Kathleen Young
IEMS 308, Professor Klabjan
February 8th, 2017

Utilizing Association Rules to Rearrange Dillard's

Executive Summary

The following report outlines the methodology and resulting recommendations of an association rules analysis on customer data from Dillard's department store.

Dillard's is a department store chain with 453 locations nationwide. A single Dillard's can house up to 1,048,576 unique items, or SKUs. In order to organize the stores to better suit their customers' needs, an association rules analysis was applied to the dataset. Association rules uncover the patterns of which items customers tend to buy together in a single market basket. The corporation can use this knowledge to make decisions about which items should go on sale when, where to put items in the store, and so on.

Problem Statement

Dillard's is rearranging its planograms. Ideally, the products (SKUs) will move to locations in the store that maximize the probability that customers will purchase them. In order to analyze how Dillard's products are related to each other, association rules were built based on the millions of Dillard's market baskets.

Assumptions

- Only a subset of the data was analyzed for the association rules. A single store was selected for analysis—store 4903 in Moline, IL. After selecting the store, the data was further parsed into departments. It is assumed that because the SKUs exist within their departments, the most accurate rules would involve SKUs in the same department moving amongst themselves, making this an appropriate way to subsection the data. Because of this approach, the same analysis could be applied to other store locations or larger sets of data.
- Quantity was disregarded for this analysis. That is, if a customer purchased two of a single item, the quantity purchased was treated as one. This simplified the one-hot encoding method used to build the association rules.
- The "trnsact" columns were identified as the following:

c1	sku
c2	store
c3	register
c4	trannum
c5	seq
c6	sale date
c7	stype
c8	quantity
c9	amt
c10	orgprice
c11	sprice
c12	interid
c13	mic

- c14 was disregarded.

Methodology

A significant amount of data exploration was required before the association rules were built. First, the columns of trnsact were identified. The headers of these columns had been removed, so it was necessary to look at the data and match up the columns with the given attribute descriptions.

The second data exploration step was a series of simple database queries. Several variables were queried to see how many existed in the database and what information they contained. For example, COUNT(*) was applied to the dept, sku, and store variables. Additionally, the specific number of SKUs in a select number of departments was examined, as well as which SKUs appeared in which departments. The following important information was uncovered:

Variable	COUNT(*)
Number of rows in trnsact table	120916896
Number of distinct departments	60
Number of distinct SKUs	1048576
Number of distinct store locations	453

At this point, the store location and the department were selected as the best ways to divide up the data. The package MLxtend was used to find the frequent item sets and association rules. The data was one-hot encoded and then the apriori function was used to identify frequent item sets. The minimum support value (the relative frequency that rule appears in a dataset) was set to 0.05. Finally, the association_rules function was used to identify the association rules.

Analysis

Frequent item sets (sets of items that are bought together) were selected based on a minimum support of 5%—that is, the set occurred in at least 5% of transactions.

Frequent item sets and association rules were created for each of the 60 unique departments in the Moline, IL store location (again, this analysis could easily be expanded for a larger dataset). The rules were based on the lift of each of the frequent item sets. The lift is the ratio of the observed support to that expected if X and Y were independent.

A lift of 1 implies that the probability of occurrence of the antecedent and that of the consequent are independent of each other. No rule can be drawn.

If the lift is > 1, the antecedant and consequent are somewhat dependent on each other and may indicate meaningful rules.

If the lift is < 1, the presence of one item has negative effect on presence of other item and vice versa. The items could be substitutes of each other, but no association rule should be drawn.

As such, the min_threshold for establishing an association rule is 1.

The frequent item sets and association rules were reviewed, and the full table of results, along with corresponding support, confidence, and lift scores can be found in appendix A. Unfortunately, the data did not include product names, so the rules are listed based on SKU number.

Conclusions

Based on the association rules provided, I would recommend focusing on reorganizing the following departments: Gary F, BE2, Bora, Coffret, and Annasui. These departments had the most rules. A full list of recommended rules can be found in Appendix A. Items in rules with the highest confidence, support, and lift should be moved closer together in the store so customers that may not have planned on buying both items will be encouraged to do so.

Next Steps

It is clear that this analysis would benefit from more time and more computing power. An obvious next step would be to expand the number of stores examined before splitting the data up into departments. I would like to look at all of the stores in Illinois. Originally, the data was supposed to be divided by department only and nothing else. However, this proved impractical as creating this table took up more space than my schema had available.

Another obvious next step is looking at how departments are placed within the store. A fascinating analysis would involve looking at departments customers shop at, rather than the exact SKUs they buy. This also may be of more general practicality to an actual department store.

Additionally, there are multiples of rules—that is, buying A leads to buying B and buying B leads to buying A. Further analysis would be required to rationalize these double rules into single rules that truly capture the support and confidence.

