Capstone Project

(SUPERVISED ML – CLASSIFICATION)

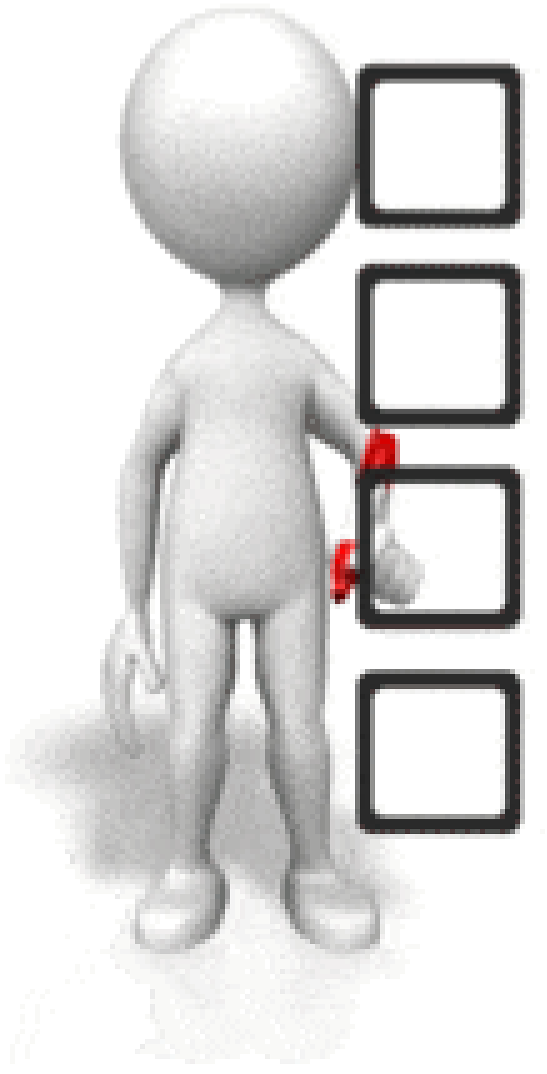
Credit Card Default Prediction

INDIVIDUAL PROJECT

- RAJAKUMARAN S

~ UNDER THE GUIDANCE OF ALMABETTER TEAM

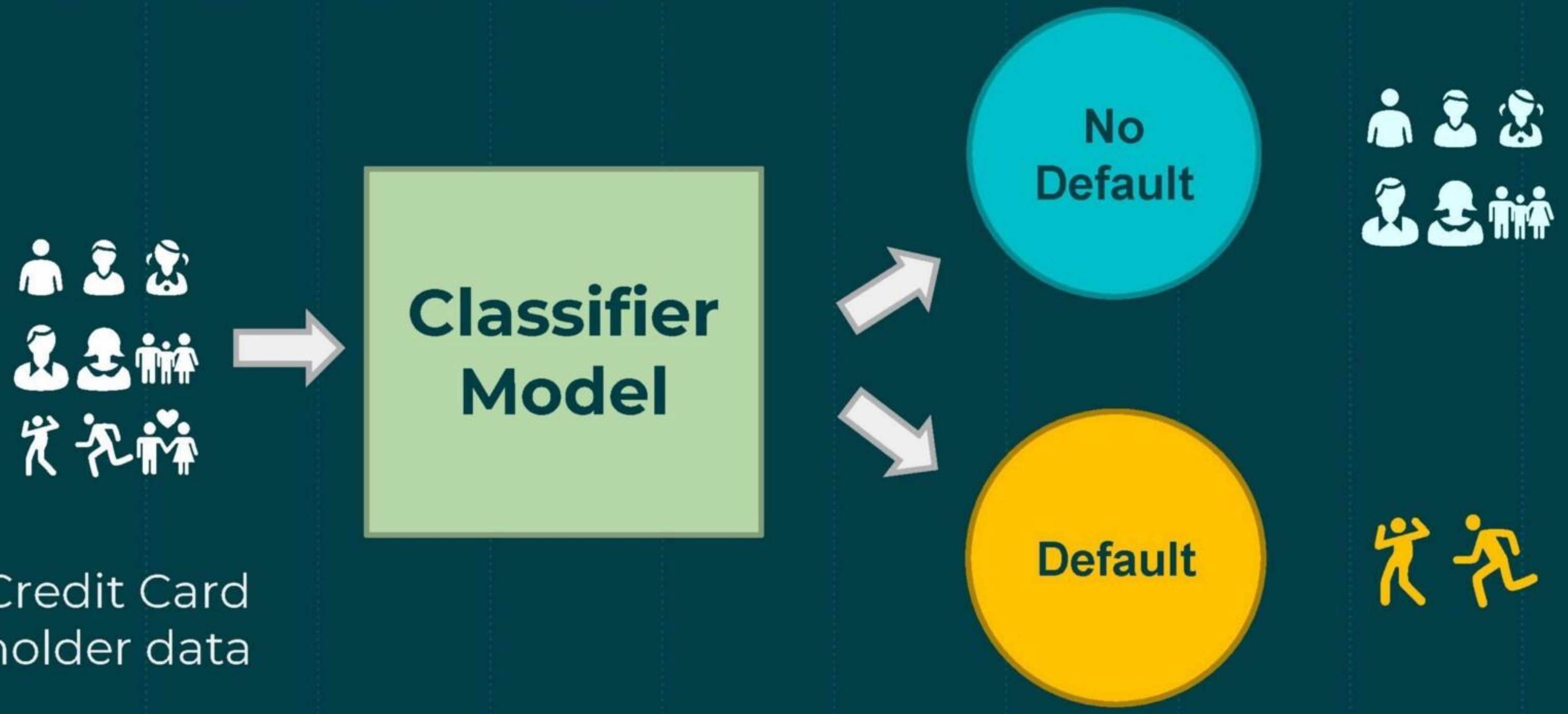
CONTENT



PROBLEM STATEMENT

- This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

OBJECTIVE



Credit Card holder data

ROAD MAP

EDA
- Data distribution of features
- Check Correlation
- Feature Engineering

Model building
- Using SMOTE Technique
- Data transformation
- Fitting & Prediction
- Evaluation matrices

Model validation
- Model selection
- Feature importance
- Conclusion

INTRODUCTION

The basic idea of this capstone project is to use the Supervised Machine Learning - Classification to predict customers default payments in Taiwan. Here we have previous 6 month transaction bills and statements as our major information to classify defaulter.

Based on these features we will be predicting our target variable i.e. credit card defaulters. By using concepts like model validation, we will came to know which features are important and how much they contribute to our target variable.

