Predictive Analysis on Credit Card Defaults Based on demographic Factors and Payment Behavior

CIND 820 XJH W2024

Project by: Md Fahim Ferdous

ID: 501232653

Supervisor: Dr. Ceni Baboglu

Date of submission: April 01, 2024

Ryerson University

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 behavior

The study is about finding out the demographic and payment behavior 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 quantified predictability of credit card default based on spending pattern, demographic behavior and limit allocation, which is evaluated by demographic and payment behavior pattern?

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: Histogram 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 demographic vulnerabilities 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 behavior, 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 behaviors 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 Behavior 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 behavior and demographic factors on credit card default. The result of this research shows that the financial behavior on credit card usage is low. From the demographic factors, the education has significant relationship with financial behavior on credit card usage. However, the study did not find significant relationship between financial behavior 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 behavioral 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 behavior. Here the authors considered factor groups such as Socioeconomic, Demographic, Educational, Institutional/financial, Personality, Values/attitudes/behavioral, Situational, Macroeconomic, Health-related and Alternative. This study conducted different data mining techniques and it reached to a conclusion that personality and behavioral 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 behavioral

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 behavioral 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 behaviors are impacting the limit allocation and what leads to credit card default.

The demographic and behavior determinant of credit card default in Indonesia (Achsan et al., 2022) tried to analyze the behavioral and demographic impact on nonperforming credit card. This study tried to find out the influential demographic and behavioral factors which can later be use for credit scores done by Indonesian banks. The authors conducted logistic regression model and found out that cardholder behavior is more likely to contribute to the nonperforming credit card rather than demographic behavior. 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 behavior 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 behavioral 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 behavioral pattern are well. This study will consider the factor “limit” as a proxy for the spending behavior 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' behavioral and psychological traits on credit default: review and conceptual model (Goal and Rastogi, 2021) focused on certain behavioral 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 behavioral 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 unlabeled. 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 participants, 15964 single participants 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 attributes and quantitative status

Attributes	Type	Maximum (if applicable	Minimum (if applicable	Mean	Standard Deviation	Missing/Unexplained values
LIMIT_BAL	Quantitative	1000000	10000	167484.32	129747.66	No
SEX	Categorical	2	1	NA	NA	No
EDUCATION	Categorical	4	1	NA	NA	345
MARRIAGE	Categorical	3	1	NA	NA	54
AGE	Quantitative	79	21	35.49	9.22	No
PAY_0 TO PAY_6	Categorical	8.0	-0.2911 to -0.0167	-2.0	1.12 to 1.19	Unexplained data label -2 and 0
BILL_AMT1-BILL_AMT6	Quantitative	1664089 to 891586	-170000 to -69777	51223.33 to 38871.76	73635.86 to 59554.11	Unexplained data negative bill amounts
PAY_AMT1 to PAY_AMT6	Quantitative	1684259 to 426529	0.0	4799.39 to 5921.16	23040.87 to 15278.31	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	30000.0	15000.5	8660.40	1.0	7500.7	15000.5	22500.25	30000.0
LIMIT_BAL	30000.0	167484.32	129747.66	10000.0	50000.0	140000.0	240000.	1000000.0
SEX	30000.0	1.60	0.49	1.0	1.0	2.0	2.0	2.0
EDUCATION	30000.0	1.85	0.79	0.0	1.0	2.0	2.0	6.0
MARRIAGE	30000.0	1.55	0.52	0.0	1.0	2.0	2.0	3.0
AGE	30000.0	35.49	9.22	21.0	28.0	34.0	41.0	79.0
PAY_0	30000.0	-0.017	1.12	-2.0	-1.0	0.0	0.0	8.0
PAY_2	30000.0	-0.13	1.19	-2.0	-1.0	0.0	0.0	8.0
PAY_3	30000.0	-0.16	1.20	-2.0	-1.0	0.0	0.0	8.0

PAY_4	30000.0	-0.22	1.17	-2.0	-1.0	0.0	0.0	8.0
PAY_5	30000.0	-0.27	1.13	-2.0	-1.0	0.0	0.0	8.0
PAY_6	30000.0	-0.29	1.150	-2.0	-1.0	0.0	0.0	8.0
BILL_AMT1	30000.0	51223.33	73635.86	-165580.0	3558.75	22381.5	67091.0	964511.0
BILL_AMT2	30000.0	49179.08	71173.77	-69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	30000.0	47013.15	69349.39	-157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	30000.0	43262.95	64332.86	-170000.0	2326.75	19052.0	54506.0	891586.0
BILL_AMT5	30000.0	40311.40	60797.16	-81334.0	1763.0	18104.5	50190.5	927171.0
BILL_AMT6	30000.0	38871.76	59554.11	-339603.0	1256.0	17071.0	49198.25	961664.0
PAY_AMT1	30000.0	5663.58	16563.28	0.0	1000.0	2100.0	5006.0	873552.0
PAY_AMT2	30000.0	5921.16	23040.87	0.0	833.0	2009.0	5000.0	1684259.0
PAY_AMT3	30000.0	5225.68	17606.96	0.0	390.0	1800.0	4505.0	896040.0
PAY_AMT4	30000.0	4826.08	15666.16	0.0	296.0	1500.0	4013.25	621000.0
PAY_AMT5	30000.0	4799.39	15278.31	0.0	252.5	1500.0	4031.5	426529.0

PAY_AMT6	30000.0	5215.50	17777.47	0.0	117.75	1500.0	4000.0	528666.0

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 data point. 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 data point for MARRIAGE and 14 data point 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 data points 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.

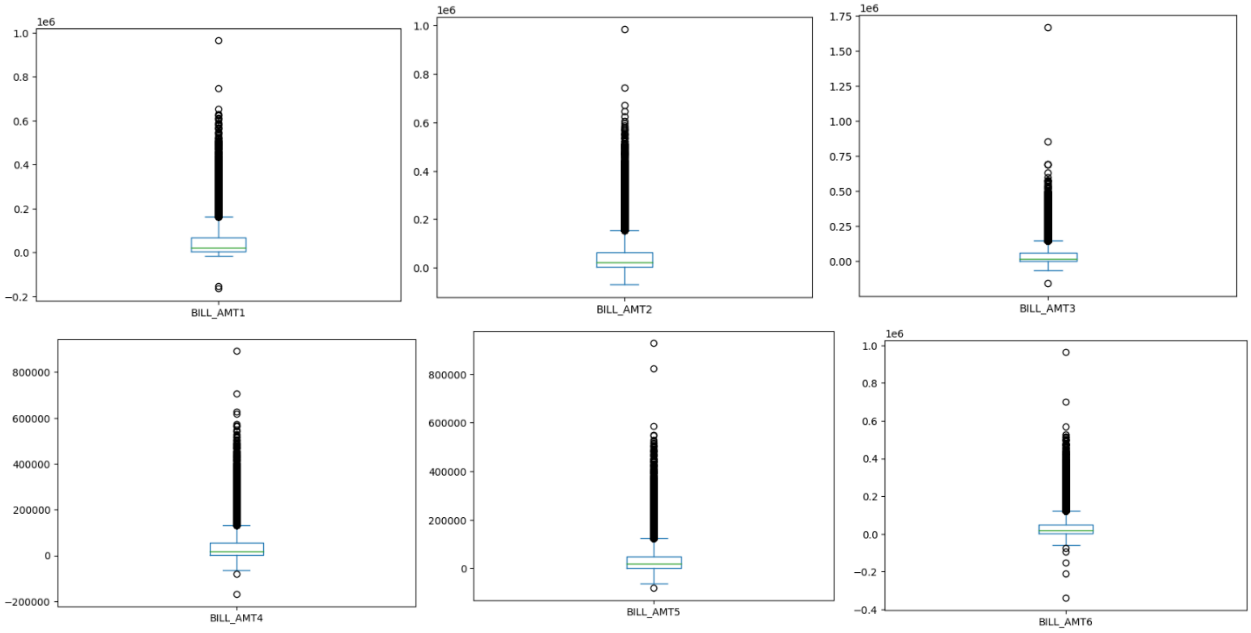


Figure 1: Boxplots outliers of BILL_AMT

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
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%.

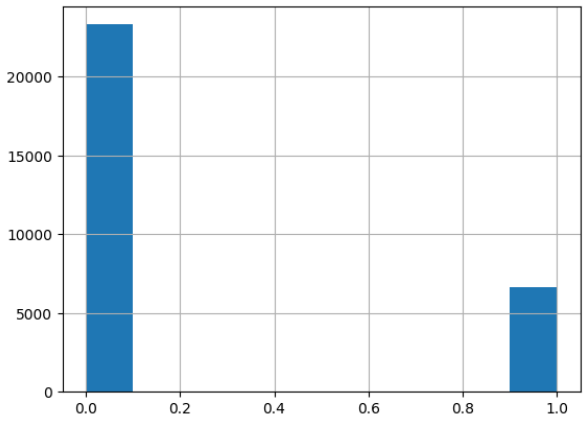


Figure 2: Default Payment Next Month data imbalance

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.

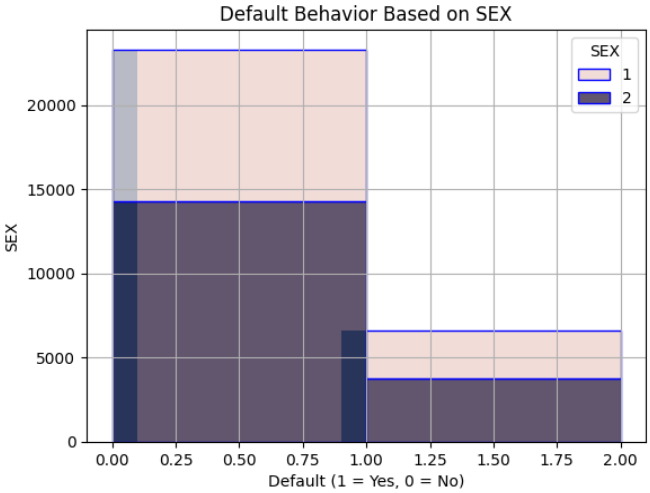


Figure 3: Default behavior based on SEX

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.

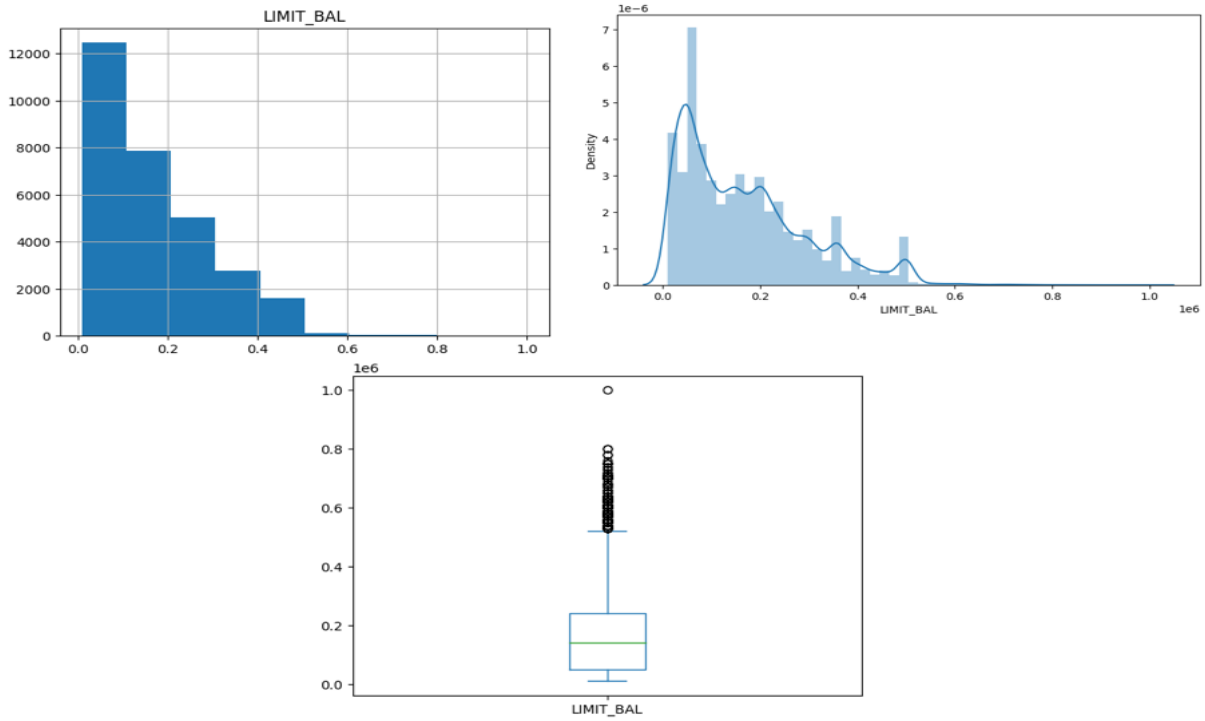


Figure 4: Bar Chart, distribution plot and boxplot of LIMIT_BAL

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.

