Capstone Project-3

Credit Card Default Prediction

(Supervised Machine Learning-Classification)

Batch- Cohort Seattle

Presented By:

Vaitul Sidhdhapara
Drashti Shah

CONTENTS:

- Introduction
- Problem statement
- Data Summary
- Exploratory Data Analysis
- Model Building
- Evaluation
- Limitations
- Conclusion

INTRODUCTION:

- Credit risk plays a major role in the banking industry business. Bank’s main activities involve granting loan, credit card, investment, mortgage, and others.
- Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate.
- As such data analytics can provide solutions to tackle the current phenomenon and management credit risks. This project discusses the implementation of a model which predicts if a given credit card holder has a probability of defaulting in the following month, using their demographic data and behavioral data from the past 6 months.

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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

Objective

- Development of a model for predicting if a given customer id has a probability to default in the following month or not.
- Benefits:
  - Detection of upcoming frauds.
  - Gives better insight of customer base.
  - Allows financial institutions to take necessary steps to minimize the lose from the possible defaults.

Data Description

There are 25 variables.

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- **Gender:**
  - 1=male
  - 2=female
- **EDUCATION:**
  - 1=graduate school
  - 2=university
  - 3=high school
  - 0, 4, 5, 6=others
- **AGE**: Age in years

Data Description

- **MARRIAGE**: Marital status
  - 1=married,
  - 2=single,
  - 3=divorce,
  - 0=others
- **PAY_0**: Repayment status in September, 2005
  - -2: No consumption;
  - -1: Paid in full;
  - 0: The use of revolving credit;
  - 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 Description

- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)

Data Description

- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default.payment.next.month**: Default payment
  - 0=no
  - 1=yes

Data Acquisition

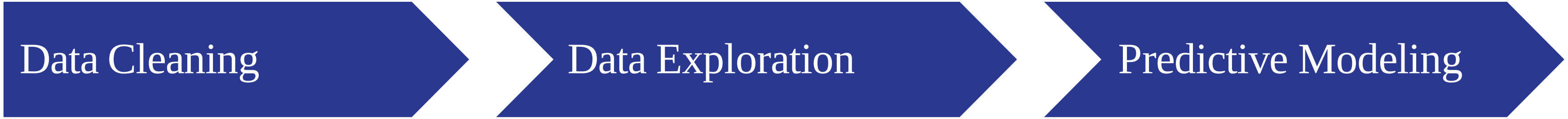
Dataset

- Default Payments of Credit Card Clients in Taiwan from 2005

Why This Dataset?

- Real credit card data
- Comprehensive and complete
- 30,000 customers
- Usage of 6 months
- Age from 20-79
- Demographic factors
- No credit score or credit history

Approach Overview



Data Cleaning

Data Exploration

Predictive Modeling

Understand and Clean

- Find information on undocumented columns values.
- Clean data to get it ready for analysis.

Graphical and Statistical

- Exam data with visualization.
- Verify findings with statistical tests.

Machine Learning

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XG Boost Classifier
- K Neighbors Classifier
- Support Vector Classifier
- Naive Bayes Classifier

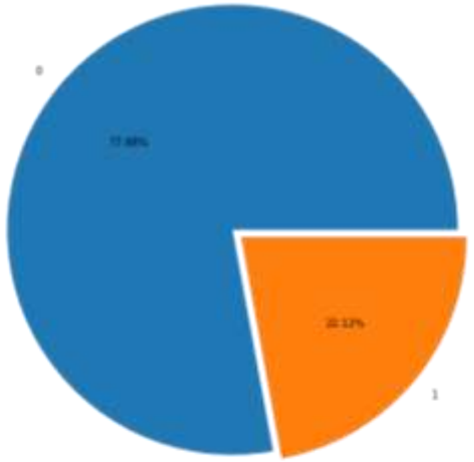
Part-1

Exploratory Data Analysis

What demographic factors impact payment default risk?

