# Capstone Project-2

## Credit card Default Prediction

Submitted by


**Aadarsh Pandey**

**Ankita Hanamshet**

**Darpan Agrawal**

**Vandana Pattnaik**

**Vinay Kulkarni**

Data Science Trainees, Almabetter

# AGENDA


Problem Overview
Dataset Overview
Data Preparation
Proposed Models
Evaluation
Conclusion
Credit Card Default Prediction

# Overview

- This project is aimed at predicting the case of customers default payments in Taiwan

- Given is the dataset wherein 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

- Given are different parameters such as Credit limit, payments done, bill amount etc. to determine the actual probability by building a comprehensive model with the best approach possible

# Goal

- The model we built here will use all possible factors to predict data on customers to find who are  defaulters and non-defaulters next month.
- The goal is to find the whether the clients are able to pay their next month credit amount.
- Identify some potential customers for the financial institution who can settle their credit balance.
- To determine if their customers could make the credit card payments on-time.
- **Default** is the failure to **pay** interest or principal on a loan or credit card payment.

# Approach Design

Preparation of data

Applying models

Comparing the applied models

Evaluation Model

Analyse the best model


Problem Overview
Conclusion
Dataset Overview
Credit Card Default prediction
Evaluation
Feature Engineering
Proposed Models

# Dataset Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 1 to 30000
Data columns (total 24 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   LIMIT_BAL                   30000 non-null  object
 1   SEX                         30000 non-null  object
 2   EDUCATION                   30000 non-null  object
 3   MARRIAGE                    30000 non-null  object
 4   AGE                         30000 non-null  object
 5   PAY_0                       30000 non-null  object
 6   PAY_2                       30000 non-null  object
 7   PAY_3                       30000 non-null  object
 8   PAY_4                       30000 non-null  object
 9   PAY_5                       30000 non-null  object
 10  PAY_6                       30000 non-null  object
 11  BILL_AMT1                   30000 non-null  object
 12  BILL_AMT2                   30000 non-null  object
 13  BILL_AMT3                   30000 non-null  object
 14  BILL_AMT4                   30000 non-null  object
 15  BILL_AMT5                   30000 non-null  object
 16  BILL_AMT6                   30000 non-null  object
 17  PAY_AMT1                    30000 non-null  object
 18  PAY_AMT2                    30000 non-null  object
 19  PAY_AMT3                    30000 non-null  object
 20  PAY_AMT4                    30000 non-null  object
 21  PAY_AMT5                    30000 non-null  object
