Default of Credit Card Clients

Great Lakes Institute of Management

ABSTRACT

This project presents and discusses data-driven predictive models for predicting the defaulters among the credit card users. Data used include details like limit balance, age, sex, amount of bill statement, repayment status and amount of previous payment. The paper discusses which variables are the strongest predictors of default, and to make predictions on which customers are likely to default. Four machine learning models were trained with repeated cross validation and evaluated in a testing set: (a) logistic regression,(b)KNN, (c) decision tree, (d) Naïve Bayes ,(e)Random Forest and Ensembling methods like Boosting and Bagging. The base model gave 81% R-square value with all the variables without any treatment.

DECLARATION

We hereby declare, that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Table of Contents

EXECUTIVE SUMMARY

The objective of the project is to predict which variables are the strongest predictors of default, and to make predictions on which customers are likely to default. Data used details of credit card customers like limit balance, age, sex, education payment status, bill amount, etc. First, base line models were formed using Logistic Regression, KNN, Naïve Bayes, Decision Tree and Random Forest, without performing any transformation on the data which gave us the best accuracy of around 87% by Logistic Regression Model. Later standardization of data was done and outlier treatment was performed along with rectification of data imbalance and solving it using SMOTE feature selection, hyper parameter tuning for all the models and them Recursive Feature Elimination for Decision Tree and Random Forest . Then found that the best model with training accuracy 88.3741% and testing accuracy 85.64% is Random Forest Model after hyper parameter tuning.

1. Introduction

1.1 Overview of the Dataset:

The credit card issuer has gathered information on 30000 customers. The dataset contains information on 25 variables, credit data, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not?

Name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4,5,6=others .
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-2, -1,0=pay duly , 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)

PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)

default.payment.next.month	Default payment (1=yes, 0=no)

The total proportion of defaults in the data is 15.66% which is 4700 out of the total data set comprising of 30,000 samples. This could be due to a large bias and therefore not a realistic representation of the bank's customer base. However, the data was collected during a debt crisis which provides an argument for the assumption that the data represents a non-biased sample of the customer base. In any case, the high amount of defaults in should be taken into consideration when making generalizations about the results or methodology of this case study. The high number of defaults will especially have an effect on estimates of the bank's financials. There are 15 numerical columns and 9 categorical columns in the dataset. There are no null values or missing value in the dataset. In the dataset the numbers of non-defaulters are 25300 and defaulters are 4700.

1.2 Problem Statement:

We have been a dataset contains data regarding details of 30,000 customer for 5 months which is taken from the bank. The task is to see which variables are the strongest predictors of default, and to make predictions on which customers are likely to default the next month.

1.3 Problem Solving Methodology Used:

The methodology used to design the project was CRISP-DM. CRISP-DM stands for crossindustry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. It is flexible and useful when using analytics to solve tricky business problems. It has 6 stages like Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment shown in figure1.

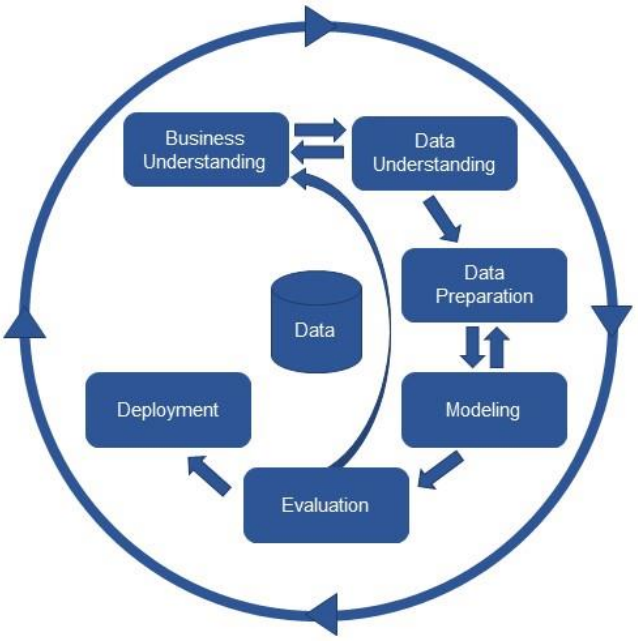


Figure 1

a. **Business Understanding:** The first stage of the CRISP-DM process is to understand what one wants to accomplish from a business perspective. The objective of this stage is to uncover significant factors that could influence the outcomes of the project. Disregarding this step can mean that a great deal of effort is put into producing the right answers to the wrong questions. In this project the domain is related to details related to credit card details of customers. The customer details like age, limit balance, sex, education, payment status, bill amount, repayment status of 5 month from April 2005 to September 2005 are collected from the bank so that prediction can be made whether the customer is likely to default in the next month or not. The objective of the project is to maximize the prediction accuracy of the energy consumption of Appliances. To achieve it, we would first understand and treat the data, then various techniques such as VIF and PCA will be used to reduce the columns or dimension of data. Lastly, various machine learning Regression models along with ensemble techniques will be used on the data. The model with the best accuracy will be the chosen model.

b. **Data Understanding:** The second stage of the CRISP-DM process requires one to acquire the data listed in the project resources. This initial collection includes data loading, if this

is necessary for data understanding. Data was collected for a span of 5 months by the bank. Defaulter is the target variable which is stating whether the credit card customer will pay the due amount next month or not. The predictor variables are

SEX ,EDUCATION ,MARRIAGE, AGE,PAY_0 ,PAY_2,PAY_3, PAY_4 ,PAY_5 ,PAY_6, default, ID,LIMIT_BAL, BILL_AMT1 ,BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5 and PAY_AMT6. When Exploratory Data Analysis was performed on the data. It was checked if there were any outliers, if data is normal and what are the patterns seen when predictor variables are plotted against the target variable. First, univariate analysis was performed to check how each variable is distributed, if the distribution is normal or not. Normal distribution is important because of the Central limit theorem. In simple terms, if you have many independent variables that may be generated by all kinds of distributions, the aggregate of those variables will tend toward a normal distribution. This universality across different domains makes the normal distribution one of the centre pieces of applied mathematics and statistics. From Univariate Analysis, we found that the most of the numerical variables is highly skewed towards the right, as a large number of values may lie in that interval and a large number of outliers were visualized in some variables. From the boxplots many outliers were visualized. For the categorical univariate analysis like education we can see that university going seem to take more credit card and they seem to default more , then comes graduate school going students taking credit card and they seem to default more then come high school going students and other , both these category seem to have taken less number of credit card and defaulters among them are also less. In categorical column, sex can see that females are taking more credit card but men seem to have more percentage defaulting than women. In marriage, more of married people have taken credit card and defaulters are also more among them , others the proportion is very less to compare male and female. In age we can see a steady growth of credit card users until 30 and then a steep fall of credit card users until 79 but defaulters among them are showing a constant count of defaulters from age of 23 to 40 then the defaulters number is falling. Then in limit balance we can see that people with high LIMIT_BAL seem to default less. In categorical column PAY_1 , PAY_2 , PAY_3 ,

PAY_4 , PAY_5, PAY_6 we can see most of the customers belong to 0,-1and -2 which are people who pay their due at the correct time and in other categories where people seem to default in 1st , 2nd up to 9th month are very less in number. As the PAY_AMT increases the number of defaulters decreases. Same as PAY_AMT, when the BILL_AMT also increases the number of Defaulters reduces. There are no null or missing values in the dataset. Also, there are no redundant columns in the dataset. So it looks like the PAY_X variables are the strongest predictors of default, followed by the LIMIT_BAL and PAY_AMT variables. To make predictions about whether a customer is likely to default - we'll train a number of different classifiers and see how well they perform. As usual, we start by splitting the data into train/test sets and rescaling.

c. **Data Preparation:** This is the stage of the project where one decides on the data that on e will use for analysis. The criteria used to make this decision includes the relevance of th e data to the data mining goals, the quality of the data, and technical constraints such as li mits on data volume or data types. For data preparation, number of outliers were checked and observed to have more number of outliers in some features and zscore was performed in the numerical columns for transforming the data. Then the data was scaled using stand ard scalar. VIF and RFE were also performed to lower the number of columns and remov e multicollinearity.

d. **Modelling:** As the first step in modelling, one selects the actual modelling technique that one will be using. Although one may have already selected a tool during the business und erstanding phase, at this stage one maybe be selecting some specific modelling technique. If multiple techniques can also be applied. After treating the data, we performed various models on our data such as KNN, Naïve Bayes, decision tree, Random forest and Logisti c Regression, along with ensemble techniques. Hyper parameter tuning was also perform ed on the tree models. The model with best accuracy achieved was with Decision Tree w hich was tuned to get the best parameters using Grid Search and then Adaptive Boosting t echnique.

e. **Evaluation:** During this step it will be assessed if the model meets your business objectiv es and to how much degree and seeking to determine if there is some business reason wh y

this model is deficient. Another option is to test the model(s) on test applications in the real application, if time and budget constraints permit. The evaluation phase also involves assessing any other data mining results one generated. Data mining results involve model s that are necessarily related to the original business objectives and all other findings that are not necessarily related to the original business objectives, but might also unveil additi onal challenges, information, or hints for future directions. To evaluate the model the data was separated into train and test. Then, the data was passed to various machine learning a lgorithms like, Decision Tree, Random Forest and Logistic Regression. Accuracy score a nd RMSE values were checked. K fold cross validation was also performed to evaluate th e data.

f. **Deployment:** In the deployment stage one may take their evaluation results and determin e a strategy for their deployment. It makes sense to consider the ways and means of deplo yment during the business understanding phase as well, because deployment is absolute c rucial to the success of the project. This is where predictive analytics really helps to impr ove the operational side of the business. For the project, deployment steps have not been performed. The scope of the procedure we followed was till model evaluation.

2. Visualization

2.1 Univariate Analysis:

LIMIT_BAL

LIMIT_BAL states the amount of given credit. This is the maximum amount a customer can spend with their credit card in a single month. The amount of balance limit is dependent on the bank's own screening processes and other unknown factors.

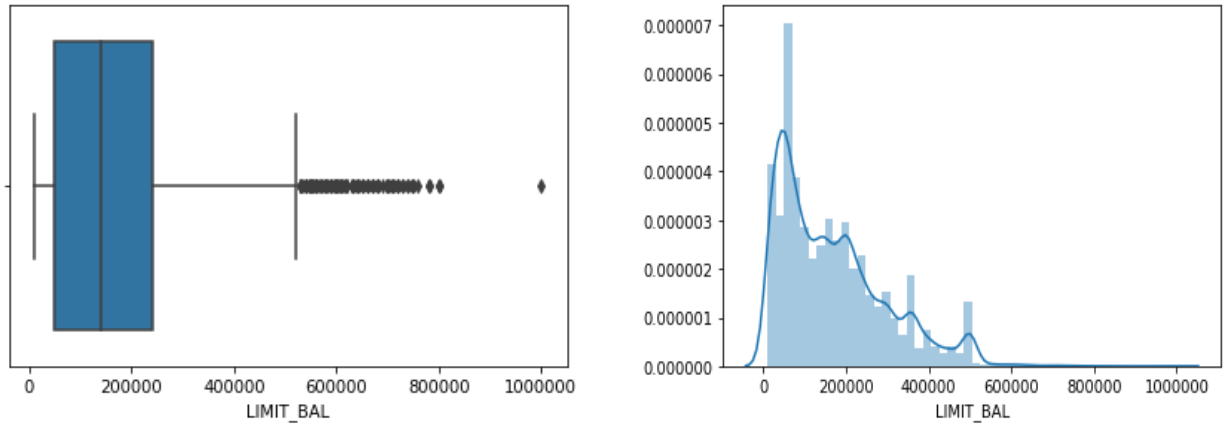


Figure 2

This is a numerical column. The mean and median values of are 1674840.322 and 140000 respectively. The LIMIT_BAL distribution of credit card dataset is between 10000 and 1000000. The distribution is right skewed. In box plot of limit balance feature we can observe many outliers which are been treated here as extreme values since it is a bank dataset which is highly sensitive.

