Probability and Statistics (IE 6200) - Sec 7 - Group-6 Final Project Report on Data Analysis of Credit Card Defaulters

Aviral Agrawal, Urjasvit Sinha and Bhavana Joshi

Objective:

To find the correlation between the probability of default commited by credit card users with its associated factors which have been extracted from the database of an important bank in Taiwan with data of 6 months. Quantifying the effects of each associated factor like, credit card user's age, gender, marital status and permited limit balance, on the probability of default payments. The obtained results are then used for extracting meaningful information and patterns of the defaulters. This information can be used by the banks to identify their customers as credible or not credible clients.

Data Description:

The raw data is obtained from UCI Machine Learning Repository. The data set contains 23 variables out of which we've identified 6 main factors to run our analysis to predict the credibility of potential credit card customers.

Refer to the table gievn below for detailed description of the attributes of the dataset and how the data has been modified to suit the analysis in R studio.

Attributes	Values
Sex	Gender (1 = male; 2 = female)
Education	Education level (1 = graduate school; 2 = university; 3 = high school; 4 = others)
Marriage	Marital status (1 = married; 2 = single; 3 = others)
Age	Age (in years)
Limit_Bal	Limit of the credit card
Pay_1 to Pay_6	History of past payment
Bill_amt_1 to Bill_amt_6	Amount of bill statement
Pay_amt_1 to Pay_amt_6	Amount of previous payment
defaulters	Default payment (Yes = 1, No = 0)

The raw data has been filtered to remove outliers in order to obtain normal distribution for some of the variables. All the processing was performed in excel as well as R.

Approach:

Following the traditional steps of data analysis, we started off with cleaning and filtering the data to get rid of unwanted noise or outliers and null values. To do so, we first studied each variable, calculated its statistics like mean, standard deviation and range, and plotted its distribution. We used the log() function

in R in order to normalize our variables. We then categorized the data using different combinations of the variables to find correlation between the target variable- defaulters and rest of the variables. This not only gave an insight of the dataset but also helped in forming a basis for the hypothesis tests to be carried out on the dataset. Based on this, we came up with three hypothesis and carried out Test of Hypothesis. We also calculated confidence intervals.

In addition to covering concepts learned in the lecture/lab, we utilized the 'rcompanion' package to plot the normal histogram for 'limit_bal' variable, 'corrplot' pakcage to plot the correlation plot of the dataset, 'rpart', 'rpart.plot', 'party' packages for plotting the decision tree.

Analysis:

Statistical analysis

We carried out basic statistical analysis of the 'limit_bal' variable to find their measure of central tendency, variablity and symmetry as given below:

Min. 1st Qu. Median Mean 3rd Qu. Max. 4.605 6.215 7.090 6.966 7.696 9.210

skewness:

[1] -0.3805244 attr(,"method") [1] "moment"

kurtosis:

[1] -0.6183416 attr(,"method") [1] "excess"

The 'limit_bal' variable was normalized by dividing by 100. We selected the required columns and categorized the data with different combinations. Then ,created a new table with total people in each category with their respective number of defaulters along with their probabilities as shown below:

```
##     category Total defaulters Probabilty
## 1 category 1  4046        894  0.2209590
## 2 category 2  6201       1271  0.2049669
## 3 category 3   901        227  0.2519423
## 4 category 4  1375        353  0.2567273
## 5 category 5  3356        872  0.2598331
## 6 category 6  4955       1173  0.2367306
## 7 category 7   778        183  0.2352185
## 8 category 8  1076        291  0.2704461
```

Visualization

We then visualized all the data created and all the target variables.

Histogram plot of 'limit_bal' variable before normalization was obtained as shown in Fig.1
The histogram represents the frequency of the individuals having specific credit limit balance per month.

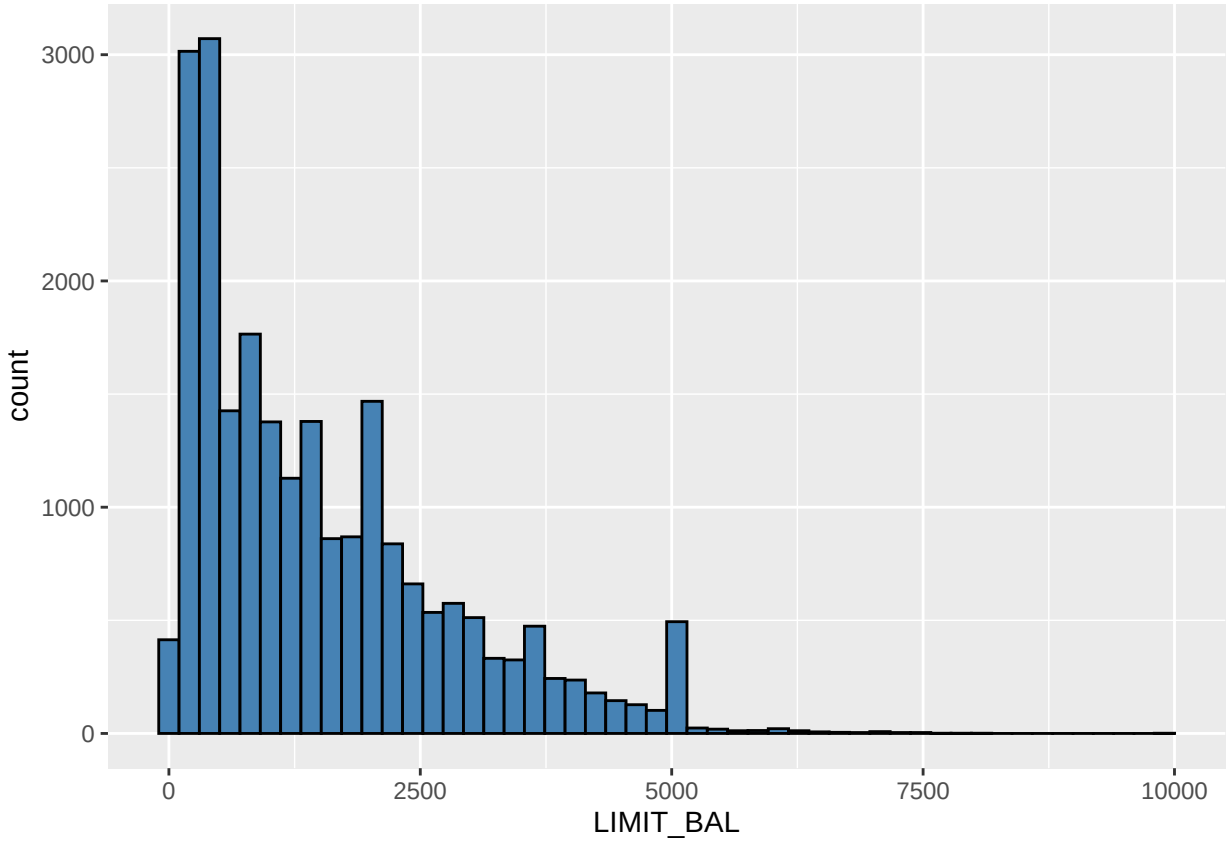

Figure 1: Before normalization

Histogram plot of 'limit_bal' variable after normalization was obtained as shown in Fig.2