 22  PAY_AMT6                    30000 non-null  object
 23  default payment next month  30000 non-null  object
dtypes: object(24)
memory usage: 5.5+ MB
```

Dataset Description:
(30000, 24)

No null value count

# Continue

…

Independent variables:

- Customer ID
- Credit limit
- Gender
- Age
- Marital status
- Level of education
- History of their past payments made (April to September) (X6 to X11)
- Amount of bill statement (X12 to X17)
- Amount of previous payment (X18 to X23)

Dependent variables:

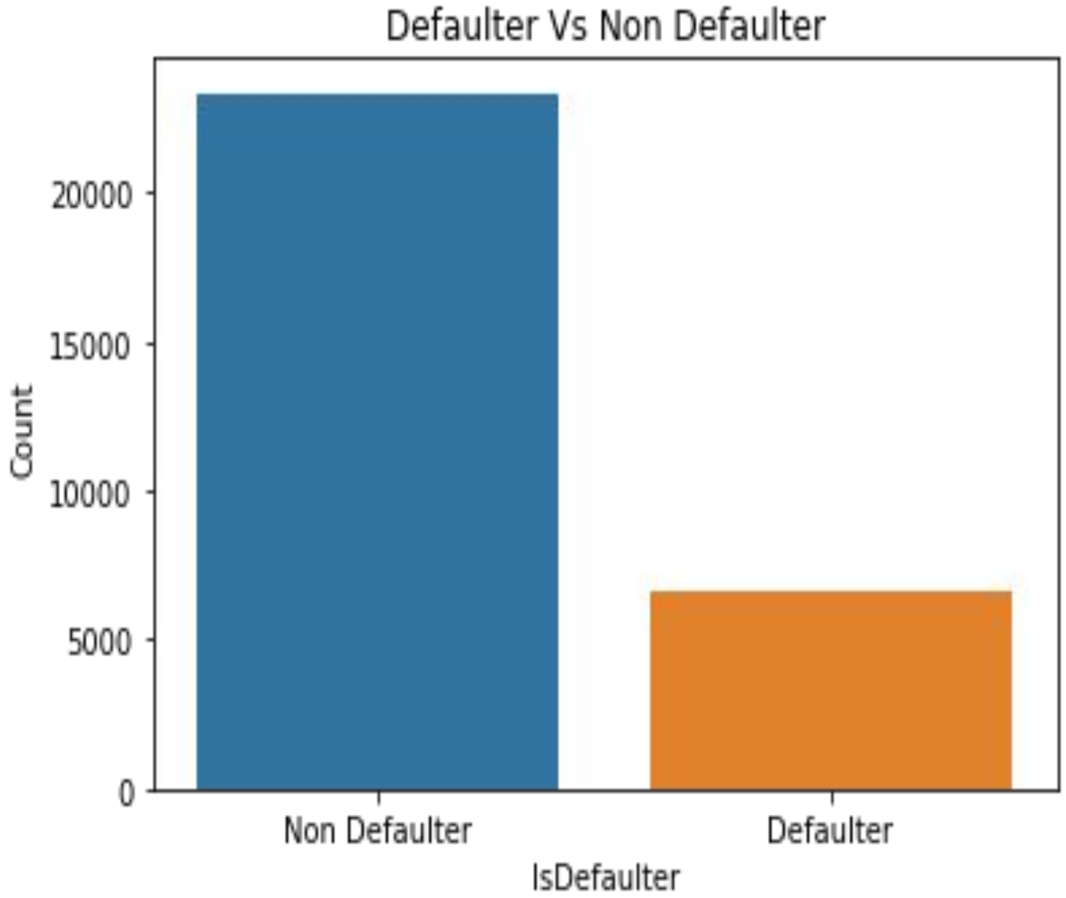
- default – A customer who will be default next month payment (0: no, 1: yes)

# Dataset overview

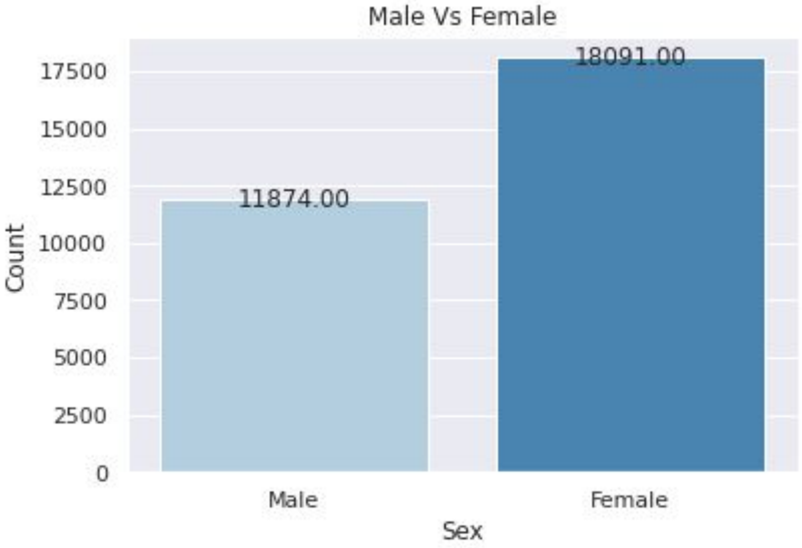
Graph shows total number of records for defaulters and non-defaulters.

If they would do payment or not (yes=1 no=0) for next month

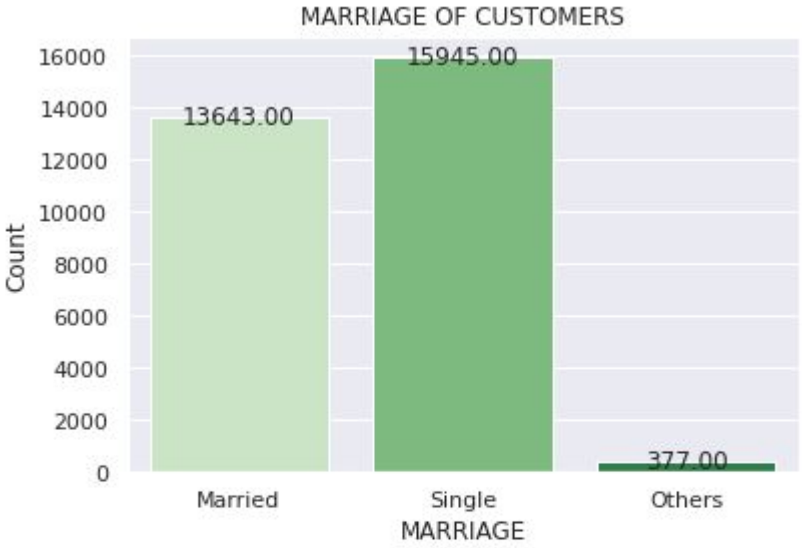
22% - default
78% - non-default



# Continue
…



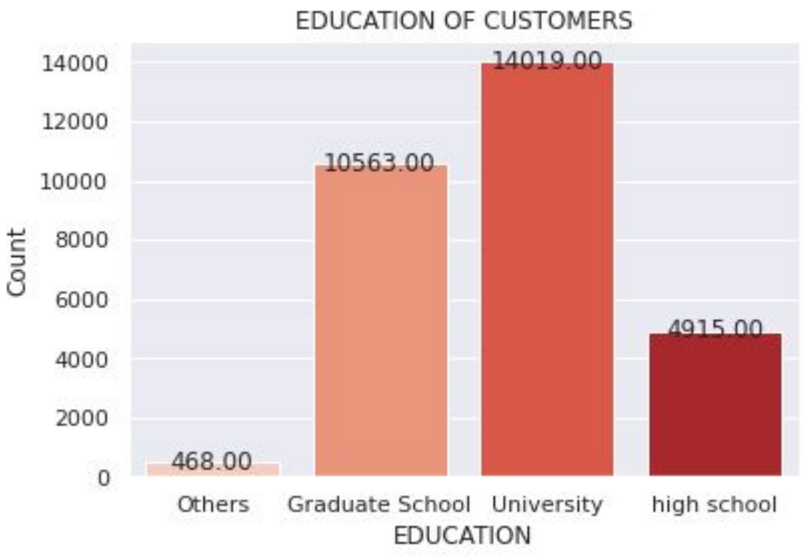

1. It shows count for ‘sex’ attribute

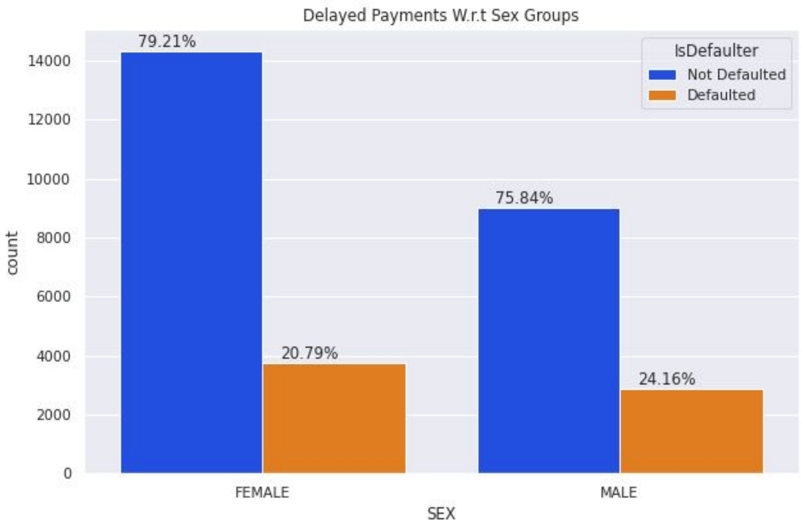



2. It shows default count for ‘marriage’ attribute

# Continue

...



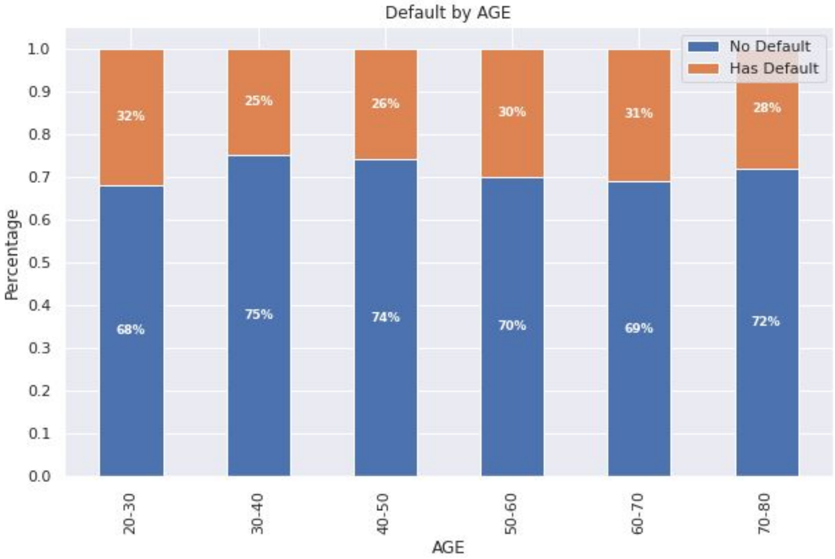




1. It shows count for Education attribute values with respect to credit card count

2. It shows Delayed Payment % wrt Sex.

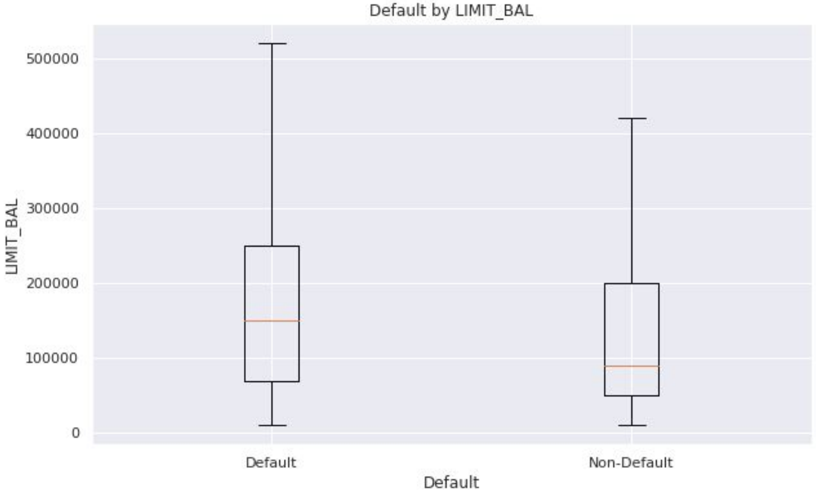
# Continue

...




1. It shows %count for Default vs Age

People aged 30-50, and >70 have least default rates




2. It shows Default vs Limit balance

