CREDIT CARD DEFAULTER PREDICTION USING MACHINE LEARNING - CLASSIFICATION

A dissertation submitted for the Post Graduate Programme in
Data Science and Engineering

Great Learning Bangalore
2023

BATCH DETAILS	**PGP-DSE Bangalore August 2022**
TEAM MEMBERS	**SMRITI AGARWAL,** **KARTHIK RAO,** **GUNJAN PREET KAUR,** **MOZAMMIL ALAM,** **SREERAG P**
DOMAIN OF PROJECT	**FINANCE**
PROJECT TITLE	**CREDIT CARD DEFAULTER PREDICTION USING MACHINE LEARNING - CLASSIFICATION**
GROUP NUMBER	**GROUP – 1**
TEAM LEADER	**SMRITI AGARWAL**
MENTOR NAME	**ANJANA AGRAWAL**

Date: 27/01/2023

ABSTRACT

Increasing Non-Performing Loans and advances by banking and financial institutions have become one of the main problems in this era. The rise in the number of willful defaulters in the banking and finance sectors has been particularly problematic. Card portfolios represent a major part of the loans and advances portfolio of a bank. Banks issue credit cards after reviewing credit card applications. Except for the fully secured credit cards, other credit cards are not required to attach any security, which increases the risk of loss due to defaulters.

Therefore, the issuers of credit cards must have a better understanding of and experience analyzing the credit card applications they receive from customers. However, the analysis techniques used by them might have a few shortcomings; they might be less accurate, or personal judgment might influence the analysis. Therefore, proper credit card application analysis and a customer credit scoring model are needed for making informed decisions.

This research study focuses on developing a machine-learning process using supervised learning methods to predict the credit card default probability of credit card applicants before issuing the credit card to customers. This model can be applied in the banking sector to reduce the risk of issuing credit cards to future defaulters.

To develop a supervised learning model and testing the model, the data for Taiwanese credit card clients is used for this study. The research study will be conducted using supervised learning algorithms like decision tree, random forest, k-nearest neighbours and a few boosting algorithms. After analyzing the performance of these algorithms, the best algorithm for this kind of prediction will be selected, and a final model will be built.

OBJECTIVES:

The objectives are:

1. To find out the best classification algorithm for predicting Credit Card default.
2. To devise a model to predict whether a customer will default in paying their credit card bill next month or not.

INDUSTRY REVIEW:

BACKGROUND:

In the face of the market economy, credit cards are the emerging trend and are swiftly replacing other forms of payment. People are more inclined towards acquiring credit cards that allow them to make purchases ahead of actual payment. However, it is often possible that consumers are unable to estimate their repayment capabilities while making a purchase. This can cause issues for the banks by increasing their loan risk as well as for the consumers as it increases their credit crisis. As this medium of payment becomes more common, the risk of credit card default is also increasing. Hence, it is imperative for credit card issuers to effectively identify high-risk credit card defaulters and mitigate debt risk. The major factors that will indicate a customer's chance of default are considered by the credit card issuers while verifying their customer’s eligibility. Another significant factor contributing to the issuers' conundrum is their choice to allocate the credit limit to specific consumers.

Hence, banks should develop mechanisms that can assess the probability of default on the consumer side.

CURRENT SOLUTION TO THE PROBLEM:

The job of identifying defaulters can be challenging. Currently, loan officers decide whether to approve your application based on your credit history and data verification. If an applicant has a bad credit score, they may be denied credit or get it with harsher terms like high-interest rates.

However, these criteria can pose a few challenges. First, it is possible that a customer might not have any prior credit history, so there will be a lack of data to base any decision on. Second, loan officers are often subjective. There is a risk of their decisions being heavily influenced by personal factors.

An example of issues faced while resorting to these methods is that if a person fails to inform the bank or issuer about an address change, it is difficult to trace the person in the event of default.

Therefore, studying credit-card charge-off rates through mathematical models seems both timely and relevant.

PROPOSED SOLUTION TO THE PROBLEM:

To deal with the problem, we propose to develop a model for credit card defaulter prediction using machine learning techniques. We will train the machine with the previous dataset, while taking into consideration all the factors that influence a person's risk for defaulting, which are often overlooked in traditional methods of examining an application.

Feeding the machine with historical data will train the machine to predict defaulters when fed with new data. Hence, it will help in the advanced detection of defaulters, thus reducing credit risk.

LITERATURE SURVEY:

The global banking and financial services industry has faced a remarkable shift in meeting new challenges. New technological innovations, advanced communication modes, the internet, and the expansion of bank - branch networks lead to intense competition in the banking industry. As a result of the Sub-Prime crisis and bank runs, strict banking rules and regulations have been imposed on banks by national and international banking authorities. These rules and regulations have further restricted the business opportunities of banks. Due to the consequences of the Sub-Prime Crisis and the Basel III Capital accord implementation, banks were compelled to have more precautionary actions than ever before in addressing high-risk exposures. Banks have no competitive advantages for a prevailing long time- period, while banks are becoming more demanding and selective in their preferences. The health of the economy is closely related to the soundness of its financial system. One of the most important participants in the financial system is the banking system. The financial sector of Taiwan is dominated by the banking enterprises, especially the commercial banks. For the development of the economy, the banks provide a greater portion of strength. With the developing and modern financial environment, the Banks also improve the quality of the service that they are providing to the customers through new technologies. Taiwan is no exception to these effects, and almost all industries, including private and nationalized banks, are providing varied services to attract customers.

The banking system is the lifeblood of the economy. Without a proper banking system, economic stability and growth cannot be achieved. Banks get deposits from depositors and grant loans and advances to the general public using those deposits. Those loans and advances are keys to the growth of investment in the country and are important parts of any organization in the country. These loans and advances are being given to individuals as well as corporations. Increasing competition within the financial industry adds more value to the loans and advances portfolio of the financial institutions. However, granting loans and advances will increase the risk of the financial institutions. According to Investopedia, credit risk is the risk or possibility of arising loss, as a result of non-repayment of loans & advances and breach of the obligations. As a result of credit risk, lenders do not receive principal amount or interest as agreed. Loans & advances portfolio of financial institutions consists of term loans, credit cards, pawning, leasing and other loans. Credit card portfolio is a most important part of the Banks' lending portfolio. In the bank's loan & advance portfolio, credit cards have a major portion and importance.

Due to the risk associated with the issuance of credit cards, it is important for banking institutions to identify which of their potential customers might default when it comes to timely payments. Our proposal for a model to predict future defaulters is to help these institutions make informed decisions regarding their customers. This will help these institutions to reduce their risk for incurring losses.

DATASET AND DOMAIN:

1) DATA DICTIONARY:

The dataset contains credit card data for cardholders in Taiwan between April 2005 and September 2005. There are 30000 rows and 25 columns. There are a total of 6636 defaulters out of 30.000 total observations. With the positive class (defaulters) accounting for 22% of all transactions, this data set is significantly imbalanced.

2) VARIABLE CATEGORIZATION (COUNT OF NUMERIC AND CATEGORICAL):

Out of the 25 columns:

1. SEX, EDUCATION, MARRIAGE,   PAY_0 (which is renamed as PAY_1), PAY_2, PAY_3, PAY_4, PAY_5 and PAY_6 appear as numerical initially but they are categorical columns. Hence, there are a total of 9 categorical columns.

2. Our target variable i.e. default.payment.next.month is categorical in nature making the final count of categorical columns as 10.

3. Columns ID,AGE,LIMIT_BAL, BILL_AMT1-BILL_AMT6 and PAY_AMT1-PAY_AMT6 are numerical in nature. Hence, a total of 14 columns are numerical.

Data 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: (0=unknown,1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

● MARRIAGE: Marital status (0=unknown,1=married, 2=single, 3=others)

● AGE: Age in years

● PAY_0: Repayment status in September, 2005 (-2=no consumption (inactive account), -1=paid duly,0=use of revolving credit, 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)

3) PRE PROCESSING DATA ANALYSIS:

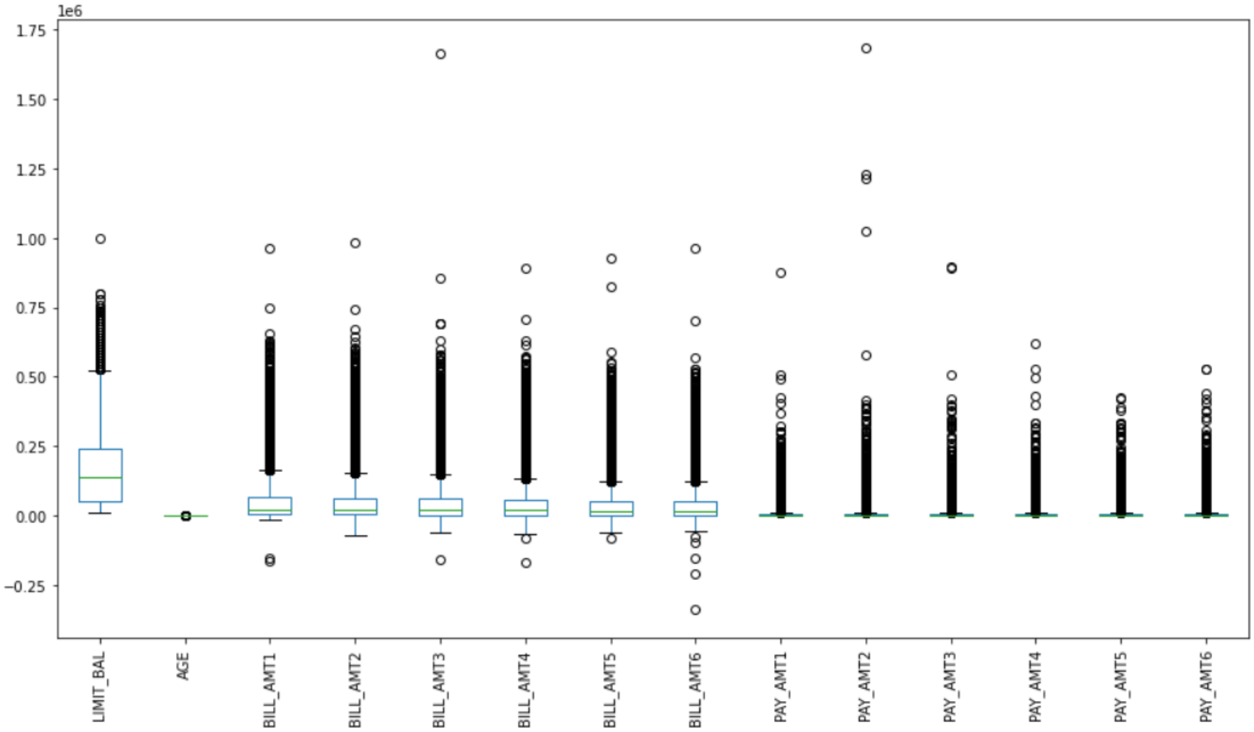
Missing values: It is found that there are no missing values in the data.

Outliers: There is presence of outliers in all the numerical columns.

Redundant columns: ID column is redundant, and hence it is dropped.

DATA EXPLORATION (EDA):

Checking for outliers:



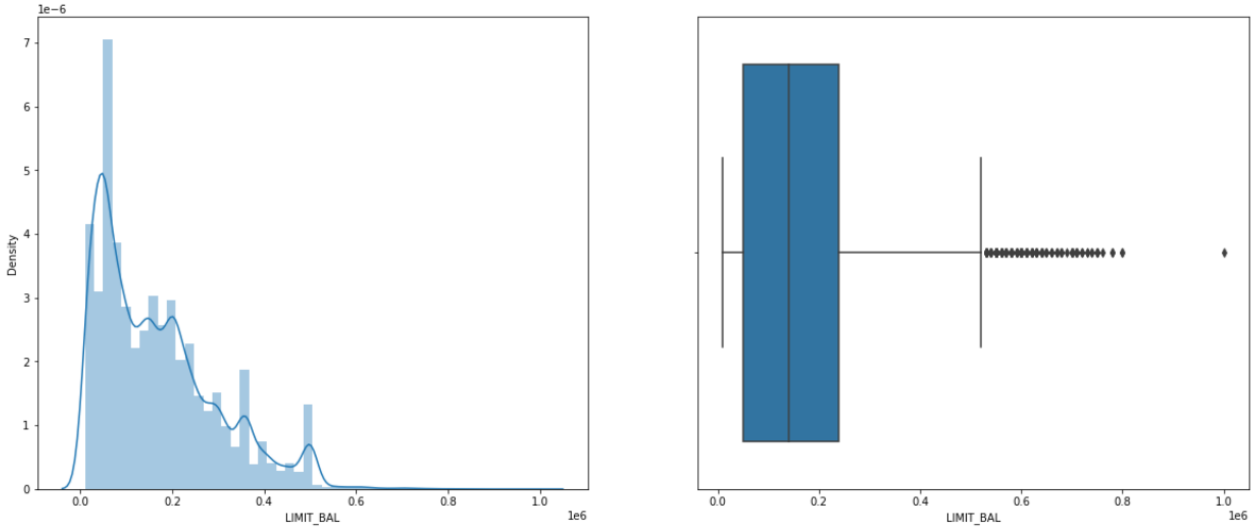
We cannot apply the IqR method for treating outliers. A credit card dataset having many outliers is a normal occurrence in real life scenarios. So, in this case, we will have to examine each column separately, and any unusual values will have to be treated as missing values.

Univariate Analysis¶

We made distribution plots and boxplots for studying the numerical variables and count plots for our categorical variables.

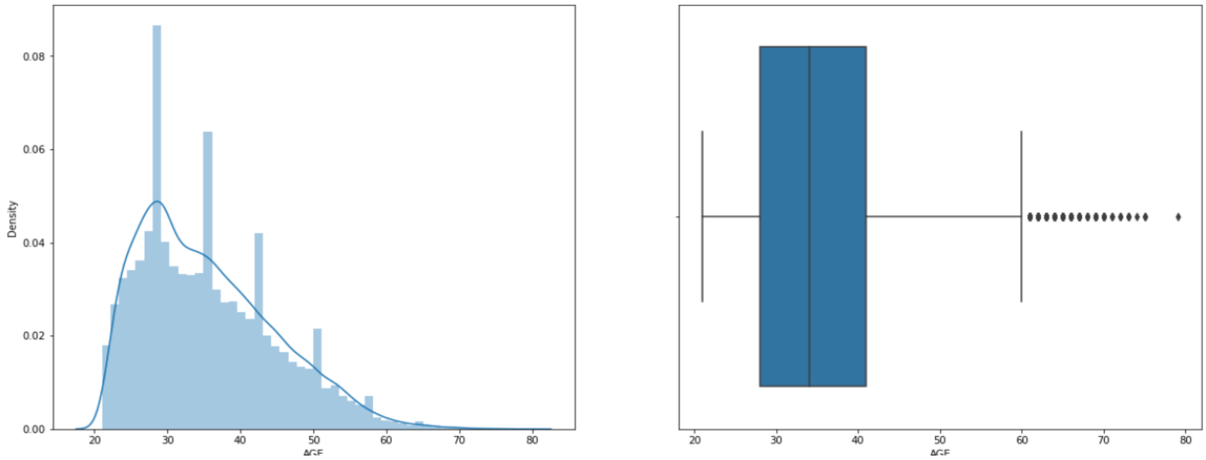
Numerical Variables:-

LIMIT_BAL



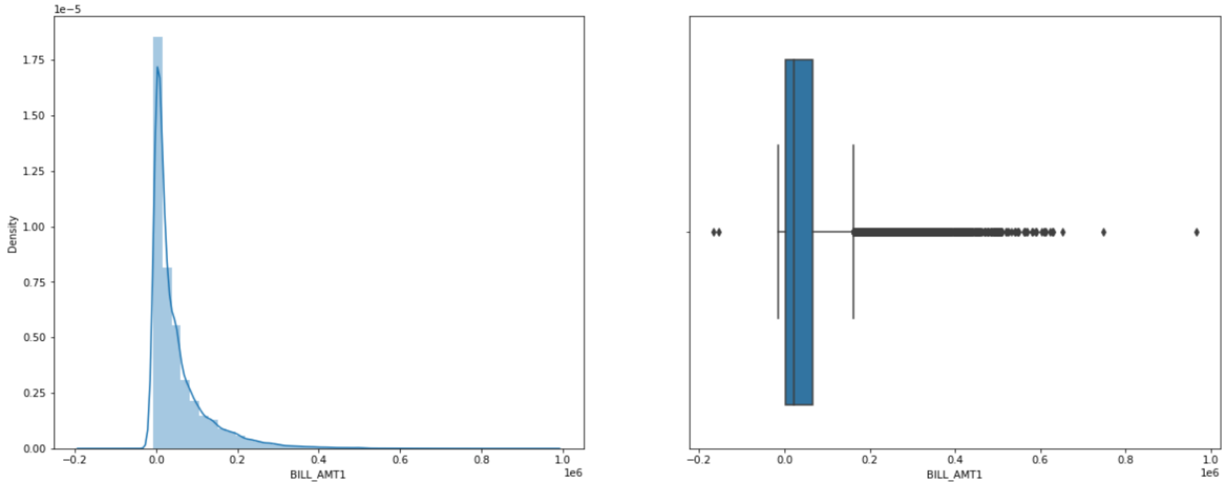
LIMIT_BAL column is right skewed and the boxplot shows the presence of outliers.

Age

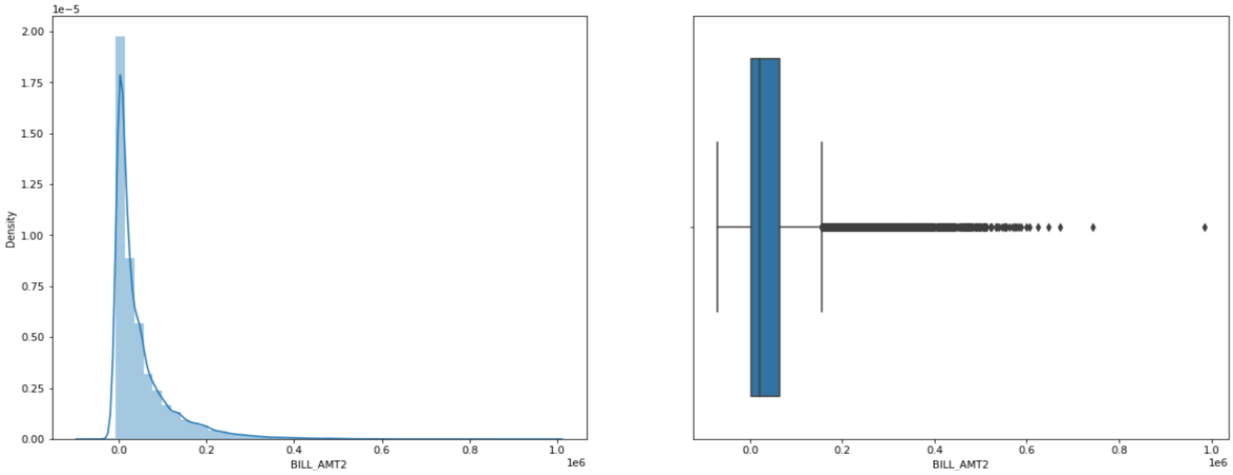


AGE column shows slight skewness on the right side. Fewer outliers are present when compared to other columns.

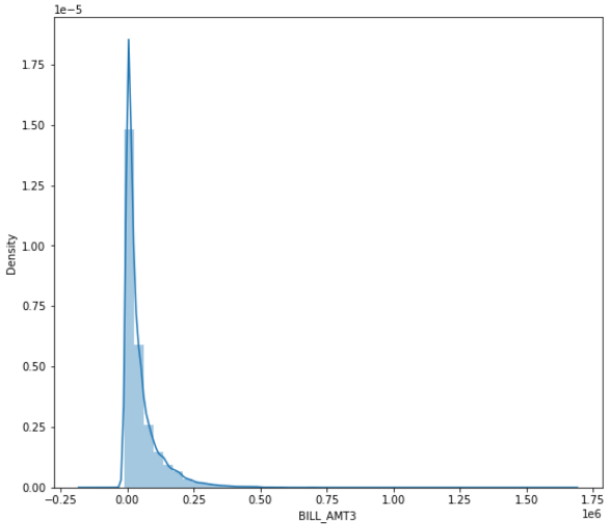
BILL_AMT1

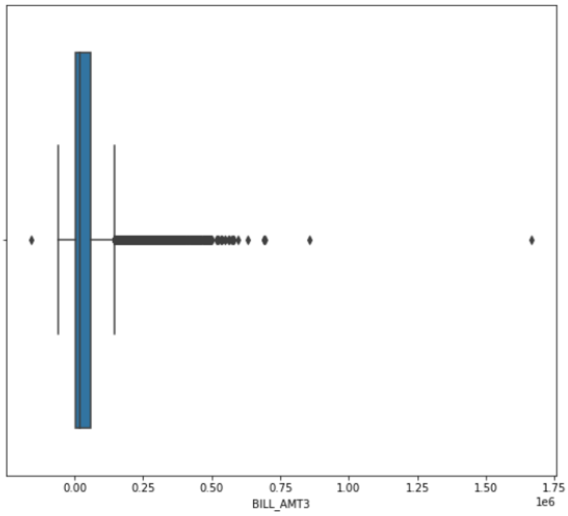


BILL_AMT2

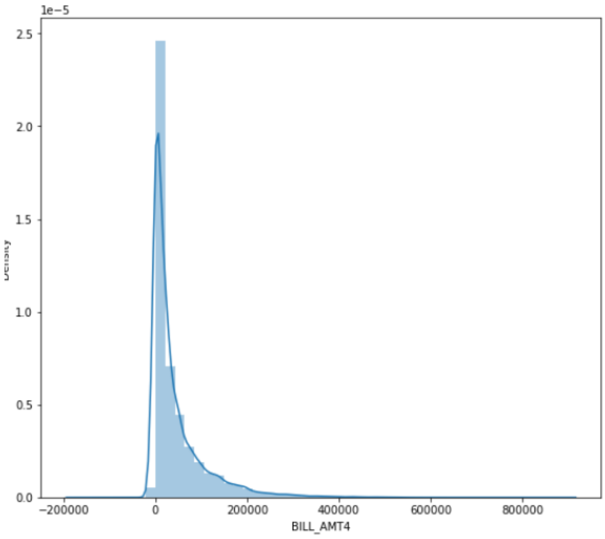


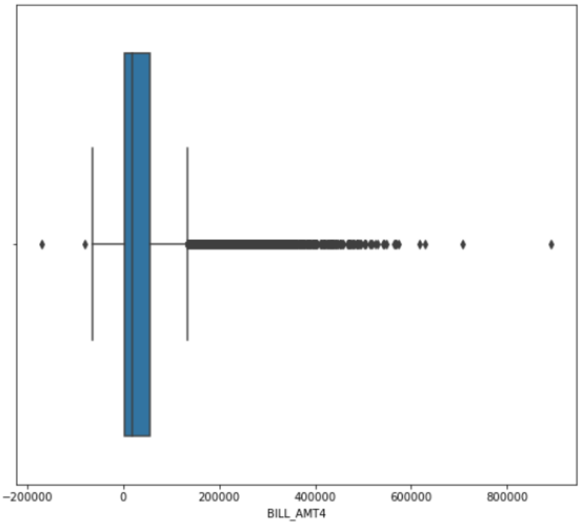
BILL_AMT3



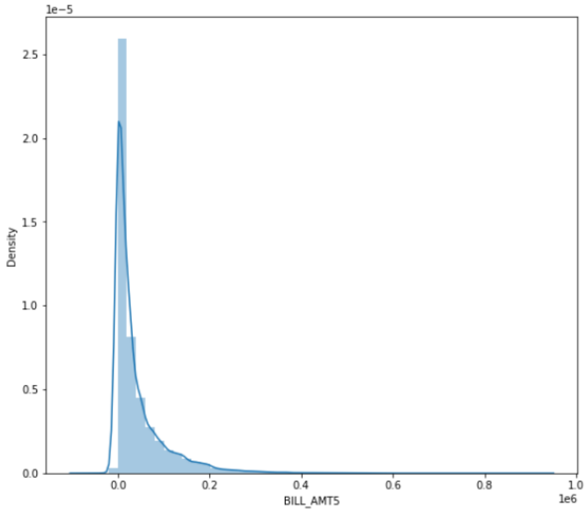


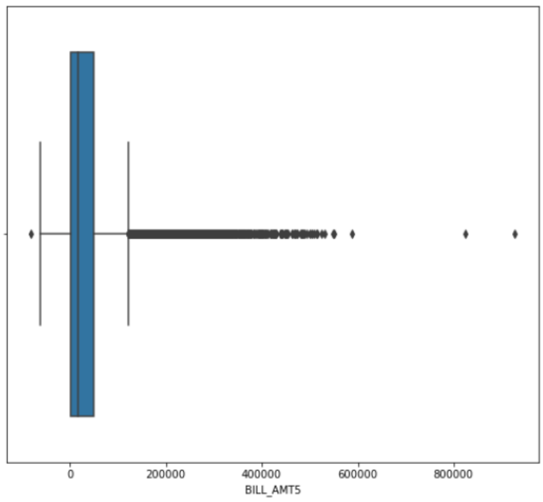
BILL_AMT4



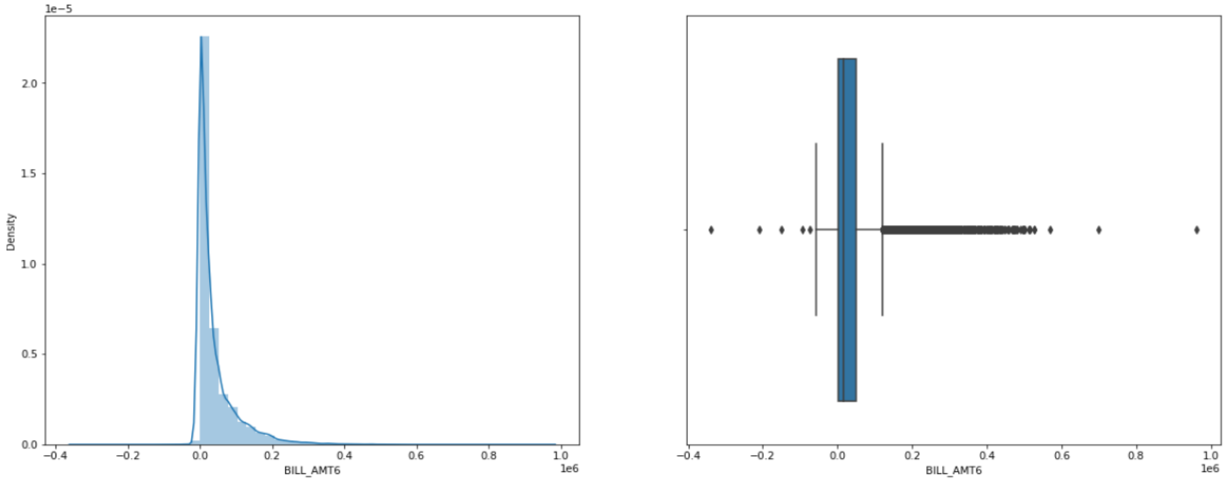


BILL_AMT5



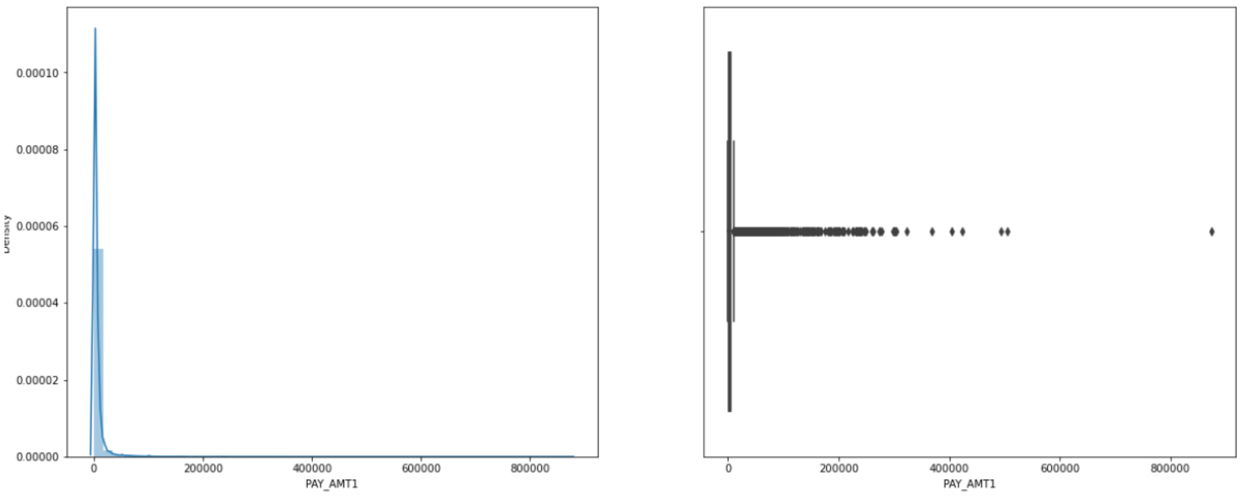


BILL_AMT6

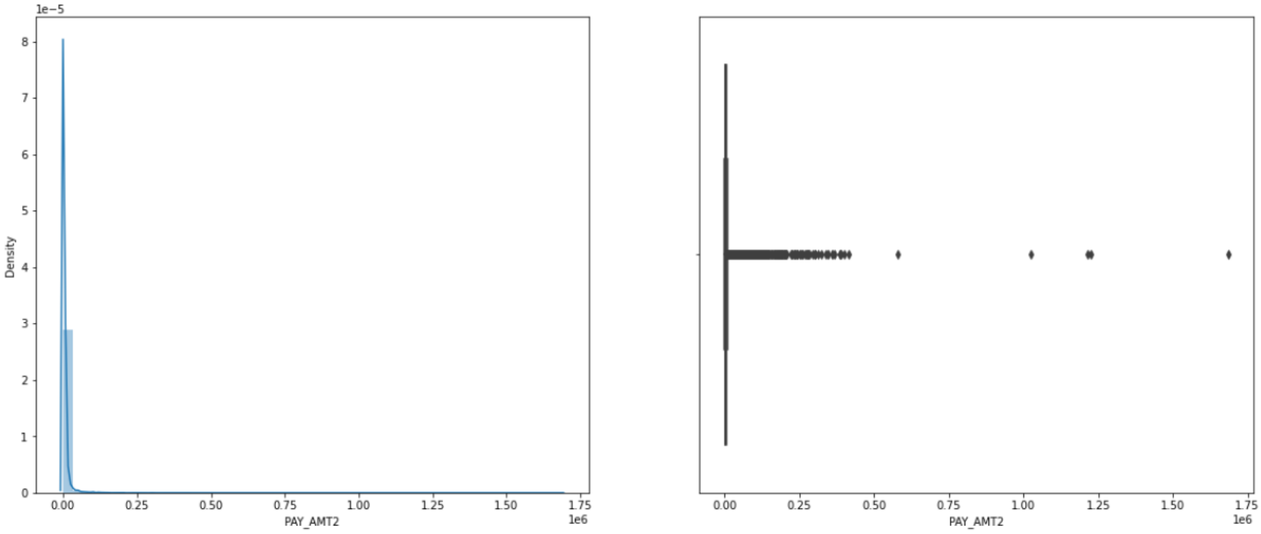


BILL_AMTX columns are highly right skewed. Most of the observations are concentrated around 0. Boxplots show the presence of numerous outliers.

