ONLINE SHOPPERS INTENTION

GALI Maikel

AZER Jérémy

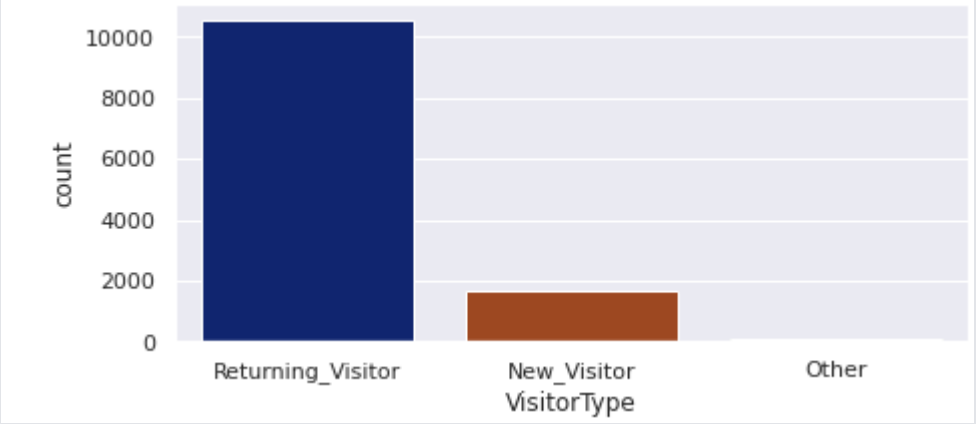
INFORMATION OF OUR DATASET :

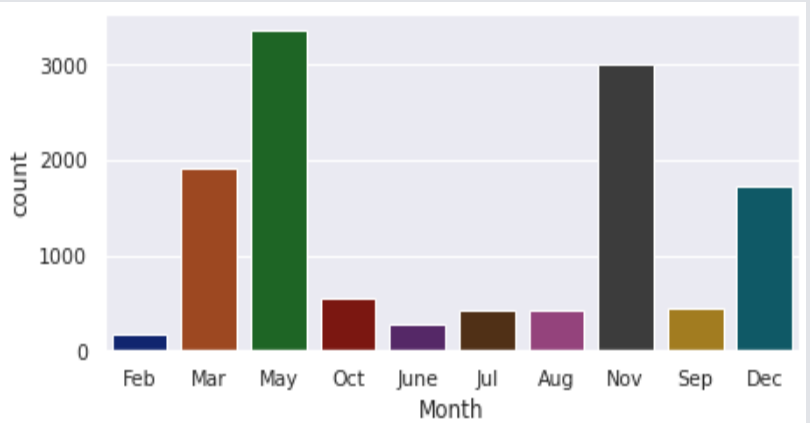
In our dataset, we have 10 numerical and 8 categorical variables. The main variables is Revenue and the two options are True or false and consist of purchase or not of shoppers online. In this dataset, we don't have any missing values but the dataset is unbalanced data. That's why we use an preprocessing of our data.

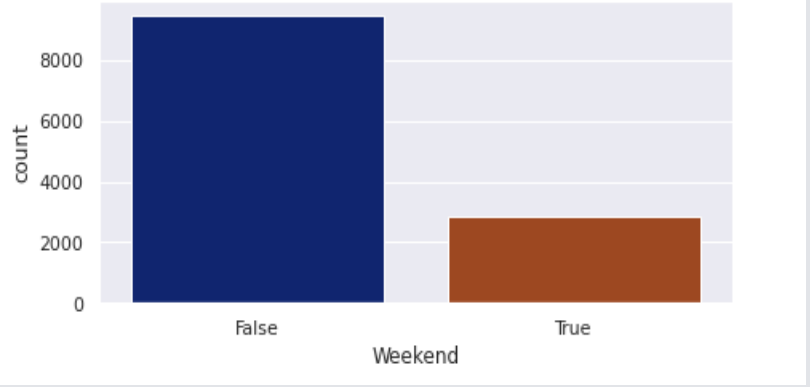
Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType
0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	1	1	1	1
0.0	2	64.000000	0.00	0.10	0.0	0.0	Feb	2	2	1	2
0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	4	1	9	3
0.0	2	2.666667	0.05	0.14	0.0	0.0	Feb	3	2	2	4
0.0	10	627.500000	0.02	0.05	0.0	0.0	Feb	3	3	1	4

GRAPHICS WITH PURCHASE

We have purchase by month, visitor type and week end. The dataset is quite unbalanced. We have to do preprocessing to the data. There is more négatives entries and most of our data is composed of returning_visitor. The second graph shows the number of no purchase and he is more important in May and november and indicates that they are more visits to the website this month. The last graph shows that people don’t really go the website on Weekend.

