Association Analysis

Matthew Olson

BIS 350 Business Analytics Theory & Practice

Overview

This is an analysis of a data set containing over 7,000 product purchases from a small grocery store. In this analysis we will discover purchasing patterns within the data and examine the relationships between products being purchased in the same transactions. To accomplish this we will be using SAS Enterprise Miner 4.2 to create an association data model, once trained, this model will create sets of rules showing how strongly correlated products are with each other based on the purchasing data. Pairs of items often purchased in the same transactions will be targeted for association and modeled. It's important to note the distinction that the percentages discussed are for predicting *future* purchases not for past purchases.

Some common terms that will be seen in this model:

Lift: Lift is a measure of performance of the association model, and is a ratio of the association found in the data vs the association output by the model. A quick example is if the data shows 5% of transaction that contain Snickers bars also contain a Coke, but the model predicts 10%, then the lift is 2.0.

Rule: A rule is a pairing of items created by the model since they are often purchased together, comprised of a Left hand rule and a Right hand rule

Confidence: Confidence is the conditional probability that if a given rule contains item A it will also contain Item B.

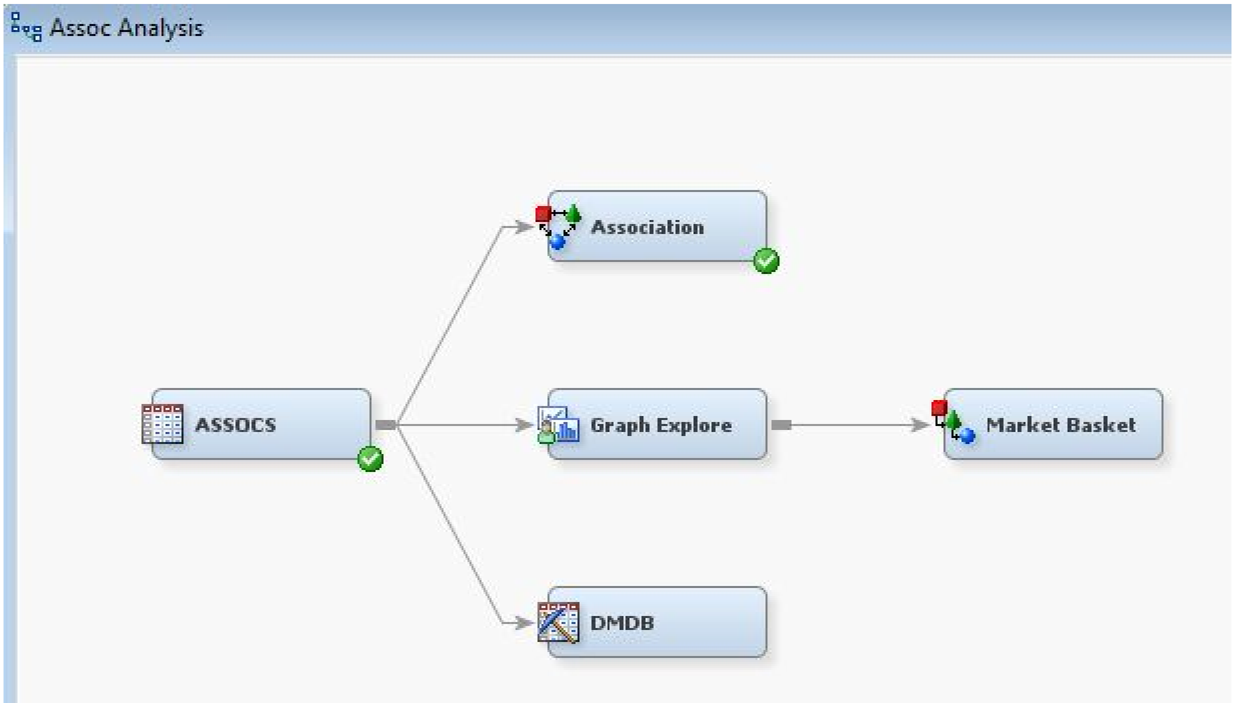
Left hand Rule: The first product of a product association rule.

Right Hand Rule: The second product of an association rule; this is what confidence level predicts.

Support: Likelihood that a purchase contains all of the items in an association rule (Left hand + Right Hand)

The Model

Here is the model we created in SAS Enterprise Miner, since this is a relatively simple association algorithm, data won't need to be standardized or transformed and after a simple analysis can be plugged right into the association model node.

