Data Mining Project

Portfolio Sample

Will Kittredge

TABLE OF CONTENTS

Associations

Objectives

The objective of this part of the project (Part 1: Associations) is to apply data mining algorithms to the *Associations.xlsx* dataset to create association rules. According to IBM, "association rules associate a particular conclusion with a set of conditions." A model from these rules could be used to predict which items might appear together, and to predict the strength of the relationship between the items (2021a). Success will be measured by the amount and quality of the association rules that a business might be interested in. Such rules would likely meet a minimum rule support requirement (enough information to support conclusions from a rule), meet a minimum rule confidence requirement (likeliness of the consequent given the antecedent), have an interesting lift (the antecedent affects the probability of the consequent), and have relatively high deployability.

Data Description

The *Associations.xlsx* dataset is relatively small, consisting of 11 attributes (`ChildBks`, `YouthBks`, `CookBks`, `DoItYBks`, `RefBks`, `ArtBks`, `GeogBks`, `ItalCook`, `ItalAtlas`, `ItalArt`, and `Florence`) and 200 records. Every record has either a one (1) or a zero (0) recorded for each attribute, with one representing the presence of an item and zero representing the absence of an item (see **Figure 1**). Outside of changing some of the default settings in the Excel source node within IBM® SPSS® Modeler, data preprocessing/preparation was not necessary in this case.

Figure 1

First Ten Rows of Associations.xlsx

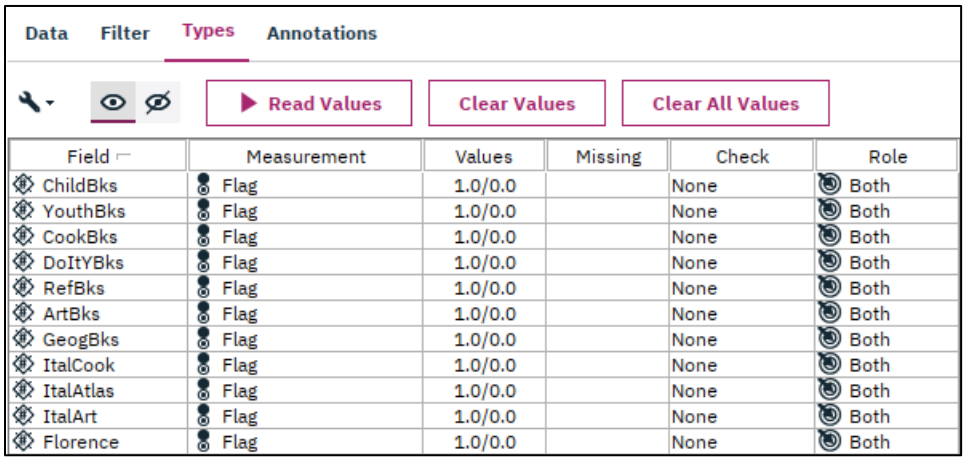
ChildBks	YouthBks	CookBks	DoItYBks	RefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt	Florence
1	1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0
0	0	1	1	0	0	1	1	0	0	0
0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	1	0	0	0	0

Upon importing the Excel document, measurement and role settings are automatically configured for each attribute (visible under the "Types" tab of the Excel node). In this case, the automatic settings were not suitable for our purposes. The "Continuous" measurement type was selected for each attribute, meaning that IBM® SPSS® Modeler is configured to expect any value within the range of 0.0 to 1.0 (visible under "Values" column). This is not representative of our data because only two values are possible: one or zero. The measurement for every attribute was changed to "Flag" to reflect this. Additionally, because each attribute can appear in multiple association rules and as either an input/antecedent or as an output/consequent in each rule, the

role of every attribute was set to "Both." After these changes were made (see **Figure 2**), the dataset was ready for the data mining session.

