

**BUSINESS CASES WITH DATA SCIENCE**

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

**Business Case 3 – Market Basket Analysis**

Group AC

Gabriel Cardoso, number: m20201027

João Lucas, number: m20200758

João Chaves, number: m20200627

Luís Almeida, number: m20200666

APRIL 2021

INDEX

[1. INTRODUCTION 1](#_Toc69309876)

[2. BUSINESS UNDERSTANDING 2](#_Toc69309877)

[2.1. Background 2](#_Toc69309878)

[2.2. Business Objectives 2](#_Toc69309879)

[2.3. Business Success criteria 2](#_Toc69309880)

[2.4. Situation assessment 2](#_Toc69309881)

[2.5. Determine Data Mining goals 2](#_Toc69309882)

[3. PREDICTIVE ANALYTICS PROCESS 3](#_Toc69309883)

[3.1. Data understanding 3](#_Toc69309884)

[3.2. Data preparation 3](#_Toc69309885)

[3.2.1. Duplicated Data 3](#_Toc69309886)

[3.2.3. Feature Engineering 3](#_Toc69309887)

[3.2.4. Outliers Detection 3](#_Toc69309888)

[3.2.5. Data Normalization and Encoding 3](#_Toc69309889)

[3.2.6. Feature Selection 3](#_Toc69309890)

[3.3. Modelling 3](#_Toc69309891)

[3.3.1. Model Selection 3](#_Toc69309892)

[3.3.2. Feature Importance 3](#_Toc69309893)

[4. Results Evaluation and Deployment of the model 4](#_Toc69309894)

[4.1. Main types of customer behaviour 4](#_Toc69309895)

[4.2. Type of products that should have an extended amount of product offerings 4](#_Toc69309896)

[4.3. Substitute type of products 4](#_Toc69309897)

[4.4. Complementary items 4](#_Toc69309898)

[5. DEPLOYMENT AND MAINTENANCE PLANS 5](#_Toc69309899)

[5.1. Deployment 5](#_Toc69309900)

[5.2. Maintance 5](#_Toc69309901)

[6. CONCLUSIONS 6](#_Toc69309902)

[6.1. Considerations for model improvement 6](#_Toc69309903)

# INTRODUCTION

Understanding the purchasing behaviours of customers is the standard rule for success of the new age of business. Rather than focusing on the product portfolio and making them the best deliverable, understanding consumer needs and tailoring our offer to those segments is more efficient and rewarding for both parties involved.

Market Basket analysis is a data mining concept used very frequently to solve this business problem. By analysing previous orders, we can obtain valuable insights about past and, with improvements to the flowing speed of data, about the present, in order to predict the future consumer behaviours and answer important questions such as: which products are complementary, substitutes or which ones should have a bigger or lower offer.

This work proposes an approach detect the afore mentioned relationships and provide a reasoning behind such behaviours. By the major use of market basket analysis algorithms such as the “apriori” algorithm, and metrics to a dataset from a retailer named Instacart, containing multiple orders, we will be able to generate useful insights that will help managers in their decision-making process by providing evidence of synergetic behaviours in a pool of baskets.

# BUSINESS UNDERSTANDING

## Background

Instacart is an online store that sells grocery products. The consumer places the order through a website or mobile app, and the products are either delivered to the customers’ house or picked up in a physical store.

The retailer has a vast portfolio of products and delivers them in the same day as the order was placed (obviously with certain time constraints due to scheduling and location), as well as personal assistance in case of problems with the delivery or related with the products themselves.

So far, the data generated by the multiple orders and transactions has been treated in a rudimentary manner and the manager team understands that it has potential for optimization and to give more insights than the ones that it is generating now.

## 2.2. Business Objectives

Aiming to provide an overview of Instacart's business as complete as possible.

* 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 sucess of this project will depend on answering objectivelly to the business objectives, therefore it will depend on:

* Defining concrete consumer behaviours that can be exploit in order to improve the service quality.
* Define exactly wich products can or should have an increased offer.
* Define exactly wich products are substitutes.
* Define exactly wich products are complementary.

