Capstone Project - 3

Credit_Card_Default_Prediction

Individual Project by:
Anish Johnson

Contents:

Introduction:

As more and more consumers rely on credit cards to pay their everyday purchases in an online and physical retail store, the amount of issued credit cards and the overwhelming amount of credit card debt by the cardholders have rapidly increased.

Therefore, most financial institutions have to deal with the issues of credit card default in addition to credit card fraud.

To tackle this these institutions rely on default analysis systems empowered with machine learning and several other techniques.

Objective:

Our objective is to conduct quantitative analysis on credit card default risk by using machine learning models with accessible customer data to assist in predicting the case of customers' default payments in Taiwan.

Dataset Preview:

This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Dependent Variable: DEFAULT (Yes = 1, No = 0)

Independent Variables:

Basic User Data	History of Past Payment	Amount of Bill Statement	Amount of Previous Payment
ID: Unique ID of each client.	**PAY_0:** Repayment status in September.	**BILL_AMT1:** Amount of bill statement in September.	**PAY_AMT1:** Amount of previous payment in September.
LIMIT_BAL: Amount of the given credit.	**PAY_2:** Repayment status in August.	**BILL_AMT2:** Amount of bill statement in August.	**PAY_AMT2:** Amount of previous payment in August.
SEX: Gender: (1=male, 2=Female)	**PAY_3:** Repayment status in July.	**BILL_AMT3:** Amount of bill statement in July.	**PAY_AMT3:** Amount of previous payment in July.
Education: Qualifications.	**PAY_4:** Repayment status in June.	**BILL_AMT4:** Amount of bill statement in June.	**PAY_AMT4:** Amount of previous payment in June.
Marriage: Marital status.	**PAY_5:** Repayment status in May.	**BILL_AMT5:** Amount of bill statement in May.	**PAY_AMT5:** Amount of previous payment in May.
Age: Age of client.	**PAY_6:** Repayment status in April.	**BILL_AMT6:** Amount of bill statement in April.	**PAY_AMT6:** Amount of previous payment in April.

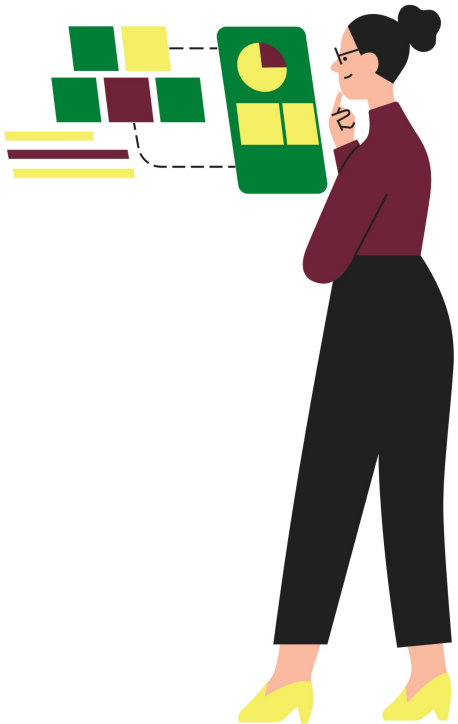
Dataset Summary:

- There are 30000 rows and 25 columns in the given data.
- There are no duplicate or null values present.
- We will have to rename a few variables to understand them easily.
- Education has a few extra values [0,5,6]; we will add them into category [4] since we are only been provided with details for [1,2,3,4].
- Marriage also has an extra value [0]; we will add it into the category [3].

Exploratory Data Analysis:

- Dependent Variable

- Independent Variables
  - Gender.
  - Education.
  - Marriage.
  - Age.
  - LIMIT_BAL
  - History of Past Payment.
  - Amount of Bill Statement.
  - Amount of Previous Payment.
  - Correlation with Dependent and Independent variables.



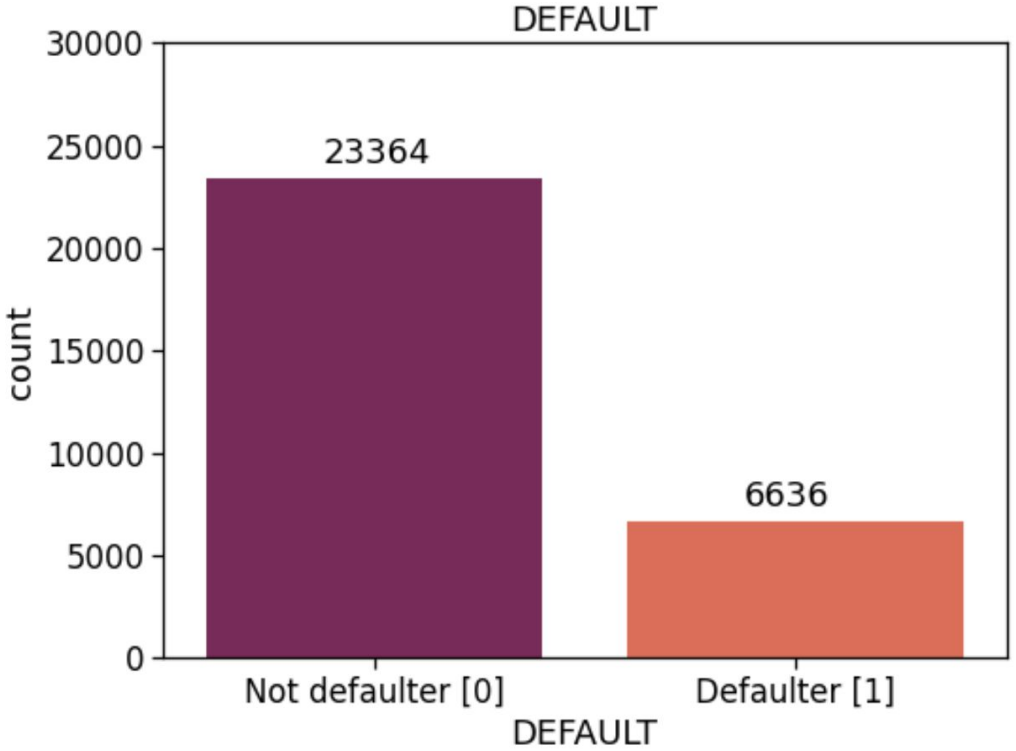
Dependent variable:

DEFAULT:

The data contains 77.88% of Non-Defaulters and 22.12% of Defaulters.

