ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Online Shoppers Purchasing Intention

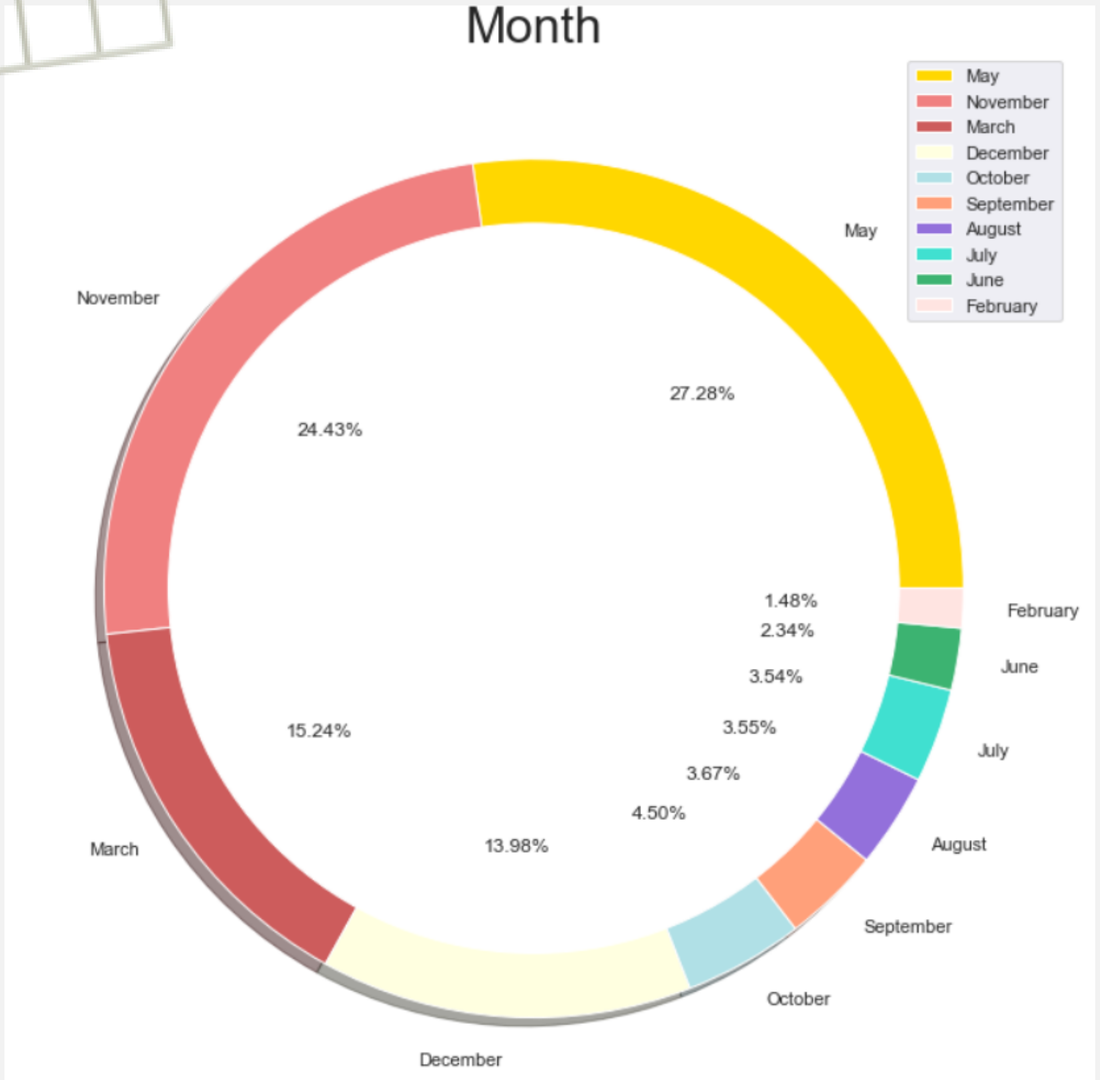
