Predictive Analysis on Credit Card Defaults Based on Demographic Factors and Payment Behaviour

CIND 820 XJH W2024

Project by: Md Fahim Ferdous

ID: 501232653

Supervisor: Dr. Ceni Babaoglu

Date of submission: February 19, 2024

Table of Contents

1.0 Abstract:

1.1 Introduction:

In today's world, the mode of transaction is taking a paradigm shift to keep up with the modern era progression. Credit card is taking the place of cash transactions and has installed the concept of contactless transaction, which opened a new door to the business world with lots of risk. With lots of promises, risk of default has been introduced as credit card is an arrangement where customer has to pay the due within a specific timeframe. It is noticed that the credit card default and due payment frequency is increasing, and it is prevalent amongst the people with specific demographic aspects. This is not only restricted to demographic factors but also to Limit allocation which is impacting the usage behavior. The study on The Relationship between Demographic Factors and Financial Behavior on Credit Card Usage in Surabaya (Memarista, Malelak and Anastasia, 2015) was done on specific demographic factors based on 105 respondents. This study was based on five demographic factors such as age, gender, education, income marital status and tried to show the impact of these demographic factors against financial factors using Chi-squared test and cross tabulation. And on the study The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients (Yeh and Lien, 2009) conducted the analysis to predict the card default based on six data mining techniques. Using different techniques, the authors tried to portray the comparative analysis of the best model output.

This project will focus on the importance and impact on credit card usage patterns and include demographic factors, limit and payment behavior; with higher data volume to conduct a predictive analysis with the impactful factors on credit card defaults.

1.2 Scope of the study:

This is a quantitative study to identify the combined factors (i.e: limit, demographic factors and payment factors) impact and predict the credit default. There were some previous works on credit default and associated factor impacts, however, there was an attempt to see the default event from as a whole point of view. There was no interaction between data dependency. And the study sample was also smaller. This project aims to predict the credit default based on two stages. On the first stage the limit will be justified based on demographic factors and payment behavior. The limits are fixed by credit analysts based on financial factors and past payment behavior. This study will focus on the impact of demographic factors on limit allocation. Limit is an important factor in credit default. The theme is, due to the limit amount, the usage amount and the bill amount go up. If the limit is allocated high while the demographic factors and payment behavior are indicating mismatch, that will lead to credit default. Upon finalizing the impacts and importance on limit, the study will further analyze how these factors are impacting credit default and the predictive analysis on credit default.

1.3 Research question:

The primary objectives for this project are:

- Impact of Demographic factors and payment trend on Limit allocation
- Conducting effective predictive analytics on card default based on impactful demographic factors and payment behaviour

The study is about finding out the demographic factors' impact on limit, and the combined impact of limit, demographic and spending pattern on credit default. The data taken for the study is for six months in Taiwan. The question for the research is:

- What is the predictability of credit card default based on spending pattern, demographic behaviour and limit allocation?

1.4 Methodology:

The study will be conducted with classification theme. The pattern of defaults will be identified with historical data with exploratory analysis. After that the predictive analysis will be conducted based on Decision tree, Logistic regression and Support Vector Machine. To find out the effectiveness of the model, confusion matrix will be developed, and accuracy, precision and recall will be calculated. Steps are provided below:

- Data collection and data preparation
- Exploratory data analysis: Bar Charts and boxplot
- Data balancing, scaling and normalization
- Dimensionality reduction
- Experiment design: training-test set
- Data modelling: Decision Tree, Logistic Regression and SVM
- Cross Validation and model output evaluation
- Result interpretation

The Analysis will be conducted using python and for illustration tableu will be used.

1.5 The Dataset:

The dataset was collected from UCI Machine Learning Repository. The dataset includes a total of 23 variables and 30000 structured data points. The database presents factors such as:

X1: Amount of the given credit (NT dollar)

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: past payment history. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar September-April)

X18-X23: Amount of previous payment (NT dollar September-April)

Dataset link: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

Github Link: https://github.com/Cattitude101/CIND-820-Project/tree/main

2.0 Literature Review:

Demographic variables can have a substantial impact on credit default rates, reflecting the intersection of personal characteristics with financial behaviors. Factors such as age, income, employment status, and education level often play pivotal roles. Younger individuals may face higher default rates due to limited credit history and financial experience. Lower income levels can contribute to financial instability, increasing the likelihood of default. Educational background influences financial literacy, affecting individuals' ability to manage credit responsibly. Additionally, marital status and family size can impact financial commitments and, consequently, default probabilities. Geographical location may influence economic conditions and job opportunities, further affecting credit defaults. Understanding these demographic variables is crucial for lenders and policymakers to develop targeted risk mitigation strategies. By incorporating demographic insights into credit risk assessments, financial institutions can tailor their lending practices and credit scoring models, ultimately reducing default rates. Additionally, policymakers can design targeted financial education programs to address specific demographicvulnerabilities and promote responsible financial behavior, contributing to overall financial stability. Credit card default is a widespread issue all over the world. This is not a conventional loan product, rather it is a type of continuous loan with secured and unsecured category for daily sage purpose. A limit is set based on the earnings and the security. However this limit allocation itself is an important factor for purchasing behaviour, which impacts the payment pattern and eventually to credit card default. In the article The Effect of Credit on Spending Decisions: The Role of the Credit Limit and Credibility (Soman and Cheema, 2002), The research studied consumer decisions about utilizing a credit line and reinforced prior findings that consumers are not aware of the value of their future incomes. The author argued that consumers use credit limit as a parameter of their future earnings potential. Specifically, the inference regarding the assigned limit is that the future income of the customer will be aligned with the assigned limit. If the allocated credit limit is high, they are likely to infer that their lifetime income will be high and hence their willingness to spending will also be high. Conversely, consumers who are granted lower amounts of credit are likely to control their spending. So, whenever the credit is allocated, it is solely based on income and security. This study will try one step further to identify the demographic and payment behaviours to on limit so that customer limit allocation justification can be measured. Because the limit setting and limit increment is also fixed based on spending pattern and amount. So, if the spending is more, the limit will be increased, which may lead to higher amount of purchase and at the end may lead to credit default.

The demographic factors have a big impact on the credit card default. If the limit is allocated to young, uneducated and unemployed person, it refers to high risk potential for credit default. The Relationship between Demographic Factors and Financial Behaviour on Credit Card Usage in Surabaya (Memarista, Malelak and Anastasia, 2015) constructed a study with five demographic factors and tried to find out the relationship between financial behaviour and demographic factors on credit card default. The result of this research shows that the financial behaviour on credit card usage is low. From the demographic factors, the education has significant relationship with financial behaviour on credit card usage. However, the study did not find significant relationship between financial behaviour and demographic factors like age, gender, income, and marital status on credit card usage. This study will combine the outcomes of these studies and try to identify with demographic factors and payment pattern to identify the impact on limit and credit card default.

The studies conducted up to now focused on the credit defaults from the broad perspective or the impact of demographic factors on card defaults only. There is the study The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients (Yeh and Lien, 2009) which conducted the analysis to predict the card default based on six data mining techniques. This study conducted based on the risk management perspective to estimate the real probability of default by using KNN, Logistic regression, Discriminant analysis, Naive Bayesian classifier, Artificial neural networks, and Classification trees. They tried to identify the best fit model to predict the credit default without getting analytical to the factors. They have used the same dataset model but considered the whole dataset without singling out based on importance. However, this study will first focus on the demographic and payment pattern to identify the limit and then use the impactful demographic factors, limit and the purchase behavioural factors to derive the credit card default prediction.

A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors (Çallı and Coşkun, 2021) is a thorough and comprehensive study on identifying the predictors of credit default based on individuals' financial behaviour. Here the authors considered factor groups such as Socioeconomic, Demographic, Educational, Institutional/financial, Personality, Values/attitudes/behavioural, Situational, Macroeconomic, Health-related and Alternative. This study conducted different data mining techniques and it reached to a conclusion that personality and behavioural variables are effective predictors of credit default. However, these two factors include broad array of sub factors. However, the study only focused on the predictors of credit default. On the contrary, this study will be focused on more specific variables and based on cause and result argument. The data taken here is fewer than the discussed paper and the data includes the payment pattern, which is the key factor to identify the appropriateness of limit allocation and credit default.

Considering the works that has been done, this study will contribute to the limit vs demographic factors and payment pattern. There are studies regarding demographic factors and behavioural factors. The implications of the limit were omitted. Limit is an important factor considering the financial worth determination. There are many studies which are focusing on the same objective but followed different process and different methodology. Most of the studies focused on the overall factors impact on the dependent variable. There is not any study that is exactly like this one. The studies did not focus on the cause and impact theme for the study. Some of those studies tried to put emphasis on the effectiveness of credit default prediction based on different data mining method. Some of the studied identified the cause of credit defaults but those are the line of credits, not the credit cards which is a high-risk product, and the risk factors are totally different. There is a study "Demographics, attitude, personality and credit card features correlate with credit card debt: A view from China" (Wang et al., 2011) which tried to correlate the demographics and behavioural aspects with credit card debt. The authors conducted a regression analysis which led to the conclusion that demographic variables and credit card features have limited explanatory powers. On the other hand, attitude variables and personality variables provide more explanation regarding credit card debt. The authors also found that some credit card features provide a wrong sense of "illusion of income" which led the customer to credit debt. But this study did not focus on how me graphic and payment behaviours are impacting the limit allocation and what leads to credit card default.

