Matthew L. Pergolski
IST 707
2/3/2022

Introduction

As defined by Investopedia.com, a Personal Equity Plan (PEP) is “an investment plan…that encourage[s] people over the age of 18 to invest in [various] companies.” Furthermore, the PEP offers incentives for those interested to sign up – and an increase of these generated accounts is seen as a beneficial move for most banks.

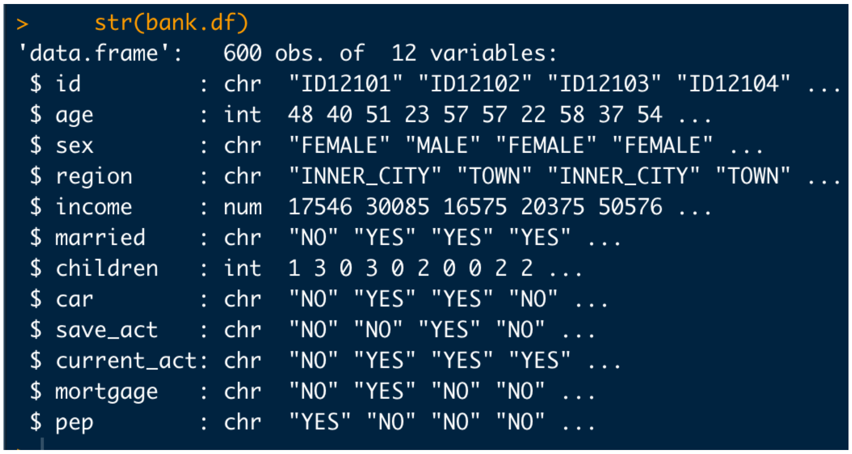
With a provided data set of historical transactions, the main objective of this study is to provide valuable insights to banks looking to expand the userbase of these accounts. The findings may suggest actionable insights in terms of who to target for continued growth of these accounts. The method in which the data set will be analyzed is through Association Rule Mining (ARM), which can detect patterns and trends in the data based on various attributes.

Overall, five (5) main association rules will be provided to banking clients looking to expand their user base of these PEP accounts.

Analysis and Models

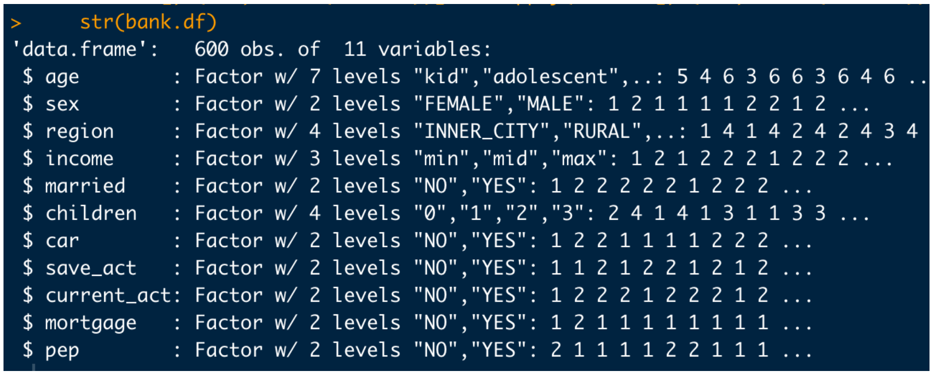
About the Data

When exploring and integrating the data set within the RStudio environment, a holistic review is examined over each variable and observation. Overall, it is observed that the data set has 12 variables/attributes with 600 observations. Most of the data types for these variables are initially categized as a ‘chr,’ or character string within R. In order to provide valuable/actionable insights for the banking clients interested in this report, data pre-processing will need to be executed. This involves selecting the relevant data types for each variable.



```
>     str(bank.df)
'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" ...
```

