CHAPTER 11: EXERCISE 1

ASSOCIATION ANALYSIS

NITIN KALÉ, UNIVERSITY OF SOUTHERN CALIFORNIA
NANCY JONES, SAN DIEGO STATE UNIVERSITY

OBJECTIVE

The objective of this exercise is to generate association rules (or affinity) for the survivability of the passengers on RMS Titanic[1].

ACTIVITIES

- Import and prepare data
- Apply data mining algorithms
- Configure predictive models
- Create data visualizations
- Analyze and interpret output from models
- Publish results

SOFTWARE PREREQUISITES

- SAP Predictive Analytics 2.2

[1] https://en.wikipedia.org/wiki/RMS_Titanic

UCC PRODUCTS REQUIRED

- None

DATA SET

- Data file titled *Titanic_E11_1.xlsx*

SCENARIO

Using an Excel data file containing information about the passengers of the Titanic, you will use association analysis to generate rules for their survivability.

ASSOCIATION ANALYSIS

Several predictive models are available in SAP Predictive Analytics. More can be integrated from the *R language*. We would like to discover the *associations* among items. These are presented as rules with values for *support*, *confidence,* and *lift* for each rule

1. We will now do an association analysis (using an *Apriori* algorithm) for the passenger data in the Titanic disaster [2]

    a. Launch SAP Predictive Analytics

    b. Click on Expert Analytics, then on Expert Analytics.

    c. Create a new document. Choose MS Excel as Data Source. *Next.*

    d. Browse for the *titanic_E11_1.xlsx* file. *Create*.

2. We will now launch the prediction capabilities of SAP Expert Analytics

    a. Click on *Predict*. You are now in the *Designer* tab.

[2] http://www.rdatamining.com/examples/association-rules

EPISTEMY PRESS

b. You will see several *Algorithms* such as Regression, Outliers, Time Series, Decision Trees, Neural Network, Clustering and Association. See Figure 1.

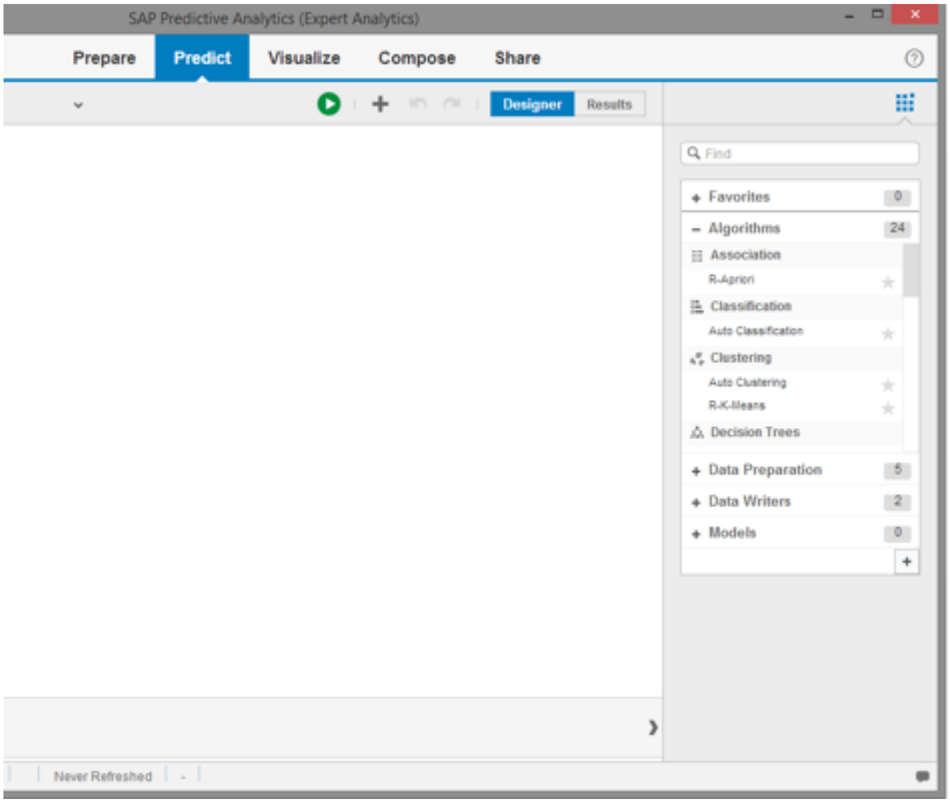


Figure 1

c. And you see the data source *titanic_E11_1.xlsx.*

d. Double-click the *R-Apriori* algorithm. The algorithm is automatically connected to the data source

e. Roll your mouse over the algorithm and click on *Configure Settings*

f. Item Column(s) – Select *Class, Sex, Age* and *Survived*

g. Support: 0.01, (leave confidence at .8)

h. Done

i. Click *Run*

3. The algorithm is now generating the association rules. After the execution is complete, click

OK to review the results.

   a. You see a table of *rules* that were generated.

   b. Now let's rerun the algorithm so that only Survived is on the right hand side of the *rules*.

