CREDIT CARD DEFAULT PREDICTION

Vaitul Sidhdhapara & Drashti Shah

Data Science Trainees,

AlmaBetter, Bangalore

Abstract

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. Thus, assessing, detecting and managing default risk is the key factor in generating revenue and reducing loss for the banking and credit card industry. Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting.

The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history. According to a literature review report on analysing credit risk using machine and deep learning models, "credit risk management problems researched have been around credit scoring; it would go a long way to research how machine learning can be applied to quantitative areas for better computations of credit risk exposure by predicting probabilities of default."

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision-making process.

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. This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information.

Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns. More details about the data cleaning can be found in this Jupyter Notebook.

We have worked on these models to get good results:

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

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.

Data Description

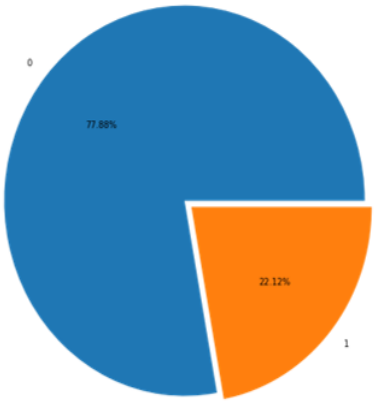
There are 25 variables:

- **ID:** ID of each client
- **LIMIT_BAL:** Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- **GENDER:** Gender
  - 1 = male,
  - 2 = female
- **EDUCATION:**
  - 1=graduate school
  - 2=university
  - 3=high school
  - 4=others
  - 5=unknown
  - 6=unknown
- **MARRIAGE:** Marital status
  - 1=married
  - 2=single
  - 3=others
- **AGE:** Age in years
- **PAY_0:** Repayment status in September, 2005
  - -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
- **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)
- **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
  - 1 = Yes
  - 0 = No

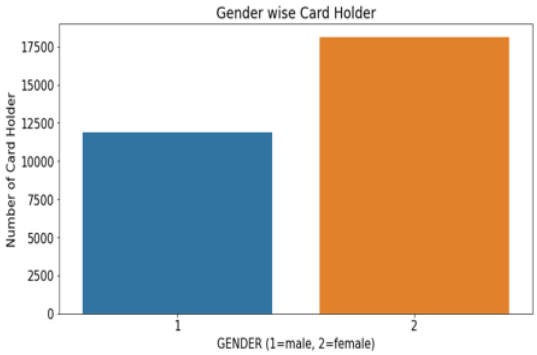
EDA (Exploratory Data Analysis)

❖ How much Credit Card defaults ?



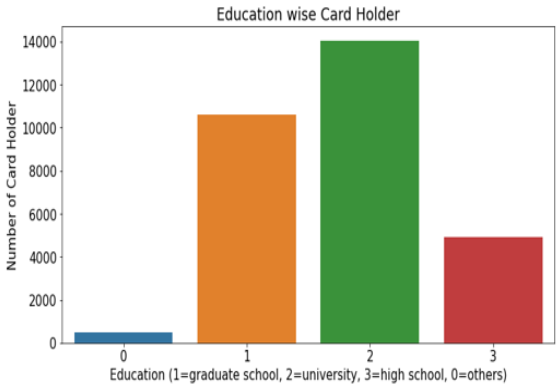
➢ 77.88% is non default while 22.12% are default .

❖ Gender wise Credit Card Holders.



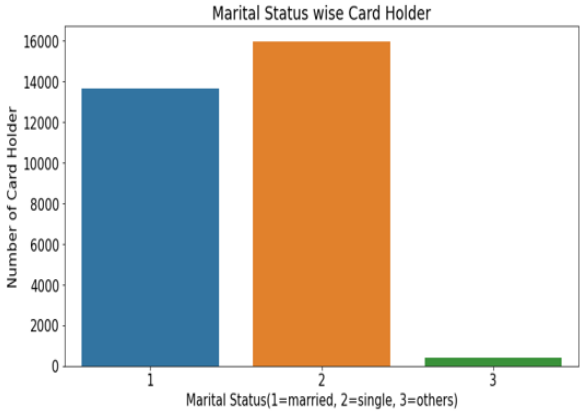
➢ Females have more number of card compare to Males.