The demographic and behaviour determinant of credit card default in Indonesia (Achsan et al., 2022) tried to analyze the behavioural and demographic impact on nonperforming credit card. This study tried to find out the influential demographic and behavioural factors which can later be use for credit scores done by Indonesian banks. The authors conducted logistic regression model and found out that cardholder behaviour is more likely to contribute to the nonperforming credit card rather than demographic behaviour. There are any studies conducted on credit card default and factor analysis, but none of those studies used the limit implication and demographic and payment behaviour importance on credit default from the cause-and-effect point of view. There are some works that are very close to this study, but the time frame and the data frame are different, and the studies were conducted form a more holistic point of view rather than zooming in to factor-based analysis. There is not any study which is exactly like this. In the study The Usage Patterns of Credit/Debit Card across Various Demographics (Rauf et al., 2022) the authors aimed to investigate usage patterns of credit/debit cards across demographics in Lahore and Kasur, distinguishing between urban and rural areas. A 225-consumer sample was initially estimated, but after data cleaning, the final analysis included 200 subjects. Using a close-ended questionnaire and SPSS for empirical analysis, the study found that gender, occupation, area, and income significantly influenced credit/debit card usage patterns during purchases. The research delved into card preferences, financial conditions, budget control, and money shortages. The results indicated that varying demographic attributes played a crucial role in shaping the behavior of Pakistani credit/debit card holders during transactions. Notably, gender, occupation, area, and income emerged as key factors influencing buying habits. This study contributes significantly to understanding the evolving patterns in the growing sector of cash-less transactions in Pakistan. As the country's economy shifts towards cashless transactions, the findings of this research can serve as a valuable foundation for future studies in the banking sector, providing insights into user behavior that can aid in policy-making and business strategies. This study outcome was quite useful for our study because this study used different algorithms and methods but found useful demographic factors which contributes directly to the spending pattern and payment patter. The allocation of the limit is considered based on financial aspects and previous credit usage. But at the end it depends on the person type, how he/she will utilize the credit. And this utilization estimation is the ultimate consequence of breaching the credit limit and event of credit default.

The difference between this study and the previous study is that this study is focused on the behavioural pattern rather than financial factors for the credit default. The previous studies were done based on the model efficiency and result accuracy. With all the factors the studies tried to find which model is better at predicting. Some studies tried to create a direct affiliation between demographic factors and the credit defaults. But all these studies omitted one important aspect that credit default does not happen just because of financial factors and demographic factors but for behavioural pattern are well. This study will consider the factor "limit" as a proxy for the spending behaviour of the customers. Limits are allocated based on financials, but because of the demographic and spending patterns, the limit ceiling breaks and limits are reallocated. That is why limit will be considered as the dependent variable and based on this important demographic factor will be identified. And the second stage will find out the predictability of the credit default based on the important demographic and spending factors. All the previous studies either stopped at only efficiency measurement of default predictability or establishing relationship with the default. But none of the studies did the cause-and-effect analysis which can add value regarding default prediction. The study Understanding the impact of borrowers' behavioural and psychological traits on credit default: review and conceptual model (Goal and Rastogi, 2021) focused on certain behavioural and psychological traits of the borrowers which have the tendency to predict the credit risk of the borrowers. The study adopted the systematic literature review to find out those traits. This study specifically focused on behavioural and psychological traits which are directly related to the spending pattern, and this is specifically relevant to our study. However, this study has used different dataset and did not use any predictive or descriptive algorithm to explain the predictability. Our study is a replication-based study which will take the theoretical and analytical outcomes as a reference for conduction and add new value to the predictability of credit default.

3.0 Data Description

Credit card default dataset (default of credit card clients.xls) contains different demographic data, payment count, bill amount, and payment amount from April to September. The dataset contains data for 30000 customers. There are 23 explanatory variables based on which the credit card default will be predicted. Out of 23 variables, limit, bill amount and payment amount are numerical data and age is nominal data. The variable "The default payment next month" is presented in the binary format where 1 is presented as YES and 0 is presented as NO. This data set has multivariate characteristics and imported into python. The attributes are explained as below:

LIMIT_BAL: This parameter is a numeric and this explains the amount is allocated for that person by the financial institution based on his/her income and spending pattern. The minimum value is 10000 and maximum value is 1000000. On an average the Limit allocated for these 30000 respondents is 167484.32. This includes both the individual consumer credit and his/her family (supplementary) credit. Limit balance will act as a filter to identify the importance of the demographic factors as these factors have impact on the credit default and the credit default occurs because of the irrational allocation of limit. Here limit balance will act as a standard for those demographic factors.

SEX: SEX indicates the gender of the individual. This categorical data has been converted to numerical data where 1 is considered as male and 2 is considered as female. In the database there are 11888 male and 18112 female participants.

EDUCATION: EDUCATION: EDUCATION: EDUCATION refers to the level of education the participant has received. This is a categorical data which has been converted to numeric data levels. The Data levels along with counts are as below:

1 = graduate school:10585

2 = university:14030

3 = high school: 4917

4 = others:123

In this dataset there are category 5 and 6 which are unlabelled. Category 0 is not documented.

MARRIAGE: This variable is referred to the marital status. This categorical data has been converted to numerical data level. Married people are categorized as 1, singles are categorized as 2 and others are categorized as 3. As per the dataset, there are 13659 married participant, 15964 single participant and 323 other participants. It has data point 0 which are undocumented.

AGE: Age is the preferred as the age of the participants. This data is in the numerical value format. The maximum data point is 79 and minimum datapoint is 21 and the average is 35.49.

PAY_0 to PAY_6: These data points refer to the history of past payments. This data points refer to payment tally from April 2005 to September 2005. It is covered to numerical value to address the counts. The measurement scale for the repayment status is: -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. It has undocumented data label like -2 and 0.

BILL_AMT1 TO BILL_AMT6: This parameter is numerical in type and refers to amount of bill statement (NT dollar). These data points spread from April 2005 to September 2005. Highest data point value is 1664089 among these columns and lowest data point is -339603 and average is between 38871.76 to 51223.33. There are negative bill amounts. They can be considered as advanced payments.

PAY_AMT1 to PAY_AMT6: This parameter refers to mount of previous payment (NT dollar) from April 2005 to September 2015. Highest amount paid is 1684259, lowest amount is 0, average is from 4799.39 to 5921.16.

default payment next month: Default payment next month is the occurrence of default of payment on the next month. This is the predictability of occurring a default for the specific customer. This is the **dependent variable,** and all the other variables are independent variable. This categorical variable is numbered as 1=Yes and 0=No.

The data attribute summary is provided below:

Table 1: Dataset Dictionary

Attributes	**Type**	**Data Description**	**Missing/Unexplained values**
LIMIT_BAL	Quantitative	Allocated Limit	No
SEX	Categorical	Gender of participants	No
EDUCATION	Categorical	Level of Education	345
MARRIAGE	Categorical	Marital Status of Participants	54
AGE	Quantitative	Age of Participants	No
PAY_0 TO PAY_6	Categorical	Payment Status of the customer	Unexplained data label -2 and 0
BILL_AMT1 - BILL_AMT6	Quantitative	Amount of Bill for the month	Unexplained data negative bill amounts
PAY_AMT1 to PAY_AMT6	Quantitative	Amount of bill payment for the month	No
default payment next month	Categorical	Default payment next month (Yes/No)	No

To understand the data pattern a comparative table is presented below based on minimum value, maximum value, standard deviation, count and quartile distribution. Based on the below table it can be said that the range of LIMIT_BAL is too broad, there is no null value, and there is negative value for BILL_AMT data. There can be an implication that the money is reimbursed from some merchant, or the cardholder paid more than the bill. For limit Balance higher amount is located on the 3rd quartile. The mean age is 35.49 years and 4th quartile age are 41 years. The mean is showing lower for BILL_AMT because of the negative values. For the uncleansed and imbalanced data, the quarterlies and means are not portraying the proper picture of the dataset.

Table 2: Statistical overview of the dataset

index	count	mean	std	min	25%	50%	75%	max
ID	30000	15000.5	8660.40	1.0	7500.75	15000.5	22500.25	30000
LIMIT_BAL	30000	167484.32	129747.66	10000.0	50000.0	140000.0	240000.0	1000000
SEX	30000	1.60	0.49	1.0	1.0	2.0	2.0	2
EDUCATION	30000	1.85	0.79	0.0	1.0	2.0	2.0	6
MARRIAGE	30000	1.55	0.52	0.0	1.0	2.0	2.0	3
AGE	30000	35.49	9.22	21.0	28.0	34.0	41.0	79
PAY_0	30000	-0.017	1.12	-2.0	-1.0	0.0	0.0	8
PAY_2	30000	-0.13	1.19	-2.0	-1.0	0.0	0.0	8
PAY_3	30000	-0.16	1.20	-2.0	-1.0	0.0	0.0	8
PAY_4	30000	-0.22	1.17	-2.0	-1.0	0.0	0.0	8
PAY_5	30000	-0.27	1.13	-2.0	-1.0	0.0	0.0	8
PAY_6	30000	-0.29	1.150	-2.0	-1.0	0.0	0.0	8
BILL_AMT1	30000	51223.33	73635.86	-165580.0	3558.75	22381.5	67091.0	964511

index	count	mean	std	min	25%	50%	75%	max
BILL_AMT2	30000	49179.08	71173.77	-69777.0	2984.75	21200.0	64006.25	983931
BILL_AMT3	30000	47013.15	69349.39	-157264.0	2666.25	20088.5	60164.75	1664089
BILL_AMT4	30000	43262.95	64332.86	-170000.0	2326.75	19052.0	54506.0	891586
BILL_AMT5	30000	40311.40	60797.16	-81334.0	1763.0	18104.5	50190.5	927171
BILL_AMT6	30000	38871.76	59554.11	-339603.0	1256.0	17071.0	49198.25	961664
PAY_AMT1	30000	5663.58	16563.28	0.0	1000.0	2100.0	5006.0	873552
PAY_AMT2	30000	5921.16	23040.87	0.0	833.0	2009.0	5000.0	1684259
PAY_AMT3	30000	5225.68	17606.96	0.0	390.0	1800.0	4505.0	896040
PAY_AMT4	30000	4826.08	15666.16	0.0	296.0	1500.0	4013.25	621000
PAY_AMT5	30000	4799.39	15278.31	0.0	252.5	1500.0	4031.5	426529
PAY_AMT6	30000	5215.50	17777.47	0.0	117.75	1500.0	4000.0	528666

4.0 Data Observation and cleaning:

The dataset has some observations and inconsistencies because of which the data summary is not showing the appropriate picture. The observations for the data are presented below:

- There are no null or missing values in the datapoint. The data attribute summary shows that the data type is **int64**, which means no mixed data.
- MARRIAGE has undocumented label 0.
- EDUCATION has undocumented label 0,5 and 6.
- BILL_AMT has negative balance
- The data has imbalance, it has fewer credit default and more non default.
- The data has the mix of categorical and continuous variables.
- PAY_0 is inconsistent with other PAY data labels.
- There are outliers in the BILL_AMT.

