"WILL THE CREDIT CARD DEFAULT?"

Supervised by, Dr. Arnab Bhattacharya

CS685A: Course Project

Submitted by,

Rajdeep Saha – 201380 – rajdeep20@iitk.ac.in

Sagnik Dey – 201397 – dsagnik20@iitk.ac.in

Shuvam Gupta – 201421 – shuvamgpta20@iitk.ac.in

Soumik Karmakar – 201428 – soumik20@iitk.ac.in

Soumyadip Sarkar – 201431 – soumyadip20@iitk.ac.in

ACKNOWLEDGEMENT

We take this opportunity to express our sincere gratitude and deep regards to our project guide *Dr. Arnab Bhattacharya* for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. His blessings, help, time to time guidance will carry us a long way in the journey of life which we are about to embark.

We would also like to thank the institute authorities for giving us the chance to do the project and for providing the environment and necessary facilities required for the completion of our project. We would also like to thank PK Kelkar Library for having provided various reference books related to our project.

Finally, we would like to extend our gratitude and thanks to our parents. Without their constant support and encouragement, it would not have been possible for us to proceed with our effort.

INDEX

1. Introduction & Data Description:

Missing credit card payments once or twice does not count as a default. A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months. In our project, we shall carry out a brief diagrammatic analysis and Machine Learning algorithmic analysis of credit card default data.

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.

There are 25 variables:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT 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_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)
- PAY_2: Repayment status in August 2005 (scale same as above)
- PAY_3: Repayment status in July 2005 (scale same as above)
- PAY_4: Repayment status in June 2005 (scale same as above)
- PAY_5: Repayment status in May 2005 (scale same as above)
- PAY_6: Repayment status in April 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

Here, we have total 30000 data points, each datapoint consists of an ID column, 23 explanatory variables and 1 response variable. Here, 'default.payment.next.month' is considered as response variable.

Out of 23 explanatory variables, we have 9 categorical variables, namely, 'SEX', 'EDUCATION', 'MARRIAGE', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6' and 14 numeric variables, namely, 'LIMIT_BAL', 'AGE', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6'.

2. Missing Value Check:

Before carrying out any analysis, we would wish to check if our dataset contains any missing values and we have found the following result, -

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

Hence, we can say that our data doesn't contain any missing values and we can safely proceed to carry out our further analysis.

3. Information Extraction and Visualization:

• Age Wise Analysis:

First, we have divided the dataset according to age groups, having age-range 20-29, 30-39, … and obtained the following figure,

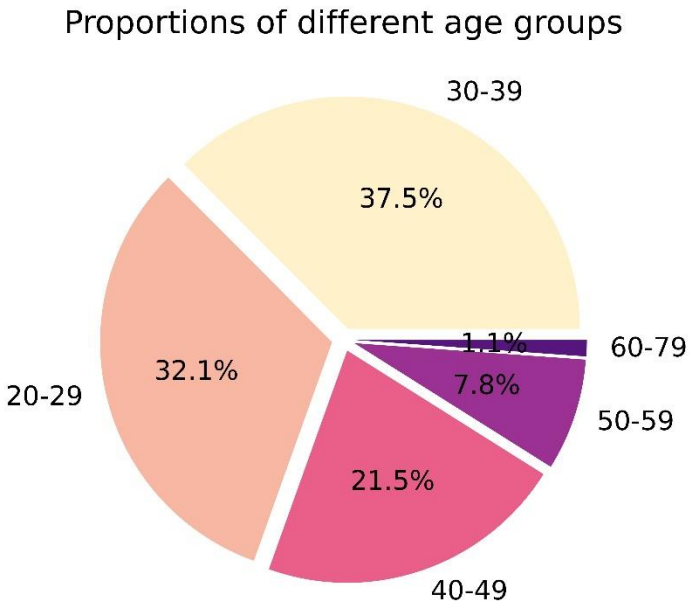

Fig 1: Proportions of Different Age Groups

We can see that, data of the people with age between 30 and 39 is the highest (37.5%) that was collected in the survey and data of the people with age between 60 and 79 is the lowest (1.1%) that was collected in the survey.

After having obtained the proportions of different age group, we now wish to check, how the number of defaulters vary across these groups. For this we can use **Multiple Bar Diagram**. Also, from the graph, we can see that, whether there is some significant difference between the number of defaulters and non-defaulters across various age groups.

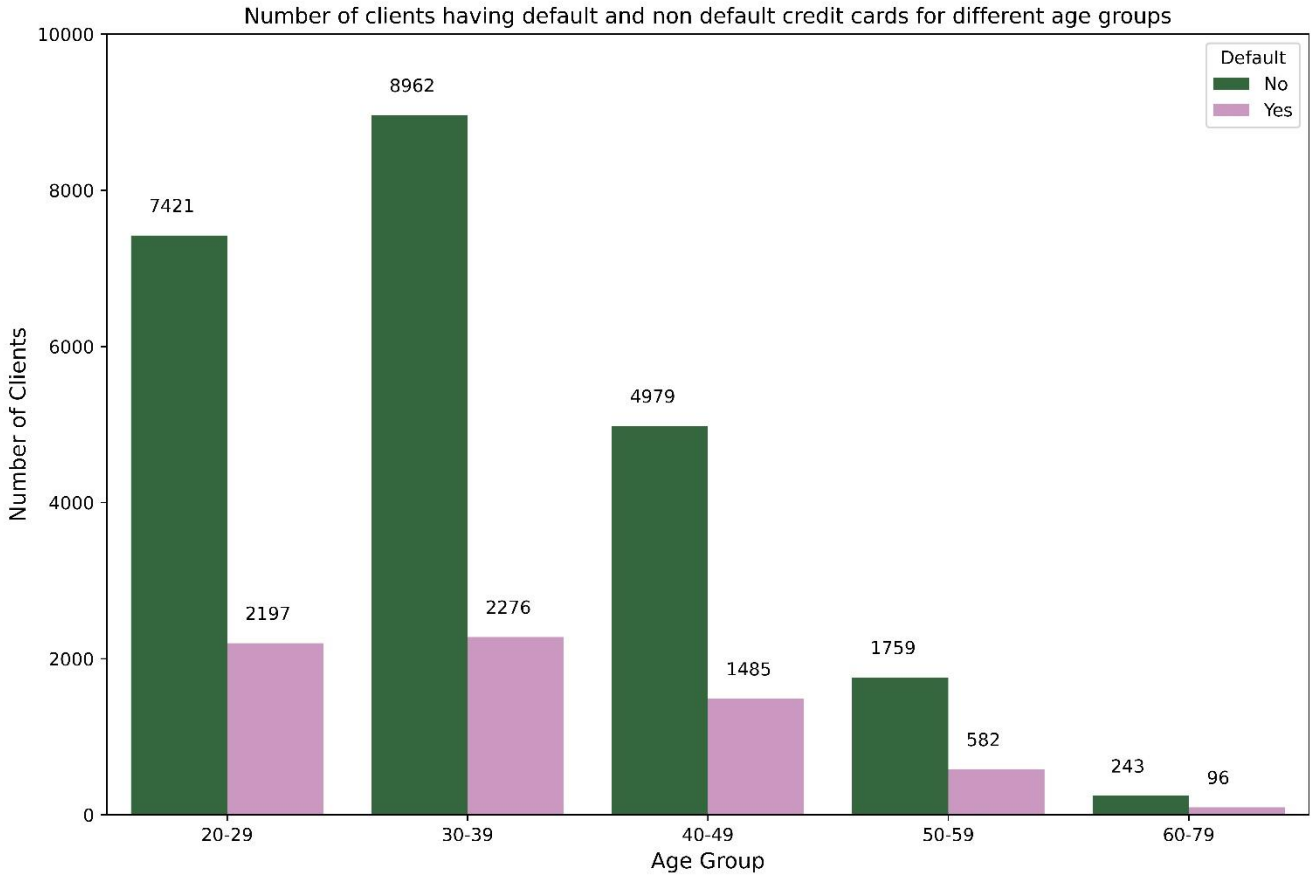

Fig 2: Distribution of the number of defaults and non- defaults across various age groups

From the figure, we can see that, across all the age groups, age group 30-39 has maximum number of defaulters (8962), age group 60-79 has the lowest number of defaulters (243). Across all the age groups, age group 30-39 has maximum number of non-defaulters (2276), age group 60-79 has the lowest number of non-defaulters (96)

Also, from the graph, we can conclude that, for all the age groups, the number of non-defaulters is significantly lesser than the number of defaulters.

Now, we shall explore the percentage of defaulters across various age groups, i.e., we will see each group contains how much percentage of total defaulters.

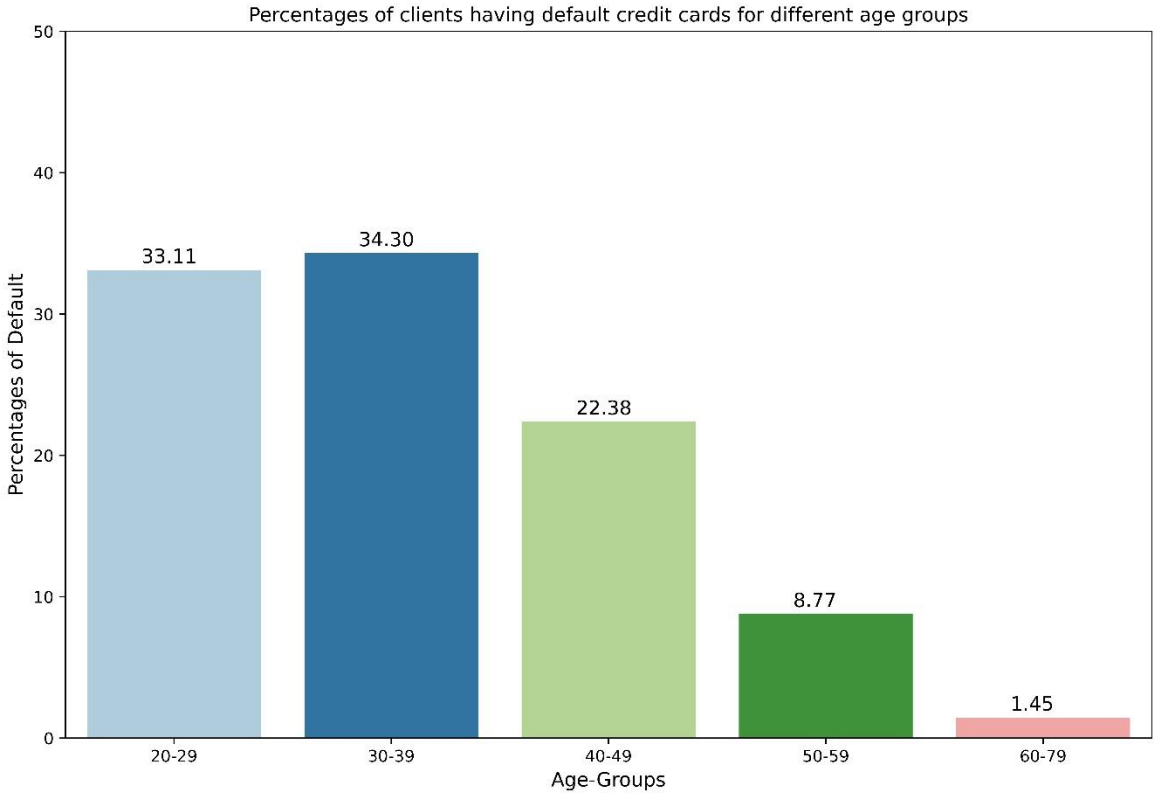

Fig 3: Percentage of defaulters in each age group

From the above graph, we can see that, among all the defaulters for all the age group, age group 30-39 has the highest percentage of defaulters (34.30%) and age group 60-79 has the lowest percentage of defaulters (1.45%) .

