Capstone Project - 3
Credit Card Risk Prediction

Team Member
Harish Kollana

Discussion Points

1. Problem Statement
2. Data Summary
3. Explorative Data Analysis
4. Feature Engineering
5. Model Fitting
6. Feature Importance
7. Model Comparison
8. Challenges
9. Conclusion

The Dilemma

How Credit Card Works

Customer's apply for credit card with their details

The company will asses customer profile and give credit card

The users Will use the credit limit to shop & e.t.c.

The credit card is good option until the customer repay on time. But when the customer spends more than his earning limit and unable to pay the loan. The credit default happens.

Problem Statement

The Taiwan Credit card issuer issues credit limits to the customer and in that there will be defaulters and non-defaulters. Based on the limit the issuer provided, Age, Education, Gender and other features the limit is provided.

We were provided with one such already classified label in our data set containing 30,000 observations with 25 columns.

Our experiments can 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

Data Set Name : default of credit card clients.xls

Data Set Information:
Number of instances: 30,000
Number of attributes: 25

Features:
'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', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6', 'default payment next month'

Data Summary

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) 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).

Data Summary

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;

X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.

Data Summary

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

EDA

Missing Values & Data Types

There are no null values and no duplicates as well

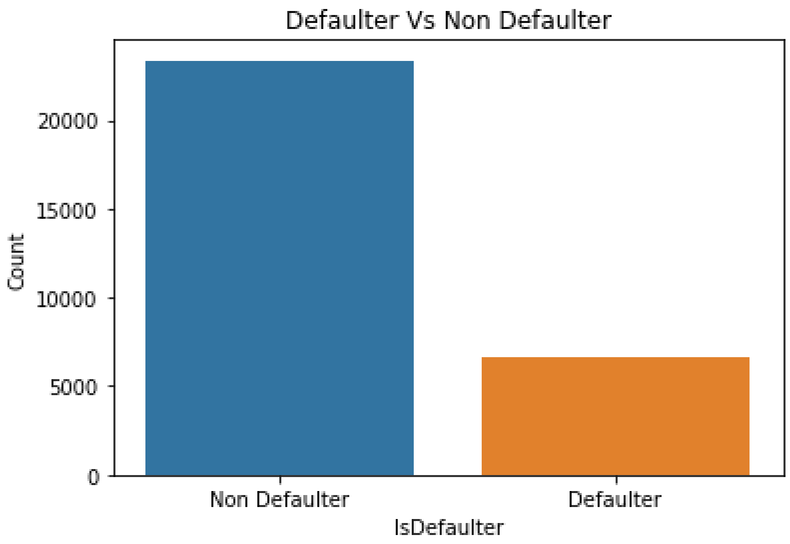
```
ID                            0
LIMIT_BAL                     0
SEX                           0
EDUCATION                     0
MARRIAGE                      0
AGE                           0
PAY_0                         0
PAY_2                         0
PAY_3                         0
PAY_4                         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
```

```
#check for duplicate values
df.duplicated().sum()
```

```
0
```

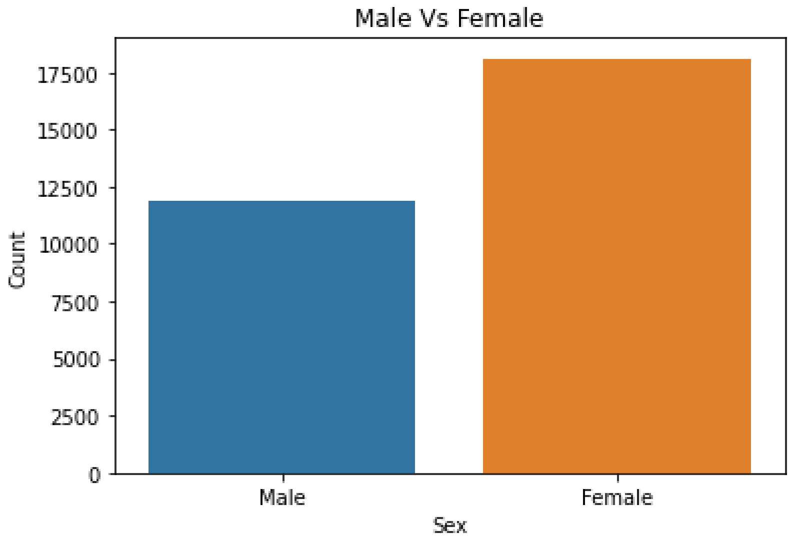
EDA (continued)

