Dillard's Point of Sale Data Association Rules Analysis

Charles Marshall
Prof. Klabjan
IEMS 308
13 February 2020

Introduction:

The premise of this business problem is to help Dillard's, a struggling retail department store, rearrange the floors of the store in a way which will place items likely to be bought together next to one another. This movement will promote sales by making it easier for customers to access items which they will be statistically likely to buy together. The analysis of this report will present potential improvements to Dillard's current store format by suggesting items which should be grouped together and provide Dillard's with data about what items are commonly bought together which might be useful when devising sales or inventory plans.

Data Exploration:

One of the most important parts of this project was to understand and explore the data. I sought to initially understand the data by viewing the relational database schema and understanding how certain tables were connected to one another. I found the two most important tables in this project were the transaction data and the SKU information. The transaction data gave raw point of sale data and was used to do the market basket association rules analysis. The SKU information was used to find out more about items included in the rules, including the style, color, size and brand. This was useful for drawing conclusions from the rules.

Additionally, the most challenging part of data exploration for this project was dealing with the large datasets. The transaction data was so large that it could not be loaded into Jupyter Notebook in its entirety, instead it was read in using only the first 7 columns including all the primary keys and the SType, the type of the transaction. The other columns, such as AMT and MIC, were left out because they are not relevant parameters to creating market baskets and doing association rules analysis.

Because the transaction dataset was so large, approximately 121 million rows, it needed to be broken up in a creative way. I decided to subset the data using specific SKU values. Instead of using random SKUs, I decided to use SKUs which corresponded to a specific department, more specifically the Polo Men's Departments. I chose this department because I was familiar with the brand and thought I would be able to more effectively analyze the rules. Additionally, this department had a large number of unique SKU values (about 142 thousand unique values). A large number of unique SKUs in this department allows for a large number of

choices. In the end, I ended up using the 100 most frequently purchased SKUs in order to effectively perform one hot encoding by not overload the apriori algorithm. This action also greatly reduced the size of the data set from over 5 million data points to 250 thousand data points. The 100 unique SKUs and their description are included in Appendices A.

Besides departments, I analyzed many other options to potentially use to subset the data. For instance, I decided against using specific store locations as a parameter because I wanted to create rules which could be applied to all locations, not just one store. Additionally, I considered subsetting the data by only considering a specific time period, specifically the summer of 2005, because I know retail stores often change their configurations when new products come in in different seasons. However, I found that using the summer still left the data set with 23 million data points, and it would therefore need to be broken up more. Both of these approaches are valid and could have been used, but in the end, I decided to go in a separate direction with the business question I answered.

Solve the Problem:

After data exploration, the first step in defining the problem was defining a market basket, which I defined as the variable orderID. The unique market basket is supplied by the composite primary key in the transaction table. In this table, no two rows can have the same SKU, store, register, transaction number, sequence number, and date combination. So, the orderID was created by adding the store, register, transaction number, sequence number, and date. From the composite primary key, I knew that if two unique SKUs had the same orderID then they had to be a part of the same market basket. The two SKUs could not have the same value because then it would break the rules of a primary key having a unique value for each row. The concept of the market basket could be scaled to any sized basket by grouping by the orderID and summing the rows after all the SKUs were one hot encoded. After dropping some no longer relevant columns, the data set is ready to be modeled with association rules.

The apriori function was used to create a set of frequent items based on their support. The important part of this function is to pick a minimum support value which is not too small so that it will blow up, but large enough so that it will not restrict too many rules. After apriori, association_rules was run on the set of frequent items with a minimum lift of one, which indicates independent items, in order to generate rules.

Analysis:

The association_rules function lead to the creation of nearly 440 rules. The description of these rules is as follows:

	support	confidence	lift
count	438.000000	438.000000	438.000000
mean	0.000471	0.073210	9.391543
std	0.000512	0.078111	9.109040
min	0.000104	0.002935	1.015415
25%	0.000129	0.015443	2.189526
50%	0.000251	0.038598	5.583817
75%	0.000531	0.107968	14.065761
max	0.002423	0.344828	37.511159

I found the most important values in this table to be the mean lift, confidence and support. With this information, I chose to look at a subset of rules which had a lift greater than 15 and a confidence greater than 0.25. These values were chosen by using round values which were greater than the mean of both the lift and confidence. The 15 rules from this subset are included in full in Appendix B. The description of these rules is as follows:

	support	confidence	lift
count	15.000000	15.000000	15.000000
mean	0.000326	0.293849	29.343860
std	0.000169	0.033044	6.350264
min	0.000104	0.253521	17.870856
25%	0.000135	0.257898	25.106399
50%	0.000427	0.295238	31.298819
75%	0.000474	0.321598	34.271343
max	0.000493	0.344828	37.511159