❖ Education wise Credit Card Holders.



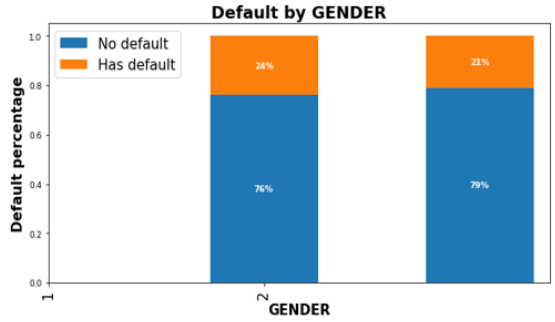
➢ More number of Credit Card holders are University students followed by Graduates and then High school students.

❖ Marital Status wise Credit Card Holders.



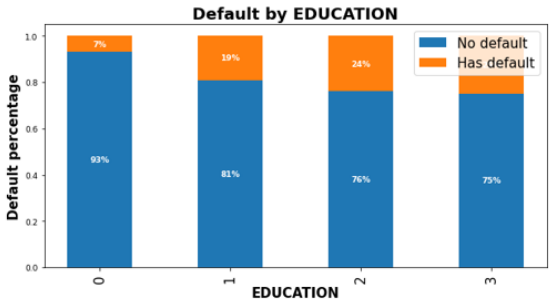
➢ More number of Credit Cards holders are Married.

(5) On average, which gender group tends to have more default payments?



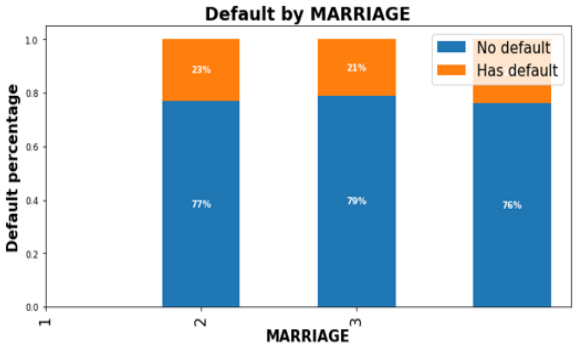
- 24% male have default payment while 21% female have default payment, the difference is not significant. (2-Female, 1-Male)

(6) Did customers with higher education have less default payment?



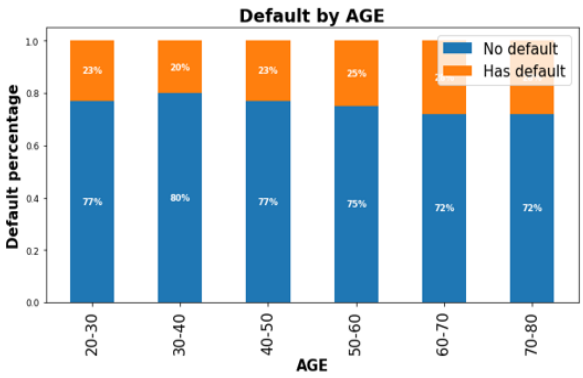
- The data indicates customers with lower education levels default more.

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



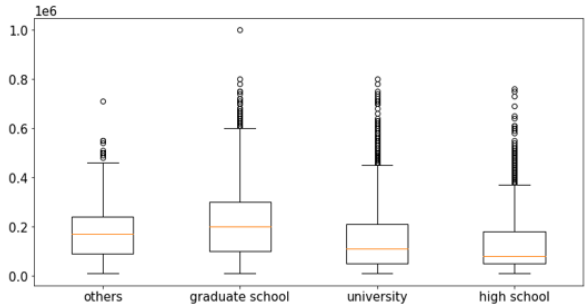
- There is no difference of default risk in terms of marital status, although the 'other' marital status group has high default percentage.

❖ Do younger people tend to miss the payment deadline?