Defaulters Vs Non Defaulters



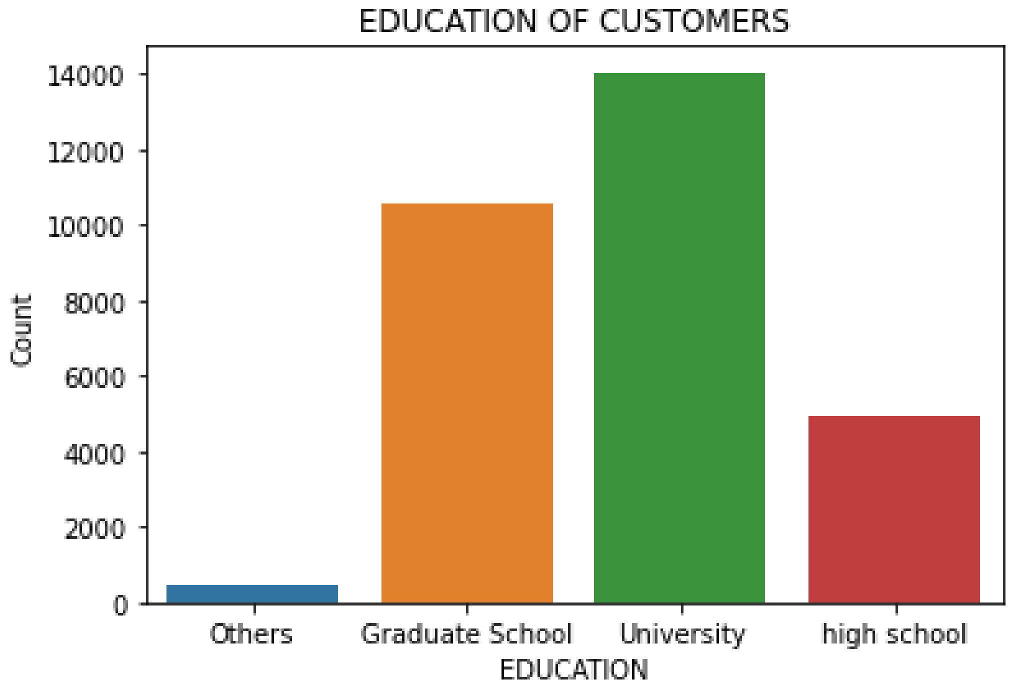
EDA (continued)

Male Vs Female Users



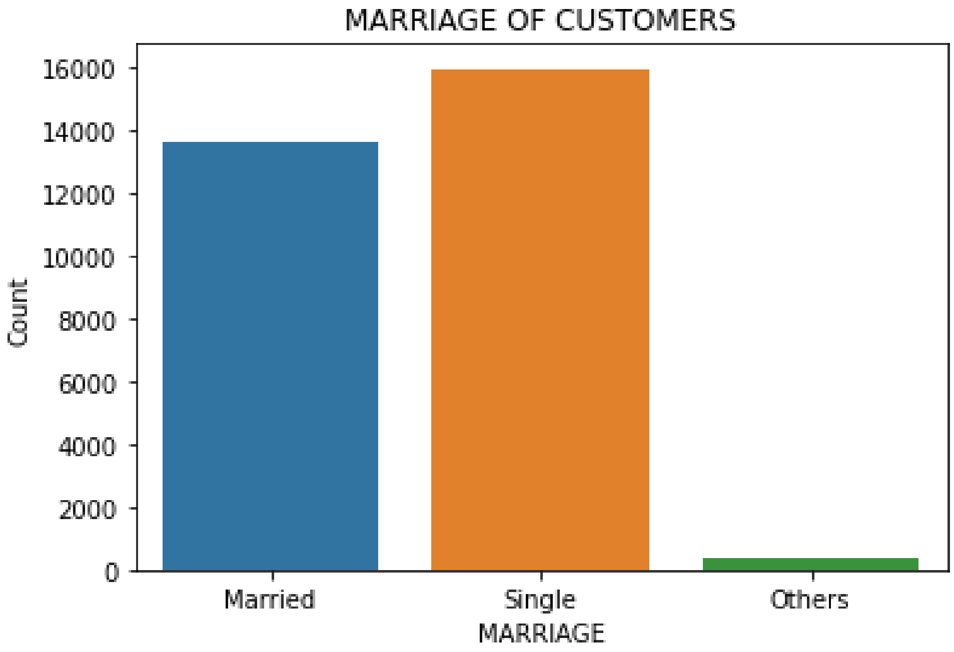
EDA (continued)

Education Stats of Customer’s



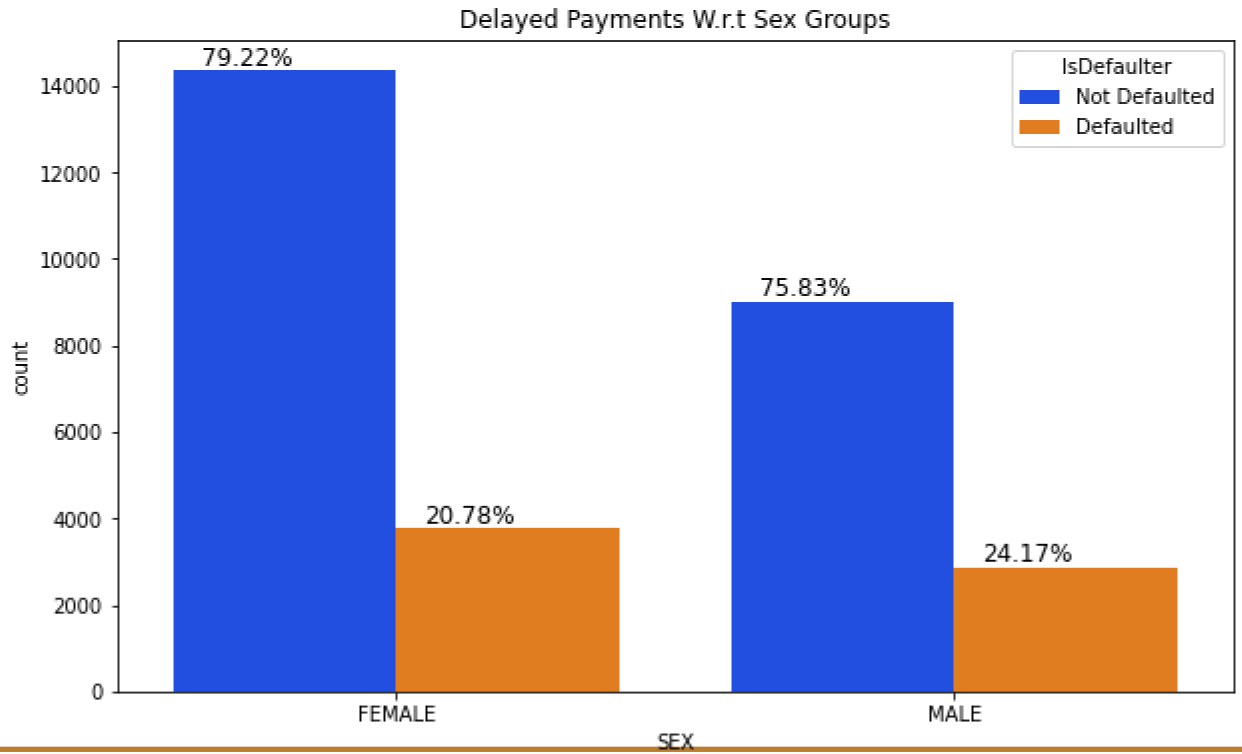
EDA (continued)