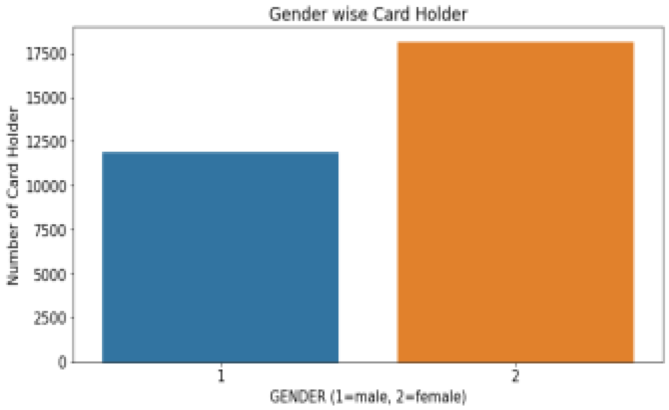
EDA

- ❖ How much Credit Card defaults ?



- ➢ 77.88% is non default while 22.12% are default .

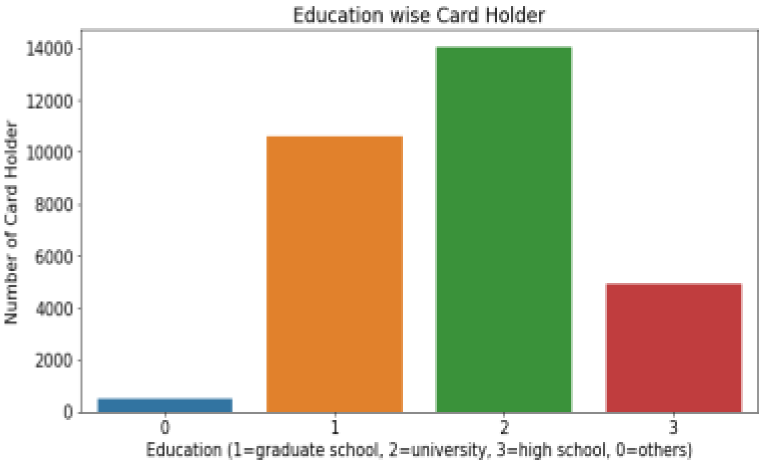
- ❖ Gender wise Credit Card Holders.



- ➢ Females have more number of card compare to Males.

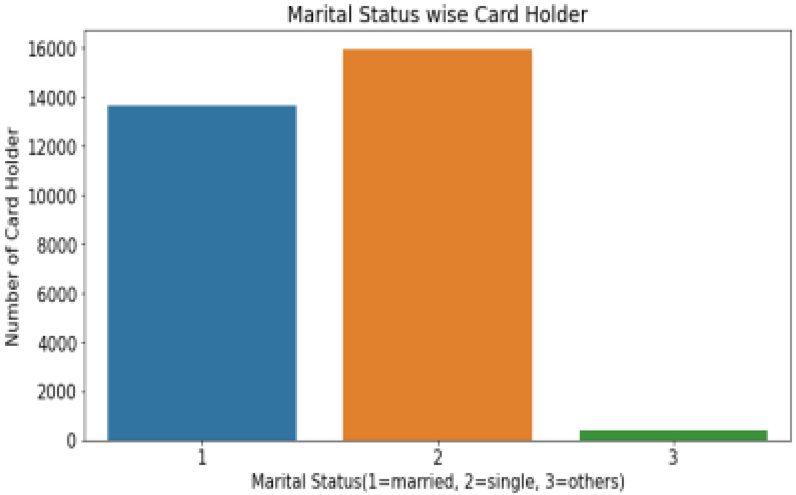
EDA

- Education wise Credit Card Holders.



- More number of Credit Card holders are University students followed by Graduates and then High school students.

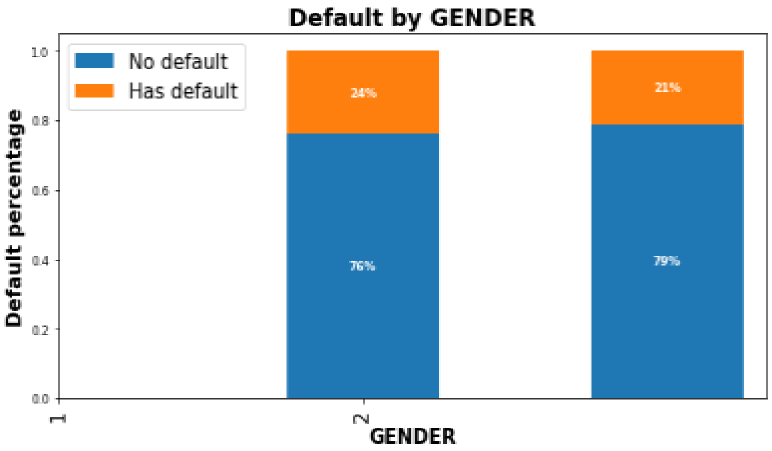
- Marital Status wise Credit Card Holders.



