ZE GERMANS LIMITED

Group G - Authors:

Li-lou Dang-Thai, m20200743

Maximilian Maukner, m20200645

Steffen Hillmann, m20200589

Ehsan Meisami Fard, m20201050

GitHub Repository

Table of Contents

1. Introduction

Data Mining is mainly uncovering hidden insights and information from the databases. Its functionality ranges from clustering, classification, prediction and link analysis. Link analysis or association rules were introduced in 1993 [1] and are utilized to identify underlying relationships among a set of items in a dataset. It largely entails analysis of market basket data which allows grocery retail stores and now even e-commerce grocery companies to understand the buying behaviour of their customer and hence a better and more sophisticated customer- and product segmentation.

In this instance, Instacart is a same-day grocery delivery and pick-up service operating in the United States, and Canada. Customers shop for groceries through the Instacart mobile app or Instacart.com from various retailer partners. The company then will connect the customer to an Instacart personal shopper that will pick up, pack, and deliver the order to their home, within the time frame requested.

The given dataset from Instacart contains information on each customers' order including the list of items that were purchased. It also entails the order details regarding the day of the week and time of the day as well as products within the each order and its respective department allocation.

The rest of this report follows the structure of the CRISP Methodology. Firstly, the business as well data mining objective of the given case is illustrated. Secondly, subsequent to an overview of the given dataset, the required steps for data preparation as the groundwork for further analyses are explained. Thirdly, an extensive exploratory data analysis is shown which depicts the initial underlying information and insights. Moreover, a Market Basket Analysis (MBA) has been implemented to find relations and associations among product types and a customer segmentation to divide the customers into similar clusters based on purchase behaviour as well as product preferences. Lastly, a more detailed attention has been paid to the appealing segments found by implementing a brief MBA.

2. Business Understanding

2.1. Business Objective

Instacart is using transactional data to understand which products a customer is likely to buy again, try for the first time, or add to their cart during their next shopping session. Due to the lack of internal analytical capabilities, Instacart has decided to hire an external consultancy to take full advantage of its data. Thereby Instacart's ultimate goal is to obtain an overview of its business as complete as possible. Moreover, the following questions should be addressed:

- What are the main types of consumer behavior in the business?
- Which types of products should have an extended amount of product offerings?
- Which types of products can be seen as substitutes?
- Which items are complementary?

Subsequently, due to the lack of extensive user data, it might be interesting to uncover groups that behave similarly as this relates to the overall goal and the identification of the main types of consumer behaviors.

2.2. Situation Assessment

The provided dataset is a relational set of files describing customers' orders over time. The dataset is anonymized and contains a sample of 200,000 grocery orders from more than 100,000 Instacart users. It is important to note that the given datasets are a fraction of the whole dataset of Instacart transactions. Although 200,000 orders information should be sufficient to identify significant insights, utilizing a fraction of the dataset could hinder this investigation by the acquisition of existing associations among the product types.

2.3. Data Mining Goals

The above-mentioned business objectives can be translated into a rather more technical data mining objective. In this case, the data mining objective is to identify relations and associations among product types, especially identifying complementary and substitute products, by applying an apriori algorithm and association rules mining. Moreover, a

subordinate objective is to cluster customers based on their buying behavior and product preferences which leads to further insight exposure. Subsequently, a *Market Basket Analysis* will be performed on top of the generated groups from the segmentation to show how they deviate from the overall population.

3. Data Mining Process

3.1. Data Understanding

The dataset was provided by Jane Doe, one of Instacart's district managers. It was delivered in csv files and separated into 4 data frames: orders, orders product, products, and departments.
The dataset contains details on the following topics:

- Orders - 200,000 orders, containing the number of items purchased, the time and date, the number of days since the prior order, and the *user_id* of the shopper who placed the order.
- Products - There are 134 different products available online, they each belong to 21 departments such as produce, dairy eggs, snacks, etc.
- Users - This dataset has a limited amount of user information containing 105,272 single users.

