**HW 3 – Karan Puran Ashar – kashar@syr.edu**

Our main objective was to perform association rule mining. In other words that is to extracts rules from the data. Rules indicate correlation and not causation.

**Preprocessing**

Our first step was to remove the ‘id’ variable from our data frame. This is because we do not need to include the variable to generate our rules. Having the ‘id’ variable as part of our rules will be wrong. The ‘id’ is just an incremental number to uniquely identify each customer.

There are a few numerical variables we would want to discretize. That is to put them in ‘bins’ - group them.  
The first is the income variable. We divided the variable into 3 bins. So, to do this we divide the range between the maximum income and minimum income by 3. We name each of these regions ‘low’, ‘medium’, ‘high’ respectively.  
Then we treat the ‘age’ variable in a similar way. Here the bins are at a space of 10 years. 0 – 10 being the first bin and above 60 being the last bin. I also named these bins appropriately. For example, ‘child’, ‘teens’, ‘twenties’, ‘thirties’, ‘forties’, ‘fifties’, ‘old’.

The ‘children’ variable is in the numeric form. In the numeric form it means that having 2 children is better/greater than having 1 child and so on. We do not want to treat this variable in this manner. So, we will convert the numeric variable to a factor variable. Here we consider the different number of children to be at par and not one superior to the other.

**Interpretation**

As I start experimenting, I come across many rules. At first, I do not fix any parameter to generate specific form of rules. For example, there can be anything on the LHS as well as the RHS. There are a few parameters set which are used to filter out the rules generated. That I support threshold is set to 0.001, the confidence threshold is set to 90%, and the maximum length the rules can be is set to 3.

To being with we can see some obvious rules like when the age is ‘teen’ the ‘income’ is in the ‘low’ category. This seems fair and obvious. We can expect teenagers not to have a high income. Another obvious rule we can see is when the income category is ‘high’ the customer has a savings account. This is also an expected rule. Where else do you expects them to save their money :p.

To get a clearer understanding of the rules I sorted the rules by confidence first and then by the count. The count is an important variable to consider. One rule we see is that when age is in the ‘twenties’ category and number of children are one, the income is ‘low’. This is also not surprising. This is because probably having a child at an early age could be financially difficult.

Now I add a condition to the rule generation process. That is, I want to generate only those rules which have the ‘pep’ variable on the RHS. This is because ‘pep’ is our target variable. We want to make decisions around this variable. For this experiment I also increased the maxlen parameter to 4. I have targeted those customers who do not have a pep account first. This because those are the people who can convert from a no to a Yes with less effort. But this can go both ways based on the business. The people with already a pep account would require more or less convincing, again a decision to be taken by clients.

Top 5 most interesting rules:

1. Children = 3 and Savings Account = No -> pep =No

Support= 0.0367

Confidence = 1

Life = 1.8

These are the customers we should not even try to go after. They should be a straight no. This seems to make sense as well. Not having a savings account when you have the responsibility of 3 other people makes you seem to be carefree.

1. Sex = Female and Children = 3 and income =medium -> pep = No

Support = 0.0283

Confidence = 1

Lift = 1.8

This is something I feel we should go after. This is because the female with 3 children has a medium income. The pep is something she could be interested in to invest and gain which could help support her.

1. Sex= Male and income = low and children = No-> pep=No

Support = 0.0233

Confidence = 1

Lift = 1.8

This for me is also a straight no. We would be wasting our time behind these customers. This is because we a low income and 3 children the man wouldn’t be having any savings left to invest in our pep.

1. Age= Fifties and income = medium and children = 3 -> pep=No

Support = 0.0150

Confidence = 1

Lift = 1.8

This seems like a good set of customers to target. The high age means that the person will be earning since a long time and probably has a good amount of savings. The 3 children probably are old enough to be by themselves. No current pep indicates that he/she could be completely to the pep concept. There is a good chance that he would like to invest his ‘medium’ income somewhere, which could help him in future.

1. Age=Forties and income= low and children= 1 -> pep=Yes

Support = 0.0283

Confidence = 1

Lift = 2.2

These are the people with low income and 1 child. They can seem to be as responsible people. We can say this because they already have a pep despite having a low income. They can be easy to approach. We simply need to provide a better plan than their current one and even a small increase would have them take the plan.

**Support** – Is the ratio of how many times the rule occurred divided by the total number of transactions.

**Confidence** – Is the ratio of how many times the rule occurred divided by the number of times the LHS appeared.

**Lift** – Is the ratio of confidence divided by expected confidence. Where expected confidence is the number of transactions with RHS divided by total number of transactions.

For example –

Rule 🡪 **Age=Forties and income= low and children= 1 -> pep=Yes**

Total number of transactions = 600

Count for the rule = 17

Support = 17/600 = 0.0283

Confidence = 17/17 = 1

Expected Confidence = 274 / 600 = 0.4567

Lift = 1 / 0.4567 = 2.2