From looking at the rules, it is evident that some of the rules “repeat” in a way. For instance, rules 1, 2, and 3 in Appendix A are just different permutations of the same group of SKUs. Using this logic, the rules were broken down into different groups or market baskets:

Basket 1 SKUs: {1049441,1589441,439441}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
68584	439441	4505	404	400009441043	2BAO 581965	NAVY	34	1	5745232	POLO FAS
163747	1049441	4505	404	400009441104	7BAO 581972	UNFR KHAKI	34	1	5745232	POLO FAS
248103	1589441	4505	404	400009441158	3BAO 581965	SAND	34	1	5745232	POLO FAS

Basket 2 SKUs: {1869441,459441,1069441}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
71637	459441	4505	404	400009441045	2BAO 581965	NAVY	36	1	5745232	POLO FAS
166877	1069441	4505	404	400009441106	7BAO 581972	UNFR KHAKI	36	1	5745232	POLO FAS
291253	1869441	4505	404	400009441186	3BAO 581965	SAND	36	1	5745232	POLO FAS

Basket 3 SKUs: {2078810,1571243,5901243,4912330,9972087,2108810}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
245379	1571243	4505	302	400001243157	3VPP 650185	SPA BLUE	ONE	1	5715232	POLO FAS
323397	2078810	4505	302	400008810207	5VPP 651649	BSC WHITE	ONE	1	5715232	POLO FAS
769457	4912330	4505	302	400002330491	3VPP 656523	VALOR RED	ONE	1	5715232	POLO FAS
925021	5901243	4505	302	400001243590	7VPP 656475	OASIS YELL	ONE	1	5715232	POLO FAS
1560073	9972087	4505	302	400002087997	2VPP 650100	RELAY BLUE	ONE	1	5715232	POLO FAS

Basket 4 SKUs: {5709431,3269431,6149446}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
510957	3269431	4505	404	400009431326	3BRG 581965	SAND	36	1	5745232	POLO FAS
894494	5709431	4505	404	400009431570	2BRG 581965	NAVY	36	1	5745232	POLO FAS
963876	6149446	4505	404	400009446614	7BRG 581972	UNFR KHAKI	36	1	5745232	POLO FAS

Basket 5 SKUs: {3279431,5719431,6169446}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
512447	3279431	4505	404	400009431327	3BRG 581965	SAND	38	1	5745232	POLO FAS
896023	5719431	4505	404	400009431571	2BRG 581965	NAVY	38	1	5745232	POLO FAS
966969	6169446	4505	404	400009446616	7BRG 581972	UNFR KHAKI	38	1	5745232	POLO FAS

Basket 6 SKUs: {4898351,5228351,5549233,5748351}

	SKU	Dept	ClassID	UPC	Style	Color	Size	Packsize	Vendor	Brand
767035	4898351	4505	107	400008351489	2AEB 793365	CLSC BLUE	L	1	5745232	POLO FAS
818845	5228351	4505	107	400008351522	5AEB 793365	CLSC WHITE	L	1	5745232	POLO FAS
869386	5549233	4505	107	400009233554	6AEB 790003	W/W INDIGO	L	1	5745232	POLO FAS
900529	5748351	4505	107	400008351574	7AEB 793365	BLU/WHT STR	L	1	5745232	POLO FAS

From the descriptions of the SKUs, one thing that can easily be seen is that it is often hard to decipher exactly what each item is because there is no intuitive label. Because of this, we are often left to make an educated guess based on the sizes. For instance, I would guess that Basket 3 items are socks, ties, or some other item which is one size fits all. I would guess that basket 6 items are shirts because of the L label and basket 5 items are shorts because of the size 38.

Another thing that is evident is that the items that are grouped together and form rules together are items which are the same in every dimension except for color or style. This would suggest that people shop for similar items at the same time. For instance, someone would go to the Polo Men's Department store looking for shorts and would be likely to buy multiple pairs of shorts in different colors or styles, but the same size. This would suggest that the department should attempt to keep similar clothing items (i.e. shorts style 1 and style 2) close to one another in the store. Additionally, the department should try to make sure similar sized items are near one another in different colors and styles so to increase the ease for the customer to look at different options.