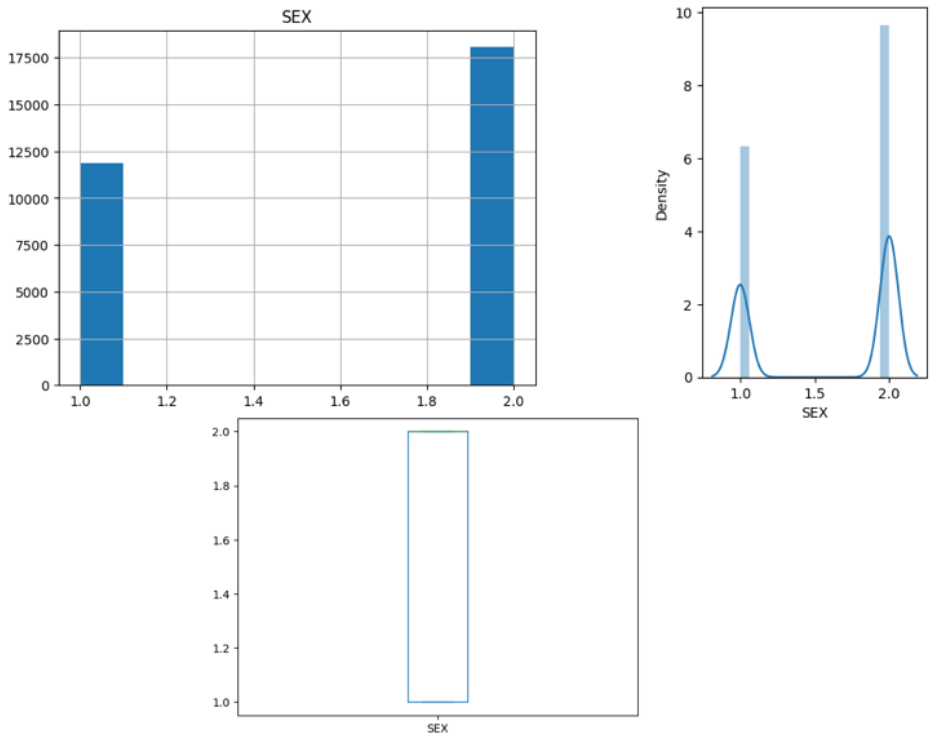


Figure 5: Bar Chart, distribution plot and boxplot of SEX

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.

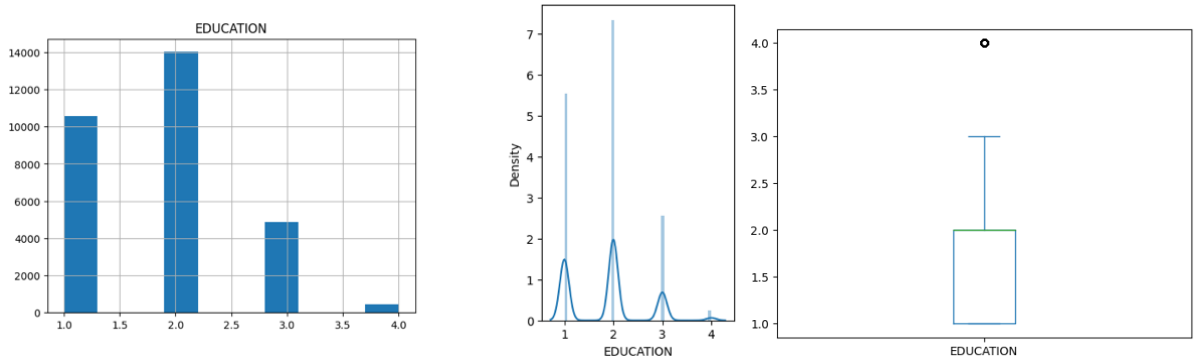


Figure 6: Bar Chart, distribution plot and boxplot of EDUCATION

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.

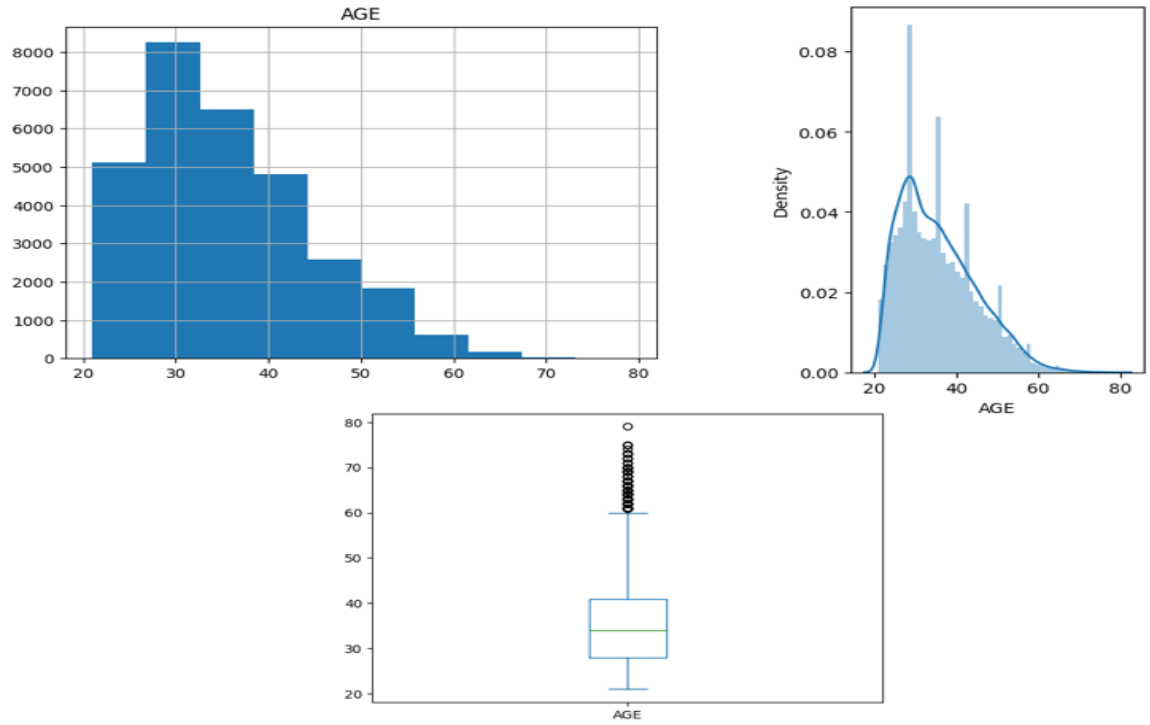


Figure 7: Bar Chart, distribution plot and boxplot of AGE

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.

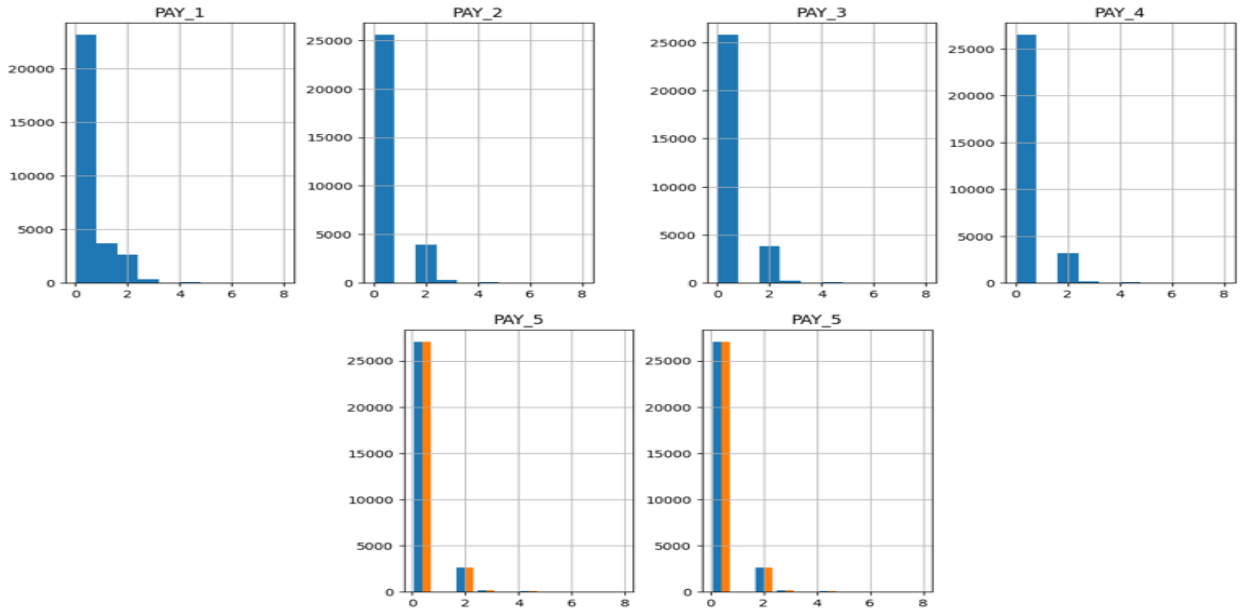


Figure 8: Bar Chart of PAY history

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.

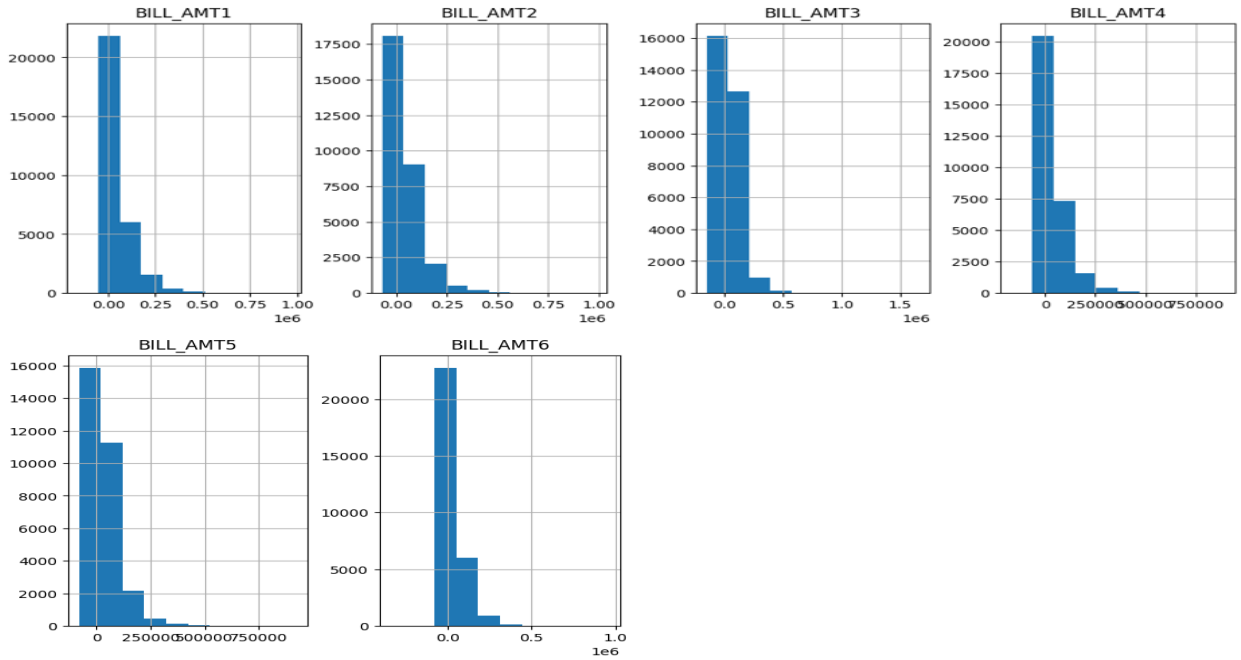


Figure 9: Bar Chart of BILL_AMT history

The PAY_AMT is most of the cases 0-10000 amount. There are some outliers for the bill payment but these outliers indicate 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.