Furthermore, the dataset is anonymized and composed of 200,000 orders from more than 100,000 clients of Instacart. After analyzing the data through the DTale library, the only missing values found were in *days since prior order* which can imply that those missing values correspond to the first purchase of the customer. Besides, the data is mainly composed of numerical variables. The categorical variables represent the given IDs for each data frame (e.g. *user_id*).

3.2. Data Preparation

Data preparation consists of selecting features, cleaning the data, and eventually implementing feature engineering to discover new information. For this project, the provided datasets do not require any major data cleaning. Only the variable *days since prior order* in the *order* table contains Nan values which have to be separated before applying a segmentation, since for these missing records an attribution of a number, even with data imputation, would bias our results. Moreover, data transformation is required to perform MBA and Clustering. Specifically, several merges and pivoting were used to prepare the data for further analyses. For the MBA the final transformed table has the *order_id* as rows and the *product types* as columns, whereas for the shopper segmentation the rows of the final table represent the unique *user_id* and the columns contain information about the buying behavior (4 features) and as well as the product types (134 features).
Furthermore, due to the different scales in our final transformed segmentation table which is utilized for initiating clustering analysis, the MinMaxScaler has been used to transform and normalize the input features. Besides, Principal Component Analysis has been implemented to reduce the dimensionality of the dataset at hand. The principal components represent each a proportion of the explained variance of the set of observations, with its first principal component explaining most of the variance. In this case, the first ten principal components have been selected which represent a cumulative explained variance of 72,2% of the 138 initial observations.

3.3. Exploratory Data Analysis

EDA is a critical process of performing initial investigations and analysis to discover patterns within the dataset which are related to the overall goal of Instacart, namely to generate as much insight as possible given the datasets provided. As a result, the team expects to identify some type of customer's behaviors and extended product offerings. The following three key questions arose from our business objectives.

3.3.1. What Do the Customers Mostly Purchase?

There are a total of 21 departments at Instacart. The figure below shows the proportionality of departments from the total of orders made and the respective amount of reorders. As seen from the department's analysis most ordered

products are from the produce department with 29% followed by dairy egg with 17%, and snacks departments with 9%. The department with the highest reorder ratio is dairy-egg with 67%, followed by beverages department with 65% and produce department with 65%. Even though pets and alcohol are the departments that represent the smaller amount of order, they have a reorder ratio higher than 55%, meaning that the small number of people that purchase them do it frequently and by habit which seems reasonable as customers who live with pets, will frequently need to buy necessary items for their care. On the opposite, personal care has the lowest reorder ratio, as those products are the ones that last the longest with the prolonged shelf life which makes it possible to easily store the product within the household and thereby are not required to be purchased frequently. To sum it up, it can be seen that the customer more often reorder organic products and daily consumed items instead of non-organic products and personal care items.

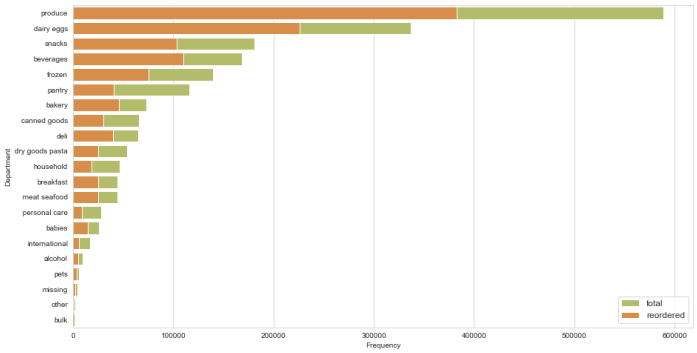

Figure 1: Total orders and reorders from departments

3.3.2. When do the customers mostly order?

After having a look at the type of product the customer usually purchases, it might be interesting to know the order time preference. The heat map below (Figure 2) shows the number of orders per day and hour. There is a clear effect of the hour of the day on order volume. Looking at the entire week most orders are made on days 0 and 1 with a value of almost 35.000 each day. Regarding the time of day, customers usually order in the same time frame during the week and weekends: between 8 am and 5 pm. Unfortunately, there is no information regarding which values represent which day, but one would assume that this is the weekend. Customers also tend to restock their supplies mostly on Saturday and Sunday, which can be seen in the notebook. Also, the least orders were placed on Wednesday. Another interesting fact is that most customers restock after a week or a month. It seems that some customers prefer to buy weekly and monthly supplies at once. Probably here 30 days represents the upper limit, and not necessarily any particular month. There is a continuous spike in orders from day 1 to day 6, showing that some customers are frequent buyers with a short window of restocking.

