



# Online Shoppers Purchasing Intention Dataset

# Sommaire



➡ **Le Dataset (Compréhension, Exploratory Data Analysis)**

➡ **MACHINE LEARNING (classification)**

➡ **API**

# Le Dataset



Le dataset consiste en 12,330 sessions.

L'ensemble de données a été formé de sorte que chaque session appartienne à un utilisateur différent sur une période d'un an pour éviter toute tendance à une campagne, un jour spécial, un profil d'utilisateur ou une période spécifique.

## Objectif du Projet :

Prédire si une session se terminera par un achat ou non.

Le dataset est composé de 17 features + 1 target

RangeIndex: 12330 entries, 0 to 12329

Data columns (total 18 columns):

#	Column	Non-Null Count	Dtype
0	Administrative	12330 non-null	int64
1	Administrative_Duration	12330 non-null	float64
2	Informational	12330 non-null	int64
3	Informational_Duration	12330 non-null	float64
4	ProductRelated	12330 non-null	int64
5	ProductRelated_Duration	12330 non-null	float64
6	BounceRates	12330 non-null	float64
7	ExitRates	12330 non-null	float64
8	PageValues	12330 non-null	float64
9	SpecialDay	12330 non-null	float64
10	Month	12330 non-null	object
11	OperatingSystems	12330 non-null	int64
12	Browser	12330 non-null	int64
13	Region	12330 non-null	int64
14	TrafficType	12330 non-null	int64
15	VisitorType	12330 non-null	object
16	Weekend	12330 non-null	bool
17	Revenue	12330 non-null	bool

dtypes: bool(2), float64(7), int64(7), object(2)

# Le Dataset - présentation des features



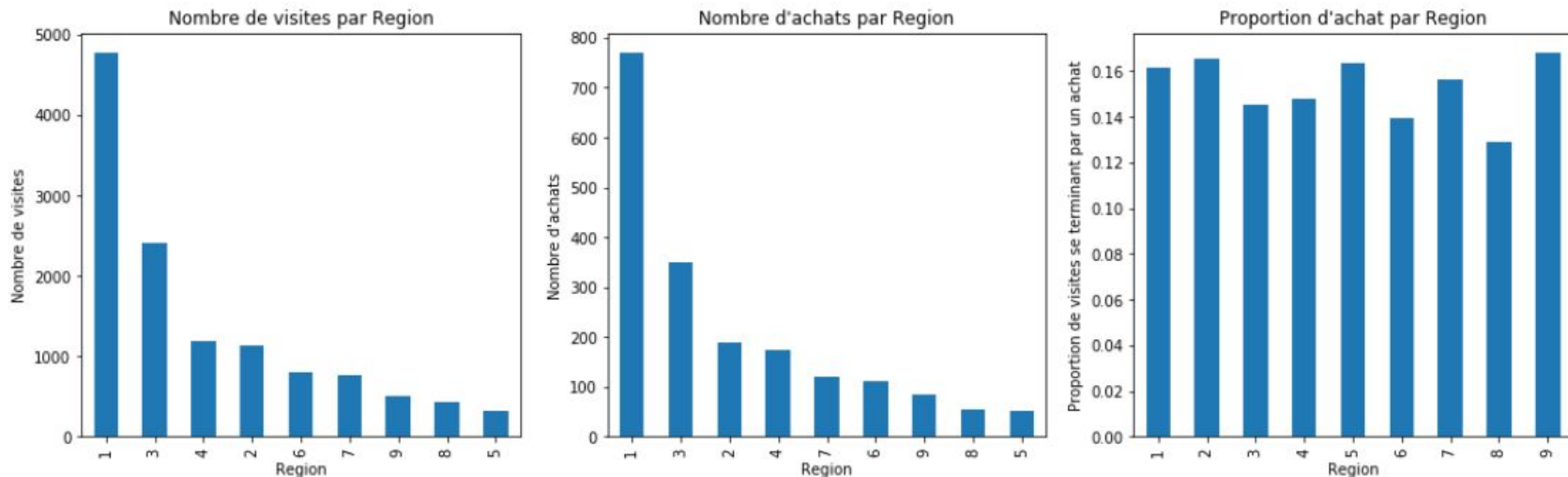
## Features du dataset :