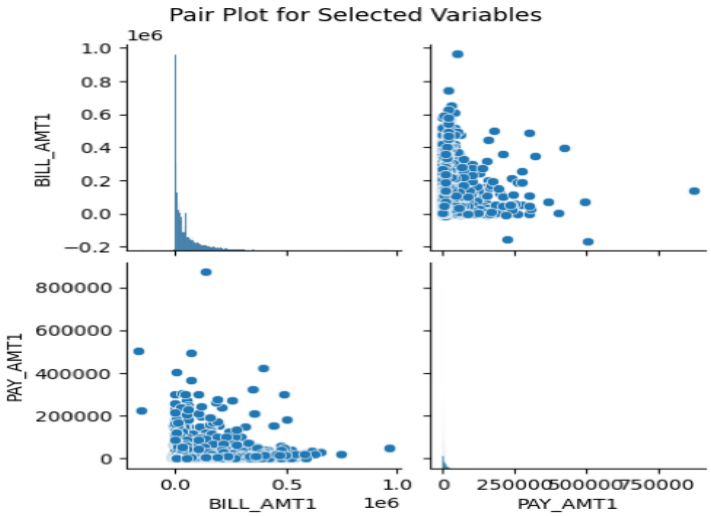


Figure 10: Pair plot between BILL_AMT1 and PAY_AMT1

The PAY_AMT bar charts are shown below:

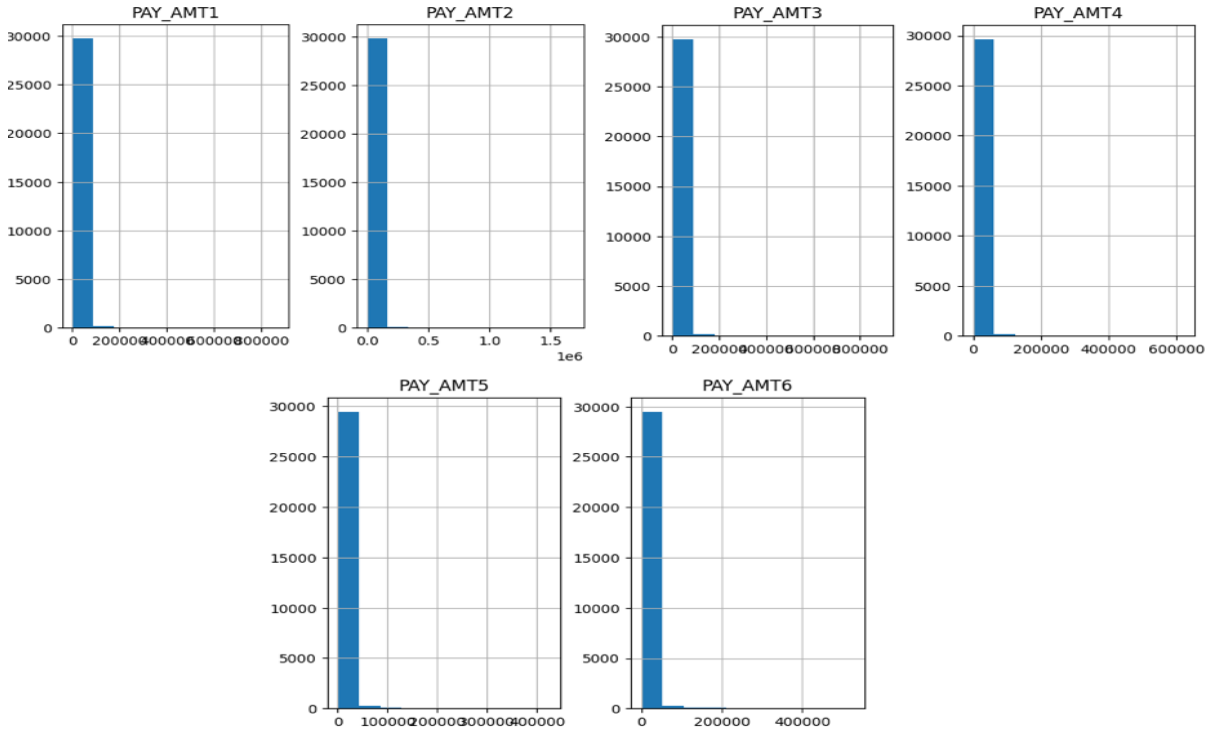


Figure 11: Bar charts of PAY_AMT

Boxplots for PAY_AMT is provided below.

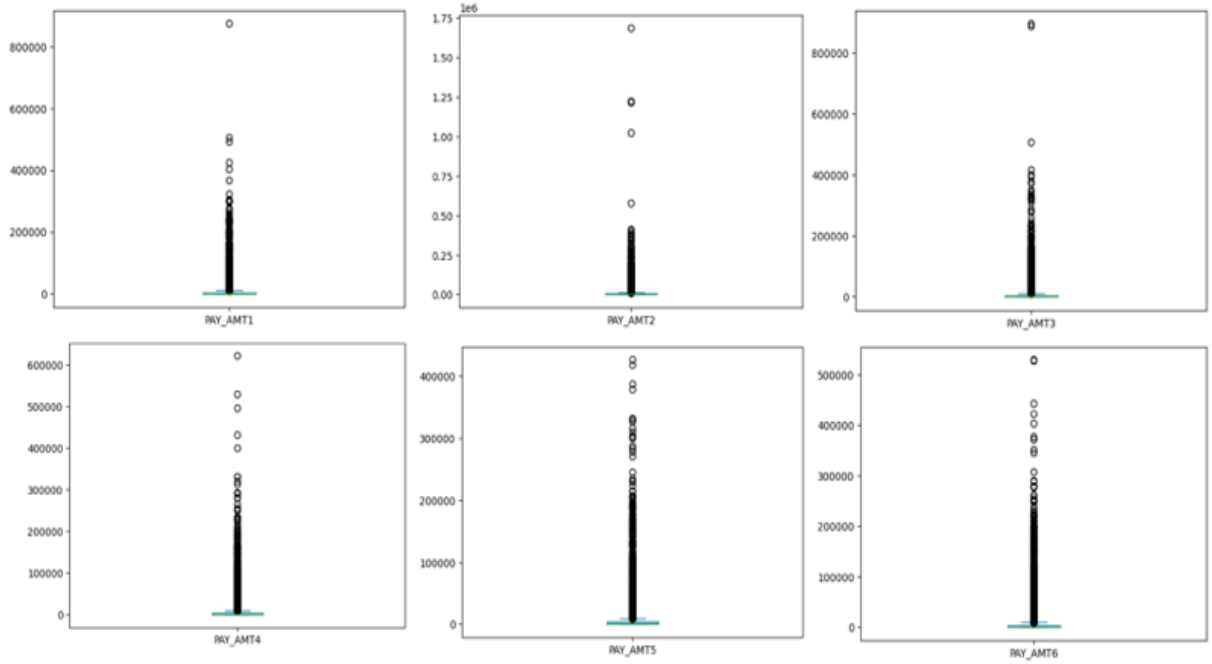


Figure 12: Boxplots of PAY_AMT

6.0 Correlation Analysis:

The variables in the dataset can be categorized into two types: demographic and payment behavior. 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	0.144802485579276 1	-0.231739764034 7224	-0.110683247373 8504	0.024952818456164105

The Correlation scatterplots are provided below:

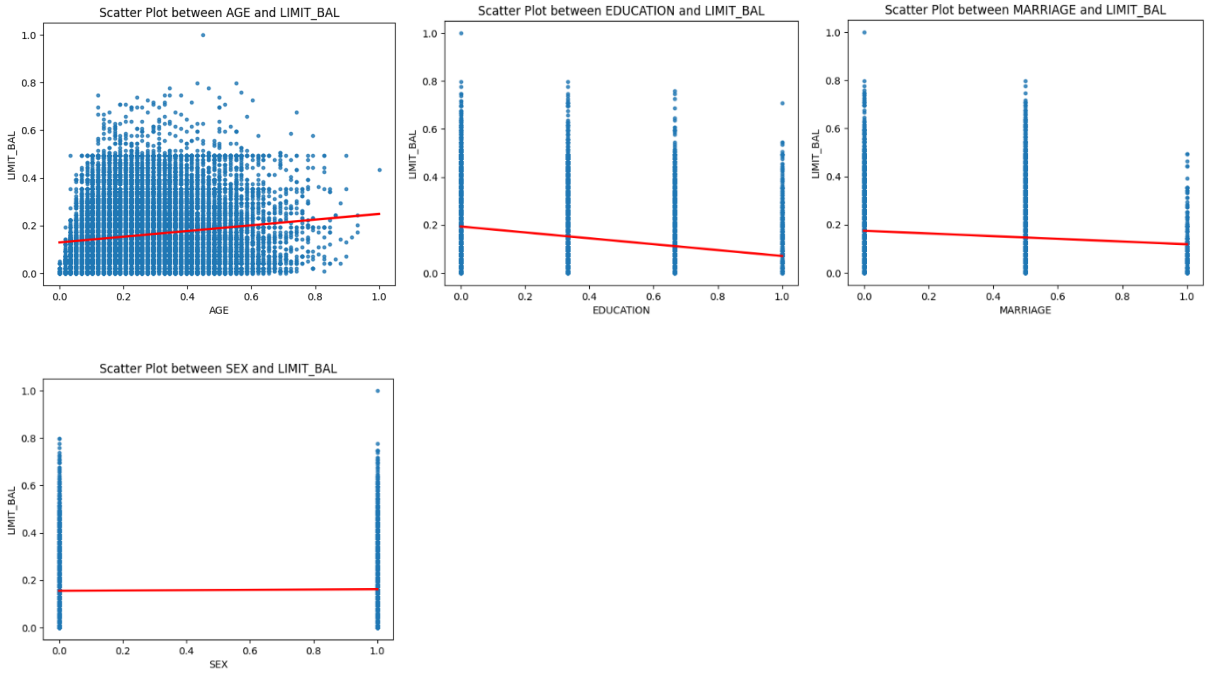


Figure 13: Pearson correlation between demographic variables and LIMIT_BAL

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.

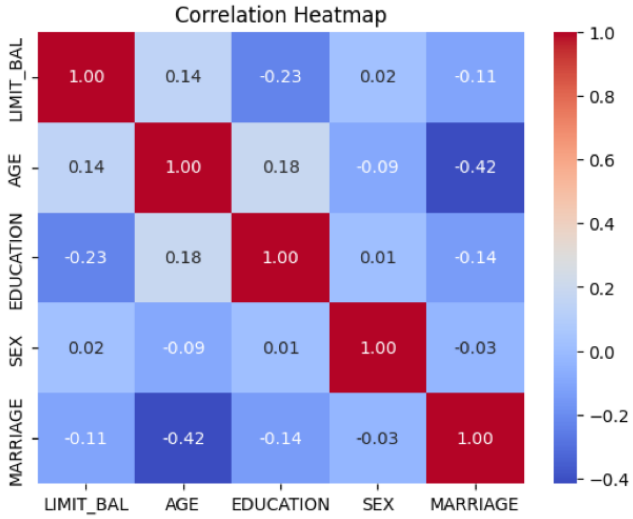


Figure 14: Heat map for Demographic variables and LIMIT_BAL

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.

