Zane Alderfer

Dr. Bolton

IST 707

October 27, 2023

Homework 2

Introduction

Investing in retirement has been a focus of concern for multiple generations now. Unfortunately, a lot of individuals have no investment in their retirement which will surely become an area of concern for that individual as they will find themselves working well past retirement age or having no wiggle room leading up to retirement in terms of having an excess of money to spend. That is why it is important to collect data on people and be able to identify those who are capable of developing a retirement plan but haven’t taken advantage of the tools readily available to them.

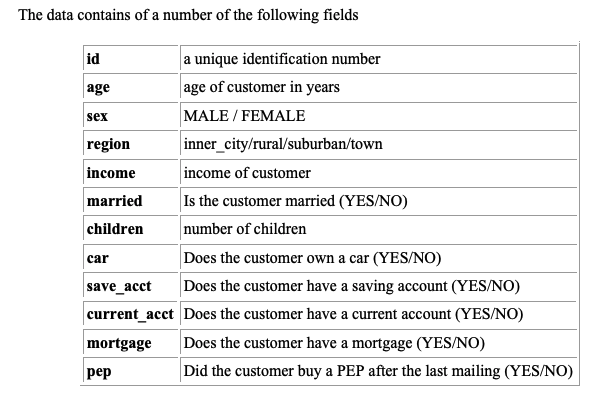
While it has since been replaced with Individual Saving Accounts (ISA) in 1999, Personal Equity Plans (PEP) were a great way for individuals 18 years and older to invest in their future. Although the primary goal of a PEP was to enhance the future for the individual investing in the plan, corporations also benefit from these plans as well in the form of a financial investment toward their company from these individuals. With that in mind, institutions would highly benefit from being able to identify what individuals would be more likely to invest in a PEP.

The dataset being analyzed contains information on people that could potentially be used to adequately pick the most prospective individuals to invest in a PEP. The goal here would be to find a cocktail of familiar attributes among these individuals that seem to have a high correlation to purchasing a PEP. Using the presented demographics and financial information, the targeted population will hopefully become apparent.

Analysis and Models

**About the Data**

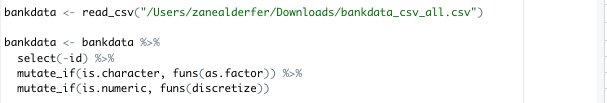
The dataset contains 600 observations of demographic data and financial information that can be used in determining candidates for purchasing a PEP. There are twelve variables presented here as shown in the table below.



*Table 1*

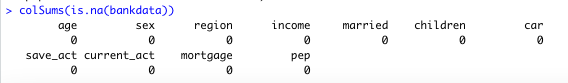
**Reading and Converting the Data**

To read in the data, a similar method was used as Tin’s in the provided record to transactions for association rule mining markdown file provided. Once the data was read in, a mutate\_if function was used to convert all the variables to factors or discretize wherever appropriate. This can be seen down below.

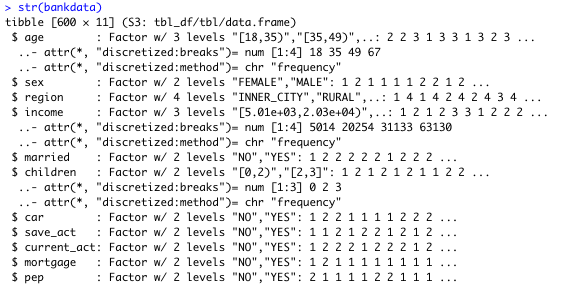


**Data Structure**

Identifying missing values as seen below:

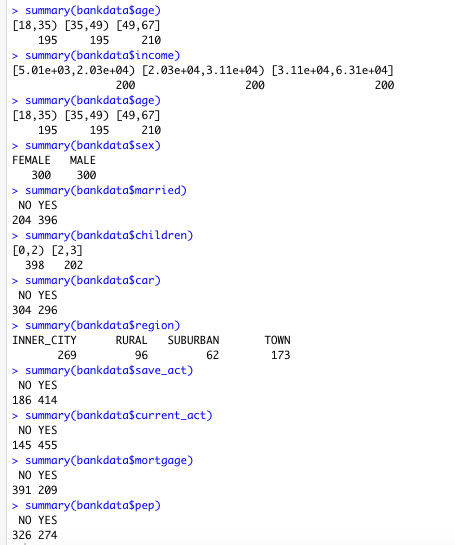


There are no missing values as can be seen and to better understand the data a look at how the variables are factorized can be seen:



**Closer Look at the Data**

A summary of all the variables:

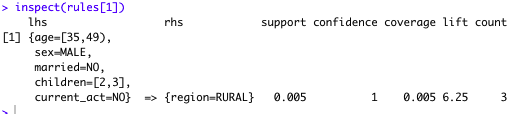


For the most part, there’s a relatively even distribution across all the variables in terms of demographics which will make for a result that effectively paints an accurate picture of the data. The dataset also has a perfect distribution of incomes and age across the three age and income ranges that came from the factorization which are presumably the 2 most important variables involved.

