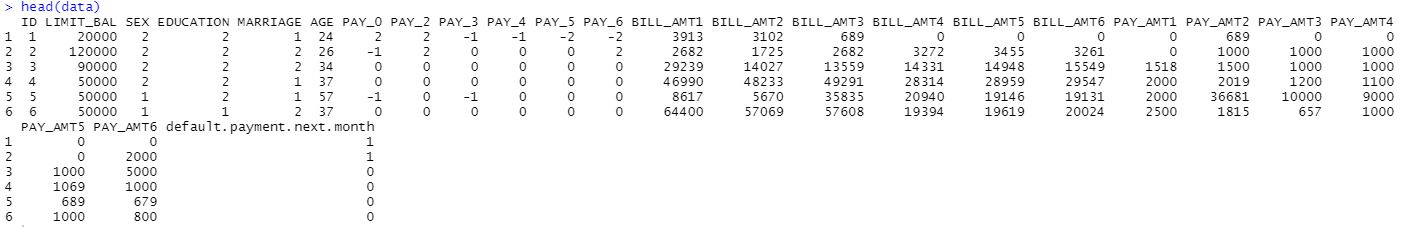
18. **PAY\_AMT5**: Amount of previous payment in May 2005 (NT dollar)

19. **PAY\_AMT6**: Amount of previous payment in April 2005 (NT dollar)

20. **default.payment.next.month**: Default payment (1=yes, 0=no)

**Total count 🡪** Rows = 30,000 Columns = 26

**1.4 Glimpse of Data**



**1.5 Problem Statement**

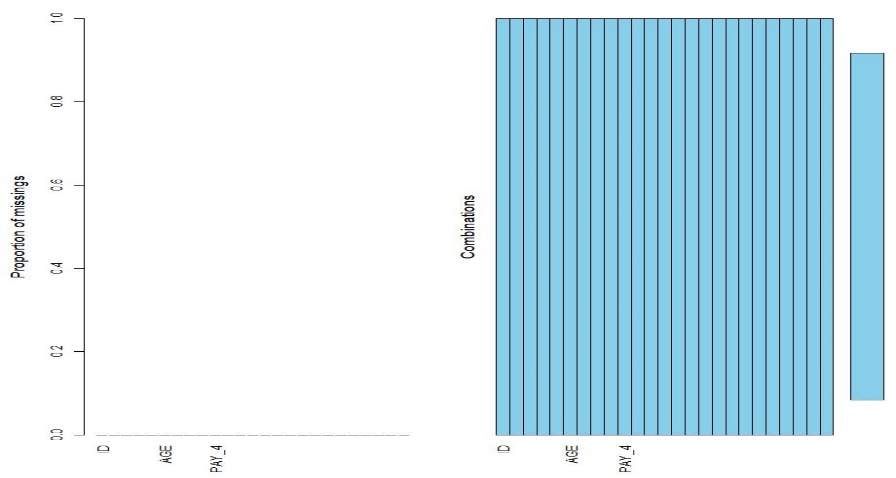
1. How does the probability of default payment vary by categories of different demographic variables?

2. Which variables are the strongest predictors of default payment?

**2.Data Cleaning**

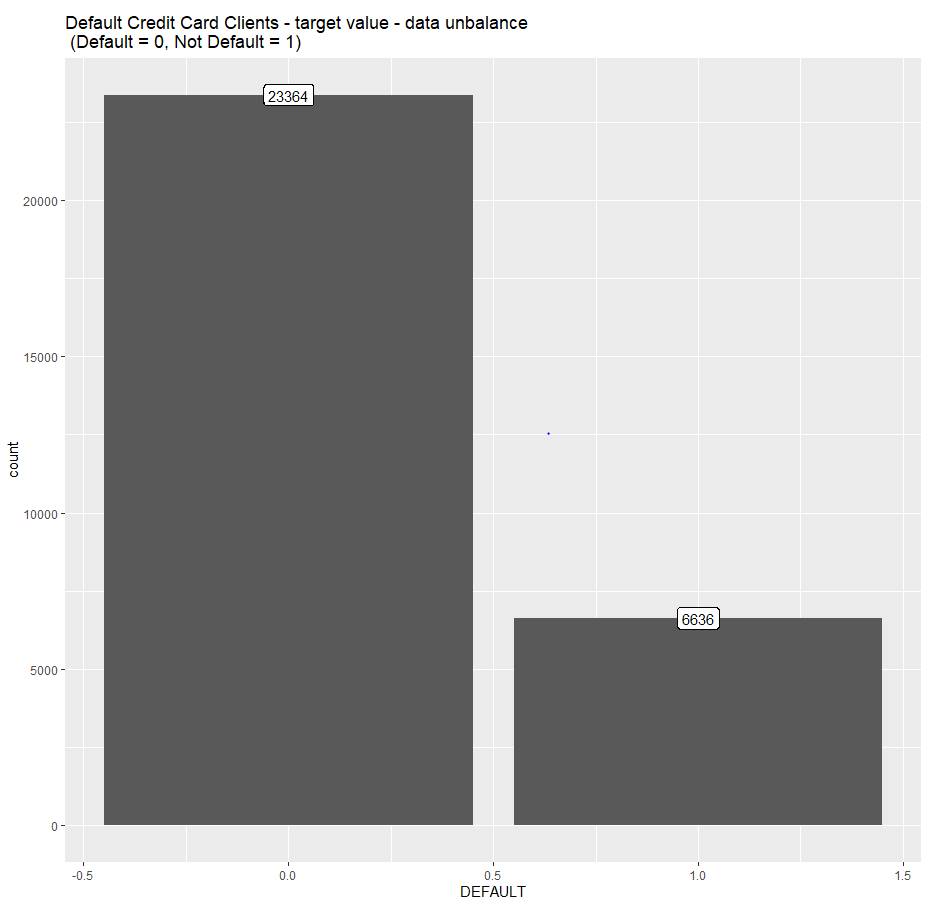
**2.1 Missing value**

Here, we are searching for the missing value.



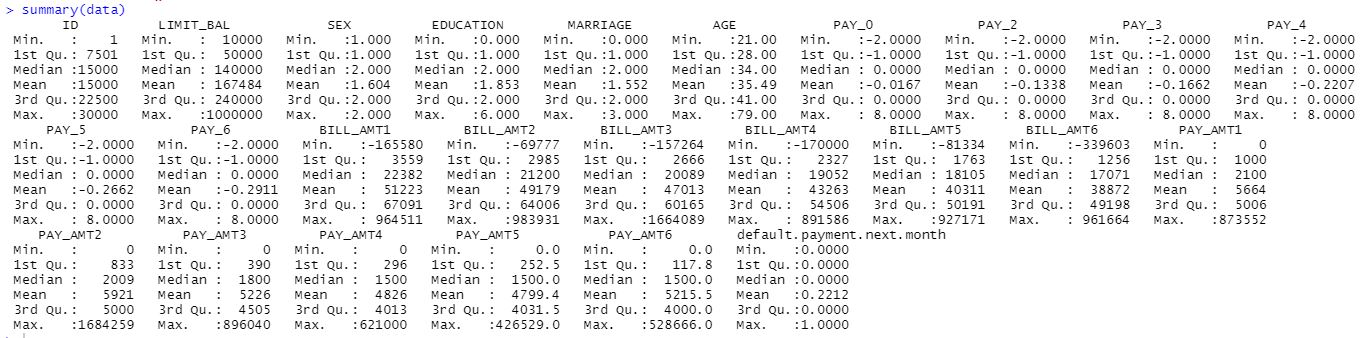
The data does not contain any missing value as we can see in the above chart.

**2.2 Unbalance Data**



We can see that 6,636 out of 30,000 (or 22%) customers will default for the next month. The information has got a huge unbalance with reference to the target (default.payment.next.month).