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: (1 = graduate school, 2 = university, 3 = high school, 0,4,5,6 = others)*
- *MARRIAGE: Marital status (1 = married, 2 = single, 3 = others)*
- *AGE: Age in years*
- Scale for PAY_0 to PAY_6 :

  (-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,

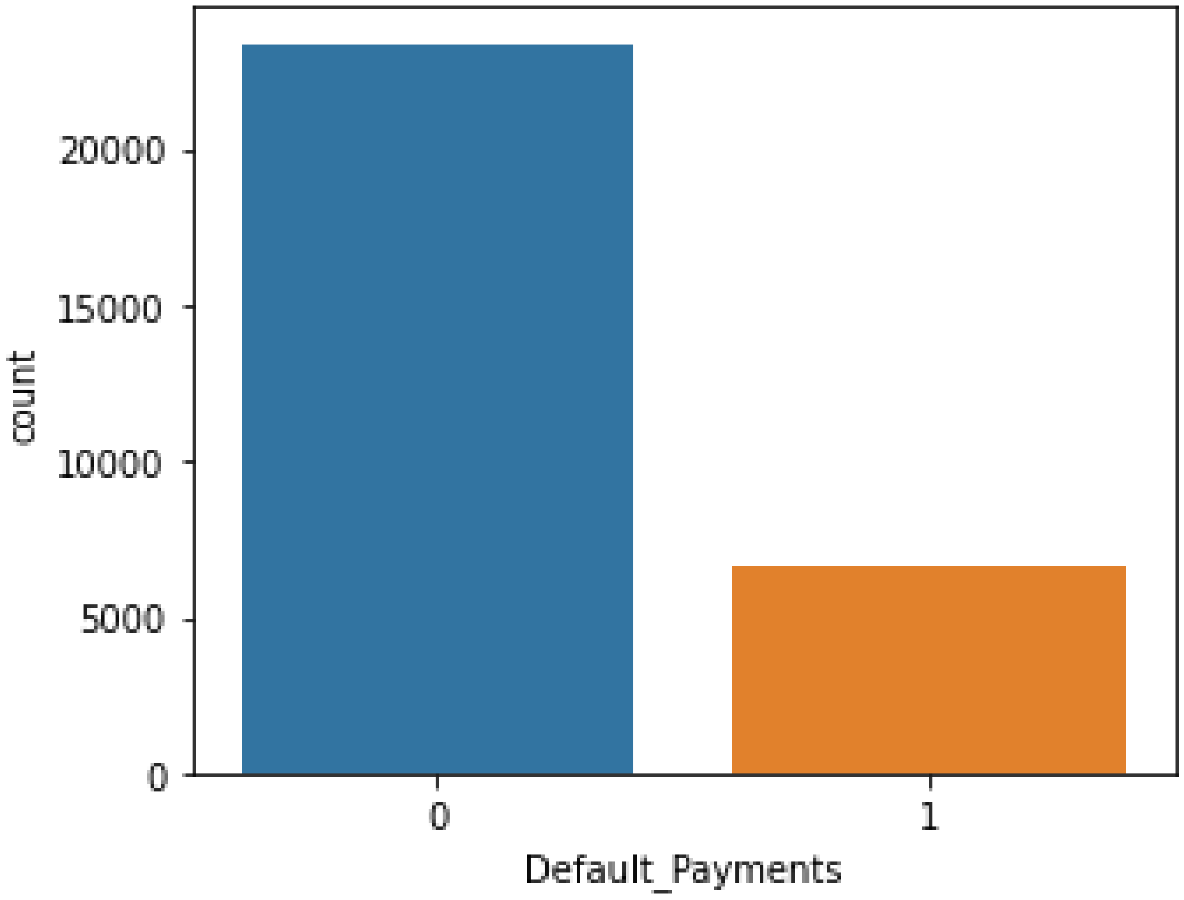
  ... 8 = payment delay for eight months, 9 = payment delay for nine months and above)
- *PAY_0 to PAY_6: Repayment status in (September, 2005), (August, 2005).....(April, 2005)*
- *BILL_AMT1 to BILL_AMT6: Amount of bill statement in (September, 2005), (August, 2005).....(April, 2005)*
- *PAY_AMT1 to PAY_AMT6: Amount of previous payment in (September, 2005), (August, 2005).....(April, 2005)*
- *Default payment next month: Default payment (1=yes, 0=no)*

