Association Rules for Retail Stores

Executive Summary

Dillard, with approximately 450 stores, is an American retail chain across different states. For large retail stores, organizing items and rearranging them in the stores can be extremely challenging as minor changes can cause huge impacts on customers' shopping behaviors. For example, if customers usually buy toothpaste and toothbrush together, putting them close to each other will boost the other's sales record.

Even though Dillard do have point-of-sales data, it is the first time the company is performing the analysis, which means we assume most stock keeping units, the number that can be used to track inventory, are not close to each other as they should be. Hence, due to the budget constraint of 20 moves, we find the best 100 SKUs for candidates to move and choose 20 moves within them.

Association Rules is a technique used in data analysis that discovers if-then statements between items. In the case of Dillard, each transaction becomes a basket of goods, while our duty is to find whether the item in the basket can lead to the purchase of another one.

Taking the brief look of the results, 20 strongest association rules were created based on three criteria: confidence, lift and support. It is shown that people who bought the product from one brand is more likely to buy another product within the same brand.

Problem Statements

Richard Thaler, the Nobel Prize winner for Economics, proposed the idea that "subtle nudging", which means small changes by definition, can alter a person's behavior without even noticing it. In the case of grocery stores, nudges are widely used to influence people's shopping decisions so that people will then have the impulse of buying more stuff. Our duty is to provide useful information to Dillard so that their floor plans with goods can not only be improved to boost sales, but also make people's life easier to get the stuff they want. 20 best moves suggestions should be found

Assumptions

1. There is no data-entry error in the database, so that every information entered is correct.
2. The complete data is over 1 GB, which is too large for us to process it efficiently. Consequently, we choose to use stores only in the State of Illinois, assuming customers across different states have very similar shopping behaviors. Therefore, the result can be and should be applied to other stores in America.

3. Since it is the first time for Dillard to try to implement such strategy, we assume that SKUs now are randomly distributed among stores, and it's highly unlikely that best candidates are close to each other.
4. The records that include returned products to the stores are not helpful for our analysis as returning usually means dissatisfaction. By contrast, we are looking for meaningful data that can provide more information about consumer's satisfying behavior, so we exclude returned records.
5. Based on Apriori principle, we eliminated singleton that appears less than 20 times, assuming the fact that "if the itemset is frequent, then its subset is also frequent"
6. Based on the best judgement, we mapped the attribute names in the transaction data into the following and we use that for further analysis:]
   Var 1 = SKU
   Var 2 = Store
   Var 3 = Register
   Var 4 = Trannum
   Var 5 = Seq
   Var 6 = seq
   Var 7 = stype
   Var 8 = quantity
   Var 9 = original price
   Var 10 = sales price
   Var 11, 14 unknown
   Var 12,13 = interid/mic

Methodology

After having the data downloaded, we have the database diagram to see how these 5 files are connected. Since we realized that the transaction file is too large to process at one time, we then need to choose a clever way to sample the stores. Randomly choosing might not be the best alternative because here we are trying to find the co-occurence of items with some presence in the store, so we might get a result that is very biased in the end. Therefore, we performed exploratory data analysis in the STRINFO table, which is a table about information related to the store. After aggregation, instead of randomly choosing, we then choose stores based on states. We realized that Illinois has three stores. Assuming people's shopping behavior should be similar to each other across states, we then filter the dataset to only have three stores' transaction data in Illinois.

Another major thing we do to understand the data is to count the frequency of SKUs to get rid of redundant information. If the item appears to be purchased at least

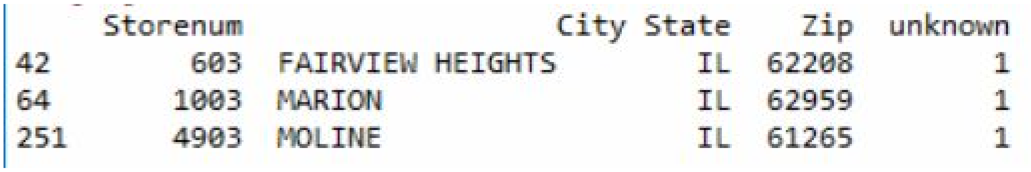
20 times, we regard this SKU number as valid number. Not only it get rids of the outliers, but also solves the memory crash problem for one-hot-encoding.

