Capstone Project : 3

Credit Card Default Prediction

TEAM MEMBERS

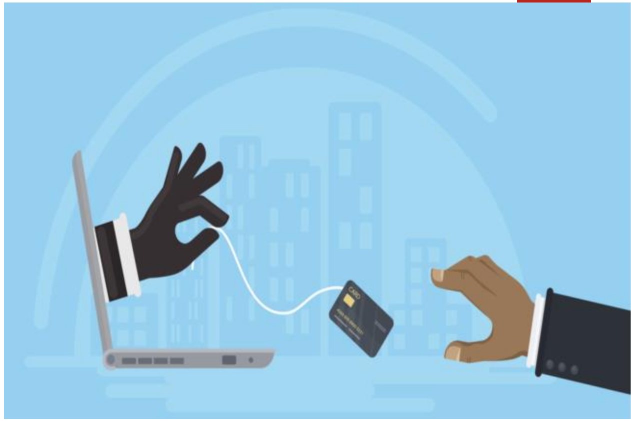
Team Member

Atul Chouhan

Content

1. ❑ Understanding Business Problem
2. ❑ Dataset Information
3. ❑ Feature Analysis
4. ❑ Exploratory Data Analysis
5. ❑ Data Pre-processing
6. ❑ Model Implementing
7. ❑ Challenges
8. ❑ Conclusions

Understanding Business Problem



→ Topic – "Credit Card Default Prediction"

→ 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".

-> **Target** is to minimize the risk that a customer being a payment defaulter, and maximize the profit of the bank

Dataset Information

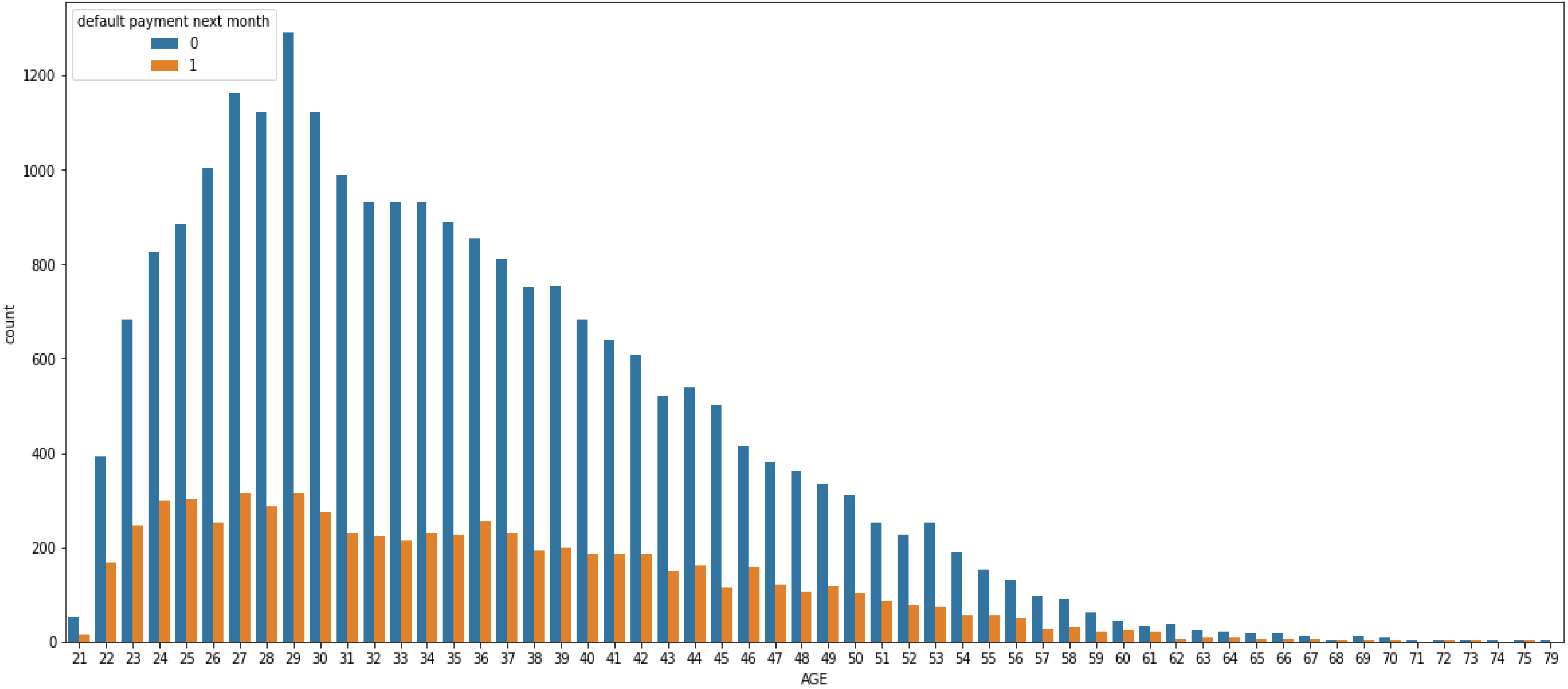
- This dataset contains 29999 observations and 23 features that contain the data of last six months of customer.
- There are 3 categorical features in our dataset.
- This dataset is from the city of Taiwan and doesn’t have any null or duplicate values.

