Brandon Croarkin

# HW #3 – Association Rule Mining

## Introduction

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. Investors received both income and capital gains free of tax. These benefits incentivize the public to participate in investing through the prospect of income and capital growth at a greater rate than leaving your money in a bank account. However, even with these great incentives to have an account, banks still need to undergo a large amount of sophisticated advertising campaigns, involving data mining, to help educate customers about these products.

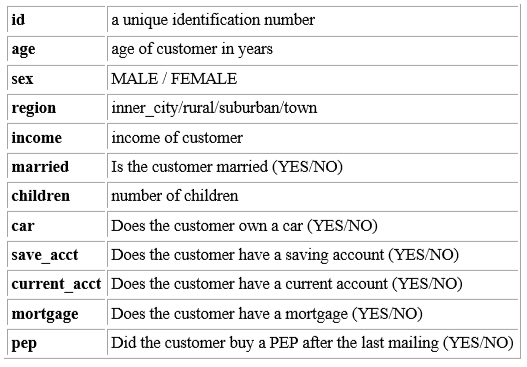
When launching products, such as a PEP, a direct mail piece is often sent to customers, and a record is kept as to whether the customer responded and bought the product. Using this information, marketing departments can mine this data, using tools such as association rule mining, to determine factors associated with obtaining the new PEP.

The analysis presented in this paper will go through some sample bank data with demographic information and whether they obtained a PEP. Using this data, association rule analysis can be performed to learn information on what factors lead to an individual buying a PEP, as well as other interesting rules that can be used for promotion and marketing. Using this information, a bank could help make more efficient marketing campaigns by targeting the groups that have the highest propensity to get a PEP and other financial services based on their demographic information.

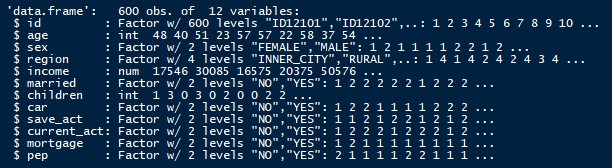
## Analysis

### The Data

The bank data file used in this analysis was obtained from DePaul University and contains attributes on each person’s demographics and banking information, including whether they bought a PEP after the last mailing. This data set consists of 600 observations on 12 variables. The full list of the variables and their descriptions is below.



An additional look at the structure of the data shows the data types of the variables. All but 3 of the variables are factors, with the remaining 3 (age, income, and children) being either an integer or number.

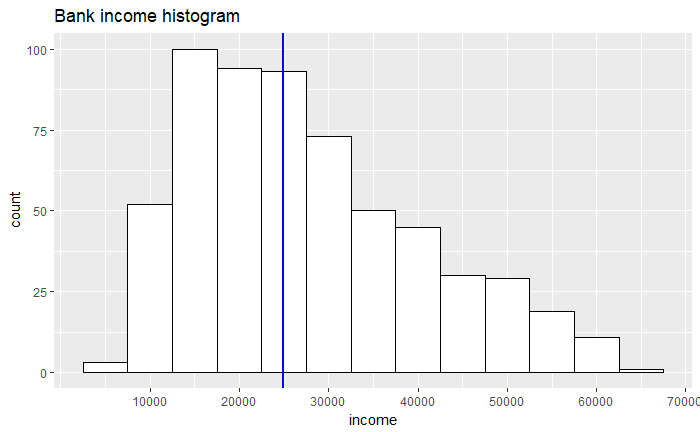
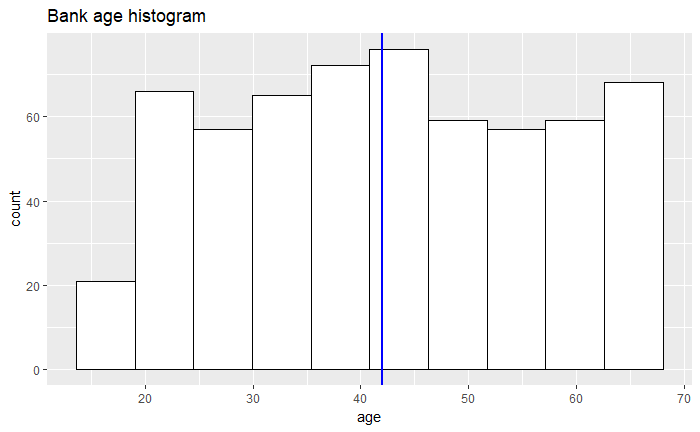


### Data Preparation

Before performing any association rule analysis, preprocessing is needed to put it in a form that is suitable for analysis. To perform association rule analysis the data first needs to be in a discrete format. Additionally, unneeded columns like the ID field can be deleted since they do not contain any important information. No missing data was found in this dataset.

