# Credit Card Default Prediction

DATA1030 project final report

Yun Li
https://github.com/YunLi1996/DATA1030FinalReport.git

Dec 7, 2021

## 1 Introduction

Credit card default is an important aspect of risk management for banks. With banks over-issuing credit cards to expand market shares and clients over-spending using their cards, credit card default has become a non-negligible risk. It may affect a bank's functionalities, and in severe cases it may even effect a whole economy. The reasons and mechanisms behind this phenomenon has been well discussed and studied in Finance and Economics literature, and researches have various types of business models to predict credit card default rates. However, with more and more relevant data available, it is also worth exploring this topic under the framework of data science, and backing up existing prediction theories with machine learning models.

This project aims at developing a predictive machine learning tool for classifying whether a credit card user will default on his or her payment in the coming month. The dataset I use in this project comes from the UCI Machine Learning Repository (Yeh,2016). This dataset was donated by two researchers Yeh and Lien and it was originally collected by a bank in Taiwan. The dataset consists of 30,000 observations, each for a unique credit card user. There are in total 24 features in this dataset, which include an ID for credit card users, users' balance limit, their sex, educational level, marital status, and age. There are 18 other features which reflect credit card activity histories of the users - 6 features track users' monthly repayment delay status for the past 6 months, 6 features track users' monthly bill statement for the past 6 months, and 6 features track users' monthly repayment amount for the past 6 months.

There are a number of papers using this dataset to predict credit card users' default behavior, and they also discuss and compare different algorithms used for prediction. The paper by Yeh and Lien (2009) compared 6 different data mining techniques and discovered that artificial neural network gave the best predictive performance, with an R-squared of 96.47 %. Another paper by Koklu and Sabanci (2016) used Multilayer Perceptron (MLP) and k Nearest Neighbors (kNN) machine learning algorithms, with the highest successful rate at 80.66 %.

## 2 Exploratory Data Analysis

In this section I present some figures from exploratory data analysis on the dataset.

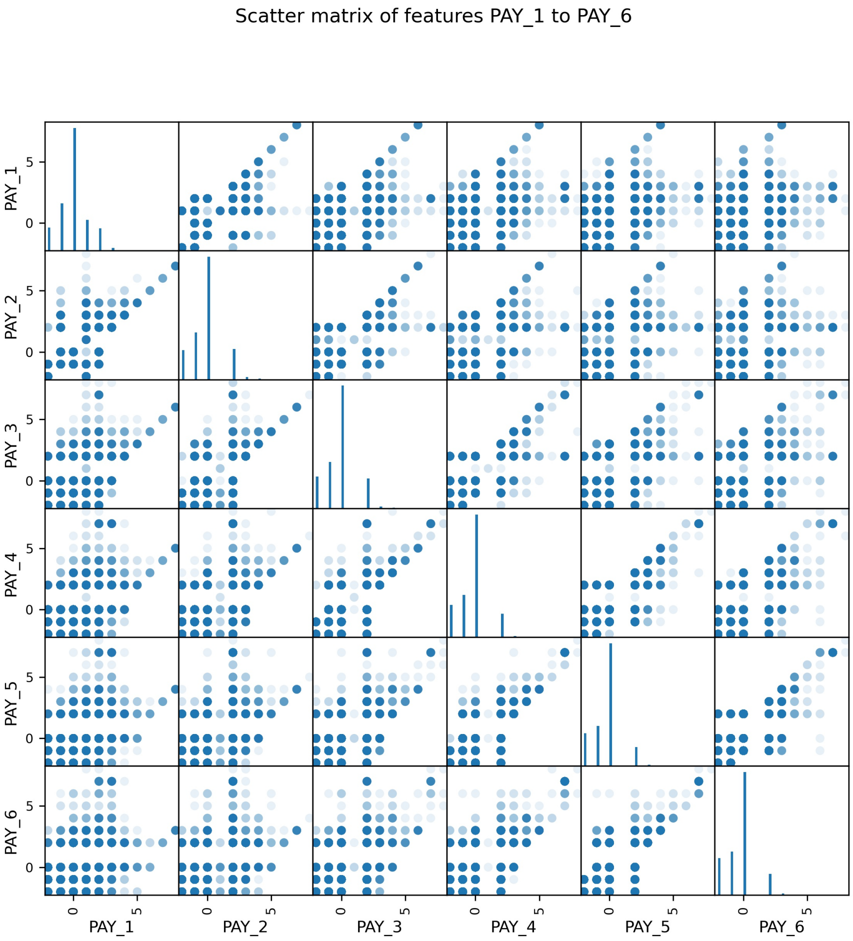
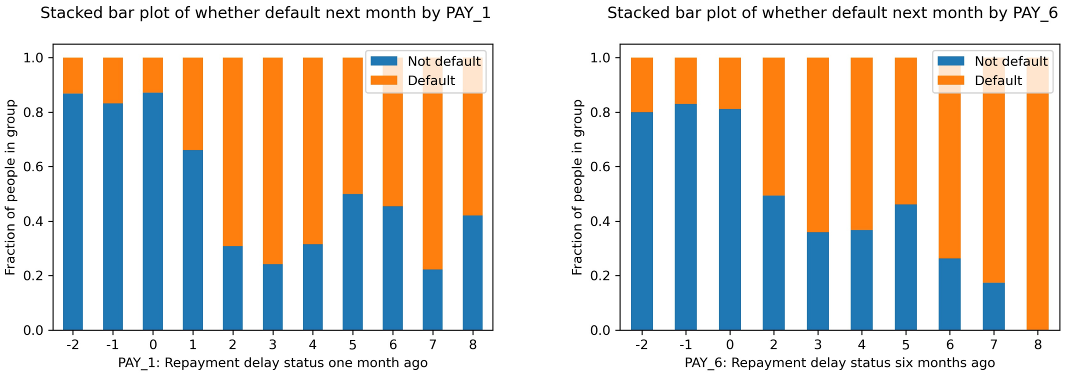


