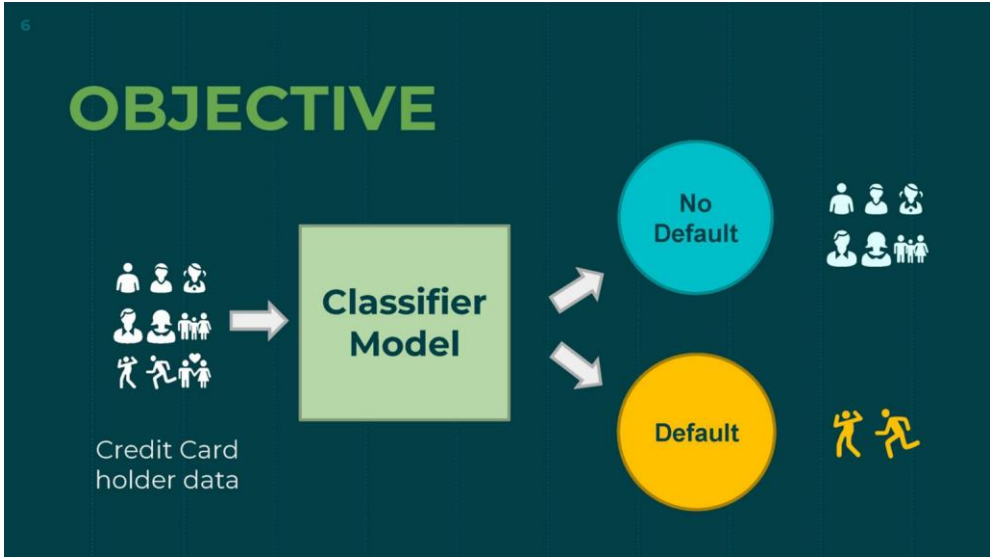
Capstone Project – 3

credit card default prediction

Shubham Bhadouria

Content

- Introduction
- Defining Problem Statement
- Data Summary
- EDA /dataset preparation
- Modelling Overview
- Model evaluation
- Challenges
- Conclusion
- Q&A



Introduction

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.

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.

Defining Problem Statement

- Identify the key features that determine the likelihood of credit card default.
- Predicting whether the customer is defaulter or non-defaulter.

Data Summary

- X1 - Amount of credit(includes individual as well as family credit)
- X2 - Gender
- X3 - Education
- X4 - Marital Status
- X5 – Age
- X6 to X11 - History of past payments from April to September
- X12 to X17 - Amount of bill statement from April to September
- X18 to X23 - Amount of previous payment from April to September
- Y - Default payment

Approach Overview

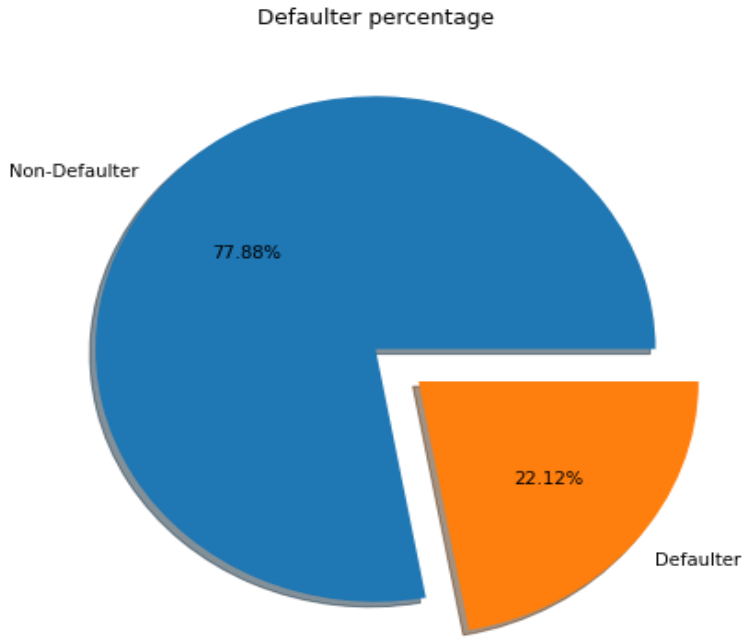
- Data inspection and cleaning
  Exploring data, checking for outliers
- Data Pre-processing & EDA
  Checking distributions of variables
  Univariate and multivariate analysis
  Checking for imbalanced dataset
- Modelling (Implementing Machine Learning Algorithms)
  KNN
  Random Forest
  XGBoost
  SVM
  LGBM
- SHAP features

Basic Data Exploration

- Dataset contains the data of credit card holders of Taiwan from April 2005 to September 2005.
- Dataset contains 30000 rows & 25 columns.
- In dataset, 6 months payment and bill details are available.
- There are no null and duplicate values in the dataset

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5
0	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0
1	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455
2	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948
3	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959
4	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146

