BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Instacart

Market Basket Analysis

Group T

Ana Amaro, number: m20200598

Filipe Lourenço, number: r20170799

Gonçalo Almeida, number: m20200594

Guilherme Neves, number: r20170749

April, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. BUSINESS UNDERSTANDING 1
   1.1. Introduction and Presentation of the Business Objectives 1
   1.2. Situation Assessment 1
   1.3. Data Mining Goal 2
   1.4. Project Plan 2
2. DATA ANALYSIS 2
   2.1. Data Understanding 2
   2.2. Data Preparation 3
3. MODELLING 4
   3.1. Modelling Technique 4
   3.2. Model Construction and Assessment 5
4. EVALUATION 6
   4.1. Evaluation Results 6
   4.2. Review Process 9
   4.3. Determine Next Steps 9
5. DEPLOYMENT 9
   5.1. Deployment Plan 9
   5.2. Plan Monitoring and Maintenance 10
   5.3. Conclusions and Brief Review of the Project 10
6. REFERENCES 10

1.Business Understanding

1.1. Introduction and Presentation of the Business Objectives

Instacart is an American company that operates both in its home country and Canada. They differentiate from the common grocery stores by providing a system of delivery and pick-up via smartphone application and their website. The deliveries are made on the same day the order is registered, and users are not only able to select from a vast range of products, but also to speak directly with the Instacart's member that will be responsible for collecting the items, allowing for a deeper customization. Due to the nature of the business, Instacart was able to collect large amounts of data and store it in a transactional database, but is having troubles extracting the most value from this data. It was in this context that our team was consulted. Jane, the owner, asked us to help her understand the main types of customers' behaviours, the products that can be seen as substitutes and as complements, while also trying to figure out, in a data-driven way, which products' segments should have a more differentiated offering.

Considering the requests presented by Jane and how the quality of the work is dependent on the true patterns existing across orders (we are just uncovering those), we will have to rely on subjective business success criteria. Firstly, we hope to be able to understand the behaviour of the customers in such a way that Instacart can leverage that knowledge in their business' operations (e.g., manage traffic allocation and employees' schedules by understanding the busiest days and hours). Moreover, we expect to uncover patterns between products and products' categories, in a way that allows for efficient marketing campaigns (e.g., bundling) and website optimization to improve the convenience for customers (which may improve the retention rate) and to increase their expenditure with Instacart.

1.2. Situation Assessment

Jane provided us with four CSV files containing the transactional database constructed by Instacart. In those, we can find information regarding the products (in an aggregated way) and the departments to which they belong, the products purchased per order (including the order by which each was added to the cart and if the item was a reorder or not) and specific information about the context of the order (which user made it, when and how much time had passed since the last order).

We agreed with the timeframe purposed to prepare a presentation of 5 minutes to the management team and to deliver a full report regarding the steps followed and the recommendations we want to provide. Moreover, the main costs agreed are the working hours our team will devote to the project and that will be billed to Instacart in function of the benefits achieved. The main risks associated with this task are related with the quality of the work being highly dependent on the need to find not just patterns, but actionable ones. Nevertheless, in this case, we can mitigate the risks by making use of the appropriated filtering criteria and taking advantage of the low-cost experimentation Instacart can make to test any of the rules made (e.g., changing the layout of the website). Furthermore, there is also the risk that some previous marketing campaigns were the cause for some association rules we might find, but this can be fixed by assessing this with Instacart at the end of the project. From the terminology point of view, the language used by Instacart is straight-forward and there does not seem to exist the need for any clarification.

1.3. Data Mining Goals

The main data mining goal for this project is to be able to uncover patterns in the behaviour of Instacart's customers, particularly in the way they purchase, by recurring at market basket analysis techniques. Although the goal seems simple, the insight found should be actionable and not artificially generated (e.g., result of previous campaigns and/or current website layout organization), and as typical from this type of problems, we expect a significant number of rules to emerge from data, being that we will have to filter them in a way we convey as much information as worth to act upon, while finding strategies to communicate the messages effectively to the Instacart's team.

To measure success, we will compare subjectively, and assess by majority vote of our consulting team, if the final insight gathered uncovers a significant part of the actionable patterns existing in the data and if our team was able to convey them in an interesting and action-driven way to the management team, meaning this, we should recur to visualization tools as often as possible.

1.4. Project Plan

Concisely, our project will be based on the following steps: 1. Retrieve the datasets and gather some superficial understanding of the data and variables; 2. Construction of some visualizations to help get a deeper insight about the data; 3. Integration of the different datasets into one and simplification of the created dataset to the points of interest (e.g., we already suspect that we will be more interested in simple co-occurrences than in knowing how much of each item was purchased); 4. Define the model(s) to use, create the conditions to apply and construct it, based on the data analysis done and on the business and data mining goals (we already expect the use of the Apriori and PrefixSpan algorithms); 5. Filter the rules found and construct visualizations that convey those rules; 6. Verify if the model meets the data mining success criteria, contextualize the importance of the findings to the problem at hand and create a visual way to communicate those (we are considering the preparation of a dashboard); 7. Present the report (and potential dashboard) to the management team and if the approval is met, deploy the plan.

2.Data Analysis

2.1. Data Understanding

The data collection, although involving the importation of four different files, was fairly straightforward. During the data analysis, mainly recurring to visualization methods, we were able to extract some insight that helps us to already understand a part of our customers' behaviour. Firstly, we have discovered we are dealing with data from 200,000 orders belonging to 105,273 customers.

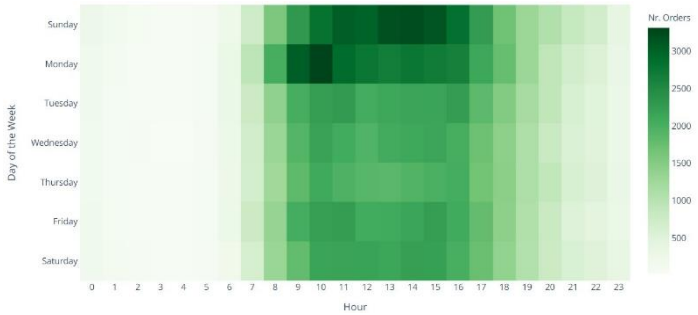

Figure 1 – Distribution of the number of orders, per day of the week and hour.

As shown above, it is also clear that the most active hours of the day are between 9:00 and 17:00 and that there seems to exist a significantly higher number of orders made on Sunday and Monday. Moreover, we also explored other aspects, as the distribution of products per department, the number of items purchased from each department, the right skewed distribution of products purchased per order and the reorder rate of different products, but for brevity reasons we will leave those out of the current report.

From a quality assessment point of view, the data seems to have been properly collected, with the main exception being some "missing" labelled products and departments and some missing values in the variable containing information about how long has passed since the customer's last purchase. In the case of the former and considering the "missing" label products were not aggregated to the category "others", we have reasons to believe the type of products to which they refer and the department to which they belong have been permanently lost due to a collection issue. Being this said, regarding the latter, we have the missing value problem with around 11.6% of our customers, which corresponds to information regarding the first-time ordering of a customer, so "actual missing values".

Still discussing the quality of the data provided, we noticed that the sample of orders provided consists of spare orders across customers, meaning this that for a particular buyer that, for instance, we know has made at least 90 purchases, we may just have 4 or 5 non-consecutive orders. Although this seems unimportant, one of the hypotheses we want to test was if it would be possible to segment the customers based on the products that they buy and then apply the market basket analysis on the top of each "cluster" encountered, which considering this information is not possible. In case we had personal data about the customers, we could still cluster based on that information and then analyse the typical purchase behaviour of each cluster, but considering that the data we have available for clustering is almost exclusively regarding their buying pattern, using a clustering technique would be incorrect: we would need representativeness of the orders made per customer, which is not possible to be assured as there is not enough amount of orders per customer (more than 25% of our data consists of only one order's customers). In practice, what this means is that while we can look at the overall transactions and attempt to find patterns, as we can assume that 200,000 orders are a large enough sample of the population, we cannot attempt to find patterns in purchasing habits of segmented users, as we lack a representative sample of each user's habits.

2.2. Data Preparation

Although the data provided was well constructed and presented a good level of quality, in the analysis we intend to do, we will not need a significant part of the variables provided. Firstly, regarding hour of the day, day of the week and other variables related with customer insight, we consider that the exploratory analysis made above already helps to uncover some insight that will be useful to answer the business goals requested to us (particularly in understanding the behaviour of the customers), but for the next steps of our process, those will not be relevant. On this same note, from the market basket analysis point of view, information regarding if a specific item was a reorder or a first-time purchase, will not be significant. Furthermore, and for reasons already mentioned (the "missing" label seems to constitute a "truly lost label"), we will also disregard from the remaining of our analysis all products/departments with this label.

Before moving on with our project, we decided to integrate the different datasets we had. Considering we were provided with the schema associated to the relational database and the relations

among tables, it was a fairly easy process to merge the different datasets by recurring to the foreign keys that established the bridges between datasets. Lastly, and already anticipating the analysis we will need to develop, we constructed a pivot table with only the co-occurrences between products (disregarding aspects like the amount of each item purchased) and an array that stores arrays reflecting the "purchase" path of the customers by departments.

3.Modelling

3.1. Modelling Technique

Considering the desire to make a marketing basket analysis, we will use modelling techniques that are usually associated with this analysis. The algorithms we will use are the Apriori and PrefixSpan algorithms, and the main requirement for these tools are to be "fed", respectively, with a co-occurrence matrix and with an array containing, for each order, an array with the departments from where an item was bought, reflecting the order by which each was added to the cart. The first model, Apriori, allows us to explore the most frequent itemset present in our data, and we will conduct this analysis at the product level. Furthermore, on the top of the Apriori algorithm, we will use association rules, recurring for this to the mlxextend's python implementation, which will allow us to shape the patterns found to get insights about the relations of complementarity and substitution, while assuring the strength of the patterns. Regarding the second model, PrefixSpan, it will help us to optimize the website's layout, by understanding the most frequent sequence of departments to which our customers go to and in which order. Lastly, and although not being a modelling technique (but being one of the requirements for the success of this project, including the data mining phase), we will construct a dashboard with the main results found and some contextualization of their meaning.

The association rules we will base our analysis on will be the support, the lift and the conviction. Although the first two are widely used and accepted in the industry, the last one is less known. The conviction appears as an attempt to face the problems associated with the lift (agnostic to the direction of the association rule, fails to capture perfect association and focuses on co-occurrence instead of implication) and with the confidence (widely dependent on the frequency of the consequent, lack of independence detection capability and does not consider both antecedent and consequent individual's supports), and measures the degree of implication of a certain rule[1], by assessing the frequency according to which a certain rule found is wrong, compared with the frequency that would be expected to be wrong if the events were independent[2]. Although the weakness of confidence is well addressed in literature (and being also one of the reasons why lift is usually used to complement it), the lift's weakness is not as well explored, so we will present one example where the conviction is a more intuitive measure: Suppose the Census data revealed that 50% of the population are men, 60% of the population have never given birth and the intercept between these two events is 50%. From a domain expertise point of view, we can clearly understand that once we know someone is a man (antecedent), we know he never gave birth (consequent), so the events should account for a perfect association scenario. Nevertheless, by computing the lift, we get a rounded value of 1.67, independently of the direction of the rule, which although indicating non-independency is an oddly low value, which can be interpreted as people who are man (or have not given birth) are 1.67 more likely to not have given birth (or being man) than if there was independency. Following the same events, but using the conviction, we reach an infinite value for the rule man → not given birth, and around 3 for the rule not given birth → man. These values seem significantly more intuitive and better

at capturing the association between events, being that people that are men are sure to not give birth, and not giving birth also allows us to understand that there is a significant association with being a man (value higher than 1). Overall, and although this metric does not legitimate causality, it helps to stablish direction, and by combining this metric with the easy-to-understand concept of lift, we will be able to limit some rules that only are relevant in one direction (as we would with confidence but considering both the probability of the antecedent and consequent [3], not allowing for an excessive importance being given to the probability of the consequent).

3.2. Modelling Construction and Assessment

With the models to use chosen, we must define the correct parameterization of the association rules thresholds we intend to define. Regarding this, we will follow different approaches according to the different analysis we want to conduct.

Firstly, in order to understand the substitute products (and here we will only look at one-to-one substitutes), we will only focus on products with an individual support higher than 6.25%, being this the threshold we considered significant to act upon (around 12,500 transactions), but not conditioning the minimum level of support for the two products combined (if the products are perfect substitutes their combined support would be 0). Furthermore, and knowing it would be necessary a lift and conviction below 1 to establish substitution, we decided to push the threshold a little further (to assure the results were more significant and better filtered) and set the maximum lift to be 0.9, with the conviction lower than 1 and/or vice-versa.

Regarding the complementarity of the products, we not only looked at products that were complementary but also at complementarity between itemset. Once more, we defined the support to be 6.25% (in this case for both individual items, but also itemset, using the Apriori algorithm in the "classic way"), while this time, to find complementarity, we decided to define either the lift or the conviction to be at least 1.5, while the other being at least 1.25.

Lastly, in order to understand the shopping path of the customers across departments, we will define the minimum number of transactions needed to constitute a "path" to be 5,000 transactions (what corresponds to a support of 2.5%), using the PrefixSpan, as mentioned above. Note that in this case, we will not limit with other measurements as we did with the products, as we are pursuing a different goal. While previously we wanted to assure strong rules of association, here we are just interested in understanding the most common sequences followed by the customers in their shopping experience, consequently considering "paths" just based on support is enough for the general layout mechanism we intend to put in place.

Regarding the results obtained, in the substitutes products, we were able to discover 27 rules, with the majority including the product "soft drink". While some of the products seem to be intuitive substitutes (e.g., "frozen meals" and "fresh herbs", as "fresh herbs" involve the cooking process that "frozen meals" replace it), others are less easily interpreted that way (e.g., "paper goods" and "fresh foods"). Nevertheless, given the thresholds we enforced, we can, with a reasonable degree of certainty, make the argument that even if not intuitive substitutes, the rules found report to cases where products are purchased together less frequently than expected, which can be motivated by things like the lifestyle of the buyer (e.g., a significant number of "substitutes" to the "soft drinks" were

food considered “healthy”, so although a “banana” may not seem to replace a “Coca-Cola”, maybe the buyers who purchase one or the other are into different food-lifestyles).

From the complementarity point of view, we were able to find 61 rules, being 8 between 2 products, 36 between 3 products and 17 between 4 products. All the 61 rules were carefully submitted to different metric’s thresholds, so we are confident that all the rules discovered provide valid insight that Instacart should take into consideration. Following this, “fresh vegetables” was the category that appeared more often on the antecedent side (24 times), while “fresh fruits” was the most present in the consequent side (32 times). Despite this, and whereas as a set “fresh fruits” alone was also the most frequent consequent (23 times), when considering the most frequent antecedent (as a set) it was not “fresh vegetables”, but “package cheese” and “fresh fruits” combined (4 times).

Finally, from the department analysis, we got 3,503 “paths”, which after some further cleaning - mainly because whenever a customer purchased twice or more from the same department it was creating paths like (“produce”, “produce”, “bakery”), which for our purpose we want to treat the same as (“produce”, “bakery”) - was narrowed down to 1,363 “paths”.

In overall terms, our team believes that the data mining goals have been achieved. Firstly, we believe that the use of the strict rules applied helped to ensure that we were only collecting valid insight and significant enough to be worth to act upon. Furthermore, we also consider having been able to find appropriated tools to communicate effectively with the management team the findings, being that the visualizations and insights are summarized in the dashboard available here (due to limitations of free version it may be required to refresh a few times). Note that this dashboard is essential to the proper achievement of our data mining success criteria, being the mechanism created to communicate with the management team, avoiding thus having to provide here deep tables that would have very little interpretability to the management team.

4.Evaluation

4.1. Evaluation Results

In a more general overview, we believe to have been able to address the concerns of the Instacart’s manager. Firstly, we were able to understand the most interesting behaviour from the customers regarding when they buy and which products they buy the most. With this information, we hope to give Instacart the tools to understand in which hours and days they should have more and less people working and how to optimize the traffic to their website, as slow response of the platform may drive customers away. Furthermore, and as provided in the tab “Products Time Analysis” of the dashboard, we have created a visualization tool that allows for Instacart to understand the most purchased product per day, creating the possibility to make day-specific discounts. Lastly, by also providing the most often “re-ordered” departments and products, at tab “Loyalty Products” of the dashboard, we believe to also have created a way for Instacart to understand the most “loyal-creation” products/departments.

On another note, and following the results presented in the previous section, we believe to have encountered products that behave as substitutes, and although some of the rules found may be more related with the type of customers each product targets than the direct substitutability of the products, the metrics used to judge the strength of the rules are enough to validate the pattern.

Knowing the existence of “substitutes”, may help Instacart to know not only which products not to bundle, but more importantly, which products to not try a cross-sale with, as it would be wasting the customer’s attention span. All the information regarding substitutes products is available in tab “Substitutes Products“ at our dashboard.

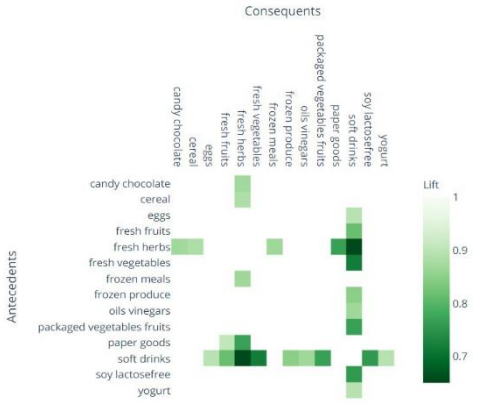

Figure 2 – Relations found for Substitute products, coloured by Lift.

In what concerns the complementarity of products and itemset, in “Complementary Products” at our dashboard, we provide a tool that Instacart’s marketing team can use to understand, according to the rules we defined, which products tend to be associated frequently and significantly. This tool can be leveraged to attempt cross-sales of products, for instance, by “popping-up” complementary products once a user adds something to the cart. Moreover, and while there is the need to discover if the rules found are due to direct complementarity between the products or due to some more indirect relation, we believe that the tool provided can also guide Instacart’s team to try to understand this dynamic, both by running promotions on some product(s) and keeping track if the demand for the “found-complementary” increases - which if it happens in one side-only (running promotion on A increases also sales of B, but running in B doesn’t increase sales of A) may suggest causality, while if it happens in a two-sided way may prove the direct complementarity (promotion on A increases demand for B, and vice-versa)-, or simply try bundling promotions combining the two or more items that seem complementary. On this bundling opportunity, we would like to add that given the existence of a significant number of rules with perishable items (e.g., fresh vegetables and/or fruits), it may create the opportunity to bundle “close to expiration date” products, lowering the price the costumer pays due to be in bundle, while allowing the store to profit from products that would otherwise be wasted.

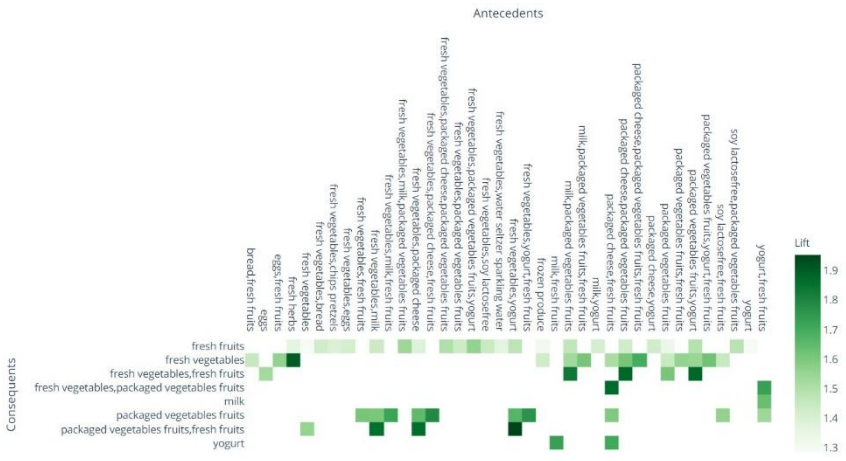

Figure 3 – Relations found for Complementary products, coloured by Lift.

As requested, we have also looked at some opportunities for a more extended offer of products. On this regard, and although the rules found do not open questions like "the customers are buying meat and vegetables from us, so why are they not buying the rest of the meal (e.g., rice or pasta)?", we were still able to gather some interesting points. From a broader point of view, it becomes clear that our customers tend to purchase especially "healthy" and/or "fresh" items (e.g., "fresh fruits" and "fresh vegetables or "cheese" and "milk"), being that these types of items seem a good section of products to increase the offering, as they not only have value for themselves, but as discussed, they tend to be associated with the purchase of other products. Moreover, the department to which these products belong ("produce" and "dairy eggs") are also the ones where the customer, that buys from those departments, tends to purchase more units per order on average (3.94 items for "produce" and 2.49 items for "dairy eggs"). On the opposite note, it seems against the "general industry trend" that costumers do not purchase as frequently from the "bakery" department as it would probably be expected [4], even more considering their appetence for fresh goods. In this regard, there may exist some lack of offerings that are appealing to customers and that may justify the relatively low support of this department (around 27%), which should be investigated, in order to find out which products customers may feel are missing from here.

Lastly, by understanding the movements across departments, we expect to be able to optimize the layout of the website/application in a more convenient way to our customers. Regarding this, and although being a frequent practice to separate as far as possible frequently purchased together aisles/departments, in the Instacart's reality this may not be the best option. By being an online platform, this dynamic is not as effective and also this would go against one of the Instacart's values, which is to give people "more time to enjoy [the food] together"[4], therefore convenience being key (it would reflect poor understanding of Instacart's culture to suggest them an approach against their values). Our suggestion regarding this would be to use a dynamic layout, based on support (with the most frequent department appearing on the top), once the customer enters the website, but that progressively changes as the shopping experience enlaces. What this would mean is that if the "produce" is the department with highest support, it would firstly appear in the first position, but after the user shops from one department, the department appearing first (after the one the user is currently on), should be the department that most frequently is visited after the current department (let us assume to be "dairy eggs" after the user has been on "produce"), and then it would proceed with this behaviour to look for the next reordering (in our example, if in fact the customer goes to "dairy eggs", we then would suggest the most frequently visited department after being firstly to "produce" and then to "dairy eggs"). Regarding this use-case, we prepared a "demo" version at tab "Layout Departments" from our dashboard.

Before proceeding to the review of the process followed, we would like to suggest, for future endeavours, Instacart to provide data with more granularity, in this case meaning that it would lead to more interesting rules and suggestions if instead of having product's departments and product's types, we had the possibility of looking at the actual products. On this same note, it would be interesting to apply a clustering technique to better understand our customers behaviour, and in order for that to happen we would propose that, in subsequent projects, personal data (e.g., age, gender,…) and/or representative purchase behaviour (either by having all transactions of the customers in study or for customers with a significant volume of purchases having just a large enough sample to assume representativeness) was provided for each customer. Moreover, the full project has been done with the consideration that any association rules was not the mere result of a previous marketing effort,

which is naïve, at best, but the only option, considering the lack of data regarding previous marketing campaigns, bundle offers and cross-promotional campaigns.

4.2. Review Process

When reviewing the data mining process, our team did not spot significant mistakes or failures. The process followed the structure initially proposed in section 1.4. of the report and complies with the CRISP-DM methodology to the best of our capabilities. On this note, we would like to add that on this project the CRISP-DM methodology is not ready to follow with perfect flow the needed steps (e.g., where should it be discussed the construction of a dashboard?), but we attempted to follow it as rigorously as possible. One of the main difficulties regarding this was with the first business goal (understand customers' behaviour), that partially did not involve the use of any model, which seemed hard to introduce as the CRISP-DM methodology appears created for the results to be almost exclusively achieved after/during the modelling phase.

4.3. Determining Next Steps

Considering everything done so far, our team decided it was appropriate to proceed to the deployment of the project. The possible steps for improvement (e.g., attempt different levels as threshold for association rules and/or ask for more data) would be time and effort demanding, and would delay the release and validation of the results, which in this type of project is crucial: being that testing is cheap (e.g., change of layout and quick-run promotions), any lag for an already good project could mean the loss of a significant number of iteration test cycles.

5.Deployment

5.1. Deployment Plan

Considering the practical nature of the results found and the need to contextualize them to provide an answer to the business goals (which includes the "intended application"), a significant part of the strategies proposed for our deployment have already been presented in section 4.1., and being complemented in the tool created: dashboard. Being this said, the suggestions provided still give room for the marketing team to explore them at a deeper level, not only with more domain knowledge, but also knowing which rules found may be due to previous marketing efforts.

On a more technical side, this endeavour should not be seen as an end in itself, as we believe the knowledge gattered, and the quality of the pattern mining techniques used would be significantly incremented if Instacart puts in place a continuous machine learning system. For this, the next steps would include the creation of an automatized pipeline that would "feed" each transaction made in real time to the algorithms and mechanisms used in this project, leading to a constant update of the rules. In this regard we would suggest, not only to use a classical storage for the orders, but also to recur to a "moving window approach" to complement it, as we believe it can be beneficial to explore trends in the food industry (e.g., different diets constantly emerging). With this, Instacart would have not only a system that would contemplate all data and keep providing the patterns found in the full database, but also one that could be used to find patterns in monthly or yearly basis, taking advantage of trends that could otherwise be dissolved in the full database. Lastly, the real-time updates should also be

linked with the dashboard created, allowing for the marketing team to have constant updates in a visual and non-technical way of what is happening under the hood.

5.2. Plan Monitoring and Maintenance

For this project, it is crucial to keep a close look at the results obtained with our deployment plan. Firstly, and considering the intention to build a real-time updated system, we expect to be able to track every single A/B test, bundling and data-driven discounts/recommendations attempted, to understand whether the quality of the rules we are targeting increases or not. Moreover, we also expect that some of the changes proposed will impact the retention rate of customers as it will open "paths of least resistance" to their shopping experience. Thus, it is important to monitor if the retention rate and the number of customers who make more than one order increases.

Lastly, we should have in place some triggers in our future automatized pipeline to detect if any rules found in this report stops being relevant, or any new ones appear, according to the criteria we defined.

5.3. Conclusions and Brief Review of the Project

Although the review of the project should be discussed with the various stakeholders - in order to not only identify possible pitfalls or processes that might be improved, but also to keep track of the results of the deployment-, being of particular relevance to have a meeting with the marketing team - to validate if some rules found were artificially generated (previous marketing efforts)-, we can already make a short overview of the process: All things considered, we believe to have created a tool - dashboard- for the marketing team to be able to quickly understand the insight found, while being able to capture the main relations among products and gather some insight about the customers behaviour. Furthermore, we believe that we have provided some useful data-driven suggestions of how to apply our findings, and to have proposed a proper deployment plan and monitoring system, that will enable for long-lasting results to be achieved from this pattern mining effort.

6.References

[1] Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. *Dynamic Itemset Counting and Implication Rules for Market Basket Data.*

[2] Fjallstrom, P. (2016). *A way to compare measures in association rule mining*.

[3] Peajapati, D. J., Garg, S. & Chauhan, N. C. (June, 2017). *Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment*. Future Computing and Informatics Journal, V.2, I.1, pp 19-30

[4] Hoeft, A., (2016, August 18). *Top 10 Grocery Items in America*. Retrieved from: https://datecheckpro.com/2016/08/18/top-10-grocery-items-in-america/