Customers with high credit limit tends to have higher default rate


Problem Overview
Dataset Overview
Conclusion
Credit Card Default Prediction
Evaluation
Feature Engineering
Proposed Models

# Feature Engineering

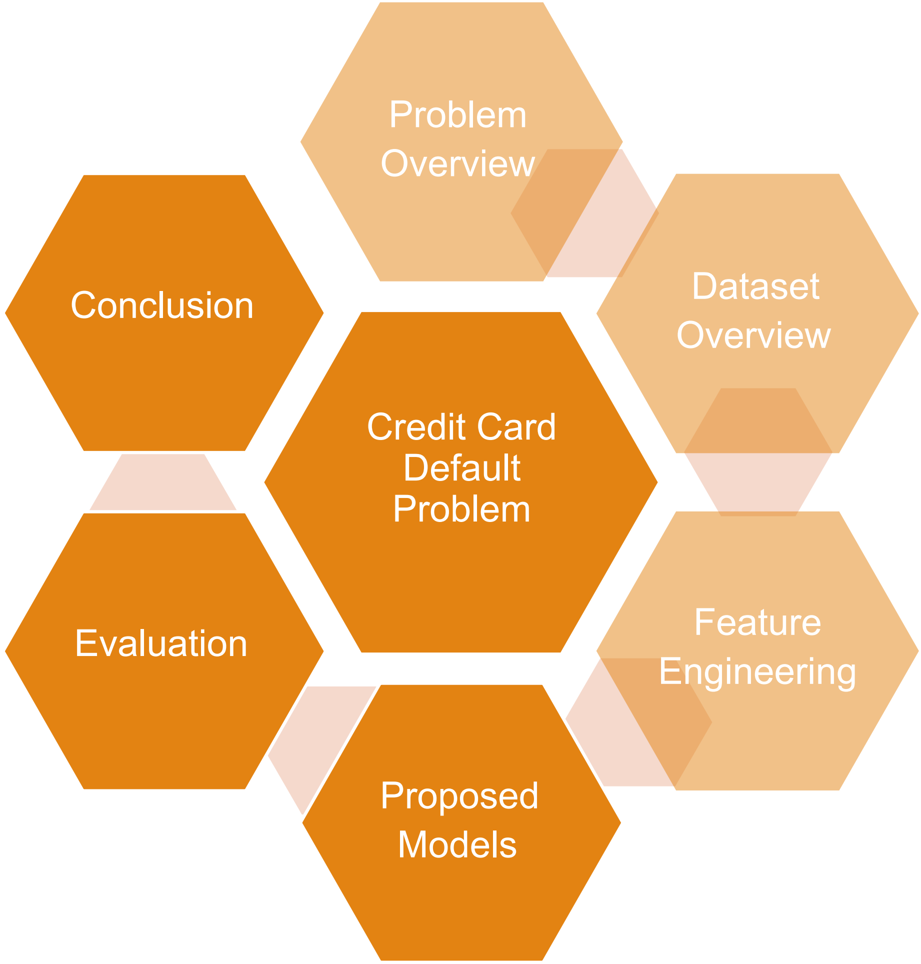
- In this Feature Engineering we have made a separate column for balance of defaulters and got that we have got the result that there are 46670 defaulters.

- We have also done the Label Encoding on the gender section which says Female as 0 and Male as 1

- We have also done One hot encoding on Education and marriage and also divided the age column into 7 groups (21, 30, 40, 50, 60, 70, 80).

- We have separated the Independent Feature and dependent Feature and made a separate columns for Independent Feature.

- Also rescaled the Independent Feature using StandardScaler.


Problem Overview
Dataset Overview
Feature Engineering
Proposed Models
Evaluation
Conclusion
Credit Card Default Problem

# Proposed Models

## Logistic Regression

- It is used for Binary classification.
- Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid over fitting.
- In logistic regression the hypothesis is that the conditional probability p of class belongs to ”1”
- if probability is greater than threshold probability, generally 0.5, else it belongs to the class ”0”.

$$\text{Ex. } Y(i) = \begin{cases} 1, p \geq 0.5 \\ 0, p < 0.5 \end{cases}$$

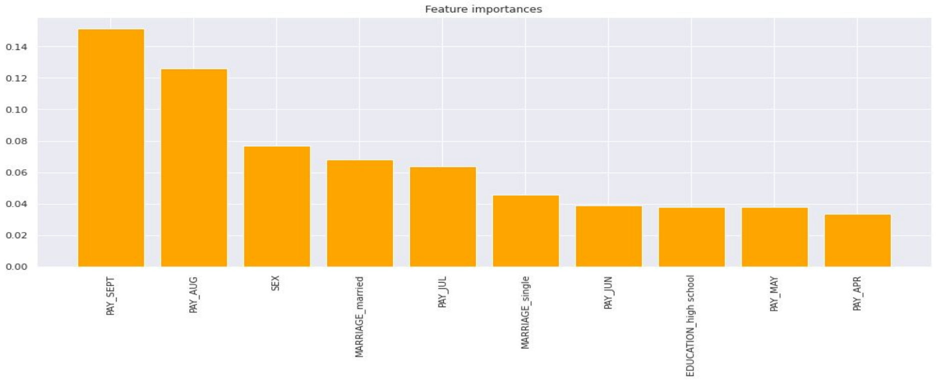