Figure 2

Corrected Types Settings for Associations.xlsx



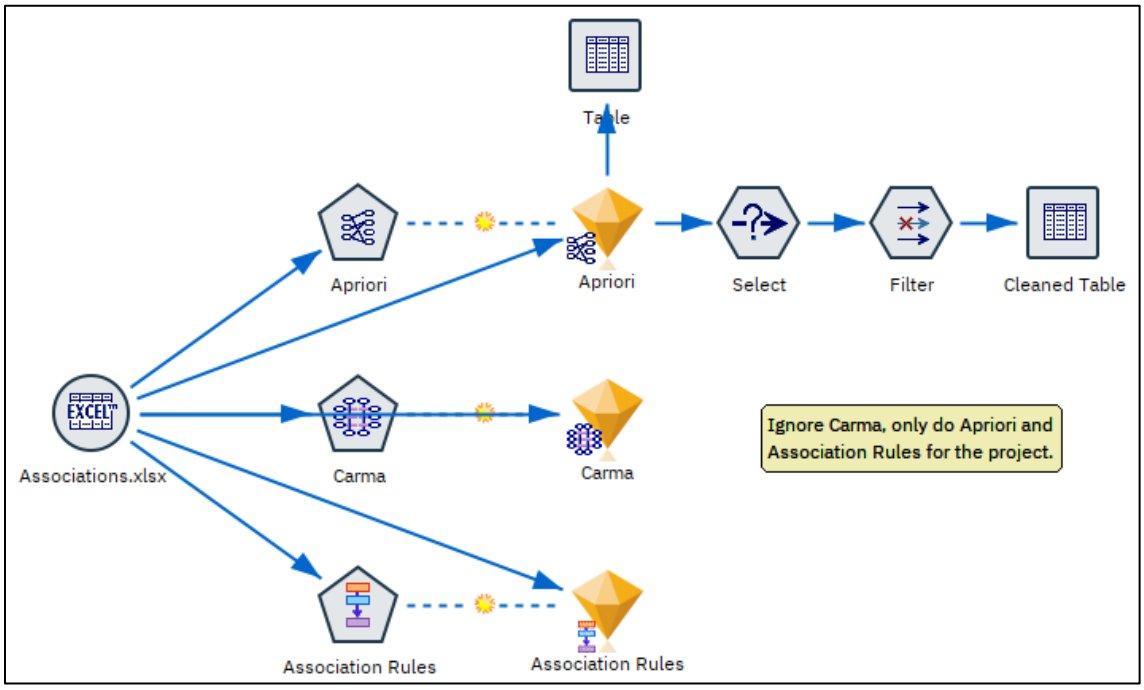
Field	Measurement	Values	Missing	Check	Role
ChildBks	Flag	1.0/0.0		None	Both
YouthBks	Flag	1.0/0.0		None	Both
CookBks	Flag	1.0/0.0		None	Both
DoItYBks	Flag	1.0/0.0		None	Both
RefBks	Flag	1.0/0.0		None	Both
ArtBks	Flag	1.0/0.0		None	Both
GeogBks	Flag	1.0/0.0		None	Both
ItalCook	Flag	1.0/0.0		None	Both
ItalAtlas	Flag	1.0/0.0		None	Both
ItalArt	Flag	1.0/0.0		None	Both
Florence	Flag	1.0/0.0		None	Both

Data Mining Session

Using IBM® SPSS® Modeler, three association models – Apriori, Carma, and Association Rules – were constructed based on the *Associations.xlsx* dataset (see **Figure 3**). Non-default parameter settings for each of the model nodes are visible under their respective subheadings.

Figure 3

Associations Stream in IBM® SPSS® Modeler



Apriori

"The Apriori node extracts a set of rules from the data, pulling out the rules with the highest information content… For large problems, Apriori is generally faster to train…" (IBM, 2021b). Images of the parameter settings for the Apriori node are visible below (see **Figure 4** and **Figure 5**), and each change is summarized in the tables below.

FIELDS		
Option	**Default**	**Current**
Predefined/Custom roles toggle	Use predefined roles	Use custom field assignments (all attributes)

MODEL		
Option	**Default**	**Current**
Minimum rule confidence	80.0%	70.0%
Maximum number of antecedents	5	3