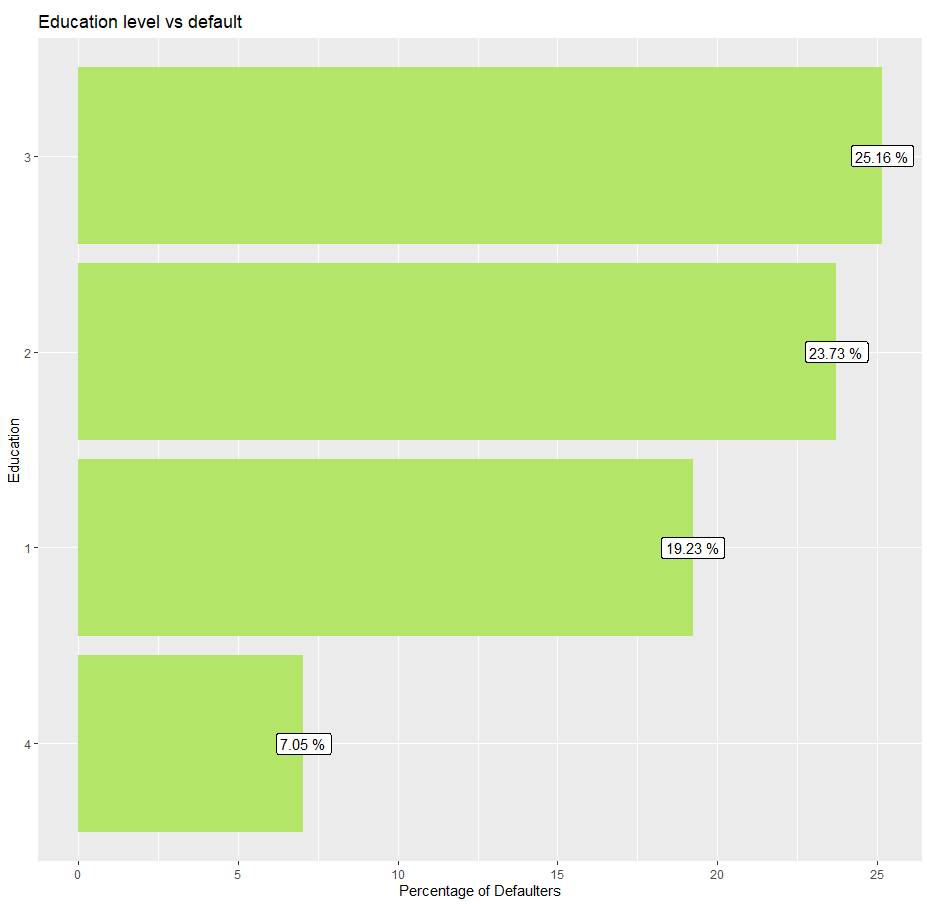
**2.3 Data Summary**



* There are 30,000 distinct credit card clients.
* The average value for the amount of credit card limit is 167,484. The standard deviation is unusually large, the max value being 1M.
* Education level is mostly graduate school and university.
* Most of the clients are either married or single (less frequent than the other status).
* The average age is 35.5 years, with a standard deviation of 9.2.
* As the value 0 for default payment means 'not default' and value 1 means 'default', the mean of 0.221 means that there are 22.1% of credit card contracts that will default next month (will verify this in the next sections of this analysis)

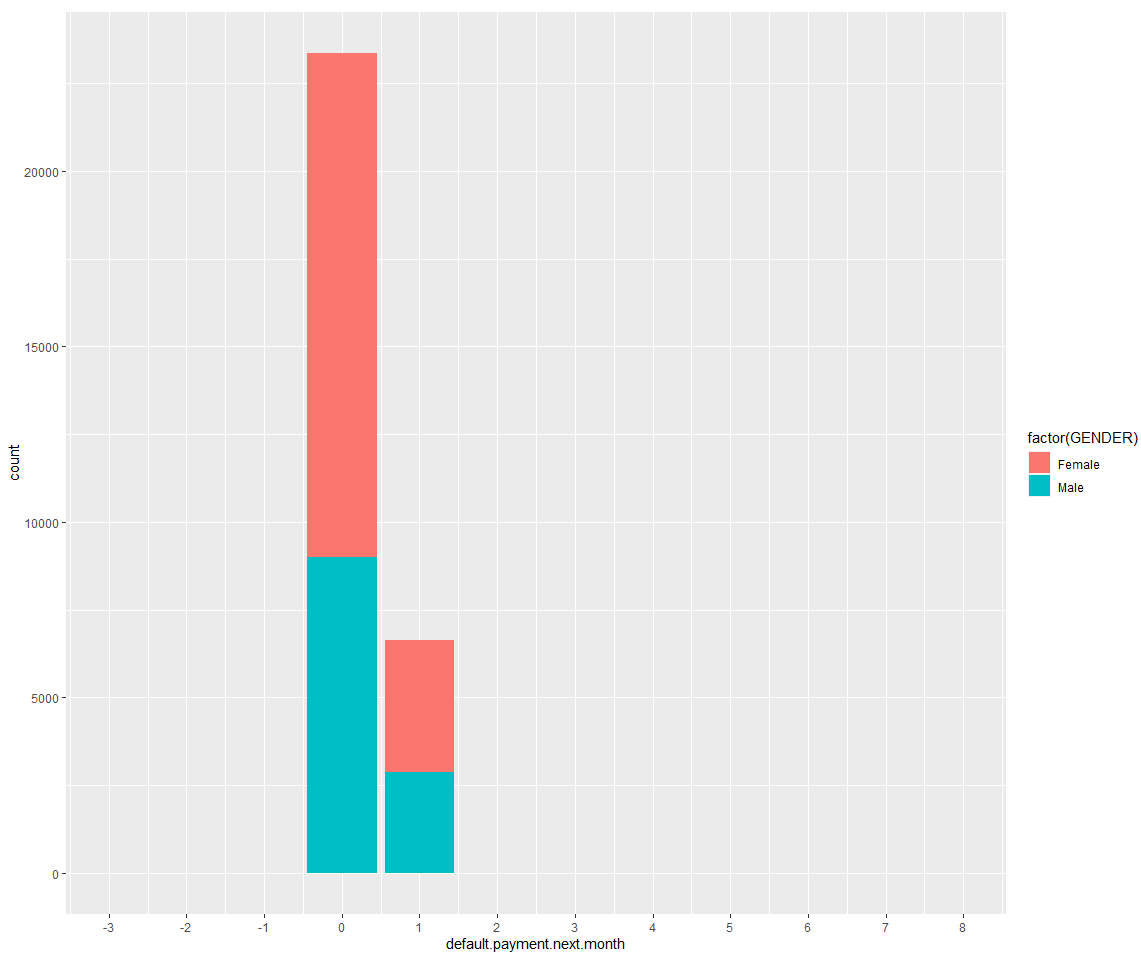
**3.Data Visualization**

**3.1 Education level vs Default**



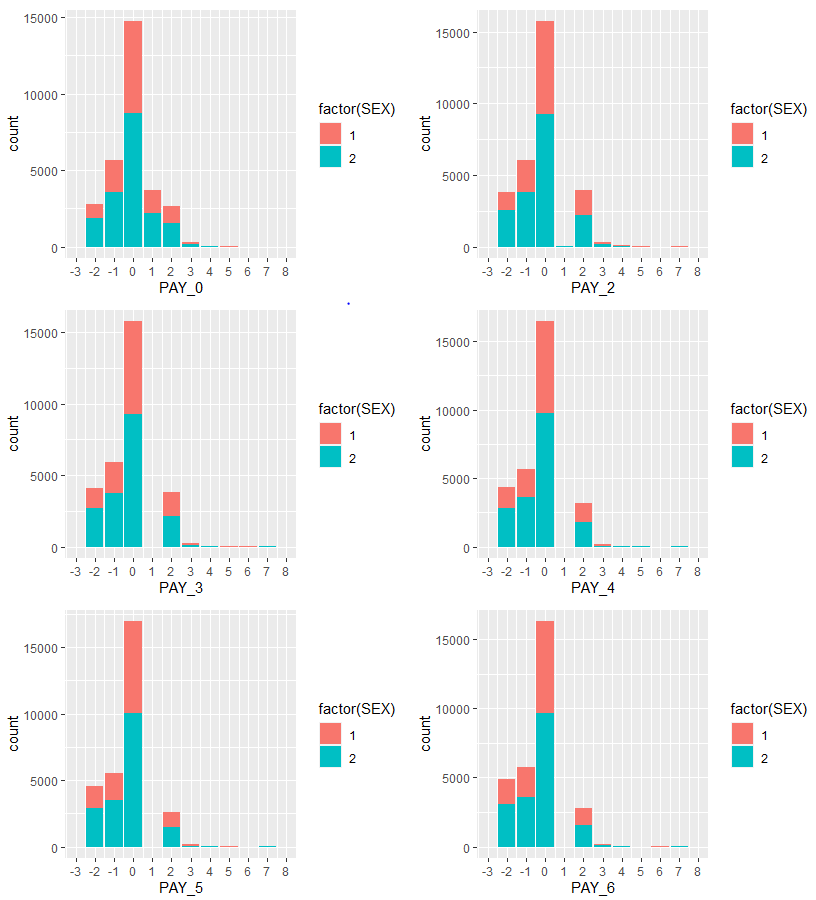
* Here we found that around 25% are defaulter clients are having High School (3) degree. Furthermore, correspondingly around 24% of defaulter clients are having University (2) degrees.