Feature Analysis - The frequency of defaults

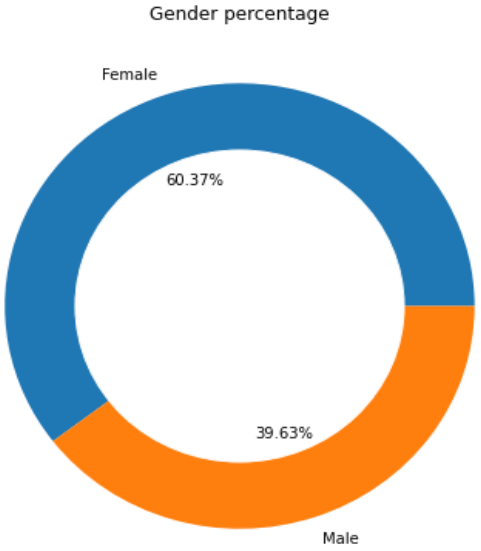


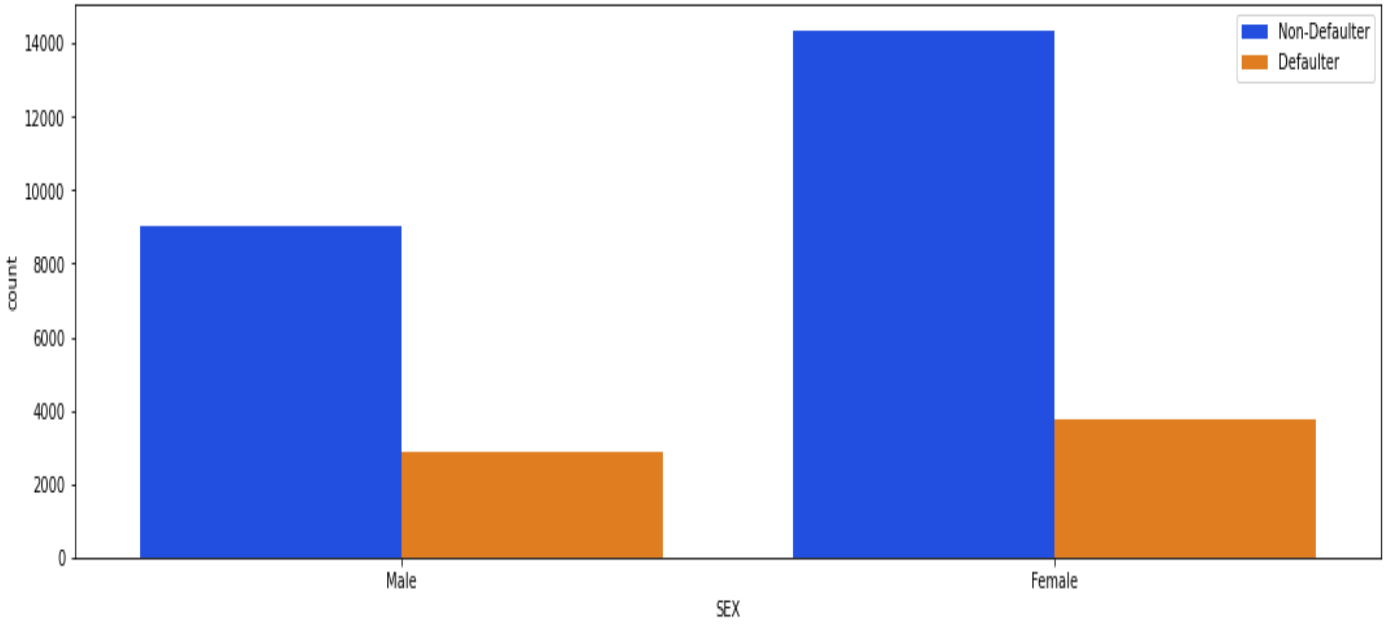
It is clear from the pie chart that it is a case of Imbalanced dataset.

Number of Non-default is much higher than total non-default.

Total non- Default data are 77.88% while total default are 22.12%.

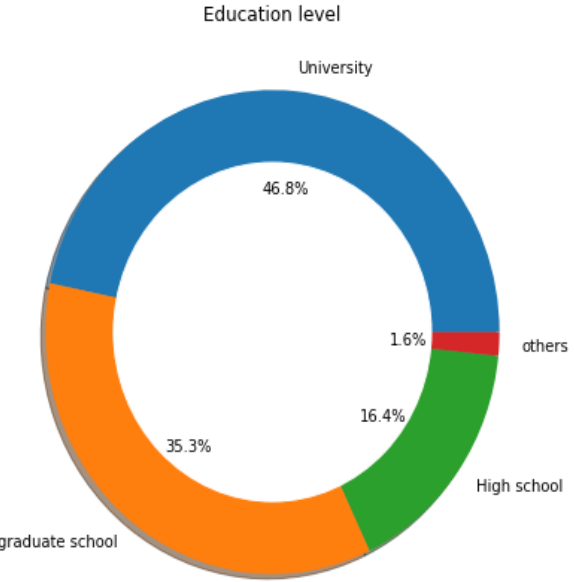
Feature Analysis - Gender wise defaulter prediction

