Capstone Project-3

Credit Card Default Prediction

Table Of Content

1. Introduction
2. Problem statement
3. Data Summary
4. Exploratory Data Analysis
5. Feature Engineering
6. Modelling
7. Model Selection
8. Challenges
9. Conclusion

Introduction

A credit card has become an indispensable part of our lives, with its ease of use and convenient pay-back options.

The discounts, offers, and deals that a credit card offers are unmatched by any other financial products and spell a bonanza for the wise user.

However, credit cards can become debt traps if not used correctly, or if you spend more than you can repay when the bill comes around.

A credit card default happens when you fail to make any payment towards your credit card outstanding bill for a long period of time.

Problem Statement

A Taiwan-based credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood.

This would inform the issuer's decisions on who to give a credit card to and what credit limit to provide.

It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers.

Data Summary

- No of observations – 30000
- No of features - 25

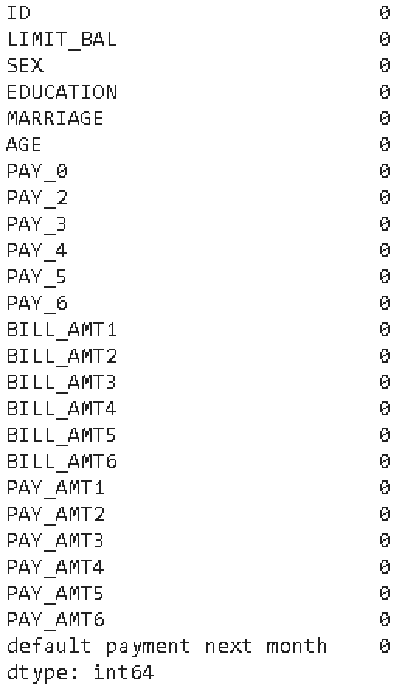
- **Independent Variables**
- X1: Amount of the given credit
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. (-1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.)
- X12-X17: Amount of bill statement
- X18-X23: Amount of previous payment
- **Dependent Variable**
- default.payment.next.month: default payment (1=yes,0=no)

Exploratory Data Analysis

Check For Missing values

- No missing values Found

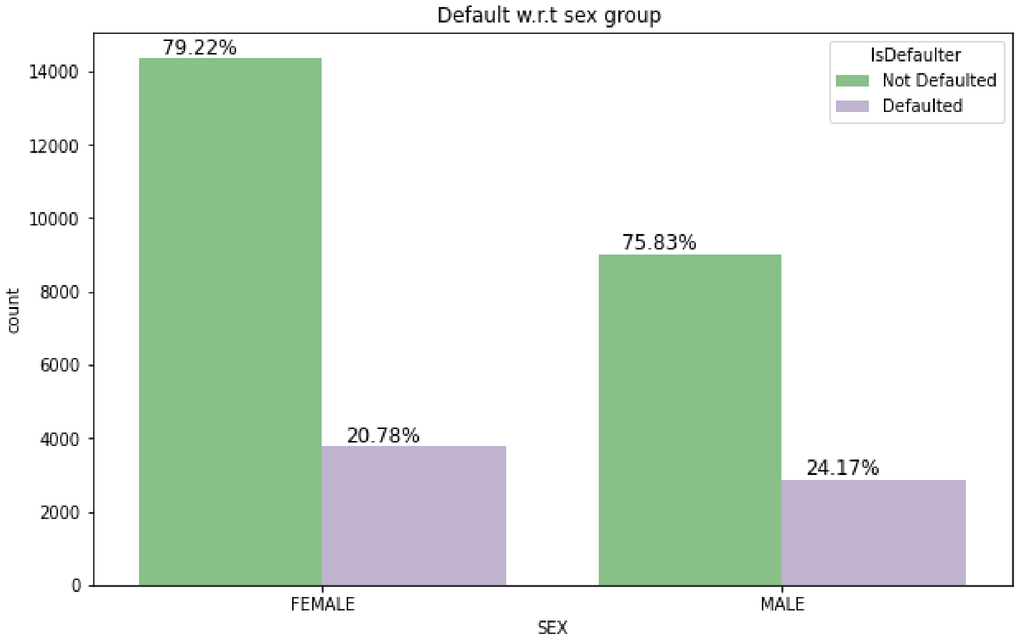
- So we are Good to go!!



```
ID                            0
LIMIT_BAL                     0
SEX                           0
EDUCATION                     0
MARRIAGE                      0
AGE                           0
PAY_0                         0
PAY_2                         0
PAY_3                         0
PAY_4                         0
PAY_5                         0
PAY_6                         0
BILL_AMT1                     0
BILL_AMT2                     0
BILL_AMT3                     0
BILL_AMT4                     0
BILL_AMT5                     0
BILL_AMT6                     0
PAY_AMT1                      0
PAY_AMT2                      0
PAY_AMT3                      0
PAY_AMT4                      0
PAY_AMT5                      0
PAY_AMT6                      0
default payment next month    0
dtype: int64
```

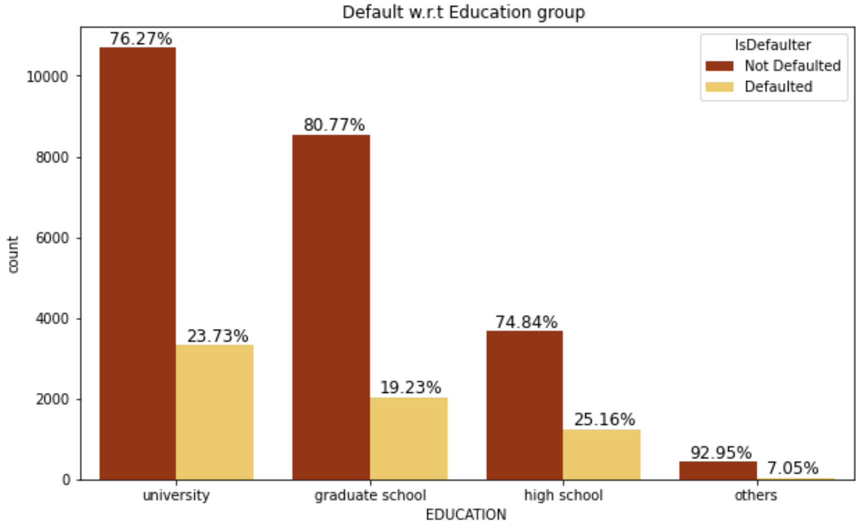
Which sex group tends to have more delayed payments?

- Females tend to have slightly more delayed payments

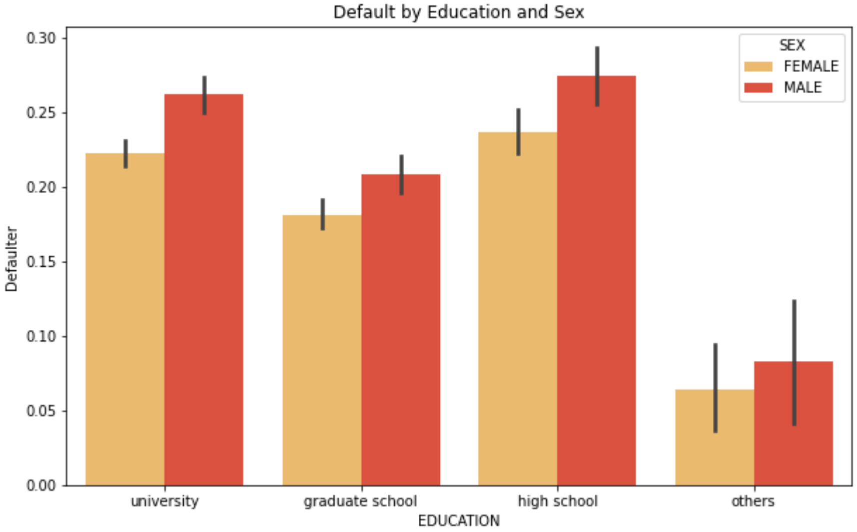


Education feature Distribution

- Did customers with higher education have less delayed payment?

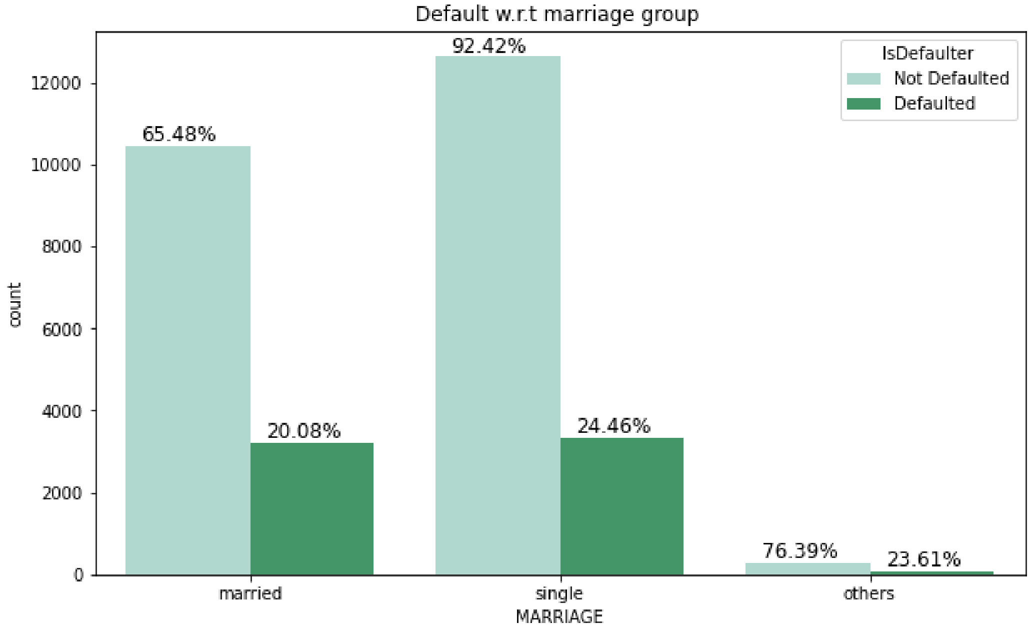


- Which SEX group has made more defaults in their respective EDUCATION category?



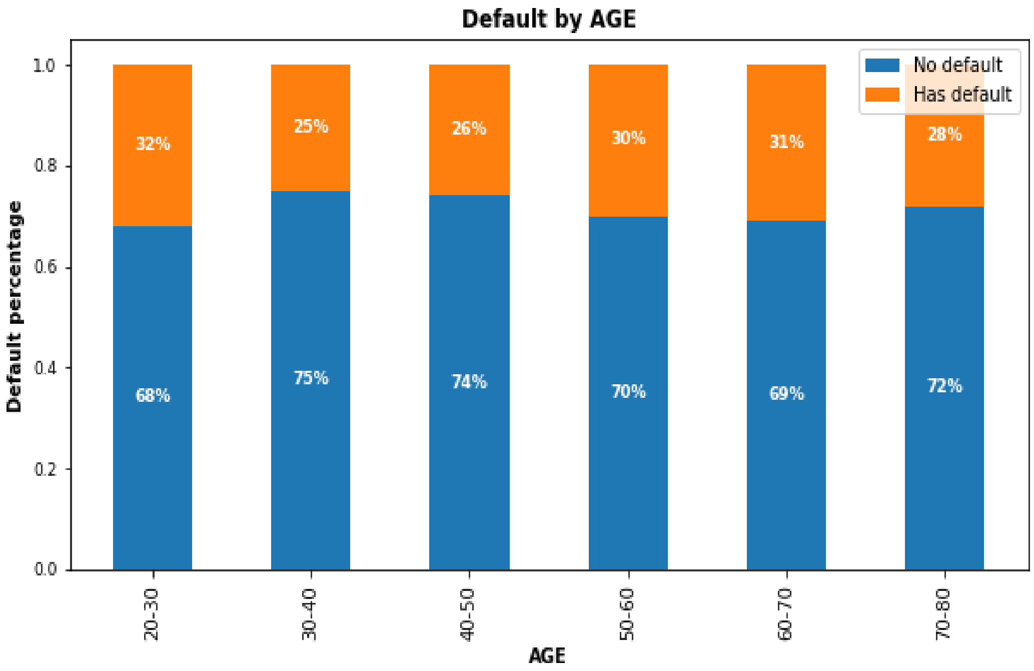
Does marital status have anything to do with default risk?

- Both married and single have almost
- same default rate



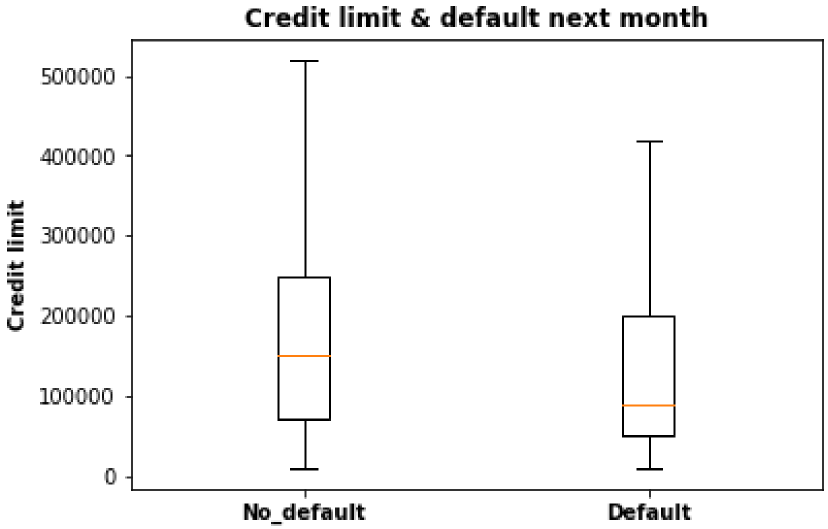
Do younger payment tend to miss the payment deadline?

Customers aged between 30-50 had the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates. However, the delayed rate dropped slightly again in customers older than 70 years.



Is there any correlation between credit limit and the default payment next month?

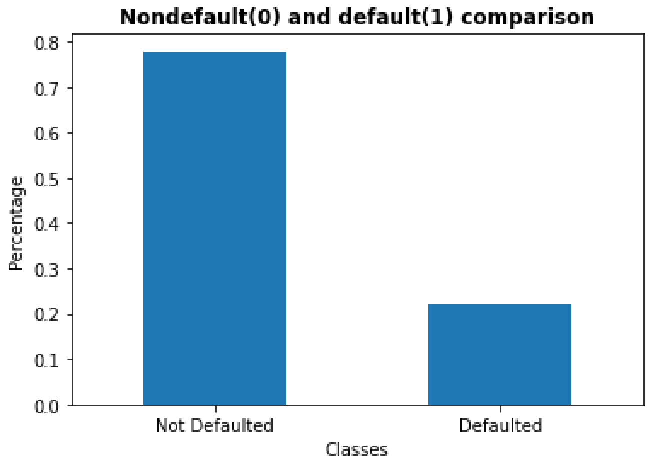
- Customers with high credit limits tend to have higher 'no-default’ rate.

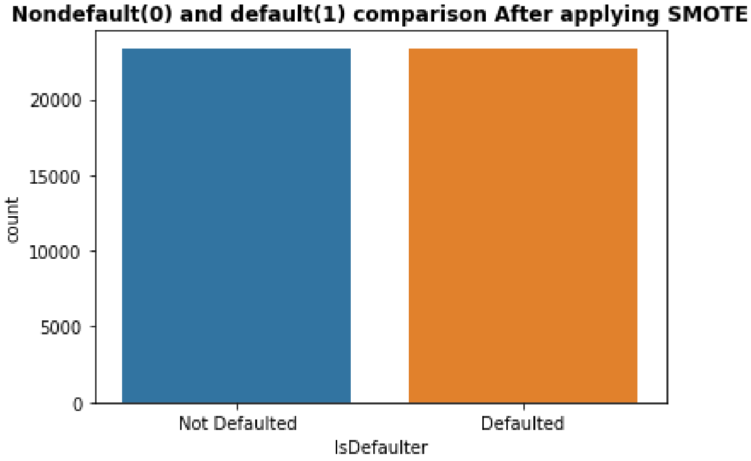


Imbalanced data

- Data is Imbalanced
- 78% - Non Default Class
- 22% - Default Class

- Used SMOTE (Synthetic minority
- Oversampling) technique to deal
- with imbalanced data





Feature engineering

- Label Encoding
- SEX
- Female = 0 , Male =1

- One Hot Encoding
- Education
- Marriage

- Feature Selection
- Rescaling data using StandardScaler
- Train Test Split

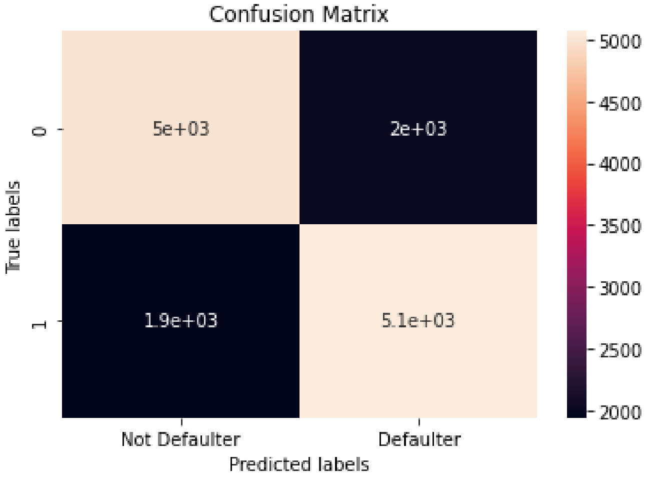
Modelling Overview

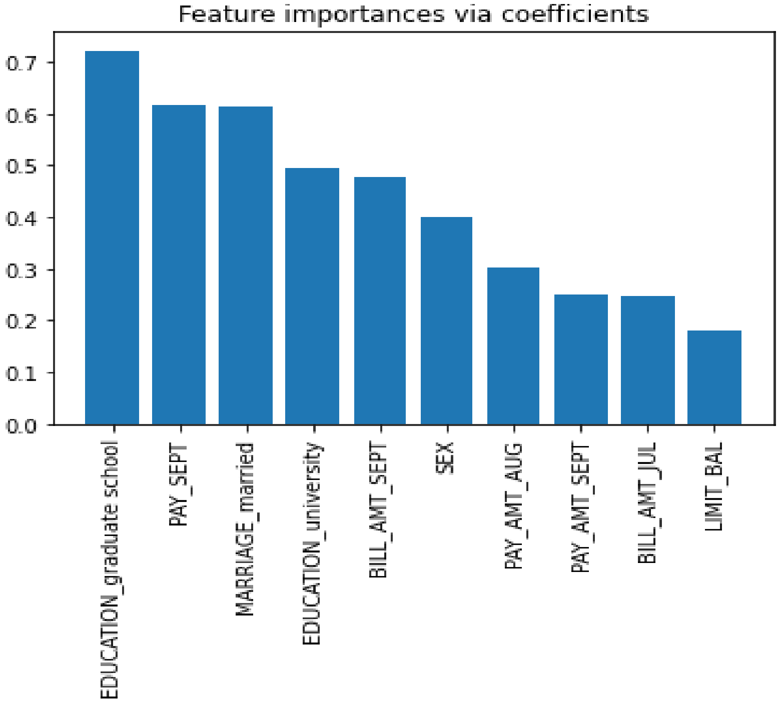
- Models used:

- 1. Logistic Regression
- 2. Random Forest Classifier
- 3. XGBoost Classifier
- 4. SVC