Analysis and Prediction

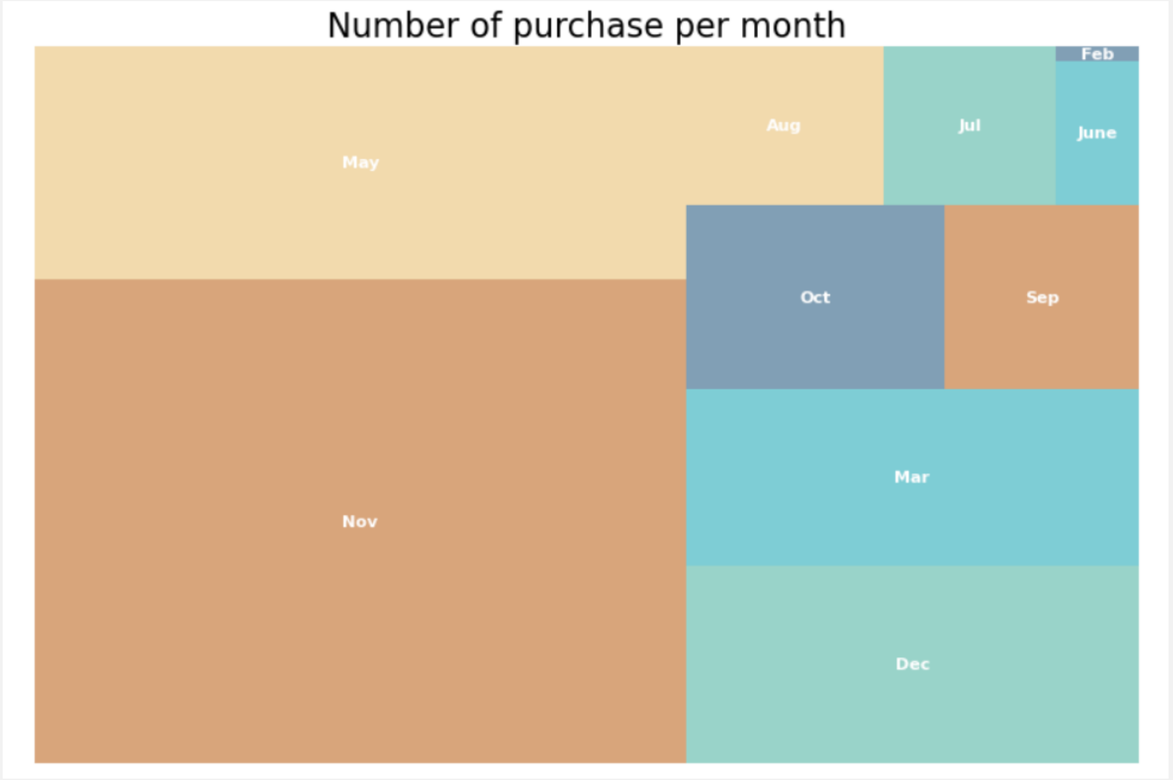
KARUNATHSAN Nilany & ELMALEH Tom & SAMBATH Sïndoumady

