2019-0703 IST 707 Data Analytics

Homework Assignment 3 (week 3)

Ryan Timbrook
NetID: RTIMBROO

Assignment Topic: Association Rules
Term: Summer, 2019

Homework Assignment 3 (week 3)

Table of Contents

1 Introduction

1.1 Purpose

Provide insights and suggestions on the kinds of potential buyers the financial institution (client) should target in their new 'Personal Equity Plan' (PEP) product launch using association rule data mining techniques.

1.2 Scope

The marketing department of a financial firm keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product.

Perform Association Rule discovery on the clients banking dataset. Experiment with different parameters and preprocessing that identifies 20-30 strong rules. A strong rule is one that has high lift and confidence while at the same time having relatively good support.

Target rule generation of the PEP class to understand the types of customers who have bought PEP in the past and those who have not. Identify the rules this targeting creates and select the top 5 most 'interesting' rules. Provide the quality measures of these rules along with explaining their patterns. Make recommendations based on the discovery that provides the client with potential business opportunities.

2 Analysis and Models

The Analysis section contains **subsections.**
The second and remaining subsections of Analysis are the model(s).
In some cases, there may only be one model. A model is any method used to analyze the data. Each Assignment specifies which models to use. Always include model details and parameter values when applicable.
***** Have Visualizations throughout the assignment.**
Include measures and comparisons.
Tables are great for comparing.

2.1 About the Data

Contains all the information about the dataset, the variables, the cleaning and prep, checking for an dealing with missing values, checking for and dealing with incorrect values, checking for an dealing with outliers, feature generation, normalization (if needed), etc. In this subsection, you will also "explore" the data.
This means that you write about each variable, **visualize** each variable (as feasible), and talk about what the variable represents. Tables are great for this as well.

The marketing department of a financial firm keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this store of prior experience, the managers decide to use data mining techniques to build customer profile models.

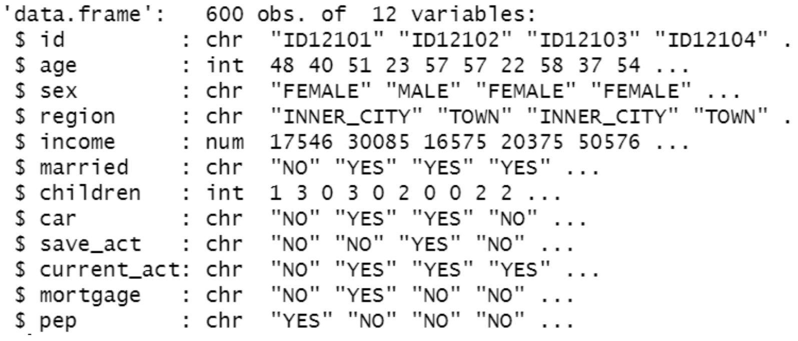
The data contains of a number of the following fields:

id	a unique identification number
age	age of customer in years
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer
married	Is the customer married (YES/NO)
children	number of children
car	Does the customer own a car (YES/NO)
save_act	Does the customer have a saving account (YES/NO)
current_act	Does the customer have a current account (YES/NO)
mortgage	Does the customer have a mortgage (YES/NO)
pep	Did the customer buy a PEP after the last mailing (YES/NO)

Each record is a customer description where the "pep" field indicates whether or not that customer bought a PEP after the last mailing.

2.1.1 *Dataset Info*

The original dataset contains 600 observations with 12 variables in record format. In order to use the Apriori algorithm, this dataset needed to be transformed to a transactional format. See Data Transformation 2.1.3 for details.



```
'data.frame':   600 obs. of  12 variables:
 $ id         : chr  "ID12101" "ID12102" "ID12103" "ID12104" .
 $ age        : int  48 40 51 23 57 57 22 58 37 54 ...
 $ sex        : chr  "FEMALE" "MALE" "FEMALE" "FEMALE" ...
 $ region     : chr  "INNER_CITY" "TOWN" "INNER_CITY" "TOWN" .
 $ income     : num  17546 30085 16575 20375 50576 ...
 $ married    : chr  "NO" "YES" "YES" "YES" ...
 $ children   : int  1 3 0 3 0 2 0 0 2 2 ...
 $ car        : chr  "NO" "YES" "YES" "NO" ...
 $ save_act   : chr  "NO" "NO" "YES" "NO" ...
 $ current_act: chr  "NO" "YES" "YES" "YES" ...
 $ mortgage   : chr  "NO" "YES" "NO" "NO" ...
 $ pep        : chr  "YES" "NO" "NO" "NO" ...
```

2.1.2 *Banking Dataset, Data Exploration & Cleaning*

There were no missing values from this dataset.