Logistic Regression

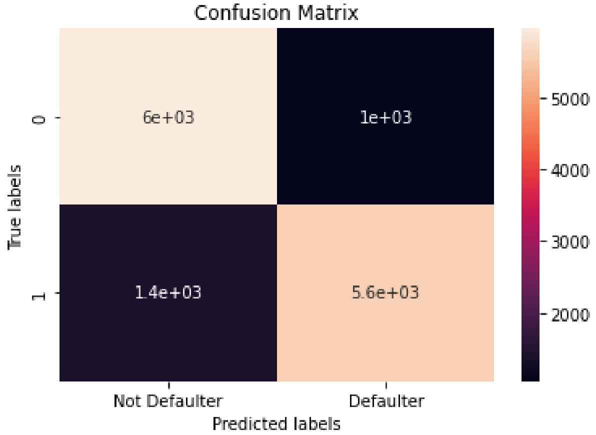
- Parameters:
- C: 0.1
- Penalty: L2

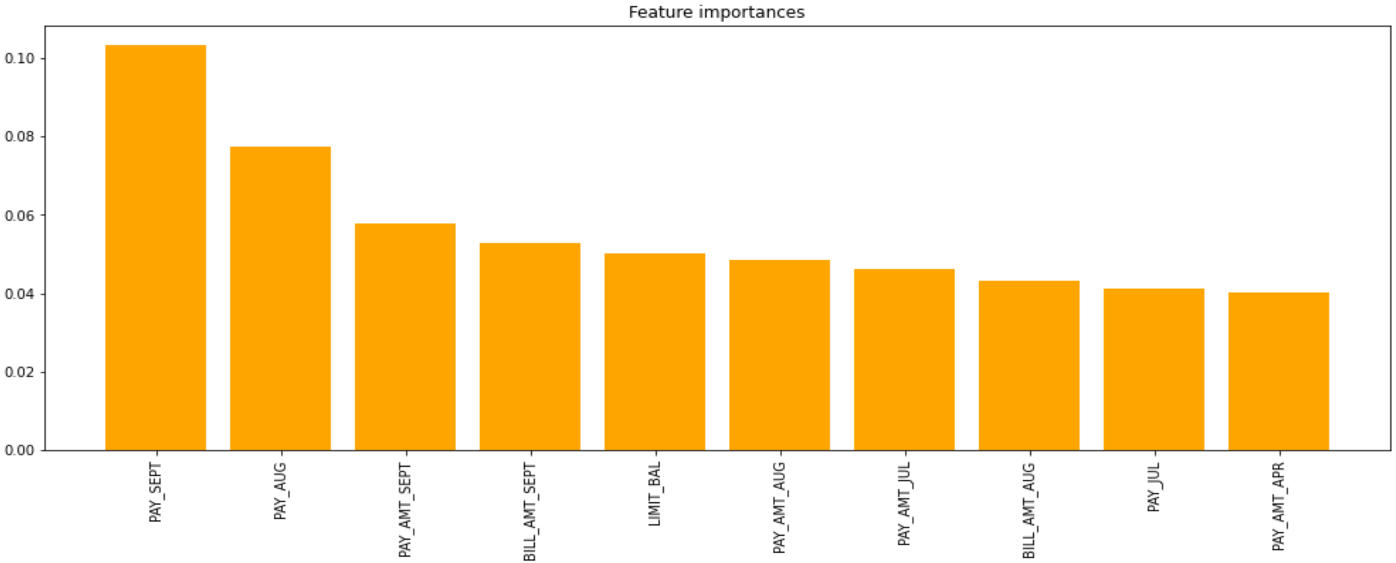




Random Forest Classifier

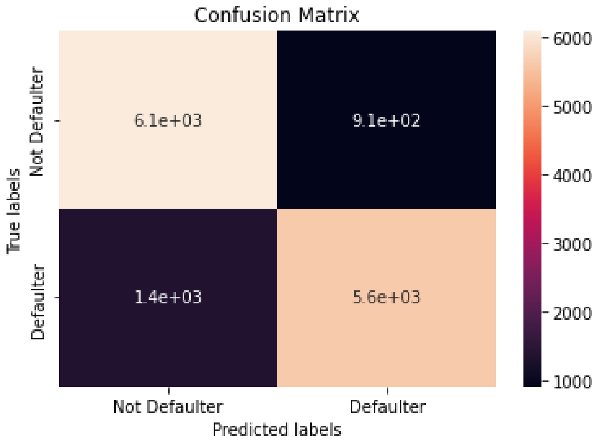
- Parameters:
- 'max_depth': 30, 'min_samples_leaf': 3,
- 'min_samples_split': 10, 'n_estimators': 150