Based on the observations the following steps have been implemented:

1. Labels rename and -1 and -2 is merged with 0: for card payment count, PAY_0 refers to paid duly. So, to keep the alignment the data label PAY_0 has been renamed to PAY_1. In the dataset there are data which are labelled as 0 and 2 which are undocumented. -1 refers to paid duly and 1 means 1 month due. To align the data, -1, -2 and 0 all are labelled as 0 which means paid duly.
2. Undocumented data label 0: For the variables MARRIAGE and EDUCATION has 0 label which is not documented. There are 54 datapoint for MARRIAGE and 14 datapoint for EDUCATION. Considering the data definition in these two variables, it can be assumed that these are missing variables which are tagged as 0. So, these datapoints are removed from the data.
3. Undocumented data label 5 and 6 for EDUCATION: for the variable EDUCATION there are undocumented data label 5 and 6. This variable has distinct and well documented data labels. So, to keep the data aligned these two data labels are covered to 4=others. 331 data points are converted to data label 4.

4. Outliers of BILL_AMT: There are many bills amounts which are high considering the means of the overall dataset. However, considering the credit card industry situation these are normal because bill amounts get high because of their financial capability and their spending pattern. If their financial capability is high, limit is set as high and vice versa.

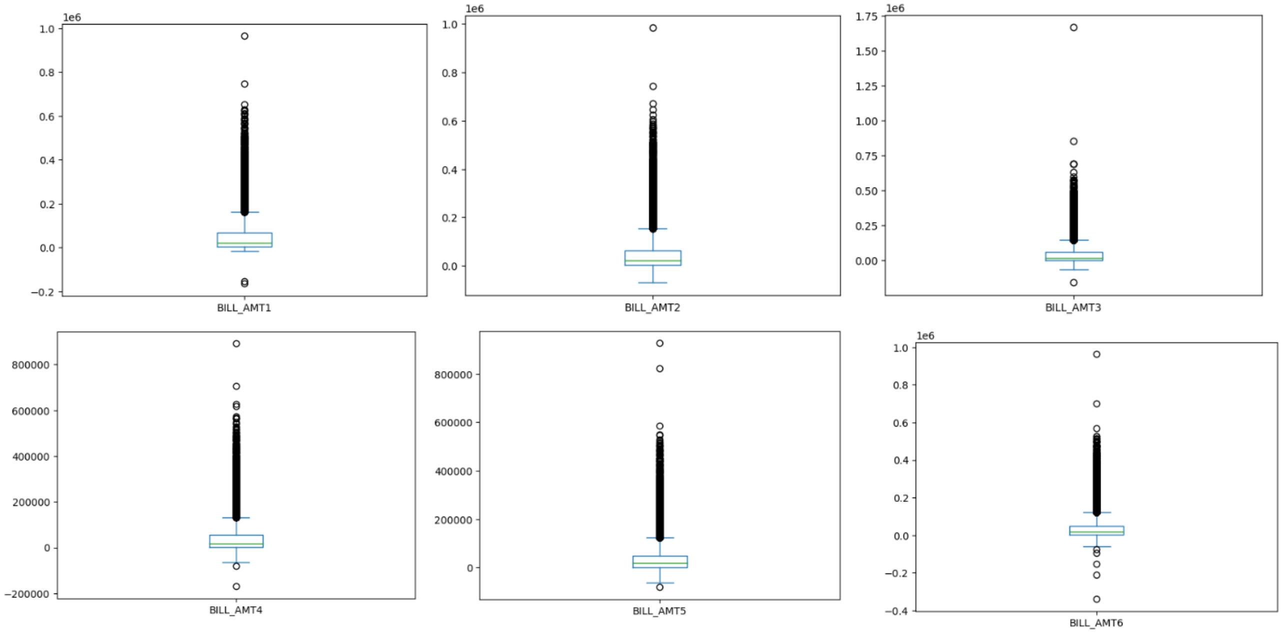

Figure 1: Boxplots outliers of BILL_AMT

To check the alignment of data, we will check the data consistency considering outlier PAY_AMT1>300000 as sample. As we can see that the variables are well aligned with the BILL_AMT. To analyze and identify a better outcome, these outliers will be kept in the dataset.

These outliers are not clerical mistakes, and the data is well aligned. So, we are keeping the outliers.

Table 3: BILL_PMT outliers justification

index	LIMIT_BAL	PAY_1	PAY_2	BILL_AMT2	PAY_AMT1	BILL_AMT1
2687	500000	0	0	367979	368199	71921
5687	480000	0	0	400000	302000	106660
8500	400000	0	0	405016	405016	6500

index	LIMIT_BAL	PAY_1	PAY_2	BILL_AMT2	PAY_AMT1	BILL_AMT1
12330	300000	1	0	324392	505000	-165580
25431	170000	0	0	167941	304815	30860
28003	510000	0	0	481382	493358	71121
28716	340000	0	0	176743	873552	139808
29820	400000	1	0	394858	423903	396343
29867	340000	0	0	331641	300039	44855
29963	610000	0	0	322228	323014	348392

Negative bill amounts may refer to reimbursement for the returned product and prepay of the bills. Since the bill amount is significant considering the frequency and amount, this is not a sign mistake. Bill amount should be kept understanding the movement and consequence as credit default.

5) Data imbalance: There is a data imbalance with the dependent variable. The default 1=yes data is way too less than the default 0=No data, which makes it difficult to make an accurate prediction. The ratio of default is 22.15%.

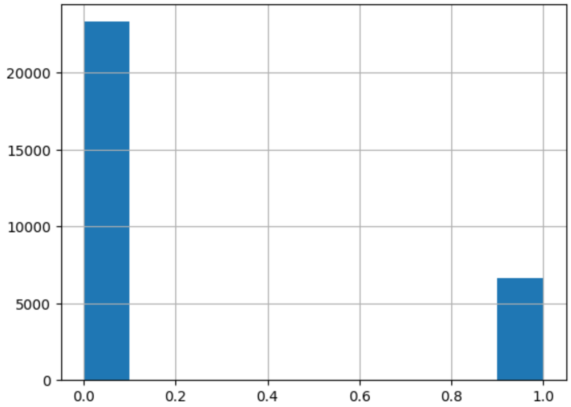

Figure 2: Default Payment Next Month data imbalance

It is important to make a balance as the other factors are important for the default payment variable. Demographic variable "SEX" is presented in perspective of default credit card payment.

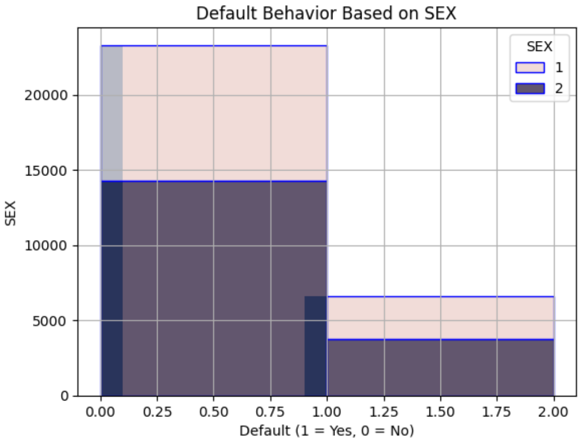

Figure 3: Default behavior based on SEX

As we can see in the graph the number of male non default is significantly lower than females. Females have higher number of credit default than the males. These data illustrations are not logical as the variables are not balanced.

5.0 Exploratory Data Analysis:

5.1 Visual Analysis:

There are numeric variables and categorical variables which are expressed as binary variables. The limit variable shows that the data is skewed to the left. There are limits which are really high and considered as outliers, but those data are legit because these customers have the money and because of that the limit is allocated. Here no data points are missing. LIMIT_BAL has all the data points well explained and placed.

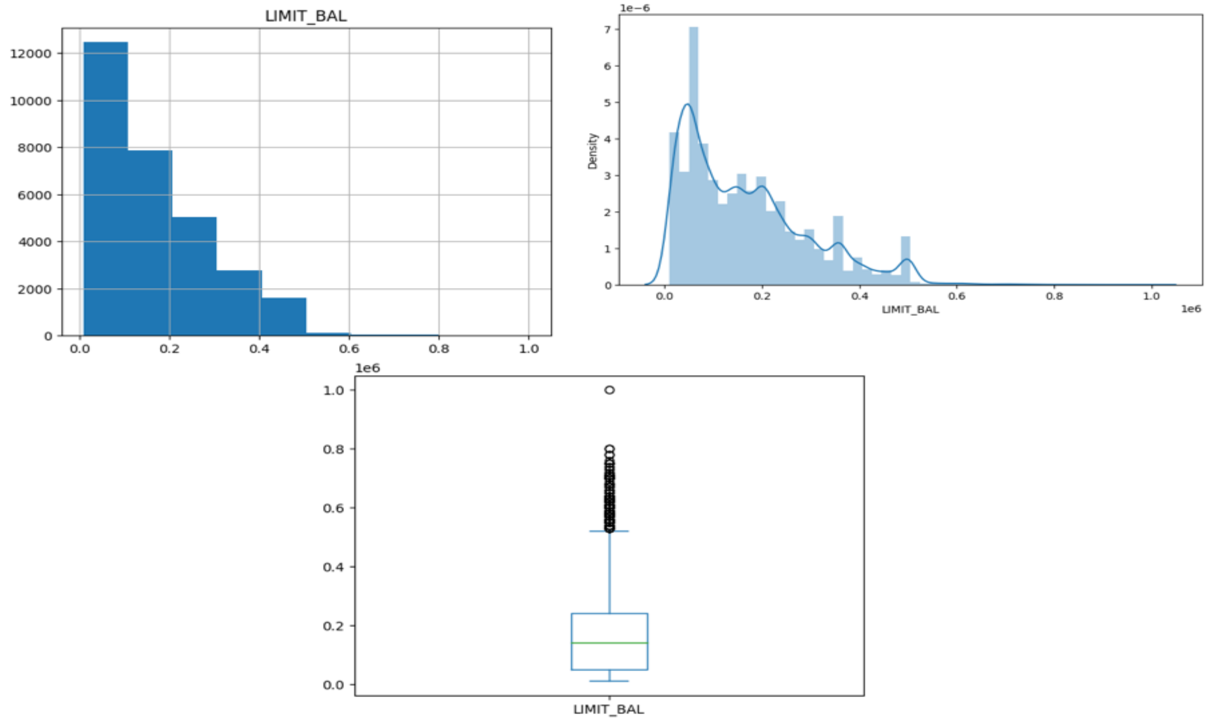

Figure 4: Bar Chart, distribution plot and boxplot of LIMIT_BAL

As it can be seen from the SEX dataset, there are more female participant data points then the male participant. There were some values labeled as 0 which was considered as missing data and removed from the dataset. From the dataset we can see that it is positively skewed.

