Predicting Credit Card Default

Ying Guo

January 21, 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:

- LIMIT_BAL: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- SEX: Gender (1 = male, 2 = female).
- EDUCATION: Education (1 = graduate school, 2 = university, 3 = high school, 4 = others).
- MARRIAGE: Marital status (1 = married, 2 = single, 3 = others).
- AGE: Age (year).
- PAY_0 _ PAY_6: History of past payment.
  - PAY_0 = the repayment status in September, 2005
  - PAY_2 = the repayment status in August, 2005
  - PAY_3 = the repayment status in July, 2005
  - PAY_4 = the repayment status in June, 2005
  - PAY_5 = the repayment status in May, 2005
  - PAY _6= 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
  - 3 = payment delay for two months
  - 4 = payment delay for two months
  - 5 = payment delay for eight months
  - 6 = payment delay for two months
  - 7 = payment delay for two months
  - 8 = payment delay for eight months
  - 9 = payment delay for nine months and above
- BILL_AMT1 - BILL_AMT6: Amount of bill statement (NT dollar).
  - BILL_AMT1 = amount of bill statement in September, 2005
  - BILL_AMT2 = amount of bill statement in August, 2005
  - BILL_AMT3 = amount of bill statement in July, 2005
  - BILL_AMT4 = amount of bill statement in June, 2005
  - BILL_AMT5 = amount of bill statement in May, 2005

  - BILL_AMT6 = amount of bill statement in April, 2005

- PAY_AMT1 - PAY_AMT6: Amount of previous payment (NT dollar).
  - PAY_AMT1 = amount paid in September, 2005
  - PAY_AMT2 = amount paid in August, 2005
  - PAY_AMT3 = amount paid in July, 2005
  - PAY_AMT4 = amount paid in June, 2005
  - PAY_AMT5 = amount paid in May, 2005
  - PAY_AMT6 = amount paid in April, 2005

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

3 Data Wrangling

3.1 Data Cleaning

Here are a few steps I performed for data checking.

- In the project, I first check the shape of the data. There are 30000 records and 24 different variables. I visually inspect the first five and last five rows of the data.
- I check the column names. I rename some of the variables to make it easy to understand.
- If I have string categorical variable, I would check if they use a consistent way to represent things. For example, "Friday“, "Fri“, and "fri“ could all be there for Friday. "United States”, "USA“, and "US“ could be the same thing.
- I check the data type of each column and the data type matches the content.
- I count the frequency for each categorical variables. I find a few strange things:
  - For Education, there are three undocumented categories (0, 5, 6). A future investigation could be done if I have access to people who created this data. Since I do not have that access, for now, I will group 0, 5, 6 to others (which is 4).
  - For Marital status, class 0 is not documented. For now, I group 0 to others (which is 3).

3.2 Missing Values

I also check if there are any missing values. I did not find any missing values in the current dataset. However, if I have missing values, I would need to investigate the nature of the missing values.

- If the records with missing values are only a small fraction of the dataset, we can remove all the cases with missing values.

- If we are dealing with a small data set, we need to see which imputation methods we could use for the data set. If the missing value is numerical and missing at random, we can use mean if we are going to fit the data with linear regression since mean is an unbiased estimator. If we want a simple measure that is robust to outliers, we could use the median.
- If the missing value is categorical, one could impute value ?missing? to make missing values as a category. We can also impute with the most frequent class.
- Another good choice for imputation is to use models to predict the missing values. A good example is MICE (multivariate Imputation via Chained Equations). In Python, FancyImpute has a function IterativeImputer, which is similar to MICE but it returns a single imputation. We can also use interpolation with linear regression, random forest, and KNN.

3.3 Outliers

I applied median-absolute-deviation (MAD) based outlier detection for all numerical features. I used a threshold of 3.5. A data point with Z score whose absolute value larger than 3.5 is labeled as an outlier. In this case, I found 10791 cases with "outliers“, which is almost 1/3 of the data. I also used box-plot to check for outliers.