This indicates a Class imbalance in the data.

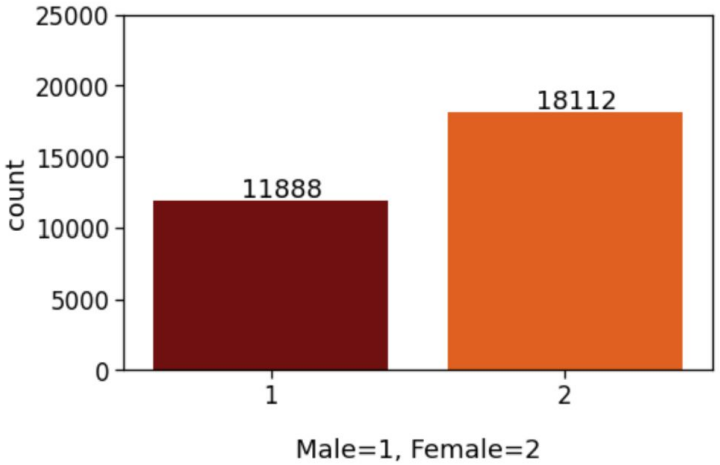
To resolve this issue, we will Oversample the data to balance both categories.

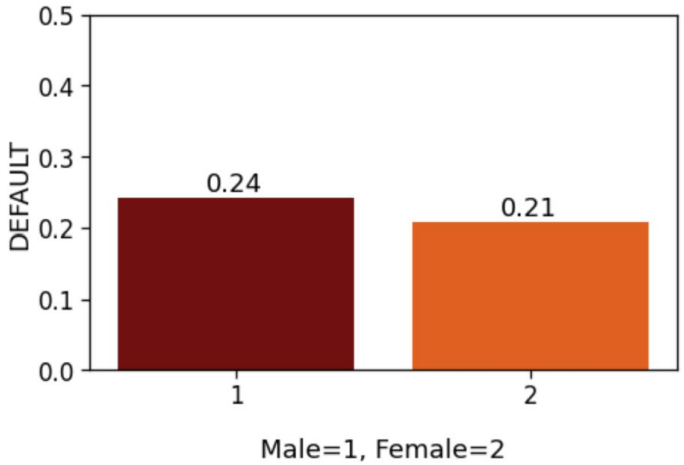


Independent Variables:

Impact of demographic factors on credit card defaults:

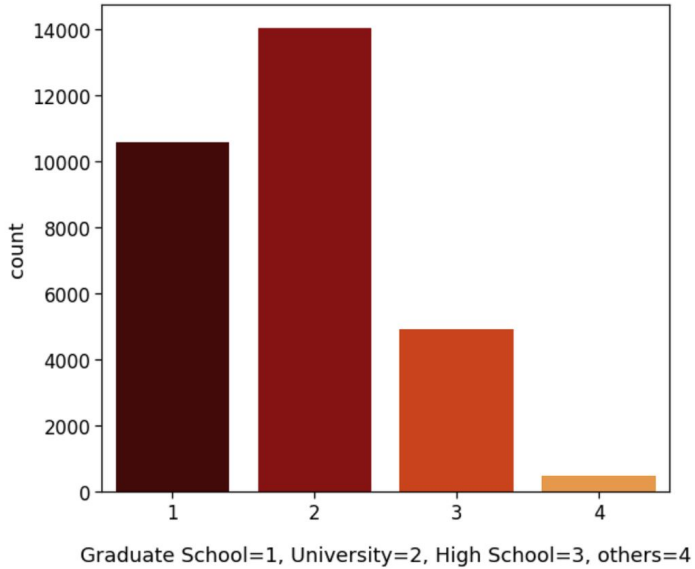
Gender:

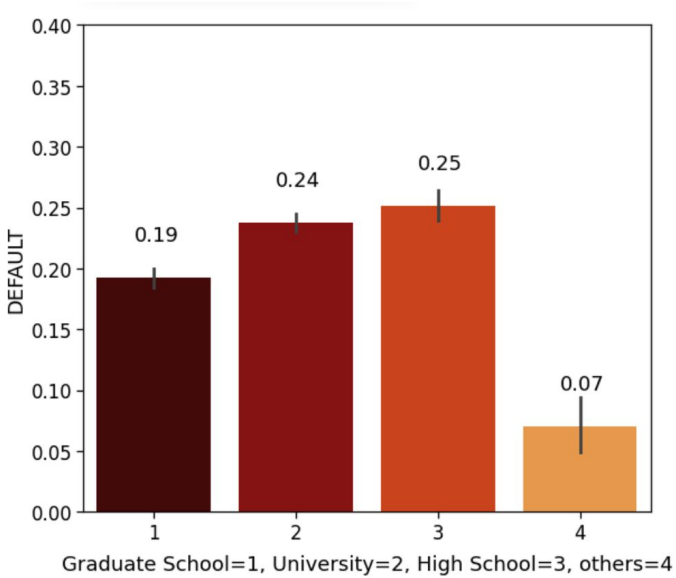




There are more female users compared to males, but the default rate for males is slightly higher than for females.