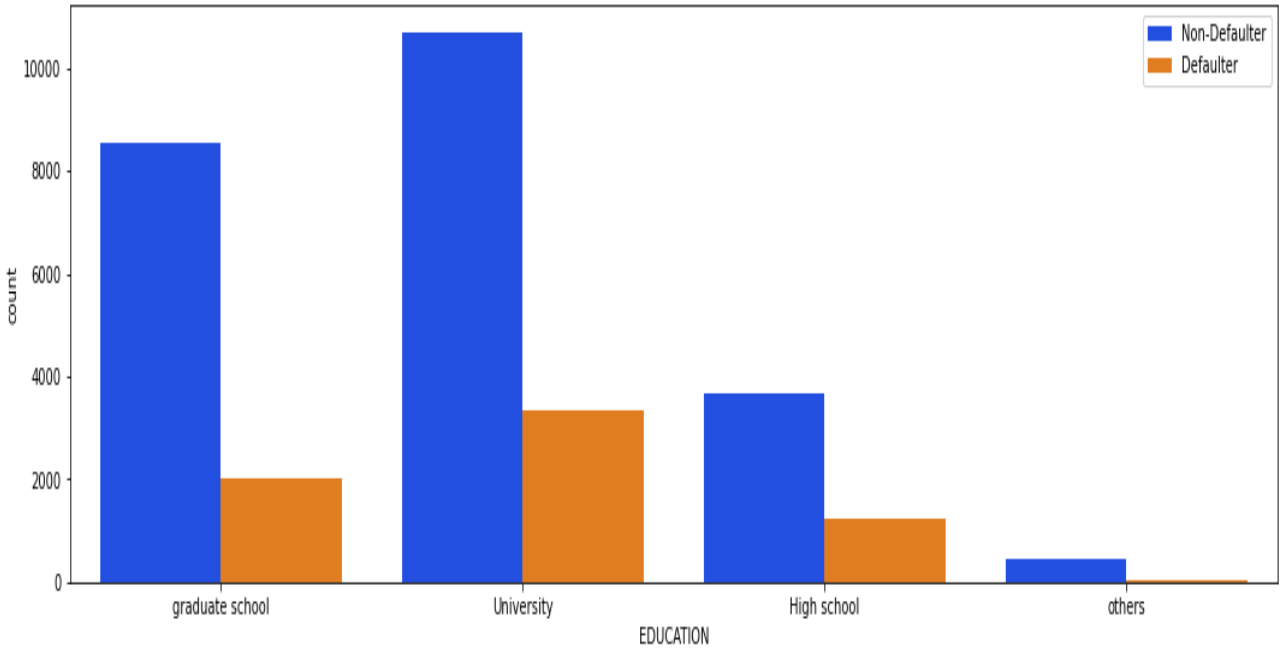




- Females are more prone to be defaulter as compared to men

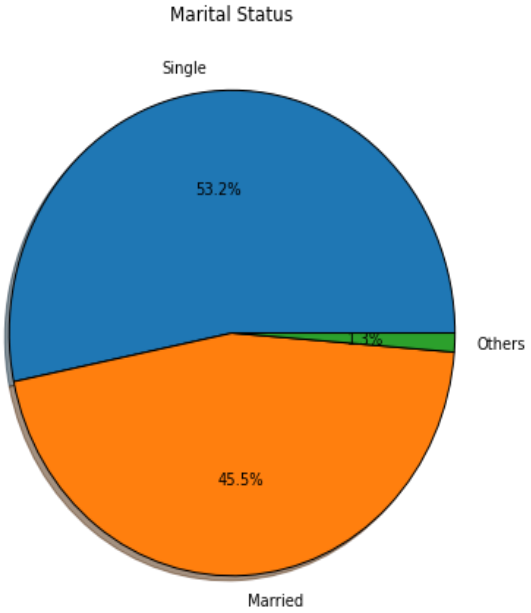
Feature Analysis - Education wise defaulter prediction

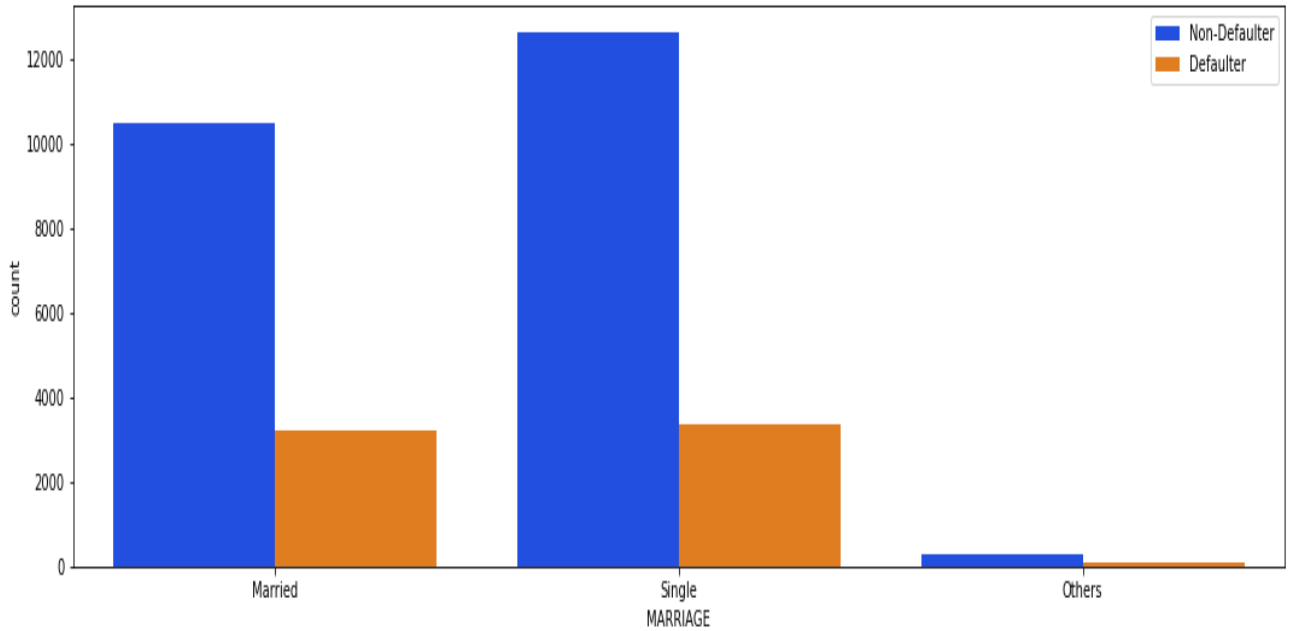




Most of the defaulters are from university followed by graduate school, high school and others.

