Homework Assignment 3

IST 707 10/26/2021

Dan Burke

2021

# Overall Analysis Process and Findings

## Preprocessing and Preparation for Association Rule Mining

Steps Taken:

1. Import “arules” & “arulesviz” libraries
2. Import “bankdata\_csv\_all.csv” as a data frame object “bankData”
3. Create new data frame object, with “id” attribute/column removed.
   1. This column is not required during this analysis as the data will be used to create a transaction data object.
4. Verify the structure of the imported data in order to check for needed type conversions.
5. Visualize data to understand distribution of numerical attributes. Specifically:
   1. Income
   2. Quantity of Children (“Range()” used as well)
6. Discretize age and income attributes, utilizing 3 breaks labeled “low”, “medium” and “high”. Income has the potential to be expanded into a larger quantity of breaks (levels) however throughout this analysis it was not found to be needed.
   1. Generate Boxplot of age and income to quickly understand their distributions.
   2. Convert attributes of type “char” to factor
      1. Sex
      2. Region
      3. Married
      4. Children
         1. Labels: “None”, “One”,”Two”,”Three”
      5. Car
      6. Save Account
      7. Current Account
      8. Mortgage
      9. Pep
   3. Verify structure of data frame (“str()”) to ensure types have been accurately modified
7. Create new transaction data object (“trans\_bankData”) for rule generation.
   1. As a record data object rules may be generated the same, however
8. Generate and visualize (via histogram) a “frequent items” data frame in order to quickly understand the population’s items’ frequency and distribution.

## Rules Generation First Pass

The first pass of rules generation included the following parameters utilizing the “Apriori” algorithm:

**rules\_transactions <- apriori(trans\_bankData,**

**parameter = list(support=0.2,confidence=0.7))**

Support of “0.2” and confidence of “0.7” was utilized as a baseline to better explore the dataset. These parameters generate generated a multitude of rules, however the right-hand side (RHS) was predominantly populated by “current\_act=YES” with approximately two (2) results populating with “save\_act=YES”.

These parameters did not result in any meaningful findings, requiring modification to the support and confidence parameters as well as defining the RHS parameter in a following pass at rules generation.

## Rules Generation Second Pass - Right Hand Side (RHS) defined as "pep=YES"

The second pass of rules generation included the following parameters as well as defining the RHS as “pep=YES”. Support and confidence parameters we lowered in order to produce a left-hand side (LHS) that was not predominantly populated with RHS of “current\_act=YES”.

**rules <- apriori(data = trans\_bankData,**

**parameter = list(supp = 0.1, conf = 0.5),**

**appearance = list(default "lhs", rhs = "pep=YES"),**

**control = list(verbose = F))**

Rules generated were then sorted by lift (decreasing). Lift is most practically defined as “normalized confidence” or in other words a conditionally probability. A lift greater than 1 shows a positive correlation, where a lift less than 1 denotes a negative correlation. All lift values generated in this pass ranged from approximately 1.29 to 1.62.

This pass yielded more meaningful results compared to previous passes:

Table

Description automatically generated

The key insight we’ve found here, is each LHS set either contains **“married=NO”** or **“mortgage=NO”** all with lift values above 1 (positive correlation) and confidence greater than 0.59.

For example the following rule: {married=NO, mortgage=NO} which possesses a confidence of 0.707 and lift of 1.54 (positive correlation) indicates that a customer who is not married and does not possess a current mortgage will have enrolled in the PEP program via our marketing mail correspondence 70.7% of the time. This is not to be taken at face value as the support or “how often can you apply this rule” is a relatively low value of 0.15.

The causation for this is unclear, however it leaves an avenue for further investigation, defining the RHS as “pep=YES”.

## Rules Generation Third Pass - Right Hand Side (RHS) defined as "pep=NO"

The third pass of rules generation included the following parameters as well as defining the RHS as “pep=NO”. Support and confidence parameters were again lowered.

**rules <- apriori(data = trans\_bankData,**

**parameter = list(supp = 0.2, conf = 0.5),**

**appearance = list(default = "lhs", rhs = "pep=NO"),**

**control = list(verbose = F))**

The RHS definded as “pep=NO” produced the following top 10 rules, when sorted by lift (descending).

A screenshot of a computer

Description automatically generated with medium confidence

Here we see that the LHS item sets predominantly contain “married=YES”. Though we witness higher support values yet with reduced confidence and lift as compared to defining the RHS as “pep=NO”, it appears that married customers are less likely to have enrolled in PEP, again causation is not clear however, all rules which include “married=YES” possess a lift value greater than 1, indicating a positive correlation.

## 5 particularly intriguing rules from this pass include:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | LHS | RHS | Support | Confidence | Coverage | Lift | Count |
| 8 | {married=YES} | {pep=NO} | 0.433333 | 0.6111111 | 0.6600000 | 1.1247444 | 242 |
| 3 | {sex=FEMALE, married=YES} | {pep=NO} | 0.2116667 | 0.6512821 | 0.3250000 | 1.1986786 | 127 |
| 10 | {sex=FEMALE, save\_act=YES} | {pep=NO} | 0.2033333 | 0.5922330 | 0.3433333 | 1.0899994 | 122 |
| 12 | {sex=FEMALE} | {pep=NO} | 0.2833333 | 0.5666667 | 0.5000000 | 1.0429448 | 170 |
| 24 | {sex=MALE} | {pep=NO} | 0.2600000 | 0.5200000 | 0.5000000 | 0.9570552 | 156 |

# Actionable Recommendation

When evaluating the above rules, it appears that women have a slight positive correlation with declining the PEP program when provided mailed marketing materials, while men possess a slight negative correlation. In addition to the slight difference of normalized estimated conditional probability (lift) between the sexes, it appears married customers possess a more significant positive correlation with declining PEP after receipt of mailed marketing materials.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | LHS | RHS | Support | Confidence | Coverage | Lift | Count |
| 6 | {sex=FEMALE,married=NO} | {pep=YES} | 0.1033333 | 0.5904762 | 0.1750000 | 1.293014 | 62 |

After revisiting the second pass of rules generation (RHS defined as “pep=YES”), there is a rule which included unmarried women enrolling in the PEP (below).

When considering the second and third pass rules it is my recommendation that the differences in PEP enrollment between sexes be disregarded, however our attention and marketing efforts should be focused demographically to capture married customers (both individuals, not only a single household).

This is advantageous to our firm as this market appears to be underpenetrated. Our lack of penetration may be due to our method of PEP marketing, specifically targeting each member of the marriage individually. Married couples often live within the same household and because of this, only one member may inspect that household’s mail. Our mailings may interest one of the partners and not the other and be discarded. It is my recommendation that we explore other marketing avenues where we can individually target each member of the household ensuring all parties of the household are aware of the PEP enrollment opportunity.