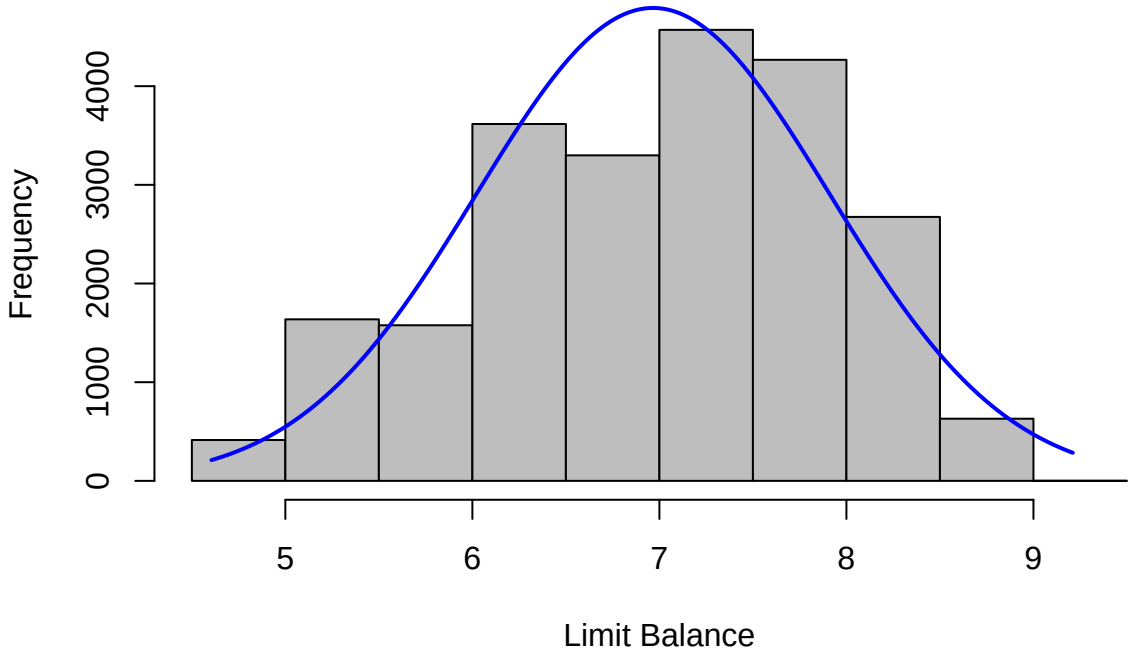

Figure 2: After normalization

Barplot of the categorical data obtained is shown in Fig.3 Here the bar plot is between the 8 categories (combination of gender, education & marriage status) and the limit balance. The plot is subdivided into total number of people falling in each category and the defaulters in them.

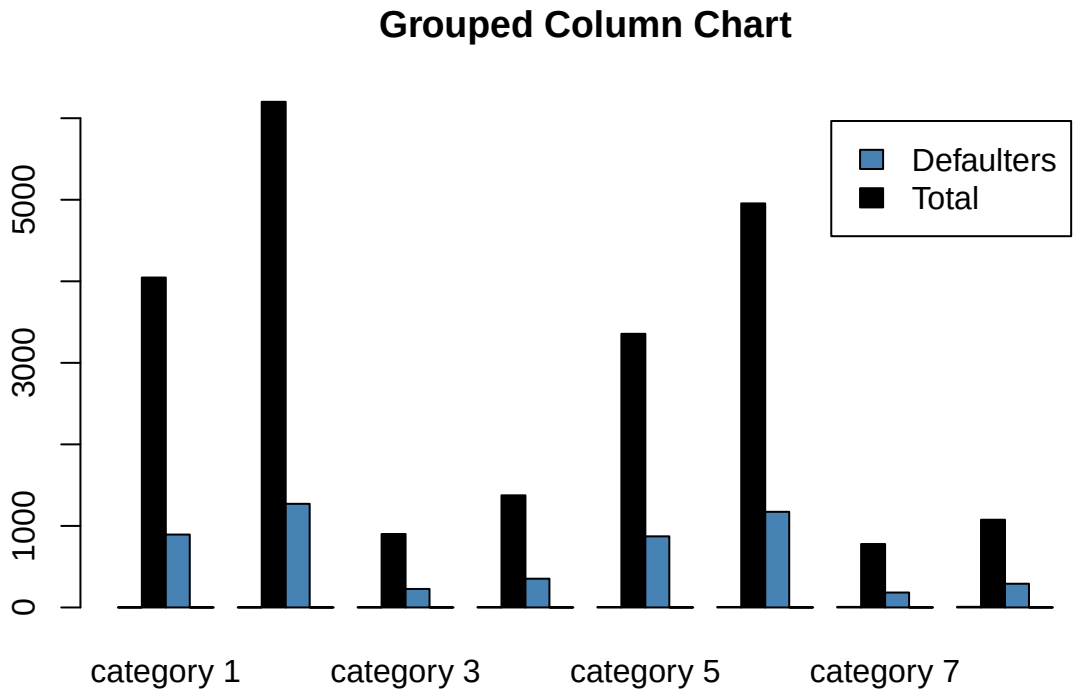

Figure 3: Categorical data

Histogram depicting the probability of each category was obtained as shown in Fig.4.