As previously mentioned, data ‘munging’ will need to occur on the data set in order to please the client. The cleansing process first involves removing a variable that will not provide any actionable value to the analysis, the ‘id’ column. Subsequently, a term known as ‘bucketing’ or discretization will be conducted on the ‘age’ and ‘income’ columns. The former will bucketed within decades spanning from 0 to 60+ while the income variable will be split into three (3) overall groups: min, mid, and max. Following this, the rest of the data set will be rebranded as a ‘factor.’ The updated data frame looks as follows:



```
>     str(bank.df)
'data.frame':   600 obs. of  11 variables:
 $ age        : Factor w/ 7 levels "kid","adolescent",..: 5 4 6 3 6 6 3 6 4 6 ..
 $ sex        : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1 1 1 2 2 1 2 ...
 $ region     : Factor w/ 4 levels "INNER_CITY","RURAL",..: 1 4 1 4 2 4 2 4 3 4
 $ income     : Factor w/ 3 levels "min","mid","max": 1 2 1 2 2 2 1 2 2 2 ...
 $ married    : Factor w/ 2 levels "NO","YES": 1 2 2 2 2 2 1 2 2 2 ...
 $ children   : Factor w/ 4 levels "0","1","2","3": 2 4 1 4 1 3 1 1 3 3 ...
 $ car        : Factor w/ 2 levels "NO","YES": 1 2 2 1 1 1 1 2 2 2 ...
 $ save_act   : Factor w/ 2 levels "NO","YES": 1 1 2 1 2 2 1 2 1 2 ...
 $ current_act: Factor w/ 2 levels "NO","YES": 1 2 2 2 1 2 2 2 1 2 ...
 $ mortgage   : Factor w/ 2 levels "NO","YES": 1 2 1 1 1 1 1 1 1 1 ...
 $ pep        : Factor w/ 2 levels "NO","YES": 2 1 1 1 1 2 2 1 1 1 ...
```

As seen above, 11 variables are now shown, since the ‘id’ column was removed – and the rest of the data set is categorized as a factor. Consequently, the data set is now viable and eligible to be analyzed with the ARM method.

Exploring the Data

Within the ‘Exploring Data’ section, various rule sets will be developed and generated in order to find the most ‘interesting’ or ‘valuable’ association rules that banking clients will be interested in reviewing.

The main component of each association rule is the following structure:

- Support
- Confidence
- Lift

As defined in the textbook, the ‘support’ for ARM determines how often a rule is applicable to a given data set while ‘confidence’ determines the frequency of any potential link between two transactions, terms, items, or any other type of factored data.

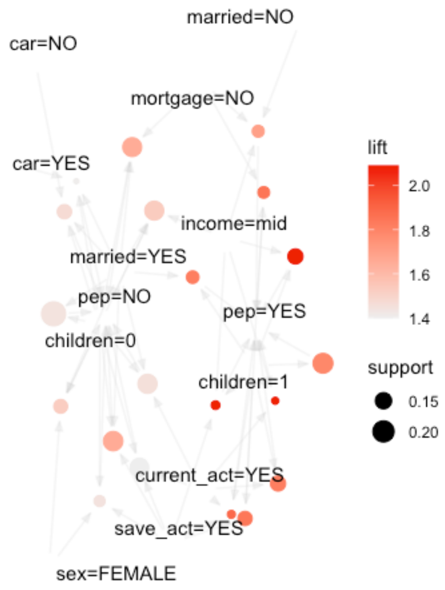
The last aspect of ARM noted in this paper is ‘lift,’ which ultimately informs the reader of the quality for a specific rule. In short, it is calculated by a ratio of confidence and expected confidence of a given rule; any lift value larger than 1 indicates a relatively ‘strong’ rule (in comparison to others), while a lift value equal to one suggests ‘independence.’ A lift value less than 1 informs the reader that the rule is not relatively strong and may not provide any useful or valuable insight.

The ARM concept can help the reader/viewer notice the establishment of relationships between two or more variables within a data set – which can be leveraged for, in this case, potentially increasing the number of PEP accounts for banking clients.

