Market Basket Analysis with Instacart Data

[200208] Business Cases – Case 3 (19 Apr 2021)

MASTER OF DATA SCIENCE AND ADVANCED ANALYTICS

Group AA

Ahmadov, Emil (m20201004)

Macean, Doris (m20200609)

Shin, Doyun (m20200565)

Tagiltseva, Anastasiia (m20200041)

NOVA Information Management School

Universidade Nova de Lisboa

INDEX

1. INTRODUCTION 1
2. METHODOLOGY 1
3. BUSINESS UNDERSTANDING 1
   3.1. Business Objectives 1
   3.2. Situation assessment 1
   3.3. Data Mining goals 2
4. DATA UNDERSTANDING 2
5. DATA PREPARATION 3
   5.1. Checking and handling of missing and correlated values 3
6. CLUSTERING 3
   6.1. Customer Clustering 3
   6.2. Product Clustering 4
7. Association Rules Analysis 4
   7.1. Baseline association Rules Analysis 5
   7.2. Associations on First Orders and First 15 Items 5
   7.3. Days of Week Analysis 5
   7.4. Times of Day per Day of Week 6
8. DEPLOYMENT AND MAINTENANCE PLANS 7
   8.1. Pre-Deployment 7
   8.2. Full Deployment 7
   8.3. Post-Deployment 8
9. CONCLUSIONS 8
   9.1. Considerations for improvement 8
10. REFERENCES 8

1. INTRODUCTION

Our team was tasked with investigating and defining the relationships between the various products offered by Instacart. Understanding complements and substitutes within the product offering can provide Instacart with a holistic view of the behavior of its customers, as well as the intelligence to target the dynamically changing customer interests.

Our research could demonstrate how machine learning can be applied to retail businesses where no user interaction detail is available explicitly. Specifically, market basket analysis is a data mining technique that is used to inform a retailer about products that are frequently bought together, and items that are substitutes to one another. The data, and the algorithms used, allow grocery suppliers to revolutionize the process as to how customers discover and purchase groceries.

2. METHODOLOGY

Cross-Industry Standard Process for Data Mining (CRISP-DM) [1] was the applied methodology in the execution of this research. CRISP-DM [2] provides a six-step process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The following sections offer a succinct description of the execution of each of these steps.

Our team began with an exploratory data analysis of Instacart's dataset. Once our data exploration was complete, we proceeded with product and customer clustering. We then developed association rules across various slices of the dataset.

3. BUSINESS UNDERSTANDING

3.1. BUSINESS OBJECTIVES

The company wants to obtain insights about their sales and customers according to the following points:

- Understand the customer base and identify the different types of consumer purchasing behavior.
- Identify which types of products are frequently sold and could be offered in a wider range for better inventory control.
- Identify substitute and complementary products and develop different marketing strategies to monetize these hidden patterns.

3.2. SITUATION ASSESSMENT

For this project, we utilized a mid-range PC of our own and a 2 018 887 sample of order history data from Instacart (200 000 orders for 134 products). The number of unique customers in the whole dataset is 105273. Addex Consulting consists of four data scientists using python as the primary coding language. Various python modules were used to analyze the data and develop various association rules. The specific tools are addressed in our enclosed notebook.

We developed a model that could accurately predict product correlations within the dataset provided. The utility of our algorithm could be improved by a complete dataset. We were unable to explore various characteristics of the company's customers which would be useful for a valuable analysis. Namely, total number of orders, average basket size and average time between orders could help us to develop a more useful application for Instacart. Furthermore, this data could help us with many decisions we made during our process.

3.3. DATA MINING GOALS

The business objectives can be translated to the following data mining goals:

- Association: The main goal of association is to establish the relationship between items which exist in the market. The typical examples of association modeling are Market basket Analysis and cross selling programs [3]. The tool used for association rule mining is the Apriori Algorithm.
- Clustering: In this, Data Mining organizes data into meaningful subgroups (clusters), such that points within the group are similar to each other, and as different as possible from the points in the other groups.

4. DATA UNDERSTANDING

Our team began with an exploratory data analysis of Instacart's dataset in order to gain a general overview of the behavior of Instacart's customers, features of the 134 product offerings, as well as any frequent patterns or correlations among items.

Some insights were expected. For example, we found that most orders take place between 10am and 3pm, and the frequency of many orders was weekly or monthly. The average basket size was around 6 items with fresh fruit and vegetables as the most ordered products.