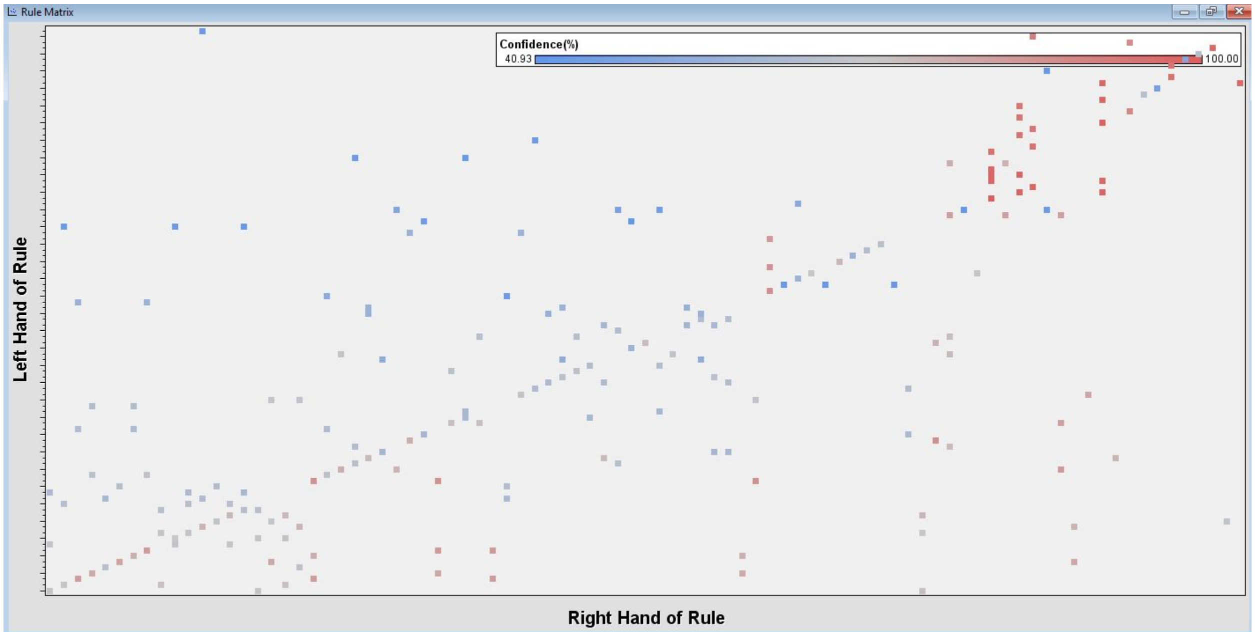


Note:

The Association Node is where the model is created. The other two branches are used for exploratory analysis, this helps to determine what type of analysis is best for this data and also helps to gain a better understanding of the structure of the data. In this case, Association was a better representation than the somewhat similar Market Basket Analysis.