**3.2 Gender Vs Default**



* In the above plot, we can observe a greater number of female customers falls under the defaulter’s category for the next month compared to the number of male customers.

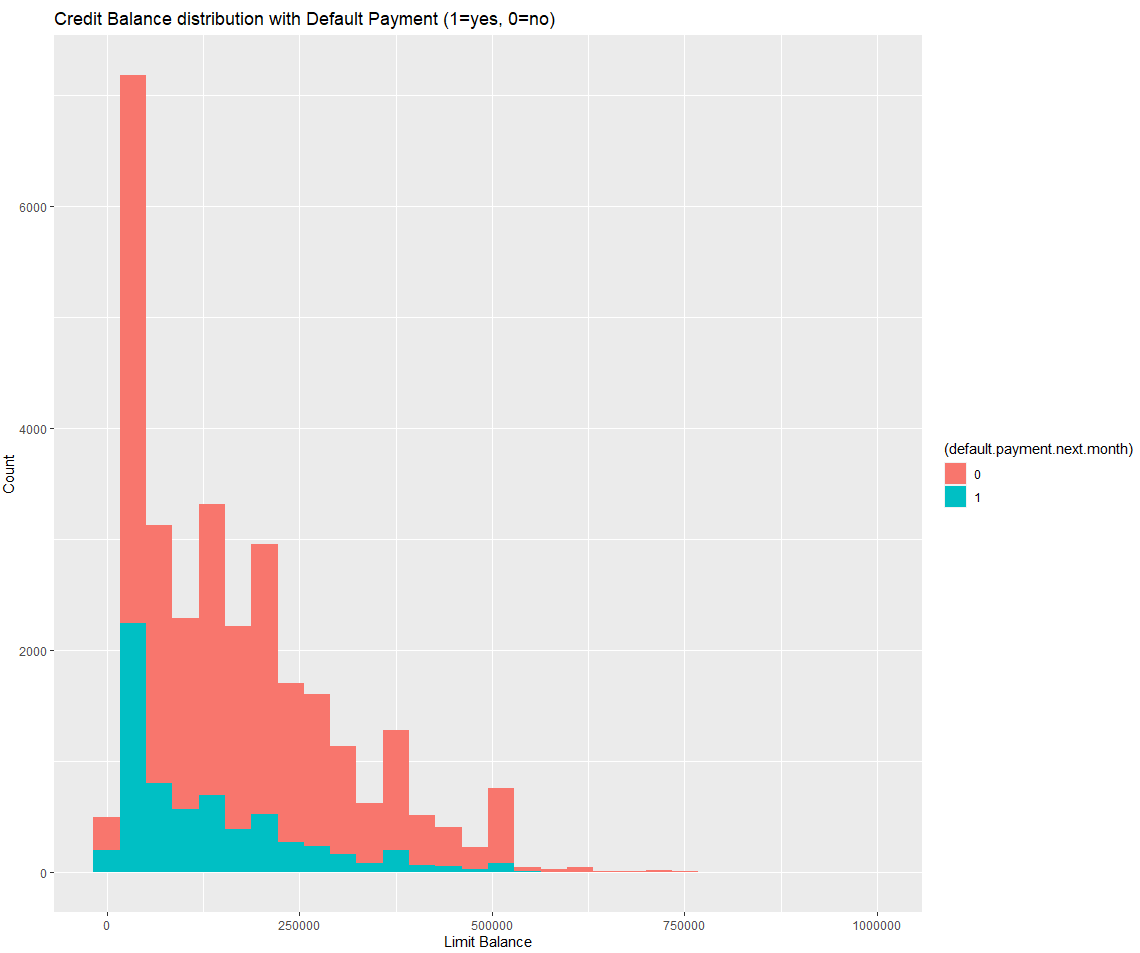
**3.3** **Plotting graph for Pay\_X to see the variation in values**



From the above graph, we can infer

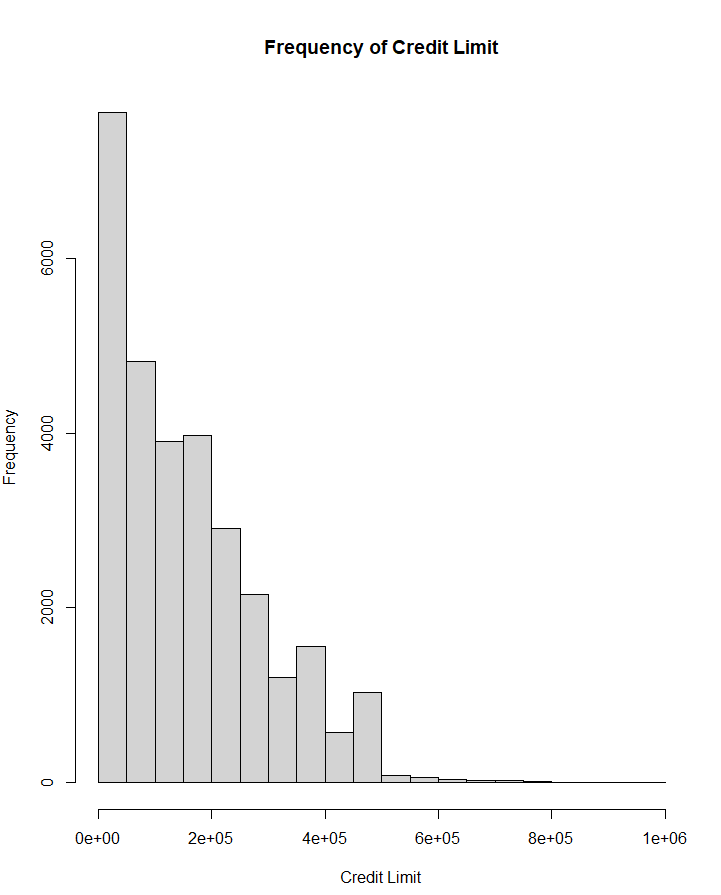
* There is an additional undocumented value of Pay\_X value of 0 and -2
* Pay\_X = 0 seem to have max value

**3.4 Credit Balance distribution with Default Payment**



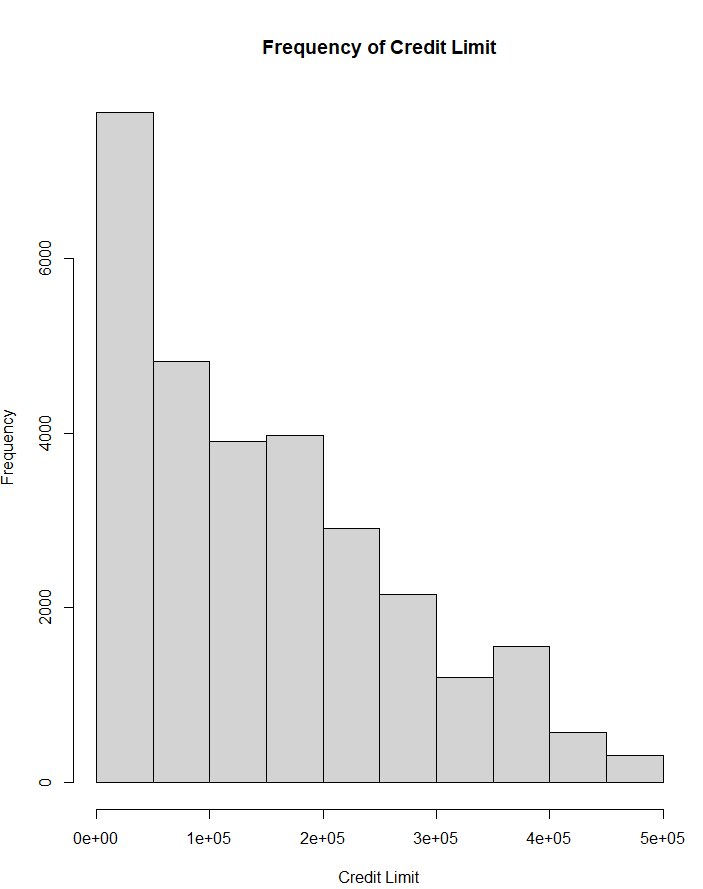
* In this plot, we can see that most of the defaulters fall under the category of low credit balance limit. We can see a huge spike in the less than 50000 limit balance range which has the majority of the defaulters.
* Also, we can see that for the majority of bars with different limit balance range, the defaulters are higher in number than the non-defaulters.