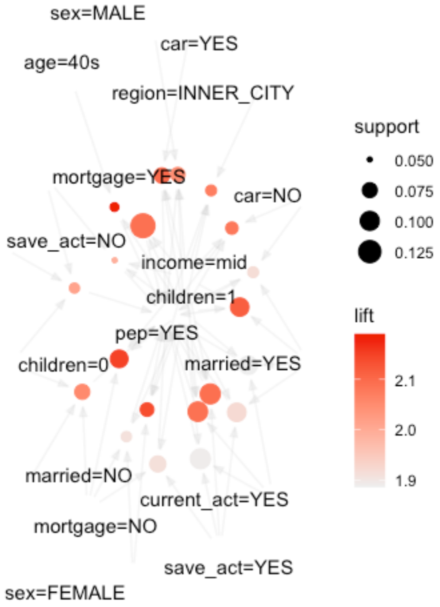
In order to conduct the ARM analysis, an ARM package titled ‘arules’ will be utilized along with the ‘apriori’ function – data visualizing packages ‘arulesViz’ will also be used to assist in visualizing these association rules. See below for rule/model development for the banking data set.

Results

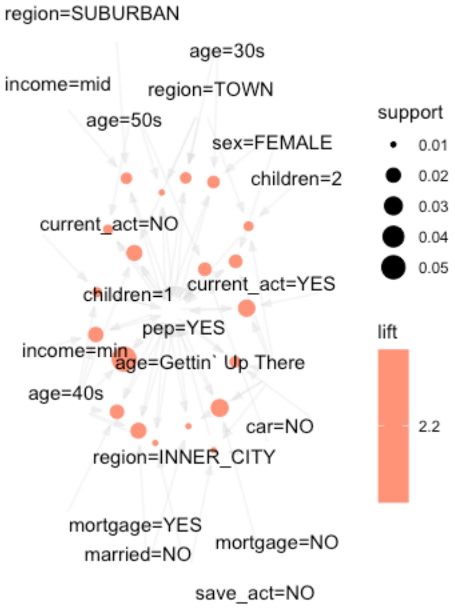
The first ruleset utilized a support value of 0.1 and confidence of 0.75, meaning that an observation that occurred in 10%+ in the data included a second, third, or more pairing 75% of the time. Overall, 24 rules were developed spanning confidence values from 0.96 to 0.75 and lift values over 1.



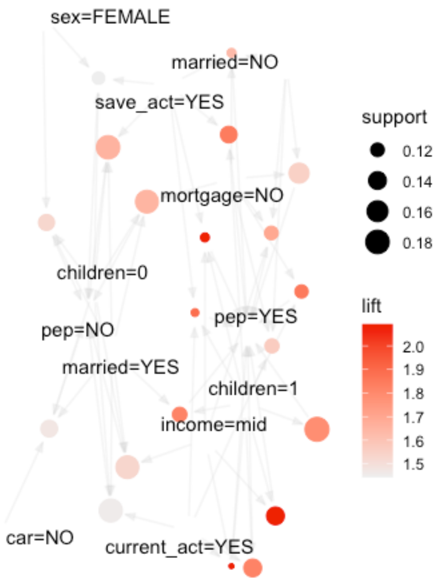
Following this, the second rule set utilized a support level of 0.05 and confidence of 0.86. This means that an observation that applied to at least 5% of the data included a second, third, or more pairing 86% of the time. The second rule set generated 26 rules, spanning confidence values from 1 to 0.86 and lift values of over 1.