Figure 4

Apriori Node: Fields Parameters

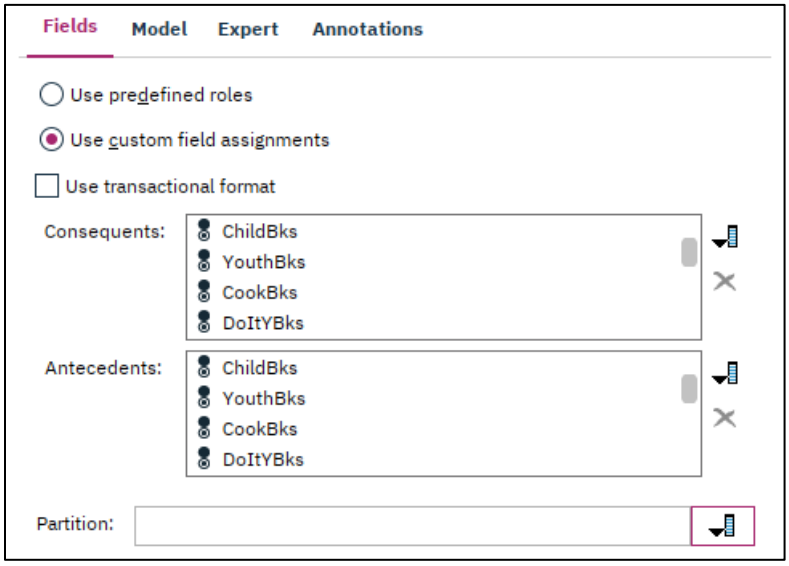
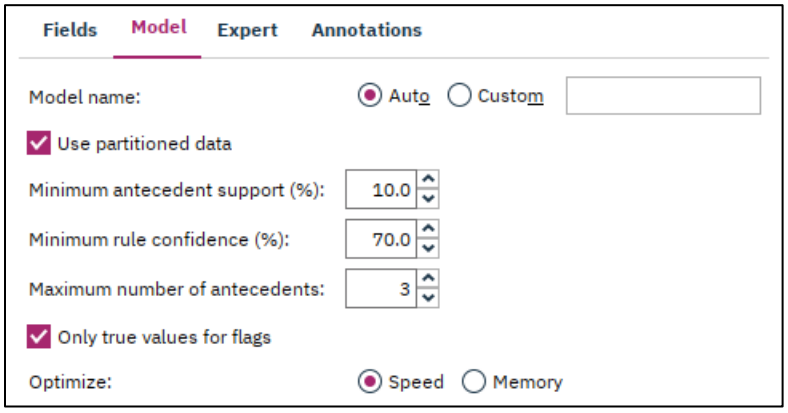


Figure 5

Apriori Node: Model Parameters



Carma

Ultimately, the Carma model produced strange results during class and was discarded as a result, but the parameters set for the model will still be included in this section. Images of the parameter settings for the Carma node are visible below on the next page (see **Figure 6** and **Figure 7**), and each change is summarized in the tables below.

FIELDS		
Option	**Default**	**Current**
Predefined/Custom roles toggle	Use predefined roles	Use custom field assignments (all attributes)

MODEL		
Option	**Default**	**Current**
Minimum rule support	20.0%	5.0%
Minimum rule confidence	20.0%	70.0%
Maximum rule size	10	5

Figure 6

Carma Node: Fields Parameters

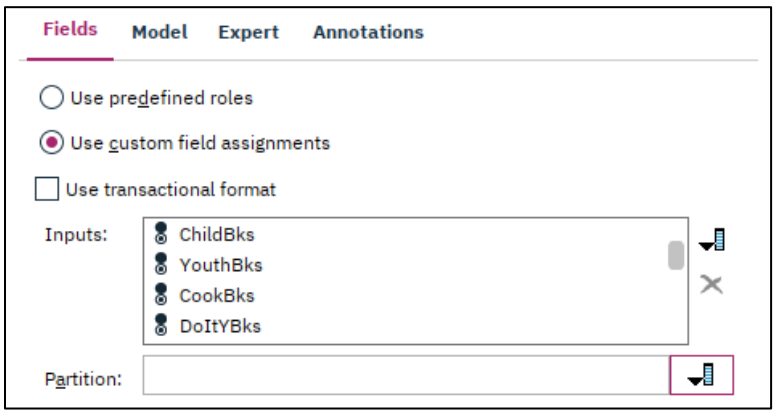
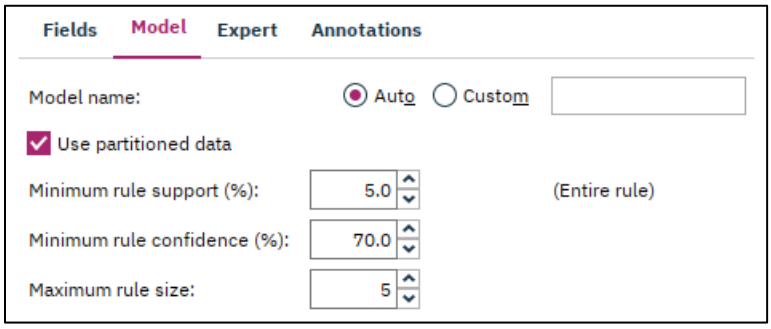


Figure 7

Carma Node: Model Parameters



Association Rules

"The Association Rules Node is similar to the Apriori Node; however, unlike Apriori, the Association Rules Node can process list data" (IBM, 2021b). Images of the parameter settings for the Association Rules node are visible below on the next pages (see **Figure 8**, **Figure 9**, and **Figure 10**); each change is summarized in the tables below.