EDA

Data distribution of target variable

- ❑ Non-Defaulter(0) -**78%**
- ❑ Defaulter(1) – **22%**

22% of customers has default payment next month and We Have Imbalance Dataset

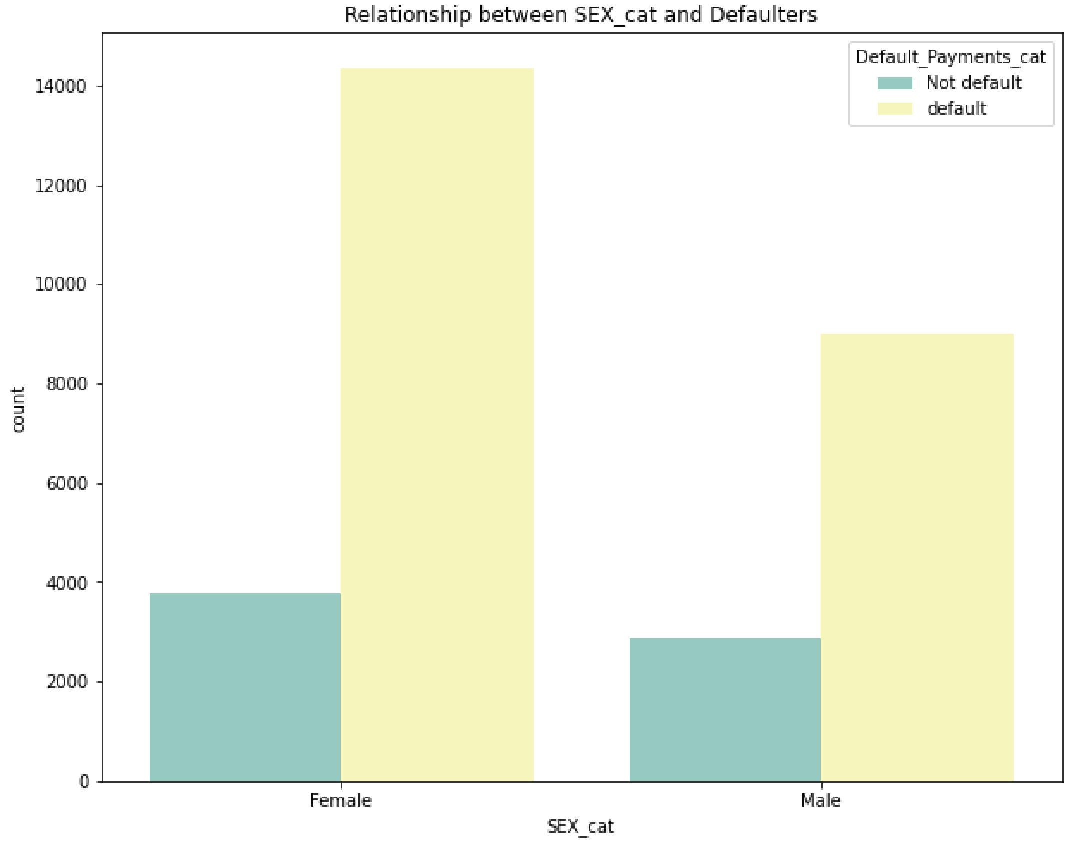


EDA

SEX Column According to Target Variable

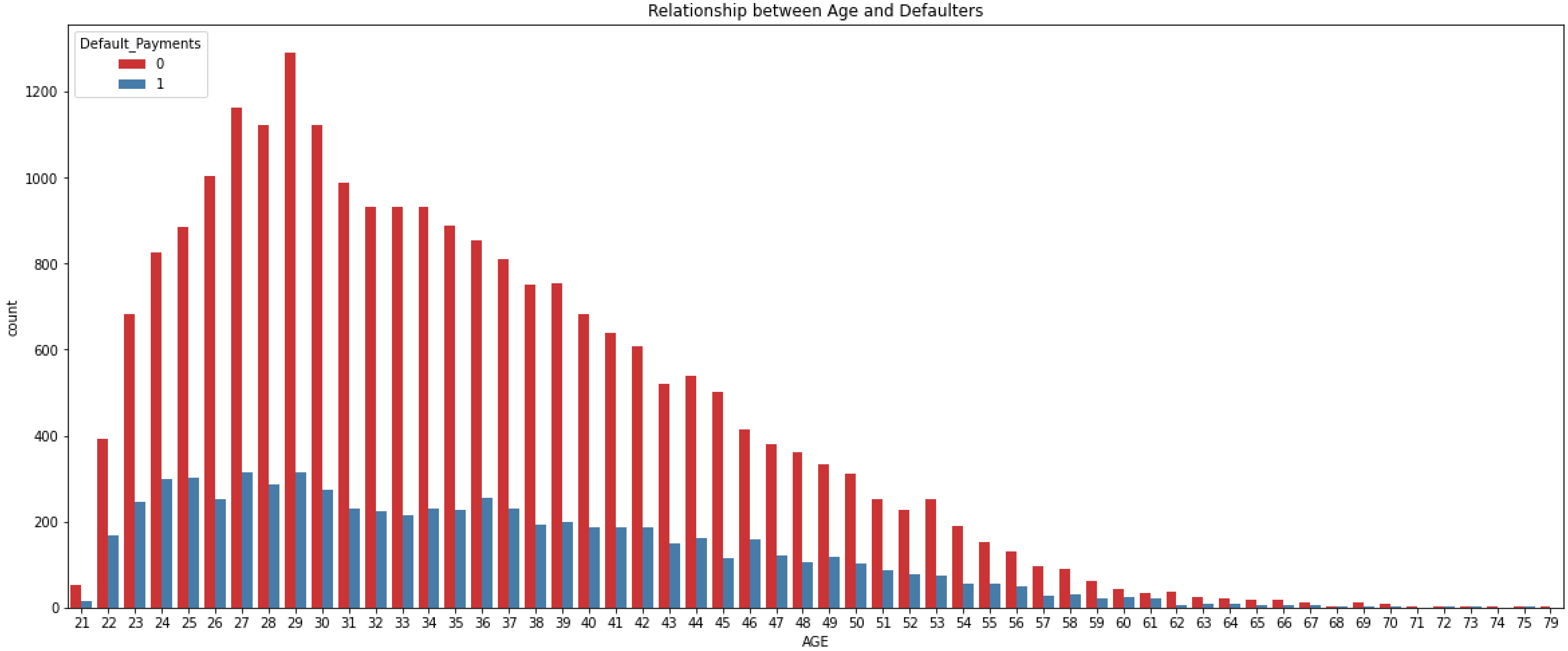
SEX	Male	Female	All
Default_Payments			
Non-default	0.758328	0.792237	0.7788
Default	0.241672	0.207763	0.2212
All	1.000000	1.000000	1.0000

Female : Non Default - 76%, Default 24%
Male : Non Default - 78%, Default 22%
Females have lower default risk than males in this dataset.