- **Administrative** : Le nombre de pages “Administratives” que l'utilisateur a visité.
- **Administrative\_Duration** : Le temps passé sur ces pages “Administratives”.
- **Informational** : Le nombre de pages “Informatives” que l'utilisateur a visité.
- **Informational\_Duration** : Le temps passé sur ces pages “Informatives”.
- **ProductRelated** : This is the number of pages of this type (product related) that the user visited.
- **ProductRelated\_Duration** : This is the amount of time spent in this category of pages.
- **BounceRates** : Pourcentage de visiteurs entrant sur le site par cette page et qui en sortent directement sans aucune autre action.
- **ExitRates** : Pourcentage de session s'étant terminé sur cette page.
- **PageValues** : Valeur moyenne d'une page qu'un utilisateur a visitée avant d'accéder à la page d'objectif ou d'effectuer une transaction e-commerce..
- **SpecialDay** : Valeur entre 0 et 1. Plus elle se rapproche de 1 plus la visite se rapproche d'un jour spécial ( Saint Valentin...)
- **Month** : Le mois pendant lequel la visite a eu lieu.
- **OperatingSystems** : L'OS utilisé par le visiteur.
- **Browser** : Browser depuis lequel la connection a été effectuée.
- **Region** : Region depuis laquelle l'utilisateur s'est connecté.
- **TrafficType** : Integer indiquant quel type de trafic l'utilisateur utilise
- **VisitorType** : New Visitor, Returning Visitor, et Other.
- **Weekend** : Booléen indiquant si la personne s'est connectée le week end ou non.

## Target du dataset :

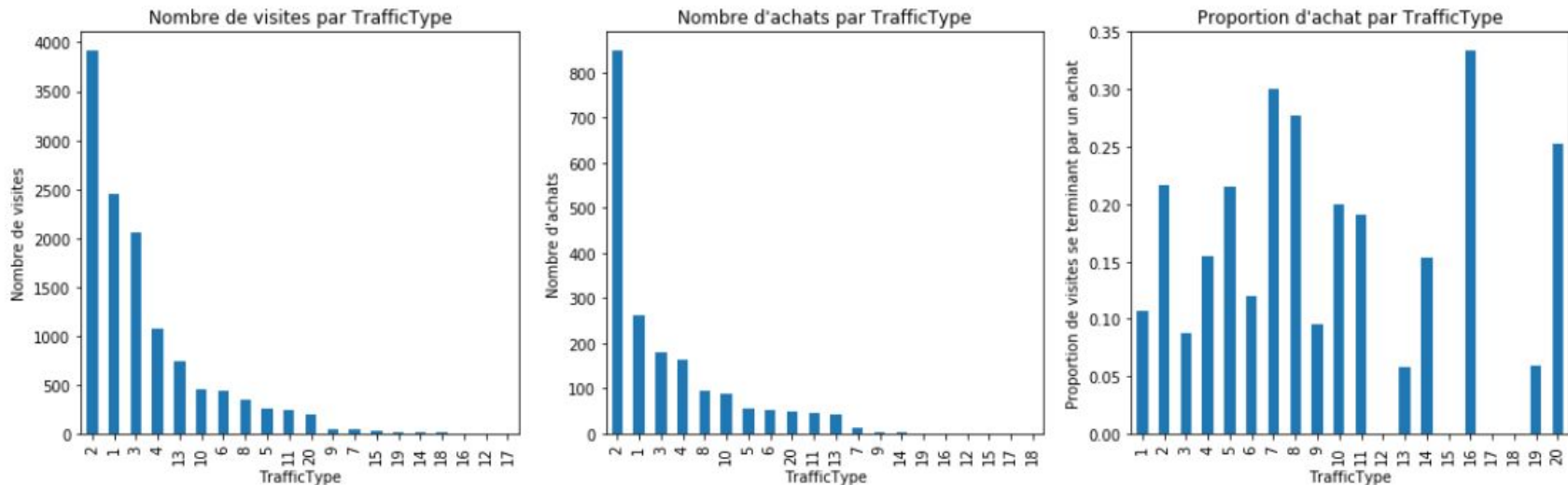
- **Revenue** : Booléen représentant si la visite s'est conclu par un achat ou non.

