

**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 J

Francisco Hermenegildo, number: 20200737

Gil Gonçalves, number: 20201066

Ikram Bouziri, number: 20200753

April 2021

INDEX

[1. INTRODUCTION 1](#_Toc66527933)

[2. BUSINESS UNDERSTANDING 2](#_Toc66527934)

[2.1. Background 2](#_Toc66527935)

[2.2. Business Objectives 2](#_Toc66527936)

[2.3. Business Success criteria 2](#_Toc66527937)

[2.4. Assess Situation 2](#_Toc66527938)

[2.5. Determine Data Mining goals 2](#_Toc66527939)

[3. PREDICTIVE ANALYTICS PROCESS 3](#_Toc66527940)

[3.1. Data understanding 3](#_Toc66527941)

[3.2. Data preparation 4](#_Toc66527942)

[3.3. Modeling: 5](#_Toc66527943)

[3.3.1. Select Modeling technique 5](#_Toc66527944)

[3.3.2. Build Model 6](#_Toc66527945)

[3.3.3. Model Assessment 6](#_Toc66527946)

[4. RESULTS EVALUATION 7](#_Toc66527947)

[5. DEPLOYMENT AND MAINTENANCE PLANS 9](#_Toc66527948)

[5.1. Plan Deployment 9](#_Toc66527949)

[5.2. Plan monitoring and maintenance 9](#_Toc66527950)

[6. CONCLUSIONS 10](#_Toc66527951)

# INTRODUCTION

Instacart is an American company that provides a grocery delivery and pick-up service via website or mobile app in the United States and Canada. Within their platform, users are assigned each time to a personal shopper with whom they can speak when selecting items from different type of products. The personal shopper then reviews the order and delivers it on the same day.

This project aims to provide the managers with the maximum knowledge possible about the preferences of the customers and the relationship between the products to enable the manager of this company to make better decisions regarding the platform’s layout and develop marketing strategies relevant to the users’ behavior.

Being the first time either element of the group is facing this specific challenge, we expect many hours of research to cement the knowledge required to perform the tasks and interpret the results obtained.

# BUSINESS UNDERSTANDING

## Background

## Instacart is an American company that provides a grocery delivery and pick-up service via a website or mobile app in the United States and Canada. While the user selects items from a wide portfolio of products through the Instacart app, he has the option to speak with the personal shopper assigned to him in order to better manage the shopping experience. Then, the personal shoppers review the order and do the in-store shopping and delivery on the same day.

As understanding the purchasing patterns of consumers is key for retail companies, Instacart is currently using its available 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 the shopping. However, the limited analytical capabilities of the company do not allow it to fully extract useful knowledge from the data.

## Business Objectives

In this project, we seek to provide a complete overview of Instacart’s business by taking full advantage of its data to identify the relationships between the different products and provide the company a view of the behaviors of its customers. The business objective of this company is to use the provided knowledge to optimize the layout of the application, develop relevant promotions and different marketing strategies for its customers.

## Business Success criteria

At the end of this project, the management of the company will have useful information regarding the purchasing patterns of the customers and the different relationships between the products thus, be better equipped to be able to customize the application layout for each user and develop relevant marketing strategies.

## Assess Situation

The company provided a dataset composed of a relational set of 4 CSV files describing customers' orders over time: ‘department.csv’; ‘product.csv’; ‘orders.csv’ and ‘order\_products.csv’. The dataset contains a sample of 200.000 grocery orders from more than 100.000 Instacart users and a total of 2.000.000 item transactions. Instacart required this task to be completed and presented to the management team by the 19th of April 2021.

## Determine Data Mining goals

An unsupervised learning technique called Market Basket Analysis will be used in this project to inform the company about relationships across different products to achieve its business objectives.The application of this technique will allow the group to reach the project’s goals and answer the following questions:

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

# PREDICTIVE ANALYTICS PROCESS

## Data understanding

**(Data Collection)** All the necessary data for the development of this project was provided by Instacart’s management team, in the form of a relational set of 4 CSV files describing a sample of 200.000 grocery orders from more than 100.000 Instacart users with a total of 2.000.000 item transactions. As complementary data, it was also provided the respective metadata and a PDF file with the business description and requirements for the project. The data was uploaded to a Jupyter notebook, which allowed all the necessary operations to discover the data and develop the association rules.

