Homework #3

Alex Hyman

10/16/2018

## Clean the data

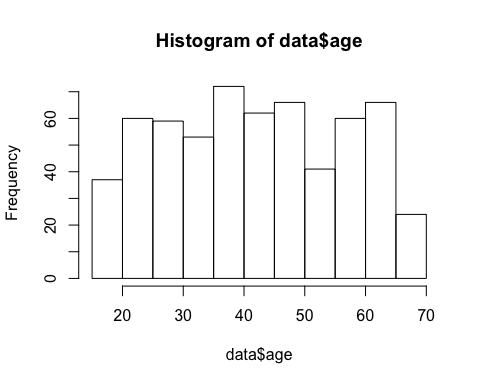
The banking csv file was loaded into the R environment with the read.csv function. The columns of the data frame included the id, age (in years), sex, region of the data object, income (in dollars), marriage status, number of children, car ownership, savings account status, whether the customer has an account with the bank, whether the customer has a mortage, and whether the customer bought the Personal Equity Plan after receiving the mailing campaign. Because association rule algorithms require all columns to be discrete, the id column needs to be deleted and the age, income, number of children columns need to be converted into discrete columns.

The age column was discretized by frequency into three seperate categories: Young, Adult, and Old. I decided to discretize age by frequency becasue a histogram of the different ages of customers who received the promotional campaign in the mail was fairly uniform, therefore both the frequency and interval should be fairly uniform.

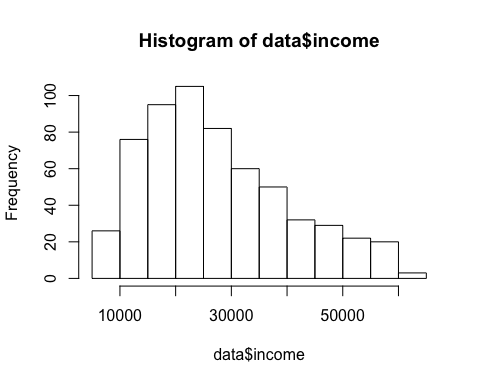
The income column was discretized by interval into three classes: Low, Middle, and Upper. I decided to disctretize income by equal intervals because the income histogram showed a heavy right-skew with a long tail on the positive side. If income was to be discretized by equal frequency, there would a lot of lower-income customers that are grouped into the middle class, and a lot of custoemrs that are middle class grouped into the upper class income group.

The number of children column was discretized into three different categories: None, One, and Multiple. I decided to separate the categories into these categories because I wanted to distinguish between customers that had children and those who had had not had children, and I wanted to see if there were any difference in the behavior of customers with one child versus customers with multiple children. There were also barely enough customers with three children to group all of these customers into one category, so I determined ir would be most appropriate to group them with customers with two children.

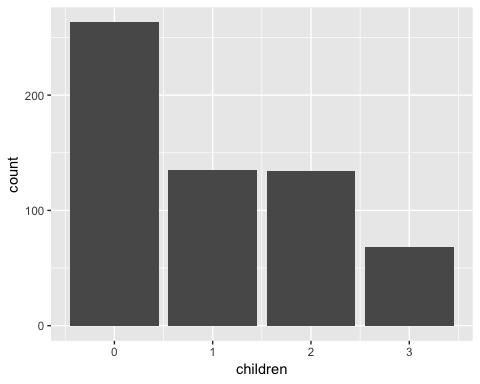
#Read data  
data <- read.csv("bankdata\_csv\_all.csv")  
#Get rid of ID  
data <- data %>% select(-id)  
#hist of age  
hist(data$age)



#Discretize Age  
data$age <- discretize(data$age, labels=c("Young", "Adult", "Old"))  
  
#hist of income  
hist(data$income)



#Discretize Income  
data$income <- discretize(data$income, method = "interval", labels=  
 c("Low", "Middle", "Upper"))  
  
#Children hist  
ggplot(data, aes(children)) + geom\_bar()



#Discretizing children  
data$children <- ifelse(data$children == 0, "None", ifelse(data$children == 1, "One", "Multiple"))  
#Making children a factor  
data$children <- as.factor(data$children)

## Creating Rules

The the apriori function from the arules package was used to mine some association rules from the data set. The target number of rules was between 20-30, and the rules had to include the Personal Equity Plan (PEP) as the RHS item. Having the PEP as the RHS item would mean that the items on the LHS are giving insight to how a customer with this profile is likely to react in terms of the PEP.

To limit the amount of rules mined from the banking dataset, minimum values for support and confidence were set at various levels. The support measurement sets a limit to how prevalent the combination of categories is within the dataset and the confidence informs us of how strong the rule is.