**3.5 Frequency of Credit Limit**



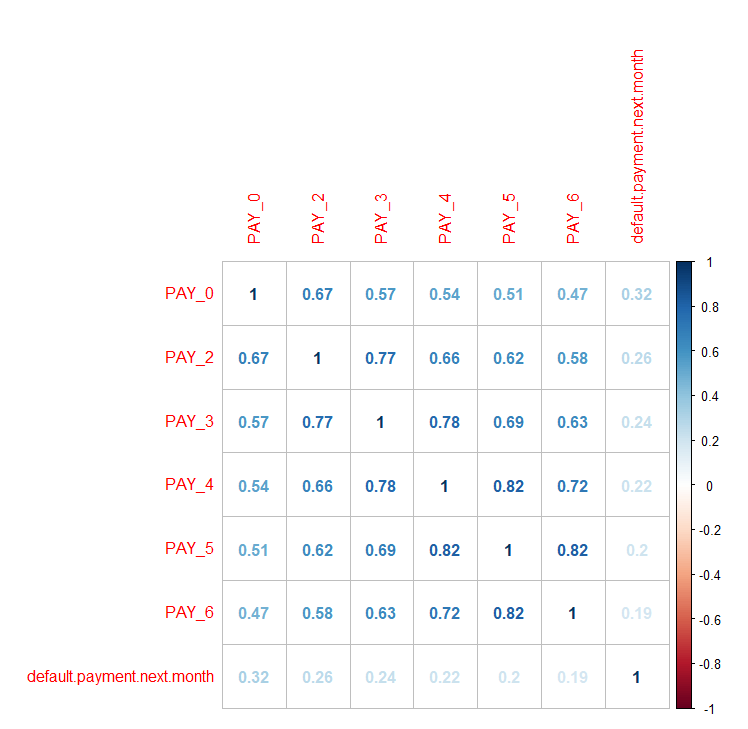
* We found that there were about 928 customers with a credit limit of more than $500000. These customers take up about 3% of the data.

Next, we have analyzed the data with customers who hold credit limit less than $500000



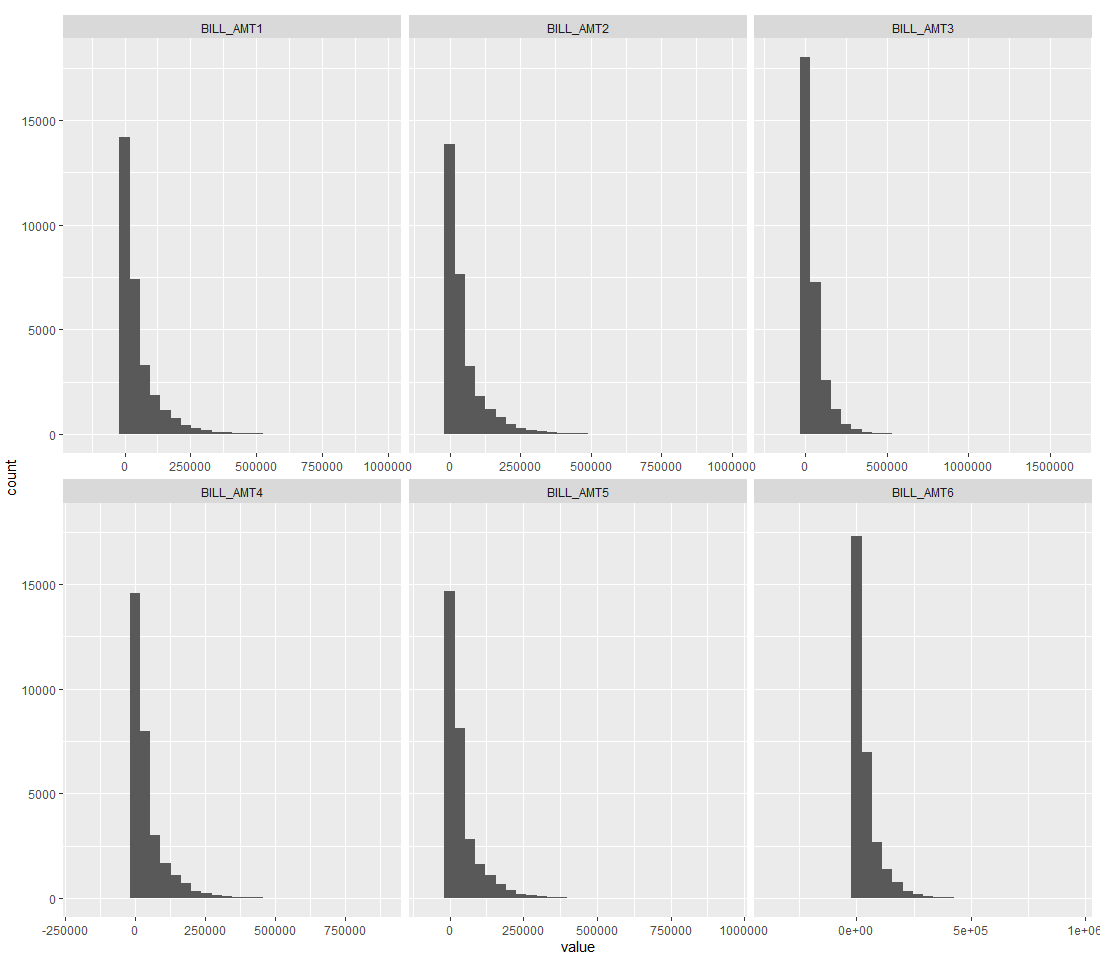
* We have now taken into account only the customers who have got less than 500000 credit limits, we can see that the distribution is right-skewed with 10 distinct groups. The majority of the customers fall under this credit limit category with very few in the 400000 – 500000 range.

**3.5.1 Correlation Test**



* In the correlation analysis, we found that there is a little correlation between the pay delay variable and default. However, there seems to be a high correlation within different months of pay delay.

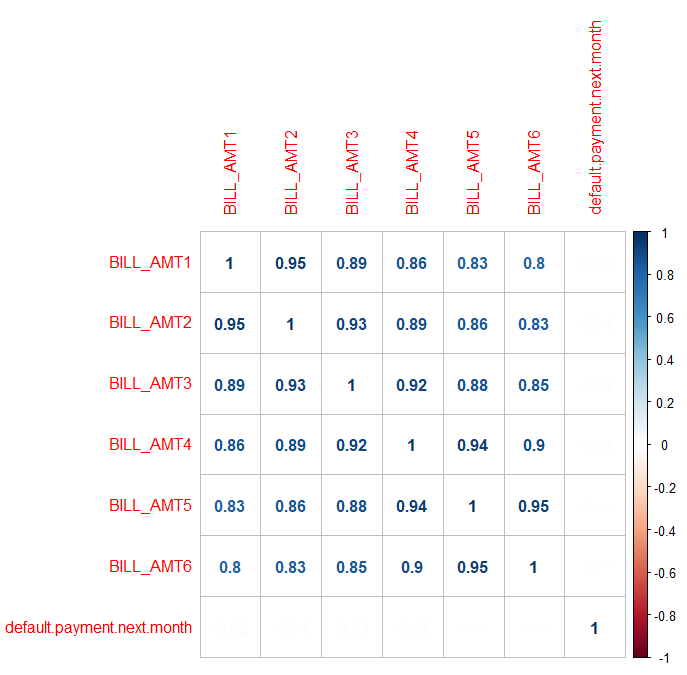
**3.6 Bill Statement**



* We have plotted the distribution of bill statements for each month, we can see that all of them are right-skewed, and the majority of bill statements are less than $250000.