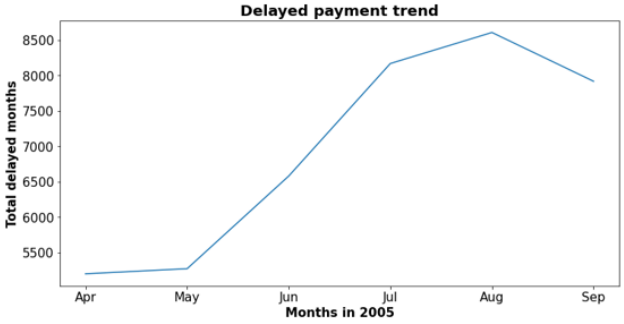
- Customers aged between 30-40 had the lowest default payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates.

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



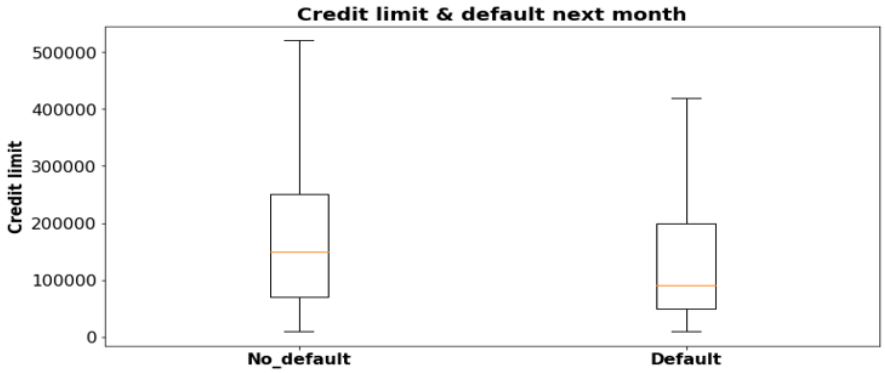
➢ From the boxplot, we can see that customers with graduate school education have the highest 25% percentile, highest, median, highest 75th percentile and highest maximum numbers, which proves that customers with higher education levels did get higher credit limits.

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



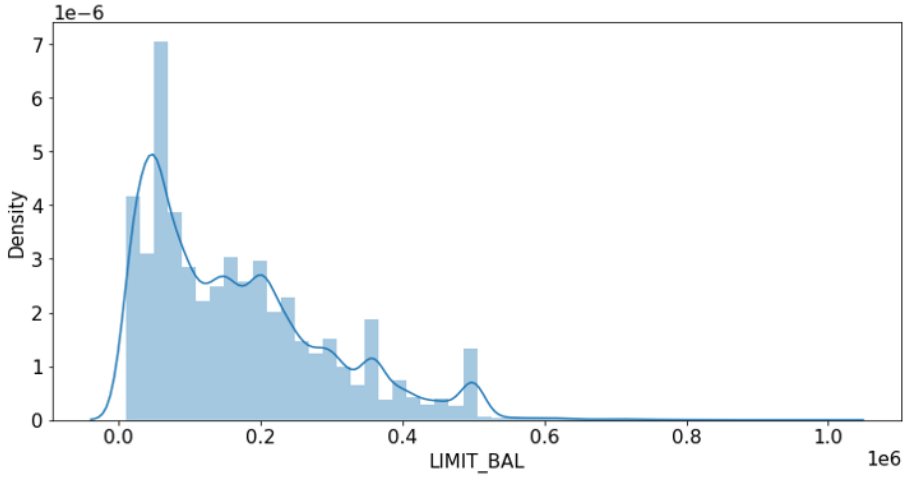
➢ There was a huge jump from May,2005 (PAY_5) to July, 2005 (PAY_3) when delayed payment increased significantly, then it peaked at August, 2005 (PAY_2), things started to get better in September, 2005 (PAY_1).

❖ 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.

❖ Checking the Data Distribution of Balance Limit.