## Random Forest Classifier

- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset
- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result
- The predictions from each tree must have very low correlations

# Continue

## XGBoost

- It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems
- It's vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon supervised machine learning, decision trees, ensemble learning, and gradient boosting.



# Evaluation Process

Evaluation Metrics:

- **Accuracy:** Accuracy determine how often the model predicts default and non-default correctly.
- **Precision:** Precision calculates whenever our models predicts it is default how often it is correct.
- **Recall:** Recall regulate the actual default that the model is actually predict.
- **Precision Recall Curve:** PRC will display the tradeoff between precision and recall threshold.

# Confusion Matrix

**True Positive – A person who is defaulter and predicted as defaulter.**
True Negative – A person who is non-defaulter and predicted as non-defaulter.  False Positive – A person who is predicted defaulter is non-defaulter.

| # | Non-defaulter (predicted) - 0 | Defaulter (predicted) - 1 |
|---|---|---|
| **Non-defaulter (actual) - 0** | TN | FP |
| **Defaulter (actual) - 1** | **FN** | TP |

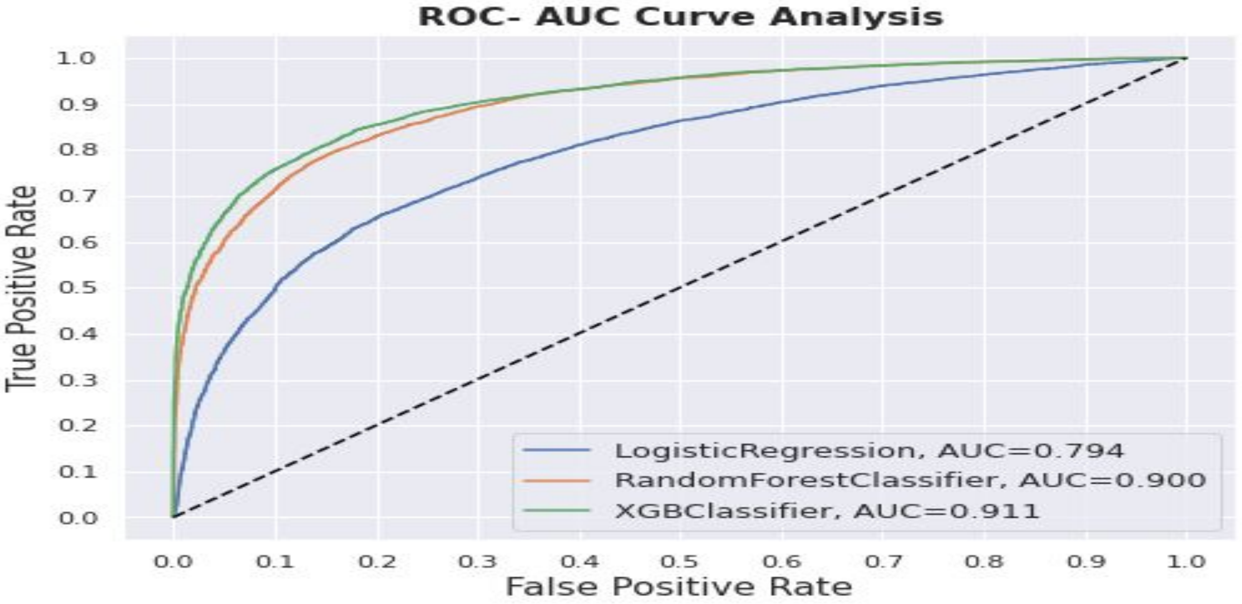

Problem Overview
Dataset Overview
Conclusion
Credit Card Default Prediction
Evaluation
Feature Engineering
Proposed Models