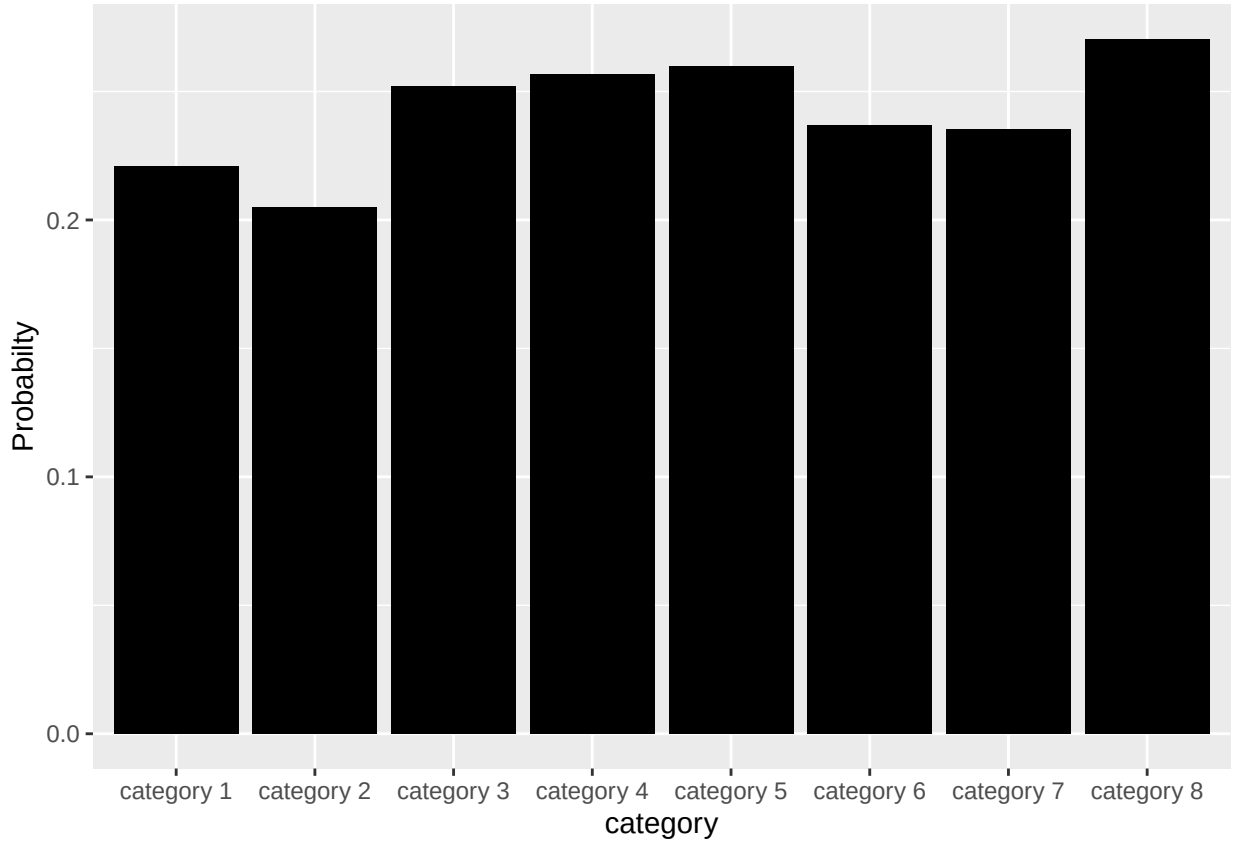

Figure 4: Probability of each category

Cullen and Frey graph of 'limit_bal' variable was obtained as shown in Fig.5 Here the observation is plotted against the kurtosis and square of skewness where all the theoretical distributions are present. The two closest theoretical distributions to the observation point are chosen and goodness of fit test is performed.

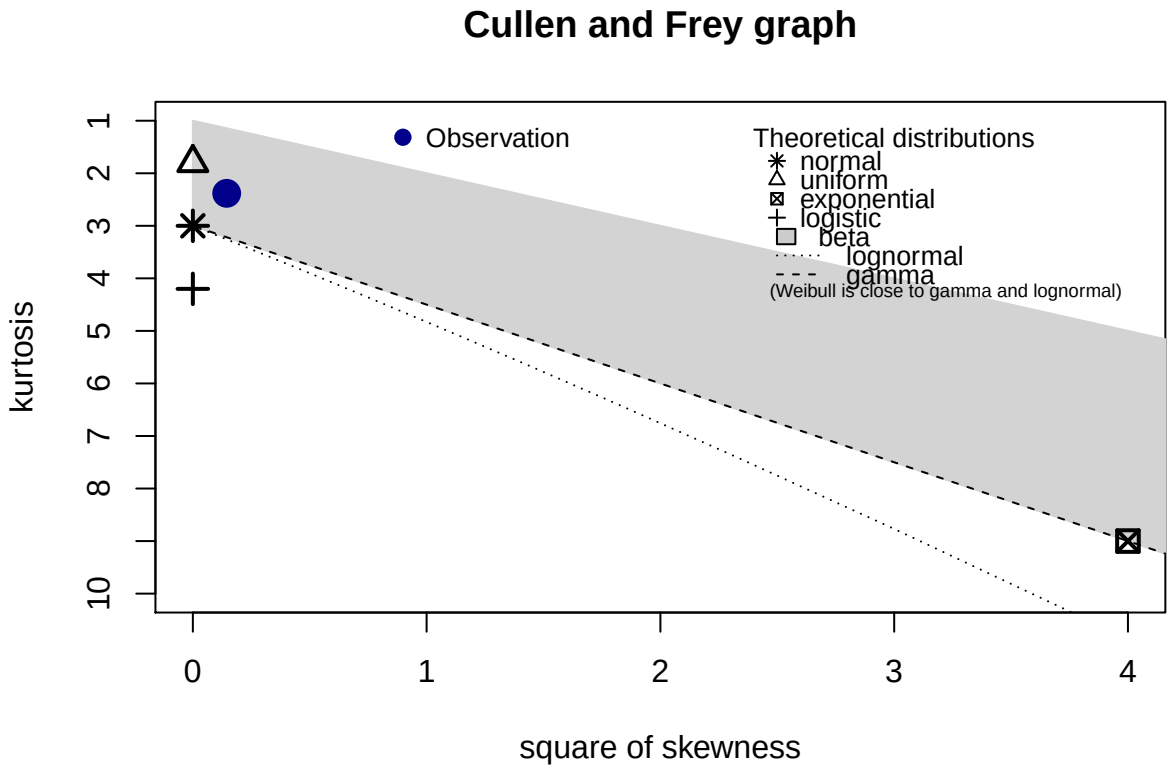

Figure 5: Cullen and Frey graph of 'limit_bal' variable

```
## summary statistics
## ------
## min:  4.60517   max:  9.21034
## median:  7.090077
## mean:  6.966007
## estimated sd:  0.9442891
## estimated skewness:  -0.3805748
## estimated kurtosis:  2.381997
```