Marriage Stats of Customer’s



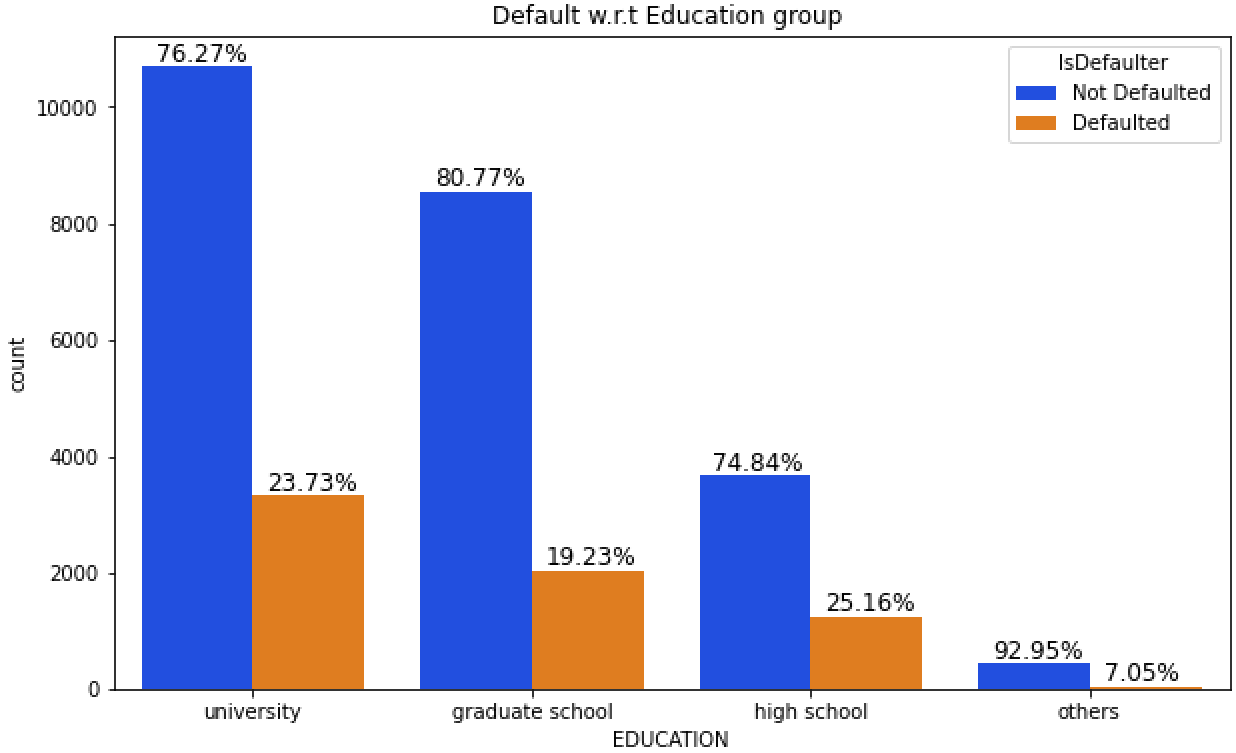
EDA (continued)

Gender Wise Defaulted Payments Percentage



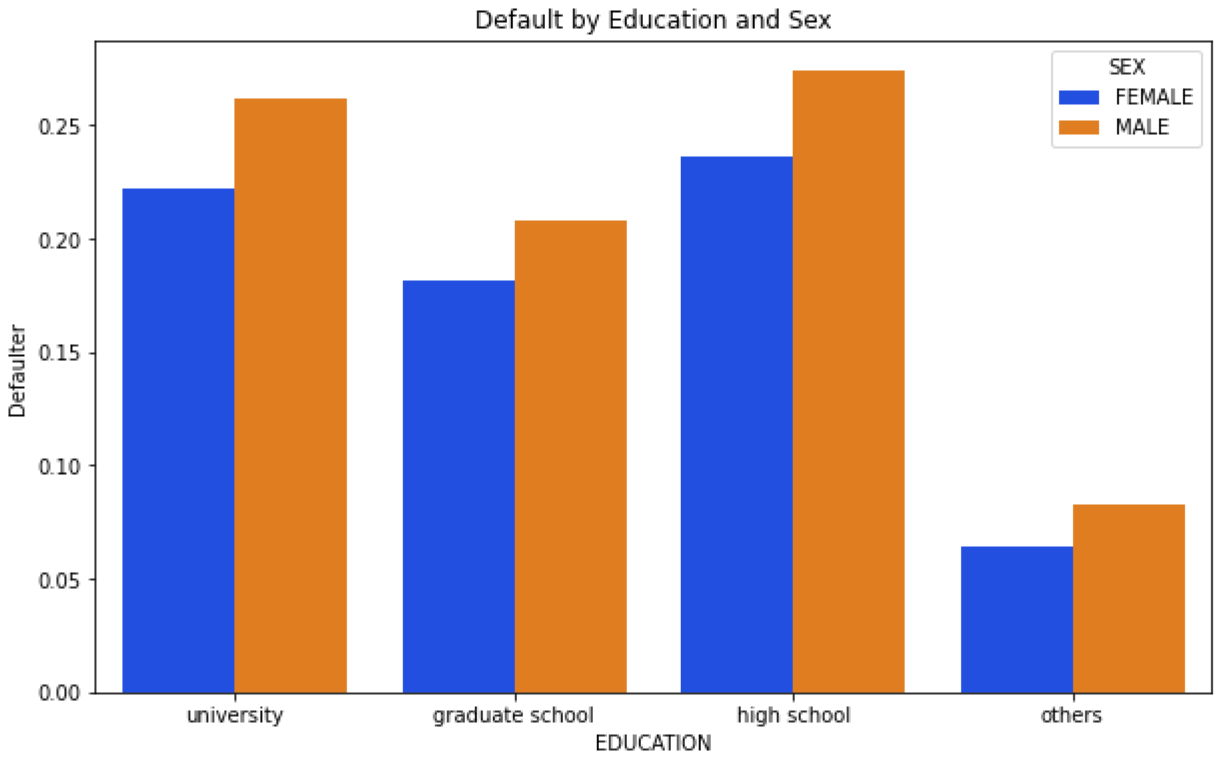
EDA (continued)