EDA

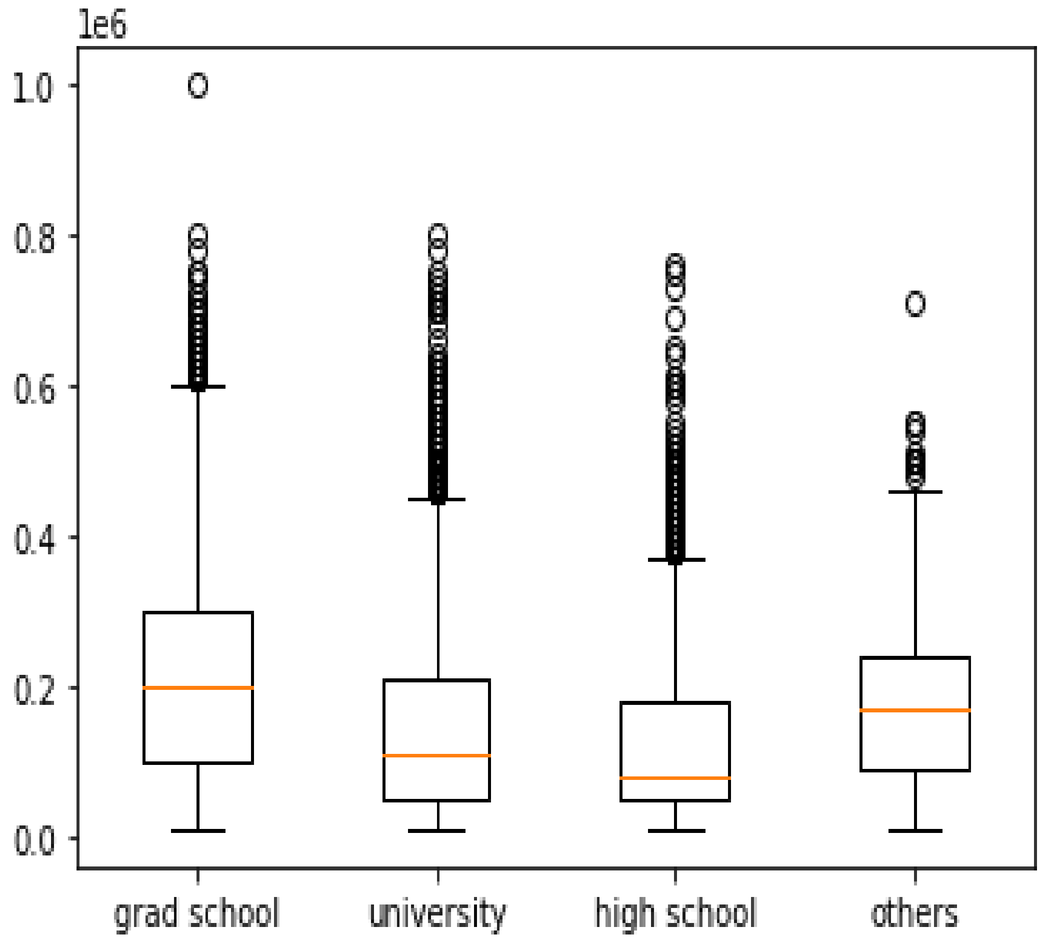
Analysis on AGE feature

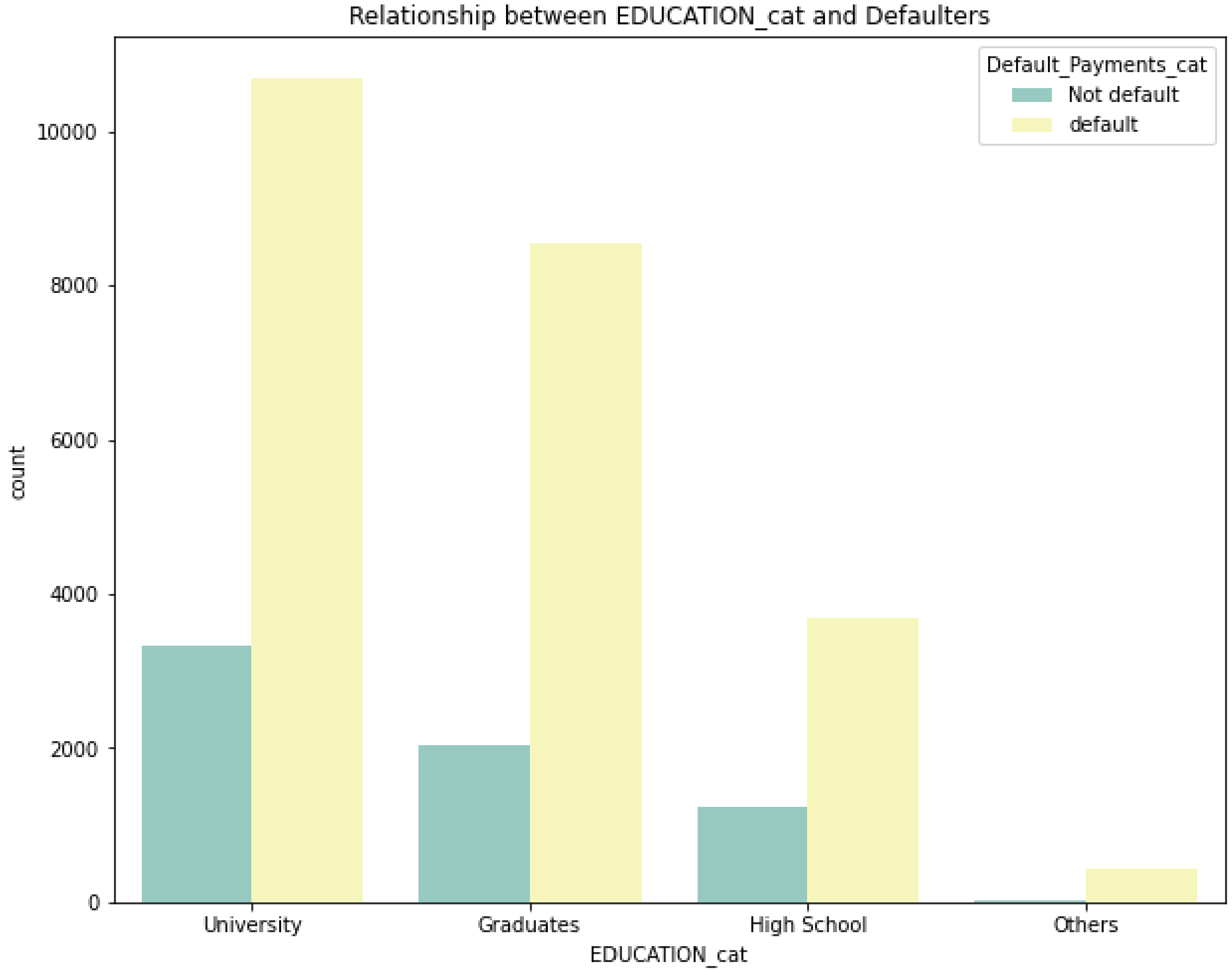


- ❑ 20 to 45 years customer are on average for defaulters
- ❑ Age above 60 years are almost defaulters

EDA

Analysis on Education feature



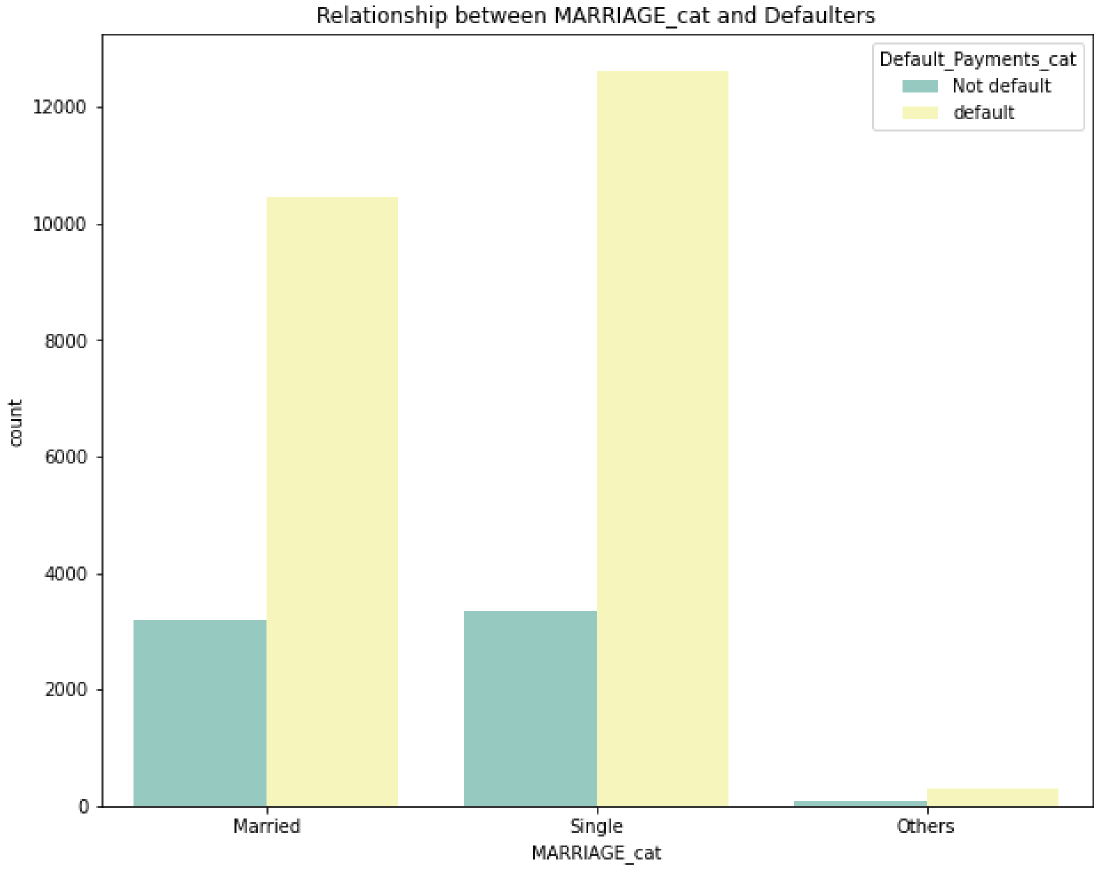


- Customer which had education at University level has more user as well as defaulters

