Python for Data

A Quick Review of our Data

The Dataset : Online Shopper's Intention

Our dataset regroups data from users on an e-commerce website. The main goal of this dataset is to show the information on how much time users spent on the website, when, and most importantly whether they spent money.

Administrative int64	Administrative_Dur… floa…	Informational int64	Informational_Dura… float…
ProductRelated int64	ProductRelated_Dur… floa…	BounceRates float64	ExitRates float64
PageValues float64	SpecialDay float64	Month object	OperatingSystems int64
Browser int64	Region int64	TrafficType int64	VisitorType object
Weekend bool	Revenue bool		

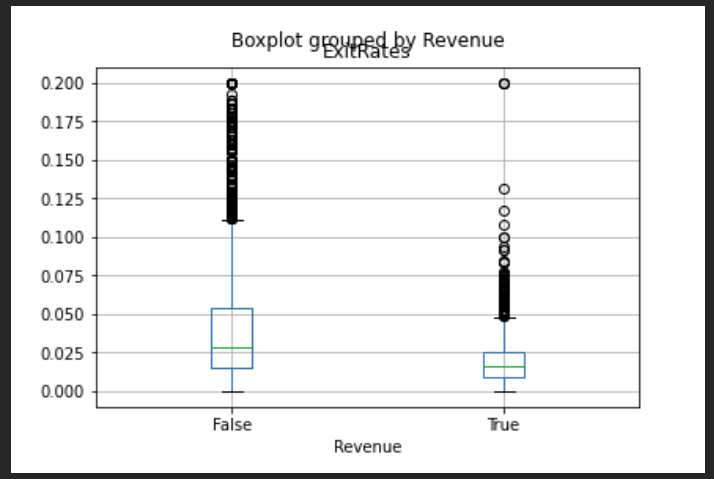
Here we can see the different variables

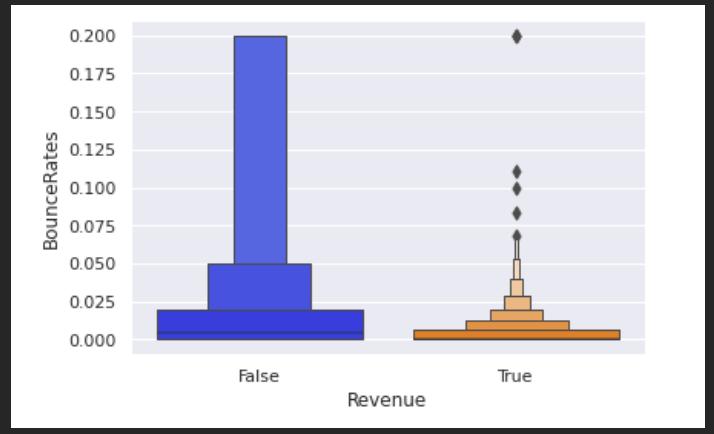
The important data : *Revenue*

By reviewing the dataset, we could tell that the variable *Revenue* is the most important one.

Indeed, making analysis on why users spent money might be the most relevant analysis.

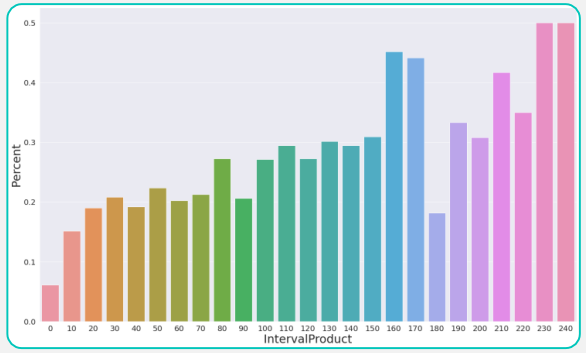
So, we started analyzing :

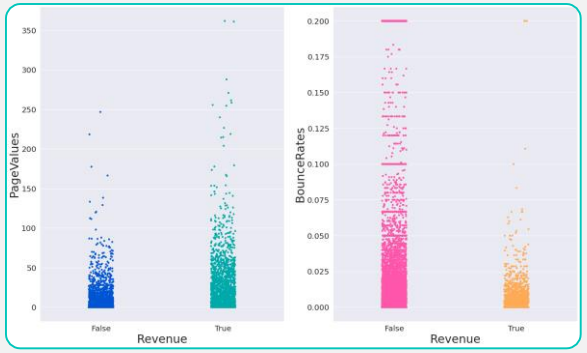




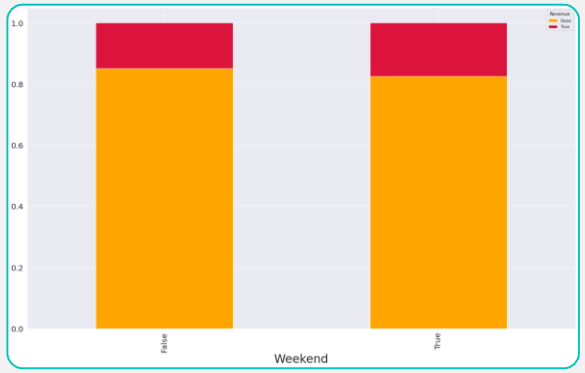
The important data : *Revenue*

By looking at the variables, we decided to work on the *Revenue* variable. We made a lot of analysis and plots to determine the importance of each variable on *Revenue*.