AGE

This is the age of the customer which is stated in years.

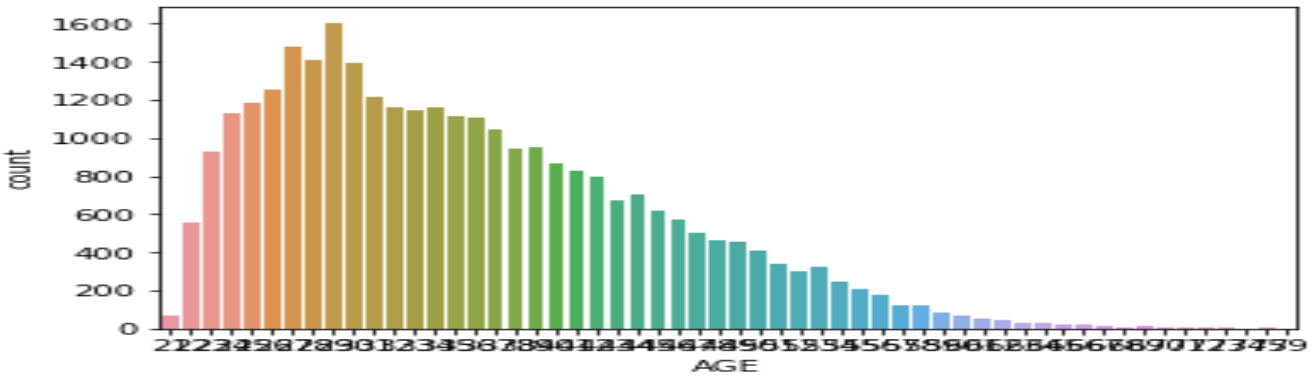


Figure 3

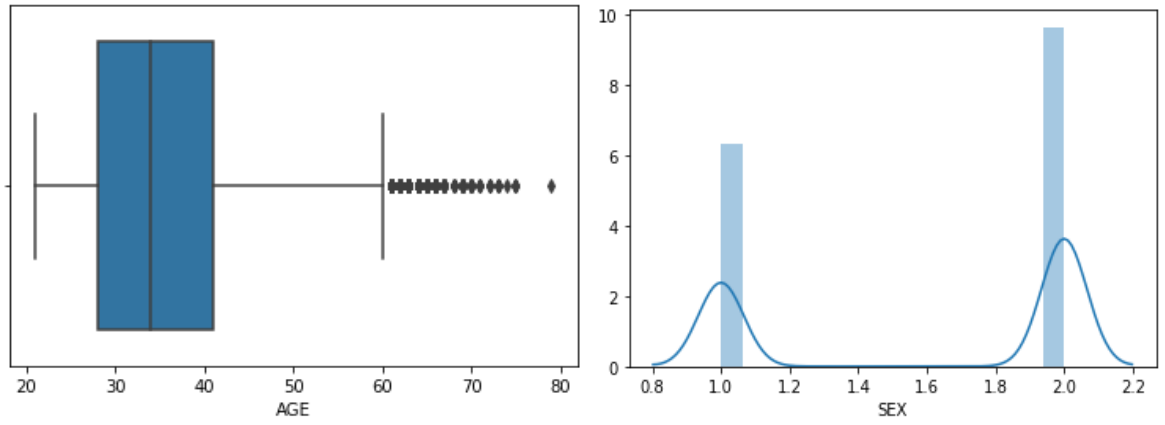


Figure 4

The mean and median of the age are 35 and 34 respectively. This is numerical column. There are no missing values in this column. There are some outliers in AGE column. The decline in number of customers starts from about 30 years among the non-defaulting group, while the number of customers of different ages stays much more constant from 25 to around 40 years. This indicates that likelihood of default among men grows with age. In box plot we can see many outliers which are considered here as extreme values.

SEX

This variable can obtain a value of 1 for male and 2 for female. In this study, sex and gender are used interchangeably to intend the same thing.

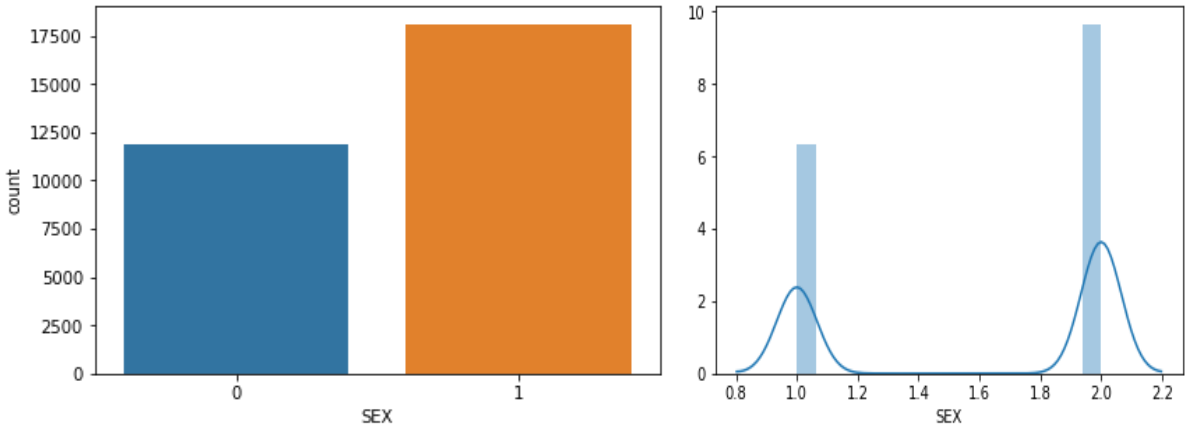


Figure 5

The SEX column in the dataset does not have any null values. This is a categorical column. Approximately 60.4% of the customers in the dataset are female, and the remaining 39.6% male. The proportion for male is more and female is less but the defaulters are more in men than women.

MARRIAGE

Referred to as "married" in the analysis, this variable can obtain three values: 1 = Married, 2 = Single, 3 = others such as divorced or widowed.

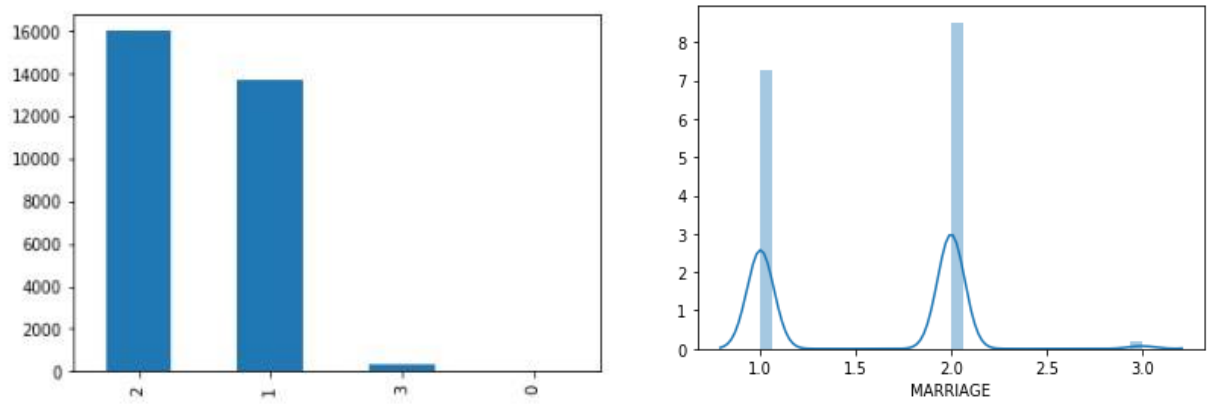


Figure 6

The MARRIAGE column has no null values. This is categorical column .Marital status is mostly divided into categories "Married" and "Single", with respective proportions of 53.2%, 45.5%, and the group "Other" containing only 1.3% of the customers.

EDUCATION

The education level of a customer is represented as one of four values: 1 = Graduate school, 2 = University, 3 = High school, 4 = Other. For the purpose of analysing customer groups, this is assumed to indicate the highest level of education completed.

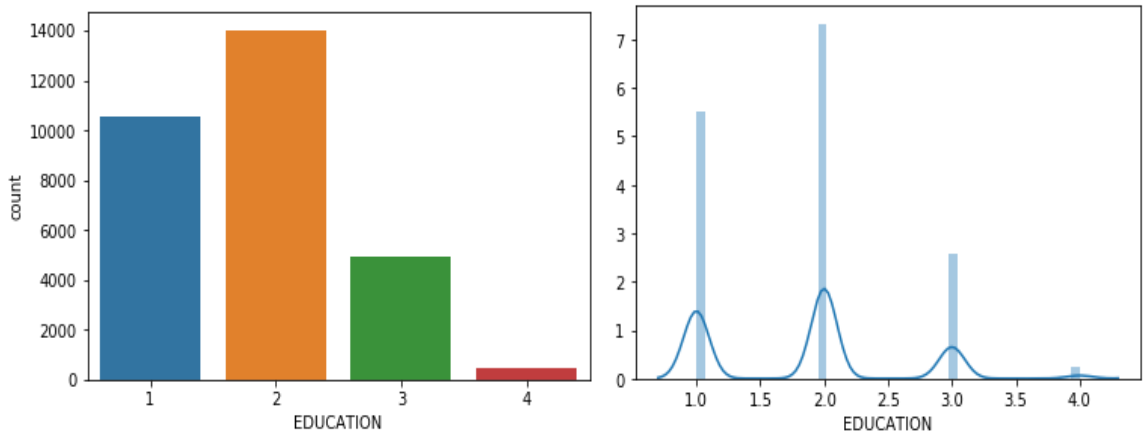
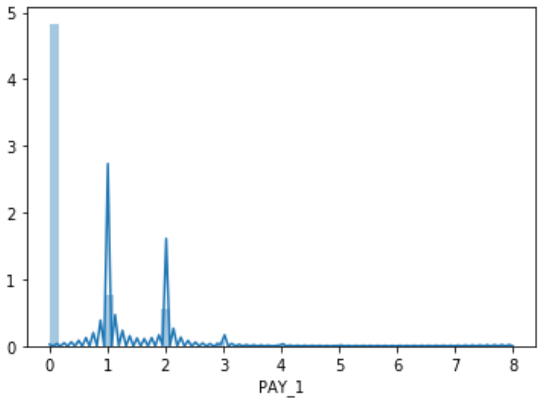


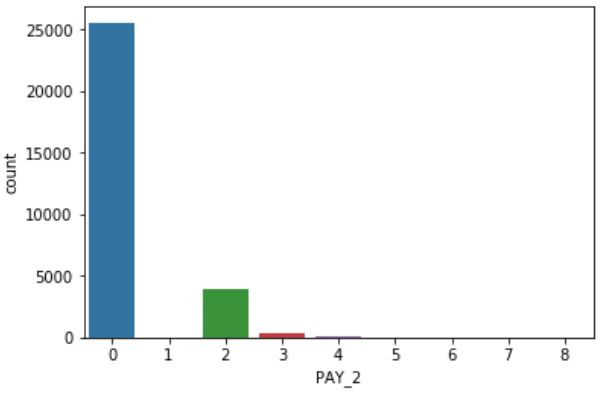
Figure 7

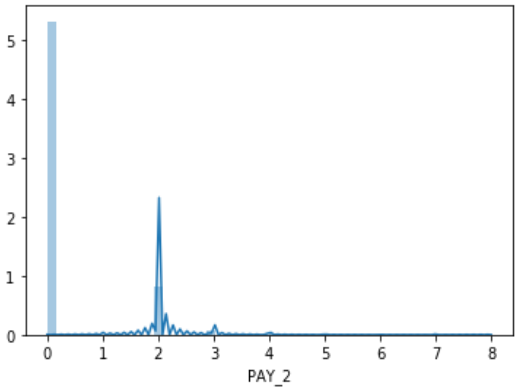
The Education column does not contain any null values. This is categorical column. We can see that University is the most common level of education in the dataset, followed by Graduate school and High school. Respectively, the proportions of customers in each category are: 47%, 35%, and 16%. The last category "Other" amounts to only 1.5% of the customers in the dataset.

