**Instacart Project Part 4**

**Assignment M12.E1**

May 4, 2023

Saul Hernandez

BIA 678-WN Big Data Technologies

Stevens Institute of Technology – Spring 2023

Table of Contents

[1 Project – Instacart Project Part 4 2](#_Toc134132599)

[1.1 Product Recommendation Analysis 2](#_Toc134132600)

[1.1.1 Introduction 2](#_Toc134132601)

[1.1.2 Partitioning 2](#_Toc134132602)

[1.1.3 Analysis of the Partitions 5](#_Toc134132603)

[1.1.4 Product Recommendation with Tableau 5](#_Toc134132604)

[1.1.5 Conclusion 8](#_Toc134132605)

[1.1.6 References 8](#_Toc134132606)

# Project – Instacart Project Part 4

## Product Recommendation Analysis

### Introduction

With a stronger understanding of the tendencies of our customers, we can now move on to figuring out how to best recommend products to them to hopefully influence more purchases. For this we will utilize a dataset that includes the combination of different products and the Confidence, Support, and Lift associated with them. To explain what these values mean, the support is simply the probability that a transaction will contain a product. The confidence is the probability of transactions that contain product a to also contain product b. Finally the lift is a ratio of these two calculations that shows how the purchasing of one product impacts the probability that the other will be purchased. This lift characteristic will act as the weight for our analysis.

### Partitioning

To begin, uploaded this dataset to Gephi, which is a clustering and visualization tool. Initially, all the data points were interconnected with no real clear understanding of how these should be recommended. However, there were some clear relationships between some products. The initial output can be found below.

Diagram

Description automatically generated

**Figure 1: Gephi Directed Graph of Products.**

The lines have varying thickness, which correspond to the weight (lift) that we had for each. There were no labels on any of the nodes or edges so it was confusing to see what product strongly promotes the purchase of another. For this reason, node labels were added. This, however, still was only a jumble of points.

Diagram

Description automatically generated

**Figure 2: Labeled Directed Graph.**

In order to understand how the different items were interconnected, we partitioned the points (and colored them) based on the modularity classes. In other words, any items that were connected to each other would fall into a modularity class, but any that did not have any chain to another product would be found in a separate one. This gave a better glimpse of how everything was connected.

In order to analyze each partition separately, we added a filter. Below are the different outputs of these.

Diagram

Description automatically generatedDiagram

Description automatically generatedRadar chart

Description automatically generatedDiagram, radar chart

Description automatically generatedDiagram

Description automatically generated with medium confidence

**Figure 3: Product Partitions.**

It can be seen that not all the partitions are the same size or as interconnected. I gave the partitions in the order of products included. The top of the page has the highest included products while the bottom right, obviously, has the lowest.

### Analysis of the Partitions

When looking at how the products are connected, it can be difficult to tell which more strongly leads to which, especially in the topmost one. One thing to note is that while one product can lead to another, that same product could increase the likelihood of buying yet another totally unrelated product. One such case begins with a bag of organic bananas. When someone orders a bag of organic bananas, they are more likely to purchase organic avocados. The addition of this influences the likelihood of purchasing organic raspberries. This makes it more likely that someone will purchase strawberries, which finally makes it more likely someone will buy hummus. It is important to note that hummus is both connected to organic avocados and strawberries, but was not a branch node from bananas.

This form of analysis is extraordinarily powerful, as from one product we were able to suggest a series of other products that are more likely to be purchased when the previous was included, some even having twice this effect due to the way they were suggested. While it may be very beneficial to recommend any branch product from these chains, those that lead to even larger branches should be recommended first. This would increase the likelihood of more products being purchased. Using the above example again, if hummus was suggested instead of raspberries, then both raspberries and strawberries would lose the opportunity to be as relevant.

If Instacart could include a sort of analysis to find the longest chain that branches from a product that was just added to cart, following this could potentially have the highest yield of successful recommendations.

### Product Recommendation with Tableau

While Gephi proved to be very useful for the purpose of product recommendation and mapping how this should be done, a similar analysis was done within Tableau. One key difference is that instead of using only the weight (lift) variable, we kept the confidence and support variables to see how the lift is influenced by the combination of these. This is because lift can be large either due to a very small support of the second product and a moderate confidence or due to a very large confidence and a smaller confidence. To ensure that the lift is useful in recommending, it is helpful to map this.

Timeline

Description automatically generated

**Figure 4: Tableau Analysis.**

The above graph uses the size of the circles as the lift, so the larger the circle the larger the lift. This is not important, however, if the support (Y-axis) is large. This would only mean that the two products were purchased together out of chance, because they are both likely to be purchased anyways. The lower and more right a product set is, and the larger the circle is, the more likely this is a truly useful relationship. One such case is showcased in suggesting grapefruit sparkling water when some purchases lime sparkling water. This has a low support, high confidence, and one of the largest lifts.

One way to combat this is to limit the support that is included in the graph. In the below graph, the support allowed was halved, only allowing for values with a support of 0.01 or less.

Timeline

Description automatically generated

**Figure 5: Tableau Analysis (Limited Support).**

Some product combinations which can now take the limelight are those such as limes from jalapeños, yellow onions from garlic, and avocados from cucumbers. As you can see, by understanding how these products are interconnected and how the likelihood of one being purchased in general impacts the lift of the two, one can make a more informed decision on how these products should be recommended.

### Conclusion

This analysis of products was fruitful, in that it allowed us to not only see how certain products can influence the purchase of others, but also how the likelihood a product being purchased in the first place may skew this calculation. For Instacart to improve on their recommendation system to drive up products in each order and make the experience easier for the target audience, they should seriously consider using these technologies and developing models to both find the longest node chains and those that lead to products that have low initial support for their purchase. This would lead to an increase in not only general sales, but sales of products that are typically not purchased as often.

### References

References

<https://www.kaggle.com/c/instacart-market-basket-analysis>