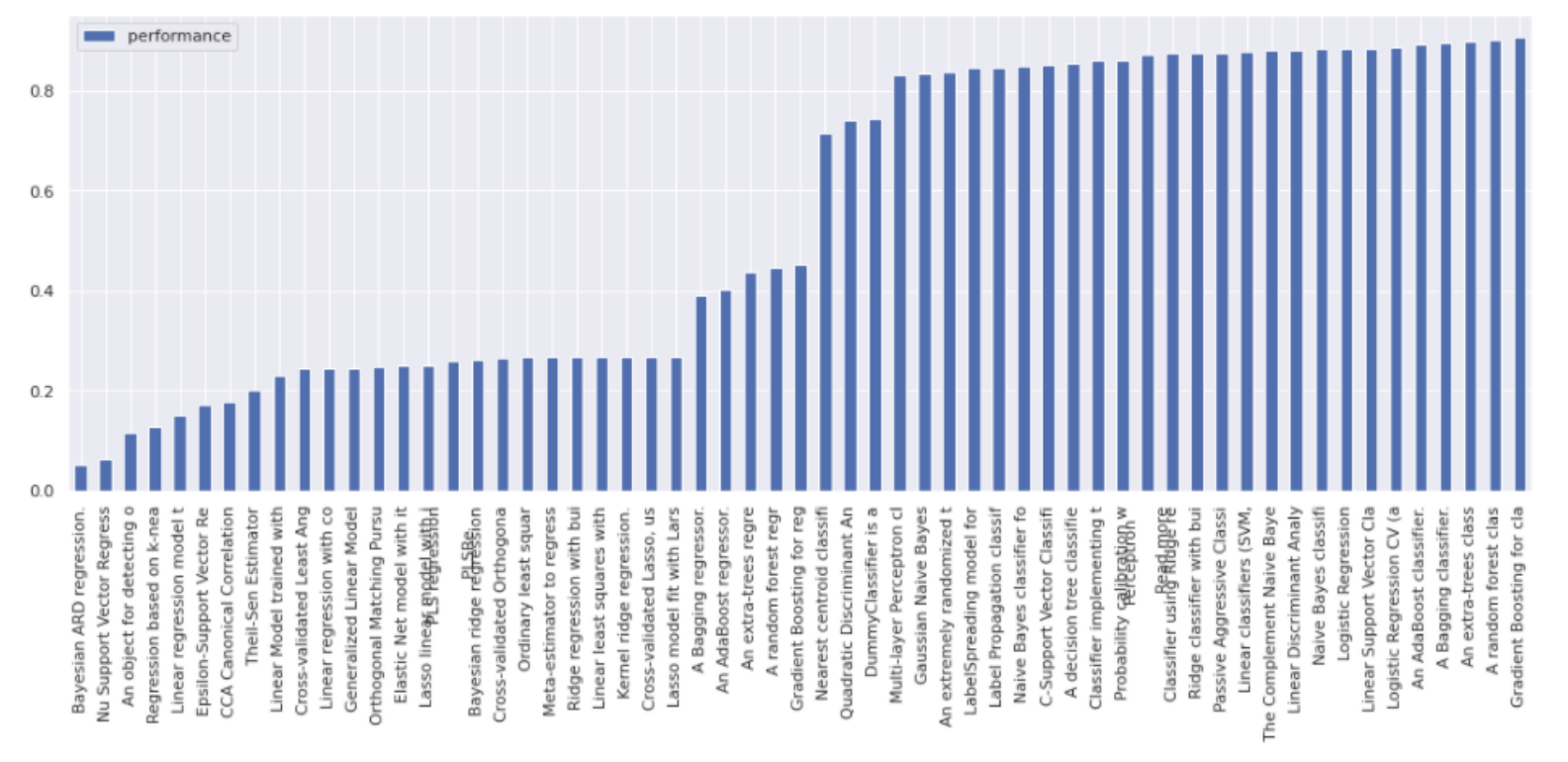
Before making predictions, we had to make sure that there weren't any superfluous data.



Finally, after this step, we ran into the predictions.