# Exploratory Data Analysis



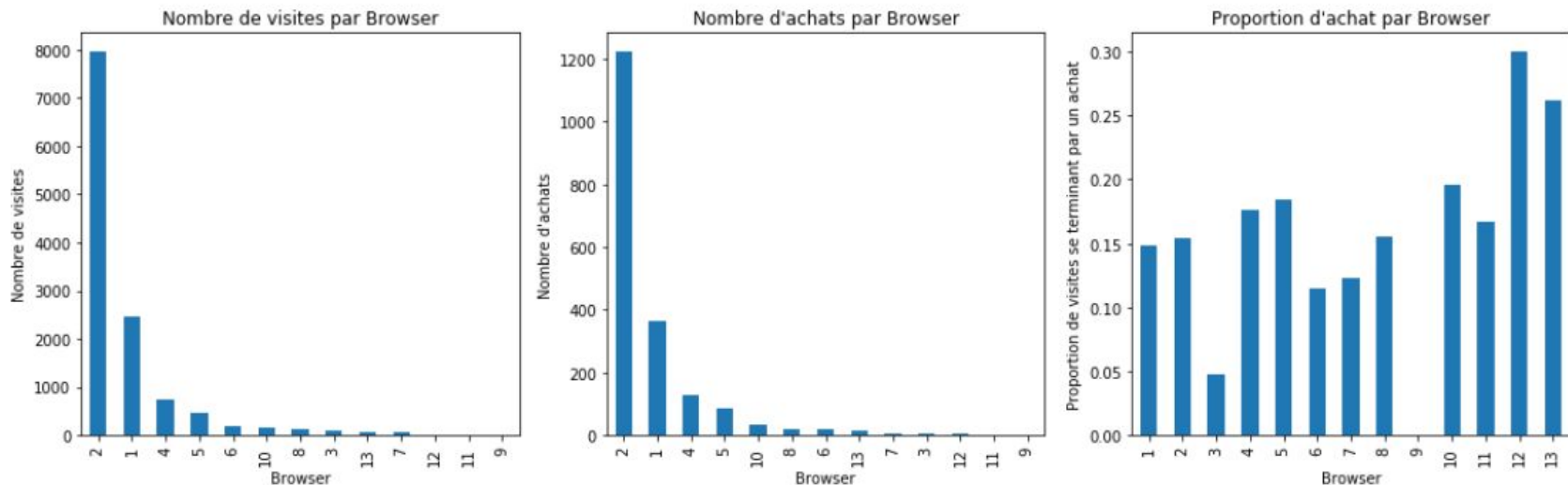
On observe que deux régions se distinguent des autres sur les deux schémas de gauche. Cependant, cette tendance n'est pas confirmée sur le troisième graphique. Nous pensons donc que la région n'aura pas un grand impact sur les différents modèles de machine learning.

# Exploratory Data Analysis



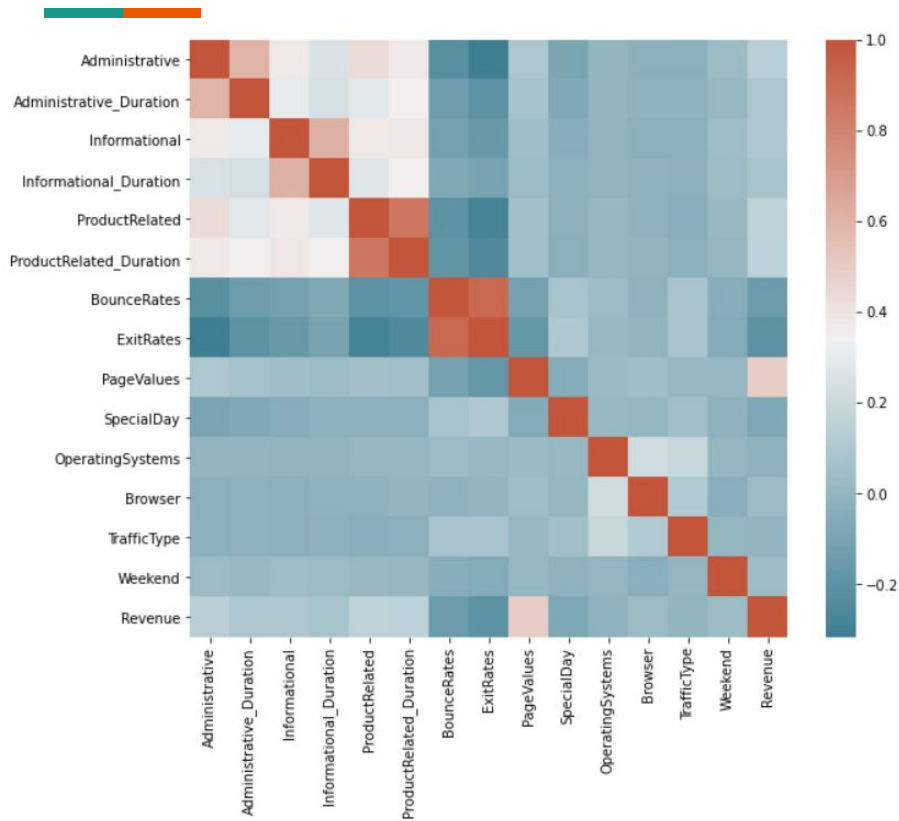
On observe ici que le TrafficType joue un rôle particulier dans le taux de conversion, puisque les taux en fonction de celui-ci diffèrent grandement. Aussi, on remarque que le TrafficType ayant le taux le plus élevé (schéma de droite) est le 16. Or, ce TrafficType ne contient que très peu de visites (schéma de gauche). Pareil pour le 20..