Some interesting findings we had was that the rate of product reorder was significantly higher for the first items added to the basket. As such, we reviewed what items are usually added to the basket first. Interestingly, the most frequently bought items are not usually the first items added to the cart, with fruits and vegetables typically occupying the 7th or 8th position in the cart. We noted that the items most frequently added early were specialty wines and spirits. We also reviewed reorder rate across department, as well as products. We noticed that some items were "one-shot" in that they were rarely reordered. These items were typically personal care items, pantry products and international goods.

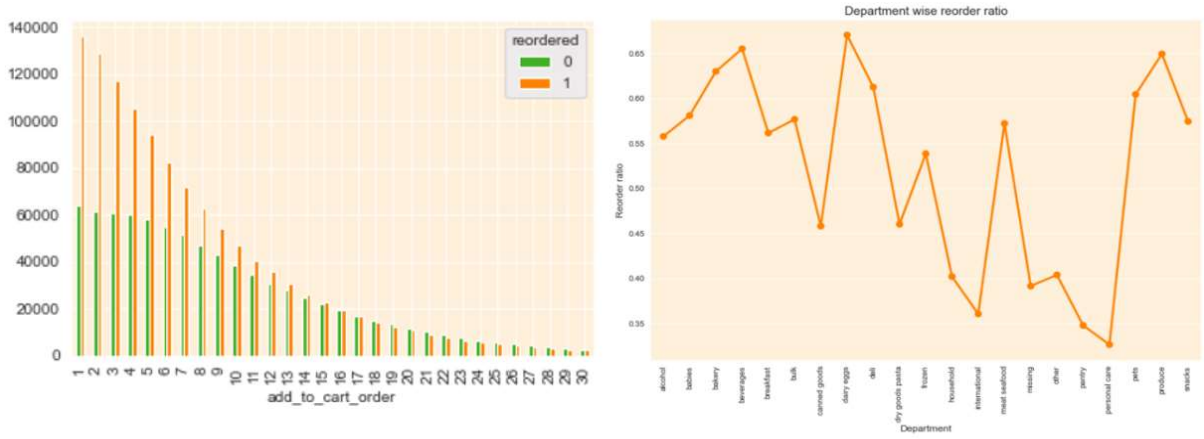

Figure 1 Reorder rate by cart position and reorder ratio by department

Finally, we did an analysis of products purchased in the morning and in the afternoon. As can be seen below, there is a reversal in the dominant products around noon. It appears that snacks, breakfast goods and household items are more common in the morning, while alcohol, desserts and personal hygiene products were more frequent in the afternoon.

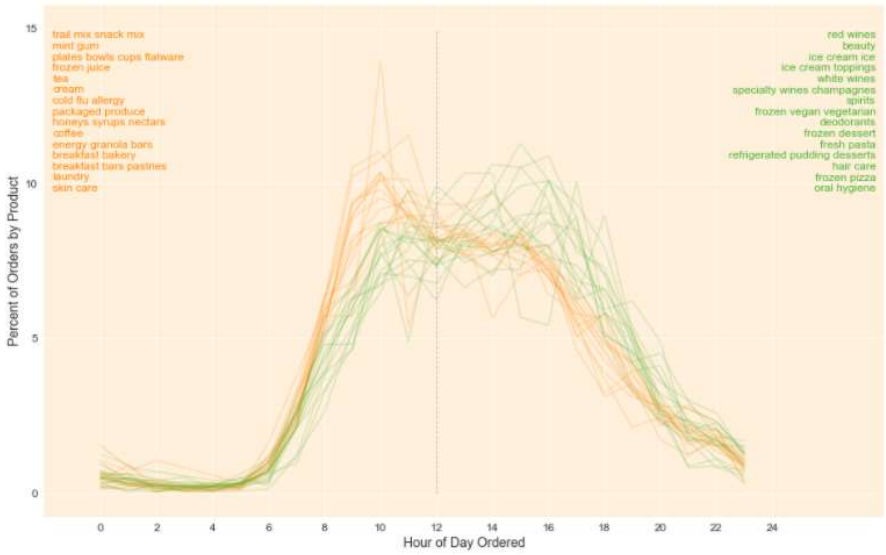

Figure 2 Percent of Orders by Product by the hour of day (top-15 for morning and afternoon)

5. DATA PREPARATION

5.1. CHECKING AND HANDLING OF MISSING AND CORRELATED VALUES

The only missing values were found in the days since last order. These values are associated with the first order of those users. As such, we did not treat the missing values, but rather conducted an analysis of these first orders separately.