Education wise Defaulted Payments Percentage



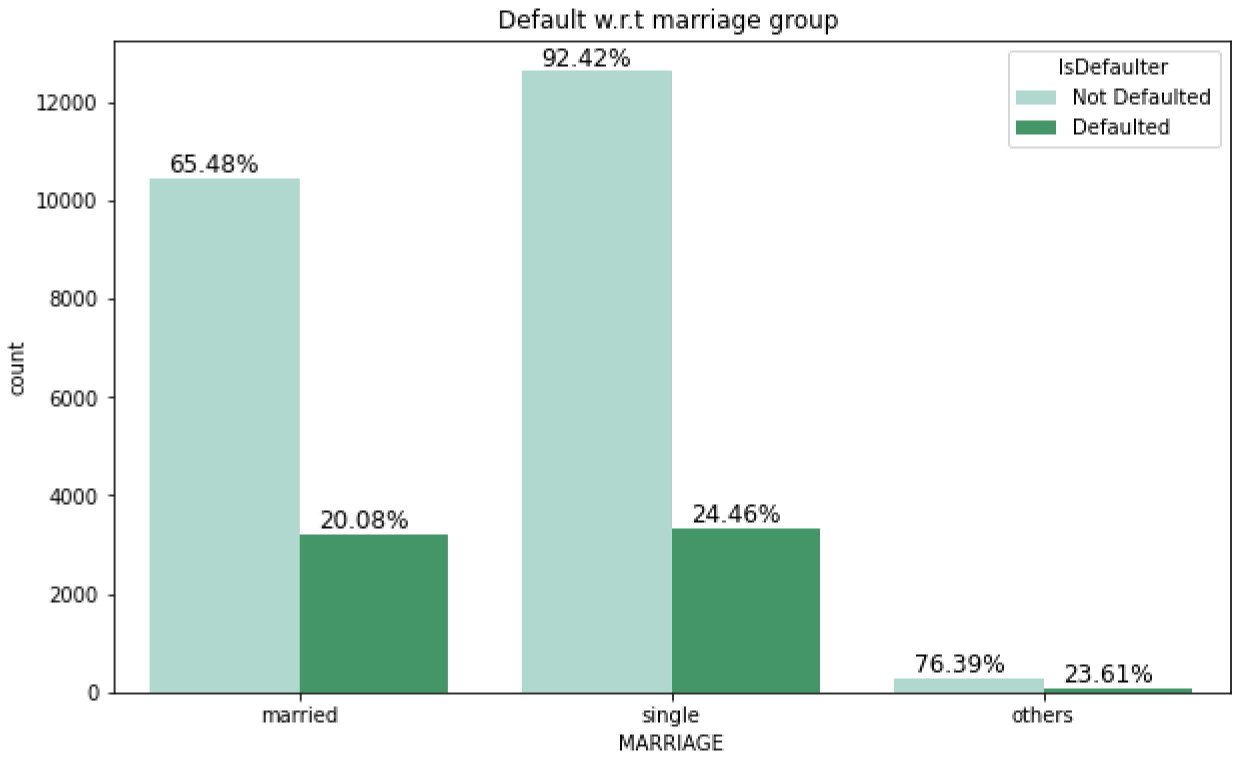
EDA (continued)