# Exploratory Data Analysis



On observe ici que le Browser joue un rôle particulier dans le taux de conversion, puisque les taux en fonction de celui-ci diffèrent grandement. Aussi, on remarque que le Browser ayant le nombre de visites le plus élevé est le 2 (schéma de gauche). Or, ce Browser ne contient qu'un faible taux de conversion (schéma de droite).

# Exploratory Data Analysis



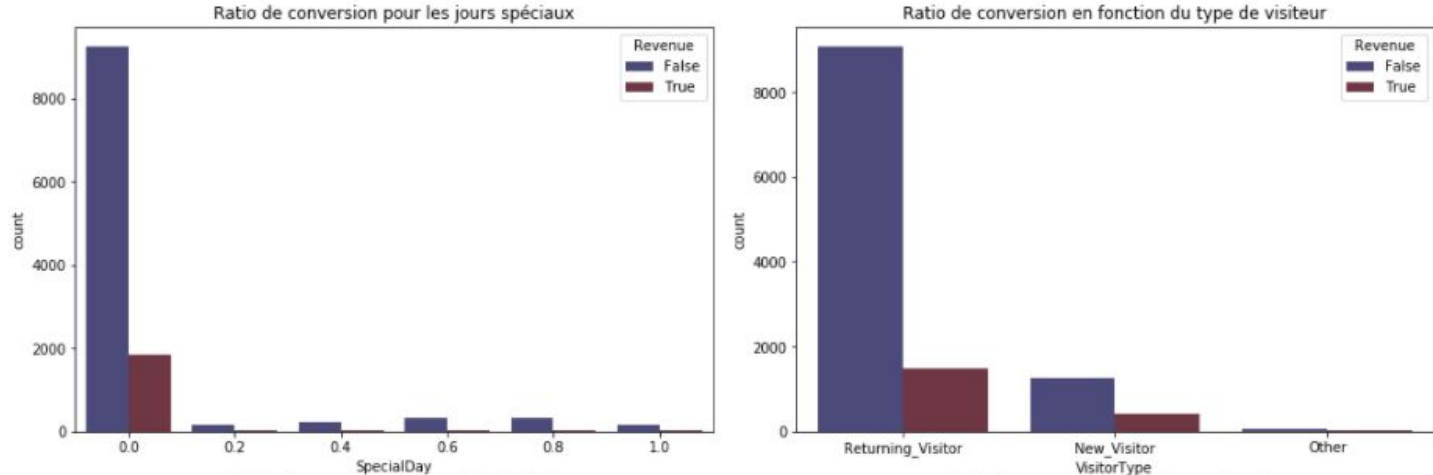
## Matrice de Corrélation

On observe dans cette matrice de corrélation que la feature PageValues est fortement reliée à la feature Revenue.

On observe aussi des "carrés" de corrélation rouges, qui sont logiques (Information est forcément corrélé à Information Duration)