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
20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

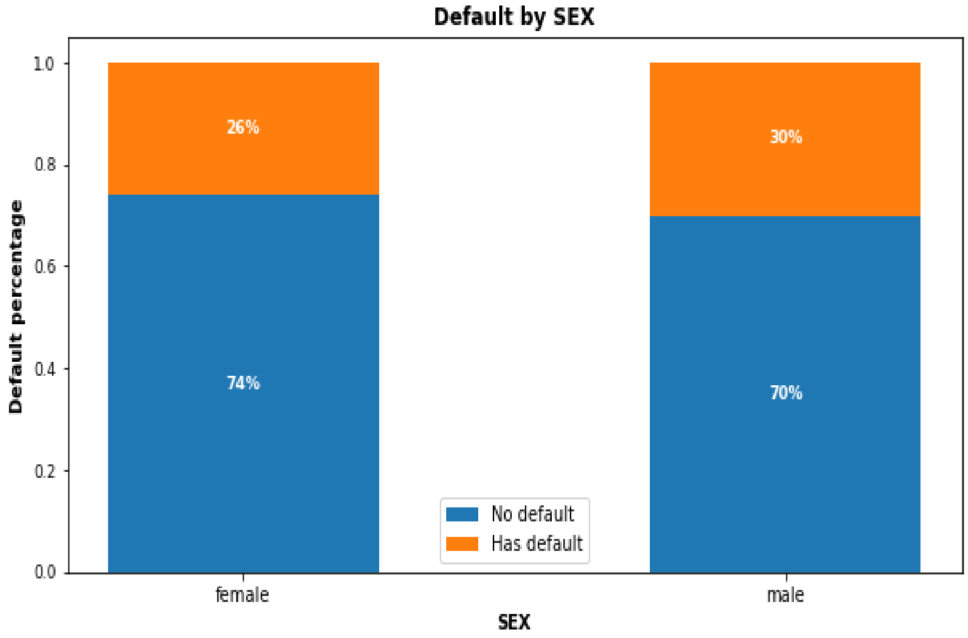
Feature Summary

- X1: Amount of the given credit, includes both individual and family credit.
- X2: Gender(1=Male and 2=Female)
- X3: Education(1=graduate, 2= university, 3= high school and 4= others)
- X4: Marital status (1= Married, 2 = single, 3= others)
- X5: Age in year.
- X6-X11: History of past payment from April to September
- X12-17: Amount of bill statement fro April to September
- X18-X23: Amount of previous payment from April to Se
- Y: Default payment