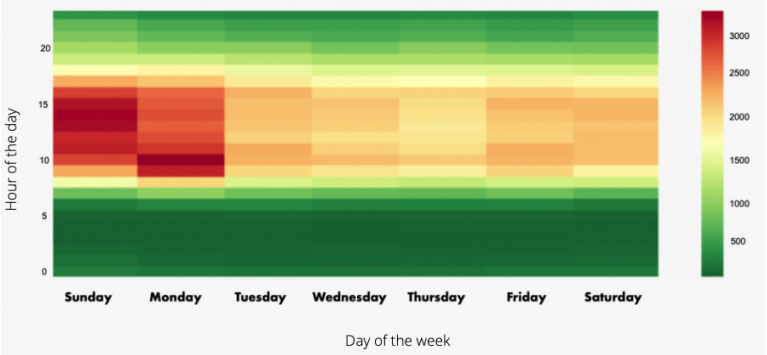

Figure 2: Number of orders by Time of Day and Days of the Week

3.3.2. Which product gets reordered frequently?

Reorders are an important part of the analysis, they represent product popularity and demonstrate products with a high probability of being re-purchased. This enables the identification of departments that require an extended range of products to satisfy Instacart's shoppers. Of all the orders, 12% have no reordered items, while the rest of them contain reorder products.

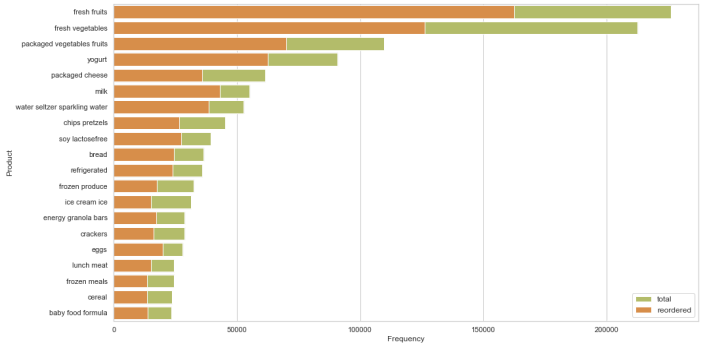

Figure 3: Total orders and reorders of products

The graph above (Figure 3) displays that customers mostly purchase fresh fruits, fresh and packaged vegetables as well as yogurt, milk, and cheese as they are items that are daily consumed. People have the tendencies of buying the same range of products that are related to food and drink. Milk, sparkling water, fruits, and eggs are the most common aisles the product is reordered from, as they are items that are daily consumed, and one rarely switches from their usual meal plan. More importantly, the high reorder rate implies that these products have a short time of shelf life. The least frequently ordered items such as haircare, skincare, kitchen supplies have a low reorder rate as they are the ones which last longer than others. For further visualizations and analysis, please refer to the Jupyter Notebook BC3_G_EDA.

3.4. Market Basket Analysis

Market Basket Analysis is one of the essential techniques used by retailers to uncover associations between different product or product categories. It is also known as association rule mining or affinity analysis which have been utilized to generate insight concerning the purchase behaviour of the customers. The analysis is mainly based on the product purchases that occur simultaneously in retail transactions and based on apriori algorithm which focuses on determining frequently occurring data and the association between them. This enables retailers to identify relationships and associations between the products that their customers buy.

Association Rule Mining (ARM) is the common practice for MBA when it concerns analyzing a large volume of sales transactions with a high number of products or product categories. There are three main metrics that unpack a substantial amount of information which are *Support, Confidence, Lift*, and *Conviction*. The output for each metric has a distinct interpretation which will be investigated thoroughly during this report.

