GROUP 5

Capstone Project – Final Report Default Analysis of Credit Card Clients

Mentor: Mr. Ankush Bansal

Submitted By,

- **Aishwarya Kumar**
- **Arjun R**
- **Hemanth Kumar**
- **Shaik Mohin**
- **Vigna Pushwini**

PROBLEM STATEMENT

To predict whether the client will pay her/his next month credit card bills on time using the dataset containing information like limit balance, sex, education, marriage, age, past payment status, bill amount, payment amount and to build a suitable model to predict credit card defaulters, so that banks can identify the right clients and improve their marketing strategy to increase profit and minimize credit risk.

PROJECT OUTCOME

The aim of this project is to construct an efficient risk prediction system to detect the possible defaults for the credit card holders. The system collects the personal and financial information about the credit card holders and then applies a suitable model to predict the default cases.

INDUSTRY REVIEW

In today's world, people are heavily depending upon financial institutions to take loans for different purposes, and applications from people are increasing day-to-day. Financial institutions cannot approve all these applications due to their reliability, because they may face huge capital loss if they sanctioned the loan without having any prior assessment of default risk.

Besides, the number of transactions in the banking sector is rapidly increasing and huge data volumes are available which represent the customer's behavior and the risks around loan are increased. Currently, most of the financial institutions use their own credit score cards and risk assessment techniques in order to analyze the loan application and to make decisions whether to approve the application or not.

Machine learning algorithms will be used to study the historical credit data to extract patterns from it, which would help in predicting the likely defaulters, thereby helping the financial institutions for making better decisions in the future.

DATA SET AND DOMAIN

- A dataset is a collection of data and it can be structured or unstructured.

- A structured data is represented in a tabular format, where every column of the table represents a particular variable, and each row corresponds to a given record of the dataset in question.

- Unsupervised data is not represented in a tabular form, data that we fetch from Facebook, Twitter, and Netflix etc. with the help of recommendation systems are all our unsupervised data.

LITERATURE SURVEY:

- During 2005 the credit card issuers started to face a cash and credit card debt crisis. In order to increase market share, card-issuing banks over-issued credit cards even to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused the credit card for consumption and accumulated heavy credit card debts. The crisis caused the blow to consumer finance confidence and it started to become a big challenge for both banks and cardholders.

- In a well-developed financial system, crisis management is on the downstream and risk prediction is on the upstream. The major purpose of risk prediction is to use financial information, such as business financial statement, customer transaction and repayment records, etc., to predict business performance or individual customers' credit risk and to reduce the damage and uncertainty.

- The datasets with same issue of credit fraud but for a different country is observed for the knowledge over the data dictionary.

PROPOSED MODEL:

As part of this study the below models are explored,

- **Logistic Regression**: Logistic regression is often used in credit risk modelling and prediction in the finance and economics literature. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables. A logistic regression model produces a probabilistic formula of classification. Logistic Regression has problems to deal with non-linear effects of explanatory variables.

- **Decision Tree Classifier**: Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

- **Random Forest Classifier**: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

- **XG Boost Classifier**: Extreme Gradient boosting is a popular machine learning algorithm that combines multiple weak learners, like trees, into a one strong ensemble model. This is done by first fitting a model into the data. However, the first model is not likely to fit the model perfectly to the data points so we are left with residuals. We can then fit another tree to the residuals to minimize a loss function that can be the second norm but gradient boosting allows the use of any loss function. This can be iterated for multiple steps which leads to a stronger model and with proper regularization overfitting.

- **Light GBM**: Light GBM is a gradient boosting framework that uses tree-based learning algorithm. Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm. Light GBM is prefixed as ‘Light’ because of its high speed. Light GBM can handle the large size of data and takes lower memory to run. Another reason of why Light GBM is popular is because it focuses on accuracy of results.

- **Ada-boost**: Ada-boost is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as base classifier if it accepts weights on the training set. Adaboost should meet two conditions: The classifier should be trained interactively on various weighed training examples. In each iteration, it tries to provide an excellent fit for these examples by minimizing training error.

- **Voting Classifier**: A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

DATA DICTIONARY

- **DEFAULT** - Default payment next month (Yes=1, No=0)
- **LIMIT_BAL** - Amount of the given credit (INR)
- **SEX** - Gender (1 = male; 2 = female)
- **EDUCATION** - Education (0=Others/Ph.D.; 1= graduate school; 2= university; 3= Diploma; 4= Highschool; 5=SSLC; 6=No Education)
- **MARRIAGE** - (1 = married; 2 = single; 3 = divorcees; 0=others)
- **AGE** - Numerical Values
- **PAY** - History of repayment status

  PAY_0 = the repayment status in September 2005
  PAY_2 = the repayment status in August 2005
  PAY_3 = the repayment status in July 2005
  PAY_4 = the repayment status in June 2005
  PAY_5 = the repayment status in May 2005
  PAY_6 = the repayment status in April 2005

  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)

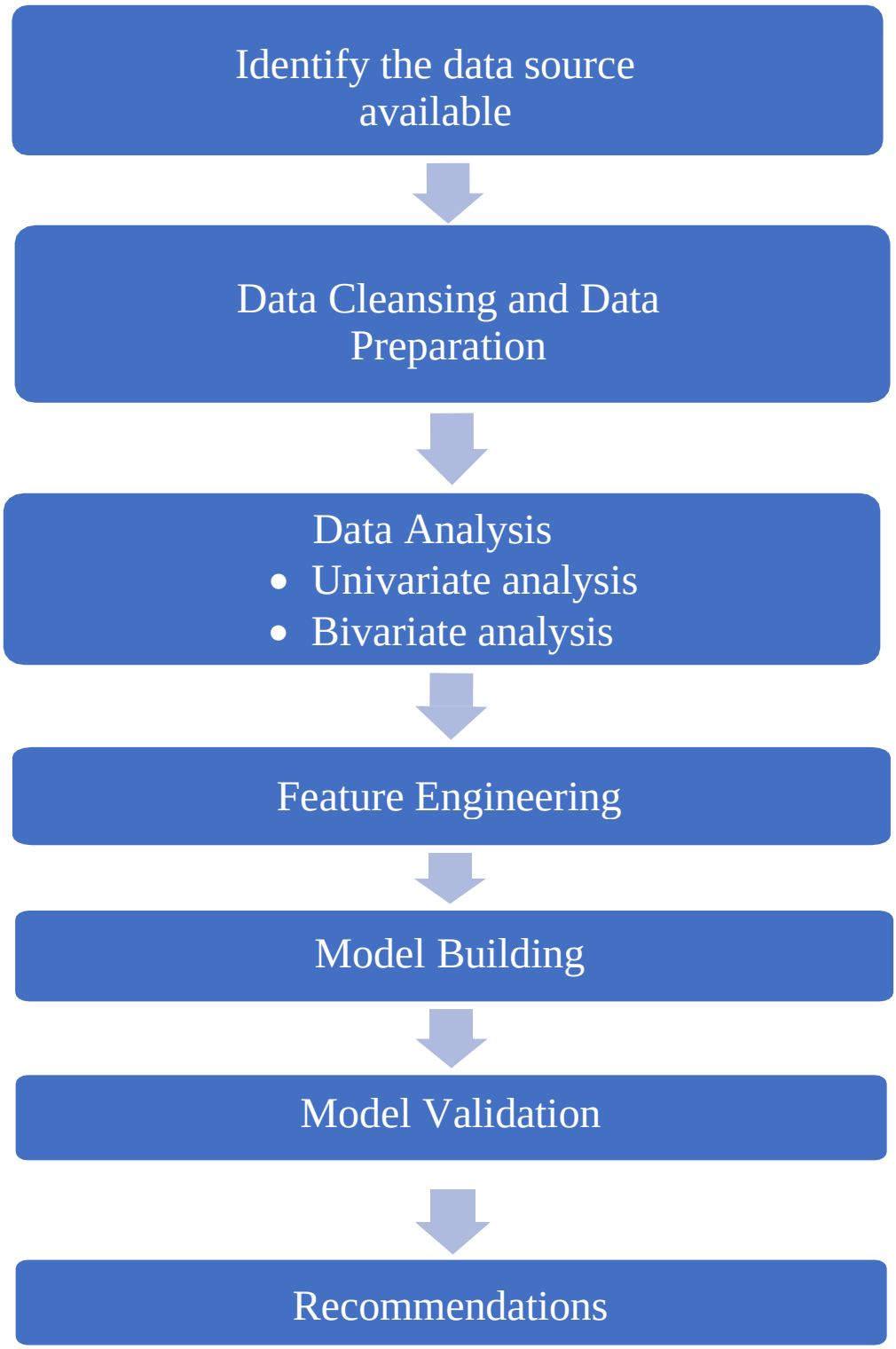
- **BILL_AMT**- Amount of bill statement (INR)

  BILL_AMT1 = amount of bill statement in September 2005
  BILL_AMT2 = amount of bill statement in August 2005
  BILL_AMT3 = amount of bill statement in July 2005
  BILL_AMT4 = amount of bill statement in June 2005
  BILL_AMT5 = amount of bill statement in May 2005
  BILL_AMT6 = amount of bill statement in April 2005

- **PAY_AMT** – Amount to be paid (INR)

  PAY_AMT1 = amount paid in September 2005
  PAY_AMT2 = amount paid in August 2005
  PAY_AMT3 = amount paid in July 2005
  PAY_AMT4 = amount paid in June 2005
  PAY_AMT5 = amount paid in May 2005
  PAY_AMT6 = amount paid in April 2005

PROJECT METHODOLOGY



PRE-PROCESSING DATA ANALYSIS

MISSING/NULL VALUES:

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

From the above table it is understood that there are no null values in the dataset.

OUTLIERS

Feature	No. of outliers	Percentage of Outliers
LIMIT_BAL	1558	5.19
BILL_AMT_1	1556	5.19
BILL_AMT_2	1562	5.20
BILL_AMT_3	1300	5.33
BILL_AMT_4	1550	5.17
BILL_AMT_5	1584	5.28
BILL_AMT_6	1571	5.25
PAY_AMT_1	625	2.08
PAY_AMT_2	445	1.48
PAY_AMT_3	565	1.88
PAY_AMT_4	652	2.17
PAY_AMT_5	631	2.10
PAY_AMT_6	645	2.15

The outlier treatment is done for the raw data by considering the upper and lower limits by setting the upper limit as Mean+ 2*Standard Deviation and the lower limit as Mean -2*Standard Deviations.

UL= Mean + 2 * Standard Deviation

LL=Mean – 2 * Standard Deviation

The number of outliers in the column BILL_AMT_2, BILL_AMT_3, BILL_AMT_5 and BILL_AMT_6 is said to be the highest at around 5%.

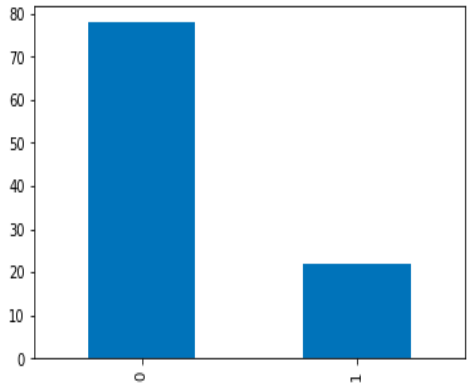
REDUNDANT COLUMNS

ID: Values under this column are unique

EXPLORATORY DATA ANALYSIS

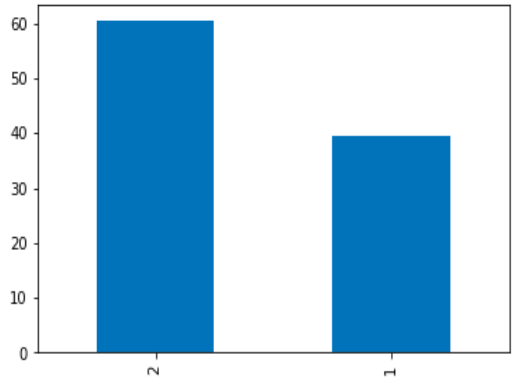
UNIVARIATE ANALYSIS:

```
0    77.88
1    22.12
Name: default payment next month, dtype: float64

0    23364
1     6636
Name: default payment next month, dtype: int64
```



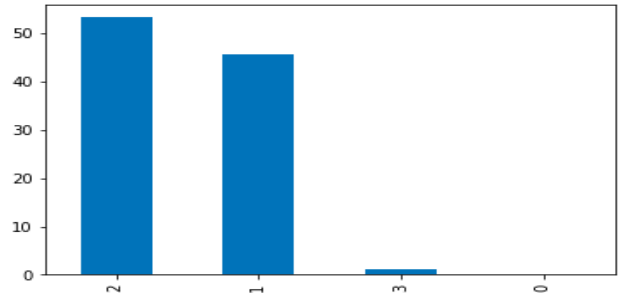
From bar chart we can understand that in the given dataset there are 77% (23364) non-defaulters and 22% (6636) defaulters, thus we understand that the dataset is highly imbalanced.

```
2    60.373333
1    39.626667
Name: SEX, dtype: float64

2    18112
1    11888
Name: SEX, dtype: int64
```



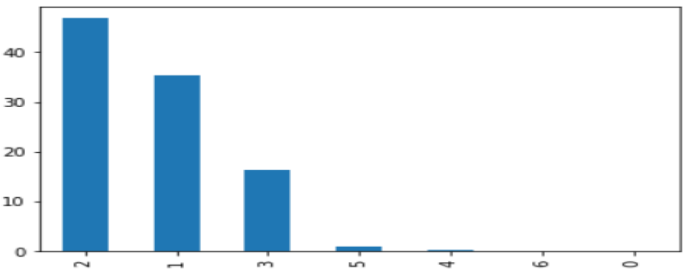
From bar chart we can understand that in the given dataset there are 40% (11888) men and 60% (18112) women, thus we understand that women own more credit cards compared to men.

```
2    53.213333
1    45.530000
3     1.076667
0     0.180000
Name: MARRIAGE, dtype: float64

2    15964
1    13659
3      323
0       54
Name: MARRIAGE, dtype: int64
```

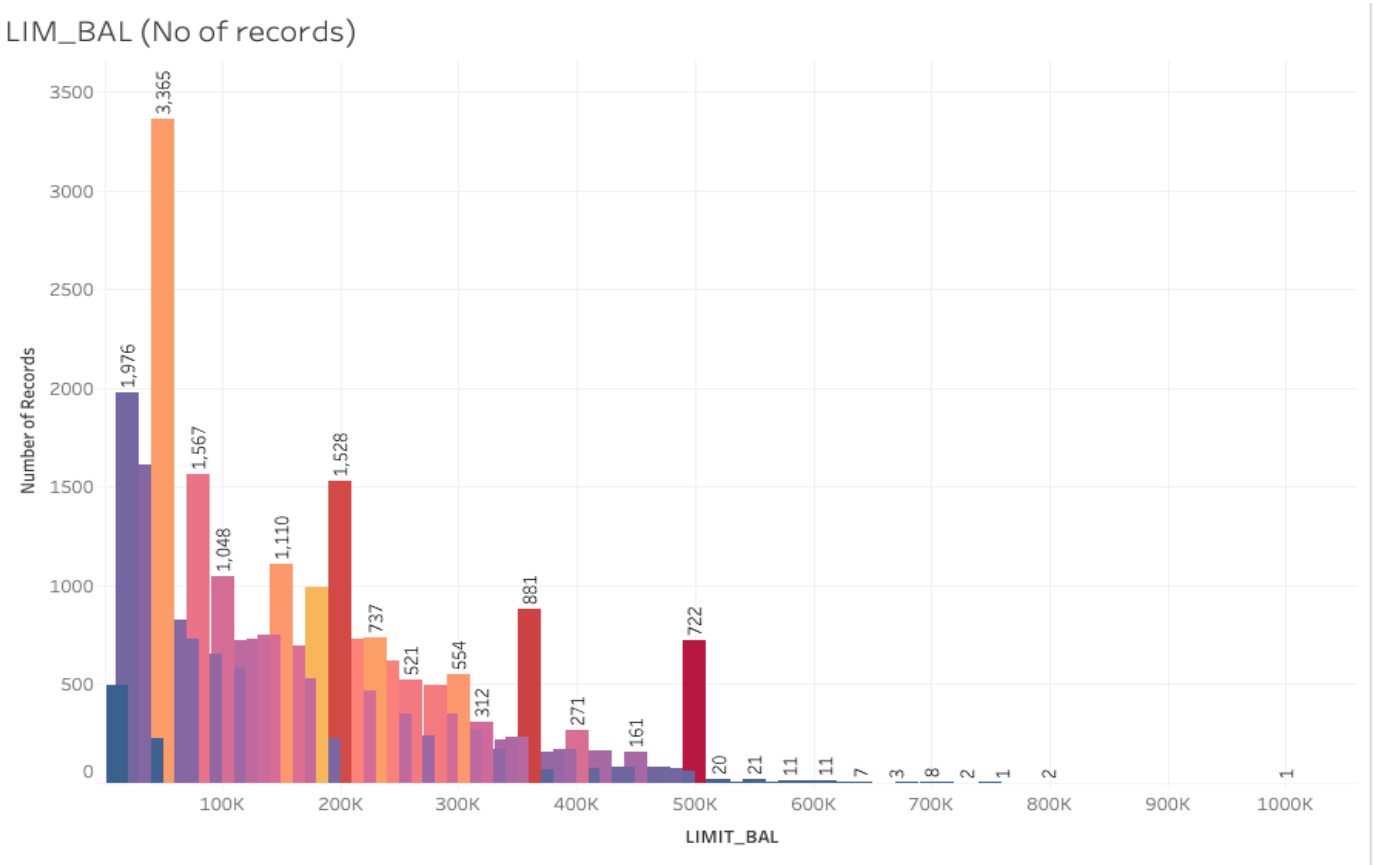


From the above bar chart, we can understand that in the given dataset looking into the marriage column there are 53% (15964) un-married,45% (13659) married, 1% (323) divorcees and 0.1% (54) others.