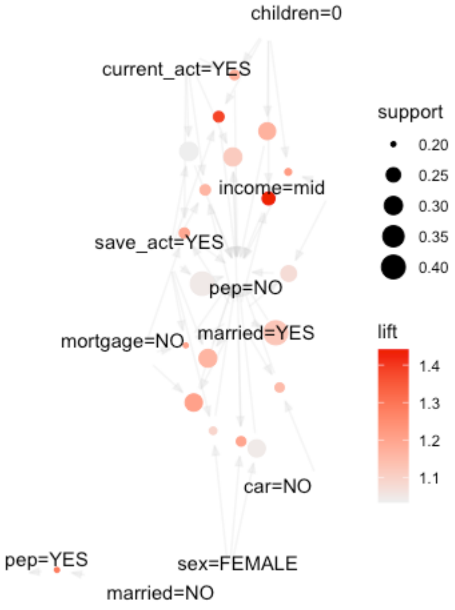
The third rule set utilized a support level of 0.01 and confidence of 1. This means that an observation that applied to at least 1% of the data included a second, third, or more pairing 100% of the time; the rule set generated 68 rules, spanning confidence values from 1 to 1 and lift values of over 1 as well.



The fourth rule set utilized a support level of 0.1 and confidence of 0.7. This means that an observation that applied to at least 10% of the data included a second, third, or more pairing 70% of the time; the rule set generated 34 rules, spanning confidence values from 0.96 to 0.70 and lift values all over 1.



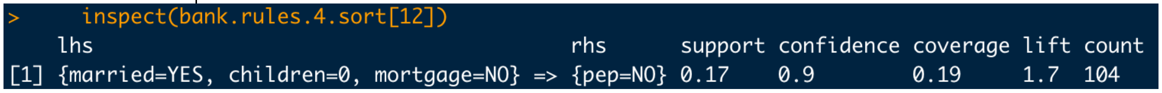
The fifth and final rule set utilized a support level of 0.2 and confidence of 0.55. This means that an observation that applied to at least 20% of the data included a second, third, or more pairing 55% of the time; the rule set generated 22 rules, spanning confidence values from 0.78 to 0.55 and lift values equal to or over 1.



The rule set that will be chosen for final conclusions will be rule set four (4). This fourth set was chosen due to the relatively high support value of 0.1 (or 10%) while the confidence met or exceeded 0.7 (or 70%). Lift ranges spanned from 2.1 to 1.3, which indicates non-independence. The range of confidence spanned from 0.96 to 0.70, which provides an optimal range of likely occurrences.

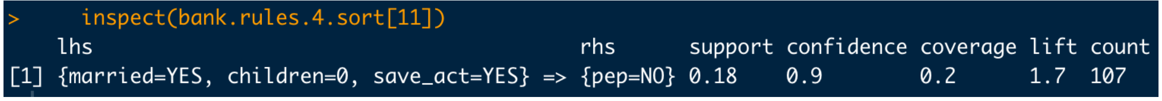
From the fourth rule set, the five (5) main rules used within the conclusion will be the following:

1. Rule Set 4 | Rule 12



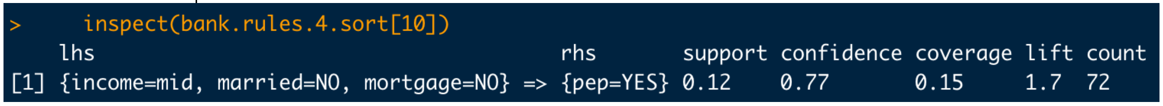
```
>    inspect(bank.rules.4.sort[12])
    lhs                                      rhs      support confidence coverage lift count
[1] {married=YES, children=0, mortgage=NO} => {pep=NO} 0.17    0.9        0.19     1.7  104
```

2. Rule Set 4 | Rule 11



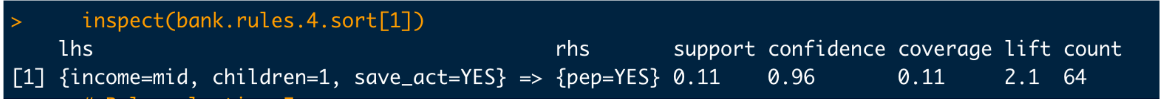
```
>    inspect(bank.rules.4.sort[11])
    lhs                                      rhs      support confidence coverage lift count
[1] {married=YES, children=0, save_act=YES} => {pep=NO} 0.18    0.9        0.2      1.7  107
```