- More number of Credit Cards holders are Married.

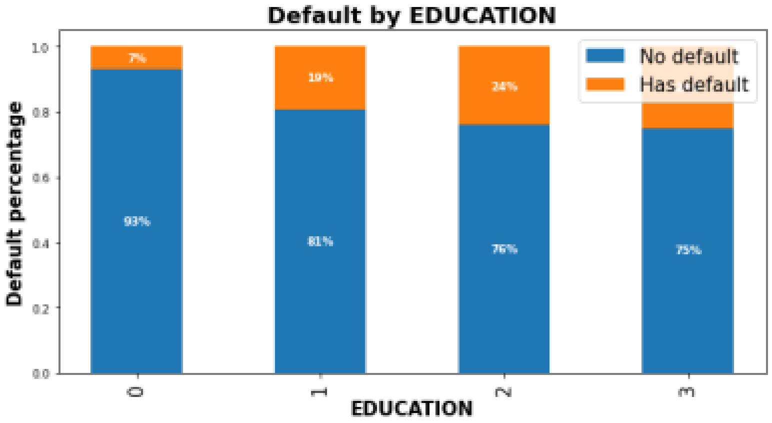
EDA

- On average, which gender group tends to have more default payments?



- 24% male have default payment while 21% female have default payment, the difference is not significant. (2-Female, 1-Male)

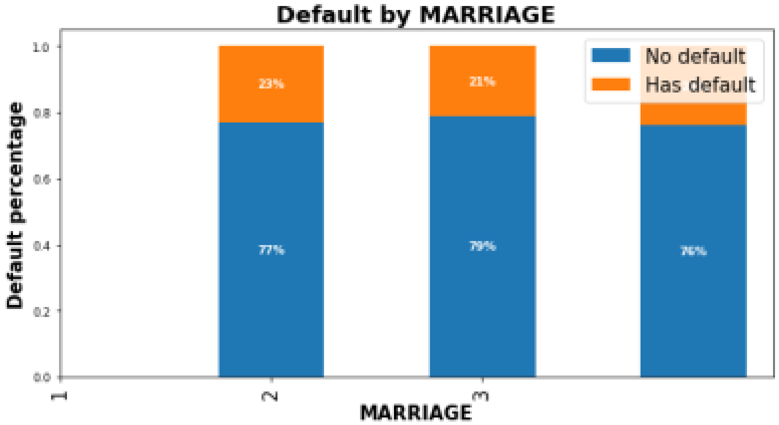
- Did customers with higher education have less default payment?



- The data indicates customers with lower education levels default more.

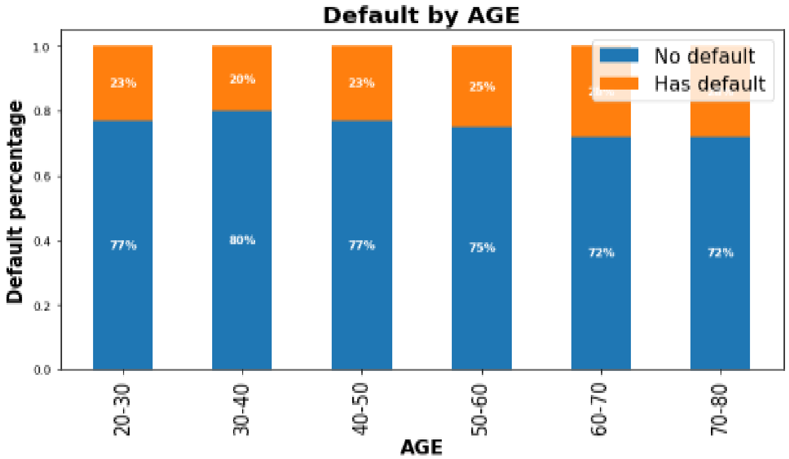
EDA

- Does marital status have anything to do with default risk? Note the credit limit includes the family's total credit.