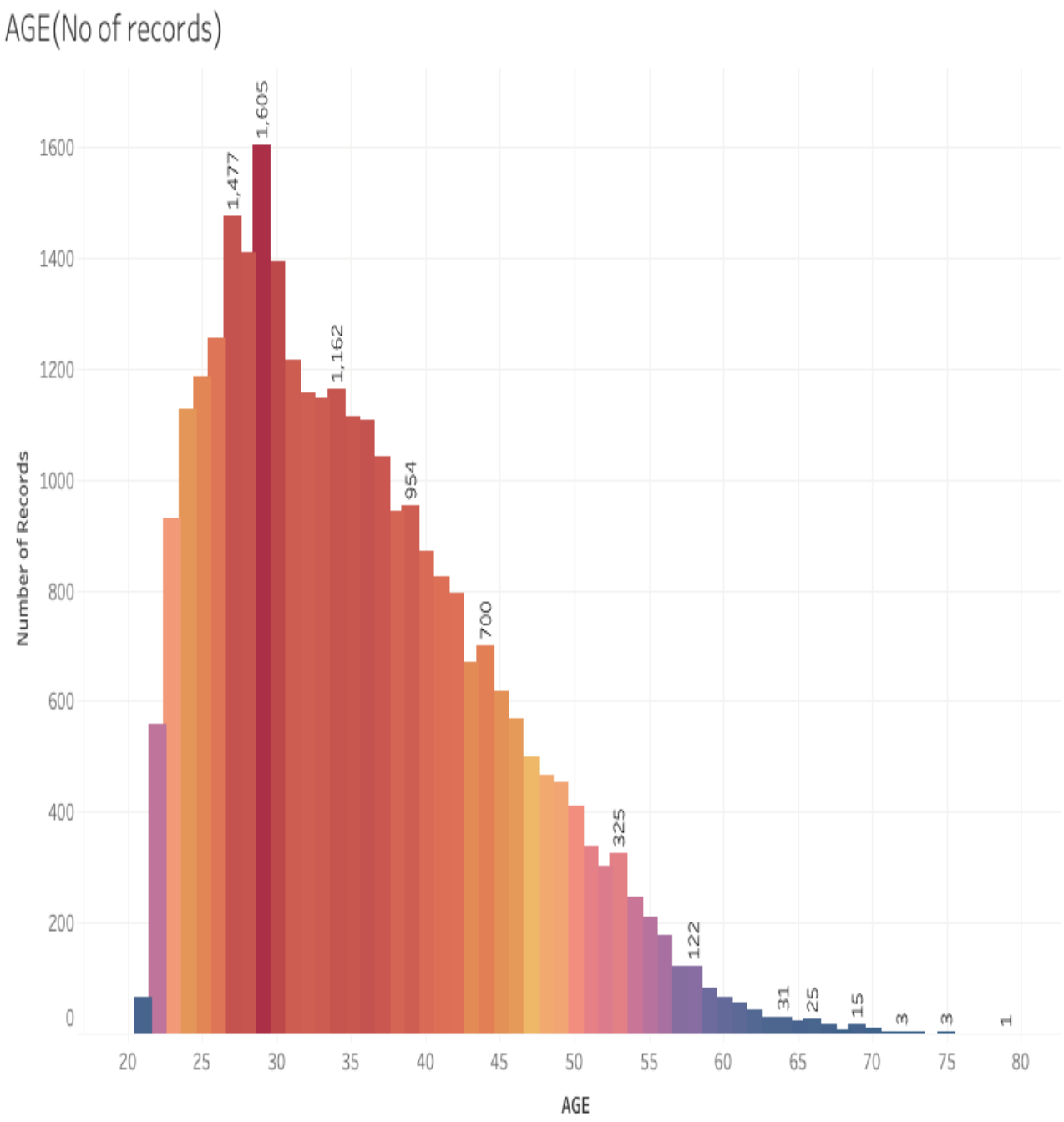
```
2     46.766667
1     35.283333
3     16.390000
5      0.933333
4      0.410000
6      0.170000
0      0.046667
Name: EDUCATION, dtype: float64

2     14030
1     10585
3      4917
5       280
4       123
6        51
0        14
Name: EDUCATION, dtype: int64
```



From the bar chart we can understand that the population have 46% (14030) University level education, 35% (10585) have Graduation, 16% (4917) have Diploma, 0.9% (280) have SSLC, 0.41% (123) have high school ,0.04% (14) have others and 0.17% (51) with no education background.

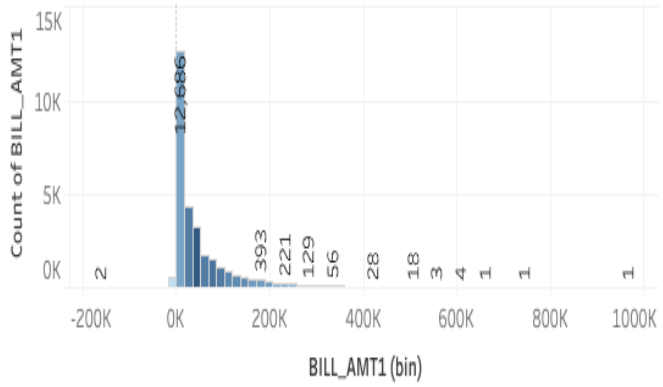


From the analysis we can say that most of the clients fall under the limit balance of less than 200k with so many of them having limit balance less than 100k and the highest limit balance of 1000k which is for only 1 client.

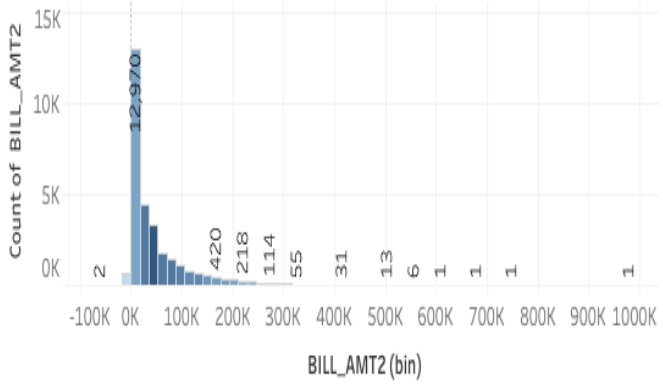


From the analysis we can say that most of the clients fall under the age group of 25 to 50 years.

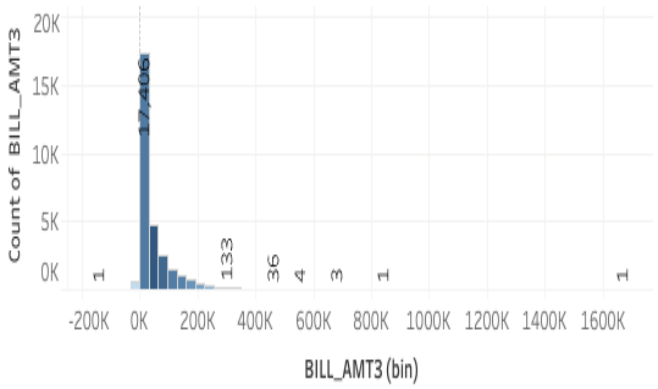
BILL_AMT1



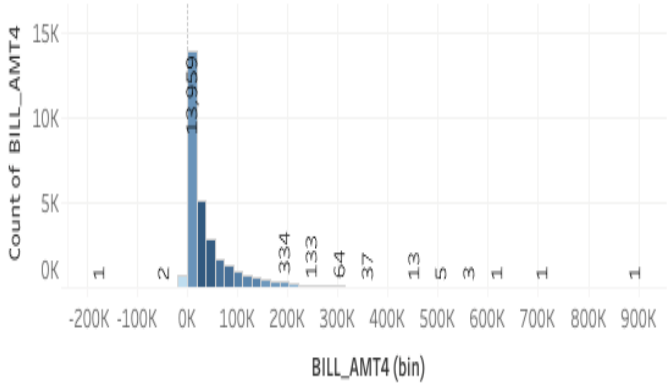
BILL_AMT2



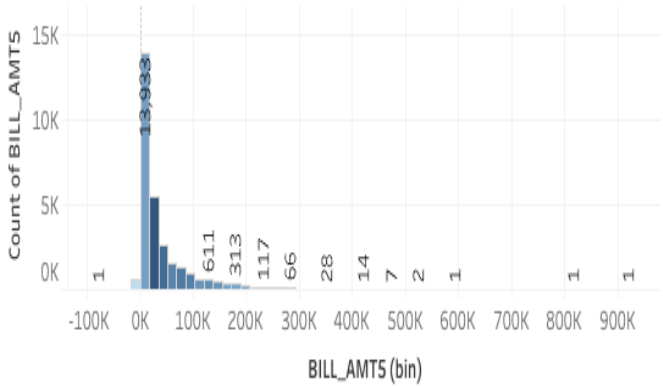
BILL_AMT3



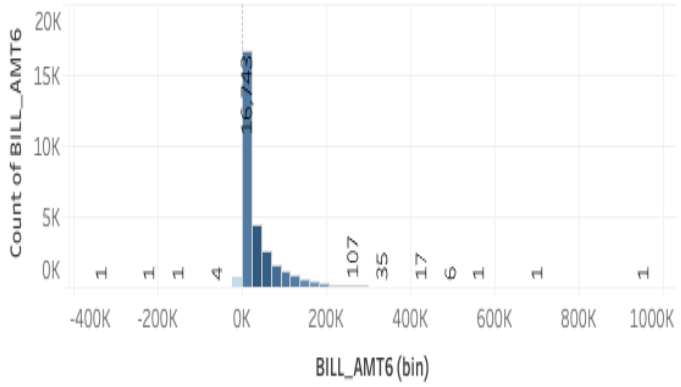
BILL_AMT4



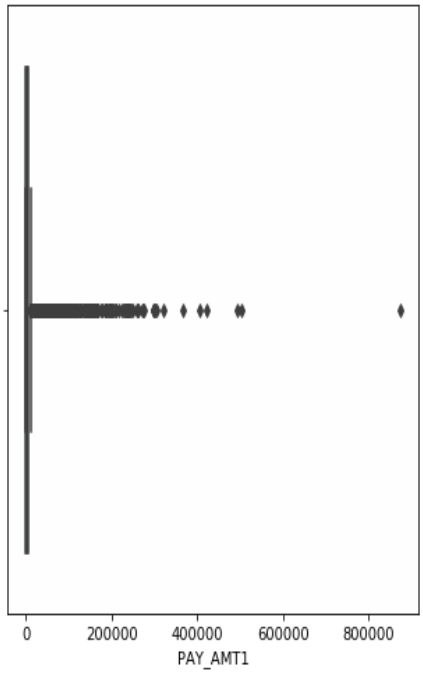
BILL_AMT5

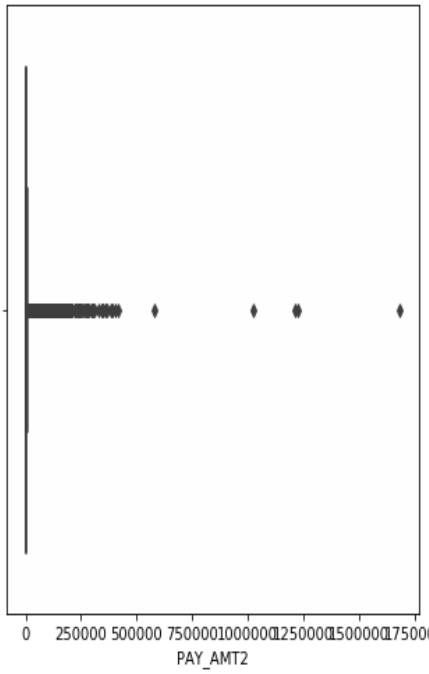


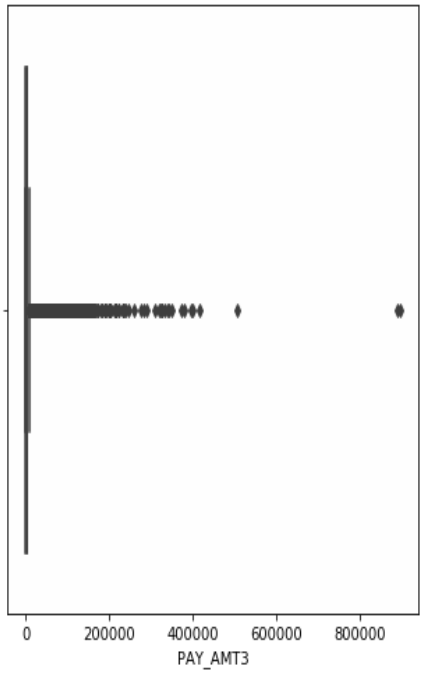
BILL_AMT6

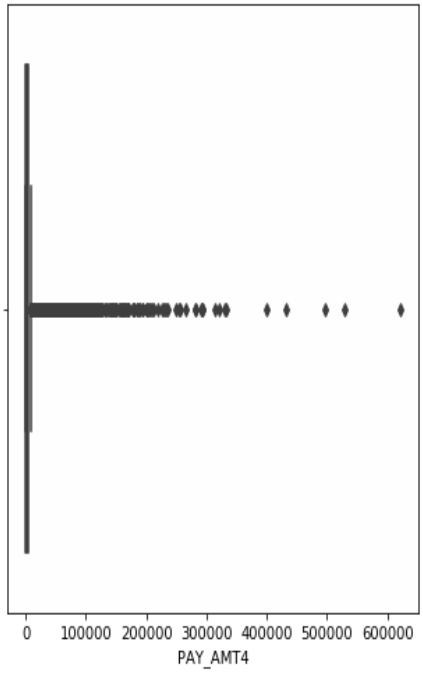


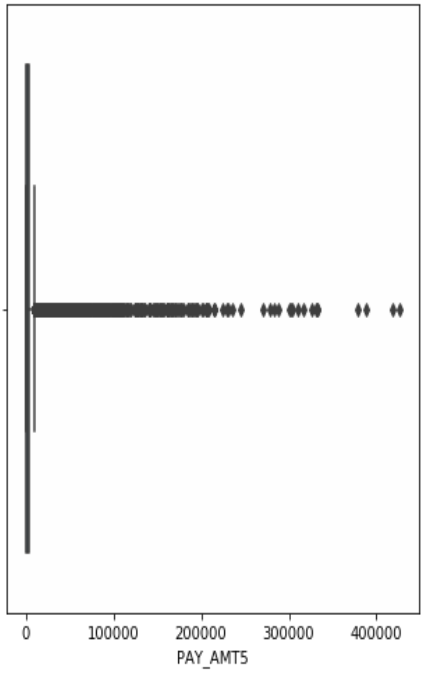
In the analysis we can see that major number of clients have their bill amount up to 200K.