Finally, based on the rules alone, it is difficult to determine which rules are actually most important. Should rules with the highest lift be considered the most important? How heavily should support be weighted? These questions could be answered with a more complete idea of Dillard's goals.

Appendix A: Association rules

Antecedants	Consequents	Support	Confidence	Lift	Department
183088	173088	0.1	0.5	1.2	GARY F
173088	183088	0.417	0.12	1.2	GARY F
2177157	2107157	0.197	0.261	1.696	BE2
2107157	2177157	0.154	0.333	1.696	BE2
4037330	2107157	0.274	0.219	1.422	BE2
2107157	4037330	0.154	0.389	1.422	BE2
2177157	3597708	0.197	0.348	3.13	BE2
3597708	2177157	0.111	0.615	3.13	BE2
2177157	4037330	0.197	0.348	1.272	BE2
4037330	2177157	0.274	0.25	1.272	BE2
2177157	5289751	0.197	0.261	4.36	BE2
5289751	2177157	0.06	0.857	4.36	BE2
2177157	6386649	0.197	0.304	1.874	BE2
6386649	2177157	0.162	0.368	1.874	BE2
4037330	3597708	0.274	0.25	2.25	BE2
3597708	4037330	0.111	0.615	2.25	BE2
6386649	3597708	0.162	0.316	2.842	BE2
3597708	6386649	0.111	0.462	2.842	BE2
4037330	4017330	0.274	0.219	1.347	BE2
4017330	4037330	0.162	0.368	1.347	BE2
4037330, 2177157	3597708	0.068	0.875	7.875	BE2
3597708, 2177157	4037330	0.068	0.875	3.199	BE2
4037330, 3597708	2177157	0.068	0.875	4.451	BE2
2177157	4037330, 3597708	0.197	0.304	4.451	BE2
4037330	3597708, 2177157	0.274	0.219	3.199	BE2
3597708	4037330, 2177157	0.111	0.538	7.875	BE2
4737469	5649840	0.545	0.167	1.146	BORA
5649840	4737469	0.145	0.625	1.146	BORA
6571028	4737469	0.091	0.6	1.1	BORA
4737469	6571028	0.545	0.1	1.1	BORA
6571028	5649840	0.091	0.6	4.125	BORA
5649840	6571028	0.145	0.375	4.125	BORA
6347532	1508645	0.186	0.313	2.443	COFFRET

1508645	6347532	0.128	0.455	2.443	COFFRET
2419753	6347532	0.081	0.714	3.839	COFFRET
6347532	2419753	0.186	0.313	3.839	COFFRET
3638860	3428013	0.102	0.6	7.35	ANNASUI
3428013	3638860	0.082	0.75	7.35	ANNASUI
3718013	3428013	0.204	0.3	3.675	ANNASUI
3428013	3718013	0.082	0.75	3.675	ANNASUI
3848088	3428013	0.102	0.6	7.35	ANNASUI
3428013	3848088	0.082	0.75	7.35	ANNASUI
3428013	5798907	0.082	0.75	7.35	ANNASUI
5798907	3428013	0.102	0.6	7.35	ANNASUI
3718013	3578013	0.204	0.3	3.675	ANNASUI
3578013	3718013	0.082	0.75	3.675	ANNASUI
3658013	3718013	0.102	0.6	2.94	ANNASUI
3718013	3658013	0.204	0.3	2.94	ANNASUI
3658013	3728013	0.102	0.6	4.9	ANNASUI
3728013	3658013	0.122	0.5	4.9	ANNASUI
3658013	5798907	0.102	0.6	5.88	ANNASUI
5798907	3658013	0.102	0.6	5.88	ANNASUI
3728013	3668013	0.122	0.5	4.9	ANNASUI
3668013	3728013	0.102	0.6	4.9	ANNASUI
3718013	3737654	0.204	0.3	2.94	ANNASUI
3737654	3718013	0.102	0.6	2.94	ANNASUI
3718013	3848088	0.204	0.3	2.94	ANNASUI
3848088	3718013	0.102	0.6	2.94	ANNASUI
3718013	5798907	0.204	0.3	2.94	ANNASUI
5798907	3718013	0.102	0.6	2.94	ANNASUI
3718013, 3848088	3428013	0.061	1	12.25	ANNASUI
3428013, 3718013	3848088	0.061	1	9.8	ANNASUI
3428013, 3848088	3718013	0.061	1	4.9	ANNASUI
3718013	3428013, 3848088	0.204	0.3	4.9	ANNASUI
3848088	3428013, 3718013	0.102	0.6	9.8	ANNASUI
3428013	3718013, 3848088	0.082	0.75	12.25	ANNASUI