STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Revision: 1.0

Table of Figures

COURSE PROJECT PART 4

ASSIGNMENT M12.E1.1

ASHLEY LAU
BIA-678 BIG DATA TECHNOLOGIES
Stevens Institute of Technology – Summer 2023

1 EXECUTIVE SUMMARY

Instacart is an American online service via website and mobile app that allows customers to stop from a network of groceries at the ease of their fingertips. A customer selects the grocery store they'd like to shop from as well as a list of items for purchase. After an order is processed, a personal shopper will retrieve the items and deliver at a pre-selected time frame. Customers can choose from a set of delivery options such as contactless delivery, pickup and club store deliveries. Additional fees are included like delivery fees ($3.99 for orders $35+ or $7.99 under that amount), pickup fees, service feeds (dependent on grocery location, alcohol purchases, etc.) There is an additional option of Instacart+ membership for other benefits. Instacart has become a popular service due to its broad network of access to grocery chains, like Kroger, Safeway, Costco, and Target, convenience, and was an immense help for families at home during the pandemic.

To suggest product suggestions for shoppers this paper will explain the technical approach towards building reference dataset and using concepts of association rule mining to come up with final product recommendations. Association rule mining is a process of discovering interesting relations between items in large databases. It is a rule-based machine learning method that identifies co-occurrence patterns among items in a dataset. Association rule mining is often used in market basket analysis to find relationships between products that are frequently purchased together. This information can be used by shoppers to select similar products for customers that fit their needs as well as boost Instacart's customer service. Three main metrics used in association rule mining are support, confidence, and lift. The support of an item set is the proportion of transactions that contain the item set. The confidence of an association rule is the proportion of transactions that contain the antecedent (first item set) as well as the consequent item (second item set). Thirdly, lift of an association rule is the ratio of the confidence of the rule to the support of the consequent. A high support value indicates that the item set is common, while a high confidence value indicates that the association rule is very likely. A high lift indicates that the association rule is interesting, because it is more likely to occur than would be expected by chance. A lift value greater than 1 is a good indication that an association rule will happen. In the data preprocessing stage, product items were grouped by order Id before calculating support, confidence, and lift metrics across the dataset.

The two tools used that helped us reach our product pairing suggestions were Tableau and Gephi. Tableau is a business intelligence (BI) and analytics platform that helps people see and understand data. Tableau is easy-to-use and has capabilities to create collaborative, scalable dashboards. Gephi is a graph visualization and exploration software tool that can be used to visualize and explore a wide variety of data, including relational data. It is a popular tool for social network analysis, but it can also be used for other tasks, such as clustering, community detection, and visualization.

After importing data into Gephi, the software identified 30 nodes of data points with 148 edges. Using statistical features to run modularity on the dataset, Gephi found 3 modularity classes; the visual output for this segmentation is shown in Figure 1. The first class revolves around bananas in green web. Bananas have possible relationships with twelve other food items; one-third of the items are organic and besides organic hummus all the food sets with bananas are either fruit or vegetable. The second class revolves around organic Haas avocado in pink web. Organic avocados also have possible relationships with twelve food items. Interestingly all the relationship pairings with the organic Haas avocados are also organic products. The only organic product that was not fruit or vegetable is organic whole milk. The third class revolves around limes in yellow web. Limes has association with three food items – large lemon, organic avocado, and organic cilantro. These relationships are further analyzed in Tableau to arrive at product recommendations.

