Rohini Shrivastava

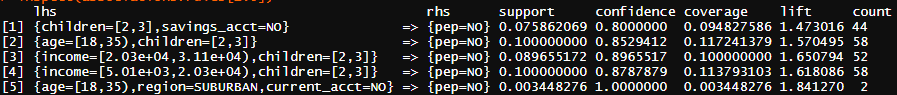
IST 707

Homework 3

The PEP data was downloaded from LMS and imported into R using read\_excel. From there I removed any duplicates in the data and NA rows. After that the PEP dataframe was converted into a transaction in order to run to find the rules. In this dataset the most common criteria were found to be the following items: yes to a current account, yes to a savings account, 0-2 children, yes to married, and no to a mortgage. The data set has 580 rows and 606 columns.

When running the model, we made the support 0.01, the confidence 0.8 with a minimum number of variables as three and maximum as 10. The right hand side used PEP and it was set to “YES” and “NO”.

The five top five rules all had pep be no. Number one was with children 2,3 and no savings account. The support was 0.075, confidence was 0.8, and lift was 1.47. Second rule had the age be between 18 and 35 with children 2,3. The support was 0.1, confidence was 0.85, and lift was 1.57. Rule number three had income between 20300 and 31100 with children 2,3. The support was 0.0896, confidence was 0.896, and lift was 1.65. Rule number four was income 5010 and 20300 with children 2,3. The support was 0.1, confidence was 0.878, and lift was 1.61. And the last was age between 18 and 35, living in a suburban area, with no current account. The support was 0.00344, confidence was 1, and lift was 1.84. Overall, in these rules the most common occurrence for a PEP=No is Children 2,3. It has shown up in four or the top 5 rules. Another big thing noticed are people who make a low income and do not have a current account. These have shown up twice.



We are going to look more detail into rule 3, where income and children affect PEP. The support for rule 3 is 0.08, which means that the popularity of product in all the transactions is about 8%. When the model ran we kept the support at anything greater than 1%. The confidence of the rule is 0.896. The confidence is the likelihood of having the left variables and the right variable. In our model we wanted to see everything above 0.8. Usually a confidence above -.8 is seen as good. The lift was 1.65. The lift tells the likelihood of getting the variables together. A lift higher than 1 means that there is an association between the variables. Since the lift is higher than 1, income, children and pep are all associated with each other.