Credit Card default prediction

Raja Chowdhury, Saaquib Mustafa, Sandipan Das, Sahil Kolambkar

Data science trainees,

AlmaBetter,Bangalore

**Abstract:**

Credit risk plays a major role in the banking industry business. Banks' 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.

1. **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](https://www.listendata.com/2019/07/KS-Statistics-Python.html) to evaluate which customers will default on their credit card payments

### Attribute Information:

### This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

### X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

### X2: Gender (1 = male; 2 = female).

### X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

### X4: Marital status (1 = married; 2 = single; 3 = others).

### X5: Age (year).

### X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.

### X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

### X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

1. **Introduction**

### A Taiwan-based credit card issuer wants to better predict the likelihood of default for its customers, as well as identify the key drivers that determine this likelihood. This would inform the issuer’s decisions on who to give a credit card to and what credit limit to provide. It would also help the issuer have a better understanding of their current and potential customers, which would inform their future strategy, including their planning of offering targeted credit products to their customers. There are times when even a seemingly manageable debt, such as credit cards, goes out of control. Loss of job, medical crisis or business failure are some of the reasons that can impact your finances. In fact, credit card debts are usually the first to get out of hand in such situations due to hefty finance charges (compounded on daily balances) and other penalties. A lot of us would be able to relate to this scenario. We may have missed credit card payments once or twice because of forgotten due dates or cash flow issues. But what happens when this continues for months? Here we will talk about what counts as credit card default and the circumstances you are likely to face when you default on your cards.

**3.Steps performed in this EDA**

Handling dataset with the fundamental steps to unveil the factors:

* Importing Libraries and Loading the Datasets
* Overview of The Datasets
  + Reading & Inspection of dataset
  + Further analysing the dataset
* Data Wrangling and Processing
  + Replacing the category 0 to 1 in order to maintain the sequence
  + Cleaning categorical values for education and marriage
  + Renaming the columns for better analysis
  + Creating age buckets for better analysis
  + Handling Outliers
* Exploratory Data Analysis
* Key Findings and Conclusion from EDA
* Feature Engineering
  + Multicollinearity
  + Dummification
  + Scaling Numerical Columns
* ML Model
  + Train-Test Split
  + Model Training and Prediction
    - Logistic Regression
    - Support vector classifier
    - Decision Tree classifier
    - Random Forest
    - XGboost
* Hyperparameter tuning
  + - Logistic Regression
    - Support vector classifier
    - Decision Tree classifier
    - Random Forest
    - XGboost
* Feature importance of tuned and untuned models
* Model performance of our best model

Conclusion

**4.Data Processing & Preparation**

* We have reduced the categories of marriage and education for better analysis.
* We have divided age into several categories.

**NULL VALUE TREATMENT**

There are no null values present in our dataset.

**Handling Outliers**

* We used boxplot to detect outliers.
* Then we used z score to remove those outliers.

**Data Analysis and Findings**

**1.What is the ratio of defaulters and non defaulters?**

The number of people who will default their payment is much less than number of people of people who will pay on time

**2.What is the ratio of males and females along with their default status?**

The Number of females are much more than males in our dataset (6,224) and we can also say that females tend to pay their default on time compared to their male counterparts.

**3.What is the history of past payment?**

From history of past payment analysis it is clear that most people who pay duly are not likely to default their payment.

**4.Which age group has the highest limit balance?**

Here, we can say from age 21 to 39 limit balance is increasing however from 39 to 61 it started to decline and then from 62 to 79 it has increased drastically.

**5.Feature Engineering:**

* Created payment values and dues column in the dataset.
* Removed some of the highly correlated columns because it might influence the data after going through the heatmap.

**6. Machine Learning Algorithms used:**

* **Logistic Regression-**In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model.
* **Support Vector Classifier-**A support vector machine (SVM) is a type of deep learning algorithm that performs supervised learning for classification or regression of data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, which are labeled for classification.

### Decision Tree- **Decision tree** builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

### Random Forest-**Random Forest Regression** is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

### XG boost- **XGBoost** is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e. how far the model results are from the real values. The most common loss functions in XGBoost for regression problems is reg: linear, and that for binary classification is reg: logistics.

### 7. Evaluation Metrics used:

### Precision Score-The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0.

### Accuracy Score-Accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions made. We calculate it by dividing the number of correct predictions by the total number of predictions.

### Recall Score-Recall score is used to measure the model performance in terms of measuring the count of true positives in a correct manner out of all the actual positive values. Precision-Recall score is a useful measure of success of prediction when the classes are very imbalanced.

### AUC ROC score- ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

### F1 Score-The F1 score is defined as the harmonic mean of precision and recall. As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate.

### 6.2 Hyper parameter tuning-

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

### 7.Conclusion:

Random Forest Classifier has the best value of accuracy score of 84%.Random Forest Classifier has the best value of precision score of 84%.Random Forest Classifier has the best value of recall score of 83%.Random Forest Classifier has the best value of f1 score of 84%.Random Forest Classifier has the best value of Roc\_auc score of 84%.Random Forest Classifier gave the highest importance to Dues then to payment value and then to limit balance columns accordingly.Random Forest Classifier had average fitting time of 1.67 according to cross validate method of sklearn. This proves Random Forest Classifier algorithm has perfectly fitted all the dataset.