We performed the goodness of fit test on the 'limit_bal' variable and plotted the density, cdf, qq and pp plots. The four graphs which are plotted are used to test the goodness of the distribution like how fair the dataset is fitted. Refer to the results below:

Density plot is shown in Fig.6

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate  Std. Error
## mean 6.9660066 0.006268986
## sd   0.9442683 0.004432820
## Loglikelihood:  -30891.84   AIC:  61787.67   BIC:  61803.73
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

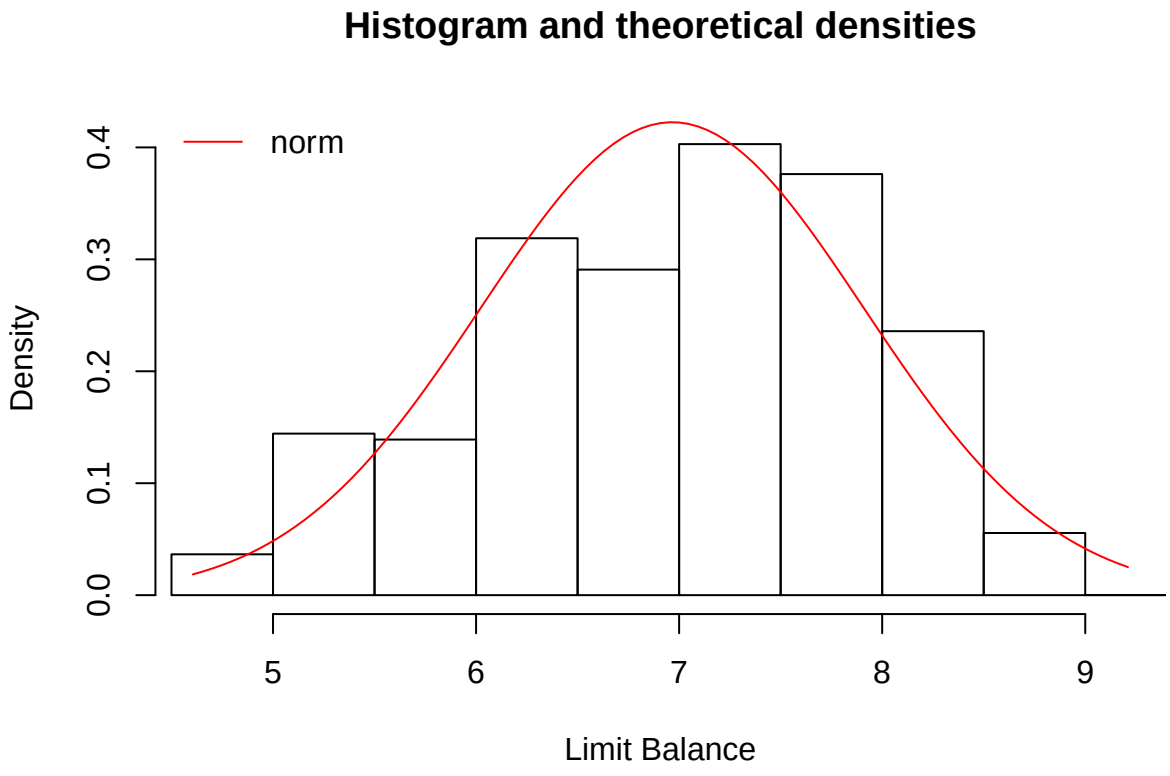

Figure 6: Density plot of 'limit_bal'

CDF plot is shown in Fig.7

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate  Std. Error
## mean 6.9660066 0.006268986
## sd   0.9442683 0.004432820
## Loglikelihood:  -30891.84   AIC:  61787.67   BIC:  61803.73
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

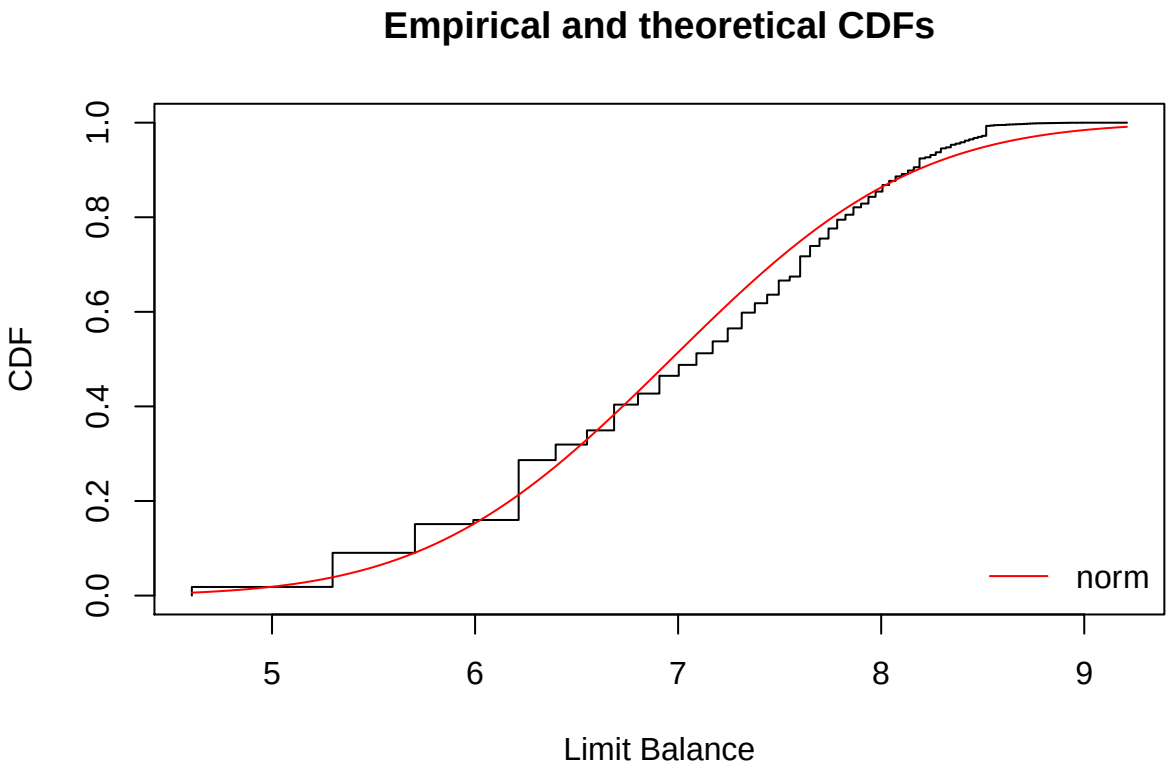

Figure 7: CDF plot of 'limit_bal'