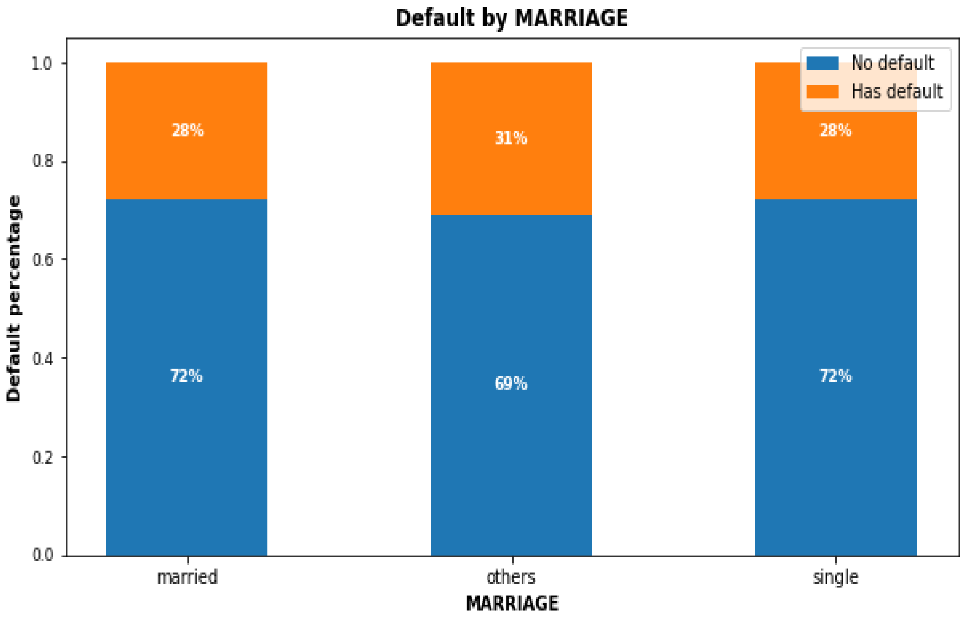
Feature Analysis Of Age column



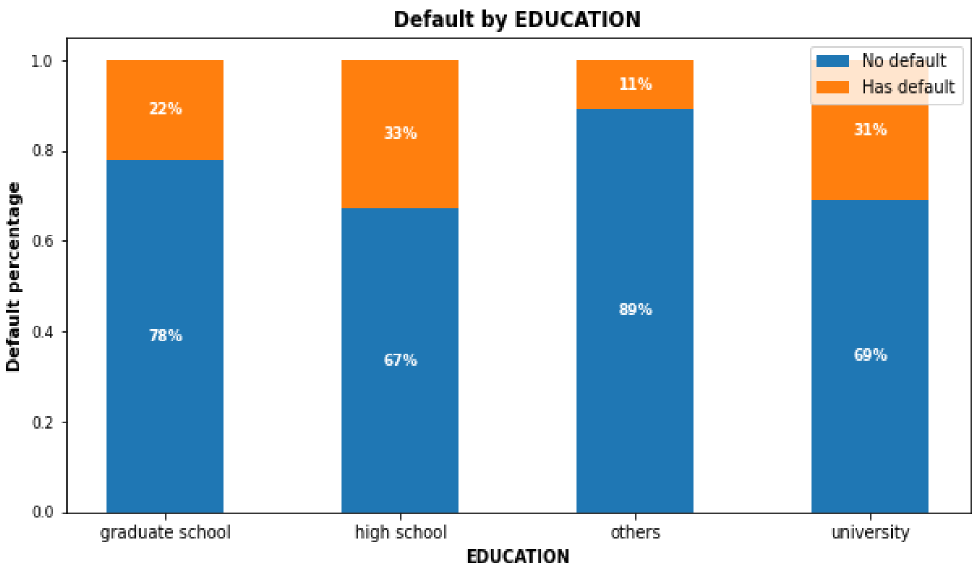
Analysis Of Gender column



Analysis Of Marriage column

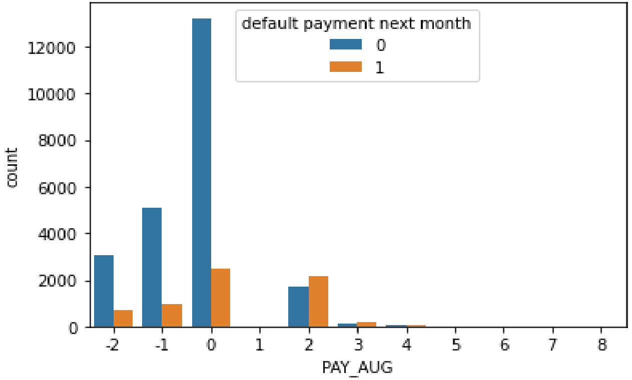


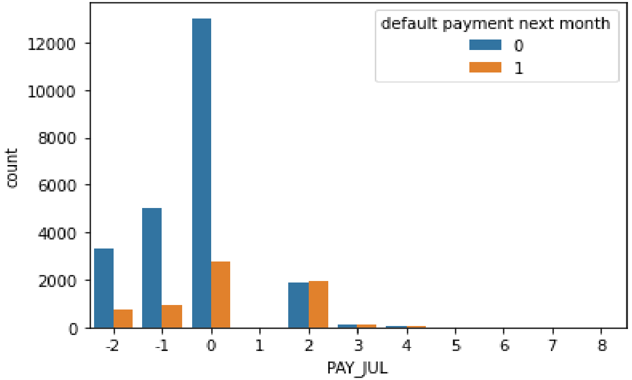
Analysis Of Education Column

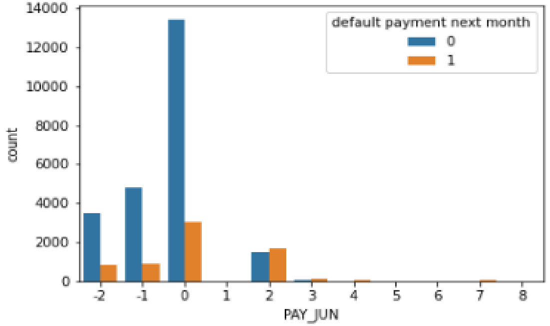


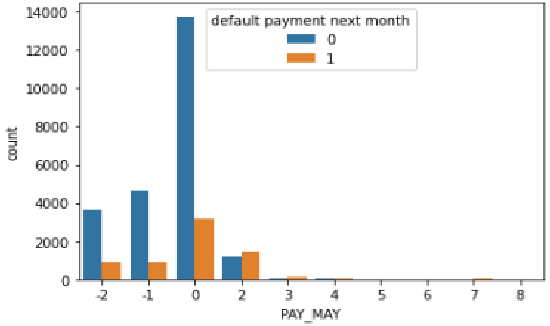
- The data indicates customers with lower education levels default low%. Customers with high school and university educational level had higher default percentages than customers with grad school education