## 2.4. Situation assessment

This 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**. For each user, we provide a few of their orders, with the sequence of products purchased in each order. Due to the fact the original dataset was too large, the products were grouped by their types, resulting in a total of **134 generalized products**. We also provide the week and hour of day the order was placed, and a relative measure of time between orders.

## 2.5. Determine Data Mining goals

Clusterization of Instacart customers based on the frequency of orders and itens bought;

Generation of association rules for the ordered products using Apriori algorithmn;

Optimization of minimum support in other to create more trustful and stronger rules;

Creation of visualizations to showcase the respective density and strength of the rules.

# PREDICTIVE ANALYTICS PROCESS

## 3.1. Data understanding

* There were 2.019.501 transactions altogether;
* 134 different items were bought overall;
* The most frequently bought item was “fresh fruits”: 226039 purchases;
* There were 124.342 single item baskets and the biggest basket included 100 items;
* Mean basket had around 17 items.

## 3.2. Data preparation

### 3.2.1. Duplicated Data, Inconsistencies and Missing Values

Since the data is related to an e-commerce business where all orders are important to get a good grasp of the relationship between each department their respective products, it was decided to leave all duplicate values as they are products/orders that represent a value input.

Missing values were checked, and they were related to a feature called days\_since\_prior\_order that represent the number of days for a customer to order a new cart. After looking at the metadata, it was possible to realize that all missing values were related to customer that didn’t order again yet. They were replaced by 45 since the value 0 can be addressed to customer that ordered two times in the same day.

### 3.2.3. Outliers Detection (Por mim, JP\_Lucas, apagava esta parte)

Despite of the fact that Outliers Detection is an important part of any data mining work, it was decided that, although there were some categories in the product and department features that will influence the analysis, they were considerate necessary to make a deep and viable scanning of the relationships that those values might have with the rest.

**3.2.4. Methodology Used**

For this project the methodology used was based on market basket analysis algorithms and the metrics associated with it.

In order to have the best insights we decided to tackle this problem in a two-side approach; firstly, we created association rules and draw conclusions from them in a broad way and regarding the full dataset, and secondly, we segmented the dataset into cluster and produced association rules and insights to each of them specifically, which provided us a holistic view from different perspectives.

In the first approach, we applied the *apriori* algorithm with a minimum support of 0.05, which was the most we could decrease the support before running into memory allocation errors, to produce frequent item sets. From this we produced association rules with a minimum confidence of 30%. This generated 217 association rules.

In order to generate more rules, we decomposed the dataset and divided it in days of the week, which allowed us to decrease the minimum support to 0,025 and produce more 1268 unique rules that were added to the original 217, giving us a total of 1485 association rules for all data set.

In the second approach we segmented the data set into 4 clusters and applied the *apriori* algorithm to each one of them to create specific item sets to each one of them and finally applied the *association\_rules* to each of this frequent item sets and generated rules to each cluster.

### 3.2.5. Metrics Used

The metrics used to evaluate and define our results were:

Support: *Support (A⇒B) = P(A ∪ B)* which is relative to the number of transactions that include items in the {A} and {B} as parts of the rule as a percentage of the total number of transactions

Confidence: C*onfidence (A⇒B) = P(B|A*) which refers to the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}.

Although *Apriori* Algorithmn let us build a lot of rules for each combination of item sets, most of the times, those rules aren’t all that relevant and might be redudant. To tackle this weakness, a correlation measure can be used to augment the support–confidence framework for association rules. Therefore, Lift was chosen to help identify some relationships between the rules and the respective products.

Lift: *Lift (A, B) = P(A ∪ B) / P(A)P(B),* where the occurrence of itemset A is independent of the occurrence of itemset B if P(A ∪ B) = P(A)P(B); otherwise, itemsets A and B are dependent and correlated as events. The most commom values that represent negative and positive relationship are less than 1 and greater than 1, respectively. In other words, a lift lower than 1 would indicate that those item sets would be substitutes and greater than 1 would lead to a complementary nature between both of them.