BUILD OPTIONS (RULE BUILDING)		
Option	**Default**	**Current**
Items per Rule Maximum conditions	5	3
Items per Rule Maximum predictions	1	2
Rule Building Maximum number of rules	1,000	100
Rule Building Rule criterion for top N	Confidence	Deployability
Rule Criterion toggle	Unchecked	Checked
Rule Criterion Confidence	10.0%	70.0%
Rule Criterion Condition Support	5.0%	10.0%
Rule Criterion Lift	2	1

MODEL OPTIONS		
Option	**Default**	**Current**
Maximum number of predictions	3	2
Rule Criterion	Confidence	Deployability

BUILD OPTIONS (OUTPUT)		
Option	**Default**	**Current**
Rule Tables Confidence toggle	Checked	Unchecked
Rule Tables Rule support toggle	Checked	Unchecked
Rule Tables Deployability toggle	Unchecked	Checked
Rule Tables Rules to display	Up to 30	Up to 15
Create a sortable Word Cloud toggle	Unchecked	Checked
Sortable Word Cloud of Rules Default sort	Confidence	Deployability
Sortable Word Cloud of Rules Max rules to display	10	15

Figure 8

Association Rules Node: Build Options (Rule Building) Parameters

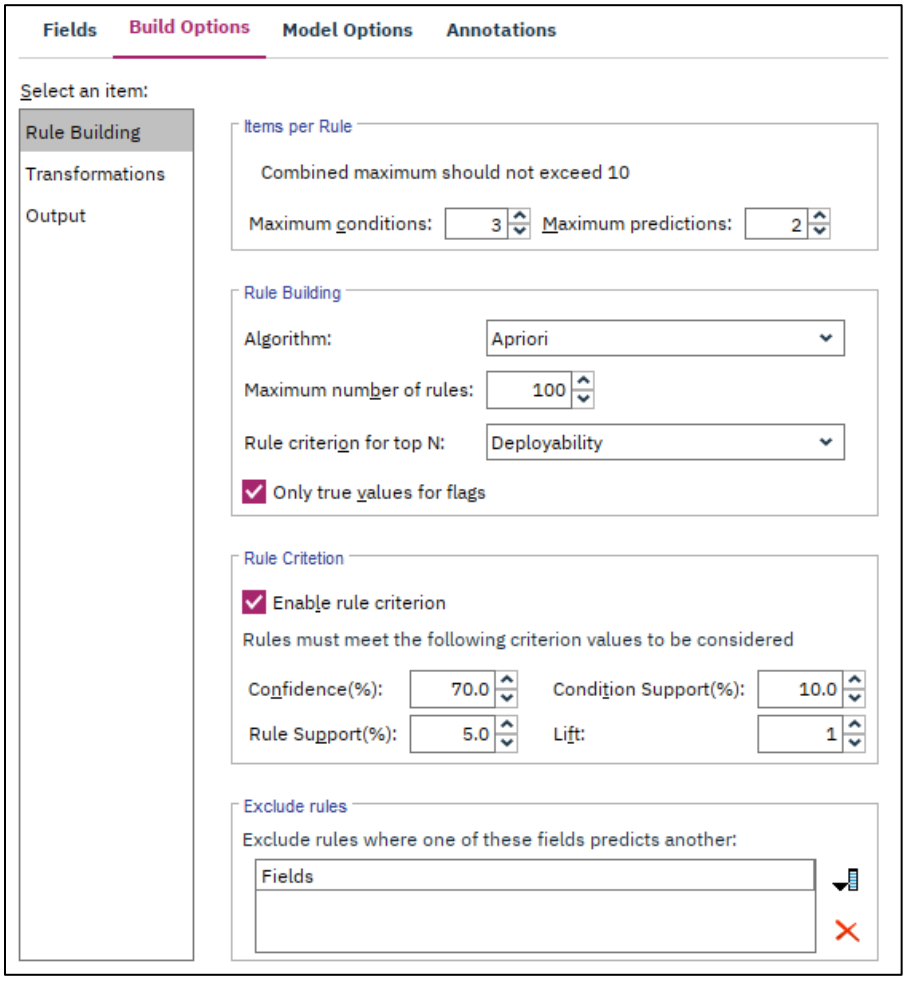


Figure 9

Association Rules Node: Build Options (Output) Parameters

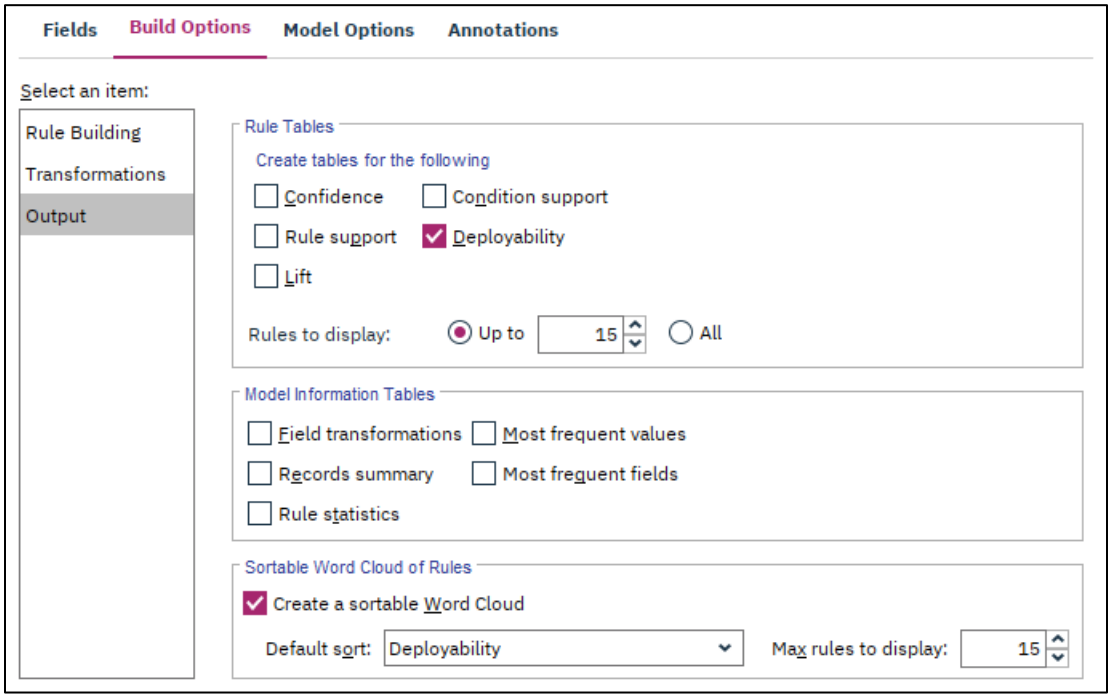
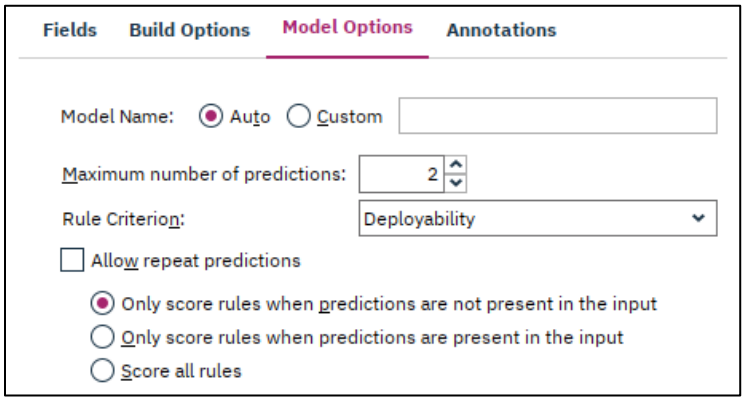


Figure 10

Association Rules Node: Model Options Parameters



Results

The results of the Apriori and Association Rules models are summarized under their respective subheadings.

Apriori

Based on the parameters that were set for the Apriori model, it produced 57 rules. Rule support ranged from 7.5% to 24.5%, confidence from 70.968% to 92.308%, lift from 1.559 to 2.382, and deployability from 1.0 to 9.0 (see **Figure 11**). A table view provides a closer look at the individual rules (see **Figure 12**).