Since the bank data contains numeric fields, such as age, children, and income, they need to be converted. Children can be converted to a factor variable easily since it is an integer ranging from 0 to 3. The age and income variables will need additional work to convert them to discrete variables.

The histograms below plot out the distribution of the age and income to give information on their distribution. Age has a very flat distribution, with the values very evenly distributed, with a drop below age 20. The mean is a little over 40. Income has a normal distribution with a positive skew. The mean is around $25,000.

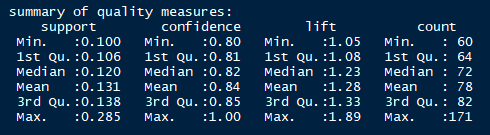


These values can be discretized by grouping the values in both columns into categories based on their values. For this analysis, the age variable is broken up into decades (under 20, 20-29, 30-39, 40-49, 50-59, and 60-69) and the income variable is broken up into five equal width categories (“very high income”, “low income”, “average income”, “high income”, and “very high income”).

### Processing

Once the data has been preprocessed, the association rule analysis can begin to find strong rules. Association rule analysis is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. The strength of an association rule can be measured in terms of its support, confidence, and lift. Support determines how often a rule is applicable to a given data set. Confidence determines how frequently items in Y appear in transactions that contain X. Lift is a measure of dependent or correlated events. Association rules should have > 1 lift to be meaningful. A value of < 1 for the lift implies a negative correlation.

An initial look at the association rules in the dataset with default parameters gives a range of the support, confidence, lift, and count.



The association rule analysis in this paper uses the apriori principle to create the rules. The apriori principle states that if an itemset is frequent, then all its subsets must also be frequent. The inverse of this is true as well. This is useful information as it reduces the number of rules the computer needs to sift through reach the association rules.

With this knowledge, tuning of the association rules can begin to find rules that pass a minimum support level and have strong lift and support. Doing this gives the strongest rules in the dataset. Additional tuning is then done to focus in on target variables, such as “PEP=YES”, “PEP=NO”, and “income=very high income”.

To find the strongest rules, the confidence and lift were each sorted in the created rule sets in descending order. This results in two lists from the given rule parameters with the rules sorted by the highest confidence in one list and the other sorted by the highest lift. Support was set at .100 during the search for the strongest rules to find all rules in the dataset since the strength of the rule was the focus.

When searching for the best rules related to having a PEP, the support was set at .12 (the median value) to only find rules that have a good level of support while also not trimming away too many potential rules. Additionally, these PEP rules set the max length of the itemset at 3 so the rules that were outputted did not have so many factors as to be uninterpretable.

## Results

From the rules outputted from the analysis, many interesting and strong rules can be found. Some of the most interesting rules are not the strongest rules, and vice versa. Below are some of the results categorized into different sections.

### Strongest Rules (Confidence)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {income=high income} | {save\_act=YES} | 0.12 | 1.00 | 1.4 | 71 |
| {children=0, save\_act=YES,  mortgage=NO, pep=NO} | {married=YES} | 0.12 | 0.99 | 1.5 | 73 |
| {sex=FEMALE, children=0,  mortgage=NO, pep=NO} | {married=YES} | 0.10 | 0.98 | 1.5 | 63 |
| {children=0, current\_act=YES,  mortgage=NO, pep=NO} | {married=YES} | 0.13 | 0.98 | 1.5 | 80 |
| {children=0, mortgage=NO, pep=NO} | {married=YES} | 0.17 | 0.97 | 1.5 | 104 |

5 rules were found that have confidence levels greater than .97, which is a very strong value. The strongest, high income implies having a savings account, has 100% confidence. This is an intuitive result, as those with a “high income” need a place to store their money. The support value means that 12% of the individuals in this dataset had a high income and a savings account, as is evident from the count of 71 which is 12% of the total observations (600). The confidence value of 1 means that out of the 71 times that high income occurred in the dataset, the individual also had a savings account. The lift can be calculated by finding the support of high income and save\_acc = YES and dividing this by the product of the support of high income and the support of save\_acc = YES. There were 414 occurrences of having a savings account, resulting in a support of .69 (414/600). The resulting equation for lift is thus.12/(.69 x .12), which equals 1.4.

