

Default credit card detection

Sibani choudhury,
Data science trainees,
AlmaBetter, Bangalore

Abstract:

This project presents and discusses data-driven predictive models for predicting the defaulters among the credit card users. Data used include details like limit balance, age, sex, amount of bill statement, repayment status and amount of previous payment. The paper discusses which variables are the strongest predictors of default, and to make predictions on

[K-S chart](#) to evaluate which customers will default on their credit card payments

- 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).

which customers are likely to default.

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

- 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.

2. 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. 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. 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.

3. Steps involved:

EXPLORATORY DATA ANALYSIS

Data Preparation: This is the stage of the project where one decides on the data that one will use for analysis. The criteria used to make this decision includes the relevance of the data to the data mining goals, the quality of the data, and technical constraints such as limits on data volume or data types. For data preparation, number of outliers were checked and observed to have more number of outliers in some features.

LIMIT_BAL

LIMIT_BAL states the amount of given credit. This is the maximum amount a customer can spend with their credit card in a single month. The amount of balance limit is dependent on the bank's own screening processes and other unknown factors.

AGE

This is the age of the customer which is stated in years. The mean and median of the age are 35 and 34 respectively. This is numerical column. There are no missing values in this column. There are some outliers in AGE column. The decline in number of customers starts from about 30 years among the non-defaulting group, while the number of customers of different ages stays much more constant from 25 to around 40 years. This indicates that likelihood of default among men grows with age. In box plot we can see many outliers which are considered here as extreme values.

SEX

This variable can obtain a value of 1 for male and 2 for female. In this study, sex and gender are used interchangeably to intend the same thing.

EDUCATION

The education level of a customer is represented as one of four values: 1 = Graduate school, 2 = University, 3 = High school, 4 = Other. For the purpose of analysing customer groups, this is assumed to indicate the highest level of education completed.

Bill_AMT

Amount of bill statement is recorded in this variable. It is represented in the data as 6 columns, one for each month. Data collected from 6 months, April to September.

Modelling: As the first step in modelling, one selects the actual modelling technique that one will be using. Although one may have already selected a tool during the business understanding phase, at this stage one may be selecting some specific modelling technique. If multiple techniques can also be applied. After treating the data, we performed various models on our data such as logistic regression, XG boost, svc, Random forest etc. Hyper parameter tuning was also performed.

This analysis uses 3 classification models - Logistic Regression, Random Forest and XG Boost. Since Random Forest and XG Boost 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 hyper parameters 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 model. After SMOTE sampling, the dataset has equal size 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 improved significantly with all models after over sampling with SMOTE. This proves SMOTE is an effective method in sampling imbalanced dataset.

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 arbitrarily 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

Compare within the 3 models .

Logistic Regression has the highest recall but also the lowest precision. Random Forest outperforms Logistic Regression and XG Boost if measured on their F1 scores, which is the balance between precision and recall. XG Boost has a decent performance but it takes the most time to tune the model.

Conclusions

Based on the exploratory data analysis, we discover that human characteristics are not the most important predictors of default, the

payment status of the most 2 months and credit limit

From the modeling, we are able to classify default risk with accessible customer data and find a decent model. Using a Logistic Regression classifier, we can predict with 73% accuracy, whether a customer is likely to default next month. Using a Random Forest classifier, we can predict with 92% accuracy, whether a customer is likely to default next month. Using a XG BOOST classifier, we can predict with 88% accuracy, whether a

customer is likely to default next month.

If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.

References-

1. Machine Learning Mastery
2. Geeks for Geeks
3. Analytics Vidhya