# Evaluation Result

| No. | Algorithms | Train/Test Accuracy(%) | Precision(%) | Recall(%) | Confusion Metrix |
|---|---|---|---|---|---|
| 1 | Logistic Regression | 72.00/72.09 | 72.04 | 72.11 | $\begin{bmatrix} 5100 & 2000 \\ 2000 & 5000 \end{bmatrix}$ |
| 2 | Random Forest | 95.23/81.88 | 79.72 | 83.31 | $\begin{bmatrix} 5900 & 1100 \\ 1400 & 5600 \end{bmatrix}$ |
| 3 | XGBoost | 94.61/83.02 | 80.98 | 84.42 | $\begin{bmatrix} 6000 & 1000 \\ 1300 & 5700 \end{bmatrix}$ |

# ROC-AUC Curve

ROC-AUC curve analysis for the Models




Credit Card Default Prediction
Problem Overview
Dataset Overview
Feature Engineering
Proposed Models
Evaluation Process
Conclusion

# Conclusion

- We investigated the data, checking for data unbalancing, visualizing the features and understanding the relationship between different features.

- We used both train-validation split and cross-validation to evaluate the model effectiveness to predict the target value, i.e. detecting if a credit card client will default next month.

- We then investigated three predictive models:
    - We started with Logistic Regression, Random Forest and XG Boost. Among them random forest and XGBoost classifier accuracy is almost same.
    - We choose based model based on **minimum value of False Negative value** i.e. the XG Boost
    - This would also inform the issuer's decisions on who to **give a credit card** to and what **credit limit** to provide.

# THANK YOU