The rest of these strong rules relate to factors that are related to being married, implying that there are some strong correlations associated with one being married. It is also a very interesting finding that all of these remaining strong rules include children = 0, as one would expect married couples to have children. Additionally, all these rules have PEP=NO as a factor in being married. In addition to high confidence, all of these rules also have a strong lift, ranging from 1.4-1.5.

### Most Interesting Rules

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {region=RURAL} | {mortgage=NO} | 0.11 | 0.71 | 1.09 | 68 |
| {children=1} | {pep=YES} | 0.18 | 0.81 | 1.8 | 110 |
| {married=NO} | {pep=YES} | 0.20 | 0.59 | 1.3 | 120 |
| {married=NO} | {income=very high income} | 0.028 | 0.083 | 1.4 | 17 |
| {children=0, mortgage=NO, pep=NO} | {married=YES} | 0.17 | 0.97 | 1.5 | 104 |

One of the most interesting rules found was that living in a rural region was associated to not having a mortgage, which is a counterintuitive finding. People living in rural areas would be assumed to be more likely to get a mortgage than those in cities due to the cheaper housing prices and less rental apartment options. This may be an opportunity though for a bank to reach out to rural customers to entice them into getting a mortgage as it may be an untapped market. However, this rule would need further investigation as the support is not that high at .11 and the lift is only a little over 1 at 1.09.

1 child leading to having a PEP is another interesting finding, since it is unclear why 1 child specifically would lead to this result. Having one child could potentially mean that a family has enough resources to buy a PEP, without being financially restricted by the costs of additional children. This rule is also very strong with a high support (0.18), confidence (0.81), and lift (1.8). Based on this rule, families with a single child who don’t have a PEP can be targeted as a group with a high potential to buy a PEP. Further investigation of this rule could potentially find interesting causal relationships.

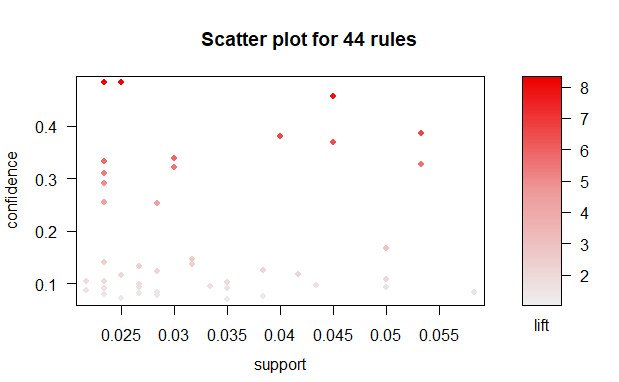
The link between not being married and having a PEP is another interesting relationship since it is interesting the marital status would affect certain financial decisions. This could be due to an individual who isn’t married having more time and resources to devote towards an investment tool like a PEP. Based on this information, unmarried individuals could be a good target for additional marketing for a PEP.

The relationship between marriage and income is another interesting finding since it is fascinating that marital status could impact someone’s financial status. This rule implies that a marriage has a negative impact on financial status. Some possible causes for this could be the costs of a marriage, or that the time spent on a marriage could take away from time spent on accumulating wealth. However, since the support and confidence on this relationship is low, additional research would be necessary. Assuming the relationship is true, unmarried individuals again seem to be a great target for PEP, for those who do not have it already, since they have a propensity to be interested in the product and a lot of wealth to put into it.

The last interesting relationship noted above involved 0 children, no mortgage, and no PEP being related to being married. This rule is interesting since it seems very counterintuitive. Having children and a mortgage both would seem on the surface to be related to marriage. This relationship also has a high confidence of .97, lift of 1.5, and a relatively high support of 0.17. The fact that these individuals do not have a PEP could be a potentially useful piece of information though since all these characteristics seem to show a customer that would have a large benefit to having a PEP since they are not financially restricted by the financial obligations of children or a mortgage. Targeted marketing at these individuals could potentially be very fruitful.

