Northwestern University

Dillard's Retail Sales Data: Association Rules with Apriori

Aarij Rehman
February 13, 2020

Introduction

Dillards is a major retail chain with stores in the continental United States. Retail chains, in an effort to drive sales, are interested in associations between their products. This information allows companies to rearrange item locations, structure promotions, and keep informed inventory.

In this project, we're interested in the former: rearranging the floor plan. We have multiple connected tables which give information about SKUs, stores, departments, and transactions. The goal is to analyze subsets of data such that we can create meaningful association rules.

Exploratory Data Analysis

The first thing to highlight is the size of the datasets. In total, there are 12.6GB of information and over 120 million samples. The largest of these tables deals with transactions, and it alone is 11.3GB. Because of how large the data is, we had to be careful in how we loaded and analyzed information.

There are 5 tables:

transct.csv – 11.6gb
skstinfo.csv – 1.07gb
skuinfo.csv – 145.3mb
strinfo.csv – 17kb
deptinfo.csv – 1kb

When reading each table, only columns that were relevant to the analysis would be parsed. This allowed me to inspect smaller amounts of data and increase processing times. I decided what was relevant by considering what elements are interesting for an exploratory data analysis and what will be necessary for executing association rules. Similarly, I discarded information that was difficult to interpret, erroneous, or not helpful.

Departments

There are 60 unique departments that exist among all the stores. Not all stores have all departments, and not all stores have the same number of departments either. Based on this, I decided it wouldn't be helpful to subset data by department, as it's possible many individual locations may find the resulting analysis useless

Stores

There are 453 stores across the country that exist in 31 different states. Many Dillards are located in just a few states. Here, we have the top five states with the most Dillards and those respective counts.

State	Count
TX	79
FL	48
AR	27
AZ	26
OH	25

Stores offer a potentially interesting method to subset the data, but similar to the issue with departments, if we consider only stores from a certain region, then many locations won't find this analysis useful.

SKUs

SKUs are Stock Keeping Units; they allow for a chain to assign values to different items to keep track of inventory and sales. This is different from the UPC code of an item which is determined by the manufacturer and identical across all retailers. In total, there are 1,564,178 SKUs that come from 2,393 different vendors and 1,960 brands.

Subsetting data based on SKUs would be very powerful, but because of the type of analysis we're running, it's difficult to run apriorpi with too many SKUs, therefore, it feels counterintuitive to subset based on them but select only a few samples.

Store SKUs

This data shows which stores carry which SKUs and what they're bought for and sold for at each location. I created a column of profit margins and then grouped the data by store. We see that the average profit value is $22 with a standard deviation of $77. It may seem odd to have a standard deviation that allows the profit to become negative, but it's an indication of Dillard's practices. Consider that the 25th percentile of all profits are $-3.76; this is because of the concept of loss leaders.

Stores are willing to sell items at a loss with the hopes that it drives the sales of their other products. Dillards, being a large retail store, likely sells many items at or below cost to enhance their overall performance. The data suggests that *many* items may be losing Dillards money, and this may offer some insight as to why their stores have performed poorly in the past decade.

mean	20.019637
std	77.545588
min	-710.320000
25%	-3.756502
50%	4.294452
75%	22.785820
max	3977.460000

Transactions

The transaction file gives a line for every SKU that was purchased in a given transaction. Because of the sheer number of entries in the table, there are 6 primary keys to help distinguish the rows. Seeing as we're interested in market baskets, we can group entries based on all primary keys except the SKUs, this will construct a single 'transaction' in the sense it is generally interpreted. Doing so, we find that we have 5,028,079 different baskets between the dates of August 8th, 2004 and August 27th, 2005.

One of the primary keys available to us is the date of the transaction. Subsetting by dates would offer incredibly valuable insight for a retail firm, and it is general enough that any results could be applied to all locations.

Subsetting the Data

For this project, I've decided to consider all transactions made in December. Choosing December would offer valuable business insight because of increased purchases due to Christmas. Dillards may be interested in how they can arrange items specifically during the holiday season to boost sales when it matters most.

To further subset the data, we'll look at only the 200 most popular SKUs. This allows us to not overload the apriorpi algorithm and also demand reasonably high support.

We trim our data frame to only include transactions involving the most popular SKUs. After that, we need to 'one-hot' encode the different products so the algorithm can interpret our data. Lastly, we need to create the baskets. As mentioned before, this means grouping by the primary keys – leaving out the SKUs.

After converting our binary encoded basket to Booleans, we're ready to run apriorpi.

Creating Association Rules & Business Insights

Apriori

We will use the **apriori** function from the **mlxtend** package. Creating the rules is as simple as running those two functions in order; however, we need to decide an appropriate minimum support.

Using too low of a value will result in a combinatorial explosion and using too high of a value won't generate enough item sets. After some trial and error, I found that using 0.004 as the minimum support generates many sets while avoiding runtime issues.