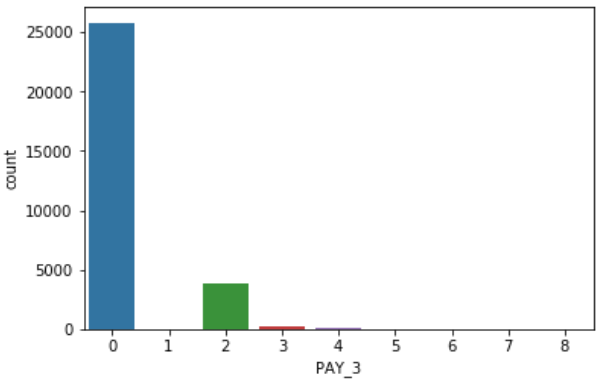
PAY

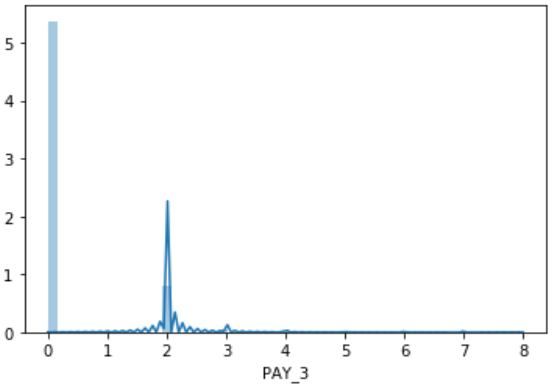
Payment status is represented as 6 different columns, one for each month. The value of payment status for a month indicates whether repayment of credit is was delayed or paid duly. A value of -1 indicates pay duly. 6 Values from 1 to 8 indicate payment delay in months, with a value of 9 defined as a delay of 9 months or more. Data collected from 6 months, April to September.

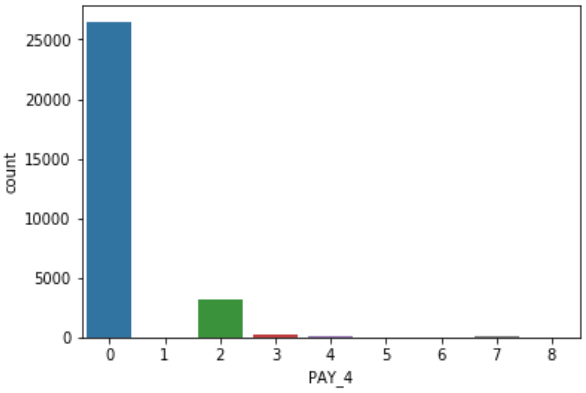


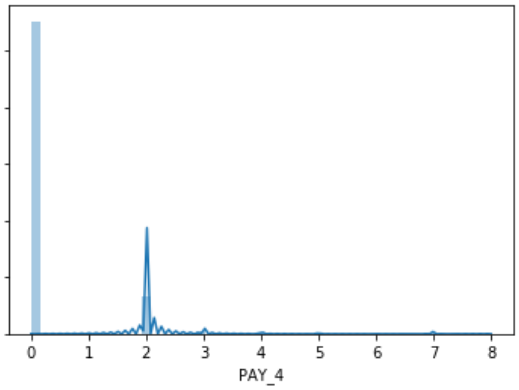

count
PAY_2


PAY_2


count
PAY_3


PAY_3


count
PAY_4


PAY_4

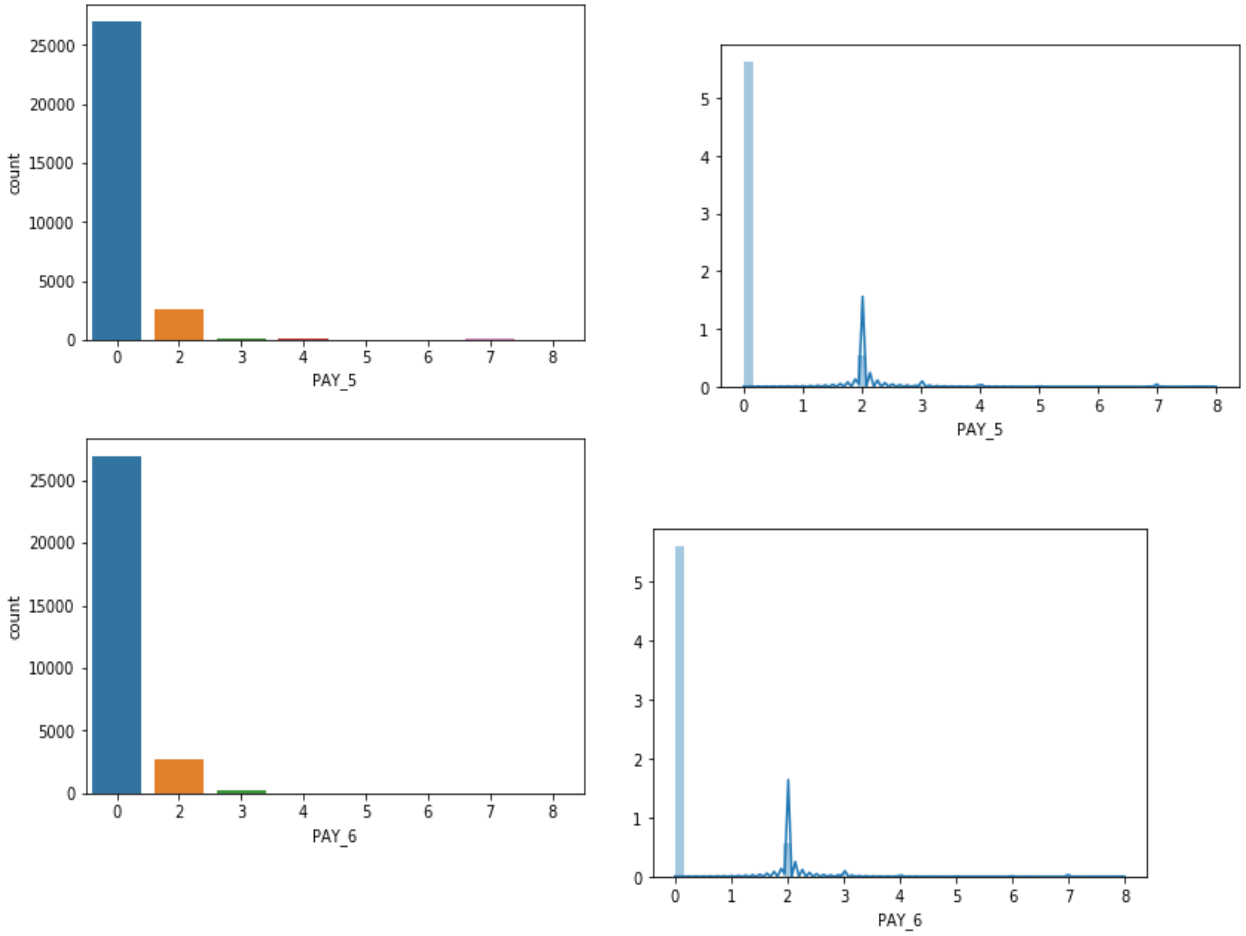


Figure 9

The PAY column does not have any null values. This is a categorical column. Majority of the values belongs to 0.Mostly people are paying the due on time which is denoted by 0 from the above graph.

Bill_AMT

Amount of bill statement is recorded in this variable. It is represented in the data as 6 columns, one for each month. Data collected from 6 months, April to September.