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

Modeling

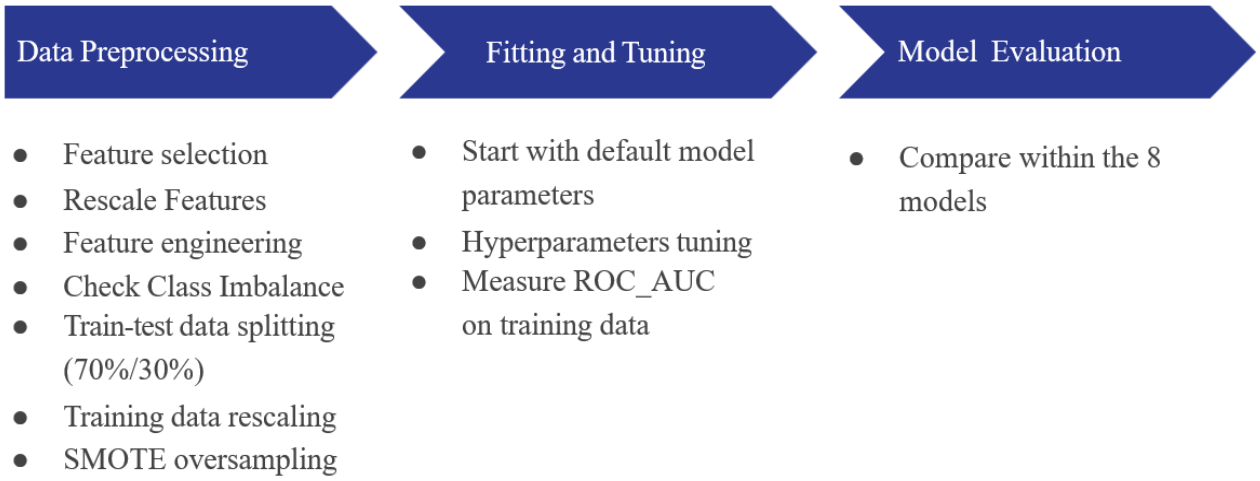
Overview of Model Building

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

We Worked 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



Modeling Preparation

Since there are labeled data and the expected outcome is the probability of customer default, we define this as supervised machine learning and it is a binary classification problem. For better model performance, we first take a few pre-processing steps to prepare for modeling.

- **Feature Selection:** There are 25 columns in this dataset and the target variable is the column 'DEF_PAY_NMO' (means "default next month"). We drop the column 'ID' and 'DEF_PAY_NMO', save the rest 23 as predictor features. Those predictor variables include categorical variables such as sex, age, education level and marital status, along with numerical variables, such as payment status, credit limit, bill amount, etc. With this dataset, we don't need to do PCA or dimensionality reduction. From this dataset, after performing whole model preparation techniques we created a dummies variable from the labeled data and after we get 125 features.

- **Check Class Imbalance:** It is common sense that most customers do not default. This dataset is likely to be dominated by 0s (non-default) with rare 1s (default). Imbalanced dataset will mislead machine learning algorithms and affect their performances. 'DEF_PAY_NMO' variable shows 22% of customers have default and 78% have no default. The class ratio is roughly

1:4. We consider this dataset is imbalanced and will use SMOTE oversampling technique after train-test data split to balance the data.

- **Data Rescaling:** The feature variables' value very vastly. For example, the credit limit value is up to 100,000 NTD and the payment status only ranges from 0 to 8. In order to make all variables have similar ranges, so the Logistic Regression model can perform well in regularization, we rescale the training data. In this process, we make sure to only fit training data (X_train) and then transform training data and test data (X_train, X_test), instead of fit and transform the entire X (consists of X_train and X_test).

- **Transform Categorical Column:** In the dataset, 'AGE' column has continuous values which are the individual customer's age. In the business context, we are more concerned about the age groups than the specific age, so we bin the 'AGE' column to 6 bins - 21~29,30~39,40~49,50~59,60~69, and 70~79. Finally, we convert this column into numerical data type because sklearn does not accept categorical data type.

- **Split Training and Test Data:** For each model, we use the same ratio for training and test data split (70% for training, 30% for test) to ensure consistency. After splitting the data, we set the test data aside and leave it for the very end, which is the final testing after hyperparameter tuning.

Predictive Modeling

This analysis uses 8 classification models

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

We have worked on these eight models. After that we get the result, conclusion, evaluation, performance metrics, feature importance, ROC-AUC curve etc.