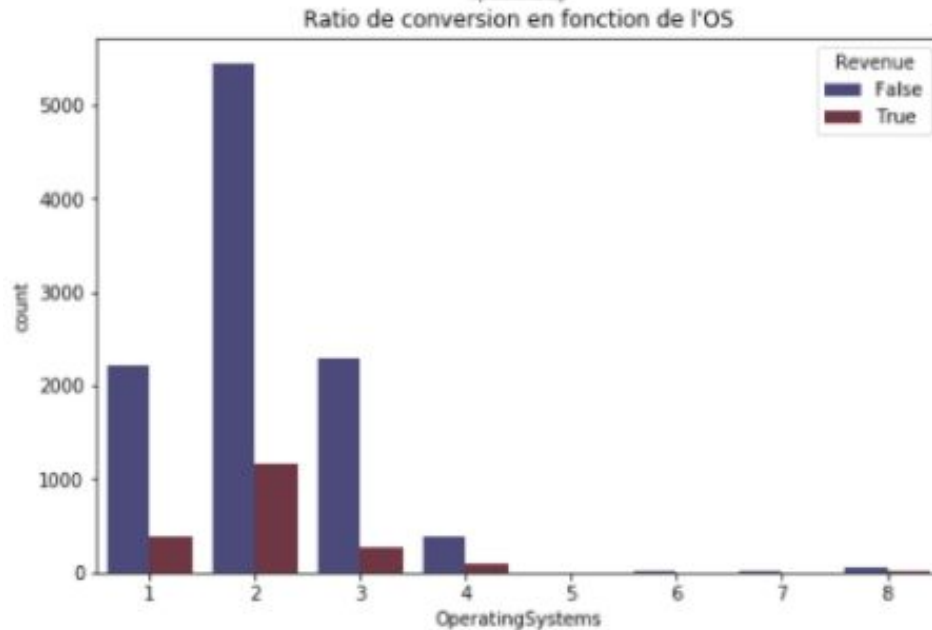
# Exploratory Data Analysis



Gauche : On observe que peu de visites sont réalisées lors des special days, ce qui est logique puisqu'il y a peu de special days dans l'année. Cependant, et ce n'est pas précisé par le dataset, nous nous demandons si les créateurs du dataset ont normalisé l'importance des facteurs SpecialDay en fonction de leur faible nombre dans l'année.

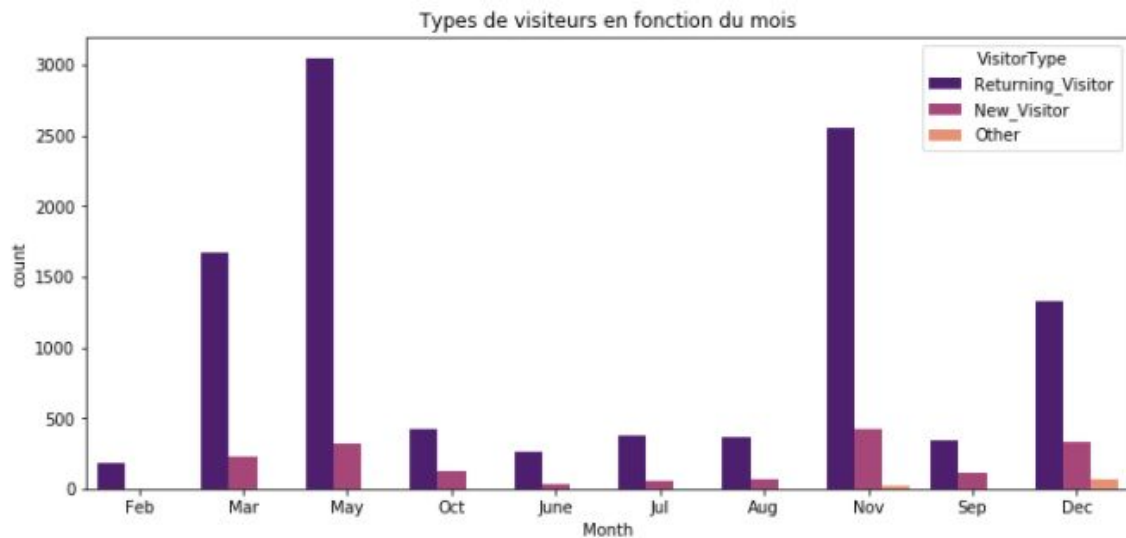
Droite : On observe que le taux de conversion est plus élevé pour les nouveaux visiteurs, qui sont tout de même en bien moins grand nombre que ceux qui reviennent (les habitués du site).

