Credit Card Payment Default Analysis Proposal

By Teresa Nan

Problems to solve:

Credit card companies are in fierce competition in attracting customers. The emerging mobile wallet makes this competition even hotter. However, payment default risk associated with the credit card industry is an important consideration in generating revenue and reducing loss. “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.”[1]

Researches have shown that certain age groups and demographic groups are more likely to have credit card default payment by analyzing data from the US[2].Further study to verify these findings using a dataset outside of the US would be beneficial. Do younger customers tend to miss bills more often? What demographic factors would be strong indicators of credit card default payment?

Potential clients:

A proper algorithm will help credit card issuers predict which group of customers have relatively higher risk of default payment. This analysis will potentially help credit card companies, loan lenders, and banks make informed decisions when targeting customers, promoting, and issuing credit cards, therefore to mitigate payment default risk.

Dataset acquisition:

This project will be using a public dataset from Kaggle:
https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
The original dataset can be found here at the UCI Machine Learning Repository. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

---

[1] https://www.creditcards.com/credit-card-news/credit-card-delinquency-statistics-1276.php
[2] TransUnion Industry Insights Report Q1 2018

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns. There is a clear explanation on these 25 variables. Overall, the cleanness and usability of this dataset are high.

Tentative approaches:

The tentative plan of analyzing the dataset is as follows: (Note: the approaches may change.)

- Exploratory data analysis
    - Diagnose data for cleaning.
    - Use data visualization to understand data and find the correlations between variables. The questions regarding which factors would be strong indicators of credit card default payment will get answered.
- Data preparation and transformation
    - Prepare data to build a model
- Data modelling
    - Predictive data analysis (logistic regression, SVM, random forests, etc)
    - Tune each model
- Model evaluation and model presentation
    - Determine metrics of model evaluation
    - Compare different models
    - Explain how the model works

Project deliverables:

The final deliverable will be a Jupyter Notebook uploaded to Github which includes codes, data visualization and a slide deck which serves as a project report.