Credit Card Default Prediction

Self Project by
Onkar A. Pawar

1. Introduction

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. This can be proven by industry business reports and statistical data. For example, "The Federal Reserve Bank of New York measures credit card delinquencies based on the percent of balances that are at least 90 days late. For the third quarter of 2019, that rate was about 8%, about the same level as in the previous quarter." Thus, assessing, detecting, and managing default risk is the key factor in 2 generating revenue and reducingthe 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 analyzing 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." 3.

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.

2. Problem Description:

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.

3. Data Exploration:

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. Data dictionary is available in Appendix A.

The main findings from the exploratory analysis are as follows:

- Males have more delayed payment than females in this dataset. Keep in mind that this finding only applies to this dataset, it does not imply this is true for other datasets.
- Customers with higher education have less default payments and higher credit limits.
- Customers aged between 30-50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. However, the delayed rate drops slightly again in customers older than 70.
- There appears to be no correlation between default payment and marital status.
- Customers being inactive doesn’t mean they have no default risk. We found 317 out of 870 inactive customers who had no consumption in 6 months then defaulted next month.

3.1 Gender Variable:

Males have more delayed payments than females in this dataset.

Which gender group tends to have more delayed payment? Since there are more females than males in the dataset, we use percentage of default within each sex group. Figure 1 shows 30% males have default payment while only 26% females have default payment. The difference is not significant. To verify if this is due to chance, we use a permutation test and a t-test on each group's default proportions and mean respectively, and the results support the findings.

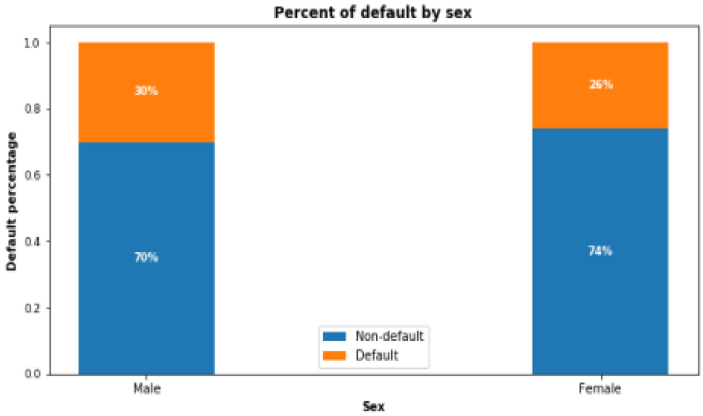


Fig3.1 Percent of default by sex.

3.2 Education Variable:

Customers with higher education have less delayed payment.

Figure 3.2 indicates customers with lower education levels default more. Customers with high school and university educational level have higher default percentages than customers with grad school education. Notice there is an education group “others” which appears to have the least default payment, but this group only has 468 (or 1.56%) customers, and we don’t know what consists of this group.

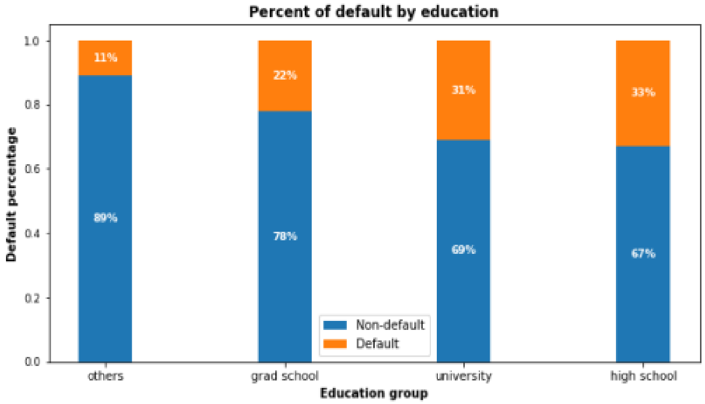
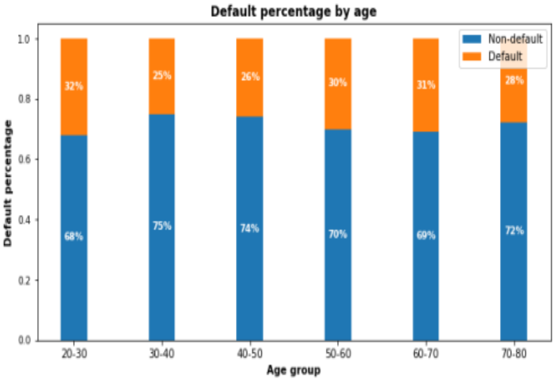


Fig 3.2: Percent of default by education level. The “others” group only consists of 1.56% of total customers.

3.3 Age Variable:

Middle-aged customers have the lowest default rate.

The bar chart in figure 4 shows the default probability increases for customers younger than 30 and older than 70. Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. This aligns with social reality that customers aged 30-50 typically have the strongest earning power. We also notice the delayed rate drops slightly again in customers older than 70. This is understandable because elder customers' consumption tends to decrease. Lastly, we use a Chi-squared test to verify this finding and the test statistics support it.



4. 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 preprocessing 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 stauts, 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.

● 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). An 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.

● 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.

5. Predective Modeling:

This analysis uses 3 classification models - Logistic Regression, Random Forest and XGBoost. Since Random Forest and XGBoost are tree based on algorithms, rescaling is only performed on Logistic Regression, not on these 2 models. For each model, we first try the model's default parameters, train each model without SMOTE and with SMOTE samplings. Then tune each model's hyperparameters to find the optimal performance. As mentioned earlier, this dataset has imbalanced classes, therefore we use precision and recall, instead of accuracy as the performance metrics.

SMOTE Oversampling:

In the initial model fitting, we start by using all models' default parameters. To compensate for the rare classes in the imbalance dataset, we use SMOTE(Synthetic Minority Over-Sampling Technique) method to over sample the minority class and ensure the sampling is not biased. What this technique does under the hood is simply duplicating examples from the minority class in the training dataset prior to fitting a mode. After SMOTE sampling, the dataset has equal sizes of 0s and 1s. In order to verify if SMOTE improves models' performance, all 3 models are trained with SMOTE and without SMOTE. Below table shows the ROC_ AUC scores on training data 9 improved significantly with all models after over sampling with SMOTE. This proves SMOTE is an effective method in sampling imbalanced dataset.

● Hyperparameters Tuning:

We utilize Scikit-Learn library's built in functions such as cross-validation, randomized search and grid search to make this process easier. In Logistic Regression, the only hyperparameter C penalizes a large number of features, reduces model complexity and prevents overfitting. We use randomized search to find the best C because C has a large search space and randomized search saves computing time. With Random Forest, there are many hyperparameters available for tuning, but we use most of the default settings in sklearn and only focus on a few. After creating a parameter grid, we use grid search to find the best parameters combinations.

● 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 is 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 abituraily high, but the precision will decrease. We use below metrics to measure model performances.

a. Confusion matrix
b. ROC_AUC curve
c. Precision_recall curve

● Feature Importance:

By plotting the feature importance on tuned Random Forest 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.

6. Conclusion:

More credit card default for limit balance about 10000. It might mean that credit cards might be too easy to be issued for people who have low credit scores. The variance of the default rate for a limit balance over 500,000 NTD is higher than other ranges of limit balance. It is a lower default rate for cardholders who have a higher education level. Moreover, the default rate for clients whose age over 60 was higher than mid-age and young people. The best-fit algorithm for predicting limit balance is the bagging approach. The best-fit algorithm for predicting whether a client default next month is a classification tree.