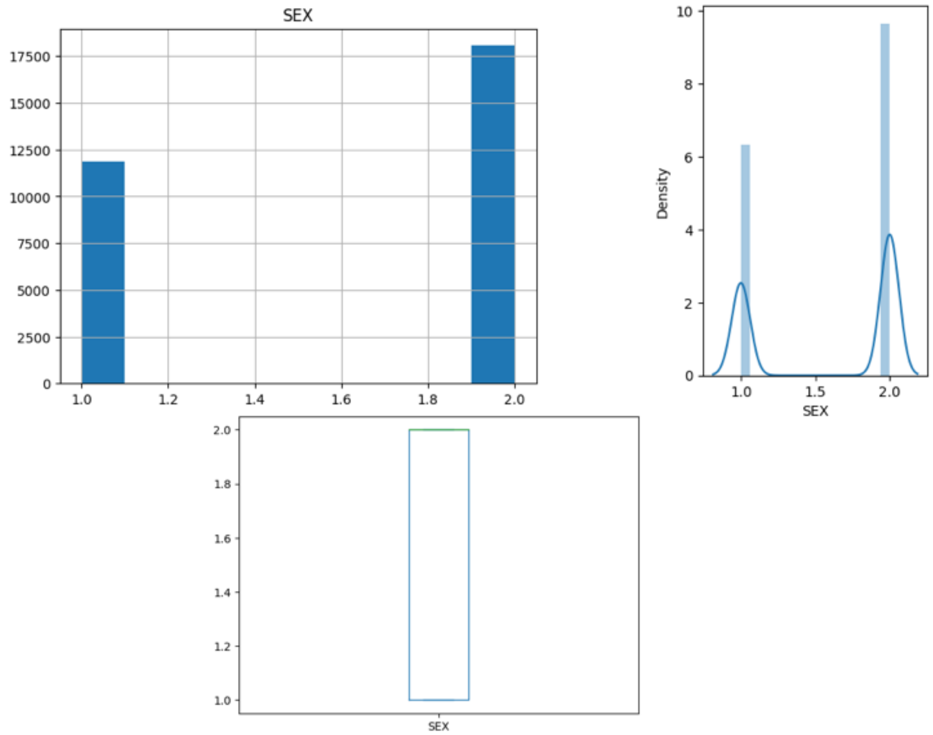

Figure 5: Bar Chart, distribution plot and boxplot of SEX

The variable Education's also a categorical data which had some unexplained values labeled as 0, 5 and 6. Label 0 was considered as the missing value and removed from the dataset.

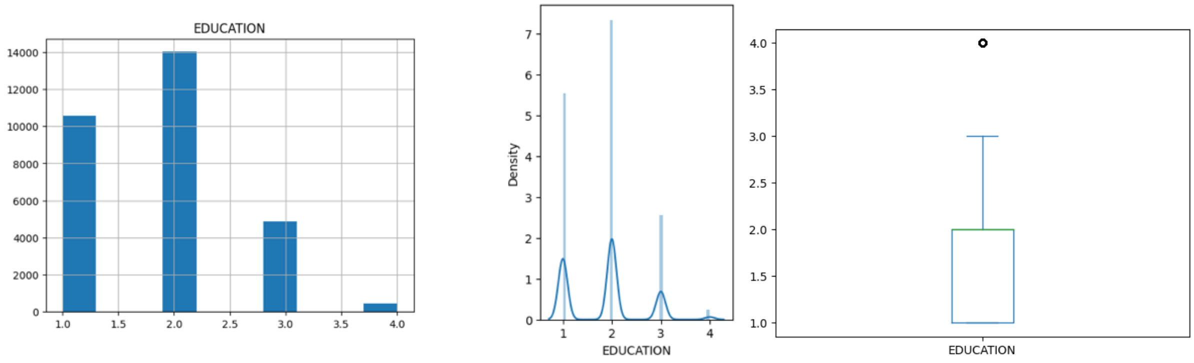

Figure 6: Bar Chart, distribution plot and boxplot of EDUCATION

As we can see that majority of the participants have completed graduate school and university label. The participants' education summary is provided below:

Table 4: Data distribution of EDUCATION

EDUCATION Details of the participants	Total	%
Graduate School	10581	35.50%
University	14024	46.85%
High School	4873	16.28%
Others	454	1.52%

In the dataset it is seen that the majority participant age is between 20 to 45 years of old. The curve will be skewed to the left. There are many outliers but these are important contributors for the analysis, and they are not any clerical mistake.

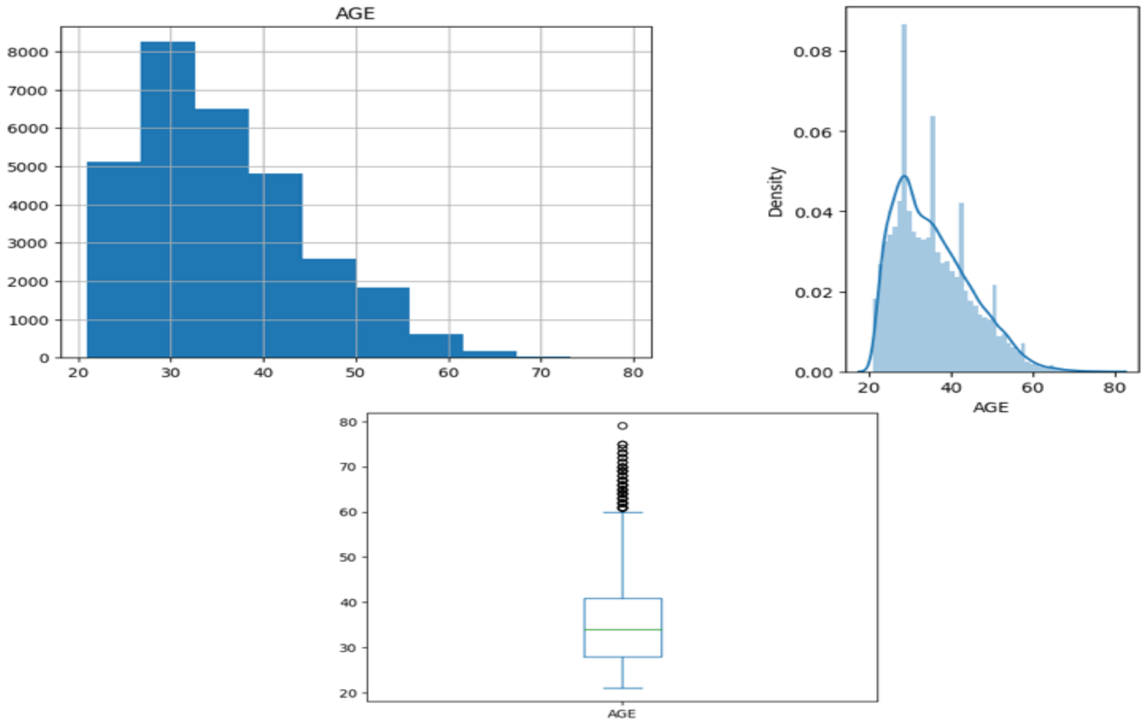

Figure 7: Bar Chart, distribution plot and boxplot of AGE

Variable PAY had some discrepancy regarding the data definition. The data label 0, -1 and -2 has been redefined as paid duly and all of them are defined as 0. After plotting we can see that most of the clients paid duly. The data is positively skewed.

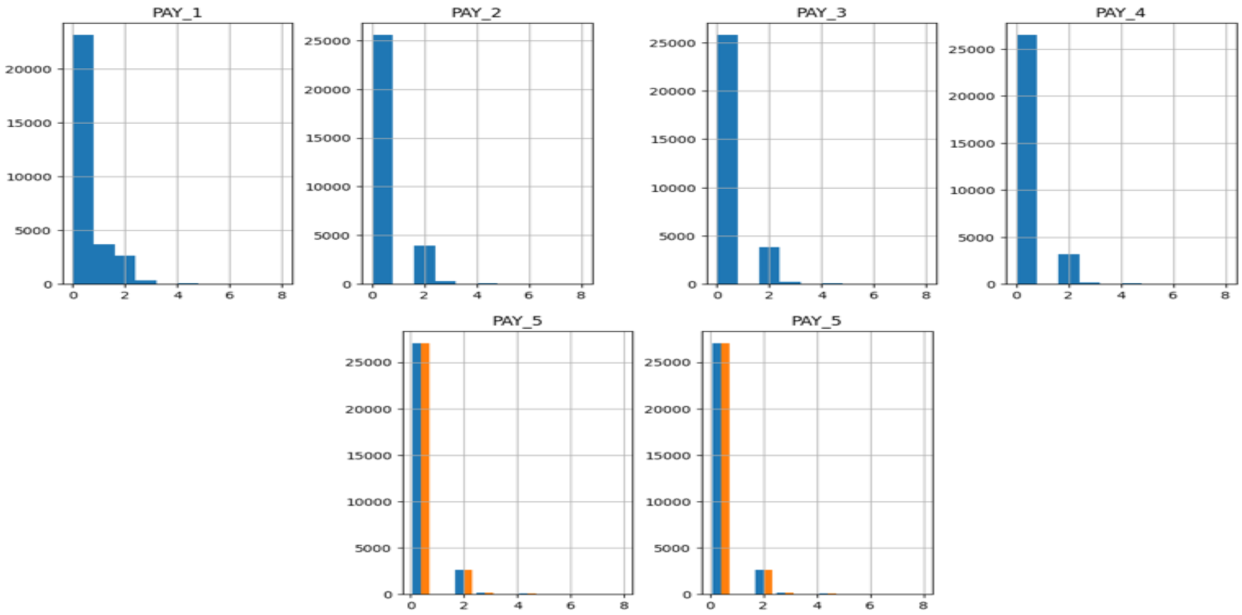

Figure 8: Bar Chart of PAY history

The BILL_AMT data shows us that most frequent bill amount is below 25,000 however there are many negative bill amounts which are prevalent in the later BILL_AMT data. We are keeping these anomalies for the purpose of a better predictive analysis. Negative bill payment occurs when there is a merchant refund or paid more than the due.