- There is no difference of default risk in terms of marital status, although the 'other' marital status group has high default percentage.

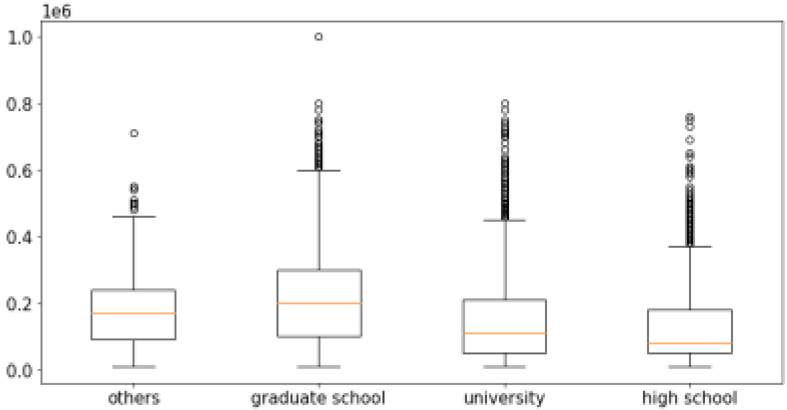
- Do younger people tend to miss the payment deadline?



- Customers aged between 30-40 had the lowest default payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates.

EDA

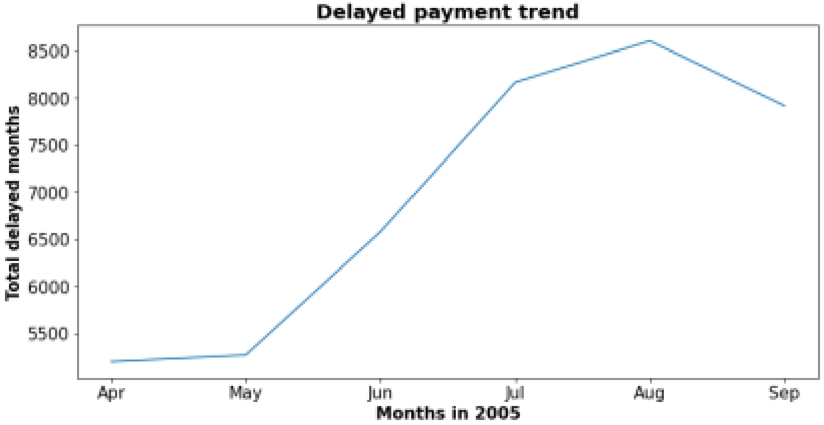
❖ Did customers with a high education level get higher credit limits?



➢ From the boxplot, we can see that customers with graduate school education have the highest 25% percentile, highest, median, highest 75th percentile and highest maximum numbers, which proves that customers with higher education levels did get higher credit limits.

EDA

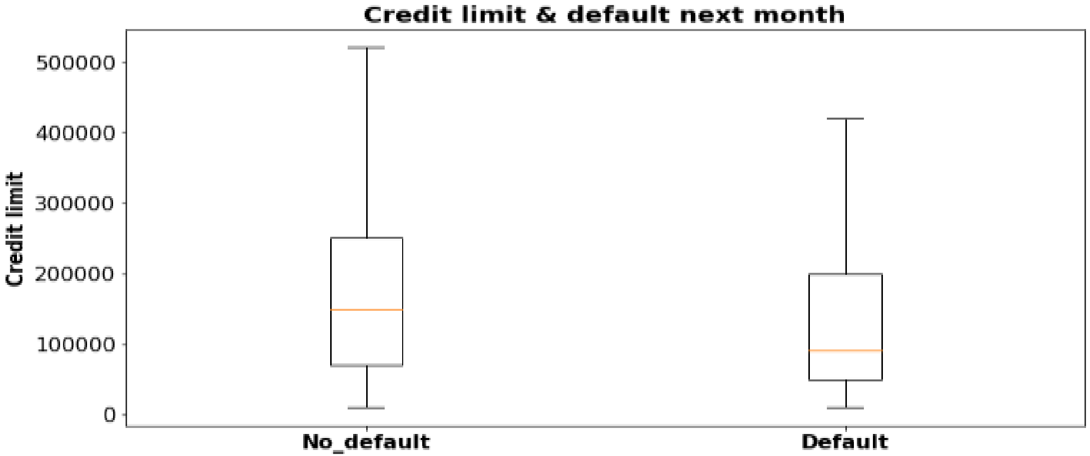
- Has the repayment status changed in the 6 month from April 2005 (PAY_6) to September 2005(PAY_0)?