**Figure 1** This figure displays the scatter matrix plot of features PAY_1, PAY_2, ..., PAY_6, which present the users' repayment delay status one month ago, two months ago, ..., six months ago, respectively. For these features, a higher value indicates a longer period of repayment delay in that month. This figure shows that all the historical repayment delay status features are roughly positively correlated with each other. For months closer in terms of time, repayment delay status exhibit a more eminent positive correlation with each other.



**(a)**

**(b)**

**Figure 2** This figure displays the stacked bar plots of whether the credit card users default next month grouped by two features: (a) the users' repayment delay status one month ago, and (b) the users' repayment delay status six months ago. For the repayment delay status features, a higher value indicates a longer period of repayment delay in that month. Both sub-figures 2.(a) and 2.(b) show that credit card users with longer repayment delay periods seem more likely to default in the next month. Comparing the sub-figures, this relationship seems more eminent for repayment delay status six months ago. Given the relationship found in this figure, the features PAY_1 to PAY_6 can potentially be very predictive of the target variable.

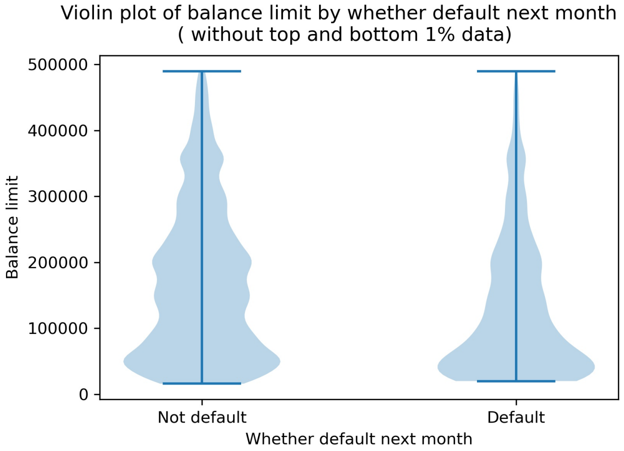


**Figure 3** This figure displays the violin plot of credit card users' balance limit grouped by whether they default next month. Since the feature balance limit has high outliers, I only keep the middle 98% of data points of this feature for this plot, so that the violin shapes are not too heavy on their tails. This figure shows that users who default on repayment next month seem to be those with lower balance limit, and it is very rare for them to have balance limit over 300,000 New Taiwan Dollars. This implies that balance limit is probably also predictive of the target variable.