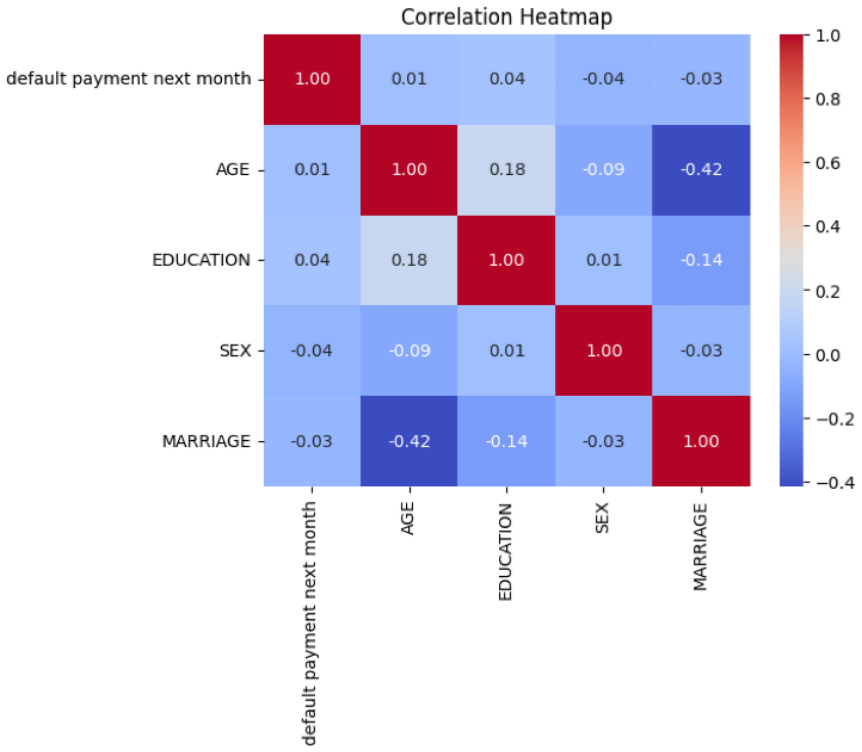


Figure 15: Heat map for Demographic variables and default payment next month

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.

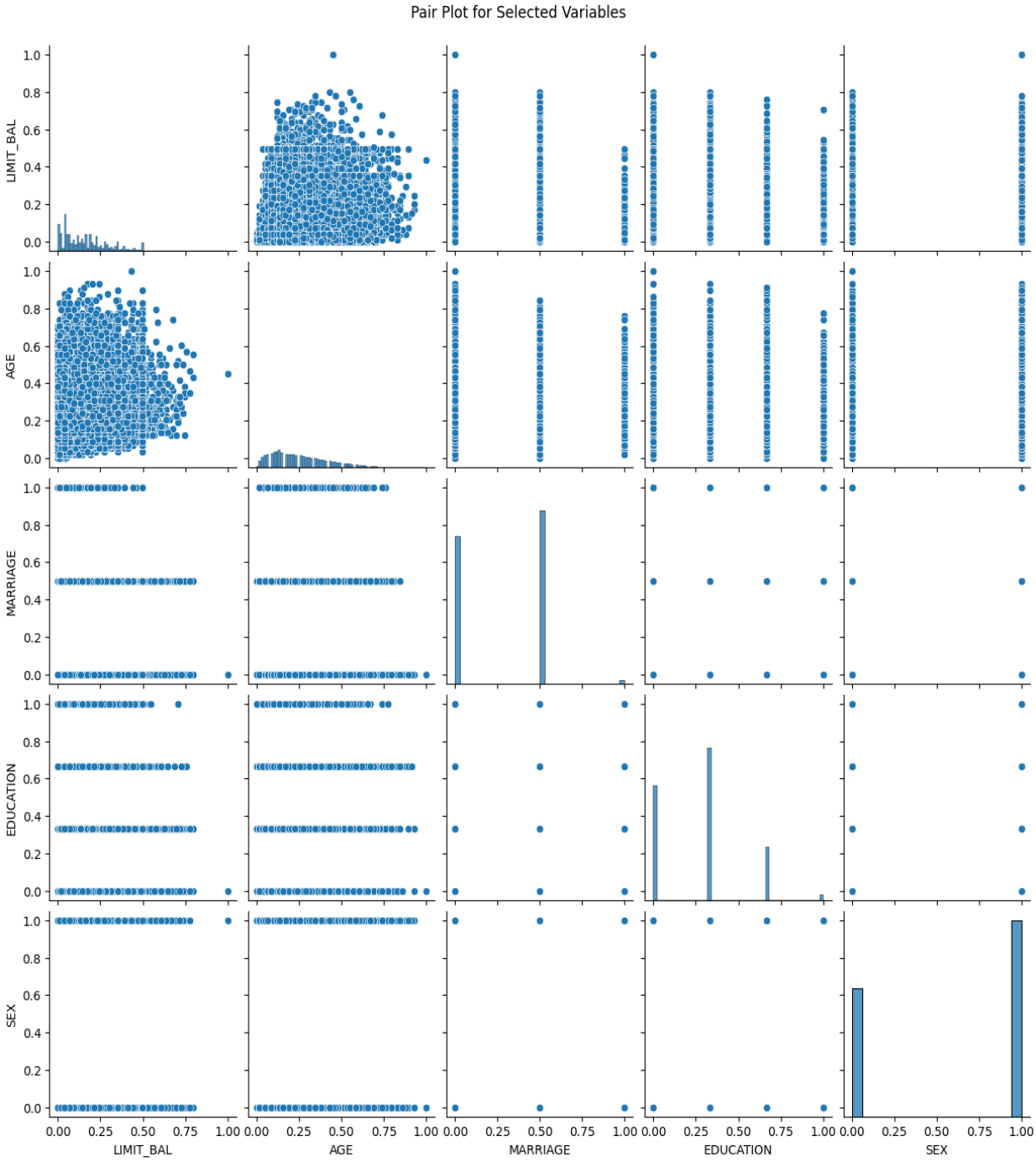


Figure 16: Pair plot for Demographic variables and LIMIT_BAL

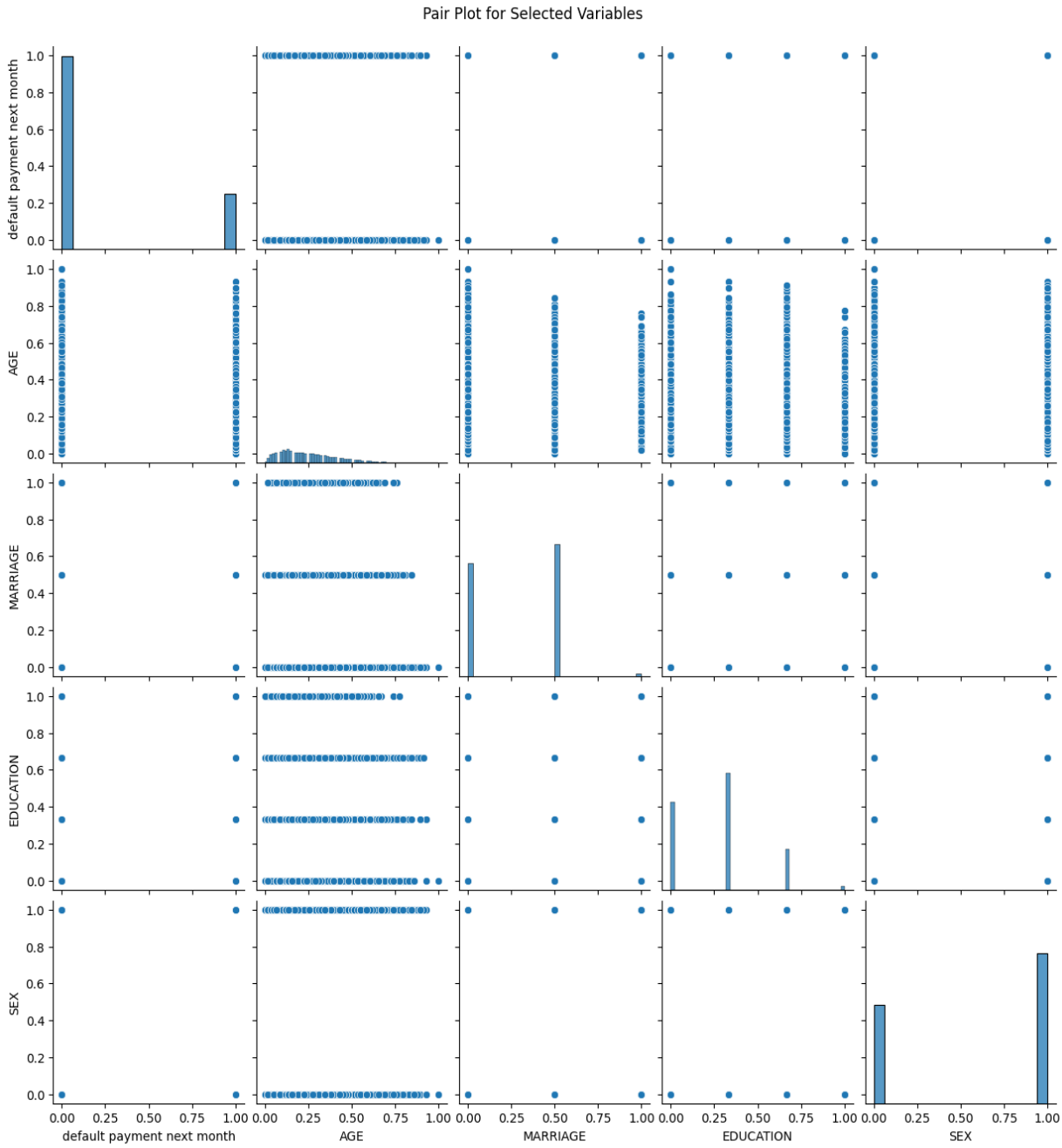


Figure 17: Pair plot for Demographic variables and default payment next month

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.

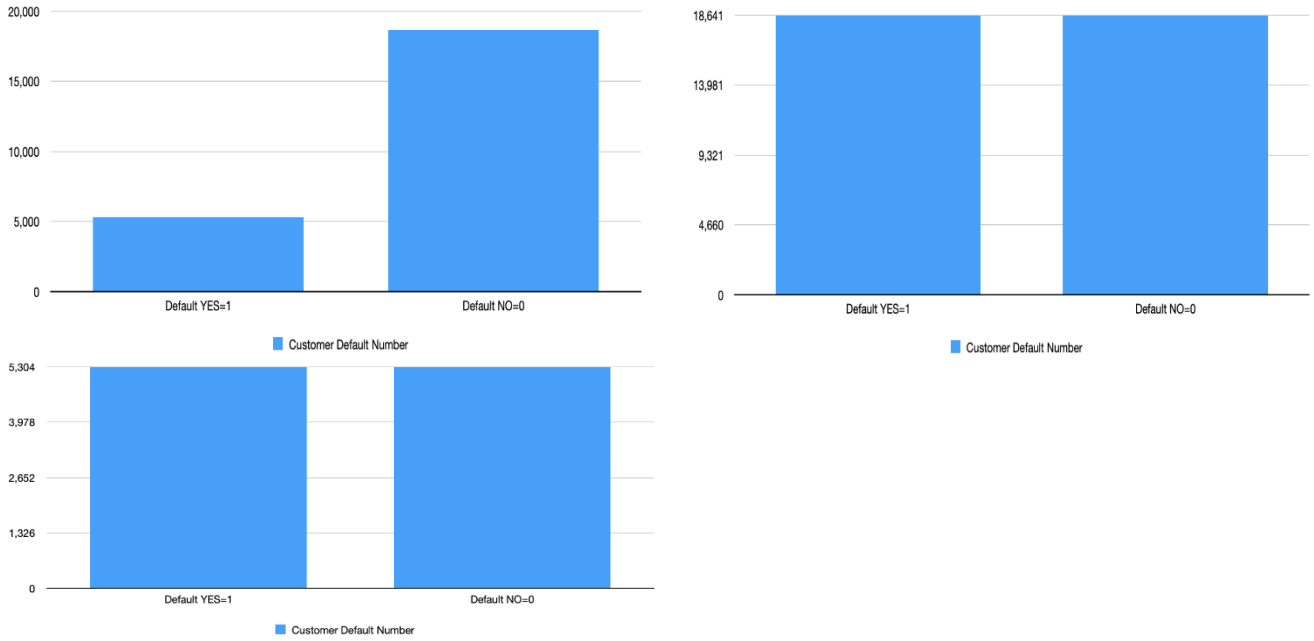


Figure 18: Data balancing of unbalanced data, data up-scaling and data down-scaling

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.