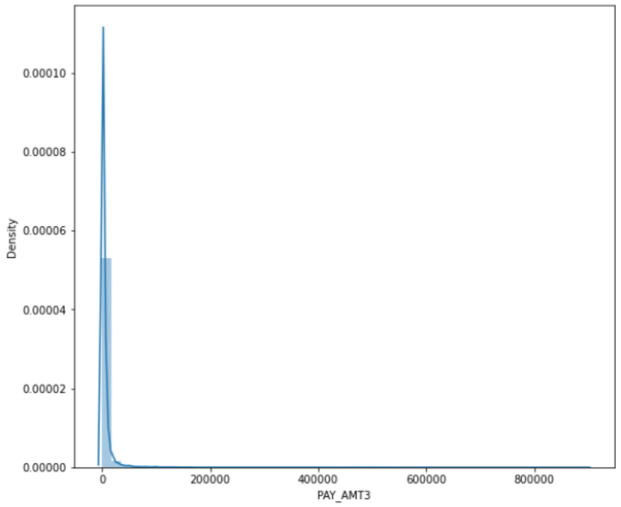
PAY_AMT1

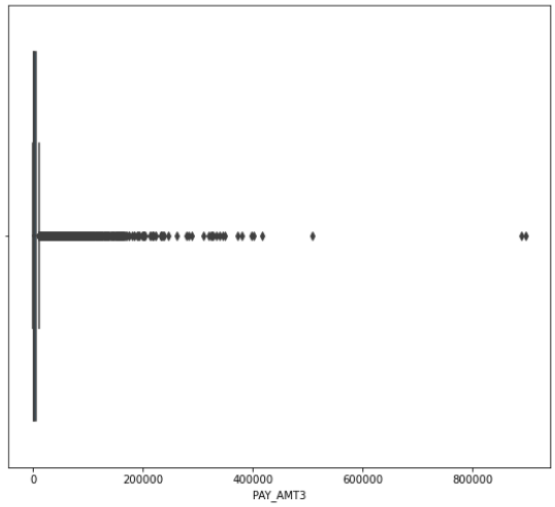


PAY_AMT2

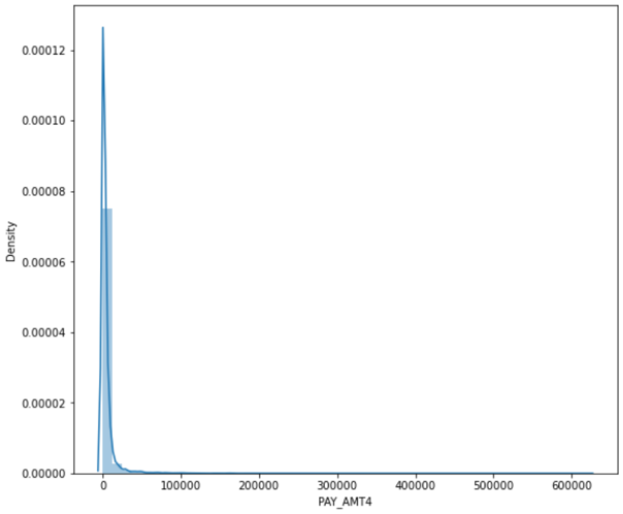


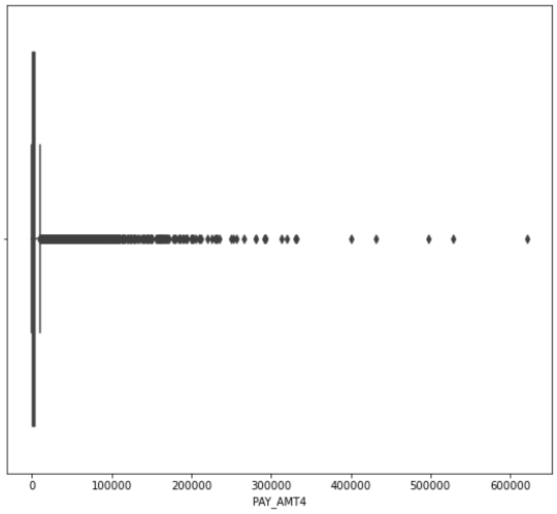
PAY_AMT3



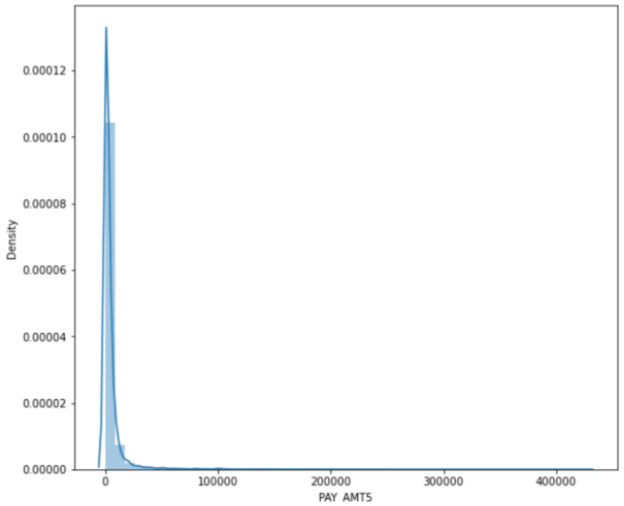


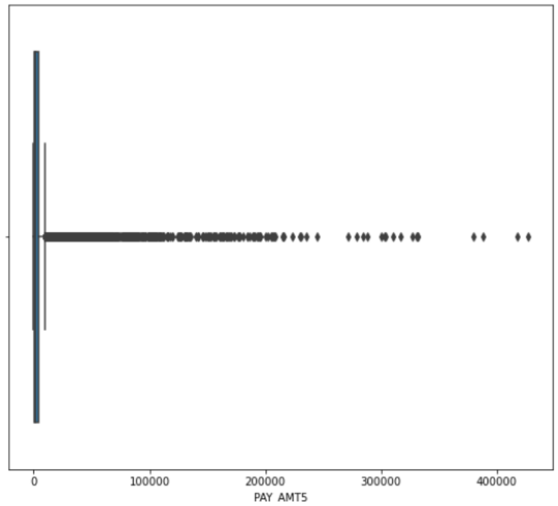
PAY_AMT4



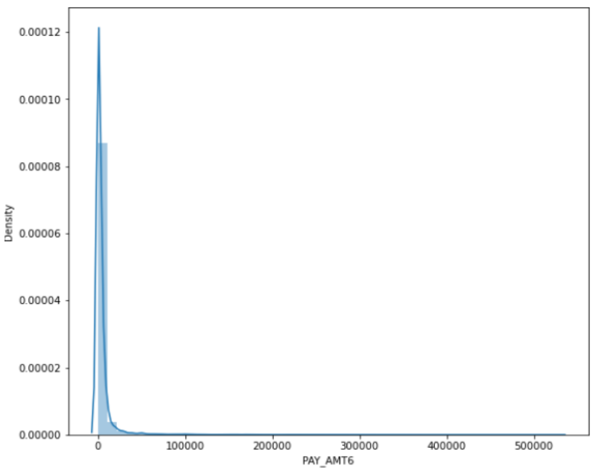


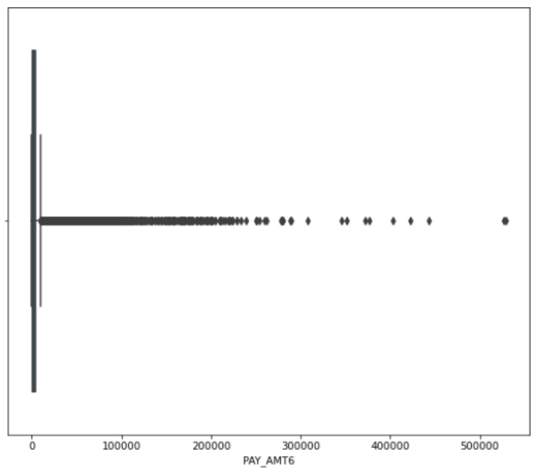
PAY_AMT5





PAY_AMT6





Inferences:

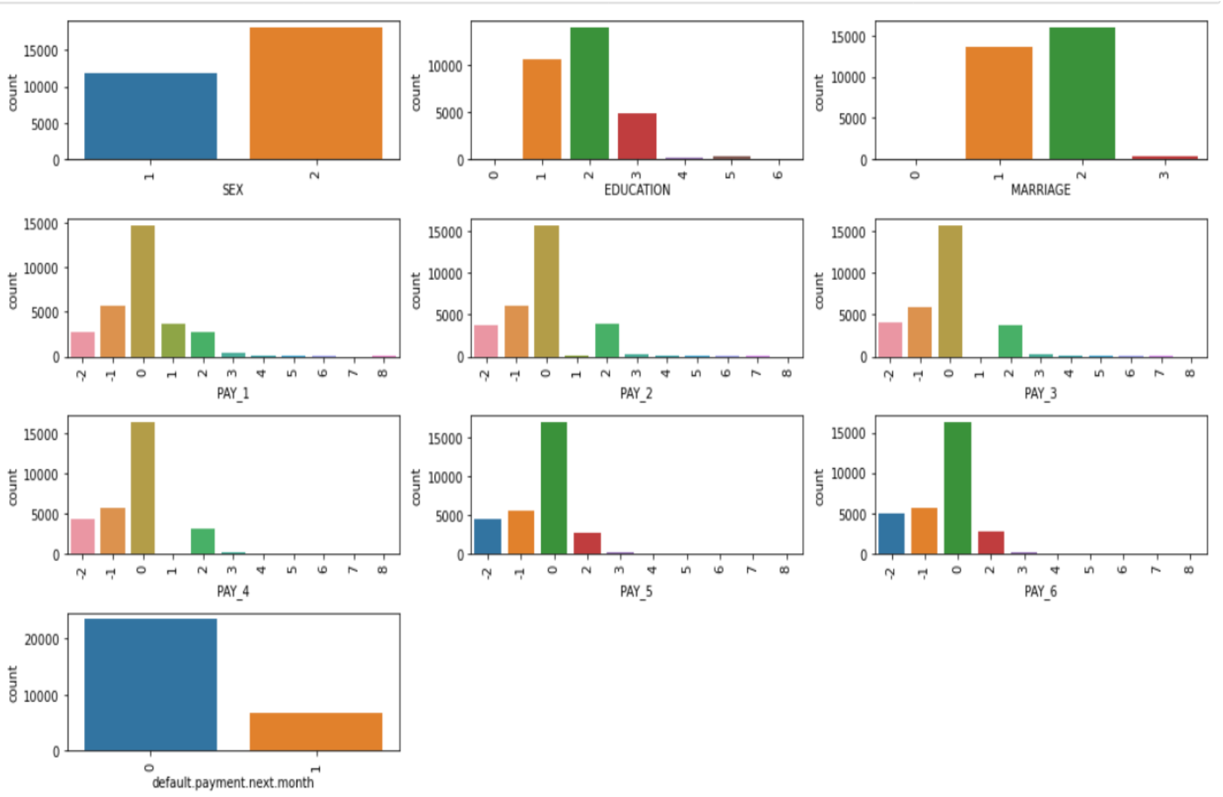
1. LIMIT_BAL column is right skewed and the boxplot shows the presence of outliers.

2. AGE column shows slight skewness on the right side. Fewer outliers are present when compared to other columns.

3. BILL_AMTX columns are highly right skewed. Most of the observations are concentrated around 0. Boxplots show the presence of numerous outliers.

4. The PAT_AMTX columns show the highest skewness amongst all the columns. Most of the values are 0 in this column. This results in right skewness and the boxplots show the presence of outliers.

Categorical Variables:-



<u>Inferences:</u>

1. The PAY_X columns have maximum records under category 0, which refers to the use of revovling credit. Most observations are present in the -2, -1 and 0 categories, indicating that most customers are either inactive (-2), have paid their bill duly (-1) or make use of revolving credit (0). Hence, delayed payments are less frequent.

2. There are less male customers (1) than female customers (2).

3. Most of the customers have university level of education (2) followed by graduate level (1) and highschool level (3). The others (4) and unknown (5,6,0) categories have the least number of customers.

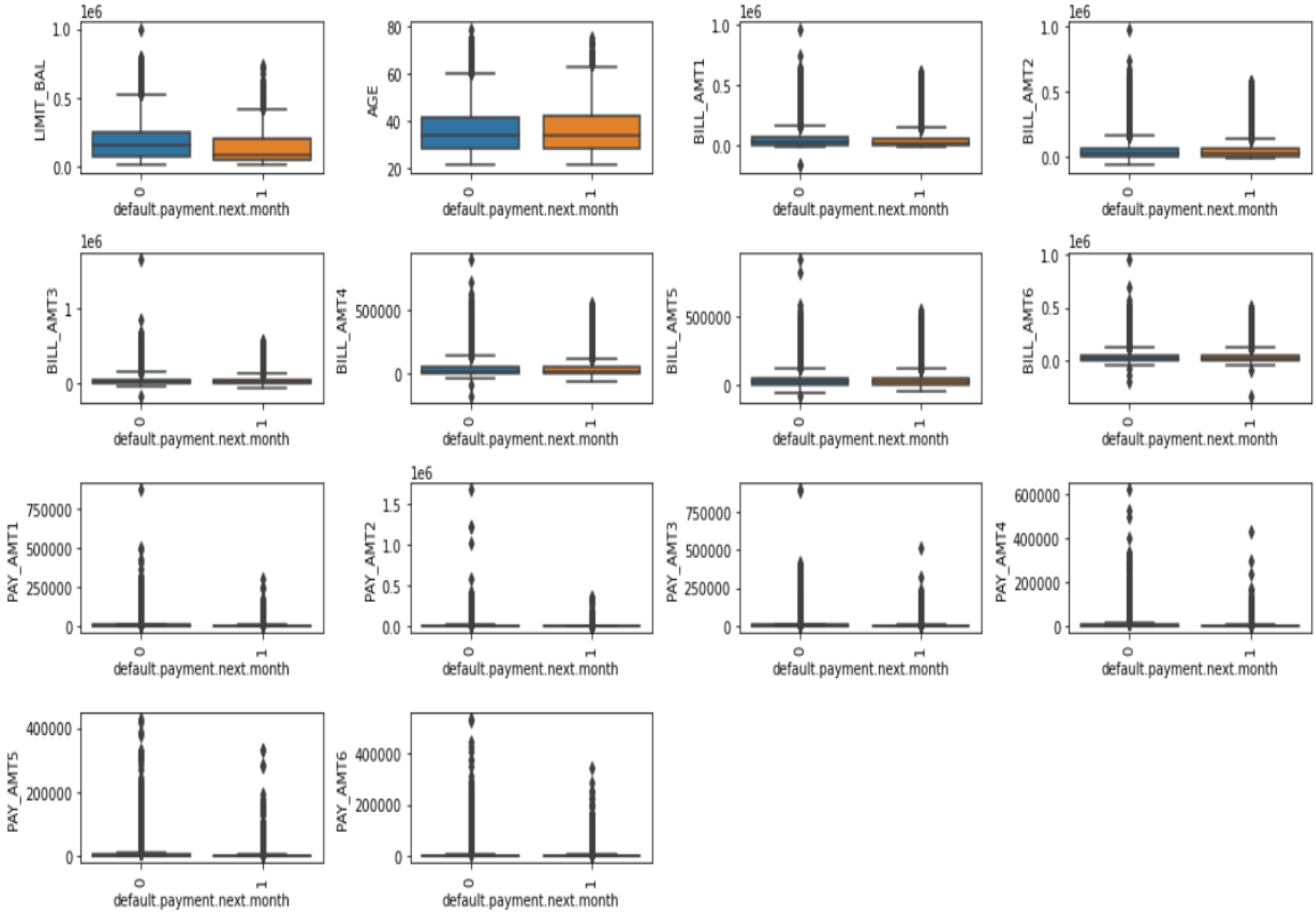
4. Most customers are single (2) followed by married (1).

5. The target variable have more non-defaulters than defaulters.

Bivariate Analysis:

We carried out bivariate analysis between all the columns and the target first. Boxplots were made between the numerical columns and the target variable and count plots were made for the categorical variables with the target as our hue.

Numerical columns vs Target:-



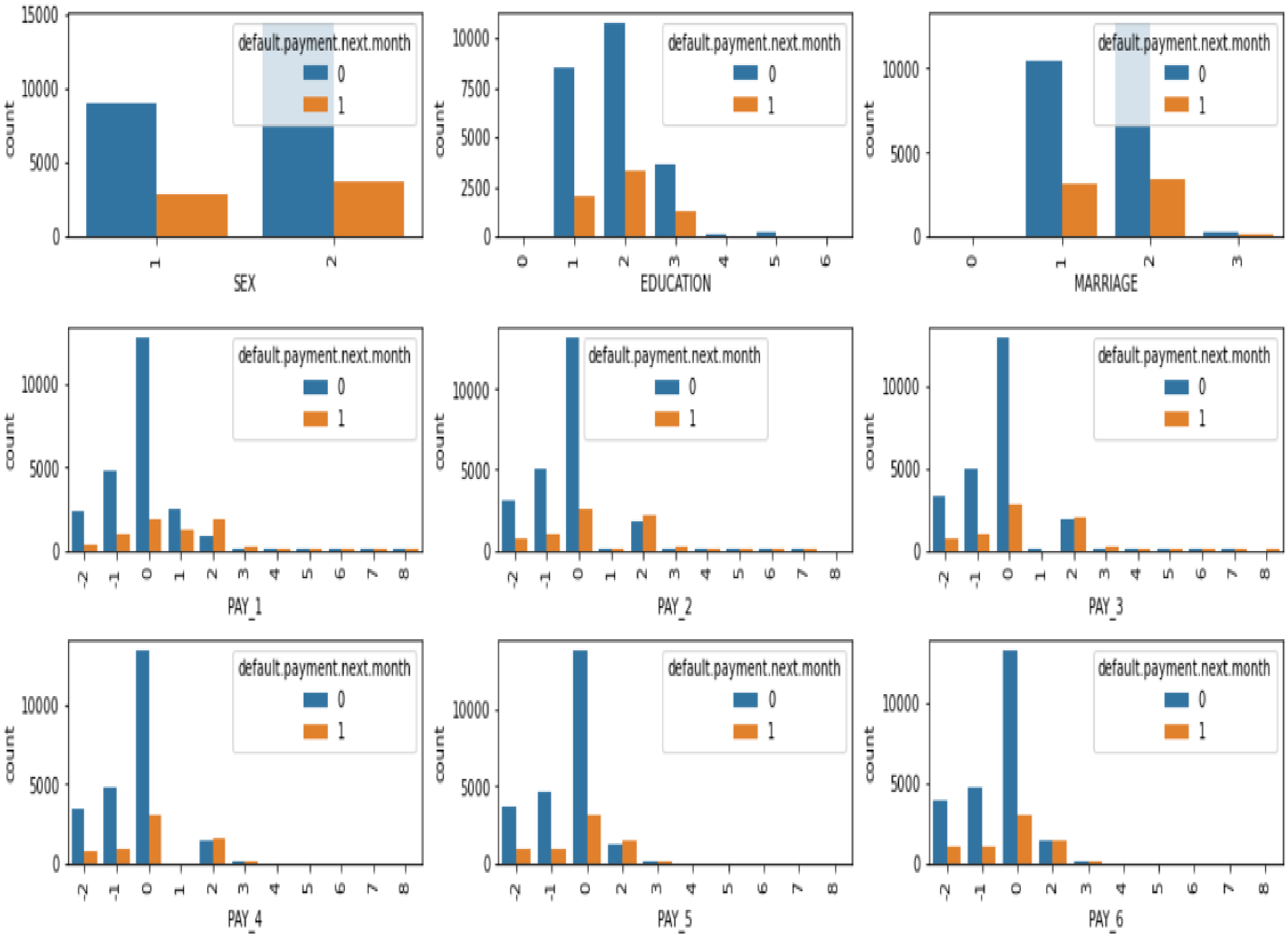
Inferences:

1. The LIMIT_BAL and AGE columns do not show much variance in either category, so we can infer that the target is not significantly influenced by these two columns.

2. The BILL_AMTX columns have slightly higher values for the non-defaulter category.

3. The PAY_AMTX columns have higher values for non-defaulter categories than defaulter categories, which is logical.

Categorical columns vs the target:-



Inferences:

1. The majority of non-defaulters in the PAY_X columns fall into the -2, -1, or 0 categories, indicating that they are either inactive customers (category -2), pay their bills on time (-1) or use revolving credit (0). Hence, they usually do not have payment delays.

2. The defaulters are almost equally distributed between both genders. However, female non-defaulters outnumber male non-defaulters.

3. Most of the people with higher educational qualifications (university and graduate levels) are non-defaulters. However, most defaulters had a university level education.

4. The defaulters are almost equally distributed among the married and single categories.

We also carried out bivariate analysis for each column with every other column in the dataset and arrived at the following important inferences:

1. In the plots between LIMIT_BAL and PAY_X columns, we see that the limit balance slightly decreases as the payment delay increases.

2. In the plots between BILL_AMTX and LIMIT_BAL, although there is not a proper linear relationship, we see that the limit balance is higher when the bill amount is higher.

3. In the plots between AGE and MARRIAGE, we see that customers who are younger are usually single and older customers belong to married or others categories.

4. In the plots between BILL_AMTX and PAY_X columns, we see that the bill amount is highest for customers using revolving credit (category 0 under PAY_X)

5. In the plots between the different BILL_AMTX columns, we see a slight linear relationship. This implies that there is some consistency in the bill amounts for customers over the 6 months.

6. In the plots between PAY_AMTX and PAY_X, we see that the pay amount is higher in the categories -1 and 0. Hence, payment delays result in lower pay amounts.

7. In plots between PAY_AMTX and EDUCATION, we see that the payment amount is slightly higher in graduate, university and high school levels.

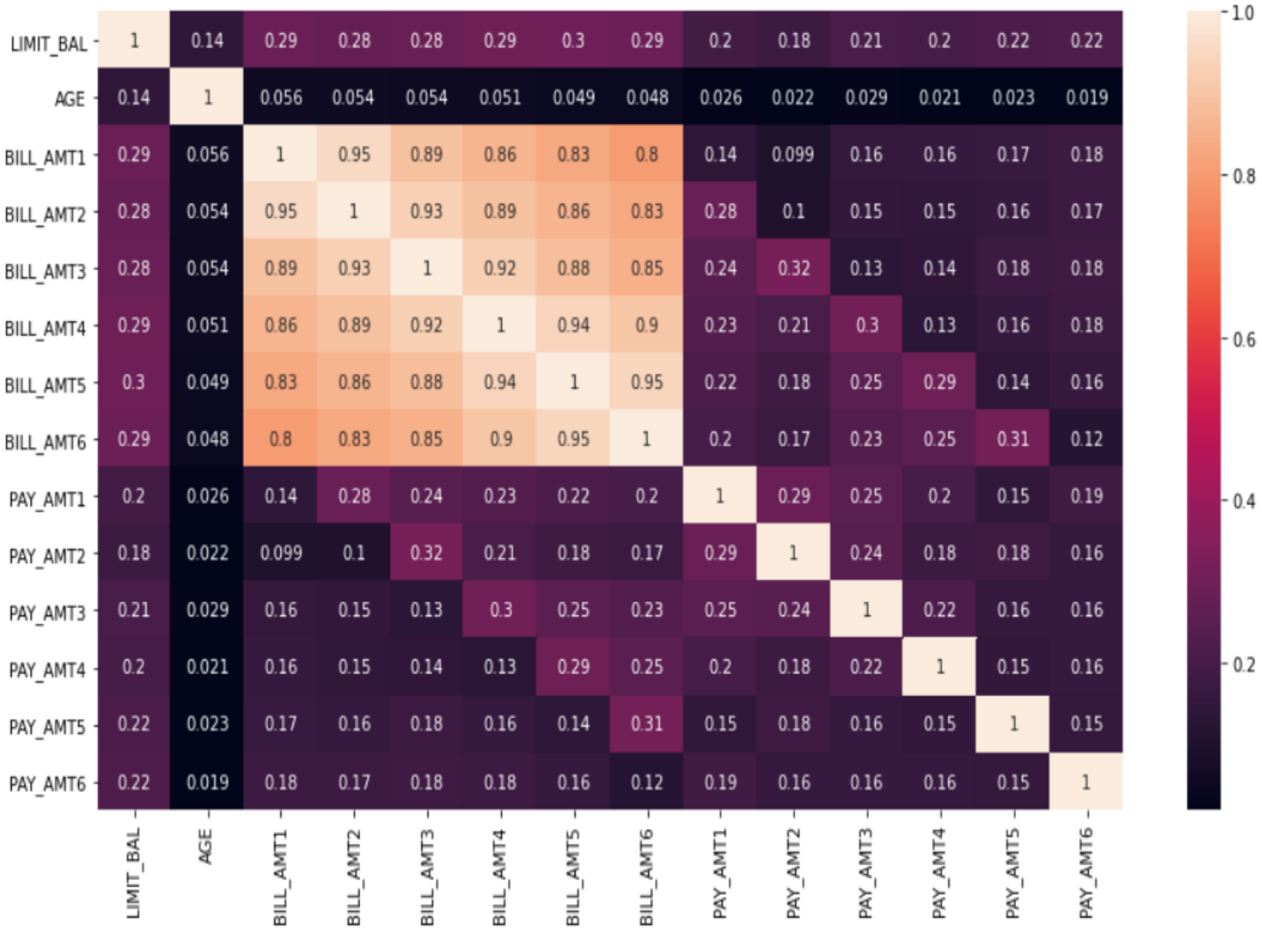
8. In the plots for PAY_X columns with sex as hue, we see that females are slightly less likely to delay the payments than males

9. In the plots for PAY_X columns with education as hue, we see that graduates are most likely to pay their bills on time and customers with university level education are the largest users of revolving credit

10. In the plots for PAY_X columns with marriage as a hue, we see that single people are the largest users of revolving credit.

Multivariate Analysis

For multivariate analysis, we first plotted a heatmap for correlation to understand how the columns are related to each other.



	LIMIT_BAL	AGE	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
LIMIT_BAL	1	0.14	0.29	0.28	0.28	0.29	0.3	0.29	0.2	0.18	0.21	0.2	0.22	0.22
AGE	0.14	1	0.056	0.054	0.054	0.051	0.049	0.048	0.026	0.022	0.029	0.021	0.023	0.019
BILL_AMT1	0.29	0.056	1	0.95	0.89	0.86	0.83	0.8	0.14	0.099	0.16	0.16	0.17	0.18
BILL_AMT2	0.28	0.054	0.95	1	0.93	0.89	0.86	0.83	0.28	0.1	0.15	0.15	0.16	0.17
BILL_AMT3	0.28	0.054	0.89	0.93	1	0.92	0.88	0.85	0.24	0.32	0.13	0.14	0.18	0.18
BILL_AMT4	0.29	0.051	0.86	0.89	0.92	1	0.94	0.9	0.23	0.21	0.3	0.13	0.16	0.18
BILL_AMT5	0.3	0.049	0.83	0.86	0.88	0.94	1	0.95	0.22	0.18	0.25	0.29	0.14	0.16
BILL_AMT6	0.29	0.048	0.8	0.83	0.85	0.9	0.95	1	0.2	0.17	0.23	0.25	0.31	0.12
PAY_AMT1	0.2	0.026	0.14	0.28	0.24	0.23	0.22	0.2	1	0.29	0.25	0.2	0.15	0.19
PAY_AMT2	0.18	0.022	0.099	0.1	0.32	0.21	0.18	0.17	0.29	1	0.24	0.18	0.18	0.16
PAY_AMT3	0.21	0.029	0.16	0.15	0.13	0.3	0.25	0.23	0.25	0.24	1	0.22	0.16	0.16
PAY_AMT4	0.2	0.021	0.16	0.15	0.14	0.13	0.29	0.25	0.2	0.18	0.22	1	0.15	0.16
PAY_AMT5	0.22	0.023	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1	0.15
PAY_AMT6	0.22	0.019	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1

1. BILL_AMTX columns are showing high correlation amongst themselves.
2. The other columns are not showing high correlation with other columns of the dataset.