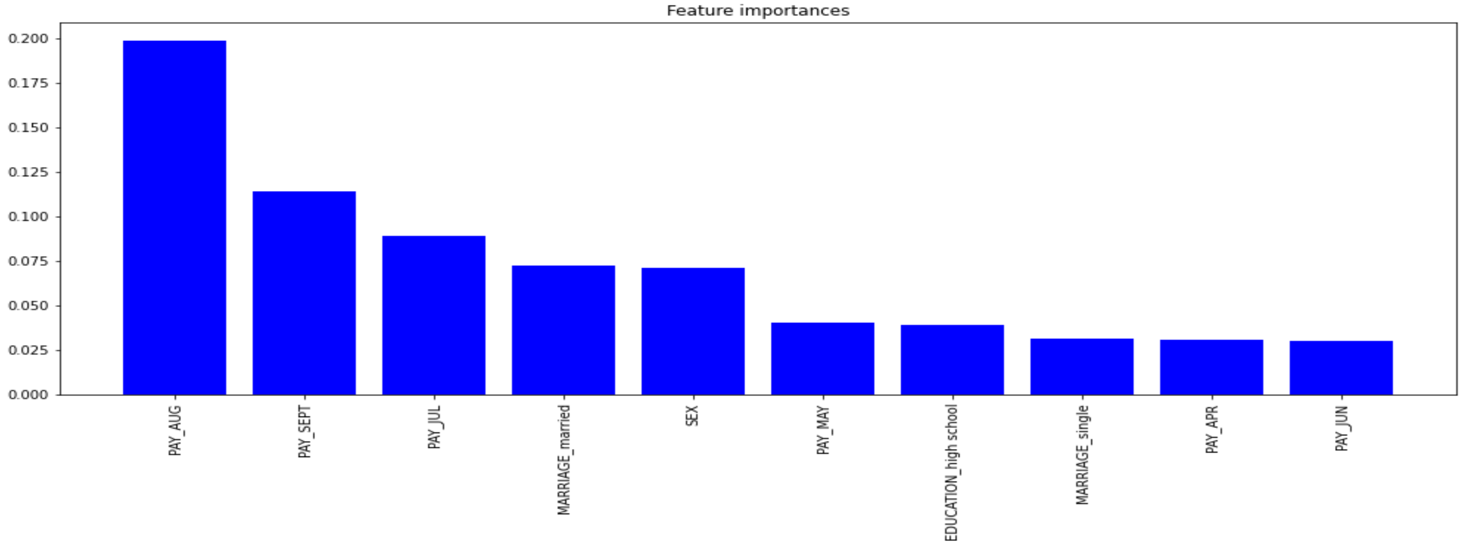

Confusion Matrix


Feature importances

XGBoost Classifier

- Parameters:
- 'max_depth': 9, 'min_child_weight': 1

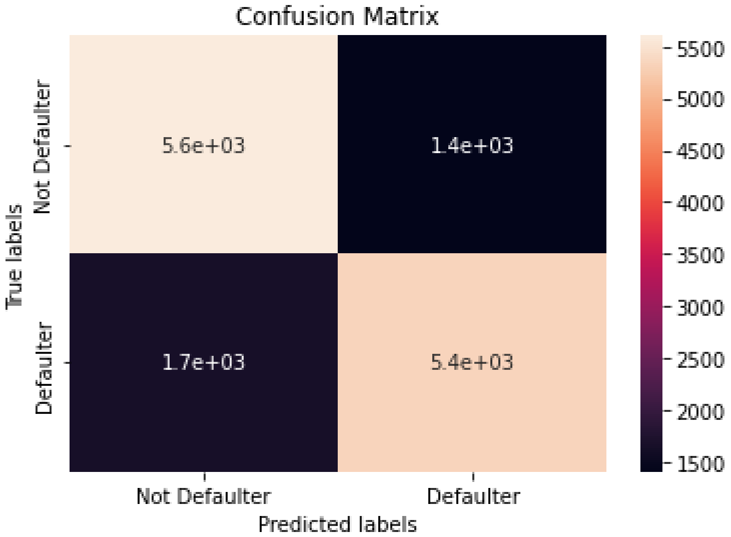




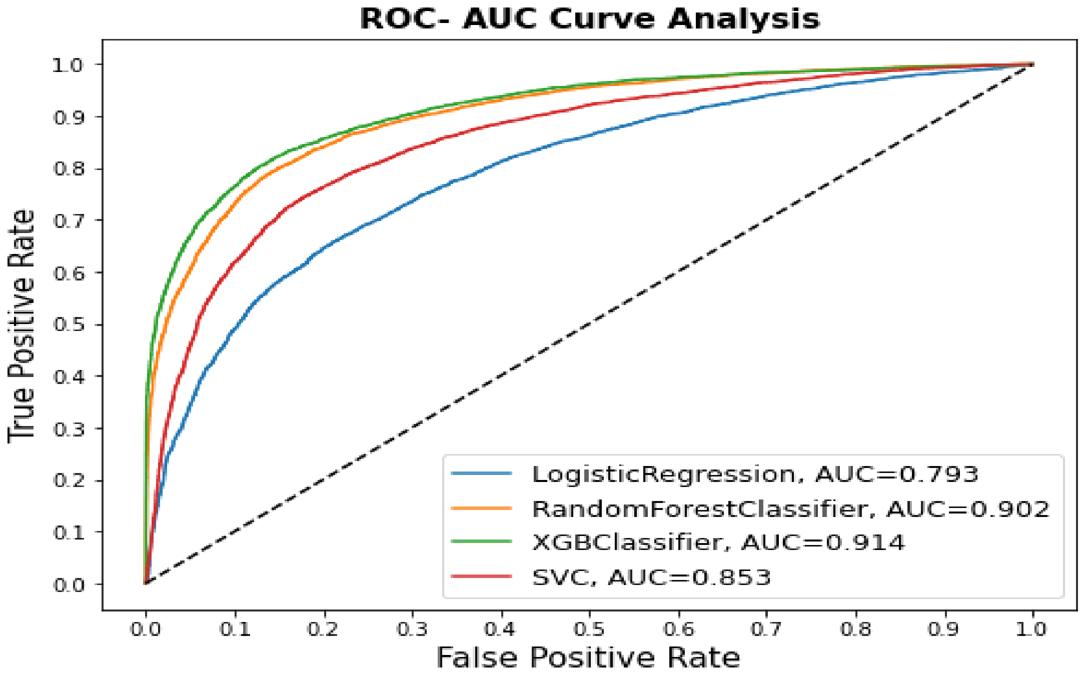
SVC – Support Vector Machine

- Parameters:
- 'C': 100,

'kernel': 'rbf'

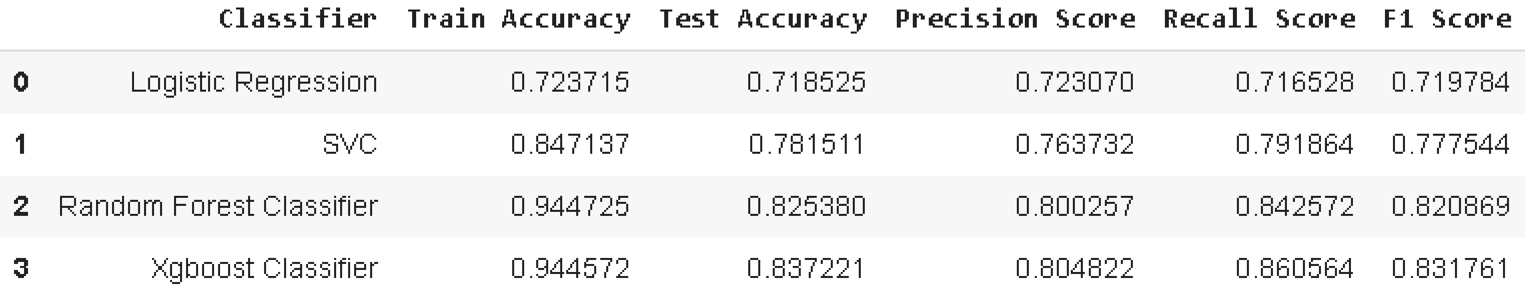


ROC-AUC Curve Analysis



Model Selection

- The business nature of credit card default prediction requires model to have a high recall.
- XGBoost has the highest Recall Rate out of all the models.



	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.723715	0.718525	0.723070	0.716528	0.719784
1	SVC	0.847137	0.781511	0.763732	0.791864	0.777544
2	Random Forest Classifier	0.944725	0.825380	0.800257	0.842572	0.820869
3	Xgboost Classifier	0.944572	0.837221	0.804822	0.860564	0.831761

Challenges

1. Feature Engineering
2. Getting high Accuracy of the models.
3. Overfitting of the model

Conclusion

- We have built predictive model for credit card agency to predict if a person would default on his/her payment of credit card.
- We have performed feature engineering, feature selection, hyperparameter tuning to prevent overfitting and decrease error rate in the model.
- Since the business nature of credit card default prediction requires model to have a high recall. Therefore we selected XGBoost as our best model.

Thank You