# Group Project

# Techscape — E-commerce

MACHINE LEARNING 2021/2022

October 28, 2021

# 1 Introduction

Welcome to TechScape. In the beginning of 2020, five Portuguese entrepreneurs built a startup to sell goods related to digital detox. Using an online store, TechScape sells products and services which will allow their customers to stay focused on the most important things and improve the balance with technology use in their lives, such as meditation kits, books, stress balls, ”dumb” phones, retreats, among others.

With the emergence of Covid-19 in March 2020, the company had some financial difficulties, and they were out of function in April 2020. But in May 2020, they restarted their activities, knowing that a digital detox is becoming even more critical in this period of time, where people need to unplug themselves and improve their quality of life.

Focused in increasing their sales, TechScape hired a team of data scientists to analyze the online behavior of their customers and to predict which customers have a high probability of buying their products depending on their online actions.

# 2 Objective of the project

Your goal is to build a predictive model that answers the question “Which customers are more likely to buy our products?” using the small quantity of data accessible from the customers data base that contains general information about the customers and their behaviour in the website from February 2020 till December 2020 (excluding April).

# 3 Datasets

You have access to two different datasets:

1. The training set should be used to build your machine learning models. In this set, you also have the ground truth associated with each user access, i.e., if the user buy at least one product (1) or not (0).

2. The test set should be used to see how well your model performs on unseen data. In this set you don't have access to the ground truth, and the goal of your team is to predict that value ("0" or "1") by using the model you created using the training set.

The available data contains the following attributes:

| Attribute | Description |
|---|---|
| Access_ID | Unique identification of the user access to the website |
| Date | Website visit date |
| AccountMng_Pages | Number of pages visited by the user about account management |
| AccountMng_Duration | Total amount of time (seconds) spent by the user on account management related pages |
| FAQ_Pages | Number of pages visited by the user about frequently asked questions, shipping information and company related pages |
| FAQ_Duration | Total amount of time (seconds) spent by the user on FAQ pages |
| Product_Pages | Number of pages visited by the user about products and services offered by the company |
| Product_Duration | Total amount in time (seconds) spent by the user on products and services related pages |
| GoogleAnalytics_BounceRate | Average bounce rate value of the pages visited by the user, provided by google analytics |
| GoogleAnalytics_ExitRate | Average exit rate value of the pages visited by the user, provided by google analytics |
| GoogleAnalytics_PageValue | Average page value of the pages visited by the user, provided by google analytics |
| OS | Operating System of the user |
| Browser | Browser used to access the webpage |
| Country | The country of the user |
| Type_of_Traffic | Traffic Source by which the user has accessed the website (e.g., email, banner, direct) |
| Type_of_Visitor | User type as ”New access”, ”Returner” or ”Other” |
| Buy | Class label indicating if the user finalized their actions in the website with a transaction<br>(Only available in train dataset) |

# 4 Deliverables

1. A Jupiter notebook with all the needed code implemented to obtain the results presented in the report and to obtain the performance obtained in Kaggle.
   The file naming format should be ”202122_Fall_AA_GroupXX_Notebook.ipynb”, where ”GroupXX” should be your group number.

2. A report that describes the analytical processes and the conclusions obtained, with at most 8 pages:

   - **Heading 1:** Arial, Size 12 pt, in bold
   - **Heading 2 (if needed):** Arial, Size 11 pt, in bold and italic
   - **Text:** Arial, Size 10 pt, line space of 1.5 points.
   - **Margins:** The default ones in word (Top, Bottom, Left and Right as 1”).

   All the figures and tables should be included in the Annexes (at the end of the document) and are not included on those 8 pages mentioned previously.
   The file naming format should be ”202122_Fall_AA_GroupXX_Report.pdf”, where ”GroupXX” should be your group number.

## 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.
- **For more information, please read the Kaggle competition rules carefully.**

# 5 Evaluation Criteria

The following table quantifies the major evaluation criteria.

| Criteria | Percentage | Maximum Grade |
|---|---:|---:|
| | | (out of 20) |
| Kaggle performance | 20% | 4 |
| Report-quality | 15% | 3 |
| Story-telling | 5% | 1 |
| Exploration | 10% | 2 |
| Pre-processing | 10% | 2 |
| Modelling | 17.5% | 3.5 |
| Performance Assessment | 7.5% | 1.5 |
| Other predictive models (not given during classes) | 5% | 1 |
| Creativity & Other Self-Study | 10% | 2 |
| 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:

- **Kaggle performace:** The performance obtained on Kaggle, on the submission selected (F1 Score).
- **Report-quality:** Each report should 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, meaningful 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 model. Involves all the steps for cleaning, transform and reduce the dataset. It also involves the creation of new variables (if any) from the original input features and the explanation of those.
- **Modelling:** the implementation of different predictive algorithms and the process of fine-tuning those models. The application of additional models not given during classes are optional and considered as points in "Other predictive models".
- **Performance Assessment:** The comparison of different models and their performance.
- **Other predictive models:** A theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages). Involves the depth and the quality of the comparative analysis provided by the different algorithms, the theoretical explanation of the algorithm itself and the justification of the chosen parameters;
- **Creativity and Other Self-Study:** If other techniques not given during practical classes are applied, a theoretical explanation of the algorithm / technique should be provided in the annex (not included in the 8 pages). This topic

includes not only the application of different techniques but also aspects of creativity, such as the the quality of visualizations, plots and others.

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