# 3 Methods

## 3.1 Data Splitting and Pre-processing

In the data splitting process, I allocate 20% of the observations to testing, the other 80% to 4-fold cross-validation (so in total 20% of observations are allocated to validation, and 60% of observations are allocated to training). I choose to use the 4-fold splitting method to increase the number of models trained, which may potentially improve the predictive power of the machine learning models used in later steps. Although this dataset includes historical credit card activities for different months, it does not exhibit time-series structure, since each unique individual only corresponds one observation (one row) of the dataset. There is also no obvious group structure in the dataset. Therefore, I assume the used dataset to be independently and identically distributed.

I apply OrdinalEncoder to 7 features - EDUCATION (educational level), and PAY_1 to PAY_6 (users' monthly repayment delay status in the past 6 months), since these features are categorical, and their categories can be ordered in a logical way. For EDUCATION, a higher value indicates a lower educational level, and for PAY_1 to PAY_6, a higher value indicates a longer period of repayment delay. I apply OneHotEncoder to two features - SEX (sex) and MARRIAGE (marital status), since these features are categorical, and their categories cannot be logically ordered. I apply MinMaxScaler to 1 feature - AGE (age), since it is continuous and have reasonal bounds. I apply StandardScaler to the remaining 13 features - LIMIT_BAL (users' balance limit), BILL_AMT1 to BILL_AMT6 (users' monthly bill statement for the past 6 months), and PAY_AMT1 to PAY_AMT6 (users' repayment amount for the past 6 months), since these features are continuous, and they are not necessarily reasonably bounded. No encoder or scaler is applied to ID (users' id) since it is not used in model training. After pre-processing, there are 27 features in the dataset excluding ID.

## 3.2 Model selection

After splitting and preprocessing, I train and compare 4 machine learning models: a logistic regression model with L1 regularization (Logistic L1), a logistic regression model with L2 regularization (Logistic L2), a random forest classifier model (RF), and a K-nearest neighbors classifier model (KNN). I tune hyper-parameters for all the models using the grid search method. I use f1 score as the evaluation metric for cross validations. Table 1 presents the models trained, their hyper-parameters tuned and search spaces.

| Model | Hyper-parameter | Search space |
|---|---|---|
| Logistic L1 | C: inverse of regularization strength | [1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4] |
| Logistic L2 | C: inverse of regularization strength | [1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4] |
| RF | n_estimators: number of trees in the forest<br>max_depth: maximum depth | [50, 100, 500, 1000]<br>[50, 100, 500, 1000, None] |
| KNN | n_neighbors: number of neighbors<br>weights: weight function used in prediction | [1, 2, 3, 5, 10, 30, 100, 200, 500]<br>["uniform", "distance"] |

**Table 1** ML models, their hyper-parameters trained and search spaces

After hyper-parameter tuning, the optimal models are recorded and used for comparison. Figure 4 shows the average f1 scores for the optimal models across the random states.

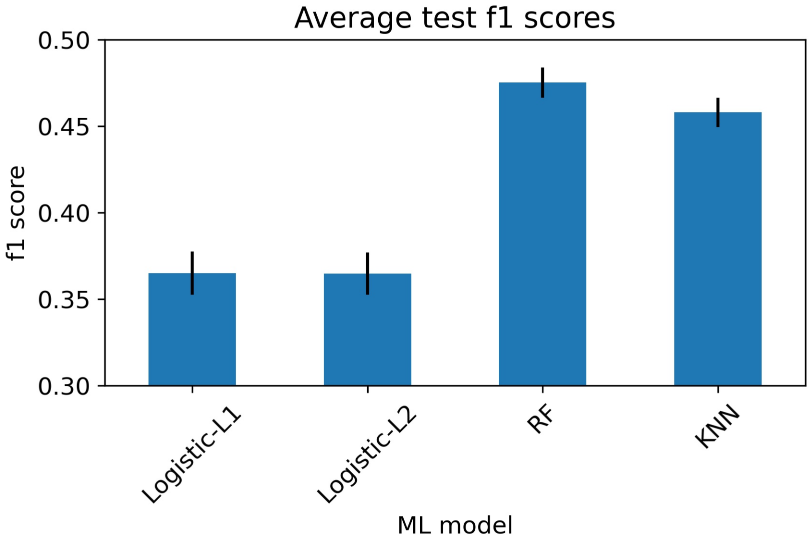


**Figure 4** Average test f1 scores for optimal models across random states

The random forest classification model performs the best on the test set so it is chosen as the model of choice. The optimal maximum depth is 500 and the optimal number of estimators is 1000.