Association Rules

For actually creating rules, we will use **association_rules** from the **mlxtend** package. We define a minimum threshold for the lift to be 1. This will eliminate creating any rules that seem counter intuitive. Namely, if a rule has a lift less than 1, it's equivalent to saying the customer is *less* likely to buy a consequent if he has already bought the antecedents.

Rules

The rules with the top hundred lift values can be found in Appendix A. An important thing to consider when creating the rules is it's expected that we see the same rule expressed multiple ways. For example, it's very common to find that Basket X implies Basket Y and Basket Y implies Basket X. If the lift of both rules is high, as is the case for many observations, it can be interpreted that these two products are often purchased together and neither one drives the sales of the either.

Regardless, we can see strong candidates for relocations. Many values of lift are over 10, suggesting that customers are ten times more likely to purchase a consequent if they've purchased the relevant antecedent. These sorts of rules would be most relevant when considering a move.

However, Dillards is interested in the top 100 candidates, not the moves themselves. That being said, they intend to make up to 20 changes. In doing so, though, it's important they consider all dimensions of a rule. Lift is important, but it can be influenced by a small support (as can confidence). Dillards will have to carefully consider their strategic goals when moving items, and with their limited options, they will have to consider the implications of every decision.

Appendix A: Top 100 Rules for Lift

antecedents	consequents	support	confidence	lift
(sku_8616048)	(sku_8956048)	0.004584	0.404255	33.428231
(sku_8956048)	(sku_8616048)	0.004584	0.379095	33.428231
(sku_3968011)	(sku_3898011, sku_3690654)	0.004901	0.169321	26.398822
(sku_3898011, sku_3690654)	(sku_3968011)	0.004901	0.764177	26.398822
(sku_8156822)	(sku_8166822)	0.005146	0.342604	21.249234
(sku_8166822)	(sku_8156822)	0.005146	0.319187	21.249234
(sku_8156822)	(sku_8146822)	0.004743	0.315752	20.761697
(sku_8146822)	(sku_8156822)	0.004743	0.311864	20.761697
(sku_8146822)	(sku_8166822)	0.004833	0.317784	19.709821
(sku_8166822)	(sku_8146822)	0.004833	0.299754	19.709821
(sku_3988011)	(sku_2716578)	0.004523	0.332187	19.345727
(sku_2716578)	(sku_3988011)	0.004523	0.263423	19.345727
(sku_3898011)	(sku_3968011, sku_3690654)	0.004901	0.133001	19.007312
(sku_3968011, sku_3690654)	(sku_3898011)	0.004901	0.700463	19.007312
(sku_3690654)	(sku_3898011, sku_3968011)	0.004901	0.192995	16.705128
(sku_3898011, sku_3968011)	(sku_3690654)	0.004901	0.424252	16.705128

antecedents	consequents	support	confidence	lift
(sku_994478)	(sku_4798193)	0.008474	0.176864	14.435895
(sku_4798193)	(sku_994478)	0.008474	0.691652	14.435895
(sku_3978011, sku_3524026)	(sku_3898011)	0.006252	0.475486	12.902482
(sku_3898011)	(sku_3978011, sku_3524026)	0.006252	0.169647	12.902482
(sku_6706135)	(sku_7596135)	0.005006	0.295619	11.936349
(sku_7596135)	(sku_6706135)	0.005006	0.202123	11.936349
(sku_3978011)	(sku_3898011, sku_3524026)	0.006252	0.118466	11.378468
(sku_3898011, sku_3524026)	(sku_3978011)	0.006252	0.600484	11.378468
(sku_3968011)	(sku_3898011)	0.011553	0.399104	10.829832
(sku_3898011)	(sku_3968011)	0.011553	0.313496	10.829832
(sku_3968011)	(sku_4440924)	0.004268	0.147425	10.400485
(sku_4440924)	(sku_3968011)	0.004268	0.301067	10.400485
(sku_7596135)	(sku_6656135)	0.012471	0.503563	9.883181
(sku_6656135)	(sku_7596135)	0.012471	0.244770	9.883181
(sku_3690654)	(sku_3968011)	0.006997	0.275525	9.518113
(sku_3968011)	(sku_3690654)	0.006997	0.241727	9.518113
(sku_6656135)	(sku_6706135)	0.007973	0.156489	9.241411

antecedents	consequents	support	confidence	lift
(sku_6706135)	(sku_6656135)	0.007973	0.470863	9.241411
(sku_6656135)	(sku_7636135)	0.005330	0.104608	8.681186
(sku_7636135)	(sku_6656135)	0.005330	0.442319	8.681186
(sku_6656135)	(sku_6696135)	0.006191	0.121501	8.023303
(sku_6696135)	(sku_6656135)	0.006191	0.408799	8.023303
(sku_3524026)	(sku_3898011, sku_3978011)	0.006252	0.082836	7.664661
(sku_3898011, sku_3978011)	(sku_3524026)	0.006252	0.578474	7.664661
(sku_3898011)	(sku_3690654)	0.006414	0.174045	6.853099
(sku_3690654)	(sku_3898011)	0.006414	0.252552	6.853099
(sku_3949538)	(sku_9323130)	0.004822	0.073725	6.120109
(sku_9323130)	(sku_3949538)	0.004822	0.400299	6.120109
(sku_3978011)	(sku_3898011)	0.010808	0.204791	5.557061
(sku_3898011)	(sku_3978011)	0.010808	0.293267	5.557061
(sku_8147564)	(sku_3949538)	0.004030	0.335734	5.132979
(sku_3949538)	(sku_8147564)	0.004030	0.061612	5.132979
(sku_8798636)	(sku_2783996)	0.005496	0.292225	5.026259
(sku_2783996)	(sku_8798636)	0.005496	0.094524	5.026259