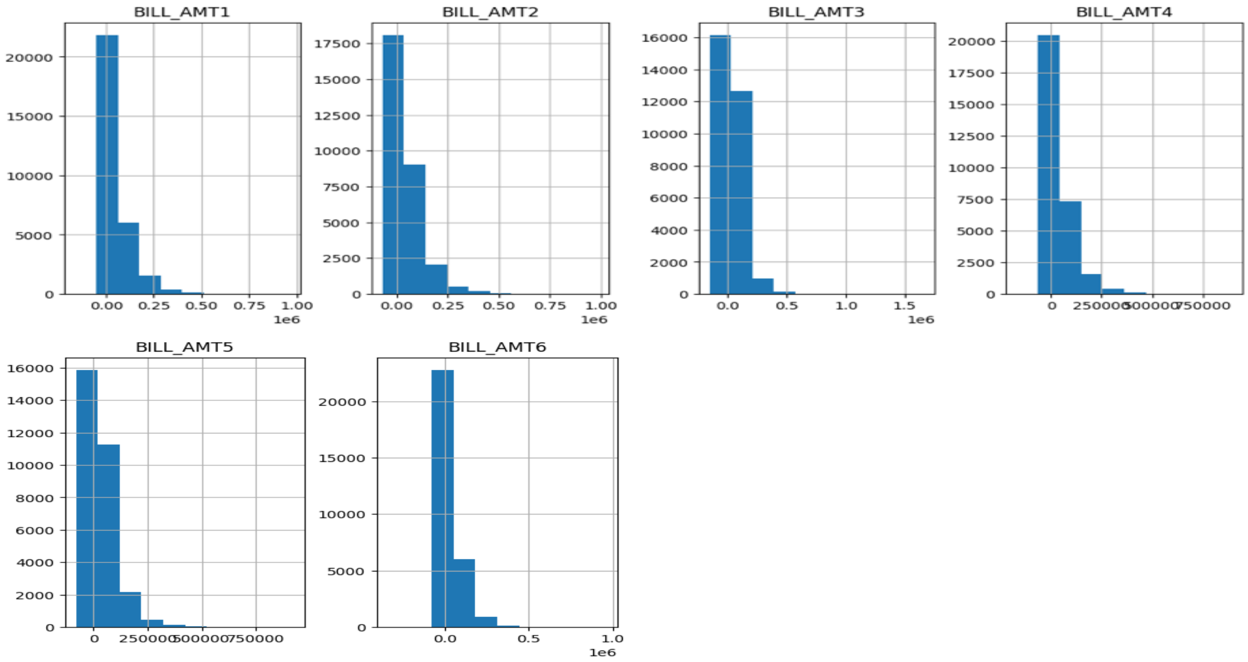

Figure 9: Bar Chart of BILL_AMT history

The PAY_AMT is most of the cases 0-10000 amount. There are some outliers for the bill payment but these outliers indicates that the customers gave high financial worth, high limit and high spending capability. The cross-table scatterplot will clear it out. Cross table pair-plot is presented below for BILL_AMT to PAY_AMT.

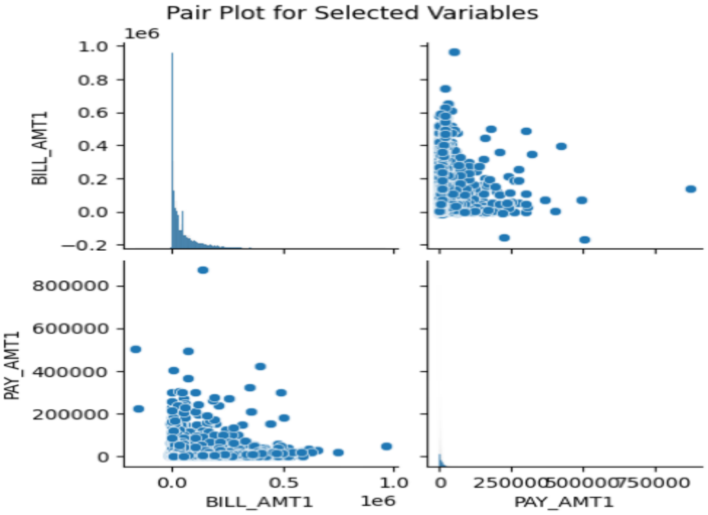

Figure 10: Pair plot between BILL_AMT1 and PAY_AMT1

The PAY_AMT bar charts are shown below:

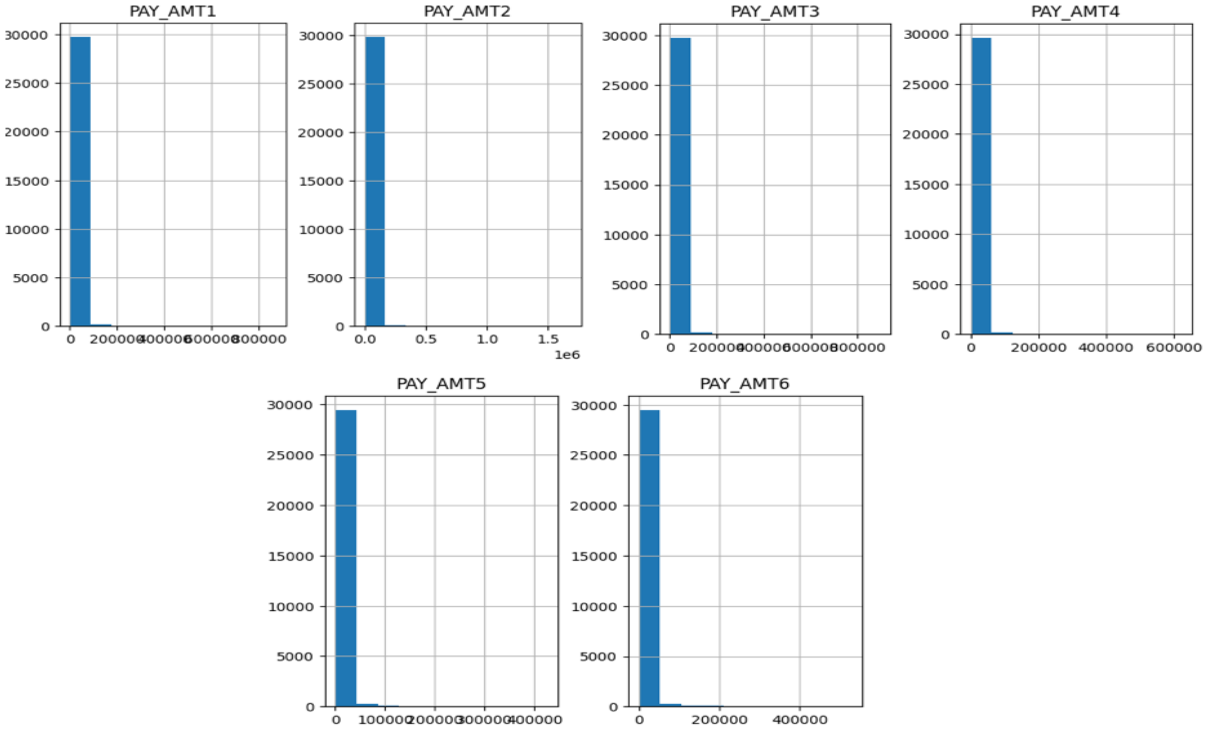

Figure 11: Bar charts of PAY_AMT

Boxplots for PAY_AMT is provided below.

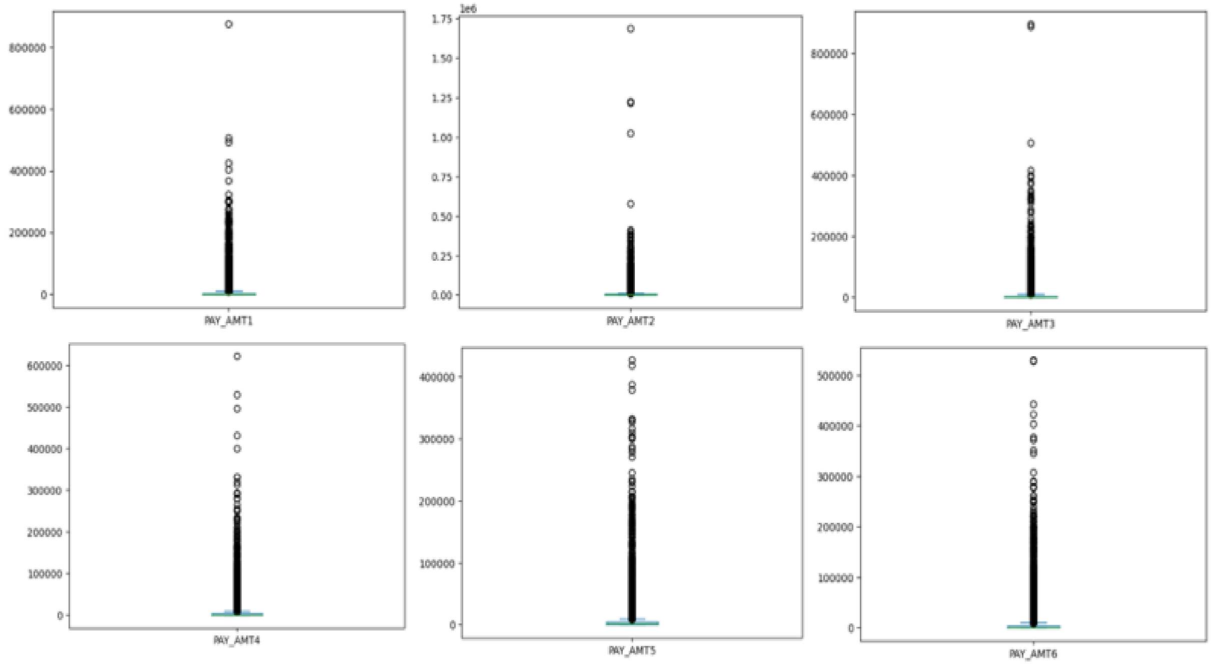

Figure 12: Boxplots of PAY_AMT

6.0 Correlation Analysis:

The variables in the dataset can be categorized into two types: demographic and payment behaviour. Previous study shows that demographic variables have some impact on the credit default. Education has the strongest affiliation with the credit default (Memarista, Malelak and Anastasia, 2015). Here LIMIT_BAL will be considered as the dummy standard variable to measure the correlation with the demographic variable and another correlation analysis will be done for the default payment on the next month. As the data types are mixed, for correlation the data was normalized.