Based on the results of the bivariate analysis, we performed multivariate analysis among different columns. We arrived at the following inferences:

1. In the plots between LIMIT_BAL and PAY_X columns with target as hue, we found that the limit balance slightly decreases for defaulters and as payment delay increases.

2. In the plots between BILL_AMTX and PAY_X columns with target as hue, we found that the bill amount increases for defaulters as the payment delay increases.

3. In the plots between PAY_X and PAY_AMTX columns with target as hue, we found that the payment amount in the -2, -1 and 0 categories of PAY_X columns was higher for non-defaulters.

4. In the plots between EDUCATION vs PAY_AMTX columns with target as hue, we found that most of the non-defaulters with an education level of graduate and university have higher payment amounts.

5. In the plots between LIMIT_BAL vs PAY_X with SEX as hue, we found that the limit balance reduces more for males with delay in payment as compared to the females

6. In the plots between LIMIT_BAL vs PAY_X with EDUCATION as hue, we found that the limit balance for graduates was higher while that for other educational qualifications reduces as the payment delay increases.

7. In the plots between LIMIT_BAL vs PAY_X with MARRIAGE as hue, we found that the limit balance for married customers was higher while that for other categories reduces as the payment delay increases.

Transformation:

As our data consists of many outliers, we used transformation techniques in order to improve the distribution of the columns. Our data consists of negative values; hence, we made use of the Yeo-Johnson transformation, which is compatible with negative values.

The transformation improved the distribution of the LIMIT_BAL, AGE, and PAY_AMTX columns. However, the BILL_AMTX columns did not show any significant improvements.
To improve the distribution of these columns, we tried a different transformation: box-cox. However, this method did not affect the distribution of these columns either.

Statistical Analysis:

H0: The variable is insignificant (no relationship between variable and target)
H1: The variable is significant (there is a relationship between variable and target)
We will carry out tests for each column w.r.t. the target column.

Categorical columns:

When categorical columns need to be tested for significance when the target is categorical, we will do a chi-squared contingency test (for sex, education, marriage, and PAY_X columns).

Obtained p-values

```
{'SEX': 4.472804335813843e-12,
 'EDUCATION': 1.2332626245415605e-32,
 'MARRIAGE': 8.825862457577375e-08,
 'PAY_1': 0.0,
 'PAY_2': 0.0,
 'PAY_3': 0.0,
 'PAY_4': 0.0,
 'PAY_5': 0.0,
 'PAY_6': 0.0}
```

Inference: p-values for all these columns are less than 0.05. Hence, we reject the null hypothesis, concluding that all these columns are significant.

Numerical columns:

When numerical columns need to be tested for significance when target is categorical, we will do a two sample mean tests (for LIMIT_BAL, AGE, BILL_AMTX and PAY_AMTX columns).
The two categories of the target are large samples; hence we can do z test.

```
0    23364
1     6636
```

Obtained p-values:

```
{'LIMIT_BAL': 1.7456891860689276e-159,
 'AGE': 0.016130835407188793,
 'BILL_AMT1': 0.000666458894348233,
 'BILL_AMT2': 0.013951754044408744,
 'BILL_AMT3': 0.01476421905131808,
 'BILL_AMT4': 0.07854544833719244,
 'BILL_AMT5': 0.24162514177412497,
 'BILL_AMT6': 0.3521150333737273,
 'PAY_AMT1': 9.234672141977337e-37,
 'PAY_AMT2': 2.8928830381533725e-24,
 'PAY_AMT3': 1.705316820236147e-22,
 'PAY_AMT4': 6.304758938546825e-23,
 'PAY_AMT5': 1.1562201413814998e-21,
 'PAY_AMT6': 2.8522640944438073e-20}
```

Inference: The p-values for BILL_AMT4, BILL_AMT5, and BILL_AMT6 are greater than 0.05. Hence, we fail to reject the null hypothesis. Hence, these columns are statistically insignificant.

Encoding

We tried two encoding methods - dummy encoding and label encoding for the categorical columns.

Train-Test Split

We carried out Train-Test Split for our data

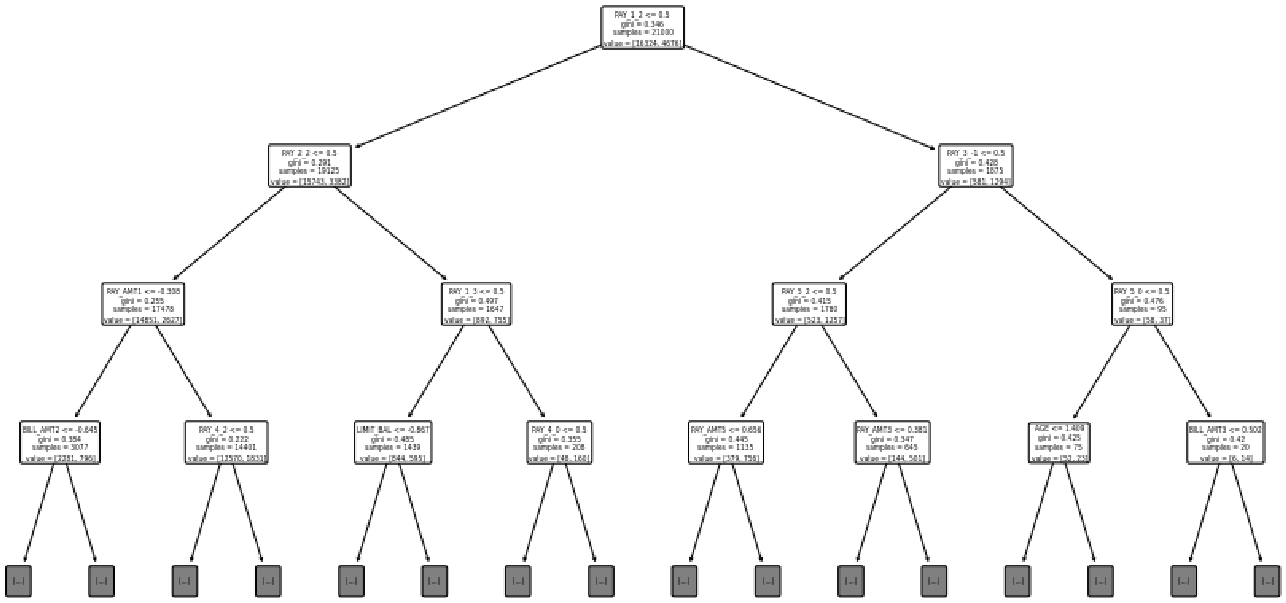
Scaling

We carried out standard scaling for our data set

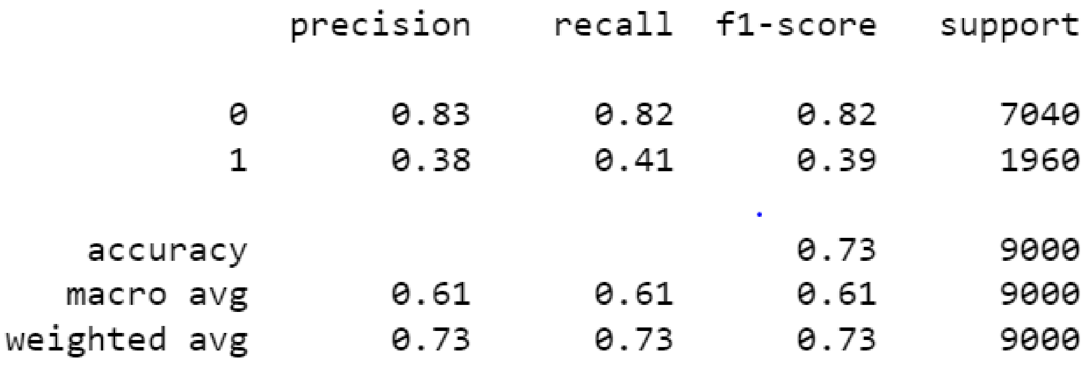
BASE MODEL

We built four models: one with the original columns and dummy encoding, one with the original columns and label encoding, one with transformed columns and dummy, and one with the transformed columns and label encoding. We will use a decision tree for our base model.

Model with original columns and dummy encoding:

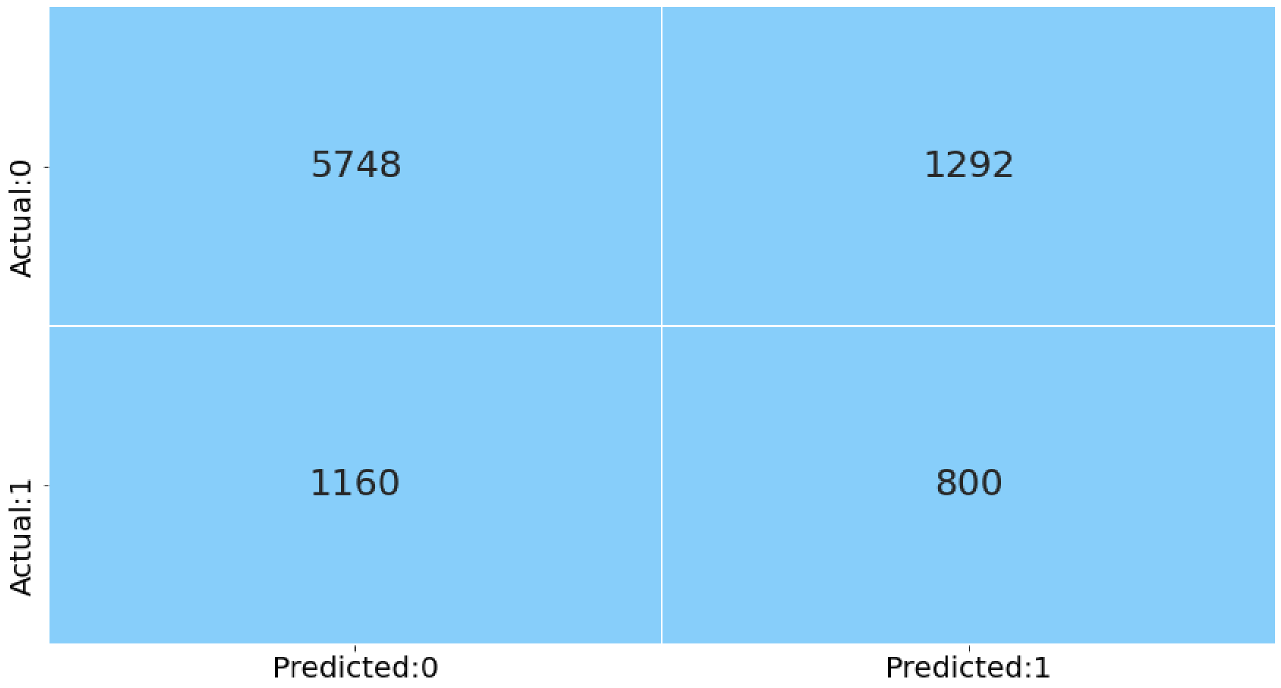


Metrics: Classification report



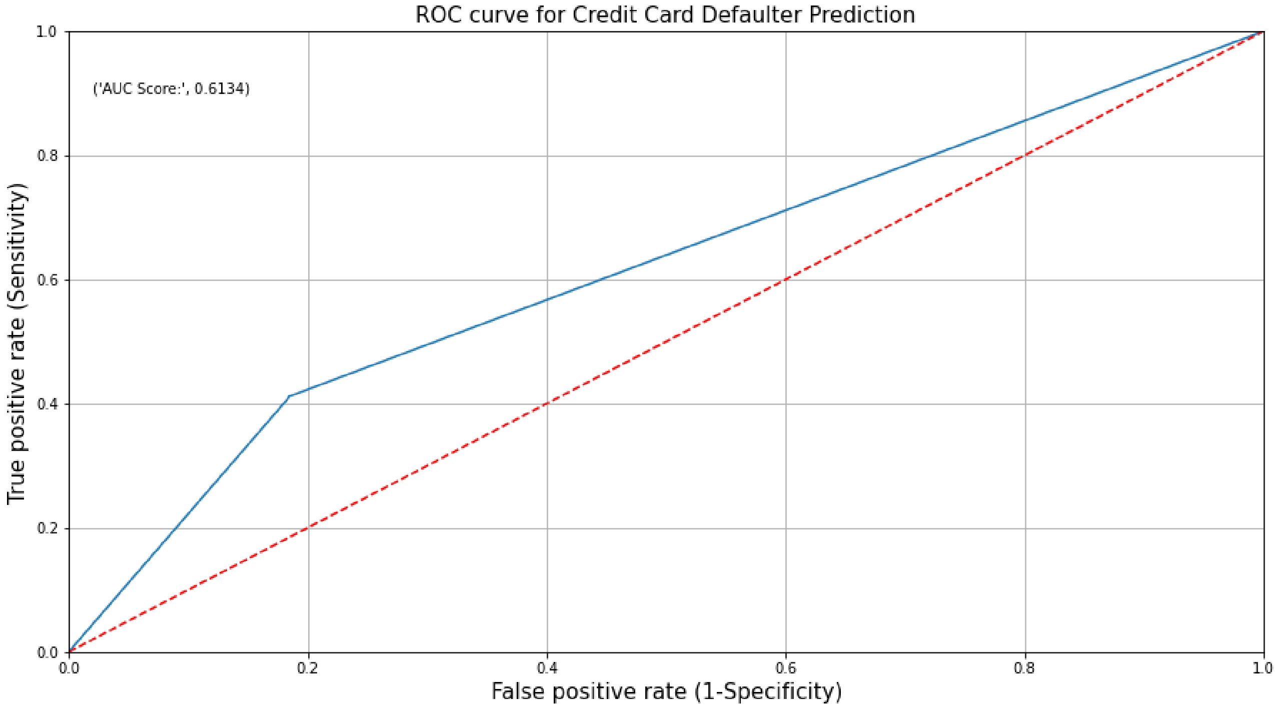
```
              precision    recall  f1-score   support

           0       0.83      0.82      0.82      7040
           1       0.38      0.41      0.39      1960

    accuracy                           0.73      9000
   macro avg       0.61      0.61      0.61      9000
weighted avg       0.73      0.73      0.73      9000
```

Inference: The model has low accuracy (0.73). The f1-score for the positive class (1) is low, indicating that the majority of the positive class is not being correctly predicted. This performance can be improved by tuning the parameters for building the model or trying other algorithms.



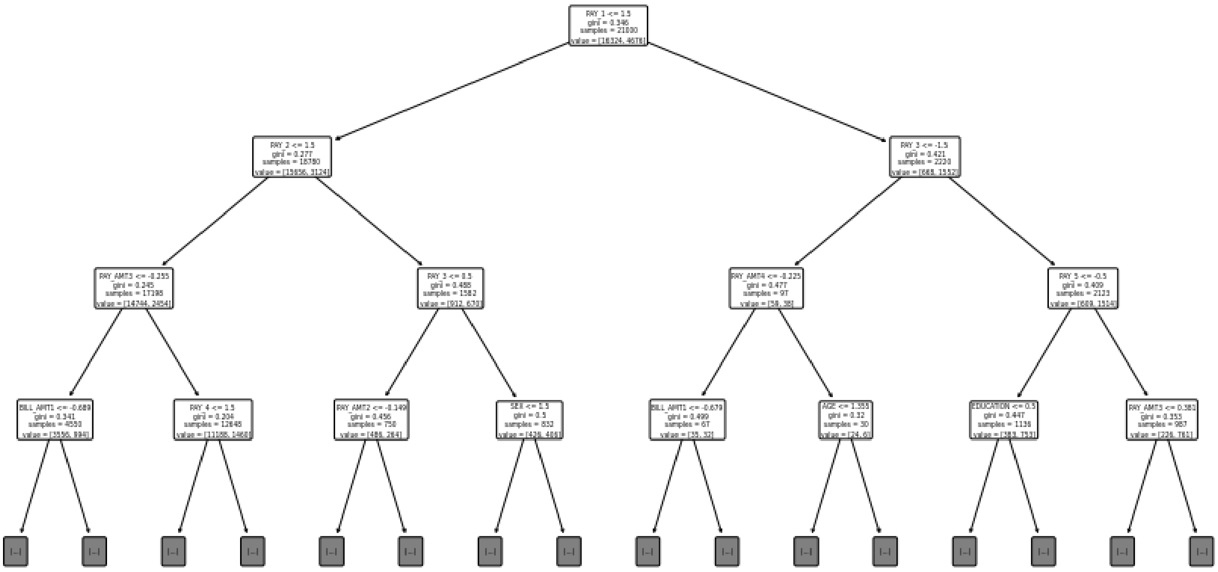
	Predicted:0	Predicted:1
Actual:0	5748	1292
Actual:1	1160	800

We can see in the above confusion matrix that the number of incorrect predictions is very high, especially for the positive class.



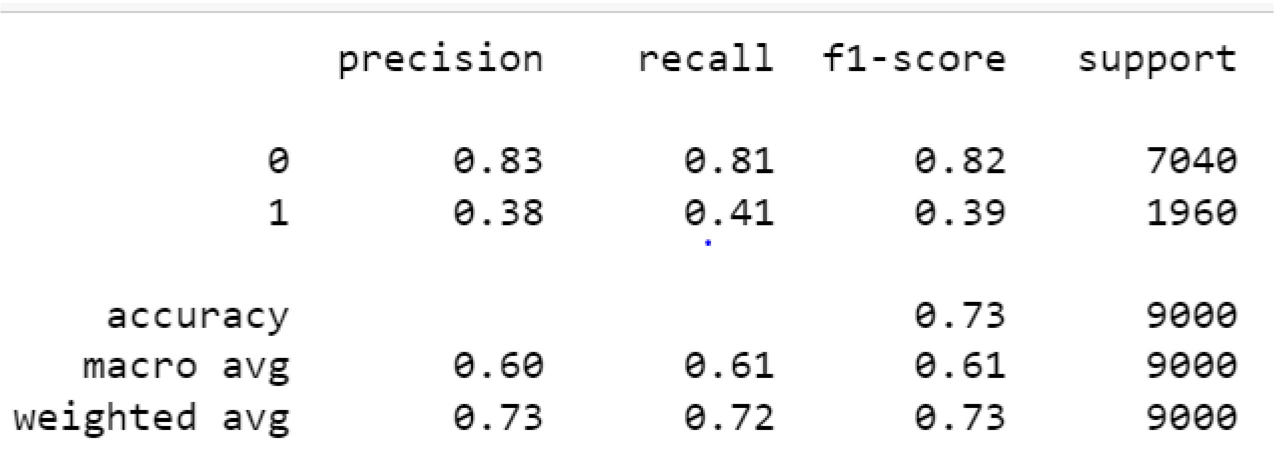
The roc_auc_score for the base model is only 0.6234, which is significantly low. Hence, we will have to work on our data and use different algorithms to improve the performance.

<u>Model with original columns and label encoding:</u>

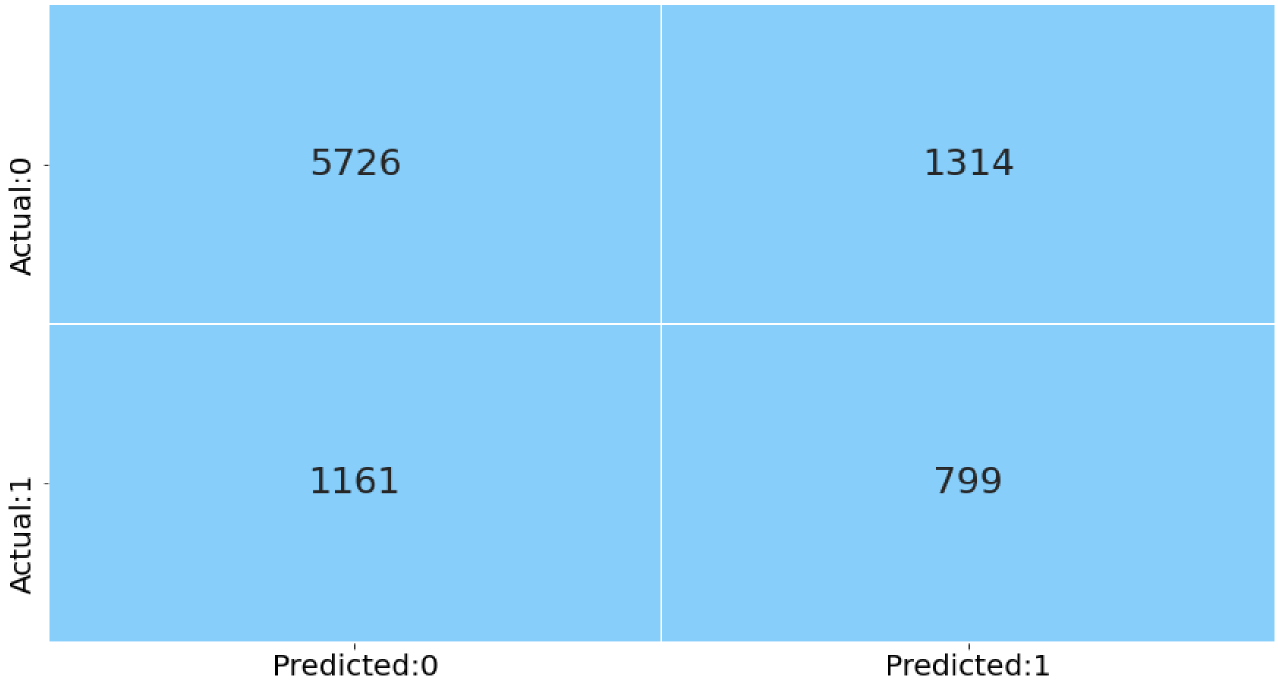


<u>Metrics:</u>

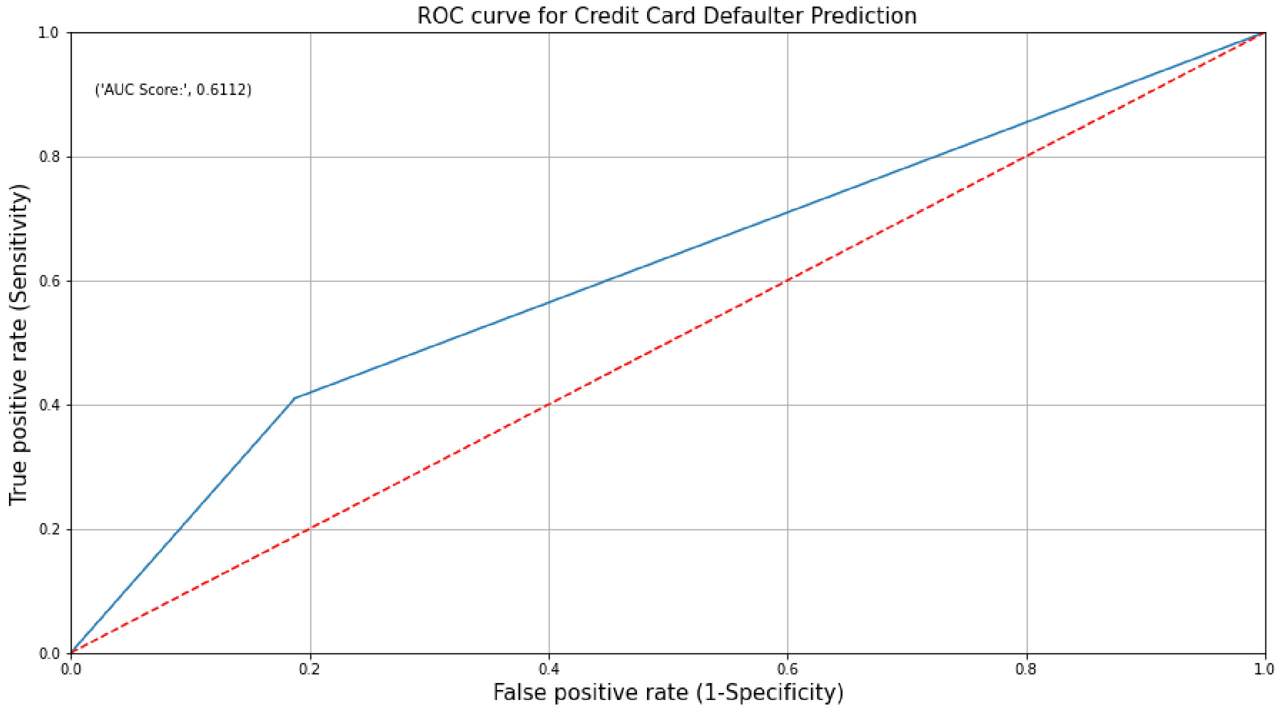
<u>Classification report</u> : It is showing the same results as with dummy encoding.



```
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      7040
           1       0.38      0.41      0.39      1960

    accuracy                           0.73      9000
   macro avg       0.60      0.61      0.61      9000
weighted avg       0.73      0.72      0.73      9000
```

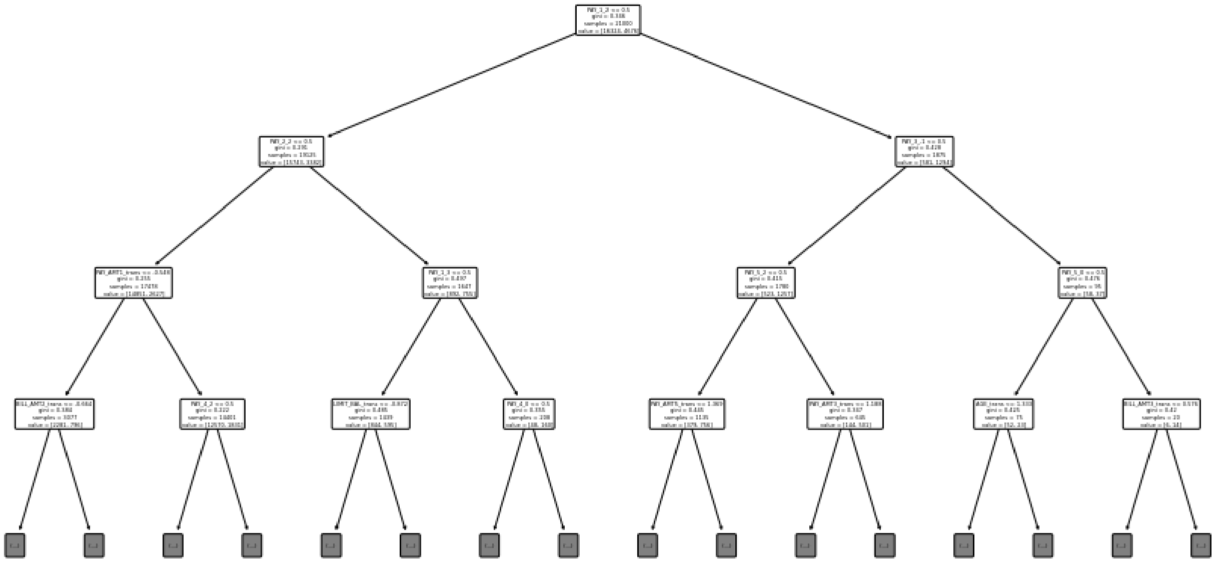


	Predicted:0	Predicted:1
Actual:0	5726	1314
Actual:1	1161	799

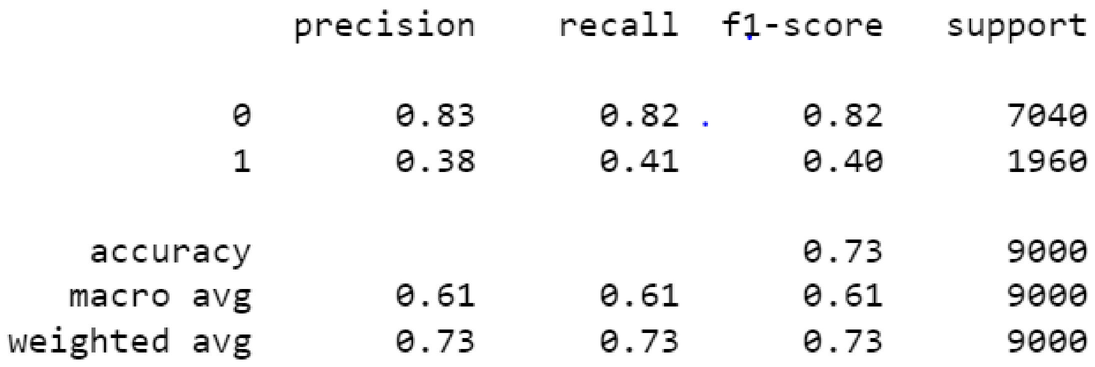


We are getting the same results as dummy encoding. Hence, type of encoding is not influencing the model performance.