Appendix A: 100 SKUs used for Polo Men's Department

100 SKUs used from Polo Men's Department with Description											
	SKU	Style	Color	Size	Brand		SKU	Style	Color	Size	Brand
0	439441	2BAO 581965	NAVY	34	POLO FAS	50	5599233	6AEB 791004	INDIGO	L	POLO FAS
1	459441	2BAO 581965	NAVY	36	POLO FAS	51	5609233	6AEB 791004	INDIGO	M	POLO FAS
2	469441	2BAO 581965	NAVY	38	POLO FAS	52	5629233	6AEB 791004	INDIGO	XL	POLO FAS
3	907523	2ANJA 796964	WHITE	L	POLO FAS	53	5679400	7 433264	CONCORDE C	L	POLO FAS
4	936370	2A 796812	BLUE/GREEN	L	POLO FAS	54	5689400	7 433264	CONCORDE C	XL	POLO FAS
5	947523	2ANJA 796964	WHITE	XL	POLO FAS	55	5699695	4A 796148	STEWART BL	L	POLO FAS
6	992340	9 430070	BLACK	M	POLO FAS	56	5709431	2BRG 581965	NAVY	36	POLO FAS
7	1049441	7BAO 581972	UNFR KHAKI	34	POLO FAS	57	5719431	2BRG 581965	NAVY	38	POLO FAS
8	1069441	7BAO 581972	UNFR KHAKI	36	POLO FAS	58	5748351	7AEB 793365	BLU/WHT STR	L	POLO FAS
9	1072340	9 430070	BLACK	L	POLO FAS	59	5753596	0A 796772	MALI YELLO	L	POLO FAS
10	1079441	7BAO 581972	UNFR KHAKI	38	POLO FAS	60	5759695	4A 796148	STEWART BL	XL	POLO FAS
11	1152340	9 430070	BLACK	XL	POLO FAS	61	5798351	7AEB 793365	BLU/WHT STR	XL	POLO FAS
12	1232340	9 430070	BLACK	XXL	POLO FAS	62	5901243	7VPP 656475	OASIS YELL	ONE	POLO FAS
13	1571243	3VPP 650185	SPA BLUE	ONE	POLO FAS	63	6093596	9A 796771	EUCALYPTUS	L	POLO FAS
14	1589441	3BAO 581965	SAND	34	POLO FAS	64	6119446	7BRG 581972	UNFR KHAKI	34	POLO FAS
15	1677524	1ANJA 790004	BSR WHITE	L	POLO FAS	65	6123596	9A 796771	EUCALYPTUS	XL	POLO FAS
16	1707524	1ANJA 790004	BSR WHITE	XL	POLO FAS	66	6149446	7BRG 581972	UNFR KHAKI	36	POLO FAS
17	1869441	3BAO 581965	SAND	36	POLO FAS	67	6169446	7BRG 581972	UNFR KHAKI	38	POLO FAS
18	2078810	5VPP 651649	BSC WHITE	ONE	POLO FAS	68	6213596	8A 796771	CLASSIC BU	L	POLO FAS
19	2108810	3VPP 651649	NUBUCK	ONE	POLO FAS	69	6246359	2BCPP 651649	NAVY	ONE	POLO FAS
20	2672329	7 047567	SOLEIL	XL	POLO FAS	70	6433893	1AJ 013506	CANYON SAG	L	POLO FAS
21	2967523	5ANJA 796983	LT BLUE MU	L	POLO FAS	71	6447365	1 431002	INK	L	POLO FAS
22	3249431	3BRG 581965	SAND	34	POLO FAS	72	6457365	1 431002	INK	XL	POLO FAS
23	3269431	3BRG 581965	SAND	36	POLO FAS	73	6486359	3BCPP 651649	NUBUCK	ONE	POLO FAS
24	3279431	3BRG 581965	SAND	38	POLO FAS	74	6673596	8A 796771	CLASSIC BU	XL	POLO FAS
25	4343643	2A 792441	WHITE	L	POLO FAS	75	6703893	1AJ 013506	CANYON SAG	XL	POLO FAS
26	4353643	2A 792441	WHITE	M	POLO FAS	76	6806359	5BCPP 651649	WHITE	ONE	POLO FAS
27	4373643	2A 792441	WHITE	XL	POLO FAS	77	6846359	6BCPP 651649	BLACK	ONE	POLO FAS
28	4403643	3A 792441	BLUE/WHITE	L	POLO FAS	78	6863893	4AJ 013507	SPRING NAV	L	POLO FAS
29	4421180	0VZ 586487	CLASSIC KH	L	POLO FAS	79	6872339	0 440010	WHITE	L	POLO FAS
30	4451180	0VZ 586487	CLASSIC KH	XL	POLO FAS	80	6952339	0 440010	WHITE	XL	POLO FAS
31	4898351	2AEB 793365	CLSC BLUE	L	POLO FAS	81	7716294	9A 796815	BRIGHT YEL	L	POLO FAS
32	4912330	3VPP 656523	VALOR RED	ONE	POLO FAS	82	7823749	7A 796771	ARMADILLO	L	POLO FAS
33	4994599	4GZ 782633	U KHAKI	34 30	POLO FAS	83	7926294	8A 796815	CHATHAM BL	L	POLO FAS
34	5024599	4GZ 782633	U KHAKI	34 32	POLO FAS	84	8013602	5AF 796635	NAVY/RED/W	L	POLO FAS
35	5104599	4GZ 782633	U KHAKI	36 30	POLO FAS	85	8057414	1 438188	CARMEL PIN	L	POLO FAS
36	5124599	4GZ 782633	U KHAKI	36 32	POLO FAS	86	8366288	6ANF 796616	BLUE/BLACK	L	POLO FAS
37	5154599	4GZ 782633	U KHAKI	38 30	POLO FAS	87	8533189	9 435866	WICKET YEL	L	POLO FAS
38	5174599	4GZ 782633	U KHAKI	38 32	POLO FAS	88	8613189	9 435866	WICKET YEL	XL	POLO FAS
39	5228351	5AEB 793365	CLSC WHITE	L	POLO FAS	89	8852339	9 440010	BLACK	L	POLO FAS
40	5238351	5AEB 793365	CLSC WHITE	M	POLO FAS	90	8932339	9 440010	BLACK	XL	POLO FAS
41	5258351	5AEB 793365	CLSC WHITE	XL	POLO FAS	91	9002262	4 045692	WHITE	L	POLO FAS
42	5291167	4BJC 583808	BASIC SAND	34	POLO FAS	92	9042262	4 045692	WHITE	XL	POLO FAS
43	5321167	4BJC 583808	BASIC SAND	36	POLO FAS	93	9272339	0 430071	WHITE	M	POLO FAS
44	5331167	4BJC 583808	BASIC SAND	38	POLO FAS	94	9352339	0 430071	WHITE	L	POLO FAS
45	5363579	6 047708	ANDOVER HE	L	POLO FAS	95	9432339	0 430071	WHITE	XL	POLO FAS
46	5441167	9BJC 586428	CLASSIC KH	34	POLO FAS	96	9502339	0 430071	WHITE	XXL	POLO FAS
47	5461167	9BJC 586428	CLASSIC KH	36	POLO FAS	97	9883749	2A 796799	CHIC CREAM	L	POLO FAS
48	5471167	9BJC 586428	CLASSIC KH	38	POLO FAS	98	9972087	2VPP 650100	RELAY BLUE	ONE	POLO FAS
49	5549233	6AEB 790003	W/W INDIGO	L	POLO FAS	99	9993749	2A 796799	CHIC CREAM	XL	POLO FAS