Confidence is technically the conditional probability of occurrence of consequent given the antecedent. Yet it's crucial to be cautious while interpreting a high confidence value for an item set. Commonly, the confidence for an association rule having a very frequent consequent will always be high. As a result, the metric *Lift* is introduced to overcome this issue and prevent false and misleading interpretations.

Support is an essential measure because a rule that has very low support value may occur simply by chance. Moreover, low support is not interesting from a business standpoint as it might not be strategically correct to prompt items that rarely get purchased together. Moreover, *Lift* measures the correlation between itemsets from antecedent and consequent of a rule. If *Lift* of itemsets X and Y is larger than 1 (positively correlated), one could assume that these itemsets are complementary products. In contrast, if the *Lift* of the itemsets are smaller than 1 and thereby negatively

correlated, one could assume that these itemsets are substitute products. Lastly, *Conviction* is a methodology to estimate the robustness of the rule. It is a measure suggested [2] for overcoming some of the weaknesses of confidence and lift. In contrast to *Lift*, *Conviction* is sensitive to rule direction which could be alternatively stated as: *Conviction* $\{X \Rightarrow Y\} \neq$ *Conviction* $\{Y \Rightarrow X\}$.

3.4.2. Support and Confidence Thresholds

Before jumping into the analysis of the complementary and substitutes products, it is vital to set a reasonable threshold for *Support* and *Confidence* with an explanation.

Concerning the *Support*, a minimum threshold of 0.025 (2.5%) has been chosen due to different reasons. First of all, while doing the analysis, it's important to have high confidence in the results generated. It is highly recommended to be cautious with results generated as the transactions and the resulting relationships can occur by chance. In this case, the 2.5% minimum threshold helps to avoid this pitfall and filter the transactions that occur in rare cases. In addition, a low threshold value for *Support* increases the amount of rules found by the algorithm which increases the computation expenses dramatically. Regarding *Confidence*, a minimum threshold of 0.15 (15%) has been set. In section 2.3. of the 'MBA' Jupyter Notebook, a scatter plot has been depicted which shows the amount of rules found for *Confidence* and *Support* at each level.

3.4.3. Substitute Product Types

Substitute product types are identified through a *Lift* value below 1 which indicates a negative purchase correlation between two product items. In this case, a total of nine substitutes have been found. The table '2.4. Substitute itemsets' in the 'MBA' Jupyter Notebook illustrate the nine substitutes found sorted after an ascending *Lift* value. For instance, in the first two rows of the previously mentioned table, one can observe a one-sided substitution of Soft drinks with fresh vegetables. Despite the *Lift* value of approximately 0.717 and 0.815 for the corresponding rules which translates to substitution itemsets, the nature of these two product types makes one less confident whether these two products actually substitute each other. However, this information could be utilized for a recommendation system given the fact that customers who have soft drinks in their baskets have a lower chance of buying fresh vegetables or fresh fruits. Furthermore, the customer food consumption preferences might influence these results. For instance, a customer that has a healthy eating habit is less likely to buy sugary drinks or frozen pizza.

3.4.4. Complementary Product Types

Complementary products are items that are frequently purchased together in the same basket. In summary, a *Lift* above 1 indicates that the presence of an item on the left-hand side will increase the probability of the items on the right-hand side.

3.4.4.1. Two to One Complementary Itemsets

First of, a two to one relationship has been investigated. In other words, two refers to the amount of items in the basket as antecedents and one refers to the following item chosen (consequents).

In section 3.4. of the 'MBA' Jupyter Notebook, the table shows the top 20 complementaries according to the significance of the rule which is based on its *Lift* value. It can be seen that customers are twice as likely to purchase fresh herbs if they have fresh vegetables and fresh fruits in their basket. The mentioned itemset has a *Support* value of 0.061 which translates to 6.1% of the total transactions from the given dataset. Furthermore, according to its *Confidence* value, the probability of a customer purchasing fresh herbs with fresh vegetables and fresh fruits in their current basket is at 19.4%. As a further example, the next itemset has a *Lift* value of approximately 2 which also indicates that the customer is twice as likely to purchase fresh vegetables if fresh fruits and fresh herbs are in the customer basket. The probability of the purchase of the consequents is at staggering 88% which is quite significant. A total of 102

complementary products with two to one association have been found. An overview is provided through a heat map in section 3.5. of the previously mentioned Jupyter Notebook.

