Predicting Credit Card Default

Ying Guo

January 20, 2019

1 Introduction

According the S&P/Experian Consumer Credit Default Index, consumer credit default indices show higher default rates for all loan types in December 2018. Residents of big cities, including Chicago, New York, Los Angeles, Dallas or Miami, have the highest default rate. For banks and other financial institute, it is important to be able to predict which customers is more likely to default on their debts. So they can avoid paying the debts for their customers in the end. Predicting credit card default is a very important topic for financial institutes, as it is similar to the problem of financial fraud detection.

Financial fraud detection is a very important topic. In Oxford English Dictionary, it defined fraud as "wrongful or criminal deception intended to result in financial or personal gain." In Economics, financial fraud is an increasingly serious problem. Enron, Cendant, and WorldCom are examples of large companies torn apart by financial fraud and scandal. The most common financial fraud includes bank fraud, insurance fraud, securities, and commodities fraud, and corporate fraud, etc. The related research paper shows that data mining techniques have been applied most extensively to the detection of insurance fraud, then corporate fraud and credit card fraud.

Ngai et al. (2011) summaries the data mining techniques in financial fraud detection and reports six main classes of methods: classification, regression, clustering, prediction, outlier detection, and visualization. The primary data mining techniques used for financial fraud detection are logistic regression, neural networks, Naive Bayes, decision trees. In this project, I would like to use most of the methods mentioned above to predict on the credit card default rate on the dataset.

2 Dataset

This data data set is about default of credit card clients. The original research aimed to compare the predictive accuracy of the probability of default among six data mining methods, including K-nearest Neighbors (KNN), Naive Bayes (NB), Neural Networks (NNs), and classification trees.

Number of observations: there are 30000 in total.

Attributes There are 23 different 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.
- 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.

The target variable is default payment (yes = 1, and no = 0).

3 Plan

I plan to carry out the following steps.

1. Exploratory Data Analysis
2. Data preprocessing and data cleaning
3. Feature engineering
4. Model training
    - KNN
    - Naive Bayes
    - Random Forest
    - Gradient Boosting Model
    - Neural Network
    - Support Vector Machine
    - Linear Discriminant Analysis + Support Vector Machine
    - Clustering
    - Outlier Detection