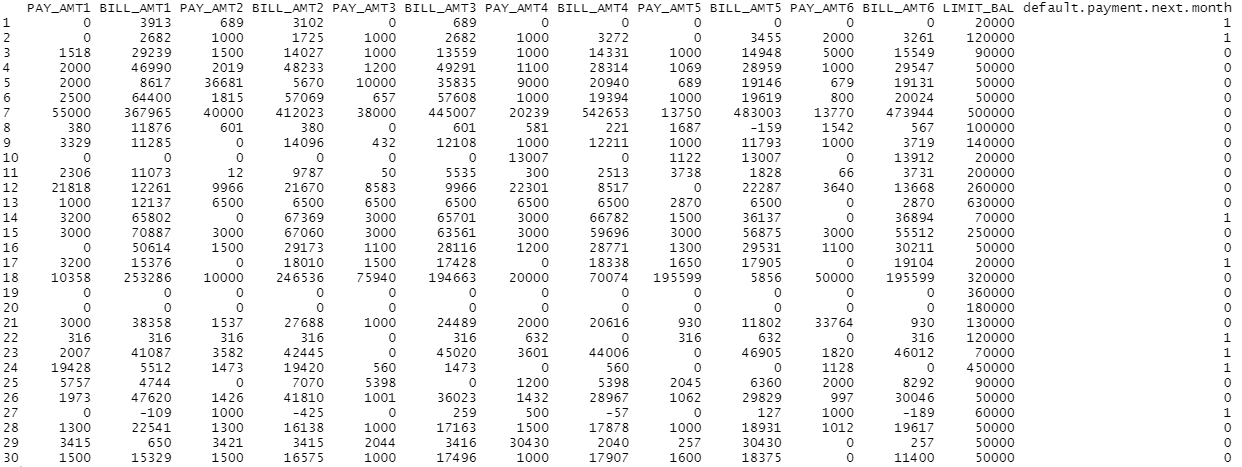
Next, we have performed a correlation test for the bill statement.

**3.6.1 Correlation Test**

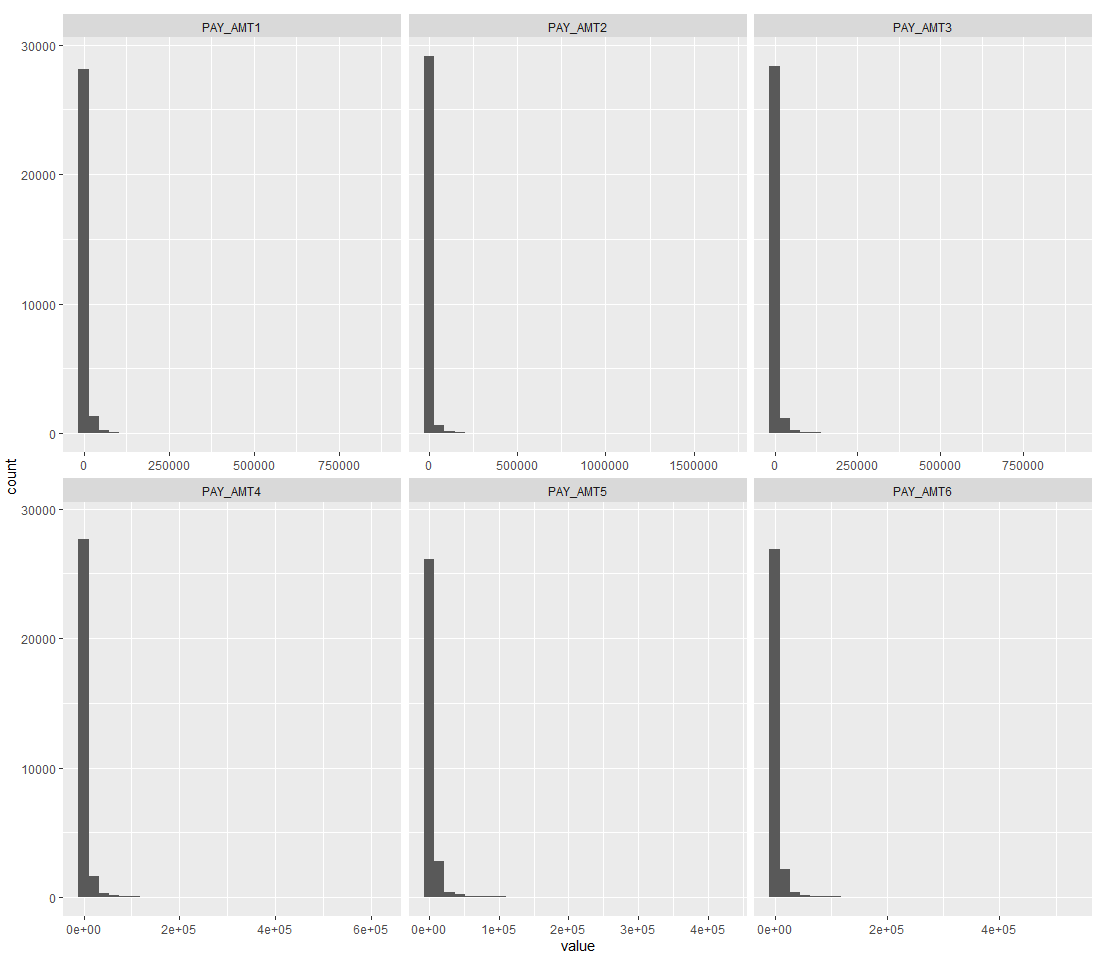


* We can observe from the correlation plot above, that bill amount is highly correlated for every month but it has close to 0 correlation with default payment.

In the next steps, we have decided to look at the relationship between pay amount for each month with limit balance and default payment next month



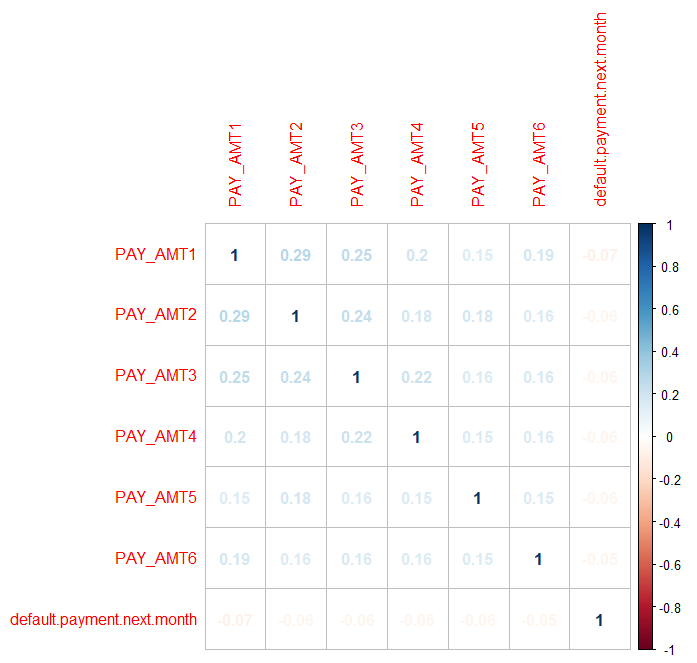
* We made some observations from the above table, there are consecutive 0’s for a few months with some at the start and end. We can interpret that it can be due to some customers who left near the end of the year and some who joined later in the year.
* We can also infer from the above table, how the variables relate to one another, the pay amount is paid for the month's bill statement and those that have not paid the full amount will be carried forward to the following bill statement.
  1. **History of Past Payment**



* From the above distribution of previous payments for each month, we observed that the distributions are right-skewed with the majority of the payments less than $50000. On comparing this observation with what we found in the bill statement analysis where a majority of the customers were billed less than $250000 and here, we can see that majority of the customers paid lesser than $50000
* We can infer from this observation that customers are likely to default on their next month's payment as they didn’t pay the full amount they were billed.

Next, we have performed a correlation test for the history of past payments.

**3.7.1 Correlation Test**



Based on the correlation plot above, we can observe that there is a low correlation between every month's pay amount (<0.50), therefore we can conclude there is no multicollinearity between the different pay amounts. Also, we observed that the correlation between the pay amount for different months and default status is close to 0.

In summary, on performing the exploratory data analysis process, we observed that there were anomalies in the variable’s education (0,5, and 6) marriage (0), and pay delay (0 and -2). Also, we observed there is a low correlation for the variables pay delay, bill amount, and pay amount with the default payment for next month.

**4. Model Fitting**

Since the response variable for this analysis “default.payment.next.month” is a binary category variable ( 0 = No default, 1 = Default), binary classification methods such as: Logistic Regression, Classification Tree, RanomForest ensembling, Naïve Bayes classification, K-Nearest Neighbour, Majority Voting ensembling and Neural Network( General training ) models are fitted. Also, Principal component analysis (PCA) was done to identify if dimension reduction is possible.