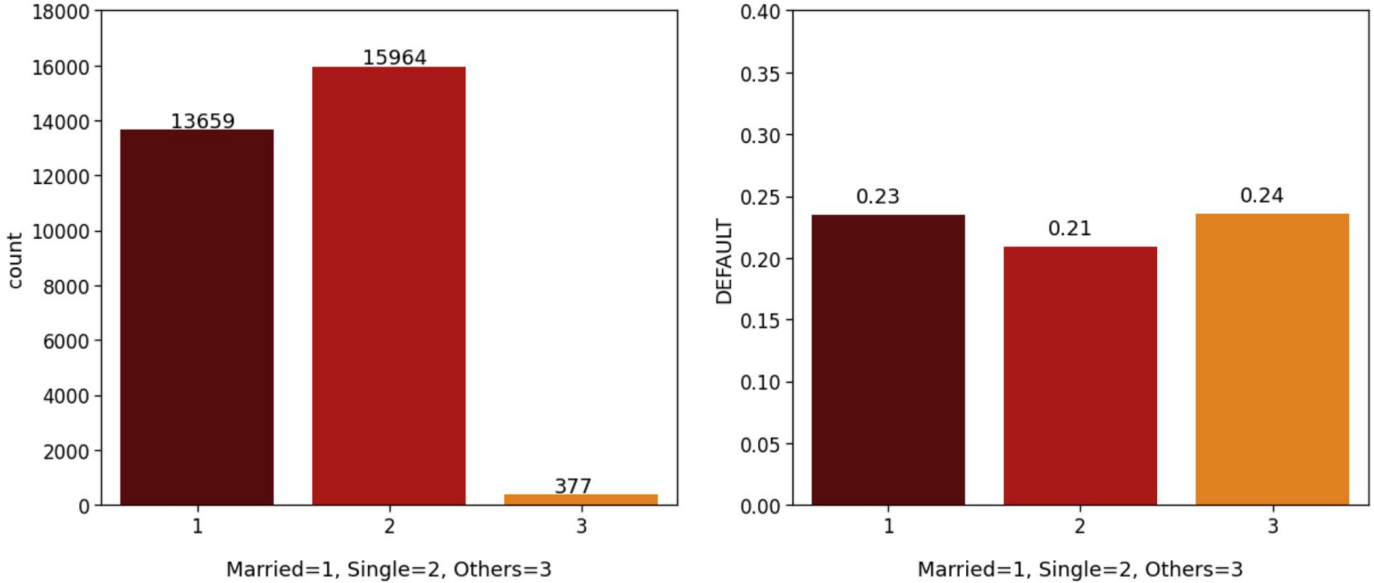
Education:





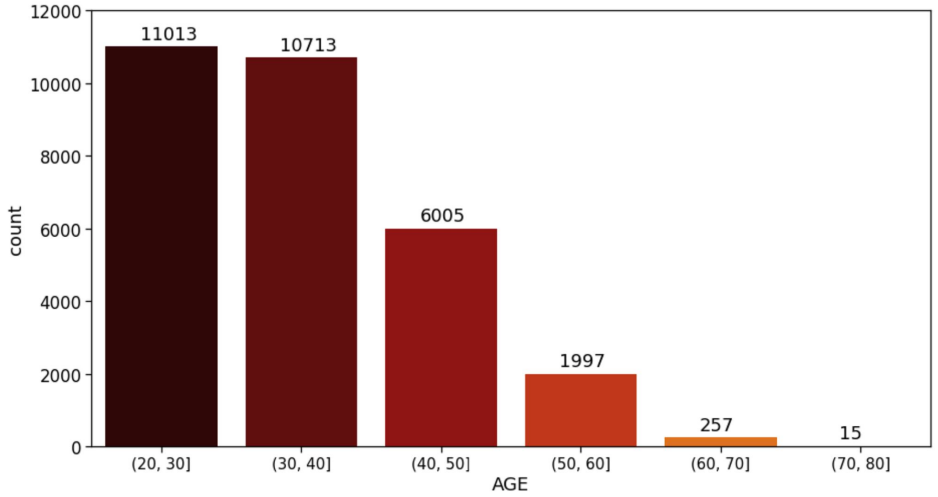
- The maximum number of users have either gone to universities or graduated.
- A higher level of education leads to a lower level of defaults.
- Others have the minimum level of defaults.

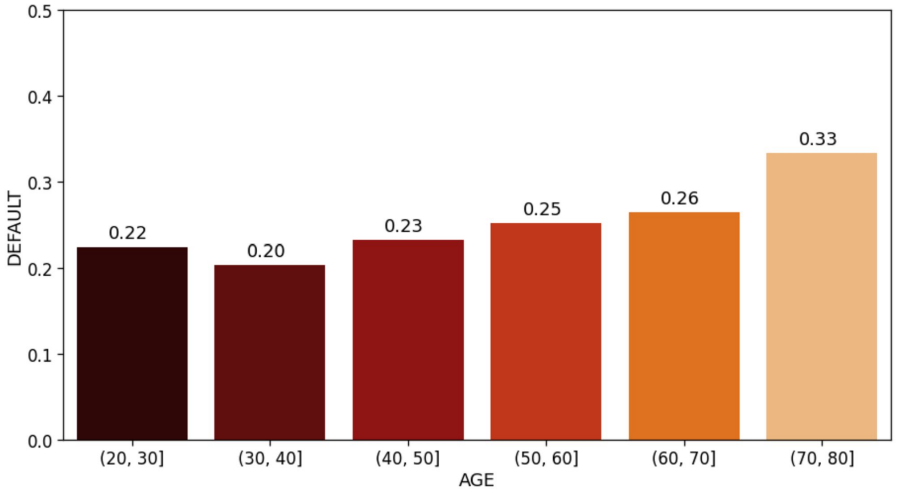
Marriage:



- Most of the users are either single or married.
- Among the three categories, **Others** has the highest percentage of defaults followed by **Married** and **Single** users.

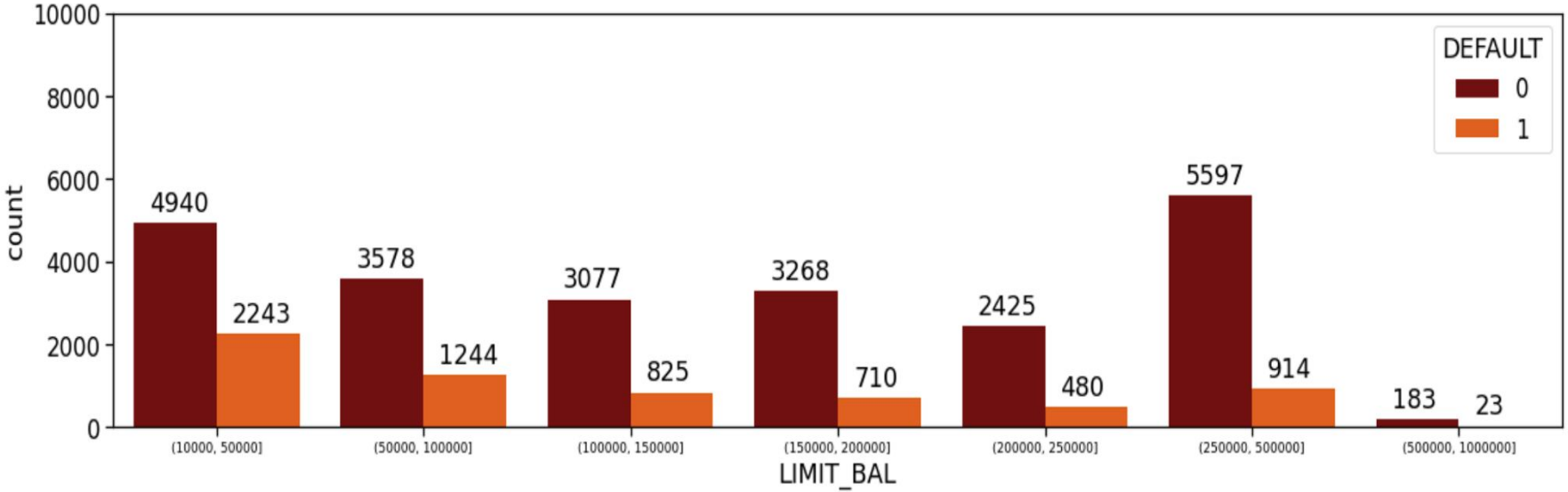
Age:





- The maximum number of users is within the age group 20 to 50.
- As the age increases, default rates increase.
- The least number of defaults are observed in the age group [30,40]

LIMIT_BAL:



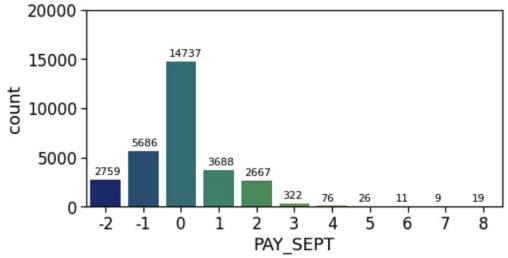
- **LIMIT_BAL** ranges from **10000 to 1000000** with most of the values **less than or equal to 200000.**
- Most of the DEFAULTS are from users within the LIMIT of 10000 to 50000.
- This indirectly means that as the LIMIT increases the number of DEFAULTS decreases.

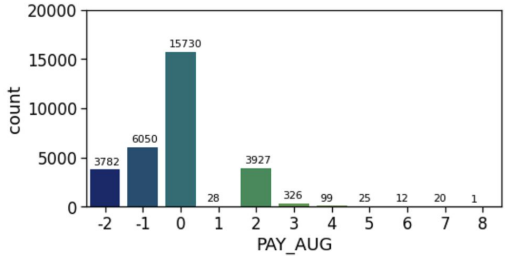
Payment details and Default rates.

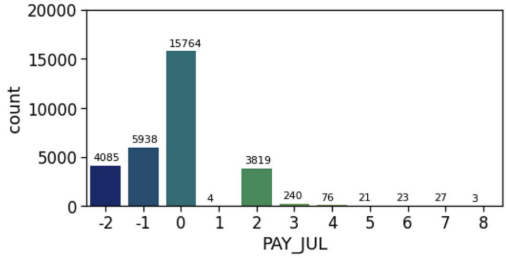
History of Past Payment:

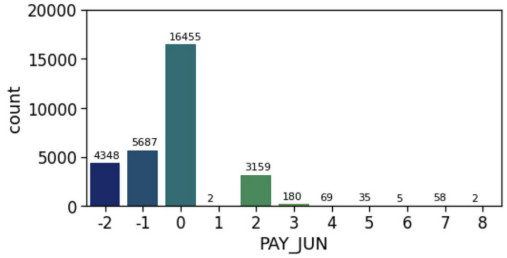
-2 = No consumption,
-1 = paid in full,
0 = use of revolving credit (paid minimum only),
1 = payment delay for one month,
2 = payment delay for two months....

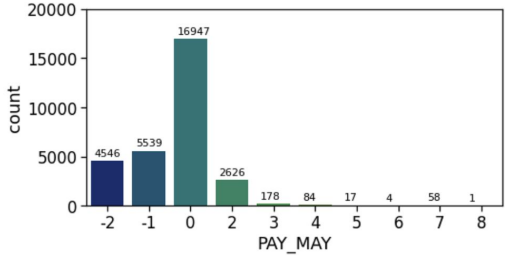
As shown in the plots, most users have paid the minimum amount in all the months.
But few users have delayed their payments for several months.