From the above graph, we can conclude that, the distribution of the percentage of defaulters out of total number of defaulters across all the age groups is right tailed, i.e., the distribution is positively skewed.

Now, we shall explore the proportion of defaulters oin each age group, ie, we shall calculate $\frac{no.\ of\ defaulters}{total\ no.\ of\ clients}$ for each age group.

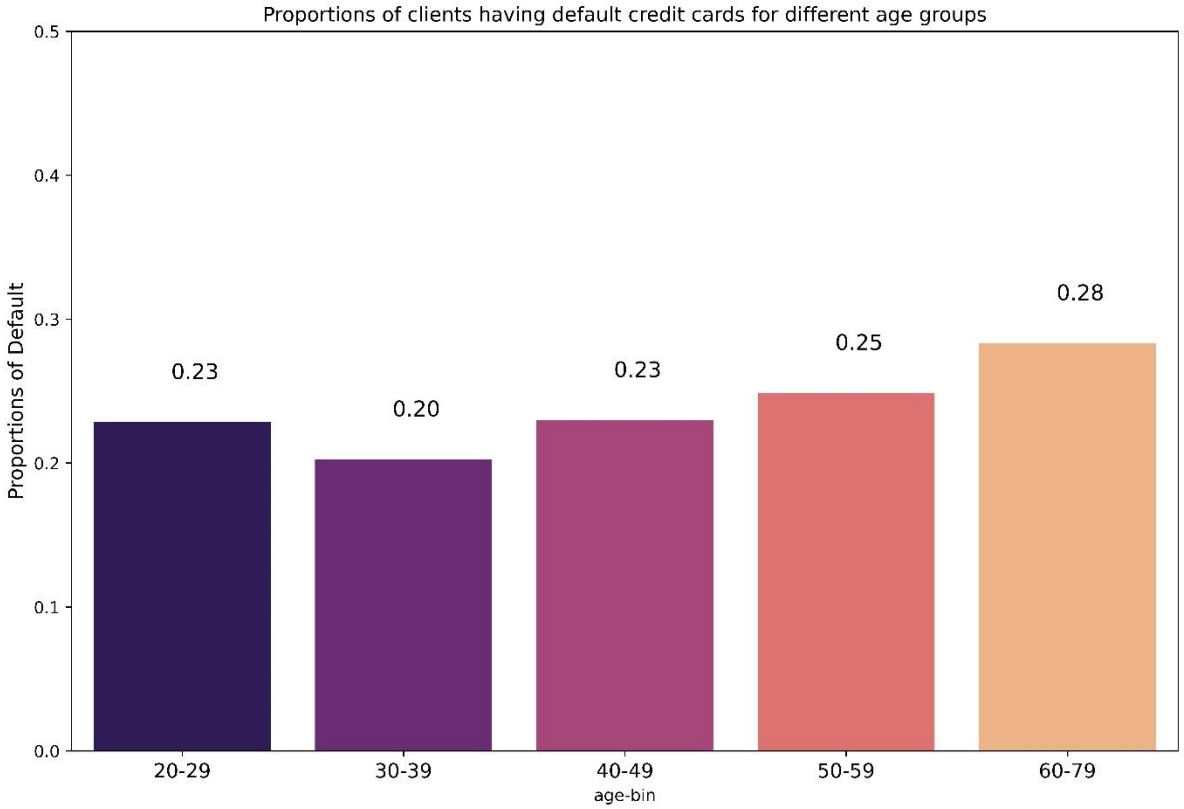

Fig 4: Proportion of defaulters in each age group

From the graph, we can see that, in each group more less 20-25% person's credit card defaults, with a maximum 28% defaulters for age group 60-79. So, we can conclude that, for all the age groups the percentage of defaulters is quite low and roughly we can estimate that out of 100 people, 24 people's credit card default.

• Gender Wise Analysis:

Now, we shall explore the data according to genders. First, we will see, how the number of defaulters and non-defaulters vary across different genders.

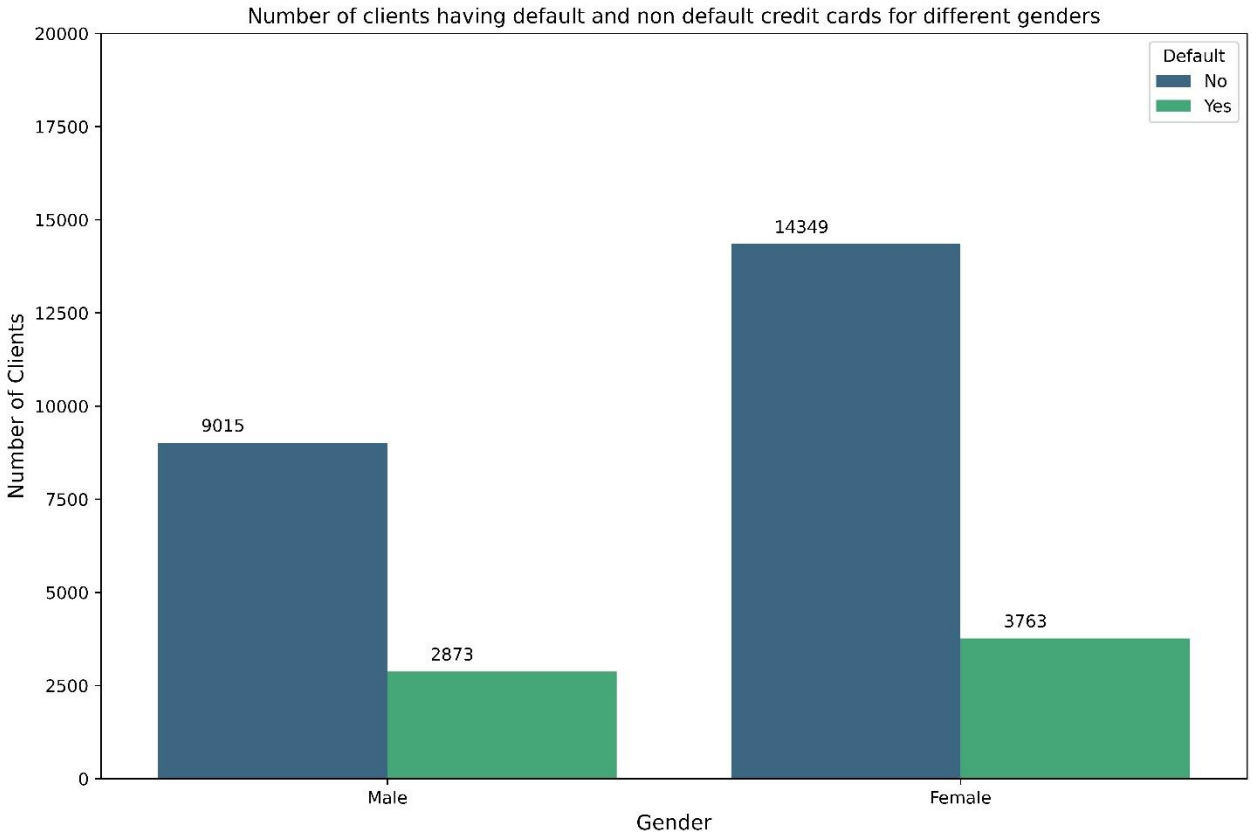

Fig 5: Number of Defaulters and Non-Defaulters for different genders

From the graph, we can see that, for both the genders, numbers of defaulters are much lower than number of non- defaulters. The number of male defaulters is higher than the number of female defaulters and the number of male non-defaulters is higher than the number of female non-defaulters.

Now, we shall see the proportion of defaulters for different genders, i.e., we will see for each gender what is the proportion of the people whose credit card defaults.

Now, we shall explore the percentage of defaulters across males and females, i.e., we will see each sex contains how much percentage of total defaulters.

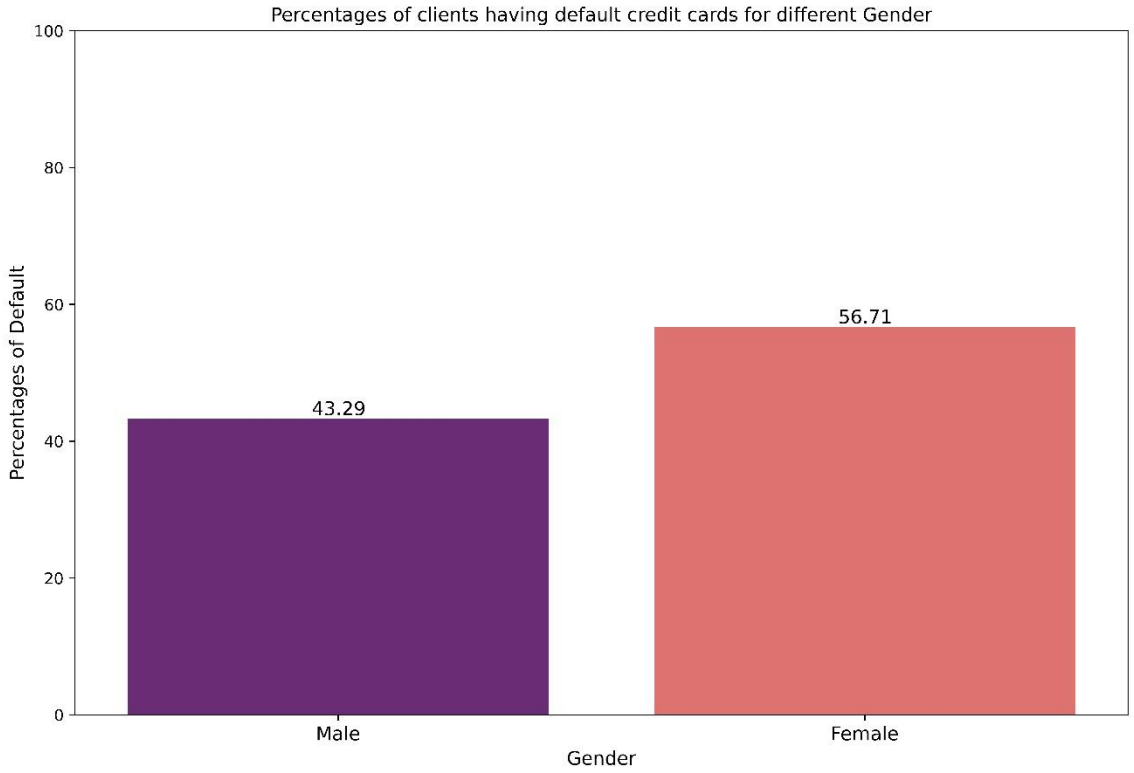

Fig 6: Percentage of defaulters in each age group

From the above graph, we can see that, among all the clients who is a credit-card defaulter, 43.29% are male and 56.71% are female, i.e., among all the persons with credit card default, total number of males with credit card default is lower than total number of females with credit card default.