The dataset is portioned into 70% Training data and 30% as Validation data, Stratified portioning is used to have equal proportion of Default, No-Default records in both Training and Validation samples.

For Education, Marriage categorical variables the factors whose definition is not available are collapsed into Unknown/Other categories.

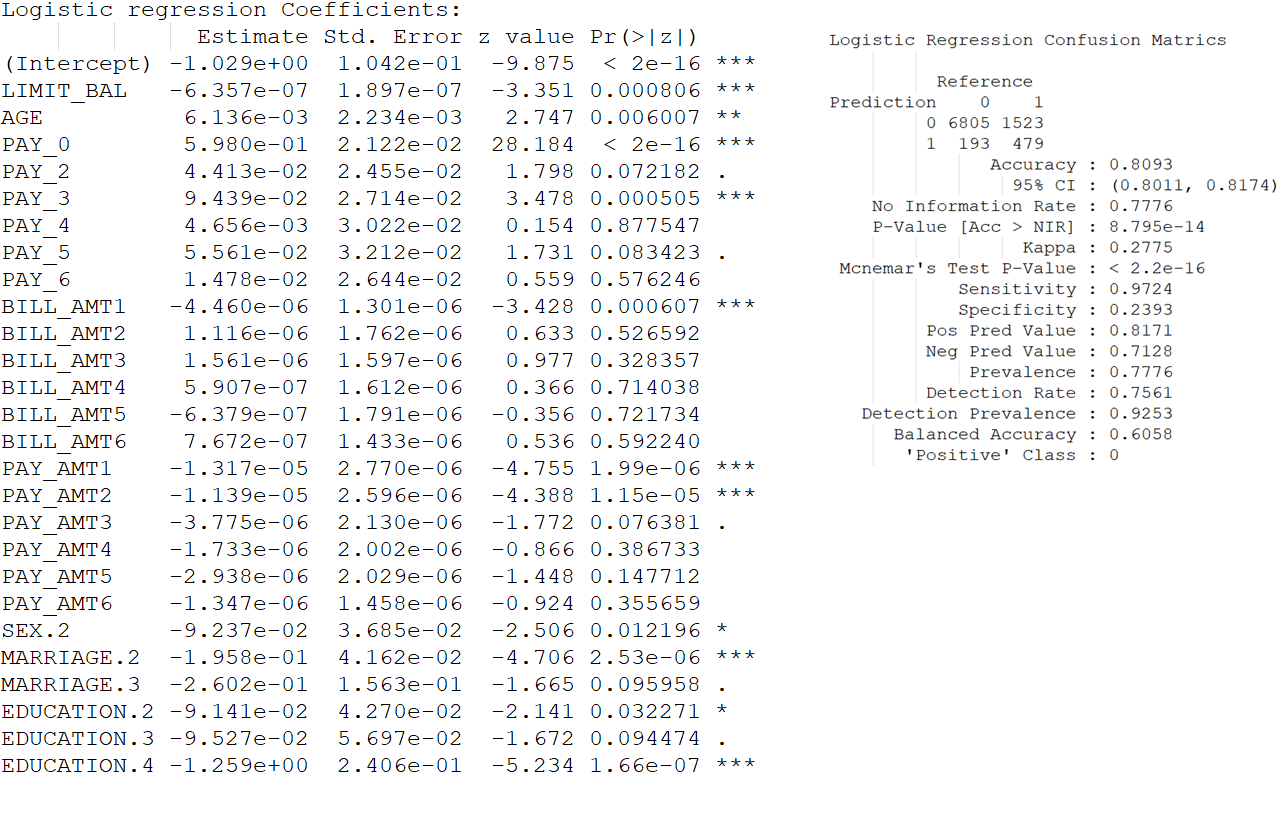
Also, Variable selection using Lasso regression found that it is enough to include only 10 variables to achieve prediction performance similar to Logistic regression.

In all the models The **“PAY\_0”** predictor is found to be the strongest predictor for response variable (default.payment.next.month).

**4.1 Logistics Regression**

In this model, the intercept term is the combination when Gender =1 (Male), Marriage =1 (Married), Education = 1(Graduate school).

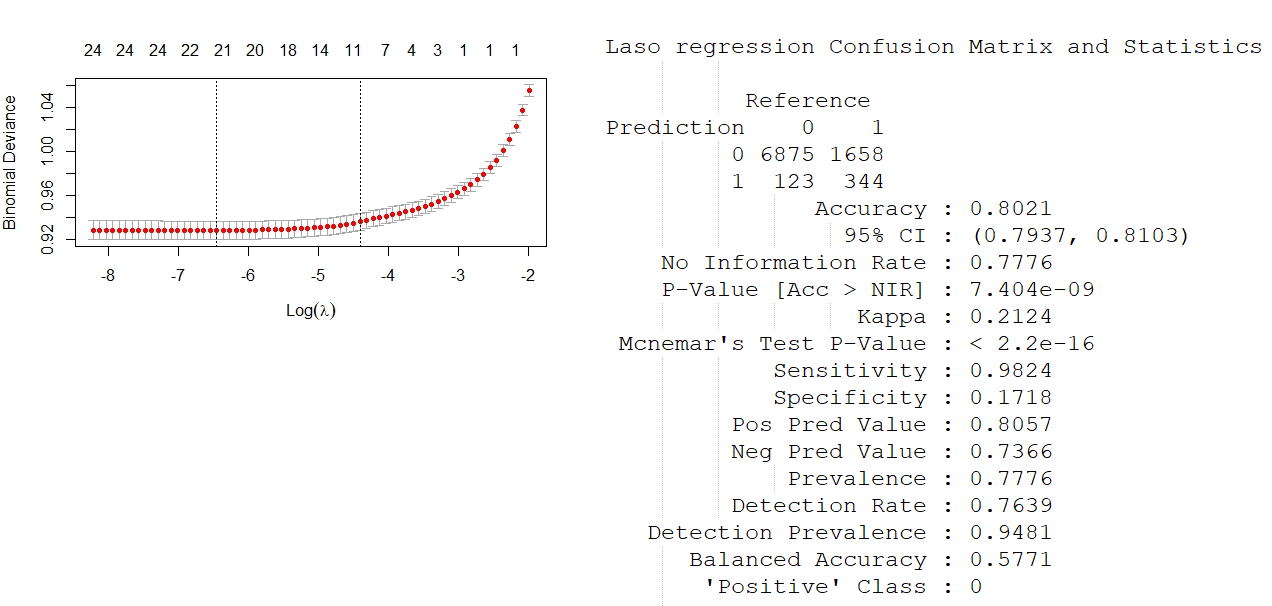
Estimated coefficients and Confusion Matrix and Statistics are (at 0.5 cut-off):



**4.1.1 LASSO Regularization**

Best Lambda ( 0.001566899, Lambda ( at 1 Standard deviation: 0.01213193.

Coefficient estimates at (), No of predictors has reduced to **10** from **26** in

Logistic regression. 

Variables identified in Lasso Regression are:

LIMIT\_BAL, PAY\_0, PAY\_2, PAY\_3, PAY\_5, BILL\_AMT1, PAY\_AMT1, PAY\_AMT2, MARRIAGE.2, EDUCATION.4

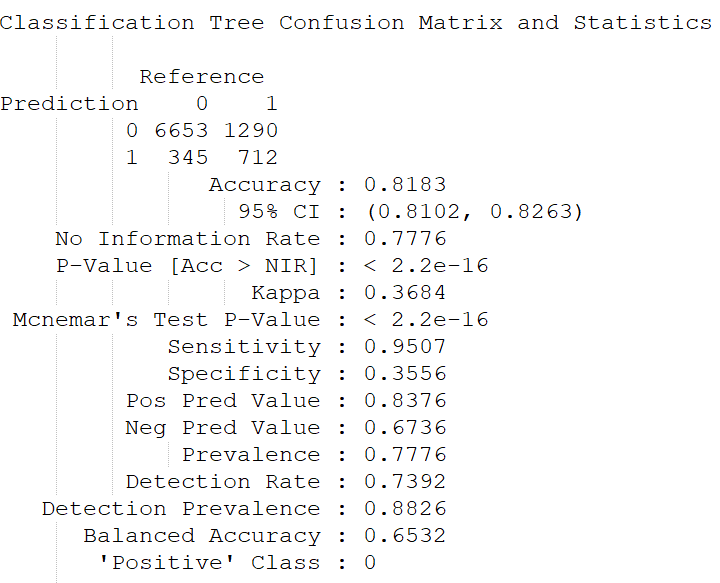
With 10 predictors, the prediction performance of Lasso regression is similar to the prediction performance of Logistic regression.