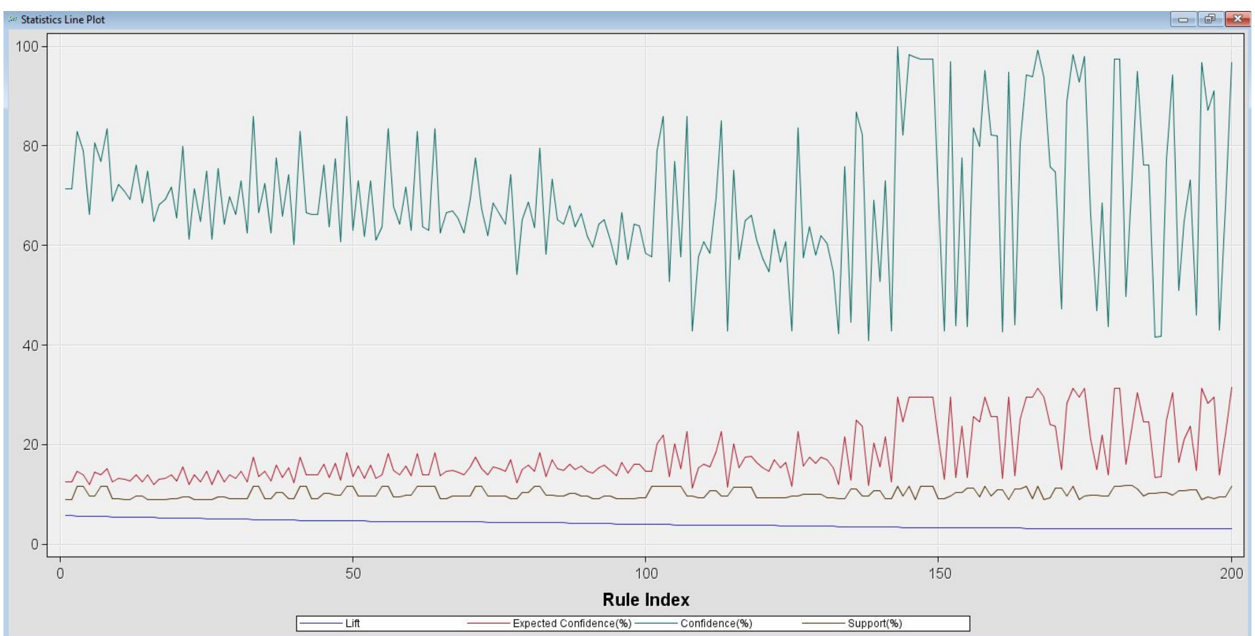
Results

After building and running the Association model we get the initial output, this will help us to understand the accuracy of the model at predicting product association and if there are adjustments that must be made. Of the initial outputs one of the most helpful plots is the Confidence level scatterplot. We talked about confidence earlier and it is one of the most important outputs of this model. Confidence level is what shows us how much of an association there is between the first side of a rule and the second side, if the model creates a pair with low confidence it is likely they weren't often purchased together and we can predict that they most likely won't be purchased together often in the future. Here is the confidence level scatterplot for our model:



As we can see we have two Axis, each containing one side of a given rule by index. The color indicator shows how strong the rule is based on confidence. From this we can interpret that there is a wide distribution of rules and they are not all high or low confidence, this distribution is a good indicator of accuracy for this model.

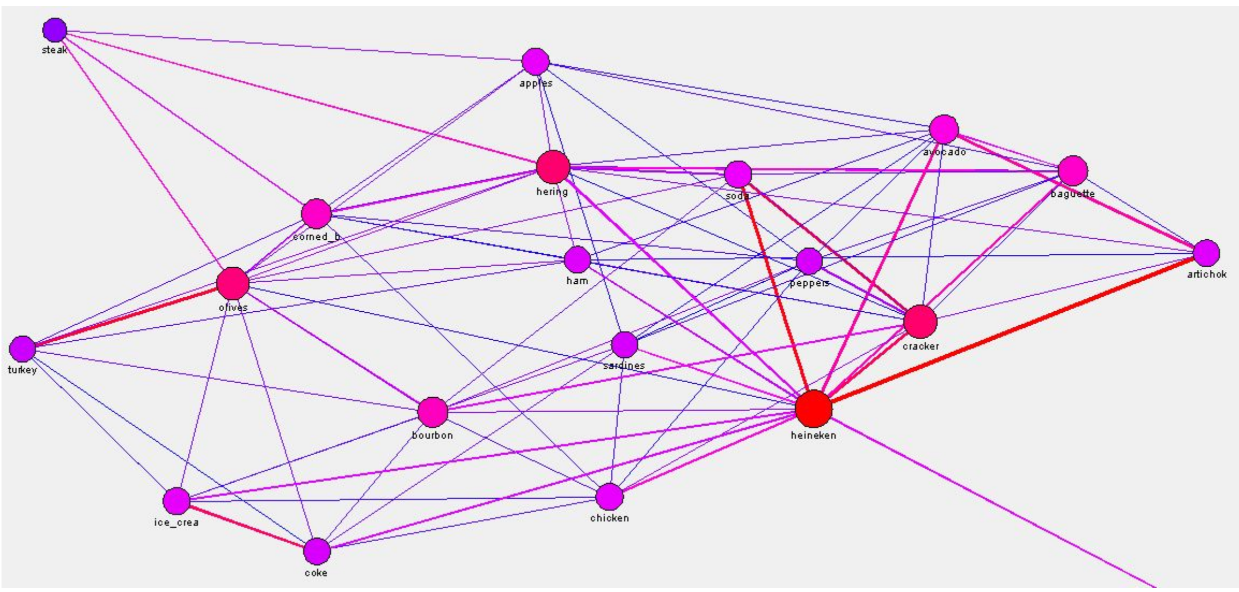
Another output that is useful for analyzing the model is the Rule index. This gives us all of our summary statistics on one place:



Here we can see that it is giving us Lift, Expected confidence, Confidence, and Support for the 200 rules created by the model. Lift is relatively steady throughout the model, which is a good sign and shows that the predictions from the model were not often incorrect when compared to the actual data for testing. As discussed earlier Lift shows how well the rules the model is predicting do when compared to actual transactions, and is essentially an error rate or form of misclassification rate.

