ONLINE SHOPPERS INTENTION

Rogério Rocha - 201805123

Tomé cunha - 201904710

SPECIFICATION OF THE WORK

- Nowadays, a large majority of businesses are supported or carried out online. Given this, it is mandatory that identifying potential buyers (much like in a real world scenario) needs to be done. And that is precisely the objective of this project.
- We will be developing a machine learning model based on supervised learning classification algorithms able to identify website visitors with a high likelihood to carry out an online transaction.
- To develop the ideal model we must experiment with different algorithms, as well as carry out an Exploratory Data Analysis.

SPECIFICATION OF THE WORK

- The dataset contains the following data:
  - **Administrative**, **Administrative Duration, Informational**, **Informational Duration**, **Product Related** and **Product Related Duration** – Continuous, these represente the number of diferente pages visited by the visitor in that session and total time spent in those pages
  - **Bounce Rate** – Percentage of visitors who enter the page and then leave (“bounce”) without triggering any other requests
  - **Exit Rate** – Percentage, calculated for all pageviews, and represents the ones that were last in the session
  - **Page Value** – Discrete, the average value for a web page that a user visited before completing na e-commerce transaction

SPECIFICATION OF THE WORK

- **Special Day** – Proximity of the site visiting time to a specific special day
- **Month** – Month value of the visit date
- **Operating System** – Operating system of the visitor
- **Browser** – Browser of the visitor
- **Region** – Geographic region from which the session was started
- **Traffic Type** – Traffic source by which the visitor has arrived to the site
- **Visitor Type** – Categorical (takes on Returning, New or Other values)
- **Weekend** – Boolean value indicating whether the date of the visit is weekend
- **Revenue** – Used as class label

RELATED WORK

- Code
  - Imbalanced Learn – over smapling and undersampling tools
  - Scikit Learn – machine learning algorithms
- Websites
  - https://machinelearningmastery.com/what-is-imbalanced-classification
  - https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imabalanced-data
- Articles
  - https://link.springer.com/article/10.1007/s00521-018-3523-0

DESCRIPTION OF THE TOOLS AND ALGORITHMS

- Tools
  - We will use Python as programming language, programming in a Jupyter Notebook environment. For machine learning algorithms, we will be using Scikit-Learn and Imbalanced Learn libraries, as well as Pandas to read and handle the data and Seaborn and Matplotlib to visualize it.
- Algorithms
  - Given the nature of our dataset we plan to use Oversampling techniques.
  - For the classification we plan on using Random Forest, Naïve Bayes and we are still unsure of the last algorithm.