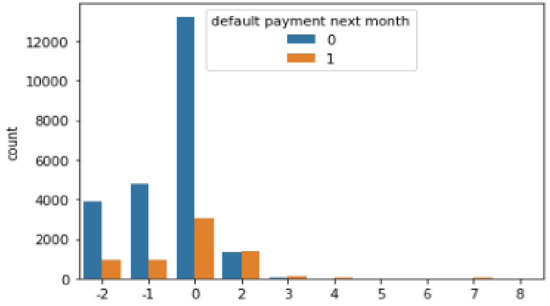
Analysis Of Repayment Month Wise



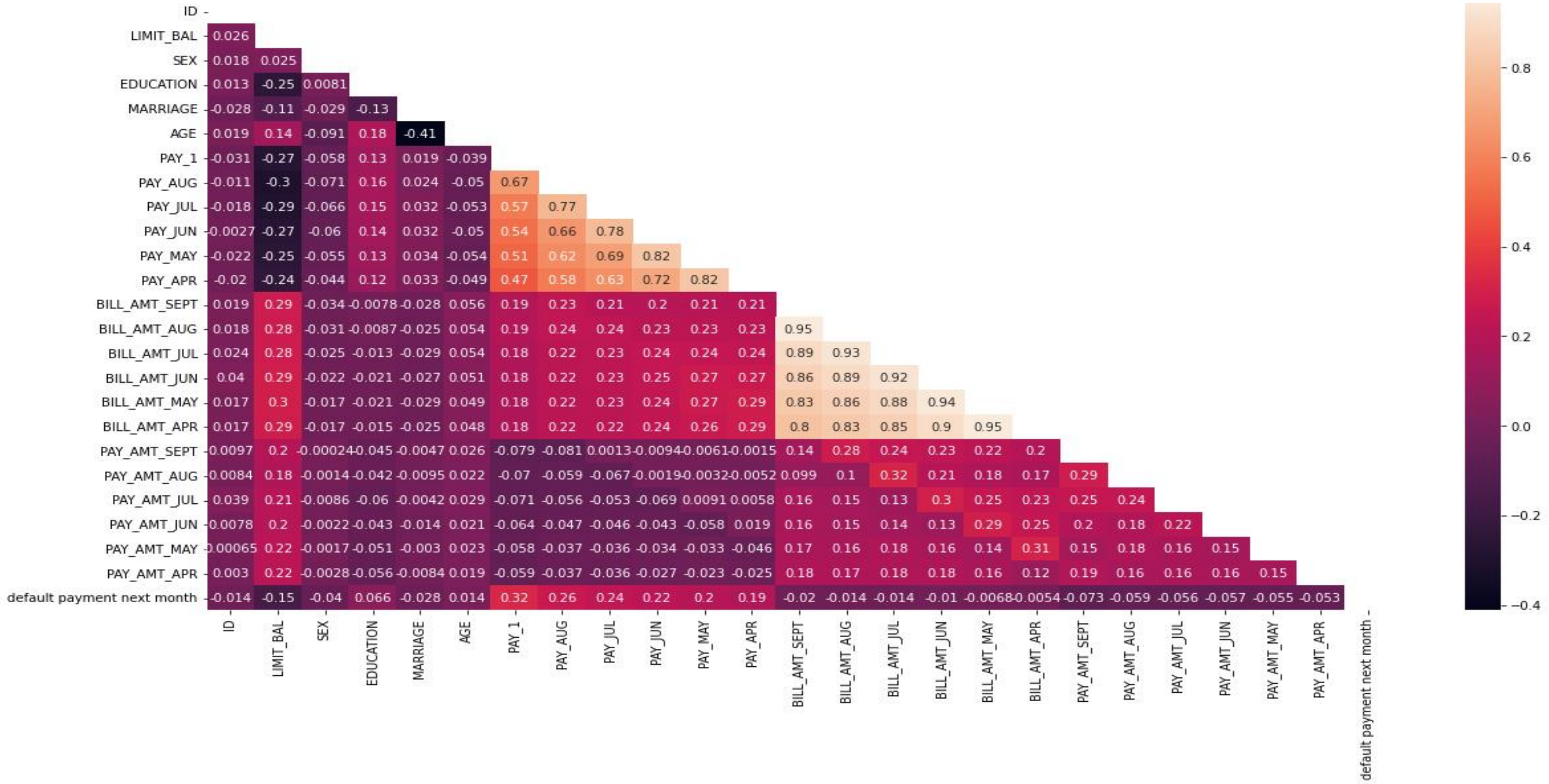






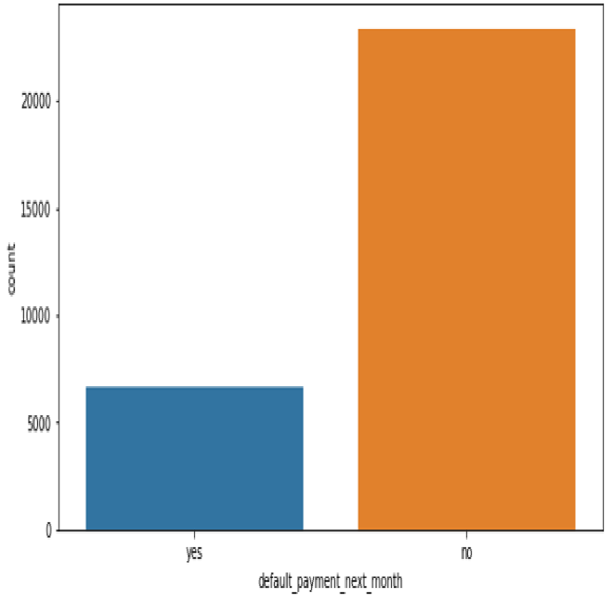


Correlation Matrix (Heatmap)

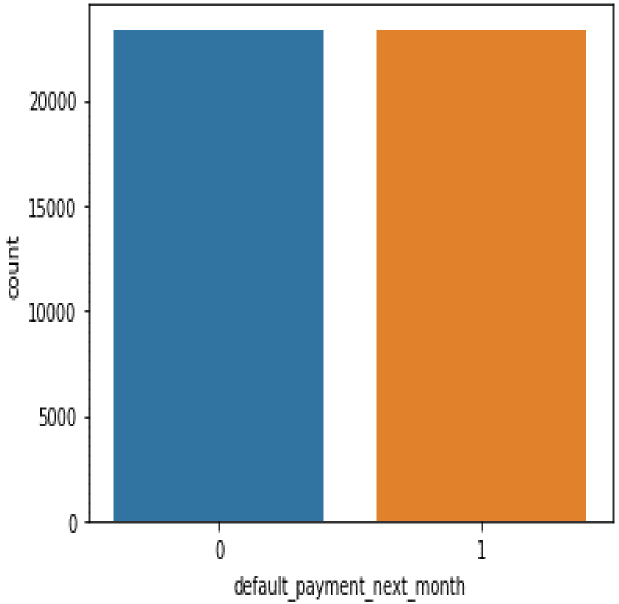


SMOTE(Synthetic Minority Oversampling Technique)