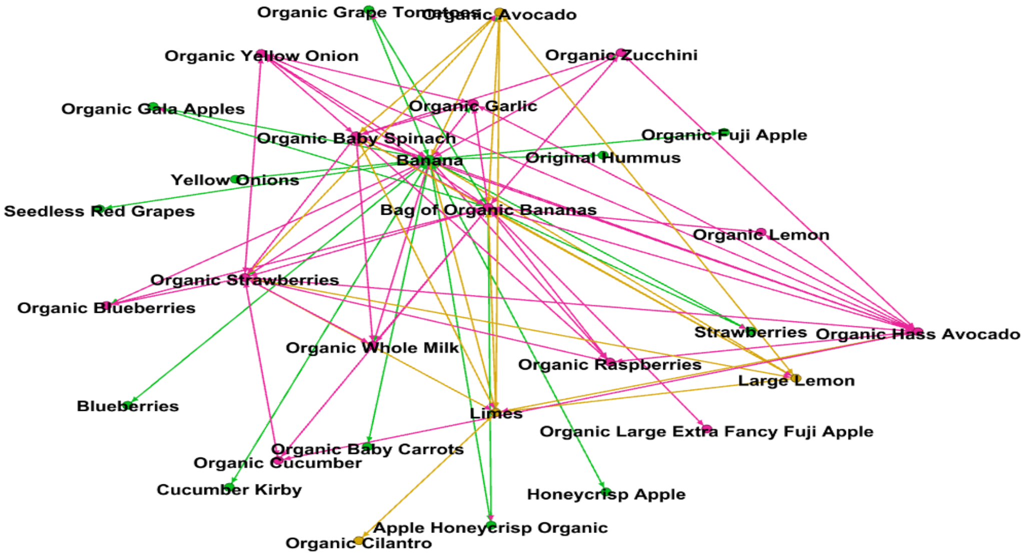

Figure 1 Gephi Relational Data for Association Rule Mining

After establishing these high-level relationships, data was put into Tableau to get visualization of these class relationships surrounding bananas, organic Haas avocado, and limes. The output for this relationship mapping comparing each set's life and support values is shown in Figure 2. As mentioned previously, for each food set pairing, sets with highest lift are strong association rule results that Instacart can recommend to their shoppers. For bananas, organic Fuji apples (lift = 2.576, support = 0.01056), honeycrisp apples, and cucumber Kirby are suggested. For organic Haas avocado, organic cucumber (lift = 3.268, support = 0.00543), organic raspberries, and organic yellow onions were top three suggested items. Likewise, bag of organic bananas has twelve possible food item relationships in the pink web; top three suggested items for bag of organic bananas are organic large extra fancy Fuji apple (lift = 2.633, support = 0.00727), organic raspberries, and organic Haas avocado. For limes, the ranking of top three suggested items resulted in organic cilantro (lift = 5.776, support = 0.005464), large lemon, and organic avocado as respective first, second, and third option.

Looking back at the relational data in Gephi, there were other strong relationships in green and pink modularity class webs. For example, organic strawberries have possible relationships with eight food items; the top three suggested items for organic strawberries are organic raspberries (lift = 3.001, support = 0.01053), organic blueberries, and organic cucumber. Organic baby spinach also has possible relationships with eight food items; the top three suggested items for organic baby spinach are organic garlic (lift = 2.542, support = 0.000653), organic avocado, and organic yellow onion (lift = 2.309, support = 0.00613). Organic yellow onion has five possible food item relationships in the pink web; the top three suggested items for organic yellow onion are organic garlic (lift = 5.699, support = 0.006866), organic Haas avocado, and organic baby spinach. The last main food relationship was large lemons found in yellow web. Large lemons can be suggested with limes (lift = 4.104, support = 0.008524), organic avocado, and organic baby spinach. These association suggestions make sense as many of these food pairings are commonly used seasonings for dishes or base ingredients for meals.

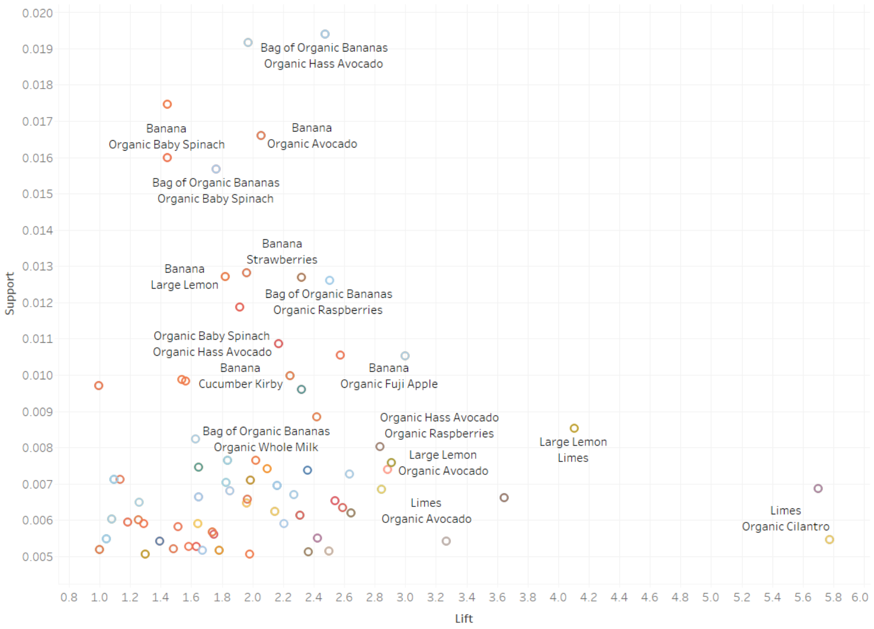

Figure 2 Tableau Lift vs Support Relationships

To conclude, association rule mining in market basket analysis is an effective method to statistically confirm product recommendations for shoppers to add to customer's Instacart order. If the data was not readily available, it would be difficult and expensive to collect; many assumptions about customer behavior would need to be decided before analysis can be completed. Market basket analysis (MBA) is also sensitive to data changes so results will not be completely accurate. MBA cannot determine cause and effect relationships between products. It can only identify associations between products. However, with association rule mining, we have been able to uncover hidden patterns in data, not simply apparent and improve product insights into customer behavior. It is a versatile, scalable tool for the online retailer to confidently suggest options for shoppers to fulfill customer's orders successfully.