To prevent formulations of rules biased relative to critical products, it was decided to implement other metrics that are known for being null-invariant (if its value is free from the influence of null-transactions). It was chosen Kulczynski measure in conjunction with the imbalance ratio[pagina 271]. The added measure substantially reduces the number of rules generated and leads to the discovery of more meaningful rules for critical products that are usually one of the main sources of revenue to Instacart.

Kulczynski: *Kulc (A, B) = (P(A|B) + P(B|A)) / 2,* it is the average of two conditional probabilities: the probability of itemset B given itemset A, and the probability of itemset A given itemset B.

Imbalanced Ratio (IR): *IR(A, B) =|sup(A) − sup(B)| / (sup(A) + sup(B) − sup(A ∪ B)),* If the two directional implications between A and B are the same, then IR(A, B) will be zero. Otherwise, the larger the difference between the two, the larger the imbalance *ratio.*

(O que está a amarelo pode ser útil para descrever as metricas acima mas depois podemos apagar?)

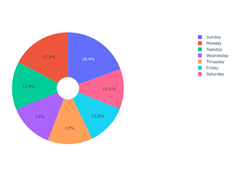
# Results Evaluation

## 4.1. Main types of customer behaviour

**Consumer Behaviour**

We analysed the behaviour of our customers from two different perspectives; first in a broad way analysing what are the general patterns of consuming and secondly, we decided to group our customers in clusters using the number of items per order and the days passed since the last order to segment our clients.

**General behaviour**



The number of orders is relatively evenly distributed throughout all the days of the week, however we can identify Sunday (19,4%) as the day with more affluence and Friday (11,8%) and Saturday (11,6%) as the days with less costumers.

Regarding the time of the day orders are placed, we notice a pick of orders between 1a.m.-2a.m., a constant flow during the morning hours and another big pick after 4 p.m. until 6 p.m.

On average customers purchase 7,27 products per order and take 13,08 days between them.

When looking at products the top five are: fresh fruits appear in 55,6% of all orders, fresh vegetables in 44,4%, packaged vegetable fruits in 36,5%, yogurt in 26,4% and milk in 24,3%. While the top five departments are: produce appearing 74,9% of all orders, dairy eggs in 67,7%, beverages in 45,7%, snack in 43,4% and frozen in 36,6%. (querem meter a lista completa?)

**Segmented behaviour**

To understand better our clients and considering that our store is online and assigns personal shoppers to each customer, we decided to segment each order into clusters to better understand the purpose of the client when they place a specific order. We got 4 distinct clusters:

Cluster 0

This cluster aggregates 43018 orders. It is the one with the highest number of items per order at an average of 13,7 items/order. On average it has been 8,96 days since the last order, which in comparison is higher than cluster 1 but much lower than cluster 2. [se calhar dizer logo 9 dias?]

Useful Insights: This cluster represents orders that although sparser in time, have a very high variety of products, indicating the possibility of being made by clients doing the “weekly shopping” filling their pantry and following a grocery shop list.

Top five bought products: fresh fruits appear in 80,5% of the orders, fresh vegetables in 70,5%, packaged vegetable fruits in 62,3%, yogurt in 48,2% and packaged cheese in 47,1%.

Cluster 1

This cluster aggregates 105925 orders making it the biggest one. It has the lowest number of items per order at an average of 4,9 items/order. It also has the lowest number of days since last order since on average has past 6,2 days.

Useful Insights: Having both the lowest number of items and the lowest number of days since last order, this behaviour could indicate less weighted customers that are buying groceries for a specific event and not following a grocery list.

Top five bought products: fresh fruits appear in 48,1% of the orders, fresh vegetables in 34,8%, packaged vegetables fruits in 27,4%, milk in 19% and yogurt in 18,6%.

Cluster 2

This cluster aggregates 38803 orders. On average each order has 6,6 items, which in comparison is slightly more than cluster 1, but less than half of cluster 0. It has the highest number of days since last order by a huge margin at an average of 26,4 days.

Useful Insights: This cluster has a huge number of days since last order and a small number of items per order, which may indicate sporadic customers that come either for a very specific reason or specific products that are either more expensive or don’t need to be buy so frequently.