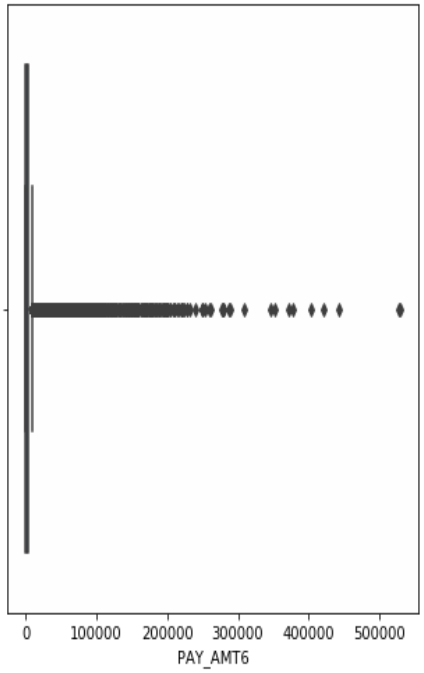






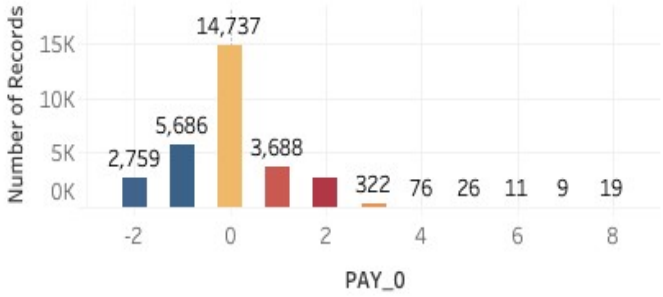




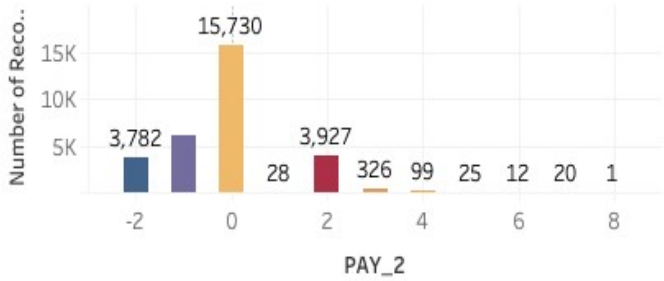


From the above shown box plots we can understand that the pay amount column has few outliers and most of the pay amount is less than 25,000.

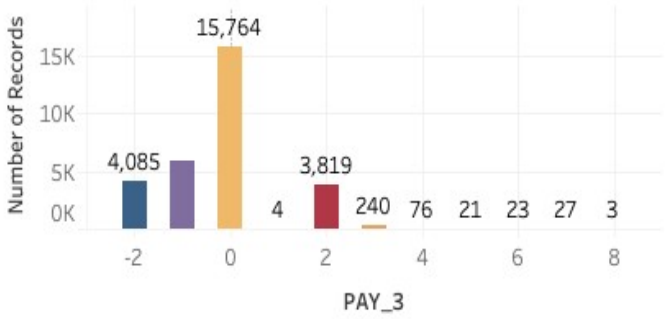
PAY_0(No of records)



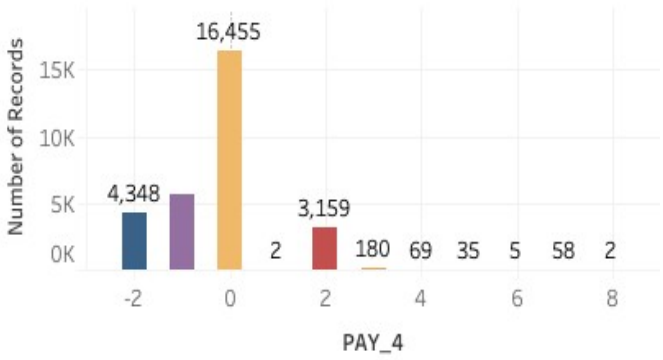
PAY_2(No of records)



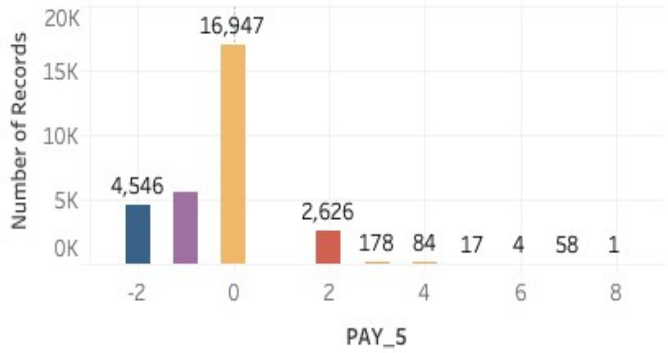
PAY_3(No of records)



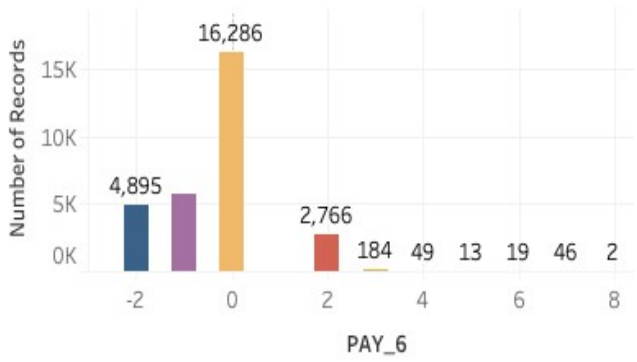
PAY_4(No of records)



PAY_5(No of records)



PAY_6(No of records)

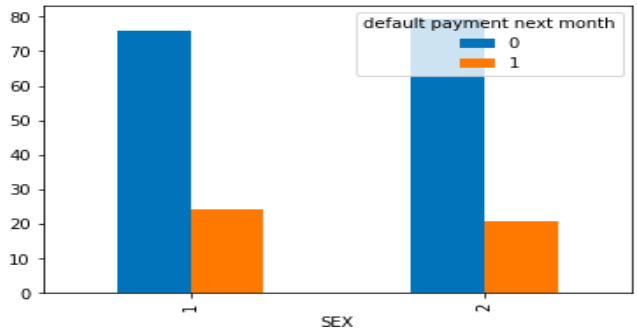


From the bar graph we can see that for all the months the Repayment status with reference to 0 is higher which tells us that most of the clients have no due amounts and they pay the bill on time ,followed by -1 shows that even some clients have a pay duly.

BIVARIATE ANALYSIS:

default payment next month	0	1	All
SEX			
1	9015	2873	11888
2	14349	3763	18112
All	23364	6636	30000

default payment next month	0	1
SEX		
1	75.832773	24.167227
2	79.223719	20.776281

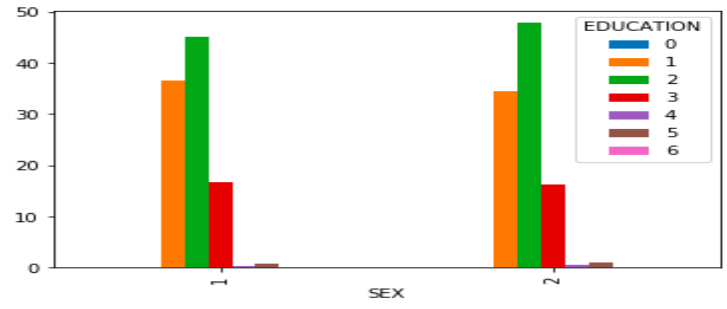


In Above graph, we can see that the Non-defaulters in female (14,349) are higher than the defaulters in number than the male (9015). For Defaulters, Females are higher in number (3763) than male (2873)

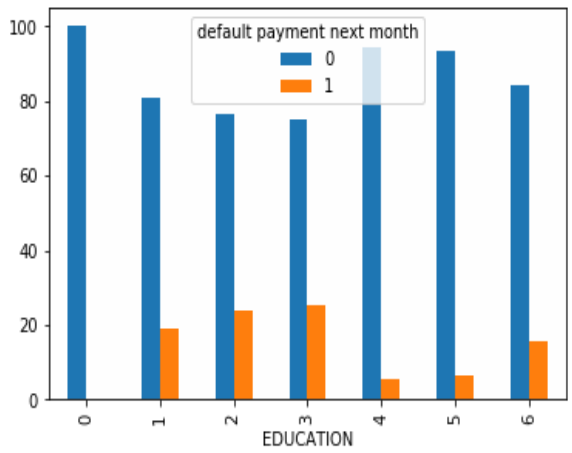
EDUCATION	0	1	2	3	4	5	6	All
SEX								
1	8	4354	5374	1990	42	95	25	11888
2	6	6231	8656	2927	81	185	26	18112
All	14	10585	14030	4917	123	280	51	30000

EDUCATION	0	1	2	3	4	5
SEX						
1	0.067295	36.625168	45.205249	16.739569	0.353297	0.799125
2	0.033127	34.402606	47.791519	16.160557	0.447217	1.021422

EDUCATION	6
SEX	
1	0.210296
2	0.143551



From the bivariate analysis of SEX vs Education, we can infer that most of the male 45% (5374), female 47% (8656) have University education followed by male 36% (4354), female 34% (6231) have done their graduation which is followed by male 17% (1990), female 16% (2927) have done Diploma.



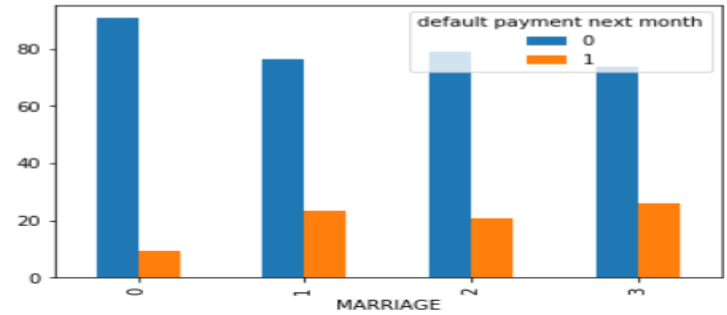
default payment next month	0	1	All
EDUCATION			
0	14	0	14
1	8549	2036	10585
2	10700	3330	14030
3	3680	1237	4917
4	116	7	123
5	262	18	280
6	43	8	51
All	23364	6636	30000

default payment next month	0	1
EDUCATION		
0	100.000000	0.000000
1	80.765234	19.234766
2	76.265146	23.734854
3	74.842384	25.157616
4	94.308943	5.691057
5	93.571429	6.428571
6	84.313725	15.686275

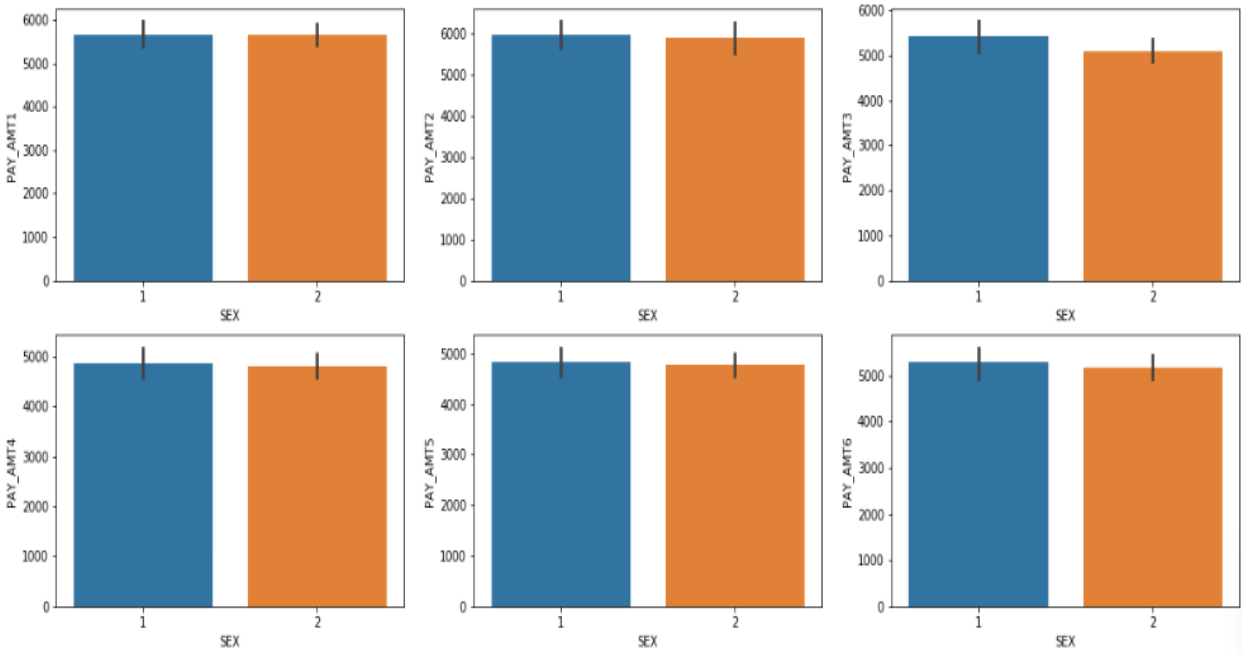
In the analysis we see that there are a greater number of non-defaulters across all the education background and the population is higher for the 2 (University) i.e. 10,700 followed by 1(graduation) i.e. 8549.

default payment next month	0	1	All
MARRIAGE			
0	49	5	54
1	10453	3206	13659
2	12623	3341	15964
3	239	84	323
All	23364	6636	30000

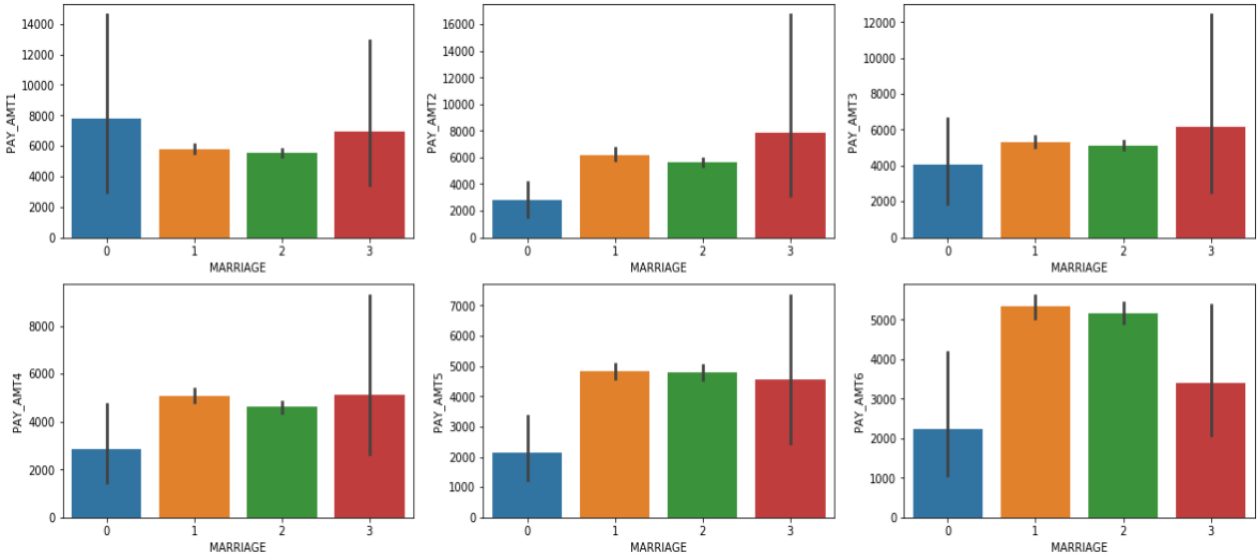
default payment next month	0	1
MARRIAGE		
0	90.740741	9.259259
1	76.528296	23.471704
2	79.071661	20.928339
3	73.993808	26.006192



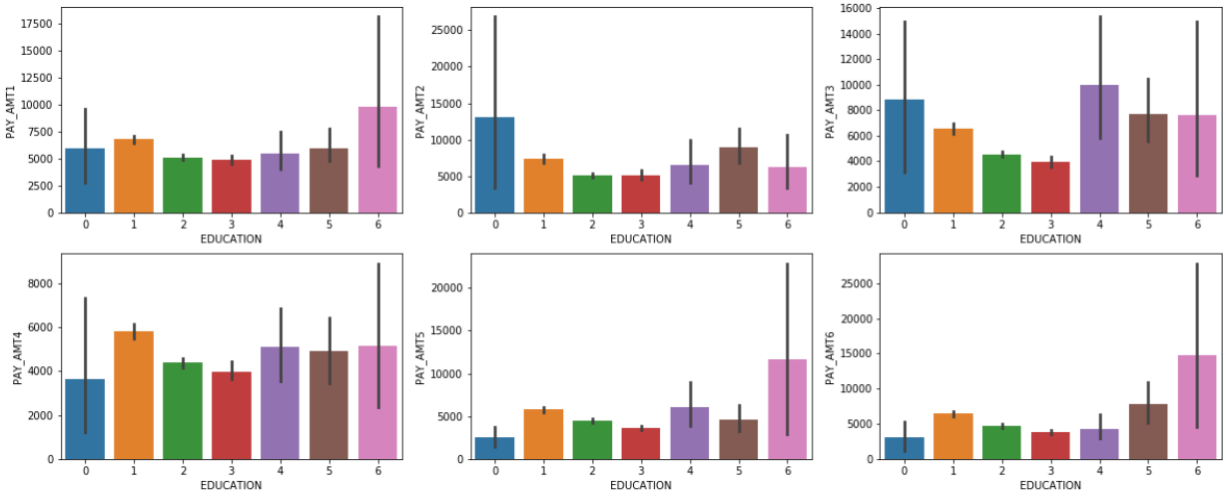
In the analysis we can see that the clients who are UNMARRIED are more in population and so as the number of non–defaulters are higher for 2 (12623) followed by the MARRIED people 1 (10453).