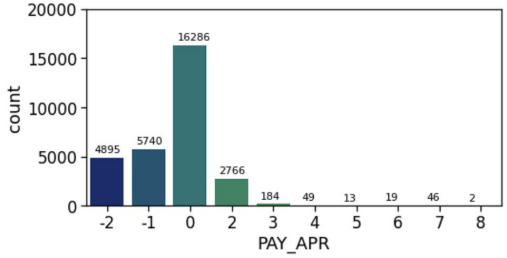






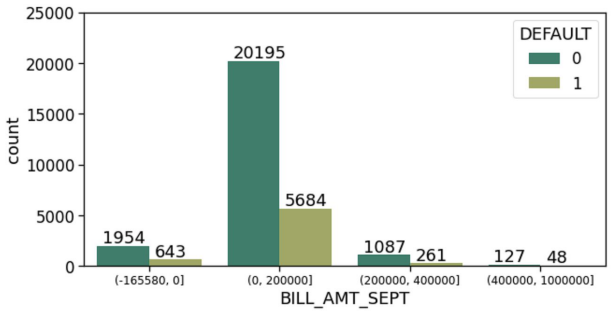


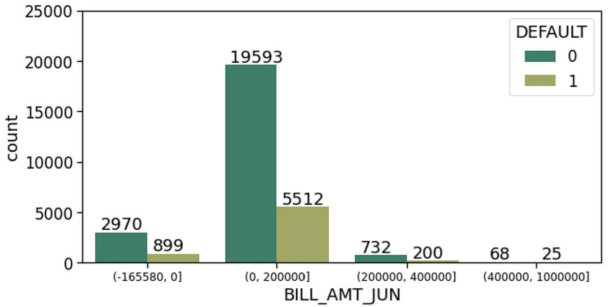


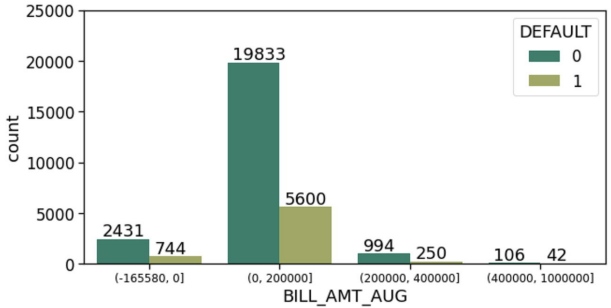


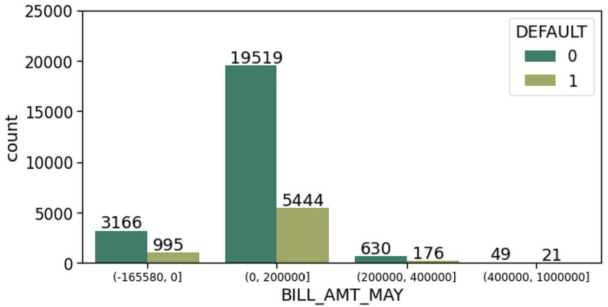
Amount of Bill Statement:

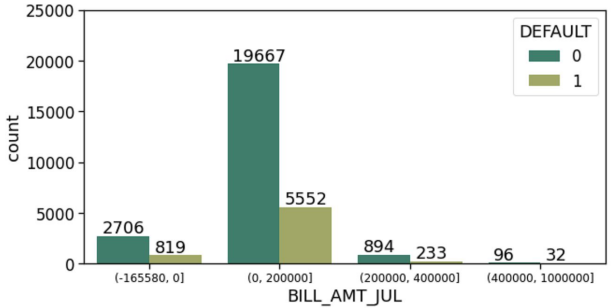
As shown in the plots, There are more DEFAULTS in the range 0 to 200000.

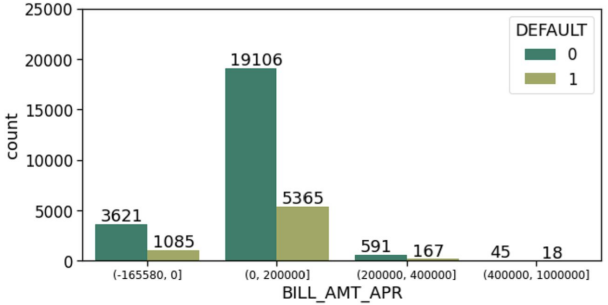








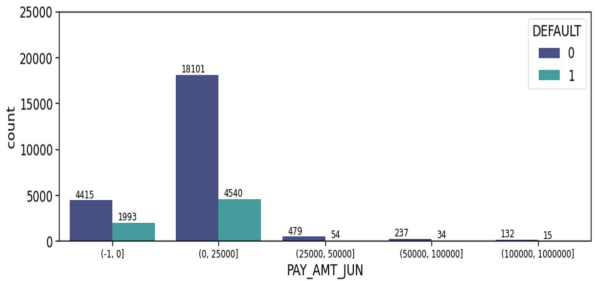


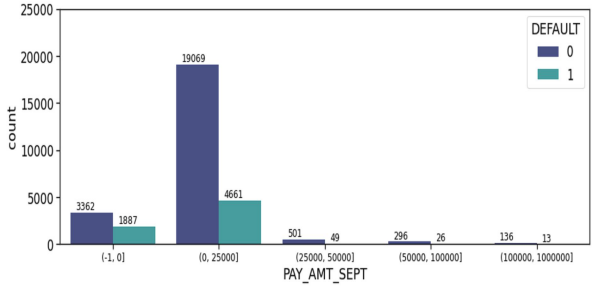


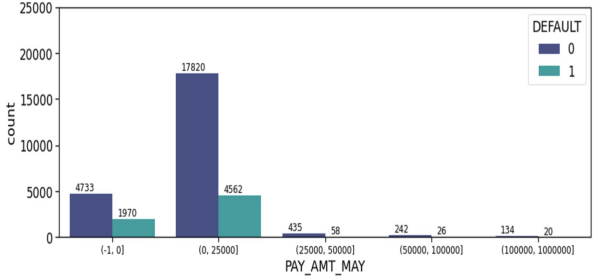
Amount of Previous Payment:

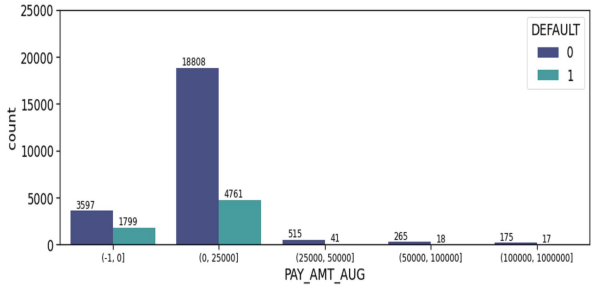
A user is less likely to default when a high amount is paid in the previous month.

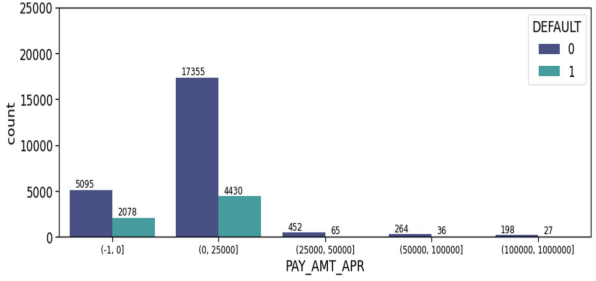
Also most of the defaults are happening in the case where less than 25000 was paid.