Originally, a minimum support value of 0.2 and minimum confidence value of 0.8 was set. This resulted in zero rules being found that had the PEP as the RHS item. The support and confidence values were then dropped to 0.15 minimum support and 0.75 minimum confidence. These values resulted in 7 rules being formed, all of which occurred more than 100 times in the bank dataset. Finally, the minimum support was reduced 0.10, and the minimum confindence was maintained at 0.75. These final parameters resulted in 28 different rules being creates having the PEP as the RHS item of the rule.

rules <- apriori(data, parameter = list("supp" = 0.1, "conf" = 0.75), appearance = list(rhs=c("pep=YES", "pep=NO")))

rules <- sort(rules, by = "confidence")  
inspect(rules)

## lhs rhs support confidence lift count  
## [1] {married=YES,   
## children=None,   
## save\_act=YES,   
## current\_act=YES} => {pep=NO} 0.1333333 0.9195402 1.692405 80  
## [2] {married=YES,   
## children=None,   
## save\_act=YES,   
## mortgage=NO} => {pep=NO} 0.1216667 0.9125000 1.679448 73  
## [3] {married=YES,   
## children=None,   
## current\_act=YES,   
## mortgage=NO} => {pep=NO} 0.1333333 0.9090909 1.673173 80  
## [4] {sex=FEMALE,   
## married=YES,   
## children=None,   
## mortgage=NO} => {pep=NO} 0.1050000 0.9000000 1.656442 63  
## [5] {married=YES,   
## children=None,   
## save\_act=YES} => {pep=NO} 0.1783333 0.8991597 1.654895 107  
## [6] {married=YES,   
## children=None,   
## mortgage=NO} => {pep=NO} 0.1733333 0.8965517 1.650095 104  
## [7] {married=YES,   
## children=None,   
## car=NO,   
## mortgage=NO} => {pep=NO} 0.1000000 0.8955224 1.648201 60  
## [8] {income=Low,   
## children=Multiple} => {pep=NO} 0.1400000 0.8936170 1.644694 84  
## [9] {income=Low,   
## children=Multiple,   
## current\_act=YES} => {pep=NO} 0.1066667 0.8888889 1.635992 64  
## [10] {children=One,   
## save\_act=YES,   
## current\_act=YES} => {pep=YES} 0.1050000 0.8630137 1.889811 63  
## [11] {children=One,   
## mortgage=NO} => {pep=YES} 0.1183333 0.8452381 1.850886 71  
## [12] {sex=FEMALE,   
## married=YES,   
## children=None,   
## current\_act=YES} => {pep=NO} 0.1000000 0.8450704 1.555344 60  
## [13] {children=One,   
## save\_act=YES} => {pep=YES} 0.1333333 0.8421053 1.844026 80  
## [14] {children=One,   
## current\_act=YES} => {pep=YES} 0.1400000 0.8316832 1.821204 84  
## [15] {married=YES,   
## children=One} => {pep=YES} 0.1233333 0.8314607 1.820717 74  
## [16] {sex=FEMALE,   
## married=YES,   
## children=None} => {pep=NO} 0.1300000 0.8297872 1.527216 78  
## [17] {children=One} => {pep=YES} 0.1833333 0.8148148 1.784266 110  
## [18] {married=YES,   
## children=None,   
## car=NO,   
## current\_act=YES} => {pep=NO} 0.1000000 0.8108108 1.492290 60  
## [19] {married=YES,   
## children=None,   
## car=NO} => {pep=NO} 0.1333333 0.8000000 1.472393 80  
## [20] {married=YES,   
## children=None,   
## current\_act=YES} => {pep=NO} 0.1750000 0.7894737 1.453019 105  
## [21] {sex=FEMALE,   
## children=None,   
## save\_act=YES} => {pep=NO} 0.1150000 0.7840909 1.443112 69  
## [22] {married=YES,   
## children=None} => {pep=NO} 0.2350000 0.7833333 1.441718 141  
## [23] {married=YES,   
## children=None,   
## car=YES} => {pep=NO} 0.1016667 0.7625000 1.403374 61  
## [24] {children=None,   
## save\_act=YES,   
## current\_act=YES} => {pep=NO} 0.1683333 0.7593985 1.397666 101  
## [25] {region=INNER\_CITY,   
## married=YES,   
## children=None} => {pep=NO} 0.1066667 0.7529412 1.385781 64  
## [26] {children=None,   
## car=YES,   
## save\_act=YES} => {pep=NO} 0.1066667 0.7529412 1.385781 64  
## [27] {children=None,   
## save\_act=YES} => {pep=NO} 0.2183333 0.7528736 1.385657 131  
## [28] {children=None,   
## car=NO,   
## save\_act=YES} => {pep=NO} 0.1116667 0.7528090 1.385538 67

## Interesting Rules

Of the rules created from the dataset, I picked five rules that I had found to be the most interesting, all of which have the personal equity plan as the RHS item. These rules and their support, confidence, and lift in the dataset is provided in Table 1.

