Credit Card Default Prediction

Adil Khan, Sunil Panigrahi, Shubham Kumar,
Vivek Singh, Sharaffin B
Data Science Trainees,
AlmaBetter, Bangalore

ABSTRACT

Credit risk plays a major role in the banking industry business. Bank's main activities involve granting loans, credit card, investment, mortgage, and others. Credit cards have 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. Also, many customers used their credit card beyond their repayment capabilities leading to high debt accumulation. This situation creates a problem for the bank. With the help of historical data, the default customers can be determined. As such, machine learning offers solutions for addressing the current issue and handling credit risk, thus this project used ML algorithms of classification to get predictions with accuracy to detect the default users based on previous six months data.

1. PROBLEM STATEMENT

. The fundamental objective of the project is to analyze credit card data collected from Taiwan-based credit card issuers and predict whether or not a consumer will default on their credit cards, as well as identify the key drivers behind this. This would inform the issuer's decisions on who to give a credit card.

2. INTRODUCTION

Default credit cards happen when clients fail to adhere to the credit card agreement, by not paying the monthly bill. Thus, the primary objective of this analysis will focus on finding the best model that predicts well the likelihood of customers' default on credit cards. In other words, it will help us to predict, with reasonable accuracy, whether the customer would fail or succeed in making the next payment. Various models will be run, compared and the best one with better prediction accuracy will be chosen. The developed models will consider all possible factors in the data set.

We will use the following several predictive models such as Logistic Regression, Support Vector Classification, Bagging Classifier, Random Forest Classification (RFC), XGBoost and compare them.

Finally, we discuss about the inferences we find from different ML models and summarize the project

3. ATTRIBUTE INFORMATION

Below there are the description of the attributes that will be used in our model for better understanding of the data:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary = credit)
- SEX: 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)

4. DATA PIPELINE

- **Data Processing**

In the first part, we have imported necessary libraries. We then used these libraries to understand the dataset.

- **Data Cleaning**

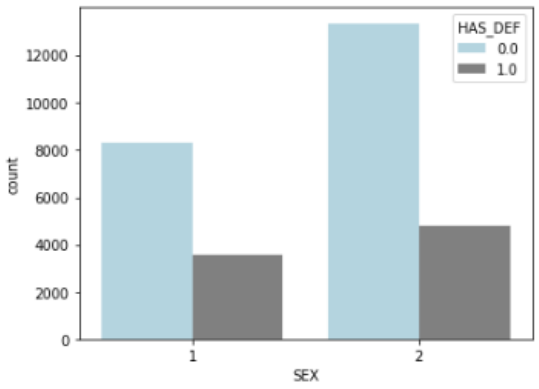
After understanding the data, we got to know that there are no missing values, null values or duplicate values, outliers in our data set and also checked for column type and column names

- **Exploratory Data Analysis**

After preparing the data set, we did some exploratory data analysis using tables and graphs to derive the observations from the data and to better understand the problem statement, and make ways to find the solution to the problem statement.

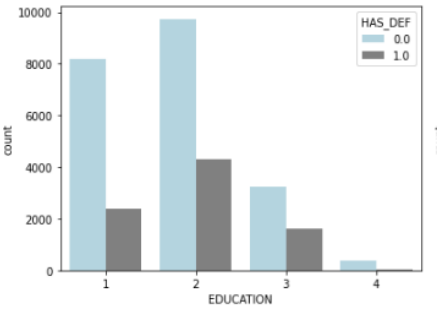
While doing EDA we tried to answer a few questions below using univariate, bivariate and multivariate analysis.

1) **Is the proportion of defaults the same for men and women?**



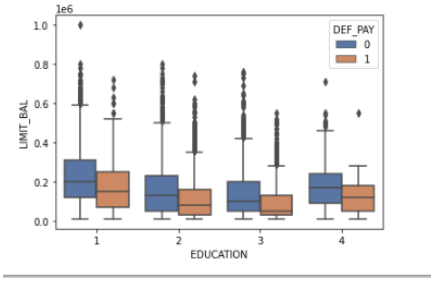
The data set contains 11888 males and 18112 females. 30% male have default payment while 26% female have default payment. The proportion of defaults for men is slightly higher than the proportion of defaults for women.

2) **Did customers with higher education have less number of delayed payment**?



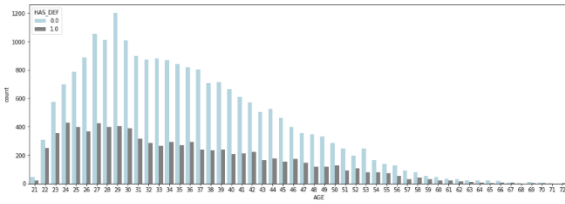
The proportion of defaults seems to decrease as the level of education increases. Customers with high school and university educational level had higher default proportions than customers with grad school education. More number of credit holders are university students followed by Graduates and then High school students

3) **Did customers with a higher education level get higher credit limits?**



clients who have lower than high school level education tend to have default payment more.