4. Go to Designer tab (on the right)

   a. Edit the R-Apriori *properties* by selecting configure settings

   b. Click on *Advanced* tab.

   c. In Rhs Item(s) type: *Survived=No,Survived=Yes* (type without spaces in between)

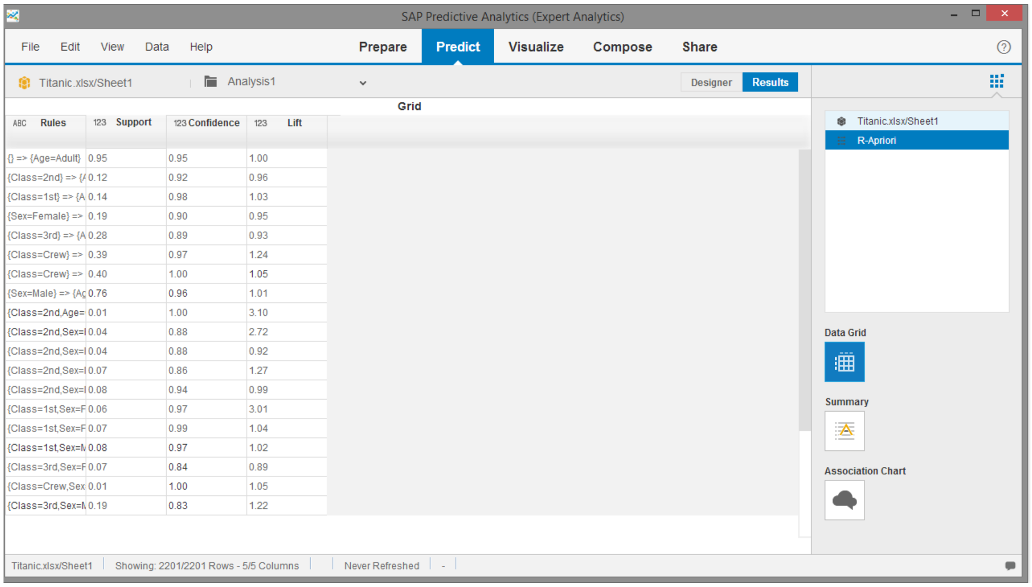
   d. Choose *Default Appearance: Lhs Items*

   e. In the Performance tab, select Sort Type: *Descending*

   f. Done. Run the analysis again.

   g. View the results

   h. You now see the results for all the Rhs (right-hand side or *consequent*) for Survived (No, Yes). See Figure 2.



ABC Rules	123 Support	123 Confidence	123 Lift
{} => {Age=Adult}	0.95	0.95	1.00
{Class=2nd} => {A	0.12	0.92	0.96
{Class=1st} => {A	0.14	0.98	1.03
{Sex=Female} =>	0.19	0.90	0.95
{Class=3rd} => {A	0.28	0.89	0.93
{Class=Crew} =>	0.39	0.97	1.24
{Class=Crew} =>	0.40	1.00	1.05
{Sex=Male} => {Ag	0.76	0.96	1.01
{Class=2nd,Age=	0.01	1.00	3.10
{Class=2nd,Sex=I	0.04	0.88	2.72
{Class=2nd,Sex=I	0.04	0.88	0.92
{Class=2nd,Sex=I	0.07	0.86	1.27
{Class=2nd,Sex=I	0.08	0.94	0.99
{Class=1st,Sex=F	0.06	0.97	3.01
{Class=1st,Sex=F	0.07	0.99	1.04
{Class=1st,Sex=M	0.08	0.97	1.02
{Class=3rd,Sex=F	0.07	0.84	0.89
{Class=Crew,Sex	0.01	1.00	1.05
{Class=3rd,Sex=M	0.19	0.83	1.22

Titanic.xlsx/Sheet1 | Showing: 2201/2201 Rows - 5/5 Columns | Never Refreshed | -

Figure 2: Association Rules

i. Click on *Association Chart*. Here you can see the results in a tag cloud format.

5. Click on *Visualize*

    a. Select component R-Apriori

    b. Convert the attributes *Confidence* , *Support,* and *Lift* to Measures (by right clicking on them and selecting 'create a measure')

    c. Change each measure's aggregation to *None* (from the default *Sum*). You may also wish to rename each measure.

    d. Create a bubble chart (available under scatter plots). X-Axis – *Support*, Y Axis – *Confidence*, Bubble width – *Lift*

    e. Add the Rules from Attributes to *Dimensions: Legend Color*

    f. You can now see the large bubbles indicating the lift for that rule. Lift indicates the strength of a rule over the random co-occurrence of the independent and the dependent variables, given their individual support.

6. We can now export the results of our association analysis

    a. Go to the predict tab

    b. Click on Designer tab

    c. Click on Data Writers drop down list

    d. Add a CSV writer to our analysis by double clicking on the CSV Writer menu.

    e. Edit its properties by selecting configure settings → properties

    f. Choose a File name and type by clicking on Browse. .csv is the default file type.

    g. Save and Close

h. Run the CSV writer

i. You can open the CSV file that was generated to review the results

Question 1: What is meant by support, confidence and lift?

Question 2: Which rule is most dependable within the rules you have found? Why?

Question 3: Why did you set a filter (see 4.c) on the *consequent* in the rules?