3.4.4.2. One to One Complementary Itemsets

In section 3.6. of 'MBA' Jupyter Notebook, the top 20 complementary items have been depicted which have a one to one association. Meaning, the antecedent and consequent consist of exactly one item. In total 111 complementary itemsets have been found.

Furthermore, in section 3.7 of the previously mentioned Jupyter Notebook, the complementary itemsets have been shown by utilizing a heatmap. It displays the complementary relationship between antecedent and consequent items by displaying the lift.

However, despite positive correlations found among product types, with a metric such as *Conviction* the range of products with high correlations can be narrowed while the significant correlations remain. Next, a new constraint has been added which should increase the strength of the rule shown. In this case, with the additional integration of a *Conviction* less or equal than 1.3 to the filtering only 31 rules remain and hence the output generated by setting a *Conviction* threshold narrows the amount of correlations substantially. The first most significant rules found have fresh vegetables as consequent and fresh herbs, canned jarred vegetables and canned meals beans as separate antecedents. As complementaries are mostly purchased together, Instacart e-commerce could utilize the information given to optimize its recommendation system once a user has an item in his or her basket.

In section 3.7, all the 31 rules have been depicted. Interesting to see that the very first and most significant rule determined by its high *Lift* value has disappeared. One can also observe that the rules left have a high *Confidence* which indicates that these rules have a higher significance than the rules found previously without the *Conviction* constraints. The information above is neatly represented in section 3.8 through a heatmap.

3.5. Shopper Segmentation

As stated above the main data mining goal is to use association rules to uncover complementary and substitute products, among others. However, it might be interesting to apply *Segmentation* to find similar patterns in the buying behavior of shoppers, as this gives Instacart even more insights into its business. In addition, MBA will be performed for the different clusters to show how they deviate from the overall population and to find the specific complementary and substitute products within each cluster.

That is, after transforming and reshaping the data, variables related to the buying behavior (e.g. number of times ordered) and the product types were used. All in all, we are left with 138 features for each shopper (user_id). Before conducting PCA for dimensionality reduction and performing clustering, it has to be noted that some orders had Nan values in the column *days since prior order,* which is due to the lack of information when a shopper has made a purchase for the first time. Hence, these orders were separated for our clustering algorithm.

3.5.1. K-Means Clustering

K-Means is a partitional clustering technique that groups unlabeled data points into K number of clusters while keeping the intra-clusters distances as small as possible. *K-Means*' advantage is that it is quite fast and can handle large datasets. However, the main drawback is that the algorithm requires to define the number of clusters prior to model execution.

One popular technique to select the optimal number of clusters is the Elbow-Method, which fits the model with a range of values for K. In our case the Elbow curve sets the optimal number of components to 4 with an R2 of 52%.

3.5.2. Cluster Analysis

Subsequently, the clusters will be examined based on the two perspectives mentioned above, namely *buying behavior* and *product types*. However, it can already be mentioned that the cluster composition is driven by features explaining the buying behavior as shown in *Figure 5*.