6. CLUSTERING

6.1. CUSTOMER CLUSTERING

Subsequently, we developed a customer segmentation that enabled us to classify customers according to their purchases. Customer segmentation allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing, and product development strategies.

We used the k-means clustering algorithm to derive the optimum number of clusters and understand the underlying customer segments based on the data provided. Since there are a lot of products in the dataset, we used Principal Component Analysis to find new dimensions along which clustering will be easier. We looked at the top 10 goods bought by individuals in each cluster, relying first on absolute data and then on the percentage among the top 8 products for each cluster.

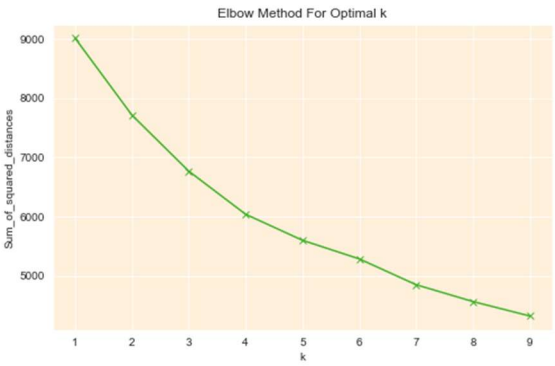

Figure 4 Elbow Method for Optimal K

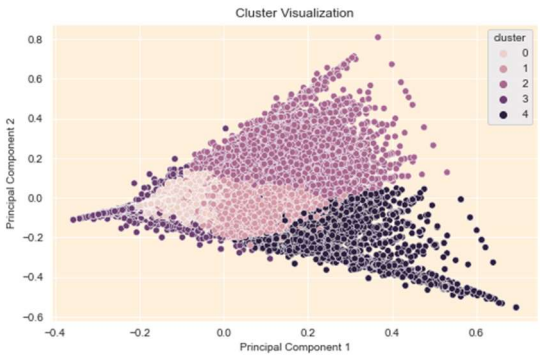

Figure 3 Cluster Visualization

It seems individuals in Cluster 2 buy more fresh vegetables than the other clusters. Using the absolute data, we find that Cluster 2 includes those customers who buy more goods. Cluster 0 is the largest cluster and the distribution of customer preferences in this cluster is relatively evenly distributed among the top 8 products. People of cluster 3 buy a lot of 'Sparkling water' and they are interested in products which are not listed in the top 8 products.

	fresh fruits	fresh vegetables	packaged vegetables fruits	yogurt	packaged cheese	milk	water seltzer sparkling water	chips pretzels
0	13.599469	14.173306	11.242351	12.745064	12.984644	10.819767	11.274102	13.161298
1	31.730155	18.433942	14.791736	13.450447	6.765386	6.365661	4.446647	4.016027
2	20.501104	50.175049	11.388252	4.836691	4.438789	3.586744	2.886549	2.186822
3	5.605433	1.841621	2.520718	3.487569	1.761050	2.624309	79.638582	2.520718
4	66.042774	8.741623	8.976584	3.879529	2.451072	4.648493	3.326836	1.933089

Table 1 The percentage of products which are generically bought by most of the customers with respect to the other top 8 in each cluster

6.2. PRODUCT CLUSTERING

For the product clusterization we explored reorder rates for clusters based on different parameters. First, we calculated count, reorder rate and average values for variables 'add_to_cart', order_dow', 'order_hour_of_day', 'days_since_prior_order for each product for all orders.

Before clusterization, we normalized the data by setting the mean to zero and variance to one (StandardScaler by scikit-learn[4]). We then used hierarchical clustering with Ward's linkage to develop our clusters.

clusters	1	2	3
count	954 807	189 408	874 672
reorder_rate	0.66	0.36	0.50
add_to_cart_mean	6.53	9.10	8.98
order_dow	2.75	2.63	2.80
order_hour_of_day	13.52	13.39	13.53
days_since_prior_order	11.21	11.63	10.73
size	17.00	34.00	82.00

Table 2 Product Clusters

As a result, we had three different clusters, each with notably different reorder rates. There are 34 products that only get reordered 36% of the time. The highest reorder rate was found for the 17 products which get reordered 66% of the time.