Figure 11

Apriori Model: Summary

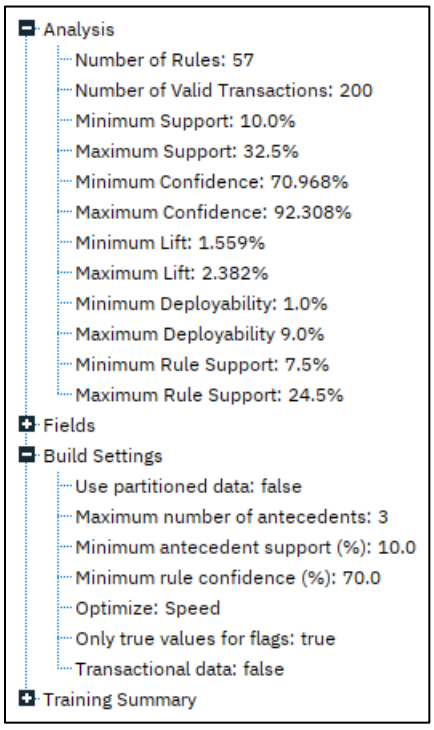
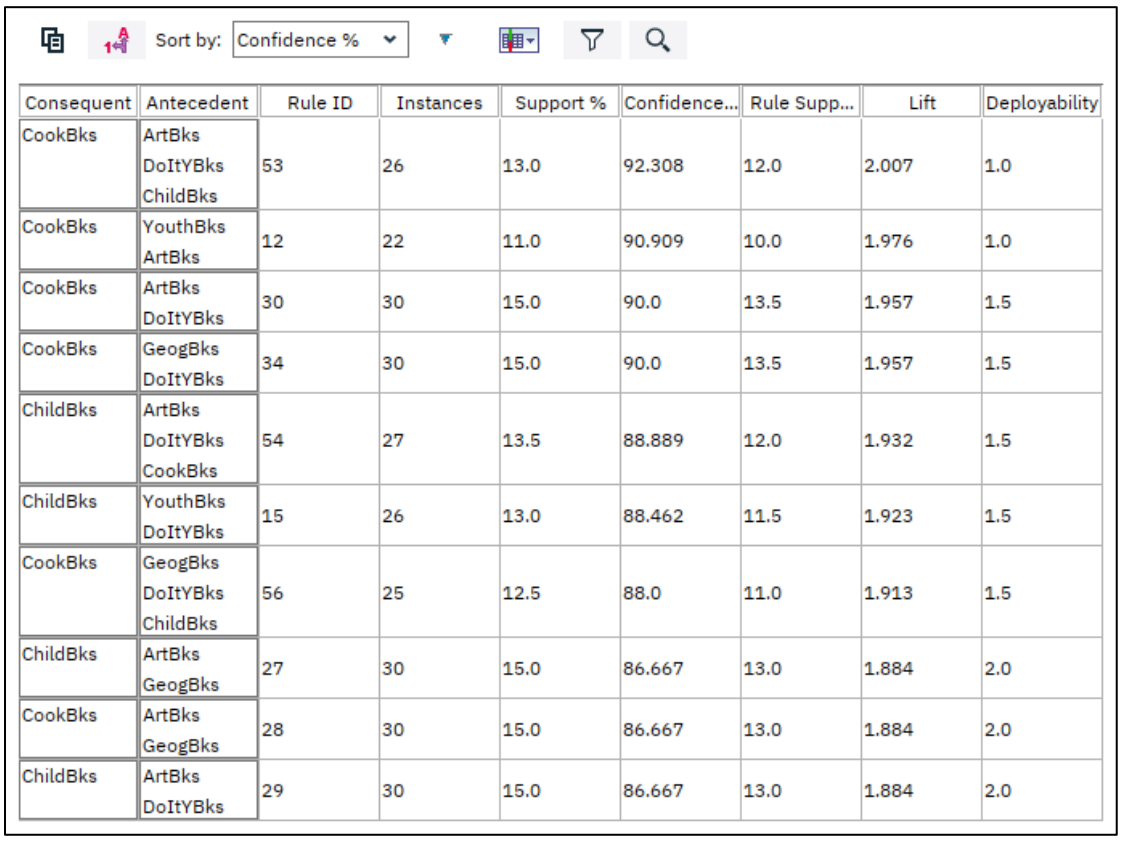