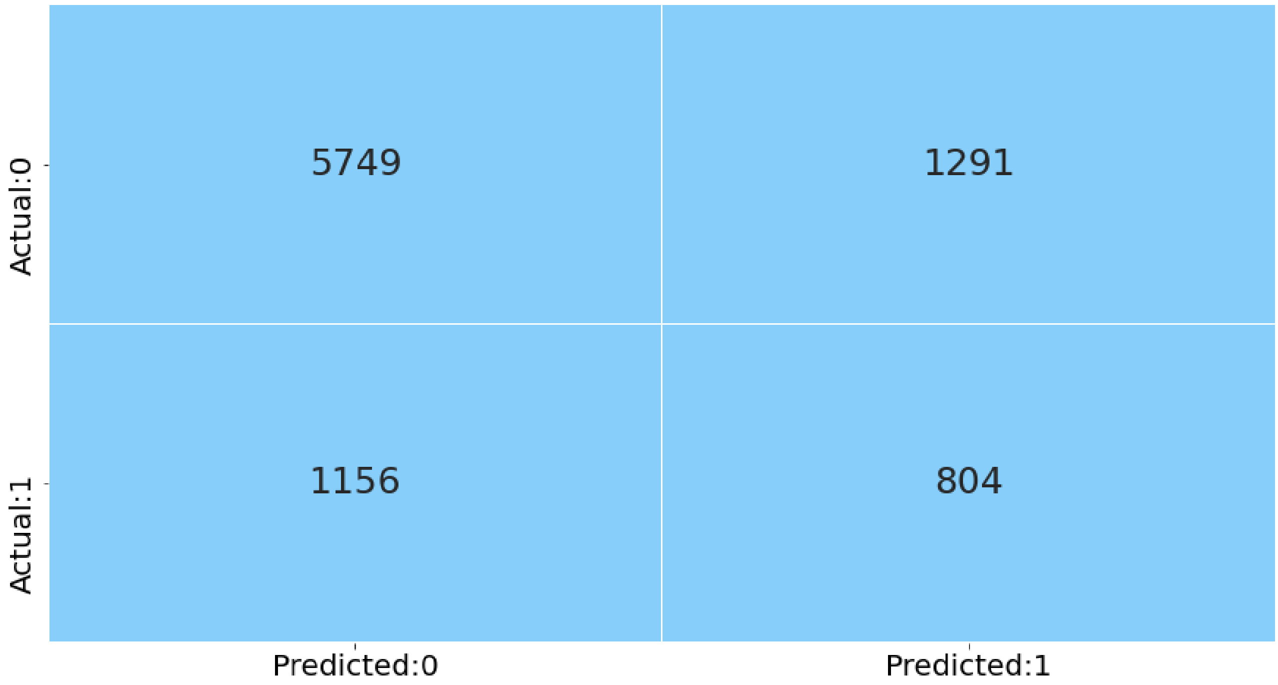
Model with transformed columns and dummy encoding:



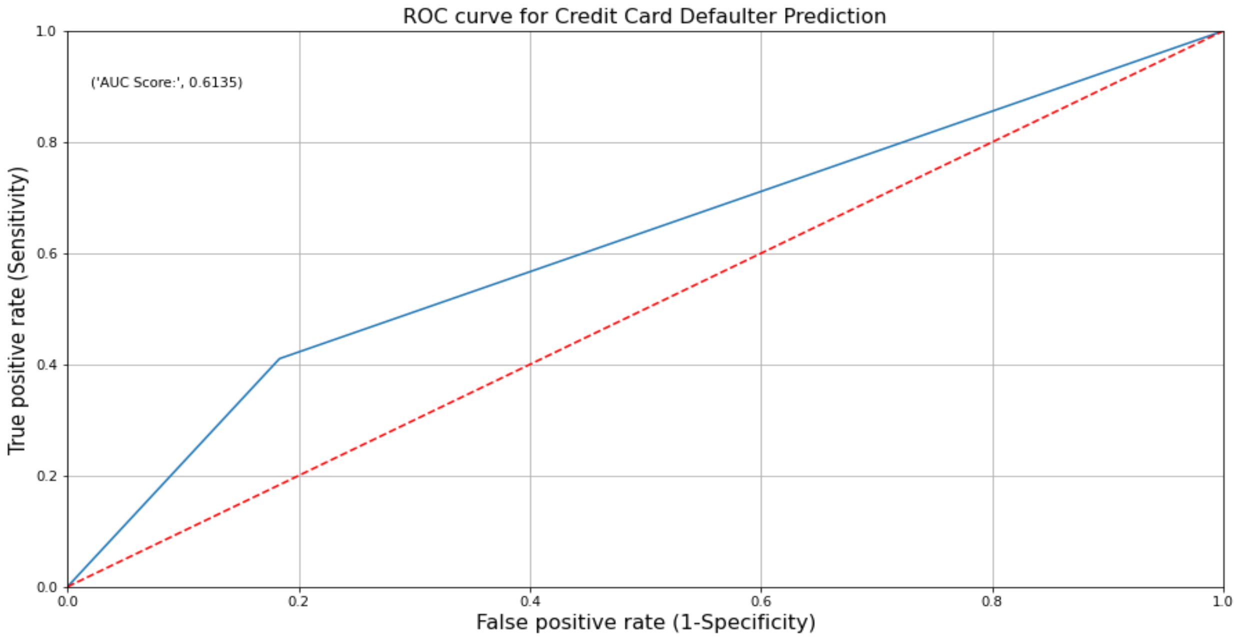
Metrics: **Classification report:-** This model is giving the same results as the previous models.



```
              precision    recall  f1-score   support

           0       0.83      0.82      0.82      7040
           1       0.38      0.41      0.40      1960

    accuracy                           0.73      9000
   macro avg       0.61      0.61      0.61      9000
weighted avg       0.73      0.73      0.73      9000
```

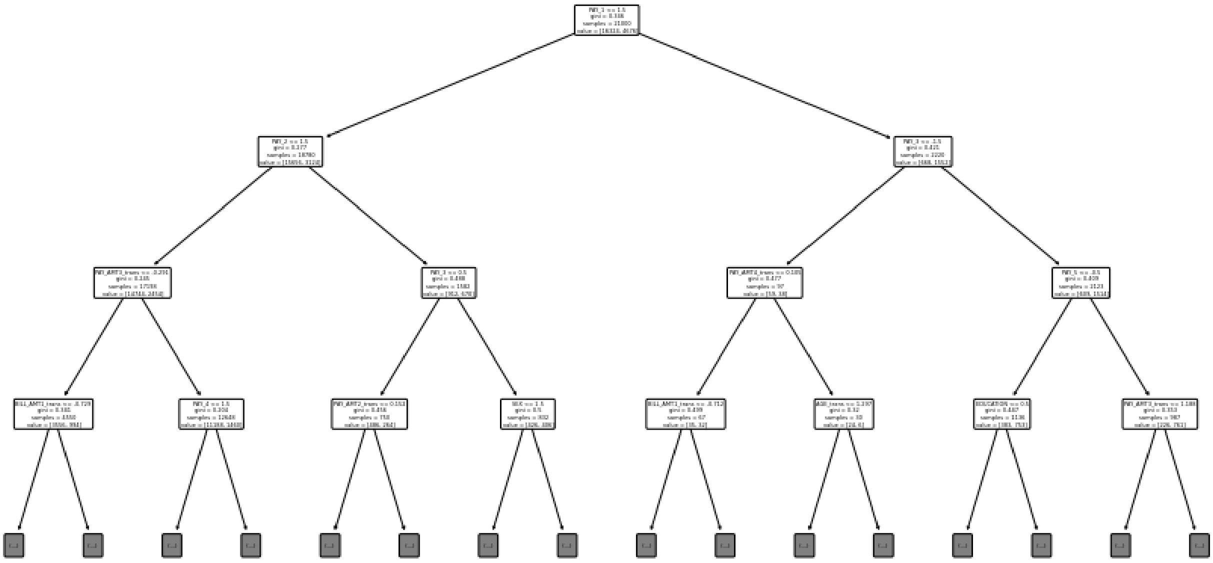


We can see in the above confusion matrix that the number of incorrect predictions is very high, especially for the positive class.

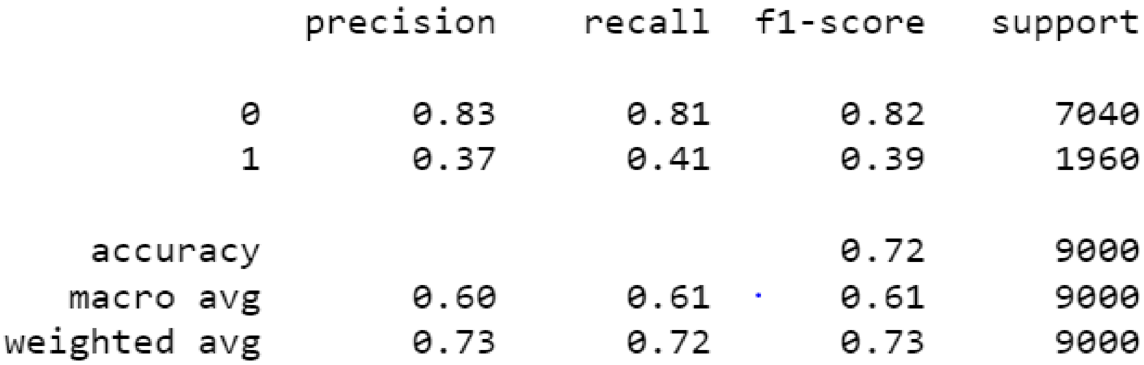


The roc_auc_score for the base model is only 0.6233, which is significantly low. Hence, we will have to work on our data and use different algorithms to improve the performance.

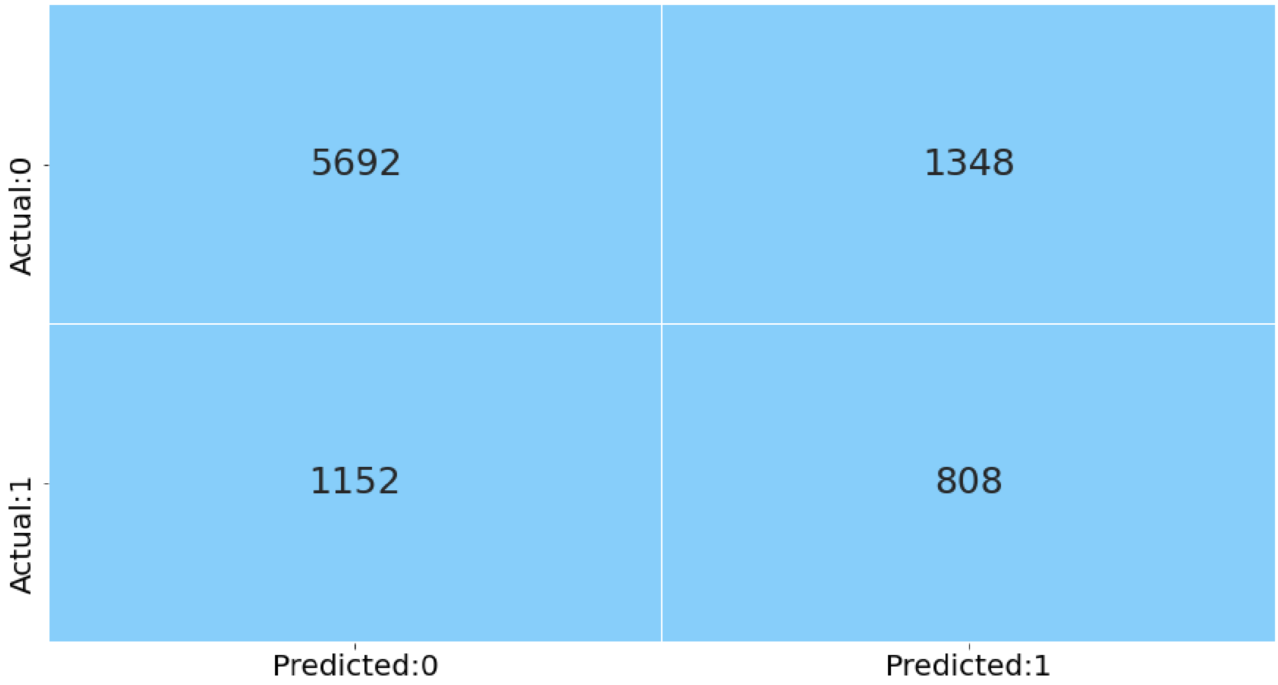
Model with transformed columns and label encoding:



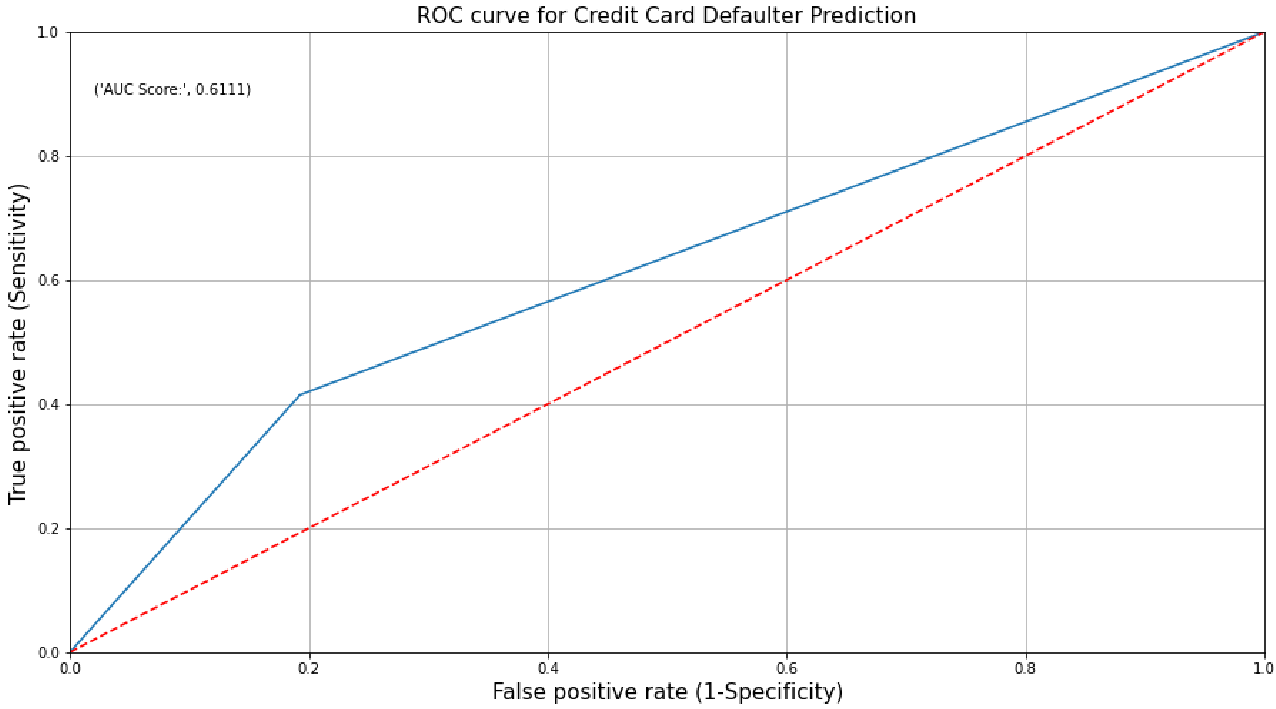
Metrics: **Classification report :-** This model is also giving the same results.



```
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      7040
           1       0.37      0.41      0.39      1960

    accuracy                           0.72      9000
   macro avg       0.60      0.61      0.61      9000
weighted avg       0.73      0.72      0.73      9000
```



	Predicted:0	Predicted:1
Actual:0	5692	1348
Actual:1	1152	808

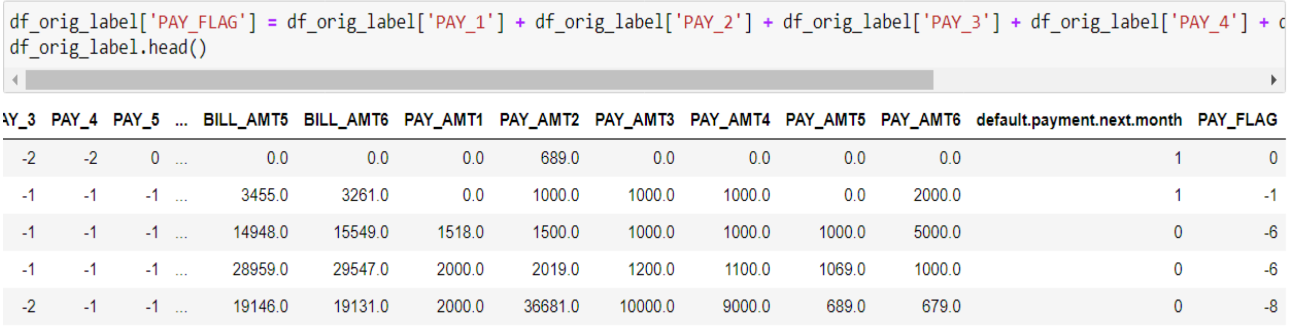


The roc_auc_score for the base model is only 0.6209, which has decreased very slightly from the previous models.

As only one model performed very slightly poorly (model with transformed columns and label encoding), we can choose any of the other three models. So, we will use original columns with label encoding for simplicity and ease of drawing inferences.

FEATURE ENGINEERING

The first feature we will form is 'PAY_FLAG' which will be the sum of all the pay flags of the different PAY_X columns for a customer.



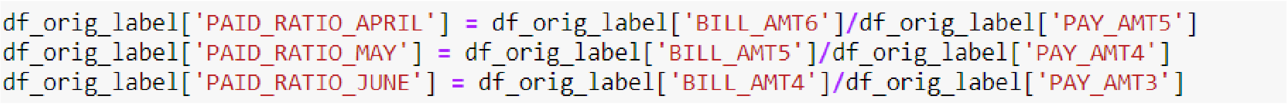
```
df_orig_label['PAY_FLAG'] = df_orig_label['PAY_1'] + df_orig_label['PAY_2'] + df_orig_label['PAY_3'] + df_orig_label['PAY_4'] + c
df_orig_label.head()
```

AY_3	PAY_4	PAY_5	...	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month	PAY_FLAG
-2	-2	0	...	0.0	0.0	0.0	689.0	0.0	0.0	0.0	0.0	1	0
-1	-1	-1	...	3455.0	3261.0	0.0	1000.0	1000.0	1000.0	0.0	2000.0	1	-1
-1	-1	-1	...	14948.0	15549.0	1518.0	1500.0	1000.0	1000.0	1000.0	5000.0	0	-6
-1	-1	-1	...	28959.0	29547.0	2000.0	2019.0	1200.0	1100.0	1069.0	1000.0	0	-6
-2	-1	-1	...	19146.0	19131.0	2000.0	36681.0	10000.0	9000.0	689.0	679.0	0	-8

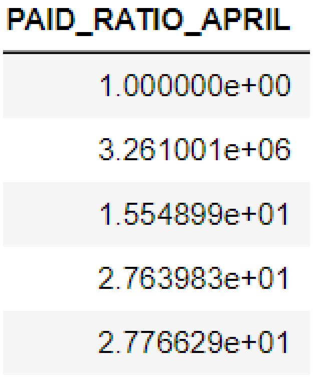
Now, we need to drop any of the PAY_X columns as the new feature PAY_FLAG is highly dependent on the PAY_X columns, dropping one of them will reduce this dependency.

The columns BILL_AMT4, BILL_AMT5, and BILL_AMT6 are insignificant according to the statistical test we performed. As a result, we can create new features based on these to see if they help improve the model's performance.

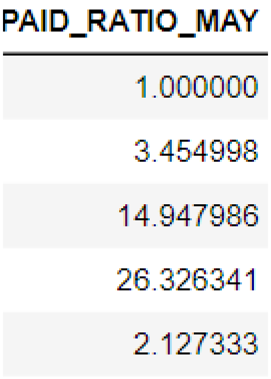
We can find a ratio between the BILL_AMT and PAY_AMT columns. BILL_AMT6 is the bill for the month of April, and PAY_AMT5 is the amount paid for this bill in May. Hence, we can find a ratio between these two columns. Similarly, we can find the ratio between BILL_AMT5 and PAY_AMT4, BILL_AMT4 and PAY_AMT3.



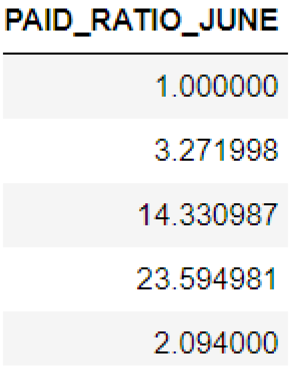
```
df_orig_label['PAID_RATIO_APRIL'] = df_orig_label['BILL_AMT6']/df_orig_label['PAY_AMT5']
df_orig_label['PAID_RATIO_MAY'] = df_orig_label['BILL_AMT5']/df_orig_label['PAY_AMT4']
df_orig_label['PAID_RATIO_JUNE'] = df_orig_label['BILL_AMT4']/df_orig_label['PAY_AMT3']
```



PAID_RATIO_APRIL
1.000000e+00
3.261001e+06
1.554899e+01
2.763983e+01
2.776629e+01

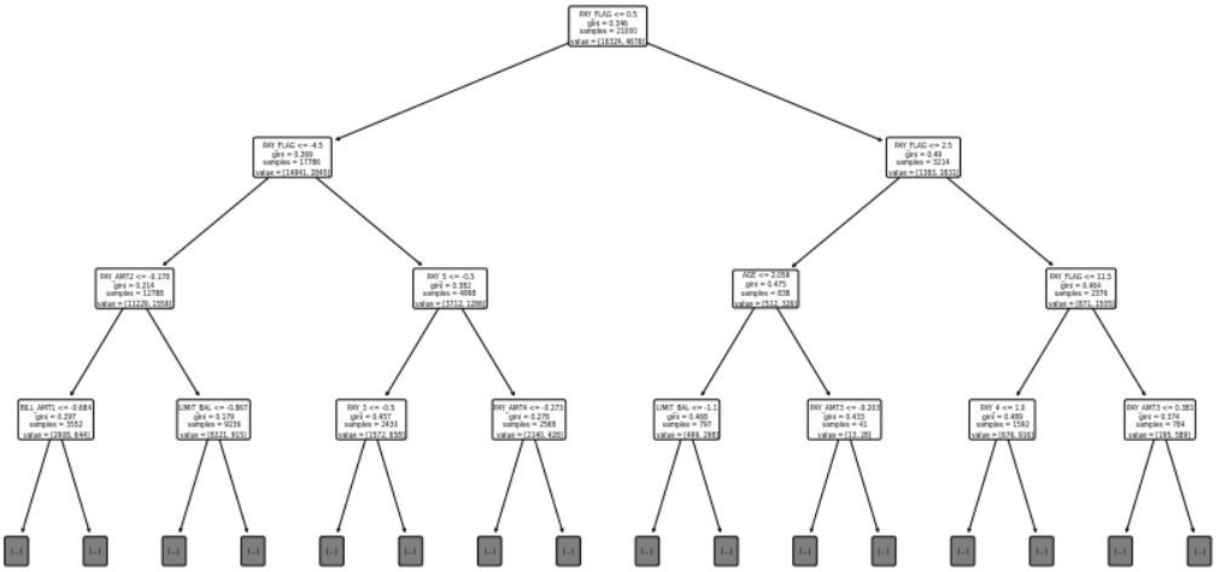


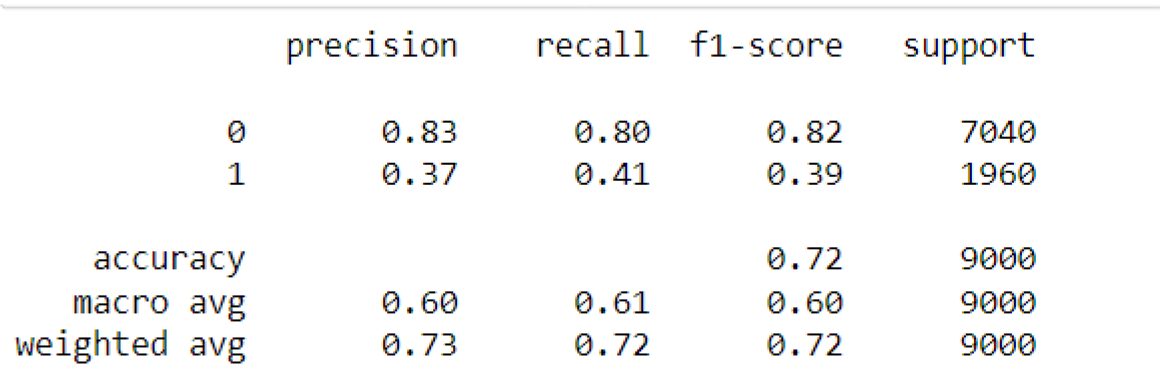
PAID_RATIO_MAY
1.000000
3.454998
14.947986
26.326341
2.127333



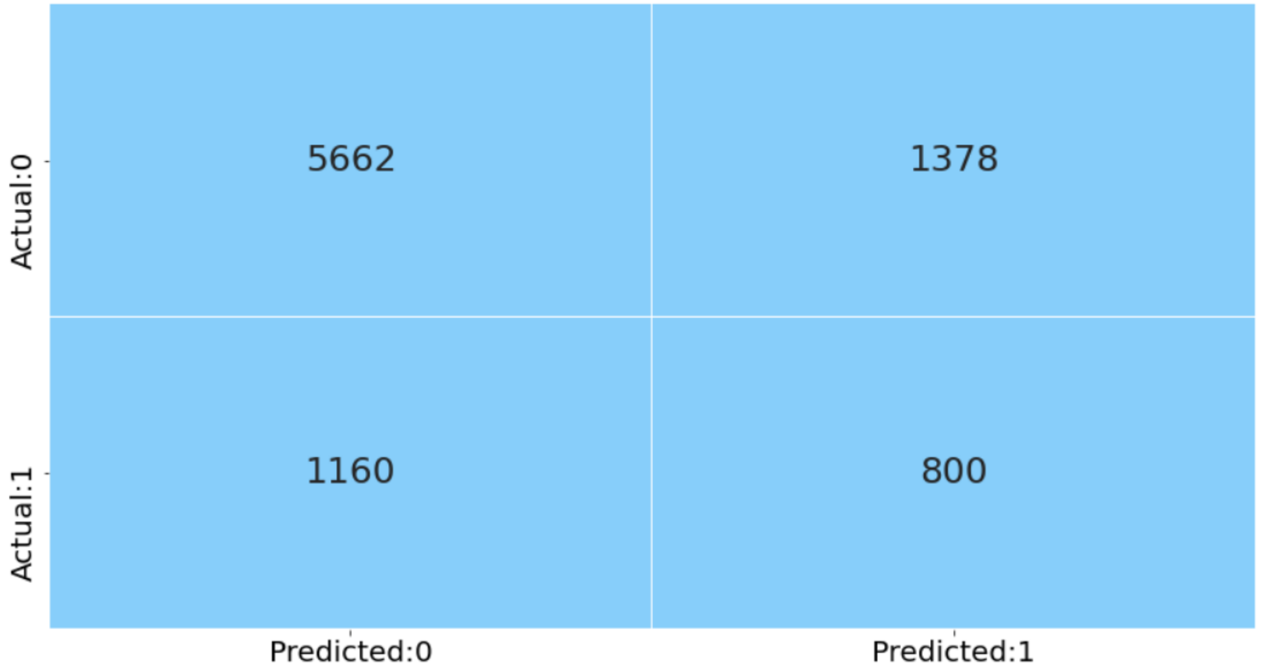
PAID_RATIO_JUNE
1.000000
3.271998
14.330987
23.594981
2.094000

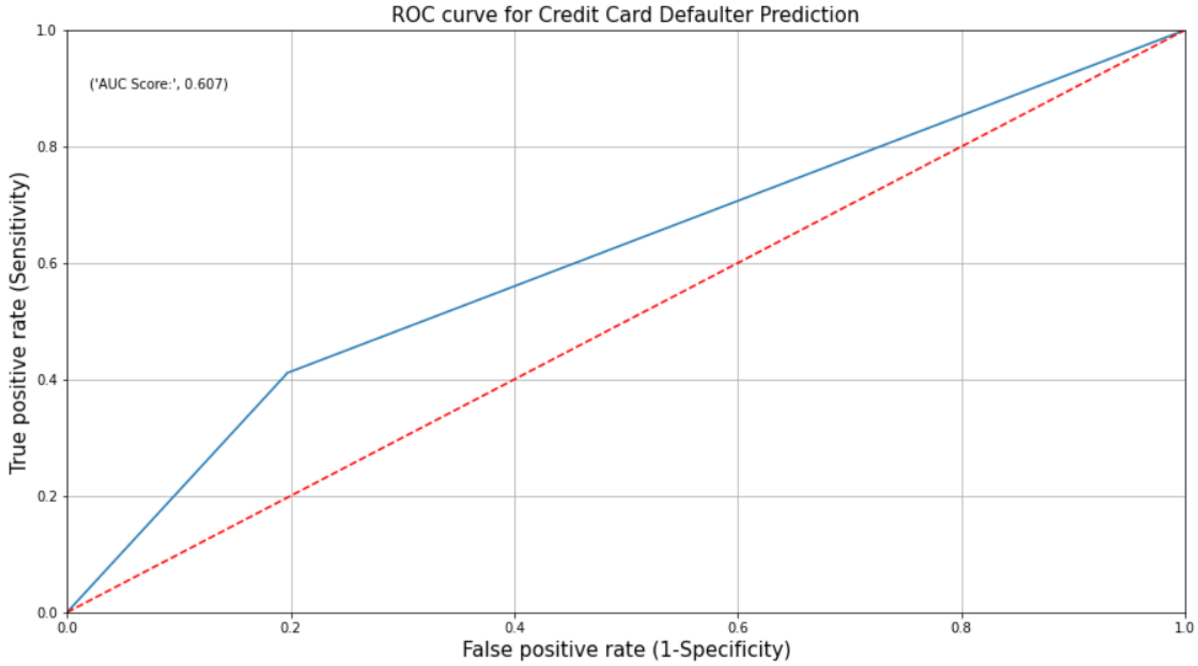
Building Initial Model on Feature Engineered Dataframe:





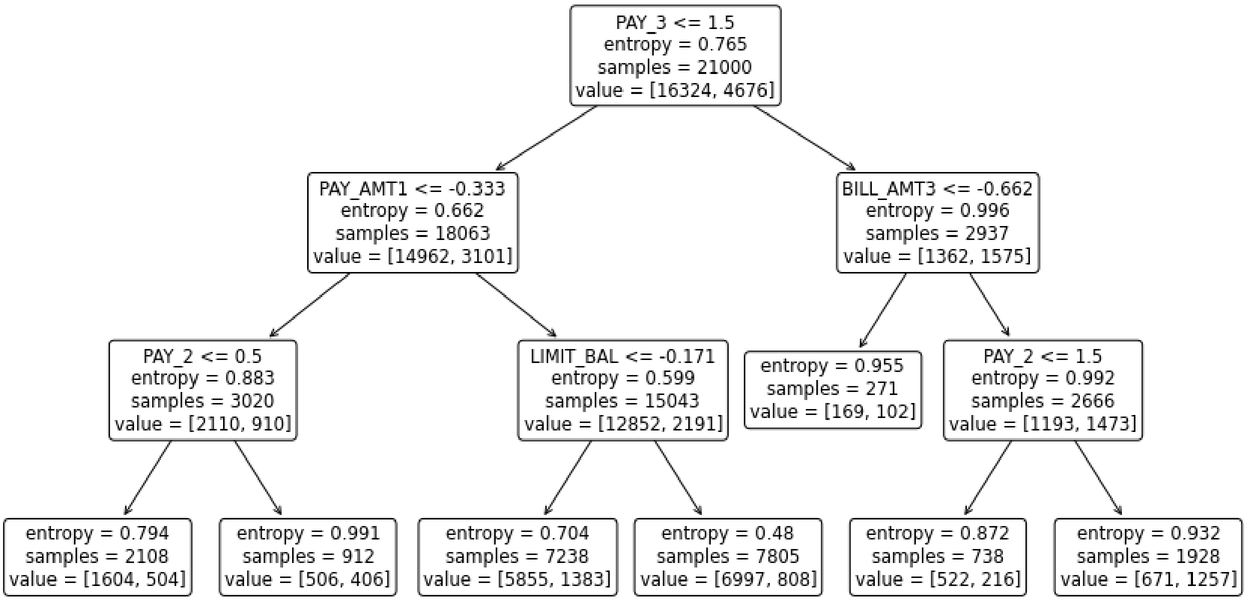
```
              precision    recall  f1-score   support

           0       0.83      0.80      0.82      7040
           1       0.37      0.41      0.39      1960

    accuracy                           0.72      9000
   macro avg       0.60      0.61      0.60      9000
weighted avg       0.73      0.72      0.72      9000
```

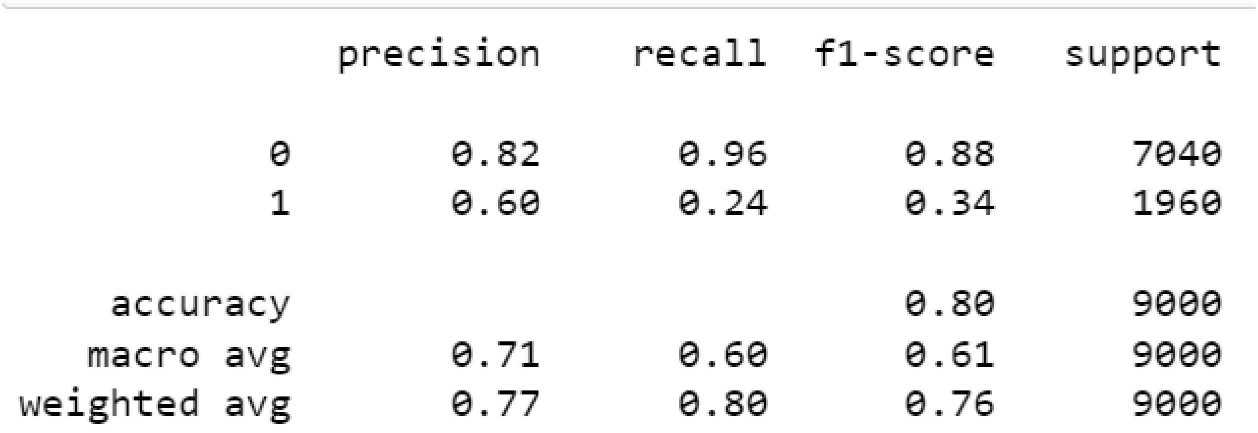




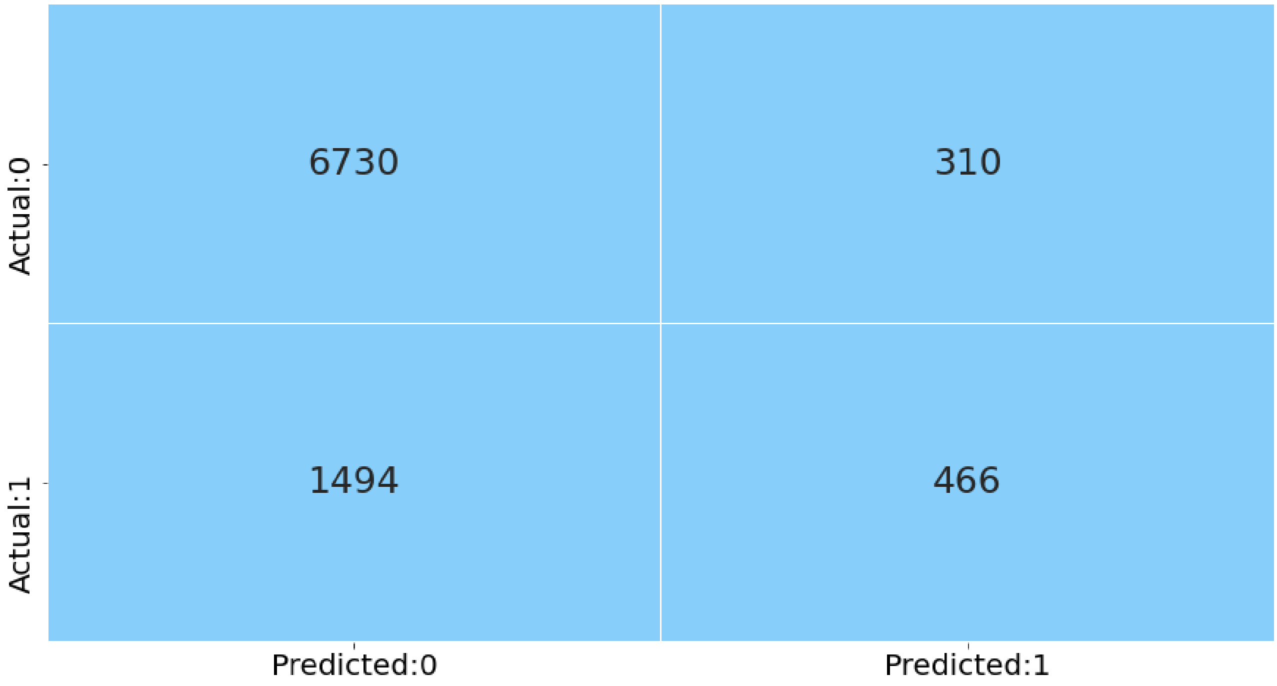
Hyperparameter Tuning Decision Tree Feature Engineered Model: -

Best parameters for decision tree classifier were found to be: {'criterion': 'entropy', 'max_depth': 3, 'max_features': 'sqrt', 'max_leaf_nodes': 7, 'min_samples_leaf': 2, 'min_samples_split': 2}

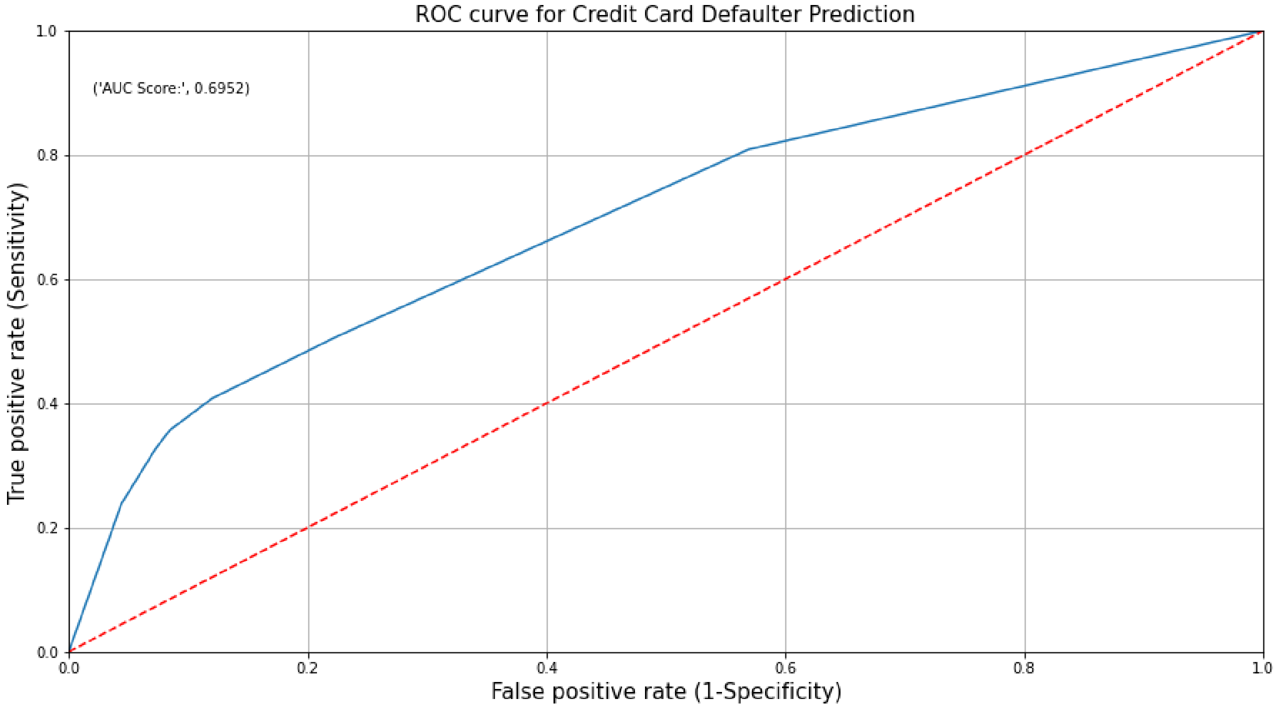




```
              precision    recall  f1-score   support

           0       0.82      0.96      0.88      7040
           1       0.60      0.24      0.34      1960

    accuracy                           0.80      9000
   macro avg       0.71      0.60      0.61      9000
weighted avg       0.77      0.80      0.76      9000
```

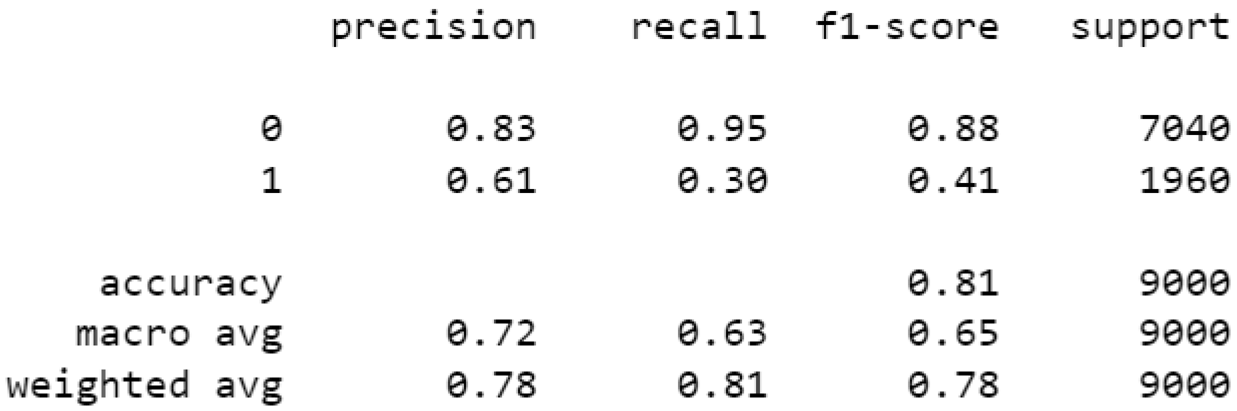


	Predicted:0	Predicted:1
Actual:0	6730	310
Actual:1	1494	466



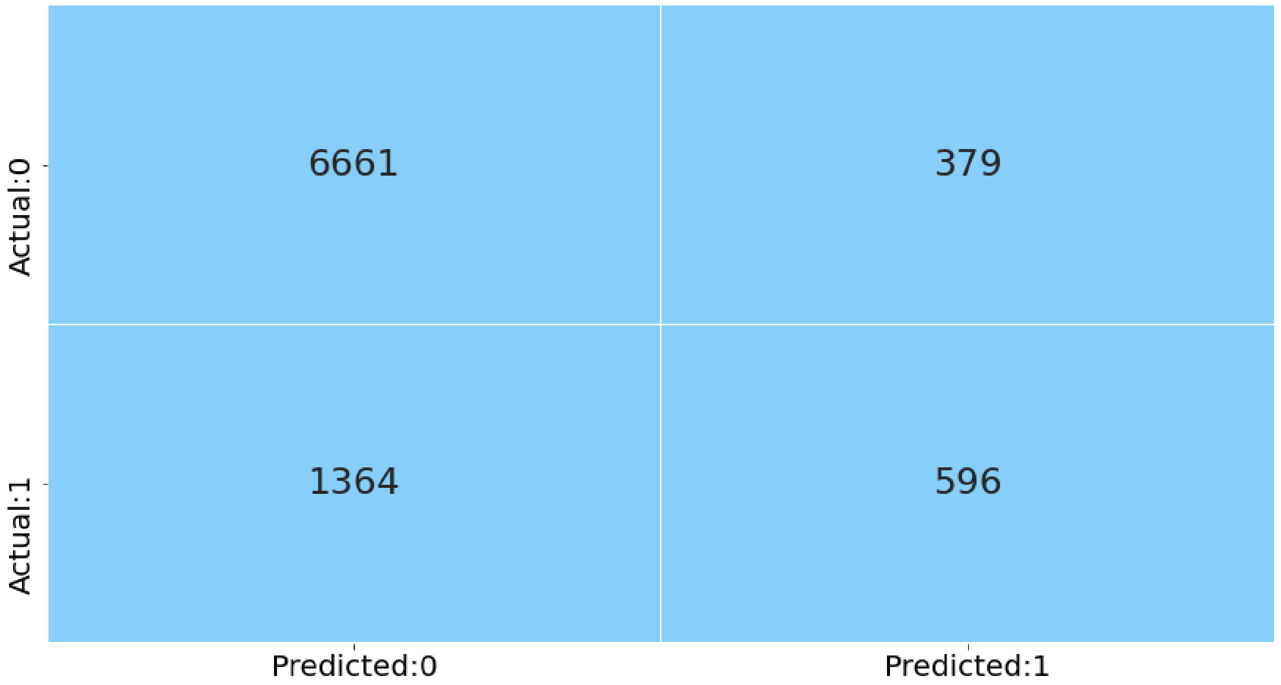
Building Boosting models - ADA Boost

AdaBoostClassifier(n_estimators=100, random_state=42)



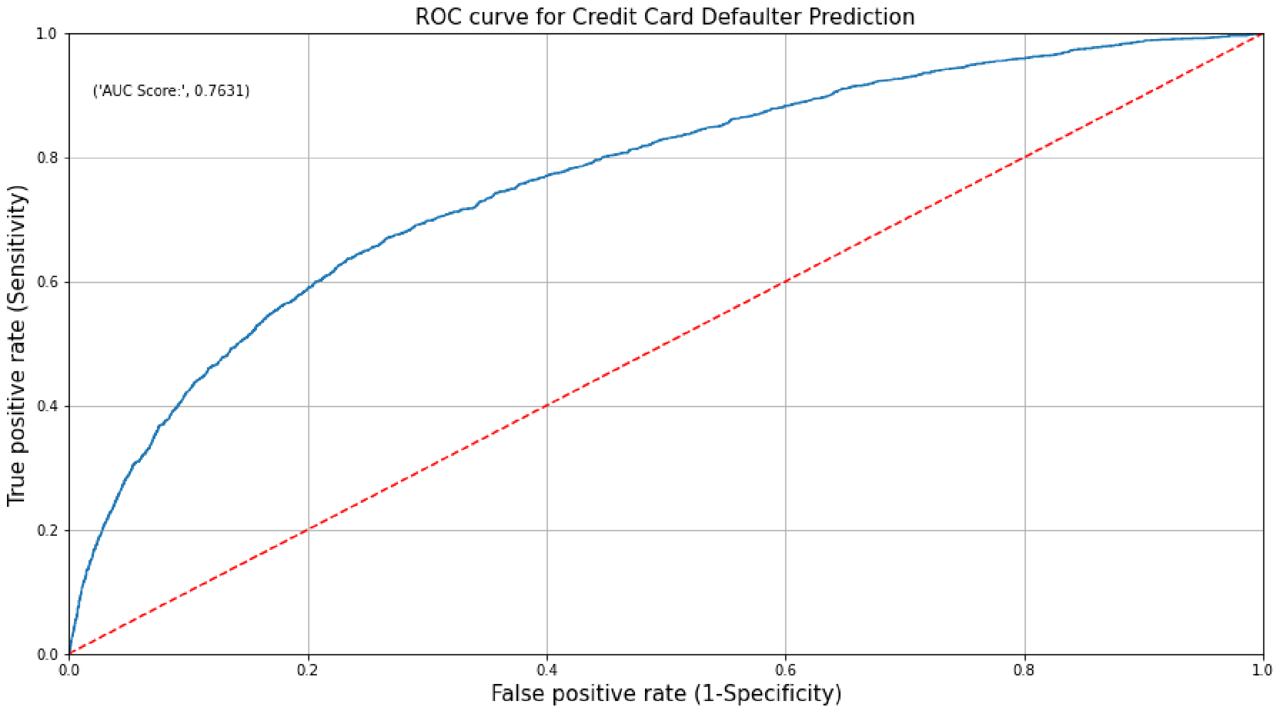
```
              precision    recall  f1-score   support

           0       0.83      0.95      0.88      7040
           1       0.61      0.30      0.41      1960

    accuracy                           0.81      9000
   macro avg       0.72      0.63      0.65      9000
weighted avg       0.78      0.81      0.78      9000
```

The results are poor for the positive class (as indicated by the f1-score). However, the accuracy has shown slight improvements from the previous models.



	Predicted:0	Predicted:1
Actual:0	6661	379
Actual:1	1364	596

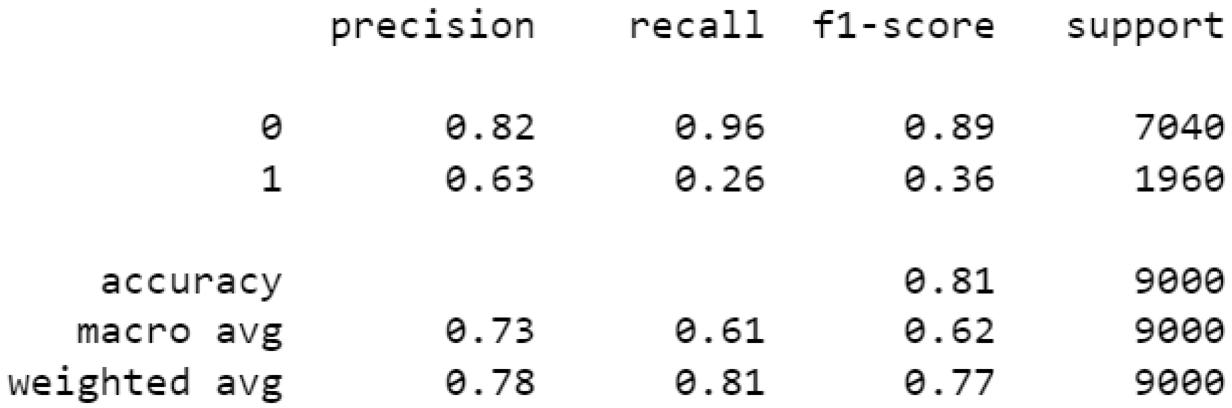
The confusion matrix shows that maximum of the positive classes are being predicted incorrectly.



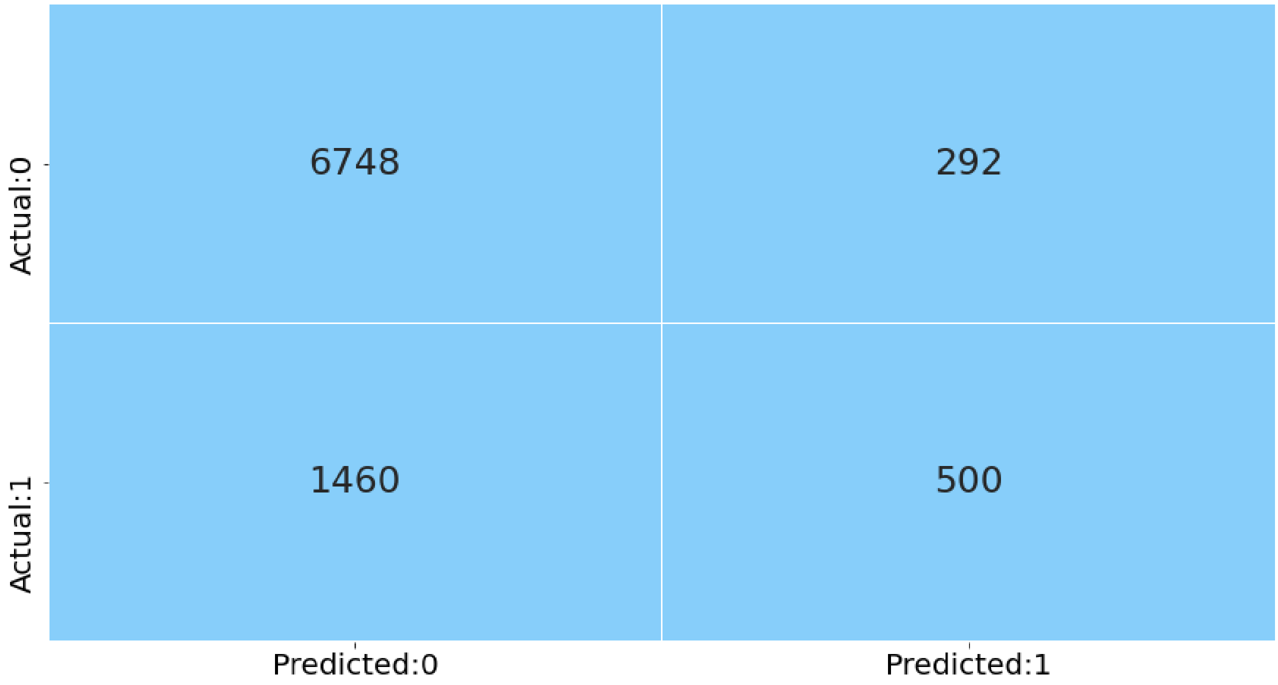
The AUC score has increased as compared to the previous models.

Hyperparameter tuning ADA Boost Model:-

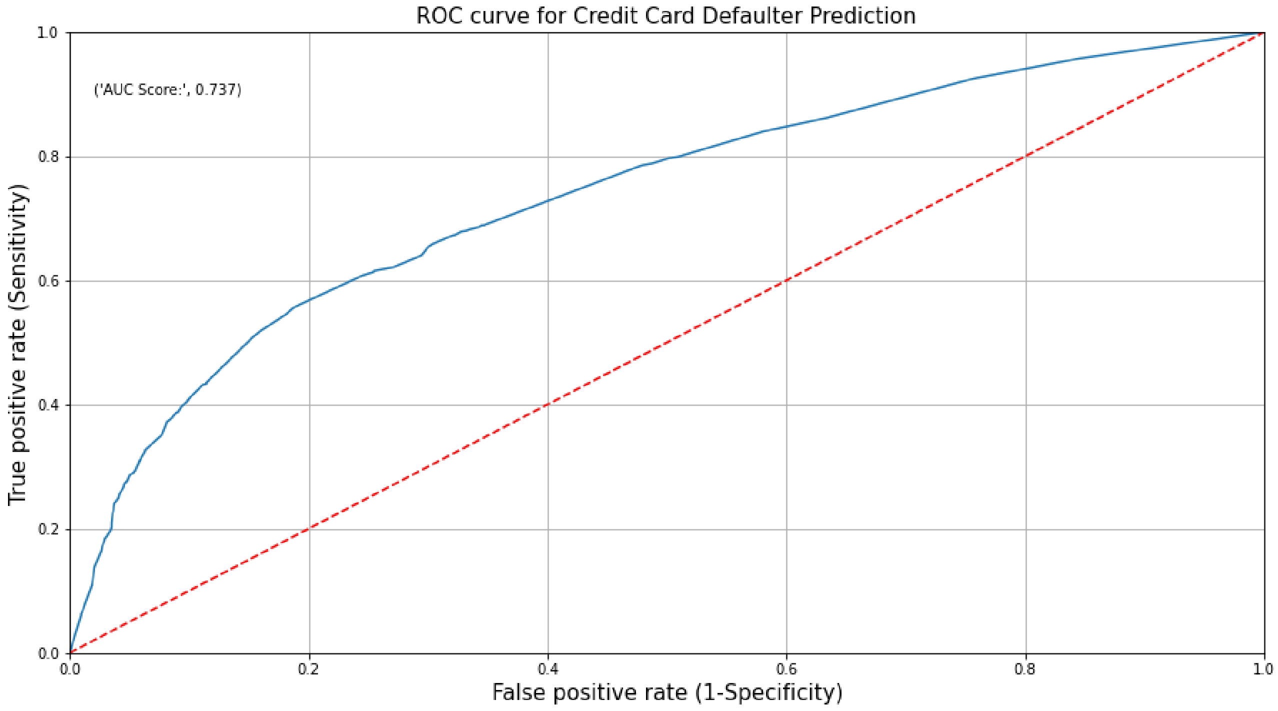
Best parameters for AdaBoost Classifier: {'learning_rate': 0.015, 'n_estimators': 150}



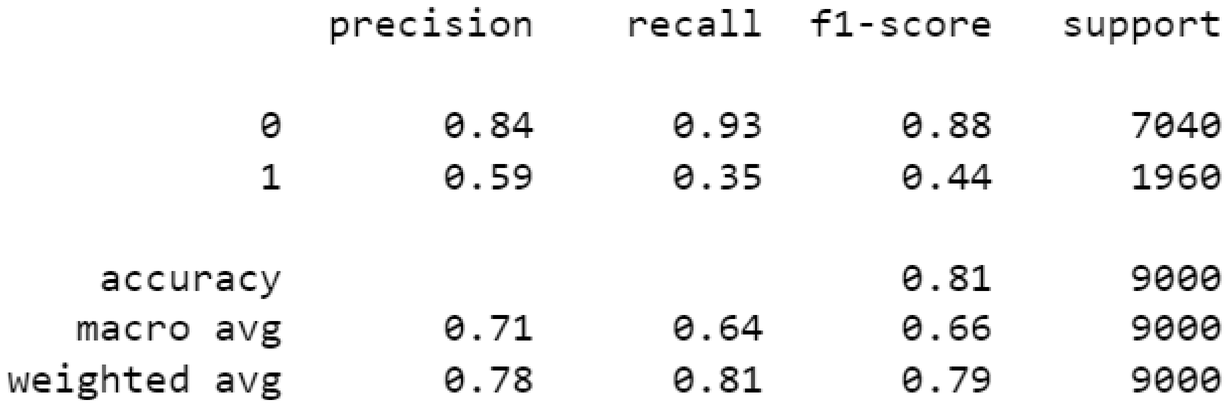
```
              precision    recall  f1-score   support

           0       0.82      0.96      0.89      7040
           1       0.63      0.26      0.36      1960

    accuracy                           0.81      9000
   macro avg       0.73      0.61      0.62      9000
weighted avg       0.78      0.81      0.77      9000
```



