Capstone Project - 3

Credit Card Default Prediction

By - Nirmesh Gupta

Let’s Catch The Defaulters

- Defining Problem Statement
- Data Summary
- Exploratory Data Analysis
- Feature Engineering
- Model Implementation
- Model Evaluation & Validation
- Model Selection

Problem Statement

- It would go a long way to research how machine learning can be applied to qualitative areas for better computations of credit risk exposure by predicting probabilities of default.
- The purpose of this project is to conduct qualitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

Data Pipeline

Data Cleaning

- Check null values and duplicate records.
- Find information on undocumented column values.

EDA

- Here, I have done some Exploratory Data Analysis and explored the data to get some insights about it.

Feature Engineering

- In this part, I have used One Hot Encoding to encode categorical features.

Build a Model

- And in this last part, I have used a number of Machine Learning algorithms to build a predictive model and used some evaluation matrices to evaluate performance of each model.

Data Summary

- This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns.

- **ID** : ID of each client.
- **LIMIT_BAL** : Amount of given credit in NT dollars.
- **SEX** : Gender(1=Male, 2=Female).
- **Education** : Education level(1=graduate school, 2=university, 3=high school, 4=others).
- **Marriage** : Marital status(1=married, 2=single, 3=others).
- **Age** : Age in years.
- **Pay_1 to pay_6** : Repayment status from September 2005 to April 2005.

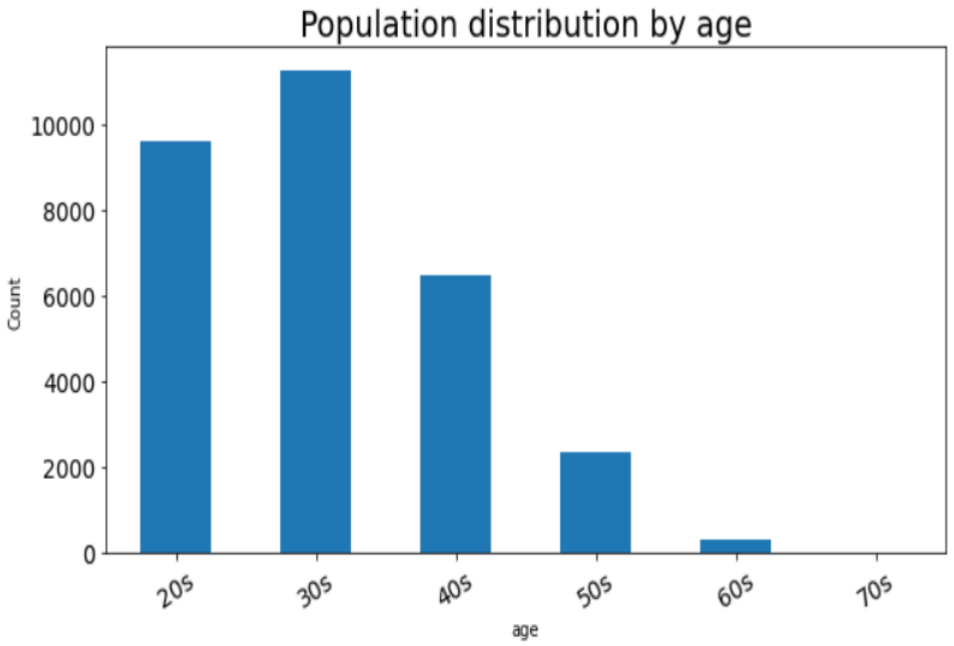
Data Summary

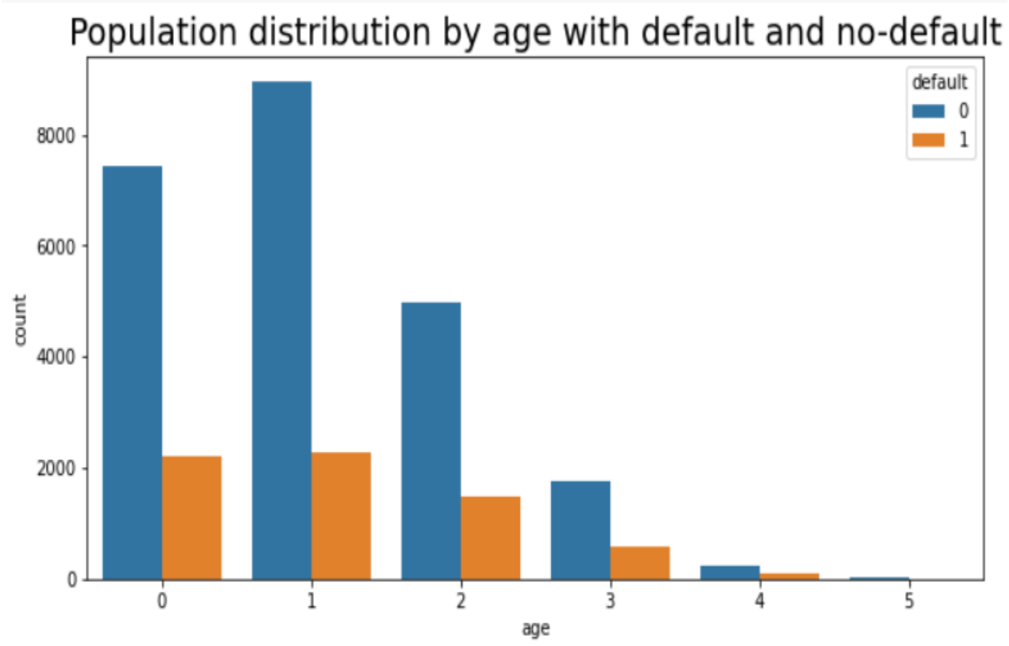
- **Bill_Amt1 to Bill_Amt6** : Amount of bill statement from September to April 2005(NT dollars).
- **Pay_Amt1 to Pay_Amt6** : Amount of previous payment in September to payment in April 2005(NT dollars).
- **Default** : This is our target feature. Default payment(1=Yes, 0=No)

Exploratory Data Analysis(EDA)

Analysis of Categorical Features

❖ Age Variable



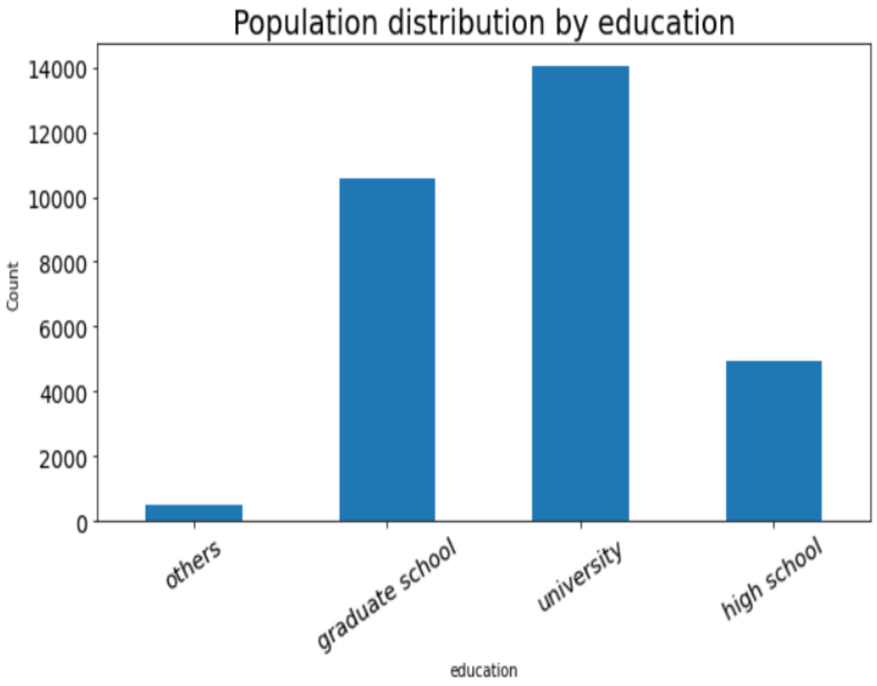


age	20s	30s	40s	50s	60s	70s	All
default							
Non-default proportion	0.771574	0.797473	0.770266	0.751388	0.716561	0.72	0.7788
Default proportion	0.228426	0.202527	0.229734	0.248612	0.283439	0.28	0.2212
All	1.000000	1.000000	1.000000	1.000000	1.000000	1.00	1.0000

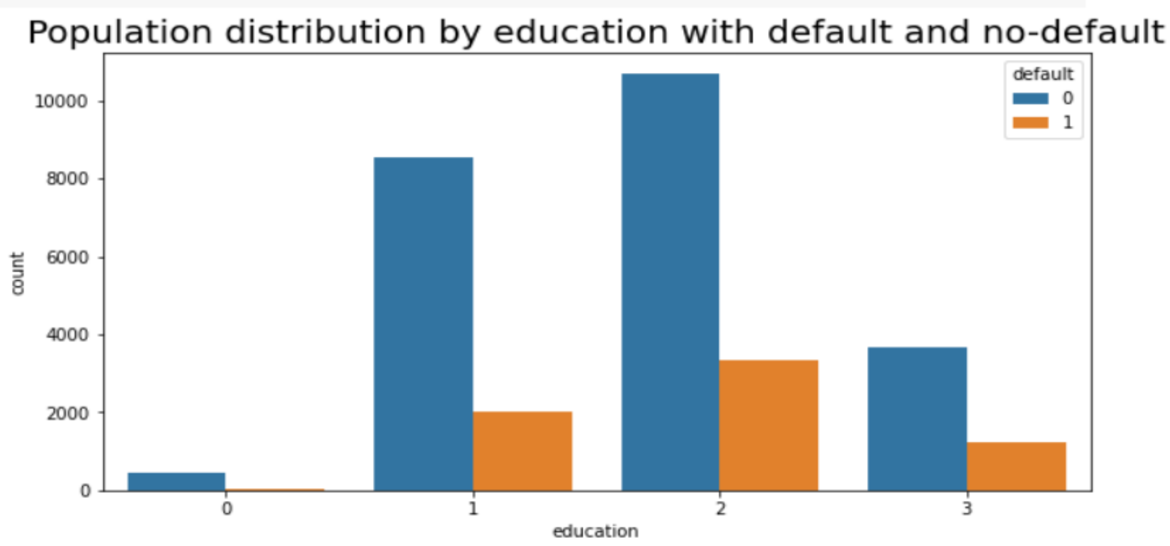
- Most of the customers are in their 30s.
- The default proportion is lowest for people in their 30s and then steadily rises with age.

Analysis of Categorical Features

❖ Education Variable



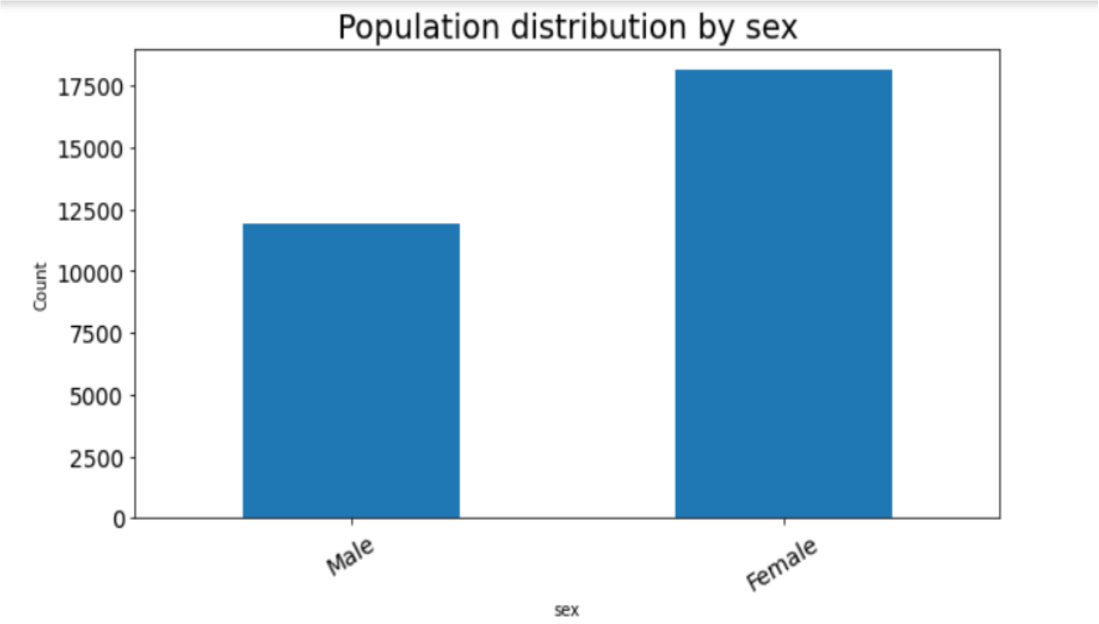
education	others	graduate school	university	high school	All
default					
Non-default proportion	0.929487	0.807652	0.762651	0.748424	0.7788
Default proportion	0.070513	0.192348	0.237349	0.251576	0.2212
All	1.000000	1.000000	1.000000	1.000000	1.0000



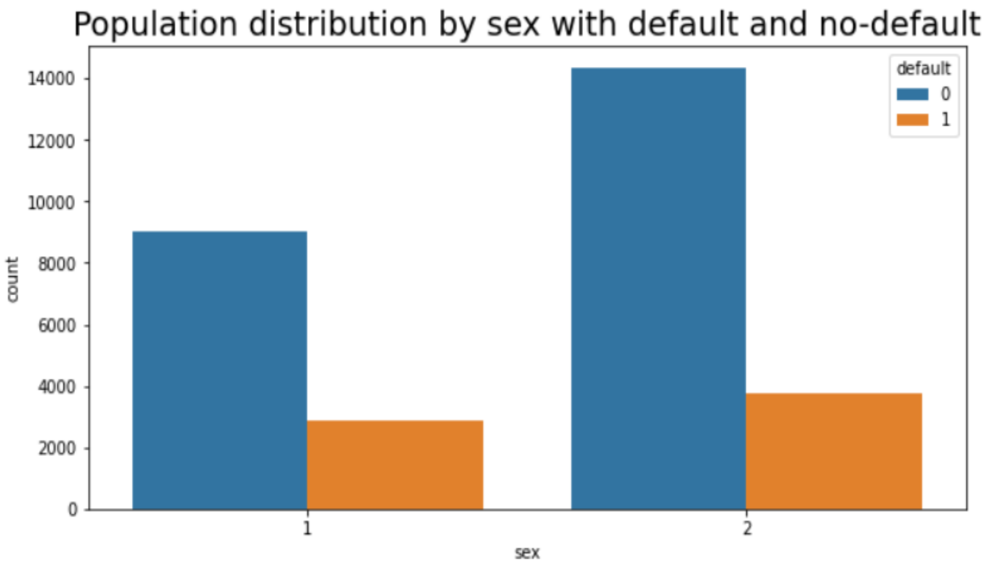
- The default proportion is least for customers with grad school level of education.
- SO, we can say, default proportion decreases with higher education level, mostly because, high educated people have higher paying jobs.

Analyzing the Categorical Features

❖ Gender Variable



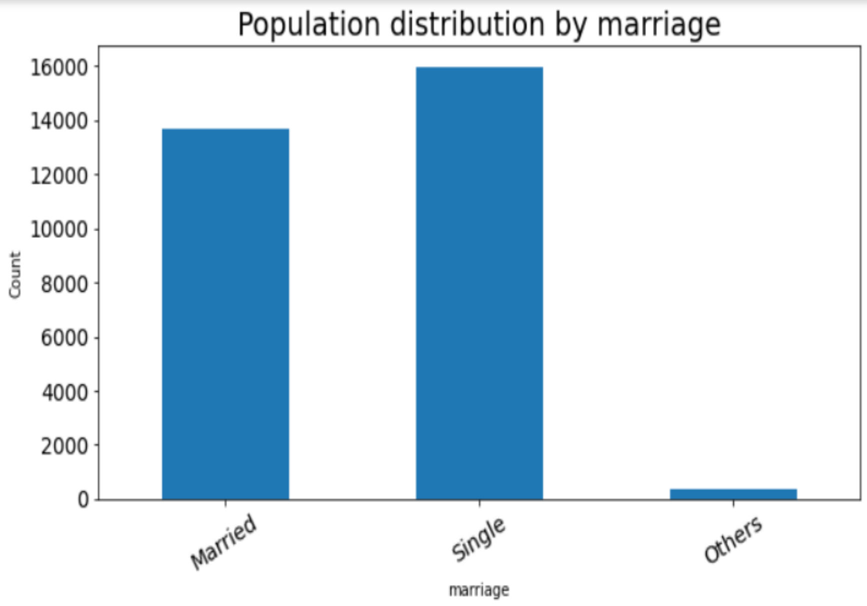
sex	Male	Female	All
default			
Non-default proportion	0.758328	0.792237	0.7788
Default proportion	0.241672	0.207763	0.2212
All	1.000000	1.000000	1.0000



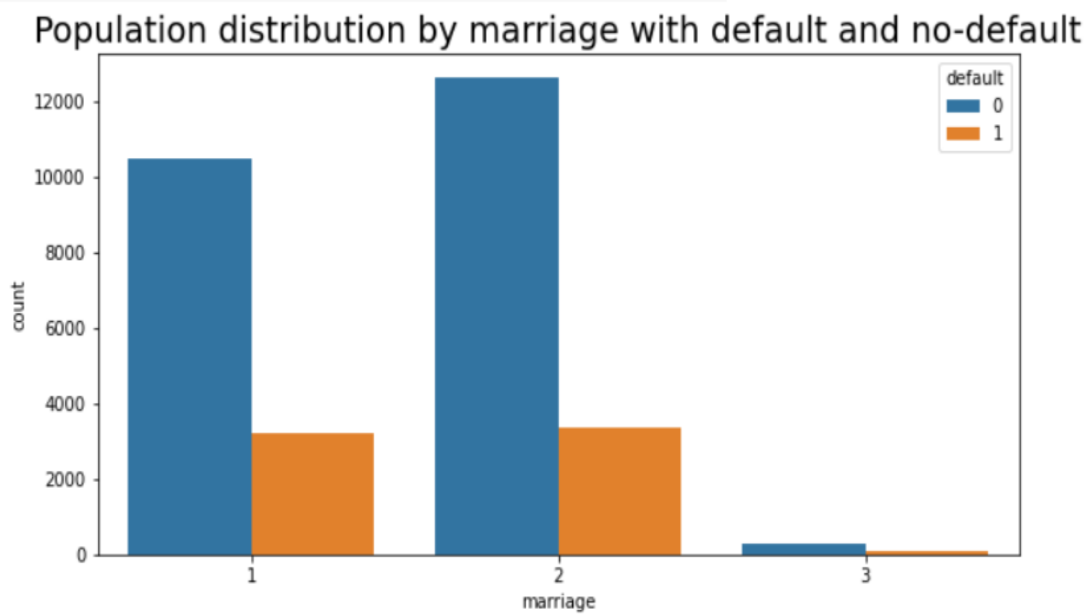
- Although, there are more female credit card holders, but the default proportion among men is higher.

Analysis of Categorical Features

❖ Marital Status Variable



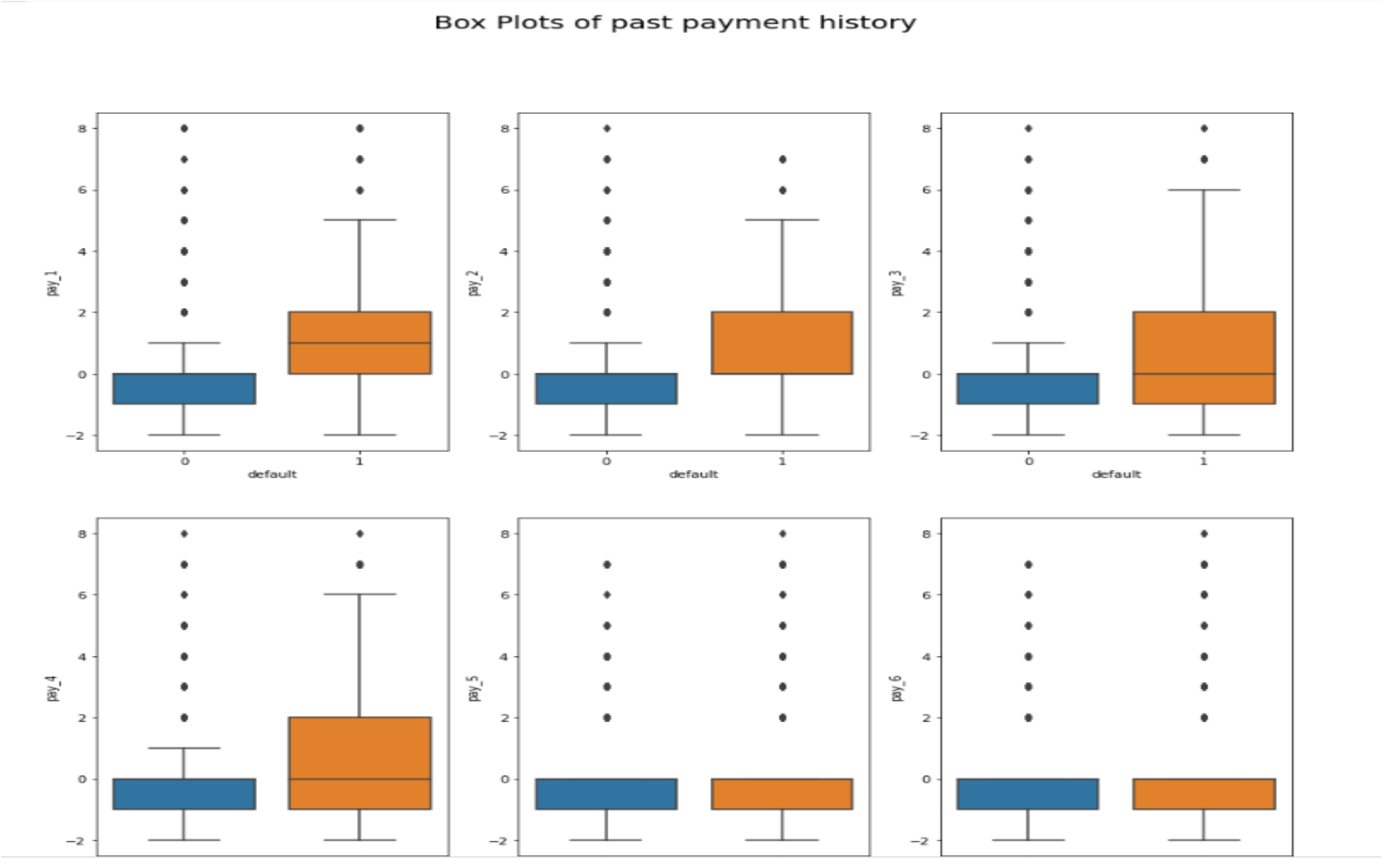
marriage	Married	Single	Others	All
default				
Non-default proportion	0.765283	0.790717	0.763926	0.7788
Default proportion	0.234717	0.209283	0.236074	0.2212
All	1.000000	1.000000	1.000000	1.0000



- Married people have higher proportion of default than single people.

Analysis of Categorical Features

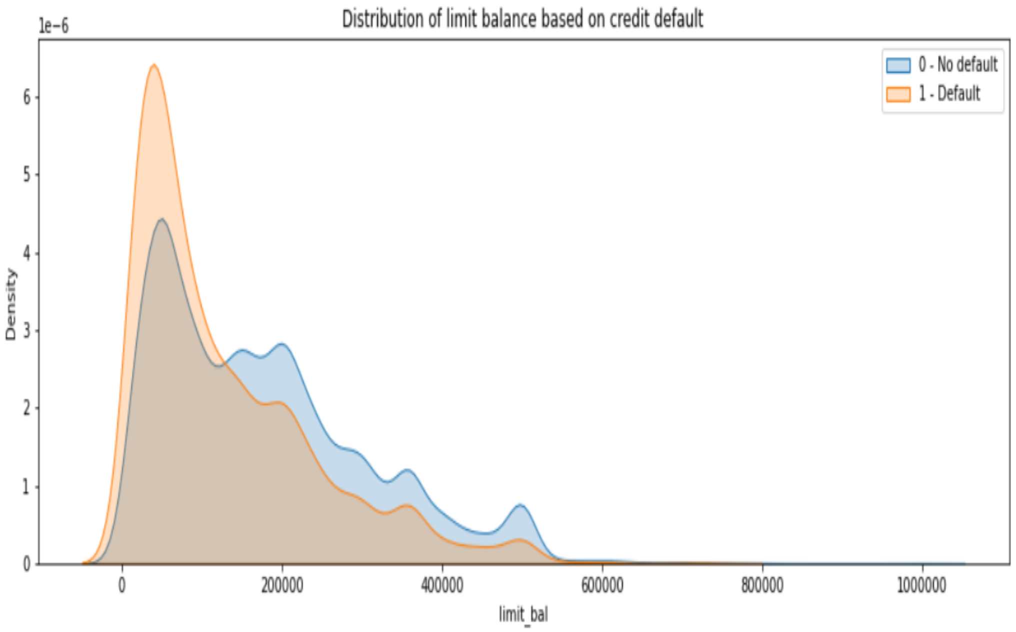
❖ Repayment Status



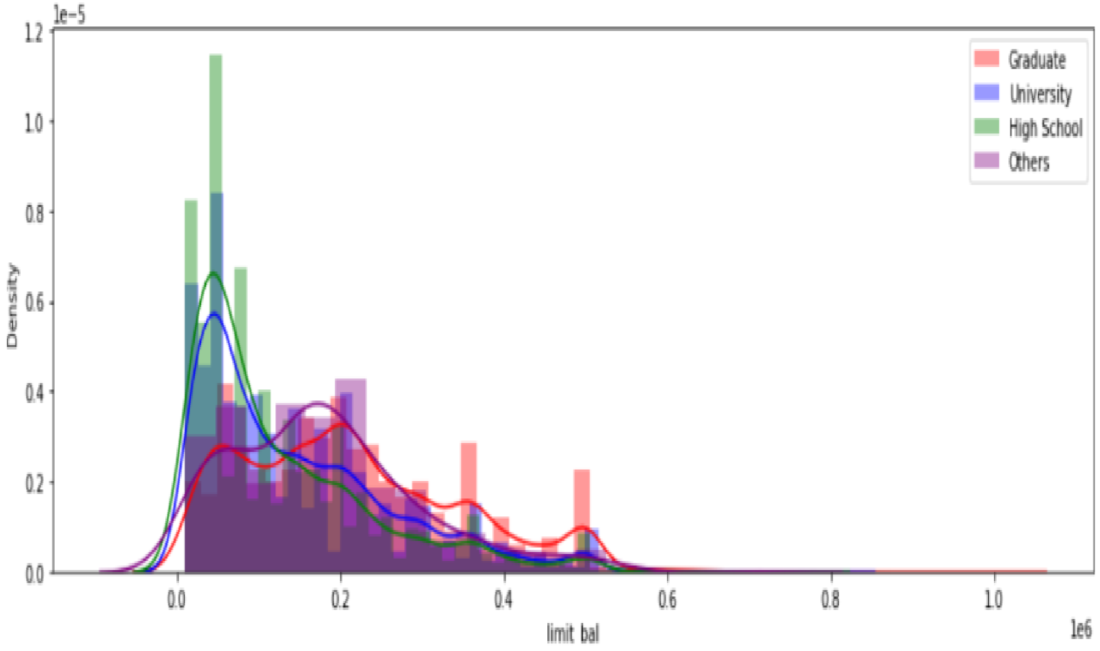
- There was a huge jump from May(pay_5) to June(pay_4) when delayed payment increased significantly, then it peaked in July(pay_3).Things started to get better in August(pay_2) and September(pay_1)

Analysis of Numerical Features

❖ Limit Balance and Default

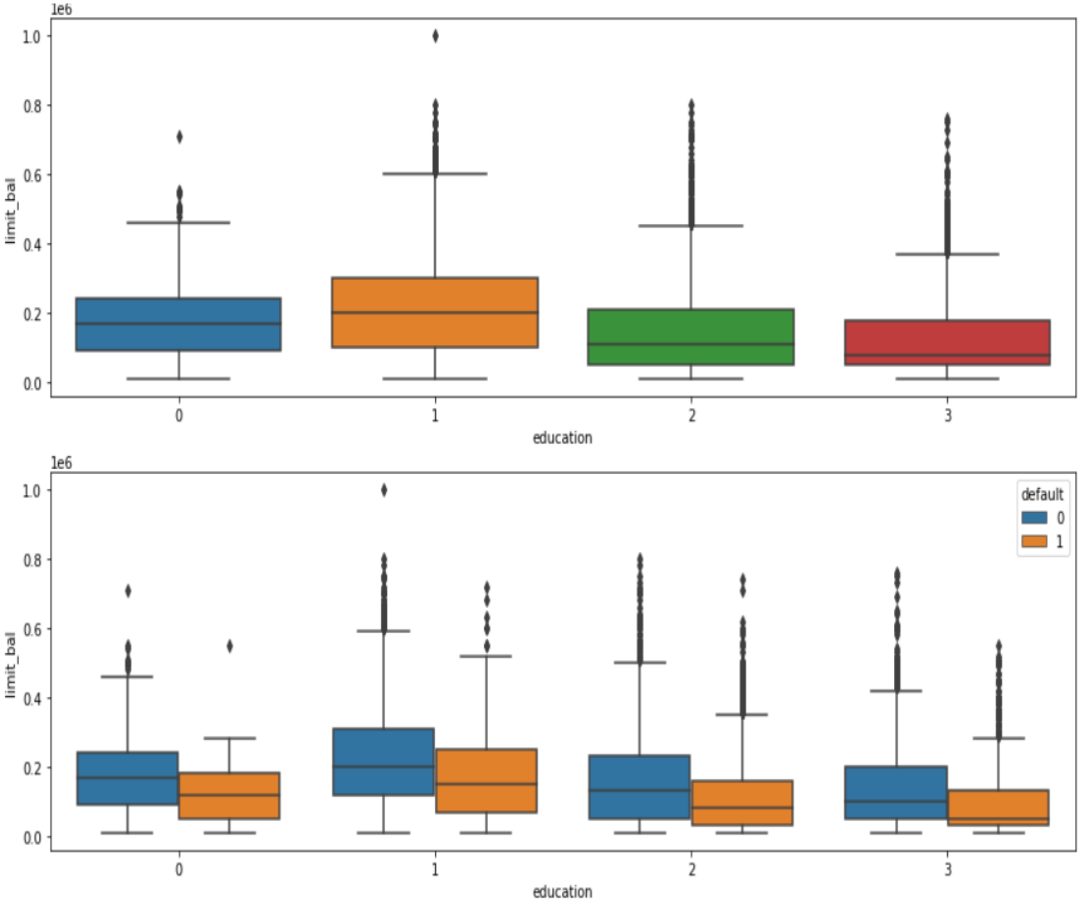


❖ Limit Balance, Default and Education



Analysis of Numerical Features

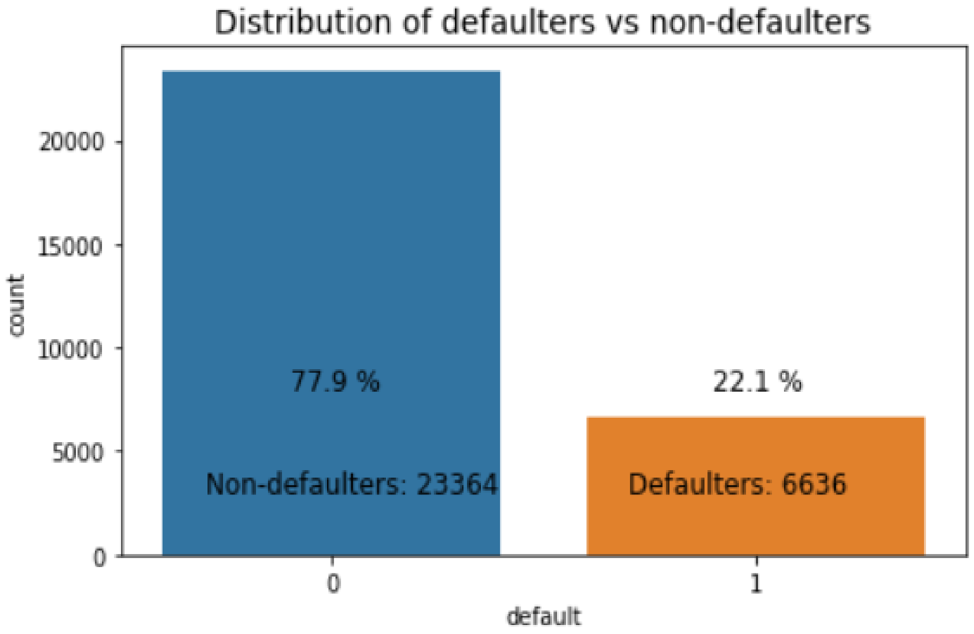
❖ Limit Balance, Default and Education



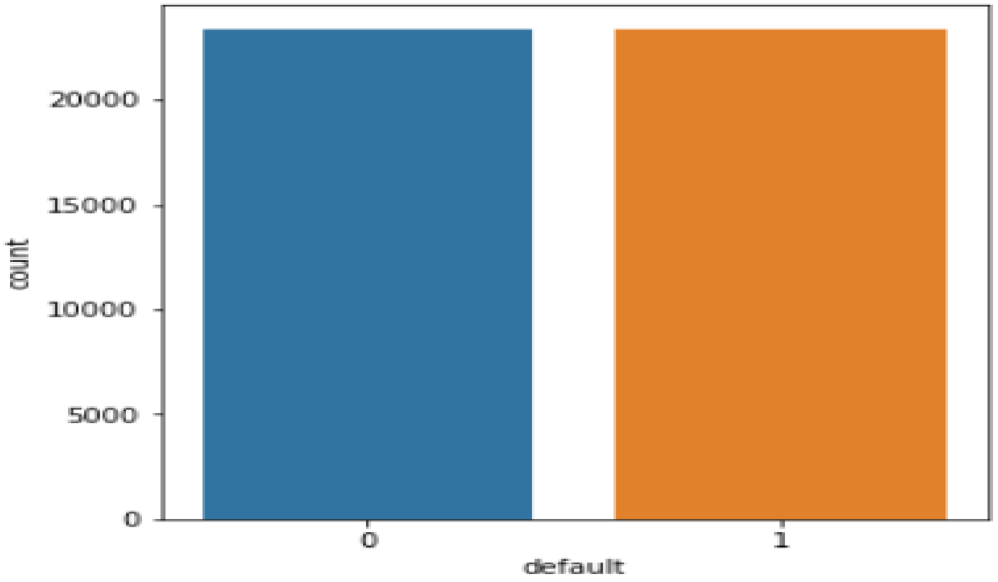
- The default proportion decreases with increase in credit limit.
- High school and University categories have a median limit balance mostly under the limit of 100K, while the graduates have median of 200K. So we can say, people with high education level get higher credit limits.

Analyzing the Target Feature

❖ Imbalanced Default Classes



❖ Balanced Default Classes



- There was class imbalance in our target feature with almost 78% of default observations and 22% of non default observations.
- I have used SMOTE oversampling technique for balancing the classes.

Feature Engineering & Scaling

- I have made a new feature “has_def” which tells if a customers has defaulted even once in the entire six month period.
- Then, I have used One Hot Encoding method to encode categorical features like Education, Sex and Marriage.
- The values of the numerical features needed to be scaled. I have used Standard Scaler to scale each feature to unit variance.

Model Implementation

- After the EDA and the data preprocessing, next task was to split the data into test and train data.
- The next step was Model Implementation. I have used following Machine Learning algorithms for modeling:

1. Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier
4. SVM Classifier

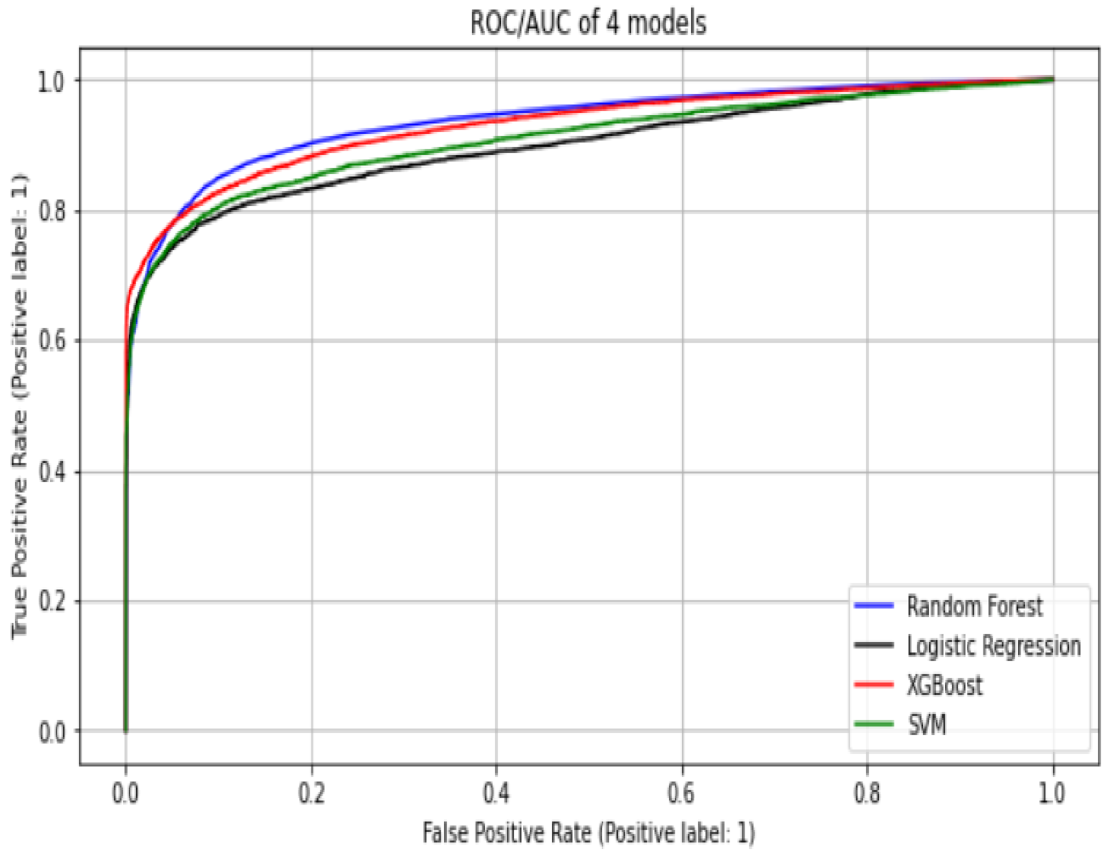
- I have evaluated each model with and without hyperparameter tuning on 4 evaluation metrics, namely,

1. ROC_AUC score
2. Precision
3. Recall
4. F1-score

- I will be considering ROC_AUC score and Recall mostly as we since we are more concerned about predicting maximum number of actual defaulters.

Model Performance

Models	ROC_AUC	Precision	Recall	F1 Score
Random Forest	0.932617	0.902716	0.839349	0.869880
Random Forest Tune	0.931066	0.899492	0.833785	0.865393
XGB Tuned	0.926598	0.920565	0.800257	0.856205
XGB	0.919091	0.916111	0.783707	0.844752
SVM	0.852181	0.928386	0.763875	0.838134
Logistic Regression Tuned	0.899292	0.932053	0.747610	0.829705
Logistic Regression	0.899348	0.936231	0.743615	0.828880

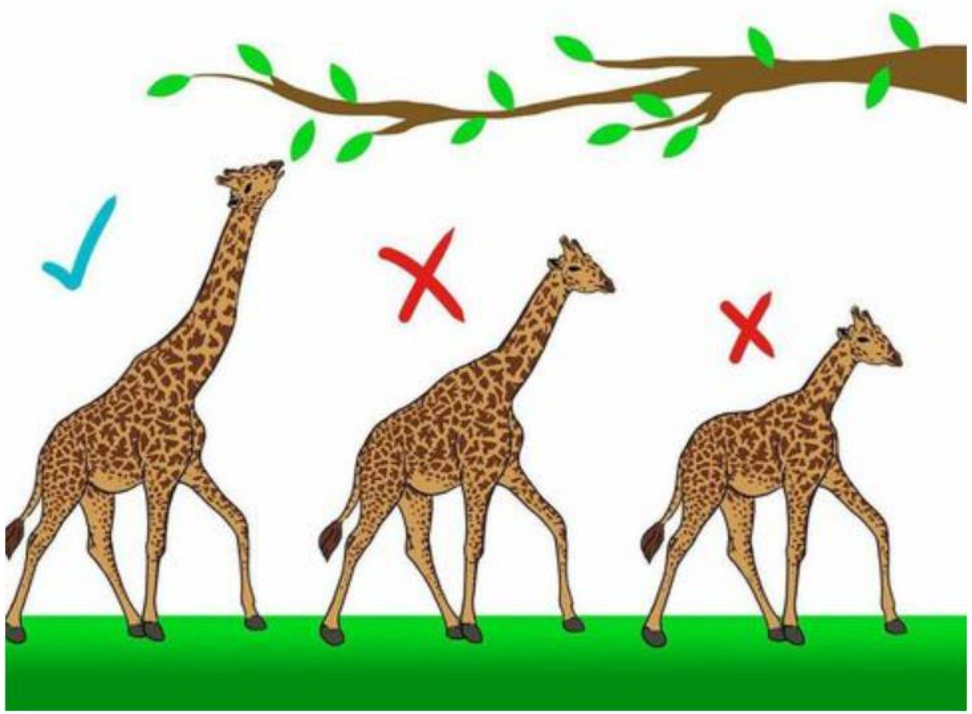


Model Validation & Selection

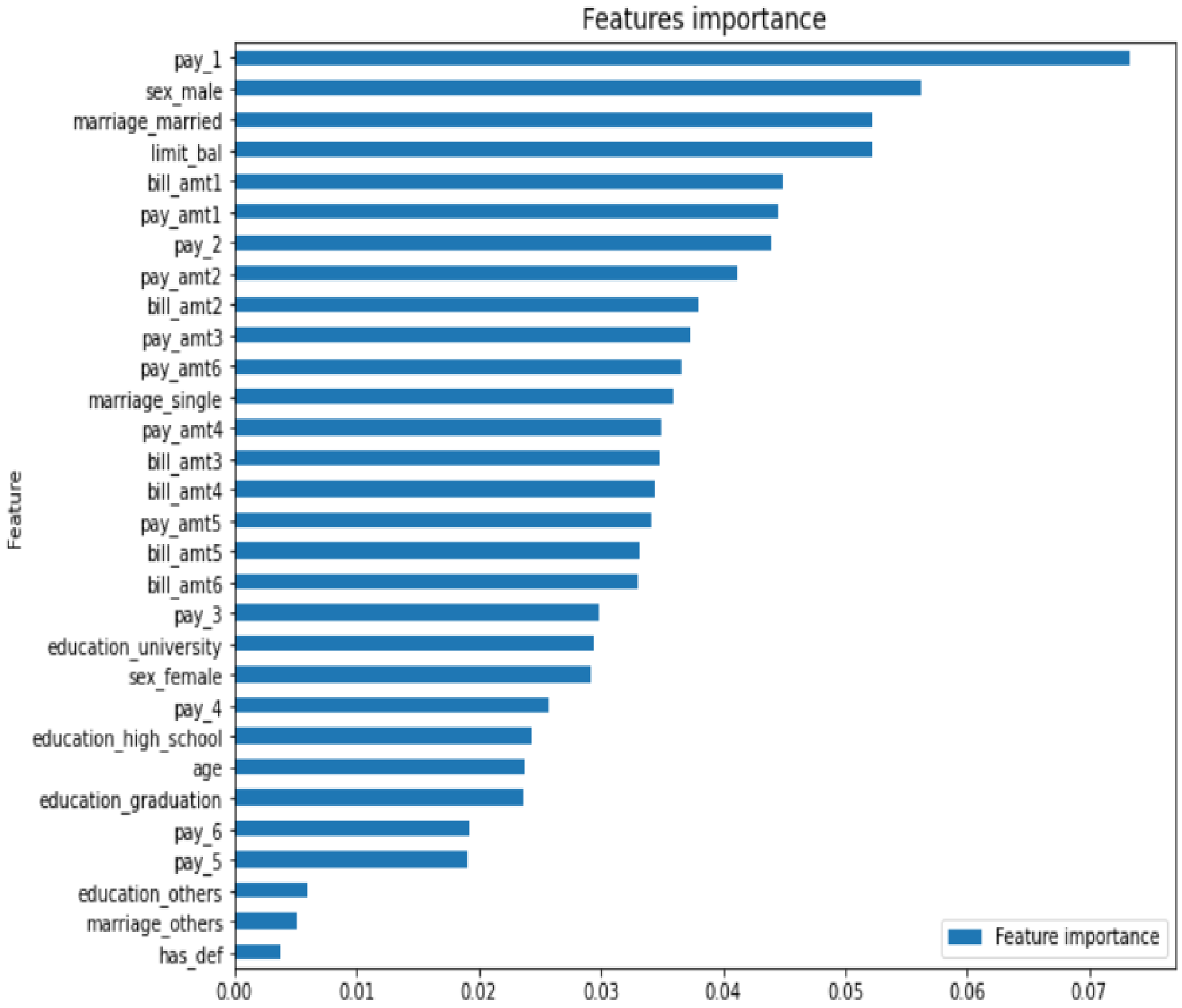
- As we can see from the performance table, all the models have given satisfactory results.
- The Logistic Regression and SVM have performed the worst among all the models with Recall of 0.74 and 0.76 respectively.
- ROC_AUC score is also least for these two models only.

Model Validation & Selection

- Random Forest Classifier with default parameters has performed the best with recall of 0.83 and also, in terms of all other evaluation metrics.
- Also, Random Forest Classifier and XGBoost Classifier have performed well at different threshold values with ROC_AUC score of 0.93 and 0.92 respectively.
- We can deploy Random Forest Classifier and XGBoost Classifier for further predictions.



Feature Importance



Conclusion

From EDA

- Although, there are more female credit card holders, the default proportion among men is higher, but the difference is not much significant.
- The default proportion decreases with higher education level. The graduate school customers had the least proportion of default. This may be because more educated people tend to have higher paying jobs which might make it easier for them to pay back their debts and also, educated people are more aware regarding the cons of defaulting on credit payments.
- Married people have higher default proportions than singles.
- Default proportion is lowest for people in their 30s and then steadily rises with age.
- Customers with high education levels get higher credit limits.
- Customers with higher credit limit have significantly lower default proportion. Intuitively, that is not surprising because the people who have higher credit limits must have displayed long periods of timely repayments to reach that place

Conclusion

From Modeling

- All of the models have pretty good AUC_ROC scores. All of the classifiers assign a higher probability of default to a defaulter over a non-defaulter with more than 85% certainty.
- Random Forest Classifier performs the best in terms of all the evaluation metrics.
- The best predictor of deliquency is the behaviour in the past months, especially the last month - pay_1.
- Random Forest Classifier and tuned XGB Classifier can be deployed to predict the defaulters

Thank You