**Case Problem**

**Predictive Analytics**

**Module 1 – Lesson 4**

**Professor: Lee**

**Rajiv Kumar**

* 1. Suppose we are examining a customer database that includes our customers who are part of our ice cream-of-the-month club. Their goal is the find groups of customers that are associated with one another. The data consists of demographic data about the customers along with their transaction data. The format of the data table is next. We decide to use association rules to learn more about the associations between customers. Are Association Rules appropriate for this application? Yes or No? Explain.

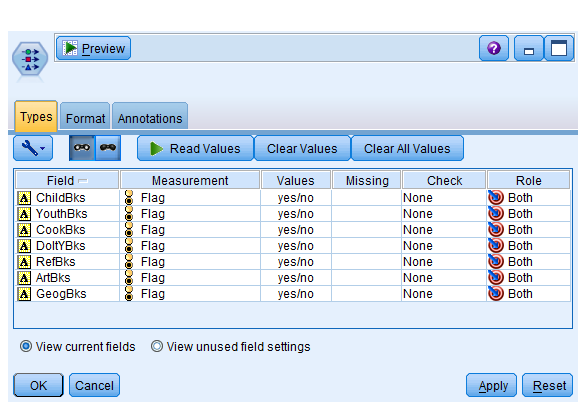
***Answer:***

No, Association rules are used to identify the relationship among the variables not among the rows. Here each customer is represented in a different row and association rule would not be appropriate to find the association between these rows, instead cluster analysis should be used if this association needs to be determined.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Age** | **Gender** | **Education Level** | **Zip Code** | **Purchase Frequency** | **Bought Chocolate?** |
| **34** | **45** | **M** | **Bach** | **80208** | **4** | **Y** |
| **35** | **39** | **M** | **HS** | **80218** | **2** | **Y** |
| **36** | **62** | **M** | **Mast** | **80208** | **7** | **N** |

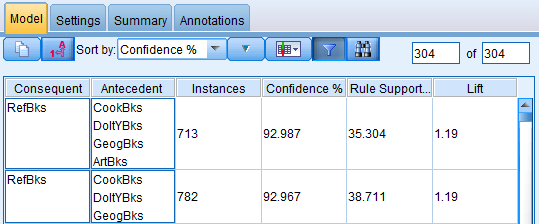
**⁞ ⁞ ⁞ ⁞**

* 1. The dataset **Books.csv** contains information about customer purchases of different types of books.
     1. Read in the csv file with the VAR Course node. Check comma as the delimiter and Read in file names from file. Use a Type node to declare all variables as Flag and set their Role as ‘Both.’



* + 1. Use the Apriori algorithm to find association rules between the different books bought. Interpret the first two rules. Find the # Instances, Confidence %, Rules Support %, and Lift for each rule. Interpret each of these values.

***Answer:***



**Rule 1:**

***Rule Support:***

Rule Support = # of times antecedent occurs / total number of transactions

There are total 663 records where each variable defined in antecedent and consequent for Rule 1 is yes. We have found this by using a select node and specifying the condition to select the records. So # of times antecedent occurs=663

Total number of transactions = 1878

Rule support % =663/1878 =0.3530\*100=35.3%

***Confidence %:***

Confidence%=(# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

We have already found out that # of transaction when both antecedent and consequent is true are 663 using select node.

Now lets find out # of rules where antecedent are true through other select node. We see that there are total 713 such observations.

So confidence % = 663/713\*100=92.98%

***Instances:***

Instances= # of instances where antecdents are true.

We have seen that there are 713 such instances by using a select node.

***Lift :***

Lift = Rule confidence /benchmark confidence where benchmark confidence=# of times consequent occurs/Total # of transactions.

We have seen that Rule confidence = .9298

Benchmark confidence = 1467/1878=.7811

Lift = .9298/.7811=1.19

**Rule 2:**

***Rule Support:***

Rule Support = # of times antecedent occurs / total number of transactions

There are total 727 records where each variable defined in antecedent and consequent for Rule 1 is yes. We have found this by using a select node and specifying the condition to select the records. So # of times antecedent occurs=727

Total number of transactions = 1878

Rule support % =727/1878 =0.3871\*100=38.71%

***Confidence %:***

Confidence%=(# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

We have already found out that # of transaction when both antecedent and consequent is true are 727 using select node.

Now lets find out # of rules where antecedent are true through other select node. We see that there are total 782 such observations.

So confidence % = 727/782\*100=92.96%

***Instances:***

Instances= # of instances where antecdents are true.

We have seen that there are 782 such instances by using a select node.

***Lift :***

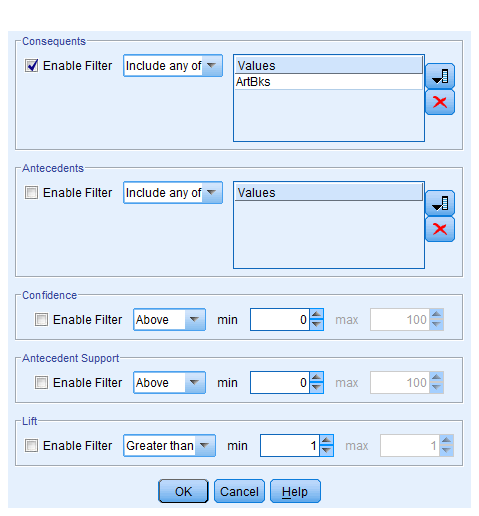
Lift = Rule confidence /benchmark confidence where benchmark confidence=# of times consequent occurs/Total # of transactions.