→It's a clear case of class imbalance, to balance both the class we apply 'SMOTE'



→
SMOTE

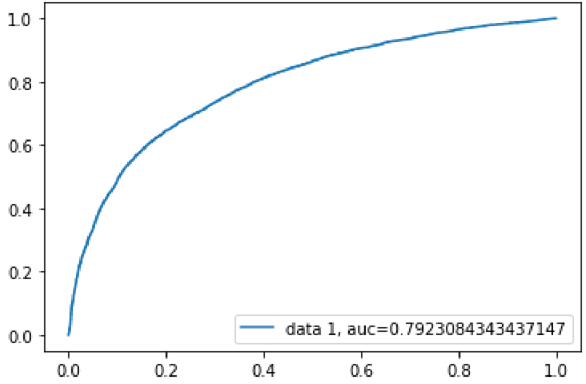


Model Implementation

→Logistic Regression

Hyper-Parameter Tuning

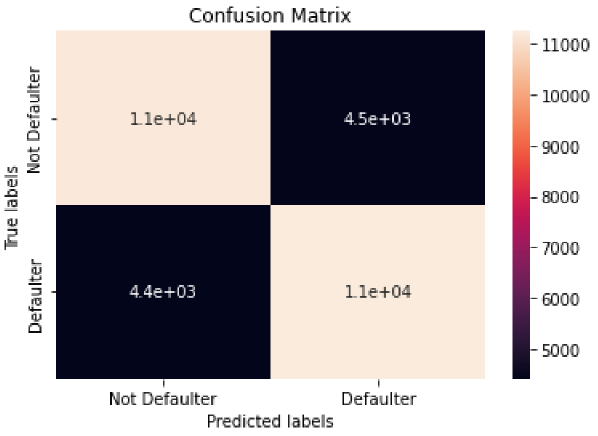
```
param_grid = {'penalty':['l1','l2'],
              'C' : [0.001, 0.01, 0.1, 1, 10, 100, 1000] }
```