# 4 Results

## 4.1 Model evaluation

After choosing the optimal model and its optimal hyper-parameters, the model of choice is retrained on new splits using 10 new random states. For each split, 80% of data is allocated to training and 20% to testing. Over the 10 splits, the average test f1 score for the baseline model is 0.3626 with a standard deviation of 0.0056, while the average test f1 score for the model of choice is 0.4542 with a standard deviation of 0.0096. Therefore, the model of choice scores on average 16 standard deviations above the baseline model.

Figure 5 shows the confusion matrix over the 10 random states. While the model classifies the negative cases pretty well, its power of classifying the positive cases is relatively weak.

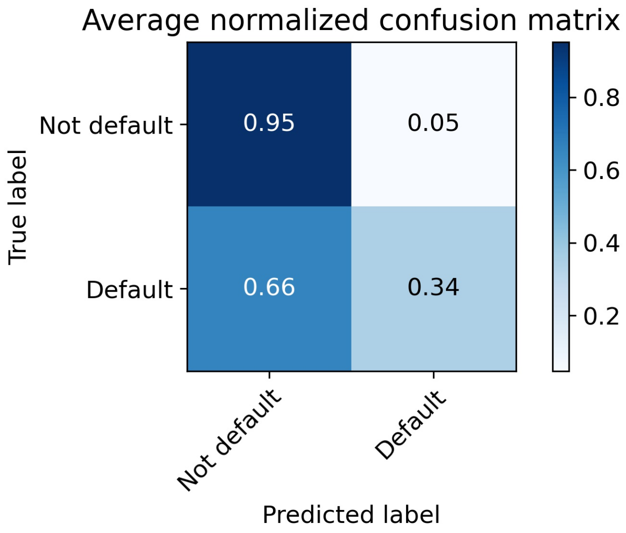


**Figure 5** Average normalized confusion matrix

## 4.2 Global feature importance

Global feature importance for the model of choice is calculated using two methods - the model coefficients and the permutation tests over 10 shuffles.