We have seen that Rule confidence = .9296

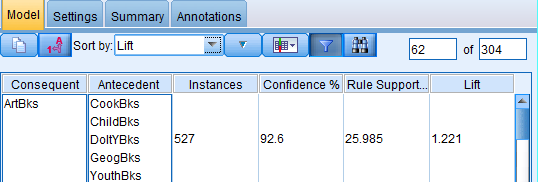
Benchmark confidence = 1467/1878=.7811

Lift = .9296/.7811=1.19

* + 1. Suppose we are most concerned with rules that have the consequent ‘ArtBks.’ Add a filter in the Association Rules Nugget to only include rules with that consequent. What is your top rule rules based on lift? Interpret the lift.



Top Rule:



***Lift :***

Lift = Rule confidence /benchmark confidence where benchmark confidence=# of times consequent occurs/Total # of transactions.

Rule confidence = (# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

=488/527 \*100 =92.59

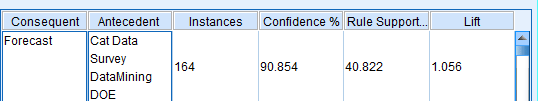
We have seen that Rule confidence = .9259

Benchmark confidence = 1424/1878=.758

Lift = .9259/.28=1.22

* 1. The dataset **StatisticsComCourses.csv** contains information about courses that students have taken from Statistics.Com. Similar to the above problem, use the Apriori algorithm and also the Carma Algorithm to find association rules between these courses.
     1. Interpret the first rule of each algorithm. Find the # Instances, Confidence %, Rules Support %, and Lift for each rule. Interpret each of these values.

**Apriori algorithm:**

****

***Rule Support:***

Rule Support = # of times antecedent occurs / total number of transactions

There are total 149 records where each variable defined in antecedent and consequent for Rule 1 is yes. We have found this by using a select node and specifying the condition to select the records. So # of times antecedent occurs=149

Total number of transactions = 365

Rule support % =149/365 =0.4082\*100=40.82%

***Confidence %:***

Confidence%=(# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

We have already found out that # of transaction when both antecedent and consequent is true are 149 using select node.

Now lets find out # of rules where antecedent are true through other select node. We see that there are total 164 such observations.

So confidence % = 149/164\*100=90.854%

***Instances:***

Instances= # of instances where antecedents are true.

We have seen that there are 164 such instances by using a select node.

***Lift :***

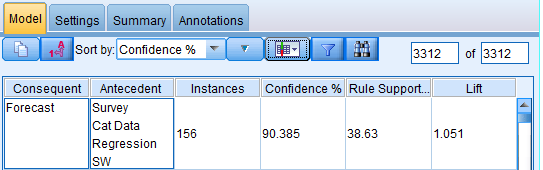
Lift = Rule confidence /benchmark confidence where benchmark confidence=# of times consequent occurs/Total # of transactions.

We have seen that Rule confidence = .9085

Benchmark confidence = 314/365=.860

Lift = .9085/.860=1.05

***Carma Algorithm:***

******

***Rule Support:***

Rule Support = # of times antecedent occurs / total number of transactions

There are total 141 records where each variable defined in antecedent and consequent for Rule 1 is yes. We have found this by using a select node and specifying the condition to select the records. So # of times antecedent occurs=141

Total number of transactions = 365

Rule support % =141/365 =0.3863\*100=38.63%

***Confidence %:***

Confidence%=(# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

We have already found out that # of transaction when both antecedent and consequent is true are 141 using select node.

Now lets find out # of rules where antecedent are true through other select node. We see that there are total 156 such observations.

So confidence % = 141/156\*100=90.38%

***Instances:***

Instances= # of instances where antecedents are true.

We have seen that there are 156 such instances by using a select node.

***Lift :***

Lift = Rule confidence /benchmark confidence where benchmark confidence=# of times consequent occurs/Total # of transactions.

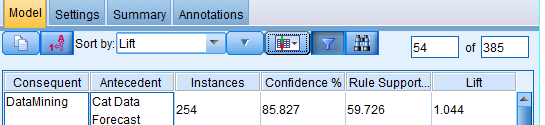
We have seen that Rule confidence = .9038

Benchmark confidence = 314/365=.860

Lift = .9038/.860=1.05

* + 1. Suppose we are most concerned with rules that have the consequent ‘DataMining.’ Add a filter in the Association Rules Nugget to only include rules with that consequent. What is the top rule for each algorithm? Interpret the lifts. What do you notice is different about the results for the 2 algorithms?

***Apriori algorithm:***

******

***Lift :***

Lift = Rule confidence /benchmark confidence , where benchmark confidence=# of times consequent occurs/Total # of transactions.

Rule confidence = (# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

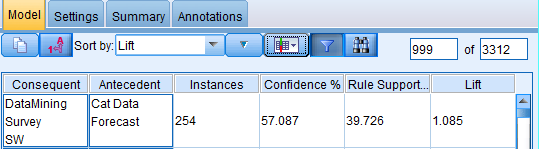
=218/254 \*100 =85.82%

We have seen that Rule confidence = .8582

Benchmark confidence = 300/365=.8219

Lift = .8582/.8219=1.04

***Carma Algorithm:***

****

***Lift :***

Lift = Rule confidence /benchmark confidence , where benchmark confidence=# of times consequent occurs/Total # of transactions.

Rule confidence = (# of transaction when antecedent and consequent is true/ # of rules where antecedents is true)\*100

=145/254 \*100 =57.08%

We have seen that Rule confidence = .5708

Benchmark confidence = 192/365=.526

Lift = .5708/.526=1.08

Both algorithms come up with different top results and Confidence % is significantly different in two algorithm for top results. CARMA also allows rules with multiple consequents in the condition but apriori does not.

**Submit your completed Word file and your stream file on Canvas. In the comments, put the names of the folks who worked on this assignment.**