Sana Khan

Homework 3 IST 707

Bank Data Homework

First perform the necessary preprocessing steps required for association rule mining, specifically the id field needs to be removed and a number of numeric fields need discretization or otherwise converted to nominal.

**#Importing Data**

**library(dplyr)**

**library(tidyverse)**

**library(arules)**

**library(arulesViz)**

**library(datasets)**

**bankdata <- read\_csv("/Users/gozi/Downloads/bankdata\_csv\_all.csv")**

**#view the data**

**str(bankdata)**

**summary(bankdata)**

**#Clean Up data**

**#Remove ID column as per instructions**

**bankdata <- bankdata[-c(1)]**

**#this is code from the data file in the class material**

**#Discretization age and income variables**

**bankdata$age <- cut(bankdata$age, breaks = c(0,10,20,30,40,50,60,Inf),labels=c("child","teens","twenties","thirties","fourties","fifties","old"))**

**min\_income <- min(bankdata$income)**

**max\_income <- max(bankdata$income)**

**bins = 3**

**width=(max\_income - min\_income)/bins;**

**bankdata$income = cut(bankdata$income, breaks=seq(min\_income, max\_income, width))**

**#convert this variable to a factor**

**bankdata$children <- factor(bankdata$children)**

**str(bankdata)**

Next perform association rule discovery on the preprocessed data. Experiment with different parameters and preprocessing so that you get on the order of 20-30 strong rules, e.g. rules with high lift and confidence which at the same time have relatively good support. Don’t forget to report in details what you have tried.

**rules <- apriori(bankdata, parameter = list(supp = 0.001, conf = 0.9, maxlen = 3))**

**summary(rules)**

**options(digits=3)**

**rules<-sort(rules, by="lift", decreasing=TRUE)**

**inspect(rules[1:20])**

**rules2 <- apriori(bankdata, parameter = list(supp = 0.001, conf = 0.8, maxlen = 3))**

**summary(rules2)**

**options(digits=3)**

**rules2<-sort(rules2, by="lift", decreasing=TRUE)**

**inspect(rules2[1:20])**

Finally, set PEP as the right hand side of the rules, and see what rules are generated.

**rulespep<-apriori(data=bankdata, parameter=list** **(supp=0.01,conf = 0.07),**

**appearance = list(default="lhs",rhs="pep=YES"),**

**control = list(verbose=F))**

**rulespep<-sort(rulespep, decreasing=TRUE,by="lift")**

**inspect(rulespep[1:30])**

I played around with the support and confidence for adjusting the RHS rules but settles on the above code for my analysis

**Rule 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **lhs** | **rhs** | **support** | **Confidence** | **lift** |
| **Income=highincome, children = 3** | **PEP = YES** | **.0133** | **1** | **2.19** |
| **Income = highincome, children=1** | **PEP = YES** | **0.0267** | **1** | **2.19** |
| **region=SUBURBAN, income=highIncome, children=2** | **PEP = YES** | **.0100** | **1** | **2.19** |
| **sex=MALE, region=SUBURBAN, income=highIncome}** | **PEP = YES** | **.0117** | **1** | **2.19** |
| **age=teens, children=2, car=YES** | **PEP = No** | **0.0100** | **1** | **1.84** |

Explanation of the rules:

For the rule **Income=highincome, children = 3,** the support means that just over 1% of the population had a high income and 3 kids and purchased a PEP. While this is low, most of the support rules were low. Confidence is 100% of high income earners with 3 kids purchased a PEP. The lift of 2.19 means that a high income earner with 3 kids is 2.19 more likely to purchase a PEP than a high income earner without kids.

**Analysis/Recommendations:**

Based on this analysis the best customers/cohorts to target would be customers who have between 1 and 3 children, are high income earners and live in the suburbs. These findings feel intuitive with what we would expect for who would purchase a PEP. It makes sense to target those who are high income earners rather than lower income, because they have more expendable income that can be used towards investments. Customers with children are also highly likely to purchase a PEP. This is most likely because these customer are planning long term for their families future, and with children they may want to have investments they can turn to in case of emergencies. Another common theme that was found in the rules was to target customers in a suburban area. This could be because those living in the suburbs are usually families who earn higher incomes since they’re closer to the city. I would also suggest to target individuals who are in their forties. Customers in this age range most likely are thinking about retirement and may not have as much debt or expenses as some of the other age ranges.

I also decided to look into who not to target which was teenagers specifically those who were parents. There could be a few reasons for this, one a teenager with a kid may not have the resources to purchase a PEP. Also, teenagers are not as likely to think about retirement or investments as older customers.