QQ-plot is shown in Fig.8

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate  Std. Error
## mean 6.9660066 0.006268986
## sd   0.9442683 0.004432820
## Loglikelihood:  -30891.84   AIC:  61787.67   BIC:  61803.73
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

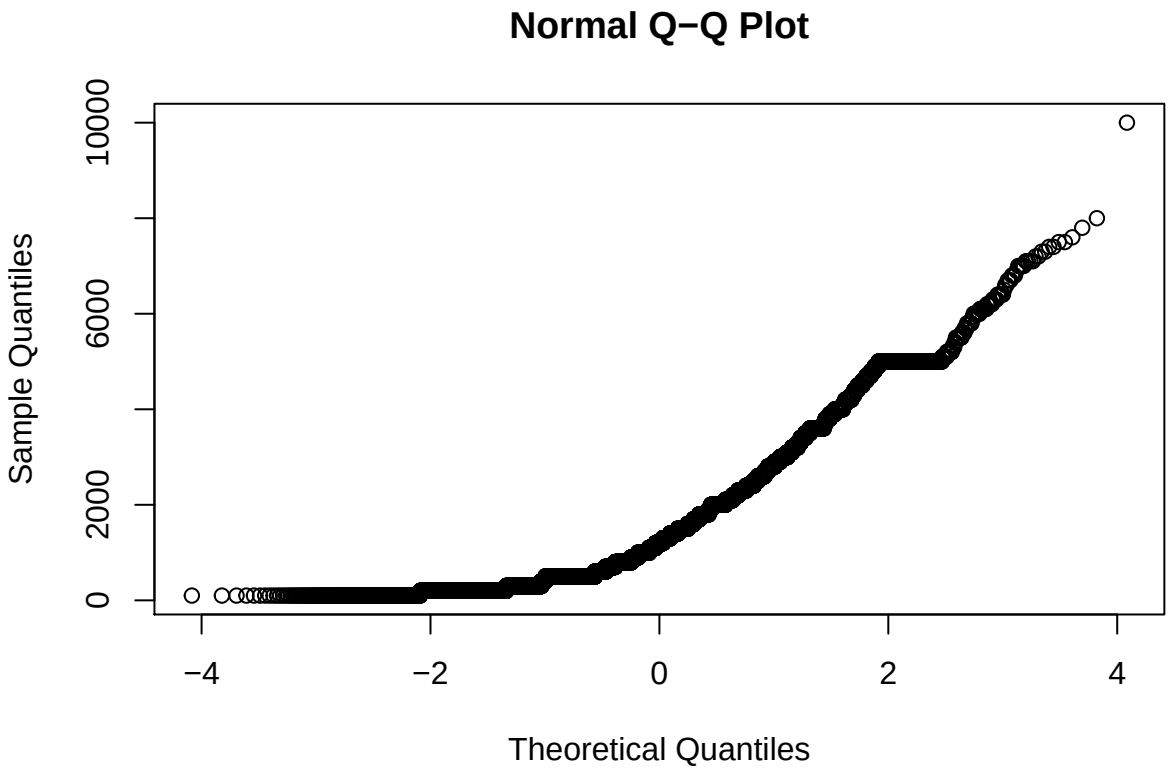

Figure 8: QQ plot of 'limit_bal'

PP-plot is shown in Fig.9

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate  Std. Error
## mean 6.9660066 0.006268986
## sd   0.9442683 0.004432820
## Loglikelihood:  -30891.84   AIC:  61787.67   BIC:  61803.73
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