Predictions

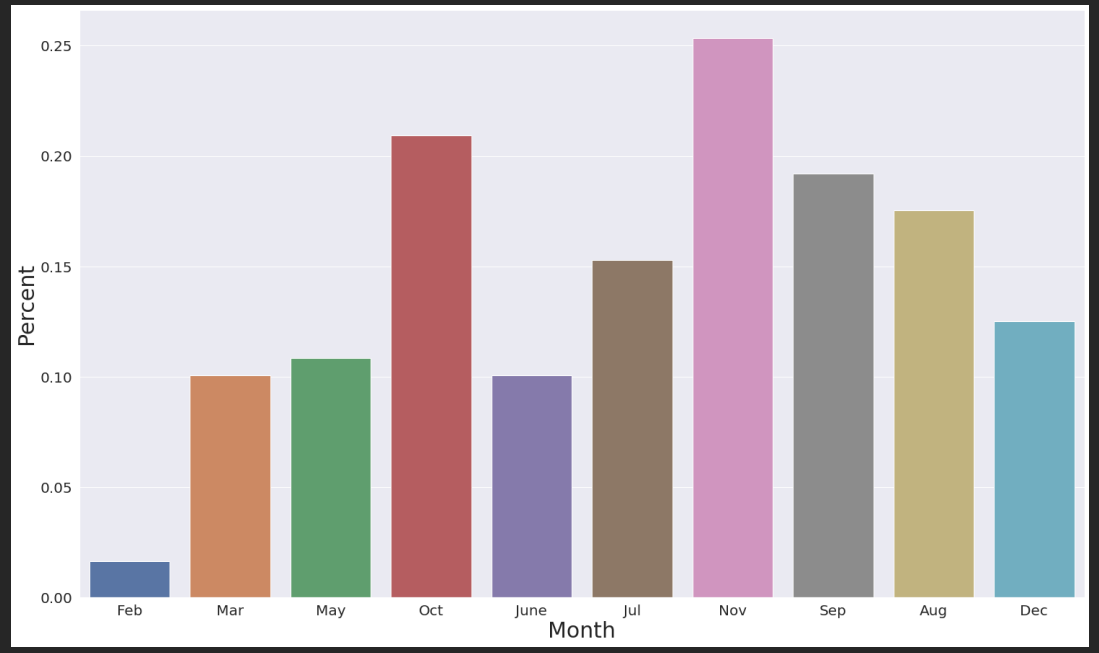
We tried different type of prediction thanks to the library scikit-learn

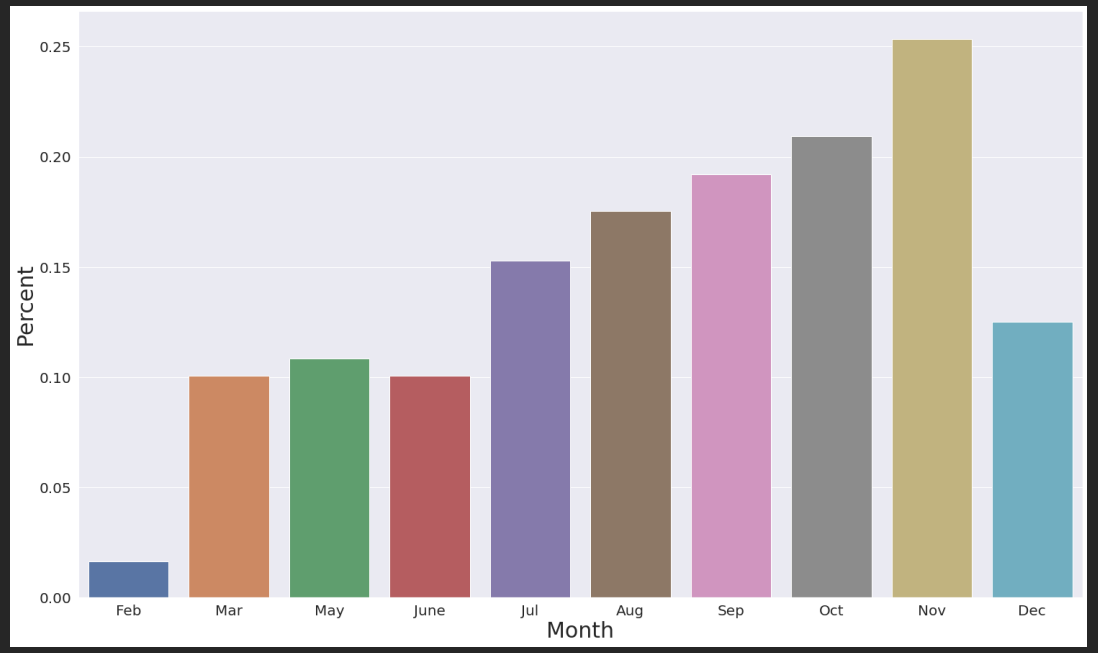


We compared the performances of each algorithm to obtain the optimal result

New variables in the dataset

- month_no (*int*): made from *month* variable. It allows to easily sort values by using *int* instead of *string*.
- IntervalProduct (*int*): we made intervals for the number of product related pages visited in the session to decrease the number of categories in a bar chart.
- New_Visitor, Other, Returning_Visitor: label binarizer for scikit-learn