**4. 2 Classification Tree**

PAY\_0, PAY\_2, PAY\_3, PAY\_5 PAY\_6 LIMIT\_BAL PAY\_AMT6 are the predictor

variables used by the classification tree.

At 82%, the rule **PAY\_0 < 2 & PAY\_2 < 2** has the highest coverage of the sample size and categorized the records as “Non -Default” category.

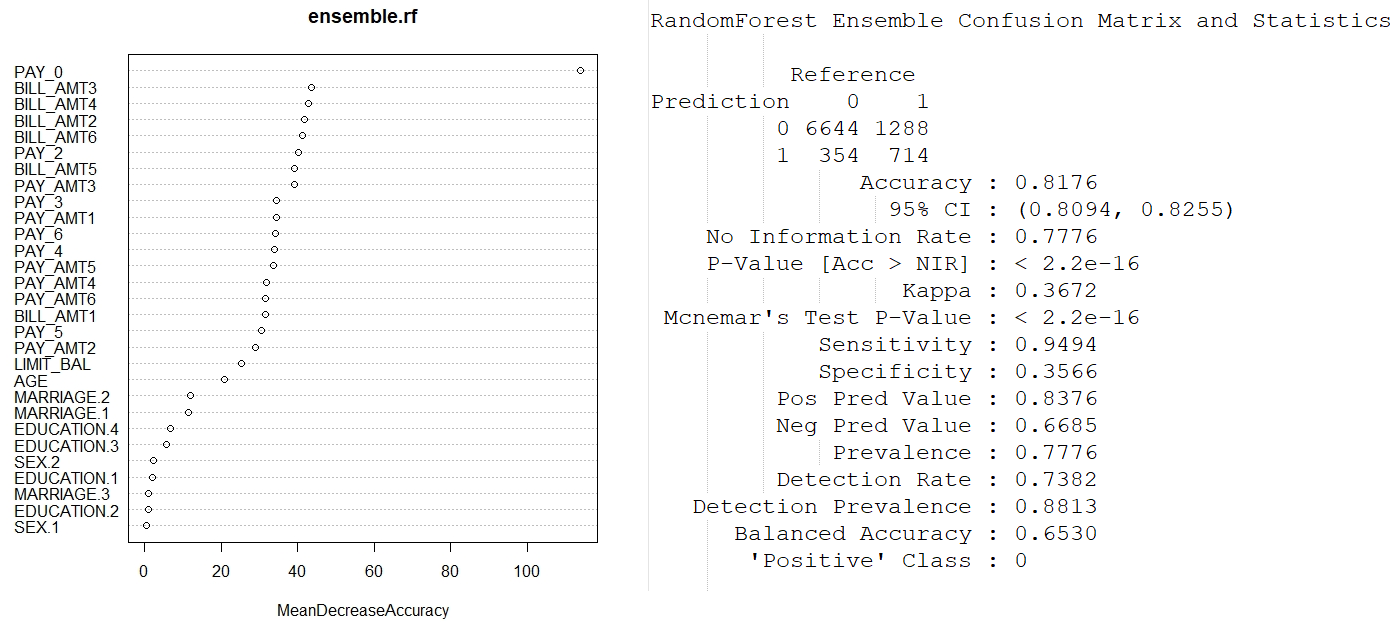


Diagram

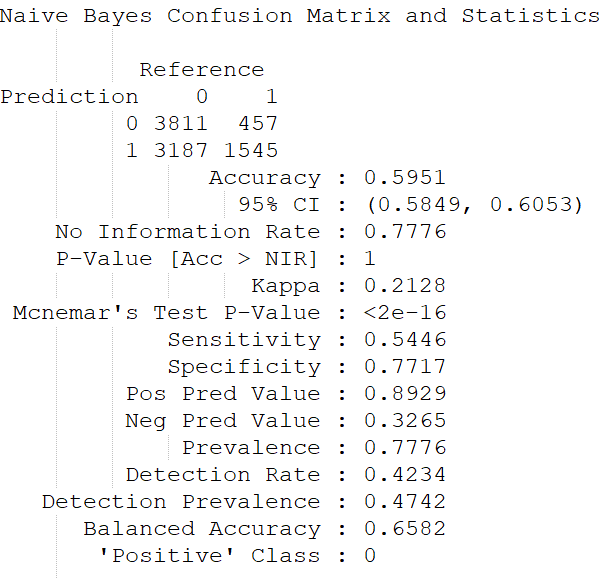
Description automatically generated

**4.2.1 Random Forest Ensemble Method**

The variable importance graph shows that th Pay\_0 variable (Loan Payment status for September 2005 month) as the most importat predictor, with demographic predictors (Age, Sex, Marriage, Education) being the least important predictors.

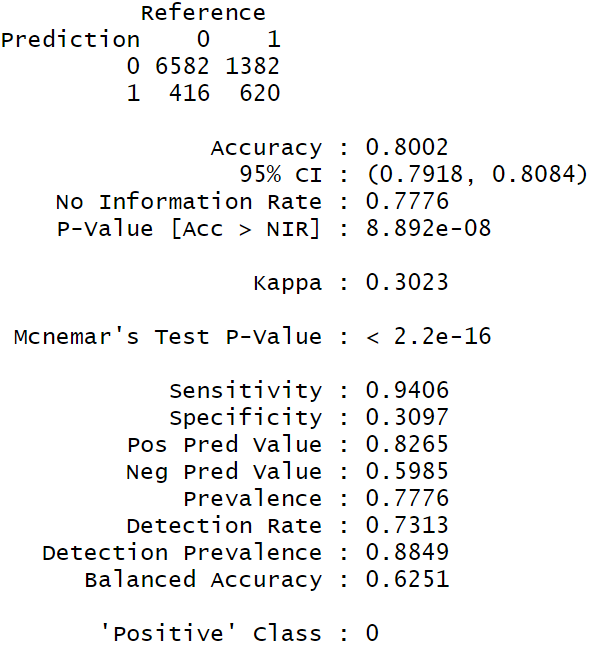


**4.3 Naïve Bayes Classification Method**



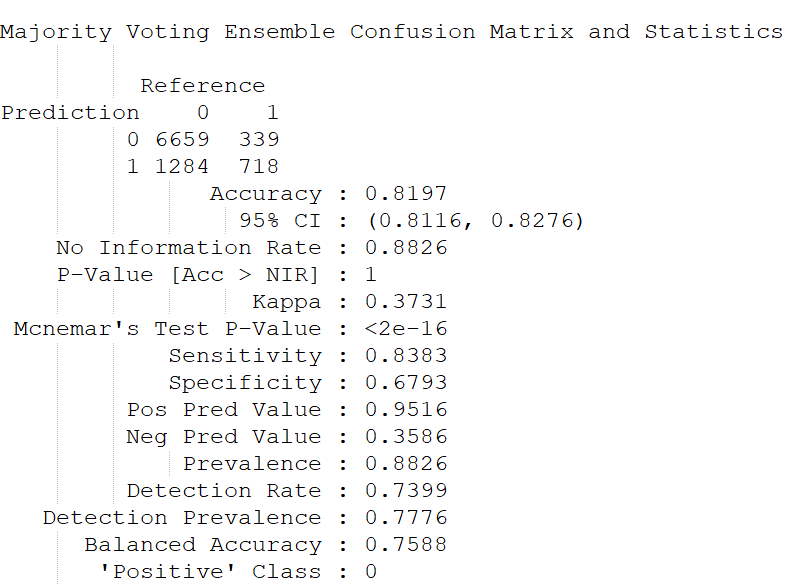
**4.4 KNN Classification**

KNN classification model was fitted with, 10-nearest neighbors’ model with normalization is fitted. The models with 10 – 30 neighbors had almost similar confusion matrix accuracy, so 10 nearest neighbor model is selected.