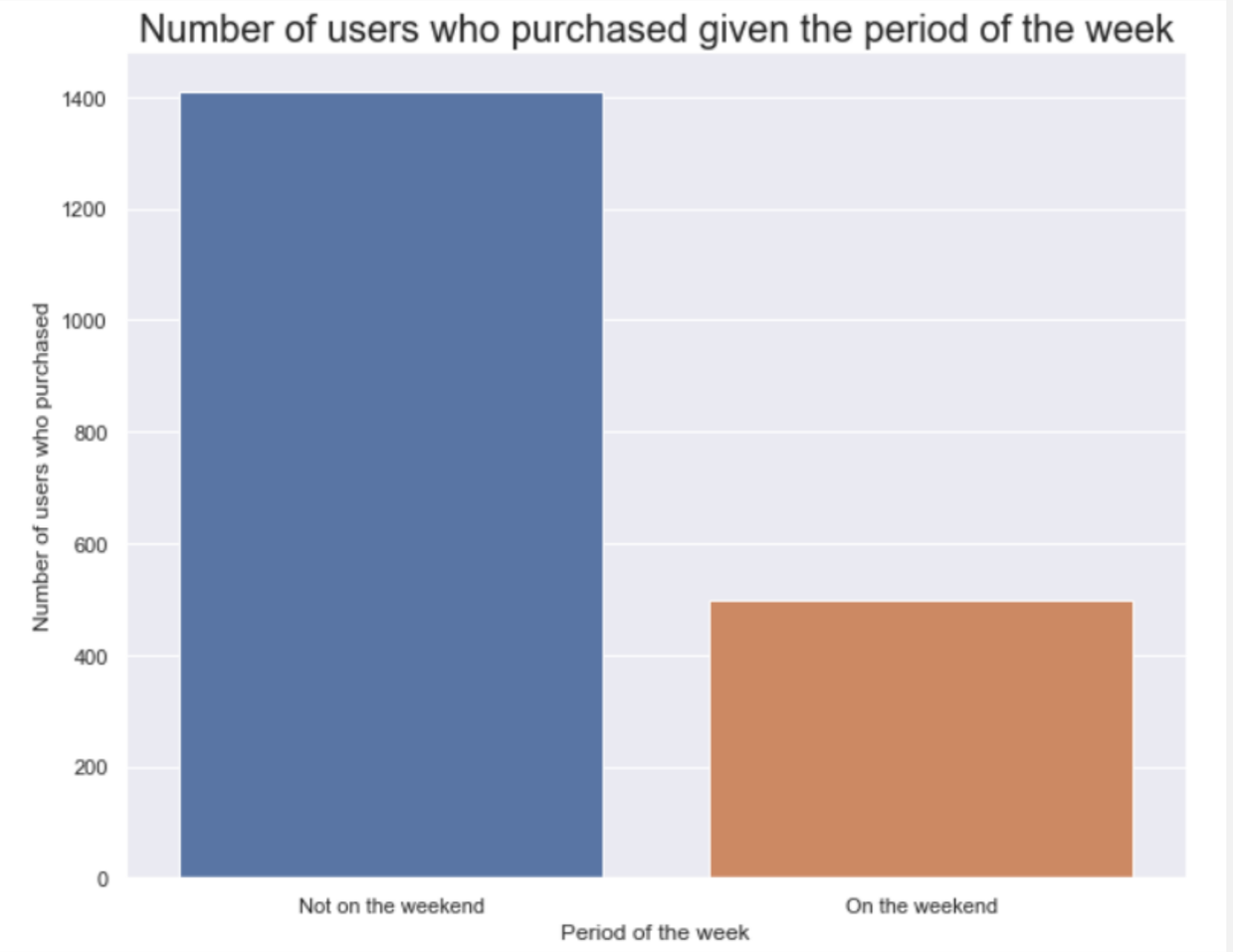
Dataset exploration & cleaning

- The dataset consists of **10 numerical** and **8 categorical features.**
- The '**Revenue**' attribute is used as class label.
- The dataset is clean, there are **no missing values**.
- The first 4 columns weren't relevant to us, so we removed them. Moreover, they are among those which contained the most null values.