After data cleaning, we then transform the transaction dataset to market baskets. Every pairing of Store-Register-Trannum-Date is unique and should be matched towards unique SKUs. From this point, there are 29250 baskets while 3291 numbers of items(SKU) in the data set. One hot encoding is used, generated by TransactionEncoder method in the mlxtend library. Using Apriori function, frequent itemsets are created and the associations rules are generate with threshold: minimum_support = 0.0005, minimum_lift = 1. Results are sorted using confidence measure, to get final 20 moves.

For analyzing which specific rules are, we first merge two tables :SKU info and Department info so that each sku's department can be listed. We then merged the rules we generated with the department description to see which categories do those SKU belong.

Analysis

We first use the strinfo chart to get descriptive statistics for Dillard's stores information.



```
     Storenum                      City State    Zip  unknown
42        603  FAIRVIEW HEIGHTS          IL  62208        1
64       1003  MARION                    IL  62959        1
251      4903  MOLINE                    IL  61265        1
```

After aggregating the stores by states, we choose stores that have number 603, 1003, 4903 as these are the only three stores in the Illinois.

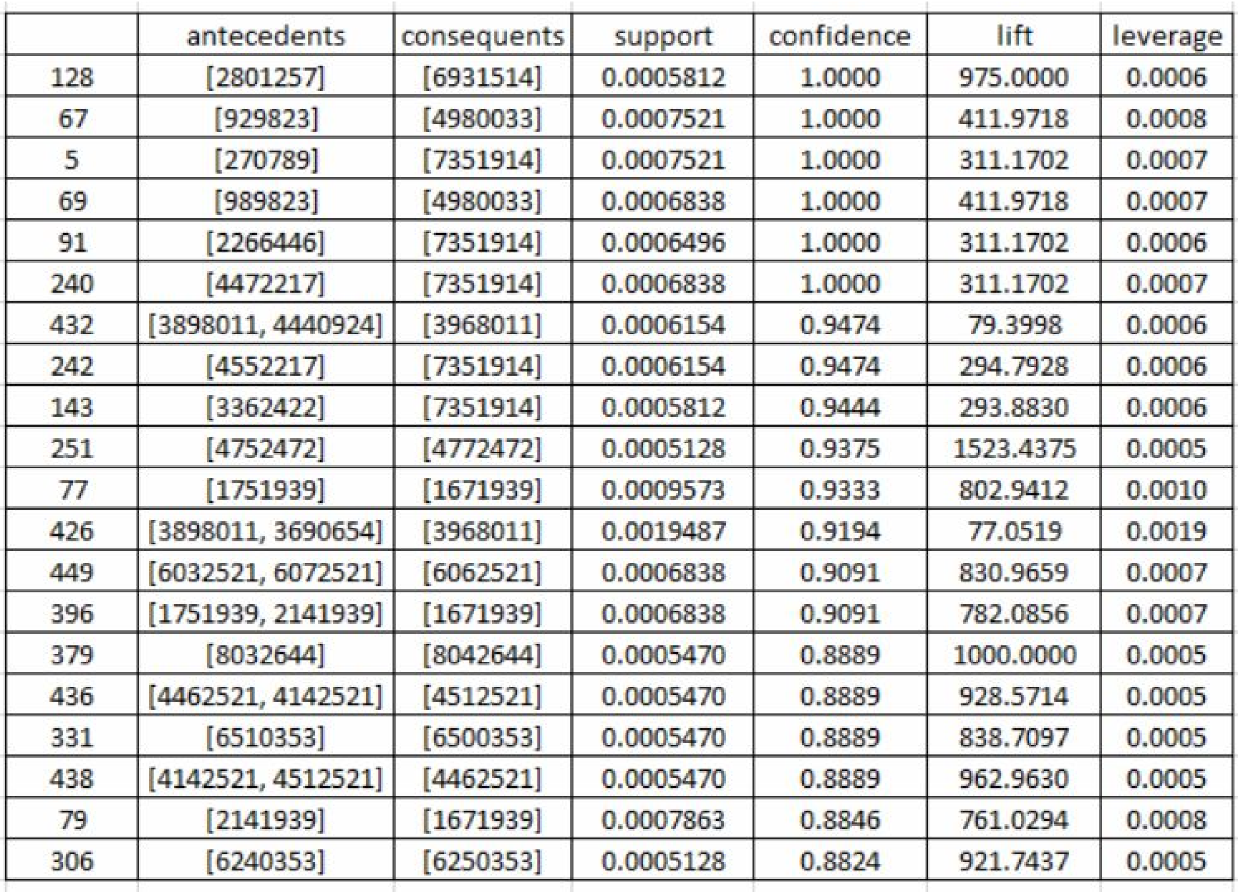
After filtering the original dataset, we did aggregation among SKU (item) to see the frequency of the data. The following table is the top 10 SKUs with largest numbers of occurrences in the transaction data. What worth mention is that this aggregation gives us the info that there are a lot of SKUs that were purchased by a small amount, so we need to get rid of these SKUs to make our computation more efficient. Therefore, we filter the data to let only items with more than 20 occurrences to remain in the dataset.