In the pearson correlation analysis between LIMIT _BAL and demographic variables it is seen that AGE, EDUCATION and MARRIAGE have relationship regarding limit allocation. Limit_BAL is used as a standard variable to understand the relationship between demographic variables and financial factor. The pearson correlation table with factors are provided below:

Table 5: Pearson Correlation between Democratic variables and LIMIT_BAL

Demographic variables	AGE	EDUCATION	MARRIAGE	SEX
Pearson correlation	0.1448024855792761	-0.2317397640347224	-0.1106832473738504	0.024952818456164105

The Correlation scatterplots are provided below:

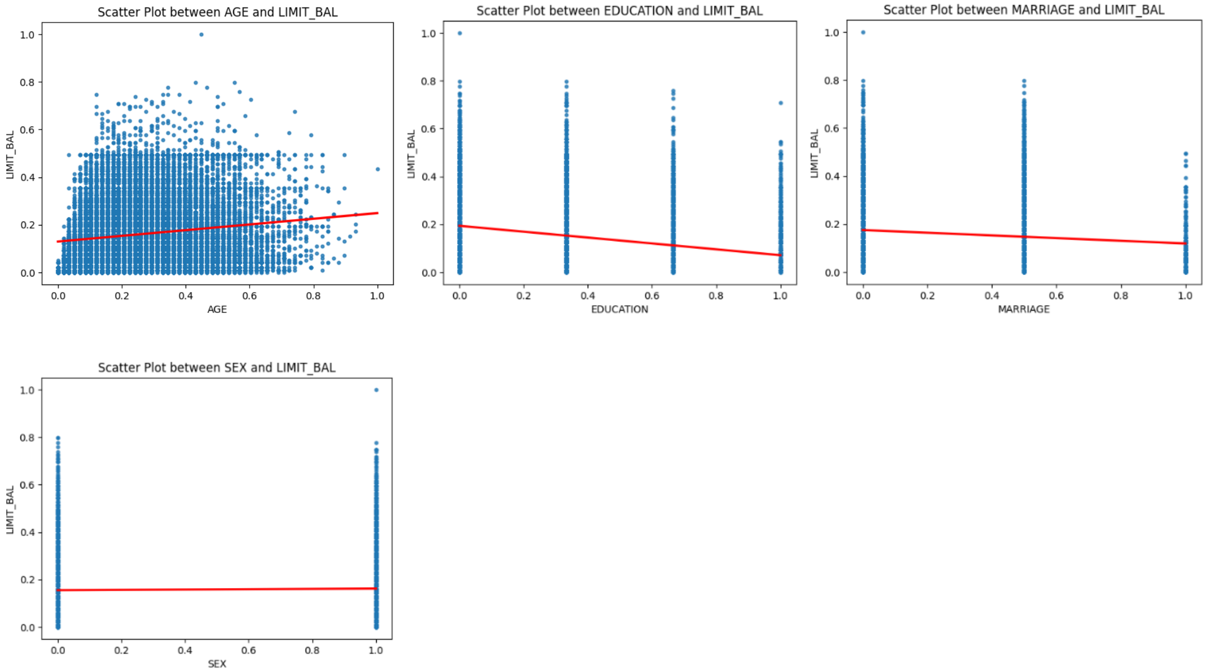

Figure 13: Pearson correlation between demographic variables and LIMIT_BAL

As per the correlation plots it is seen that AGE has positive correlation with LIMIT_BAL and EDUCATION and MARRIAGE has negative correlation. SEX has very low correlation with LIMIT_BAL and the correlation is negative. So, we will omit SEX in our further analysis.

6.1 Heat map for Demographic variables and LIMIT_BAL:

To identify the relationship between demographic variables and LIMIT_BAL, and correlation heat map is constructed where the data is normalized. In the heat map we can see that AGE and SEX have positive correlation with LIMIT_BAL, and EDUCATION and MARRIAGE have negative relationship with LIMIT_BAL. However, the correlation of SEX and LIMIT_BAL is too low. So, for the modelling we will not consider sex from this heat map perspective.

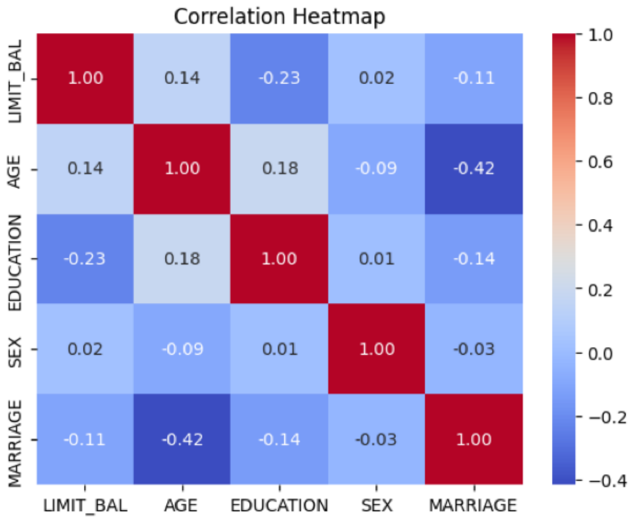

Figure 14: Heat map for Demographic variables and LIMIT_BAL

6.2 Heat map for Demographic variables and default payment on the next month:

The heat map shows negative correlation with default payment and SEX, all the other correlations are positive. However, all the correlations are too small to consider. So, we are going to consider the heat map of the LIMIT_BAL.

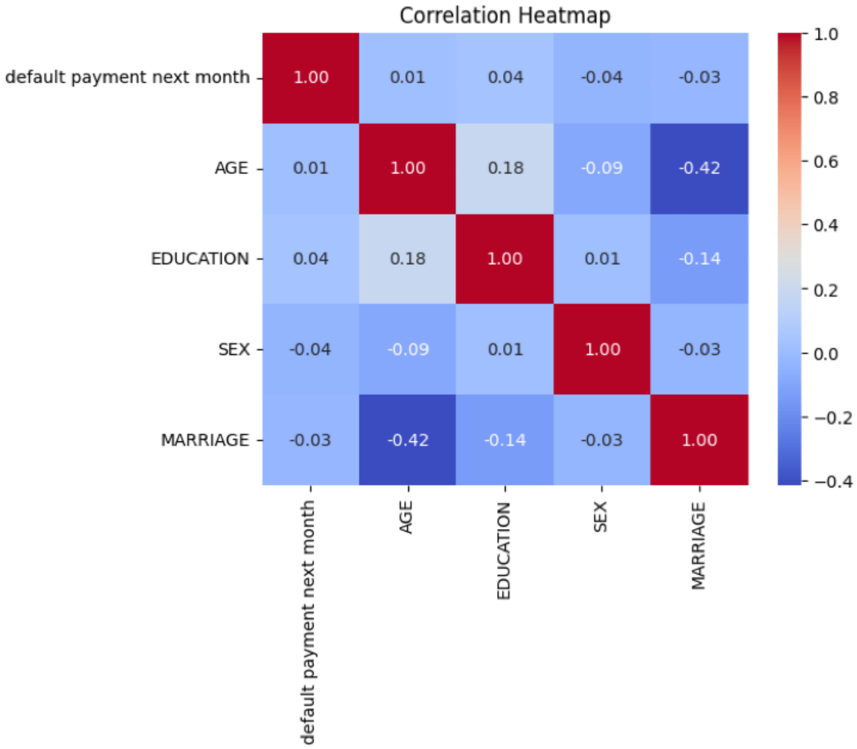

Figure 15: Heat map for Demographic variables and default payment next month

The pair plot visualizes the relationship between variables. Below the pair plots are placed to see the correlation. However, this is not the final visualization as the data is imbalanced. So, we will have to balance the data before running the model.