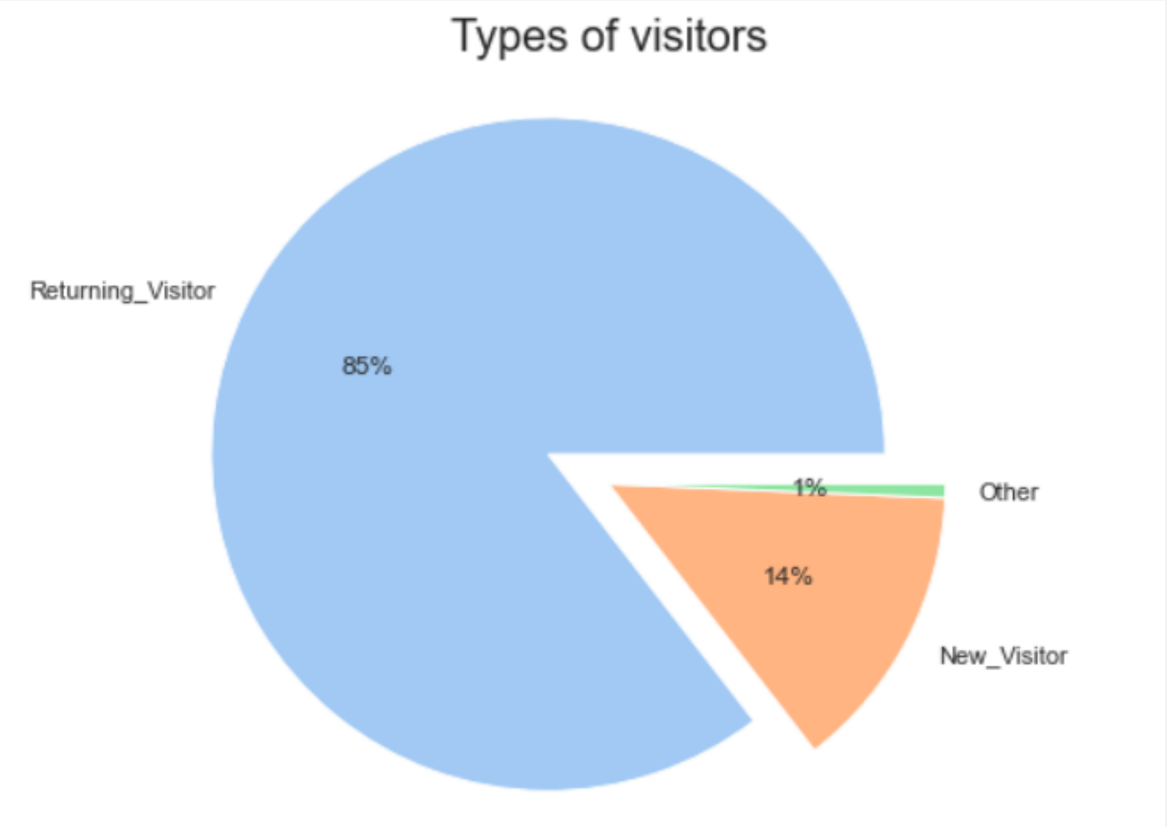
Data visualisation

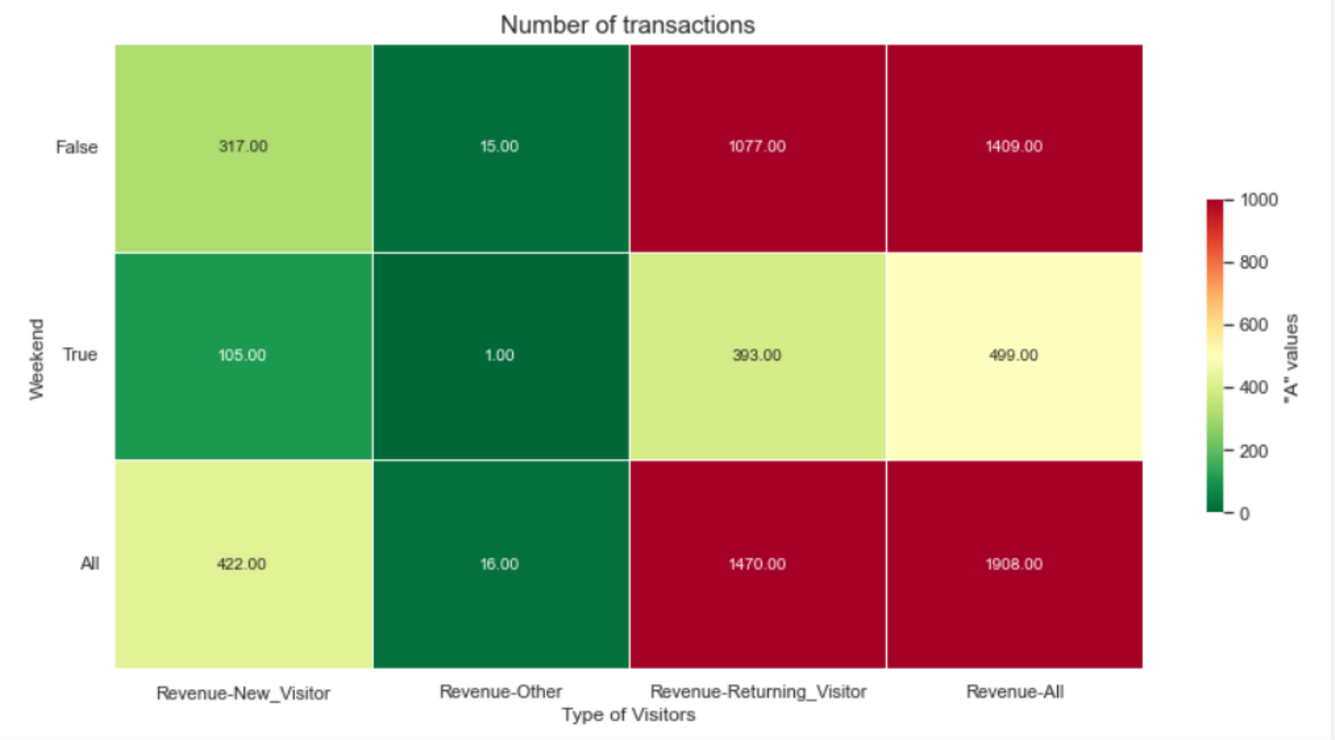





Number of users who purchased given the period of the week
1400
1200
1000
800
600
400
200
0
Number of users who purchased
Not on the weekend
On the weekend
Period of the week


Page visited on weekend
False
(9346)
True
(2859)