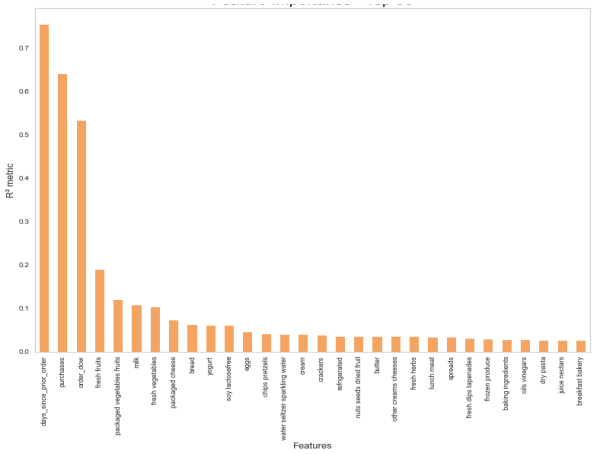

Figure 5: Feature Importance - Top 30

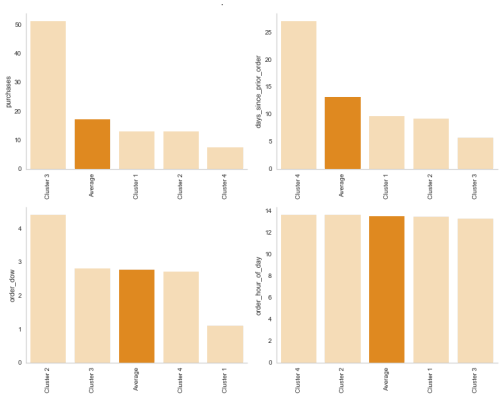

Figure 6: Cluster comparison to mean of variable

The *buying behavior* perspective includes the variables, *purchases (number of times ordered), Day since prior order, Order day of the week, Order hour of the day.* Here, two major discrepancies can be observed. Shoppers of Cluster 2 have ordered the most - on average 51 times - and made their last purchase the most recently compared to the other clusters - on average in the last 6 days. Thus, one can define these shoppers as frequent and loyal, ordering every week.

In contrast, shoppers of Cluster 3 are ordering the least (*Figure 6)* but made their last purchase on average 27 days ago. Thus, it can be argued that these are occasional shoppers who only order when they need something specific. This assumption is reinforced when taking a look at the *product type's* perspective.

No clear differences in product preferences are evident in the *product type* perspective between clusters. As previously mentioned the cluster composition is driven by features explaining the buying behavior. However, when taking a closer look at the average sum of items bought by each cluster, it can be seen that shoppers of Cluster 2 clearly outperforms the others - on average they buy 45 - whereas shoppers of Cluster 3 are buying the least items, on average 13. These findings strengthen the assumption previously made since the frequent shoppers also bought the most items whereas the occasional shoppers bought very few items.

Finally, one could argue that frequent shoppers of Cluster 2 are the most valuable for Instacart. That is, it might be interesting to identify the most consumed product types for shoppers of this cluster. Those are fresh fruits, fresh vegetables, packaged vegetable fruits, yogurt, and milk.

3.5.2. MBA for frequent and occasional shoppers

In this section, the most items purchased (Support) for each cluster is analyzed with the addition of examination of the complementary and substitute items for each cluster. Important to note that all of the rules are composed as one to one relationships.

To narrow the magnitude of the total amount of complementary itemsets found, the following filter has been implemented: *minimum Support* of 0.05, *minimum Conviction* of 1.3. In an effort to find complementaries and substitutes, the *minimum Lift* is set above and below 1, respectively.

3.5.2.1. Frequent shoppers

Due to the importance of the frequent shopper (second cluster) for Instacart, a brief and exclusive Market Basket Analysis for the second cluster has been implemented.

Firstly, in the table 4.2.2 'Support itemsets - Frequent shoppers' the most frequent itemsets are depicted. As a result, *fresh fruits* is by far the most purchased among frequent shoppers with a *Support* value of 0.608 followed by *fresh vegetables*, *packaged vegetables fruits* and *fresh vegetables with fresh fruits*.

Furthermore, the complementary and substitutes product types among the frequent shoppers is investigated. Please do take into consideration that the thresholds mentioned in the section above have been implemented.

All of the 42 complementary itemsets found can be seen in the 'Clustering Analysis and MBA' Jupyter Notebook in the table in section 4.2.3 and in section 4.2.4 'Heatmap - Complementary itemsets for frequent shoppers' through a heatmap.

Concerning the substitute itemsets among frequent shoppers, 12 substitutes have been found, however, similar to the examination of the substitutes for the whole population of shoppers, these substitutes have a minimum *Support* of 0.025. A heatmap for substitute itemsets for frequent shoppers has been plotted in section 4.2.8.