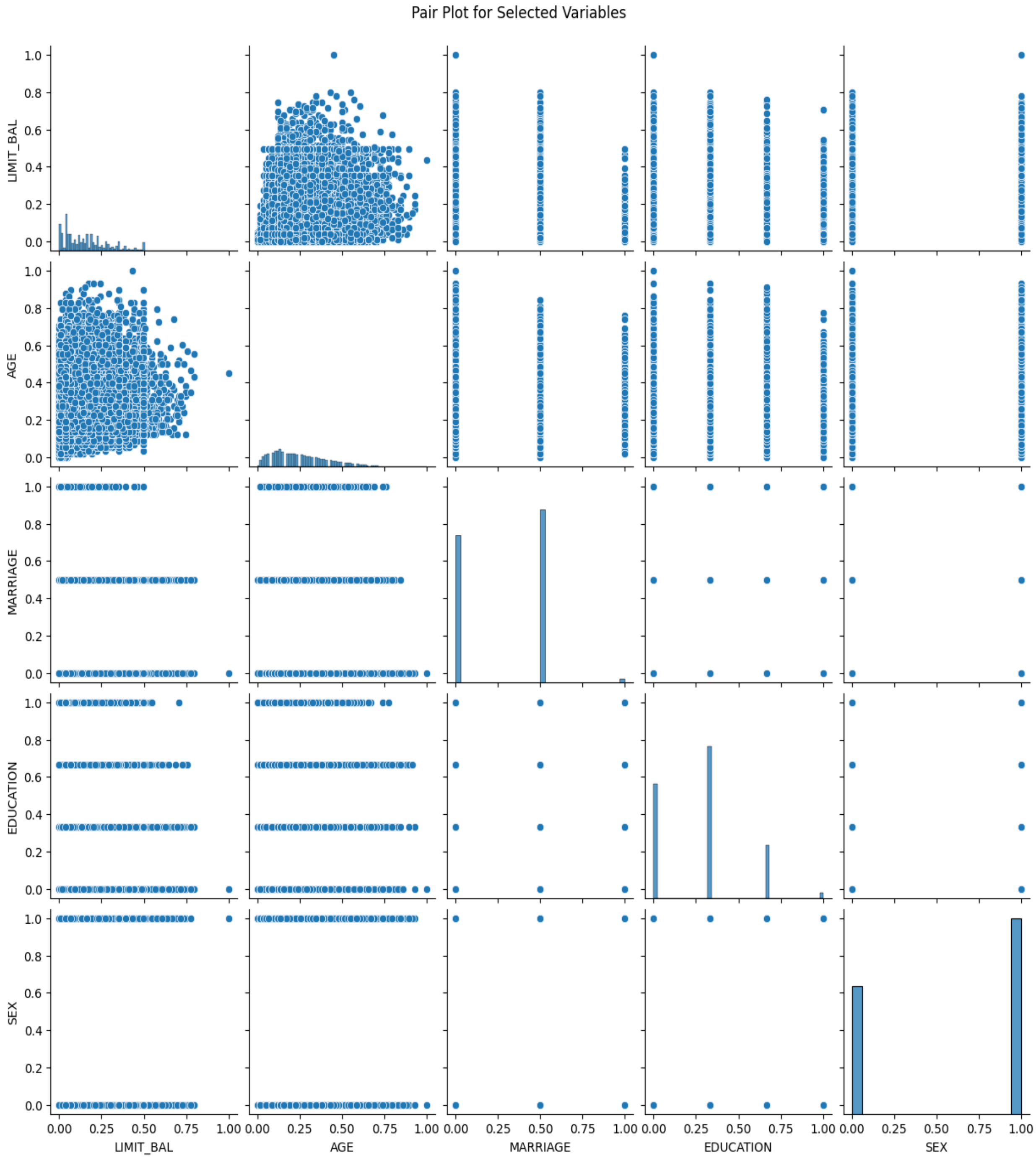

Figure 16: Pair plot for Demographic variables and LIMIT_BAL

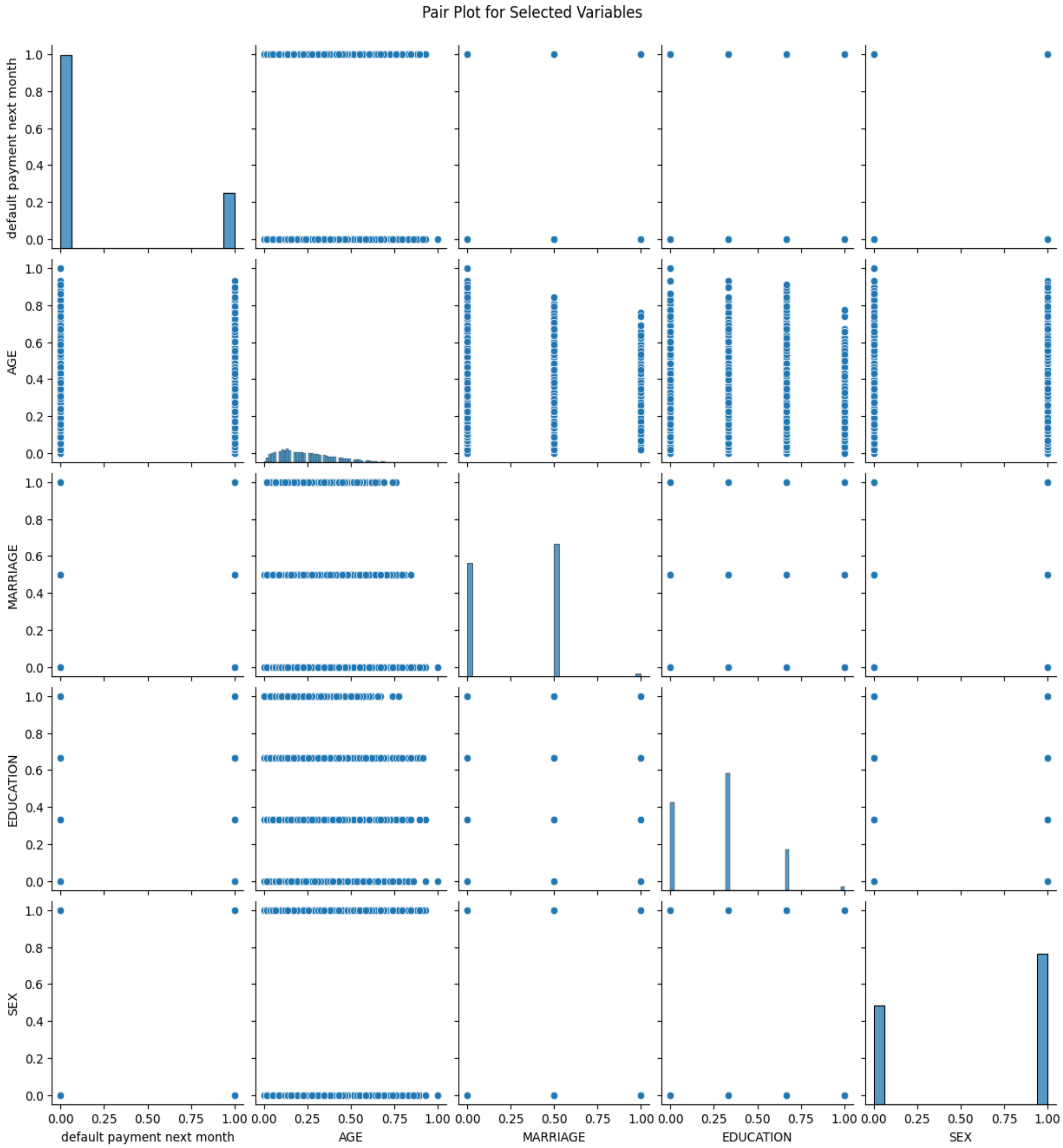

Figure 17: Pair plot for Demographic variables and default payment next month

7.0 Data Balancing:

After data cleaning and data aligning, there are 23945 data points. In the dataset, there are 18641 data which is 0=not default and 5304 data which is 1= default. As it is seen that there is a huge data imbalance, so the model will provide wrong interpretation of the data, thus provide the wrong result. So, it is necessary to balance the data so that the data will have the balance to provide more accurate result. The data is divided into train-test dataset (80/20) to conduct the research model. As it is seen that the data is imbalanced, there are two approaches to balance the data. One approach is up-sampling and other one is down-sampling the data. There are some drawbacks for up-sampling such as the data which are combined with the original one is the synthetic data. And down-sampled data has the issue of sample bias as most of the data is not there and not all the attributes are captured. So bias is created. We will consider both datasets to see the effective results with all three algorithms.

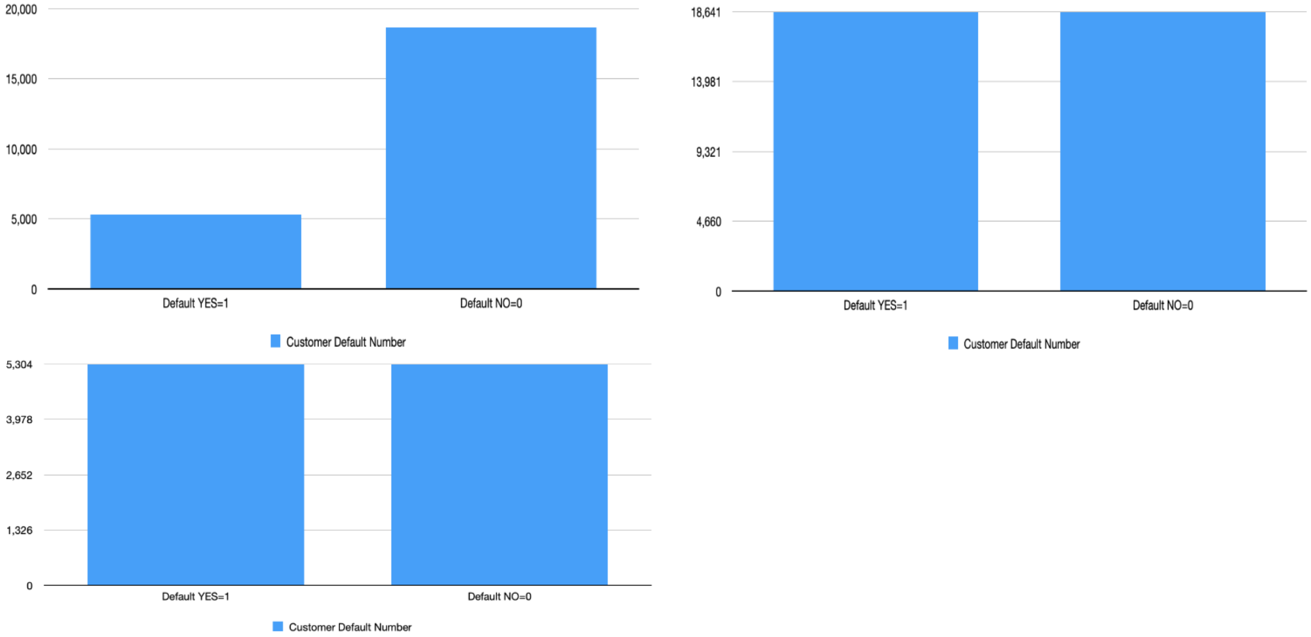

Figure 18: Data balancing of unbalanced data, data up-scaling and data down-scaling

8.0 Principal Component Analysis:

This is a dimension reduction method which considers large dataset with large number of variables and transform the dataset with smaller number of variables which contains most of the information and major attributes of the dataset. For principal component analysis the data was scaled using StandardScaler.

8.1 Feature Selection

In the data preprocessing stage, the features need to be selected for the machine learning algorithm. In the heat map of correlation we have already seen the variables which are strongly correlated or not. In the correlation analysis we have seen that BILL_AMT features are very strongly co-related, so we are discarding the BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6. The heat-map for BILL_PMT is provided below.

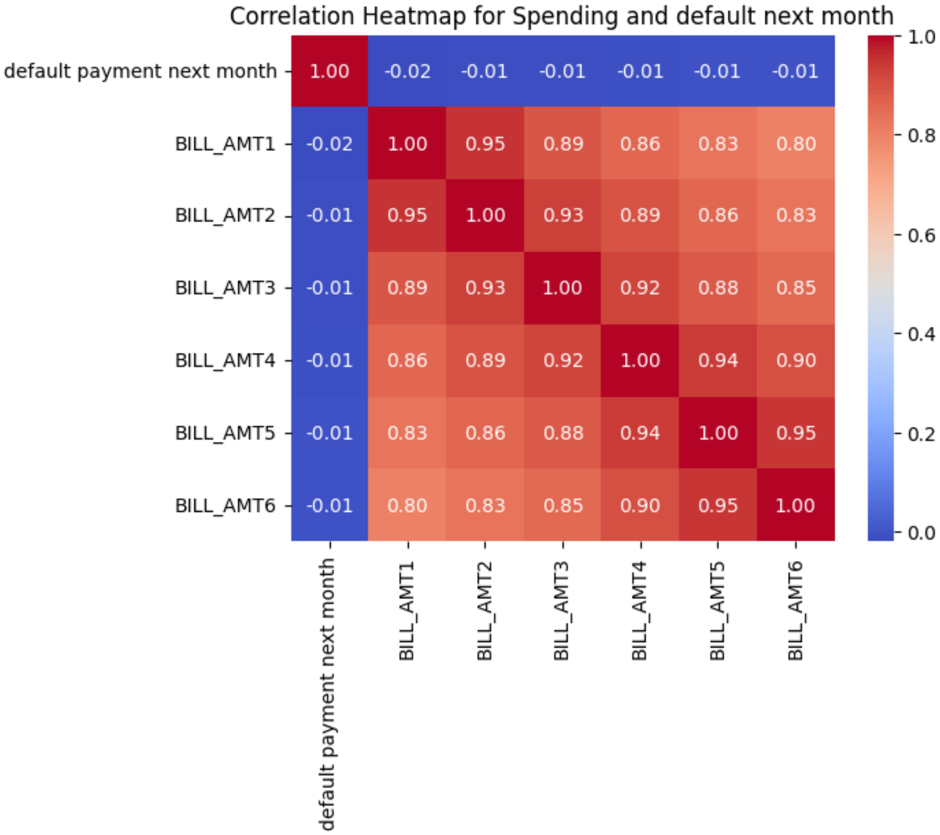

Figure 19: Heat map of BILL_AMT and default payment next month

The study will conduct PCA with 5 components. The variance ratios are provided below along with the graph.