As such, we strongly recommend Instacart pay attention to Cluster 1, because the shortage of products in this cluster can affect customer loyalty, as well as lead to lost profits

product_name	count	reorder_rate	add_to_cart_mean
soap	55150	0.782629	5.565204
red wines	52564	0.731813	5.983259
body lotions soap	226039	0.718261	7.160176
mint gum	27986	0.707425	6.426356
meat counter	39389	0.691843	6.743736

Table 3 Top-5 Products from Cluster 1 according to reorder rate

7. Association Rules Analysis

We conducted association rules analyses on different sets (split) of the data. The idea of association rules is to use support, confidence, lift, and conviction values to identify popular sets of items, substitutes, and complements. Nonetheless, these values are not definitive indicators and qualitative analysis must be conducted.

First, it is common for customers to shop for groceries for the whole household for several days. For example, say person A and B are in the same household. Person A plans to cook with pepper paste, and person B plans to snack on chocolate. Alternatively, person A plans to cook with pepper paste today and snack on chocolate tomorrow. In either case, these products would end up in the same basket, which would distort the association metrics indicating complementary relationship between chocolate and pepper paste.

Second, partially due to the first issue above, lift value is not a reliable concept for making final conclusions on the nature of products. How the relationship between two products can go from below 1 to above 1, depending on the number of unrelated products within the dataset, is demonstrated by a simple example by S. N Ramesh, 2019 [6].

Third, a substitute is defined by the intended use of the product. However, the *purpose* of an item is extremely difficult to quantify perfectly, if at all possible, causing a significant loss of information.

For these reasons, it is critical to complement the quantitative results with qualitative analysis. We also take the metrics, especially lift, rather leniently.

7.1. Baseline association Rules Analysis

The first step for association rules analysis was to conduct a simple apriori study on the whole dataset to capture the most dominant trends. At a minimum support level of 0.05 (an arbitrary starting point), all three focuses of the study (popularity defined by support, substitutes defined by low lift, complements defined by high lift and high confidence) are dominated by a combinations of vegetables, fruits, and water. Without qualitative analysis, the model indicated [water and vegetables], [water and fruits], [fruits and ice cream ice], and [chips pretzels and vegetables] to be substitutes of each other. Complementarity between [herbs and vegetables], and [fruits and vegetables] were also indicated. This dominance of fruits, vegetables, and water was expected, as they are bought almost habitually by most customers, regardless of the main items that the customer intended to buy. This behavior is rather well-known by supermarkets which influenced them to place vegetables and fruits at the entrance of the shop.

7.2. Associations on First Orders and First 15 Items

We then focused on customer behaviors in their "first": 1) First ever order from Instacart, and 2) items that were constantly added within the first 15 items in the cart. The former could hint at "Why is the customer trying online groceries shopping at all?" and the latter could hint at "What are the primary reasons the customer is making this order?" However, as shown in the notebook, the results did not differ significantly from the baseline and the dominance of fruits, water, and vegetables continued. This reinforces our hypothesis that the three products are strong attractors for customers to online groceries. Therefore, more quantity and diversity of these products would be advised.

7.3. Days of Week Analysis

As our third associations analysis, we sought to observe the changing customer behaviors over the days of the week. To do so, we separated the dataset by days of week, with 0 being Sunday. In doing so, we observed Wednesday to be the least busy traffic day and Monday to be the busiest.

For substitutes, we lowered the minimum support level by 0.01 (to 0.04), then extracted the lowest 10 lift values item sets. As discussed previously, lift value is rather unreliable and it is not appropriate to say Set 1's substitutional tendency is stronger than Set 2 just because the lift value of Set 1 is lower than Set 2's, especially when the difference is minimal. We believed ranking these 10 sets per day based on the popularity (support as the proxy metric) would be more meaningful insight for Instacart. Afterwards, we conducted a qualitative analysis and went through each sets of substitutes.