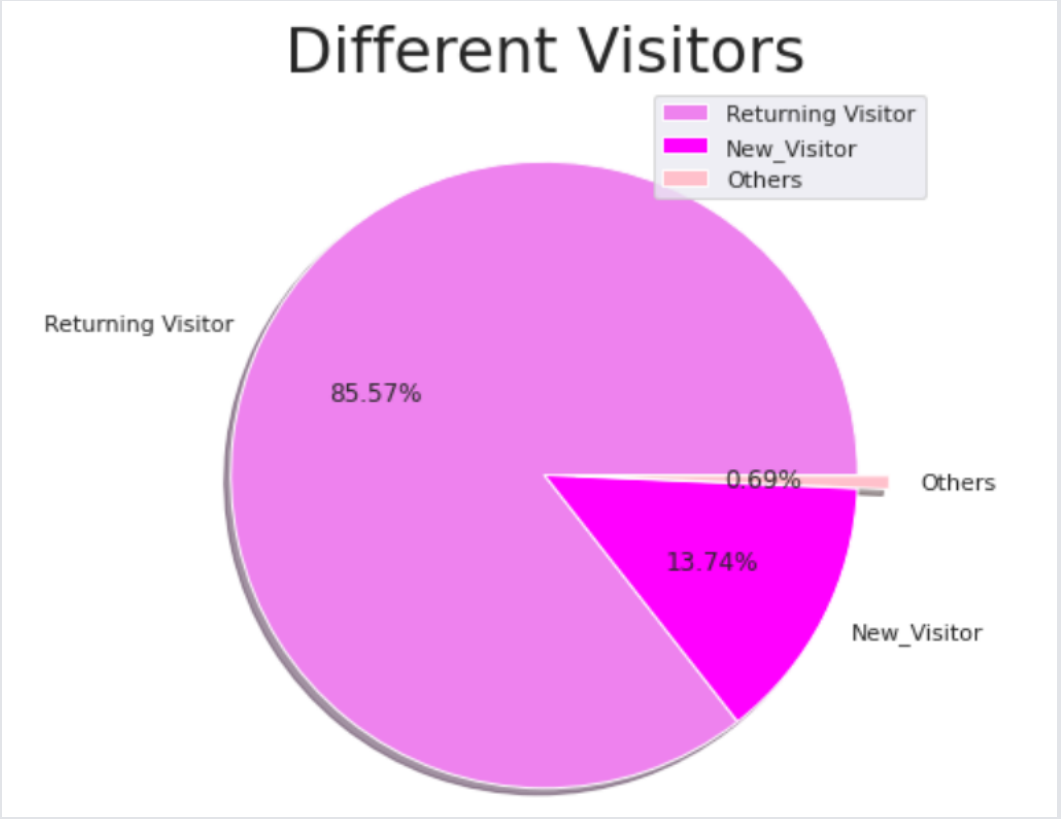






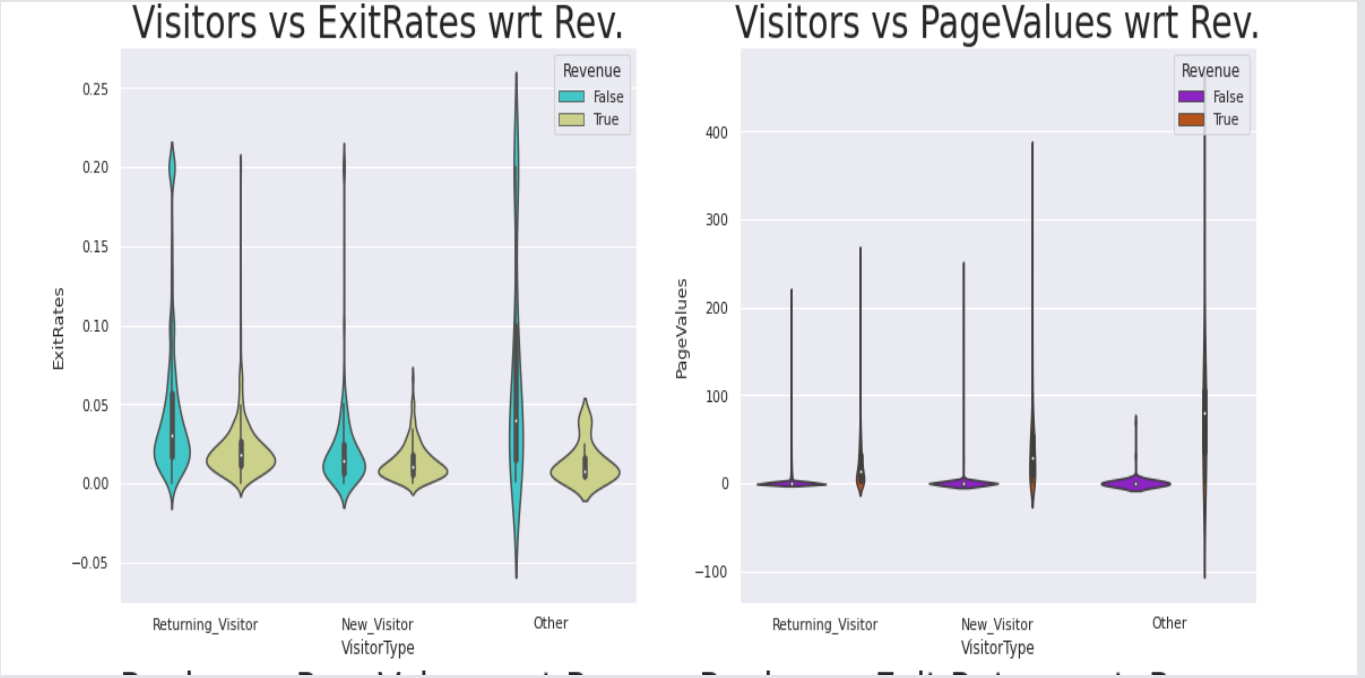
VISITORS ON ONLINE SHOPPERS

We can also see that we have more retourning visitor than other thing because of unbalanced data and people maybe not be interesting in the product.



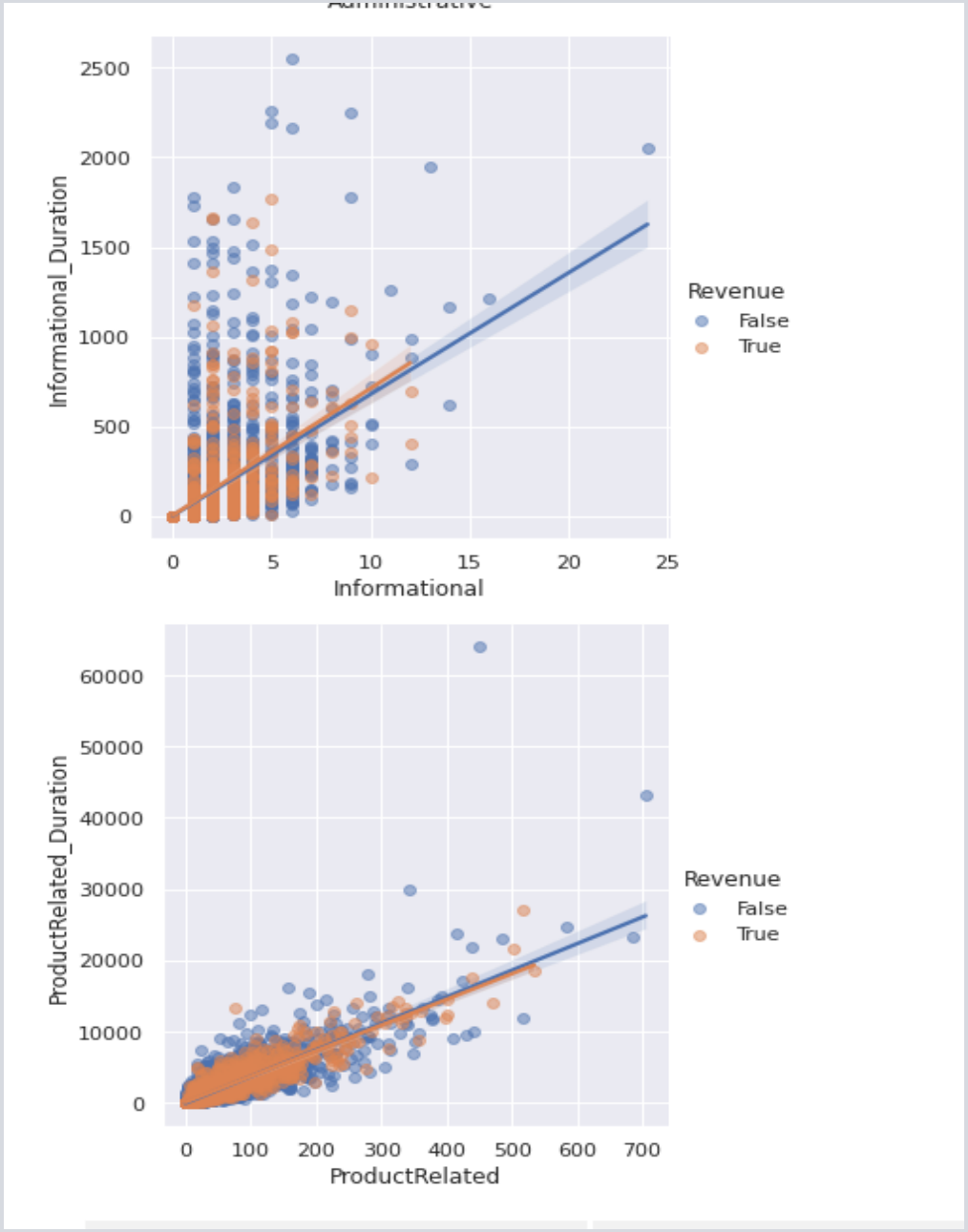
MULTIVARIATE ANALYSIS

Visitors tend to visit less pages and don't have time to browse the pages and because of this, people are not going to buy something.
The number of product related pages, and the time spent on them, is way higher than that for account related or informational pages.
The first 3 features look like they follow a skewed normal distribution.