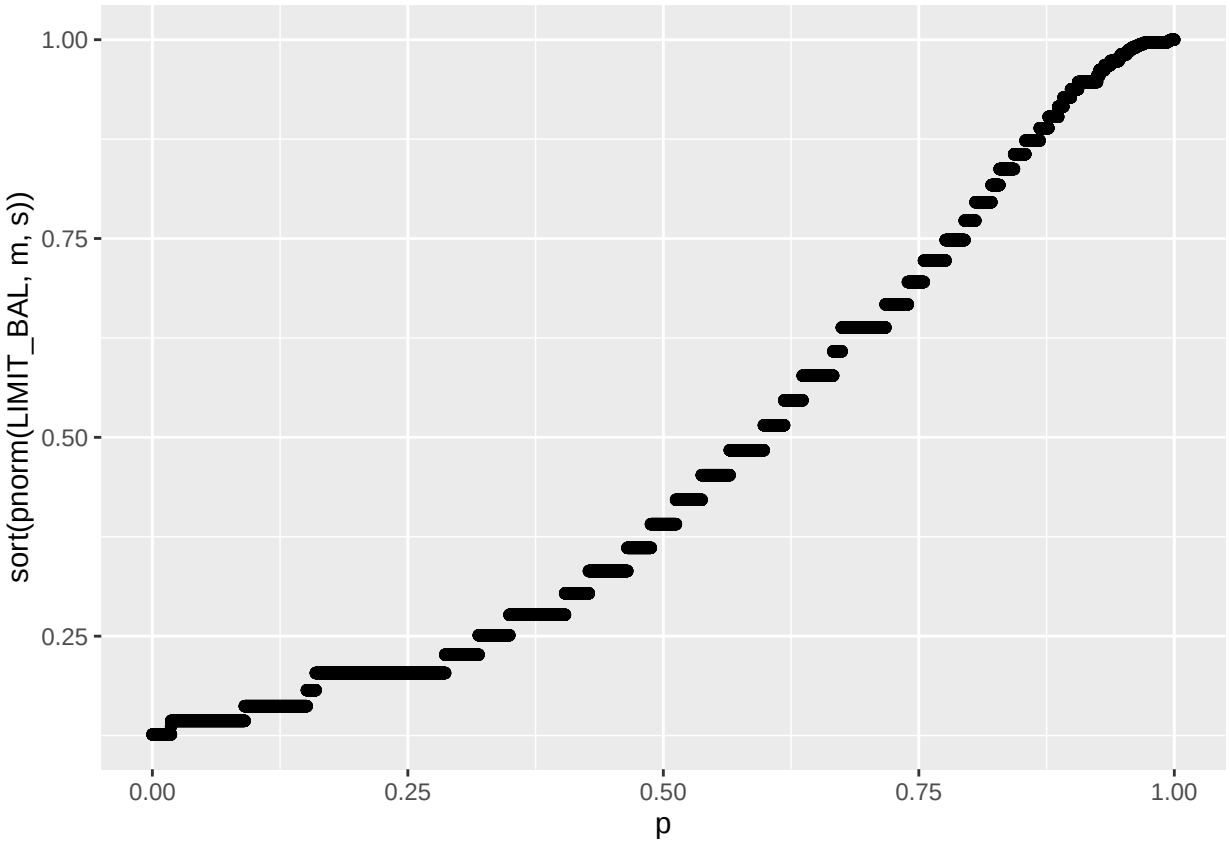

Figure 9: PP plot of 'limit_bal'

Scatter plot of the 'limit_bal' variable was obtained as shown in Fig.10. The three variable scatterplot shows the relation between age, limit balance and marital status of the individual. The plot shows that married individuals aged between 20 to 35 and limit balance lesser than 200000 tend to default more as compared to others in this group.

Figure 10: Scatterplot(Limit bal, age and marriage)

Scatter plot of the 'limit_bal' variable was obtained as shown in Fig.11. This scatterplot is between limit balance and age with sex as third variable. It shows that men aged between 20 to 30 and with limit balance lesser than 200000 have more tendency of doing default as compared to others.

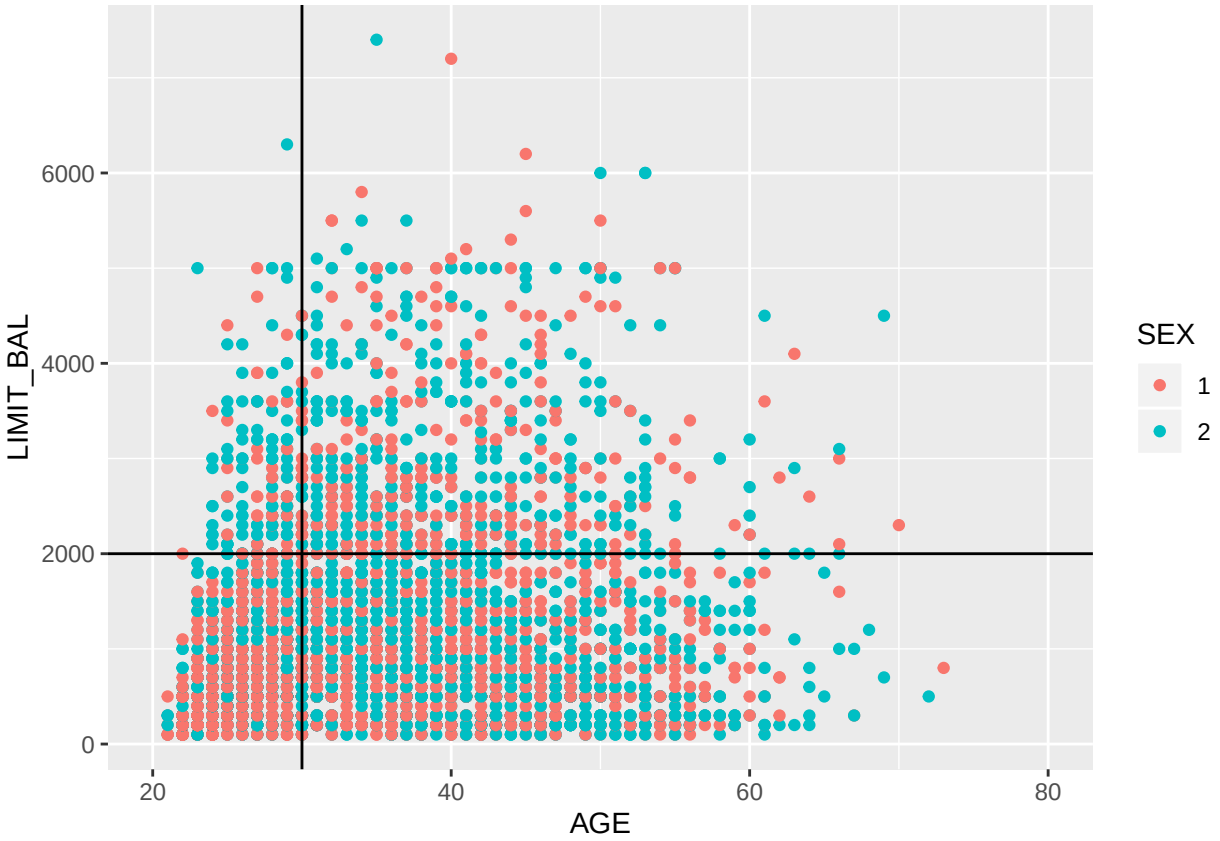

Figure 11: Scatterplot(Limit bal, age and sex)

Boxplot comparing male defaulters who are highly educated with the male defaulters with less education is shown in Fig.12