**(Data Description)**

The csv files are described as following:

|  |  |  |
| --- | --- | --- |
| File Name | Number of Rows | Column Names |
| Departments | 21 | Department\_id  Department |
| Products | 134 | Product\_id  Department\_id  Product\_name |
| Orders | 200000 | Order\_id  User\_id  Order\_number  Oder\_dow  Order\_hou\_of\_day  Days\_since\_prior\_order |
| Order\_Products | 2019501 | Order\_id  Product\_id  Add\_to\_cart\_order  Reordered |

Table 1 – Description of the data

**(Data Exploration)** Pandas Profiling was useful to explore the data:

* Duplicated rows: the amount of the exact same order
* Distribution of the variables
* Check for possible correlations

**(Data Quality)** The exploration of the provided data demonstrated its overall good quality, with no missing values identified or other types of errors.

## Data preparation

**(Data Cleaning)** The good quality of the CSV files eliminated the time spent on the data cleaning process. The only variable with missing values was “days\_since\_prior\_order” but the group decided to not do anything about it since the first purchase of the customers means there were no prior purchases, therefore, a missing value.

**(Data transformation and engineering)** In this project, any data transformation or engineering was needed since association rules were the main focus and the provided data was sufficient to build the rules and extract useful information for the company’s managers.

**(Data Formatting)** To apply the association rules methods, the data was transformed into a pivot table with the order\_id as an index and the product names in the columns to create a table in which each row represents the items present in each order. The same procedure was applied when working in the departments.

**(Data Selection)** Since the group found no real use for the variable “add\_to\_cart\_order” we decided to drop it. Also, after an initial analysis to assess the number of reorders each product registered, that variable was also dropped.

## Modeling:

### Select Modeling technique

The data mining technique used in this project is the association rules which is a tool used by retailers to analyze their basket or transactional data. This technique allows to identify strong rules between items and look for combinations of items that occur together frequently using different metrics.

The metrics used are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Metric |  | Formula | Description |
| Support |  |  | The output is the frequency of the Item regarding the total of transactions |
| Confidence |  |  | The percentage of a consequent appearing given that the antecedent occurred |
| Lift |  |  | Depending on if the value is greater or smaller than 1 gives a sense of association or dissociation respectively |
| Leverage |  | Support(X&Y) – Support(X)Support(Y) | Depending on if the value is positive or negative gives a sense of whether those products are more common to be bought together or independently |

To answer some of the proposed questions, certain rules were created aiming to trim down the overall combination of different antecedents and consequents.

To create those filters firstly the apriori algorithm was introduced to define a specific minimum threshold of support and afterward, the association rules algorithm was applied to further trim down the results in order to achieve a good enough representation of the desired complementary or substitute itemsets.

The table below contemplates the filters applied to find the relation between the products.

|  |  |  |
| --- | --- | --- |
|  | **Complementary** | **Substitute** |
| **Support** | Greater than 0,05 | Greater than 0,02 |
| **Confidence** | Greater than 0,5 | Not applicable |
| **Lift** | Greater than 1 | Smaller than 0,95 |

Table 2. Filters applied to dataset

### Build Model

Since we are dealing with an unsupervised learning Market Basket Analysis problem that requires an association via rules, the model used were Apriori and Association\_rules, both from the mlxtend library.

### Model Assessment

Since we are dealing with a Market Basket Analysis the group did not find a tool to analyze the results and instead approached through a more subjective analysis checking if the results made sense, which proved to be a good approach since some of the complementary and substitute relationships identified by the rules did not represent the respective relationship.

To avoid falling into the Beers and Diapers impossible correlation a careful analysis was made and from there were selected the complementary and substitute items present in chapter 4, Results Evaluation.