**Analysis of the Models**

To develop rules and decide which of the 5 rules are most interesting, the apriori function will be used. This uses support which is essentially the popularity of a set of items, confidence which is how often these items can be seen together and then finally lift which defines how strong a rule is.

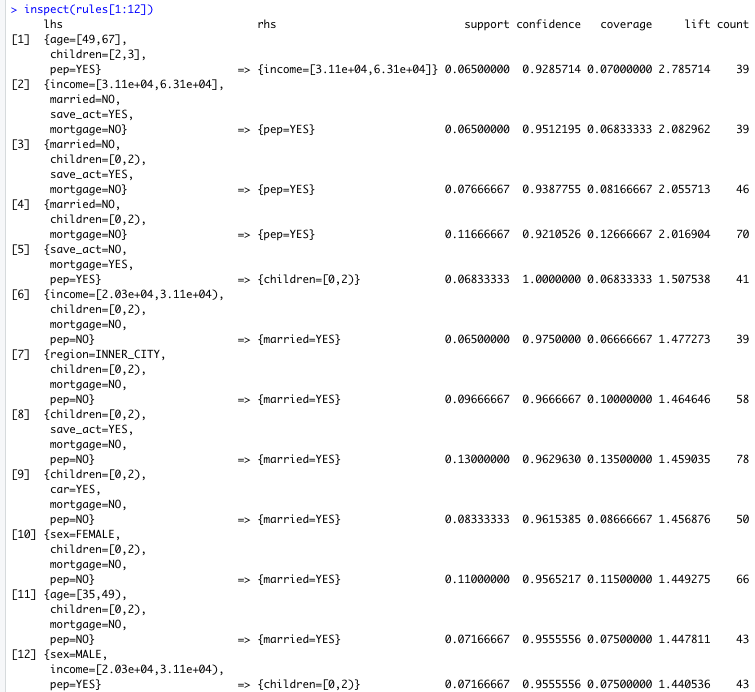
As a precursor to what is being looked for, a variable called “rules” using the apriori function was developed using a support of .005, confidence of .8, max length of 6 and in descending order based on lift. Due to the support being so low, the variables that occur will probably have a high confidence as these variables probably don’t occur very often together. The first rule result is shown below:

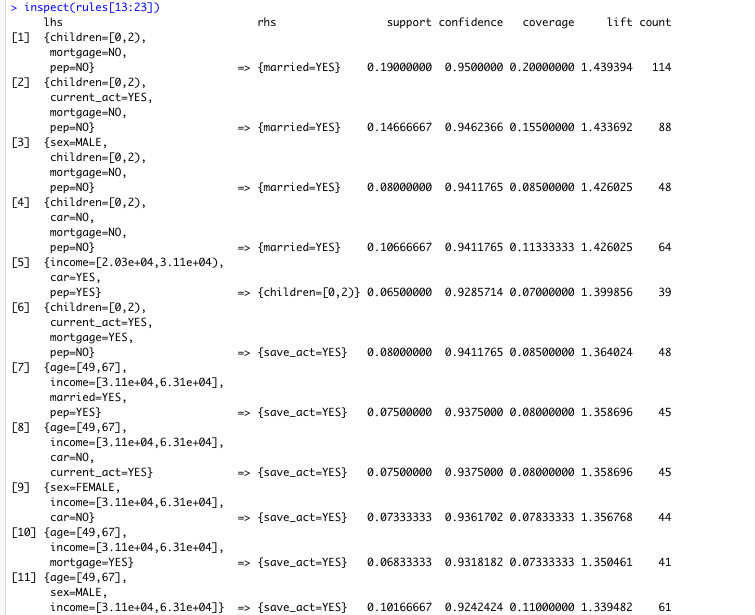


The results above indicate that a person who is between the ages of 35-49, male, not married, having between 2 and 3 children, with no current account is 100% living in a rural region. There are over 42,000 thousand rules developed using this criteria, the goal is to generate around 20-30 rules and distinguish five that are the most interesting so the criteria will be adjusted.