SKU	Count
4108011	1773
3524026	1064
3978011	934
2783996	829
5528349	799
4208011	655
6318344	618
803921	600
3161221	574

2698353 558

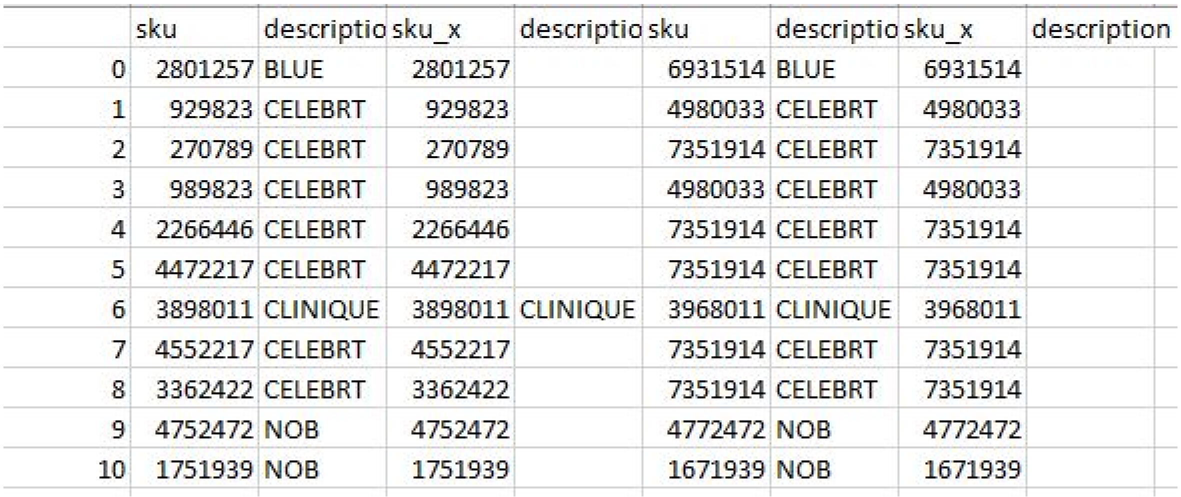
We then use the one-hot encoding to transform the data. Consequently, we get frequent itemsets with minimum support of 0.0005 as the threshold. Here, minimum support of 0.0005 means that we only consider itemsets which occur at least 50 times out of a total of 10,000 transactions. We get 1807 frequent itemsets.
After obtaining frequent itemsets, we then use lift as another criteria to give us association rules. Lift is the rise in probability of having item Y with the knowledge of item X being present over the probability of having Y without priori knowledge related to X. More specifically, if lift is bigger than 1, this means that the presence of item X truly leads to item Y. Therefore, we used this as the threshold and get 454 rules.