	Predicted:0	Predicted:1
Actual:0	6748	292
Actual:1	1460	500

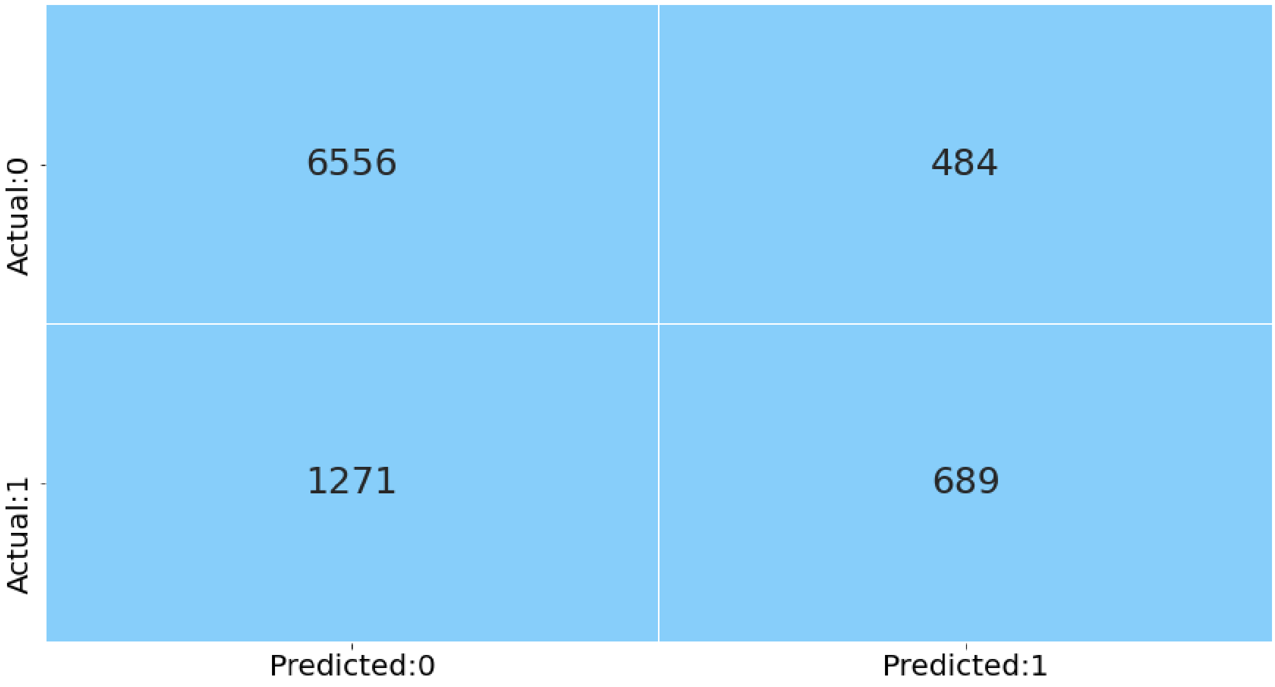


Gradient Boosting:-   GradientBoostingClassifier(max_depth=10, random_state=42)

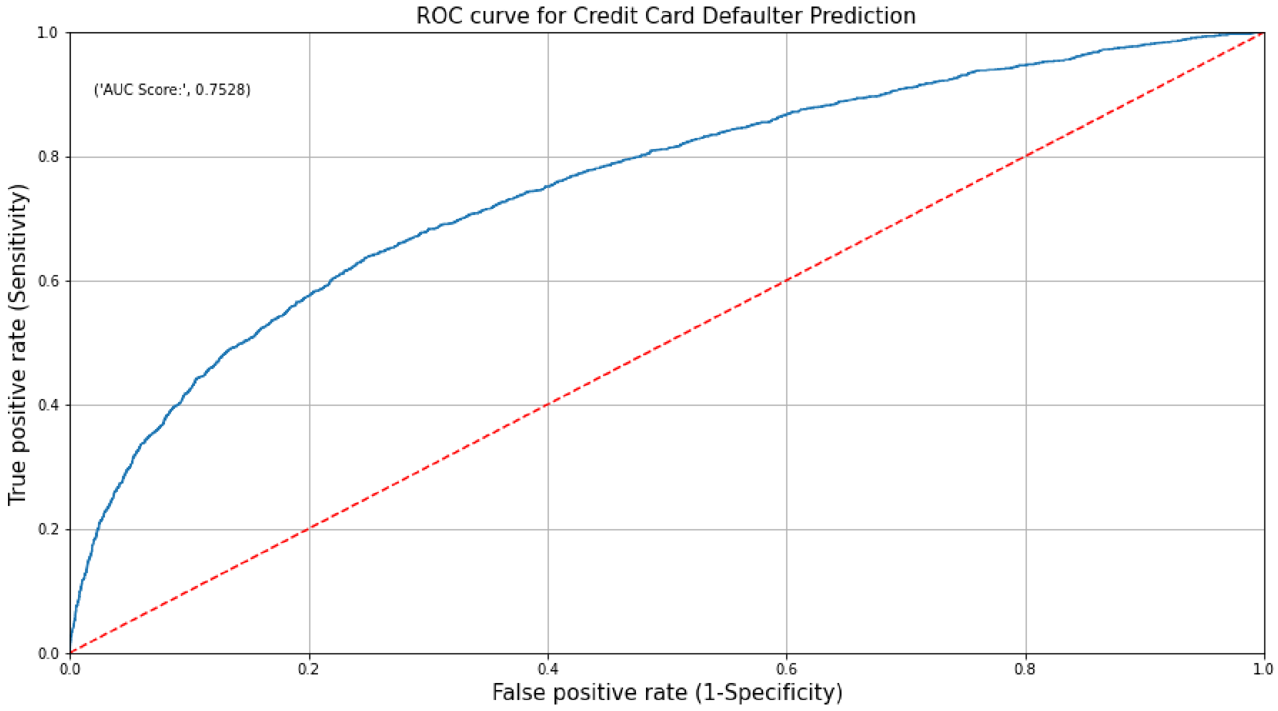


```
              precision    recall  f1-score   support

           0       0.84      0.93      0.88      7040
           1       0.59      0.35      0.44      1960

    accuracy                           0.81      9000
   macro avg       0.71      0.64      0.66      9000
weighted avg       0.78      0.81      0.79      9000
```

The f1-score for the positive has increased slightly in this model, the accuracy is the same as previous.



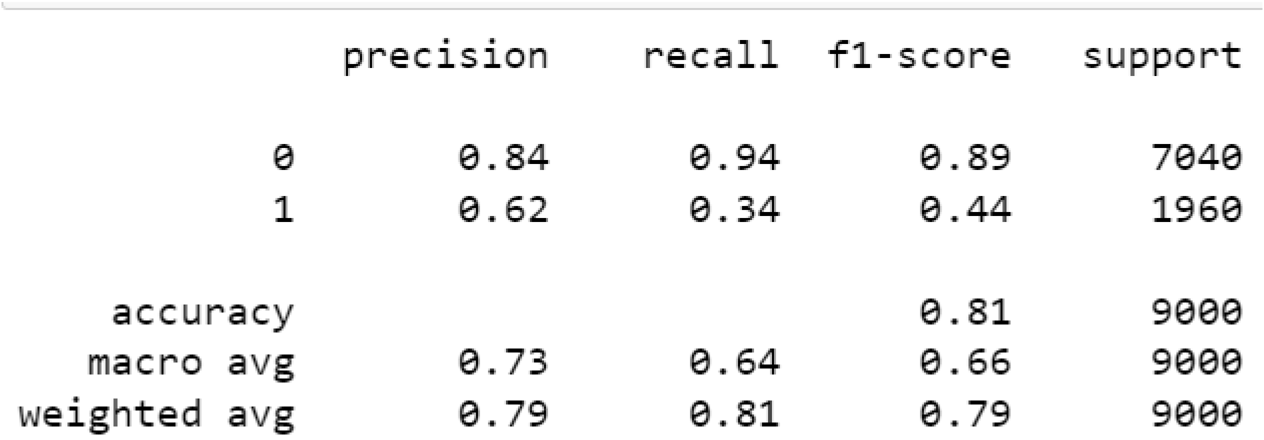
	Predicted:0	Predicted:1
Actual:0	6556	484
Actual:1	1271	689



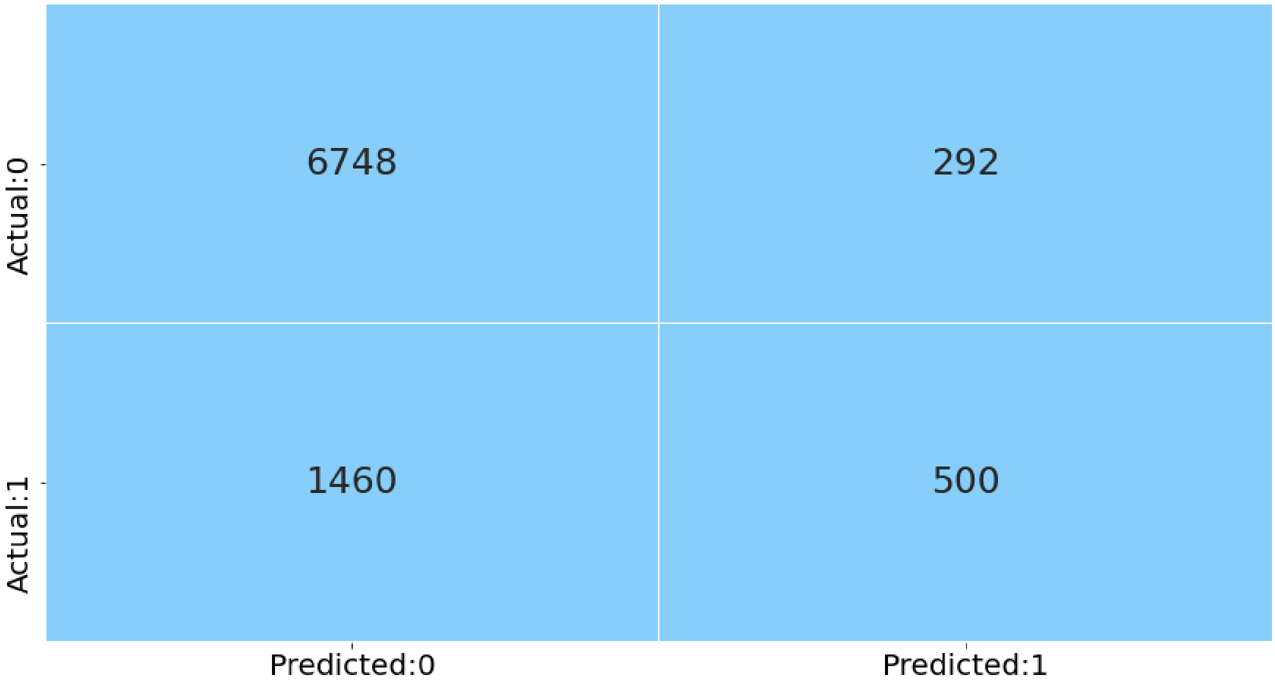
Maximum of the positive class is still incorrectly predicted.
The AUC score is similar to that of the last model.

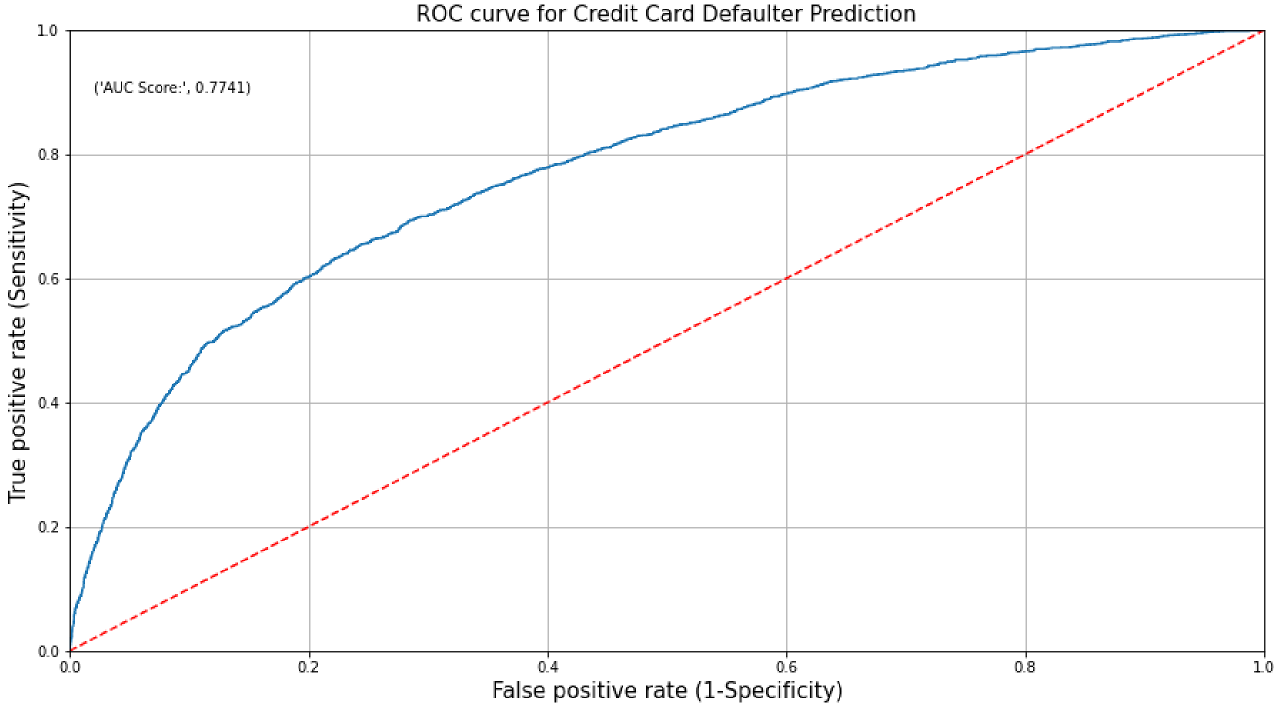
Hyperparameter tuning Gradient Boost Model:-

```
GradientBoostingClassifier(learning_rate=0.15, max_depth=4, n_estimators=120,
                           random_state=8)
```



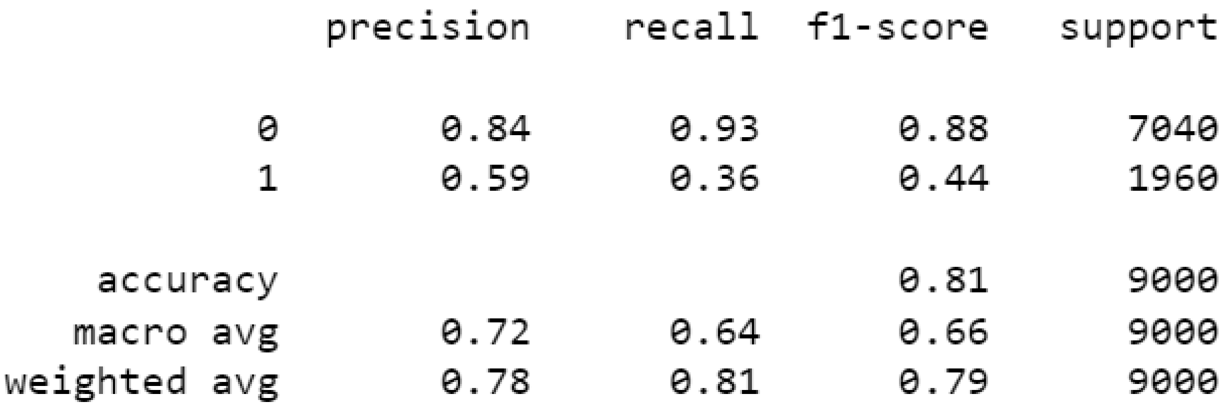
```
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      7040
           1       0.62      0.34      0.44      1960

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



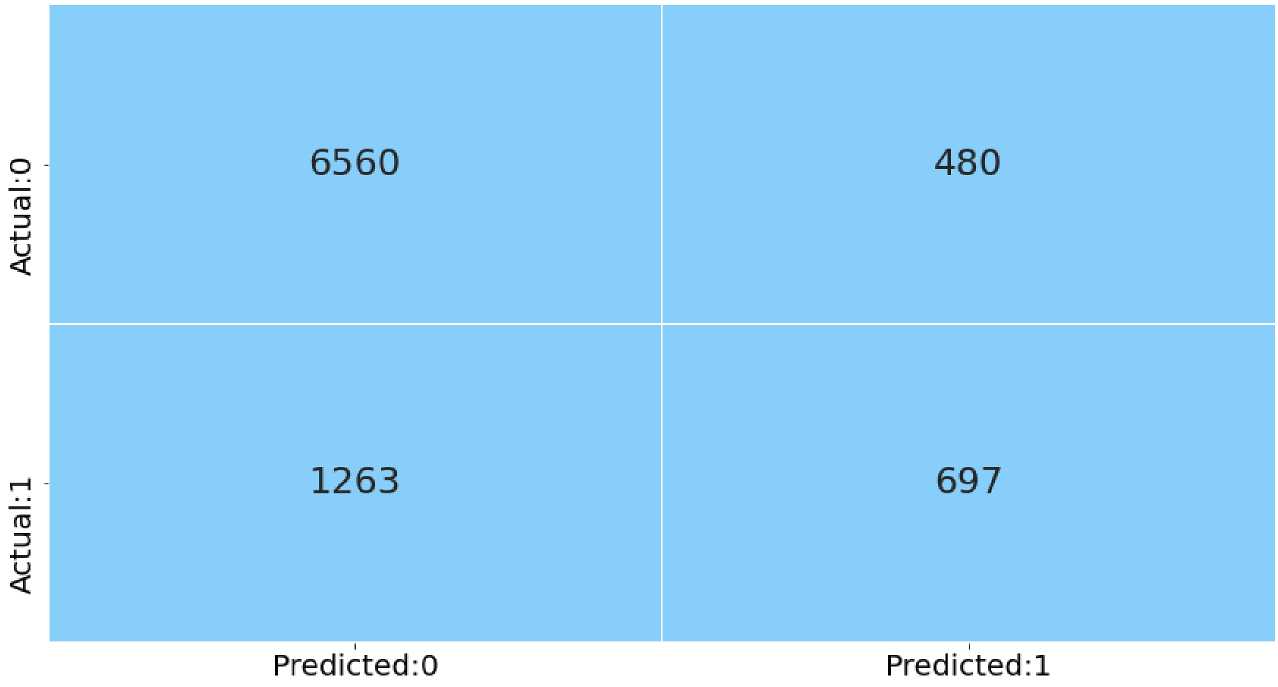


<u>XGBoost Model</u> :-

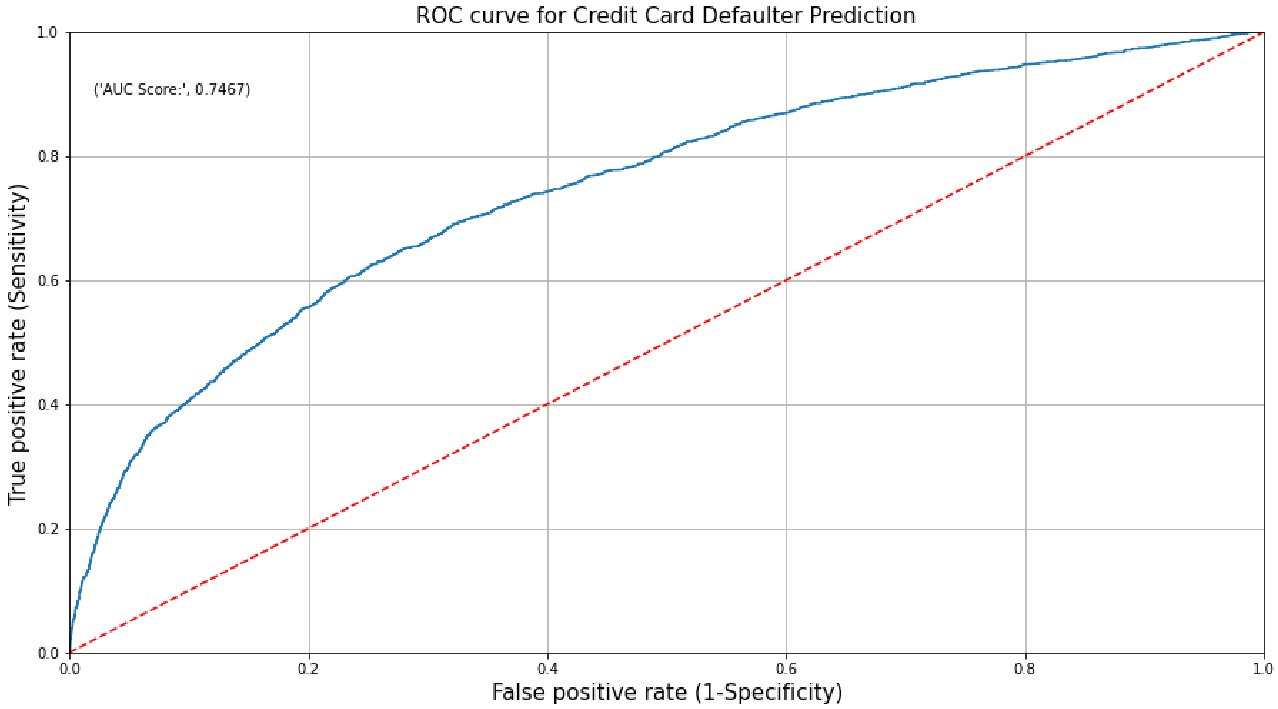
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
 colsample_bynode=1, colsample_bytree=1, eval_metric='mlogloss',
 gamma=1, gpu_id=-1, importance_type='gain',
 interaction_constraints='', learning_rate=0.300000012,
 max_delta_step=0, max_depth=10, min_child_weight=1, missing=nan,
 monotone_constraints='()', n_estimators=100, n_jobs=12,
 num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
 scale_pos_weight=1, subsample=1, tree_method='exact',
 validate_parameters=1, verbosity=None)



```
              precision    recall  f1-score   support

           0       0.84      0.93      0.88      7040
           1       0.59      0.36      0.44      1960

    accuracy                           0.81      9000
   macro avg       0.72      0.64      0.66      9000
weighted avg       0.78      0.81      0.79      9000
```



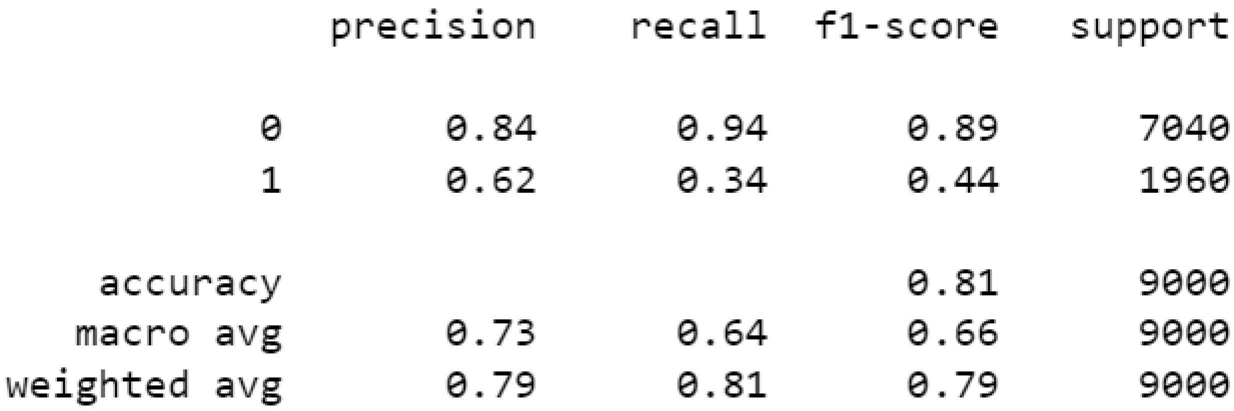
	Predicted:0	Predicted:1
Actual:0	6560	480
Actual:1	1263	697



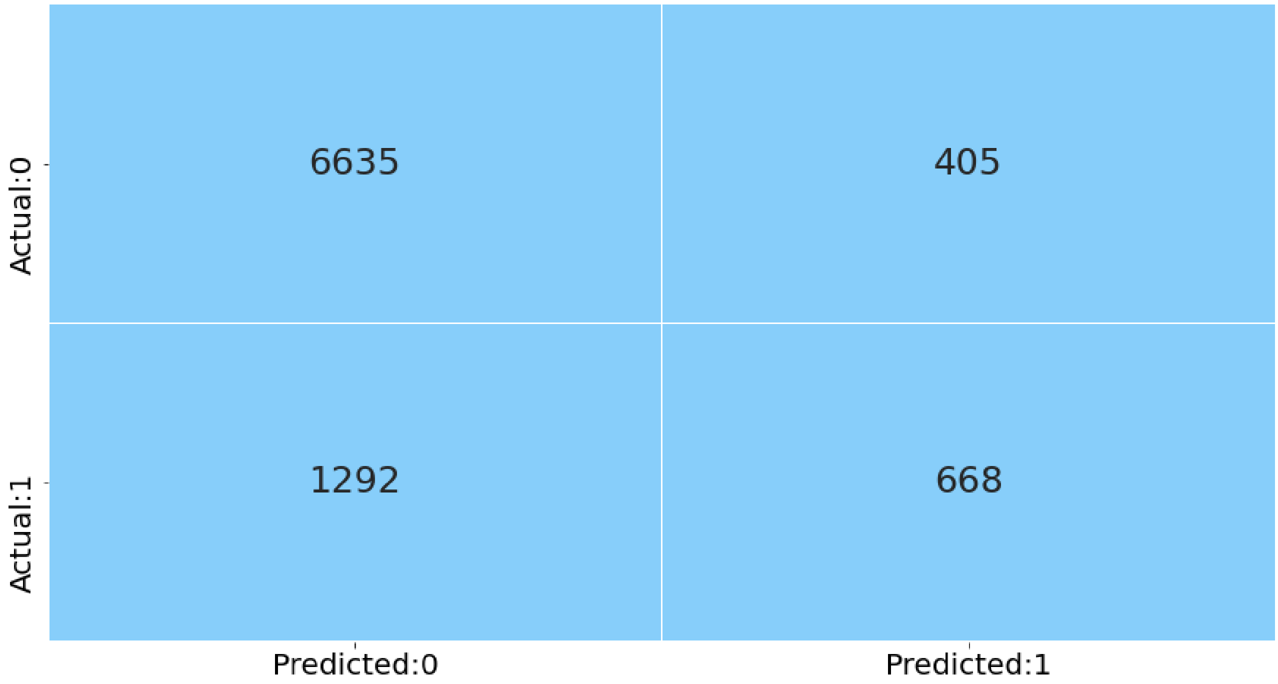
The AUC score for this model has reduced compared to the previous models.

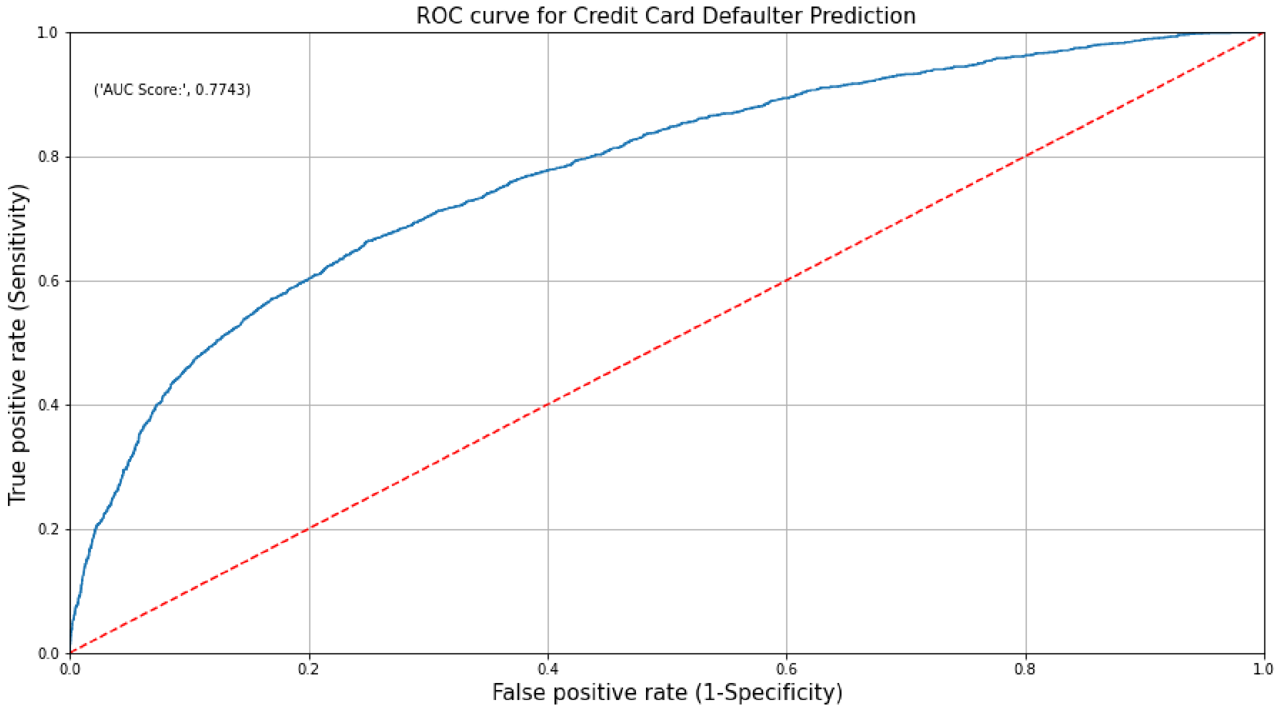
Hyperparameter Tuning XGBoost Model -

Best parameters for XGBoost classifier: {'gamma': 4, 'learning_rate': 0.1, 'max_depth': 6}



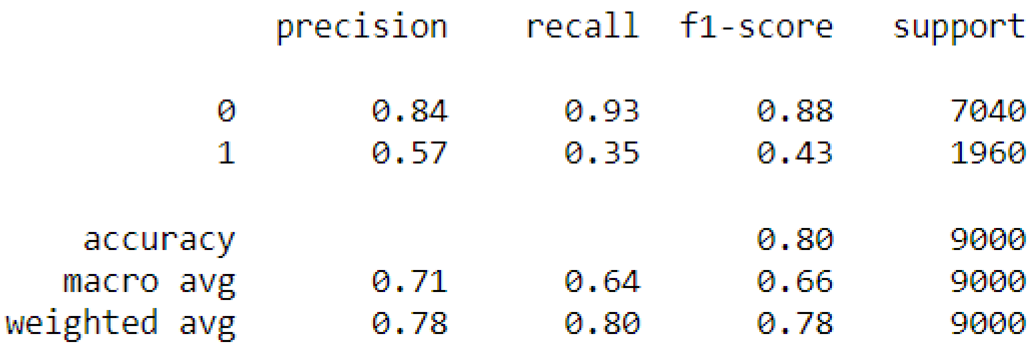
```
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      7040
           1       0.62      0.34      0.44      1960