Combination of Gender and Education w.r.t Defaulters



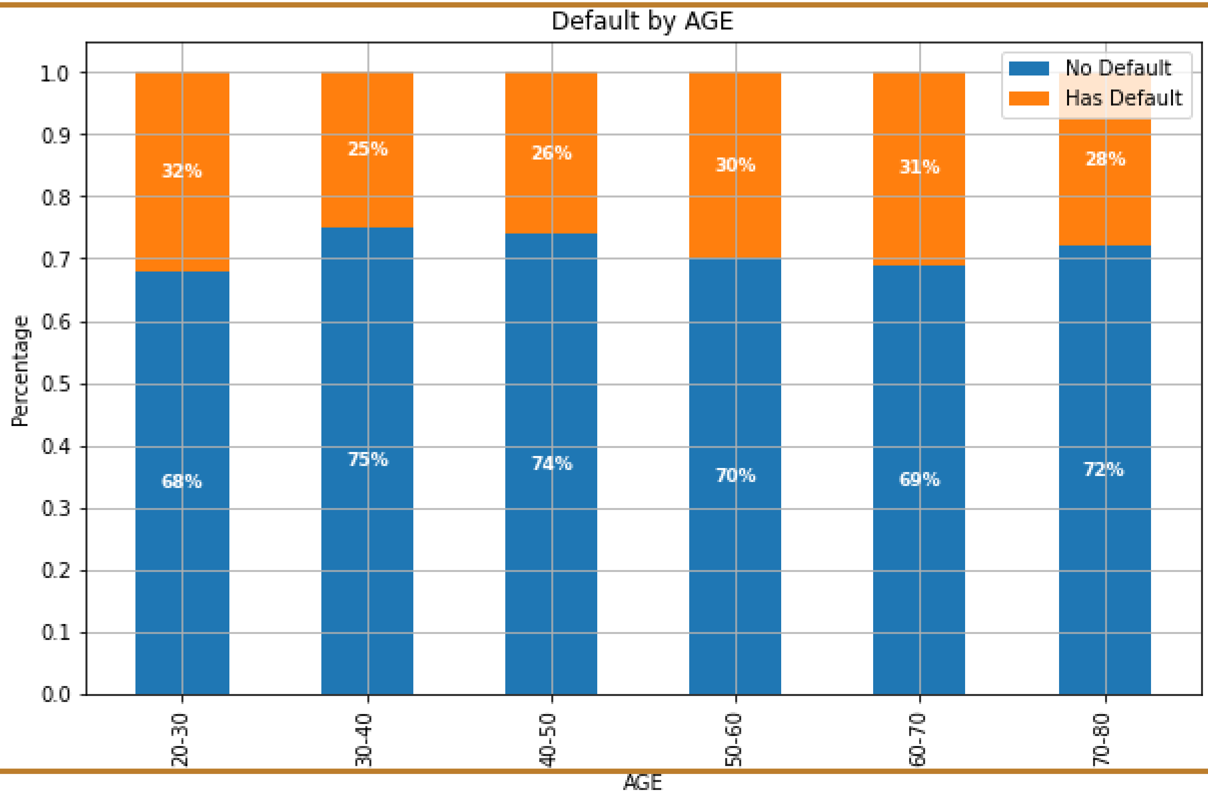
EDA (continued)

Marriage Stats w.r.t Defaulters



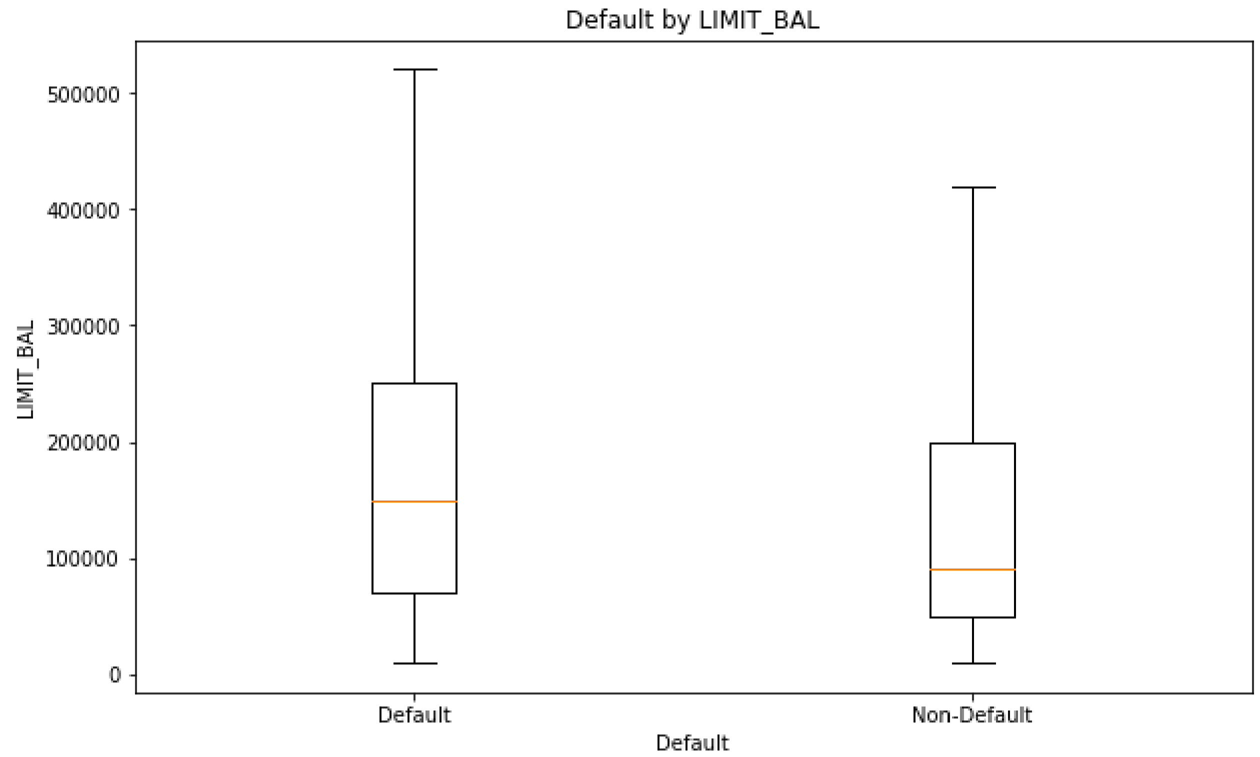
EDA (continued)