# Exploratory Data Analysis



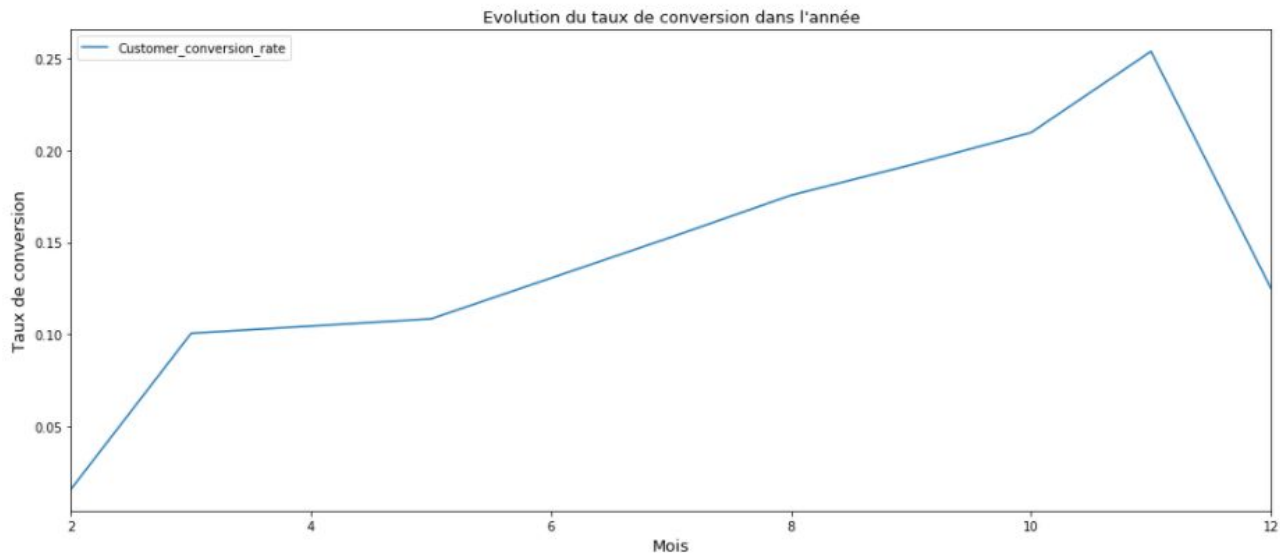
L'OS 2 se démarque des autres dans son nombre d'utilisations. Pas d'information particulière concernant les taux de conversion.

# Exploratory Data Analysis



Les mois de mars, mai, novembre et décembre se démarquent des autres.

# Exploratory Data Analysis

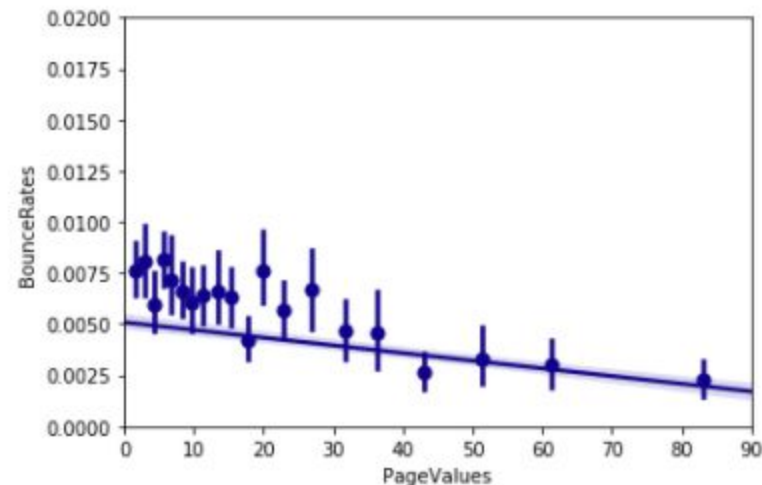
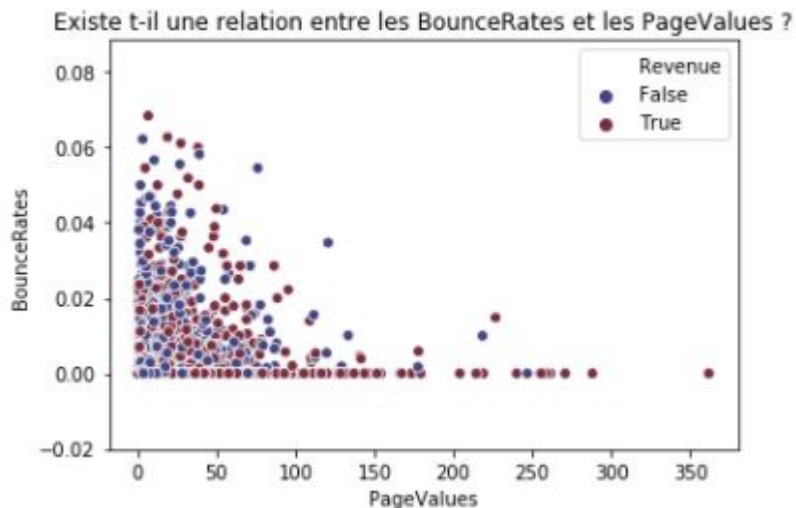


Note : Les mois d'avril et janvier ne sont pas présents dans le dataset.

Le taux de conversion atteint son pic en novembre, ce qui est logique sachant que les fêtes de fin d'année arrivent.