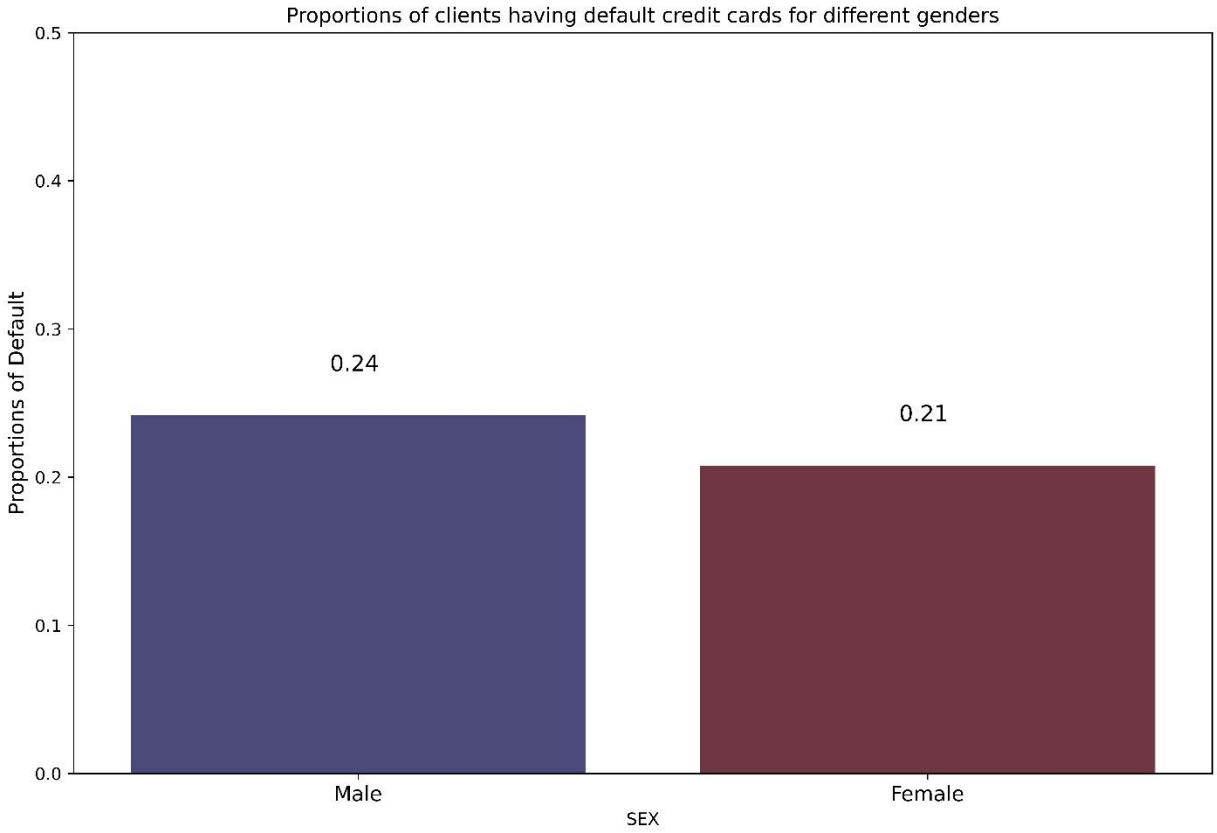

Fig 7: Proportion of defaulters for males and females

From the above figure we can see that, for male, out of 100 people, 24 are credit card defaulters and for female, out of 100 people, 21 people are credit card defaulters. Proportion of credit card defaulters for males is slightly higher than that of females.

• Education Wise Analysis:

Now, we shall explore the data according to various levels of 'Education'. 'Education' is a categorical variable in this dataset, which consists of 5 categories namely, 'Grad School', 'University', 'High School', 'Others' and 'Unknown'. First, we will see, how the number of defaulters and non-defaulters vary across different levels of education.

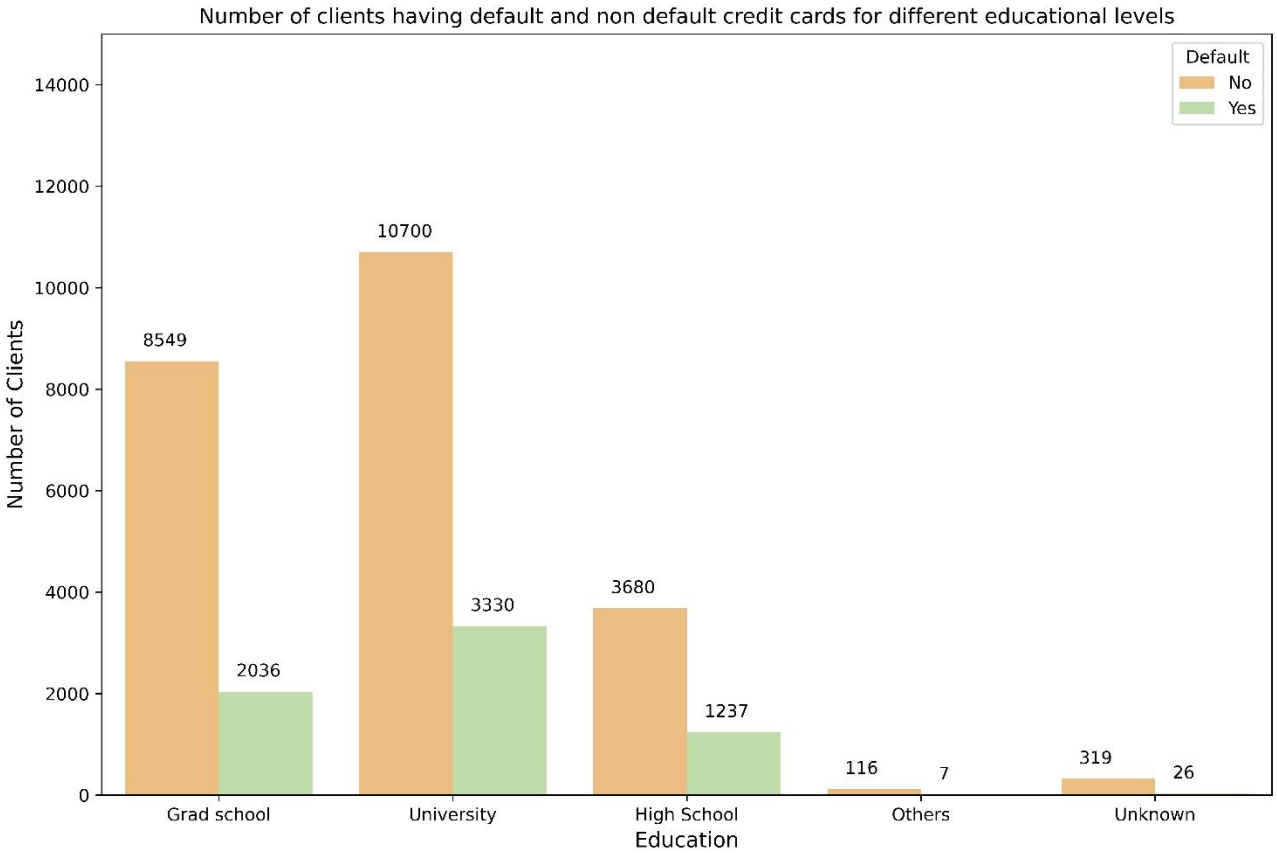

Fig 8: Number of Defaulters and Non-Defaulters for different levels of Education

From the above graph, we can see that most of the people who are credit card defaulters, belong to the education level either 'Grad School' or 'University'. Out of these two categories, the number of people who are credit card defaulters and belong to 'Grad School' is greater than the number of people who are credit card defaulters and belong to 'University'. The number of people who are credit card defaulters and belong to education level 'Others' can be considered as negligible.

Now, we shall see the proportion of defaulters for different education levels, i.e., we will see for each education levels what is the proportion of the people whose credit card defaults.

Now, we shall explore the percentage of defaulters across different education levels, i.e., we will see each education level contains how much percentage of total defaulters.

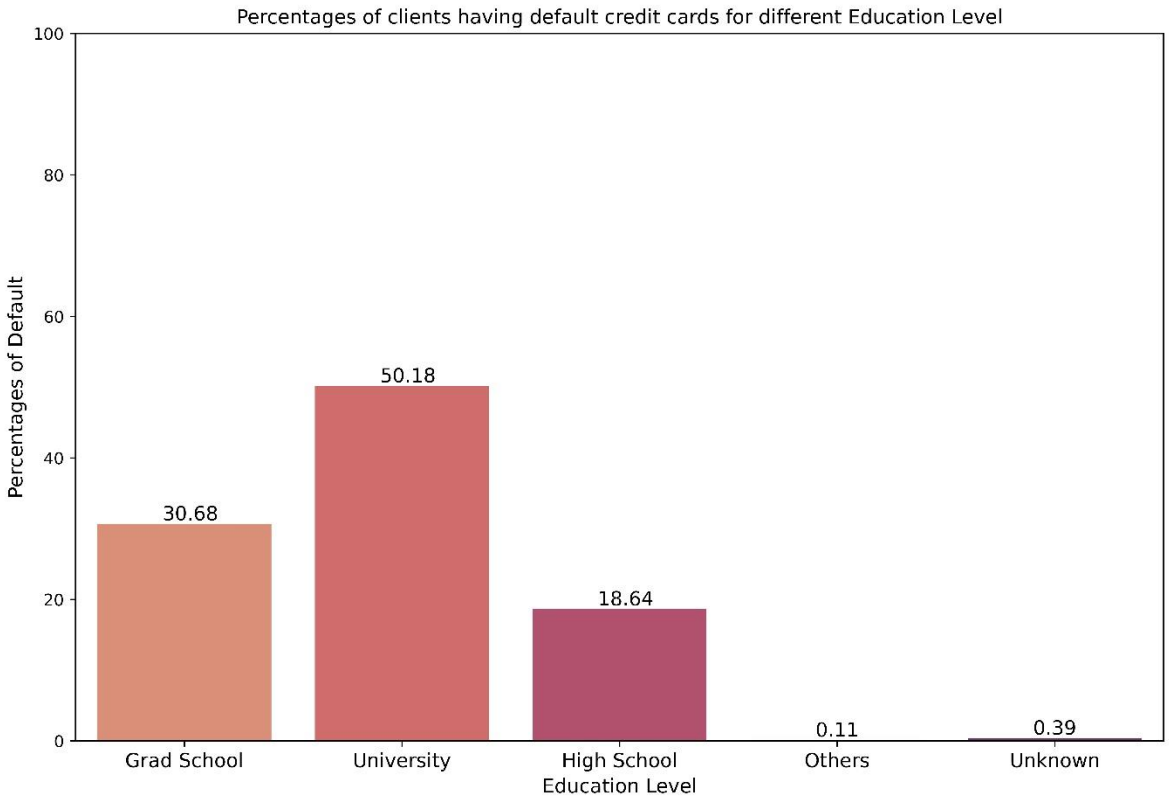

Fig 9: Percentage of Defaulters among total number of defaulters for different levels of Education

From the above figure, we can see that, 50.18% of the defaulters belong to the education level 'University' which is the highest and 30.68% of the defaulters belong to the education level 'Grad School' which is the second highest. Percentage of defaulters belonging to education level ' Others' and 'Unknown' can be considered as negligible.