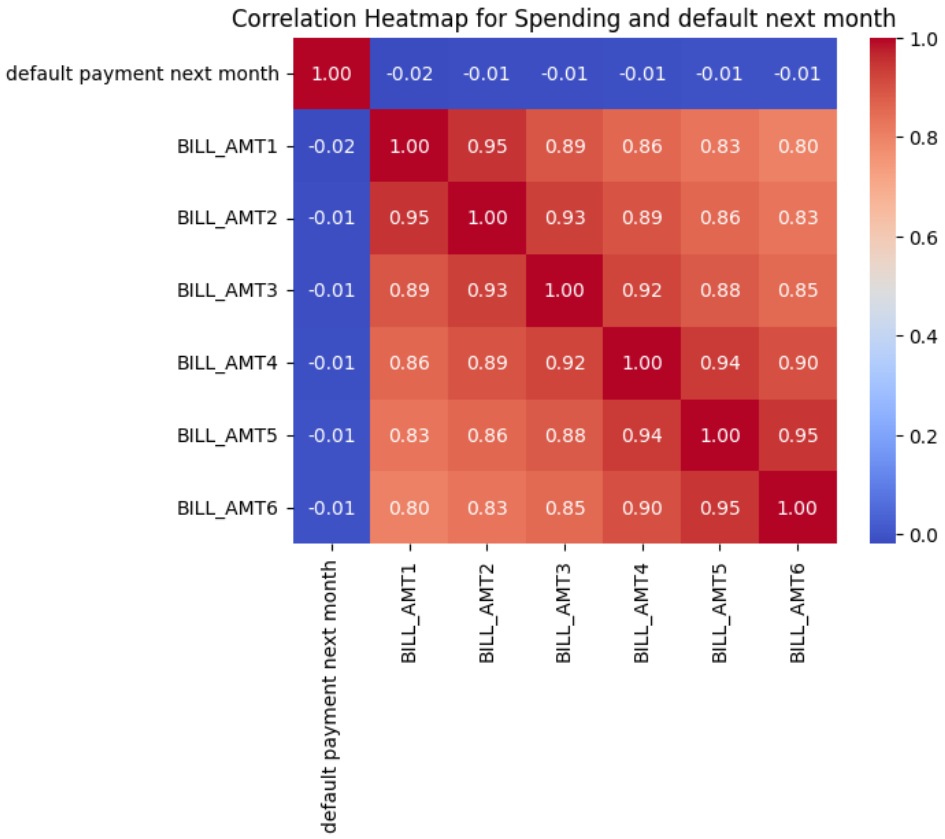


Figure 19: Heat map of BILL_AMT and default payment next month

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

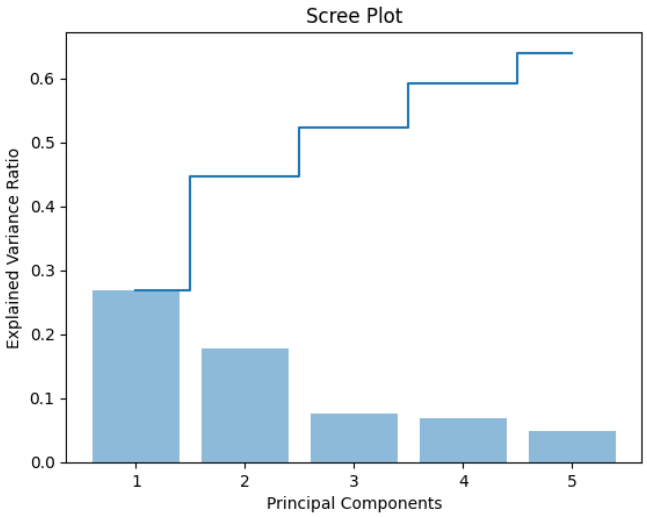


Figure 20: PC Explained variance ratio plot

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 Analysis process:

The Analysis process consists of 3 stages:

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

Data Preprocessing:

In the data preprocessing stage 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.

Model Training and Evaluation:

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 on the up-sampled dataset as the up-sampled dataset will capture majority of the characteristics of the original dataset. 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. After evaluation, to avoid the overfitting, all the datasets will be run based on cross validation. As the dataset is unbalanced, stratified k fold cross validation will be conducted so that the proportionate nature of

the train and test set is prevailed. For cross validation the K =5. After the cross validation the model result will be compared and based on the best model output, the particular model result will be taken.

Comparison and conclusion:

After model result evaluation, the up-sampled data model analysis and the cross-validation model analysis will be compared and based on that the best model outcome will be taken based on explanation and based on the overall model output and exploratory data analysis the conclusion will be sorted out.

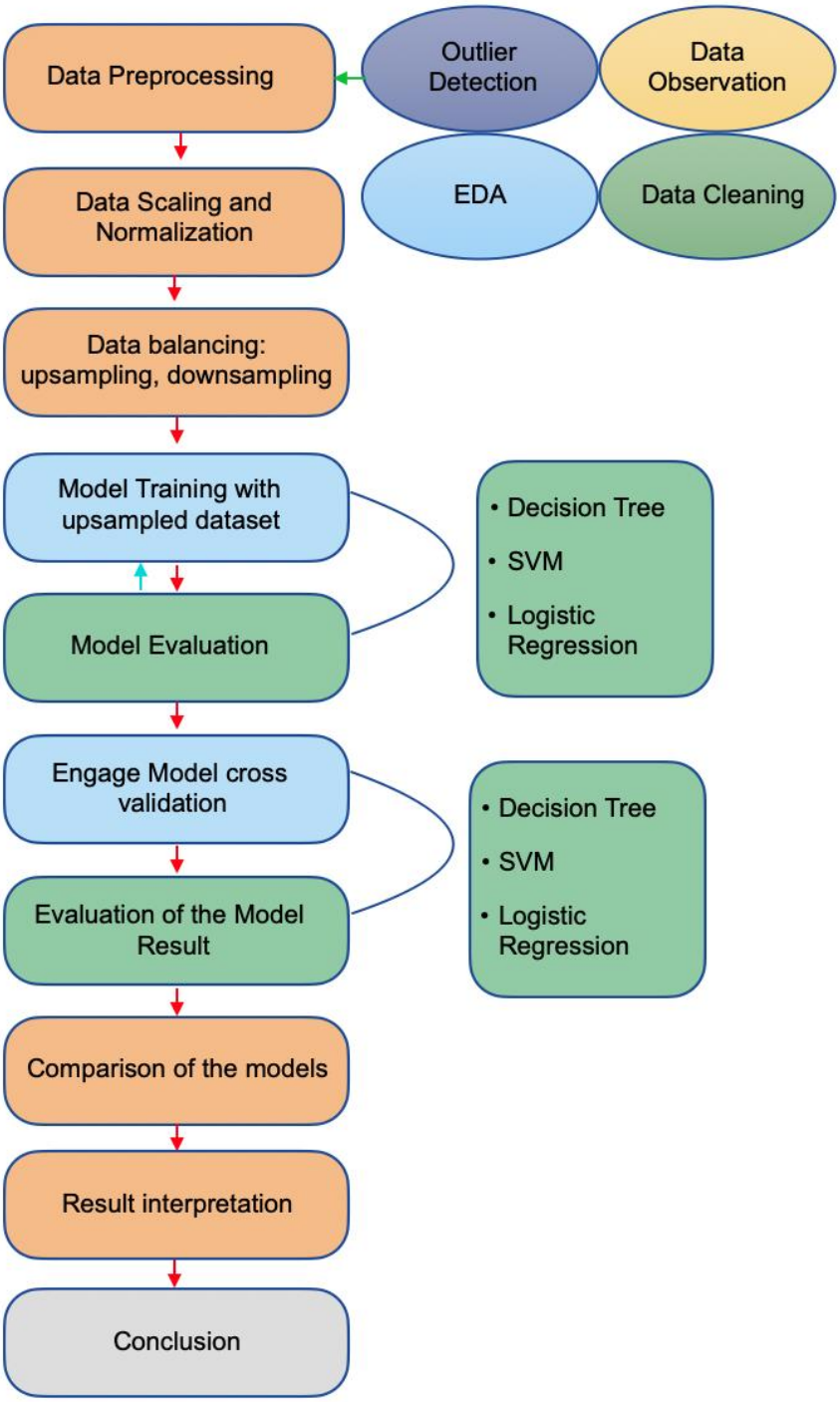


Figure 21: The Data Analysis approach and process

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 a dependent variable which is binary, and the target will be to reach the final leaf node.

9.1.2 Logistic regression:

Logistic regression is the algorithm that estimates the probability of an event occurring 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 use this algorithm to better explain credit defaults in future.

10.0 Model Output:

10.1 Model output with up-sampled data

10.1.1Decision Tree:

Classification Report :

Precision: Indicates the proportion of true positive cases out of all positive predictions. For class 0, it's 0.95, and for class 1, it's 0.83.

Recall: Also known as sensitivity, it indicates the proportion of true positive cases that were correctly identified. For class 0, it's 0.80, and for class 1, it's 0.96.

F1-score: The harmonic mean of precision and recall. It provides a balance between precision and recall. For class 0, it's 0.87, and for class 1, it's 0.89.

Accuracy: The proportion of correctly classified instances out of total instances. It's 0.88.

Support: The number of actual occurrences of each class in the test set.

Feature Importance:

Indicates the relative importance of each feature in the decision tree model. Features with higher importance values contribute more to the model's decision-making process. In this case, the most important feature is PAY_1, followed by PAY_AMT1, PAY_AMT2, AGE, PAY_AMT6, and so on.

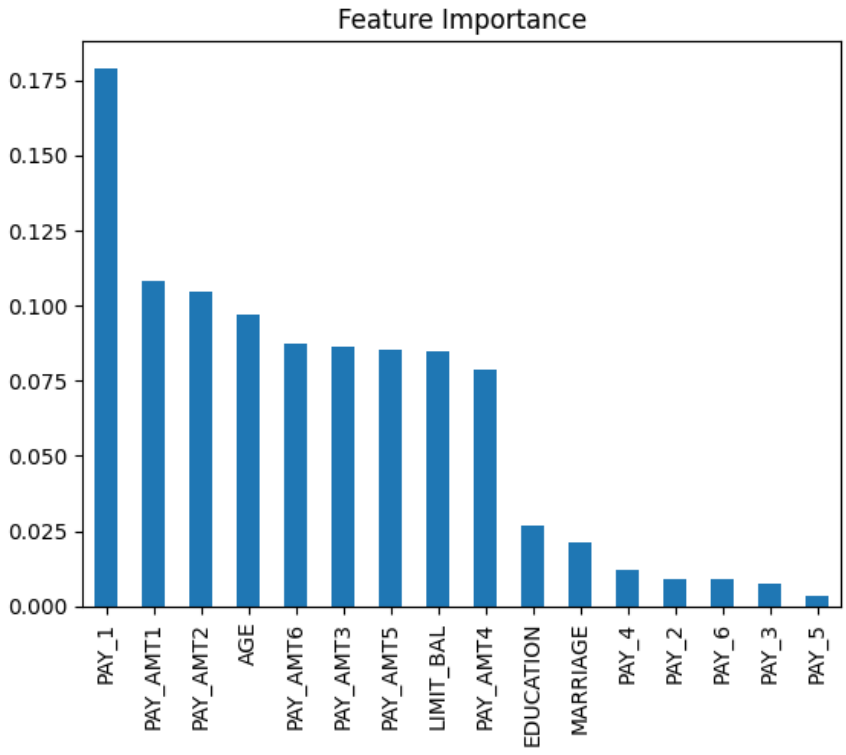


Figure 22: Feature importance

Table 7: Feature importance

Feature	Importance
PAY_1	0.178944
PAY_AMT1	0.108281
PAY_AMT2	0.104394
AGE	0.096777
PAY_AMT6	0.087340
PAY_AMT3	0.086293
PAY_AMT5	0.085451
LIMIT_BAL	0.084660
PAY_AMT4	0.078665
EDUCATION	0.026706
MARRIAGE	0.021402
PAY_4	0.012312
PAY_2	0.008881
PAY_6	0.008838
PAY_3	0.007529
PAY_5	0.003526

Accuracy:

Accuracy represents the proportion of correctly classified instances out of total instances. In this case, the accuracy is 0.8797 or 87.97%, indicating that the model correctly predicts the target variable for approximately 87.97% of the cases.