However, outliers do not mean errors. It could be a valid point as well. By checking the data for balance limit records, they look valid points. Since the ultimate goal of this project is to predict whether people default, it is better to include these extreme values since it truly exists in the real world. So in this dataset, I keep all the extreme values in the data. I will also train the final model on the no-outlier data to compare results.

4 Exploratory Data Analysis

In this section, I want to find out what variables impact whether a person defaults in the next month significantly. I create a new age variable, which is a categorical variable and each value is labeled based on its group. I also create a new variable which is the proportion of bill amount compared to the limit of balance, in other words, the percentage of people using their credit limit each month. In the dataset, I have the following categorical variables, including gender, education level, marital status, age, percentage of credit usage. Here are the graphs showing their relationship with whether a person defaults next month.

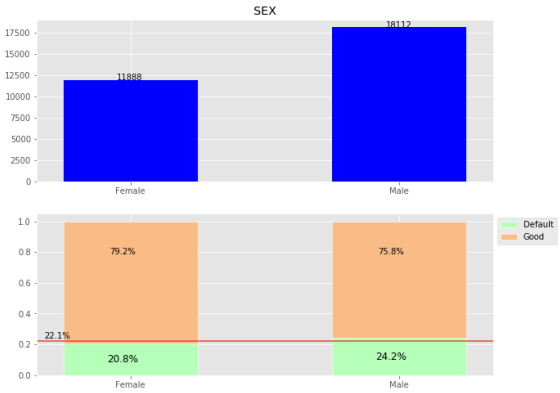

Fig.1 Gender vs. Default Rate

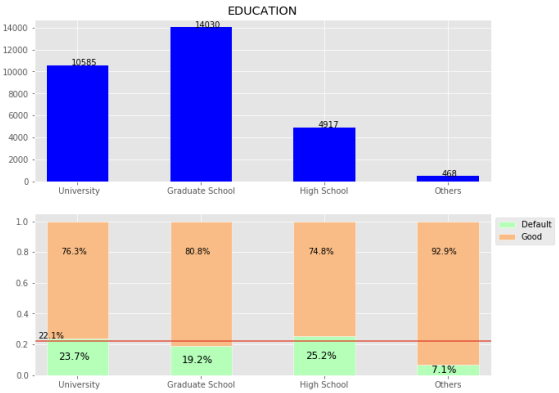

Fig.2 Education vs. Default Rate

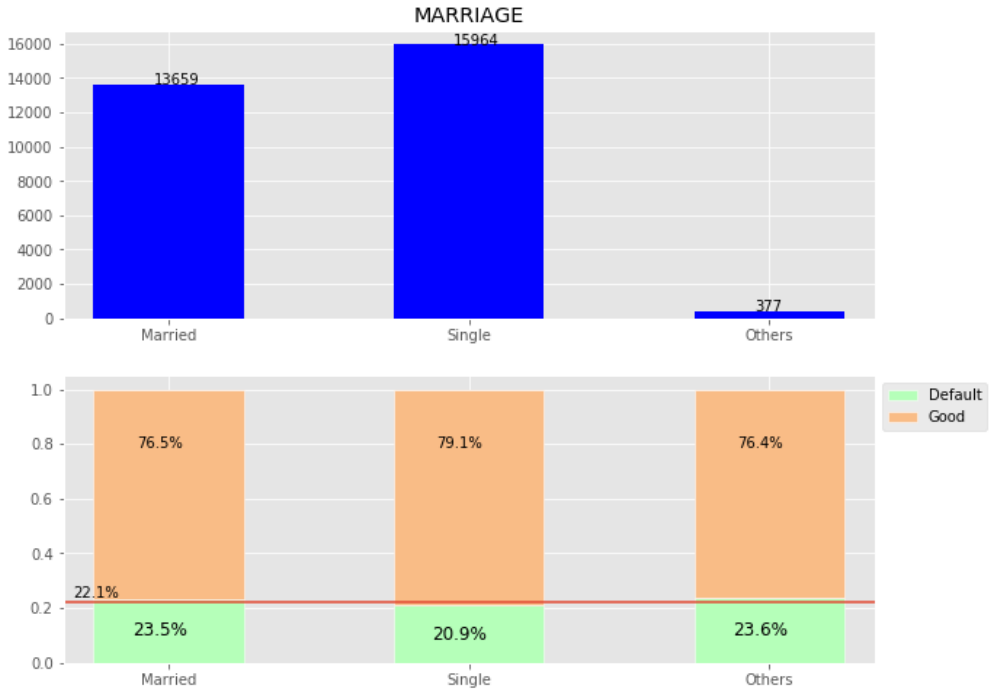

Fig. 3 Marital Status vs. Default Rate

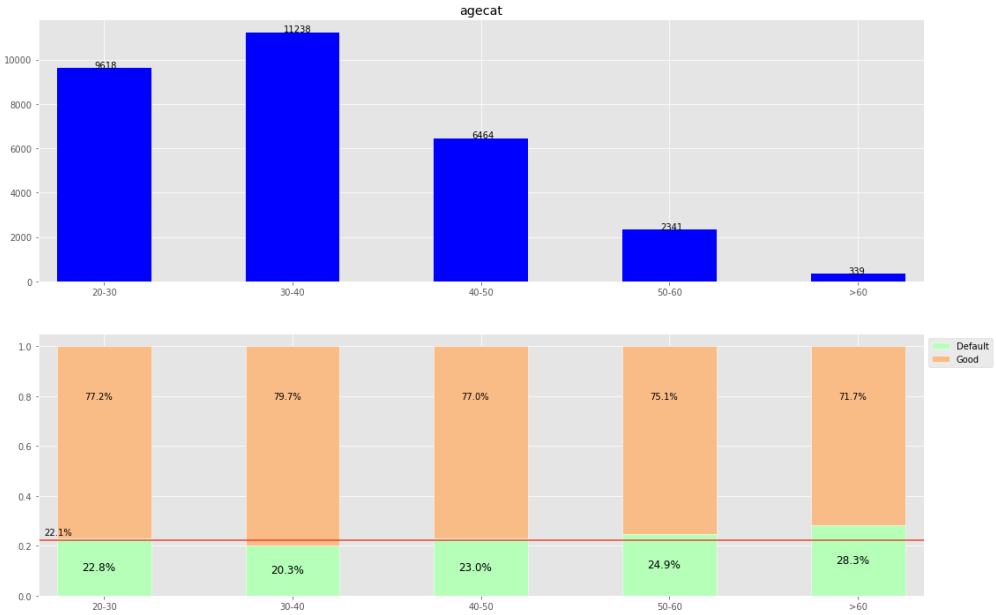

Fig. 4 Age vs. Default Rate