Top five bought products: fresh fruits appear in 49,2% of the orders, fresh vegetables in 41,5%, packaged vegetable fruits in 33,1%, yogurt in 23,3% and packaged cheese in 21,1%.[Dado que o top5 tem sempre os mesmos itens, se calhar não dizer e mostrar as association rules em plot?]

Cluster 3

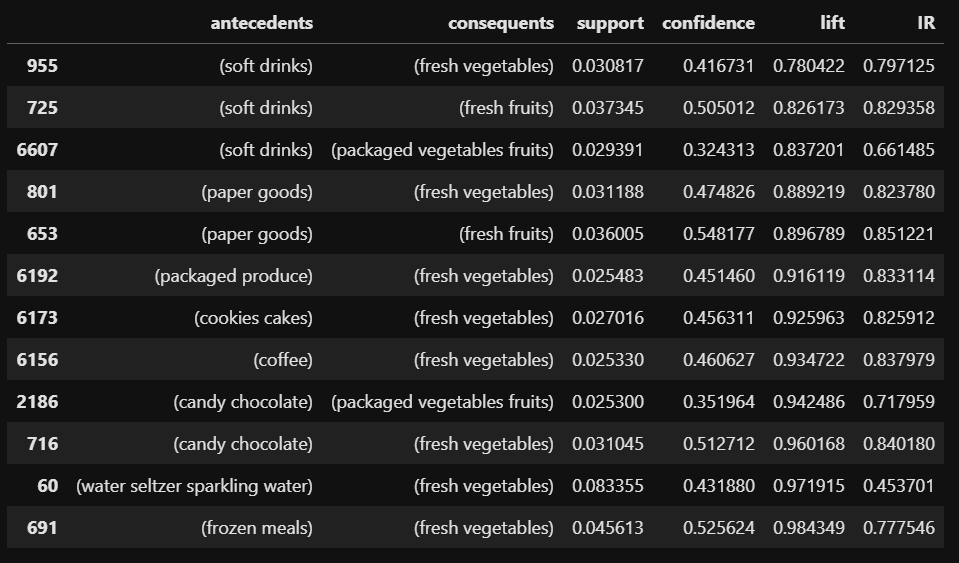
This last cluster aggregates 124342. It represents the newcomers, (eu depois acabo).

## **4.2. Type of products that should have an extended amount of product offerings**

## 4.3. Substitute type of products

We considered substitute products when an association rule generated betwen them has a lift lower than 1, meaning that usually when one person buys one of them it os less likelly it will buy the other.

Therefore, for the general data set we found the folowing substitute products:

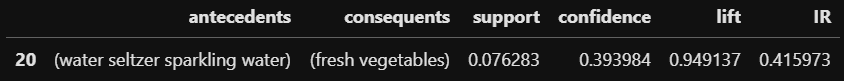


For cluster 0 we found the folowing substitute products: no código temos de tirar antecedentes ou consequentes que tenham mais do que 1 item



For cluster 1 we didn’t found any substitute products.

Finally for cluster 2 we found the following substitute products:



## 4.4. Complementary items

# DEPLOYMENT AND MAINTENANCE PLANS

## 5.1. Deployment

The main aim of this model is to give insights to higher management helping them with decision making regarding the products in the portfolio and display of them, which can be done with a report or a simple dataset containing all the association rules found as well as the metrics used to characterize them.

However, the association rules found can also be deployed directly to the customer as a recommender system, where products are recommended to the client according to strong association rules and where the product(s) the client chose work as an antecedent.

## 5.2. Maintenace

It’s important to keep the model updated, generating new association rules regularly and using the most recent data, having special attention about changes in:

* Association rules, which can indicate a change in consumer behaviour, such has new complementary or substitute products.
* Changes in the Imbalanced Ratio (IR) of strong association rules, specially in substitute products, which can indicate lack of quality, variety or any other problem about one of our products that incetivates na abnormal consumption of the other.
* A decrease in the confidence of complementary rules also may indicate that either the products are not complementary anymore, or one of them is lacking quality or other related problem.

# CONCLUSIONS

## 6.1. Considerations for model improvement