Yijing Barry Zhang
IEMS 308

Association Rule

Executive Summary:
Dillard is a large retail chain with stores across most states in the US. The retailer is considering a rearrangement of their store planogram and this analysis produced 100 candidate items (based on their SKUs) based on 2004-2005 point-of-sales data.

With an enormous dataset, we first conducted exploration on the features of the data. To control for regional and store differences, we chose transactions from one specific store in Oklahoma City for our analysis. We then proceeded with association rules mining using Apriori algorithm and found a list of association rules, which show us items that are frequently bought together.

We identified a total of 150 rules in the selected store and we identified 16 of them with high lift and high support to be prioritized for planning planogram changes. The rest of the association rules can be arranged based on either lift or support depending on the goal of the planogram change. We believe similar studies could be done at other stores and some common items across stores could be further analyzed to discover regional and temporal trends for further insights.

Problem Statement:
Dillard's is a major retail chain with several stores. The retailer is interested in rearranging the floors of the stores. For budgetary reasons, at most 20 moves across the entire chain can be made. This study intends to find 100 SKUs that are best candidates for planograms modification by analyzing Dillard's point-of-sales data.

Methodology:
Data Exploration:
The given data is very large, with more than 120 Million transactions across 453 stores. The data spans transactions from 08/01/2004 and 08/27/2005. It is also found that there are stores in most (but not all) states in the US, with a significantly more prominent presence in the South.

Feature Identification:
Most of the columns are easily identifiable. The last column of each table was ignored since it did not provide any significant information nor does it match any columns in the schema. The SKU information table is not entirely loaded as there are commas present within the csv files and the information did not immediately prove helpful for our analysis. For the transaction table, we assumed that the combination of store, saledate, and trannum constitute the composite primary key for the table and uniquely identify the transactions.

Data Selection:

Since each store might have fairly different products due to regional and demographic differences, we chose to do our analysis on only one store in Oklahoma City, which happened to be a store I frequented while I lived there. This selection method is chosen largely because of computational power constraints, but it also makes more sense to look for products within a store so that we have more concentrated association rules to support planogram rearrangement. Alternative methods are discussed in the Limitations and Next Steps section.

We also removed any return transactions to focus only on initial purchasing decisions. Adjusting for the returned items would be interesting but we have no concrete ways of distinguishing the reason for returning.

Association Rules Mining:

We used the association rules package from mlxtend. The operation is as follows. The transactions are grouped into baskets based on the composite primary key. The baskets are one-hot encoded and then frequent items were identified using apriori algorithm. Based on both intuition and computational constraints, we chose a minimum support of 0.2%, a minimum confidence of 5% and a minimum lift of 1. Detailed decisions of these choices are explained in the analysis section.

Analysis:

The subset of the data that we chose had 434365 transactions, constituting 35369 purchase baskets. These baskets had 127238 unique items, which means that our one-hot encoded matrix is particularly sparse. Because we are averaging around 3.4 purchases per item, we do not expect to have high support for most of the baskets. With some experimentation of the minimum support, we chose 0.15%, or roughly 70 baskets, as a cutoff point.

Confidence is calculated as the proportion of the frequently bought together items within the supporting (rule head) item(s). A soft threshold of 5% could be imposed as a cautionary note to the rules generated, but since we have a very sparse basket matrix, we expected the confidence to be fairly low and therefore will not rely on this metric too much as a hard limit.

Lift calculates how likely the supported item (rule body) is purchased given the supporting item (rule head) is purchased. This is the metric that we want to focus on more. Since the goal is to look for items that are often purchased together in the same visit (i.e. we are looking for complement products instead of substitutes), we are only looking for association rules where the lift is at least 1, indicating a positive correlation between the supporting items and supported items.

After obtaining around 150 rules using these criterias (found in rules.csv), we analyzed the rules by putting them into bins based on support and lift quartiles. We found 16 rules to be in the top

quartile for both support and lift. The items involved in these rules also all have fairly high confidence and should be prioritized. There are no rules with a very low confidence but certain rules with low confidence should be more carefully considered. The rest of the rules can be ranked in order of either lift or support depending on whether the goal is to optimize for the largest lift or to increase the number of co-occurring purchases. These groupings can be found in the ranked_rules.csv document

Conclusions:
Using the support and lift threshold, we found 150 association rules in total for store 9204 in Oklahoma City. These rules have a minimum support of 0.1% and they have lift ranging from 1.04 (roughly no effect) to 365.278 (significant increase in sales of the rule body). A full list of these rules can be found in the corresponding documents and they can be used according to the marketing strategy of Dillards. For example, if the goal were to increase the visit time, associated items could be kept further apart, and if the goal were to capitalize and encourage these shopping trends, these items should be kept closer together

Limitations & Next Steps:
This study was significantly limited by the amount of computational power available on a local machine, which prevented us from both storing (such as one-hot encoded baskets) and computing the data (such as large amounts of support calculation for association rules.) We believe that the methodology used is sound and the same analysis could be performed for not only individual stores but also to discover trends across the nation.

An additional feature that was not used was the department information. This was stored in the skuinfo table and we did not succeed in joining them to the transaction table. Since department store planogram changes likely happen more within their departments than across departments, looking at inter- and intra- departmental association rules separately might prove fruitful for different goals for Dillards; i.e. if we want to arrange the entire store, then we want to look at interdepartmental associations. If we want to rearrange products within a department, then intradepartmental associations are more useful.

Lastly, a good consideration for any retail operation should be to control for regional and seasonal patterns. The association rules found should be considered in the framework of their respective time-of-the-year as well as locality of the stores involved before they are used for recommendations on a planogram change.