	antecedents	consequents	lift	support	index
1	(fresh vegetables)	(refrigerated)	1.034967	0.077199	0
2	(ice cream ice)	(fresh fruits)	1.022712	0.074406	0
3	(fresh vegetables)	(ice cream ice)	1.003870	0.063801	0
4	(cereal)	(fresh vegetables)	1.003836	0.053652	0
5	(milk)	(water seltzer sparkling water)	0.994827	0.049205	0
6	(fresh vegetables)	(juice nectars)	1.026134	0.049205	0
8	(frozen meals)	(fresh vegetables)	0.984349	0.045613	0
9	(milk)	(soy lactosefree)	0.872122	0.041279	0
14	(milk)	(water seltzer sparkling water)	1.039780	0.050425	1
16	(energy granola bars)	(fresh vegetables)	0.973196	0.042448	1
18	(candy chocolate)	(fresh fruits)	1.016374	0.042215	1
24	(fresh fruits)	(ice cream ice)	1.067395	0.055541	2
25	(water seltzer sparkling water)	(milk)	0.987431	0.044946	2
28	(candy chocolate)	(fresh fruits)	1.035930	0.040804	2
33	(fresh fruits)	(ice cream ice)	1.062251	0.057499	3
35	(yogurt)	(water seltzer sparkling water)	1.097309	0.050721	3
37	(ice cream ice)	(fresh vegetables)	1.038007	0.043163	3
38	(milk)	(water seltzer sparkling water)	0.959463	0.041527	3
43	(fresh fruits)	(ice cream ice)	1.052562	0.057300	4
46	(fresh fruits)	(baking ingredients)	1.098130	0.044807	4
47	(milk)	(water seltzer sparkling water)	0.982703	0.044406	4
53	(fresh fruits)	(ice cream ice)	1.051474	0.067365	5
56	(water seltzer sparkling water)	(milk)	0.951843	0.045901	5
57	(soft drinks)	(fresh fruits)	0.821515	0.044320	5
58	(fresh fruits)	(candy chocolate)	1.042333	0.042249	5
61	(ice cream ice)	(fresh fruits)	1.034761	0.080281	6
62	(ice cream ice)	(fresh vegetables)	0.990400	0.066600	6
63	(packaged vegetables fruits)	(ice cream ice)	1.038094	0.054874	6
65	(cereal)	(fresh vegetables)	1.023846	0.048896	6
66	(fresh vegetables)	(juice nectars)	1.028017	0.048513	6

Table 4 DOW Substitutes, index column representing days of week, starting from 0 = Sunday.

During the process we noticed different types of substitutes. One, the usual sense of substitution i.e., substitution by purpose. Two, substitution by underlying decision. For example, on Mondays, the lift value between fresh fruits and soft drinks were very low (0.82) compared to lift values of other sets (rarely below 0.95). Fruits and soft drinks can hardly be considered as substitutes. Instead, when a customer decides to have healthy snack, the customer is likely to complement it with healthier beverages instead of soft drinks. If unhealthy snacks like chips were decided, it would be more common to pair with soft drinks. Hence, we excluded these items as well as obviously non-substitutional products (e.g., cheese and water) from the final substitution sets. The raw result of pre- qualitative analysis is provided in the notebook (dataframe 'dfs_subs_final'). Table 4 summarizes the final substitutes list.

Finally, we conducted an experimental study on the complements. From the baseline study on the whole dataset, we established complementary relationships among fruits, herbs, and vegetables. Therefore, we sought to exclude the most dominant consequent one by one and adjust the parameter slightly, to capture less obvious complementary sets. We first excluded [fresh vegetables], which resulted in dominance of fresh fruits as the consequent, against antecedents of different combination of frozen sets, which led us to suspect the reliability of the result. Again, , these sets appear to show high lift and high confidence value, simply due to the fact that fruits are often purchased regardless of the primary items the customer intended to buy. Nevertheless, we witnessed the items were usually related to breakfast goods, including yogurt, milk, cracker and cheese. We could also conclude that the decreased presence of fresh herbs indicates a very strong complementary relationship between fresh herbs and fresh vegetables. We continued with exclusion of fresh fruits and relaxed the parameter, lowering minimum confidence and lift thresholds. The breakfast set trend continued to be visible, with [frozen produce, packaged cheese], [yogurt, milk, packaged cheese], and [fresh dips tapenades, yogurt] found to show complementary pattern with [packaged vegetables fruits].

