**Problem#3: Churn Detection with Account’s Activities and Transactions**

Contents

[Data: 2](#_Toc489215853)

[Observations: 2](#_Toc489215854)

[Program & Detail: 2](#_Toc489215855)

[Solution: 3](#_Toc489215856)

[Data Preparation: 3](#_Toc489215857)

[Tune and Run the model: 3](#_Toc489215858)

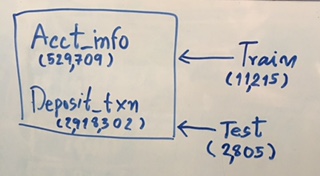
[Prediction and Output transformation: 3](#_Toc489215859)

[Improvement Plan: 3](#_Toc489215860)

[Reference: 3](#_Toc489215861)

# Data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **File Name** | **Table name** | **Size (rows)** | **Filed Number** |
| Account info Data | tj\_03\_account\_info.csv | acc | 529,709 | 9 |
| Transaction Data | tj\_03\_deposit\_txn.csv | txn | 2,918,302 | 7 |
| training data | tj\_03\_training.csv | train.org | 11,215 | 2 |
| testing data | tj\_03\_test.csv | test.org | 2,805 | 1 |



# Observations:

* File name: tj\_03\_account\_info.csv
  + Each account has more than 1 records. I have to filter only the latest updated record for each account.
* 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 acc.
* Train table – add column name.
* Test table – add column name.
* Table acc - Extract only maximum transaction date from each account number.
* Create new column from info\_txn\_dt, info\_open\_dt, info\_last\_acctive\_dt and txn\_date by finding difftime in days from today (07/30/2017).
* Merge 2 tables – txn and acc
* Extract more 2 columns from txn\_type – is\_DR, is\_CR. This technique is help to increase more accuracy of the XGBoost model.
* 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 account\_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 account\_no. The output will be probability of Churn. So, I have to convert this data to be Churn flag.
* To convert probability to be Churn 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 account \_no using mode function.
* Rearrange the Churn flag order by original test file before exporting.

# Improvement Plan:

* N/A

# Reference:

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