- **Performance Metrics:** Since this is a classification problem with imbalanced classes, accuracy is not the best metric because the data is dominated by non-default class, thus precision and recall are a better choice. In the credit card default risk business context, detecting as many defaults as possible is our ultimate goal because misclassifying a default as non-default is costly, therefore a high recall score is the best metric. However, there is a known trade-off between precision and recall. We can raise recall to arbitrarily high, but the precision will decrease.
  We use below metrics to measure model performances.

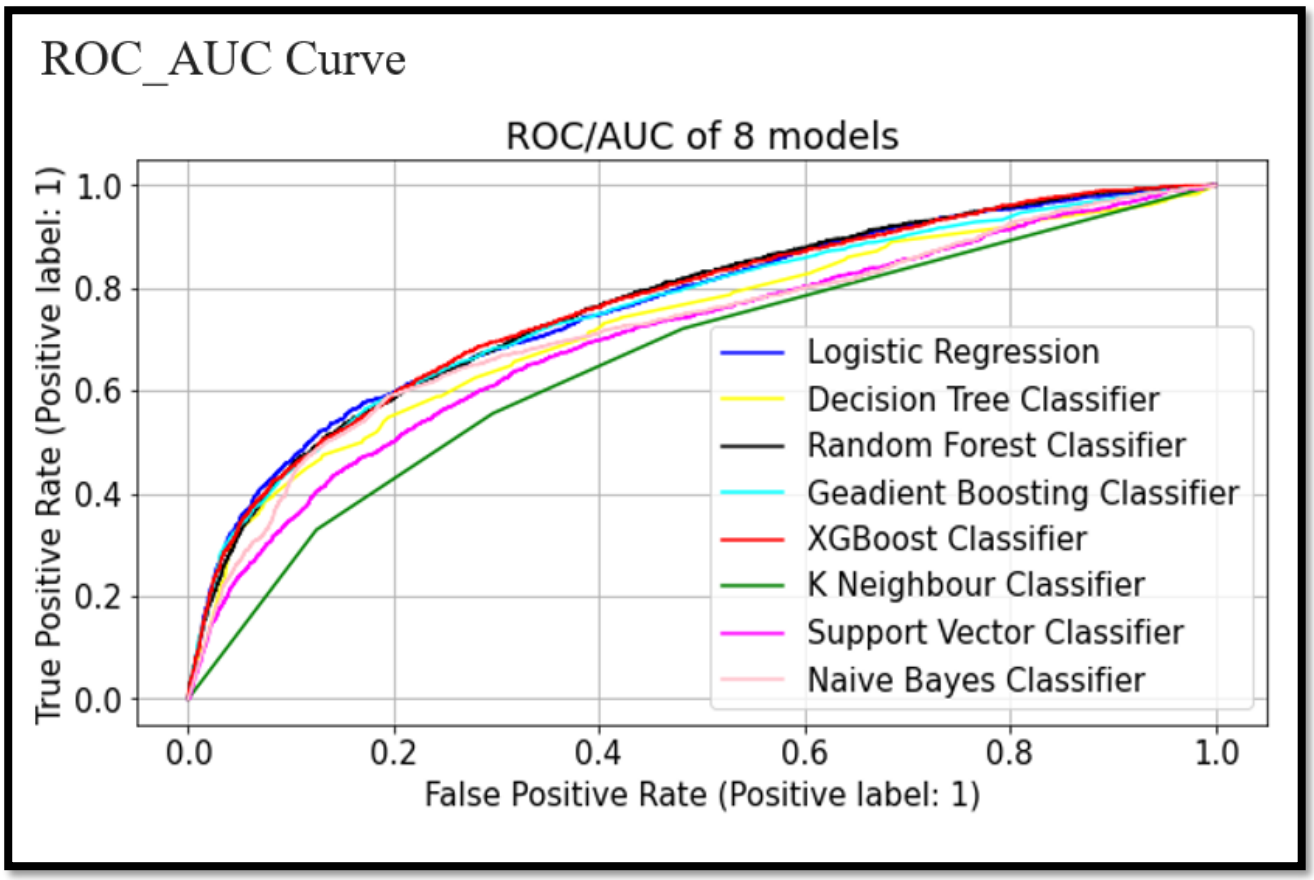
  - Confusion matrix
  - ROC_AUC curve
  - Precision recall curve

  We will see whole Performance Metrics in graph in next topic.