Table 1 Interesting Rules and Their Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** |
| children=One | pep=YES | 0.1833333 | 0.8148148 | 1.784266 |
| income=Low  children=Multiple | pep=NO | 0.1400000 | 0.8936170 | 1.644694 |
| married=YES  children=None | pep=NO | 0.2350000 | 0.7833333 | 1.441718 |
| children=None  car=NO  save\_act=YES | pep=NO | 0.1116667 | 0.7528090 | 1.385538 |
| children=One  save\_act=YES  current\_act=YES | pep=YES | 0.1050000 | 0.8630137 | 1.889811 |

The first rule I found most interesting is that customers that have only one child are likely to purchase the Personal Equity Plan. I found this rule to be the most interesting because this is such a basic rule, and there is a fairly high confidence, and a very high lift. This is not the most intuitive rule, however it may suggest that adults that have only had one child want to attempt to plan a secure future for that one child. Customers with multiple children may have already began planning for their children and customers without children may not be as worried about their financial security. This rule could be of interest to the business objective because it is generalized enough where it can be used to outreach to new customers instead of only retained customers.

The support of 0.183 that the first interesting rule has given us means that 18.3% of the customers in the dataset have only one child and purchased the Personal Equity Plan after receiving it in the mail. The confidence of 0.814 means that 81.4% of the customers that have only one child, purchased the personal equity plan as well. This was calculated by dividing the support value by the fraction of customers in the dataset that have only one child. The lift of 1.78 means that the occurrence of the rule in the dataset is 1.78 times more common than it would be given that the rule occurs due to random chance. This value was derived by dividing the support value by the probability of a customer having a child and purchasing the PEP, which would be the fraction of people that have only one child multiplied by the fraction of people that bought the PEP. The calculations for each of these statistics are provided below:

**Recommendation:** *Create a marketing campaign targeted towards customers that have given birth to their first child, or customers that have recently given birth. Mailing lists could potentially be acquired by businesses that offer classes for pregnant couples. The clients of these businesses would likely be giving birth to their first child, as customers that have already given birth to a child would have likely already taken the course and would not need to pay for it again.*

The next rule that I found to be the most interesting was that low income customers with multiple children will not purchase the personal equity plan. This would likely be due to the customers not having disposable income to plan for the long-term, and instead spending the money they would have wanted to invest on their children. This rule had a high confidence of 0.894, which would suggest that this is a strong rule. This is an interesting in terms of the business objective because it segments out a large portion of the population that would not be interested in the service.

**Recommendation:** *Bank stops advertising PEP to customers in the lower income bracket that have multiple children.*

The rule that married customers without children do not purchase the PEP was an interesting rule because of the high support and because it was not something I would have guessed. The reason behind this may be because both partners in the marriage are working and donating to their 401k independently, and may be less worried about their financial future than someone with children to support. This is interesting in terms of the business objective because I this customer segment could be a whole new demographic to convince that the PEP is a good product.

**Recommendation:** *Exclude newlyweds from the promotion. Or, depending on how profitable the PEPs are to the bank, a marketing campaign could rollout stating why a PEP is important, especially if the couple plans to have a child.*

Another rule that I found to be interesting is that customers without children, without a car, and with a savings account are not likely to purchase the PEP. This was interesting because I would think that customers that do not have a car or children, but do save money in the form of a savings account would be interested in long-term savings for retirement. However, this rule could be due to customers saving up for big, short-term expenses such as a car, down payment, or a child. This is an interesting rule in terms of the business objective because this demographic is likely to use some form of a banking service (potentially a loan), so a marketing campaign for these customers to use some other service may be a good idea.

**Recommendation**: *Experiment by sending mailers marketing their mortgages and car loans to these customers and experiment on whether these customers respond to any form of additional services.*

The final rule that I found to be interesting was that customers with one child, a savings account, and a current account will purchase the PEP. I found this rule to be interesting because it had the highest lift of all the rules. I also found it interesting because of how similar the LHS is to the previous rule described, but the RHS of the rule is the opposite result. The rule does make sense as these customers have already shown an inclination to save some money, but now that they have a child, they also need to save money for the long -term. This is an interesting rule in terms of the business objective because of the high lift. This lift would mean that this is a fairly intentional rule.

**Recommendation**: *Place a heavy emphasis on current customers with this profile, and maybe spend additional resources trying to retain these customers. They should be a primary target for the PEP.*