- There was a huge jump from May,2005 (PAY_5) to July, 2005 (PAY_3) when delayed payment increased significantly, then it peaked at August, 2005 (PAY_2), things started to get better in September, 2005 (PAY_1).

EDA

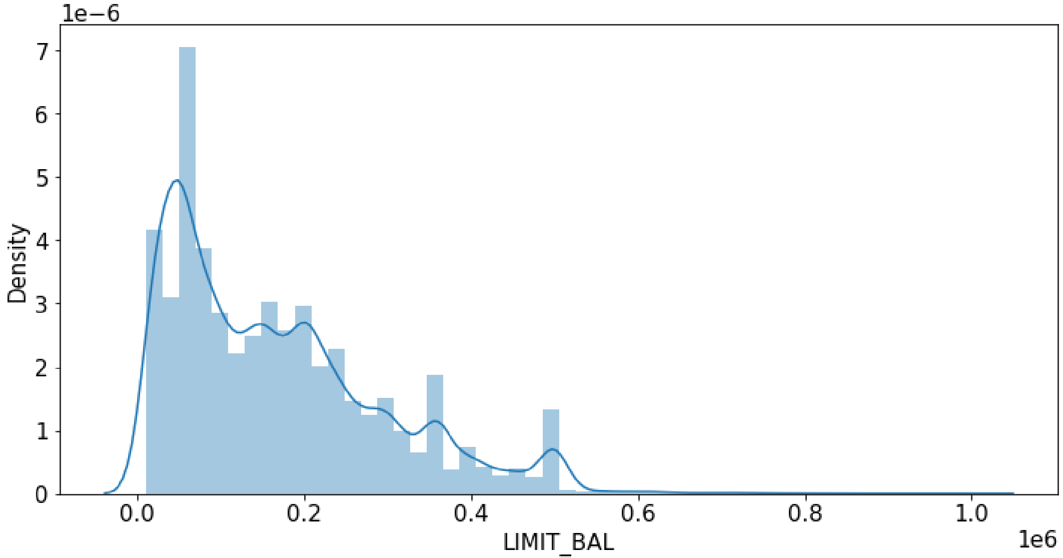
- ❖ Is there any correlation between credit limit and the default payment next month?



- ➢ Unsurprisingly, customers who had higher credit limits had lower delayed (default) payment rates.

EDA

❖ Checking the Data Distribution of Balance Limit.



➢ Here **"LIMIT_BAL"** has positively skewed distribution.

CONTENTS:

- Demographic factors that impact default risk are:
    - Education: Higher education is associated with lower default risk.
    - Age: Customers aged 30-50 have the lowest default risk.
    - Gender: Females have lower default risk than males in this dataset.
    - Credit limit: Higher credit limit is associated with lower default risk.

Part-2

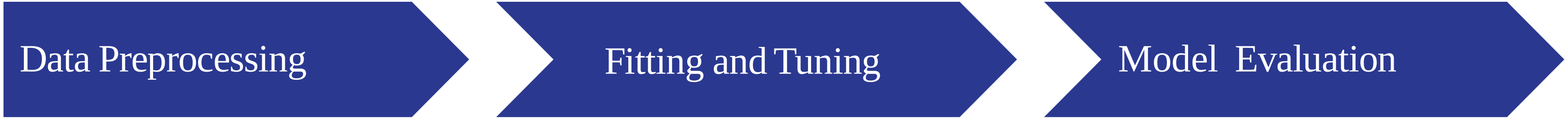
Predictive Modeling

Find the Best Classification Predictive Model.