- **Feature Importance:** By plotting the feature importance Random Forest model and XGBoost model, it is clear that 'PAY_1','PAY_2' (the most recent 2 months payment status), along with credit limit (LIMIT_BAL) are the most important predictors. Since we don't have customer income data, generally speaking, higher credit limits are associated with lower default risk.
  We will see Feature Importance graph in next topic.

Evaluation

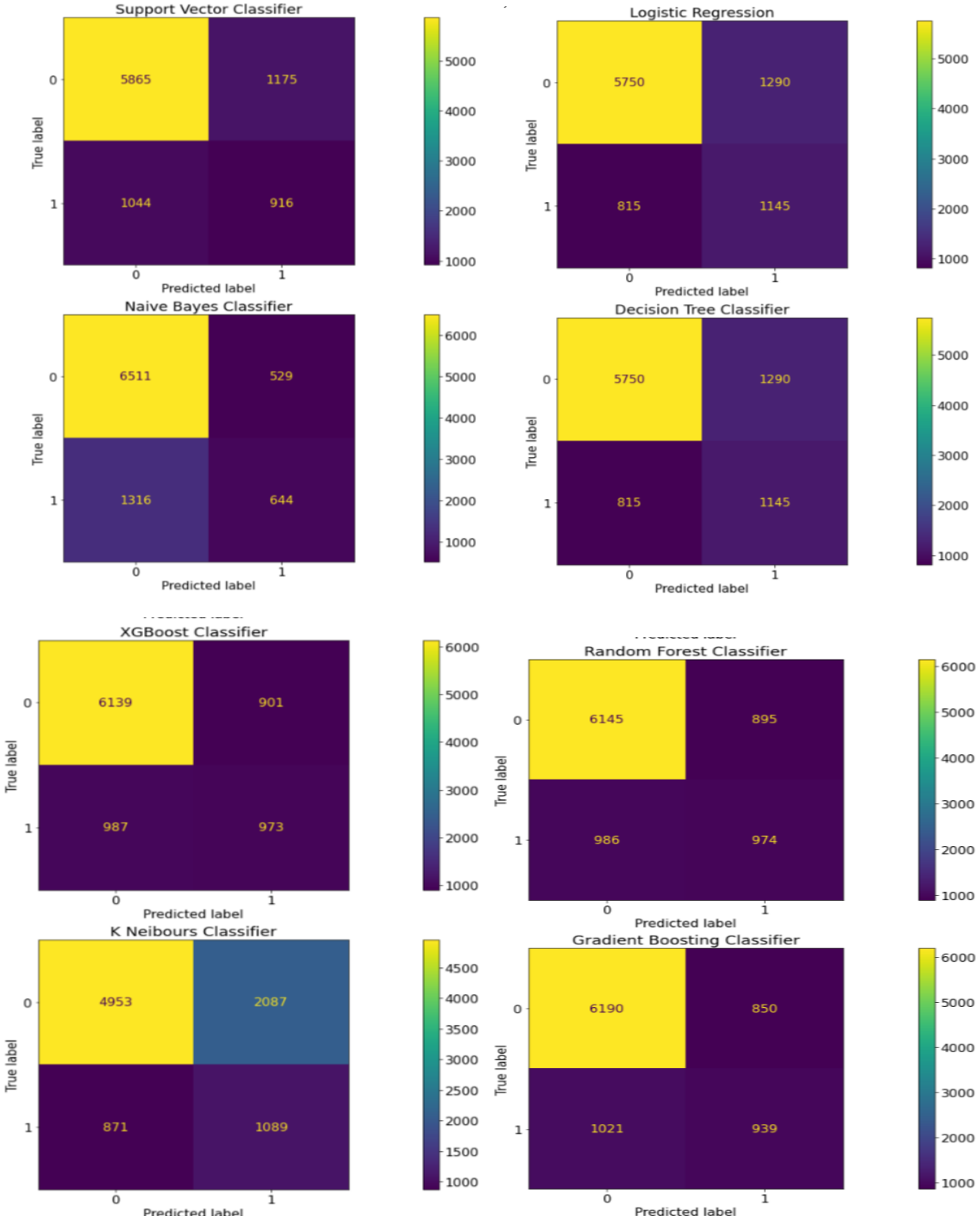
- This plot shows ROC-AUC curve for whole 8 model.
- XGBoost classifier algorithm shows best performance compare to other.



- Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate (1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.
- ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

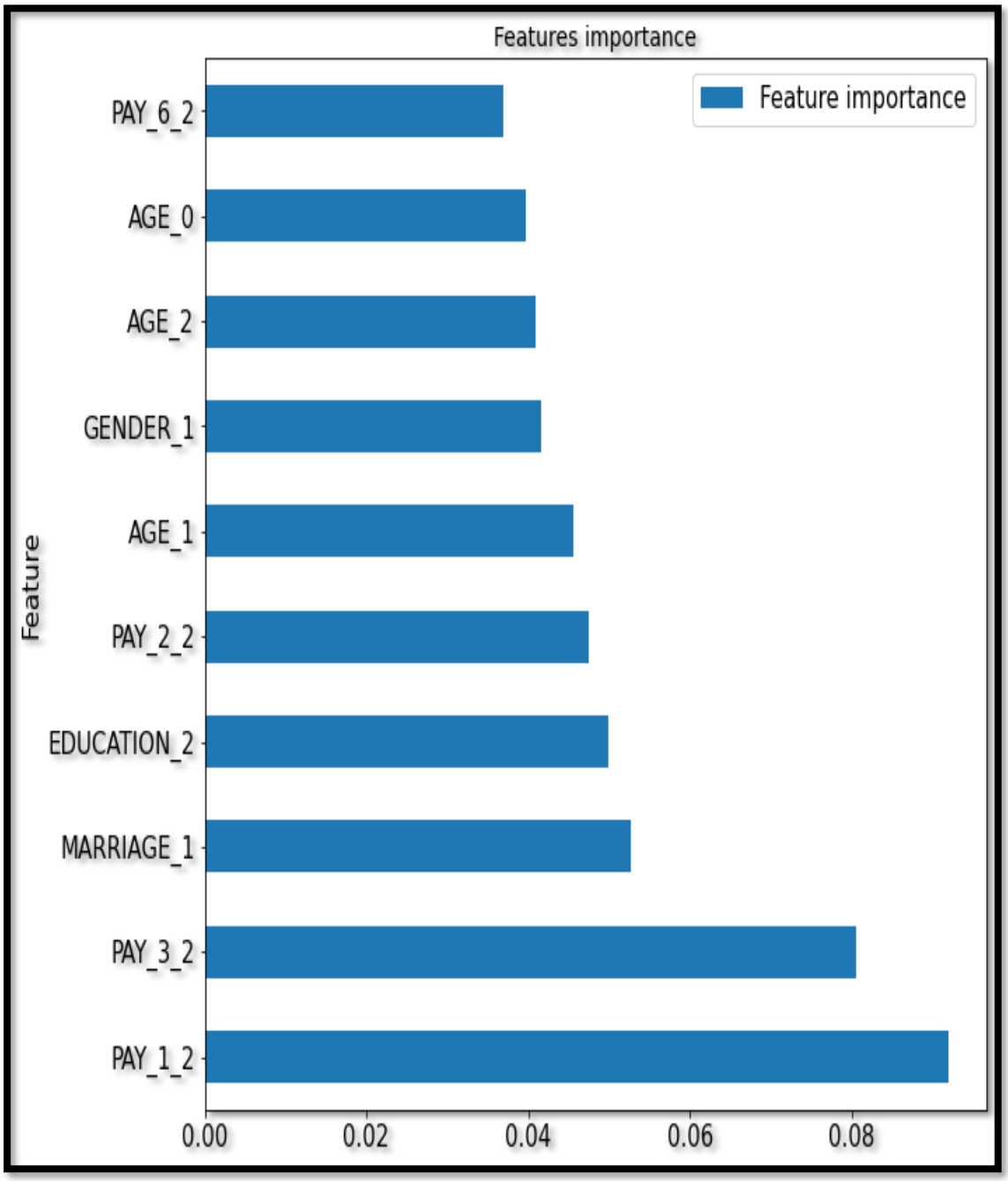
Confusion Matix

- A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.