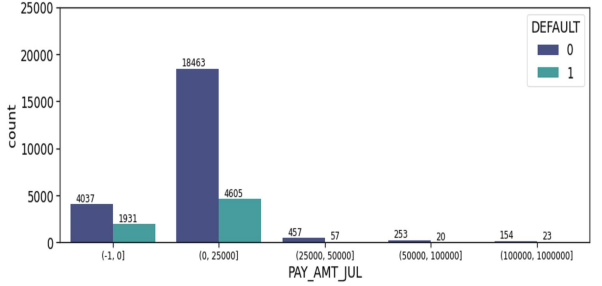






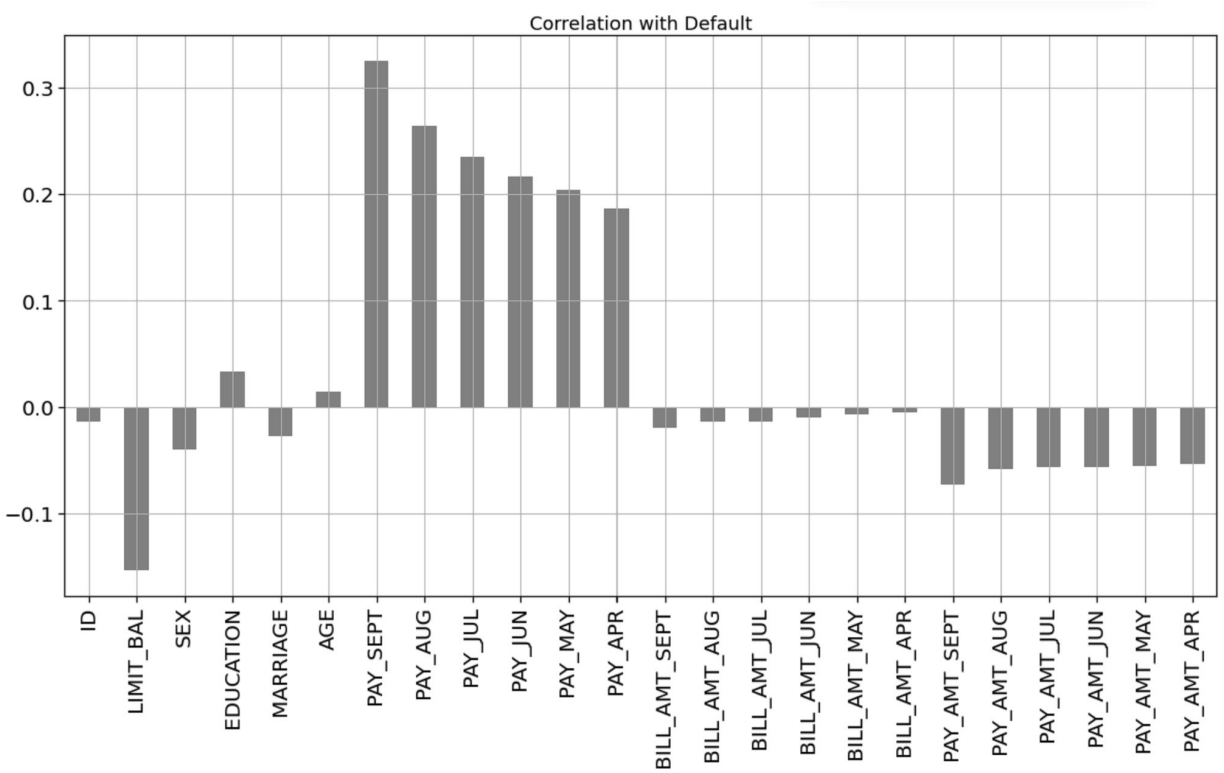




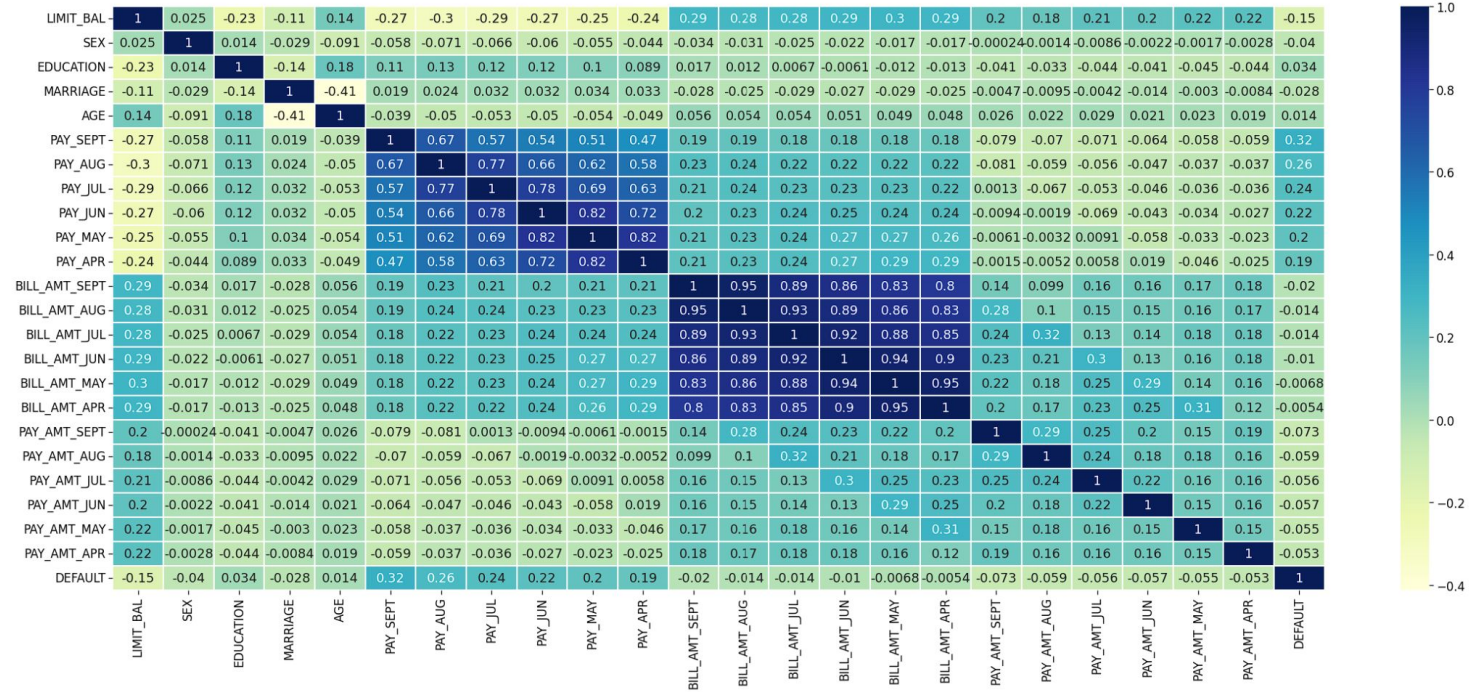


Correlation with dependent variable.