To get down to about 20 rules, the support was adjusted to 0.065 and confidence to 0.92 which left a total of 23 rules.

With only 23 rules, it should be easy to inspect all of them at once and decide which five are the most interesting as seen below:



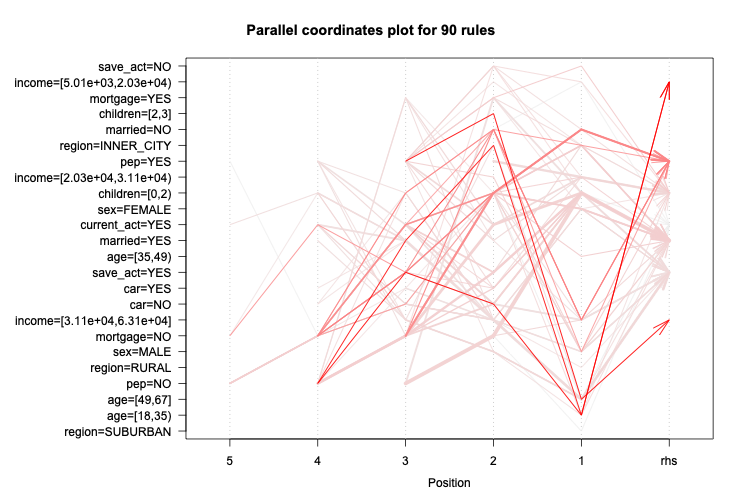


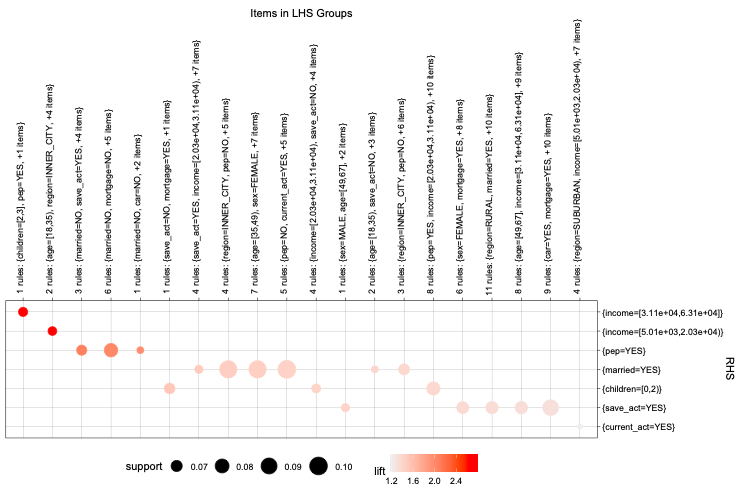
As mentioned above, there’s a total of 23 rules and to have the rules hold value, the rule should contain the PEP variable in some fashion.

The five most interesting rules are as follows:

1. *{income = [3.11e+04, 6.31e+04], married = NO, save\_act = YES, mortgage = NO} => {pep = YES}*
   1. *Support = 0.065*
   2. *Confidence = 0.9512195*
   3. *Lift = 2.082962*
2. *{married = NO, children - [0,2), save\_act = YES, mortgage = NO} => {pep = YES}*
   1. *Support = 0.07666667*
   2. *Confidence = 0.9387755*
   3. *Lift = 2.055713*
3. *[region = INNER\_CITY, children = [0-2), mortgage = NO, pep = NO} => {married = YES}*
   1. *Support = 0.09666667*
   2. *Confidence = 0.975*
   3. *Lift = 1.477273*
4. *{sex = MALE, income = [2.03e+04, 3.11e+04), pep = YES} => {children = [0,2)}*
   1. *Support = 0.07166667*
   2. *Confidence = 0.9555556*
   3. *Lift = 1.440536*
5. *{age = [49,67], children = [2,3], pep = YES} => {income = [3.11e+04, 6.31e+04]}*
   1. *Support = 0.065*
   2. *Confidence = 0.9285714*
   3. *Lift = 2.785714*

There seems to be a lot of rules that involve children, marriage, and a mortgage so to further inspect that, a few graphs were created to see a possible correlation between these variables and pep investments. To get an overall better gauge of all the data and a more meaningful graph, the support and confidence was adjusted a little to allow for more rules and in this case it will be 90 rules. A graph based on support and lift was created below:

*Graph 1*

*Graph 2*

The two graphs above are represented by either the darker shade of arrows or darker shade of circles that have the highest lift in terms of rules. In graph 2, however, if the circles are larger, this indicates that the rules have more support or more appearances in the dataset.