4) **Does age has any relation with Credit Limit and Default Payments?**

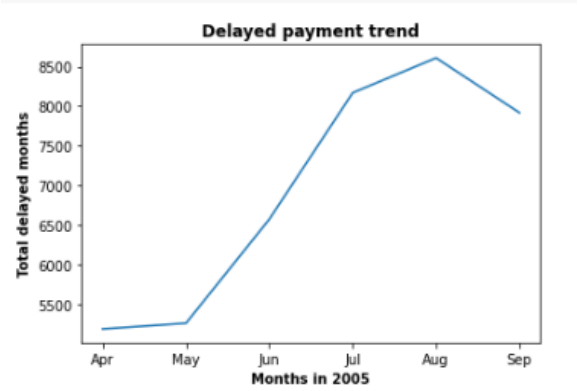


There are more adults as compared to old people above 40 and below 60

There are very low senior citizens

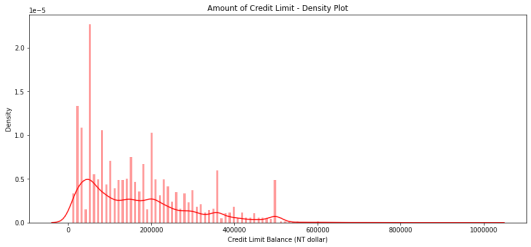
Customers aged between 30-50 had the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates. However, the delayed rate dropped slightly again in customers older than 70 years.

5) **Has the repayment status changed in the 6 month from April 2005 (PAY_6) to September 2005(PAY_1)?**



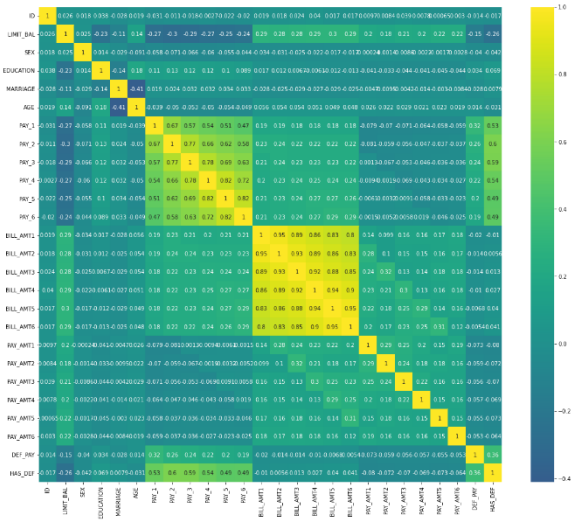
There was a huge jump from May,2005 (PAY_5) to July, 2005 (PAY_3) when delayed payment increased significantly, then it peaked at August, 2005 (PAY_2), things started to get better in September, 2005 (PAY_1).

6) Distribution of credit limit to customers



the credit limit of customers are usually lower than 2,00,000 TWD. Credit limit at 50,000 TWD has the highest density among all the credit limit levels, followed by 20,000 TWD credit limit. 2,00,000 TWD credit limit reaches another peak after the credit limit at 80,000 TWD

MULTICOLLINEARITY HEATMAP



- "age" is inversely correlated with "Marriage".
- the bill amounts from April to September are highly positively correlated.
- Multicollinearity exists between the payment status variables.
- Moreover, the bill amounts of each month are weakly positively correlated or even NOT correlated to the payment amount of the same month, meaning that the customers are not paying the exact amount of their bills.
- The payment status(PAY_1 TO PAY_6) and credit limit(LIMIT_BAL) are negatively correlated, i.e. the later the repayment, the lower the credit limit.
- The credit limit is inversely correlated with the customer having a default payment next month.

5) MODEL BUILDING

This section focuses on building and choosing the best model for predicting the defaulters. Before building the model, the dataset was divided into training and test data set.

- Train Data Set = 80%
- Test Data Set = 20%

- **Fitting different models**
  For modelling we tried various classification algorithms like:

1. **Logistic Regression**
2. **Support Vector Classification**
3. **Bagging Classifier**
4. **Random Forest Classification**
5. **Gradient Boosting Algorithm**
6. **XGBoost**

- **Tuning the hyperparameters for better accuracy**
  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models like Random Forest Classifier and XGBoost classifier.

5.1. ML Algorithms:

The best parameters, the accuracy, precision, recall and the KS chart of each model has been given below

Logistic Regression:

Logistic Regression is a classification algorithm that has been given the name regression due to its mathematical resemblance to linear regression.

Accuracy on train data: 0.81025

Accuracy on test data: 0.80883

Precision: 0.7163461538461539

Recall: 0.22456669178598343

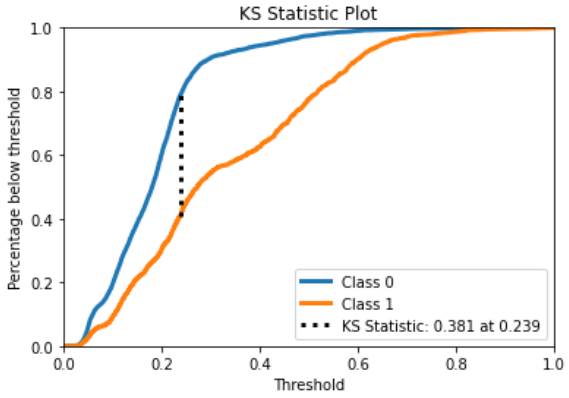
The best parameters were :

'C': 1.0,

'penalty': 'l2',

'solver': 'newton-cg'

The KS chart had a value of 0.381



.