Figure 12

Top Ten Rules by Confidence



Consequent	Antecedent	Rule ID	Instances	Support %	Confidence...	Rule Supp...	Lift	Deployability
CookBks	ArtBks DoItYBks ChildBks	53	26	13.0	92.308	12.0	2.007	1.0
CookBks	YouthBks ArtBks	12	22	11.0	90.909	10.0	1.976	1.0
CookBks	ArtBks DoItYBks	30	30	15.0	90.0	13.5	1.957	1.5
CookBks	GeogBks DoItYBks	34	30	15.0	90.0	13.5	1.957	1.5
ChildBks	ArtBks DoItYBks CookBks	54	27	13.5	88.889	12.0	1.932	1.5
ChildBks	YouthBks DoItYBks	15	26	13.0	88.462	11.5	1.923	1.5
CookBks	GeogBks DoItYBks ChildBks	56	25	12.5	88.0	11.0	1.913	1.5
ChildBks	ArtBks GeogBks	27	30	15.0	86.667	13.0	1.884	2.0
CookBks	ArtBks GeogBks	28	30	15.0	86.667	13.0	1.884	2.0
ChildBks	ArtBks DoItYBks	29	30	15.0	86.667	13.0	1.884	2.0

These rules have the highest confidence values (and lift values closer to the maximum of the range), but the rule support values are towards the lower end of the range and the deployability values are low. The bottom ten rules by confidence, however, seem to have much higher deployability values and some lift values closer to the maximum of the range, along with moderately improved rule support values (see **Figure 13**).

Figure 13

Bottom Ten Rules by Confidence

Sort by: Confidence %

Consequent	Antecedent	Rule ID	Instances	Support %	Confidence...	Rule Supp...	Lift	Deployability
GeogBks	ArtBks ChildBks CookBks	52	31	15.5	70.968	11.0	2.366	4.5
ChildBks	ArtBks	3	53	26.5	71.698	19.0	1.559	7.5
ChildBks	YouthBks CookBks	17	36	18.0	72.222	13.0	1.57	5.0
DoItYBks	RefBks ChildBks	23	36	18.0	72.222	13.0	2.222	5.0
DoItYBks	ChildBks CookBks	39	54	27.0	72.222	19.5	2.222	7.5
CookBks	DoItYBks	8	65	32.5	72.308	23.5	1.572	9.0
ChildBks	RefBks CookBks	26	40	20.0	72.5	14.5	1.576	5.5
ChildBks	YouthBks ArtBks	11	22	11.0	72.727	8.0	1.581	3.0
ChildBks	GeogBks CookBks	36	46	23.0	73.913	17.0	1.607	6.0
CookBks	YouthBks GeogBks ChildBks	41	23	11.5	73.913	8.5	1.607	3.0

Association Rules

The Association Rules model returned the top 15 rules ranked by deployability as requested, and seems to have produced some rules with higher deployability values than the maximum of the Apriori model rules (see **Figure 14** and **Figure 15**). The rules have mixed confidence values in the range of around 70% to 80%, rule support in the range of roughly 17% to 30%, and lift values mostly near 1.2 to 1.3 (but with two outliers with lift values of ≈1.7).