Analysis

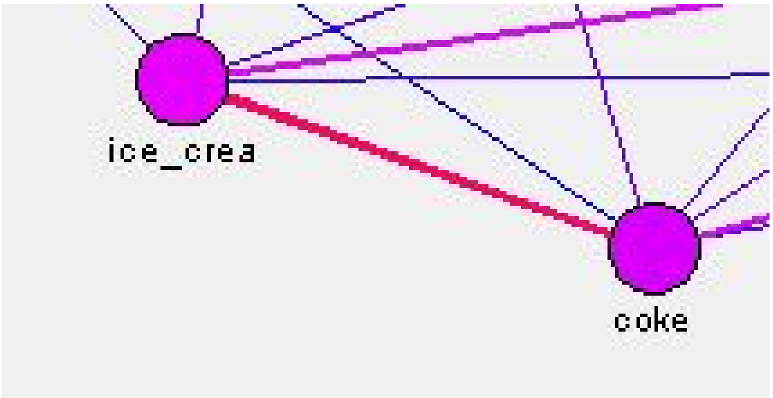
An Overview of associations:



This is a view of the rules generated by the model and their relationships to each other. Clusters of rules demonstrate grouping of products that will be bought together in purchases.

The Larger and more red a product node is, the more this product was purchased in total

The Thicker and more red a node link is, the more the pairing of products was bought together



This shows the strong relationship between Ice Cream and Coke

Now if we view this rule in the Model output we can see it does indeed have very strong associations between Coke and Ice Cream

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
2	29.57	70.29	21.98	2.38	220.00	ice_crea ==> coke
2	31.27	74.32	21.98	2.38	220.00	coke ==> ice_crea

If a given transaction contained Ice cream, there is a 70.29% likelihood that it will also contain Coke.

Ice cream and Coke were one of the strongest rules generated by this model. Let's take a look at some others

Turkey ⇔Olives

2	28.27	46.72	22.08	1.65	221.00	olives ==> turkey
2	47.25	78.09	22.08	1.65	221.00	turkey ==> olives

If a person buys Turkey there is 78.09% likelihood they will also be purchasing Olives

Coke ⇔ Chicken

2	29.57	44.13	13.89	1.49	139.00	chicken ==> coke
2	31.47	46.96	13.89	1.49	139.00	coke ==> chicken

If a person buys Chicken the model predicts 44.13% likelihood of also purchasing a Coke

Soda ⇔ Chicken

2	48.75	78.93	25.07	1.62	251.00	soda ==> cracker
2	31.77	51.43	25.07	1.62	251.00	cracker ==> soda

Soda was purchased much more often than crackers, hence the large differential in confidence levels in the inverse rule(78% vs 51%). Nonetheless, the model predicts a 78.93% chance a person buying Soda will also buy Crackers.

Applications

This data reveals a lot about the purchasing patterns of consumers in this store, but how can we take advantage of this information? Here are some suggested applications of this data and insights.

Strategic Sales: Put the first Item of a highly correlated product paring on sale, It will drive sales of the second product in the pairing even if its not on sale.

Example: If a transaction has Soda there a 78% chance it will contain Crackers. Putting Soda on sale should cause more people to also buy Crackers at their full price

Product Placement: Stock highly correlated Items on opposite ends of the store. Since consumers will often purchase both Items this will increase time spent in store and will increase overall product exposure

Example: Stock the bottles of Coke in the East side of the store and the Chicken on the west side. Customers will spend more time in the store since they will be walking between the two

Advertising: Heavily advertise products that have a highly associated item pair/rule. Advertising an item with no correlations with drive sales of that item alone. Advertising an item with a few highly correlated pairs will drive sales of both that Item and its correlated items.

Example: Advertise Chicken more since it has Soda Coke and Sardines often purchased with it. Increased chicken sales will also mean increased soda coke and sardines sales according to our model.