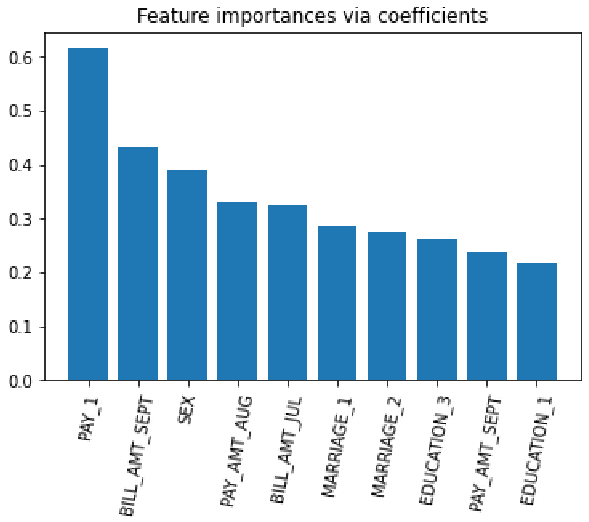

```
The accuracy on train data is  0.7169003737183377
The accuracy on test data is  0.7177225860839116
The accuracy on test data is  0.7177225860839116
The precision on test data is  0.7247730220492866
The recall on test data is  0.7146693950633073
The f1 on test data is  0.719685749243351
The roc_score on test data is  0.7177661629354947
```

Logistic Regression (continue)

Confusion matrix




```
[[11199  4454]
 [ 4409 11245]]
```

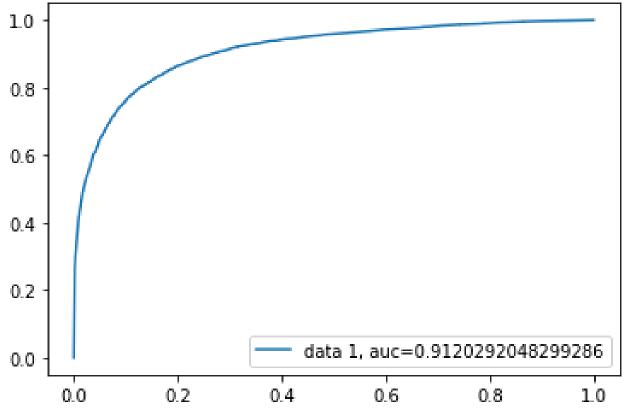


Random Forest Classifier

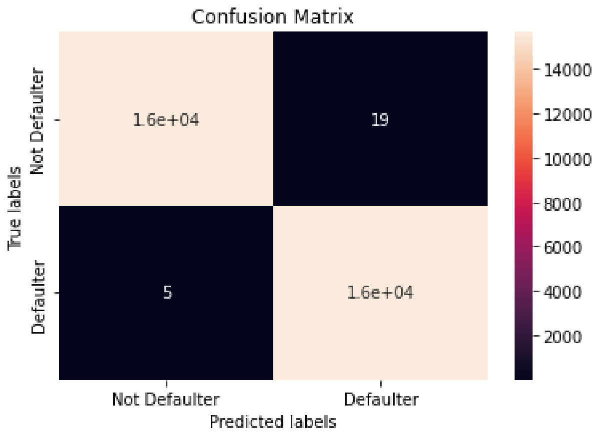
Hyper-Parameter Tuning

```
[ ]  #set the parmeter
     param_grid = {'n_estimators': [150,200], 'max_depth': [20,30]}
```