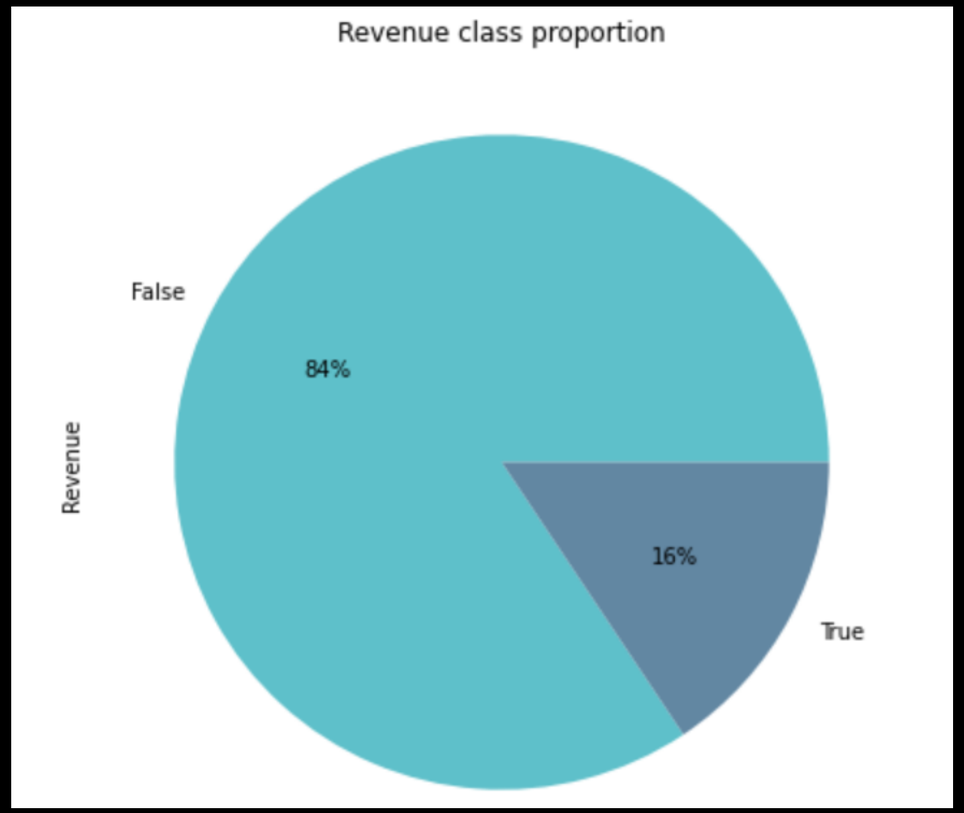

Types of visitors
Returning_Visitor
85%
1%
Other
14%
New_Visitor


Number of transactions
False
True
All
Weekend
317.00
15.00
1077.00
1409.00
105.00
1.00
393.00
499.00
422.00
16.00
1470.00
1908.00
Revenue-New_Visitor
Revenue-Other
Revenue-Returning_Visitor
Revenue-All
Type of Visitors
1000
800
600
400
200
0
"A" values

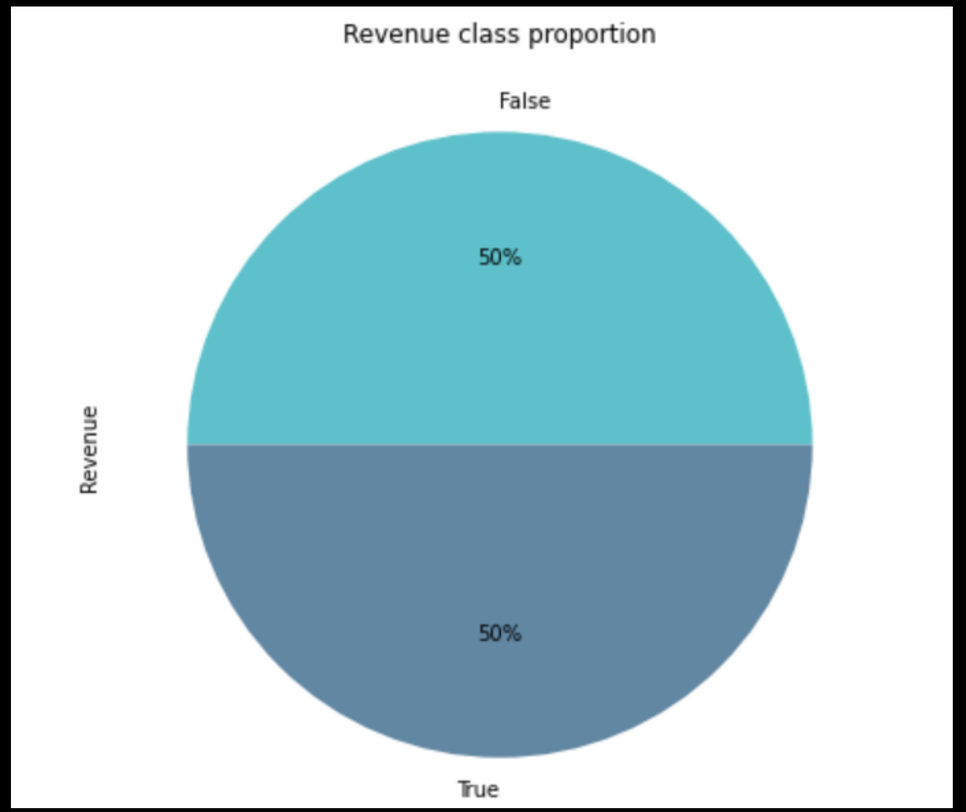
Data pre-processing

- Variable encoding
- Handling unbalanced datasets by implementating a resampling strategy
- Dataset splitting
- Scaling