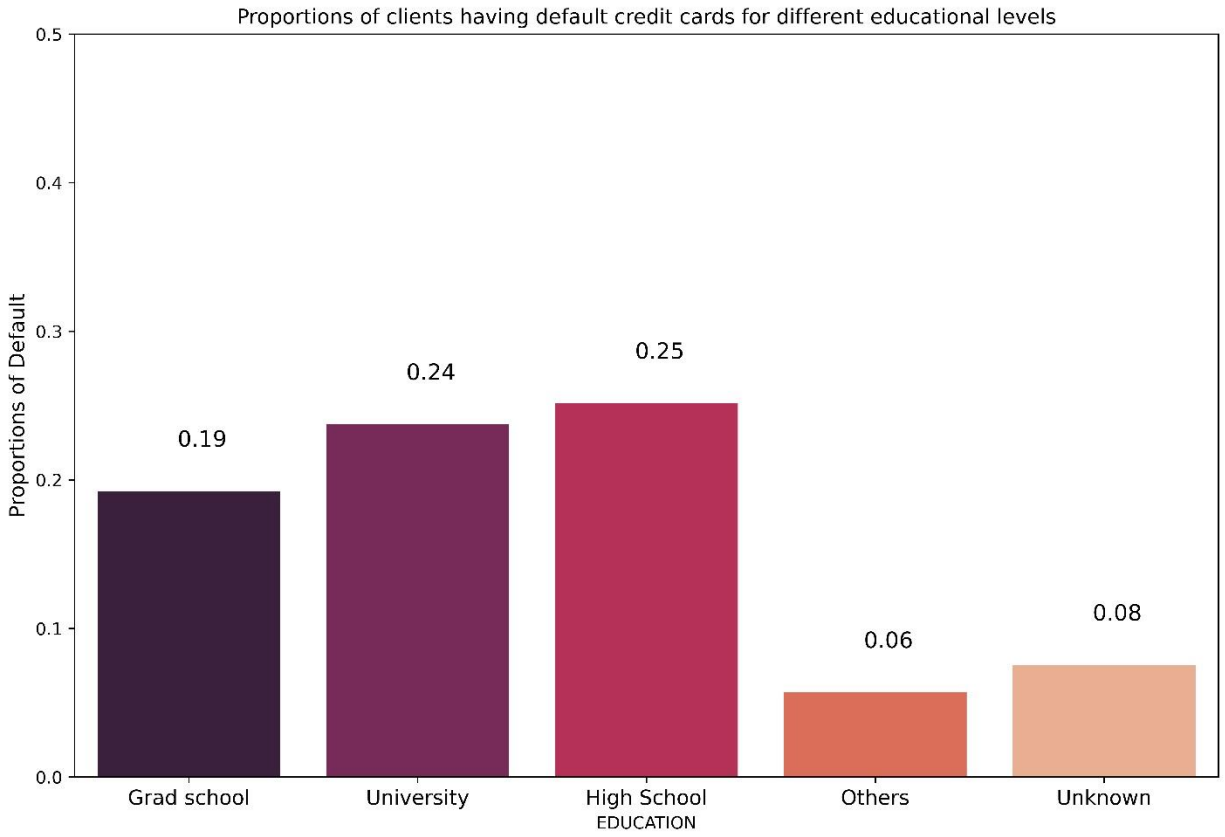

Fig 10: Proportion of defaulters for different levels of education

From the above plot, we can see that, out of 100 people belonging to 'Grad School', 19 people are defaulters, out of 100 people belonging to 'University', 24 people are defaulters, out of 100 people belonging to 'High School', 25 people are defaulters. We can see that, for the education group 'Others' and 'Unknown', number of defaulters out of 100 people is very less, 6 and 8 respectively.

• Marital Status Wise Analysis:

Now, we shall explore the data according to various levels of 'Marriage'. 'Marriage' is a categorical variable, reflecting marital status of a person- married, single, divorce and others.

Firstly, we will see, what is the number of defaulters and non-defaulters for different marital status

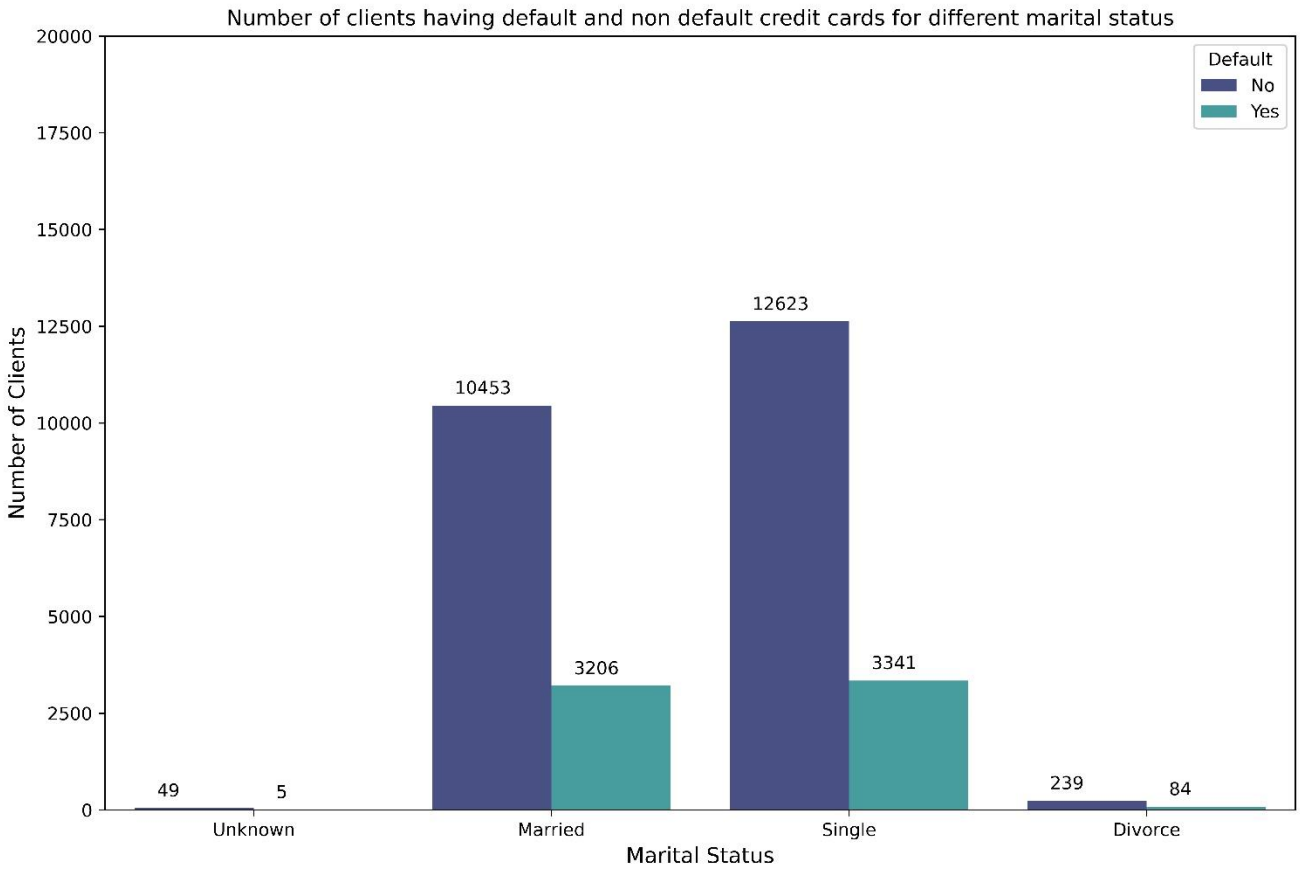

Fig 11: Number of Defaulters and Non-Defaulters for different marital levels

From the above graph, we can see that most of the people who are credit card defaulters, has the marital status either 'Married' or 'Single'. Out of these two categories, the number of people who are credit card defaulters and are single is greater than the number of people who are credit card defaulters and are married. The number of people who are credit card defaulters and has the marital status 'Unknown' or 'Divorce' can be considered as negligible.

Now, we shall explore the percentage of defaulters across different marital status, i.e., we will see each marital status contains how much percentage of total defaulters.

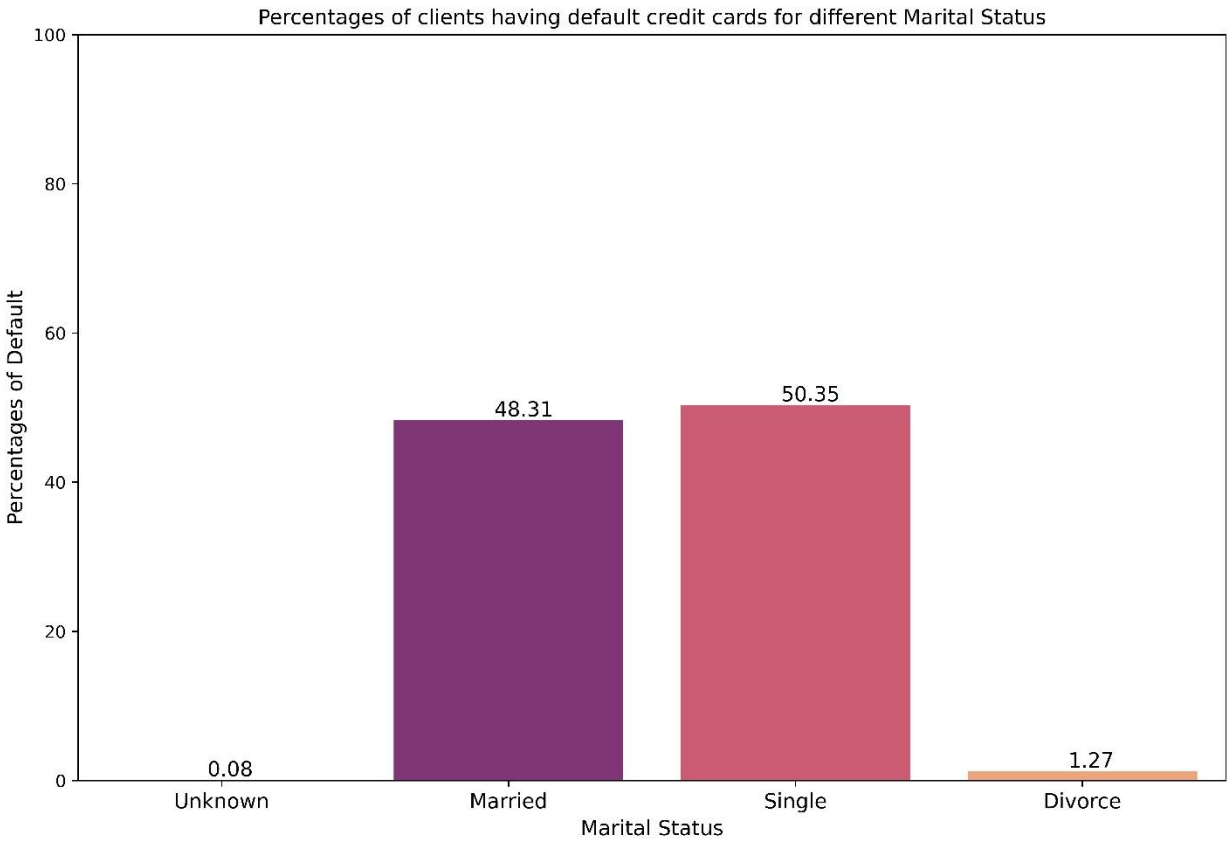

Fig 12: Percentage of Defaulters among total number of defaulters for different marital status

From the above figure, we can see that, 50.35% of the defaulters have the marital status 'single' which is the highest and 48.31% of the defaulters belong to the status 'married' which is the second highest. Percentage of defaulters belonging to the marital status 'Divorce' and 'Unknown' can be considered as negligible.

