Association Rules
Eric Xi
IEMS 308
Feb 11, 2019

Executive Summary

In the retail industry, rearranging a store's floor plan is an important factor that management teams should consider in improving revenue and customer experience. One feature of an optimal floor plan is that items that are frequently bought together are placed together. This lead to multiple benefits. For example, a customer who wouldn't have bought an associated item might choose to buy it because he/she sees it. Another benefit is that a shopper would spend less time looking for items that he wants to buy, leading to higher customer satisfaction. According to a blog written on updating floor plans, the author advises retailers to switch up floor plans once a year, and re-evaluate the original plan every six months to ensure efficiency (dotActiv). More frequent than that would cause shopper frustration, and less frequent than that would lead to encompassing data that doesn't fit more recent trends in the re-evaluations.

Dillard's is a major chain with several stores across the United States, and it is interested in rearranging the floors of the stores. For budgetary reason, they can only make at most 20 moves across the entire chain, and we are tasked to find the 100 SKUs that are the best candidates to modify the planograms. Therefore, we have decided to approach this problem by running market basket analyses across transactions that were made at the top **10 stores** based on volume in the **past 6 months** to identify items that are frequently bought together.

The Apriori Algorithm that is employed here is a data mining algorithm used for mining frequent item sets that leads to relevant association rules. It is devised to operate on a database containing a lot of transactions (Hackerearth). The resulting output of the algorithm would enable Dillard's to re-arrange their floor plan in an efficient way.

Taking a brief look at the result, we see that three main modifications should be made in each of these three brands: Chanel, Lancome, and Clinique. 20 SKUs would need to be rearranged to be closer to other SKUs that are associated with the same brand.

Problem Statement

Dillard's is a major retailer with multiple stores across the United States, and they have provided us with transaction data across its 400+ stores. Our job is to utilize the Apriori Algorithm to mine association rules that identifies items frequently bought together. Using the algorithm, we hope to provide Dillard's with a plan that would enable them to maximize the benefits from making 20 moves across the entire chain.

Assumptions

1. The data provided by Dillard is accurate and truly reflects all transactions.
2. The combination of 'STORE', 'REGISTER', 'TRANNUM', 'DATE' is sufficient in identifying a unique transaction.
3. Purchasing behavior top 10 largest stores by transaction volume is representative of behavior of all stores.

4. Analyzing data in the past 6 months is sufficient in understanding the most recent purchasing behavior of consumers.

Methodology

Our first step after reading the data is to create a new feature that would represent a unique transaction. After multiple trial-and-errors of combinations of primary keys in the database, we were able to identify the combination of 'STORE', 'REGISTER', 'TRANNUM', and 'DATE' provides us with the best estimation of a unique 'transaction_code'.

The second step involving filtering down the dataset so that only relevant information to the business problem is kept. We took the following steps to filter the data set:

1. Only include purchases where STYPE = 'P'. This would keep data that are purchase events.
2. Subset dataset to only include transactions that happened in the most recent 6 months using the 'DATE' feature. The most recent transaction is from 2005-08-27. Therefore we have filtered the data to only include transactions from 2005-2-27 onwards (~40% of data).
3. Subset dataset to only include top 10 stores by transaction volume using the 'STORE' feature
4. Assign min_sup based on price of SKUs and filter down dataset to only 100 SKUs that are best candidates to modify the planogram.
    a. Divide the dataset based on ~$60 intervals in ORGPRICE.
    b. Calculate the mean support and standard deviation of support in each of the intervals.
    c. Specify min_support for each interval as (mean + 15.55 * std). The reason for why a multiplier of 15.55 is chosen is because this figure would help trim the data down to the 100 SKUs requested by the client.
    d. Subset data to only SKUs that have support > min_support.
5. Drop columns: 'REGISTER', 'TRANNUM, 'SEQ', 'DATE', 'STYPE', 'STORE', 'ORGPRICE'

The third step involves transforming the dataset into the appropriate format that the Apriori Algorithm can accept. We used the groupby function to group 'transaction_code' and 'SKU', sum up the number of quantities in each 'transaction_code'-'SKU' pair, then unstack the data such that each row represents a unique transaction, and each column represent a unique SKU. We then reset the index to transaction_code and filled the NA with 0s. Since we are only concerned with the number of transactions and not the quantity of each item, we then converted the sum of quantity into 1s and 0s where 1 would indicate a transaction of an item occurred and 0 would indicate a transaction did not occur.

The Forth step is to load the filtered data into the Apriori algorithm using the mlxtend package, and utilize the association rules function to generate the table that will provide the respective support, confidence, and lift information of each rule.

Lastly, we recalculated support and lift using the actual transaction volume rather than the filtered down transaction volume based on min_sup, sorted the table by lift and removed rules that have duplicate lifts. The next section will further discuss the outputs of the analysis.

Analysis

1. Top 100 SKUs Analysis:

As mentioned in step 4 in the methodology section, we assigned min_sup based on the price of each SKU and filter down the dataset to keep only 100 SKUs that are best candidates to modify the planogram. Taking a brief look at how these 100 SKUs are distributed in terms of 'BRAND' we found that they are mainly concentrated in Clinique and Lancome. This suggests that these brands have the highest support within its respective price range.