Appendix B: Association Rules for Polo Men's Department

Table of Rules for Polo Men's Department					
	antecedents	**consequents**	**support**	**confidence**	**lift**
1	({'SKU_1049441', 'SKU_439441'})	({'SKU_1589441'})	0.000474093	0.275482094	33.56855144
2	({'SKU_1049441', 'SKU_1589441'})	({'SKU_439441'})	0.000474093	0.284900285	31.29881885
3	({'SKU_439441', 'SKU_1589441'})	({'SKU_1049441'})	0.000474093	0.344827586	37.51115932
4	({'SKU_1869441', 'SKU_459441'})	({'SKU_1069441'})	0.000493057	0.318042813	31.74843947
5	({'SKU_459441', 'SKU_1069441'})	({'SKU_1869441'})	0.000493057	0.258064516	29.82646045
6	({'SKU_1869441', 'SKU_1069441'})	({'SKU_459441'})	0.000493057	0.255528256	26.12618488
7	({'SKU_2078810', 'SKU_1571243'})	({'SKU_5901243'})	0.0001043	0.328358209	35.40913531
8	({'SKU_4912330', 'SKU_5901243'})	({'SKU_1571243'})	0.000251269	0.325153374	34.97413364
9	({'SKU_4912330', 'SKU_1571243'})	({'SKU_5901243'})	0.000251269	0.335443038	36.17314139
10	({'SKU_9972087', 'SKU_1571243'})	({'SKU_5901243'})	0.000146969	0.295238095	31.83756451
11	({'SKU_4912330', 'SKU_2108810'})	({'SKU_9972087'})	0.000123264	0.254901961	29.62325933
12	({'SKU_5709431', 'SKU_3269431'})	({'SKU_6149446'})	0.000426684	0.253521127	18.2570699
13	({'SKU_3279431', 'SKU_5719431'})	({'SKU_6169446'})	0.000450389	0.305466238	24.08661237
14	({'SKU_4898351', 'SKU_5549233'})	({'SKU_5228351'})	0.000109041	0.315068493	21.8465096
15	({'SKU_5549233', 'SKU_5748351'})	({'SKU_5228351'})	0.000118523	0.257731959	17.87085612