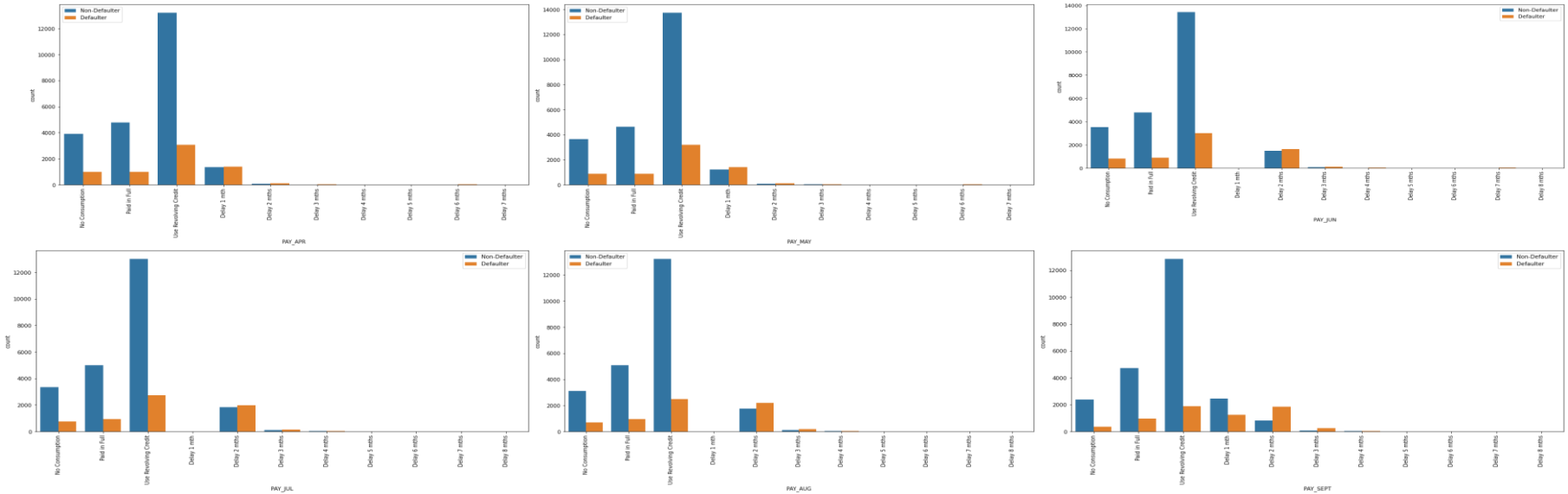
Feature Analysis – Marital status wise defaulter prediction





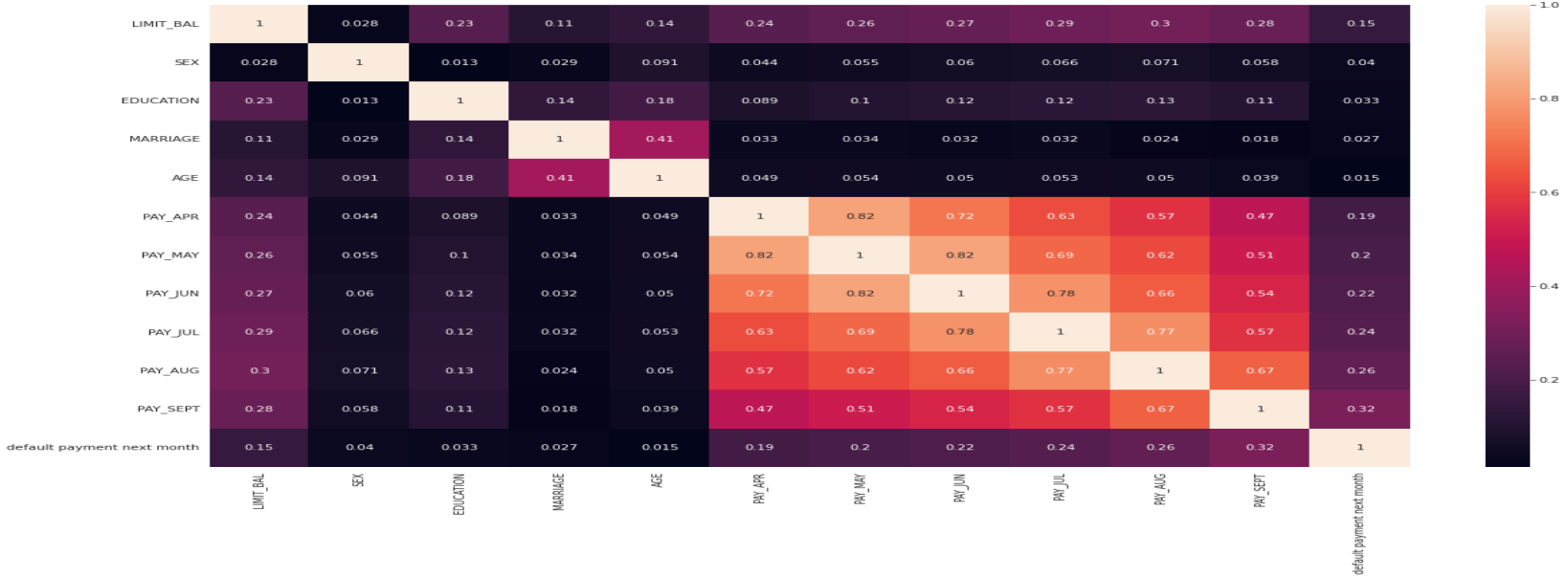
Total number of defaulters are almost same for single as well as married person.