Finally, we excluded the [packaged vegetables fruits] and relaxed the confidence level down to 0.5. Strictly speaking, this value is rather low to define certain item sets to be regarded as complementary. Nonetheless, as a part of exploration and topic for discussion we believed it was a meaningful approach. The result qualitatively made sense and lift values of these items were found to be very high even when we set the minimum threshold to 1.1. Most of these findings were especially pronounced on Sundays (index=0). Considering the low support levels throughout, Instacart can use this information to make more paired offerings to customers depending on the profitability of these products.

antecedents	consequents	lift	confidence	support	index
(milk, packaged cheese)	(yogurt)	1.702227	0.514581	0.049103	0
(bread, milk)	(yogurt)	1.673986	0.506044	0.036407	0
(yogurt, lunch meat)	(packaged cheese)	1.916885	0.518519	0.026378	0
(frozen produce, packaged cheese)	(yogurt)	1.669658	0.504735	0.026262	0
(yogurt, crackers)	(packaged cheese)	1.879861	0.508503	0.026001	0
(frozen produce, milk)	(yogurt)	1.739756	0.525926	0.024696	0
(milk, lunch meat)	(yogurt)	1.712066	0.517555	0.023073	0
(milk, lunch meat)	(packaged cheese)	1.898902	0.513654	0.022899	0
(milk, crackers)	(yogurt)	1.733295	0.523973	0.022175	0
(bread, lunch meat)	(packaged cheese)	1.956100	0.529126	0.022117	0
(yogurt, lunch meat)	(packaged cheese)	2.179116	0.504640	0.021047	1
(baby food formula)	(yogurt)	1.824236	0.517688	0.023373	6
(frozen produce, milk)	(yogurt)	1.877361	0.532764	0.021853	6

Table 5 Complement Detection by Exclusion, after 3rd element, and minimum confidence = 0.05

7.4. TIMES OF DAY PER DAY OF WEEK

As our last analysis, we sought to utilize the same technique, but further restricting time period. We categorized the hours the order was made into morning, afternoon, evening, and night. Then we split the original data into days of week, then into times of day. During the process we noticed a cyclical pattern, which can be used together with Instacart website traffic record for more meaningful insight (e.g., traffic pattern vs order rate pattern to handle the server more efficiently). On every day, there was a sharp uptick from night to morning, then a sharp turn after reaching afternoon was visible.

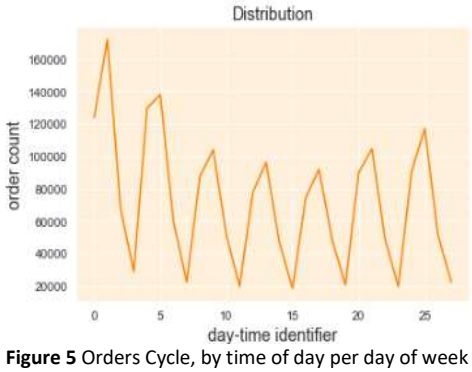

Figure 5 Orders Cycle, by time of day per day of week

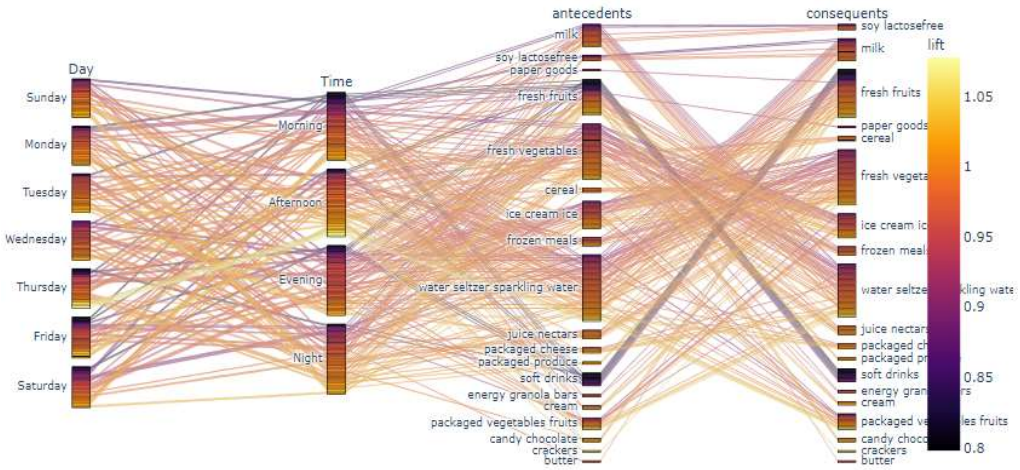

Figure 6 Substitutional relationship chart, by time of the day per day of week

After producing association-rule-based item sets, we stopped our study, due to the rising complexity of relationships among different items. To conduct another qualitative analysis on this data would require more domain knowledge, available to Instacart. As such, we prepared the data for further exploration. Ideally, a discussion between Instacart's subject matter expert and one of our data analysists to go through the potential substitute sets together could draw some more meaningful conclusions. Interactive version of Figure 6 is also included in our codes attached.