Details of past payments show more correlation with the dependent variable. Education and Age also show a positive correlation, whereas most of the other variables either show no or negative correlation.



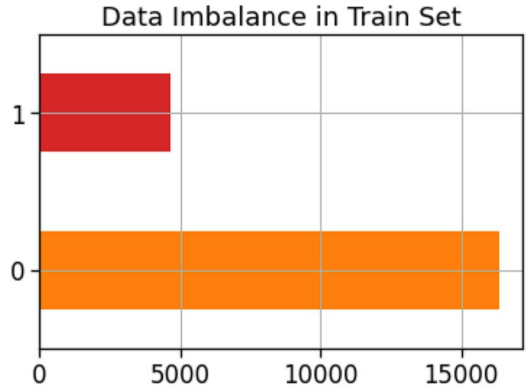
Multicollinearity among variables.

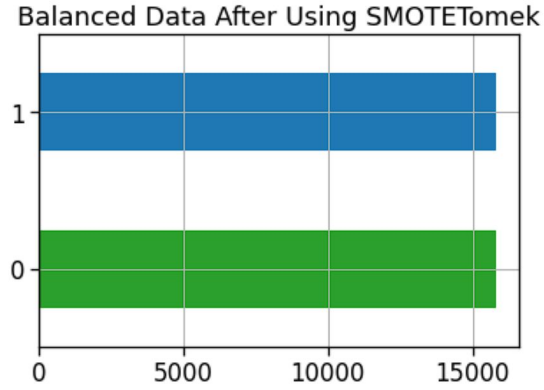


Multicollinearity exists between variables of History of past payment.
Multicollinearity exists between variables of Amount of Bill Statement.
Only few variables seem to be correlated with the dependent variable.

Data Preparation:

We need to prepare the data before we put them through algorithms.
We start by **dropping Id** and performing one-hot encoding for the categorical features.
Next, separate the dependent and the independent variables, where y is the dependent variable DEFAULT and X contains the rest of the features in our dataset.
Now do the train test split to separate the training and the testing data that we will use to build and validate our models.
(70% training and 30% test data.)
We will also oversample our training data using SMOTETomek to remove the class imbalance.
As shown in the bar plots the class imbalance has been treated after using SMOTE.





Model Building and Selection:

Now that we have prepared our data, it's time to use them on algorithms.

The models we will be using are:

- LogisticRegression
- XGBClassifier
- KNeighborsClassifier
- RandomForestClassifier
- ExtraTreesClassifier

We will compare them and select the best-performing model for the classification.

Imbalanced Data:

	Model Name	Train Accuracy	Test Accuracy	Precision Train	Precision Test	Recall Train	Recall test	ROC AUC Train	ROC AUC Test	AUC Train	AUC Test	F1 Score Train	F1 Score Test
0	LogisticRegression	0.82	0.82	0.68	0.69	0.36	0.36	0.65	0.66	0.654963	0.655654	0.47	0.47
1	XGBClassifier	0.77	0.76	0.48	0.46	0.66	0.63	0.73	0.71	0.730832	0.710229	0.56	0.53
2	KNeighborsClassifier	0.85	0.79	0.73	0.54	0.48	0.35	0.71	0.63	0.713114	0.634276	0.58	0.43
3	RandomForestClassifier	1.00	0.82	1.00	0.66	1.00	0.37	1.00	0.66	0.998970	0.658223	1.00	0.47
4	ExtraTreesClassifier	1.00	0.81	1.00	0.60	1.00	0.36	1.00	0.65	0.998816	0.646680	1.00	0.45

Oversampled Data:

	Model Name	Train Accuracy	Test Accuracy	Precision Train	Precision Test	Recall Train	Recall test	ROC AUC Train	ROC AUC Test	AUC Train	AUC Test	F1 Score Train	F1 Score Test
0	LogisticRegression	0.72	0.77	0.78	0.48	0.60	0.59	0.72	0.71	0.715081	0.705841	0.68	0.53
1	XGBClassifier	0.81	0.80	0.86	0.55	0.74	0.51	0.81	0.69	0.811108	0.694797	0.80	0.53
2	KNeighborsClassifier	0.87	0.67	0.81	0.36	0.96	0.62	0.87	0.65	0.869054	0.650029	0.88	0.45
3	RandomForestClassifier	1.00	0.80	1.00	0.55	1.00	0.49	1.00	0.69	0.999779	0.686130	1.00	0.52
4	ExtraTreesClassifier	1.00	0.79	1.00	0.53	1.00	0.44	1.00	0.66	0.999779	0.661981	1.00	0.48

On both, Imbalanced and Oversampled data, Logistic regression, XGB, and Random Forest Classifiers give better results than others.

Using XG Boost with imbalanced data while applying the scale_pos_weight=3.521 gives the best recall of **63%.**

Random Forest is overfitting the data which can be solved by tuning the hyperparameters.

Hyperparameter tuning and cross validations.

To perform the hyperparameter tuning and cross-validations, we will use Grid Search CV and Randomized Search CV. Three Cross validations for each set will be conducted to find the best parameters.

	Model Name	Accuracy Scores	ROC AUC Scores	Precision	Recall	F1 Score
0	Logistic Regression	0.760556	0.703333	0.455036	0.362632	0.530858
1	XGB with Imbalanced Data	0.751000	0.709604	0.455036	0.635359	0.530287
2	XGB with SMOTETomek	0.820000	0.656277	0.672880	0.362632	0.471279
3	Random Forest Classifier	0.792222	0.698853	0.530326	0.531391	0.530858

XG Boost Classifier with SMOTETomek gives the best accuracy of **82%** but lacks recall which is crucial in classifying the defaulters.

Using XG Boost with imbalanced data while applying the scale_pos_weight=3.521 gives the best recall of **64%** approx.