### Rules Related to High Income

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {age=60s} | {income=very high income} | 0.053 | 0.327 | 5.6 | 32 |
| {pep=YES} | {income=very high income} | 0.050 | 0.109 | 1.9 | 30 |
| {married=NO} | {income=very high income} | 0.028 | 0.083 | 1.4 | 17 |
| {age=60s, married=NO} | {income=very high income} | 0.025 | 0.48 | 8.3 | 15 |
| {age=60s, save\_act=YES} | {income=very high income} | 0.053 | 0.39 | 6.6 | 32 |

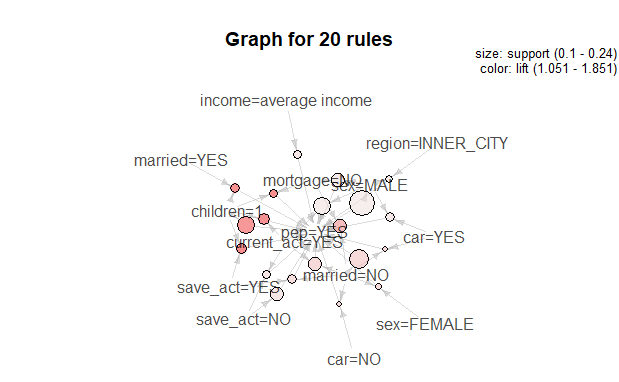


Within the category of rules related to having a very high income there are some additional interesting trends. Being unmarried in your 60s has the highest lift and confidence towards very high income, which is an interesting finding. The relationship between age and high income makes sense since being older allows for more time to accumulate wealth and thus reach this category. Not being married being related to a very high income is a more interesting finding. It is not immediately obvious what the causal relation between these variables is, but one could surmise that not being married gives someone more time to focus on their career and thus reach this high income. Additionally, it is interesting to find that our variable of interest (pep = YES), has a positive lift on having a high income, although the confidence is not very high. This piece of information could be used to help market the PEP product.

### Rules Related to Having a PEP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {children=1, mortgage=NO} | {pep=YES} | 0.12 | 0.85 | 1.9 | 71 |
| {children=1, save\_act=YES} | {pep=YES} | 0.13 | 0.84 | 1.8 | 80 |
| {children=1, current\_act=YES} | {pep=YES} | 0.14 | 0.83 | 1.8 | 84 |
| {married=YES, children=1} | {pep=YES} | 0.12 | 0.83 | 1.8 | 74 |
| {children=1} | {pep=YES} | 0.18 | 0.81 | 1.8 | 110 |

## 



The rules above are the ones that are of the highest interest in determining how to structure further marketing to get more clients with PEP’s. Based on the rules found above, it would be advisable to target individuals with 1 child and without a mortgage. Having 1 child seems to be the most widespread rule, appearing in each of the top 5 rules related to having a PEP. Although, the linkage between these two is not immediately evident, these rules all have very high confidence (>.8) and very high lift (>1.8). Besides having 1 child, some other factors that showed up in determining whether someone has a PEP is having a savings account, having a current account, and being married. However, additional analysis of these additional factors would be needed as their high confidence and lift could simply be due to their connection with having 1 child.

### Rules Related to NOT Having a PEP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {married=YES, children=0} | {pep=NO} | 0.23 | 0.78 | 1.4 | 141 |
| {income=very low income, save\_act=YES} | {pep=NO} | 0.11 | 0.77 | 1.4 | 68 |
| {children=0, save\_act=YES} | {pep=NO} | 0.22 | 0.75 | 1.4 | 131 |
| {income=very low income, married=YES} | {pep=NO} | 0.11 | 0.74 | 1.4 | 67 |
| {income=very low income} | {pep=NO} | 0.16 | 0.70 | 1.3 | 97 |

In addition to obtaining rules related to having a PEP, it is also important to find rules that are related to not having a PEP. This knowledge can be used to help limit marketing to individuals that are unlikely to get a PEP. Based on the rules listed above, being married with 0 children is the strongest indicator that someone will not get a PEP with a support of 0.23, confidence of 0.78, and lift of 1.4. Each of these indicators indicate a strong relationship. Based on a holistic view of these rules, the factors with the biggest correlation to not having a PEP are being married, having 0 children, having a very low income, and having a savings account.

## Conclusions

Individuals with 1 child are shown to be the ones most likely to buy a PEP and are thus a very strong candidate for additional targeting on those that do not have a PEP already. Further research of the causal relationship between these two factors could also be considered to understand the nature of this pattern.

Some of the rules found on the factors leading to high income, including having a PEP and a savings account, are two very interesting findings that could be used to help promote bank services. This connection between these two financial products and financial success could help persuade additional customer to come to the bank.

Lastly, it is evident that having a very low income dissuades many people from getting a PEP. However, these are often some of the people that could get the greatest benefit from these products due to the tax incentives. It is important to educate all incomes of the benefits of financial tools, such as a PEP, so that all individuals can have access.

Knowledge of these trends is very important for a bank to understand so that further marketing can be more customized and targeted to result in better success for both the bank’s bottom line and that of its customers.