Resampling strategy SMOTE



Before SMOTE



After SMOTE

Machine learning & Modeling

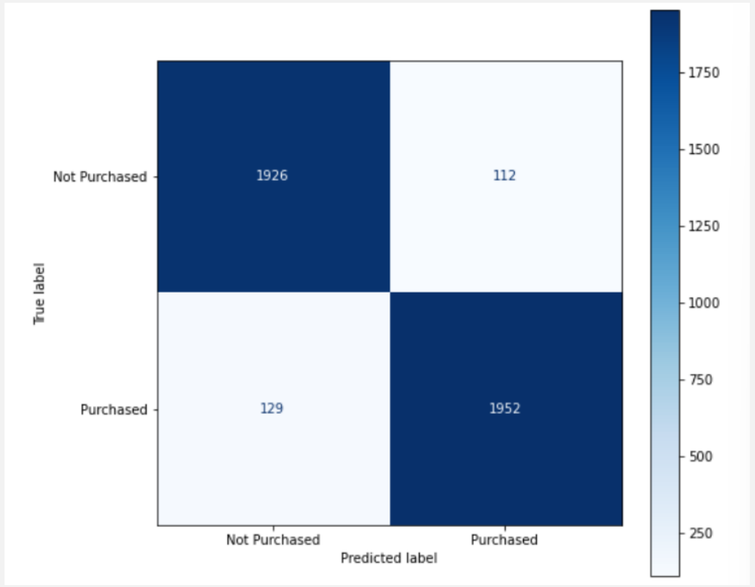
For each model, we executed the following steps :

1. **Import the necessary libraries**
2. **Define the grid parameters**
3. **Apply GridSearchCV and evaluate the result**
4. **Save the best model obtained in a specified variable using best_estimator**
5. **Save the score obtained on the test set using the model we just defined earlier**
6. **Plot the confusion matrix to evaluate the performance of the model**

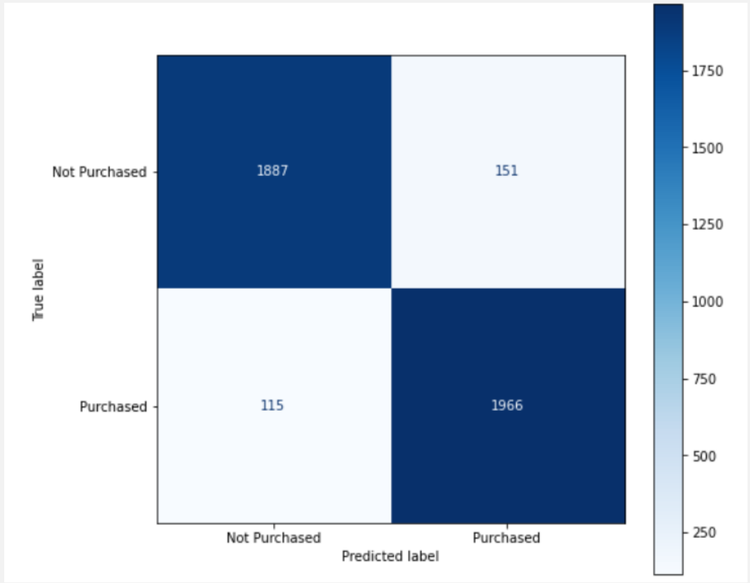
	Model	Score
0	XGBoost	0.941491
1	Random Forest	0.935421
2	Decision Tree	0.918184
3	KNN	0.911386
4	SVC	0.891964
5	Naive Bayes Classifier	0.883224
6	Stochastic Gradient Descent	0.737800
7	Gradient Boosting Score	0.737800