8. DEPLOYMENT AND MAINTENANCE PLANS

8.1. PRE-DEPLOYMENT

The identified association rules can be used for better organizing the layout of the mobile app and website. Complementary products that were identified can be shown on the same page for a more comfortable app experience. Also, for substitute products, a detailed comparison tool can be provided. Another approach that these association rules can be used for is to increase the sales of complementary products.

Additionally, the recommendation system that was developed by our company will be also deployed inside the app. This system will use previous purchasing history of the user to predict their next purchases. Interactive user interface must be built for this purpose.

In order to have a real-life experience and feedback, a pilot program must be deployed. A specific target group must be identified for testing purposes. This program should be tested with a group of 2-3 thousand people. Based on the feedback from users, the company will need to adjust their interface accordingly.

For implementing these changes, we will need to work with the user interface engineers from the company. We will work together with them to connect the back-end of the association rules and recommendation system to the front-end.

8.2. FULL DEPLOYMENT

After the pilot program is conducted, UI engineers will need to gather feedback from the users and make necessary changes to the interface. When the final version of the interface is decided, full deployment of the project can be done.

As our world is constantly changing, it is highly probable that customer behaviors will be changing too. Also, as new products emerge in the market, new complementarities or association rules may arise between products. During full deployment, this recommendation system and association rules will need to be updated to match the changing consumer behavior and market dynamics. This system must be continuously evaluated to allow for better performance.

For the purpose of performance analysis and continuous maintenance of the system, we will be providing training to the workers on the company side. Other than that, our company will also complete maintenance every year.

8.3. POST-DEPLOYMENT

In the post deployment, the company can adjust this recommendation system with all of the data and use association rules for better recommendation. Right now, our recommendation system is only using previous purchasing behavior of a specific customer; however, in the future the company can use association rules and customer clusters for better recommendations. Also, based on the association rules that also consider day and time, even more advanced recommendation systems can be built.

9. CONCLUSIONS

Our team has conducted Exploratory Data Analysis to better understand the data and gain interesting insights from the data. These insights include finding the most popular products of the company, reorder ratio of the products and popularity of the products across the time. We also conducted customer and product clustering to identify different customer behaviour patterns as requested. Our team also conducted Association Rule Mining on the given data and provided insights about the complement and substitute products across the day and time of order. Additionally, a product recommendation system based on the previous purchases of the customer is also provided.

In conclusion, the business needs of the company were fulfilled by our team by providing most popular products, customer clusters and Association Rules. A deployment plan has also been provided to the company for the application of this project.

9.1. CONSIDERATIONS FOR IMPROVEMENT

The data provided by company was incomplete as the data was randomly sampled from the total dataset. In the future, a more advanced analysis can be provided on all the data for randomly selected customers. Additionally, the given data did not include demographic information about the customers, which may be useful for model improvements in the future.

10. REFERENCES

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Retrieved September 10, 2015, from https://themodeling-agency.com/crisp-dm.pdf
[2] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.
[3] Manpreet Kaura, Shivani Kang. Market Basket Analysis: Identify the changing trends of market data using association rule mining. Bhai Gurdas Institute of Engineering and Technology, Sangrur 148001, India. 2016
https://core.ac.uk/download/pdf/82094674.pdf
[4] About Feature Scaling and Normalization. Sebastian Raschka'sWebsite. July 11, 2014
https://sebastianraschka.com/Articles/2014_about_feature_scaling.html
[5] Sohaib Zafar Ansari. Market basket analysis: trend analysis of association rules in different time periods. February 2019
https://run.unl.pt/bitstream/10362/80955/1/TEGI0458.pdf
[6] Ramesh, S. N. Market Basket Analysis - Why 'Lift' is an odd metric 2019, August 18
https://www.linkedin.com/pulse/market-basket-analysis-why-lift-odd-metric-nadamuni-ramesh/?articleId=6548498481789661184