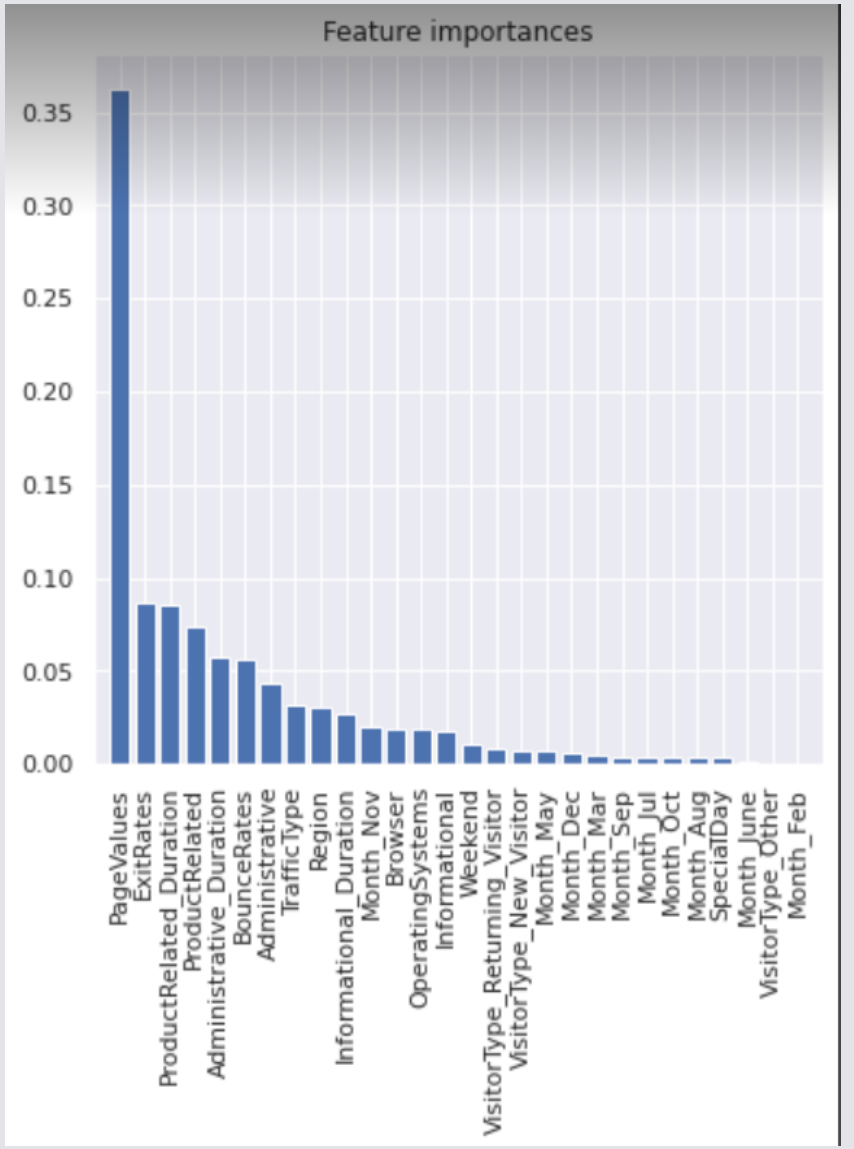
GRAPHICS OF LINEAR REGRESSION

We can see that point when Revenue is true is close to the line and it means to people who purchase on the website and people who not purchase is not close to the line.



UNNECESSARY VARIABLES :

We use this graphics to know what variables is unnecessary for model like random forest, gaussian naive bayes and extra trees

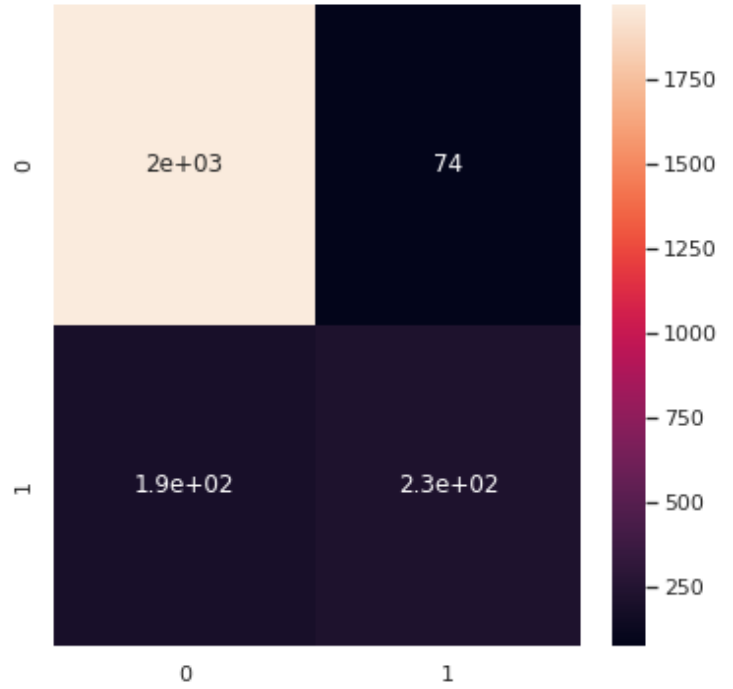


RANDOM FOREST 1:

For the random forest, we do two preprocessing, the first preprocessing consists of convert the categorical variables into dummy/indicator variables

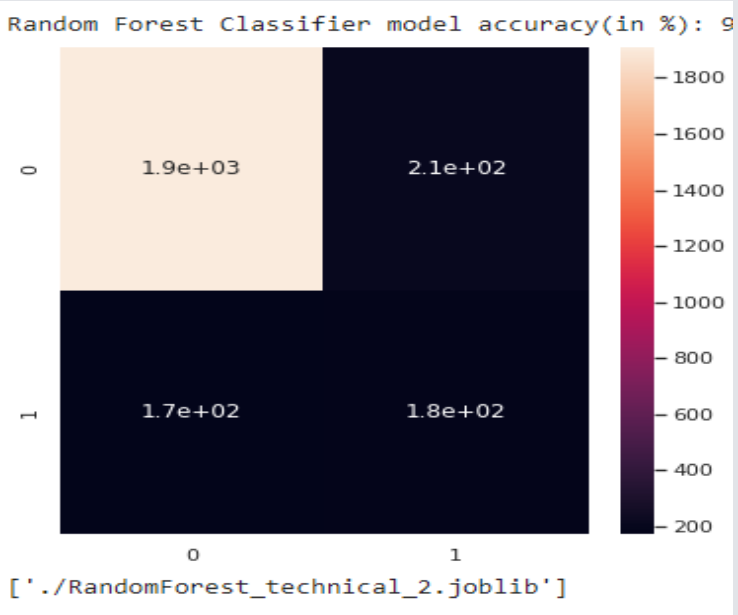
Metrics RandomForest :
Training Accuracy : 1.0
Testing Accuracy : 0.8917274939172749

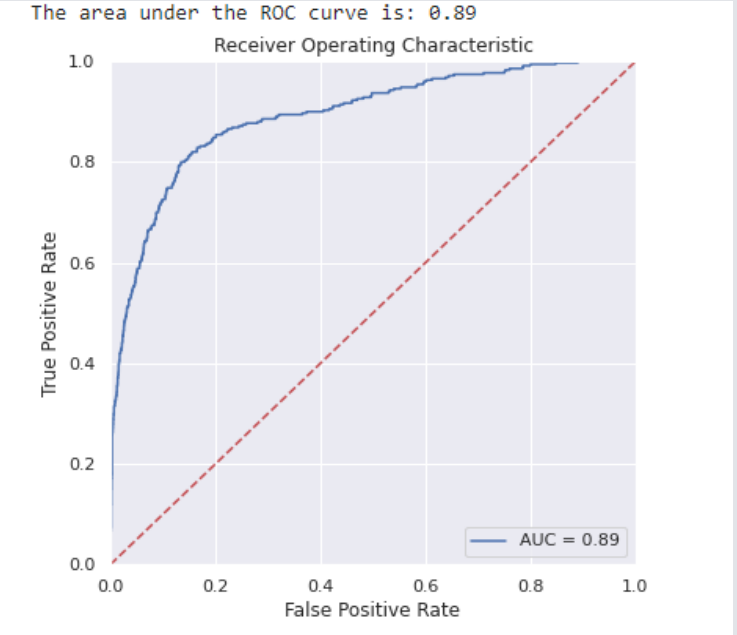
	precision	recall	f1-score	support
False	0.91	0.96	0.94	2044
True	0.76	0.54	0.63	422
accuracy			0.89	2466
macro avg	0.83	0.75	0.78	2466
weighted avg	0.88	0.89	0.88	2466



RANDOM FOREST 2:

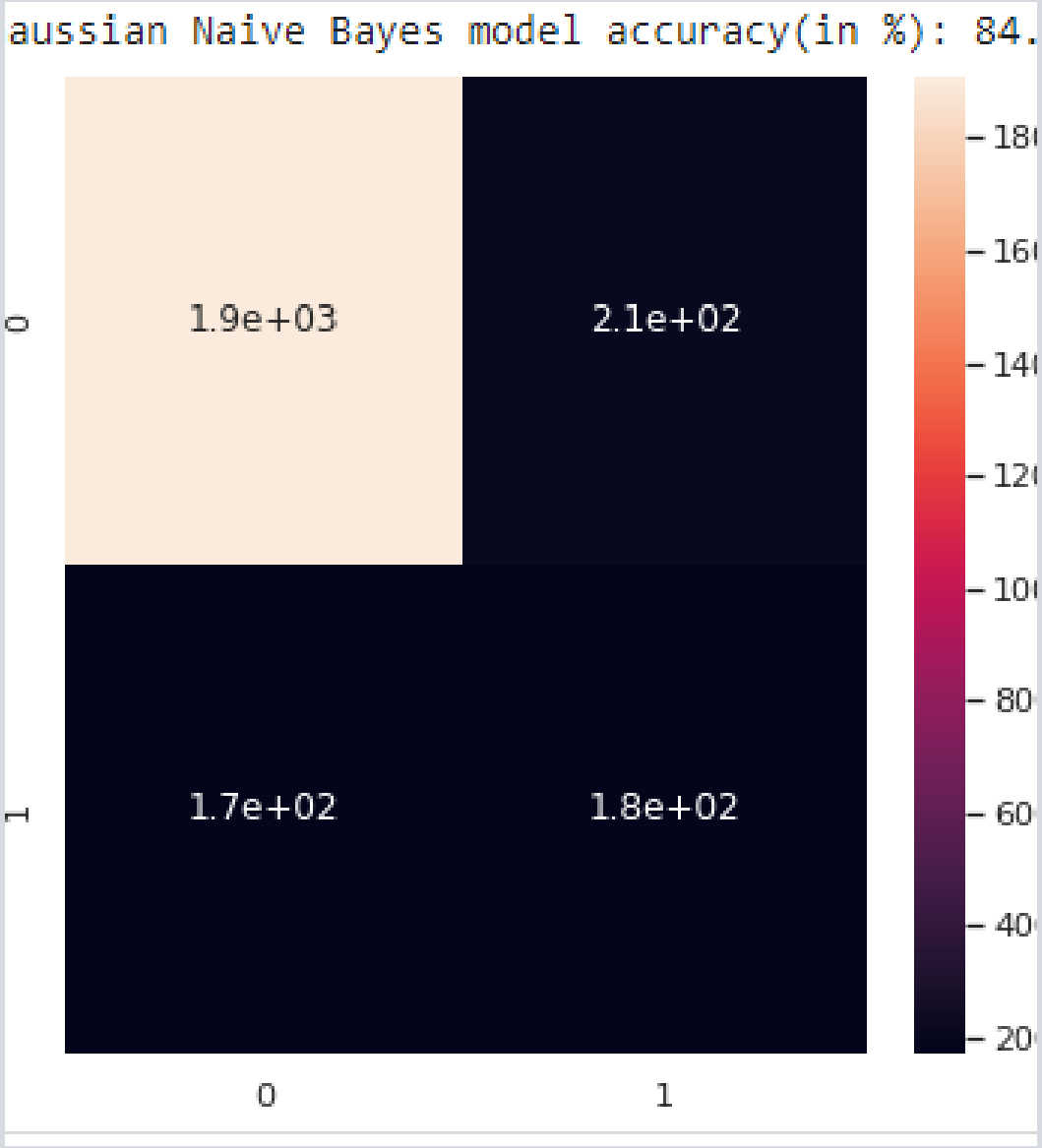
For the random forest, we do two preprocessing, the first preprocessing consists of convert the categorical variables into dummy/indicator variables and the second preprocessing consists of removing unnecessary columns. It is the best model because he has the best accuracy.