Figure 14

Top 15 Rules by Deployability: Word Cloud

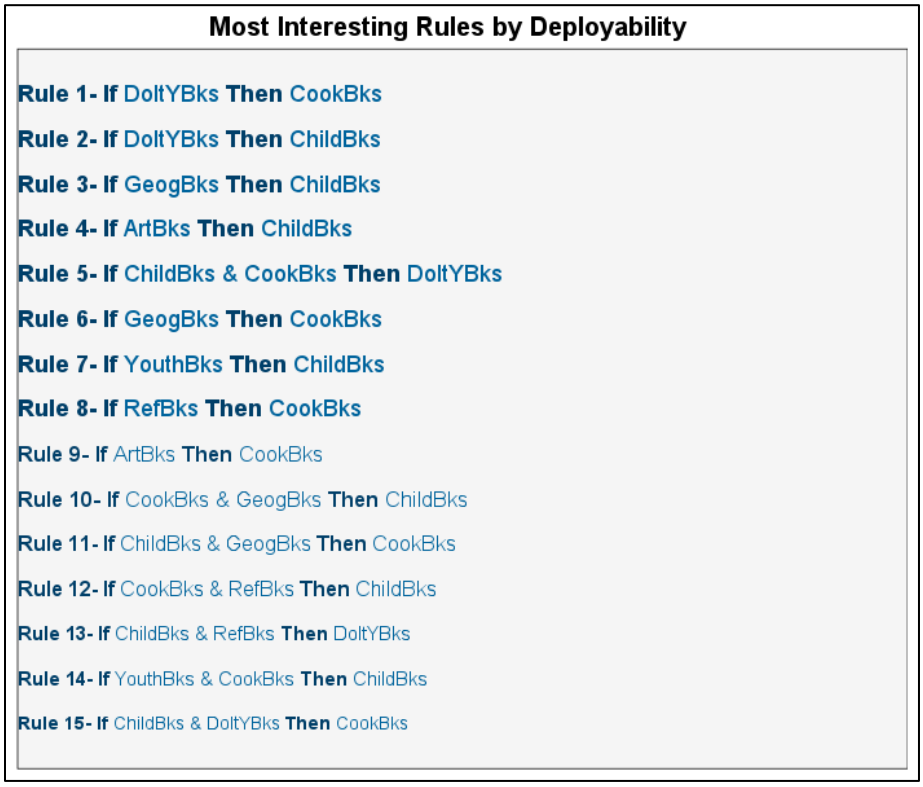


Figure 15

Top 15 Rules by Deployability: Table

Most Interesting Rules by Deployability

					Other Evaluation Statistics			
Rank	Rule ID	Condition	Prediction	Sorted By Deployability(%)	Condition Support (%)	Confidence (%)	Rule Support (%)	Lift
1	1			11.39	41.14	72.31	29.75	1.24
2	2			10.13	41.14	75.38	31.01	1.29
3	3			9.49	37.97	75.00	28.48	1.29
4	4			9.49	33.54	71.70	24.05	1.23
5	5			9.49	34.18	72.22	24.68	1.76
6	6			8.86	37.97	76.67	29.11	1.32
7	7			8.86	34.18	74.07	25.32	1.27
8	8			8.86	34.18	74.07	25.32	1.27
9	9			8.23	33.54	75.47	25.32	1.30
10	10			7.59	29.11	73.91	21.52	1.27
11	11			6.96	28.48	75.56	21.52	1.30
12	12			6.96	25.32	72.50	18.35	1.25
13	13			6.33	22.78	72.22	16.46	1.76
14	14			6.33	22.78	72.22	16.46	1.24
15	15			6.33	31.01	79.59	24.68	1.37

Recommendations

If the objective was to discover association rules that a business may be interested in for some reason such as market basket analysis, I think that the project could be considered successful. Although none of the models produced perfect association rules, there were definitely some rules that were more interesting than others due to higher performance in at least one metric.

Specifics will likely vary depending on the situation and the business question(s) being addressed, but selecting a most important or least important metric to sort rules by (lift, for example) and then pruning the top performers to leave the best all-arounder performers across the other metrics may be a good strategy for selecting rules of further interest. Rule ID 39 in the bottom ten Apriori model rules by confidence (see **Figure 13**) is a potential example. While it has a low confidence value compared to the other rules in the set, all of the values for the other metrics (rule support, lift, deployability) are near the maximum values in the set of all the rules.

Note: Parts 2 and 3 (completed by Brock Byard and Jacob Derenzy) are omitted from this sample.

References

Chen, J. (2024, February 7th). *What is a neural network?* Investopedia. https://www.investopedia.com/terms/n/neuralnetwork.asp

Cart (classification and regression tree) in machine learning. (2023, December 4th). Geeks for Geeks. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/

Hayasaka, S. (2022, February 11). *How many clusters?*. Medium. https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5

C5.0 node. (2024, January 17th). IBM. https://dataplatform.cloud.ibm.com/docs/content/wsd/nodes/c50.html?context=cpdaas

IBM. (2021a, March 4). *Building an association model*. IBM Documentation. https://www.ibm.com/docs/en/sdm/18.0.0?topic=oms-association-rule-scoring-options

Kohonen node. IBM. (n.d.-a). https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=models-kohonen-node

TwoStep cluster analysis. IBM. (n.d.-b). https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=features-twostep-cluster-analysis

IBM. (2021, March 4). *Types of models*. IBM Documentation. https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=mining-types-models

What is random forest? (N.A). IBM. https://www.ibm.com/topics/random-forest

Littler, S. (2024). *CHAID (chi-square automatic interaction detector).* Select StatisticalServices. https://select-statistics.co.uk/blog/chaid-chi-square-automatic-interaction-detector/

Sharma, P. (2024, February 20). *The Ultimate Guide to K-means clustering: Definition, methods and applications*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Sandhu, S. (2016, June 28th). *How does quest compare to other decision tree algorithms?* Stack Exchange. https://datascience.stackexchange.com/questions/12449/how-does-quest-compare-to-other-decision-tree-algorithms