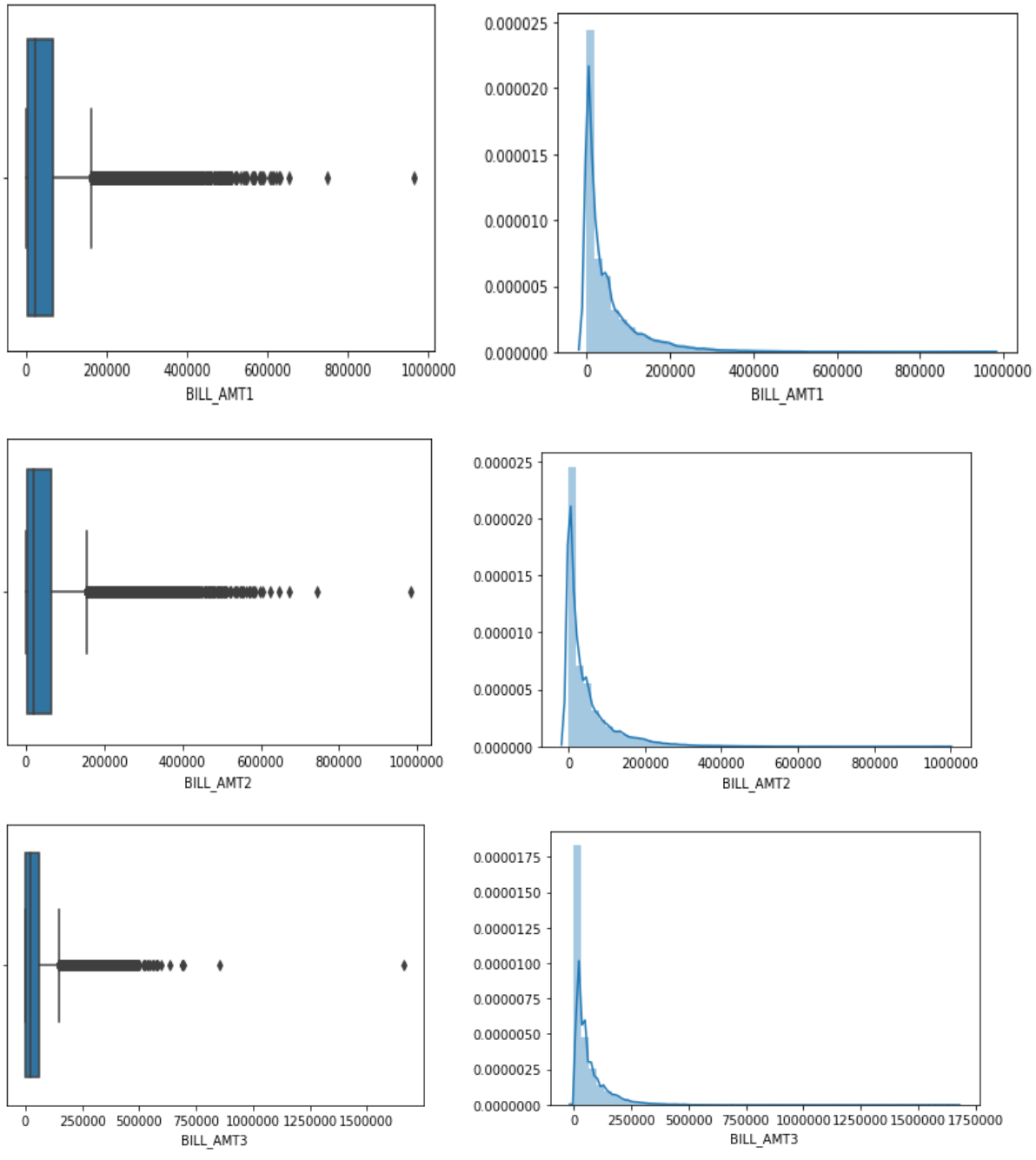


Figure 10

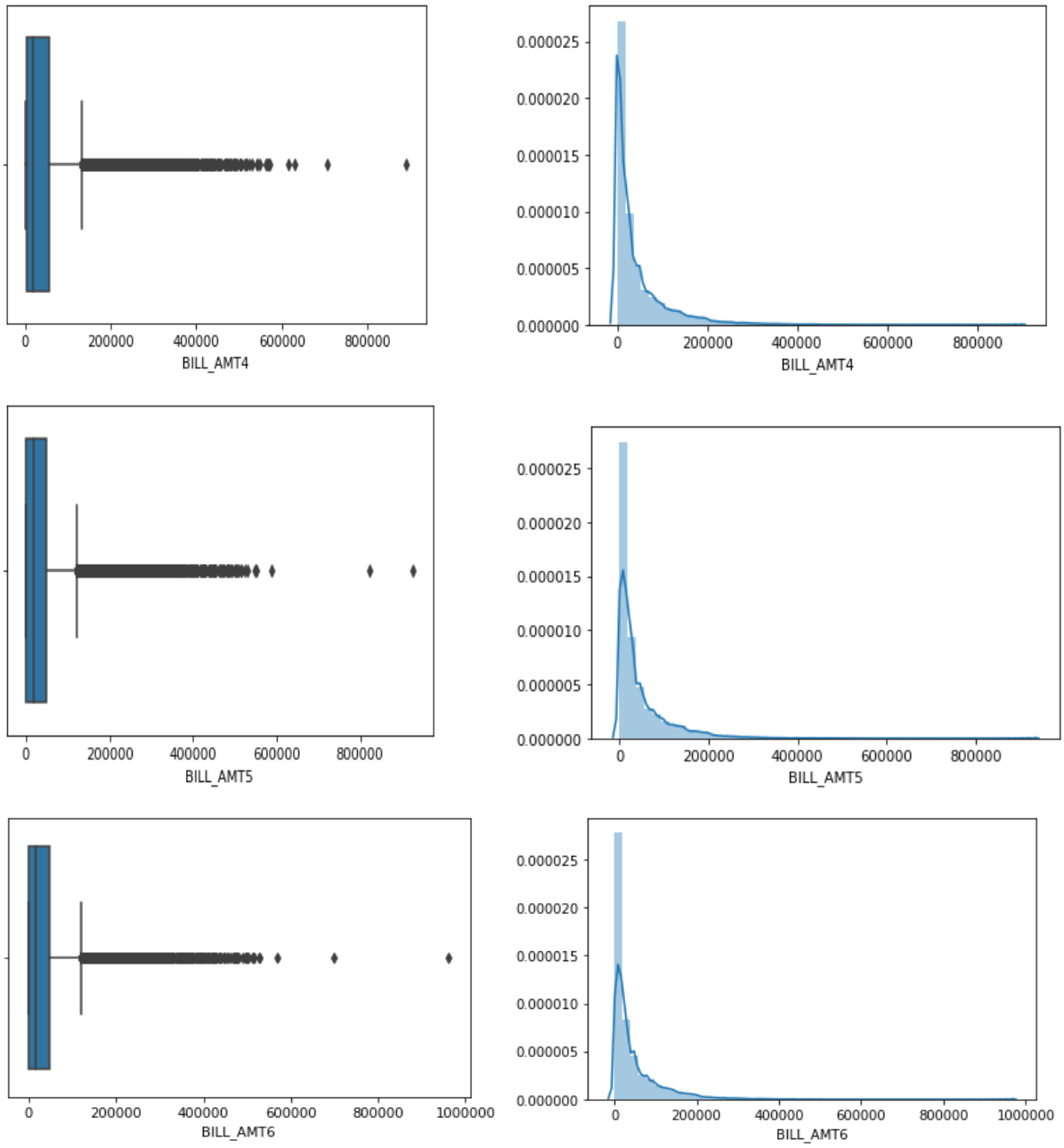


Figure 11

BILL_AMT column has no missing values. This is a numerical column. All the plots are right skewed.From the box plot we can observe the presence of some outliers which are treated as extreme values as this is a bank dataset.

PAY_AMT :

Amount of previous payment stored in 6 different columns for each month, similarly to payment status and bill amount. The payment amounts correspond to the same months as payment status and bill amount. For example, the payment amount for April indicates amount paid in April.