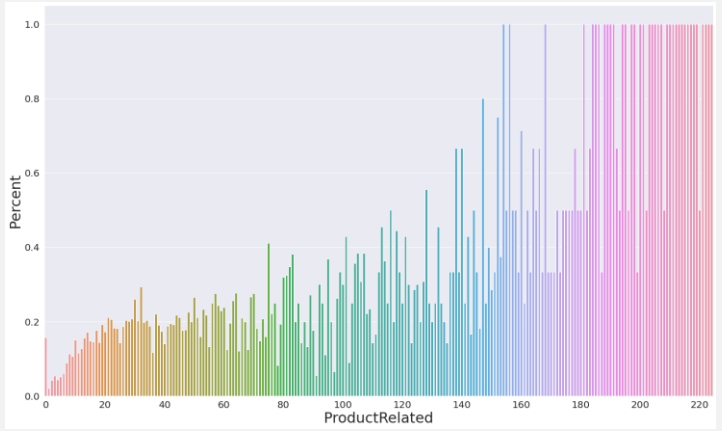


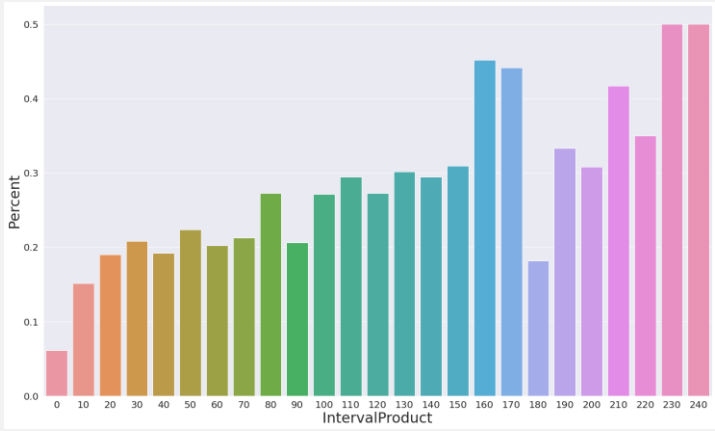
month_no

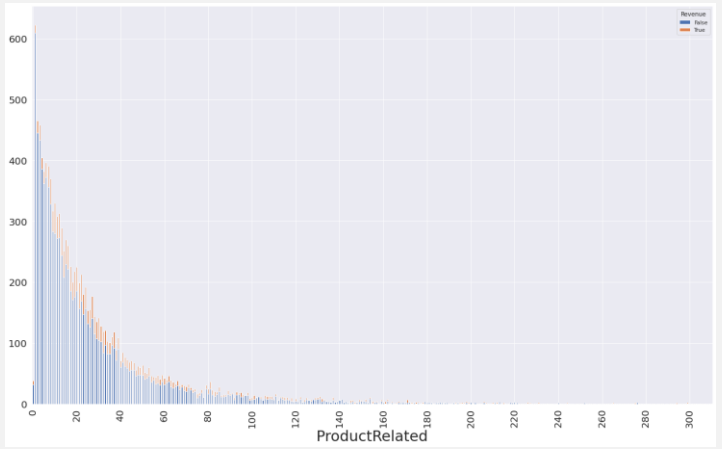
Without the *month_no* variable, the plots with month in x axis could'nt be ordered correctly, and so, were more difficult to read and interpret.

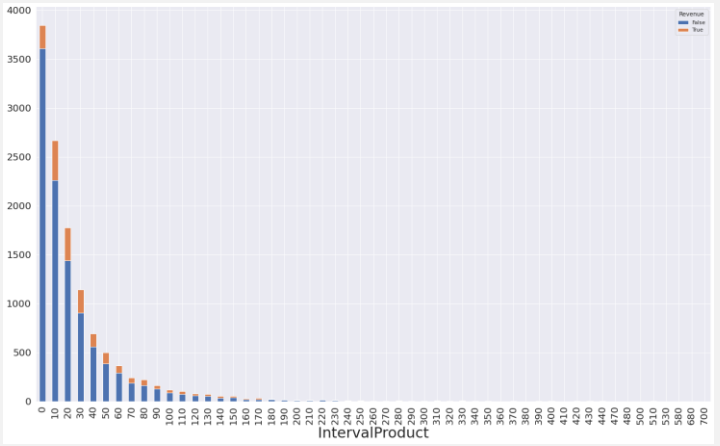
IntervalProduct

We created intervals of 10 pages to reduce the number of bars in the chart. The first line of plots represents the percentage of customers who buy and the second line the absolute frequency.









The Flask API