RG Selector optimization

Country: INDIA

Product: POS

Date of preparation: 8-Nov-18

# Background

Set up new RG Selector for POS CD (Consumer durables), TW (Two-wheelers) and HA (Home appliances). There is a new scorecard which was developed by HQ Development center for all the above mentioned products in June-October 2018[[1]](#footnote-1).

The final scorecard is based on GM (general model) method and contains multiple pillars. Specifically, two of them Product pillar and Aldi pillar are subject to analysis whether they can be omitted and used in RGS instead.

# Proposed method, data

* Data is prepared as all new clients who entered Pre RGS between 02/2018 and 04/2018 and belong to the product groups in scope of analysis
* Dimensions are analyzed in terms of prediction potential (on top of information contained in score)
* Additional dimensions needed from business point of view are added
* Cutoffs for respective score version are estimated as the quantiles of score rejecting the same percentage of clients on each row of new RG Selector
* Cutoffs are finally expertly tuned to get lower risk on “big” rows, reflect strategy of AR in Aldi/non-aldi and specific product groups

# Codes/Scripts

Due to legacy reasons, multiple software was used in the process of RGS design.

## Data preparation

All relevant dimensions from IN risk datamart (ap\_uwi.daily\_data\_all) were used for analysis of their statistical significance. Data from 02/2018-04/2018 with target FSTQPD120 were used for estimation of dimensions.



The data need to contain:

* Applications which enter the Pre strategy (enter the RGS)
* Observable risk
* Variables which can add value to score
* New score for all the cases

For setup of cutoffs, we need

* Same as above
* Also rejected applications, typically on shorter target

## Grouping + WOE

Data sample was split to training/validation sample for all following statistical estimations, enriched by grouping (fine setting of groups as they can be merged later in RGS design). Grouping of GSPN was done based on expert input of GSPN similarity.



GSPN model will be redeveloped by the end of 2018, i.e. RGS will be adjusted accordingly.

Additionally, WOE was calculated for each of group. Both version of score described above were included in data.

For grouping and woe calculation, first section of PSW[[2]](#footnote-2) was used (csv files serve as input = list of variables to be binned/grouped). Result is exported (all even non observable cases) to csv to be used in R. Filtering is done in R afterwards.



## Dimensions selection

Selection of dimensions was calculated in R. Below the codes are in 2 versions, version \_npp is for score with No Product and aldi Pillars.

* At first calculate the added value of each variable on top of information contained in score. This is done in function *compute\_Gini\_Extra\_Var.* The function is called repeatedly by adding variables one by one.
* Next, the id of each group is added to data. For numerical variables by ordering, for categorical randomly. It is done by codes like
  + *pom <- data %>% group\_by(ACC\_EXTRA\_DP\_WOE) %>% summarise(avg = mean(ACC\_EXTRA\_DP)) %>% arrange(avg) %>% mutate(GRP\_ACC\_EXTRA\_DP = row\_number()) %>% ungroup() %>% select(-avg); data <- left\_join(data, pom)*
  + *pom <- data %>% distinct(PRODUCT\_GOODS\_NAME\_WOE) %>% mutate(GRP\_PRODUCT\_GOODS\_NAME = row\_number()) ; data <- left\_join(data, pom)*
* For GSPN it is manually corrected to have group names corresponding to content.
* Dimension analysis is calculated for each of the variables in these groups  
  
* *computeGiniLessVar –* calculates the gini of each “removal of variable” from selected dimensions
* Most important section is marked as *Selection of final RG* in the code, here we define 3 functions: dimension.agg, join.bins.row and new.bins and start grouping of rows. Dimensions are defined in object *dimensions* in order which is then used for adding one by one. The first calling exports the first 2 dimensions to dimension\_agg2.csv where user selects rows to merge for second dimension (first dimension cannot be merged anymore. If you need to merge it, redefine it in data)
* After setting the first column in the dimension\_agg2.csv file (according to desired merging), the file is loaded back, added next dimension and next round exported dimension\_agg3.csv
* The output of this process is on particular sheets in



* Whole process of this section is in following files



## Simulation of row ID of new RGS

* Data is created for both approved and rejected cases by scripts  
  
* RG Simulator is used. In this code first part prepares tables to load RGS in for of “Blaze table shape”. This was created in previous step  
  
* Then RGS is loaded to these tables. FK\_RGS\_RULES\_all and FK\_RGS\_RULES\_npp for the version with score with all pillars and no product/aldi pillar npp.
* Section *A. BLAZE\_OPERATOR* defines functions for assigning rows. This part of code needs to be run separately in SQL to get loaded. Then next part CREATE TABLE FK\_app\_RGS\_row\_exact is run as a separate part where IDs of new RG selector are assigned to the data.

## Cutoff setting

* Cuttoffs are set based on approved and rejected cases. Data from previous step are exported by  
   
* The exported data is loaded to R



* Function set.cutoffs sets the cutoff (quantile of score) by approval rate on each row of new RGS. AR as kept same on each row by observing number of rejected/approved cases. August 2018 data and FPD30 target was used for tuning of cutoffs.
* Results are exported to cutoffs\_proposition(\_npp).csv, in this file desired AR can be set on each row and defined in needed.AR

   
- this last step was finally not done in R but in excel simulator which is used by local IN risk tem for standard cut offs setup

# Results

Two versions of RGS are considered as a solution, version with all pillars in scorecard and version without the 2 pillars which use the additional information (from pillars) in RGS as dimensions.  


# Recommendations

Following points should be considered in next redesign

1. Variable insurance (FLAG\_INSURANCE or more detailed PRODUCT\_INSURANCE) was significant in the analysis, however not used as the insurance data was fully in line with the expectations (penetration). Consider inclusion after proper analysis.
2. GSPN model is correlated with Aldi flag variable. Both are considered in the model as final dimensions as they are needed for managing the risk (and to fulfill sales agreements). However Aldi is used in scorecard already and some GSPN codes contain mainly aldi clients. It would be better to use only one of the dimensions due to simpler management and less RGS rows.

1. Documentation is not attached here to save space, it is stored on HQ server \\proliant10.office.hci\HCI\_A\SCORING\1 HQ Development center\1 Scorecards\2018\06 IN POS CD + TW [↑](#footnote-ref-1)
2. https://wiki.homecredit.net/confluence/display/SW/Scoring+Workflow+Home [↑](#footnote-ref-2)