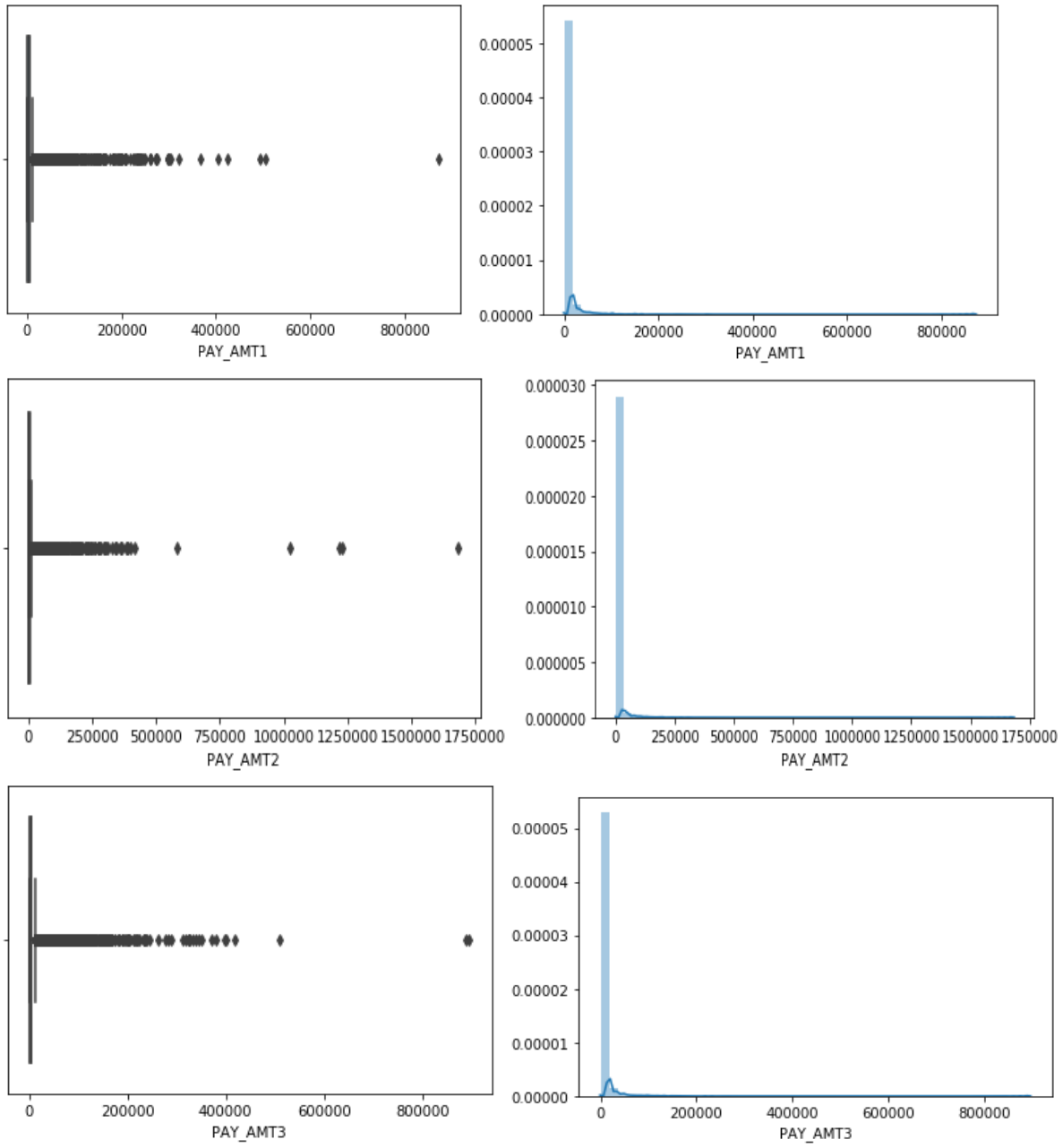


Figure 12

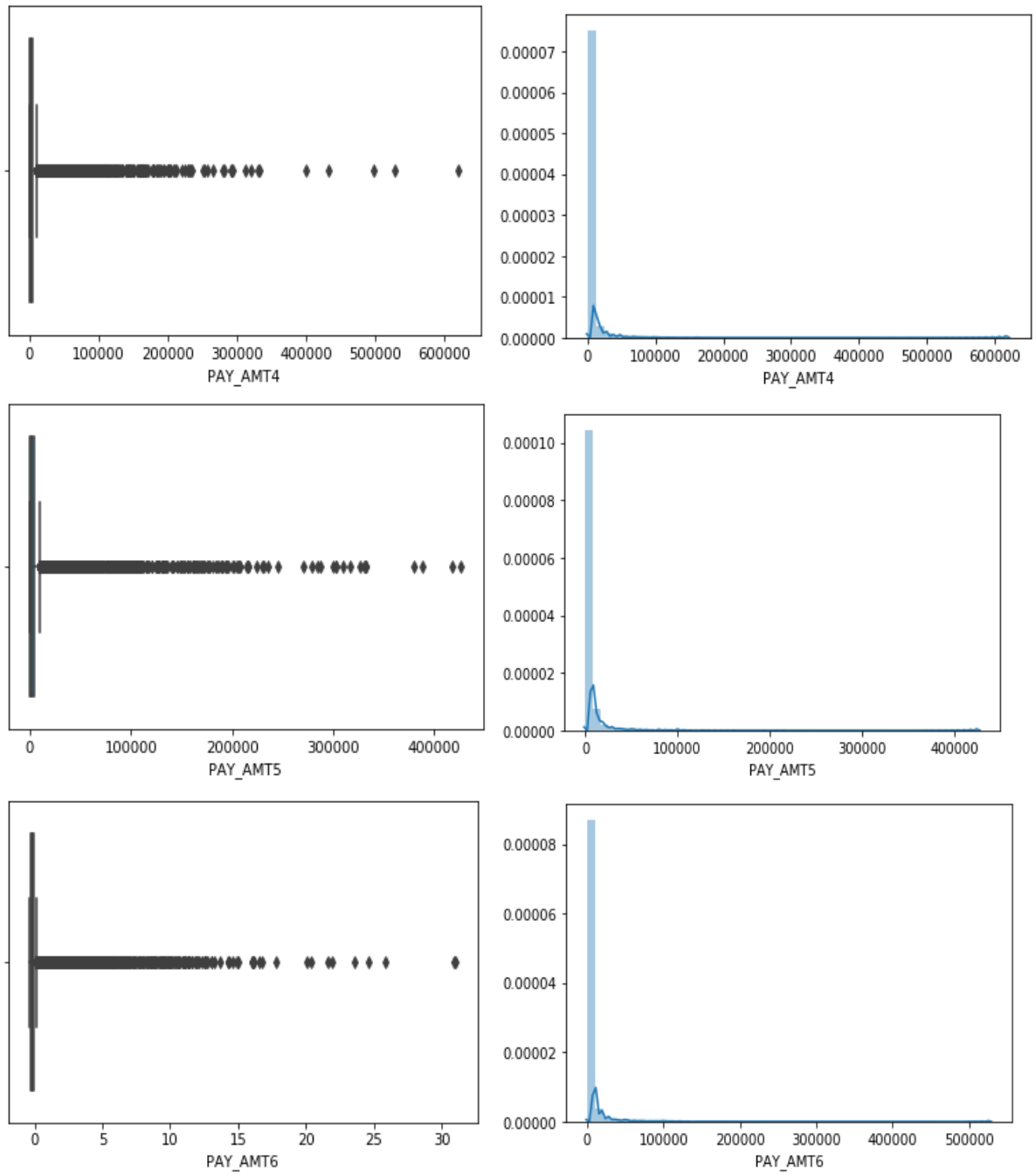

0
100000
200000
300000
400000
500000
600000
PAY_AMT4
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
100000
200000
300000
400000
500000
600000
PAY_AMT4
0
100000
200000
300000
400000
PAY_AMT5
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
100000
200000
300000
400000
PAY_AMT5
0
5
10
15
20
25
30
PAY_AMT6
0.00008
0.00006
0.00004
0.00002
0.00000
0
100000
200000
300000
400000
500000
PAY_AMT6

Figure 13

PAY_AMT column has no missing values. All the above plots are right skewed. This is a numerical column.

2.2 Bivariate Analysis

LIMIT_BAL

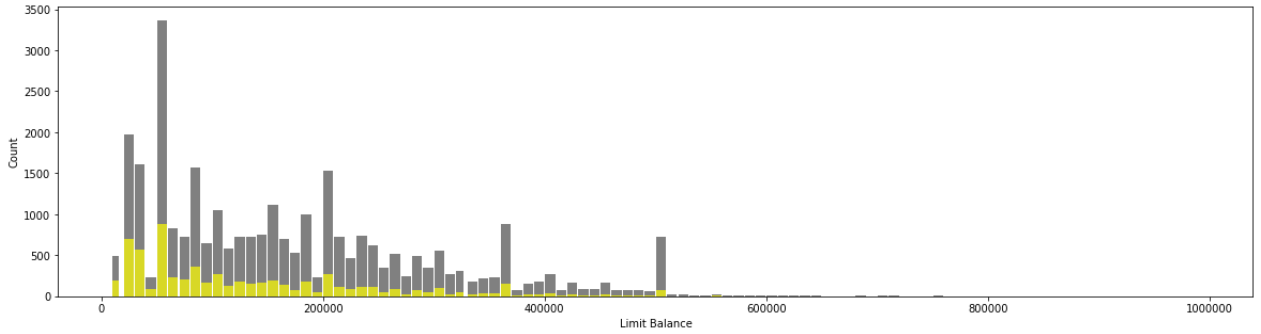


Figure 14

The above plot was used to visualize the relation between LIMIT_BAL and the target Variable. The Yellow colour shows the defaulters and rest are non-defaulters. From the above plot we can assume that the people with high LIMIT_BAL have less chances of defaulting

AGE

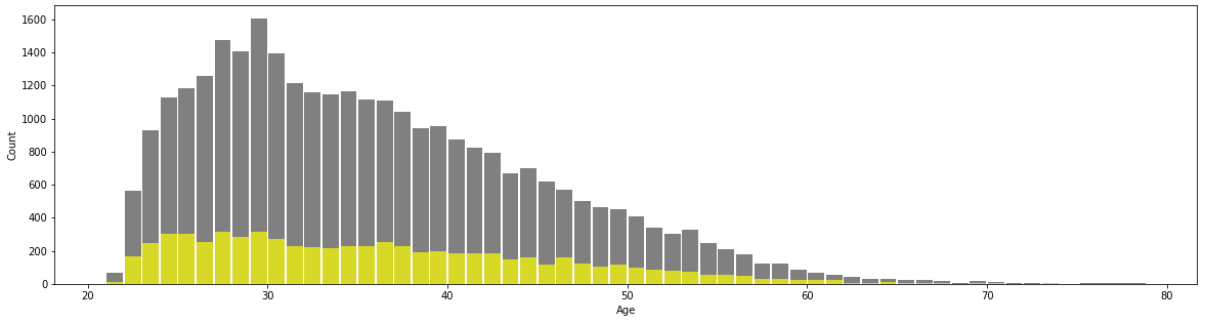


Figure 1 3

The above plot was used to visualize the relation between AGE and the target Variable. The Yellow colour shows the defaulters and rest are non-defaulters. From the above plot we can assume that most of the customers are having age nearer to 30 years and most of the defaulters are also for the same range.

SEX

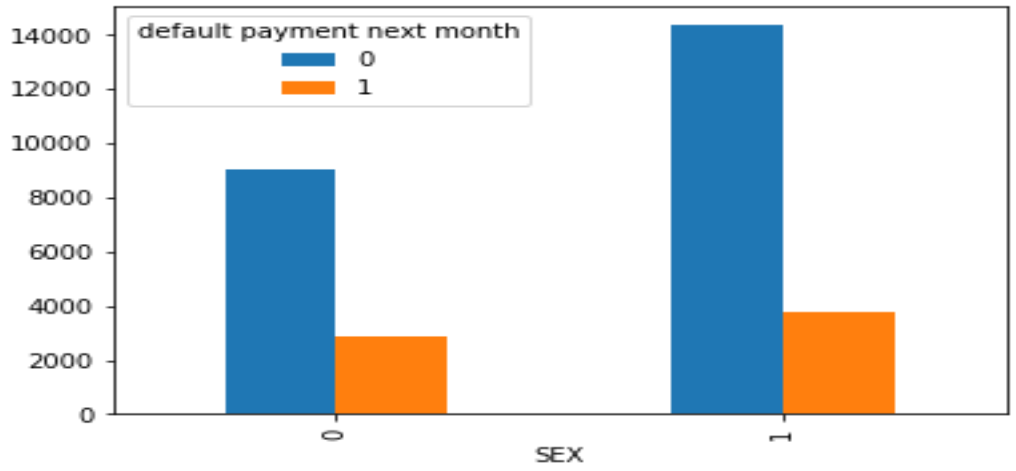


Figure 16

The above plot was used to visualize relationship between AGE and the target Variable. From the above plot we can find that the count of females are larger as compared to that of the males and the proportion of defaulters are higher for male customers.

MARRIAGE

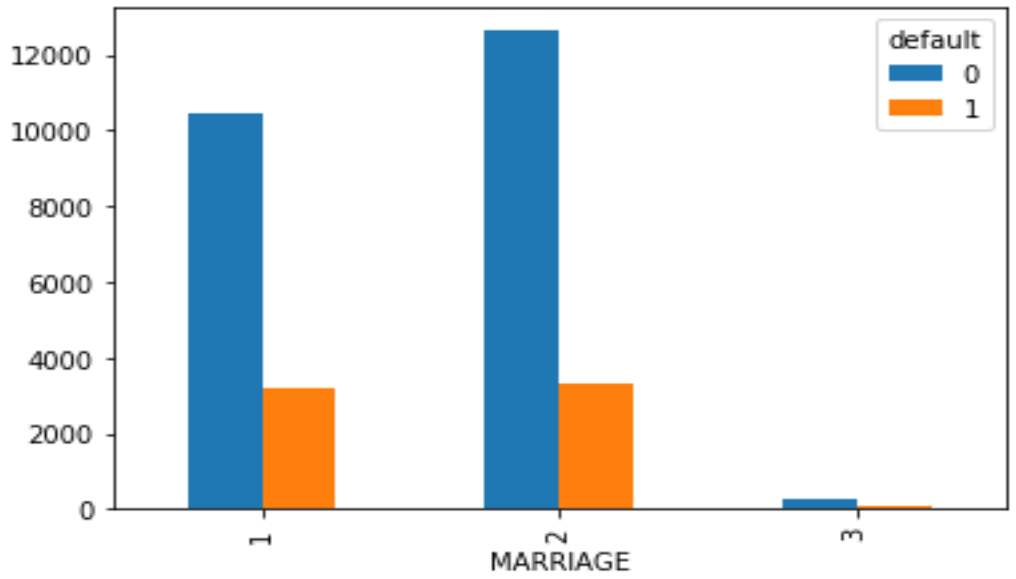


Figure 17

This plot shows the relation between Marriage and the target variable. Married customers are more likely to default, with 23.5% defaults. Single customers default are slightly less likely to default with 20.9% defaults. The small subset of "Other" has a ratio of 23.6% defaults

EDUCATION

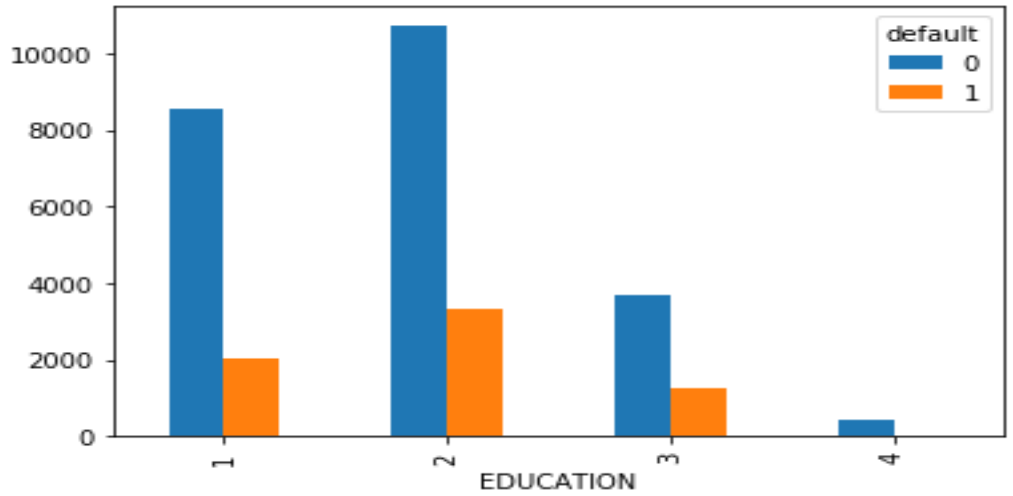


Figure 18

There are small differences in proportions of defaults in the categories, with "High school" at a default rate of 25.1%, "University" at 23.7%, and "Graduate school" at 19.2%.