On observe une tendance croissante tout au long de l'année avec une décroissance lors du dernier mois de l'année.

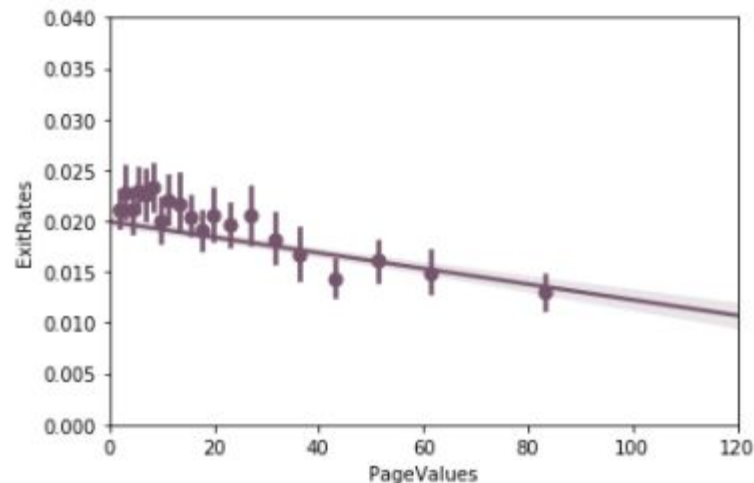
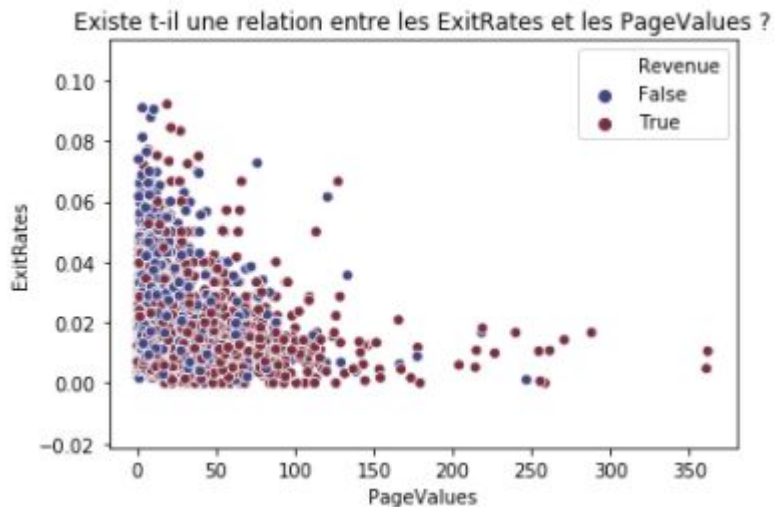
# Exploratory Data Analysis



Une limite (90 en abscisse) a été fixée pour construire le modèle. Cela nous a permis d'éviter les outliers.

On observe que donc l'exit rate diminue en fonction du nombre de PageValues de façon linéaire. Soit, plus l'utilisateur reste connecté, moins son taux d'arrêt de visite est élevé, ce qui est logique et confirme nos opinions préalable.

# Exploratory Data Analysis



Une limite (120 en abscisse) a été fixée pour construire le modèle. Cela nous a permis d'éviter les outliers.

On observe que donc l'exit rate diminue en fonction du nombre de PageValues de façon linéaire. Soit, plus l'utilisateur reste connecté, moins son taux d'arrêt de visite est élevé, ce qui est logique et confirme nos opinions préalable.