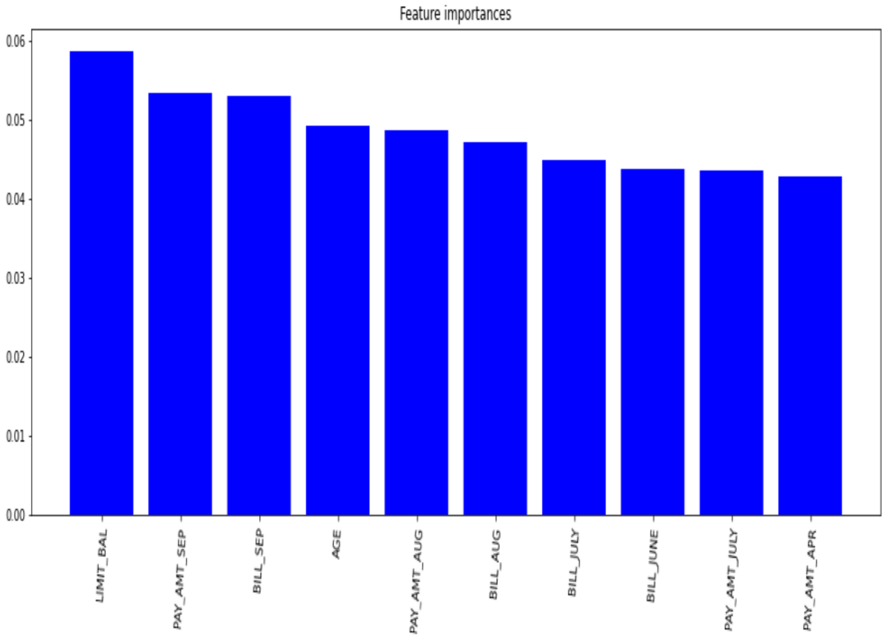
```
The accuracy on test data is  0.8358083133389533
The precision on test data is  0.8238651102464332
The recall on test data is  0.8440074408716449
The f1 on test data is  0.8338146495143082
The roc_score on test data is  0.8359999205624848
```



Random Forest Classifier(continue)




```
[[15634    19]
 [    5 15649]]
```



XGBoost Classifier

Test dataset result

```
The accuracy on test data is  0.8370403994552883
The precision on test data is  0.8189364461738002
The recall on test data is  0.8496837572332122
The f1 on test data is  0.8340268146093389
The roc_score on train data is  0.8374826796178576
```

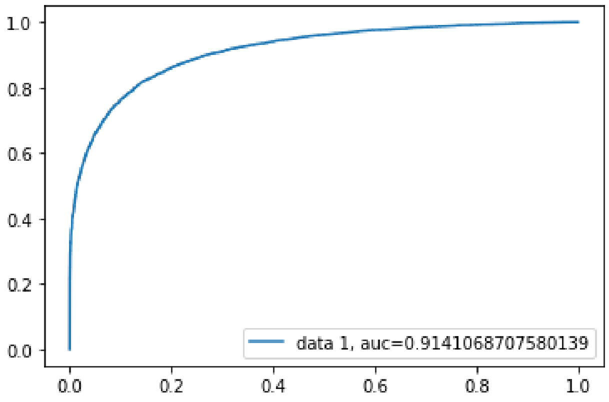
Hyper parameter

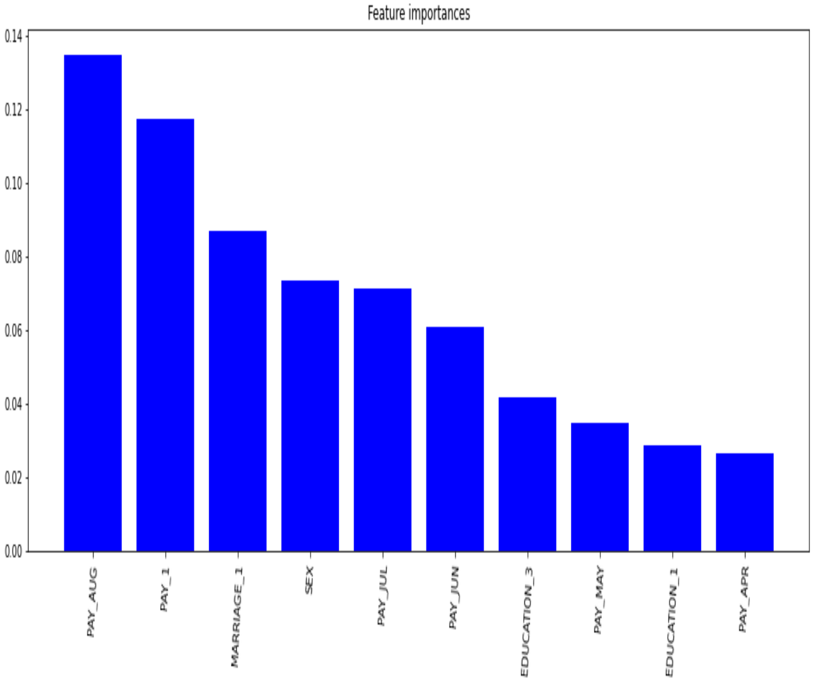
HyperParameter tuning

```
[ ] param_test1 = {
     'max_depth':range(3,10,2),
     'min_child_weight':range(1,6,2)}
    gsearch1 = GridSearchCV(estimator = XGBClassifier( learning_rate =0.1, n_estimators=140, max_depth=5,
     min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8,
     objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27),
     param_grid = param_test1, scoring='accuracy',n_jobs=-1, cv=3, verbose = 2)
    gsearch1.fit(X_train, y_train)
```