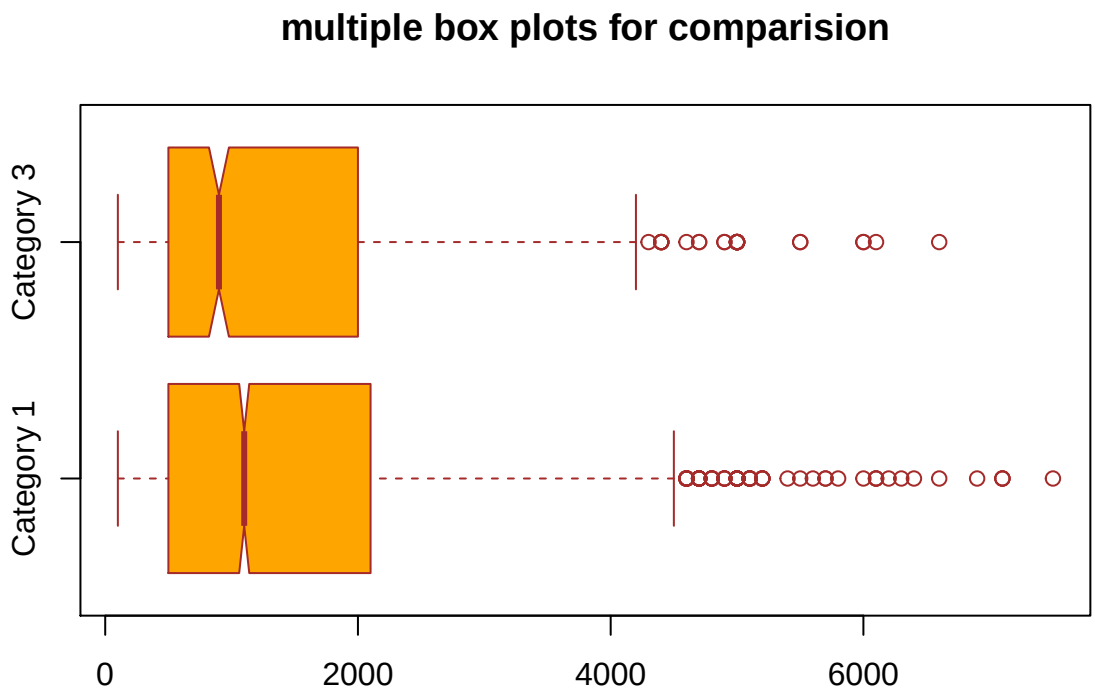

Figure 12: Boxplot comparing category 1 and 3

Correlation plot of the entire dataset is shown in Fig.13. The correlation plot shows the correlation of each column. It uses correlation matrix for plotting. The blue circles represent positive correlation whereas the red ones represent negative correlation.

```
##              SEX EDUCATION MARRIAGE   AGE LIMIT_BAL PAY_1 PAY_2 PAY_3 PAY_4
## SEX         1.00      0.02     0.01  0.00      0.00 -0.02 -0.05 -0.02 -0.02
## EDUCATION   0.02      1.00     0.00  0.00     -0.04  0.09  0.10  0.07  0.07
## MARRIAGE    0.01      0.00     1.00 -0.41     -0.10  0.00  0.00  0.01  0.01
## AGE         0.00      0.00    -0.41  1.00      0.12 -0.02 -0.02 -0.03 -0.02
## LIMIT_BAL   0.00     -0.04    -0.10  0.12      1.00 -0.23 -0.24 -0.24 -0.23
## PAY_1      -0.02      0.09     0.00 -0.02     -0.23  1.00  0.73  0.59  0.54
##            PAY_5 PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5
## SEX        -0.02 -0.02     -0.02     -0.02     -0.02     -0.02     -0.02
## EDUCATION   0.06  0.06      0.05      0.04      0.05      0.05      0.04
## MARRIAGE    0.01  0.02     -0.04     -0.04     -0.04     -0.04     -0.04
## AGE        -0.03 -0.03      0.08      0.08      0.08      0.08      0.08
## LIMIT_BAL  -0.21 -0.22      0.42      0.41      0.41      0.42      0.42
## PAY_1       0.51  0.47      0.12      0.12      0.12      0.12      0.13
##            BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6
## SEX            -0.02     0.00     0.00     0.00     0.00     0.00     0.00
## EDUCATION       0.04    -0.03    -0.01    -0.01     0.00    -0.01    -0.01
## MARRIAGE       -0.04     0.00     0.00     0.00    -0.01     0.00    -0.01
## AGE             0.07     0.03     0.03     0.03     0.02     0.03     0.02
## LIMIT_BAL       0.41     0.23     0.23     0.24     0.24     0.25     0.25
## PAY_1           0.13    -0.10    -0.09    -0.08    -0.08    -0.06    -0.06
##            defaulters
## SEX             -0.01
## EDUCATION        0.03
## MARRIAGE        -0.03
## AGE              0.01
## LIMIT_BAL       -0.18
## PAY_1            0.39
```

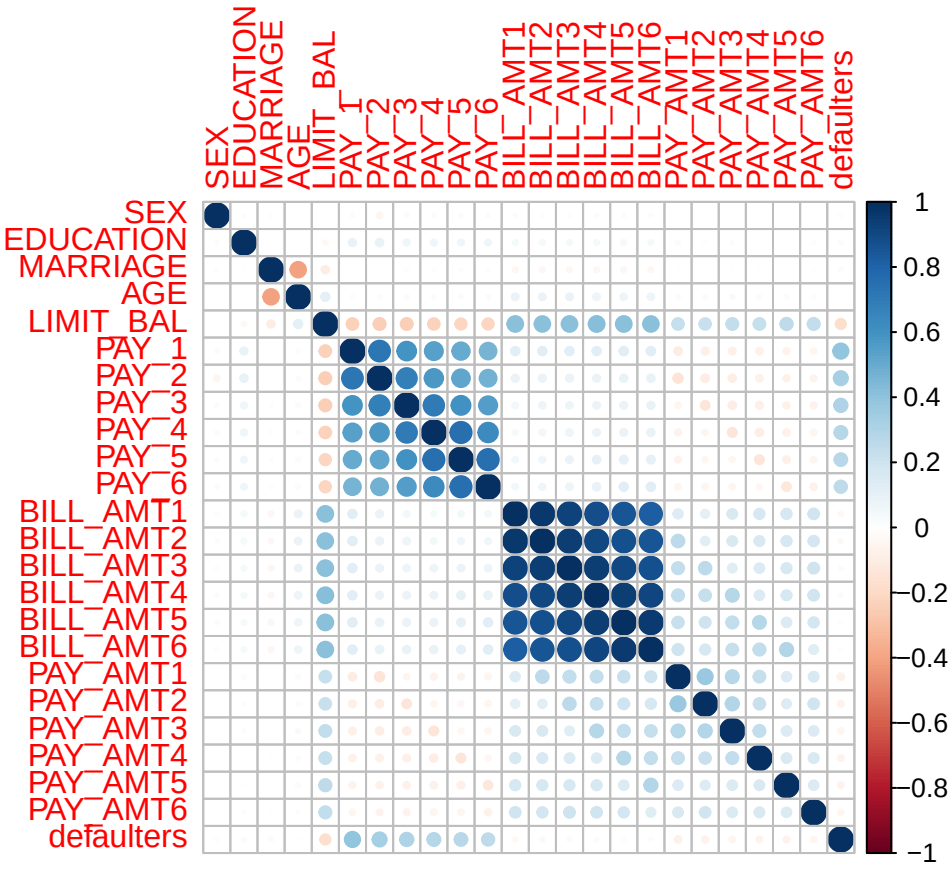

Figure 13: Correlation plot

Hypothesis Testing

We tested for three different hypothesis stated as follows:

Hypothesis 1

We are going to test if unmarried men with higher level of education default more.