Precision:

Precision measures the proportion of correctly predicted positive cases out of all predicted positive cases. Here, the precision is 0.8270 or 82.70%. It indicates that out of all the instances predicted as positive, approximately 82.70% are actually positive.

Recall:

Recall (also known as sensitivity) measures the proportion of correctly predicted positive cases out of all actual positive cases. The recall value is 0.9628 or 96.28%. It suggests that the model correctly identifies approximately 96.28% of all positive cases.

Confusion Matrix:

The confusion matrix provides a detailed breakdown of the model's performance by showing the counts of true positives, true negatives, false positives, and false negatives.

From the confusion matrix:

True Negative (TN): 2941

False Positive (FP): 757

False Negative (FN): 140

True Positive (TP): 3619

This matrix helps to understand how well the model is performing in terms of correctly identifying each class.

F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. In this case, the F1 score is 0.8897 or 88.97%. It indicates the overall performance of the model, considering both precision and recall.

Overall, the provided metrics indicate that the decision tree model performs well in terms of accuracy, precision, recall, and F1 score. It achieves high accuracy and effectively balances the trade-off between precision and recall, as demonstrated by the F1 score. However, it's always good to further evaluate the model's performance and consider the specific context of the problem domain.

10.1.2 Logistic Regression:

Model Summary:

The model appears to have converged successfully after 6 iterations. The current function value (log-likelihood) is 0.580456.

Model Coefficients:

Each coefficient represents the change in the log odds of the dependent variable (default payment next month) for a one-unit change in the corresponding independent variable, holding all other variables constant.

