Group Project

BUY&GO

DATA SCIENCE AND MACHINE LEARNING 2021

February 28, 2021

1 Introduction

Welcome to the BUY&GO company. This organization is a well-established company in Ireland, operating in the retail sector in three different regions: Cork, Killarney, and Kinsale, inside coworking areas. Presently they have around 10,000 registered customers and serve more than 120,000 consumers a year. They sell products from 5 major categories: Groceries, Stationery, Housekeeping, Wellness & Beauty and Animals Plants. These five categories can further be divided into Limited Edition and Basic collection. The Customers can order and acquire those products through 3 channel groups: Physical stores, yearly catalogs, and the companies' website. Globally, the company had stable revenues and a healthy bottom line in the past three years, but the profit growth perspectives for the next three years are fickle. A few strategic initiatives are being considered to invert the situation. One of those is a Marketing efficiency program to improve marketing activities, focusing on boosting the marketing campaigns' efficiency tremendously.

1.1 At the Marketing Department

The marketing department is under pressure to spend more wisely its annual budget. The CMO knows the importance of having a more quantitative approach to marketing decisions. The department requested a small team of 5 data scientists (your group) with a clear objective in mind: try to cluster the different types of customers that the company has to create more efficient campaigns. Desirably, these activities' success will prove the approach's value and convince the more skeptical within the company.

2 Objective of the project

The team's objective in this project is to identify actionable segments within the company's Customer base. These segments must be determined by looking at data available and through the usage of quantitative technics. A priori, two visions are considered essential – the customer value segmentation and the product usage segmentation. Nonetheless, other perspectives will be valued. This project's main output will be a report that identifies the main customer segments and a first draft of a marketing plan.

3 Datasets

You have access to two different sources: a file with sociodemographic data and other with firmographic data corresponding to 5000 customers.

The sociodemographic data contain the following attributes:

Attribute	Description
Card_ID	Unique identification of the customer card
Name	Customer's name
Year_Birth	Customer's Birth year
Education	Costumer's level of education
Marital_Status	Costumer's Marital status
Income	Yearly Income of costumer's household
Kidhome	Number of kids in household
Teenhome	Number of teenagers in household
Region	Store associated with the card registration
Country	Customer's country

The firmographic data contain the following attributes:

Attribute	Description
Card_ID	Unique identification of the customer card
Dt_Customer	Card's Registration
Recency	Number of days since last purchase
MntGroceries	Amount spent on groceries items
MntStationery_&_Books	Amount spent on stationery products and books
MntHouseKeeping	Amount spent on Housekeeping products
MntWellness_&_Beauty	Amount spent on wellness and beauty items
MntElectronics_&_Supplies	Amount spent on electronics and supplies
MntLimitedEdition	Amount spent on limited edition items (from the 5 previous departments)
NumDealsPurchases	Number of purchases made with discounts
NumWebPurchases	Number of purchases made through web
NumCatalogPurchases	Number of purchases made through monthly catalog
NumStorePurchases	Number of purchases made on Store
NumWebVisitsMonth	Average number of web visits a month to the company site
AcceptedCmp1	Flag indicating customer accepted offer in campaign 1
AcceptedCmp2	Flag indicating customer accepted offer in campaign 2
AcceptedCmp3	Flag indicating customer accepted offer in campaign 3
AcceptedCmp4	Flag indicating customer accepted offer in campaign 4
AcceptedCmp5	Flag indicating customer accepted offer in campaign 5
Complain	Flag indicating if customer has complained

4 Deliverables

1. A Jupiter notebook with all the needed code implemented to obtain the results presented in the report.

   The file naming format should be ”GroupXX_DSML2021_Notebook.ipynb)”.

2. A report structured similarly to the provided structure that describes the analytical processes and the conclusions obtained, with at most 6000 words (not including references, captions, tables and titles). Please check the document ”report_structure.pdf”.

   The file naming format should be ”GroupXX_DSML202021_Report.pdf”.

4.1 Notes

- We will evaluate all the topics mentioned based on the report - a well-structured and succinct report will have a big weight on the evaluation.

- The jupyter notebook will be analyzed only if some doubt arises during the report evaluation. If some steps were done in the Jupyter notebook but not described in the report, we will not evaluate those. As an example, imagine you check the outliers, and at the end of your project, you decide to keep them. In the report, you should mention how you check if you had outliers, what the steps were to remove them and why you decide to keep them at the end, among other insights that can be relevant. The jupyter notebook should be delivered with all the cells already run.

- The report and the code will pass through a process of plagiarism checking.

5 Evaluation Criteria

The following table quantifies the major evaluation criteria.

Criteria	Percentage	Maximum Grade (out of 20)
Report-quality	15%	3
Story-telling	5%	1
Exploration	10%	2
Pre-processing	15%	3
KMeans	15%	3
Description of Customer Segments	15%	3
Marketing Plan	10%	2
PCA	2.5%	0.5
KNN Imputer or other imputation method	2.5%	0.5
Other Clustering technique	5%	1
Creativity & Other Self-Study	5%	1
TOTAL	100%	20

A project that focus only on the techniques and methodologies approached during the practical classes will have at most 17 values. The remaining 3 values are possible to achieve if contributions based on self-study and creativity are applied, and clearly explained on the report.

This bullet-list provides some details about each aspect:

- **Report-quality:** Each report should follow the provided report structure and describe the steps and main insights along the process. Clarity, synthesis, objectiveness, and business-contextualization are very welcome;
- **Story-telling:** Your decisions and steps must be reasonably justified by the

previous findings (when this is possible and feasible), your hypothesis and findings must be related to the problem's business-context, etc..

- **Exploration:** Describe the studied population using statistical measures, business insights and visualizations representative of the major insights.

- **Pre-processing:** Includes all the needed steps to transform the raw data into the data prepared to cluster. Involves all the steps for cleaning, transform and reduce the dataset. It also involves the business-related transformations of the original input features and the explanation of those.

- **KMeans:** the reasoning behind k's selection, the existence of a comparative study with different ks, etc. Two perspectives are obligatory: Customer perspective and product usage perspective. More perspectives are optional and considered as points in "Creativity and other self-study". You should explain the feature selection for each perspective for the sake of each perspective segmentation.

- **Description of Customer Segments:** Each segment for each perspective should be explored (statistically and visually) and described, focusing on the main aspects that differentiate each one.

- **Marketing Plan:** You should provide a succinct but well-oriented marketing plan that will answer the main insights obtained during clustering.

- **PCA:** A theoretical explanation of the algorithm should be provided in the annex. This algorithm allow dimensionality reduction, and even if not used in the final solution, the application of PCA implies interpreting the results and the clear identification of the number of components to be used (and the process used to quantify the final number of components).

- **KNN Imputer or other Imputation methods:** A theoretical explanation of the algorithm should be provided in the annex.

- **Other Clustering technique:** A theoretical explanation of the algorithm should be provided in the annex. Involves the depth and the quality of the comparative analysis between the clustering solutions provided by the different algorithms, the existence of a short comparative study between different customers' profiles obtained by different methods, etc.;

- **Creativity and Other Self-Study:** If other algorithms not given during classes are applied, a theoretical explanation of the algorithm should be provided in the annex. This topic includes not only the application of different techniques but also aspects of creativity, such as the creation of other perspectives during clustering besides the obligatory ones.

All topics are evaluated through a comparison of the work provided by the different groups.