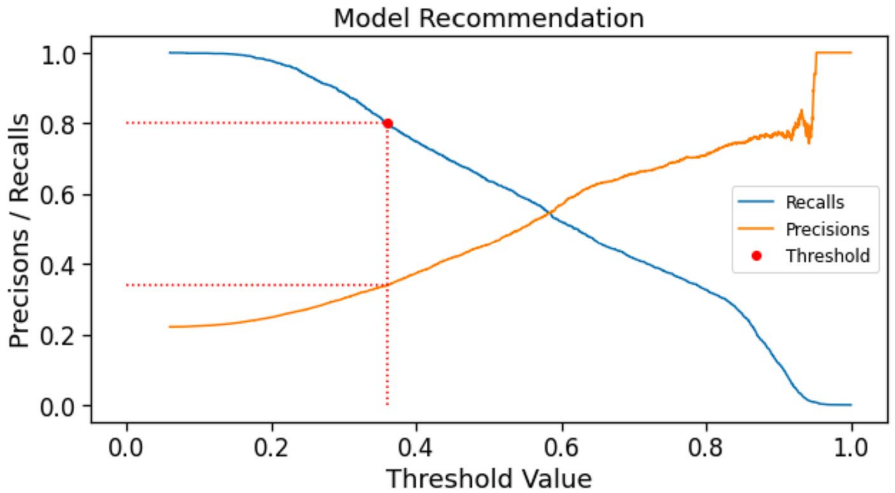
In terms of overall balance, Random Forest Classifier provides the best results and logistic regression the worst.

Threshold Moving:

From a business perspective, if a higher recall is required, it can be achieved by changing the threshold (default threshold = 0.5) for classifying the defaults.

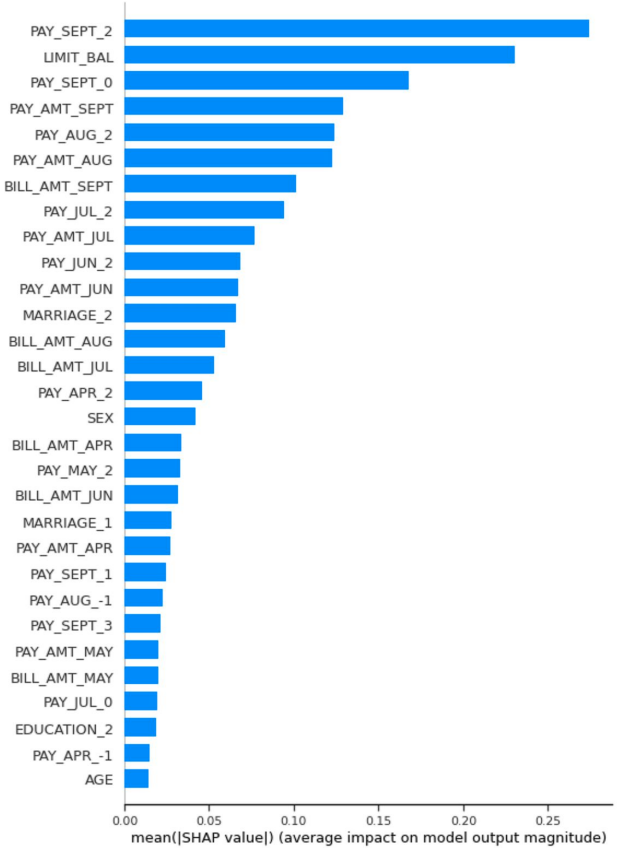
```
Recall after threshold moving: 0.7995981918633852
Precision after threshold moving: 0.33995302156737134
Threshold after threshold moving: 0.3605385
```

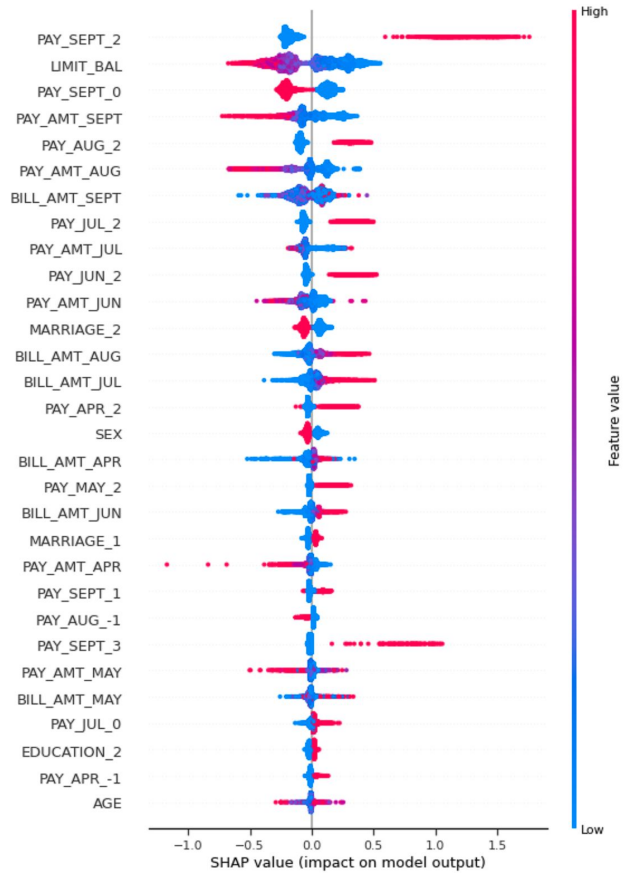
An improved recall of 80% approx can be obtained after changing the threshold value to 0.3605385.
As a result, we have lower precision.



Feature importance using SHAP:

The top 10 features crucial in classifying the DEFAULTS are **PAY_SEPT_2**, **LIMIT_BAL**, **PAY_SEPT_0**, **PAY_AMT_SEPT**, **PAY_AUG_2**, **PAY_AMT_AUG**, **BILL_AMT_SEPT**, **PAY_JUL_2**, **PAY_AMT_JUL**, **PAY_JUN_2**.





Conclusions:

- PAY_SEPT_0, PAY_SEPT_2, and LIMIT_BAL are the most vital features in predicting the defaults.
- XG Boost Classifier with SMOTETomek gives the best accuracy of 82% but lacks Recall which is crucial in classifying the defaulters, whereas **XG Boost** with imbalanced data while applying the scale_pos_weight=3.521 gives the best Recall of **64%** approx.
- Higher recall can be achieved if low precision is acceptable.
- An improved recall of 80% approx can be obtained after changing the threshold value to 0.3605385.
- In terms of overall balance, Random Forest Classifier provides the best results.

Challenges Faced:

- Data imbalance had to be treated keeping in mind not to lose anything of value.
- Data had to split and oversampled carefully to avoid data leakage.
- Tuning the hyperparameters in order to achieve best results was time consuming.

Thank You.