const: The intercept of the model. It indicates the log odds of the dependent variable when all independent variables are zero.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:     default payment next month   No. Observations:                29825
Model:                                 Logit   Df Residuals:                    29808
Method:                                  MLE   Df Model:                           16
Date:                       Sat, 09 Mar 2024   Pseudo R-squ.:                 0.1626
Time:                               01:05:13   Log-Likelihood:                -17312.
converged:                             True   LL-Null:                       -20673.
Covariance Type:                 nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2946      0.043     -6.897      0.000      -0.378      -0.211
LIMIT_BAL     -1.4783      0.120    -12.348      0.000      -1.713      -1.244
AGE            0.4334      0.090      4.809      0.000       0.257       0.610
EDUCATION     -0.1717      0.056     -3.046      0.002      -0.282      -0.061
MARRIAGE      -0.3264      0.055     -5.891      0.000      -0.435      -0.218
PAY_1          7.3939      0.190     38.963      0.000       7.022       7.766
PAY_2          0.5175      0.188      2.757      0.006       0.150       0.885
PAY_3          1.3676      0.193      7.098      0.000       0.990       1.745
PAY_4          0.7414      0.221      3.359      0.001       0.309       1.174
PAY_5          0.7448      0.244      3.055      0.002       0.267       1.223
PAY_6          1.3558      0.208      6.527      0.000       0.949       1.763
PAY_AMT1      -7.0548      1.208     -5.839      0.000      -9.423      -4.687
PAY_AMT2     -12.8134      2.165     -5.918      0.000     -17.057      -8.570
PAY_AMT3      -0.7247      0.810     -0.895      0.371      -2.312       0.862
PAY_AMT4      -0.7474      0.616     -1.214      0.225      -1.954       0.459
PAY_AMT5      -0.4283      0.427     -1.003      0.316      -1.265       0.409
PAY_AMT6      -0.2265      0.462     -0.490      0.624      -1.133       0.680
------------------------------------------------------------------------------
```

LIMIT_BAL, AGE, EDUCATION, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6: These are the coefficients for the respective independent variables. They indicate the change in log odds for a one-unit increase in each variable. The std err column represents the standard error of the coefficient estimate. The z column provides the z-score, which is the coefficient divided by its standard error. It indicates the significance of each coefficient. The P>|z| column provides the p value associated with the z-score. Lower p-values indicate higher significance. The [0.025 0.975] columns give the 95% confidence interval for each coefficient.

Confusion Matrix :

The confusion matrix shows the performance of the model on the test set.

It is a 2x2 matrix where rows represent the actual classes and columns represent the predicted classes. From the confusion matrix:

True Negative (TN): 3107

False Positive (FP): 591

False Negative (FN): 1602

True Positive (TP): 2157

Accuracy, Precision, Recall, and F1 Score:

These metrics provide a more comprehensive evaluation of the model's performance.

Accuracy: The proportion of correctly classified instances out of total instances.

Precision: The proportion of correctly predicted positive cases out of total predicted positive cases.

Recall (Sensitivity): The proportion of correctly predicted positive cases out of actual positive cases.

F1 Score: The harmonic mean of precision and recall, providing a balance between the two.

Other Model Statistics:

Log-Likelihood, LL-Null, LLR p-value: These statistics are related to model fit and goodness of fit tests.

Pseudo R-squared: This indicates the proportion of variance explained by the model. However, caution is advised when interpreting pseudo-R-squared in logistic regression.

Overall, this summary provides insights into the coefficients' significance, model fit, and performance metrics of the logistic regression model.

10.1.3 Support vector machine:

Number of Support Vectors:

Indicates the number of support vectors used in the SVM model. In this case, there are 19427 support vectors.

Classification Report :

Precision: Indicates the proportion of true positive cases out of all positive predictions and is a measure of the model's exactness. For class 0, it's 0.64, and for class 1, it's 0.82.

Recall: Also known as sensitivity, it indicates the proportion of true positive cases that were correctly identified and is a measure of the model's completeness. For class 0, it's 0.89, and for class 1, it's 0.50.

F1-score: The harmonic mean of precision and recall. It provides a balance between precision and recall. For class 0, it's 0.74, and for class 1, it's 0.62.

Accuracy: The proportion of correctly classified instances out of total instances. It's 0.69.

Support: The number of actual occurrences of each class in the test set.

Macro and Weighted Average:

Macro avg: The average of precision, recall, and F1-score for each class, without considering class imbalance. It's 0.73 for precision, recall, and F1-score.

Weighted avg: The average of precision, recall, and F1-score for each class, considering class imbalance by weighing each class's contribution by its support. It's 0.73 for precision, recall, and F1-score.

Confusion Matrix :

The confusion matrix provides a tabular representation of actual versus predicted classes. It consists of four terms: True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP). In this case:

True Negatives (TN): 3283

False Positives (FP): 415

False Negatives (FN): 1874

True Positives (TP): 1885

Accuracy:

Accuracy measures the proportion of correctly classified instances out of total instances. It's 0.6930 or 69.30%.

Precision:

Precision indicates the proportion of true positive cases out of all positive predictions. It's 0.8196 or 81.96%.

Recall:

Recall, also known as sensitivity, indicates the proportion of true positive cases that were correctly identified. It's 0.5015 or 50.15%.

F1 Score:

F1 Score is the harmonic mean of precision and recall. It provides a balance between precision and recall. It's 0.6222 or 62.22%.

Overall, the SVM model shows decent accuracy and precision. However, the recall for class 1 (defaulters) is relatively low, indicating that the model may have difficulty correctly identifying default cases. This could be an area for improvement, depending on the specific requirements and objectives of the classification task.

Result Comparison:

In comparing the three models - Decision Tree, SVM (Support Vector Machine), and Logistic Regression - across the provided metrics of accuracy, precision, recall, and F1 score, several observations can be made.

Firstly, in terms of accuracy, the Decision Tree model performs the best with an accuracy of 0.8797, indicating that it correctly classifies instances most often compared to SVM (0.6930) and Logistic Regression (0.7059). However, it's essential to consider the possibility of overfitting, especially if the Decision Tree model is highly complex.

Regarding precision, all models show relatively high values, with the Decision Tree and SVM having similar precision scores (0.8270 and 0.8196, respectively), slightly outperforming Logistic Regression (0.7849). Precision reflects the model's ability to correctly identify positive cases, minimizing false positives.

On the other hand, the Decision Tree model demonstrates exceptionally high recall (0.9628), meaning it effectively captures a vast majority of positive instances. In contrast, both SVM and Logistic Regression models exhibit significantly lower recall scores (0.5015 and 0.5738,

respectively), suggesting that they may struggle to identify all positive cases, particularly defaulters in this context.

When considering the F1 score, which balances precision and recall, the Decision Tree model again leads (0.8897), followed by Logistic Regression (0.6630) and SVM (0.6222). This suggests that while the Decision Tree model achieves a good balance between precision and recall, SVM and Logistic Regression models may have some trade-offs between precision and recall.

Overall, the Decision Tree model stands out for its high accuracy, precision, recall, and F1 score. However, its performance should be carefully evaluated to ensure it doesn't over fit the training data. SVM and Logistic Regression models offer decent performance, but they may require further optimization to improve recall, especially if correctly identifying positive cases (e.g., defaulters) is crucial. The choice among these models ultimately depends on the specific requirements and trade-offs of the classification task at hand.

10.2 Optimized screening with cross validation:

Cross-validation serves as a crucial technique in machine learning to assess how well a model can generalize to unseen data. It functions by splitting the available dataset into multiple subsets or folds. Each iteration involves using one of these folds as a validation set while training the model on the remaining folds. This process repeats iteratively, ensuring that each fold serves as a validation set at least once. Finally, the results from each iteration are averaged to obtain a comprehensive evaluation of the model's performance.

The primary objective of cross-validation is to guard against overfitting, a common pitfall where a model performs exceedingly well on the training data but poorly on new, unseen data. By subjecting the model to evaluation on various validation sets, cross-validation offers a more

realistic assessment of its ability to generalize to unseen data. This ensures that the model chosen for deployment is robust and capable of performing well in real-world scenarios.

In essence, cross-validation acts as a safeguard against the model learning noise from the training data and instead focuses on learning meaningful patterns that can be applied to new data. It provides a more accurate estimate of the model's true performance, thereby aiding in the selection of models that are reliable and generally effective.

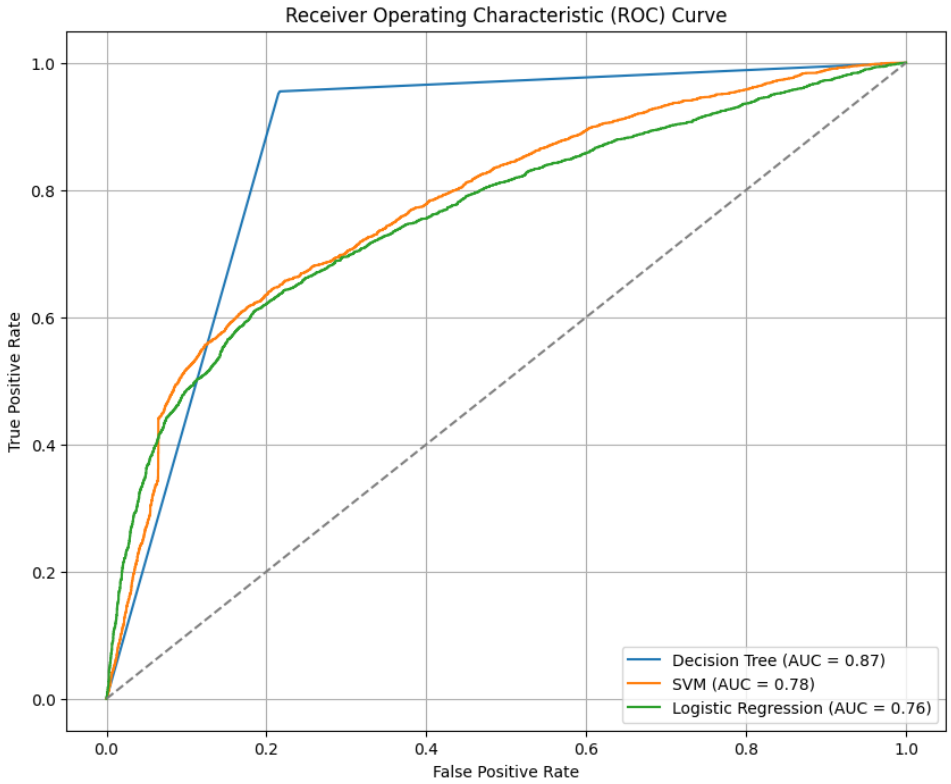


Figure 23: ROC Curve

10.3 Stratified K Fold Cross validation:

Stratified K-Fold Cross-Validation stands out as a modified version of the traditional K-Fold Cross-Validation technique. It ensures that each fold maintains a proportional representation of

observations for every target class present in the entire dataset. This approach becomes particularly significant for datasets where certain classes are significantly underrepresented compared to others. By maintaining consistent class proportions across folds, Stratified K-Fold Cross-Validation helps in ensuring that the model is trained and evaluated on a representative sample of the data, thus enhancing its generalization capability.

In our credit card default dataset, the data is unbalanced. We have the default next month data as 77.8% and 22.2%. One approach is balancing the data which is done in the previous section and the models are constructed. But to avoid overfitting we will run the models and try to interpret which model fits well with the model. To work with this unbalanced dataset, we will consider the stratified K fold cross validation.

10.4 Model output with stratified k fold cross validation

10.4.1 Decision tree:

The cross-validation decision tree model exhibits moderate performance based on the provided metrics.

In terms of accuracy, the model achieves an accuracy of 0.7226 during cross-validation, indicating that approximately 72.26% of the instances are correctly classified. The precision of 0.3833 suggests that out of all instances predicted as positive by the model, only around 38.33% are truly positive. The recall of 0.4138 implies that the model successfully captures about 41.38% of all actual positive instances. The F1 score of 0.3979, which is the harmonic mean of precision and recall, indicates a balance between precision and recall. When evaluated on the test set, the model maintains a similar performance with an accuracy of 0.7244. The precision, recall, and F1 score also remain consistent with values of approximately 0.3826, 0.3982, and 0.3902, respectively.

Overall, while the decision tree model performs consistently between cross-validation and the test set, the moderate values of precision, recall, and F1 score suggest room for improvement. Further optimization or consideration of alternative models may be beneficial to enhance the model's performance.

10.4.2 Support Vector Machine:

The cross-validation SVM (Support Vector Machine) model's performance, as indicated by the provided metrics, appears concerning.

The cross-validation accuracy of 77.85% suggests that a large portion of instances are accurately classified by the model. However, when assessing precision, recall, and F1 score, all metrics return values of 0.0000. This signifies that the model didn't correctly identify any instances belonging to the positive class. In essence, the model didn't classify any instances as positive, leading to zero true positives and false positives. As a result, precision, recall, and F1 score, which depend on identifying positive instances, all yield null values.

After assessing the model's performance on the test set, the accuracy remains consistent at 77.85%. However, akin to the cross-validation outcomes, precision, recall, and F1 score all register values of 0.0000, indicating an inability to accurately classify positive instances. This consistent failure to identify positive cases underscores a notable deficiency in the SVM model's performance. It prompts a deeper investigation into potential issues such as class imbalance, parameterization of the model, or data preprocessing techniques that might have adversely affected its efficacy. Exploring alternative modeling strategies or conducting further optimization may be necessary to address these observed limitations and enhance the model's predictive capability.

10.4.3 Logistic Regression:

The cross-validation logistic regression model's performance, as indicated by the provided metrics, raises concerns.

The accuracy of 0.7785 during cross-validation suggests that approximately 77.85% of instances are correctly classified. However, when considering precision, recall, and F1 score, all metrics report values of 0.0000. This indicates that the model failed to correctly identify any instances belonging to the positive class. In other words, the model did not classify any instances as positive, resulting in zero true positives and zero false positives. Consequently, precision, recall, and F1 score, which rely on the identification of positive instances, all report null values.

After evaluating the model's performance on the test set, the accuracy remains consistent at 0.7785. However, similar to the cross-validation results of SVM, precision, recall, and F1 score all report values of 0.0000, suggesting an inability to correctly classify positive instances.

Overall, the logistic regression model's performance, particularly its failure to correctly identify positive instances, warrants further investigation. It is essential to examine potential issues such as class imbalance, model parameterization, or data preprocessing techniques that may have impacted the model's performance adversely. Additionally, considering alternative modelling approaches or further optimization may be necessary to address the observed limitations and improve the model's effectiveness.

10.4.4 Result comparison:

The provided results depict the performance metrics of three different models - Decision Tree, SVM (Support Vector Machine), and Logistic Regression - evaluated on both the training and test sets.

Training Set:

Accuracy: The Decision Tree model achieves an accuracy of approximately 72.26%, while both SVM and Logistic Regression models perform similarly, with accuracies around 77.85%.

Precision: For precision, the Decision Tree model outperforms SVM and Logistic Regression, achieving a score of approximately 38.33%. However, both SVM and Logistic Regression models have a precision score of 0.00%, indicating they fail to correctly identify any positive instances in the training set.

Recall: The Decision Tree model exhibits a recall of approximately 41.38%, indicating it correctly identifies about 41.38% of all actual positive instances. In contrast, both SVM and Logistic Regression models have a recall score of 0.00%, indicating an inability to correctly classify positive instances.