Another criteria that is really important for us to narrow down important 100 rules and 20 rules is the confidence, which is the likelihood of consequent given the antecedents. Confidence value of 0.5 means that 50% of the transactions that contains item X will lead to having item Y. Hence, we sorted by ranking the highest confidence level to the lowest and get the following table for best 20 rules.



	antecedents	consequents	support	confidence	lift	leverage
128	[2801257]	[6931514]	0.0005812	1.0000	975.0000	0.0006
67	[929823]	[4980033]	0.0007521	1.0000	411.9718	0.0008
5	[270789]	[7351914]	0.0007521	1.0000	311.1702	0.0007
69	[989823]	[4980033]	0.0006838	1.0000	411.9718	0.0007
91	[2266446]	[7351914]	0.0006496	1.0000	311.1702	0.0006
240	[4472217]	[7351914]	0.0006838	1.0000	311.1702	0.0007
432	[3898011, 4440924]	[3968011]	0.0006154	0.9474	79.3998	0.0006
242	[4552217]	[7351914]	0.0006154	0.9474	294.7928	0.0006
143	[3362422]	[7351914]	0.0005812	0.9444	293.8830	0.0006
251	[4752472]	[4772472]	0.0005128	0.9375	1523.4375	0.0005
77	[1751939]	[1671939]	0.0009573	0.9333	802.9412	0.0010
426	[3898011, 3690654]	[3968011]	0.0019487	0.9194	77.0519	0.0019
449	[6032521, 6072521]	[6062521]	0.0006838	0.9091	830.9659	0.0007
396	[1751939, 2141939]	[1671939]	0.0006838	0.9091	782.0856	0.0007
379	[8032644]	[8042644]	0.0005470	0.8889	1000.0000	0.0005
436	[4462521, 4142521]	[4512521]	0.0005470	0.8889	928.5714	0.0005
331	[6510353]	[6500353]	0.0005470	0.8889	838.7097	0.0005
438	[4142521, 4512521]	[4462521]	0.0005470	0.8889	962.9630	0.0005
79	[2141939]	[1671939]	0.0007863	0.8846	761.0294	0.0008
306	[6240353]	[6250353]	0.0005128	0.8824	921.7437	0.0005

Further analysis related to which department SKU belongs is made, noticing that most rules are generated within one brand.



	sku	descriptio	sku_x	descriptio	sku	descriptio	sku_x	description
0	2801257	BLUE	2801257		6931514	BLUE	6931514	
1	929823	CELEBRT	929823		4980033	CELEBRT	4980033	
2	270789	CELEBRT	270789		7351914	CELEBRT	7351914	
3	989823	CELEBRT	989823		4980033	CELEBRT	4980033	
4	2266446	CELEBRT	2266446		7351914	CELEBRT	7351914	
5	4472217	CELEBRT	4472217		7351914	CELEBRT	7351914	
6	3898011	CLINIQUE	3898011	CLINIQUE	3968011	CLINIQUE	3968011	
7	4552217	CELEBRT	4552217		7351914	CELEBRT	7351914	
8	3362422	CELEBRT	3362422		7351914	CELEBRT	7351914	
9	4752472	NOB	4752472		4772472	NOB	4772472	
10	1751939	NOB	1751939		1671939	NOB	1671939	

Conclusions

20 final rules were found, using association rules. Dillard can find the detailed rules in the ‘top20rules.csv’ as well as “complete100rule.csv”. Dillard should rearrange the shelves and put correlated products together. More specifically, we can see that oftenly, if the customer choose to buy one brand’s product, he/she is more likely to get another product within the same brand. Since our assumption that it is the first time Dillard doing this still holds, Dillard may want to group the product with more brand recognition.

Next Steps

One step for next is ideally to get market basket analysis for all states, not just the state of Illinois. Even though people should exhibit similar behaviors, there might be some items that have better sales record in one place than others. Therefore, in order to maximize the profit of each store, Dillard should use this strategy to gradually adapt floor plans for each stores.

Another thing to mention is that the analysis could be weighted in the future. Right now, the marginal profits are not calculated in the strategy. Two items can be closely related to each other, while they might generate little revenue. Hence, another future step is to consider economic costs.