Now, we shall see the proportion of defaulters for different marital status, i.e., we will see for each marital status what is the proportion of the people whose credit card defaults.

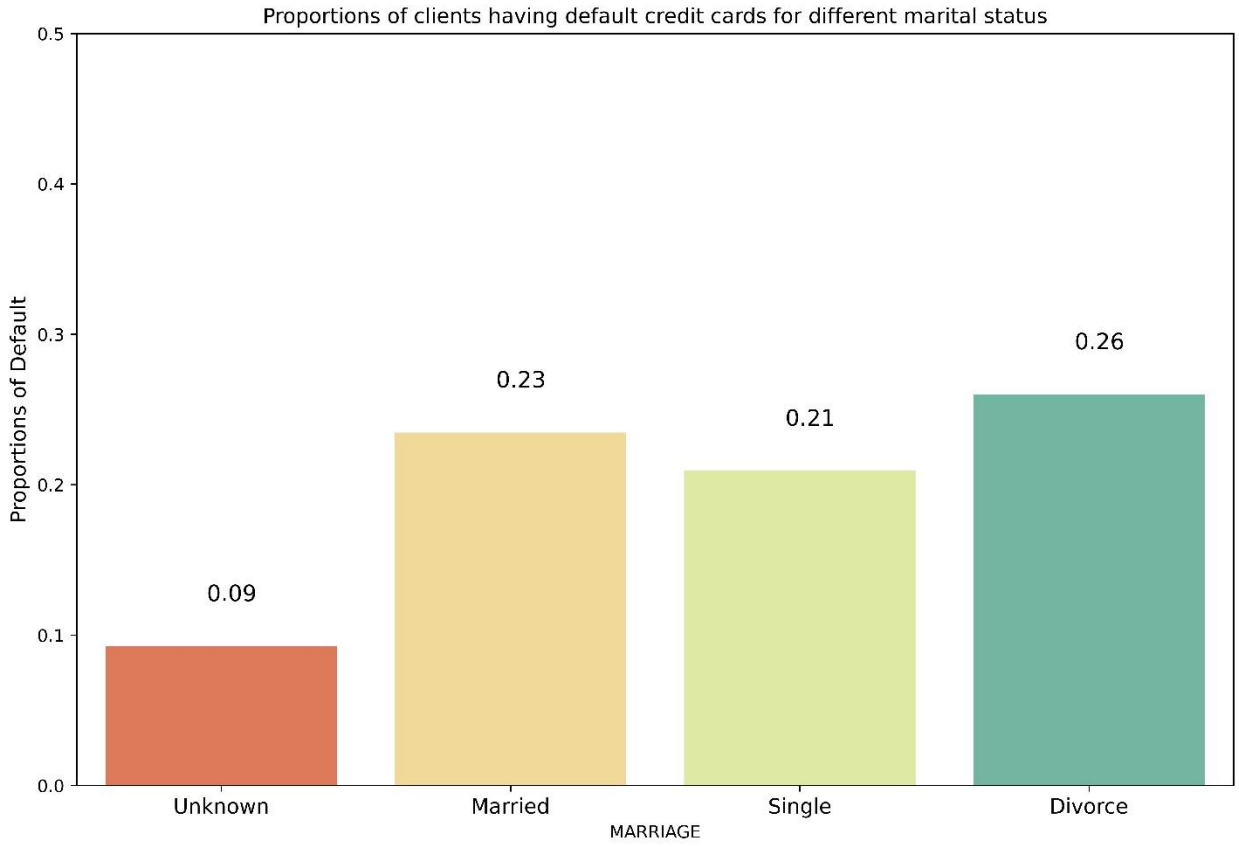

Fig 13: Proportion of defaulters for different marital status

From the above plot, we can see that, out of 100 people who are married, 23 people are defaulters, out of 100 people who are single, 21 people are defaulters, out of 100 people who had divorce, 26 people are defaulters. We can see that, for the marital status 'Unknown', number of defaulters out of 100 people is very less, 9.

Now, we shall see how proportion of defaulters vary across different gender-marital status together.

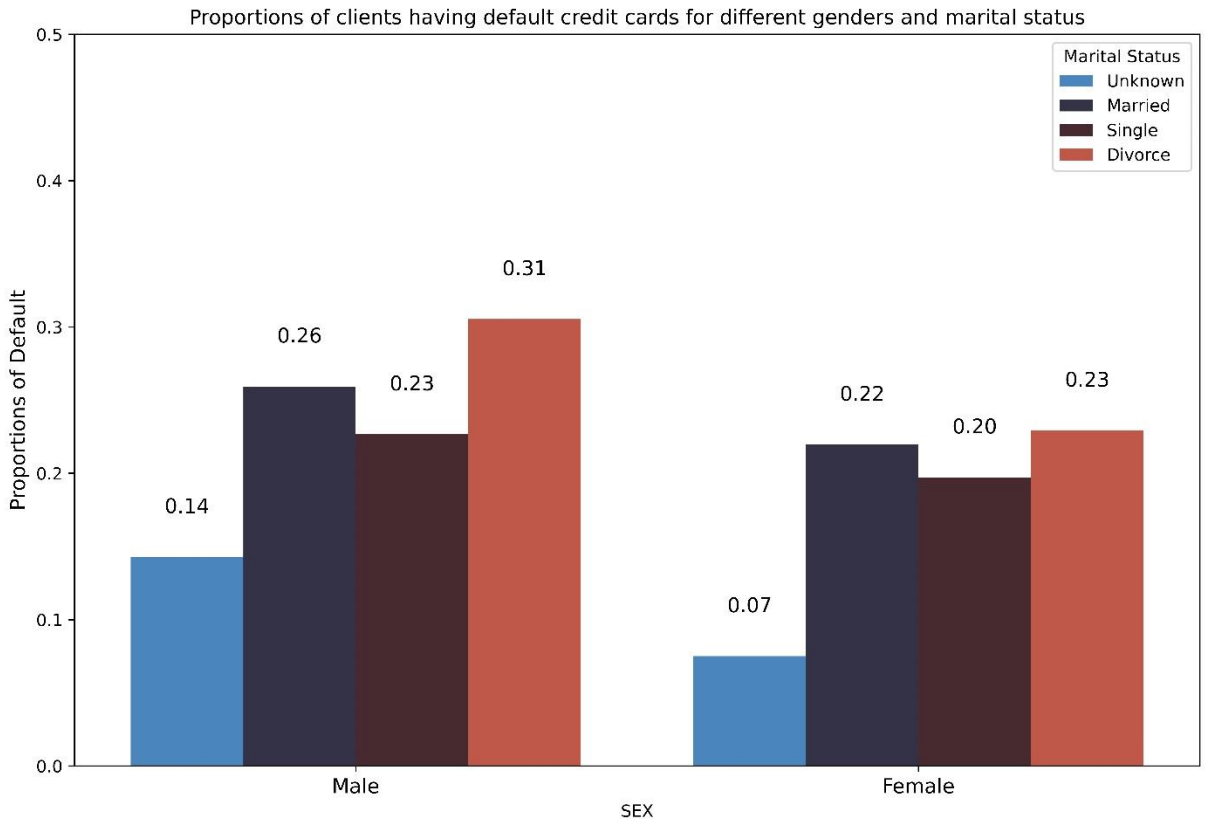

Fig 14: Proportion of Defaulters across different gender and marital status

From the above figure, we can see that, for the gender group male, the person, who had divorce has the highest proportion of defaulters (0.31) and also for the gender group female, the person, who had divorce has the highest proportion of defaulters (0.23). So, we can see that, for both the genders, proportion of defaulters is the highest for the marital status 'Divorce'. For the gender group male, the person, with 'Unknown' marital status has the lowest proportion of defaulters (0.14) and also for the gender group female, the person, with 'Unknown' marital status has the lowest proportion of defaulters (0.07). So, we can see that, for both the genders, proportion of defaulters is the lowest for the marital status 'Unknown'.

Now, we shall see how proportion of defaulters vary across different gender-education level together.