EDA

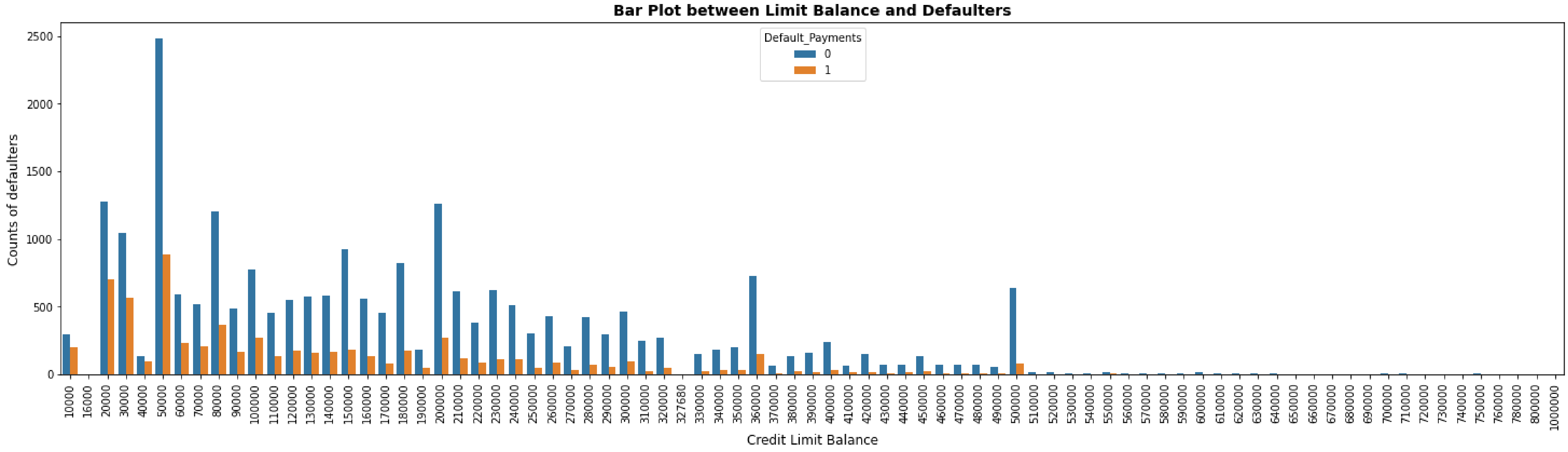
Analysis on MARRIAGE feature

- ❑ Married customer count is greater of all
- ❑ Married and single defaulter customers does not have much difference but, married customers takes lead for defaulters



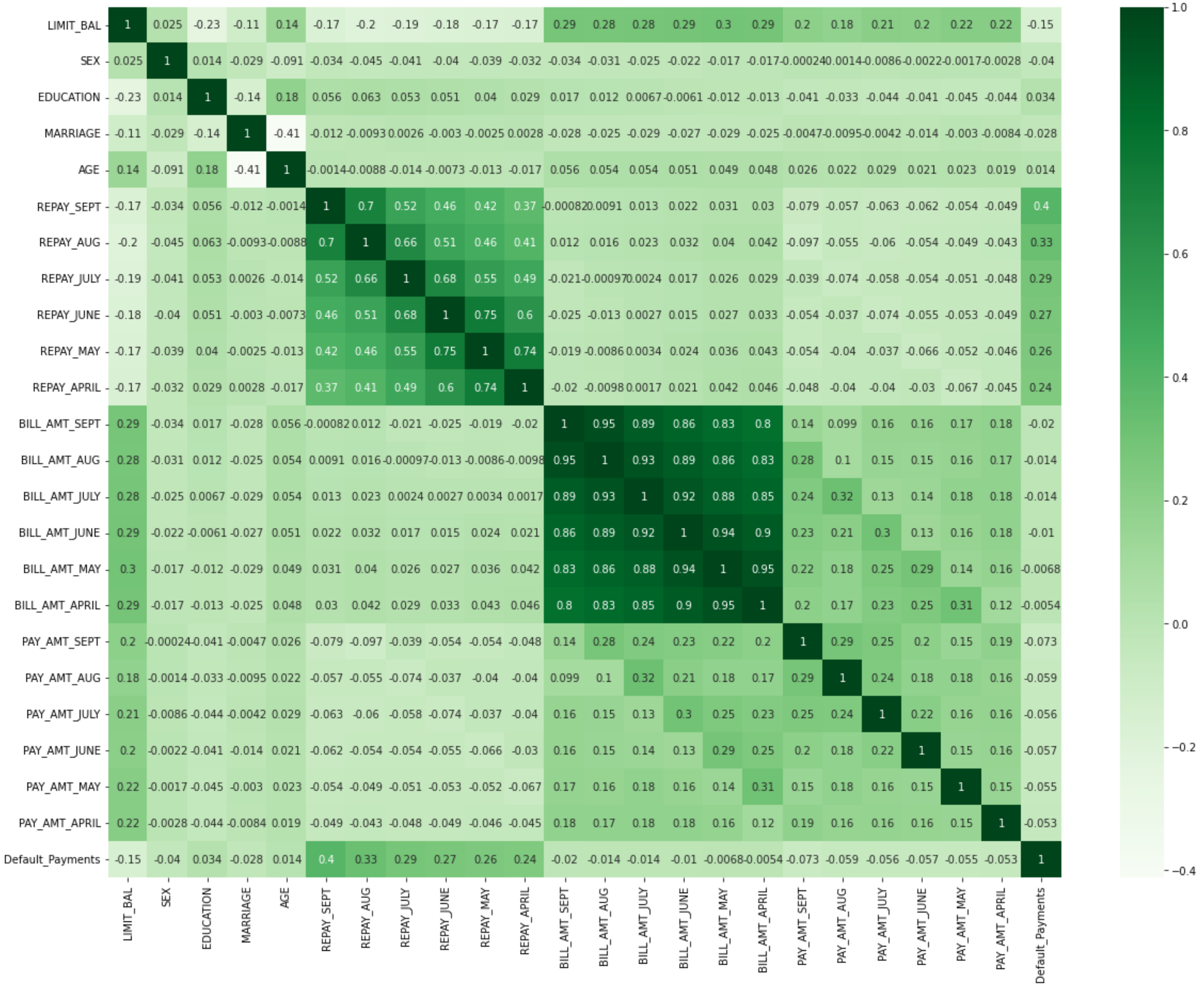
EDA

Analysis on Limit Balance feature



- ❑ Most of the Defaulter are lies between 20K to 50K. And Then Average defaulter are in between 60K to 3L.

Correlation Heat Map



- Here most of the categories have correlated with each other, because all those are previous transaction of customer
- Bill amount of 6 months have high Correlation with each other.

Feature Engineering:

STANDARD SCALER:

Scaling Independent variable with StanardScaler()

SMOTE - Synthetic Minority Oversampling Technique:

Over Sampling of Target Variable with SMOTE Technique to over come Imbalance of the Dataset.

TRAIN TEST SPLIT:

Splitting Dataset into Training Dataset for Model Training.

Testing Dataset for Model Testing.