Results

When looking at rule number 1, it is clear that those with higher incomes will probably purchase a Personal Equity Plan. Starting with graph 2, the first rule is one of the darker orange circles with a higher lift indicating that these two tend to appear in the data together very frequently. This can also be backed up by the high confidence of over 0.95. It is also interesting to note that it is one of the two rules of the five identified that has the status of not married.

Rule number 1 almost segways into rule number 2 in that once again a marital status of “no” is a criteria in this high lift rule. Someone who falls within the lines of this rule also doesn't have a mortgage with minimal children. There seems to be familiar symptoms of having more money tends to lead to being able to purchase a PEP. This can once again be seen in graph 1 as one of the darker shaded arrows indicating a high lift score.

Rule number 3 is interesting in that it has PEP as an LHS variable instead of being the resulting RHS variable. This rule, however, doesn’t seem to deviate from the “norm” as the RHS variable is marital status being “yes.” But with that being said, the LHS PEP variable has a “no” status. Of the five rules presented, this has the highest support with this combination of LHS variables and RHS result being seen in about 1% of all the data which may not seem like much but when considering large datasets, this can be significant. An interesting attribute is presented here as well in that their region is an inner city which could be something to look further into.

Rule number 4, like rule number 3, presents the PEP indicator as a LHS variable. This time, however, as a “yes.” This also introduces “sex” as a variable in the LHS variables. It should be mentioned that although the RHS variable is the individual having 0-2 children, they very well could have 0 children which could support the results of the previous 3 rules. This is of course speculation but with the lowest lift score and income being on the lower side, the results could be inconclusive.

Now we approach rule number 5 with the highest lift score. These individuals are older, with more than 2 or 3 children, a higher income and purchased a PEP. This rule seems to be a summary of the previous 4 rules as although it doesn’t happen too often, a PEP to be purchased in this scenario is almost expected. This is the darkest circle in graph 2 as well which once again supports its high lift score.

Conclusion

Based on the results, there’s not much of a secret here as to who the best prospects to sell PEP’s to should be. Before the results are broken and possible suggestions are made, addressing uncommon attributes among the five rules presented is necessary. In rule number 3, living in the inner city seems to have a possible correlation with not purchasing a PEP. This is interesting as inner cities tend to have both extremes of demographics in terms of personal income. The RHS for this rule is a marital status of “yes.” This status of marriage was not a common theme among PEP purchasers. So, it could be possible that although they live in the inner city, supporting a spouse may be the leading factor as to why they can’t afford to purchase a PEP. It is important to remember that this is a dataset of 600 observations with not even half being from the inner city so it would be interesting to see a larger dataset with more inner city individuals.

With that, the other unusual attribute was “sex” being “male” in rule number 4. There are a lot of interesting factors here as these individuals don’t seem to make as much as the other people in the other rules but they do however have 0-2 children. So, once again it seems that although being male is a contributing LHS attribute, there are other factors involved here that would suggest reasoning behind purchasing a PEP. 0-2 children could very well mean 0 children and based on 600 observations with exactly half being male, there’s not much of a reason to insist that males are more likely to purchase a PEP. Based on previous rules, one could probably assume that most of these males are not married as well so their income, although smaller, can comfortably afford to purchase a PEP.

It would be easy to suggest that looking for people with high incomes, no spouse, no children, no mortgage, and a savings account is the safest bet to be selling a PEP to. It is beyond a reasonable doubt that those selling the PEP’s already know this. So, based on the data and the five interesting rules picked, some unique suggestions could be:

* if a prospect has 2 or 3 children, it would be preferable that they’re older due to longer time to accumulate wealth
* if a prospect doesn’t make as much as others, it would be preferable that they don’t have a lot of dependents in their life
* if a prospect is married, be sure that they are comfortably supporting themselves and their spouse or that their spouse also potentially has an income

This is definitely an interesting dataset and could potentially be very enlightening if it were applied to 600,000 people instead of 600. This could also be applied to something else even more profound than a PEP such as just a 401k in general.