IEMS 308 HW 2

Association Rules

Taige Hong

**Executive summary**

Dillard’s is a major retail chain with several stores across the state. It needs to find out what goods are usually purchased together, so they can move them together to boost the sales.

Association rules is an analytical strategy that detects for co-occurrence based on past records. In this case, I am using Dillard’s past POS data to generate a list of association rules. With this output, Dillard’s can make managerial decisions of moving items around.

The outcome is 100 pairs of SKUs that are recommended to be moved together. They are put in a separate csv file in the same folder.

**Problem Statement**

The retailer is interested in rearranging the floors of the stores to increase sales. For budgetary reasons they can make only at most 20 moves across the entire chain, and 100 SKUs as best candidates should be selected to modify the planograms.

**Assumptions**

All data entries are accurate. Each row indicates a unit of SKU being sold at a certain store on a certain date.

Consumers in the same state share similar shopping habits. This primarily means it is sensible to put together and analyze the data from the same state. It helps stores to develop the local strategy.

Our target is to increase sales, this means we do not need to take the ‘return’ rows in account.

Entries with same register number, transaction number and same date are considered from a single transaction. This helps to develop the baskets.

**Methodology**

The first step is to explore the data. There are 1,564,178 SKUs so it is impossible to directly analyze the association rules within these elements. I choose to use the subset of all Illinois store data entries. This decreased the SKU number to 143,111, while the number of distinct baskets drop to 371,353. In this case, I choose to set the individual support minimum to be 0.01%. This means for an SKU to be considered, it has to appear in more than 37 rows. This drastically decreased the number of baskets to 62,511.

This process does not take the price for each SKU into consideration. It might cause some SKUs with high price but low occurrence to be ignored. However, given the type of goods Dillard’s are selling, such situation is not very likely to happen.

So far, we have finished the preprocessing of the data. I rearrange the structure of the data and let the basket number be the primary index and SKUs are its elements. In this way, I can feed the input data in the apriori function from the apyori package. The minimum support is set to be 0.0002, minimum confidence is 0.02 and minimum lift is 2. Part of the reason that the support is relatively low is because about 70% of the baskets contain only one SKU. This does not help us develop the association rules, but it will deflate the individual support for SKUs and inflate the lift.

The 243 rules have rows that share the same antecedent SKUs. I sort the data frame based on lift and keep the one with highest lift from duplicates. The highest 100 antecedents are chosen. They along with their consequents are stored in a tuple and exported as a CSV file.

**Analysis**

To understand and analyze the output we have to go through the process that bring us here. A very important process is where SKUs that does not appear in more than 37 baskets are not considered. The data is from one year’s transaction history from three locations across Illinois, having ~37 occurrences basically mean selling one unit per store per month. This is a rather reasonable criteria to choose based on the significance.

I map the results to the SKU info csv and want to know more about the characteristics of these products.

NOBLE CH 15

SUMMER S 12

NOBLE EX 12

GREAT AM 10

MILCO IN 7

CLINIQUE 6

CABERNET 5

CROSCILL 5

WESTPOIN 4

HUE/KAYS 4

HUE 4

NUTMEG M 3

NOBILITY 2

CAPELLI 2

LANCOME 2

ROYCE HO 2

CHRISTY 2

MAIN ING 1

SATURDAY 1

SAN FRAN 1

Name: 9, dtype: int64

The result suggests most of the rules come from the brands:

NOBLE CH and NOBLE EX may refer to Noble Excellence, a company targeting in bedding. CLINIQUE is the makeup company, CABERNET may refer to wines. SUMMER S, GREAT AM and MILCO IN are hard to parse.

The strategy that Dillard’s should adopt is to put products from these brands close together according to the association rules.

**Next Steps**

One major next step with the analysis is to use other state’s POS data to observe the pattern of the rules. If they are similar, we can expand the scope and try to use a general rule for such moves across the states. If not, we can consider running a separate store-level analysis as the three stores in Illinois are still geographically far from each other so it might not be able to generalize the consumer behavior across different locations.

This report already succeeded in finding out the best 100 candidates for moving, but we need to figure out what the brand names actually are if we want to go deeper and analyze the consumer’s behavior.

We might want to look at the return entries as buying and returning will not generate any profits to the company. In this report, returns are completely ignored but in reality, there might be certain products that have a high purchase rate as well as a high return rate. We can take this in consideration when improving the model.