**Problem#1: NPL Flag Prediction from Credit Card Information**

Contents

[Data: 2](#_Toc489213759)

[Observations: 2](#_Toc489213760)

[Program & Detail: 2](#_Toc489213761)

[Solution: 3](#_Toc489213762)

[Data Preparation: 3](#_Toc489213763)

[Tune and Run the model: 3](#_Toc489213764)

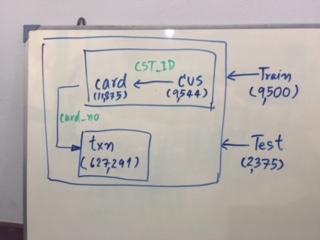
[Prediction and Output transformation: 3](#_Toc489213765)

[Improvement Plan: 4](#_Toc489213766)

[Reference: 4](#_Toc489213767)

# Data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **File Name** | **Table name** | **Size (rows)** | **Filed Number** |
| Card Data | tj\_01\_creditcard\_card.csv | cc | 11,875 | 11 |
| Customer Data | tj\_01\_creditcard\_customer.csv | cus | 9,544 | 6 |
| CC Transaction Data | tj\_01\_creditcard\_transaction.csv | txn | 627,291 | 7 |
| training data | tj\_01\_training.csv | train.org | 9,500 | 2 |
| testing data | tj\_01\_test.csv | test.org | 2375 | 1 |



# Observations:

* File name: tj\_01\_creditcard\_card.csv
  + Column exp\_dt seems to be MMYY format. I have to convert this column to be date format.
* File name: tj\_01\_creditcard\_customer.csv
  + There are some accounts show age younger than 15 years. According to research in credit card application's conditions, the minimum age that is able to apply for additional card is 15years and main card is 20years.
* All data files:
  + There is no duplicated data found.

# Program & Detail:

|  |
| --- |
| platform x86\_64-w64-mingw32  arch x86\_64  os mingw32  system x86\_64, mingw32  status  major 3  minor 4.1  year 2017  month 06  day 30  svn rev 72865  language R  version.string R version 3.4.1 (2017-06-30)  nickname Single Candle |

# Solution:

## Data Preparation:

* Add prefix to column name for table cc and cus.
* Train table – add column name.
* Test table – add column name.
* Convert Expired Date data from ‘MMYY’ to be date format.
  + I use the 1st for specifying the date. This date can be used for all months. If I use the end of month, it can be 28,29,30 or 31 depend on month and year.
* Create new column from open\_dt, pos\_dt and txn\_date by finding difftime from today (07/30/2017)
* Exclude all customer records that age is less than 15.
* Combine 3 tables together – cus, cc and txn
  + Merge cus with cc table first and then merge it to txn table.
* Setup station of columns setting to be easy when tuning the model.
* Update all character data to be numeric. (XGBoost can accepted only numeric.)
* Update all NULL to be zero.
* Update all logical data (TRUE,FALSE) to be (1,0).
* Merge train and test data with table that included all data (matched by card\_no field).
* Convert data to xgboost format.

## Tune and Run the model:

* Set up Param as a list of parameters set up for xgboost.
* Set seed to ensure that the random number generated can be reproduced.
* Set up xgb.train such as data and nrounds.

## Prediction and Output transformation:

* Do prediction on the test card\_no. The output will be probability of NPL. So, I have to convert this data to be NPL flag.
* To convert probability to be NPL flag.
  + Setup cutoff number using mean of all probability data.
  + Convert each record to be logical.
    - If probability > = cutoff number, then logical =1
    - If probability < cutoff number, then logical =0
  + Assign NPL flag for each card\_no using mode function.
* Rearrange the NPL flag order by original test file before exporting.

# Improvement Plan:

* N/A

# Reference:

https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/