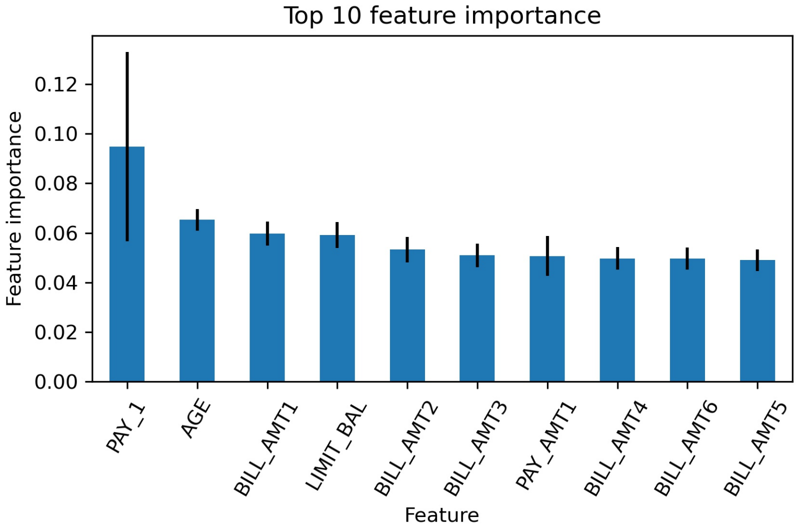


**Figure 6** Top 10 feature importance scores in order

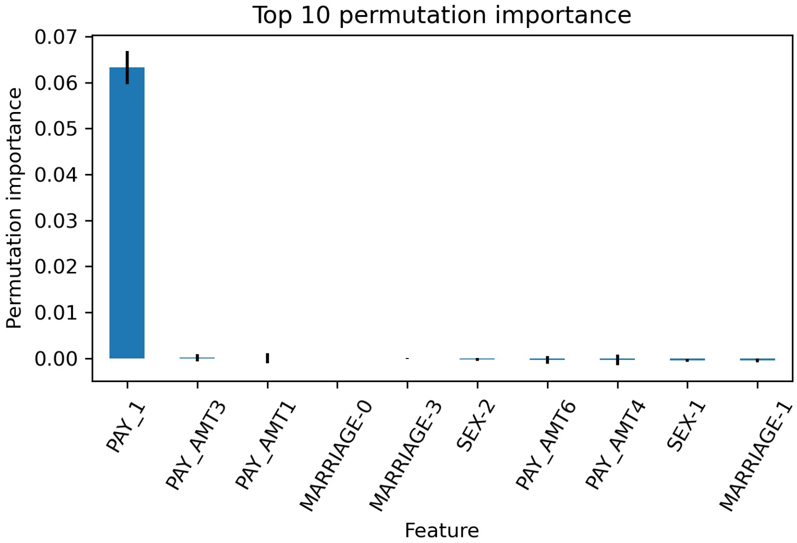


**Figure 7** Top 10 permutation importance scores in order

Figure 6 shows the 10 most predictive features in terms of model coefficients, and figure 7 shows the 10 most predictive features in terms of permutation test scores. These figures show that the three sets of features on credit card activity histories are more predictive than the demographic variables. Moreover, activity histories further in the past are less predictive than activity histories closer to the present. This is quite intuitive, since individuals' habits of using credit cards are likely to change gradually over time.

## 4.3 Local feature importance

Local feature importance for the model of choice is calculated using the SHAP values. Figure 8 shows the SHAP values for the most predictive feature PAY_1, which is the credit card repayment delay status one month ago. A higher value of PAY_1 indicates a longer period of repayment delay. The figure shows in general a positive relationship between last months' payment delay and the current month's credit card default behavior. This is quite intuitive, since individuals who have not been able to pay their credit card bills for a longer period in the past are less likely to be able to pay credit card bills now.

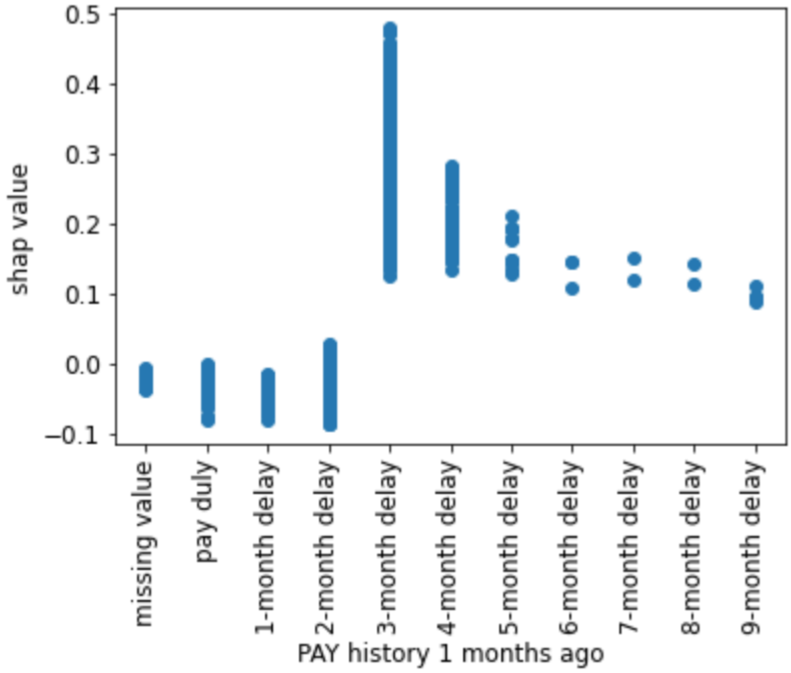


**Figure 8** SHAP values of payment delay status 1 month ago

## 5 Outlook

The major contribution of this study is that it provides evidence about the predictive power of credit card activity histories on credit card default behavior. Based on this result, bank managers are advised to conduct thorough credit investigations on customers before issuing credit cards to them.

The main weakness of the model I choose lies in its relatively low predictive power for positive cases. There are multiple ways through which I could improve this in the future. I could include other different machine learning models (such as support vector machine classifiers) in the training process. I could also collect more data to retrain the models. I believe additional data on individuals' historical general financial activities (such as histories of income streams, investments and tax payments) would be very helpful, since credit card activities are inevitably correlated with one's general financial status.

# References

Koklu, M. and Sabanci, K. (2016). Estimation of Credit Card Customers Payment Status by Using kNN and MLP. *International Journal of Intelligent Systems and Applications in Engineering*, 4:249–251.

Yeh, I.-C. (2016). UCI Machine Learning Repository. The data is available online at this address. `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`.

Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36:2473–2480.