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



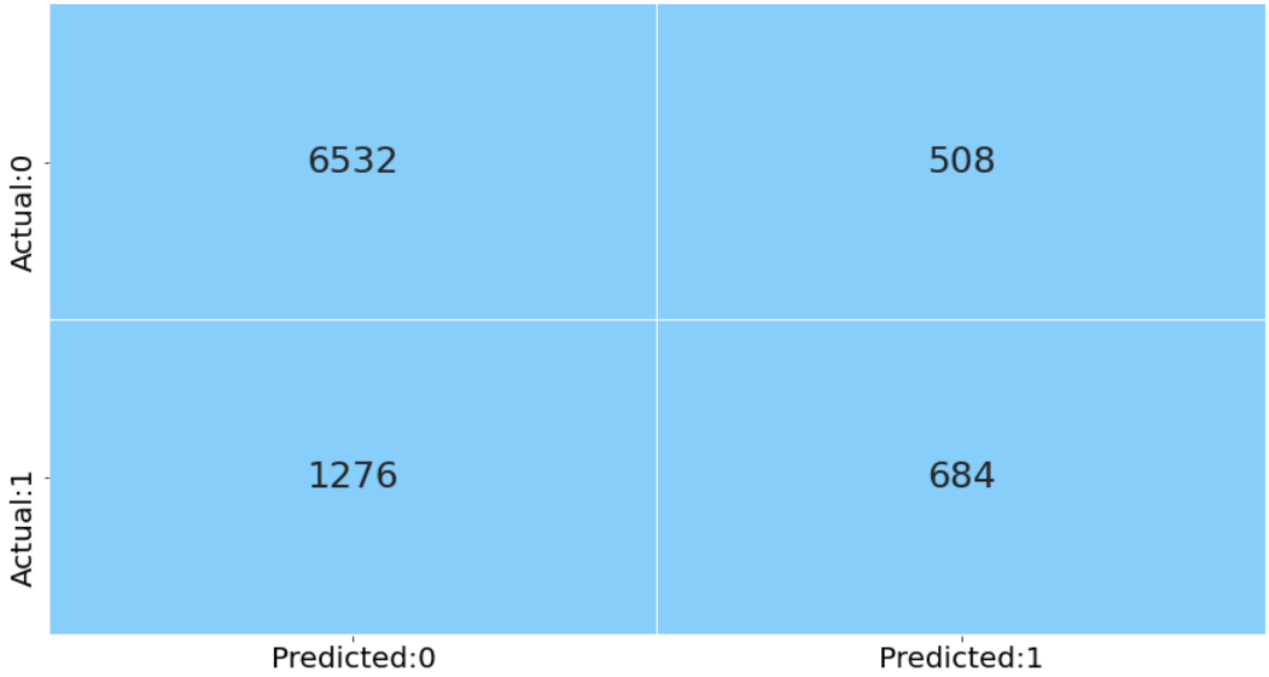


KNN Model:-

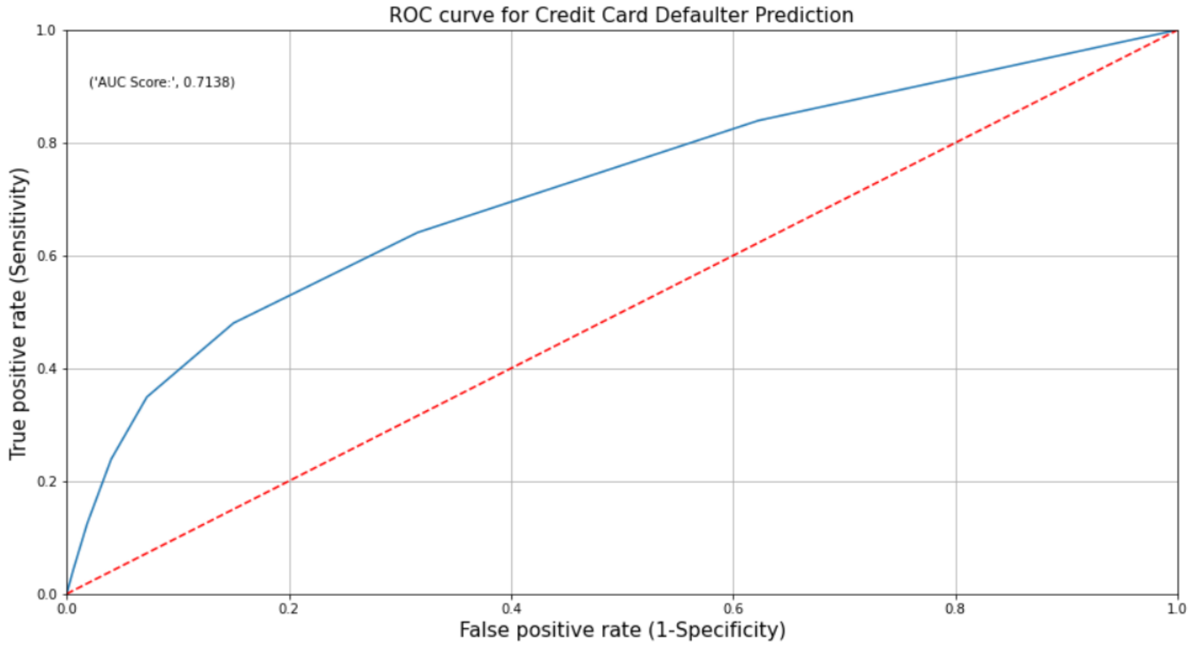
KNeighborsClassifier(n_neighbors=7,metric='euclidean')



```
              precision    recall  f1-score   support

           0       0.84      0.93      0.88      7040
           1       0.57      0.35      0.43      1960

    accuracy                           0.80      9000
   macro avg       0.71      0.64      0.66      9000
weighted avg       0.78      0.80      0.78      9000
```



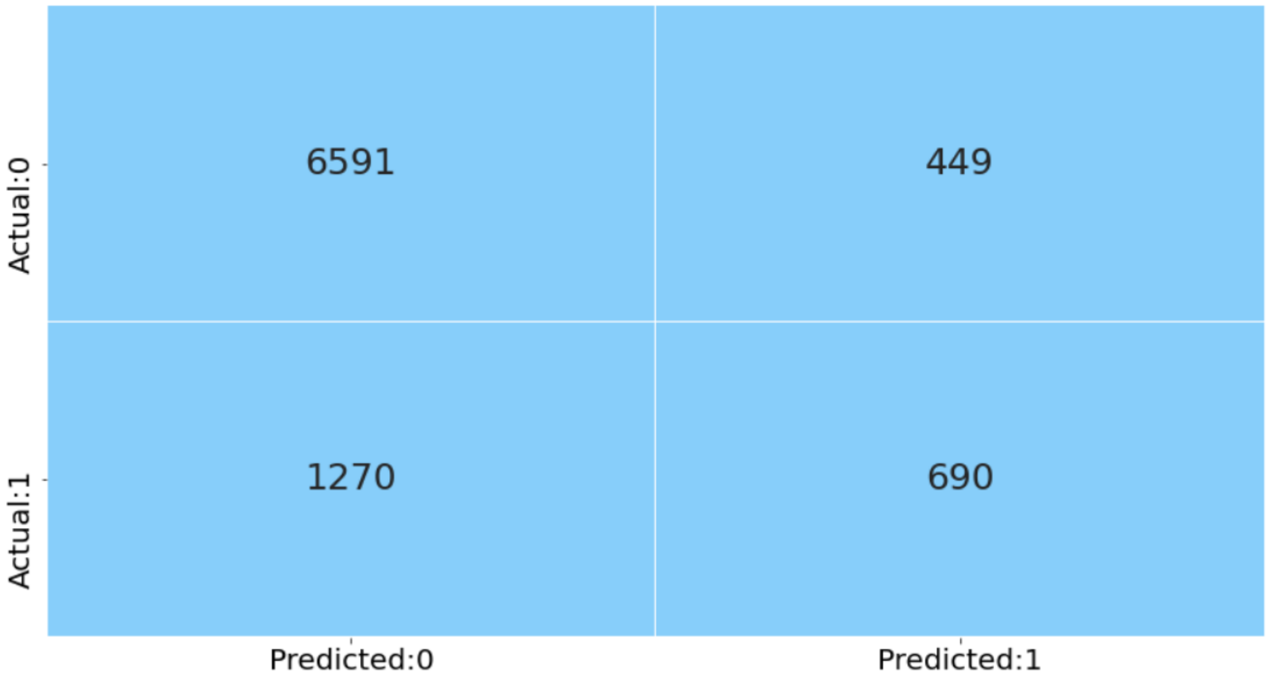
	Predicted:0	Predicted:1
Actual:0	6532	508
Actual:1	1276	684

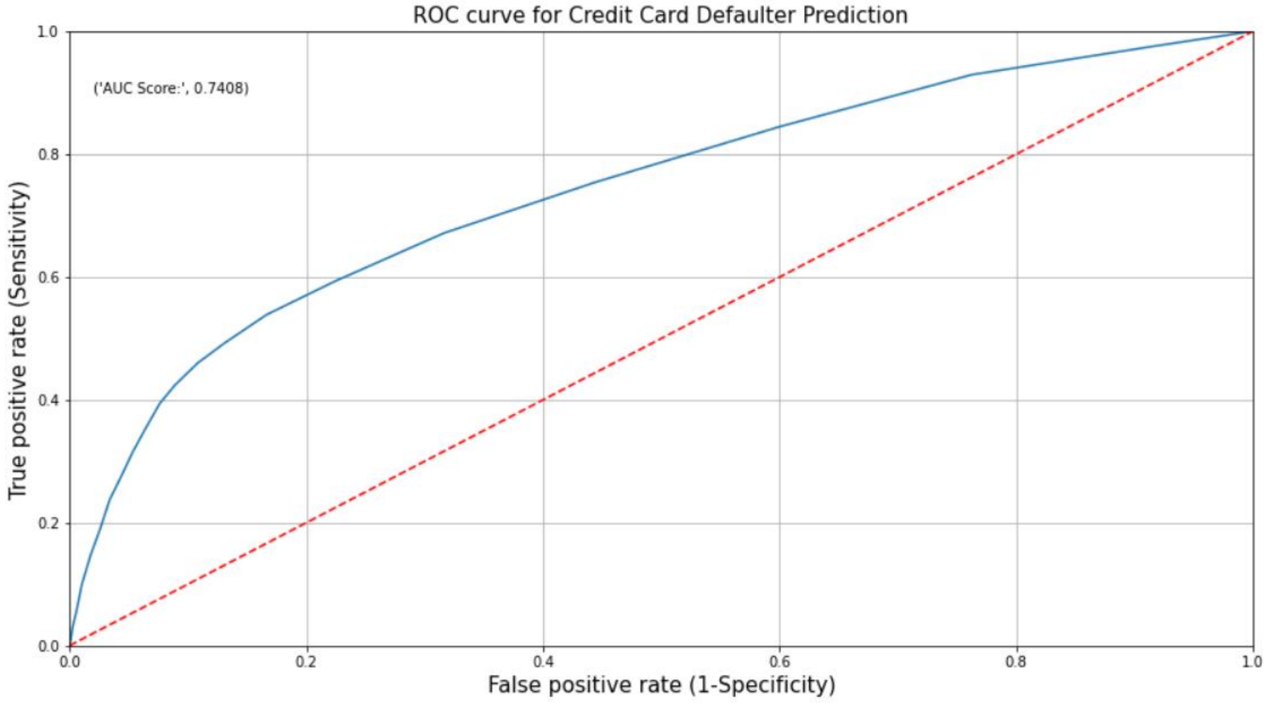


Hyperparameter Tuning KNN Model:-

Best parameters for KNN Classifier: {'metric': 'euclidean', 'n_neighbors': 23}

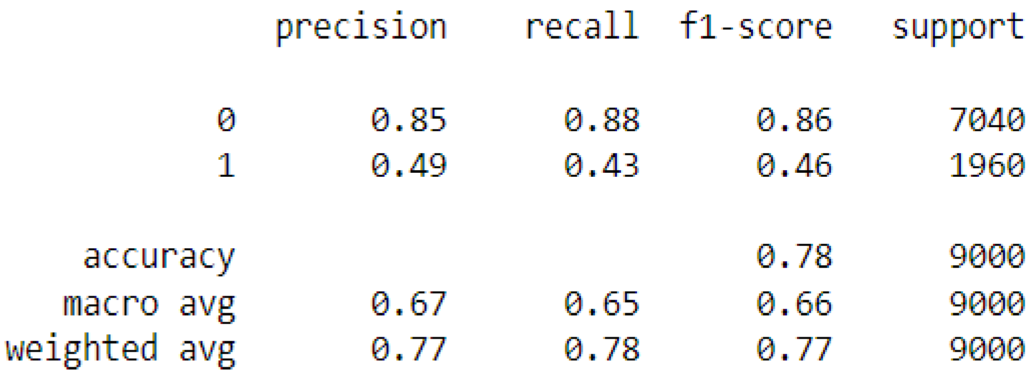
```
              precision    recall  f1-score   support

           0       0.84      0.94      0.88      7040
           1       0.61      0.35      0.45      1960

    accuracy                           0.81      9000
   macro avg       0.72      0.64      0.66      9000
weighted avg       0.79      0.81      0.79      9000
```



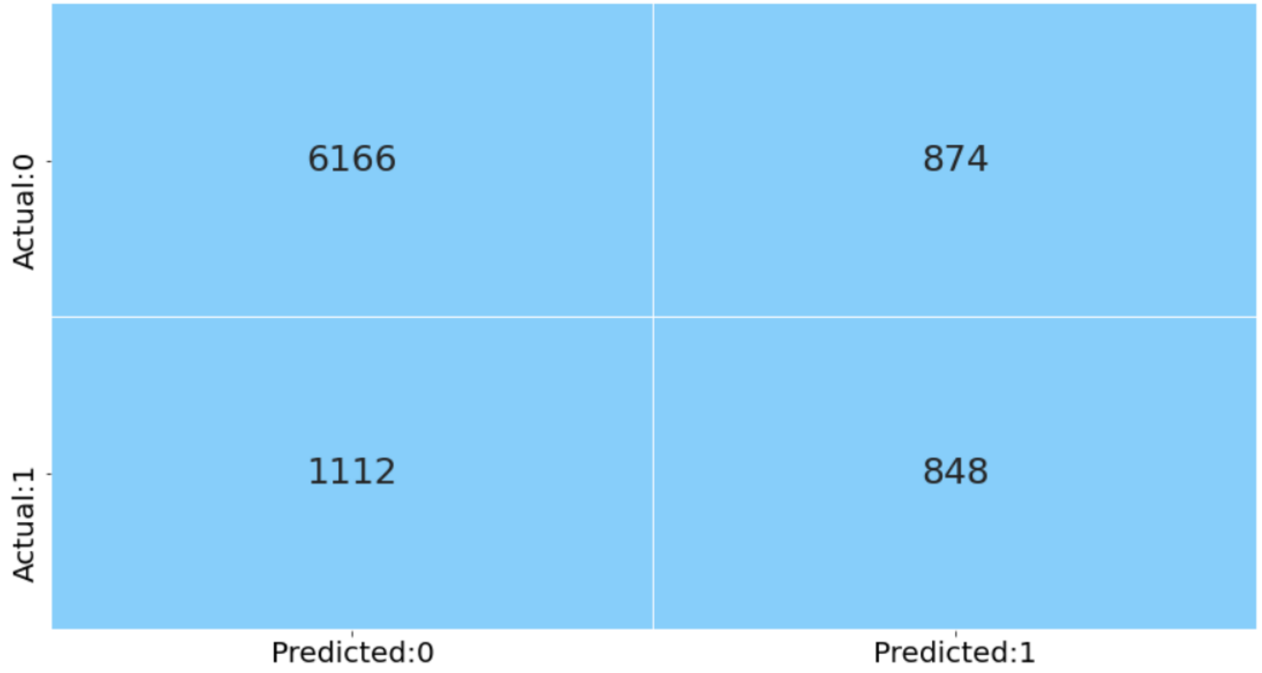


Gaussian Naive Bais Model:-

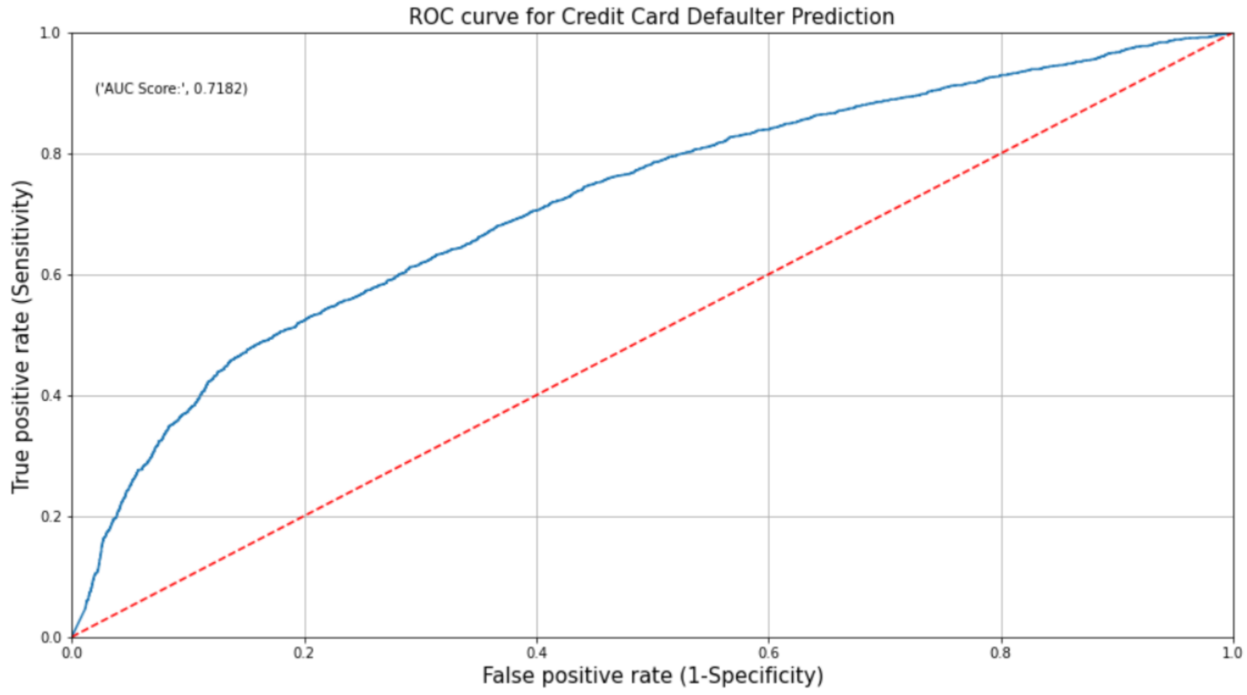
GaussianNB()



```
              precision    recall  f1-score   support

           0       0.85      0.88      0.86      7040
           1       0.49      0.43      0.46      1960

    accuracy                           0.78      9000
   macro avg       0.67      0.65      0.66      9000
weighted avg       0.77      0.78      0.77      9000
```

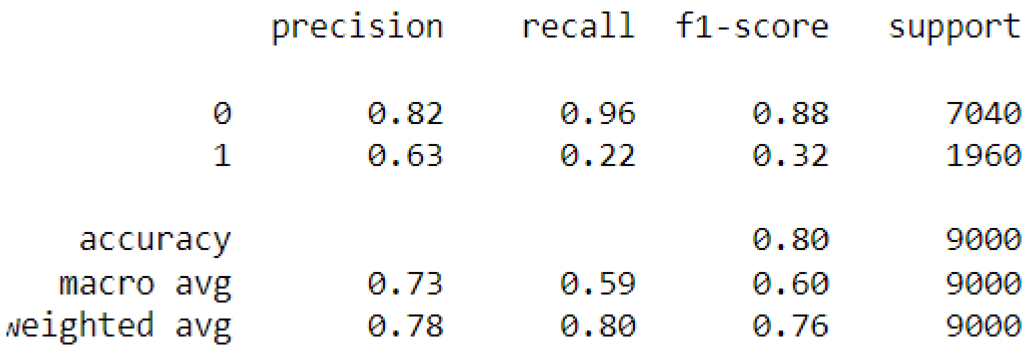


	Predicted:0	Predicted:1
Actual:0	6166	874
Actual:1	1112	848

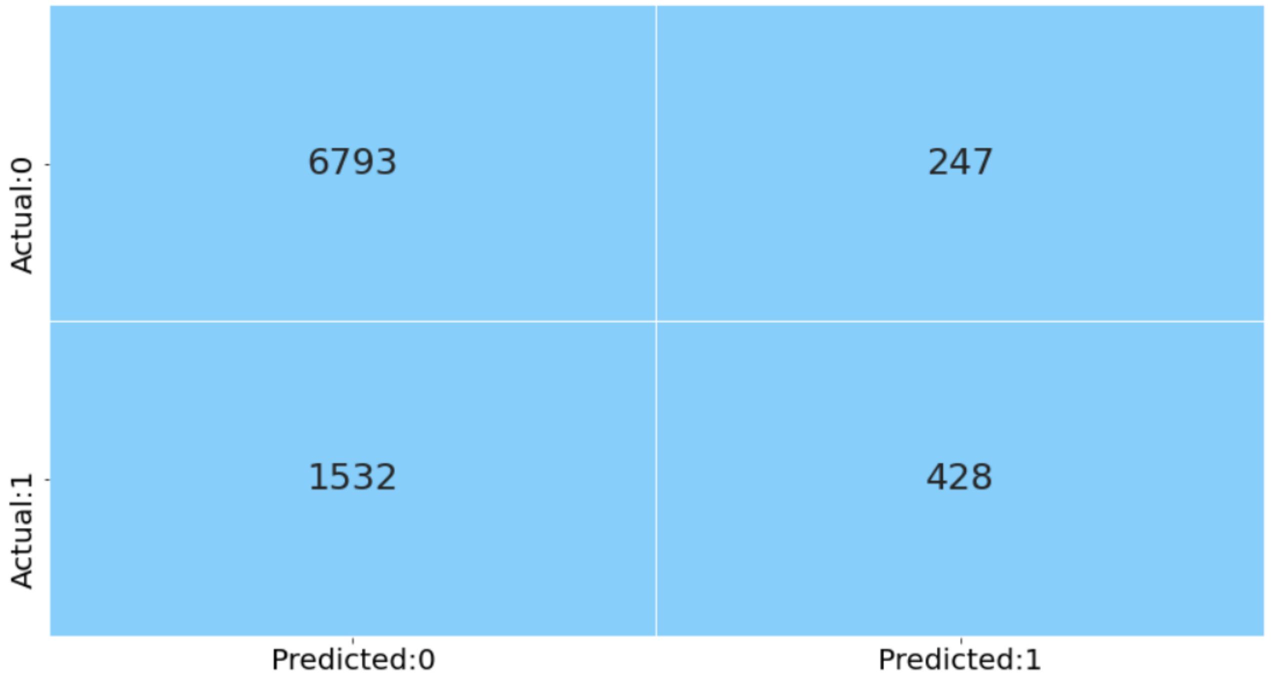


Hyperparameter tuning for Gaussian Naive Bais Model:-

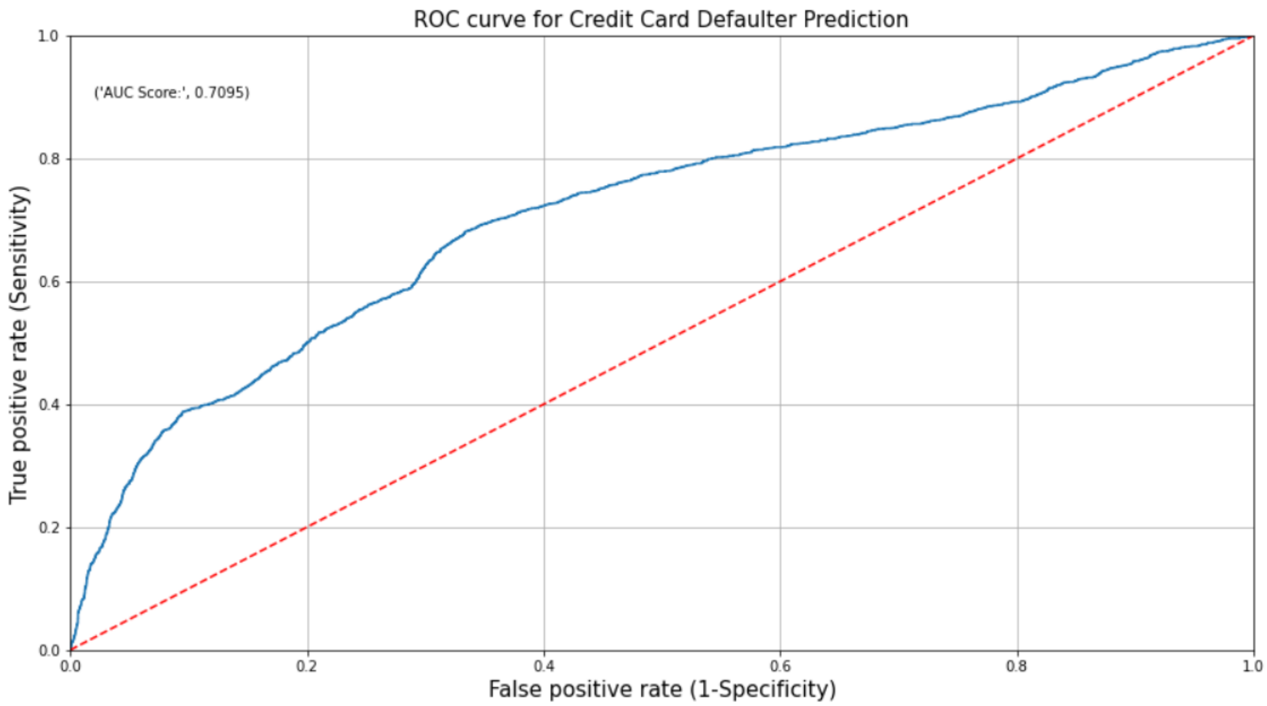
Best parameters for KNN Classifier: {'var_smoothing': 0.3511191734215131}



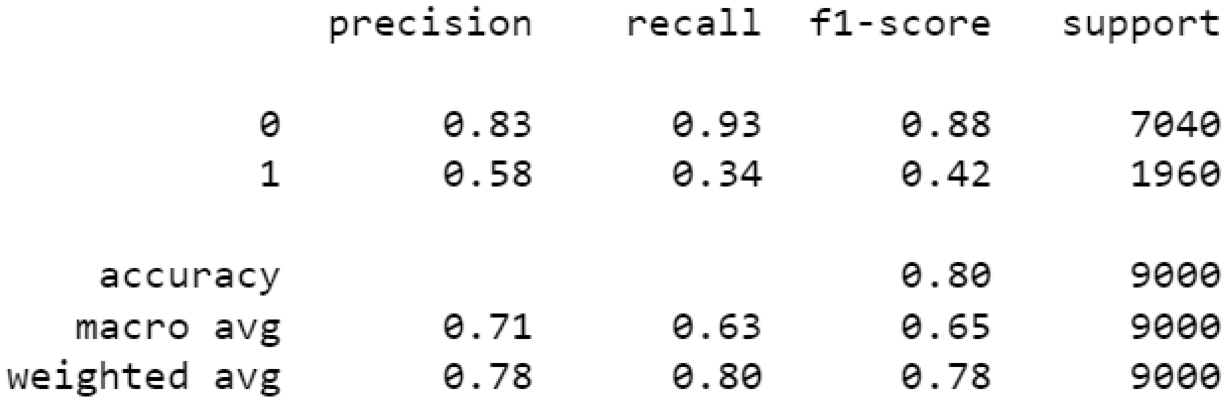
```
              precision    recall  f1-score   support

           0       0.82      0.96      0.88      7040
           1       0.63      0.22      0.32      1960

    accuracy                           0.80      9000
   macro avg       0.73      0.59      0.60      9000
weighted avg       0.78      0.80      0.76      9000
```