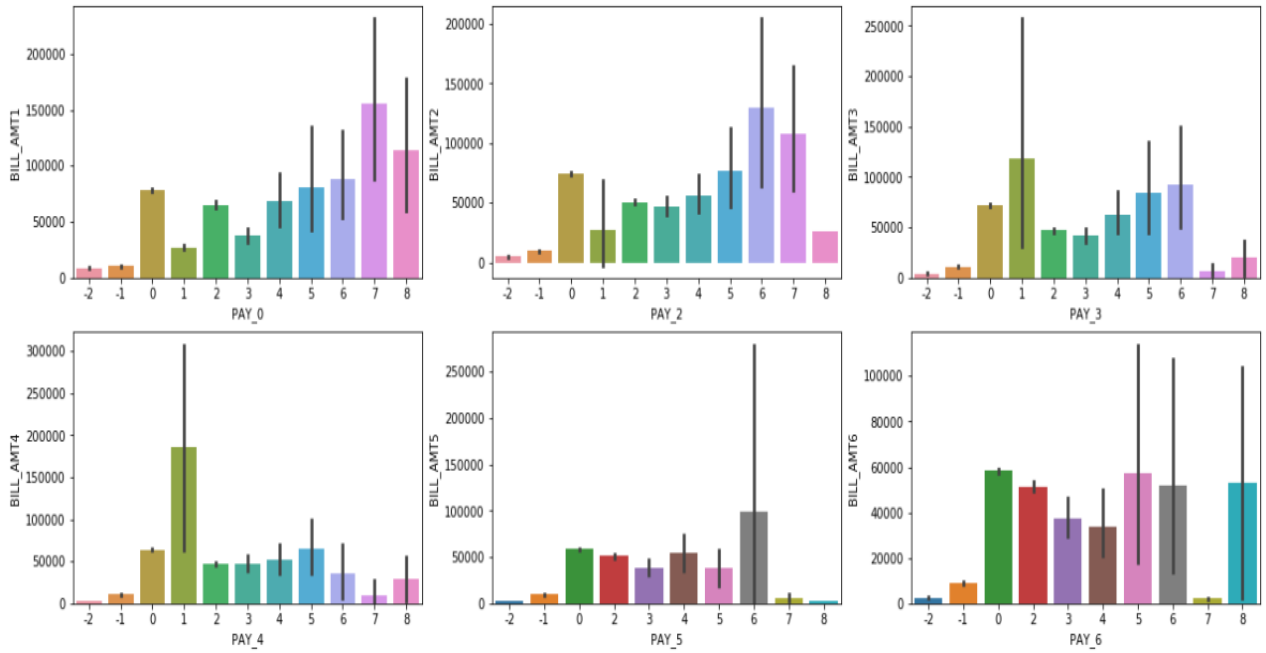
From the above analysis of payment amount and Sex for different months from April to September we can see that the pay amount is almost same for both the gender with a slight variation.



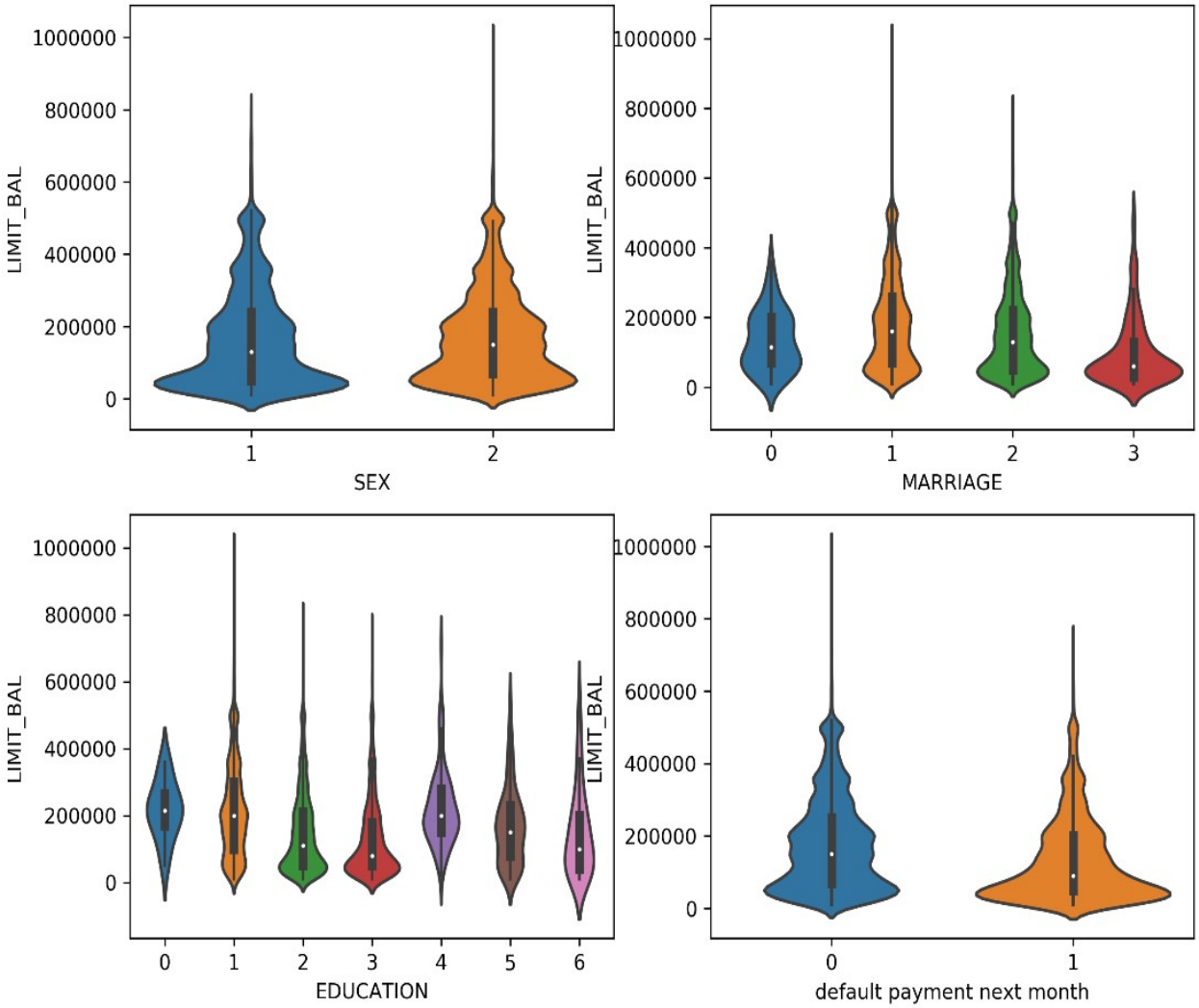
From the above analysis of payment amount and Marriage for different months from April to September we can see that the pay amount is around 4000-6000 for married and un-married people. Hence targeting such customers would give a consistent revenue.



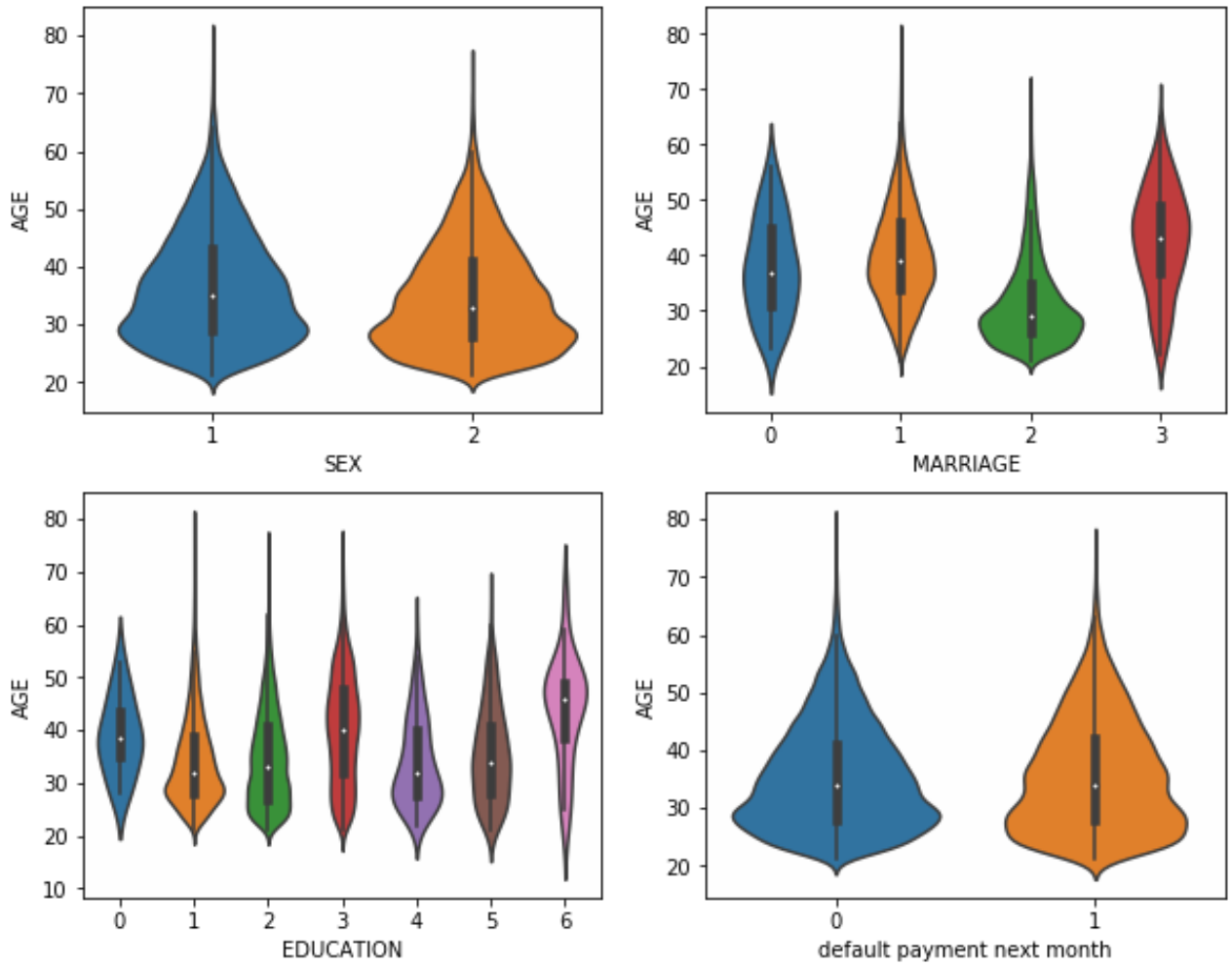
From the above analysis between pay amount and Education for different months from April to September we can state that the pay amount for people who have an education background as No Education have utilized the credit cards more than the others. Hence, they are one of our potential customers.



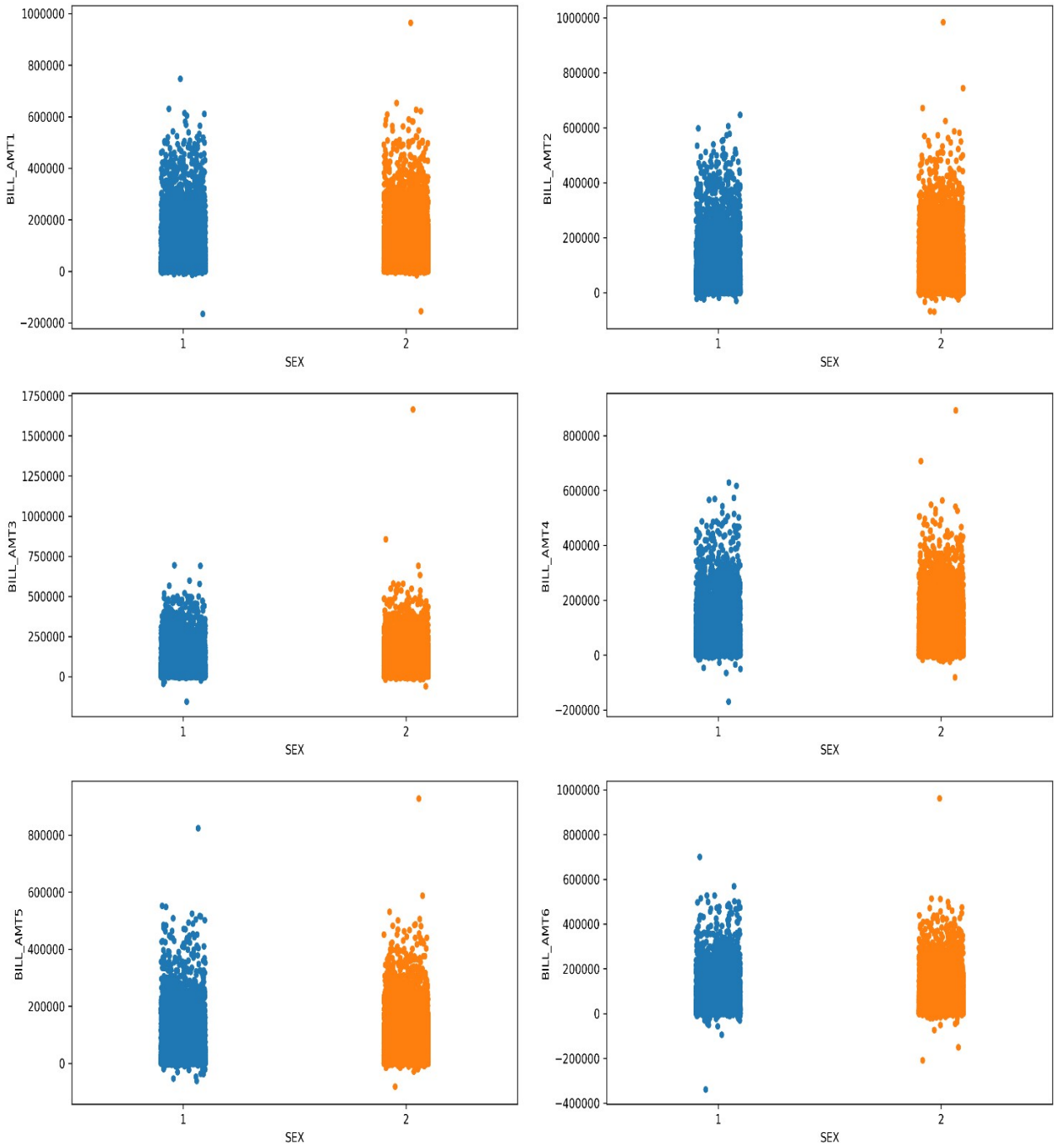
From the graph, it can be inferred that the people who have not been paying their credit card bills for more than 3 months are utilising their credit cards more than the people who are paying on time. Hence some fine's or some restrictions should be made for the best interest of the bank.