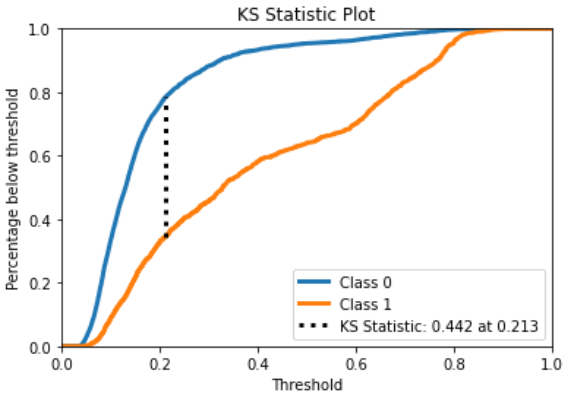
Support Vector Classification

Parameters

- C': 1.0,
- 'gamma': 'auto',
- 'kernel': 'rbf

Accuracy: 0.7886666666666666
Precision: 0.7322834645669292

Recall: 0.07008289374529013



Bagging Classifier:

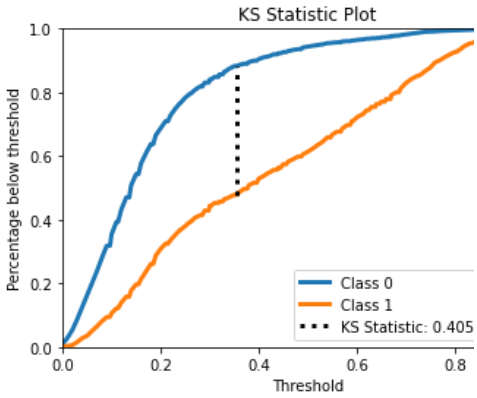
Accuracy: 0.8125
Precision: 0.6268844221105527
Recall: 0.3760361718161266

PARAMETERS

'n_estimators': 45

Random Forest Classifier:

Accuracy: 0.8135
Precision: 0.6357702349869452
Recall: 0.36699321778447624

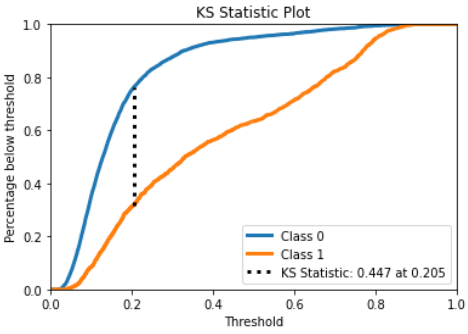


BEST PARAMETER

n_estimators': 1000

XGBoost

The accuracy on test data is
0.8213333333333334
The precision on test data is
0.36623963828183875
The recall on test data is
0.6778242677824268



Gradient Boosting Algorithm

Accuracy: 0.8186666666666667
Precision: 0.665742024965326
Recall: 0.3617181612660136

learning_rate: 0.1,
max_depth: 3,
n_estimators: 50,
subsample: 0.5

Hyper parameter tuning:

A hyperparameter is a set of information that gives an algorithm some control over how it learns. In order to learn from these hyperparameters, their definitions impact the parameters of the models. This set of values impacts performance, stability, and interpretation of a model. Each algorithm requires a specific grid of

hyperparameters that can be customized to fit the business problem. The hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate output.

We used **Grid Search CV** for hyperparameter tuning. This also results in cross validation and in our project we divided the dataset into different folds.

6. CONCLUSION

Let's take a look at what we done in particular project. First, we loaded the dataset and performed data cleaning and null value treatment. Thankfully there were no null values and duplicate values are present in the dataset. We did outlier treatment and did some EDA and data visualization. We plot different graphs of univariate analysis and bivariate analysis and make different inferences from our dataset.

At the end part of EDA we plot the correlation heatmap and find the correlation of each independent variable with our target variable.

.

In modeling part we tried six different classification models and we got the following results:

a) The dataset was not normally distributed and mostly it was right skewed

b) Male customers have more default payments than female customers.

c) There is NO significant difference in the proportion of default payment across different education levels. But clients who have lower than high school level education tend to have default payment more.

d) Married clients have a higher default payment rate than single or other marital status clients.

e) People who have the payment delay for two months have a high ratio of default next month (October). In September, a quarter of customers who repay one month later have default payment next month in October. This situation does not exist in other months as almost no one repays one month later.

The developed models took into account all possible factors and data. This final chosen model would benefit the bank before they make any decisions against that customer.

XGBOOST has the highest accuracy of 82.5 % with a recall of 67.7 % and KS chart value of 0.447 and proved to be the best model.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Stack Overflow