3. Rule Set 4 | Rule 10



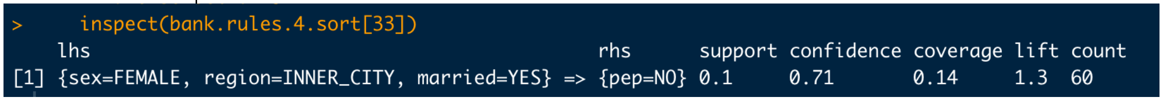
```
>    inspect(bank.rules.4.sort[10])
    lhs                                     rhs       support confidence coverage lift count
[1] {income=mid, married=NO, mortgage=NO} => {pep=YES} 0.12    0.77       0.15     1.7  72
```

4. Rule Set 4 | Rule 1



```
>    inspect(bank.rules.4.sort[1])
    lhs                                     rhs       support confidence coverage lift count
[1] {income=mid, children=1, save_act=YES} => {pep=YES} 0.11    0.96       0.11     2.1  64
```

5. Rule Set 4 | Rule 33



```
>    inspect(bank.rules.4.sort[33])
    lhs                                            rhs      support confidence coverage lift count
[1] {sex=FEMALE, region=INNER_CITY, married=YES} => {pep=NO} 0.1     0.71       0.14     1.3  60
```

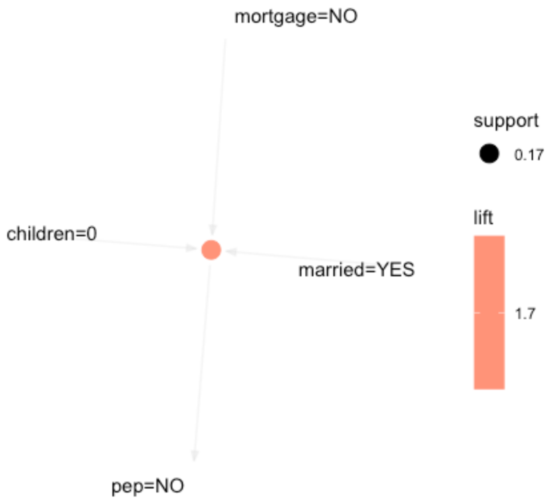
See the ‘Conclusion’ section for interpretation of these rules.

Conclusions

The following five (5) rules were chosen as a method to provide valuable insight to banking clients regarding PEP accounts. The ARM method was conducted on the banking data set provided. The support, confidence, and lift measures of the ARM method were calibrated to provide the association rules listed below. Visualizations are provided for each rule to assist in comprehension of the models.

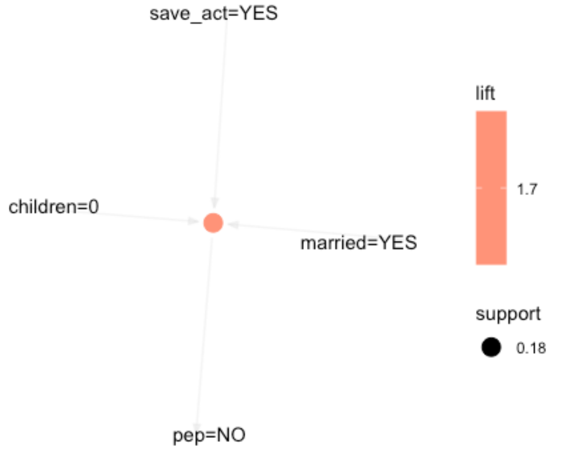
The first rule indicates that those who are married, do not have a mortgage, and do not have children are likely to not hold a PEP account. The confidence for this demographic is high at 96%. Based on the model, there is major growth opportunity for this type of demographic.