XGBoost Classifier(continue)

→AUC Curve



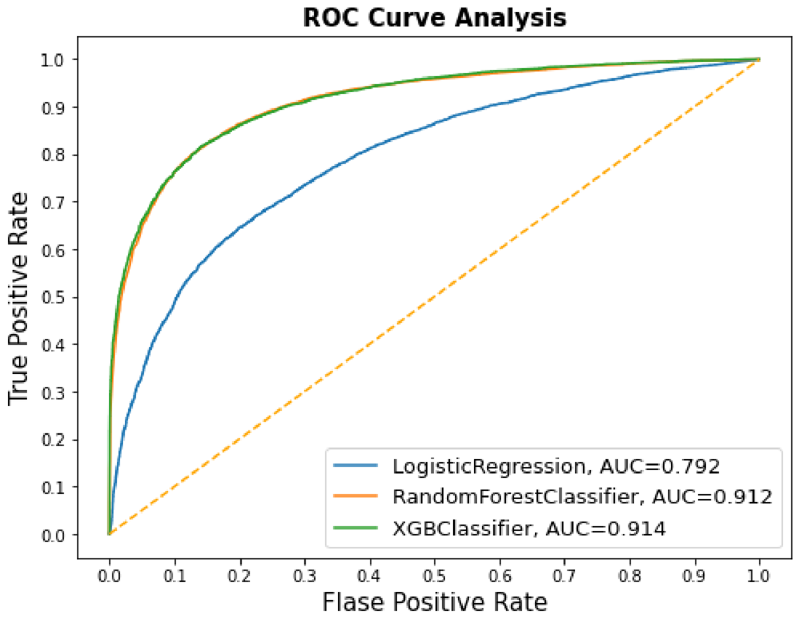


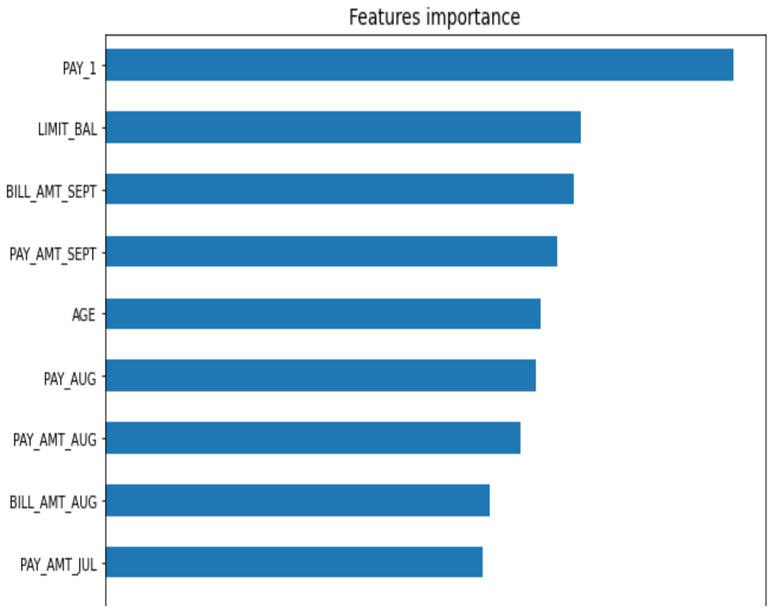
Summary

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.716900	0.717723	0.724773	0.714669	0.719686
1	Random Forest	0.999233	0.835808	0.823865	0.844007	0.833815
2	Xgboost	0.947679	0.837040	0.818936	0.849684	0.834027

- we can conclude from here that **XGboost** is the best model as it gives recall score of ~85%

AUC Curve For All models & Key Features





Challenges

- Large Dataset to handle
- Need to analyze lot of variable
- Feature engineering
- Feature selection
- Optimizing the model
- Deciding the flow of the presentation

Conclusion

- Labels of the data were imbalanced and had a significant difference.
- There were not huge gap but male clients tended to default the most.
- Labels of the data were imbalanced and had a significant difference.
- The data indicates customers with lower education levels default low%. Customers with high school and university educational level had higher default percentages than customers with grad school education
- Gradient boost gave the highest accuracy of 83% on test dataset and best recall score of ~85%.
- Repayment in the month of September (i.e. pay_1 column) tended to be the most important feature for our machine learning model.
- The best accuracy is obtained by XGBoost classifier
- **XGBoost Classifier** having Recall, F1-score, and ROC Score values equals ~85%, 83%, and 83%

Q&A

Thank you