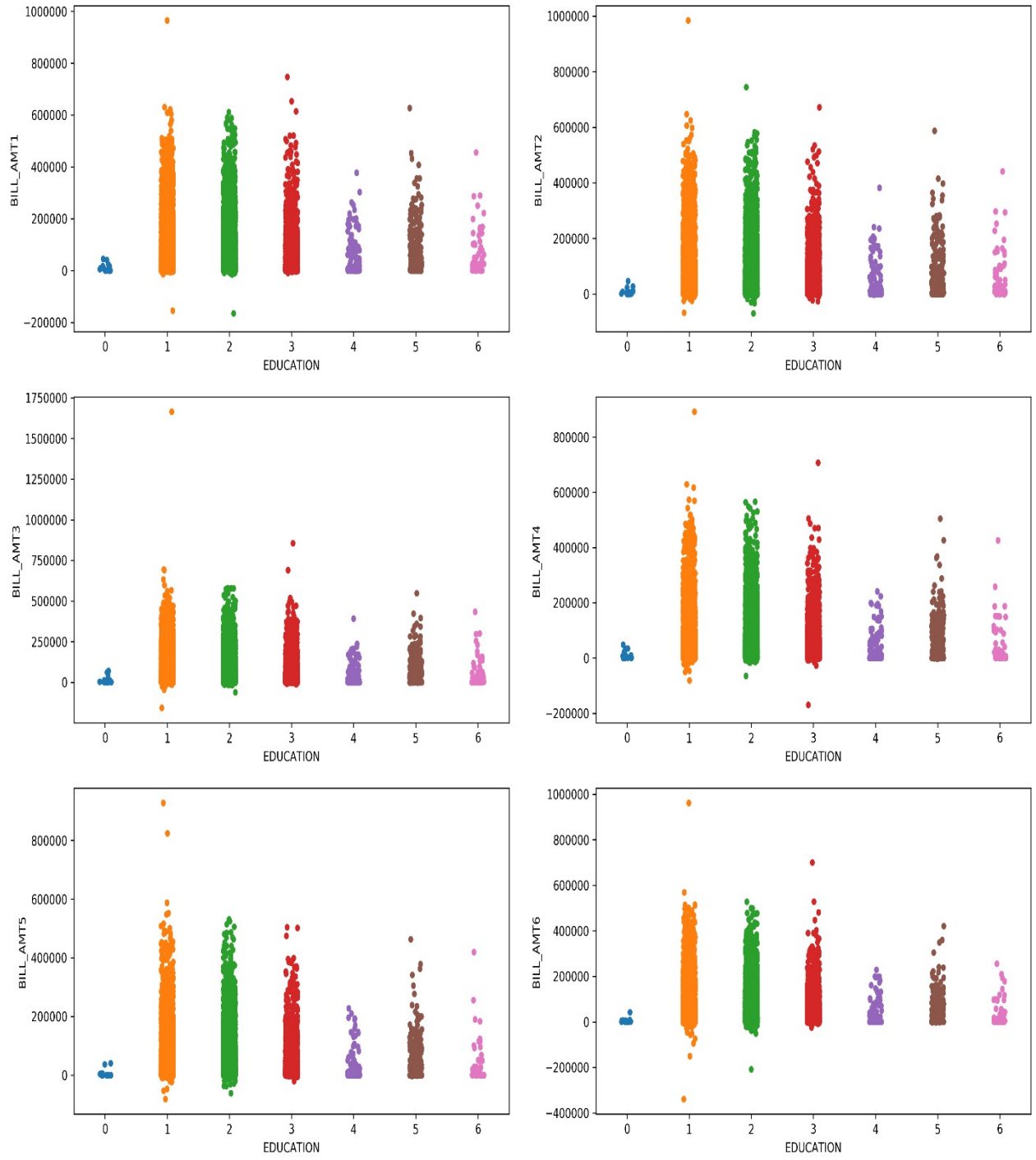
- The graph one denotes that the male and female have almost the same amount of limit balance.
- The Married People tend to have the highest number of Limit Balance.
- The people with their highest education as Graduation have the highest Limit Balance.
- The number of 0s are more than 1s in the target column, indicating the dataset is highly imbalanced.



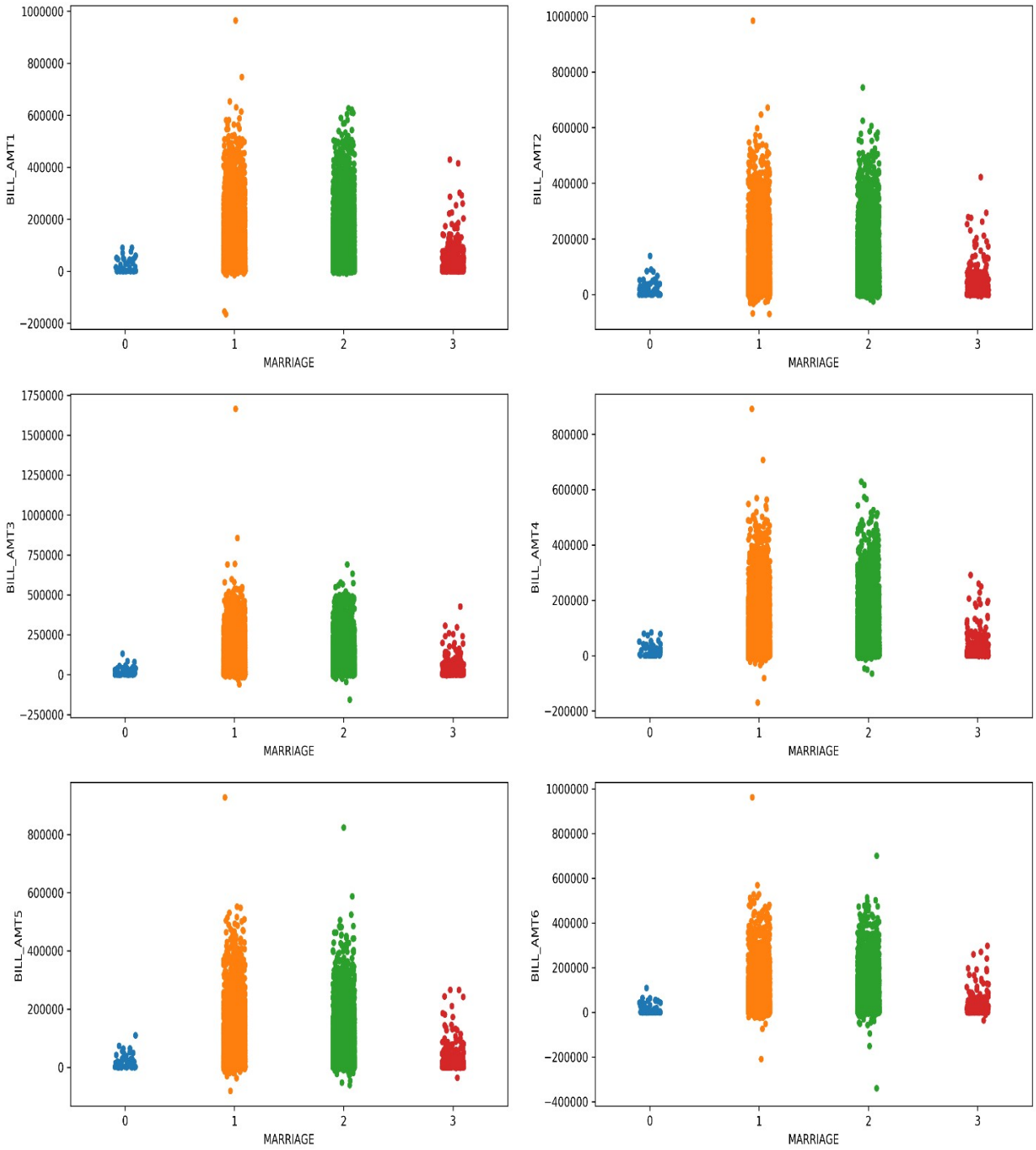
- The graph one denotes that the male clients are more at the age group of 30 to 40 compared to female clients.
- The Married and single clients tend to have more or else same age band width.
- The people with their highest education as Graduation have the highest age band width.
- The number of 0s are more than 1s in the target column, indicating the dataset is highly imbalanced.



From the above analysis of Bill amount and Sex for different months from April to September we can see that the pay amount is almost same for both the gender with a slight variation.

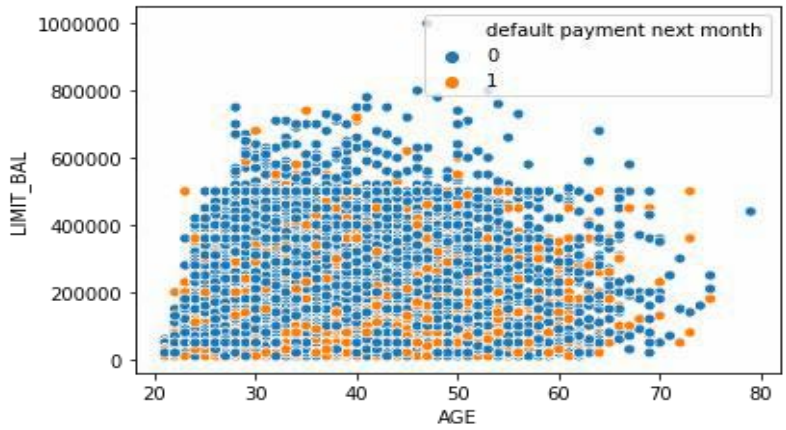


From the above strip plot it can be inferred that the people who have Educational qualification as Graduate and University are said to be using the credit card more than others from the month of April to September.

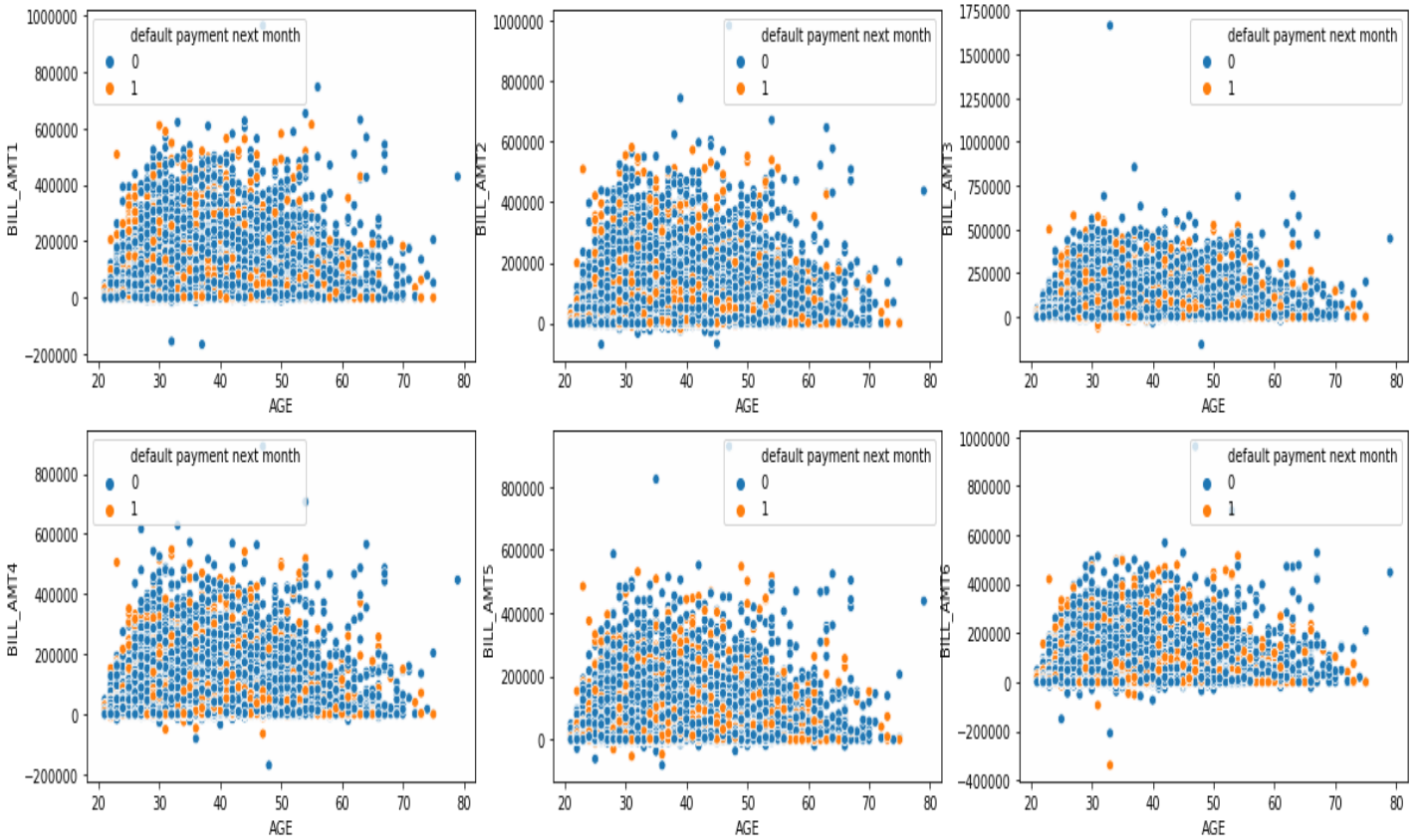


From the above strip plot it can be inferred that the people who are married are said to be using the credit card more than the unmarried people from the month of April to September.

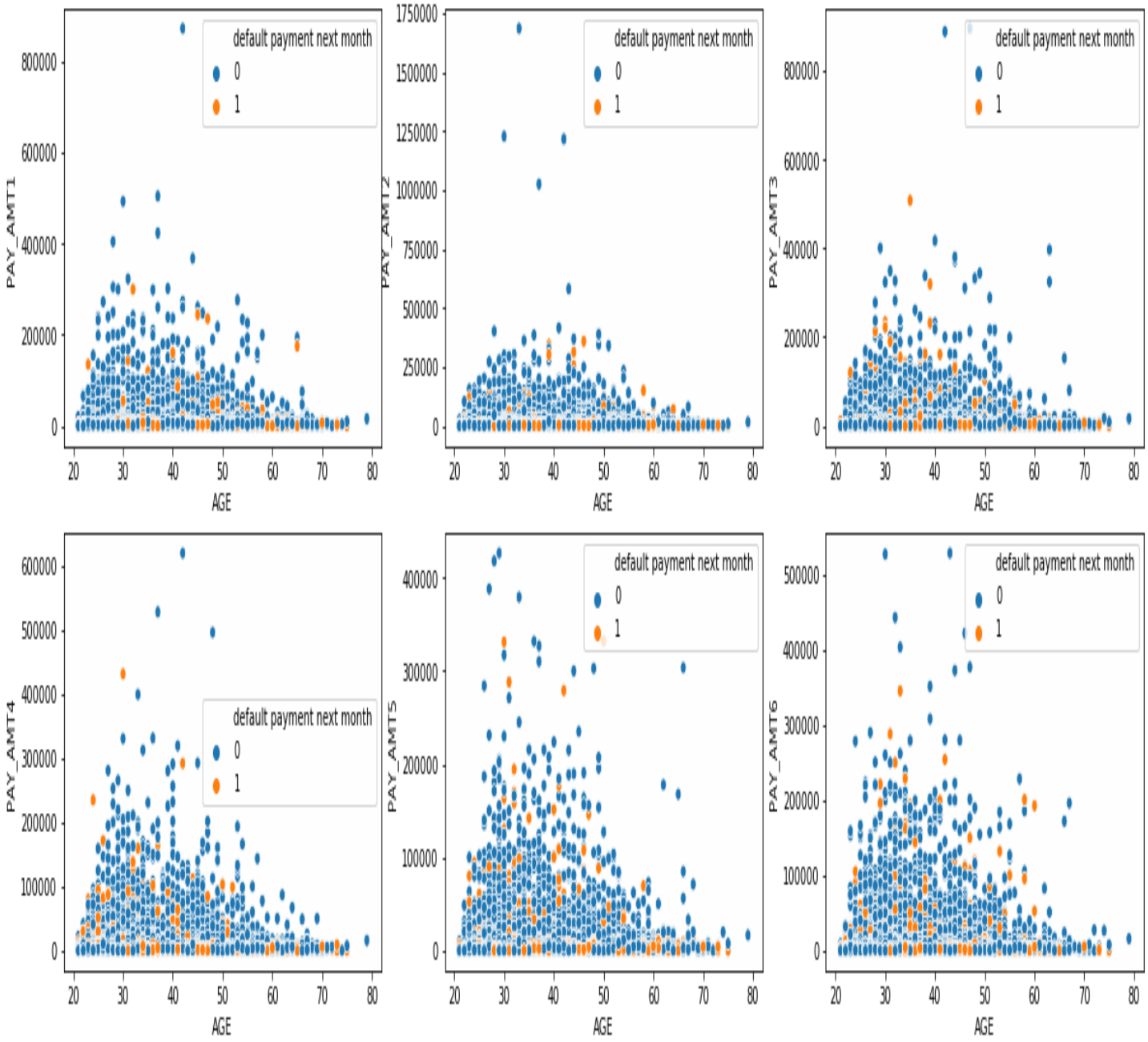
MULTIVARIANT ANALYSIS:



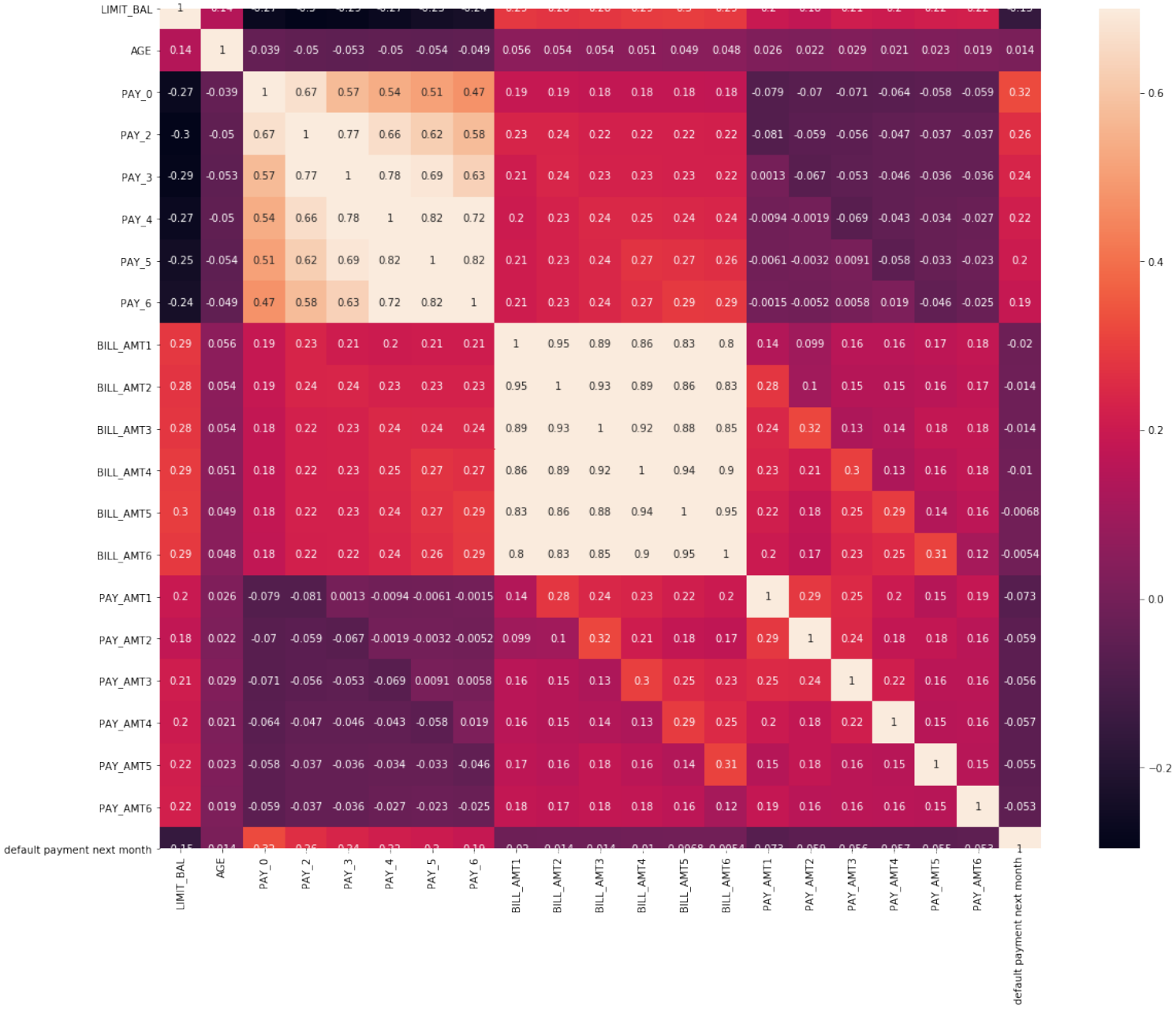
From the analysis we can see that a greater number of defaulters lie between the age group of 30 –50 since there is large population around the age group in the data. They are the one who have been sanctioned the heist limit balance.



From the Scatter plot we can see that the highest population is around the age group of 20 – 60, they are clients who spend more having the highest bill amount. Non-defaulters are more in the very same age group compared to the defaulters.



From the Scatter plot we can see that the highest population is around the age group of 20 – 53, beyond which the payment is declining. The number of defaulters are also high in the same age group, as the non-defaulters are more than the defaulters in this particular age group they also contributing to the highest pay amount hence we can they this age group is the potential customer.



The above image represents the correlation between all the features, indicating the features that are highly correlated to target variable(default_payment_next_month). Here PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6 is correlated more with the target feature compared to the other features.

STATISTICAL TEST

HYPOTHESIS TESTING - 1:

To check the correlation among categorical features we need to carry out chi-square test.

Hypothesis Statements:

- Step 1: State the Null Hypothesis and Alternate Hypothesis

  H_0: There is no dependence between categorical and target feature
  H_α: There is a dependency between categorical and target feature

- Step 2: State the level of significance

  Here we select significance level (α)=0.05

- Step 3: Choose the appropriate test

  Here we have selected the chi2_contingency test to find if the categorical features are significant or not.

- Step 4: Calculate the p-value

```
Feature SEX is significant and the pvalue = 9.876093788177908e-10
Feature EDUCATION is significant and the pvalue = 1.312857313478719e-30
Feature MARRIAGE is significant and the pvalue = 8.879828628366667e-05
```

- Step 5: Decision on whether to accept or reject the NULL hypothesis

  From the Hypothesis test we can see that all Categorical columns like SEX, EDUCATION and MARRIAGE have p-value <0.05 and hence we can reject H0 and say that the categorical features are significant.

HYPOTHESIS TESTING - 2:

To check the correlation between the means of numerical and target feature.

Hypothesis Statements:

- **Step 1:** State the Null Hypothesis and Alternate Hypothesis

  H0: Mean of the non-defaulters in the numerical columns = Mean of the defaulters in the numerical columns
  H_a: Mean of the non-defaulters in the numerical columns! = Mean of the defaulters in the numerical columns.

- **Step 2:** State the level of significance

  Here we select significance level (α)=0.05

- Step 3: Choose the appropriate test

  Here we have selected the T test to find if the numerical features are significant or not.

- Step 4: Calculate the p-value

  Feature LIMIT_BAL is significant and the pvalue = 1.3022439532597397e-157
  Feature AGE is significant and the pvalue = 0.01613684589016383
  Feature BILL_AMT1 is significant and the pvalue = 0.0006673295491221741
  Feature BILL_AMT2 is significant and the pvalue = 0.013957362392434761
  Feature BILL_AMT3 is significant and the pvalue = 0.014769982710723002
  Feature BILL_AMT4 is not significant and the pvalue = 0.07855564157651403
  Feature BILL_AMT5 is not significant and the pvalue = 0.24163444291382874
  Feature BILL_AMT6 is not significant and the pvalue = 0.3521225212306479
  Feature PAY_AMT1 is significant and the pvalue = 1.1464876142241624e-36
  Feature PAY_AMT2 is significant and the pvalue = 3.1666567628387115e-24
  Feature PAY_AMT3 is significant and the pvalue = 1.841770291503132e-22
  Feature PAY_AMT4 is significant and the pvalue = 6.830941601370003e-23
  Feature PAY_AMT5 is significant and the pvalue = 1.2413447727776169e-21
  Feature PAY_AMT6 is significant and the pvalue = 3.033589072770243e-20

- Step 5: Decision on whether to accept or reject the NULL hypothesis

  From the test we can say that the numerical features like BILL_AMT 4, BILL_AMT 5, and BILL_AMT 6 have the p-value > 0.05 and hence we can say that these numerical features are not significant.

MODELS CONSIDERED:

We have built the following base models: Logistic Regression and Decision Tree Classifiers to find the accuracy of the given data as follows: -

- Raw Data
- Outlier Treated Data
- Scaled Data

RAW DATA

Logistic Regression Scores:

```
Train score 0.7791428571428571
Test score 0.7774444444444445
0.6362201909366063
              precision    recall  f1-score   support

           0       0.78      1.00      0.88     16364
           1       0.25      0.00      0.00      4636

    accuracy                           0.78     21000
   macro avg       0.51      0.50      0.44     21000
weighted avg       0.66      0.78      0.68     21000
```

The basic model of Logistic Regression for the raw data has given the score of Train and Test as 0.779 &0.777.

Decision Tree Classifiers:

```
Accuracy score on Train: 1.0
Confusion Matrix on test:
[[16364     0]
 [    0  4636]]
AUC of Train: 1.0
```

```
Accuracy score on Test: 0.7204444444444444
Confusion Matrix on test:
[[5664 1336]
 [1180  820]]
AUC of Test: 0.6095714285714285
```

The basic model of Decision Tree Classifier for the raw data has given us the score of Train as 1.0 and test as 0.720, Which tells us that the model is overfit.

OUTLIER TREATED DATA:

Logistic Regression:

```
Train score 0.7791904761904762
Test score 0.7777777777777778
0.6488845875086392
              precision    recall  f1-score   support

           0       0.78      1.00      0.88     16364
           1       0.00      0.00      0.00      4636

    accuracy                           0.78     21000
   macro avg       0.39      0.50      0.44     21000
weighted avg       0.61      0.78      0.68     21000
```

The outlier treatment is done for the raw data by considering the upper and lower limits by setting the upper limit as Mean+ 2*Standard Deviation and the lower limit as Mean -2*Standard Deviations.

UL= Mean + 2 * Standard Deviation
LL=Mean – 2 * Standard Deviation

All the values above the upper limit have been mapped to the upper limit and the values below lower limit have been mapped to the lower limit.
The model of Logistic Regression with the outlier treatment has given the score of Train and Test as 0.77 9 & 0.777, which shows there is no impact of outliers in the data set.

Decision Tree Classifier:

```
Accuracy score on Train: 1.0
Confusion Matrix on test:
[[16364     0]
 [    0  4636]]
AUC of Train: 1.0
```

```
Accuracy score on Test: 0.7222222222222222
Confusion Matrix on test:
[[5682 1318]
 [1182  818]]
AUC of Test: 0.6103571428571428
```

The Decision Tree Classifier model for the outlier treated data has given us the score of Train and test as 1.0 & 0.722, Which tells the model is still overfit model but these scores can be still improved by scaling.

SCALED DATA:

Logistic Regression:

```
Train score 0.7791904761904762
Test score 0.7777777777777778
0.6488845875086392
              precision    recall  f1-score   support

           0       0.78      1.00      0.88     16364
           1       0.00      0.00      0.00      4636

    accuracy                           0.78     21000
   macro avg       0.39      0.50      0.44     21000
weighted avg       0.61      0.78      0.68     21000
```

The data is scaled by the min and max scaling techniques and the categorical columns are converted in to dummies. The model of Logistic Regression with Scaling & dummies has given the score of Train and Test as 0.779 & 0.777, which shows still there is no change in the score. This Score can be further improved by the feature selection process.

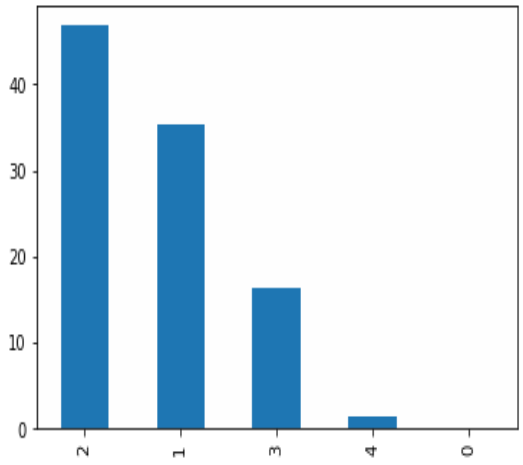
Decision Tree Classifier:

```
Accuracy score on Train: 1.0
Confusion Matrix on test:
[[16364     0]
 [    0  4636]]
AUC of Train: 1.0
```

```
Accuracy score on Test: 0.7216666666666667
Confusion Matrix on test:
[[5679 1321]
 [1184  816]]
AUC of Test: 0.609642857142857
```

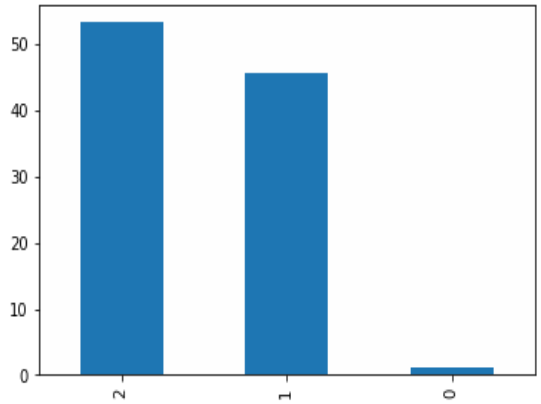
The Decision Tree Classifier model for the Scaled data has given us the score of Train and test as 1.0 & 0.72, Which tells the model is still overfitting and no significant improvement in the scores have been observed.

FEATURE ENGINEERING:



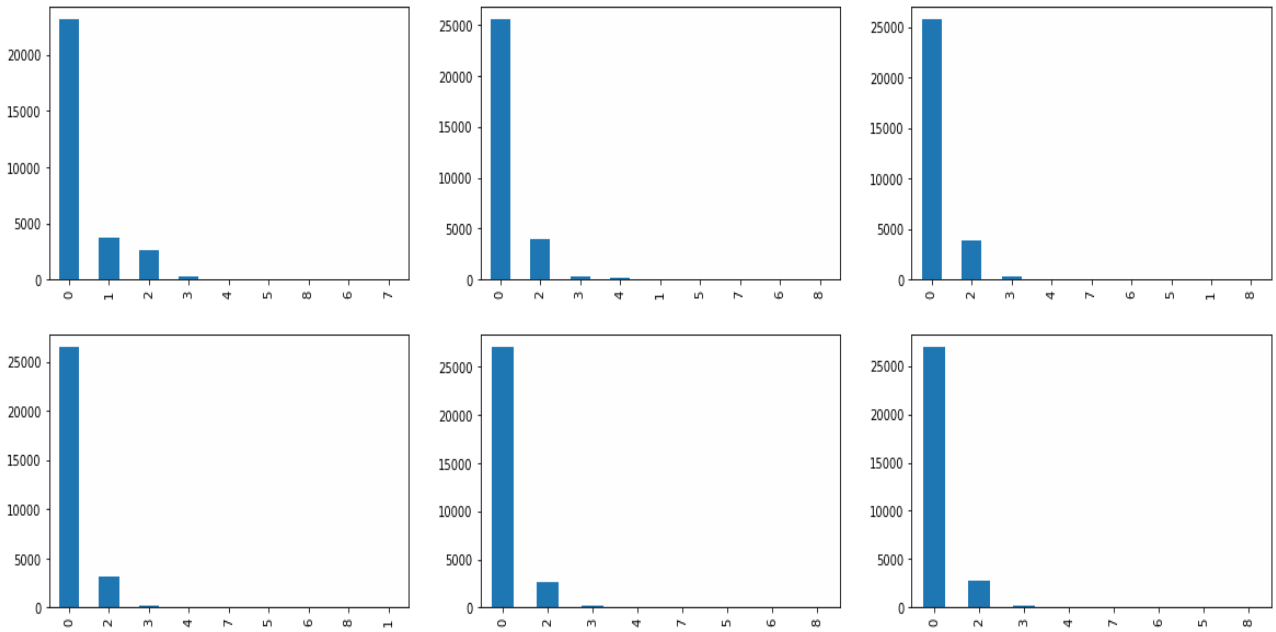

```
2    46.766667
1    35.283333
3    16.390000
4     1.513333
0     0.046667
Name: EDUCATION, dtype: float64

2    14030
1    10585
3     4917
4      454
0       14
Name: EDUCATION, dtype: int64
```

In the Education column we are replacing the Sub categories of 5 (SSLC)& 6 (No education) with 4 (High School) and now from the graph we can see that 1.5% (454) come under 4 (High School).




```
2    53.213333
1    45.530000
0     1.256667
Name: MARRIAGE, dtype: float64

2    15964
1    13659
0      377
Name: MARRIAGE, dtype: int64
```

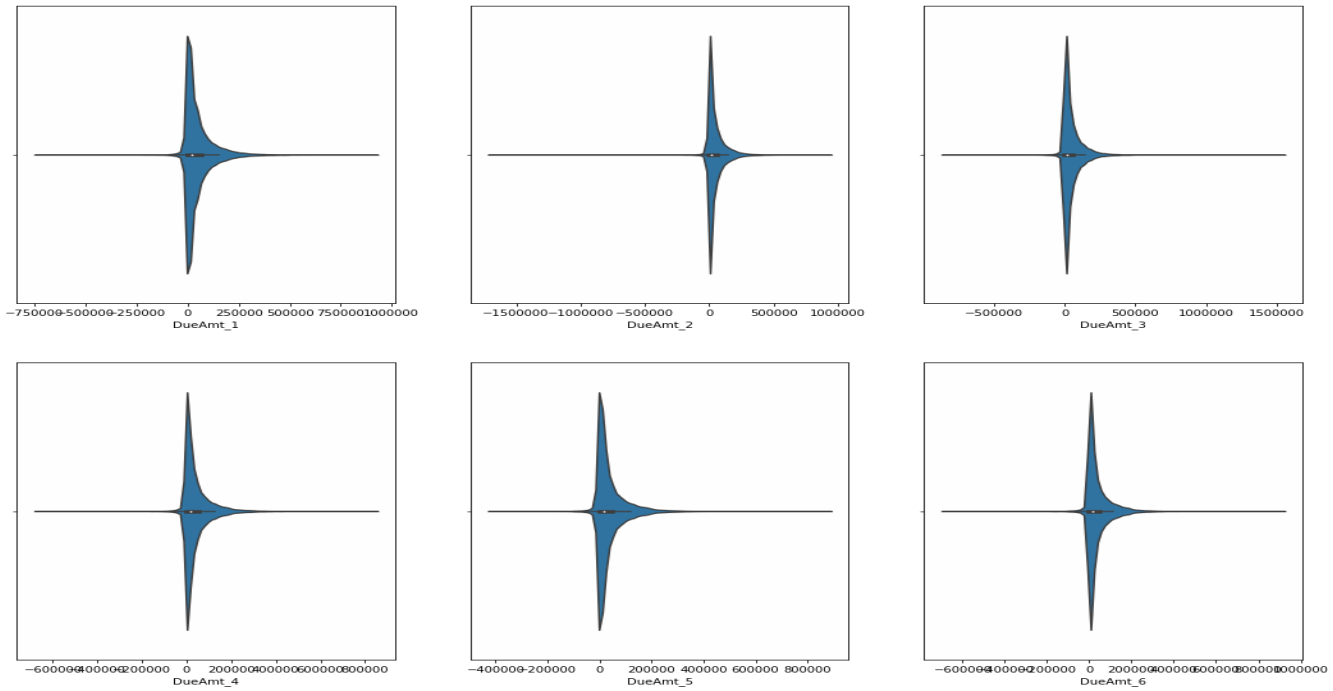
In the Marriage column we are replacing the Sub category 3 (divorce) with 0 (others) and from the graph we can see that 1.2% (377) come under Others.