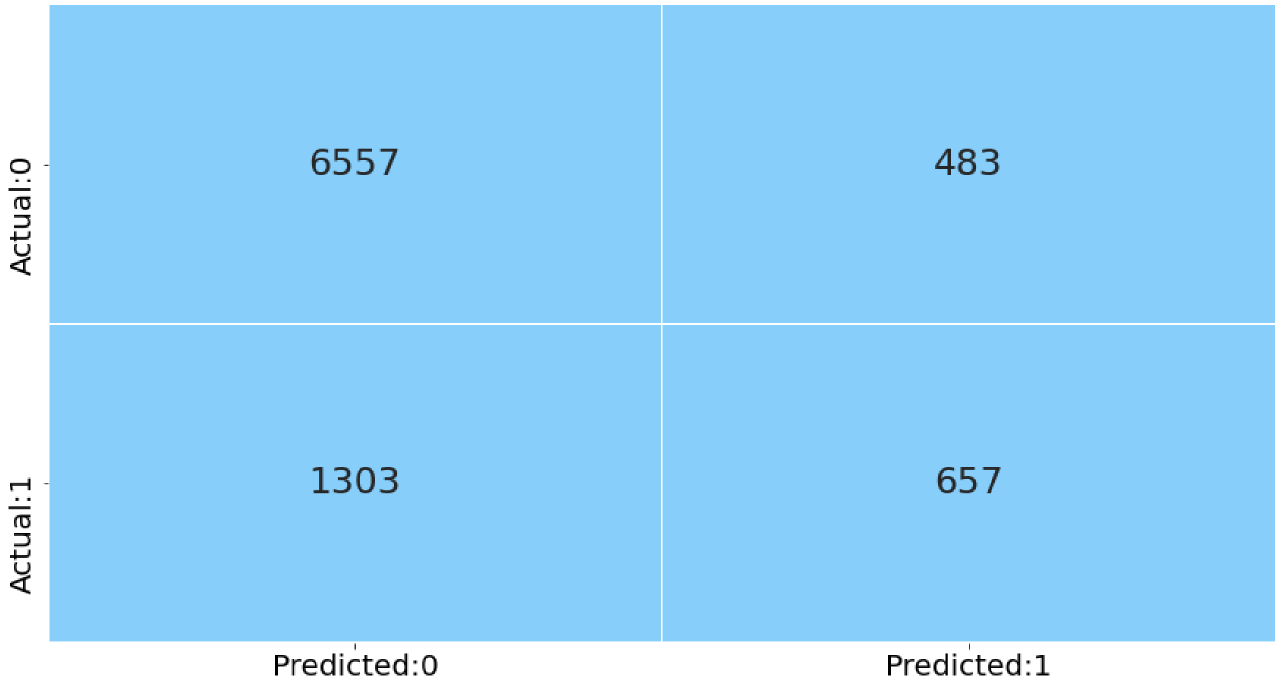
	Predicted:0	Predicted:1
Actual:0	6793	247
Actual:1	1532	428



Random Forest Model :-

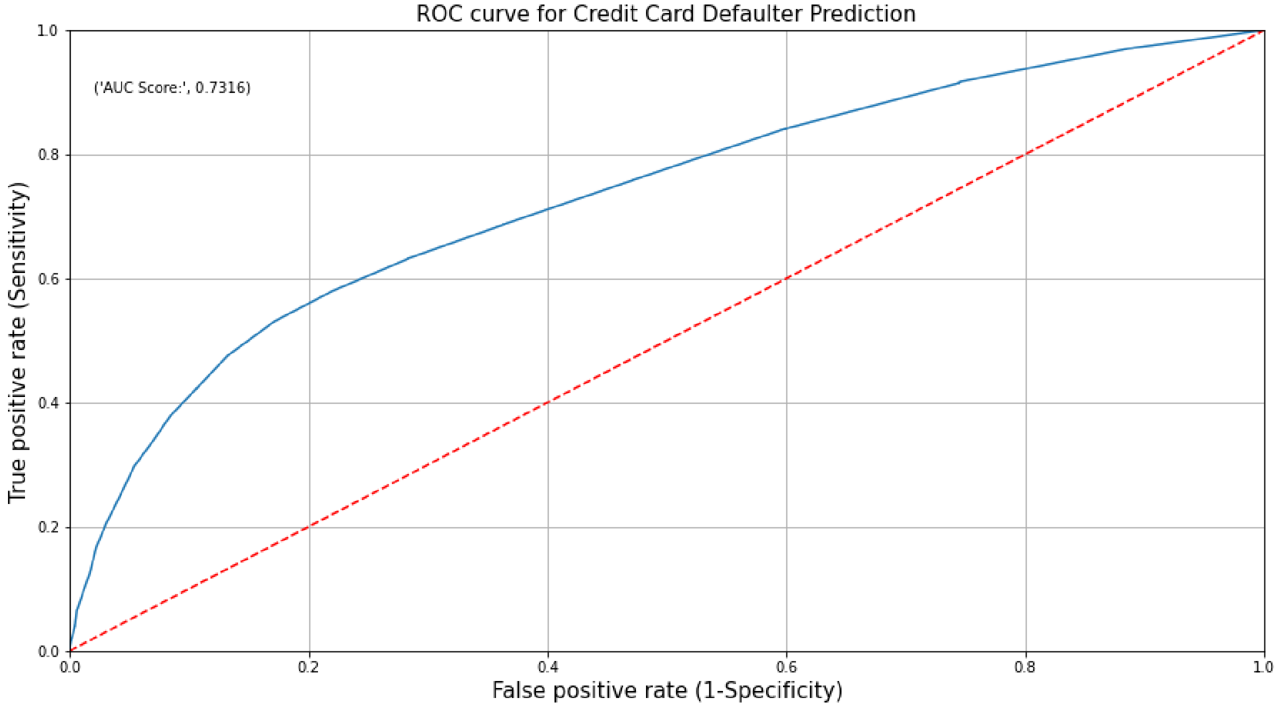


```
              precision    recall  f1-score   support

           0       0.83      0.93      0.88      7040
           1       0.58      0.34      0.42      1960

    accuracy                           0.80      9000
   macro avg       0.71      0.63      0.65      9000
weighted avg       0.78      0.80      0.78      9000
```



	Predicted:0	Predicted:1
Actual:0	6557	483
Actual:1	1303	657

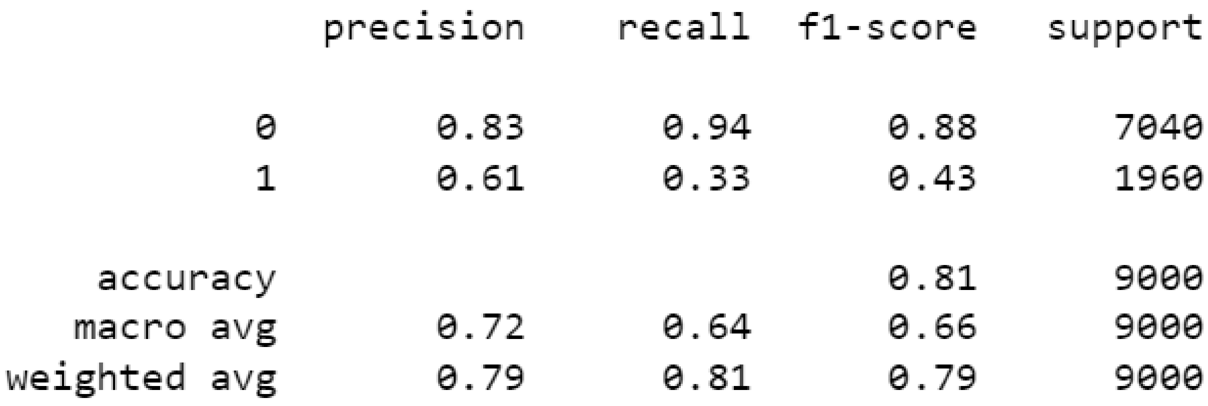
Random forest is performing poorly as compared to the boosted models.



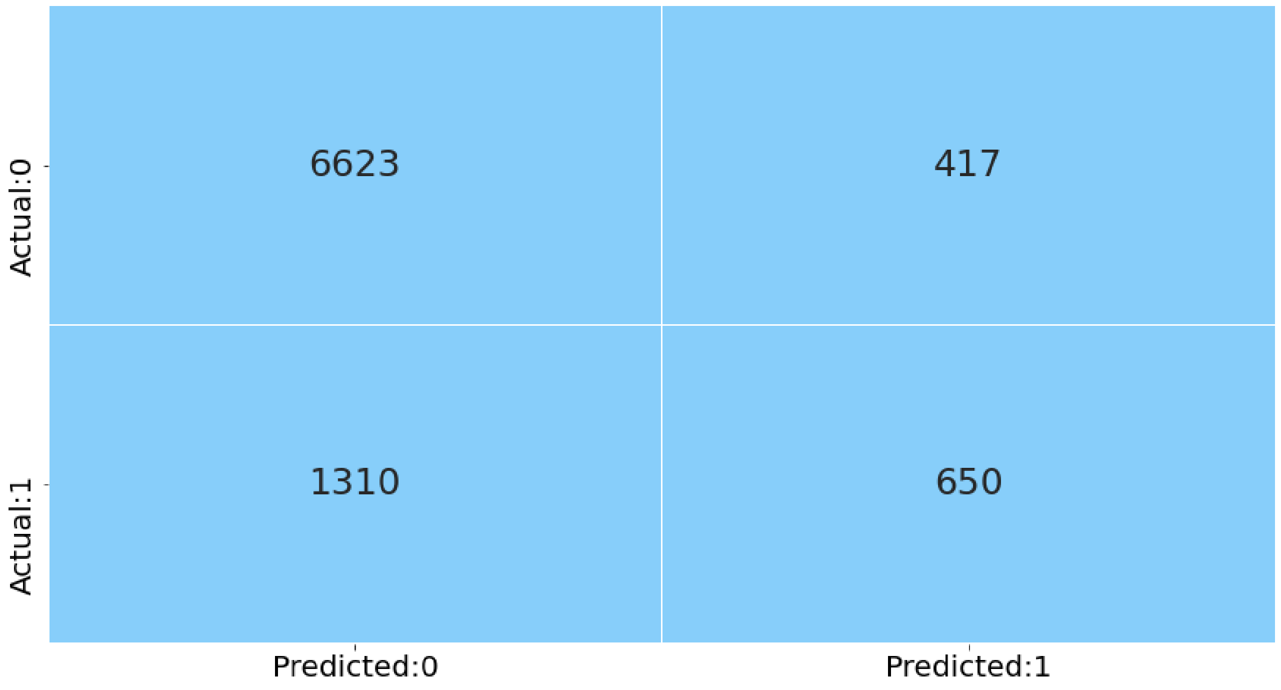
The AUC score for this model has decreased compared to the previous model.

Hyperparameter Tuning Random Forest Model:-

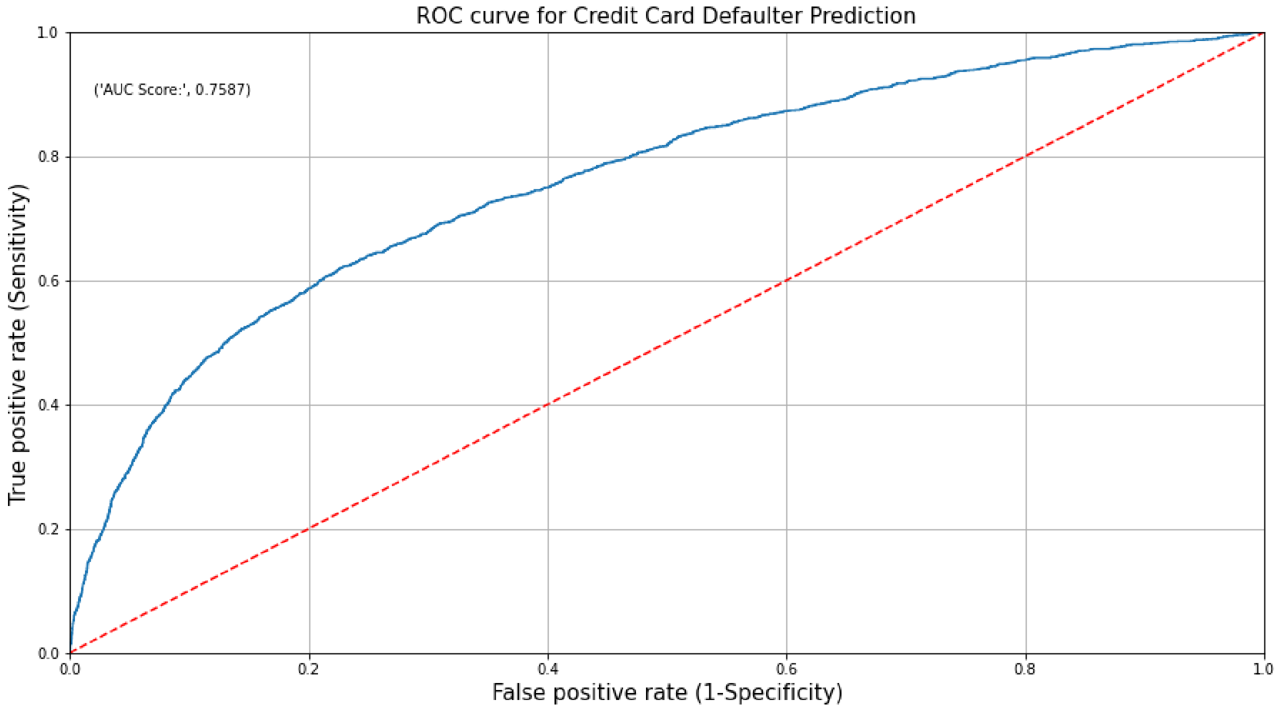
Best parameters for random forest classifier: {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_split': 11, 'n_estimators': 50}



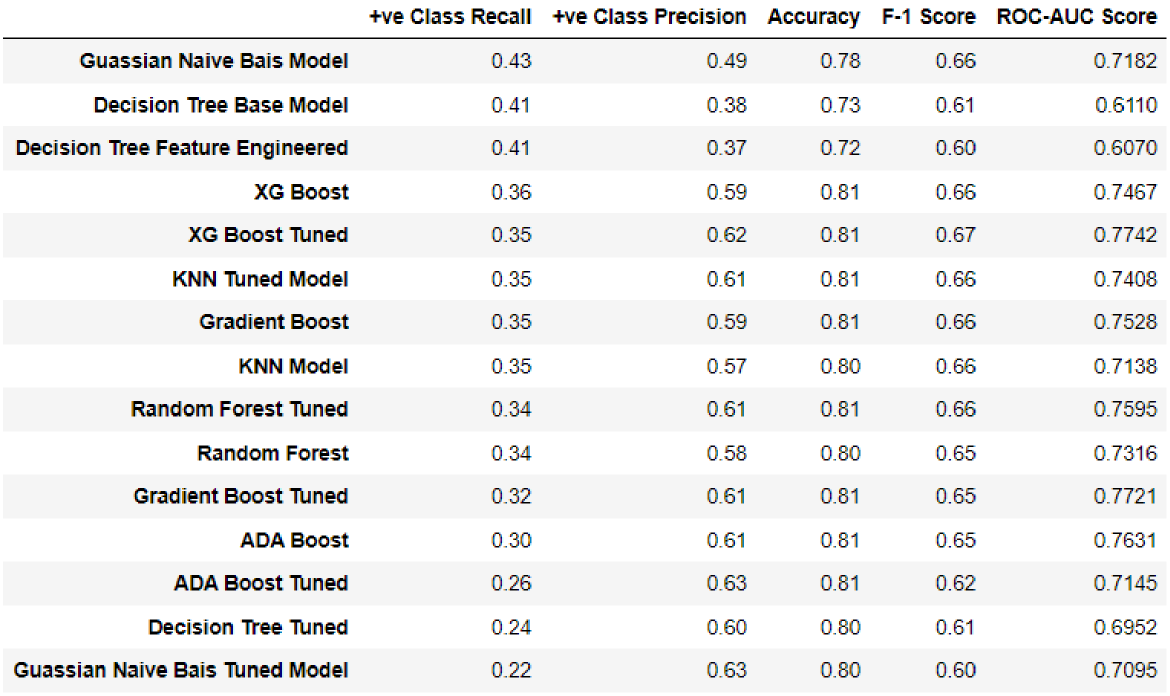
```
              precision    recall  f1-score   support

           0       0.83      0.94      0.88      7040
           1       0.61      0.33      0.43      1960

    accuracy                           0.81      9000
   macro avg       0.72      0.64      0.66      9000
weighted avg       0.79      0.81      0.79      9000
```



	Predicted:0	Predicted:1
Actual:0	6623	417
Actual:1	1310	650


ROC curve for Credit Card Defaulter Prediction
('AUC Score:', 0.7587)
True positive rate (Sensitivity)
False positive rate (1-Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

CONCLUSION:



	+ve Class Recall	+ve Class Precision	Accuracy	F-1 Score	ROC-AUC Score
Guassian Naive Bais Model	0.43	0.49	0.78	0.66	0.7182
Decision Tree Base Model	0.41	0.38	0.73	0.61	0.6110
Decision Tree Feature Engineered	0.41	0.37	0.72	0.60	0.6070
XG Boost	0.36	0.59	0.81	0.66	0.7467
XG Boost Tuned	0.35	0.62	0.81	0.67	0.7742
KNN Tuned Model	0.35	0.61	0.81	0.66	0.7408
Gradient Boost	0.35	0.59	0.81	0.66	0.7528
KNN Model	0.35	0.57	0.80	0.66	0.7138
Random Forest Tuned	0.34	0.61	0.81	0.66	0.7595
Random Forest	0.34	0.58	0.80	0.65	0.7316
Gradient Boost Tuned	0.32	0.61	0.81	0.65	0.7721
ADA Boost	0.30	0.61	0.81	0.65	0.7631
ADA Boost Tuned	0.26	0.63	0.81	0.62	0.7145
Decision Tree Tuned	0.24	0.60	0.80	0.61	0.6952
Guassian Naive Bais Tuned Model	0.22	0.63	0.80	0.60	0.7095

We built a total of 15 models using algorithms like KNN, Gaussian Naive Bayes, Decision Tree, Random Forest, and boosting algorithms like XGBoost, ADABoost and Gradient Boosting.

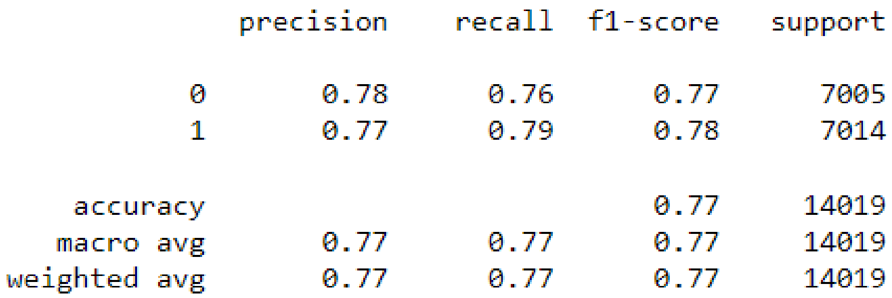
We used techniques like transformation, hyperparameter tuning, Feature engineering, scaling, and advanced algorithms like boosting for enhancing model performance.

After observing all the models built using these techniques, we came to the conclusion that these techniques did not show any significant changes in terms of the chosen metric.

The primary metric for our project was chosen to be recall for the positive class, as it gives the ratio of the number of correct predictions for the positive class to the total number of observations in the positive class. It is important for the bank to correctly identify the customers who are more likely to default. If they are unable to correctly predict the defaulters, they are at risk of losing money. Thus, recall is an appropriate metric to look at for this dataset.

According to this metric, the best-performing model is the Gaussian Naive Bayes Model. The recall score for this model is 0.43. The next best model is the Decision Tree Base Model (0.41), followed by the Decision Tree with Feature Engineering (0.41), and XGBoost (0.36).

We also carried out SMOTE for the decision tree model to demonstrate the effect of balanced data on model performance:



```
              precision    recall  f1-score   support

           0       0.78      0.76      0.77      7005
           1       0.77      0.79      0.78      7014

    accuracy                           0.77     14019
   macro avg       0.77      0.77      0.77     14019
weighted avg       0.77      0.77      0.77     14019
```

As we can see, the recall for the positive class increased from 0.41 to 0.79 after applying SMOTE.

This shows that with a balanced target column, our models will perform better. However, that is not possible in a real-life scenario. What is possible is to ask the clients for more data to better train our model.

Limitations:

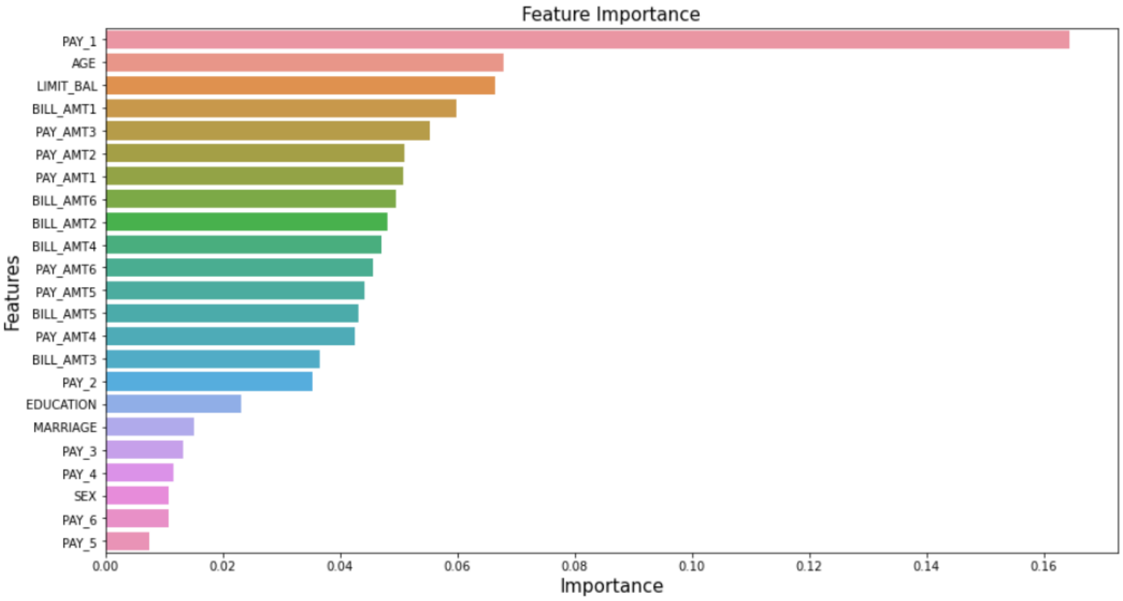
1. Data contains a lot of anomalies like outliers. However, we could not drop them as they are representative of real world scenarios and the model should be able to make predictions for such "outliers" as well.
2. The recall for our final model is on the lower end, indicating that our model has potential for improvement.
3. The techniques used for improving model performance did not have a significant impact on model performance.

Business Insights:

We determined the feature importance for the top three models, excluding Gaussian Naive Bayes, as this algorithm does not allow for conventional evaluation of feature importance.
Hence, we used the next three best models

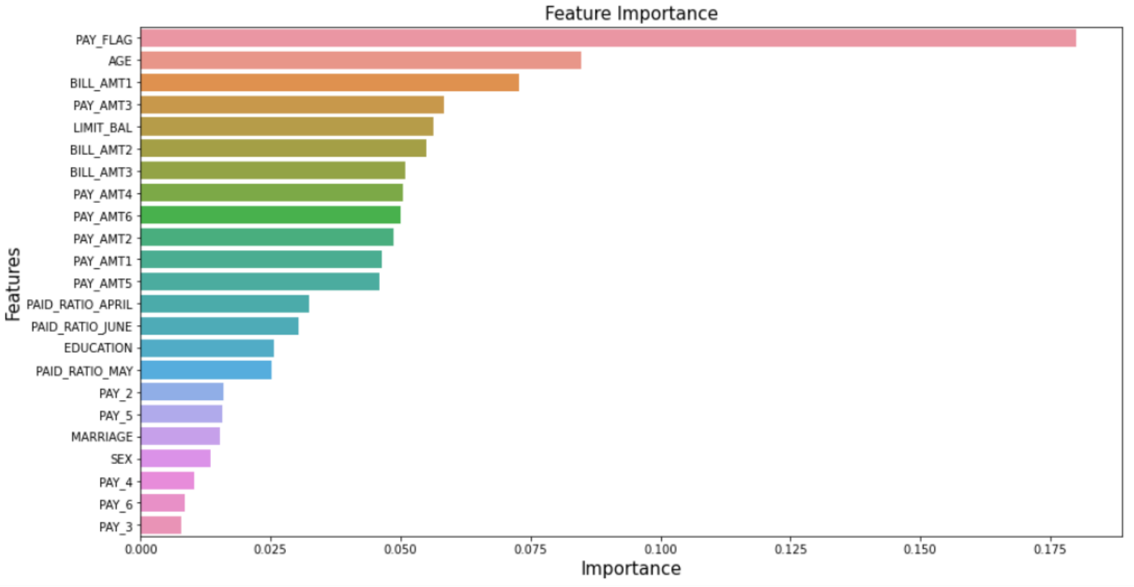
Following was obtained:

1. For Decision tree base model:



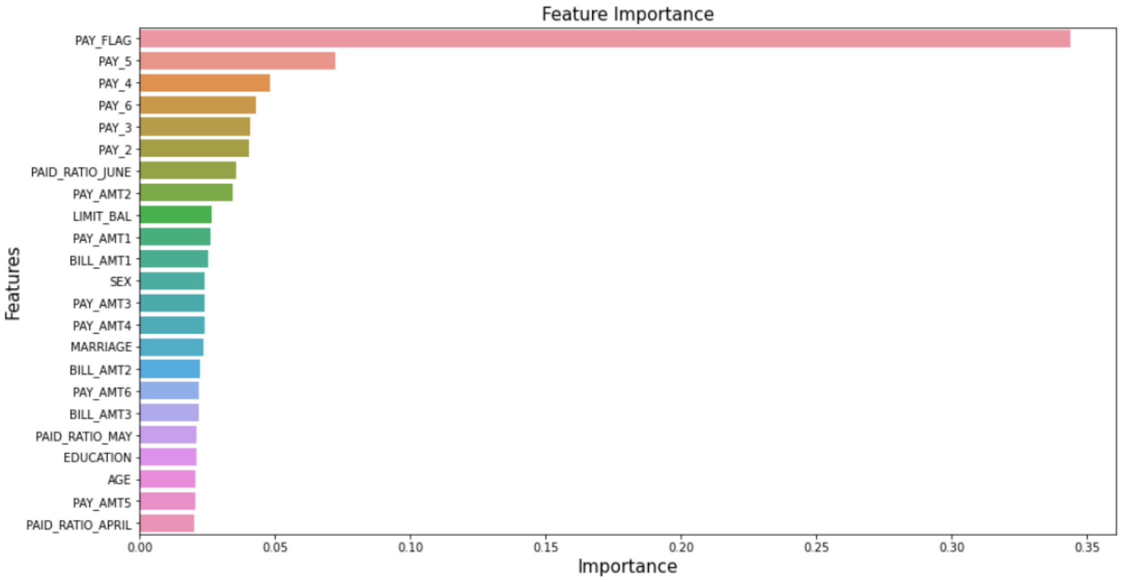
PAY_1, AGE, LIMIT_BAL, BILL_AMT1, PAY_AMT3, PAY_AMT2 are the top 6 features according to this model

2. For Decision tree model with feature engineering:



PAY_FLAG (new feature created from all PAY_X features), AGE, BILL_AMT1, PAY_AMT3, LIMIT_BAL, and BILL_AMT2 are the top 6 features according to this model.

3. For XGBoost Model:



PAY_FLAG, PAY_5, PAY_4, PAY_6, PAY_3, PAY_2 are the top 6 features according to this model.

When taking into consideration the above outcomes, we can see that the common top features for these models are the PAY_X columns, LIMIT_BAL, AGE, BILL_AMT1, and PAY_AMT3.
Hence, these features are the ones that should be focused on to make business decisions, as these are the features that affect the probability of defaulting.

Insights based on EDA:
Based on the EDA performed, we drew the following insights:

1. People with higher educational qualifications are less likely to default.
2. People with higher educational qualifications have a higher limit balance as they are less likely to default.
3. Married customers are also allowed higher limit balance.
4. Single customers are the largest users of revolving credits.
5. Female customers pay their bills on time compared to males.

From the above, we can conclude that married customers with higher educational qualifications are the safest customers, as they are less likely to default and they make their full payments on time.

Closing Reflections:

1. We have learned many things from Data Extraction, Data Cleaning, Data manipulation, EDA, Feature Engineering, Statistical aspects to Model building and further Hyperparameter tuning in order to boost the model performances.
2. Based on the steps taken and results obtained, we would like to make the following suggestions for further work on the dataset:
    - We need additional features to improve the model performance. This can be obtained by further EDA or asking the client for any additional features that can be added.
    - As our data has many outliers, it is possible that the data has been collected from two separate populations. As a result, we would need two different models for the dataset.
    - There are more complex models that can be tried such as neural network, SVM, stacked, etc, which might result in better performance.