The three best models

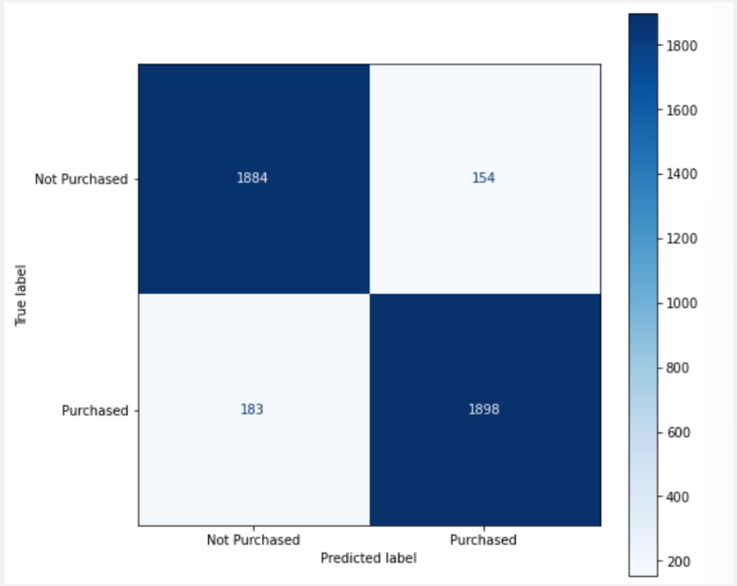
Scores Ranking



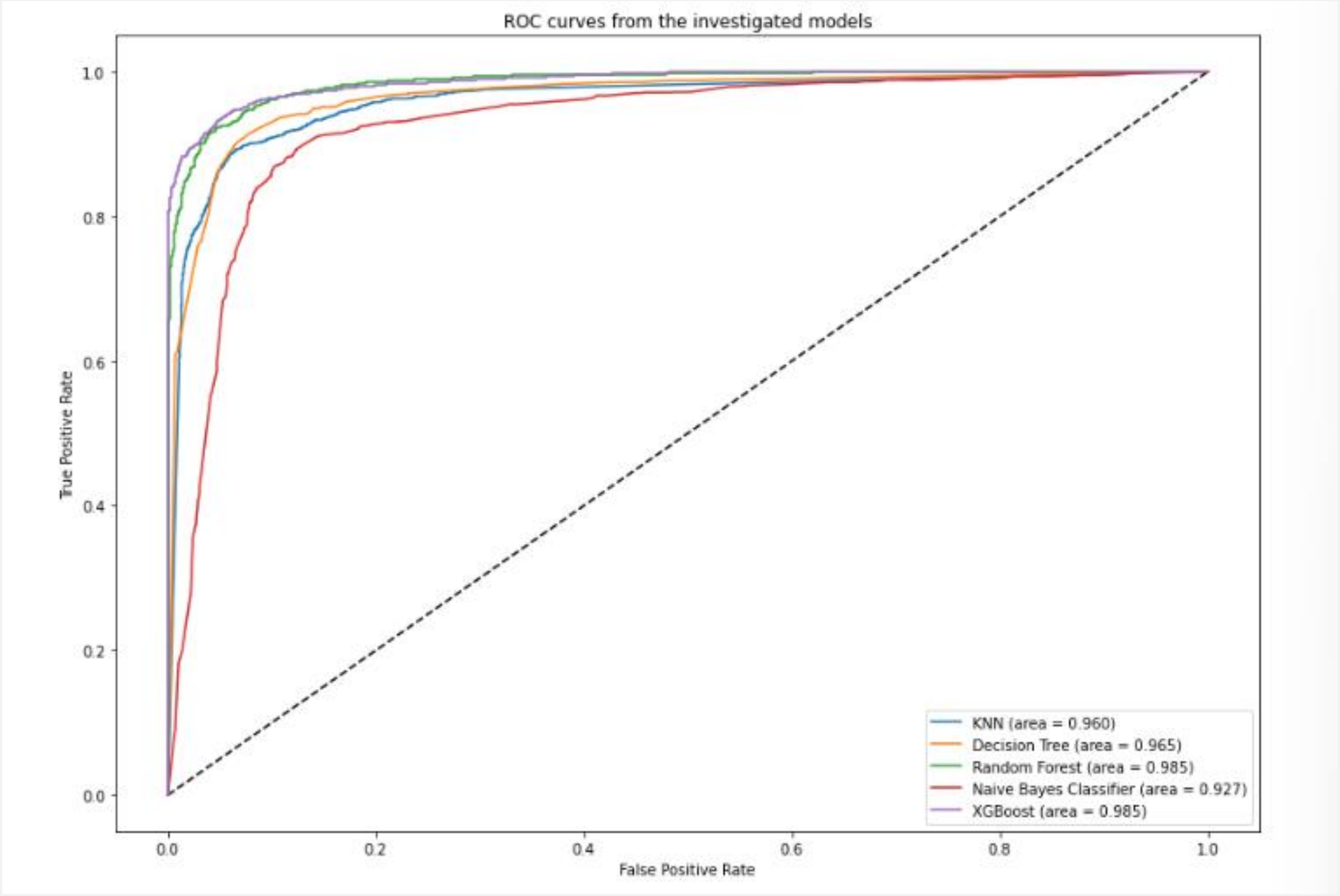
1 XGBOOST



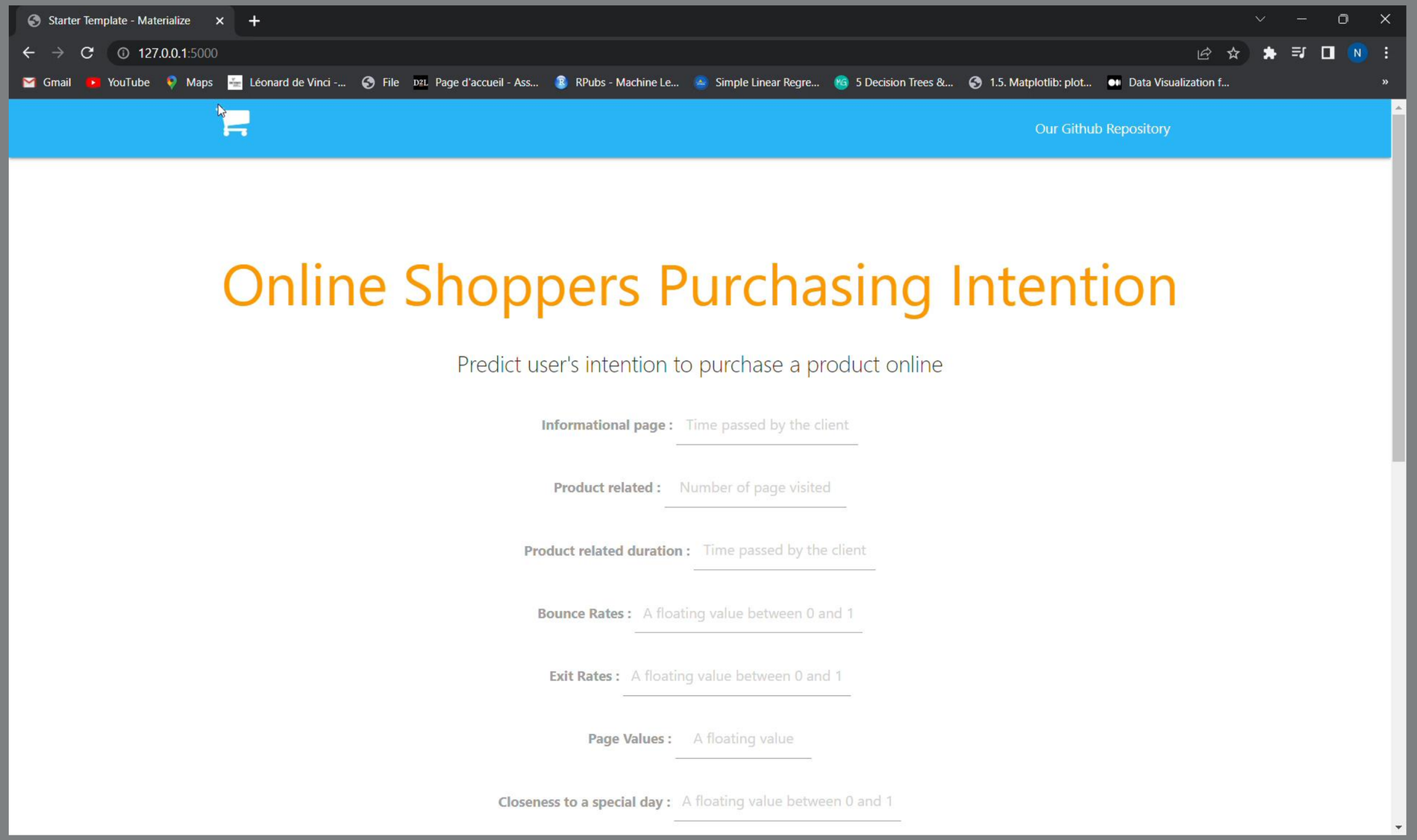
2 RANDOM FOREST



3 DECISION TREE


ROC curves from the investigated models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
KNN (area = 0.960)
Decision Tree (area = 0.965)
Random Forest (area = 0.985)
Naive Bayes Classifier (area = 0.927)
XGBoost (area = 0.985)

API


Starter Template - Materialize
127.0.0.1:5000
Gmail
YouTube
Maps
Léonard de Vinci -...
File
Page d'accueil - Ass...
RPubs - Machine Le...
Simple Linear Regre...
5 Decision Trees &...
1.5. Matplotlib: plot...
Data Visualization f...
Our Github Repository
Online Shoppers Purchasing Intention
Predict user's intention to purchase a product online
Informational page : Time passed by the client
Product related : Number of page visited
Product related duration : Time passed by the client
Bounce Rates : A floating value between 0 and 1
Exit Rates : A floating value between 0 and 1
Page Values : A floating value
Closeness to a special day : A floating value between 0 and 1