PAY STATUS

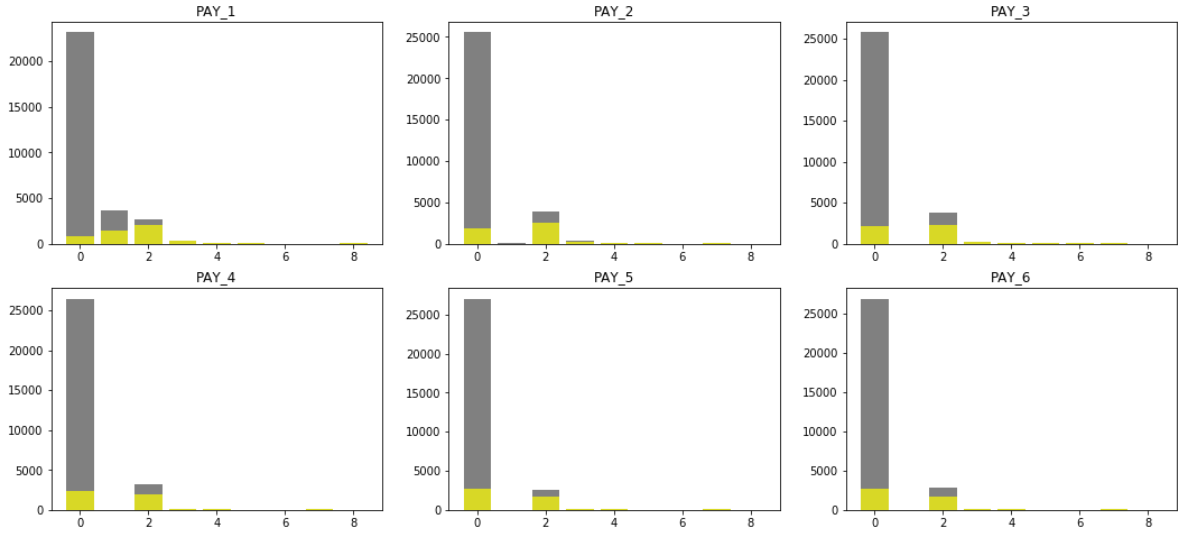


Figure19

The above plot shows the relationship between PAY and Target variable. Most of the customers have payment status as 0 which means that they have already paid for the previous month.

BILL_AMT

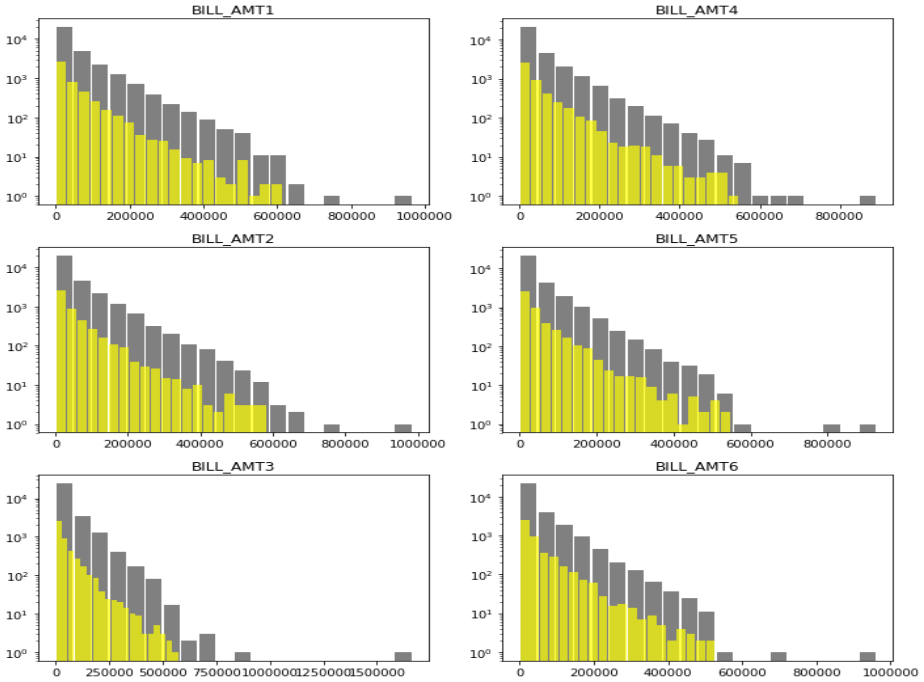


Figure 20

This shows the relationship between the BILL_AMT and target variable. As the bill amount increases the number of defaulters decreases.

PAY_AMT:

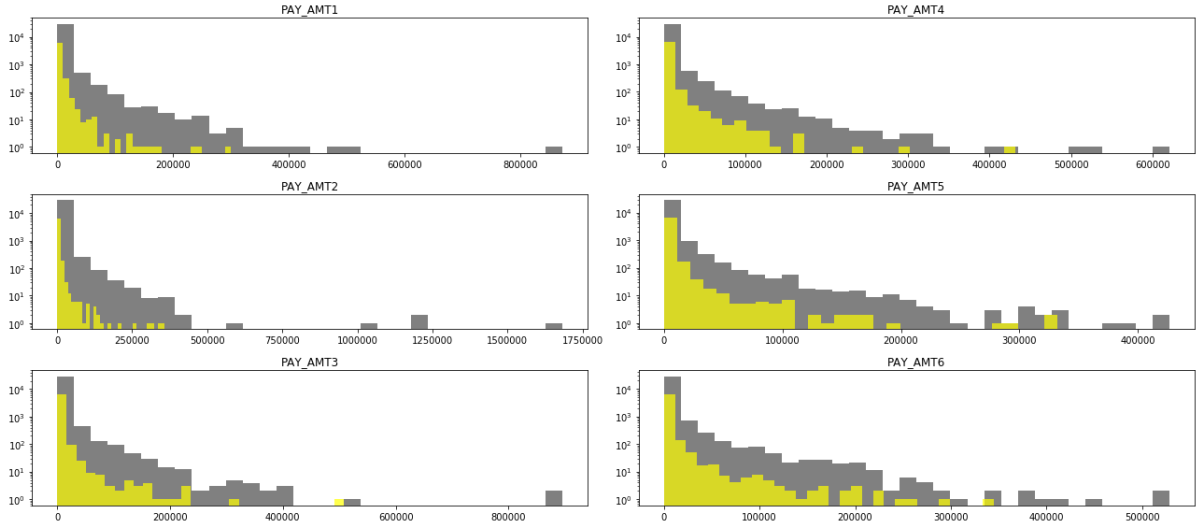


Figure 21

This shows the relationship between PAY_AMT and target variable. As the payment amount increases the number of defaulters decreases.

2.3 Correlation Matrix

BILL_AMT

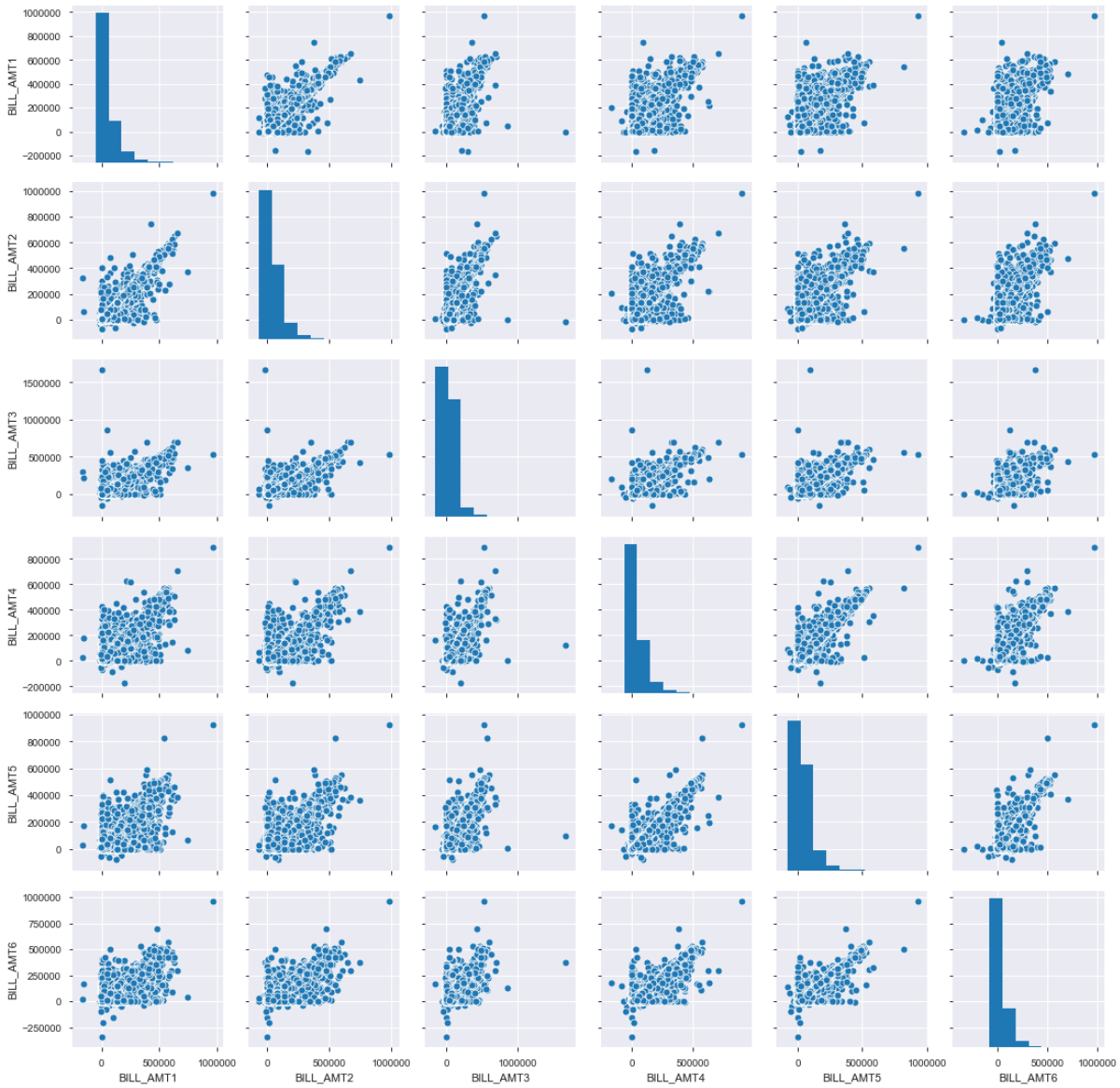


Figure 22

This plot shows the correlation between BILL_AMT variables. We can find that there is high correlation between the variables.