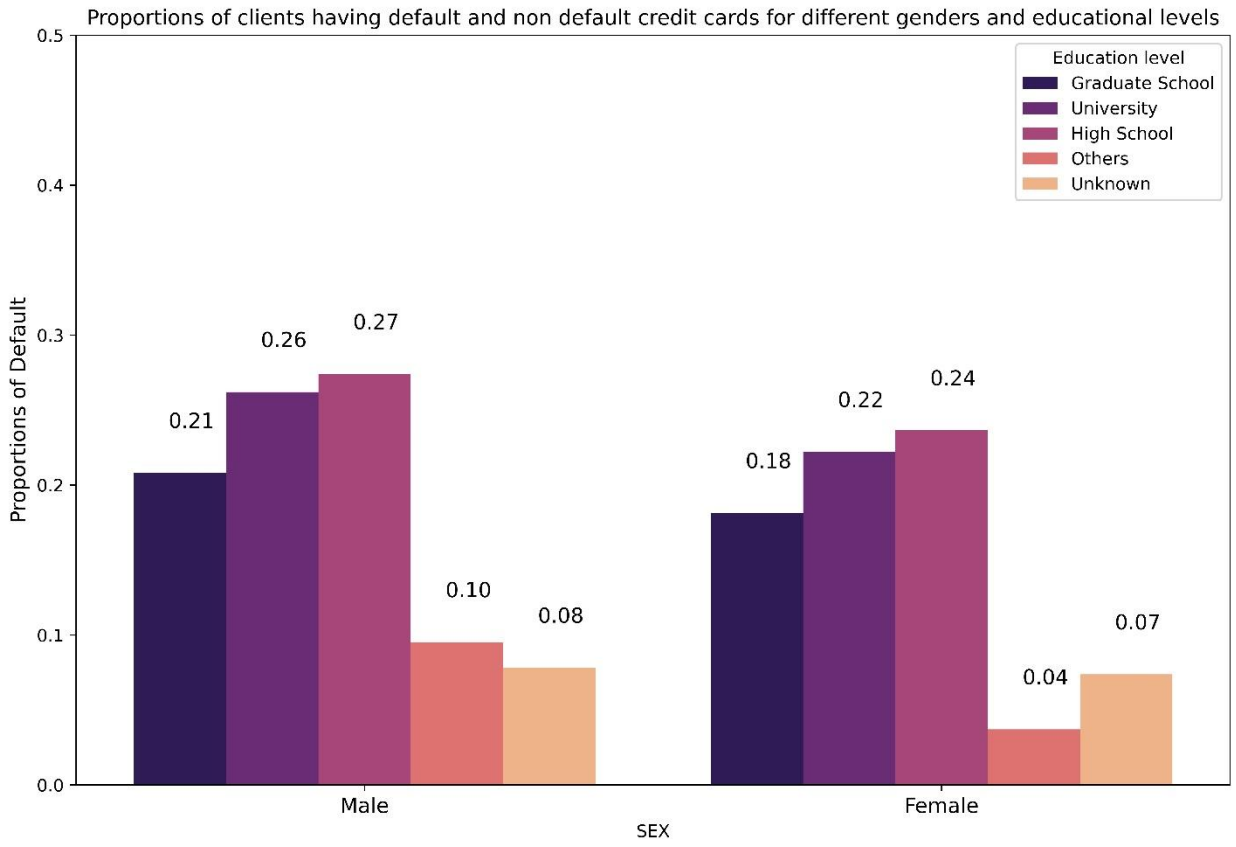

Fig 15: Proportion of Defaulters across different gender and education level

From the above figure, we can see that, for the gender group male, the person, who has education level 'High School' has the highest proportion of defaulters (0.27) and also for the gender group female, the person, who has education level 'High School' has the highest proportion of defaulters (0.24). So, we can see that, for both the genders, proportion of defaulters is the highest for the education level 'High School'. For the gender group male, the person, with education level 'Unknown' has the lowest proportion of defaulters (0.08) and for the gender group female, the person, with education status 'others' has the lowest proportion of defaulters (0.04).

•Boxplot Analysis of Limit Balance:

Now, we shall carry out boxplot analysis of the variable Limit Balance. It is the amount of given credit in NT dollars (includes individual and family/supplementary credit.

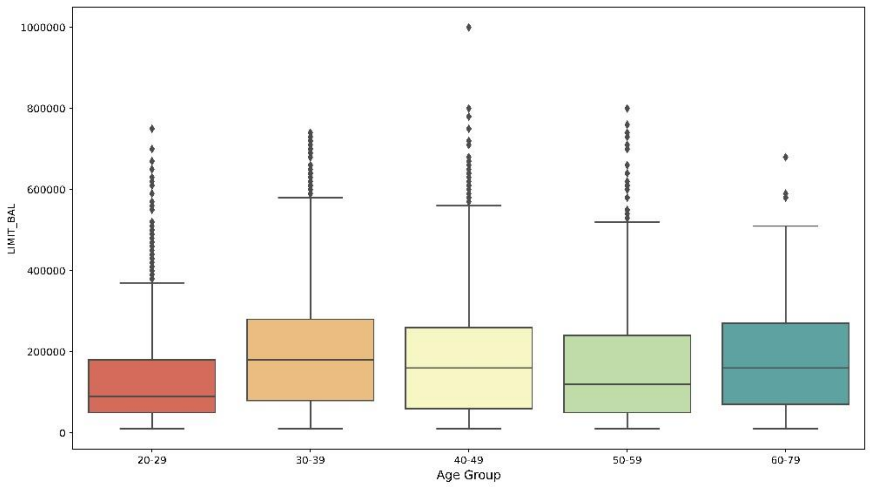

Fig 16: Boxplot of Age Group-Limit Balance

From the plot we can see that, median value of limit balance is highest for age group 30-39 and 60-79 and they are almost the same. Also, except for the age group 20-29, the median value is around 2,00,000.

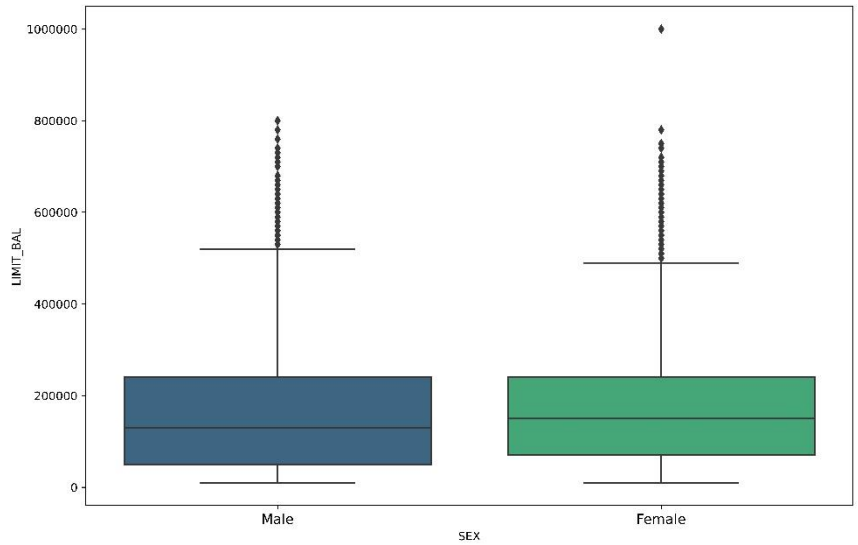

Fig 17: Boxplot of Gender-Limit Balance

From the boxplot, we can see that, the median value of the limit balance is almost same for both males and females.

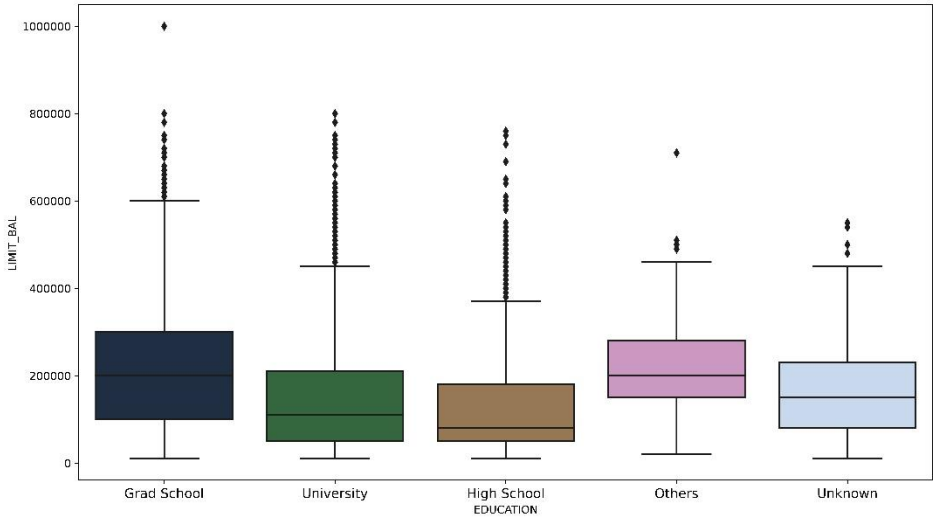

Fig 18: Boxplot of Education Status-Limit Balance

From the boxplot, we can see that, the median value of the limit balance is the highest for the education status 'Grad School' and it is almost the same as education status 'Others'. The median value of the limit balance is the lowest for the education status 'High School'.

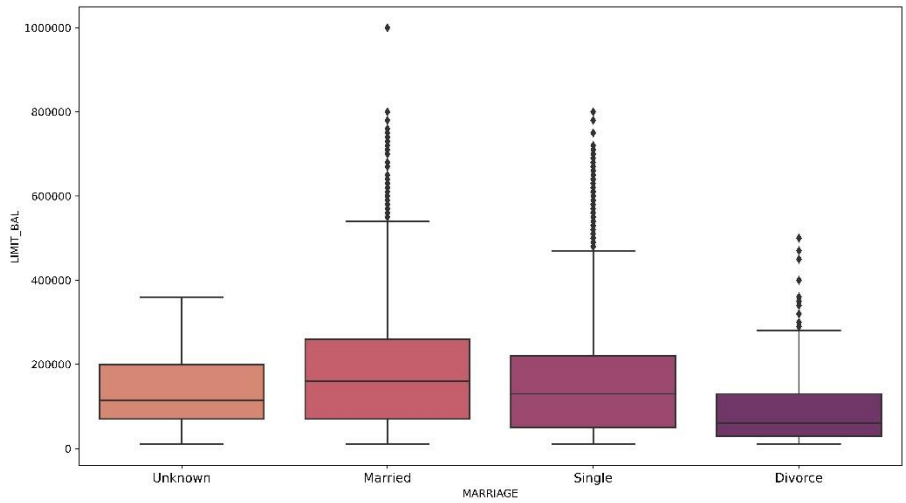

Fig 19: Boxplot of Marital Status-Limit Balance

From the boxplot, we can see that, the median value of limit balance is the highest for the clients who are married, and the median value of limit balance is the lowest for the clients who had divorced.

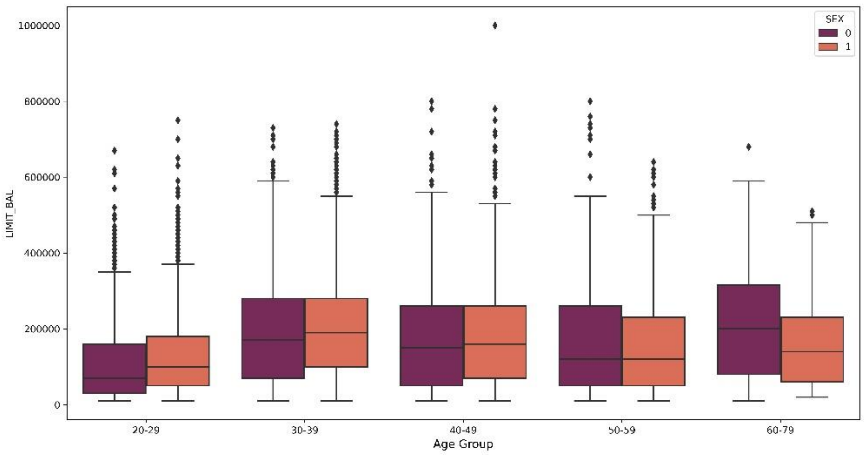

Fig 20: Boxplot of Age Group-Gender-Limit Balance

From the plot, we can see that, for age group 40-49 & 50-59, the median value of limit balance is almost the same for both males and females. For age group 20-29 and 30-39, the median value of limit balance is higher for female and for age group 60-79, the median value of limit balance is higher for male.