Target Variable before SMOTE

Target Variable after SMOTE

FITTING VARIOUS MODEL

1. Logistic Regression
2. Support Vector Classifier
3. K-Nearest Neighbors Classifier
4. Random forest Classifier
5. XG Boosting Classifier

MODEL PERFORMANCE COMPARISION

Evaluation matrices for all the models without Hyperparameter Tuning

	Model	accuracy	precision	recall	f1_score_	roc_auc_score
0	Logistic Regression_train	70.49	78.06	56.71	65.70	70.44
1	Logistic Regression_test	70.08	77.76	56.91	65.72	70.19
2	Support Vector Machine_train	72.35	77.94	62.08	69.11	72.31
3	Support Vector Machine_test	71.37	76.94	61.71	68.49	71.45
4	KNN_train	85.01	79.77	93.68	86.17	85.04
5	KNN_test	76.79	72.21	87.72	79.21	76.70
6	RandomForest_train	99.97	99.98	99.96	99.97	99.97
7	RandomForest_test	84.34	86.35	81.87	84.05	84.36
8	XGBClassifier_train	79.48	84.76	71.70	77.69	79.45
9	XGBClassifier_test	78.33	83.42	71.15	76.80	78.39

- Logistic Regression, SVM and XGB Classifier have good performance.
- KNN and Random Forest Overfitting with the model.

MODEL PERFORMANCE COMPARISION

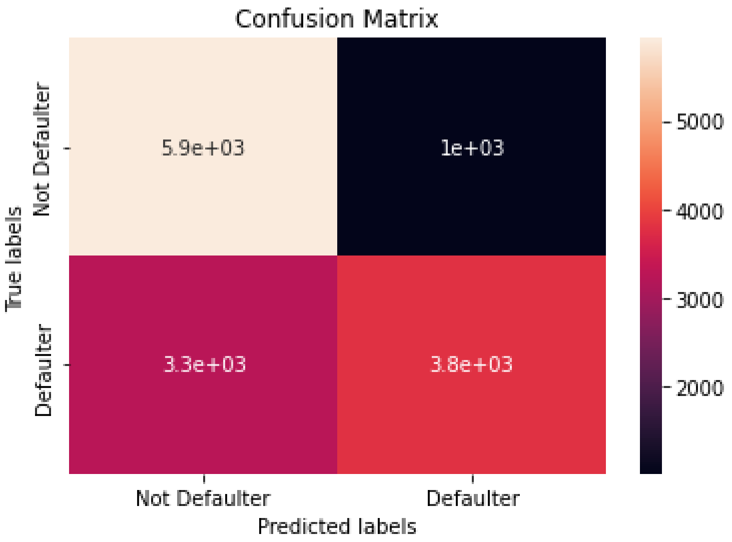
Evaluation matrices for all the models with Hyperparameter Tuning

Model	accuracy	precision	recall	f1_score_	roc_auc_score
Logistic_train_hyper	70.17	79.44	54.13	64.39	70.11
Logistic_test_hyper	69.48	78.88	53.90	64.04	69.61
knn_train_hyper	89.36	84.21	96.79	90.07	89.39
knn_test_hyper	78.59	73.60	89.71	80.86	78.50
Random Forest_train_hyper	76.63	81.70	68.42	74.47	76.60
Random Forest_test_hyper	74.76	79.56	67.19	72.85	74.82
XG Boosting Train_hyper	79.48	84.76	71.70	77.69	79.45
XG Boosting Test_hyper	78.33	83.42	71.15	76.80	78.39

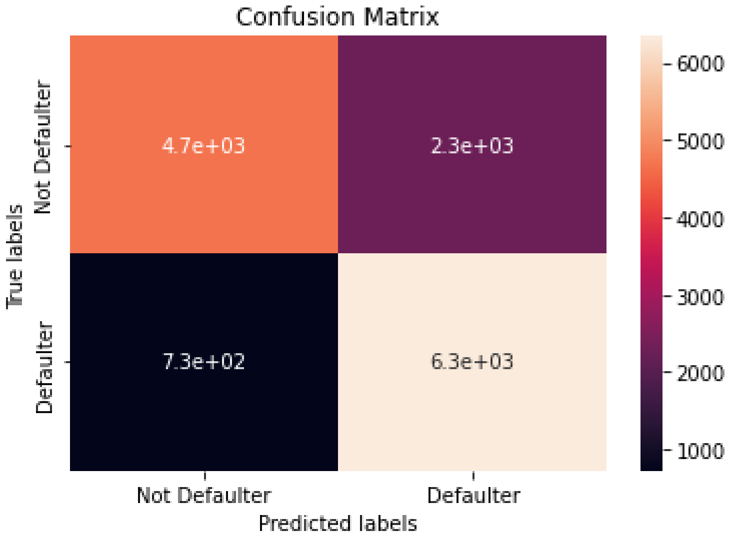
- ❑ Random Forest and XGB Classifier have good performance.
- ❑ KNN Overfitting with the model and Logistic Regression have low performance compared to others.