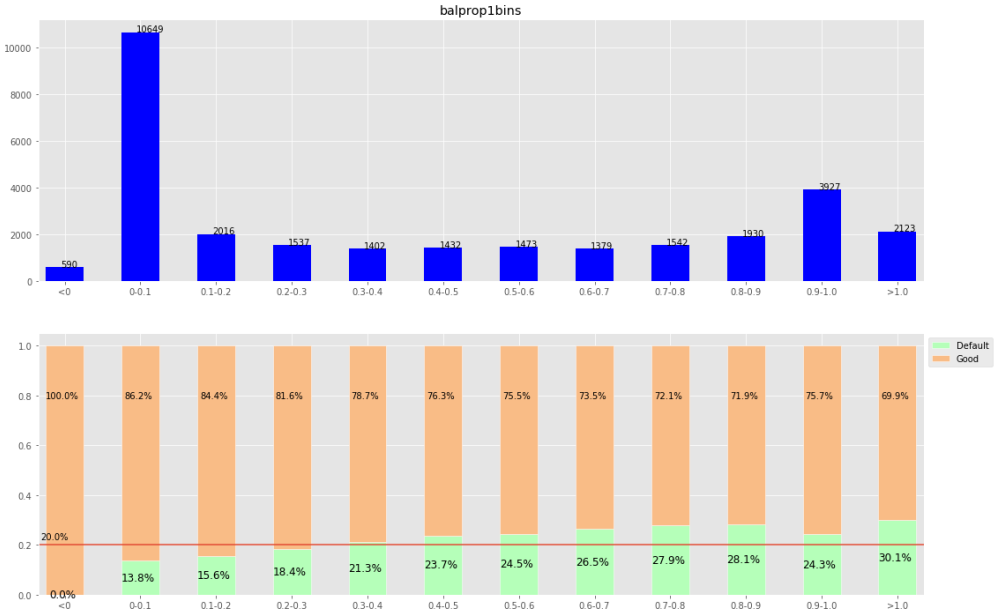

Fig. 5 Proportion of usage of balance limit vs. Default Rate

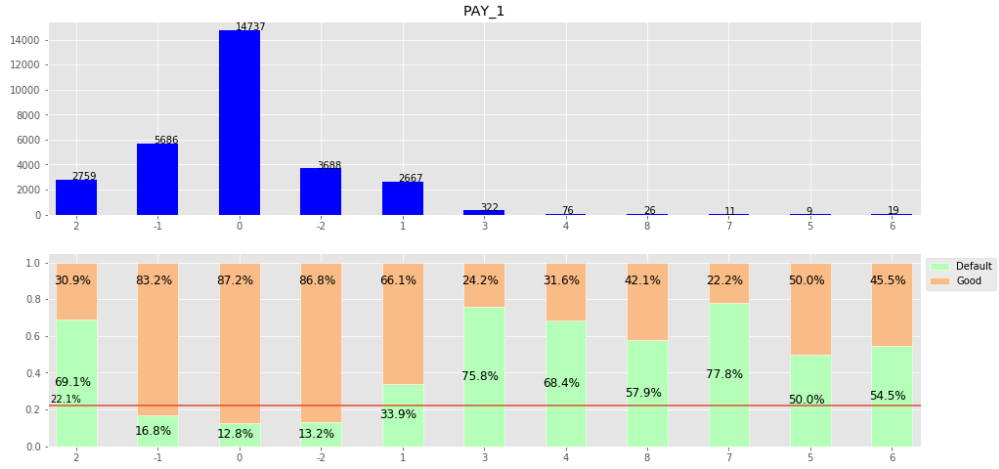

Fig. 6 Repayment Status in September, 2015 vs. Default Rate

All of the categorical variables are related to default. However, some variables have more than 2 categories. I want to find out which category is significant different from the other.

- Graduate School vs. (University and High school)
- Married vs. Single
- 20-30 Age Group vs. 30-40 Age Group
- 30-40 Age Group vs. $>$ 40 Age Group
- Normal Usage vs. Overdrawn

For the two age group tests, I use Bonferroni correction, since I am doing two separate tests on the same data. The new critical alpha is $1 - (1 - \frac{.05}{2})^2 = 0.049$. I carried χ^2 tests, t-tests, and permutation tests. All of the tests are significant at a 0.05 significant level.

The overall message is

- Male is more likely to default than females
- The highest education is university or lower is more likely to default
- Married people is more likely to default than single people
- People younger than 30 or older than 40 is more likely to default, especially for people older than 40
- People who defaulted the previous month is more likely to default the next month
- People who overdraws their balance is more likely to default.
- When a person uses less than 30% of their limit, they are less likely to default.

There are some strange things about this dataset, including undocumented categories. How they label repayment status remains a puzzle. It seems like there are some obvious errors in terms labeling the repayment status. Future investigation should be made about how they label repayment status. What is more, more information could be collected, including the income level, the location, and more.

What is more, I used all the variables above to build a logistic regression using statsmodel and sklearn in Python. The conclusion is the same as above. All parameters are significant at 0.05. The parameters estimated from logistic regression tells us how a unit increase or decrease in a variable affects the odds of default next month. For example, we expect the odds of default to decrease by 81% if the customer is female. We expect the odds of default to increase by 126% if the customer's highest education is university, compared to graduate school, while holding other things constant. However, as a machine learning algorithm, the logistic model have a zero recall score for default. In this dataset, I care about recall more than precision, as in fraud analysis, we want to identify as many frauds as possible. Since frauds costs more, I care more about false negative than false positive. I will add more variable into the model and try other machine learning algorithms.

5 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