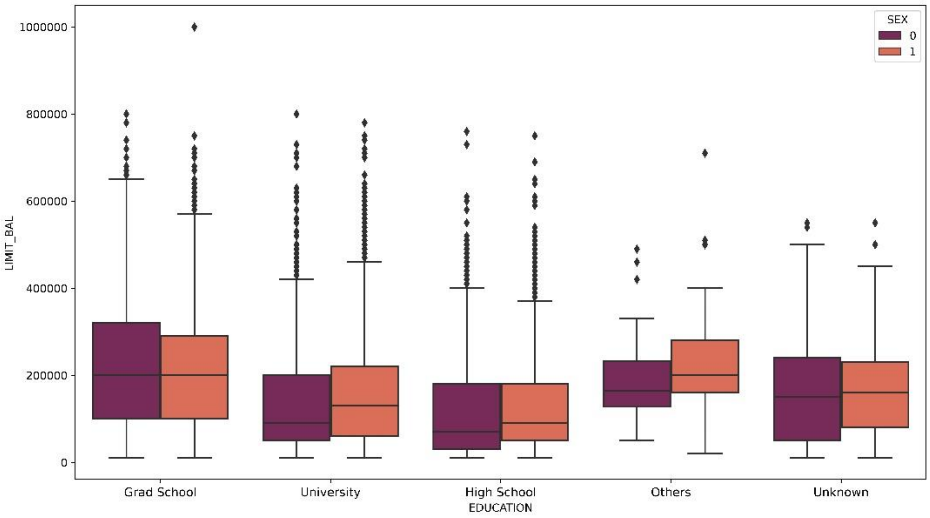

Fig 21: Boxplot of Education-Gender-Limit Balance

From the boxplot, we can see that, for the clients who has education level either 'Grad School' or 'Unknown', the median value of limit balance is almost the same. For the other education level, median value of limit balance for female is higher than that of male.

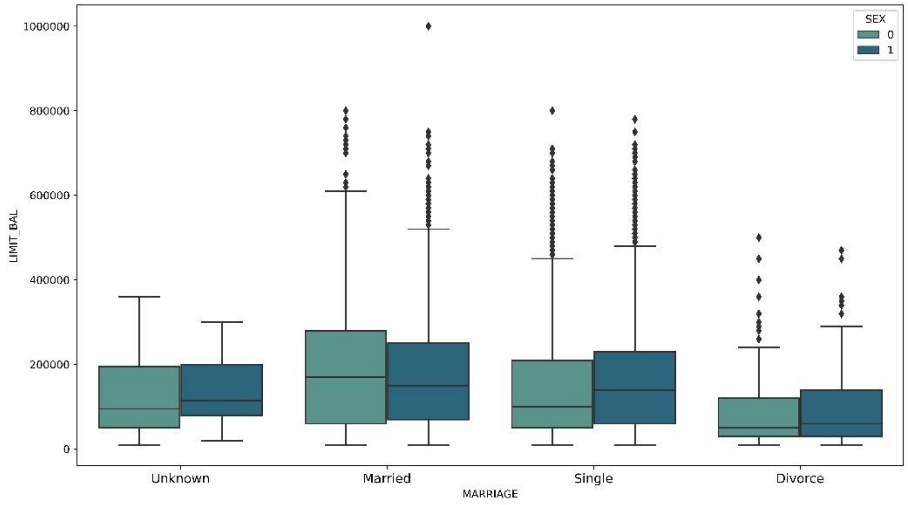

Fig 22: Boxplot of Marital Status-Gender-Limit Balance

From the boxplot, we can see that, the median value of limit balance is almost same for those males and females who had divorce. Median value of limit balance is more for female for the persons whose marital status is either unknown or is single and for married persons, median value of limit balance is more for male.

4. Testing of Hypothesis:

We want to test whether proportions of different levels for each of the categorical variables are significantly different or not using 2-proportion Z-test.

In 2-proportion Z-test we want to test,

$H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$

And the test statistic is,

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Under H_0 , the test statistic Z follows N(0,1).

We reject null hypothesis at level α if $|Z_{obs}| > \tau_{\alpha/2}$ or if p- value for the test is less than α

Now we want to test the following hypothesis using our data:

1. **Whether the proportions of defaulters for male and female clients are significantly different or not.**
   For this test, we got Z_{obs} = 6.9213 and p value = 4.47e-12
   As p value < α = 0.05, we reject null hypothesis at 5% level of significance and conclude that the proportions for male and female are significantly different.
2. **Whether the proportions of defaulters for married and not married clients are significantly different or not.**
   For this test, we got Z_{obs} = 5.1323 and p value = 2.86e-07
   As p value < α = 0.05, we reject null hypothesis at 5% level of significance and conclude that the proportions for married and not married are significantly different.
3. **Whether the proportions of defaulters for clients of age group of 35-60 and below 35 are significantly different or not.**
   For this test, we got Z_{obs} = 2.6380 and p value = 0.008
   As p value < α = 0.05, we reject null hypothesis at 5% level of significance and conclude that the proportions for clients of age group 35-60 and below 35 are significantly different.
4. **Whether the proportions of defaulters for clients having credit balance in the range 10k – 100k and above 100k are significantly different or not.**
   For this test, we got Z_{obs} = 25.94 and p value = 0
   As p value < α = 0.05, we reject null hypothesis at 5% level of significance and conclude that the proportions for clients having credit balance in the range 10k – 100k and above 100k are significantly different.
5. **Whether the proportions of defaulters for clients having educational qualification upto high school and graduation and above are significantly different or not.**
   For this test, we got Z_{obs} = -4.4801 and p value = 7.45e06
   As p value < α = 0.05, we reject null hypothesis at 5% level of significance and conclude that the proportions for clients having educational qualification up to high school and graduation and above are significantly different.

5. Feature Selection:

Before proceeding into further analysis, we will wish to check which of the variables will be relevant for our analysis. The idea behind feature selection is that, there may be variables which are highly correlated among themselves, so it is redundant to work involving all such variables.

For feature selection, we have used correlation plot. Correlation plots can be used to quickly find insights. It is used to investigate the dependence between multiple variables at the same

time and to highlight the most correlated variables in a data table. In this visual, correlation coefficients are colored according to the value.

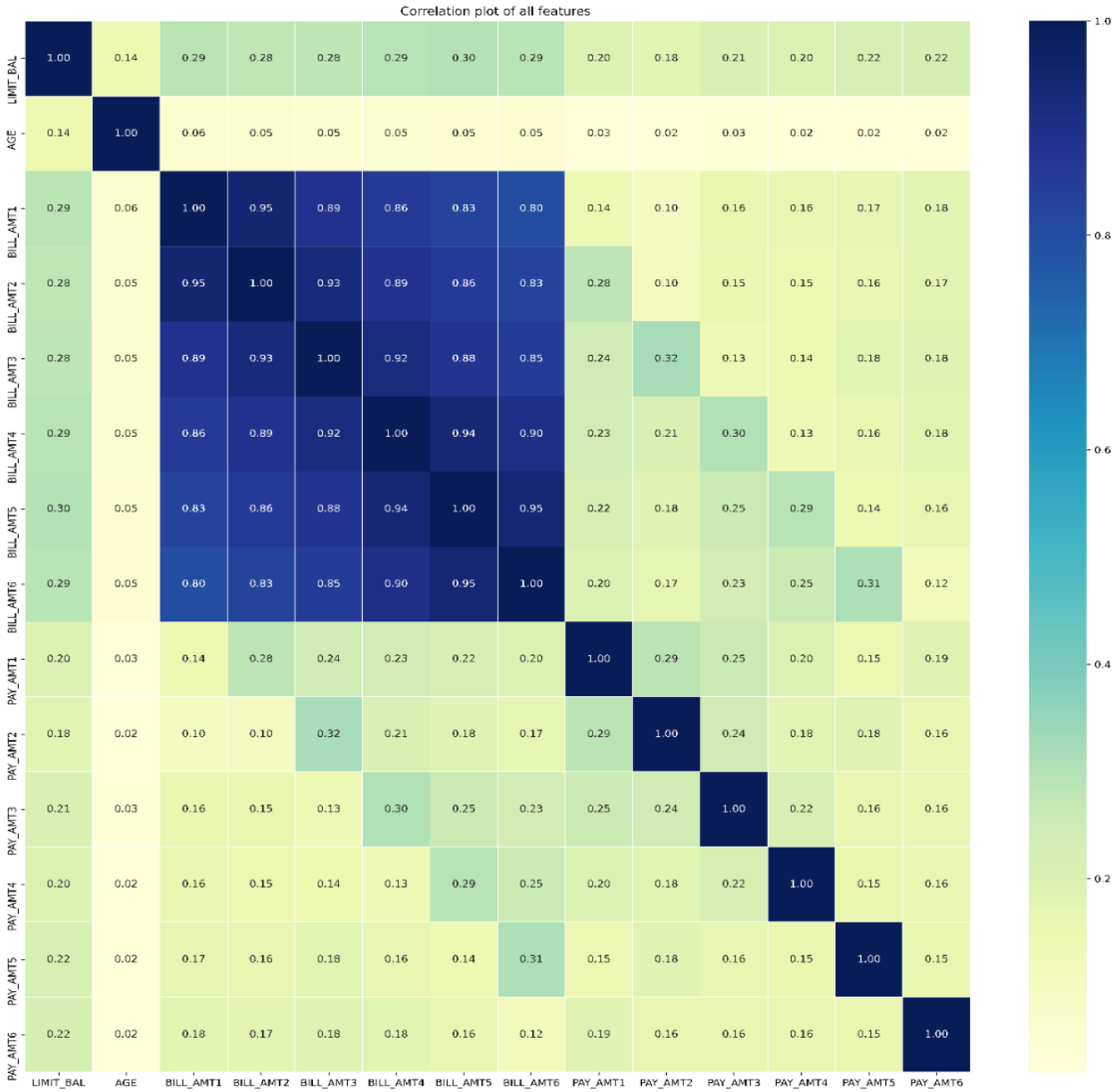

Fig 23: Correlation Plot of the features

From the above plot, we can see that, all the bill amounts are highly correlated among themselves. So, we keep the recent bill amount, i.e., the variable 'BILL_AMT1' and discard the others. Finally, we have, 18 variables to carry out our analysis.

6. Data Handling:

We have 30,000 data points and among them only 22% are credit card defaulters. So, our data is imbalanced. To make the data balanced, we have applied **Synthetic Minority Oversampling Technique (SMOTE).** SMOTE is an algorithm that performs data augmentation by creating **synthetic data points** based on the original data points. SMOTE can be seen as an advanced version of oversampling, or as a specific algorithm for data augmentation. The advantage of SMOTE is that you are **not generating duplicates**, but rather creating synthetic data points that are **slightly different** from the original data points.

The SMOTE algorithm works as follows:

1. We draw a random sample from the minority class.
2. For the observations in this sample, you will identify the k nearest neighbors.
3. We will then take one of those neighbors and identify the vector between the current data point and the selected neighbor.
4. We multiply the vector by a random number between 0 and 1.
5. To obtain the synthetic data point, we add this to the current data point.