Observation on payment history



For those who have no consumption, paid in full every month and delay 1 month, the number of no default is more than that of default. For those who use revolving credit, which means people who only pay the minimum amount every month, the non-default far exceeds the default. However, for those who delay the payment for more than one month, it turns out that the likelihood of default would then surpass the non-default,

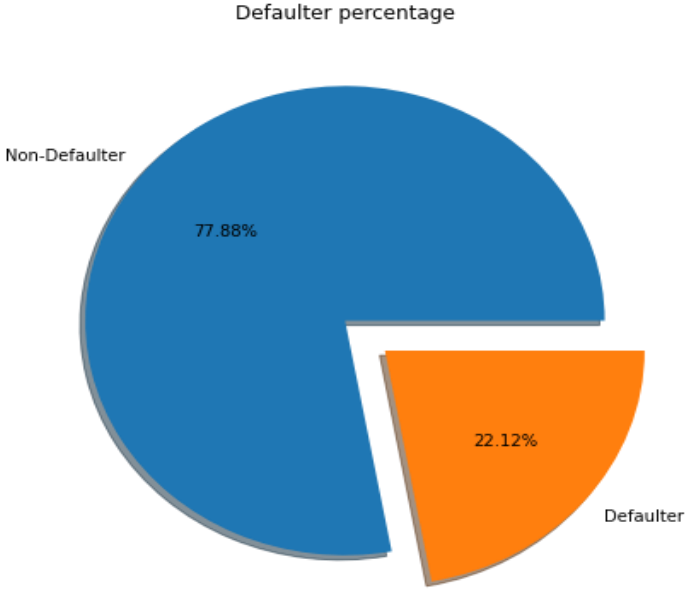
Feature Analysis - Correlation between parameters

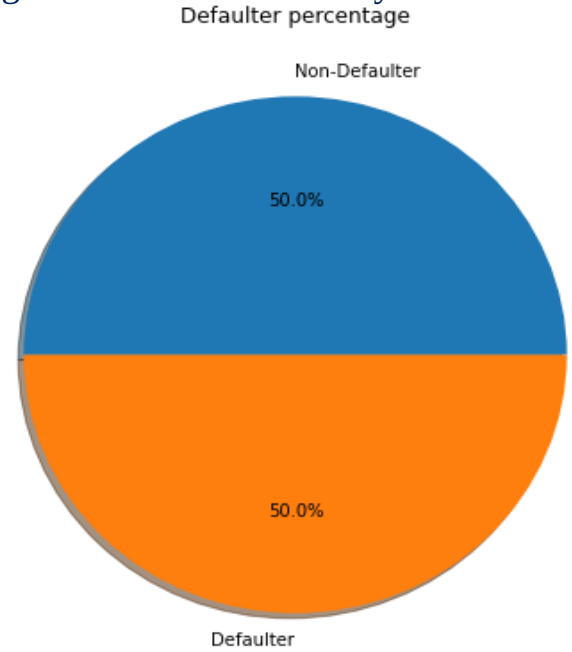


LIMIT_BAL, SEX, EDUCATION, AGE and MARRIAGE are not highly correlated to each other. PAY_APR','PAY_MAY','PAY_JUN','PAY_JUL','PAY_AUG','PAY_SEPT have correlation value > 0.5 with each other which means that people who can pay the bill on time will have high possibility to pay next bill on time.

SMOTE(Synthetic Minority Oversampling Technique)

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class.





Before SMOTE

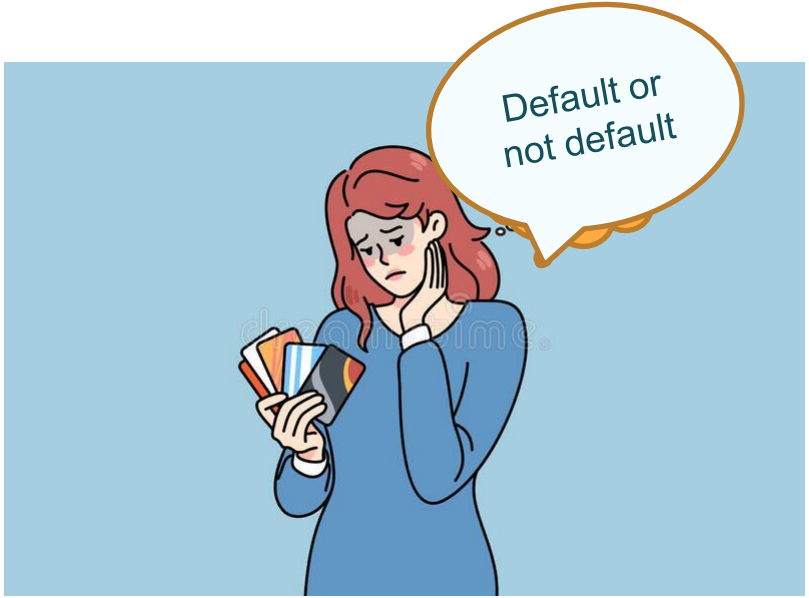
After SMOTE

Modelling Overview

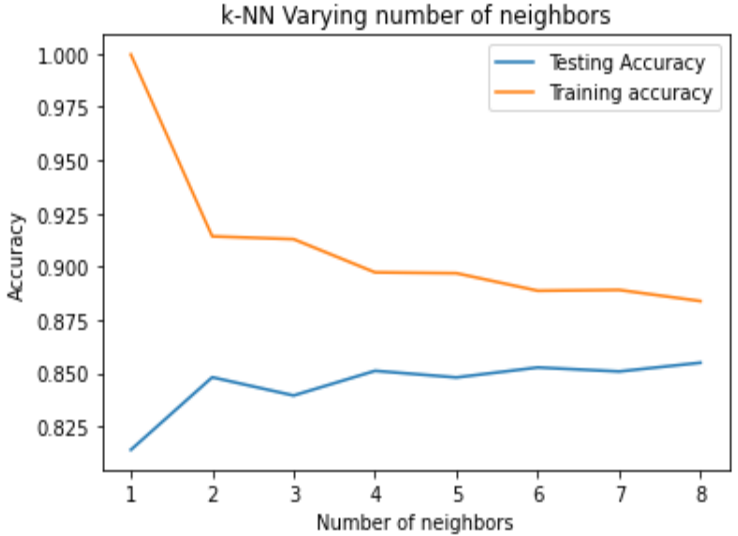
This is Classification problem statement. The data is split in training and test in 70:30.