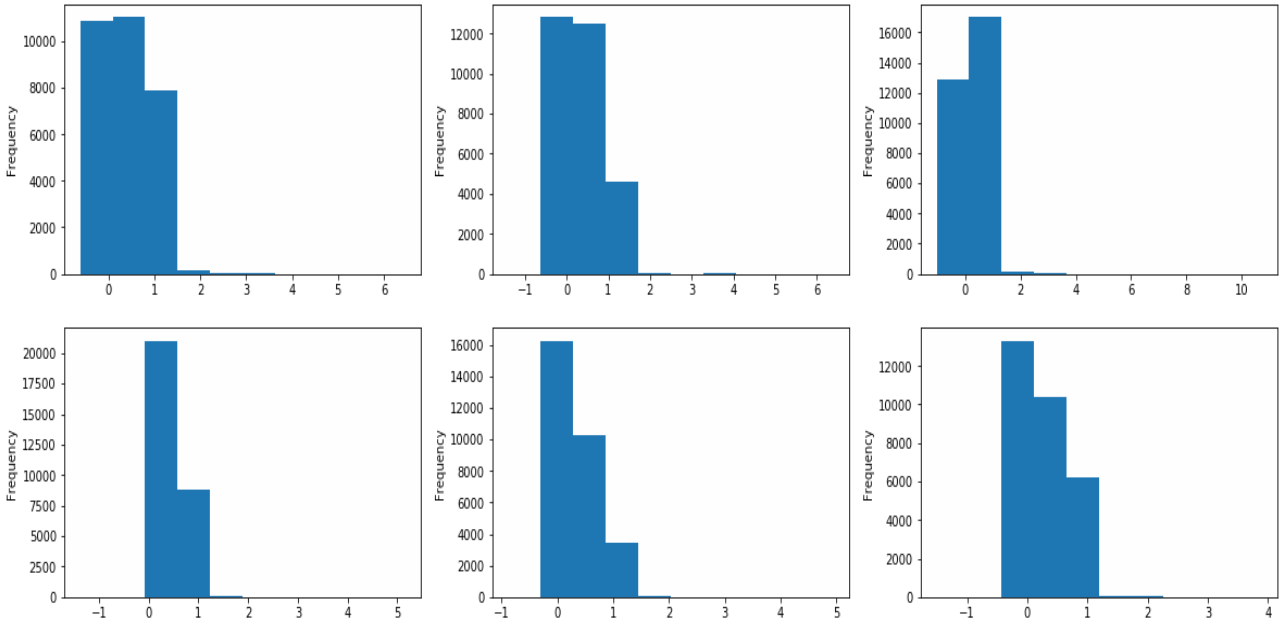
```
0    23182
1     3688
2     2667
3      322
4       76
5       26
8       19
6       11
7        9
Name: PAY_0, dtype: int64
0    25562
2     3927
3      326
4       99
1       28
5       25
7       20
6       12
8        1
Name: PAY_2, dtype: int64
0    25787
2     3819
3      240
4       76
7       27
6       23
5       21
1        4
8        3
Name: PAY_3, dtype: int64
```

```
0    26490
2     3159
3      180
4       69
7       58
5       35
6        5
8        2
1        2
Name: PAY_4, dtype: int64
0    27032
2     2626
3      178
4       84
7       58
5       17
6        4
8        1
Name: PAY_5, dtype: int64
0    26921
2     2766
3      184
4       49
7       46
6       19
5       13
8        2
Name: PAY_6, dtype: int64
```

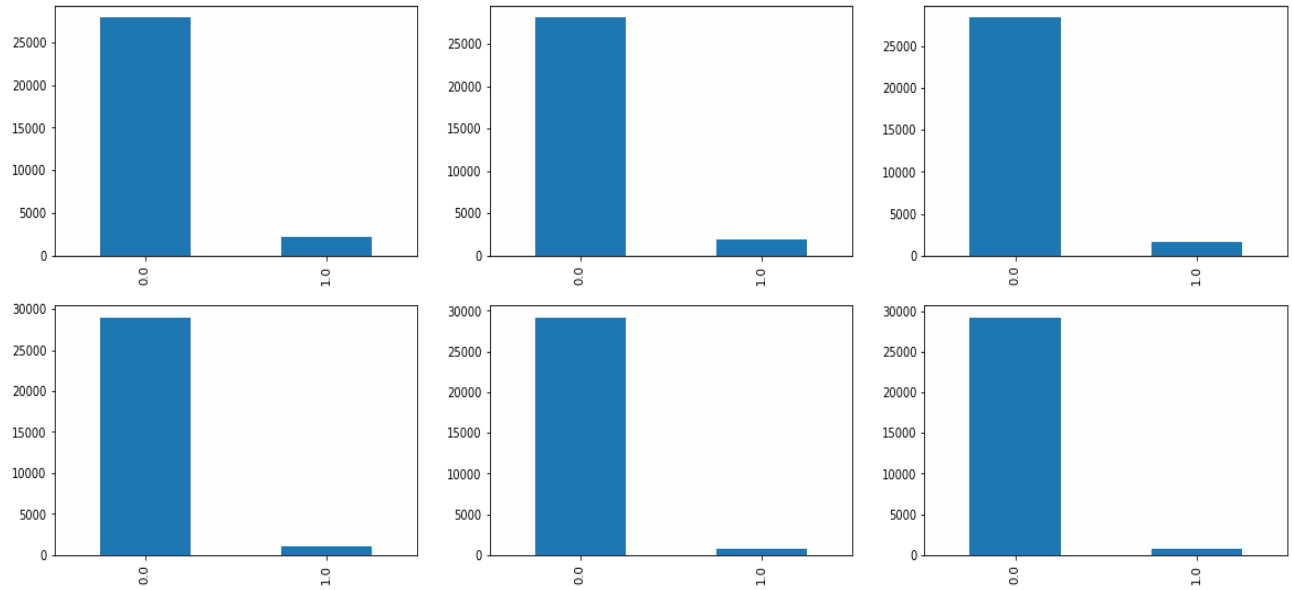
In PAY column since we know that the Sub categories –2, -1 are the number of clients who have pay duly which tells us they are non–defaulters, we have combined them with Subcategory 0 which shows the clients pay bill on time so all the non-defaulters are taken in to a single Sub category 0.



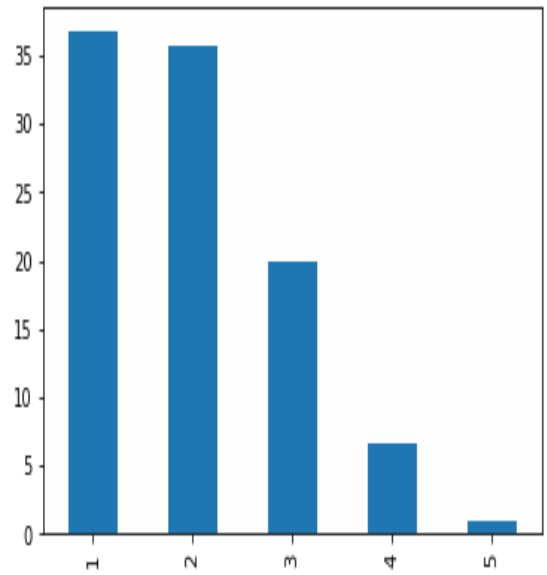
From the violin part we can see range of the Due_Amt for all the months and how the population is spread. The due amount is calculated by finding the difference between bill amount and pay amount.



A new column of Credit Limit is created by dividing the Bill Amount column with Limit Balance column to see if the clients have used beyond their credit limit or not, where values 0 to 1 indicates if the clients are using within their limit balance or not.



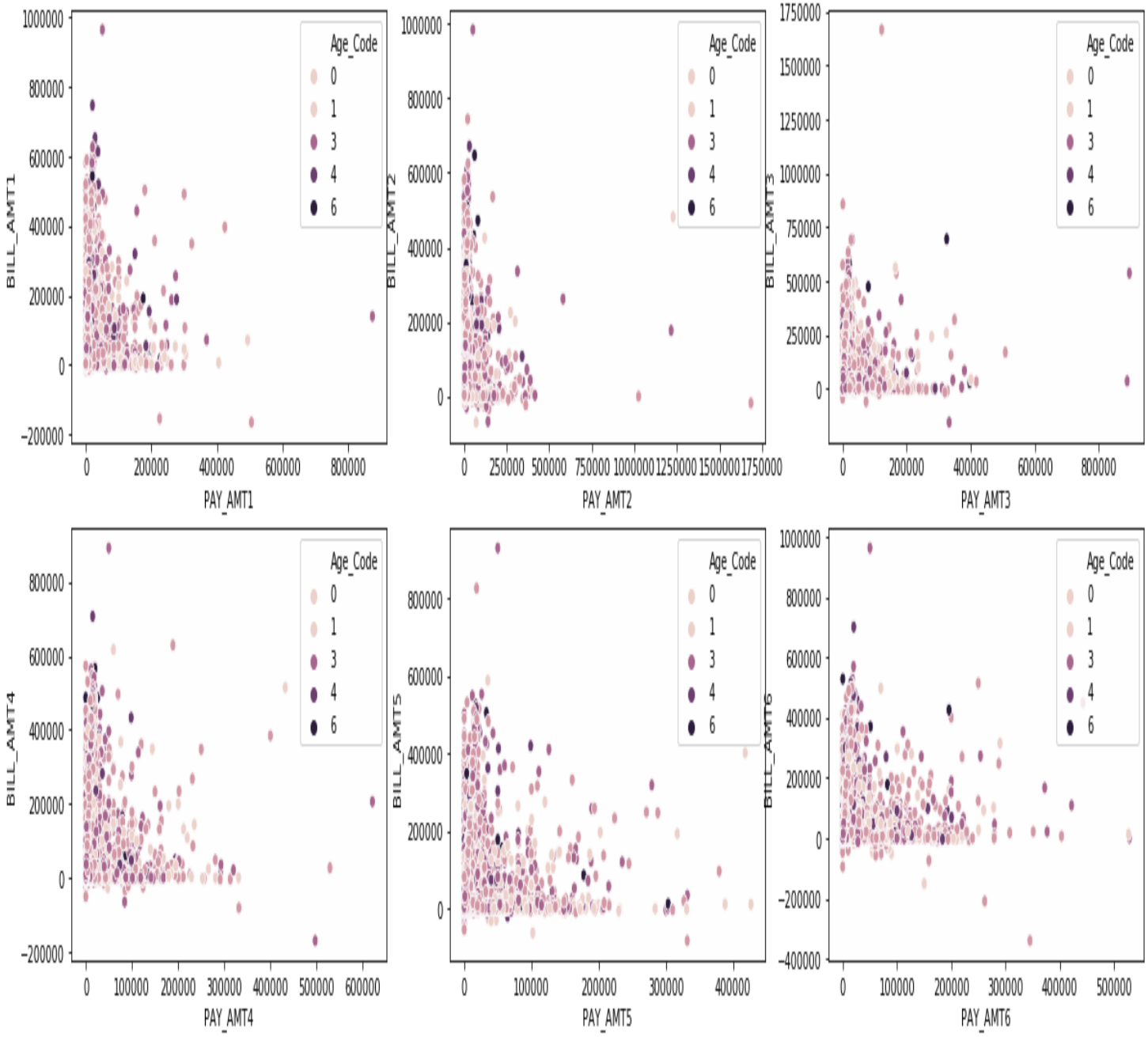
The above graph indicates a new feature engineered column credit_limit_edv, where 1 represents the number of clients using their credit card more than their limit balance and 0 representing the clients using within their limit balance. In the month of July, August and September.



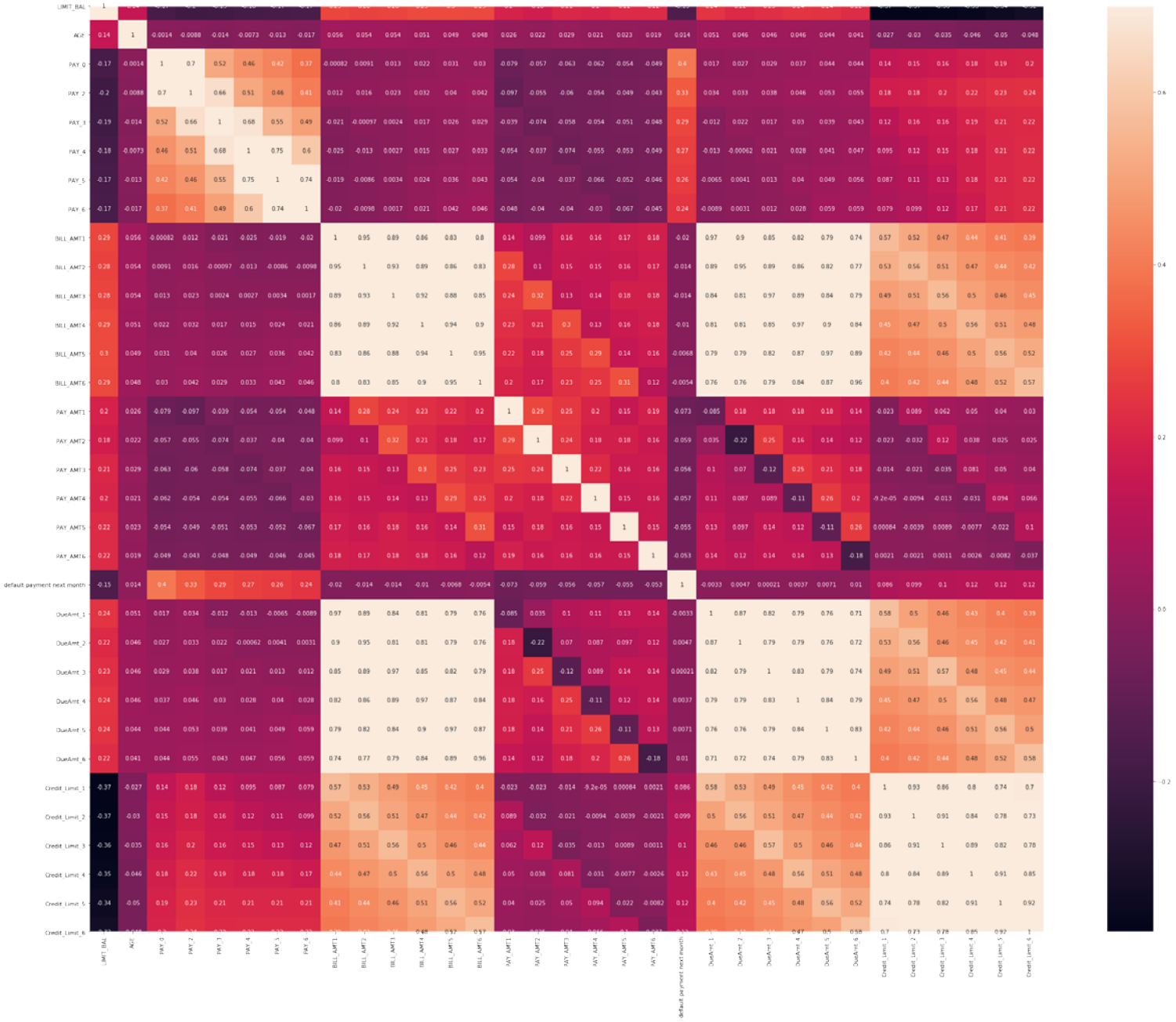

```
1    36.710000
2    35.710000
3    20.016667
4     6.656667
5     0.906667
Name: Age_Code, dtype: float64

1    11013
2    10713
3     6005
4     1997
5      272
Name: Age_Code, dtype: int64
```

```
def age_code(age):
    if age<31:
        return 1
    elif age<41:
        return 2
    elif age<51:
        return 3
    elif age<61:
        return 4
    else:
        return 5
```

The Age column of the clients is feature engineered by splitting them into different categories depending on their ages and we can see that 37% (11013) of the clients have their age less than31 followed 36% (10713) of the clients having age less than 41, So we can say that major number of clients are of age limit between 21-40, as the client with age 21 is the youngest.



