

**BUSINESS CASES WITH DATA SCIENCE**

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

Data4Business Consulting

**Business Case #4 – Market basket analysis**

Débora Santos, number: 20200748

Pedro Henrique Medeiros, number: 20200742

Rebeca Pinheiro, number: 20201096



April, 2021

INDEX

[1. INTRODUCTION 1](#_Toc65425090)

[2. BUSINESS UNDERSTANDING 1](#_Toc65425091)

[2.1. Background 1](#_Toc65425092)

[2.2. Business Objectives 2](#_Toc65425093)

[2.3. Business Success criteria 2](#_Toc65425094)

[2.4. Situation assessment 2](#_Toc65425095)

[2.4.1. Inventory of resources 2](#_Toc65425096)

[2.4.2. Requirements, assumptions and constraints 3](#_Toc65425097)

[2.4.3. Risks and contingencies 3](#_Toc65425098)

[2.5. Determine Data Mining goals 4](#_Toc65425099)

[2.6. Project Plan 4](#_Toc65425100)

[3. ASSOCIATION RULES ANALYSIS 5](#_Toc65425101)

[3.1. Data understanding and preparation 5](#_Toc65425102)

[3.2 Evaluation. 5](#_Toc65425103)

[4. RESULTS EVALUATION 9](#_Toc65425106)

[5. DEPLOYMENT AND MAINTENANCE PLANS 9](#_Toc65425107)

[6. CONCLUSIONS 9](#_Toc65425108)

[6.1. Considerations for model improvement 9](#_Toc65425109)

[7. REFERENCES 9](#_Toc65425110)

[8. APPENDIX (OPTIONAL) 10](#_Toc65425111)

# 

# 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 C.

• Jupyter Notebook with the code of the entire process.

All files can be accessed in Github:

*https://github.com/Debs86/Business\_Cases\_Projects/tree/main/BC2*

We are excited to be a part of this challenge.

# BUSINESS UNDERSTANDING

## 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.

## 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?

## Business Success criteria

The expected outcome will be well defined customers’ preferences based on the relationship founded in the products according to the purchases. With this, is possible to have an accurate overview about the whole business, making assertive combinations and providing a customized product portfolio strategy.

Another expected outcome of this report is giving visualizations that enables to give business insights and business applications for the findings.

The success of the proposed task will be evaluated by Instacart’s district manager and, if needed, we will go back to the model until we get an outcome that matches with the board’s expectation.

## Situation assessment

### 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.*

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 XXX 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.

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 XXX and XXX for some important visualizations. We also used Plotly and Dash packages to produce visualizations and a final application.

### Requirements, assumptions and constraints

The completion date of this 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.

COMPLETAR

### Risks and contingencies (mudar tudo)

Table 2.1 identifies a list of risks and contingency proposed.

|  |  |
| --- | --- |
| **Risk** | **Contingency** |
|  |  |
|  |  |
|  |  |

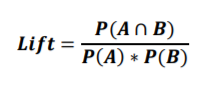
Table 2.1 - Risks and contingency.

#### Terminology

***Business glossary :***

* Department: Contains a generalized grouping of the products available.
* Order:

***Data mining glossary***

* Market basket analysis: It is a technique that consists in apply some algorithms that will find associations between items making possible to analyze customer behaviour. The goal is identifying the strength of association by the pair of purchased items analysing if there are patterns of co-occurrence.
* Association rules: Based on “If-Then” statements, association rules show probability relationship between item in a dataset. The rules also reveal the frequency of a itemset occurs in a transaction. It is a data mining approach for exploring and interpret large transactional dataset in order to find unique patterns and rules.
* Transaction: The term transaction can have different meanings according to the context applied. Considering business, transaction is defined as an event which involves goods interchange, money or services between two or more parties that involves physical exchange of value as sales, payments and purchases.
* 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.
* Causality: In general terms, causality is the center of any human decision making once provide basis for choosing to some action that will lead for a desire result. Considering associations rules, three types of causality are considered according to
  + Coincidental: Two things happen to describe the same object and have no determinative relationship between them;
  + Functional: There is a generative relationship;
  + Causual: One thing causes another thing to happen.
* Lift: Provides information if the association exist or not or if it is positive or negative.The support value can be obtained once the existence association value given by Lift. If Lift is greater than one, it suggests that two items appear more together than separated, having positive aspect. If Lift is smaller than one, means that two items appear less often together, having negative aspects. For last, if Lift value is 1, means that has almost effect between then (independent). For business analysis, it is important to look for rules with Lift >1.
* Support: Percentage of transactions that contain item or itemset. The rule that has low support not be interesting for business, once implies that the fact occurs by chance being not profitable. Support value is mostly used to eliminate rules with low support.
* Confidence: Percentage that shows how many times the rule head occurs in all groups that contains the rule body. In other words, reveal how reliable a specific rule is. In example, For given a rule X 🡪 Y, the higher the confidence more likely is for Y be in transactions that contains X.
* Complementary and substitute:

## Determine Data Mining goals

The data mining goals states project objectives in technical terms:

1. Identifying customer’s types according to their willingness to purchase products considering purchases combinations into products.

*Success criteria*:

1. Analyzing products combinations and categorizing them as substitute or complementary.

*Success criteria*:

1. Provide visualization with potential products combinations.

*Success criteria*:

1. Develop application which provides products combinations that is profitable for the business

*Success criteria*:

## Project Plan

Figure 2.2 - 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 state as the quality of the associations rules process will be directly connected with the quality of the input data. For this reason, we identify the Modelling stage dependent of the Data preparation stage. During the project, we must go and back between Data preparation and Modelling many times, repeat this iteratively until we get the desired outcome.

For the Modelling stage we aim to build an unsupervised model (market basket) using associations rules algorithm. Due to the timescales we opted for using this algorithm making adjustments in order to build a model as it is fast and efficient in terms of computational cost, simple to implement and the interpretation of the visualizations which shows the products combinations results is easy and straightforward. The associations rules quality evaluation will be made using (COMPLETAR)…..

# ASSOCIATION RULES ANALYSIS

In this section we go through the process of understanding and preparing the data for modelling, the modelling itself and algorithms used and, finally, the results evaluation.

## Data understanding and preparation

Instacart’s IT team provided 4 datasets *(Departments, Products, Order products, Orders)* that they considered important to complete the task designed. After analyzing each separately and not finding any relevant inconsistence the group decided to keep all data and not making any change in order to maintain as much information as possible.

Association rules algorithm can be applied in a binary data. Regarding data preparation, it was needed to create pivot table before applying apriori algorithm. Another action need in order to build a dashboard which contains the association rules analysis was build functions that made possible to create visualizations that will compose the application. To build those interactive charts, the package used was Plotly. For more details about both preparations regarding associations rules code and dashboard, the code is available in the Github repository.

## Modeling and evaluation

The first stage was applying apriori algorithm in order to discover the frequent items purchased. We started the process defining the support = 0.03. The main reason for setting this value was due the computational cost for that operation, once trying values below the adopted the code wasn´t able to run considering the dataset size. The maximum frequent items considered was 2, giving a total of (….) products.

Following the process, in order to have a better understanding about the consequent item products, for each antecedent in a transaction the consequent item has a confidence equal 0.5. It is important to mention that the consequent item is composed for only one product while the antecedent could have more than one product considered.

According to the literature, for business matters, the lift to be considered in order to increase sales should be above one, suggesting that the relationship between products were positive. In the case, the lift value stipulated was equal or above 1.5. Therefore, it was noticed that for some products relationships with high lift value, the value for consequent was not so high as expected which implies that(….).

Last, in order to get the complementary and substitute products for each category different metrics were considered. Regarding the complementary products the metric considered was both high lift and confidence, given a total of 33 complementary products. Regarding the substitute products, the metric considered was lift below one, given a total of 1550 substitute products.

# REFERENCES

https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL\_rstat/source/marketbasket49.htm

<https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining>

<https://www.geeksforgeeks.org/association-rule/>

<https://www.sciencedirect.com/topics/computer-science/association-rules>

<https://smallbusiness.chron.com/business-transaction-definition-examples-25244.html>

<https://www.accountingverse.com/dictionary/b/business-transaction.html>

<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.333.1386&rep=rep1&type=pdf>

<https://www.ibm.com/docs/en/db2/9.7?topic=SSEPGG_9.7.0/com.ibm.im.model.doc/c_confidence_in_an_association_rule.html>

<https://www-users.cs.umn.edu/~kumar001/dmbook/ch5_association_analysis.pdf>