Credit Card Default Research

by Sumaira Afzal, Viraja Ketkar, Murlidhar Loka, Vadim Spirkov

Abstract Credit card default might very well be a life altering event. It happens when a client have become severely delinquent on his/her credit card payment. It's a serious credit card status that not only affects person's standing with that credit card issuer, but also individual's credit standing in general and his/her ability to get approved for credit cards, loans, and other credit-based services. This research will make yet another attempt to predict if a client goint to default on the next payment. Employing verious machine learning technique we also will make an attemt to estime the amount a client would be able to pay when the bill comes. The authors of this study will try to discover who is more likely to default on the payment.

```
#> Warning: package 'ggplot2' was built under R version 3.5.3
```

```
#> Warning: package 'corrplot' was built under R version 3.5.3
```

Background

Overdepandance on credit card debt has been an ongoing theme in many countries around the word. For example US consumers started 2018 owing more than $1 trillion in credit card debt (Ref: Comoreanu) It is projected that by the end of 2019 US consumers will increase their collective debt by another 60 billion dollars. Unfortunately many consumers overestimate their ability to pay the debt on time, or the unforeseen circumstances and luck of savings make people default on their payments. This is the least desirable outcome for all parties. Unpaid debt leads, in most cases, to default on the whole outstanding balance causing financial loss for the credit institutions. Majority of the clients go through tremendous emotional and financial stress, risking their credibility. The financial institution make significant efforts to evaluate the prospective client ability to sustain the debt and pay in time to avoid the credit default.

Objective

This study pursues a few goals. First of all employing the client personal characteristics and the last six month payment history we would like to predict ax accurate as possible if the client makes the next month payment or defaults. We will employ a few supervised learning models to attack the problem.

Another objective is to understand which features of the data set have the most impact on the next payment success/ failure.

We are also motivated to unearth, if possible, any trend that might shed light on what make people to default on the payment. And lastly the authors of this study will try to estimate how much a client could pay when the next bill comes

Data Analysis

This research employs the data set sourced from UCI Machine Learning Repository. This real-life data comprises 30000 observations of the credit card payment history of Taiwan consumers.

Data Dictionary

Column Name	Column Description
ID	Customer ID
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)

Column Name	Column Description
PAY_1	PAY_1 - PAY_6 are payment statuses over a course of the last six months, where -1 = pay duly. Positive numbrs denote payment delay in months. PAY_0 - Payment status in September
PAY_2	Payment status in August
PAY_3	Payment status in July
PAY_4	Payment status in June
PAY_5	Payment status in May
PAY_6	Payment status in April
BILL_AMT1	BILL_AMT1 - BILL_AMT6 are bill amounts (NT dollar) from April till September. BILL_AMT1: September bill
BILL_AMT2	August bill
BILL_AMT3	July bill
BILL_AMT4	June bill
BILL_AMT5	May bill
BILL_AMT6	April bill
PAY_AMT1	Amount of previous payment (NT dollar). PAY_AMT1: paid in September (August bill)
PAY_AMT2	Amount paid in August (July bill)
PAY_AMT3	Amount paid in July (June bill)
PAY_AMT4	Amount paid in June (May bill)
PAY_AMT5	Amount paid in May (April bill)
PAY_AMT6	Amount paid in April (March bill)
DEFAULT_NEXT_MONTH	**Target label that denotes whether the client paid next bill or did not (Yes = 1, No = 0)**

Data Exploration

```
original = read.csv("../data/default-cc.csv", header = T,
                na.strings = c("NA","","#NA"),sep=",")
str(original)
```

```
'data.frame':   30000 obs. of  25 variables:
 $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL         : int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
 $ SEX               : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION         : int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE          : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE               : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_1             : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2             : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3             : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4             : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5             : int  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6             : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1         : int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
 $ BILL_AMT2         : int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
 $ BILL_AMT3         : int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
 $ BILL_AMT4         : int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
 $ BILL_AMT5         : int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
 $ BILL_AMT6         : int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
 $ PAY_AMT1          : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
 $ PAY_AMT2          : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
 $ PAY_AMT3          : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4          : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
 $ PAY_AMT5          : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
 $ PAY_AMT6          : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
 $ DEFAULT_NEXT_MONTH: int  1 1 0 0 0 0 0 0 0 0 ...
```

Data correlation and other observations

Takeaways from Data Exploration Excersize

Data Preparation

Data Imputing

Modeling and Evalutation

Feature Selection

Data Upsampling

Decision Tree Model

Naive Bayes Model

Random Forest Model

Logistic Regression Model

Model Comparison

AUC - ROC perfomance

Model interpretibility

Data Preparation

Verdict Despite sensitivity to data quality Logistic Regression outperforms other models in all other major categories. This is our choice!

Model Deployment

Conclusion

Bibliography

A. Comoreanu. Credit card debt study. trends and insights. URL https://wallethub.com/edu/credit-card-debt-study/24400/. [p1]

Note from the Authors

This file was generated using *The R Journal* style article template, additional information on how to prepare articles for submission is here - Instructions for Authors. The article itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts.

Sumaira Afzal
York University School of Continuing Studies

Viraja Ketkar
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies

Vadim Spirkov
York University School of Continuing Studies