3.5.2.2. Occasional shoppers

Moreover, the purchase behavior of the occasional shopper is investigated. Similar to the frequent shoppers, occasional shoppers' the most frequent purchase is *fresh fruits* with a *Support* value of 0.5 which is slightly lower than the frequent shoppers followed by *fresh vegetables*, *packages vegetables fruits* and *fresh vegetables* and *fresh fruits*. However, the table 4.3.2 'Support itemsets - Occasional shoppers' shows that the purchase behaviour of these two clusters differentiates themselves from the 7th itemsets below. On the one hand, the 7th itemsets of the frequent shoppers according to the support value is *milk*, yet on the other hand, the 7th itemsets of the occasional shoppers is *yoghurt*. The overlap among frequent shoppers and occasional shoppers for the first 6 itemsets indicates that healthy stock management for the first 6 product types are crucial. Unavailability of these itemsets could lead to the disappearance of the frequent shoppers. Moreover, it is essential to notice that the occasional shoppers have always the potential to be converted into frequent shoppers, and therefore the stock management of their favourite items should not be neglected.

Next, the complementary and substitute itemsets among occasional shoppers are investigated in conjunction with the minimum threshold mentioned in section 3.5.2 and a summary of the complementaries and substitutes have been depicted in sections 4.3.4 and 4.3.6, respectively.

4. Evaluation

4.1. Business Objectives Review

Besides maximizing the information and insights from the datasets provided, the four main objectives of this business case were to identify main types of consumer behaviour, determine the types of products that should have an extended amount of product offerings, and lastly to discover types of products that work as either complementary or substitute itemsets. In this report, the most and less frequent types of products and departments have been identified through an extensive exploratory data analysis.

Furthermore, the main types of customers have been identified through customer segmentation (clustering) based on the features that are associated with the frequency and recency behaviour patterns.

Moreover, the complementary and substitutes for the whole population of Instacart customer base has been analyzed and in addition for two clusters which are identified as frequent shoppers and occasional shoppers.

4.2. Limitation and Future Work

Throughout this report, while utilizing the apriori algorithm to find complementary and substitute itemsets, a 1:1 relationship among the product types has been investigated. This approach narrows the scope of this analysis and

focuses on antecedents and consequents with only one item.
As future work, a 2:1 or 2:2 product types relationship could result in different output and further insights. The information generated can be utilized as groundwork for building a recommendation system.

5. Deployment

Instacart's management has hired us to conduct this ad-hoc analysis to generate as many insides of its business as possible. However, the provided data probably represents only a certain period of customer transactions. Therefore, we believe that Instacart should invest in Business Intelligence solutions to keep track of the buying behavior of their shoppers since those change dynamically in time. Hereby, one provider of such services could be Microsoft PowerBI as it is a powerful data analytics tool and easy to implement.
Furthermore, we felt that too few variables were made available for shopper segmentation. Therefore, we believe that an investment on the data collection process, such as socio-demographics, will gather substantial benefits for the advised BI solution.

6. Conclusion

Overall, we were able to help Instacart and its management to achieve their initial business objectives, which was to obtain an overview of its business as complete as possible. Before performing MBA, we analyzed the entire dataset to identify some type of customer's behaviors and extended product offerings.
Furthermore, the complementary and substitute itemsets of all Instacart customers, have been inspected and subsequently depicted through a heat map.
Besides, we performed K-Means for Clustering to find similar patterns in the buying behavior of shoppers, as this gives Instacart even more insights into its business. Out of the four clusters, two interesting groups were identified, namely the frequent and the occasional shoppers. Subsequently, MBA was applied again for these groups to show if they deviate from the overall population.
Finally, we recommend Instacart to develop and implement a BI solution to dynamically monitor shopper transactions over time.

7. References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, Mining Association Rules Between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, Washington, D.C., May 1993

[2] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In J. Peckham, editor, Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, pages 255–264, Tucson, Arizona, 13–15 June 1997.