Age Intervals of customers w.r.t Defaulter's



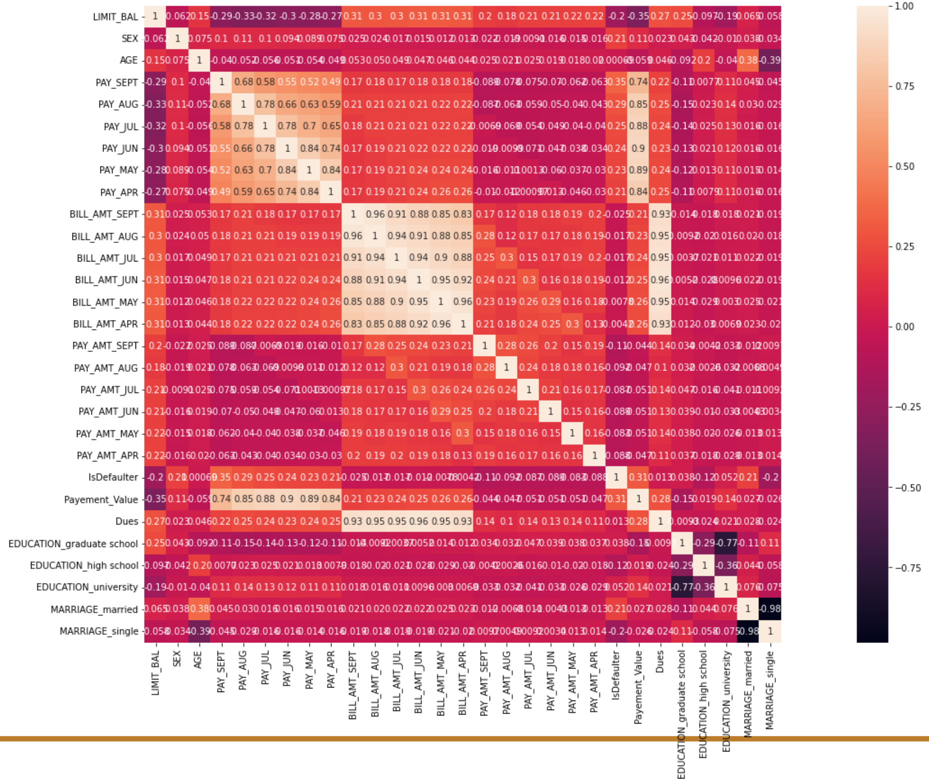
EDA (continued)

Credit Limit w.r.t Defaulters



EDA (continued)

Correlation Matrix



Feature Engineering

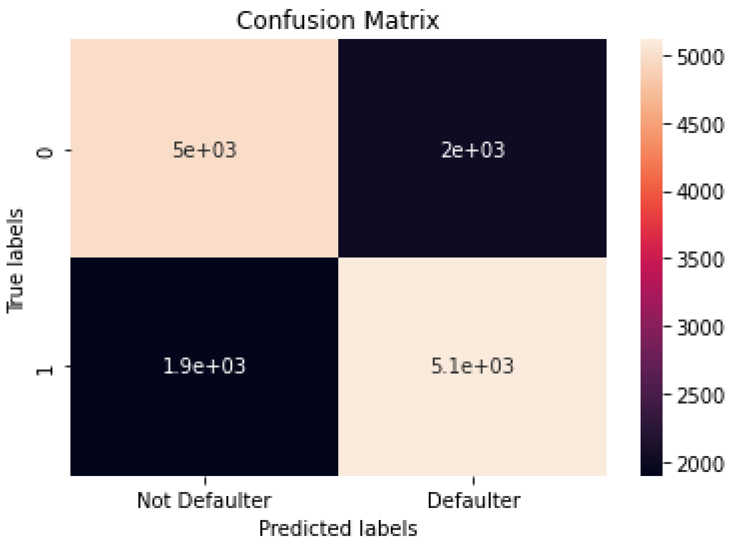
1. IsDefaulter
2. Label encoding
3. One hot encoding
4. Separating Independent and Dependent variables
5. Rescaling values using StandardScaler
6. Train test split

Model Fitting

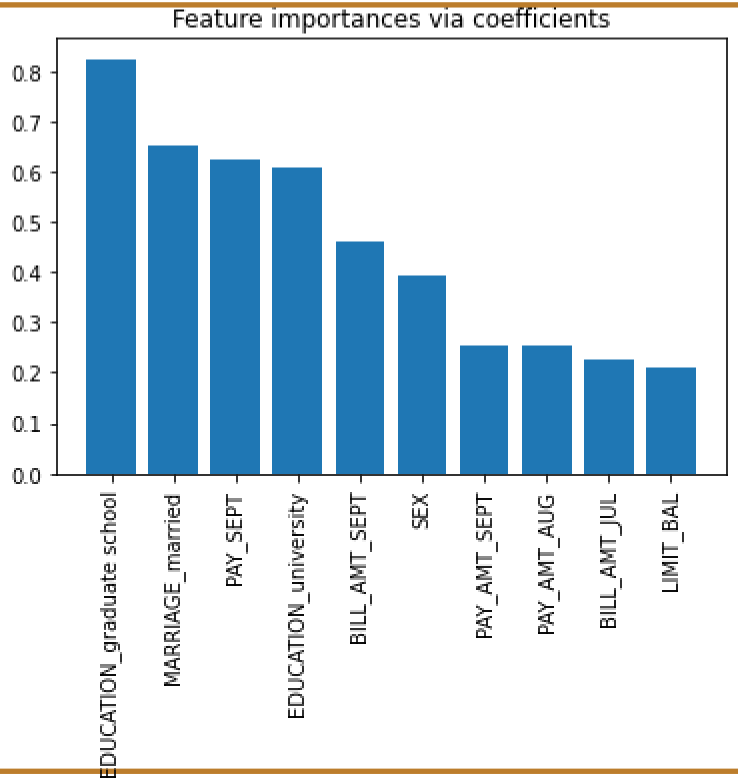
1. Logistic Regression Model
2. Random Forest Model
3. XG Boost Model