Confusion matrices of all Testing Model

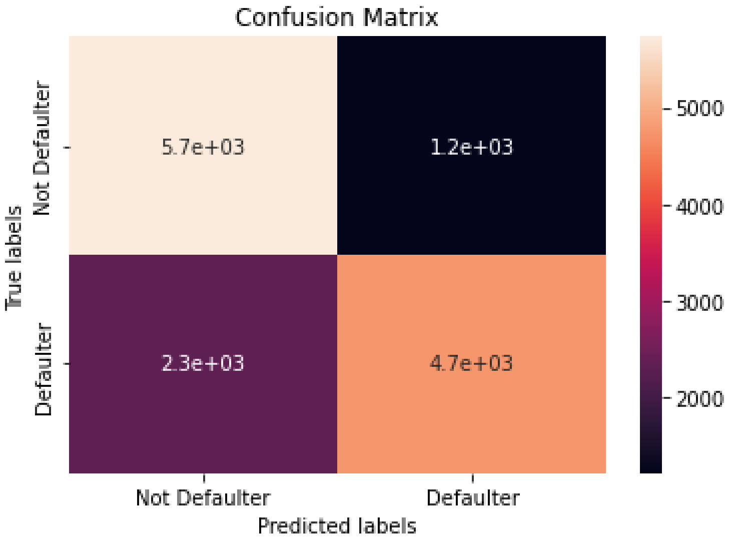
LR



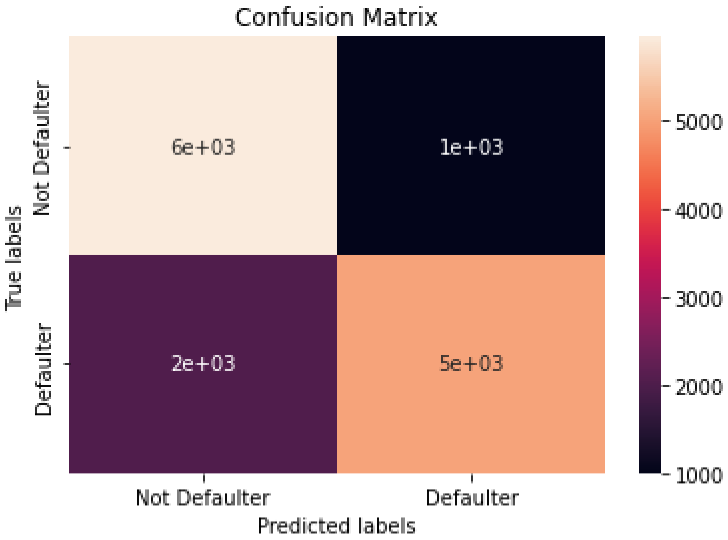
KNN



RF

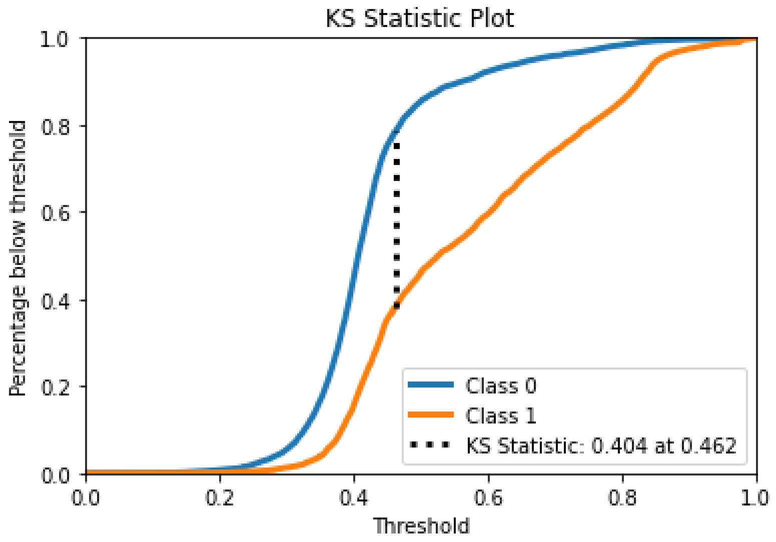


XGB

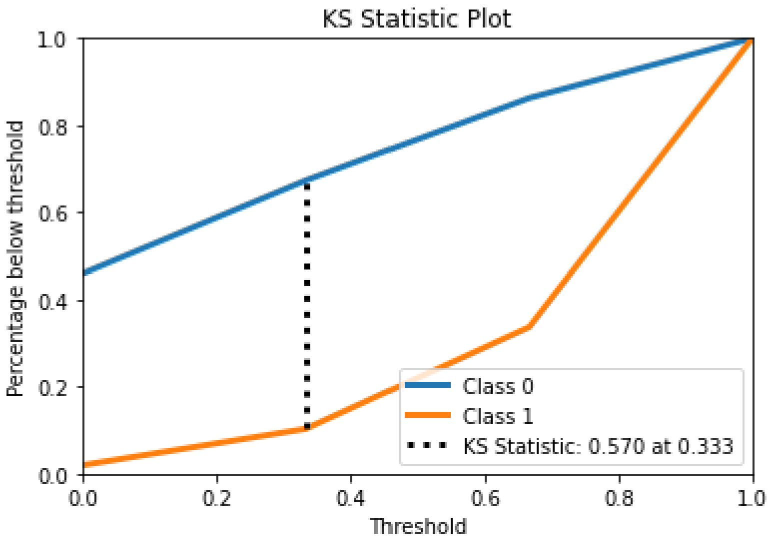


KS Statistics of all Testing Model

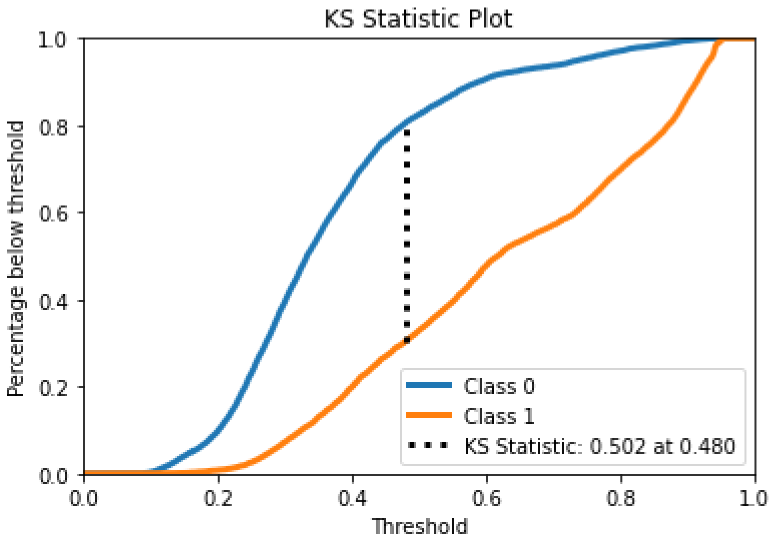
LR



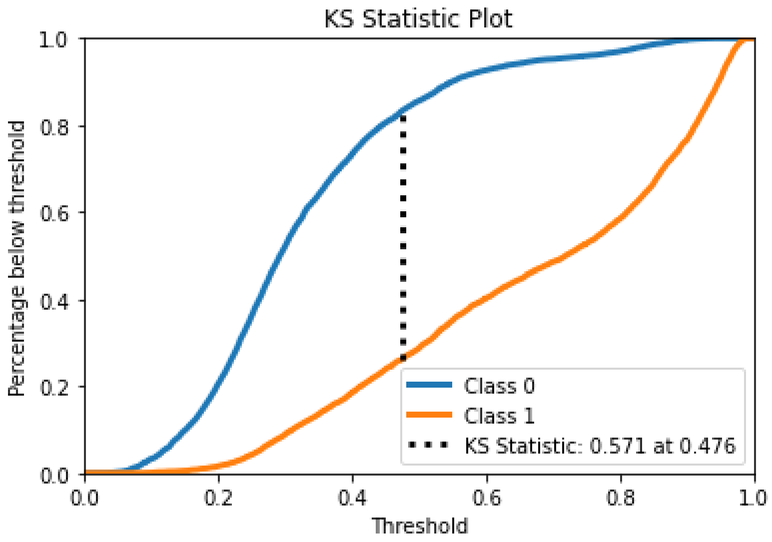
KNN



RF



XGB



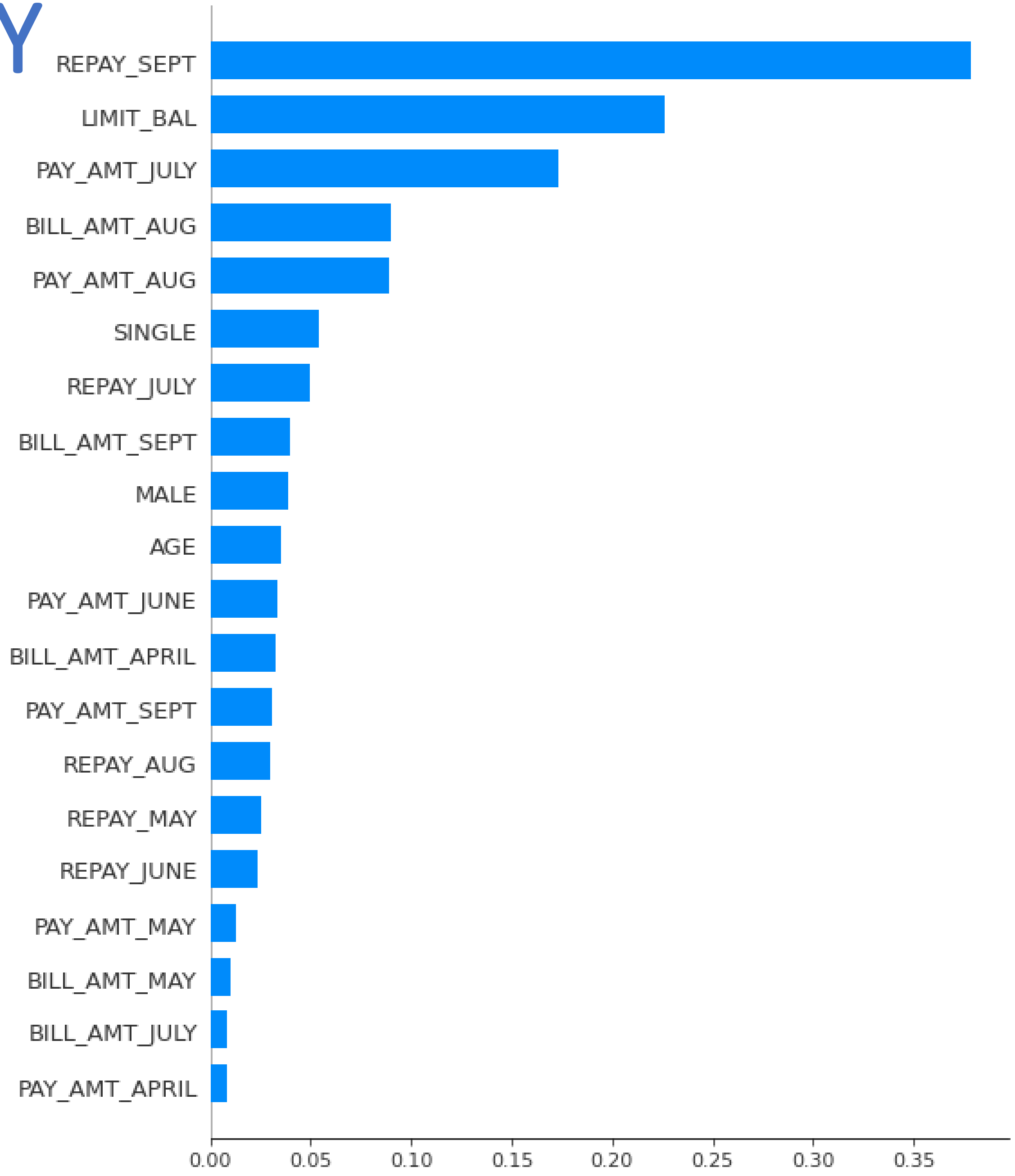
MODEL VALIDATION

- By observing Evaluation matrices for all the models-
  - ❑Logistic Regression model scores Low Performance as compared to other
  - ❑KNN looking at the scores this models are over fitting and we can't conclude with respect to accuracy.
  - ❑SVC Model is good with accuracy, but they are not best with its recall score compare to others.