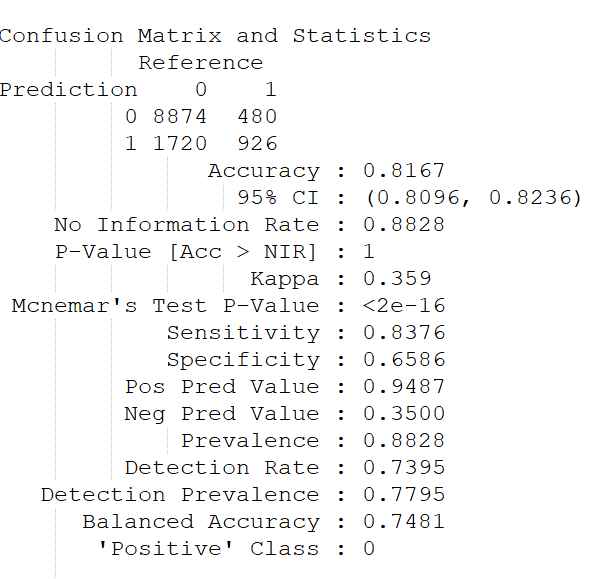
**4.5 Majority Voting Ensemble**

The classifications from Logistic regression, Classification Tree and Naïve Bayes are used for the Majority voting ensemble method.

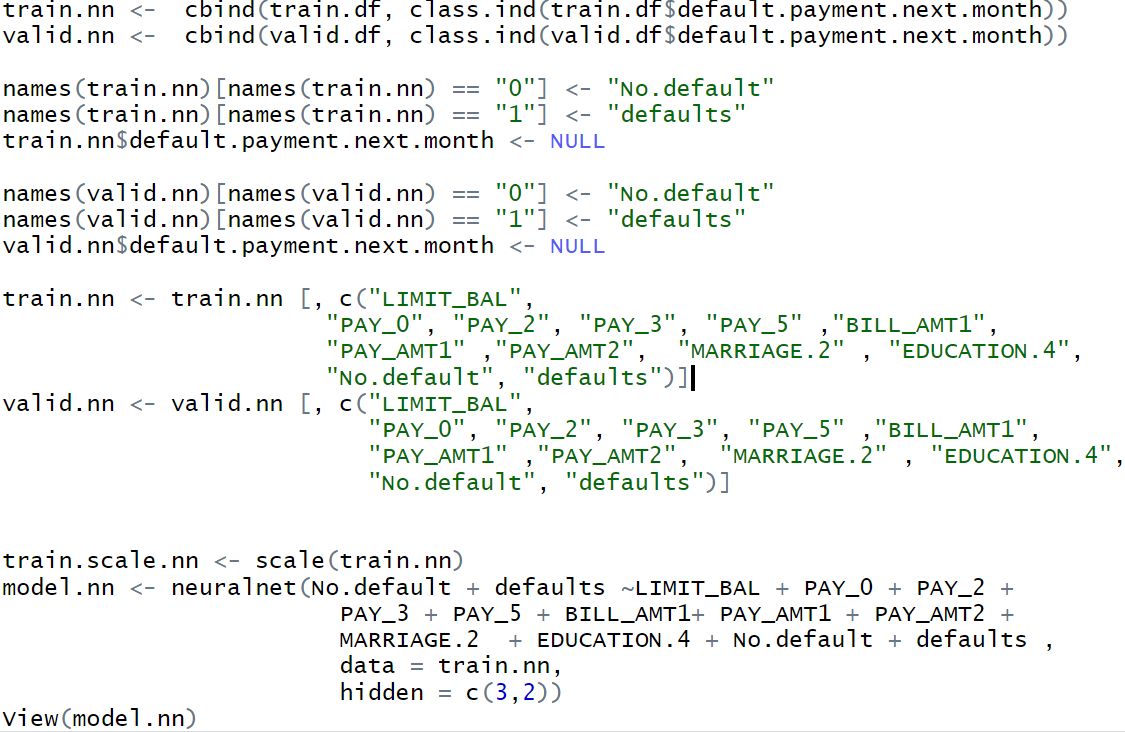


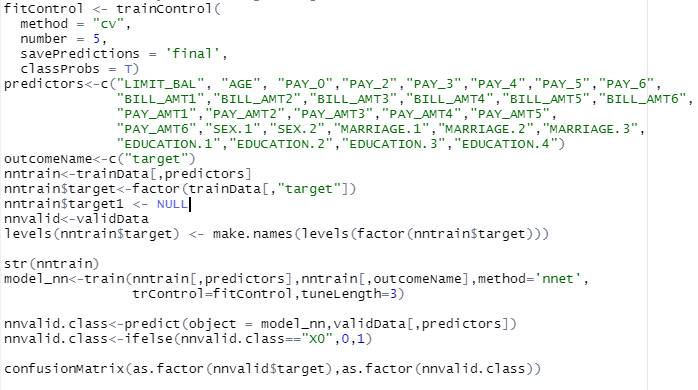
**4.6 Neural Network**

Neural network model using all the predictor variables (Category variables as dummy variables) is fitted using General train function.



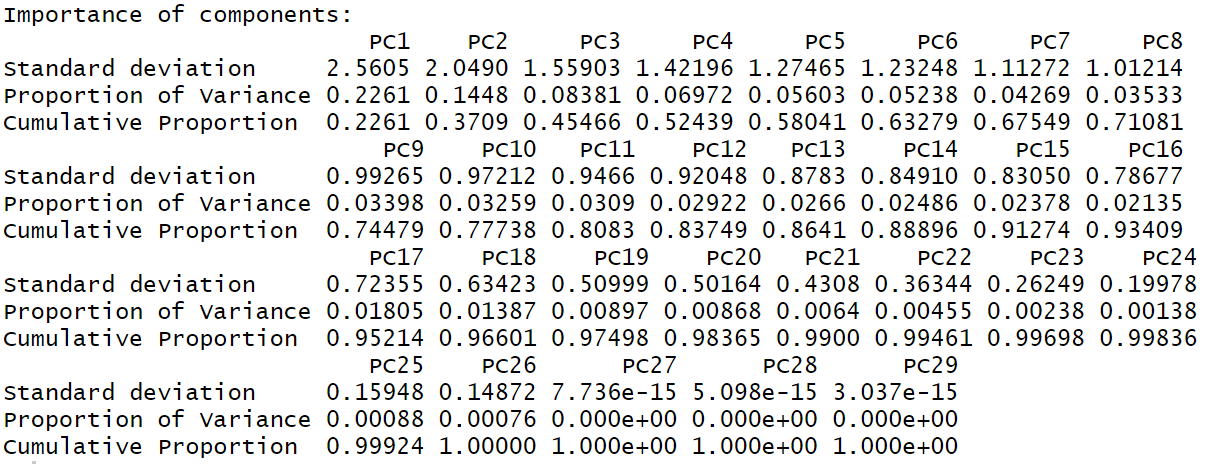
The general train function was used as using neuralnet library implementation is throwing errors.





**5. Dimension reduction using PCA**

The Principal component analysis is used to identify if the predictor variable dimensionality is possible. The Scaled PCA analysis shows that 15 variables contribute to 93% of the variance in predictor variables, hence the dimension reduction is possible.



**6. Results and discussion**

ROC Curve of RandomForest, Logistic Regression, Naïve bayes models:Chart, line chart

Description automatically generated

The RandomForest Roc curve (Red) is the steepest curve and covers the most area compared to other two models (Logistic and Naïve Bayes model).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic  Regression | Naïve  Bayes | Random  Forest | Majority  Voting | KNN (k =10) | Neural  network |
| Accuracy | 0.8093 | 0.5951 | 0.8176 | 0.8197 | 0.8002 | 0.8167 |
| Sensitivity | 0.9724 | 0.5446 | 0.9494 | 0.8383 | 0.9406 | 0.8376 |
| Specificity | 0.2393 | 0.717 | 0.3566 | 0.6793 | 0.3097 | 0.6586 |

In predictors, the PAY\_0 variable is the strongest predictor for responsible, this variable indicates the payment status for Sept 2005 month, which is the last month of observation in data set, this indicates the recent memory bias in marking a card holder as default and non-default.

If the costs associated with wrong prediction is equal in both the cases, then **“Majority voting”** has the best performance.

If identifying a regular paying customer as default customer is costlier to compared to wrongly marking a default customer as non-default customer, the model with high Sensitivity would be a better model, which in this case is a “RandomForest ensemble method”.

If identifying the defaulting customers is more important compared to wrongly identifying a non-default customer as default customer we would want to go with a model that has a high specificity which in this case happens to be a “Naïve Bayes” model, but its accuracy is lowest among all models.

**7. Learnings**

The Lasso regression has shown that models with fewer variables do match in performance with full models, the dimension reduction is also supported by the PCA analysis.

Also, simpler models like Logistic and Majority voting ensemble had equally good performance compared to more complex and computing intensive `Neural network model.

**8. References**

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients