BUSINESS CASES WITH DATA SCIENCE

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

Business Case #4 – Market basket analysis

Débora Santos, number: 20200748
Pedro Henrique Medeiros, number: 20200742
Rebeca Pinheiro, number: 20201096

April, 2021

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

INDEX

1. INTRODUCTION

Thank you for choosing **Data4Business Consulting(D4B)** to help you in the challenge of understanding the purchasing patterns and behaviors of consumers. Our main objective is identifying the relationships between different types of products and present an overview about possible combinations among goods, considering the customers' choices.

The world is experiencing a great technological and digital revolution. Understanding business data, customer's needs, and which products to offer are essential for the business success. The exponential technological advances, such as data mining techniques, artificial intelligence, internet of things, can help taking the business to have a great advance.

Through innovative technological programs, well-referenced data mining methods and right business visualizations, the present report intends to give an overview of the process behind the analysis, presents the results and provides insights you need to be successful.

In addition to the present report, the following deliverables will be submitted:

- Outcomes presentation to Instacart business management.
- Jupyter Notebook with the code of the entire process.
- Dashboard application in a py file.

All files can be accessed in Github:
https://github.com/Debs86/Business_Cases_Projects/tree/main/BC3.
We are excited to be a part of this challenge.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Instacart is an online platform that provides a grocery delivery and pick up service in United States and Canada. With a wide products portfolio, the company provides to the customer different ways of combining products and the goods are delivered in the same day ordered. This e-commerce became even more popular during the pandemic period, once in addition to a fast shopping it is also safe.

The customers can purchase the groceries via a website or mobile app. The platform allows the users to better manage their shopping by speaking with the personal shopper assigned to them.

The key person in this business is the district manager Jane Doe. She has been trying to use the data to understand more about the business. She is looking to find as much useful information as possible, therefore she has reached D4B.

2.2. BUSINESS OBJECTIVES

At this point, Instacart uses transactional data to understand which products a user is likely to buy again, try for the first time, or add to their cart next during a session. The company is not taking full advantage of this data.

The principal goal is to provide an overview of Instacart's business as complete as possible, by answering the follow questions:

- What are the main types of consumer behaviour 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?

2.3. BUSINESS SUCCESS CRITERIA

The expected outcome will be to define customers' preferences based on their purchase patterns. D4B is responsible to present and accurate overview about the whole business, making assertive analysis about products combinations and understanding the products portfolio.

Another expected outcome of this report is to provide visualizations and also an application that could bring insights and useful information to Instacart.
The success of the proposed task will be evaluated by Instacart's district manager and, if needed, we will go back to the analysis and the app solution until we get an outcome that matches with the board's expectation.

2.4. SITUATION ASSESSMENT

2.4.1. Inventory of resources

This project was made following the CRISP-DM reference model (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process built in the end of 90's and it was built by more than 200 members lead by a consortium of big companies. *CRISP-DM succeeds because it is soundly based on the practical, real-world experience of how people conduct data mining projects.[1]*
From Instacart side, the project has the support of Management as well as the IT team.
On the D4B Consulting side, this project will be conducted by a team of 3 Data Scientists and Business Analysts.
We have been provided by the Instacar's IT team with a database composed by 200,000 orders, 134 products and 21 departments. The data consists in combining information about more than 100,000 customers and their purchases. The datasets provided are divided in four main areas:

- Departments
- Products
- Order products
- Orders

It was also provided a metadata file of the datasets.
In order to find meaningful patterns on the data, we use one of the most useful data mining technique: Market Basket Analysis (MBA). It involves the analysis of Association rules and from the analysis we could extracted the frequent items and detailed information about products bought together. *Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross sales campaign. [2]*
The main technology used to achieve the objectives of this report was Python. Python is one of most important and commonly used program languages in data science projects. The main packages for market basket analysis are mlxtend and AlgorithmX for network visualizations. We also used Plotly and Dash packages to produce visualizations and a final application.

2.4.2. Requirements, assumptions and constraints

The completion date of the present phase of this project is April 19th, 2021, but we expect to continue giving support and helping Instacart to achieve the next goals for the growth of the business.
The dataset does not provide the quantity of sales of each item, but we assume that the quantity sale is one item of the respective product by transaction. Also, the dataset doesn't provide sales values, so the analysis was done based only in the quantities.

2.4.3. Risks and contingencies

Table 2.1 identifies a list of risks and contingency proposed.

Risk	Contingency
There are no sales values in the dataset	Make the analysis based only in quantity
Insufficient number of observations	Ask for more observations (transactions)
Generalized products (product types)	Ask for more features (detailed products)

Table 2.1 - Risks and contingency.

2.4.3.1. Terminology

Business glossary:

- Reordered indicates that the customer has a previous order that contains the product.

Data mining glossary

Association rules: It is a data mining approach for exploring and interpret large transactional dataset in order to find unique patterns and rules. The rules are a method for findings interesting relationships between products. The rules also reveal the frequency of an itemset occurs in a transaction. All rules have an antecedent and consequent.

- Antecedent and Consequent: The IF component in the association rule sentence as known as Antecedent while the THEN is known as consequent. An antecedent, also called as rule body, is an item found in data and consequent, or rule head, is an item found in combination with the antecedent.

The most important measures to assess the rules are:

1. Support: Percentage of transactions that contain an item or a combination of itens . The rule that has low support not be interesting for business, once implies that the fact occurs by chance, being not profitable.

$$Support = \frac{Number\ of\ transactions\ that\ item\ or\ the\ combination\ appears}{Total\ number\ of\ transactions}$$

2. Confidence: Percentage that shows the percentage of a consequent appears given that the antecedent has occurred. It reveals how reliable a specific rule is. In example, given a rule X → Y, the higher the confidence more likely is for Y be in transactions that contains X.

$$Confidence = \frac{Number\ of\ transactions\ that\ the\ combination\ appears}{Number\ of\ transactions\ that\ the\ antecedent\ appears}$$

3. Expected Confidence:

$$EC = \frac{Number\ of\ transactions\ that\ the\ consequent\ appears}{Total\ number\ of\ transactions}$$

4. Lift: Lift is a factor by which the likelihood of consequent increases given an antecedent. If Lift is greater than one, it suggests that two items appear more together than separated. These products can be considered complementary. On the other hand, if Lift is smaller than one, means that two items appear less often together. These products can be considered substitutes. For last, if Lift value is 1, means there is no relationship between the products (independent).

$$Lift = \frac{Confidence}{Expected\ Confidence}$$

A credible rule must have a good confidence factor, a high level of support and lift higher than 1.

2.5. DETERMINE DATA MINING GOALS

The data mining goals states project objectives in technical terms:

1. Identifying the main types of consumer behaviors.
   Success criteria: Provide useful information and visualization to the findings. Also provide an application with possible update with new data.
2. Creation and Analyzing of association rules.
   Success criteria: Implement the a priori algorithms and be able to describe the main rules.
3. Evaluate the association rules and define substitute or complementary products.

Success criteria: Support threshold of 0.3. Confidence threshold 0.5. Lift higher than 1.3 for complementary products and lower than 1 for substitute products.

2.6. PROJECT PLAN

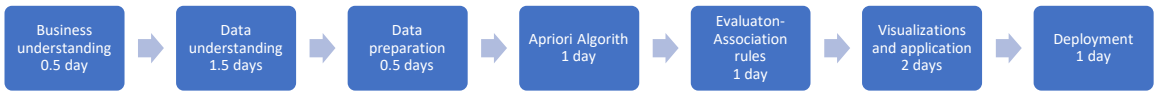

Figure 2.1 - Project's timeline.

Resources wise, for the business understanding we plan to use all the information provided in the kickoff meeting's presentation. For the core stages of the project, we plan to use Python to work the data provided by Instacart's IT team. To present the results, we expect to use Word for the report, PowerPoint for the presentation and Dash app system coded in Python to provide a user-friendly visualization of the results.

We consider the Modelling dependent of the Data preparation and data understanding state as the quality of the association rules will be directly connected with the quality of the input data. We also consider the evaluation stage dependent of the modeling step. During the project, we must go and back between Modelling and Evaluation many times, repeat this iteratively until we get the desired outcome and a good evaluation for the association rules.

For the modelling stage we aim to using Apriori and association rules libraries in python to create the association rules. We also used the library Algorithmx to demonstrate a network visualization that shows the relationship between the products.

We opted for using Apriori algorithm because it is very efficient, simple to implement and uses pruning techniques to avoid measuring certain item sets, while guaranteeing completeness [3]. The associations rules quality evaluation will be made using the follow measures: Confidence, Support and Lift.

3. MARKET BASKET ANALYSIS

In this section we go through the process of understanding and preparing the data, exploratory data analysis, the different algorithms used and, finally, the results evaluation.

3.1. DATA UNDERSTANDING

At this stage we analyzed the dataset to understand its potential and limitations. The data provided was split in four files: orders, orders_products, products and departments. We have used the Pandas library and the metadata provided to have an overview of the datasets: what are the variables, what they mean, number of variables and observations. We also check if there is noise and if there are missing values. We could find in the variables, *product_name* and *department,* one observation called missing, but we did not change it and kept the observation in our analysis. Also, in the variable *days_since_prior order,* there are 12.254 missing values, but it is not a noise. It means that the customer had not have a prior order on the data provided.

The four datasets combined presented more than 2million of items sold, of 200.000 orders, from 105.273 customers, with 134 products of 21 departments. It also contains information about day of week and hour of the day of each order, days since prior order, the order that the product was add to the cart and if the product was reordered or not.

3.2. DATA EXPLORATION

Understanding the purchasing patterns and behaviors of consumers is a key task for any retail company. At this stage, we are going into more details in the data and provide some insights to Instacar.

Regarding consumers behaviors, we could identify some important habits such as: Most of the orders were acquired in the day 0 (17,5%) and day 1(17,2%). In the other hand, day 4 is the one with less orders placed (12,5%). Regarding hour of the day, customers usually shop with more intensity from 10 am until 16 pm. On the data provided, most of the customers placed only one order, and if they placed more than one, most of them took 30 days to place the next order. In addition, the orders often have between 3 till 8 products.

Regarding the products preferences, some important notes should be considered in order to keep products that are profitable for the business.

First, we made a better overview about the products and ranked those with more frequent items sold having as result the products listed:

Products	Itens sold	%of total	Number of orders	%of orders
fresh fruits	226 039	11.19%	111 199	56%
fresh vegetables	212 611	10.53%	88 872	44%
packaged vegetables fruits	109 596	5.43%	73 083	37%
yogurt	90 751	4.49%	52 735	26%
packaged cheese	61 502	3.05%	46 199	23%

Table 3.1 – Overview products.

In terms of departments, the figure bellow shows the distribution of items sold by department.

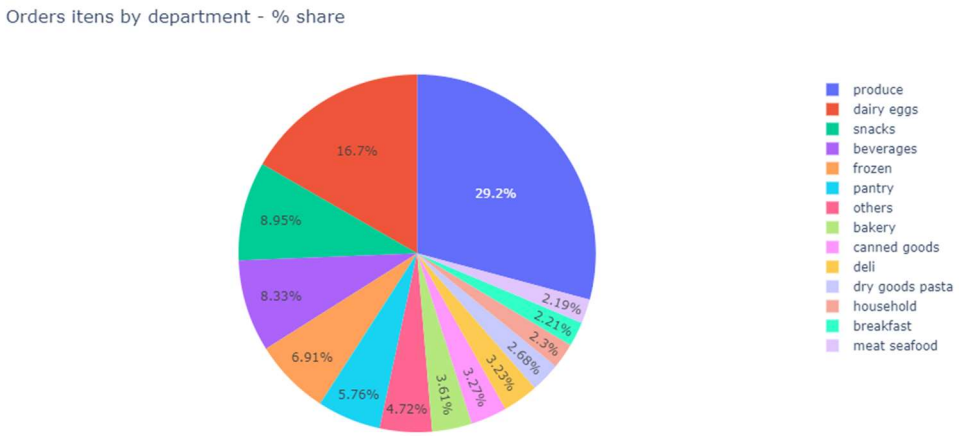

Fig. 3.1 – Overview departaments.

Getting deeper in the products analysis, the graphic bellow shows that more than 55% of items sold are represented by few products and they are present at least in 10% of the orders.

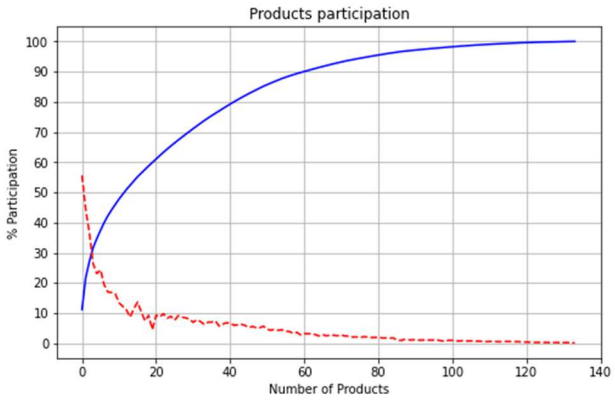

Fig 3.2 – Products participation.

So, we suggest to Instacar focus more on actions for these products: fresh fruits, fresh vegetables, packaged vegetable fruits, yogurt, packaged cheese, milk, water seltzer sparkling water, chips pretzels, soy lactose free, bread, refrigerated, frozen produce, ice cream, crackers, eggs, lunch meat.

3.3. APRIORI ALGORITHM, ASSOCIATION RULES AND EVALUATION.

In order to prepare the data to apply the apriori algorithm, we created a matrix in a binary format. Each row represents one order, and each column corresponds to a product. If the item was purchased, the matrix position is equal to 1, otherwise, the position is equal to 0. Binary datasets are interesting and useful due to its computational efficiency and minimal storage capacity [2].
We start the process applying apriori algorithm to discover the frequent items purchased. We set the support threshold equal 0.03. The main reason for setting this value was due the computational problems. Once trying values below the adopted, the code was not able to run considering the dataset size.
In the second step, we set the minimum level confidence equal to 0.50.
In the table below the number of rules generated are presented:

Metrics Threshold	***Number of rules***
Support >= 0.03 (unlimited number of products)	*1550*
Support >= 0.03 + Confidence >=0.50 (unlimited number of products)	*405*
Support >= 0.03 + Confidence >=0.50 + Lift >=1 (only 2 products, 1 antecedent and 1 consequent)	*404*
Support >= 0.03 + Confidence >=0.50 + Lift <1 (unlimited number of products)	*1*

Table 3.2 – Metrics.

Last to identify the complementary and substitute products different filters were added. Regarding the complementary products, we set the lift higher than 1.3 and in order to provide useful information, we consider only rules with 2 items. We could identify 33 rules.
Regarding the substitute products, the only filter added was lift below 1 that correspond for only one rule already presented on the table above.

4. RESULTS EVALUATION

To present the relationship between products, we used the Algorithmx library. It provides a network visualization where the edges show the relationship between the items bought together. The numbers presented next to the edges are related to the lift of that rule.
The figure bellow presents the complementary products.

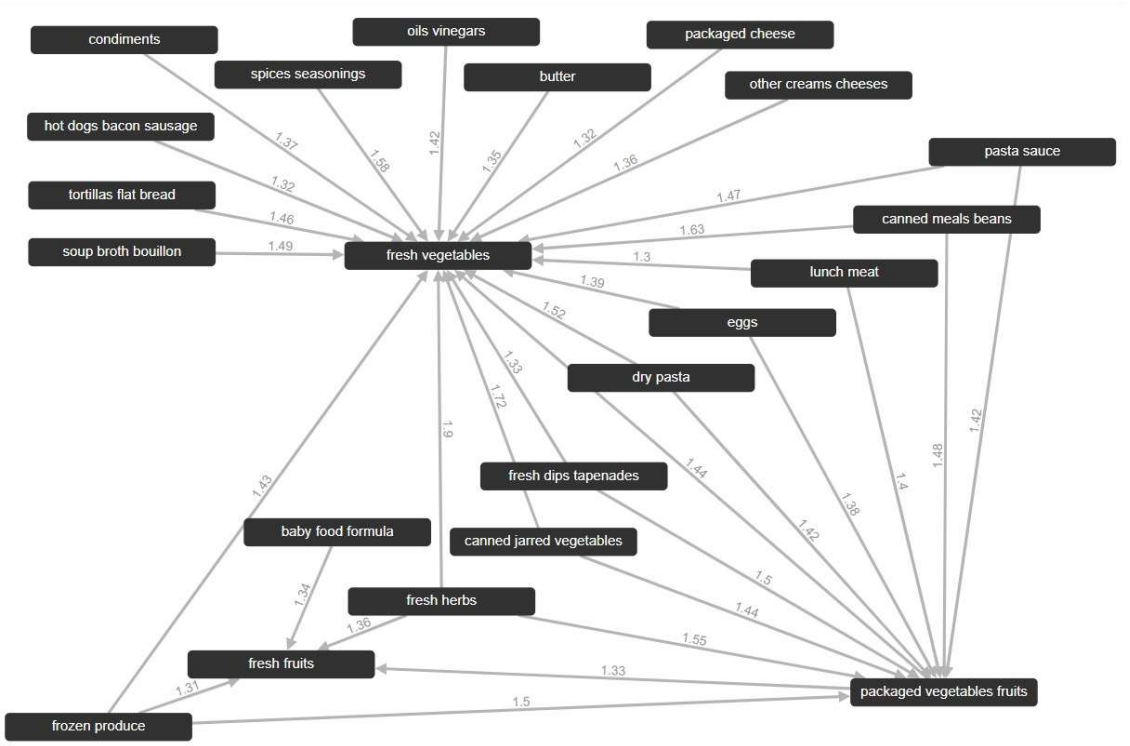

Fig 4.1 – Complementary products network

The 3 products that have the most of items sold are the ones that present relationship with many products.
The top 4 combinations of 2 products with the highest lift are:

- Fresh Vegetables + fresh herbs
- Fresh Vegetables + canned jarred vegetables
- Fresh Vegetables + canned meals beans
- Fresh Vegetables + spices seasonings.

More detail can be founded on the notebook provided.
In terms of substitute products only rule was identified:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
54	(paper goods)	(fresh fruits)	0.083575	0.555995	0.03209	0.504758	0.907847	-0.003257	0.896542

Finally, in order to provide interactive visualizations, D4B developed an application that compiles all the information regarding ”Consumer behaviors” and “Products and Department Analysis”. The dashboard consists in two interactive tabs where the Instacart business management can analysis separately the main important aspects about those two topics mentioned. The first tab, “Consumer behaviors”, four different charts were provided. The first graph displays about the number of orders by period and the user can choose if wants to analyze the information according to the day of the week or the hour of the day. The second bar chart shows how long it takes for the customer make the second order normally. The third graph informs how many products items consists in the basket order and the last one is a line chart that shows the number of orders that costumers placed. It is important to notify that all charts provided are interactives, which means that if the user pass the cursor by the charts, it is possible to see the value for the aspect desired.

The second tab “Product and Departments Analysis” is divided by two main sections: “Product analysis” and “Department participation”. Regarding the section “Product analysis”, four visualizations were created which the first one is a treemap visualization that displays products according two different criteria: “Items sold by products” or “Number of orders by product”. The user can select which treemap prefers to analyze. The second visualization is a bar chart which shows the products most reordered in percentage. In the third graph, the user selects one product to analyze the quantity sold. The user can choose if wants to see the items sold by day of the week or the hour of the day. The last visualization is a heatmap which shows the lift intensity of complementary products. The darker color label refers to products items with high lift.
Regarding the “Department participation” section, two different visualizations were developed. The first one is a donut chart, which informs in percentage the items sold according to the department. The last visualization is a treemap which displays how many orders were made by departments. As in the first tab (“Consumer behaviors”), this tab is also interactive, which also means that if the user passes the cursor on the visualizations, it is possible to analyze a particular value about the aspect.

5. DEPLOYMENT AND MAINTENANCE PLANS

We understand that the process of implementing the market basket analysis is very critical to your business. Retails shops market is very dynamic and every day, there are new purchases on the system. So, we understand that it would be very important to implement the analysis provided integrated with the grocery system. Integrating both, the results will be generated in a daily basis and the Instacart could react faster and implement promotion actions to the customers according with their preferences.
Including on this dataset the period of purchase (month), we could also include another kind of information such as change customer habits along time and the impact on sales. It would be also good include sales amount to see which products are more profitable and focus on increase their sales.
Ideally, the application developed would be integrated with Instacart system. But if it is not possible, the update of the application is made by simply updating the files with new data. The better way is creating a github to save all the files. D4B is going to provide a training session to the users of the application.
We are going to monitor the performance of the application after it is implemented for some months and make some adjustments if necessary. Also, we are going to monitor the efficiency of the rules generated.
In addition, we set some risks on this project. One of these risks is the data provided has only types of products. It would be very useful if we could implement the same analysis at the product level.

6. CONCLUSIONS

As state in the section **Error! Reference source not found.** – Business objectives, the main objective of this project was to provide a business overview to Instacart with main findings about purchasing patterns and consumer behaviors.
The results presented answer the 4 questions made by Instacart. On the section 3, we were able to describe consumer behaviors. In this same section, the products that represent the major of items sold were presented and they should be the ones with extend number of products offerings and the complementary products as well.
Regarding association rules, the process of discover relationship between products and the results presentation can be founded on sections 3 and 4. We could detect 33 rules of complementary products and only one rule of substitute products.
All the information presented can be used as a basis for decisions such as promotional support, inventory control and cross sales campaign.
We hope Instacart is satisfied with D4B work and we can continue working together.

7. REFERENCES

[1] Chapman, P, Clinton, J, Kerber, R., Khabaza, T., Reinartz, T, Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0*, CRISP-DM consortium

[2] Verma, N. (2017). *Market Basket Analysis with Network of Products*.

[3] Zafar Ansari, S. (2019). *MARKET BASKET ANALYSIS: TREND ANALYSIS OF ASSOCIATION RULES IN DIFFERENT TIME PERIODS.*

[4] H.Witten, I , A.Hall, E, J.Pal C (2016) Chapter 3 - Output: Knowledge representation , viewed at 12 April 2021, <https://www.sciencedirect.com/topics/computer-science/association-rules>.

[5] Gleeson, P (2019). Business Transaction Definition & Examples viewed at 15 April 2021, <https://smallbusiness.chron.com/business-transaction-definition-examples-25244.html>.

[6] Fever P. Business Transaction, viewed at 15 April 2021, <https://www.accountingverse.com/dictionary/b/business-transaction.html >.

[7] J. Mazlack, L. Considering Causality In Data Mining, viewed at 15 April 2021, <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.333.1386&rep=rep1&type=pdf>.

[8] Kumar, V (2005). Association analysis: Basic concepts and algorithms, viewed at 16 April 2021, < https://www-users.cs.umn.edu/kumar001/dmbook/ch5_association_analysis.pdf>.