Confusion matrix by Logistic Regression

[[4991 2019]
 [1897 5112]]

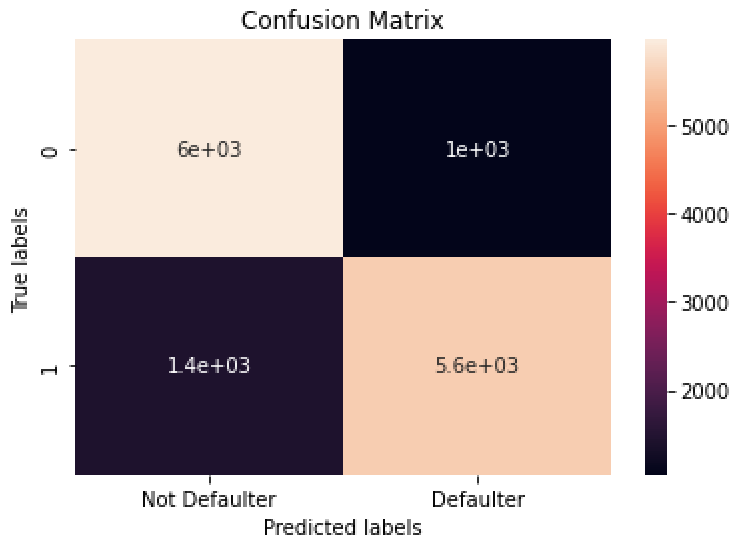


Feature Importance by Logistic Regression

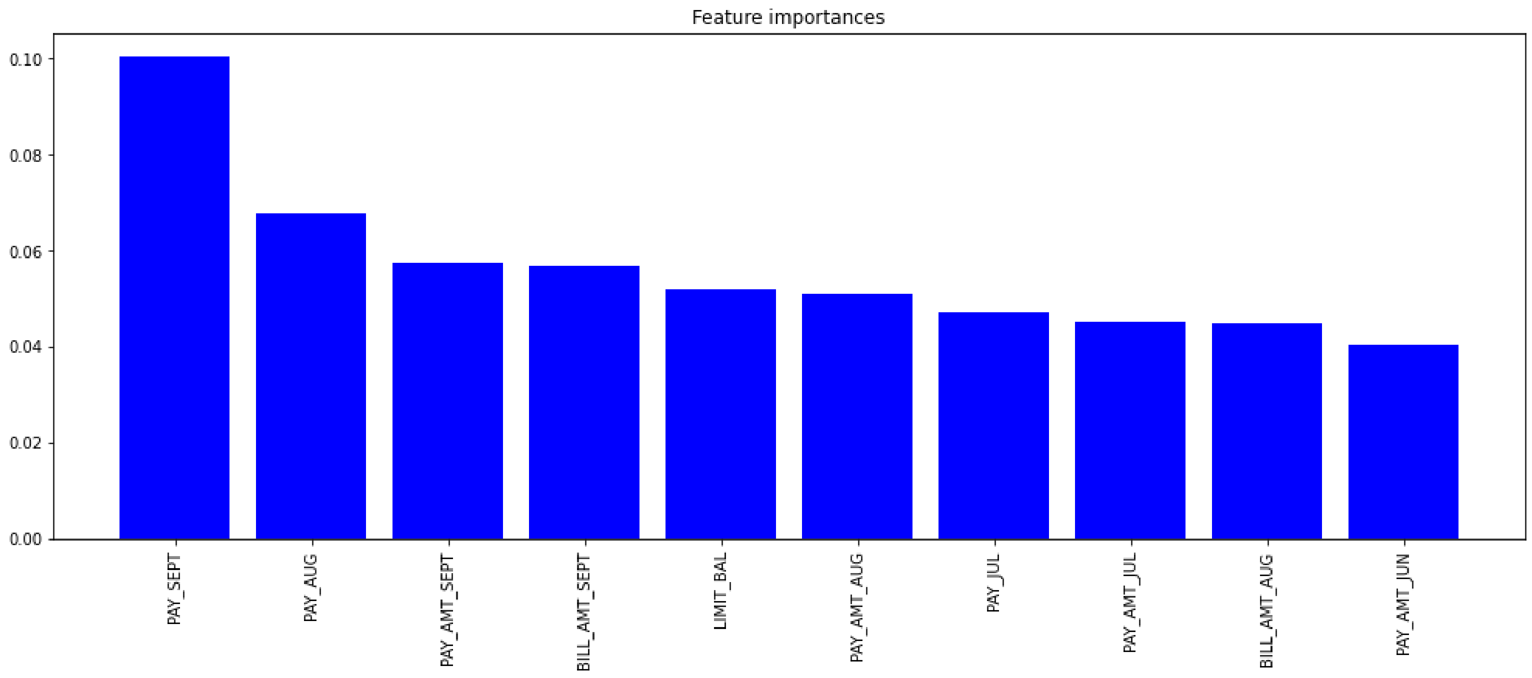


Confusion matrix by Random Forest Classifier

[[5968 1042]
 [1432 5577]]