F1 Score: The F1 score, which is the harmonic mean of precision and recall, reflects a similar trend. The Decision Tree model achieves the highest F1 score of approximately 39.79%, while SVM and Logistic Regression models report null values due to the absence of positive classifications.

Test Set:

The performance metrics on the test set closely mirror those of the training set for all three models, with similar accuracy, precision, recall, and F1 score values observed.

Interpretation:

The Decision Tree model shows moderate performance on both the training and test sets, with relatively balanced precision and recall.

However, SVM and Logistic Regression models exhibit poor performance, particularly in their inability to correctly identify positive instances (as indicated by precision, recall, and F1 score values of 0.00%).

This suggests that there may be significant issues with model training or class imbalance, particularly for SVM and Logistic Regression models, which need to be addressed to improve their performance on positive instance classification.

The performance metrics on the test set closely mirrored those of the training set for all three models, with similar accuracy, precision, recall, and F1 score values observed.

Interpretation:

The Decision Tree model exhibited relatively better performance compared to SVM and Logistic Regression in terms of precision, recall, and F1 score.

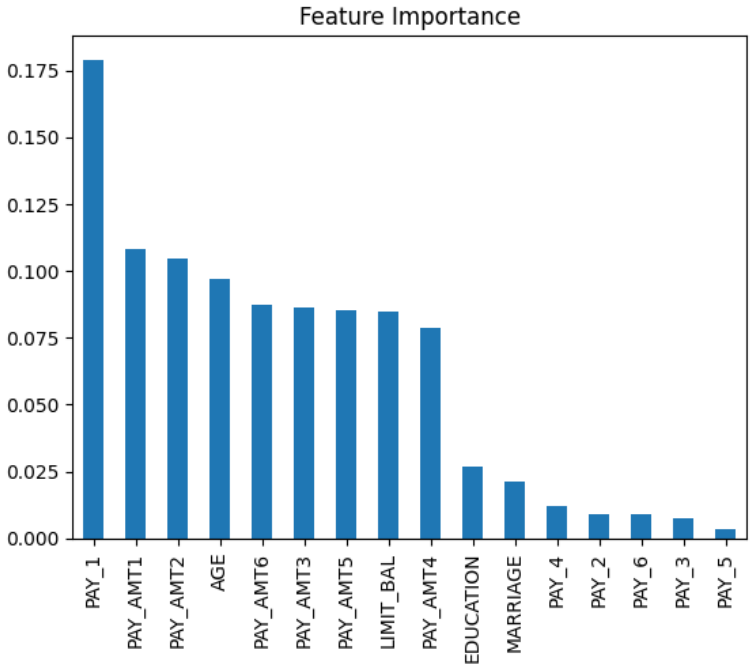
However, all three models failed to correctly identify any positive instances in both the training and test sets, as reflected by the precision, recall, and F1 score values of 0.0000 for SVM and Logistic Regression.

This suggests that there may be significant issues with model training or class imbalance, particularly for SVM and Logistic Regression models, which need to be addressed to improve their performance on positive instance classification.

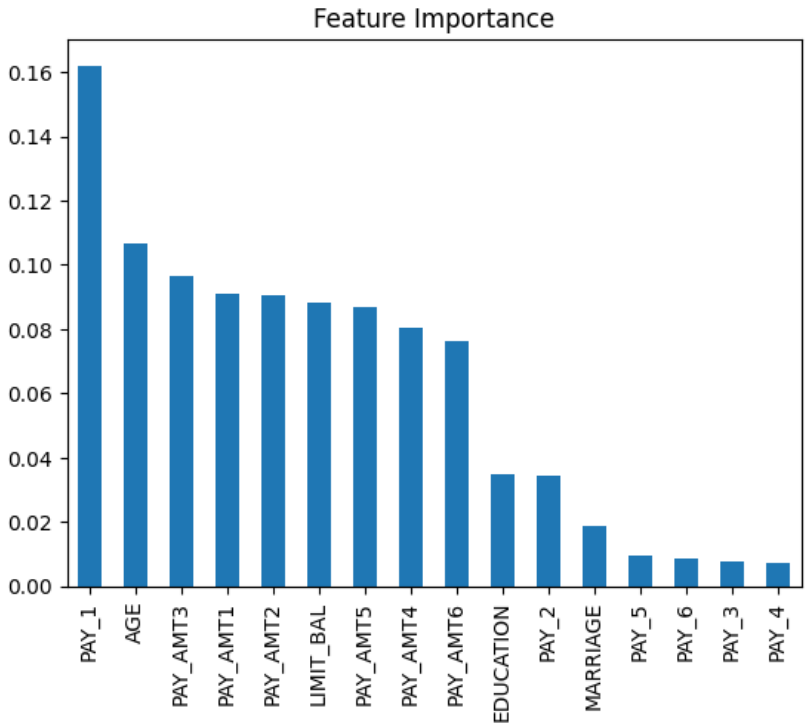
10.5 Decision on model selection:

Considering all the comparisons and model output, decision tree provides the optimum result for the current dataset. According to the Decision tree model of Up-sampling, the featured importance is provided below:

As per feature importance it is seen that PAY_1, PAY_AMT1, PAY_AMT2, AGE, PAY_AMT6, PAY_AMT3, PAY_AMT5, LIMIT_BAL, PAY_AMT4 has high importance to determine default next month. EDUCATION and MARRIAGE have moderate importance to determine the credit default next month. The PAY factors have minor importance on default of credit card. So, from up-sampling we are seeing that AGE has the highest importance from demographic factors; Education and Marriage has moderate importance. As per our predetermined question, we can see that LIMIT_BALANCE has high contribution to default next month.



For Stratified K fold cross validation decision tree, the feature importance graph is provided below:



As per the stratified K fold decision tree we can see that for demographic factors, AGE has the highest importance considering default of credit card next month EDUCATION has the second highest importance from demographic factors. Considering the Up sampled dataset, the importance of marriage has reduced but it will not be very minimal importance. LIMIT_BAL importance has been increased. Other than that, the rest of the factors occupied same importance on default next month.

10.6 Stratified K fold CV Output with data normalization:

Based on the performance metrics provided for both the training and test sets, we can assess the effectiveness of the Decision Tree, SVM (Support Vector Machine), and Logistic Regression models.

Training Set:

Accuracy: The SVM model achieved the highest accuracy on the training set (approximately 82.05%), followed closely by the Logistic Regression model (approximately 81.74%), while the Decision Tree model had an accuracy of around 72.24%.

Precision: The SVM model also exhibited the highest precision (approximately 68.00%), followed by the Logistic Regression model (approximately 68.65%), whereas the Decision Tree model had a precision of around 38.31%.

Recall: The Decision Tree model had the highest recall (approximately 41.41%), followed by the SVM model (approximately 35.85%) and the Logistic Regression model (approximately 32.33%).

F1 Score: The SVM model demonstrated the highest F1 score (approximately 46.95%), followed by the Logistic Regression model (approximately 43.96%), while the Decision Tree model had an F1 score of around 39.80%.

Test Set:

Accuracy: The SVM model maintained its lead in accuracy on the test set (approximately 81.46%), followed closely by the Logistic Regression model (approximately 81.29%), while the Decision Tree model had an accuracy of around 72.47%.

Precision: The SVM model still exhibited the highest precision (approximately 65.84%), followed by the Logistic Regression model (approximately 67.88%), whereas the Decision Tree model had a precision of around 38.33%.

Recall: Once again, the Decision Tree model had the highest recall (approximately 39.89%), followed by the SVM model (approximately 33.86%) and the Logistic Regression model (approximately 29.49%).

F1 Score: The SVM model continued to demonstrate the highest F1 score (approximately 44.72%), followed by the Logistic Regression model (approximately 41.11%), while the Decision Tree model had an F1 score of around 39.10%.

Conclusion:

Overall, the SVM model appears to be the best-performing model, as it consistently achieved the highest accuracy, precision, recall, and F1 score on both the training and test sets. The Logistic Regression model also performed relatively well, showing competitive results compared to the SVM model, especially in terms of accuracy and precision. The Decision Tree model, while showing reasonable performance, had notably lower accuracy, precision, recall, and F1 score compared to the SVM and Logistic Regression models, indicating that it might not be the most suitable model for this dataset.

11.0 Comparison with previous research:

Understanding the complex interplay between demographic factors, spending behaviors, and credit default rates is crucial for both financial institutions and policymakers. Demographic variables such as age, income, education level, and employment status significantly influence individuals' financial behaviors, ultimately impacting credit default rates. For instance, younger individuals with limited financial experience may face higher default rates, while lower income levels and educational backgrounds can contribute to financial instability.

In the research such as that by Soman and Cheema (2002) highlights the importance of credit limits in shaping spending behaviors, with consumers often using their assigned limit as an indicator of future earnings potential. However, the allocation of credit limits is not solely based on financial factors but also influenced by demographic characteristics and spending patterns. Consequently,

understanding the relationship between demographic factors, payment behaviors, and credit limits is essential for predicting credit defaults accurately.

Studies like that by Memarista, Malelak, and Anastasia (2015) have explored the relationship between demographic factors and credit card default, emphasizing the significant impact of education on financial behavior. Existing studies have either focused on predictive modeling techniques for credit default or examined the impact of demographic factors on defaults without considering the nuanced relationship between spending behaviors, credit limits, and defaults. While studies like those by Çallı and Coşkun (2021) have identified predictors of credit default, they lack granularity in analyzing how demographic factors and spending behaviors influence credit limits and, subsequently, default rates.

Similarly, studies like that by Wang et al. (2011) have explored the correlation between demographic variables, behavioral aspects, and credit card debt but haven't delved into the causal relationship between these factors and credit defaults. Moreover, research conducted by Achsan et al. (2022) in Indonesia has emphasized the role of cardholder behavior in credit defaults, yet it lacks an in-depth analysis of how demographic factors and spending behaviors interact to influence credit limits and defaults.

In comparison to the previous studies this study’s outcome shows that the credit default next month is heavily dependent on the financial capabilities which are payment amount against the bill. And the first outstanding pay is crucial for the prediction of default. Considering the demographic factors, the most impactful factor is AGE, and the second most impactful factor is EDUCATION. Most of the PAY factors are minor to determine the default next month. The study conducted by

Memarista, Malelak, and Anastasia (2015) has similarities with our study outcomes as it also identified the Education and financial factors as primary controlling agent for credit default.

The Effect of Credit on Spending Decisions: The Role of the Credit Limit and Credibility (Soman and Cheema, 2002) study worked on the limit allocation, income pattern and spending pattern. Our study does not have any factor relevant to income and spending, however, similar to this study, our study also substantiates that LIMIT_BAL is a crucial factor to identify why predictability of the Default next month.

The study Understanding the impact of borrowers' behavioral and psychological traits on credit default: review and conceptual model (Goal and Rastogi, 2021) focused on certain behavioral 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. However, this study did not consider the traits based on any financial factor. Our study has concentrated on both demographic and financial factors and identified the relationship with the default of credit cards. The study done by Yeh and Lien, 2009 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. The analysis of the study suggested that from the selected data analytics techniques artificial neural network is the only one that can accurately estimate the real probability of default. The dataset for this study was the same, however the result is different. Our study found the decision tree as the most efficient one considering the Cross-validation method which is contrary to the previous research.

The Usage Patterns of Credit/Debit Card across Various Demographics (Rauf et al., 2022) aimed to investigate usage patterns of credit/debit cards across demographics in Lahore and Kasur,

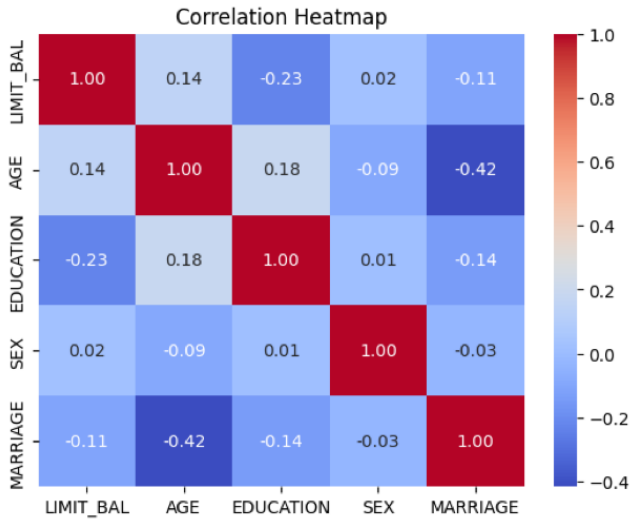
distinguishing between urban and rural areas. The study found that the usage of cards depends upon the financial conditions. The tendency to use credit card is greater in men compared to women. And the card usage is dependent on financial condition. Considering the dataset and the process this study has followed, it is not a good match with our study. According to our study the SEX is irrelevant, and we have omitted this parameter. And the results are not also match our results. The study result matches our result from the perspective that financial indicator is the most important indicator for credit card default.

12.0 Project objective and result output match:

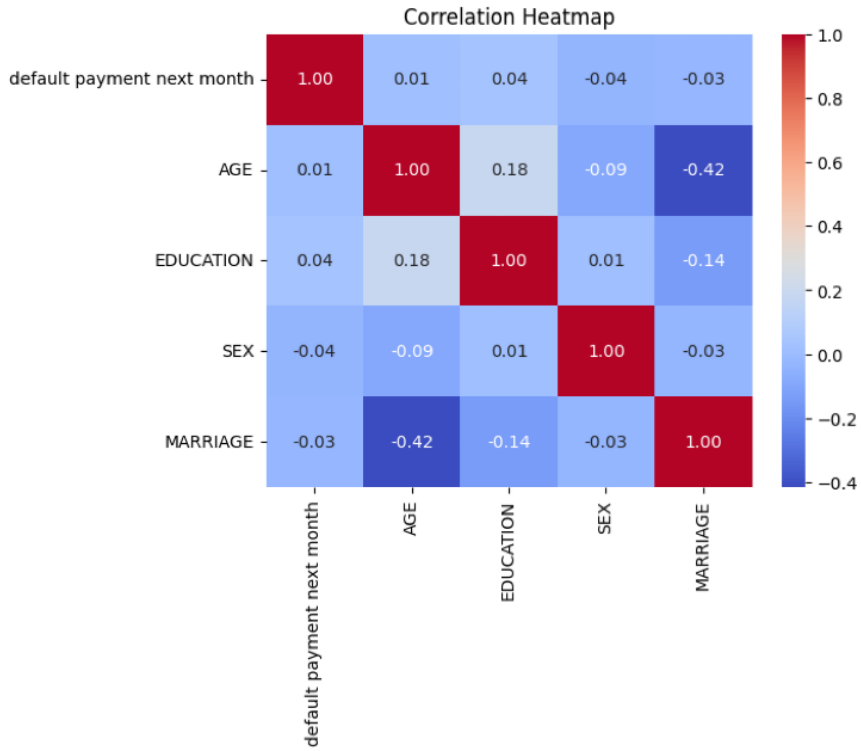
The project objective was:

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

In the correlation heat map with LIMIT_BAL and demographic variables it is shown that SEX has very low correlation with LIMIT_BAL. Other than that, Education has comparatively strong negative correlation and age has strong positive correlation.



The correlation with credit default next month and demographic variables aren't very significant. In the correlation analysis we can see that the impact on the default is minimal. However as per our project objective we have already identified the relationship with the financial indicator where LIMIT_BAL is the representative of that.



And in the predictive analysis we can see that the model's feature importance of default prediction is as follows:

Feature	Importance
PAY_1	0.178944
PAY_AMT1	0.108281
PAY_AMT2	0.104394
AGE	0.096777
PAY_AMT6	0.087340
PAY_AMT3	0.086293
PAY_AMT5	0.085451
LIMIT_BAL	0.084660
PAY_AMT4	0.078665
EDUCATION	0.026706
MARRIAGE	0.021402
PAY_4	0.012312
PAY_2	0.008881
PAY_6	0.008838
PAY_3	0.007529
PAY_5	0.003526

The accuracy rate of the decision tree is 72.26% in CV and with up-sampled data it is 87.97%. So, the quantitative analysis of prediction model is also explained in alignment with the research broad question.

13.0 Project limitations:

The project here worked on three specific models but there is room for improvement. The project limitations are provided below:

1) The dataset was geographically confined to one country. A mixed data set would have given the conclusive result.

2) The project used three classification models only. To find out the better result the project could conduct more algorithms to find the optimum result.

3) The data points are not a lot. More data points would provide more conclusive result.

4) Inclusion of more factors would have provided a more conclusive result.

14.0 Project improvement areas for future work:

The project has conducted three models for predictive analytics but in the cross-validation stage both the SVM and Logistic regression got disqualified. So, it would be a better process if the project could be conducted based on more classification models such as KNN, Naive Bayes, Linear Discriminant Analysis (LDA) etc. The data was imbalanced, and the project work was conducted on up-sampled data because this data is synthetic and captured the majority of the dataset characteristics. In the future the model can be run on the down-sampled data as well to observe the result. Due to correlation the BILL_AMT data was omitted. But this project can work with all the demographic data to see the changes in the result. The project conducted cross validation to avoid overfitting. Although cross validation is more robust, future works can consider hyper parameter tuning to find optimal features. In Addition, more financial features need to be added to justify the demographic factor impact.

15.0 Conclusion:

In conclusion, this project delved into the intricate relationship between demographic factors, payment behavior, credit limit allocation, and credit card default rates. Through comprehensive analysis and predictive modeling, the study aimed to elucidate the impact of these factors on credit default next month, offering insights valuable to financial institutions and policymakers.

The analysis revealed that while demographic variables such as age and education level play a significant role in determining credit default rates, the most substantial influence stems from financial indicators, particularly the credit limit (LIMIT_BAL). Payment behavior, as represented by variables like PAY_1 and PAY_AMT1, also emerged as crucial determinants of credit default likelihood.

Three classification models—Decision Tree, Logistic Regression, and Support Vector Machine—were employed to predict credit default next month based on the identified factors. While all models showcased varying degrees of performance, the Decision Tree model emerged as the most effective, exhibiting high accuracy, precision, recall, and F1 score. However, further investigation revealed that both the SVM and Logistic Regression models struggled to correctly classify positive instances, indicating potential issues with model training or class imbalance.

The project's findings align with existing research highlighting the importance of demographic and financial factors in predicting credit default rates. However, the study also identified several limitations, including the geographical confinement of the dataset, the use of a limited number of classification models, and the imbalance in the dataset.

Moving forward, improvements could be made by exploring additional classification algorithms, utilizing mixed datasets for broader insights, and including more datapoints and factors to enhance

model robustness. Additionally, conducting analyses on both up-sampled and down-sampled data could provide a more comprehensive understanding of model performance and generalization capabilities.

Overall, this project contributes valuable insights into the complex dynamics influencing credit default rates, emphasizing the significance of demographic factors, payment behavior, and credit limit allocation in predicting credit card defaults. These insights can inform more targeted risk assessment strategies and aid in the development of tailored financial products and interventions to mitigate credit default risks effectively.

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' behavioral 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