Hypothesis:

H0: p1 = p0

H1: p1 != p0

Population default percentage, p0:

[1] 0.03940409

sample percentage, p1:

[1] 0.042

sample size, n:

[1] 5000

Since n is greater than 30, we applied z test.

We consider 95% confidence interval. Hence, its z value is 1.96

Z score:

[1] -0.5192789

The z score does not fall in the rejection region, hence we fail to reject the null hypothesis. So we conclude that unmarried men with higher education level default more.

Hypothesis 2

We are going to compare the limit balances given for male and female customers and test for who is given a greater limit balance. We shall use two sample left tailed test.

Hypothesis:

H0: mu1 - mu2 >= 0

H1: mu1 - mu2 < 0

Male sample size: 2000

Male sample mean: 1599.95

Male sample standard deviation: 1282.129

Female population mean: mu1

Female sample size: 3000

Female sample mean: 1510.988

Female sample standard deviation: 1260.555

Female population mean: mu2

We consider 95% confidence interval. Hence, its z value is 1.96

Z score:

[1] 2.129849

The z score does not fall in the rejection region, hence we fail to reject the null hypothesis. So we conclude that male customers are given greater limit balance as compared to female customers.

Constructing confidence interval

We computed a 95% confidence interval for the difference between the proportions of male defaulters who are married and have higher level of education from two samples.

We first calculated the error and then found the lower and upper limits of the confidence interval.

The 95% confidence interval obtained is : [1] -0.0597 0.0116

Advanced Analytics using Decision Tree

Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables.

In our analysis, we have used classification trees to identify the trend of the credit card defaulters based on their alloted limit balance, age, sex and marital status.

Here is the plot of the decision tree shown in Fig.14. The decision tree given above classifies the defaulters by partitioning the columns (age, sex, limit balance, education) using set of rules. the classification provides the best split using the specified columns.

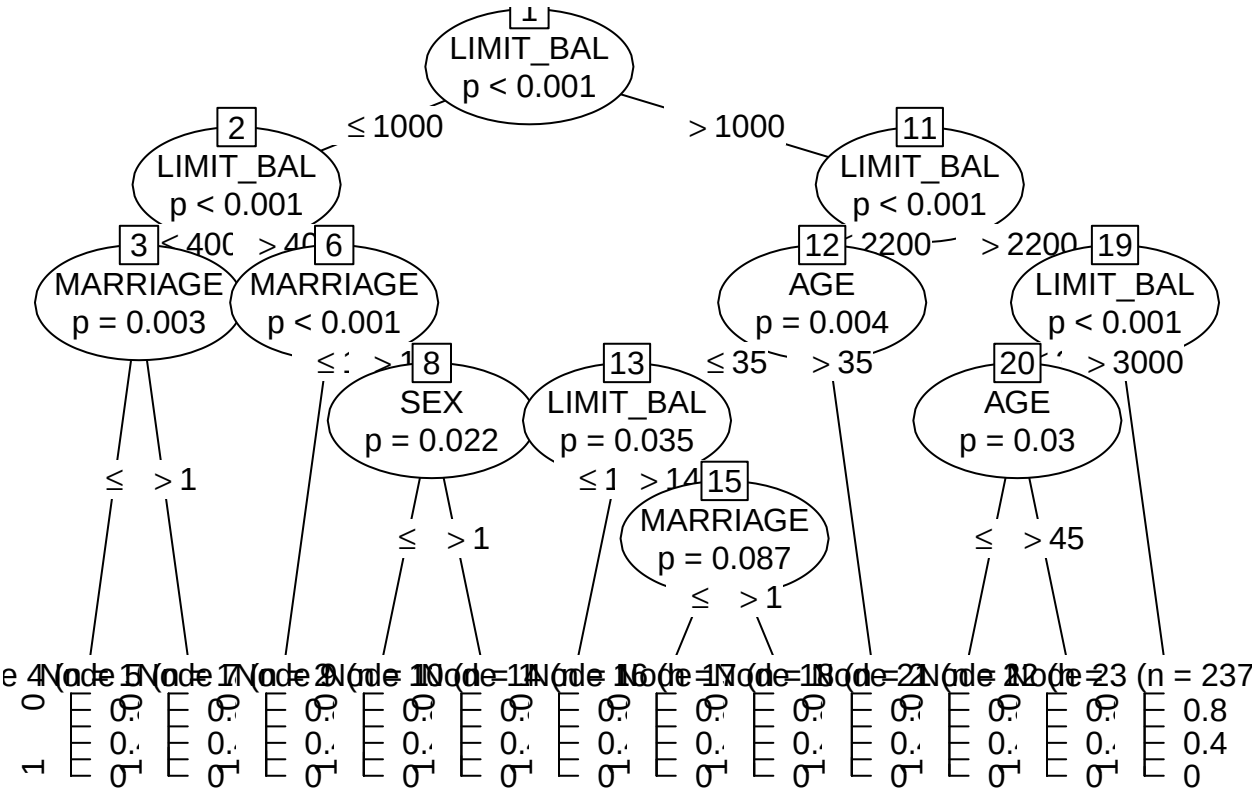

Figure 14: Decision Tree plot

1. Misclassification error in training data:

The error table for training data was obtained as follows:

	0	1
0	13943	4247
1	0	0

Hence the misclassification error is 0.2334799 i.e. the predictions made differ from the reality by 23.34%

2. Misclassification error in validation data:

The error table for validation data was obtained as follows:

testpred	0	1
0	3481	1017
1	0	0

Hence the misclassification error is 0.2261005 i.e the predictions made differ from the reality by 22.61%