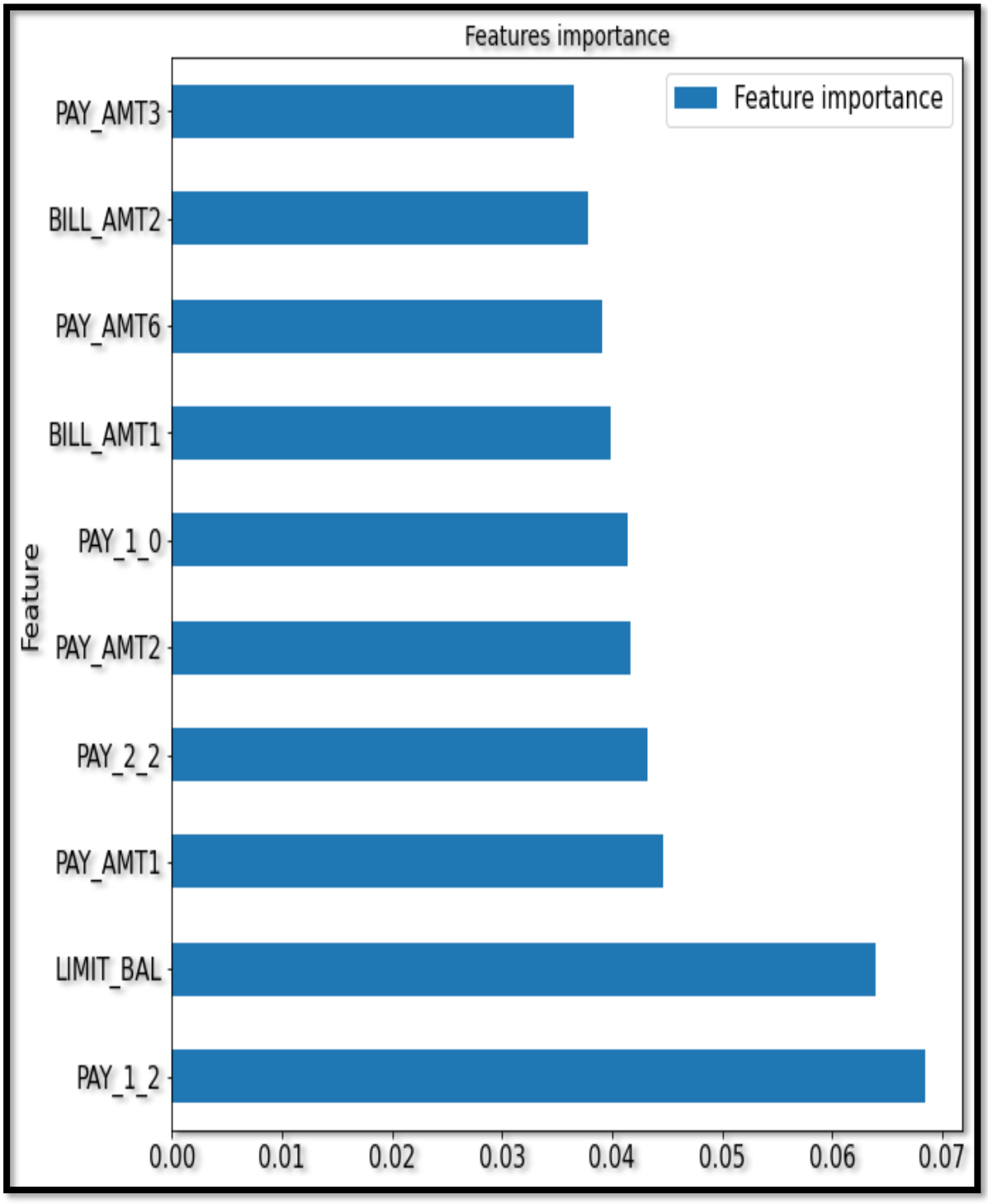
Feature Importance

Feature Importance Graph for XGBoost Model



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

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 (i.e., 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%