After applying SMOTE, we have the following result, -

```
Default
0          0.5
1          0.5
```

Now, the dataset has become balanced. So, we can come up with fruitful prediction using this synthetic balanced dataset.

7. Data Analysis:

• Data Split:

We have divided the dataset into train data, cross-validation set and test set with proportion of 7:2:1.

i.e., we have 70% datapoints as training set, 20% datapoints as Cross Validation Set, 10% datapoints as test set.

•Logistic Regression:

Logistic Regression is used for classification of data. We use logistic regression for binary classification of data. In logistic regression we use **sigmoid function** to predict the class for a data point. For a model with k parameters, suppose there are 2 classes viz 0 and 1 and if we denote p = P(Y=1) where Y is the data point for which its class is to be predicted, then we can write,

$$p = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}}$$

Now if p >= 0.5, then Y is predicted to belong to class 1 and if p < 0.5 then Y is predicted to belong to class 0.

We have used Logistic Regression to predict whether credit card of a client is default or not using parameters.

From confusion matrix we got the measures for True positive (TP), True negative (TN), False positive (FP), False negative (FN) observations and measured the **accuracy** of the model,

$$\text{Accuarcy} = \frac{\text{TP} + \text{TN}}{\text{Total number of observations}}$$

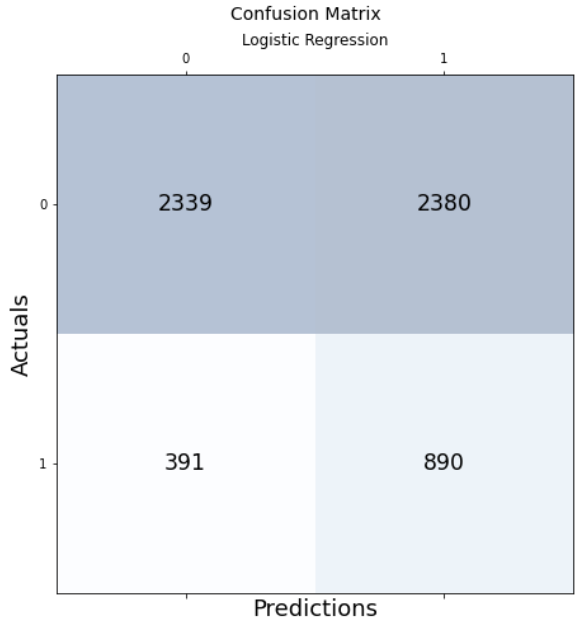

Fig 24: Confusion Matrix for Logistic Regression

The following is the measure of efficacy of the model fitted on the train data tested on cross-validation dataset,

The **Accuracy** of this model on CV set is approximately 54%.

•Decision Tree:

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

We have used Decision Tree algorithm for binary classification of our data. We fitted the model on train data, tested on the CV Dataset and obtained its measure of accuracy,

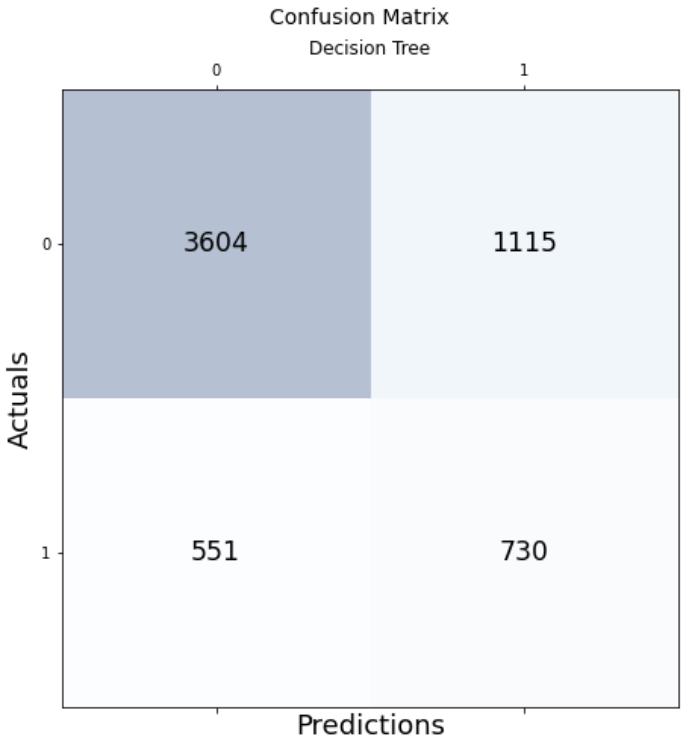

Fig 25: Confusion Matrix for Decision Tree

The **Accuracy** of Decision Tree model on the CV dataset is approximately 72.23%

•Random Forest:

A random forest algorithm consists of many decision trees. A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees.

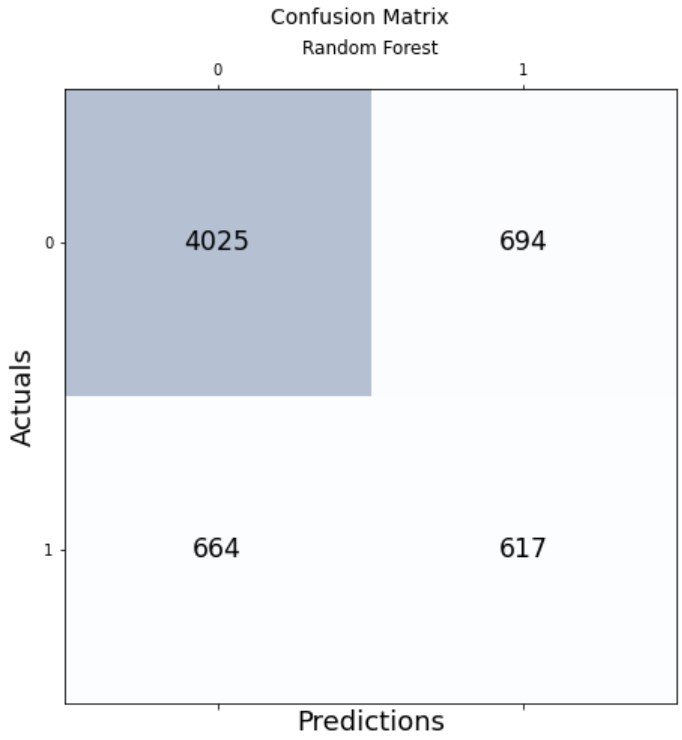

Fig 26: Confusion Matrix for Random Forest

The **Accuracy** of Random Forest model on the CV dataset is approximately 77.6%

8. Final Model:

So, comparing the Accuracy scores for different models, we selected **Random Forest** model as our final model. On fitting this model on test set we get these measures of efficacy,

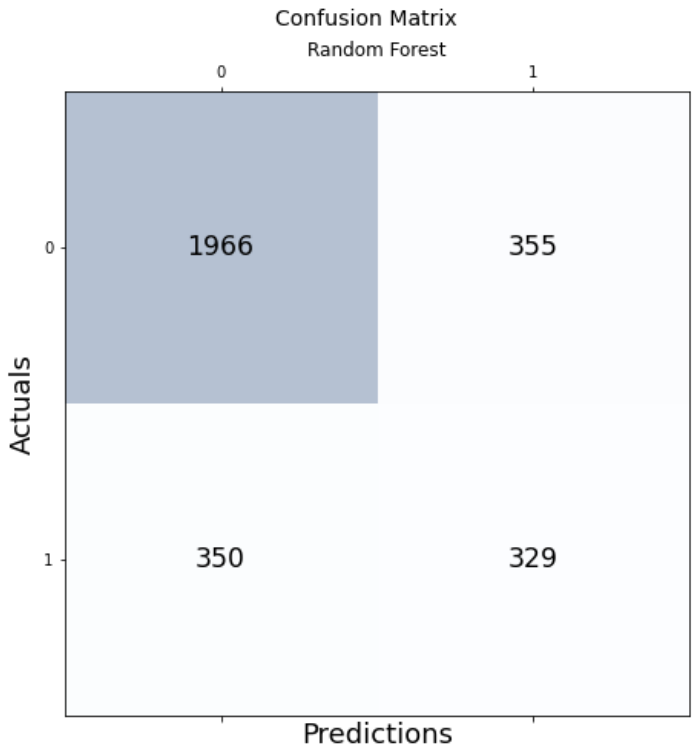

Fig 28: Confusion Matrix for Final Model

Accuracy: 0.765 (approximately 76.5%)

The ROC curve is a popular graphic for simultaneously displaying the ROC curve two types of errors for all possible thresholds. The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (AUC). An ideal ROC curve will hug the top left corner, so the larger area under the (ROC) curve the AUC the better the classifier.

ROC_AUC_Score: 0.6658

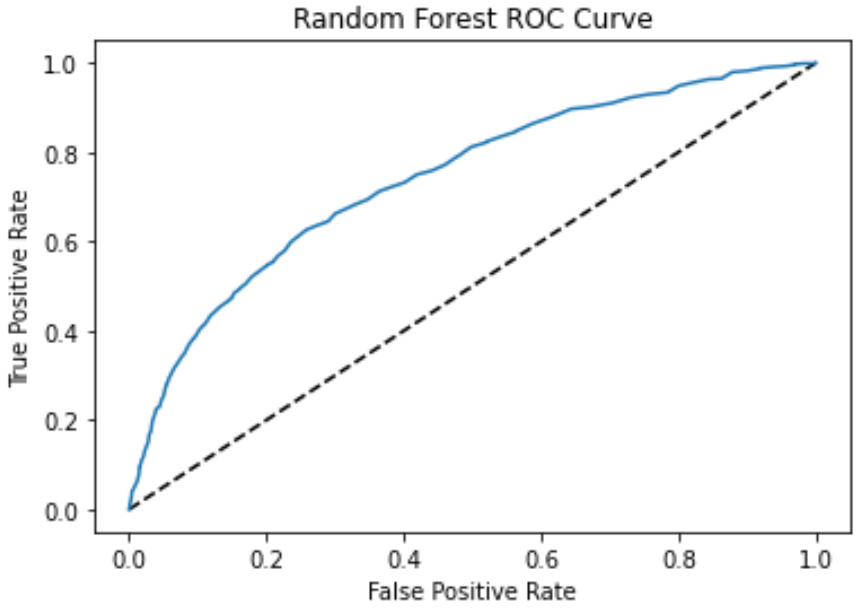

Fig 27: Confusion Matrix for Decision Tree

9. Conclusion:

After fitting Logistic Regression, Decision Tree and Random Forest on the training data, and checking their accuracy on the Cross Validation set, we found that Random Forest is the best fitted model. After final fitting the Random Forest on the test data we obtained an accuracy of 76.5% and roc_auc score of 0.6658. Considering an imbalanced dataset with only 22% default rate, the obtained result is quite satisfactory. Thus, we conclude our project here.

10. References:

(i) https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
(ii) https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis
(iii) https://towardsdatascience.com/smote-fdce2f605729
(iv) https://appsource.microsoft.com/en-us/product/power-bi-visuals/wa104380814?tab=overview
(v) An Introduction to Statistical Learning with Application in R