PAY_AMT

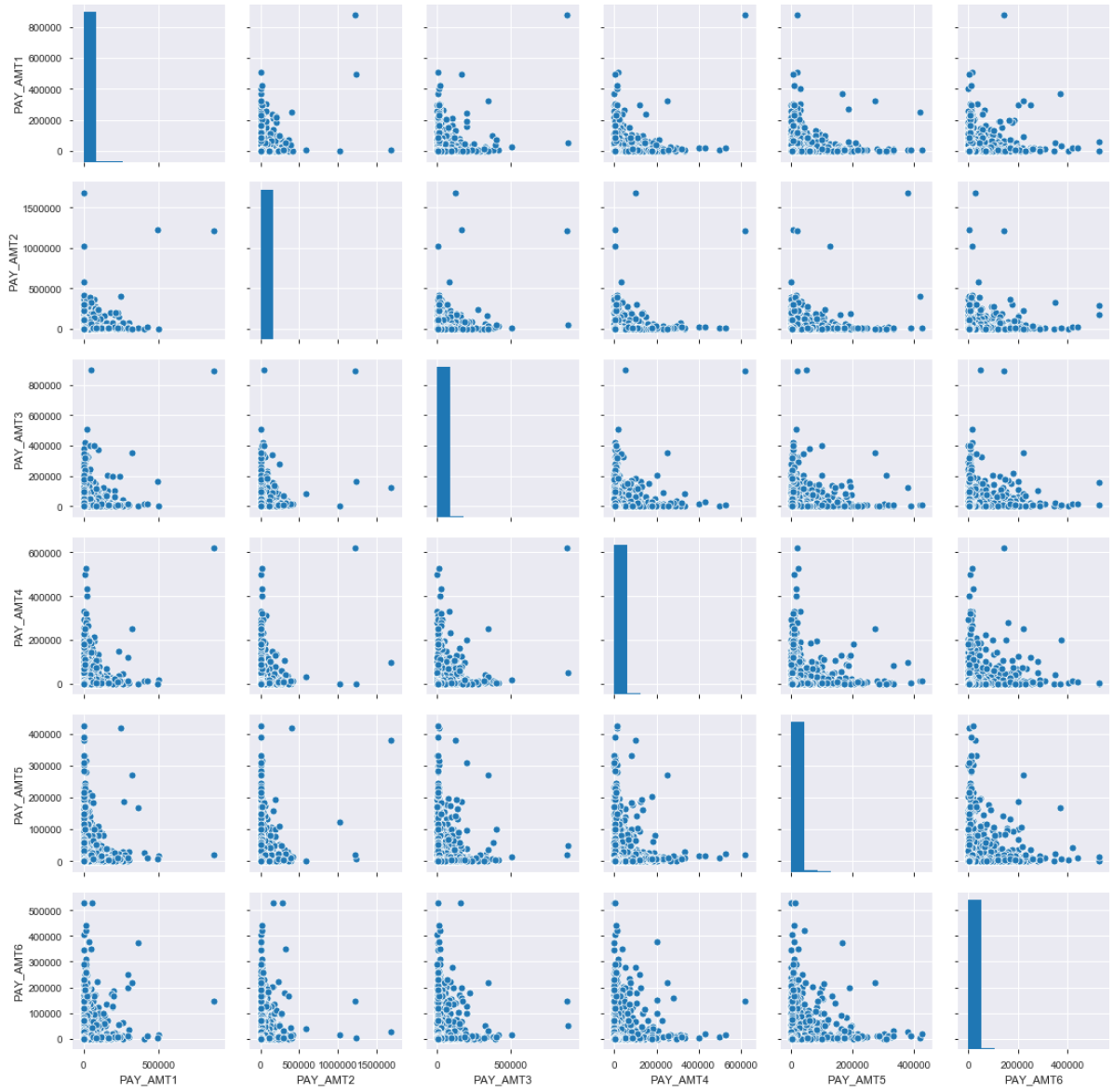


Figure 23

This plot shows the correlation between PAY_AMT variables. We can find that there is high correlation between the variables.

3. Overview of the Approach

3.1 Step-by-step walk through of the solution

After the exploratory data analysis & preparation of final data, the next steps we took towards solving our problem were as follows:

i. Splitting the data into training & test datasets, i.e. training dataset : the actual dataset that we use to train the model, model sees and learns from this data; test dataset : the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the test dataset is incorporated into the model configuration.

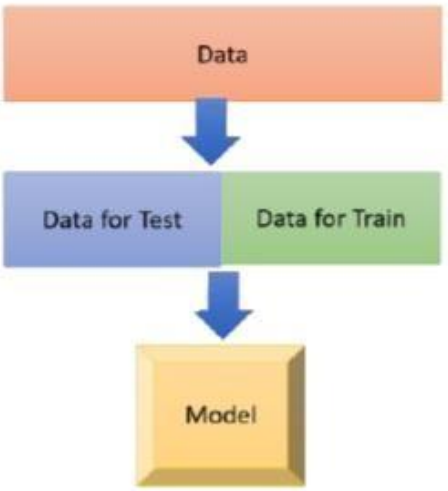


Figure 24

3.2 Model Evaluation

The final model which we have chosen as a solution for our problem statement was Random Forest with Ada Boosting. Before evaluating the final model, lets understand how the ensemble techniques, bagging & boosting works :

• **Bagging:** Decision Trees over fit the model and increases the variance. This makes the model vulnerable. Bagged Trees averages many models to reduce the variance. Although, Bagging is often applied to various Decision Tree but it can be used with any type of method. In addition to reducing variance, it also helps avoid over fitting.

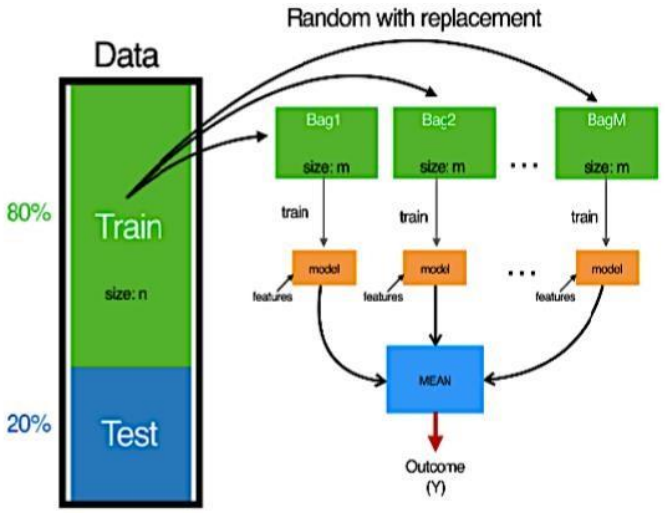


Figure 25

Draw m Samples from with replacement from original training set where m is a no less than or equal to N. Train the Decision Trees on the newly created bootstrapped samples. This Step 1 and Step 2 can be repeated n no of times. Typically, more trees, the better the model. To Generate the Prediction, we would simply average the predictions from these models created on m samples, to get a final prediction. Bagging can dramatically reduce the variance of unstable models (e.g. Decision trees) leading to improved Prediction.

 Boosting : Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous

model. Boosting gives mis-classified samples higher preference/weight. It is a method to "boost" weak learning algorithm ("single tree") into strong learning algorithm.

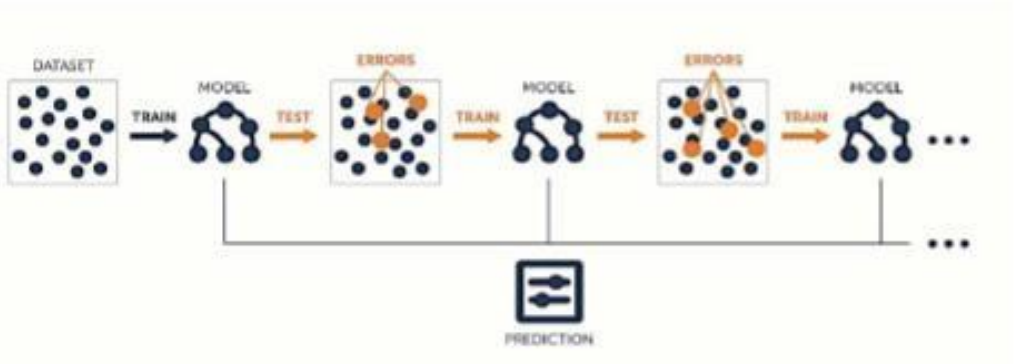


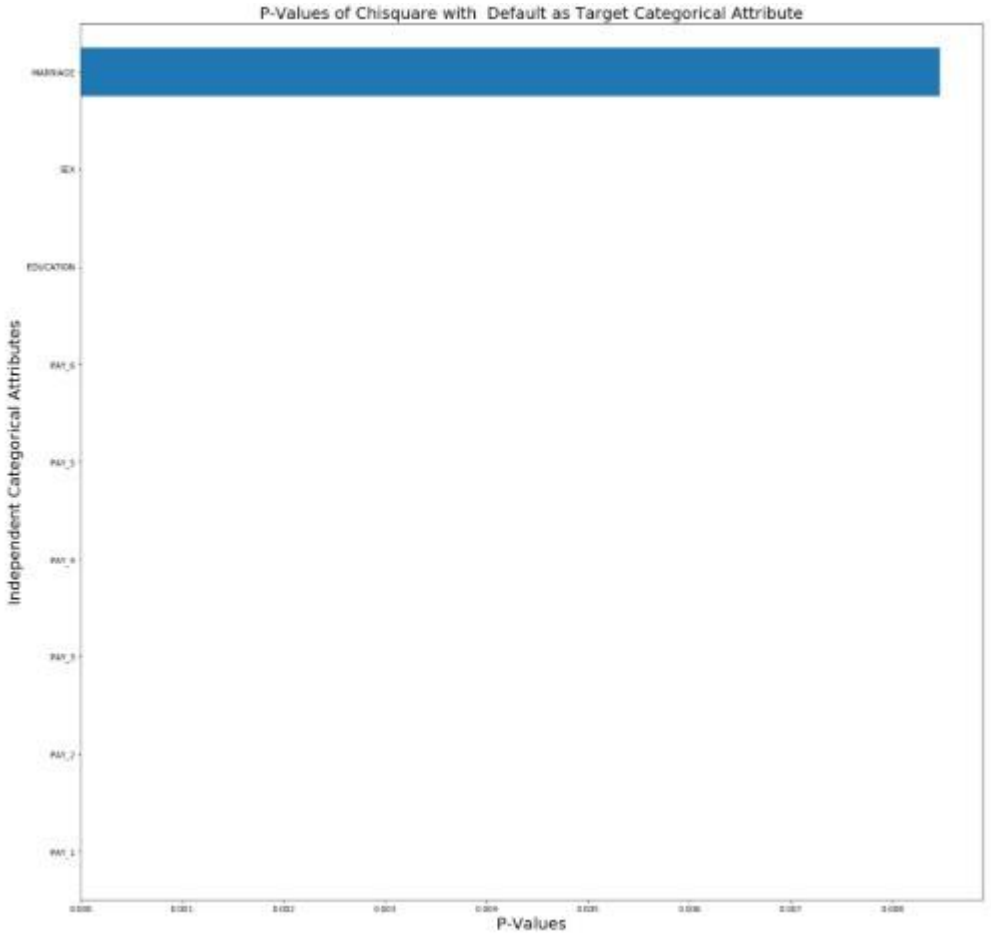
Figure 26

A subset is created from the original dataset. Initially, all data points are given equal weights & a base model is created on this subset. This model is used to make predictions on the whole dataset. Errors are calculated using actual values & predicted values. The observations which are incorrectly predicted, given higher weights. Before selecting the final model we did hyper parameter tuning & the prominent hyper parameters for the final model were chosen through GridSearchCV. After tuning our model with the best parameters we applied various ensemble techniques to reach at the best model which was Random Forest with Ada Boosting & got the highest accuracy with the combination.

4. Comparison and Implications

4.1 Comparison to Benchmark

Chi-Square test :



Since, we observe from chi2 test that Marriage column is of least importance as p_val is > 0.05 while all the other columns have p_val < 0.05 and hence, the other columns are important for our modeling.