# RESULTS EVALUATION

Throughout the analysis of the datasets, and since the early stages of data exploration, numerous conclusions were taken:

|  |  |
| --- | --- |
| Number of unique customers | 105.273 |
| The average number of orders per customer | 1,9 |
| The average number of items per order | 10,09 |
| The average number of unique items per order | 7,26 |

Table 3. Initial Findings

Some of other findings related to the customer behavior were the frequency, distribution and peridiocity of purchases from our customers:

|  |  |  |
| --- | --- | --- |
| **Number of days since last purchase** | **Number of orders per day of the week** | **Total orders per hour of the day** |
| Most orders were made with a peridiocity of 8 days | Most frequent on Sundays and Mondays | Most frequent from 9AM-5PM |
| 53% of orders | 34,7% of orders | 71,7% of orders |

Table 4. Additional Findings concerning Customer Behavior

Since the group proposed to answer the four initial questions present in the Data Mining Goals, the next step was analyzing the relationship between department support and the respective array of product offerings. The findings of the process:

|  |  |  |
| --- | --- | --- |
| **Department** | **Support** | **Number of products available** |
| Produce | 0,75 | 5 |
| Bakery | 0,27 | 5 |
| Beverages | 0,46 | 8 |

Table 5. Relation between department support and product offering

Next we aimed to find the complementary and substitute items.

The complementarity products were easy to spot since there are several itemsets with values of lift, leverage and confidence that pointed towards a relationship of such nature. The following products were deemed complementary by the group[[1]](#footnote-2):

|  |  |
| --- | --- |
| **Antecedents** | **Consequents** |
| Fresh Herbs | Fresh Fruits |
| Yogurt | Fresh Vegetables |
| Packaged Cheese | Packed Vegetables Fruits |
| Milk |  |
| Soy Lactose Free |  |
| Eggs |  |
| Frozen Produce |  |

Table 6 - Complementary items

On the contrary, regarding the substitute items, they were more challeging to define. Starting from the minimum support threshold on the first apriori algorithm run that had to be lowered from a default value of 0,05 to 0,02 to allow some itemsets to come up with results that sustain such nature.

After the initial support filtering, the group decided to focus mostly on the itemsets whose relationship revealed values of lift below 1 and negative values of leverage. In the midst of the resulting itemsets, there were some that through a subjective analysis were discarded since there was no substitute relationship between either items. In order to not incur on a impossible correlation problem as mentioned above, a careful analysis was made and the resulting substitute products are the following:

|  |  |
| --- | --- |
| Soft Drinks | Fresh Fruits |
| Yogurt |
| Milk |
| Fresh Vegetables | Packaged Produce |

Table 7.Substitute Itemsets

# DEPLOYMENT AND MAINTENANCE PLANS

With the findings identified in this Market Basket analysis project, Instacart can, from now on, implement several strategies in order to improve the overall performance of the company.

## Plan Deployment

Our team suggests various measures that can be deployed easily and in a very short time. As a first measure, the company should adapt the layout of the online platform in such a way that engages the customer to buy the complementary products identified in this project, and also implement a direct to consumer marketing approach, by suggesting complementary items to those that are already present in the basket. On another hand, Instacart can improve the product offering of the departments “Produce”, “Beverages” and “Bakery”, since these already have a high demand compared to the remaining ones.

## Plan monitoring and maintenance

The implementation of this type of analysis, requires a constant monitoring of the results. All the new transactions made by Instacart, should be introduce into the “model”, since they can alter the results of the current product relationships and also generate new ones.

# CONCLUSIONS

This third business case was the first approach to the implementation process of a Market Basket Analysis in a real case scenario. With this project, it was possible to identify the power that associations rules mining can have in this types of business in order to, in a simple and fast way, customize product offerings, change online shop layout and ultimately potentiate custumers engagement and revenue.

In the end, we managed create severel rules that ultimately enabled the identification of various complementary and substitute items, which can be addressed in different ways by Instacart, maximizing profitability.

Overall, it was a captivating and enriching experience to once again work on a real-life case and in a different field of Data Science.

1. The relationship of this products are the same for all. Each one of the antecedents is complmentary to all of the consequents [↑](#footnote-ref-2)