Online Shoppers Intention

G29

Rogério Rocha up201805123
Tomé Cunha up201904710

Work specification

- 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 classificatio 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 carry out an Exploratory Data Analysis.

Work specification

- 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 la the session
    - **Page Value** - Discrete, the average value for a web page that a user visited before completing na e-commerce transaction

Work specification

- **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

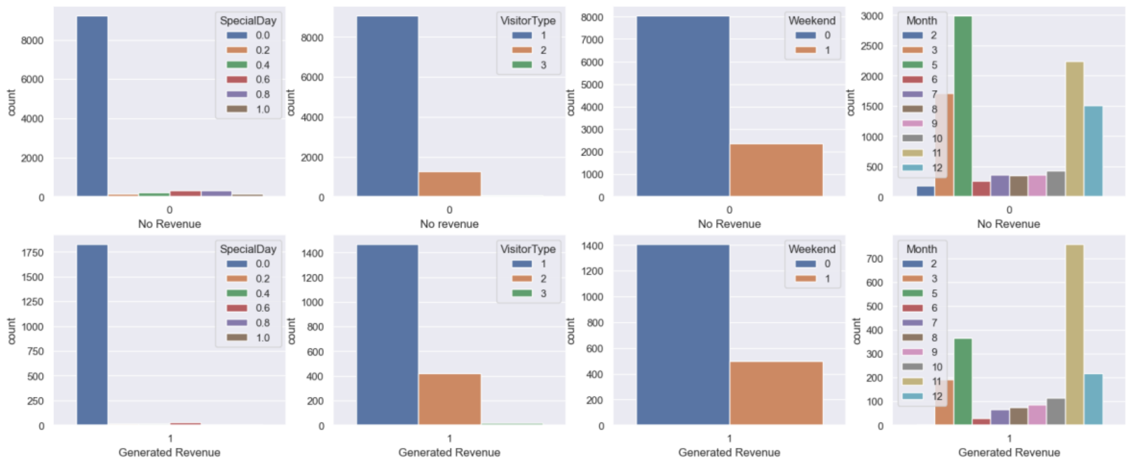
Tools and Algorithms

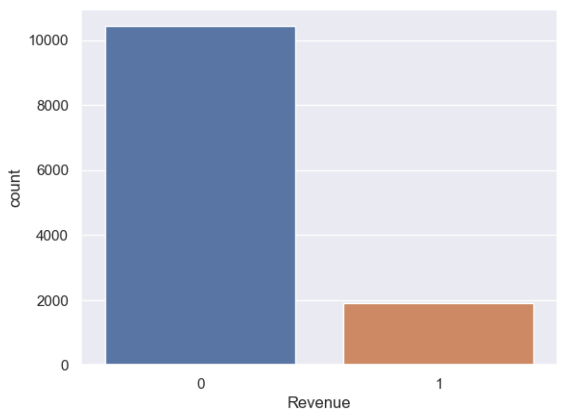
- **Tools:**
  - We have used Python as programming language, programming in a Jupyter Notebook environment. For machine learning algorithms, we have used Scikit-Learn and Imbalanced Learn libraries, as well as Pandas to read and handle the data and Seaborn and Matplotlib to visualize it.

- **Algorithms:**
  - Given the nature of our dataset we decided to use Oversampling techniques. More specifically an oversampling using cross validation (stratified 10 fold).
  - For the classification we have use the Decision Trees model, the Neural Networks model, the K-Nearest Neighbours model and the Naives Bayes model. This last one is not very effective on our dataset however we decided to compared it to the others as we already knew it would perform worse.

Data pre-processing

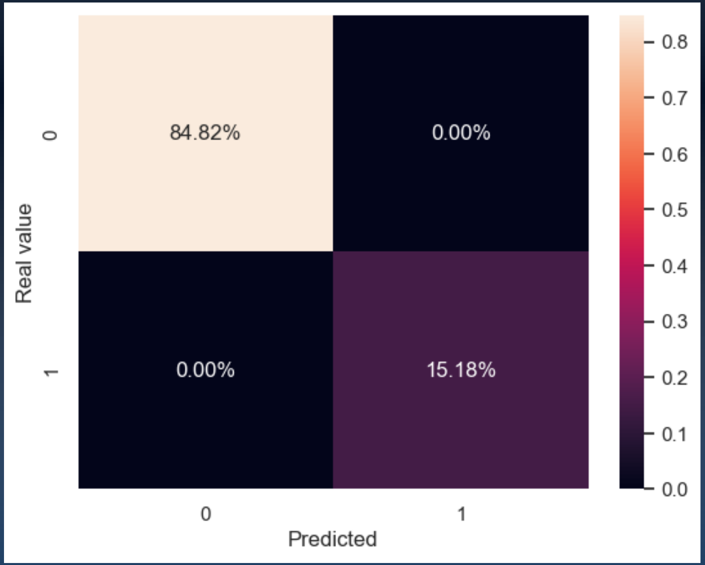
- Firstly we had to take a proper look at the data we were provided with, we cleaned up any non numerical values for our model's approach and we analyzed all the attributes for the correlation it had to the "Revenue" label.