From the scatter plot we can see that the population with age code <=3 are higher and the PAY_AMT for this range of people is around 200k.



The above image is a heat map to find out the extent of correlation with each feature present in the dataset and also with the additional feature that have been added recently.

After The addition of few new features we again carried out statistical test to find if there is dependence between numerical and target feature or not.

STATISTICAL TEST

HYPOTHESIS TESTING - 3:

To check the dependence between categorical and target feature.

Hypothesis Statements:

- **Step 1:** State the Null Hypothesis and Alternate Hypothesis

  H_0: There is no dependence between categorical and target feature.
  H_α: There is a dependency between categorical and target feature.

- **Step 2:** State the level of significance

  Here we select significance level (α) = 0.05

- **Step 3:** Choose the appropriate test

  Here we have selected the chi2_contingency test to find if the categorical features are significant or not.

- **Step 4:** Calculate the p-value

  Feature SEX is significant and the pvalue = 9.876093788177908e-10
  Feature EDUCATION is significant and the pvalue = 2.1874714431994113e-29
  Feature MARRIAGE is significant and the pvalue = 8.879828628366667e-05
  Feature Cred_Lim1_adv is significant and the pvalue = 3.156500623509219e-17
  Feature Cred_Lim2_adv is significant and the pvalue = 3.2006771492276304e-22
  Feature Cred_Lim3_adv is significant and the pvalue = 6.760945492949865e-15
  Feature Cred_Lim4_adv is significant and the pvalue = 8.611239128831722e-18
  Feature Cred_Lim5_adv is significant and the pvalue = 9.487540176335394e-19
  Feature Cred_Lim6_adv is significant and the pvalue = 7.914678867117956e-18
  Feature Age_Code is significant and the pvalue = 3.866600682620662e-05

- **Step 5:** Decision on whether to accept or reject the NULL hypothesis

  From the Hypothesis test we can see that all numerical columns mentioned above have p-value <0.05, and hence we can reject H_0 and say that the categorical features are significant

HYPOTHESIS TESTING - 4:

To check the dependence between the means of numerical and target feature

Hypothesis Statements:

- **Step 1:** State the Null Hypothesis and Alternate Hypothesis

  H0: Mean of the non-defaulters in the numerical columns = Mean of the defaulters in the numerical columns
  H_a: Mean of the non-defaulters in the numerical columns! = Mean of the defaulters in the numerical columns.

- **Step 2:** State the level of significance

  Here we select significance level (α) = 0.05

- **Step 3:** Choose the appropriate test

  Here we have selected the statistical 2 sample T test to find if the numerical features are significant or not.

- **Step 4:** Calculate the p-value

  Feature LIMIT_BAL is significant and the pvalue = 1.3022439532597397e-157
  Feature BILL_AMT1 is significant and the pvalue = 0.0006673295491221741
  Feature BILL_AMT2 is significant and the pvalue = 0.013957362392434761
  Feature AGE is significant and the pvalue = 0.01613684589016383
  Feature BILL_AMT3 is significant and the pvalue = 0.014769982710723002
  Feature BILL_AMT4 is not significant and the pvalue = 0.07855564157651403
  Feature BILL_AMT5 is not significant and the pvalue = 0.24163444291382874
  Feature BILL_AMT6 is not significant and the pvalue = 0.3521225212306479
  Feature PAY_AMT1 is significant and the pvalue = 1.1464876142241624e-36
  Feature PAY_AMT2 is significant and the pvalue = 3.1666567628387115e-24
  Feature PAY_AMT3 is significant and the pvalue = 1.841770291503132e-22
  Feature PAY_AMT4 is significant and the pvalue = 6.830941601370003e-23
  Feature PAY_AMT5 is significant and the pvalue = 1.2413447727776169e-21
  Feature PAY_AMT6 is significant and the pvalue = 3.033589072770243e-20
  Feature DueAmt_1 is not significant and the pvalue = 0.5722968045722805
  Feature DueAmt_2 is not significant and the pvalue = 0.4177344706040942
  Feature DueAmt_3 is not significant and the pvalue = 0.9715497385109686
  Feature DueAmt_4 is not significant and the pvalue = 0.5228081630917012
  Feature DueAmt_5 is not significant and the pvalue = 0.21746754266546503
  Feature DueAmt_6 is not significant and the pvalue = 0.07167810064878968
  Feature Credit_Limit_1 is significant and the pvalue = 1.5126308863936907e-50
  Feature Credit_Limit_2 is significant and the pvalue = 2.8601138535671145e-66
  Feature Credit_Limit_3 is significant and the pvalue = 8.730469672794275e-73
  Feature Credit_Limit_4 is significant and the pvalue = 2.9220006782812582e-90
  Feature Credit_Limit_5 is significant and the pvalue = 2.7348033032328747e-95

Feature Credit_Limit_6 is significant and the pvalue = 4.5661946655837264e-102

- **Step 5:** Decision on whether to accept or reject the NULL hypothesis

  From the Hypothesis test we can see that all numerical columns mentioned above have p-value <0.05, except BILL_AMT 4, BILL_AMT 5, BILL_AMT 6 and Due_Amt 1, Due_Amt 2, Due_Amt 3, Due_Amt 4, Due_Amt 5, Due_Amt 6 and hence we can reject H_0 and say that the numerical features are significant.

DATA PRE-PROCESSING

DUMMIES FOR CATEGORICAL VARIABLES:

- For all categorical columns like SEX, EDUCATION, MARRIAGE , Age_Code, Cred_Lim1_adv, Cred_Lim2_adv, Cred_Lim3_adv, Cred_Lim4_adv, Cred_Lim5_adv, Cred_Lim6_adv dummies were created by using get dummies and the original columns were dropped.

- The numerical columns like PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6 was conceited into "default payment next month".

- Features like DueAmt_1, DueAmt_2, DueAmt_3, DueAmt_4, DueAmt_5, DueAmt_6, BILL_AMT4, BILL_AMT5, BILL_AMT6 was dropped.

- Features like PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, Due_Amt 1, Due_Amt 2, Due_Amt 3, Due_Amt 4, Due_Amt 5, Due_Amt 6 and BILL_AMT4, BILL_AMT5, BILL_AMT6 was dealt in this manner because we could find from the last statistical test that they were not significant with the target column.

- Then we split the data set into X and y where y contains the target column and x contains rest of the columns other than the target column.

VIF Test:

	const	LIMIT_BAL	BILL_AMT1	BILL_AMT2	AGE	BILL_AMT3	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
vif	2368.088756	2.381361	26.251851	42.451954	12.424264	20.977374	1.700104	2.129841	1.359181	1.284004	1.257242	1.128894

Credit_Limit_1	Credit_Limit_2	Credit_Limit_3	Credit_Limit_4	Credit_Limit_5	Credit_Limit_6	SEX_2	EDUCATION_1	EDUCATION_2	EDUCATION_3
18.384709	24.904771	14.955065	11.143809	12.613884	7.863112	1.029744	490.375719	534.751693	294.922266

EDUCATION_4	MARRIAGE_1	MARRIAGE_2	Age_Code_2	Age_Code_3	Age_Code_4	Age_Code_5	Cred_Lim1_adv_1.0	Cred_Lim2_adv_1.0	Cred_Lim3_adv_1.0
32.968472	20.489287	20.933454	3.719262	8.78278	7.796188	2.839655	1.879211	2.066299	1.923384

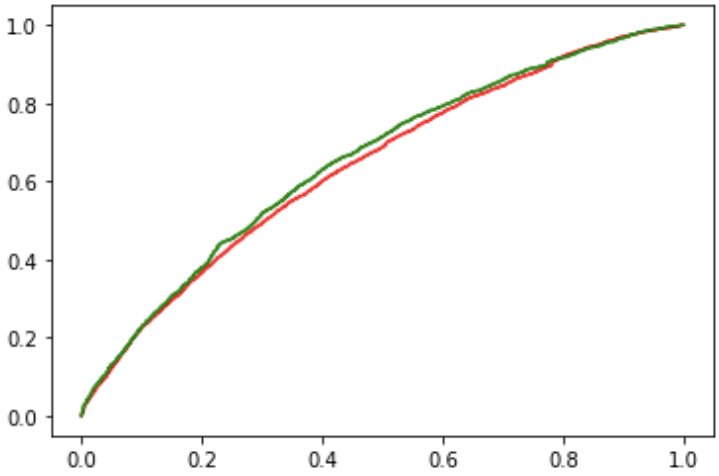
Cred_Lim4_adv_1.0	Cred_Lim5_adv_1.0	Cred_Lim6_adv_1.0	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
1.827951	1.783378	1.530714	2.036267	2.667391	2.584411	3.017551	3.283132	2.335765

From the above Vif scores we can infer that columns like AGE, MARRIAGE_1, BILL_AMT2, EDUCATION_2, Credit_Limit_2, Credit_Limit_5, Credit_Limit_3, Credit_Limit_4, BILL_AMT1 have very high Vif scores and can be dropped before modeling. High Vif value indicates that they don't not affect the target column much and that they are independent.

MODELING:

LOGISTIC REGRESSION: (AFTER FEATURE ENGINEERING)

```
Train score 0.7792380952380953
Test score 0.7777777777777778
ROC and AUC Score: 0.640103467933672
              precision    recall  f1-score   support

           0       0.78      1.00      0.88     16364
           1       0.50      0.00      0.00      4636

    accuracy                           0.78     21000
   macro avg       0.64      0.50      0.44     21000
weighted avg       0.72      0.78      0.68     21000
```



The Logistic Regression with feature engineering has given the score of Train and Test as 0.779 & 0.777. The typical ROC AUC curve is showing the performance of the classification model with ROC and AUC score is 0.64.

XG BOOST:

```
Accuracy score of train:  0.887095238095238
Confusion Matrix of train:  [[16045   319]
 [ 2052  2584]]
Auc of train:  0.9400489595102277
Train:                precision    recall  f1-score   support

           0       0.89      0.98      0.93     16364
           1       0.89      0.56      0.69      4636

    accuracy                           0.89     21000
   macro avg       0.89      0.77      0.81     21000
weighted avg       0.89      0.89      0.88     21000

Accuracy score of test:  0.8087777777777778
Confusion Matrix of test:  [[6588  412]
 [1309  691]]
Auc of test:  0.7621997142857143
Test:                precision    recall  f1-score   support

           0       0.83      0.94      0.88      7000
           1       0.63      0.35      0.45      2000

    accuracy                           0.81      9000
   macro avg       0.73      0.64      0.66      9000
weighted avg       0.79      0.81      0.79      9000
```

The XG Boost model has given the accuracy of train and test as 0.88 & 0.80 respectively.

RANDOM FOREST CLASSIFIER:

```
Accuracy score of train:  0.8267142857142857
Confusion Matrix of train:  [[15630   734]
 [ 2905  1731]]
Auc of train:  0.819806780873185
Accuracy score of test:  0.8195555555555556
Confusion Matrix of test:  [[6681  319]
 [1305  695]]
Auc of test:  0.7818299642857143
Test:                precision    recall  f1-score   support

           0       0.84      0.95      0.89      7000
           1       0.69      0.35      0.46      2000

    accuracy                           0.82      9000
   macro avg       0.76      0.65      0.68      9000
weighted avg       0.80      0.82      0.80      9000
```

The Random Forest Classifier model was Hypertuned with Randomsearch using the following parameters n_estimators:sp_randint(5,150),max_features:sp_randint(1,33),max_depth:sp_randint(2,10),min_samples_leaf:sp_randint(1,50),min_samples_split:sp_randint(2,50),criterion:['gini','entropy'] and has given the accuracy of train and test as 0.82 & 0.81 respectively.

ADABOOST CLASSIFIER:

```
Accuracy score of train:  0.8194285714285714
Confusion Matrix of train:  [[15681   683]
 [ 3109  1527]]
Auc of train:  0.7876119589730525
Accuracy score of test:  0.8148888888888889
Confusion Matrix of test:  [[6714  286]
 [1380  620]]
Auc of test:  0.768648142857143
Test:                precision    recall  f1-score   support

           0       0.83      0.96      0.89      7000
           1       0.68      0.31      0.43      2000

    accuracy                           0.81      9000
   macro avg       0.76      0.63      0.66      9000
weighted avg       0.80      0.81      0.79      9000
```

The Ada Boost Classifier model has given the accuracy of train and test as 0.81 & 0.81

LGBM:

```
Accuracy score of train:  0.8459047619047619
Confusion Matrix of train:  [[15766   598]
 [ 2638  1998]]
Auc of train:  0.8840355172626879
Accuracy score of test:  0.8174444444444444
Confusion Matrix of test:  [[6647  353]
 [1290  710]]
Auc of test:  0.7783363214285713
Test:                precision    recall  f1-score   support

           0       0.84      0.95      0.89      7000
           1       0.67      0.35      0.46      2000

    accuracy                           0.82      9000
   macro avg       0.75      0.65      0.68      9000
weighted avg       0.80      0.82      0.80      9000
```

The Light GBM model has given the accuracy of train and test values of the model as 0.84 & 0.81 respectively.

LGBM with RANDOM SEARCH CV:

```
Accuracy score of train:  0.8270952380952381
Confusion Matrix of train:  [[15616   748]
 [ 2883  1753]]
Auc of train:  0.820016255774318
Accuracy score of test:  0.8203333333333334
Confusion Matrix of test:  [[6683  317]
 [1300  700]]
Auc of test:  0.7819450357142856
Test:                precision    recall  f1-score   support

           0       0.84      0.95      0.89      7000
           1       0.69      0.35      0.46      2000

    accuracy                           0.82      9000
   macro avg       0.76      0.65      0.68      9000
weighted avg       0.80      0.82      0.80      9000
```

The Light GBM model has been Hypertuned using Randomsearch with the following parameters n_estimators:sp_randint(50,200),num_leaves:sp_randint(10,50),max_depth:sp_randint(2,15),learning_rate:sp_uniform(0,1),min_child_samples:(2,50) given the accuracy of train and test values of the model as 0.82 & 0.82 respectively.

VOTING CLASSIFIER:

```
Accuracy score of train:  0.8873809523809524
Confusion Matrix of train:  [[16252   112]
 [ 2253  2383]]
Auc of train:  0.973148907015948
Accuracy score of test:  0.8127777777777778
Confusion Matrix of test:  [[6686  314]
 [1371  629]]
Auc of test:  0.7721243571428571
Test:                precision    recall  f1-score   support

           0       0.83      0.96      0.89      7000
           1       0.67      0.31      0.43      2000

    accuracy                           0.81      9000
   macro avg       0.75      0.63      0.66      9000
weighted avg       0.79      0.81      0.79      9000
```

The voting classifier model where all the models were given weightage and checked gives the accuracy of train and test as 0.88 and 0.81 respectively.

LIMITATIONS:

Customer information such as Income, No. of dependencies, Location, could be possible indicators of credit card default behavior. Additionally, these features could be used for a better customer segmentation which could provide additional insights on the spending and repayment behavior.

Further, information such as the customers investment portfolio, and his/her other tangible & non-tangible assets could also be a positive contributor to the risk associated with each customer.

CONCLUSION:

This study aimed at applying the machine learning techniques to predict credit card default in banks. Based on the analysis of the results, LGBM with hyper tuning has a prediction accuracy of more than 82%. Banks can use machine learning to assess credit risk of customers before granting them credit card. Banks major concern in to offer valuable products and services to their clients and in order keep up with their competitors they must stay innovative and creative.

Machine learning techniques allow banks to approach their customer base in a more customized manner. By applying analytics in the business, banks can benefit in several ways. By studying the customer in terms of their risk level and applying the results from the model, it allows the bank to ingrain smart decision making into a business. It also provides greater insight into data visualization. Banks can make the most of the machine learning algorithm which can contribute in boosting their performance and image in the industry.