2.1.3 *Banking Dataset, Data Transformations*

Preprocessing steps to convert data into transactional format before it can be used in the Apriori Algorithm for Association Rule discovery.
All numeric variables were converted to nominal through discretization or transformation into factors.
The id field was removed from the dataset prior to converting it from a record type to a transaction type using the function: *as(dataset,'transactions')*

Specifically:

Age: was discretized to seven bins with labels
("CHILD","TEENS","TWENTIES","THIRTIES","FORTIES","FIFTIES","SENIORS")

Frequency Table:

CHILD	TEENS	TWENTIES	THIRTIES	FORTIES	FIFTIES	SENIORS
0	37	119	125	128	101	90

Income: was discretized to three equal width bines with labels
('LOW','MID','HIGH')

Frequency Table:

LOW	MID	HIGH
284	235	80

Children: was changed to a factor with four levels
('0', '1', '2', '3')

Frequency Table:

0	1	2	3
263	135	134	68

Married: was changed to a factor with two levels
(‘NO, ‘YES)

Frequency Table:

NO	YES
204	396

Car: was changed to a factor with two levels
(‘NO, ‘YES)

Frequency Table:

NO	YES
304	296

Save_act: was changed to a factor with two levels
(‘NO, ‘YES)

Frequency Table:

NO	YES
186	414

Current_act: was changed to a factor with two levels
(‘NO, ‘YES)

Frequency Table:

NO	YES
145	455

Mortgage: was changed to a factor with two levels
(‘NO, ‘YES)

Frequency Table:

NO	YES
391	209

Sex: was changed to a factor with two levels
(‘FEMAL’, ‘MALE’)

Frequency Table:

FEMAL	MALE
300	300

Region: was changed to a factor with two levels
(‘INNER_CITY, ‘RURAL’,’SUBURBAN’,’TOWN’)

Frequency Table:

INNER_CITY	RURAL	SUBURBAN	TOWN
269	96	62	173

Pep: was changed to a factor with two levels

('NO, 'YES)

Frequency Table:

NO	YES
326	274

2.1.4 *Data Tables & Visualizations*

2.2 Models

The second and remaining subsections of Analysis are the model(s).
In some cases, there may only be one model. A model is any method used to analyze the data. Each Assignment specifies which models to use. Always include model details and parameter values when applicable.

The banking dataset was transformed into a transaction object to be modeled using the Apriori Algorithm. The structure of the dataset as a transaction object showed it to be of 600 transactions (rows) and 32 items (columns) in sparse format.

2.2.1 *Model x1 Details*

2.2.2 *Model x1 Parameters*

2.2.3 *Model x2 Details*

2.2.4 *Model x2 Parameters*

3 Results

The Results section of the Assignment will have a subsection for results for each model (assuming that you have more than one).
Results are technical.
They offer technical information about what was found in the analysis. For example, if you performed a correlation in the analysis between all pairs of numeric variables, then your results would discuss the r-value and relationship of each pair. Similarly, if you looked at measures of center and variation, the results talk about what those measures are and what they reveal. For example, if the mean is less than the median, the data is skewed, which means….
Each model we will use in this class has results and parameters associated with it. For example, association rule mining will offer the top ten rules for sup, conf, and/or lift if you you code it to do so. These would go into the results along with the sup, conf, and lift for each rule. The meaning would also be discussed.
**** Always have visualizations**

3.1 Model x1 Results

Technical Analysis, discoveries found…

3.1.1 *Model x1 Visualizations*

3.2 Model x2 Results

Technical Analysis, discoveries found…

3.2.1 *Model x2 Visualizations*

4 Conclusions

General : 3 paragraphs.
This area is not technical at all.
This area explains what was actually found in a way that would make sense to anyone. For example, if your discovered in the analysis that association rule mining with a conf of .2 and a sup of .3 offered 10 rules, you would talk about the measures and values and rules in the *results*. In the Conclusions, you would talk about what it all means. So you would not include the rules themselves or mention of technical measures such as conf or sup. Rather, you would say that you found (as a random example) that people who buy diapers are very likely to by beer and that this means that a store should consider placing these items “near” to each other.