  - ❑XGB Classifier and Random forest have High precision, high Recall value and high KS Statistic compared to others .

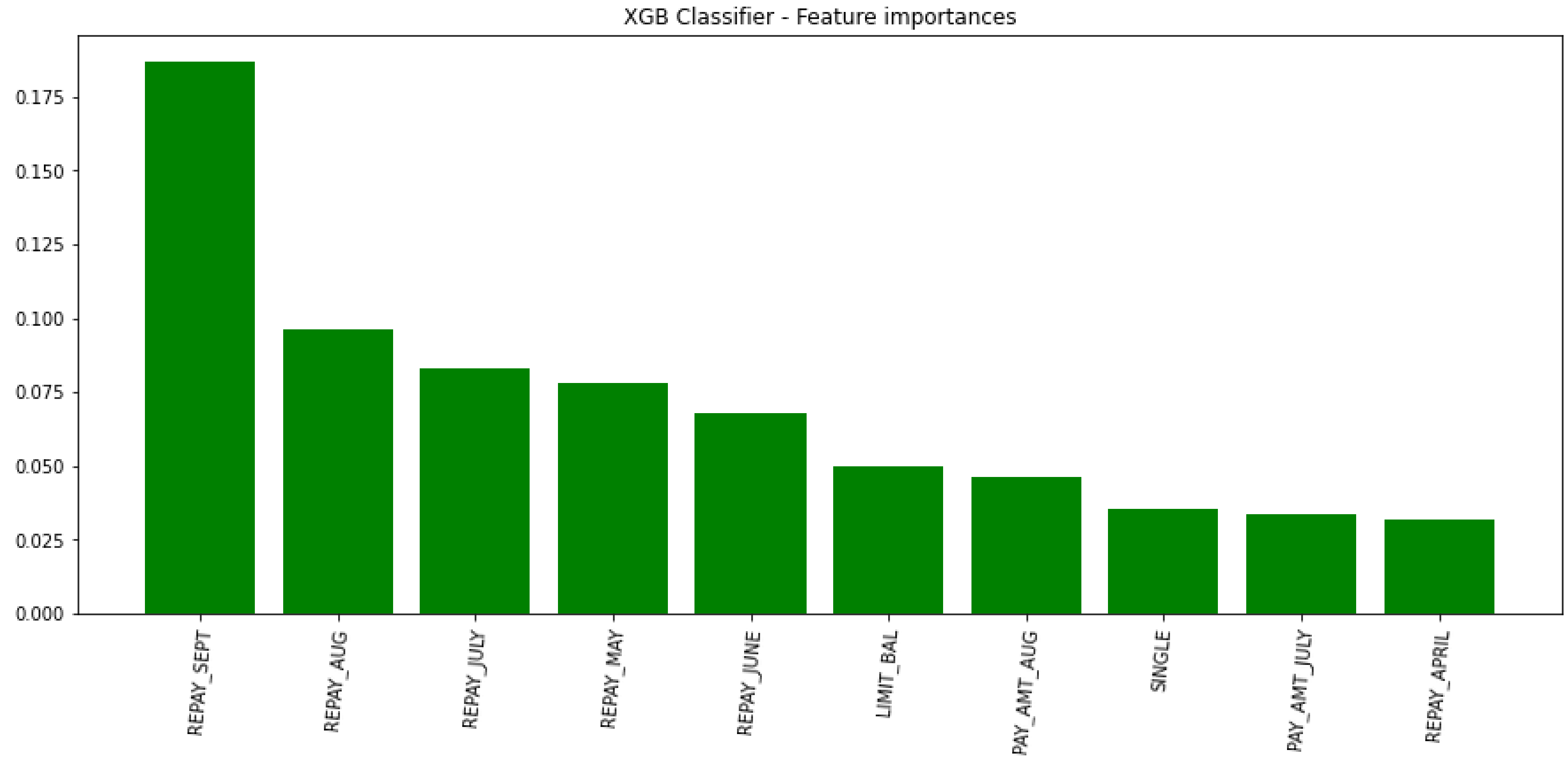
MODEL EXPLAINABILITY

1. Using SHAP

as per the SHAP technique REPAY_SEP an
d LIMIT _BAL are the most important feat
ure to predict target variable.



Feature Importance



CONCLUSION

- From entire Project analysis of ML Model, we got some evident that XGB Classifier will perform better among all the models for the Credit Card Default Prediction, since the recall score was best for this model.

- Repayment of September and Limit balance are the features contributes heavily to predict our target variable.

Thank You