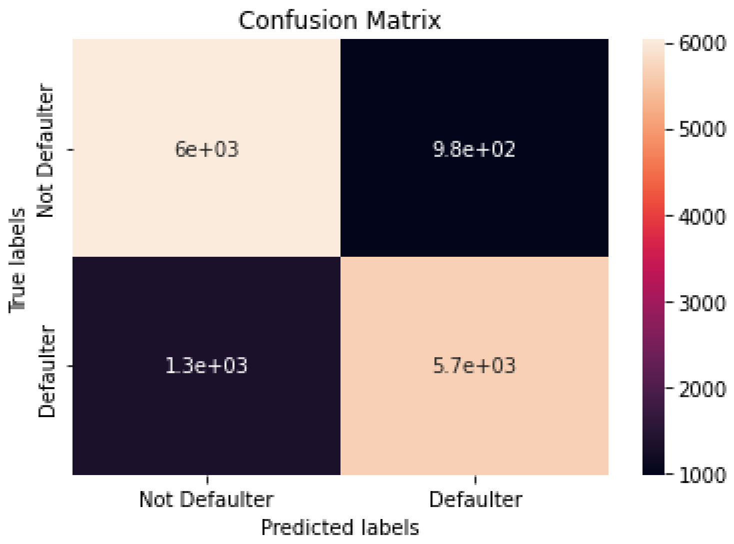


Feature Importance by Random Classifier

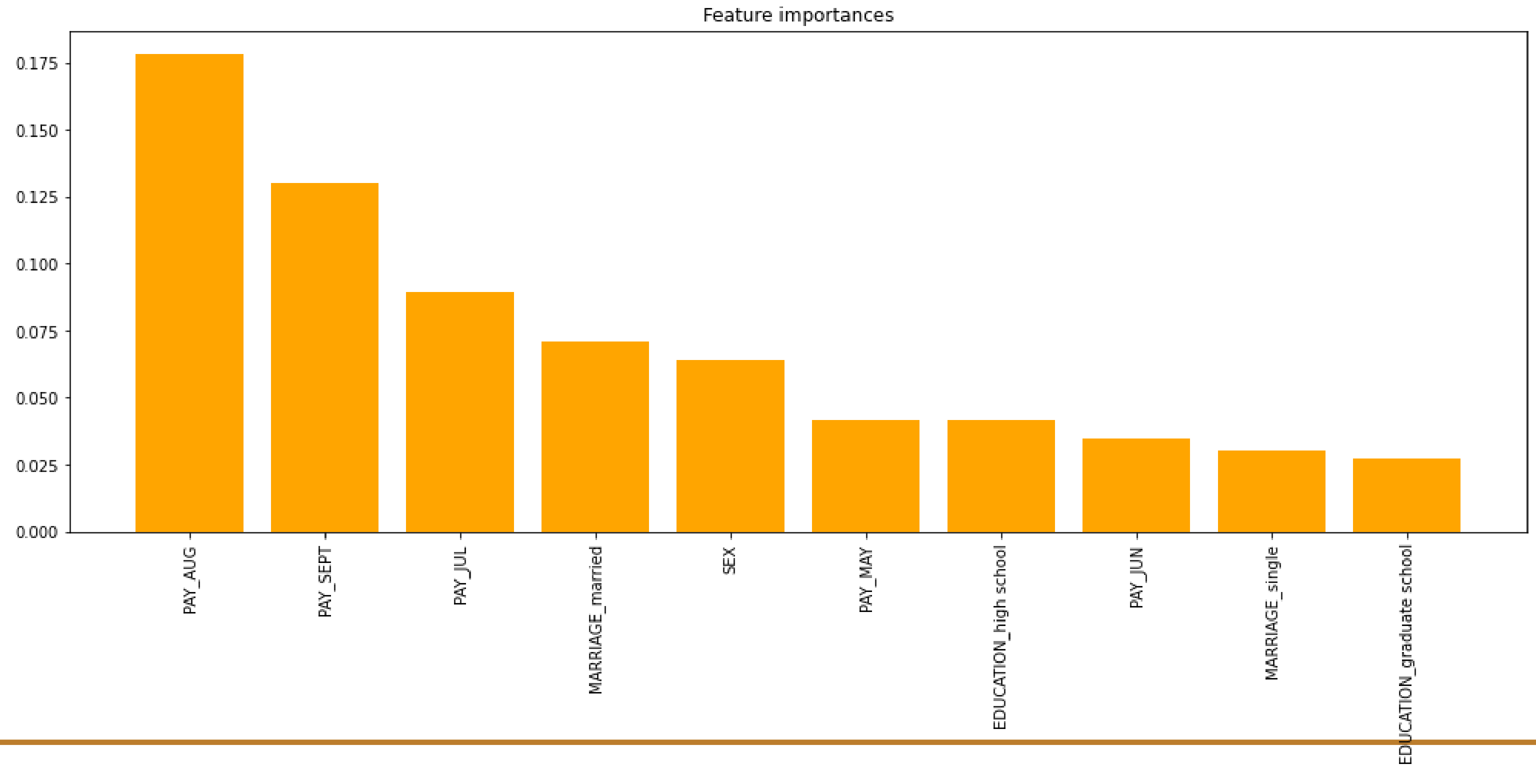


Confusion matrix by XGBOOST Classifier

[[6028 982]
[1346 5663]]



Feature Importance by XGBOOST Classifier



Model Comparison

Model	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
Logistic Regression	0.723593	0.720665	0.729348	0.716870	0.723055
Random Forest Classifier	0.956862	0.823525	0.795691	0.842574	0.818462
XGBOOST Classifier	0.947385	0.833940	0.807961	0.852220	0.829501

Challenges

1. Dataset has lot of features contains a categories in it.
2. Adding new features
3. Outliners In numerical Numbers
4. Selection comparison of models

Conclusion

We came to end stage by successfully building a model to predict whether the customer will default his / her payment

We have performed feature engineering, feature selection, hyperparameter tuning to prevent overfitting and for decreasing error.

The recall is the measure of our model correctly identifying True Positives. Thus, for all the Customers who actually default, recall tells us how many we correctly identified as is default.

As we had considered recall, XGBoost is our best model as we can see roc aoc curve is maximum.

Thank You