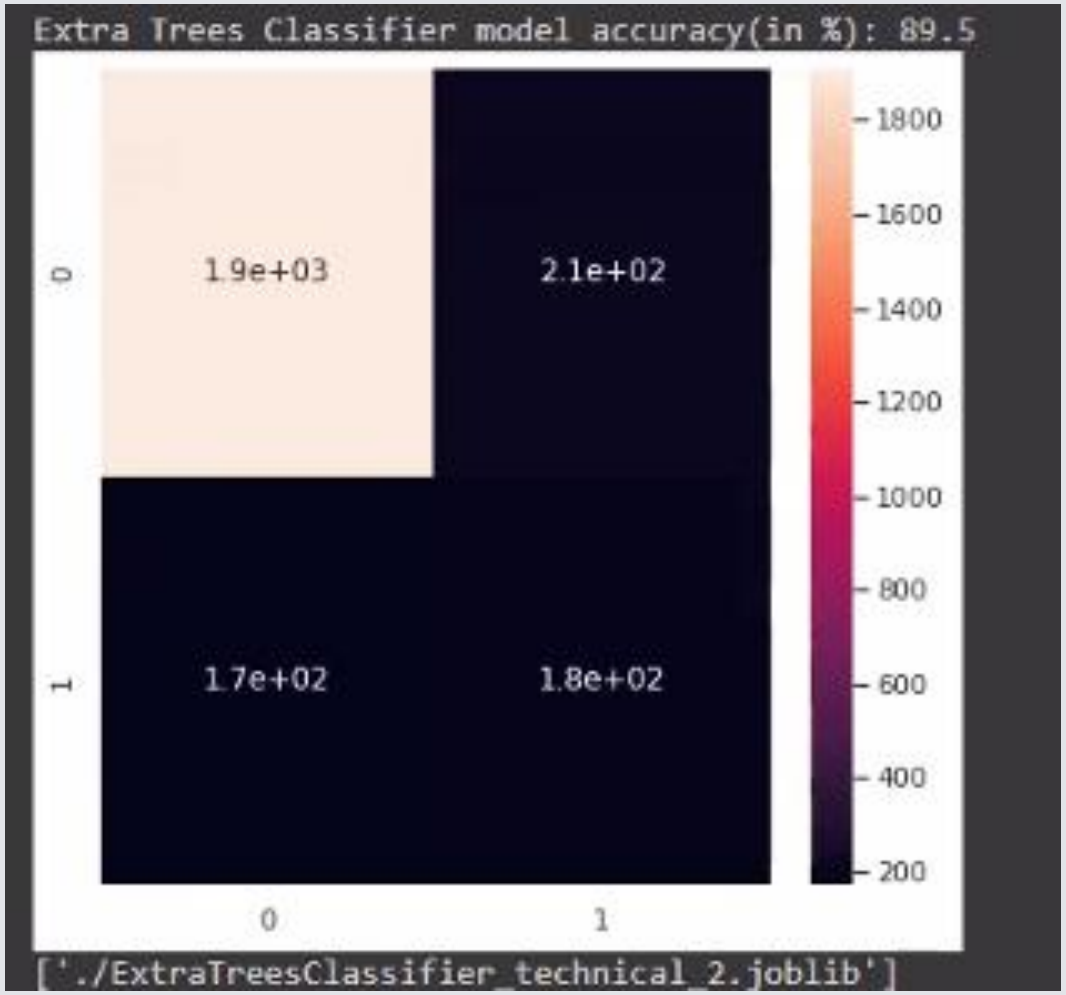
GAUSSIAN NAIVE BAYES :

For the Gaussian Naive Bayes, we do two preprocessing, the first preprocessing consists of convert the categorical variables into dummy/indicator variables and the second preprocessing consists of removing unnecessary columns.

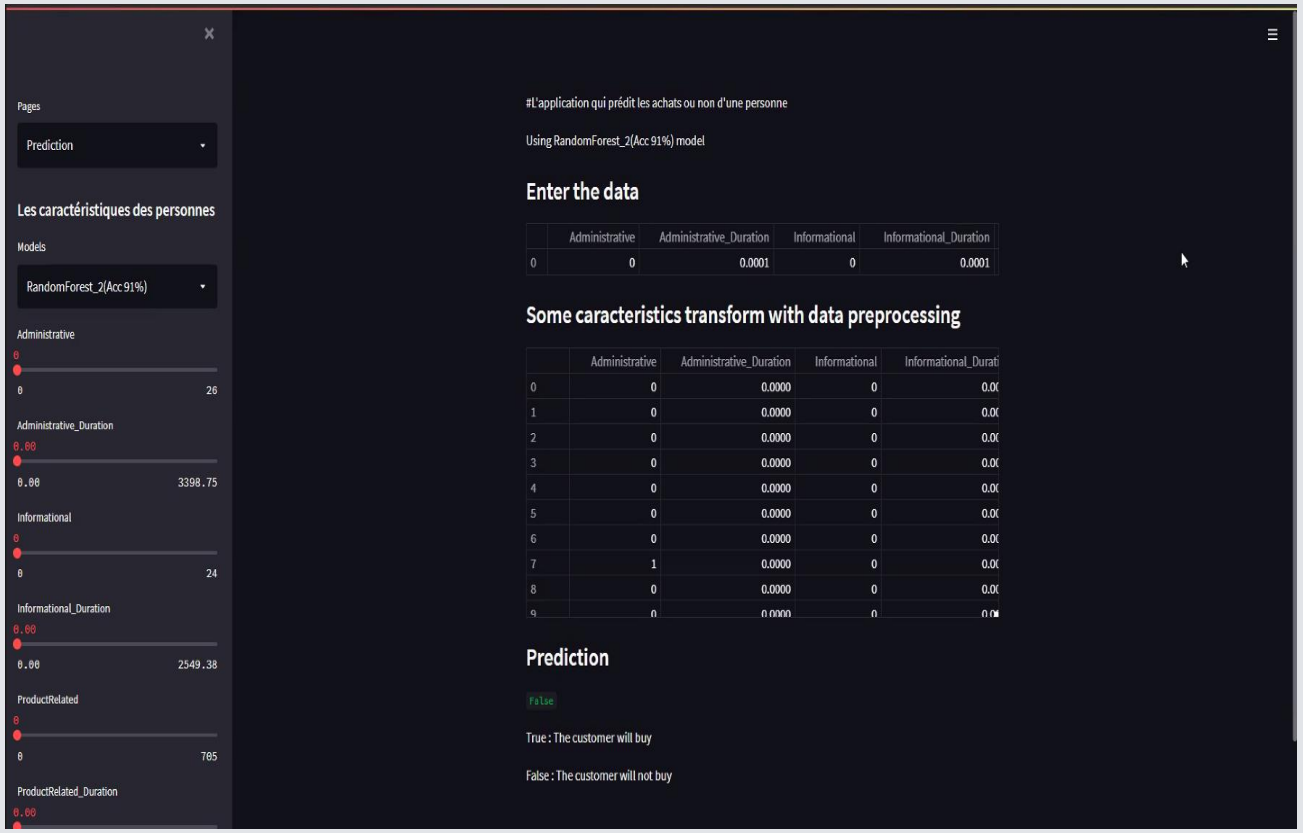


EXTRA TREES :

For the Extra trees, we do two preprocessing, the first preprocessing consists of convert the categorical variables into dummy/indicator variables and the second preprocessing consists of removing unnecessary columns.



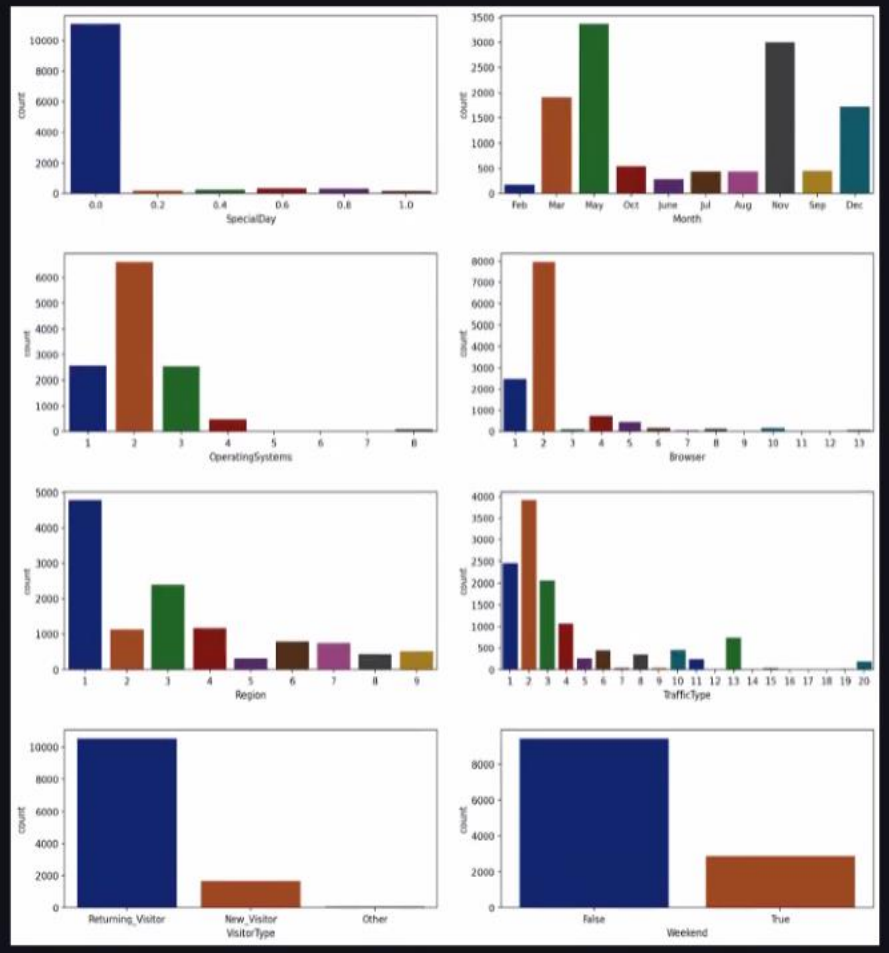
API :



After the export of our model with pickle and preprocessing, we built a rest API with streamlit. Thanks to the input of data in the data, we can predict if someone purchase or not. We have the API streamlit with checkbox and you choose prediction or see data vizualisation. On the left, you can choose your own data. Below, we have the requests to launch the API.

```
PS C:\Users\maikel\PycharmProjects\final-project-python> clear
PS C:\Users\maikel\PycharmProjects\final-project-python> streamlit run main.py
```

Pages

Dataset online_shoppers_intention

ProductRelated_Duration	12,330.0000	1,194.7462	1,913.6693	0.0000	184.1375
BounceRates	12,330.0000	0.0222	0.0485	0.0000	0.0000
ExitRates	12,330.0000	0.0431	0.0486	0.0000	0.0143
PageValues	12,330.0000	5.8893	18.5684	0.0000	0.0000
SpecialDay	12,330.0000	0.0614	0.1989	0.0000	0.0000