Rule 1 Visualization Figure:



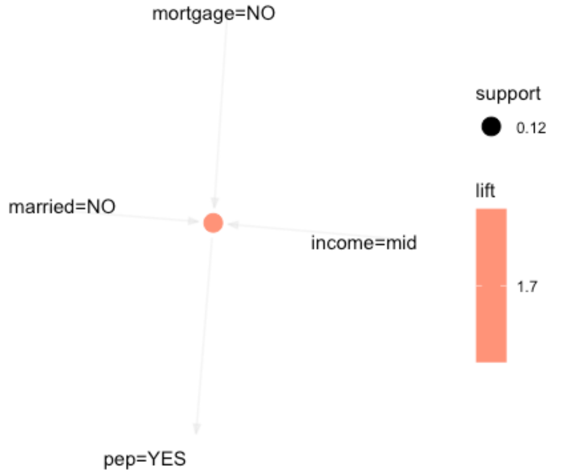
The second rule indicates that those who are married, hold a savings account, and do not have children are also not likely to hold a PEP account. The confidence for this demographic is high at 90%. Similarly to rule 1, there is major growth opportunity for this type of demographic.

Rule 2 Visualization Figure:



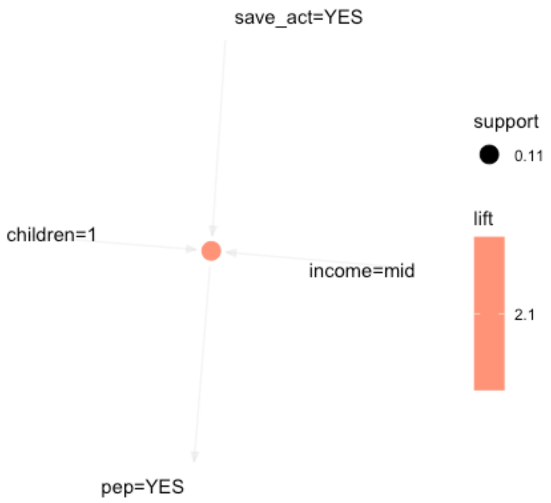
The third rule indicates that those who are not married, do not have a mortgage, and have a middle-ranged income are relatively likely to hold a PEP account. The confidence for this demographic is somewhat high at 77%. Based on the model's findings, there is lower growth opportunity for this demographic compared to the first two.

Rule 3 Visualization Figure:



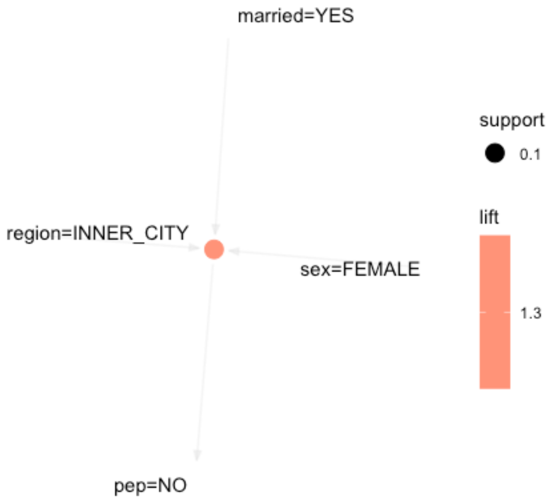
The fourth rule indicates that those who have at least one child, have a savings account, and have a middle-ranged income are relatively likely to hold a PEP account. The confidence for this demographic is high at 96%. Based on the model's findings, there is an extremely low growth opportunity for this demographic moving forward.

Rule 4 Visualization Figure:



The fifth and final rule indicates that females who living in the inner city and register as married are not likely to hold a PEP account. The confidence for this demographic is relatively high at 71%. Based on the model's findings, there is ample growth opportunity for this demographic moving forward.

Rule 5 Visualization Figure:



All in all, to reiterate, banking clients concerned with the data set provided can gain actionable insight into which demographic groups to target in potential ad campaigns (or incentivized programs) to ultimately increase the number of overall PEP growth rates.