USING BANK DEMOGRAPHY INFORMATION TO DETERMINE WHO WILL OBTAIN THE NEW PERSONAL EQUITY PLAN.

Tajudeen Abdulazeez
School of Information Studies, Syracuse University
toabdula@syr.edu

1. INTRUDUCTION

A personal equity plan (PEP) was an investment plan introduced in the U.K. that allowed people over the age of 18 to invest in shares of British companies. It was done through an approved plan, qualifying unit trust, or investment trust.

The income from a personal equity plan was tax free, so long as the invested funds remained in the plan. As with other types of equity investments, the value of the shares invested in through a personal equity plan could rise or decline with market fluctuations. It was believed that to see the best return on investment from a personal equity plan, the funds should have remained in place for upwards of five years, if not ten years. Due to certain management fees and other charges that may have been applied, withdrawing funds early could have negated the gains they accrued.

PEPs are a good way to save for a child's education. You cannot open one in a child's name, though. If your children are over 18 and UK residents, they can open one for themselves. You can fund their PEP, although the cheques into it must come from them, not you.

The question now is, can we use the demography information of bank customers to determine who will obtain the personal equity plan.

2. ANALYSIS AND MODEL

2.2. ANALYSIS

The analysis is carried out using R-programming language with the following packages:

- plyr : data pre-processing
- dplyr : pre-processing and aggregation
- arule : association algorithm (apriori)
- arulesviz : visualization of the rules generated by the apriori algorithm

The techniques used in this analysis is called association rules, this technique can be used to generate a rule for item in association. The algorithm used is apriori, the important parameters for this algorithm are support, confidence and lift.

Support is the fraction of an item occur in a dataset, Confidence is the probability that a rule is correct for a new transaction with item on the left, lift is the ration by which the confidence of a rule exceeds the expected confidence, lift has to be greater than 1 for it to be meaningful.

The supported data format for apriori algorithm to perform association rule mining in R programming are:

- Transaction data (format: basket)
- Transaction data (format: single)
- Sparse Matrix

2.2.1. About the Data: The datasets used is bank data from Syracuse university LMS, the datasets contain attributes on each person's demography and banking information. It's a record data with

600 observation and 12 variables. Each variable is described below:

- Id : used to identify each person, Its nominal variable with 600 factor. The id column is unique for each of the bank customers.
- Age : The age of each customers. It's a numeric variable with no missing records
- Sex : it's the gender of each bank customer, its male or female. The datasets contain 300 female and 300 male.
- Region : the location description of the customers. It's a factor with four level
- Income : the income for each customer. Its numeric variable
- Married : Yes if married and NO if not married. Its factor with two possible level. 204 customers are not married, and 396 customers are married.
- Children : the number of children for each customers. The data type is interger
- Save_acct : record if customer has saving account or not. If YES, customer has saving account and NO if customer did not have a saving account.
- current_acct : record if customer has current account or not. If YES, customer has current account and NO if customer did not have a current account.
- Mortgage : Record if a customer has a mortgage or Not. Factors with two possible level.
- Pep : This column contain information about each customer if they have a personal equity plan or not. Factor with two possible level (YES/NO)

The datasets have no duplicate id, no missing values but the data format is a record data which need to be transform to the format supported by apriori algorithm as discuss above.

Details of how the data is transform for it to be suitable for association mining is described in the section below.

2.2.2. Data Transformation: The datasets are transforming from record data to transaction data using label encoding discretization.

- Id: This is an identify and its is completely ignored for this analysis. It's a nominal variable
- Age: the age is discretized using the following labels: child (0-12), young_age(13 - 30), middle_age(31-50), senior_age(51 -70) and old (above 70)
- Income: numeric column can not be used for association analysis. This column is discretized into 3 equal bins.

- Children: this column is converted to factor
- Married: recode it as character, since there are many entries of YES/NO in multiple variable, its will be easy to label YES as “married=YES” and NO as “married=NO”
- Car: recode it as character and labeled YES as “car=YES” and NO as “car=NO”
- Save_acct: recode and label YES as “sav_acct=YES” and NO as “save_acct=NO”
- Current_acct : recode as character and labeled YES as “current_acct=YES” and NO as “current_acct=NO”.
- Mortgage: recode as character and label YES as “mortgage=YES” and NO as “mortgage=NO”
- Pep: recode as character and label YES as “pep=YES” and NO as “pep=NO”

Now the whole column has been converted to factor and discretization. the id column is ignored, now we have a transaction data ready for association analysis.

2.3. MODEL

Apriori algorithm from the arule packages is used to perform the association rue mining with the following parameters.

- Support: 0.002
- Confidence: 0.8

The model generated 37,808 rules, the support has a minimum of 0.002 and a maximum of 0.18, the lift has a minimum value of 1.75 and a maximum value of 2.19.

The details statistics summary of the rules is shown below.

```
set of 37808 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
    1    18   211  1383  5066 10384 11676  7000  2069

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     7.0     8.0     7.6     8.0    10.0

summary of quality measures:
    support         confidence        lift
 Min.   :0.002   Min.   :0.80   Min.   :1.75
 1st Qu.:0.002   1st Qu.:1.00   1st Qu.:2.19
 Median :0.002   Median :1.00   Median :2.19
 Mean   :0.004   Mean   :0.99   Mean   :2.16
 3rd Qu.:0.005   3rd Qu.:1.00   3rd Qu.:2.19
 Max.   :0.183   Max.   :1.00   Max.   :2.19

mining info:
   data ntransactions support confidence
 Tester           600   0.001        0.8
```

Fig: 1 Summary of the generated rules statistics and parameters.