We applied the following models :

- KNN
- Random Forest
- XGBoost
- SVM
- LGBM



Implementing K-NN




```
The accuracy on test data is  0.8526240115025162
The precision on test data is  0.7624730409777138
The recall on test data is  0.9301876863708122
The f1 on test data is  0.8380214917825538
The roc_score on test data is  0.8644725738468753
```

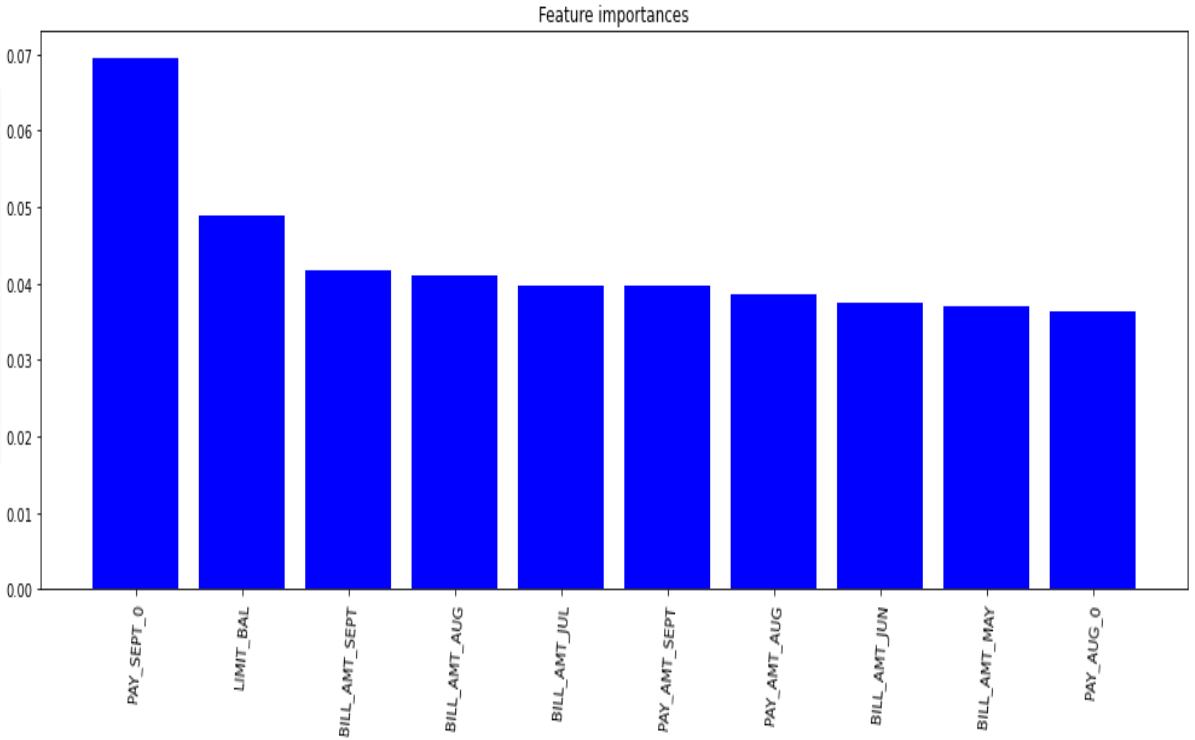
we choose n_neighbors = 8, because at n_neighbors = 8, accuracy of both training and testing are almost equal

Implementing Random Forest Classifier

Best Parameters : max_depth = 50, n_estimators = 300

```
The accuracy on test data is  0.8389647735442128
The precision on test data is  0.7679367361610352
The recall on test data is  0.8950896597955421
The f1 on test data is  0.8266522210184182
The roc_score on test data is  0.8459459377915141
```

Pay_sept_0(paid revolving credit) is the major reason of defaulter as per Random forest.