Brand	Brand Counts
CLINIQUE	42
LANCOME	26
NOBLE EX	5
CHANEL 1	4
ROSE TRE	4
DESIGNER	3
POLO FAS	3
BEAUTE P	2
EUROITAL	2
LA PRAIR	2
UNILEVER	2
BVLGARI	1
FRANCISC	1
KENNETH	1
ROMANCE	1
ROYAL AL	1

2. Rules Analysis

Using the Apriori Algorithm, we were able to generate 1517 unique rules after removing rules that have duplicate lifts. However, upon manual inspection, we found that there were several overlaps in rules. For example, there were cases of rules illustrated below:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
'144717', '788874'	'2494717', '9546798'	3.6E-05	2.3E-05	5.2E-06	1.4E-01	6.3E+03
'144717', '9546798'	'788874', '2494717'	2.5E-05	3.4E-05	5.2E-06	2.1E-01	6.1E+03
'144717', '788874', '9546798'	'2494717'	6.9E-06	1.5E-04	5.2E-06	7.5E-01	5.0E+03

'144717', '2494717'	'788874', '9546798'	4.8E-05	2.5E-05	5.2E-06	1.1E-01	4.4E+03
'2494717'	'144717', '9546798’	1.5E-04	2.5E-05	1.3E-05	8.6E-02	3.4E+03
'144717', '788874', '2494717'	'9546798'	1.3E-05	1.3E-04	5.2E-06	4.0E-01	3.1E+03
'2494717', '788874', '9546798'	'144717'	8.7E-06	2.1E-04	5.2E-06	6.0E-01	2.9E+03

Therefore, we had to sort the rules by lift and manually go through rules from the top to identify “buckets” of SKUs that are often associated with each other. We were able to identify 3 groups of SKUs that will constitute the 20 items moves requested by the client.

Group 1: '144717', '788874', '9546798', '2494717'
Group 2: '9836218', '5772500', '8963391', '8568532', '656219'
Group 3: '3524026', '3559555', '3898011', '264715', '3690654', '3968011', '2716578', '803921', '5957568', '5618966'

These 3 groups represent items that are frequently associated with each other in the top rules produced by the Apriori algorithm. The lift for the rules generated within these groups are high, which means that the combination of purchases within a group is found to be more often than expected if bought independently. Therefore, the client should act on this information and move the items within each group close to each other.

3. Analysis on the 3 Groups

After generating the 3 groups that should be move together, we wanted to understand better if there are any patterns associated with the moves. To be more specific, we want to identify if items within each group share similar characteristics. Using the ‘skuinfo’ database, we were able to trace back to each SKU’s information, and found the following:

1. Group 1 items are all products from ‘CHANEL I’
2. Group 2 items are all products from ‘LANCOME’
3. Group 3 items are all products from ‘ CLINIQUE’

The finding from this analysis is not surprising since a shopper probably has an affinity to a particular brand, so he/she might purchase multiple products from the same brand. Therefore, it would make sense to place products from the same brand together.

Conclusions

Using a dynamic min support based on price intervals, we were able to identify 100 SKU that we want to further feed into the Apriori Algorithm. We found that Clinique and Lancome products made up a huge part of the 100 SKUs analyzed, which means that these two brands are frequently purchased within products in its price range.

Since the ultimate goal of this project is to identify 20 moves across the entire chain, we utilized the Apriori Algorithm and generated 1500+ rules. Through manual inspection after sorting the rules based on lift, we were able to identify 3 buckets of items that constitutes the 20 SKU moves. Further analysis reveals that these 3 buckets are associated with 3 brands, which are 'CHANEL I', 'LANCOME', and 'CLINIQUE' respectively.

Next Steps

The client should first implement the strategy recommended in the conclusion in a random subset of stores, which would move items within the brand 'CHANEL I', 'LANCOME', and 'CLINIQUE' close to each other. However, the detailed arrangement of items within the brands would require a further dive into the lift numbers in the association rules. SKUs that are associated together by rules with high lifts should be placed even closer together.

The next step for the client should be to conduct a test vs. control analysis, where a comparison is ran between the random set of stores mentioned above and a control set of store to understand if there is an actual incremental benefit with the moves. If the analysis demonstrates a positive incremental benefit, then the client can go ahead and roll out the strategy to more stores.

One assumption that we made in the beginning is that the top 10 largest stores would represent the purchasing behavior of all stores. However, this might not be true in reality. Therefore, further analysis needs to be conducted to identify if customization is needed for stores within a region. For example, customers in some states might prefer to purchase native brands.

Finally, as mentioned in the beginning of this report, it is important for the client to continuously re-evaluate and update the floor plans based on new association rules found in more recent transaction histories.

Works Cited

Gilbert, Darren. "When To Update Store Floor Plan." *What Retail Analytics and Loyalty Card Data Can Teach You About Shoppers*, www.dotactiv.com/blog/when-to-update-store-floor-plan.

"A Beginner's Tutorial on the Apriori Algorithm in Data Mining with R Implementation." *HackerEarth Blog*, 15 Sept. 2017, www.hackerearth.com/blog/machine-learning/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/.