3. RESULTS

By Sorting the rule using confidence in a decreasing order we have the following top five rules as shown in the figure below.

```
  lhs
rhs            support confidence lif
t
[1] {income=(4.38e+04,6.31e+04],chil
dren=3}                 => {pep=pep=
YES} 0.0133  1          2.2
[2] {income=(4.38e+04,6.31e+04],chil
dren=1}                 => {pep=pep=
YES} 0.0267  1          2.2
[3] {age=middle_age,region=SUBURBAN,
children=3}             => {pep=pep=
```

Fig 2: Top five rules sorted by confidence

customer with income level between 4.38e+04,6.31e+04 and have three children are likely to obtained PEP. The rule support level is 0.013, the confidence is 1, and a lift of 2.2.

The rule will be very strong and customer belonging to this category can be target for personal equity plan because they are most likely to obtain it.

The high level of confidence and lift with a strong support level of 0.0033. The customer with middle age bracket and that come from the region classified as SUBURBAN with three children are likely to obtain a personal equity plan. Customer that belong to this category can be targeting to increase the sales of PEP.

The customer from SUBURBAN region with income bracket between 4.38e+04,6.31e+04 and at least one child is likely to buy PEP. Customer belonging to this category can be targeted.

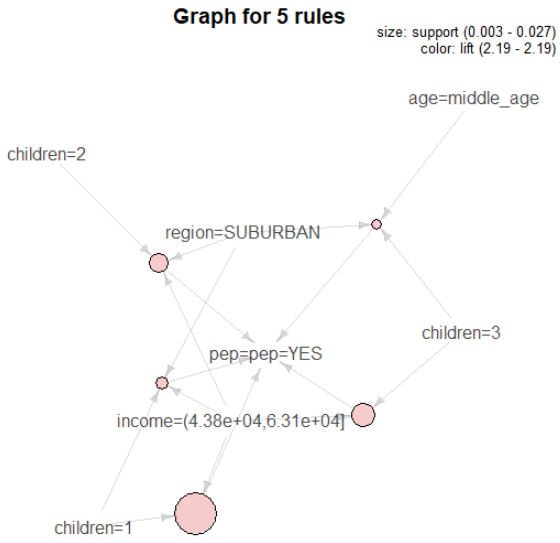

Fig 3: Graph for 5 rules sorted by confidence

By sorting the rules by lift, a new set of top five rules is generated as shown below:

```
     lhs
[1] {income=(4.38e+04,6.31e+04],children=3}
[2] {income=(4.38e+04,6.31e+04],children=1}
[3] {age=middle_age,region=SUBURBAN,children=3}
[4] {region=SUBURBAN,income=(4.38e+04,6.31e+04],c
[5] {region=SUBURBAN,income=(4.38e+04,6.31e+04],c
```

Fig 4: top five rule sorted by lift

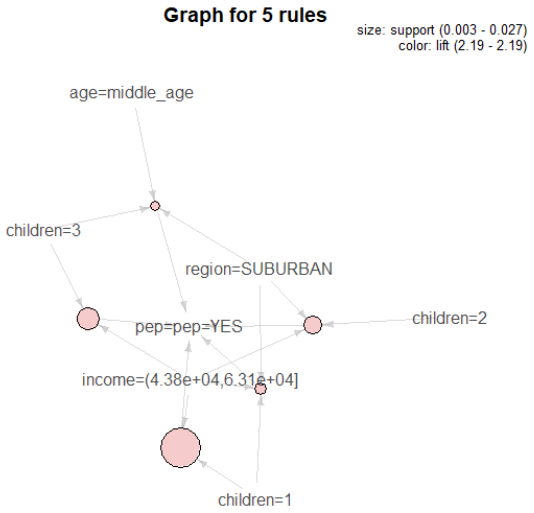

Fig 5: top five rule sorted by lift

By sorting the rule based on support, the top five rule that is generated is shown below:

```
lhs                                                rhs           support confidence lift
[1] {children=1}                                    => {pep=pep=YES} 0.18    0.81
1.8
[2] {children=1,current_act=current_act=YES} => {pep=pep=YES} 0.14    0.83
1.8
[3] {children=1,save_act=save_act=YES}       => {pep=pep=YES} 0.13    0.84
1.8
[4] {married=married=YES,children=1}         => {pep=pep=YES} 0.12    0.83
1.8
[5] {children=1,mortgage=mortgage=NO}        => {pep=pep=YES} 0.12    0.85
1.9
```

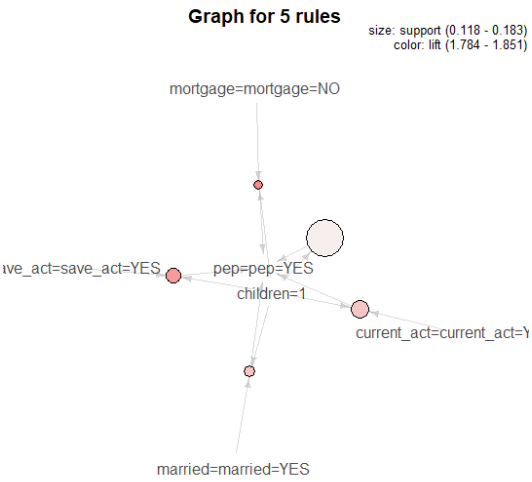

Fig 6: Top five rule sorted by support

4. CONCLUSION

Base on the result of the analysis, customer with income bracket between 43800 and 63100, and with at least one child are likely to obtained personal equity plan.

Customer that fall between the age of 31 to 50, that reside in suburban region with at least three children are likely to buy personal equity plan.

Customer from suburban region with income bracket between 4.38e+04,6.31e+04, with at least one child are likely to buy personal equity plan.

Customer that is married with at least one child are likely to buy personal equity plan.

Customer with only one child without mortgage are likely to buy a personal equity plan.