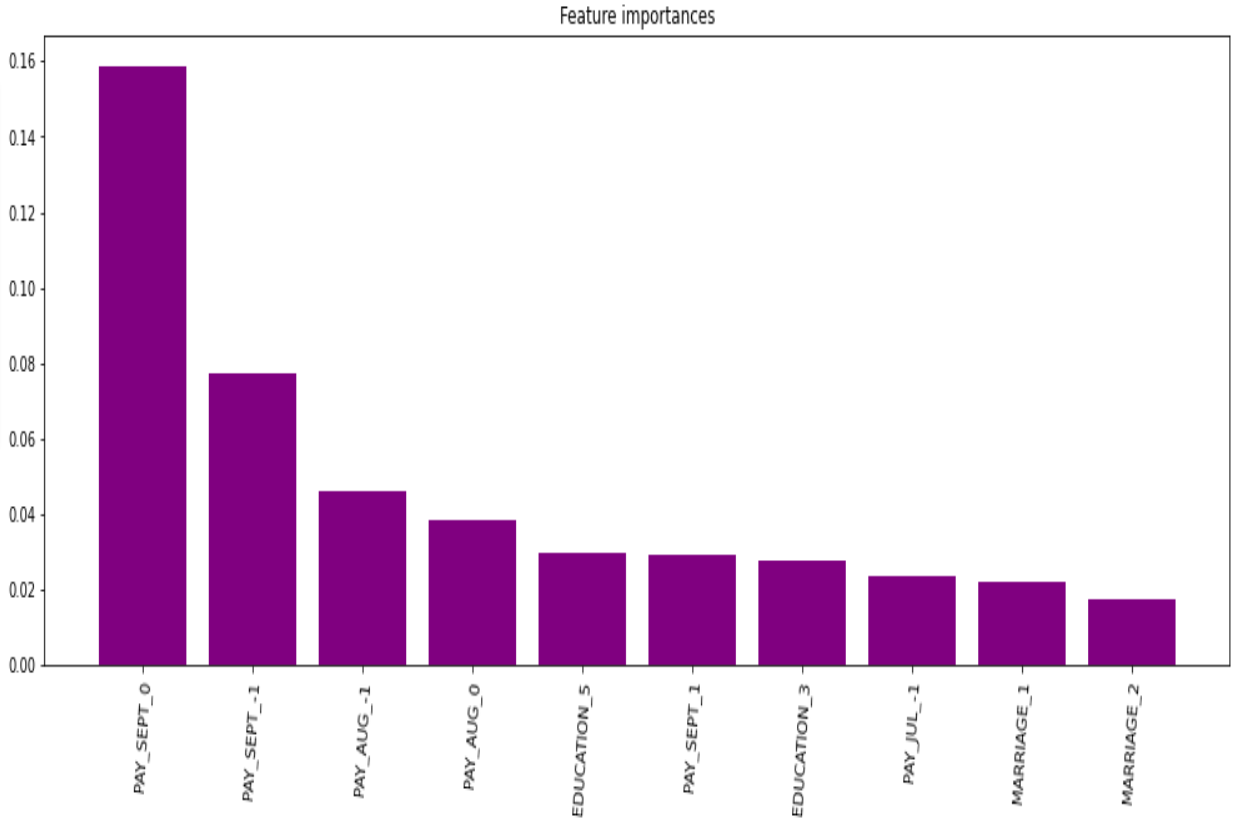


Implementing the XGBoost

Best Parameters : max_depth = 9, min child weight = 1

```
The accuracy on test data is  0.857872034507548 6
The precision on test data is  0.8015815959741194
The recall on test data is  0.9032728451069345
The f1 on test data is  0.8493943779995429
The roc_score on test data is  0.8624660943032735
```

Pay_sept_0(paid revolving credit) is the major reason of defaulter as per XGBoost.



SVM (support vector machine)

Best parameters : C = 1, gamma = 0.1, kernel = rbf

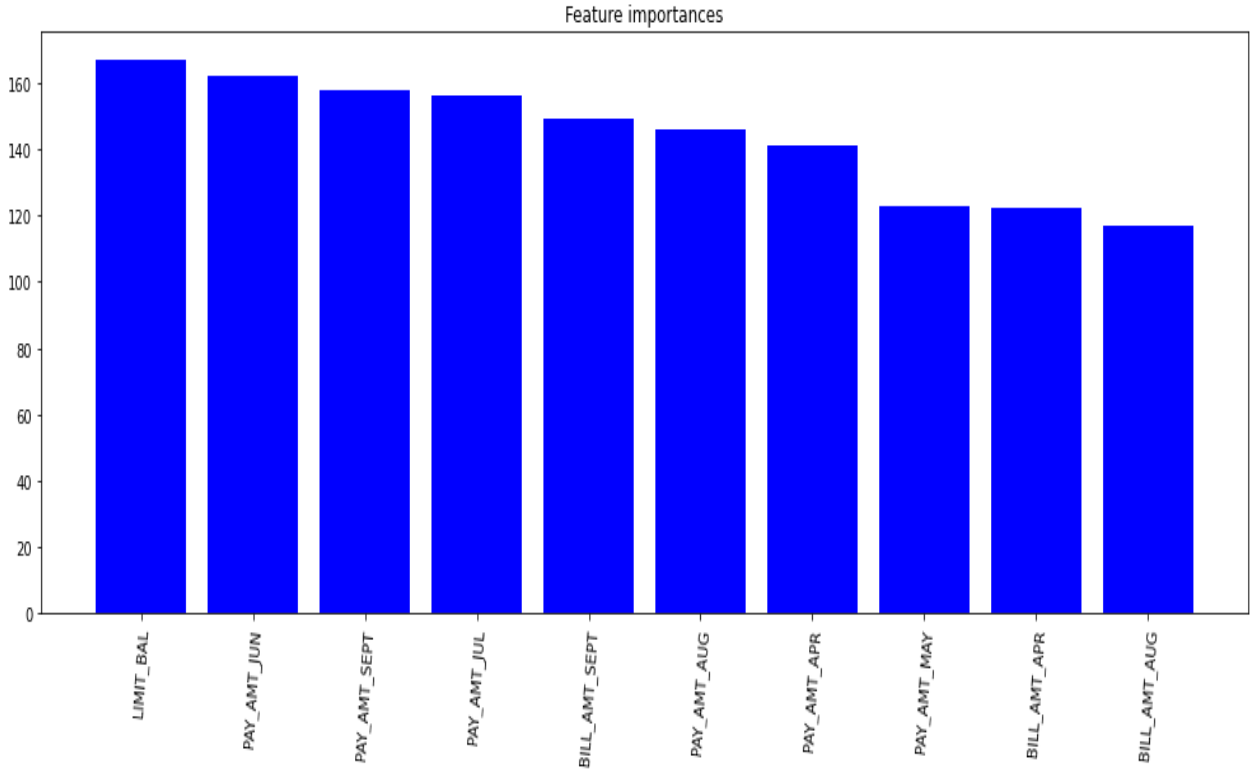
```
The accuracy on test data is  0.8465132997843278
The precision on test data is  0.725233644859813
The recall on test data is  0.9574791192103265
The f1 on test data is  0.8253292972265401
The roc_score on test data is  0.8681748755042606
```

Implement the LGBM

Best Parameters : n_estimators= 100, max_depth = 11 , learning rate = 0.3

```
The accuracy on test data is  0.8616822429906542
The precision on test data is  0.8050323508267434
The recall on test data is  0.9078968704394357
The f1 on test data is  0.8533760097546106
The roc_score on test data is  0.8663854751267306
```

LIMIT_BAL(credit limit) is the major reason of defaulter as per LGBM.