Modeling Overview

Define Problem:	Supervised learning, Binary classification
Imbalanced Classes:	78% non-default vs. 22% default
Tools Used:	Scikit learn library and imblearn
Models Applied:	Different types of Classification Models

Modeling Steps



Data Preprocessing

- Feature selection
- Rescale Features
- Feature engineering
- Check Class Imbalance
- Train-test data splitting (70%/30%)
- Training data rescaling
- SMOTE oversampling

Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure ROC_AUC on training data

Model Evaluation

- Compare within the 8 models

Working on these 8 Models

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XG Boost Classifier
- K Neighbors Classifier
- Support Vector Classifier
- Naive Bayes Classifier

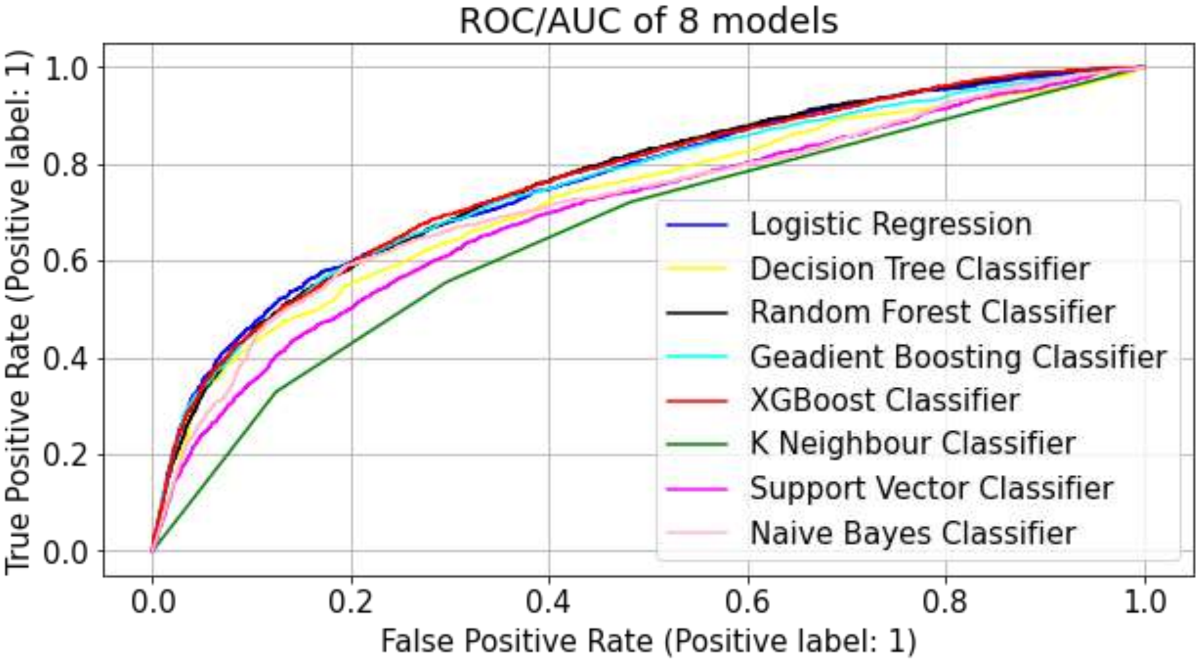
Part-3

Evaluation

- ✓ Evaluation
- ✓ Limitations
- ✓ Conclusion

Evaluation

ROC_AUC Curve

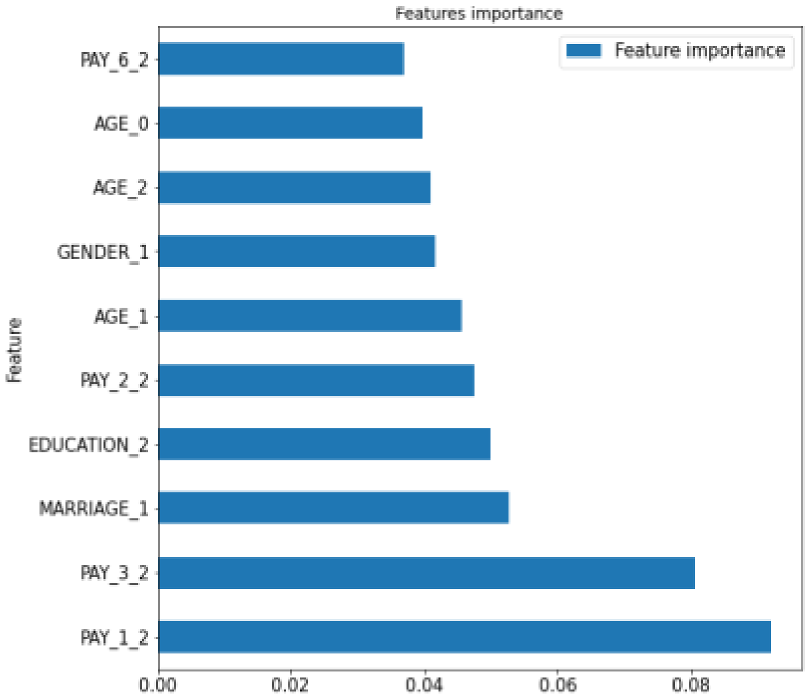


- This plot shows ROC-AUC curve for whole 8 model.

- XGBoost classifier algorithm shows best performance compare to other.

Evaluation

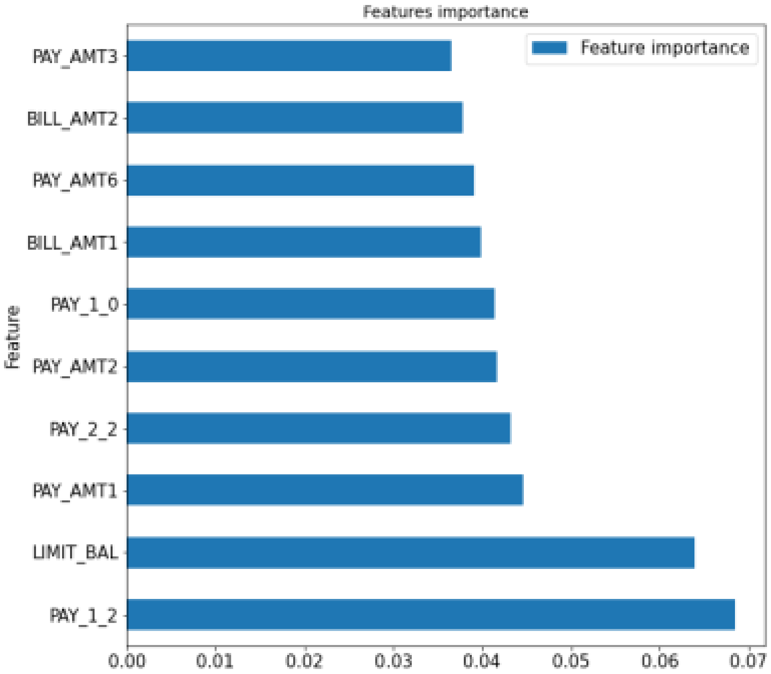
Feature Importance Graph for XGBoost Model



- ➢ This graph shows the feature importance of XGBoost model.
- ➢ In this model most important feature is "PAY_1".

Evaluation

Feature Importance Graph for Random Classifier Model



- This graph shows the feature importance of Random Forest Classifier model.

- In this model most important features are "PAY_1" and "LIMIT_BAL".

Limitations

- Best model Random Forest and XGBoost Classifier model can only detect 50%-52% of default.
- Model can only be served as an aid in decision making instead of replacing human decision.
- Here 30,000 record is not sufficient for better prediction of our model.

Conclusion

- After observing Precision, Recall, ROC-AUC curve and Accuracy score i would recommend XGBoost and Random Forest Classifier Model.
- The balance of recall and precision is the most important metric, then XGBoost and Random Forest Classifier Model are the ideal model.
- The strongest predictors of default are the PAY_X (ie the repayment status in previous months), the LIMIT_BAL & the PAY_AMTX (amount paid in previous months).
- We see that being Female, More educated, Single and between 30-40years old means a customer is more likely to make payments on time.
- Best accuracy score:

  1) Random Forest Classifier: (a) Test Data= 94% (b) Train Data= 80%

  2) XGBoost Classifier : (a) Test Data= 81% (b) Train Data= 80%

Thank You!