Table 6: Principal components variance ratios

PC	PC1	PC2	PC3	PC4	PC5
Variance Ratio	0.2467	0.1298	0.0952	0.0654	0.0597

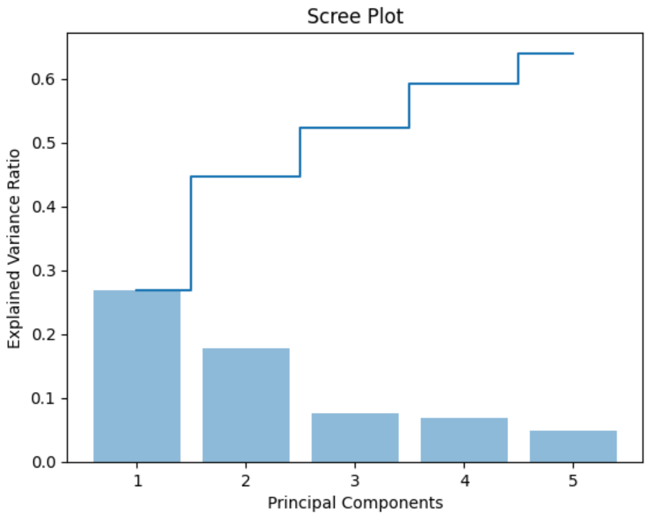

Figure 20: PC Explained variance ratio plot

For better illustration, the PCA components are unified with the original rectified data. The data heads for PCA components are provided below:

Table 7: Principal components dataset head

Serial No.	PC1	PC2	PC3	PC4	PC5
1	-1.426846	-0.932080	-0.056233	0.537666	-2.200063
2	-1.376642	-0.560051	1.136600	0.346869	0.120018
3	0.633666	-0.972101	0.545118	0.511656	0.094228
4	0.581719	-0.944610	-0.881759	0.297509	0.264787
5	0.826469	0.208075	-2.189227	0.603895	0.521653

9.0 Data Analysis process:

The Analysis process consists of 3 stages:

1) Data preprocessing
2) Model Training and evaluation
3) Comparison and conclusion

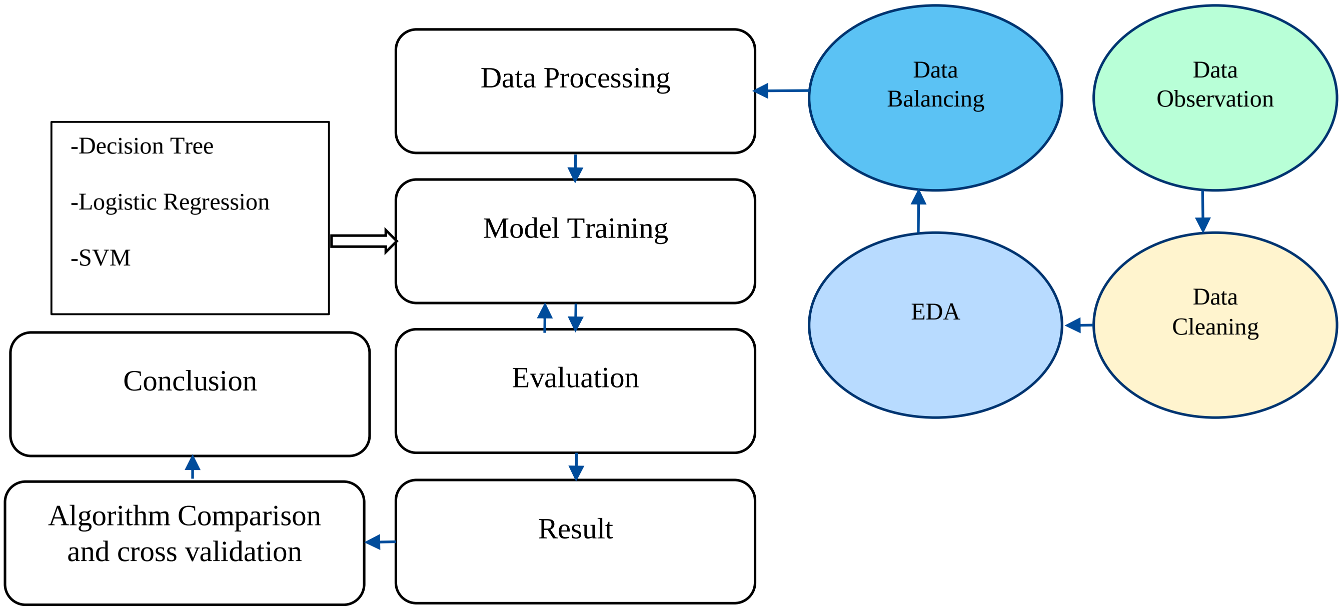

Figure 21: The Data Analysis approach and process

The study will be conducted based on six broad steps. Before proceeding on the broad steps, the dataset is prepared for the model training. First the data was fetched from the source and observed to check the data types, numbers, attributed and missing data/anomalies. For fetching and observing and analyzing the data, python 3 was used. After identifying the anomalies, unexplained data and data errors the data was cleaned, data labels were fixed, and data noises were removed. After having a clean data, Exploratory data analysis was conducted to check the data trend, curves, data quartiles, outliers, mean and median, correlation with the LIMIT_BAL and default pay next month. Based on the data pattern and correlation, in the data pre-processing stage, data was prepared for the model run. For the model run, data was divided into train-test split (80%-200%). To preserve the accuracy of the data analysis, up-sampling and down-sampling was done to the datasets so that data imbalance was averted. For this, Sklearn, matplotlib, panda, numphy and seaborn packages were used. The principal component analysis was also done so that the data features stayed intact but the data dimensions can be reduced. Before doing these, the dataset is normalized and scaled so that the data points fit within appropriate scale and higher value data points will not dominate when distances will be calculated. This is how the dataset was prepared. The second stage of the study is to construct the model for data analysis. On the next stage the study will conduct three model algorithms to identify the predictability. Decision tree, logistic regression and Byes Classifier will be conducted, and the models will be interpreted based on the output. After having the result, the data will be evaluated. We have the up-sampled and down-sampled dataset to compare the results when it will be required. If there is any anomaly found in the model result evaluation, the model will be conducted again for better accuracy. After proper evaluation there will be a cross validation and algorithm comparison. After the algorithm comparison the proper algorithm and predictability will be identified with conclusion.

9.1 Modelling algorithms:

9.1.1 Decision Tree:

Decision tree is a flowchart-based algorithm where variables of the datasets are placed as nodes which were processed to the next staged nodes based on the decision branched. With the concept of decision making to the nodes the conclusion is derived from all the variables. Our dataset has dependent variable which is binary, and target will be to reach the final leaf node.

9.1.2 Logistic regression:

Logistic regression is the algorithm that estimates the probability of occurring an event based on the independent variables. The outcome of logistic regression will be yes and no. It can be addressed as binary variable. In our study we will try to identify the predictability of credit card payment default on the next month based on demographic and payment variables which are independent variables.

9.1.3 Support Vector Machine (SVM):

SVM is a supervised learning model for classification problem analysis. The purpose of SVM is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space (GeeksforGeeks, 2023). For this study we can us this algorithm for better explain the credit default in future.

References

1. Achsan, W., Achsani, N. A., & Bandono, B. (2022). The demographic and behavior determinant of credit card default in Indonesia. *Signifikan: Jurnal Ilmu Ekonomi*, *11*(1), 43–56. https://doi.org/10.15408/sjie.v11i1.20215
2. Çallı, B. A., & Coşkun, E. (2021). A longitudinal systematic review of Credit Risk Assessment and credit default predictors. *SAGE Open*, *11*(4), 215824402110613. https://doi.org/10.1177/21582440211061333
3. GeeksforGeeks. (2023, June 10). *Support Vector Machine (SVM) algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/
4. Goel, A., & Rastogi, S. (2021). Understanding the impact of borrowers' behavioural and psychological traits on credit default: Review and Conceptual Model. *Review of Behavioral Finance*, *15*(2), 205–223. https://doi.org/10.1108/rbf-03-2021-0051
5. Memarista, G., Malelak, M., & Anastasia, N. (n.d.). *The Relationship between Demographic Factors and Financial Behavior on Credit Card Usage in Surabaya*. https://core.ac.uk/ download/pdf/32453075.pdf
6. Rauf, A., Razi, A., Khalid, A., & Hassan, Y. (2022). The usage patterns of credit/debit card across various demographics. *Pakistan Journal of Humanities and Social Sciences*, *10*(2). https://doi.org/10.52131/pjhss.2022.1002.0253
7. Soman, D., & Cheema, A. (2002a). The effect of credit on spending decisions: The role of the credit limit and credibility. *Marketing Science*, *21*(1), 32–53. https://doi.org/10.1287/mksc.21.1.32.155
8. Wang, L., Lu, W., & Malhotra, N. K. (2011). Demographics, attitude, personality and credit card features correlate with credit card debt: A view from China. *Journal of Economic Psychology*, *32*(1), 179–193. https://doi.org/10.1016/j.joep.2010.11.006

9. Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

10. Yeh,I-Cheng. (2016). Default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H

Github Link

Github Link: https://github.com/Cattitude101/CIND-820-Project/tree/main