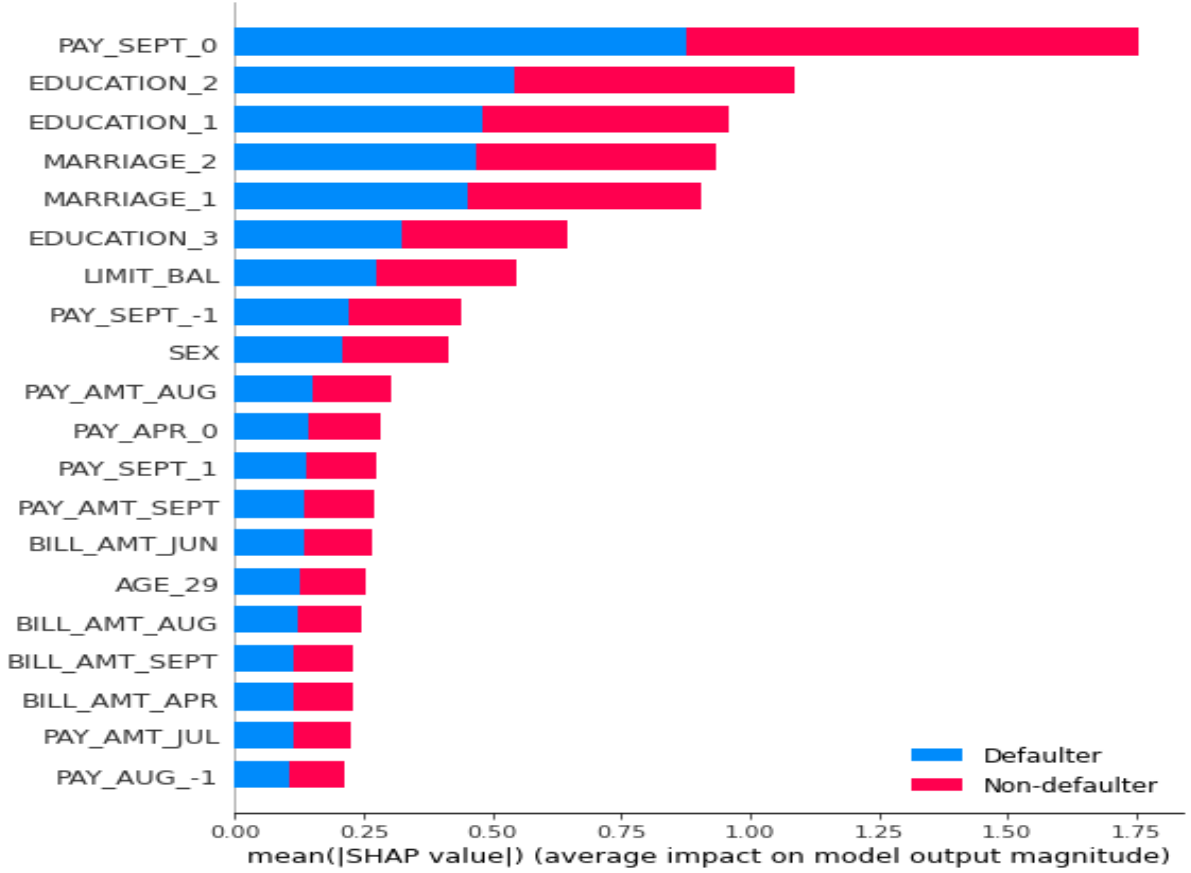


Model Evaluation

	Classifier	Test Accuracy	Precision Score	Recall Score	F1 Score	ROC_score
0	knn	0.852624	0.762473	0.930188	0.838021	0.864473
1	Random Forest	0.838965	0.767937	0.895090	0.826652	0.845946
2	Xgboost	0.857872	0.801582	0.903273	0.849394	0.862466
3	SVM	0.846513	0.725234	0.957479	0.825329	0.868175
4	LGBM	0.861682	0.805032	0.907897	0.853376	0.866385

LGBM has highest test accuracy(0.861682), highest precision score (0.805032), highest F1 score (0.853379) and highest ROC score (0.866385) among all the models. LGBM outperforms all the other models.

SHAP features



In this plot, the impact of a feature on the target class is stacked to create the feature importance plot. From here we can see that PAY_SEPT_0 has highest importance on the target 'Defaulter', while other features such as EDUCATION_2, EDUCATION_1 and subsequently all other features showed importance in decreasing order.

Challenges

- Reading the dataset and understanding the problem statement.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Dealing with Imbalanced Dataset
- Careful tuning of hyper parameters as it affects accuracy.
- Computation time was a big challenge for us.

Conclusion

Descriptive Analytics

Important takeaways from the are:

- Most of the defaulters are university passout.
- Most of the defaulter has the balance limit in the range 50000 to 200000.
- Credit cardholders who have no consumption or paid in full every month or delay 1 month, the number of no default is more than that of default. For those who use revolving credit, which means people who only pay the minimum amount every month, the non-default far exceeds the default. However, for those who delay the payment for more than one month, it turns out that the likelihood of default would then surpass the non-default, which also means the longer the payment delay, the higher risk for that person on default.

MODELS

- The ROC score of KNN Classifier, Random Forest model, XGBoost model, SVM model and LGBM model are 0.864, 0.845, 0.862, 0.868 and 0.866 respectively.
- LGBM has highest test accuracy(0.861682), highest precision score (0.805032), highest F1 score (0.853379) and highest ROC score (0.866385) among all the models. LGBM outperforms all the other models.

Q & A