Applied SelectKBest to get the top 20 features

	Specs	Score
0	LIMIT_BAL	1.057492e+08
18	PAY_AMT2	9.582460e+06
17	PAY_AMT1	8.290986e+06
22	PAY_AMT6	6.986573e+06
20	PAY_AMT4	6.067629e+06
19	PAY_AMT3	6.039725e+06
21	PAY_AMT5	5.947016e+06
16	BILL_AMT6	2.106857e+06
15	BILL_AMT5	1.726431e+06
11	BILL_AMT1	9.045943e+05
14	BILL_AMT4	5.777365e+05
12	BILL_AMT2	8.573634e+04
5	PAY_1	1.890309e+04
6	PAY_2	1.848116e+04
7	PAY_3	1.588063e+04
8	PAY_4	1.584077e+04
9	PAY_5	1.580411e+04
10	PAY_6	1.393191e+04
13	BILL_AMT3	1.922741e+02
2	EDUCATION	1.476439e+01

- The above are the top 20 features of the dataset

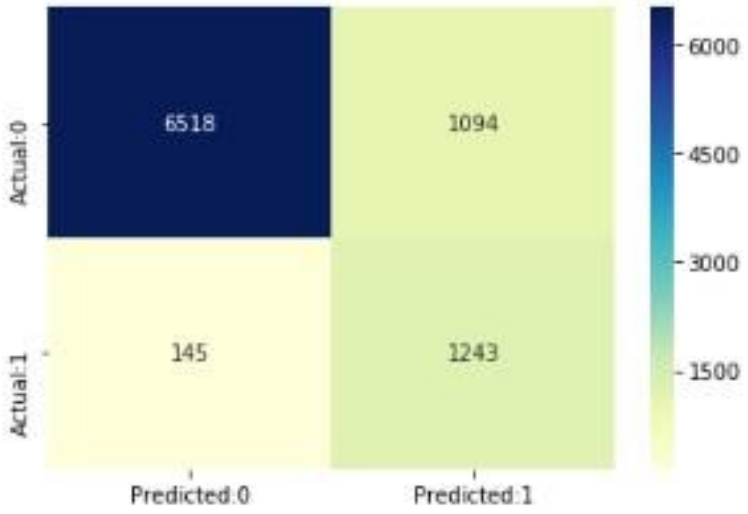
Base Line Model : Logistic Regression model taking all variables into consideration.

Logit Regression Results

Dep. Variable:	default	No. Observations:	21000
Model:	Logit	Df Residuals:	20976
Method:	MLE	Df Model:	23
Date:	Thu, 28 Nov 2019	Pseudo R-squ.:	0.4960
Time:	12:14:04	Log-Likelihood:	-4612.4
converged:	True	LL-Null:	-9151.4
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8386	0.163	-17.405	0.000	-3.158	-2.519
LIMIT_BAL	-0.1420	0.038	-3.786	0.000	-0.216	-0.069
SEX	-0.0738	0.057	-1.284	0.199	-0.186	0.039
EDUCATION	-0.2070	0.042	-4.962	0.000	-0.289	-0.125
MARRIAGE	-0.1558	0.060	-2.603	0.009	-0.273	-0.039
AGE	0.0734	0.031	2.392	0.017	0.013	0.134
PAY_1	1.6822	0.039	42.827	0.000	1.605	1.759
PAY_2	0.1918	0.038	5.063	0.000	0.118	0.266
PAY_3	0.3691	0.043	8.526	0.000	0.284	0.454
PAY_4	0.3588	0.048	7.471	0.000	0.265	0.453
PAY_5	0.2774	0.051	5.447	0.000	0.178	0.377
PAY_6	0.6498	0.043	15.102	0.000	0.565	0.734
BILL_AMT1	-0.5782	0.178	-3.241	0.001	-0.928	-0.229
BILL_AMT2	0.6335	0.204	3.100	0.002	0.233	1.034
BILL_AMT3	-0.2263	0.185	-1.221	0.222	-0.590	0.137
BILL_AMT4	0.0523	0.178	0.294	0.769	-0.296	0.401
BILL_AMT5	0.2896	0.177	1.633	0.103	-0.058	0.637
BILL_AMT6	-0.1589	0.135	-1.174	0.240	-0.424	0.106
PAY_AMT1	-0.2100	0.073	-2.867	0.004	-0.354	-0.066
PAY_AMT2	-0.2552	0.110	-2.315	0.021	-0.471	-0.039
PAY_AMT3	-0.0409	0.053	-0.771	0.441	-0.145	0.063
PAY_AMT4	-0.1301	0.062	-2.101	0.036	-0.251	-0.009
PAY_AMT5	-0.1136	0.067	-1.699	0.089	-0.245	0.017
PAY_AMT6	-0.2342	0.058	-4.021	0.000	-0.348	-0.120

Confusion Matrix



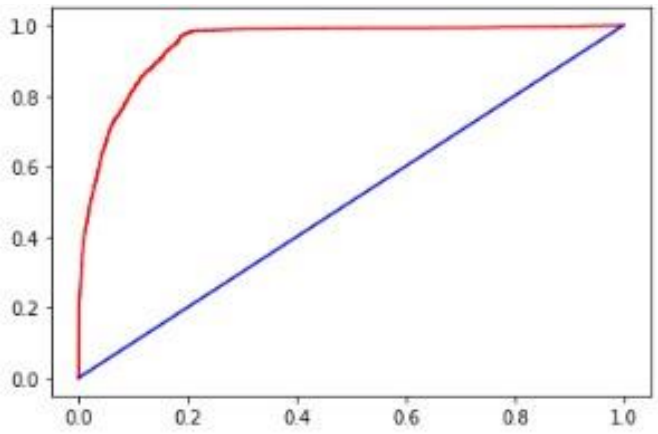
Classification Report

	precision	recall	f1-score	support
0	0.98	0.85	0.91	7611
1	0.52	0.91	0.66	1389
accuracy			0.86	9000
macro avg	0.75	0.88	0.79	9000
weighted avg	0.91	0.86	0.87	9000

ROC Curve

	Train Accuracy	Test Accuracy	F1 Score	Recall Score	Precision Score	Roc Auc Score
Logistic Regression Model	0.870993	0.878889	0.681659	0.840173	0.573464	0.949730
K Nearest Neighbor Model	0.943779	0.862000	0.654808	0.848092	0.533273	0.915355
Decision Tree Model	0.999774	0.877111	0.630841	0.680346	0.588052	0.797256
Random Forest Model	0.997569	0.888556	0.659423	0.699064	0.624036	0.939688
Naive Bias	0.872576	0.851111	0.645127	0.876890	0.510264	0.910826

- From the base models we observe that Logiostic Regression is giving the best accuracy considering all the metrics.



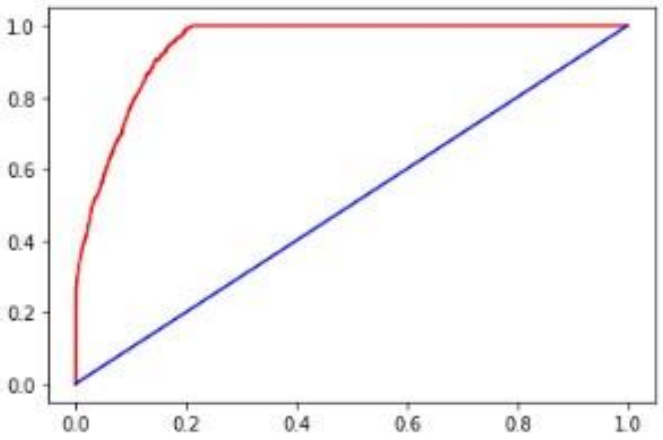
Final Model:

	Train Accuracy	Test Accuracy	F1 Score	Recall Score	Precision Score	Roc Auc Score
Logistic Regression Model	0.870993	0.878889	0.681659	0.840173	0.573464	0.949730
K Nearest Neighbor Model	0.943779	0.862000	0.654808	0.848092	0.533273	0.915355
Decision Tree Model	0.999774	0.877667	0.632387	0.681785	0.589664	0.798174
Random Forest Model	0.998163	0.885556	0.653665	0.699784	0.613249	0.939685
Naive Bias Model	0.872576	0.851111	0.645127	0.876890	0.510264	0.910826
Logistic Regression with Hyper Parameter Tuning	0.871135	0.878889	0.681659	0.840173	0.573464	0.910826
Decision Tree with Hyper Parameter Tuning	0.922098	0.863111	0.668103	0.892729	0.533793	0.946468
Random Forest with Hyper Parameter Tuning	0.887896	0.853556	0.659255	0.917927	0.514320	0.939685
Decision Tree with RFE	0.912429	0.912556	0.670297	0.575954	0.801603	0.931975
Random Forest with RFE	0.899000	0.894444	0.575134	0.462923	0.759150	0.718060
Decision Tree with Bagging	0.892730	0.832222	0.638583	0.960403	0.478308	0.884616
Decision Tree with Adaboosting	0.949432	0.886222	0.648352	0.679626	0.619829	0.801776
Random Forest with Adaboosting	0.949036	0.890000	0.681057	0.760979	0.616327	0.837263
Model with Gradient Boosting	0.881989	0.857667	0.659036	0.891289	0.522804	0.871410
Hyper parameterized Model with Gradient Boosting	0.881932	0.867667	0.663464	0.845212	0.546047	0.928605

Random Forest Classifier with Hyper Parameter tuning gives the best test accuracy of 85.64% with an excellent recall score of 91.28% and a precision score of 51.98% while the f1 score is 66.24% and the roc_auc score is 94.79%.

Confusion Matrix:

ROC Curve:



4.2 Implications

Creditors and lenders utilize a number of financial tools to evaluate the credit worthiness of a potential borrower. When both lender and borrower are businesses, much of the evaluation relies on analysing the borrower's balance sheet, cash flow statements, inventory turnover rates, debt structure, management performance, and market conditions. Creditors favour borrowers who generate net earnings in excess of debt obligations and any contingencies that may arise. Following are some of the factors lenders consider when evaluating an individual or business that is seeking credit:

Credit worthiness: A history of trustworthiness, a moral character, and expectations of continued performance demonstrate a debtor's ability to pay. Creditors give more favourable terms to those with high credit ratings via lower point structures and interest costs.

Size of debt burden: Creditors seek borrowers whose earning power exceeds the demands of the payment schedule. The size of the debt is necessarily limited by the available resources. Creditors prefer to maintain a safe ratio of debt to capital.

Loan size: Creditors prefer large loans because the administrative costs decrease proportionately to the size of the loan. However, legal and practical limitations recognize the need to spread the risk either by making a larger number of loans, or by having other lenders participate. Participating lenders must have adequate resources to entertain large loan applications. In addition, the borrower must have the capacity to ingest a large sum of money.

Frequency of borrowing: Customers who are frequent borrowers establish a reputation which directly impacts on their ability to secure debt at advantageous terms.

Length of commitment: Lenders accept additional risk as the time horizon increases. To cover some of the risk, lenders charge higher interest rates for longer term loans.

Social and community considerations: Lenders may accept an unusual level of risk because of the social good resulting from the use of the loan. Examples might include banks participating in low-income housing projects or business incubator programs.

5. Limitations and Scope

5.1 Limitations:

- One of the limitations was that the data cleaning process was hectic.
- Most of the categorical columns had unknowns in them.
- During data entry some of the defaulters were stated as non-defaulters and vice versa, so we had to transform the data.
- There are many outliers in the data which cannot be treated as the data is sensitive, if treated may result in information loss.

5.2 Scope

- If data for the whole year was taken into account better prediction models and patterns can be formed.
- The loss caused by defaulters to banks can be drastically reduced if the above considered features are given more importance.