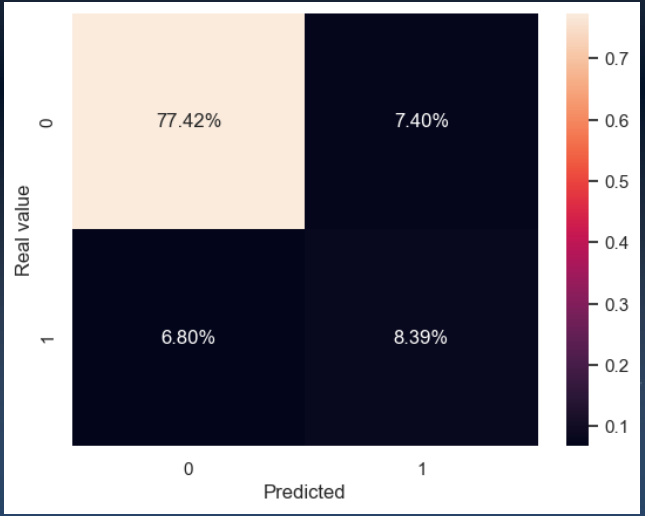


Decision Trees

Balanced dataset

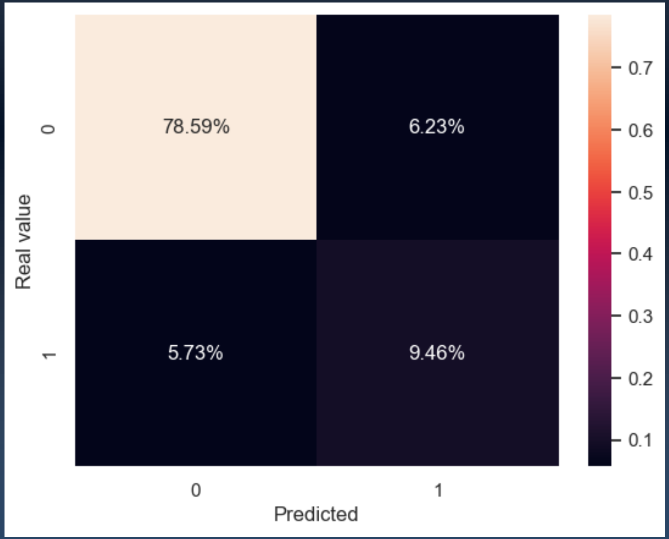


Imbalanced dataset

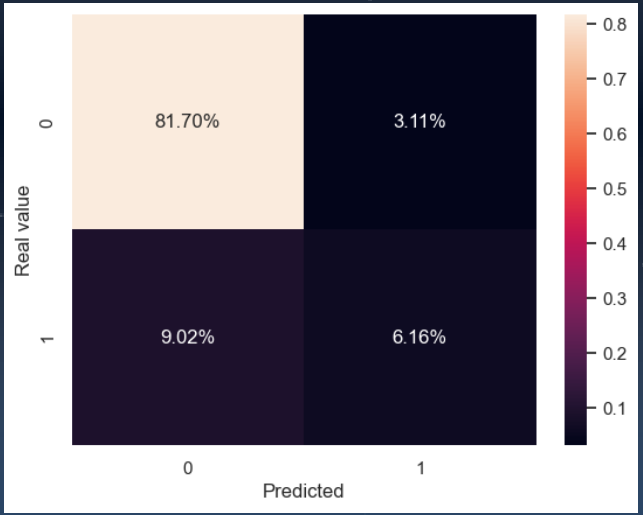


Neural Networks

Balanced dataset

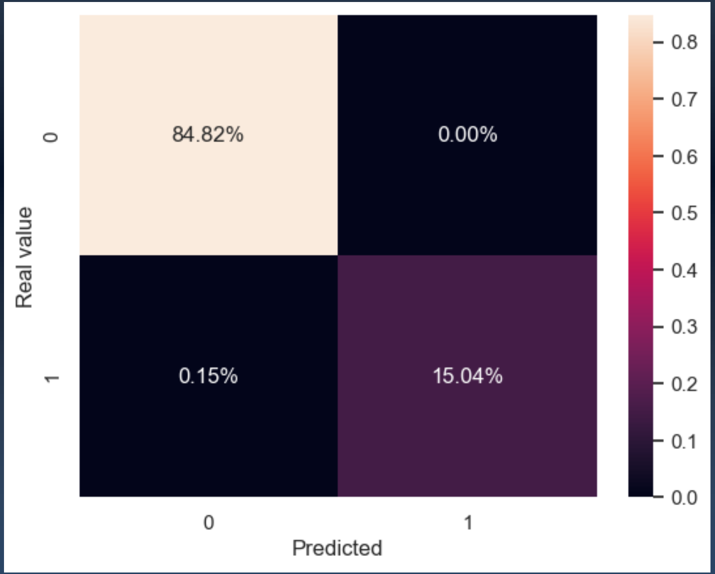


Imbalanced dataset

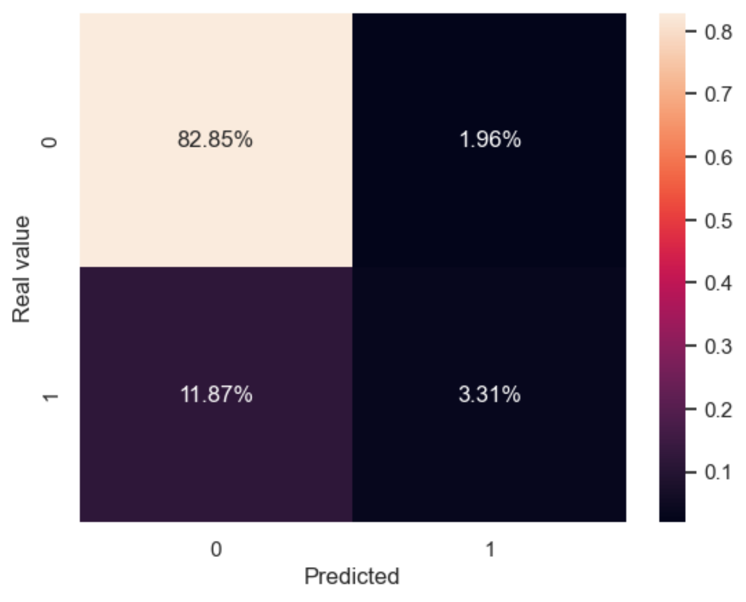


K-Nearest Neighbour

Balanced dataset

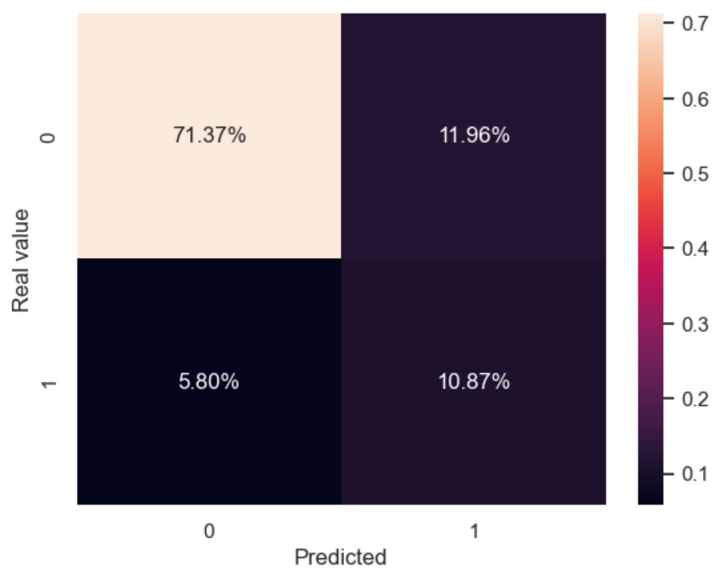


Imbalanced dataset

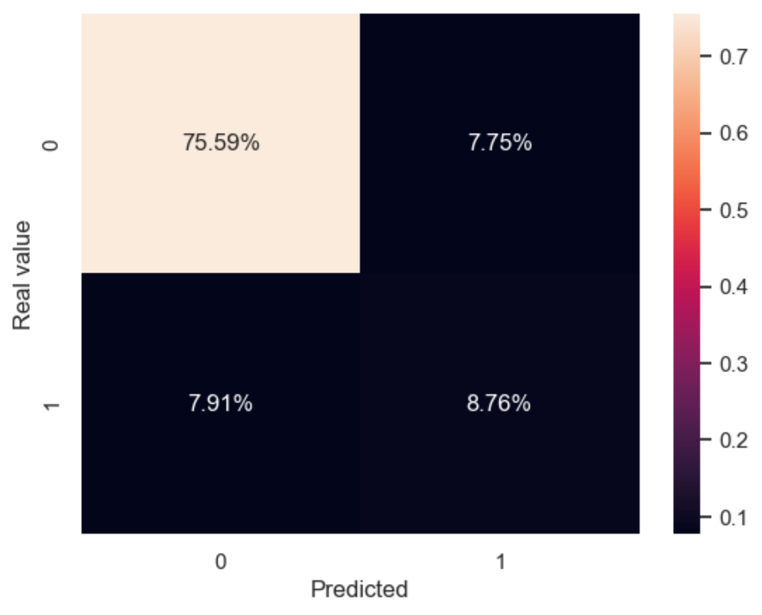


Naives Bayes

Balanced dataset



Imbalanced dataset



Results and Conclusion

- From all the models above used, we've determined that for the problem at hand the models that showed us the best results were the Decision Trees with a balanced dataset and the K-Nearest Neighbours with a balanced dataset aswell. Obviously, it is normal that a balanced dataset would produce better results as it eliminates any possible bias by the ML model.

- As for execution time, from the models used the least intensive was the Naives Bayes and the most intensive was Neural Networks.

- A closing note: the original dataset provided was somewhat small, ence some of the results may not be 100% accurate, however, it was big enough to satisfy and train our model and it produced good results.