antecedents	consequents	support	confidence	lift
(sku_8798636)	(sku_5528349)	0.006194	0.329376	4.097842
(sku_5528349)	(sku_8798636)	0.006194	0.077064	4.097842
(sku_3524026)	(sku_3898011)	0.010411	0.137949	3.743298
(sku_3898011)	(sku_3524026)	0.010411	0.282517	3.743298
(sku_3978011)	(sku_3524026)	0.013148	0.249147	3.301147
(sku_3524026)	(sku_3978011)	0.013148	0.174214	3.301147
(sku_3978011)	(sku_6318344)	0.005427	0.102839	3.252377
(sku_6318344)	(sku_3978011)	0.005427	0.171640	3.252377
(sku_803921)	(sku_3898011)	0.004033	0.105551	2.864159
(sku_3898011)	(sku_803921)	0.004033	0.109450	2.864159
(sku_803921)	(sku_3978011)	0.005244	0.137216	2.600083
(sku_3978011)	(sku_803921)	0.005244	0.099359	2.600083
(sku_8718362)	(sku_2783996)	0.004592	0.149350	2.568811
(sku_2783996)	(sku_8718362)	0.004592	0.078977	2.568811
(sku_9288505)	(sku_3949538)	0.004534	0.162515	2.484659
(sku_3949538)	(sku_9288505)	0.004534	0.069321	2.484659
(sku_1310252)	(sku_9708505)	0.004901	0.078435	2.454299

antecedents	consequents	support	confidence	lift
(sku_9708505)	(sku_1310252)	0.004901	0.153369	2.454299
(sku_8718362)	(sku_5528349)	0.005867	0.190816	2.373993
(sku_5528349)	(sku_8718362)	0.005867	0.072987	2.373993
(sku_4108011)	(sku_3978011)	0.008928	0.123150	2.333538
(sku_3978011)	(sku_4108011)	0.008928	0.169169	2.333538
(sku_4108011)	(sku_6318344)	0.005337	0.073621	2.328350
(sku_6318344)	(sku_4108011)	0.005337	0.168793	2.328350
(sku_2783996)	(sku_5137642)	0.004577	0.078729	2.271287
(sku_5137642)	(sku_2783996)	0.004577	0.132052	2.271287
(sku_5528349)	(sku_5137642)	0.006209	0.077244	2.228436
(sku_5137642)	(sku_5528349)	0.006209	0.179117	2.228436
(sku_3559555)	(sku_3524026)	0.004516	0.159481	2.113092
(sku_3524026)	(sku_3559555)	0.004516	0.059837	2.113092
(sku_2698353)	(sku_2938210)	0.004937	0.098209	2.049630
(sku_2938210)	(sku_2698353)	0.004937	0.103044	2.049630
(sku_803921)	(sku_3524026)	0.005910	0.154651	2.049092
(sku_3524026)	(sku_803921)	0.005910	0.078303	2.049092

antecedents	consequents	support	confidence	lift
(sku_1467737)	(sku_3949538)	0.005161	0.134000	2.048711
(sku_3949538)	(sku_1467737)	0.005161	0.078901	2.048711
(sku_3524026)	(sku_6318344)	0.004743	0.062843	1.987470
(sku_6318344)	(sku_3524026)	0.004743	0.150000	1.987470
(sku_1310252)	(sku_3949538)	0.008024	0.128400	1.963090
(sku_3949538)	(sku_1310252)	0.008024	0.122674	1.963090
(sku_2938210)	(sku_3978011)	0.004927	0.102818	1.948289
(sku_3978011)	(sku_2938210)	0.004927	0.093353	1.948289
(sku_3949538)	(sku_6926816)	0.004149	0.063429	1.931434
(sku_6926816)	(sku_3949538)	0.004149	0.126330	1.931434
(sku_1310252)	(sku_1467737)	0.004642	0.074285	1.928864
(sku_1467737)	(sku_1310252)	0.004642	0.120535	1.928864
(sku_803921)	(sku_4108011)	0.005298	0.138630	1.912278
(sku_4108011)	(sku_803921)	0.005298	0.073075	1.912278
(sku_3968011)	(sku_3524026)	0.004163	0.143817	1.905544
(sku_3524026)	(sku_3968011)	0.004163	0.055161	1.905544