# Machine Learning



1. Régression Logistique

2. Arbre de décision

3. Random Forest

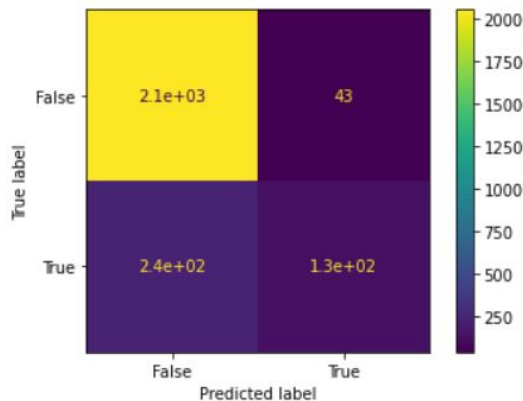
4. Gaussian Naive Bayes

5. Réseau de neurones

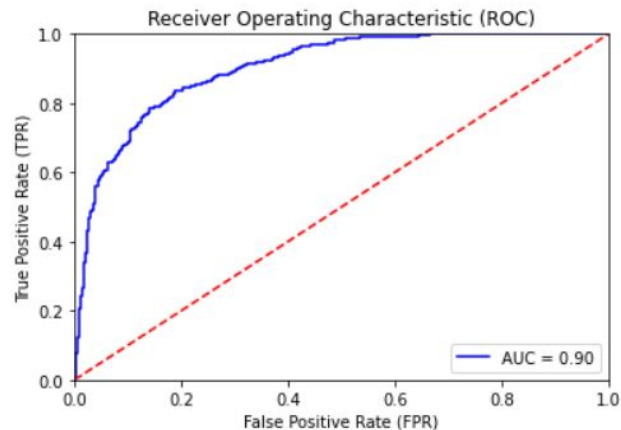
# Régression logistique

Modèle statistique qui utilise une fonction logistique afin d'effectuer une classification binaire.

Confusion Matrix



Courbe ROC

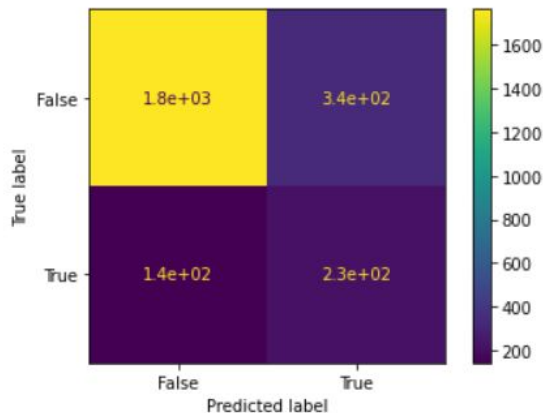




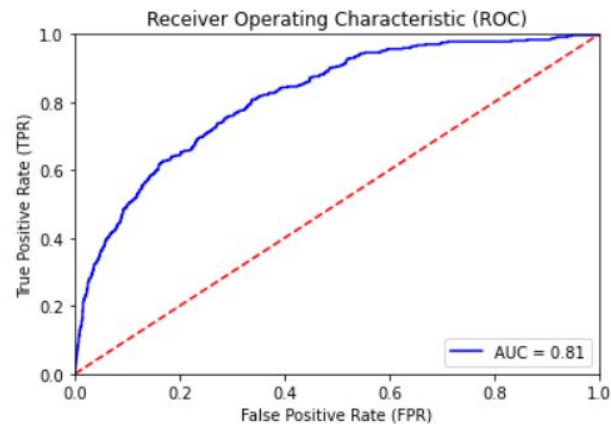
# Gaussian Naive Bayes

La **classification naïve bayésienne** est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des hypothèses.

Confusion Matrix



Courbe ROC



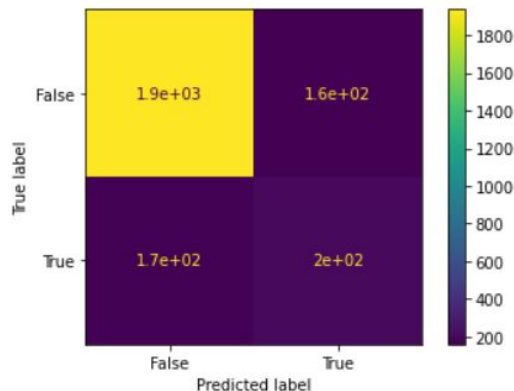
Accuracy : 0.806

# Arbre de Décision



Outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape.

**Confusion Matrix**



**Accuracy : 0.868**

# Random Forest

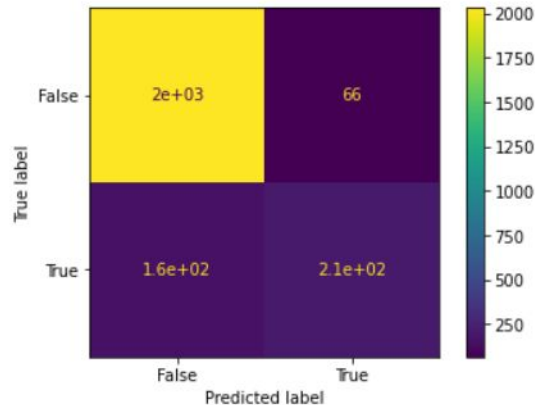


Algorithme de classification qui réduit la variance des prévisions d'un arbre de décision seul, améliorant ainsi leurs performances. Pour cela, il combine de nombreux arbres de décisions dans une approche de type bagging.

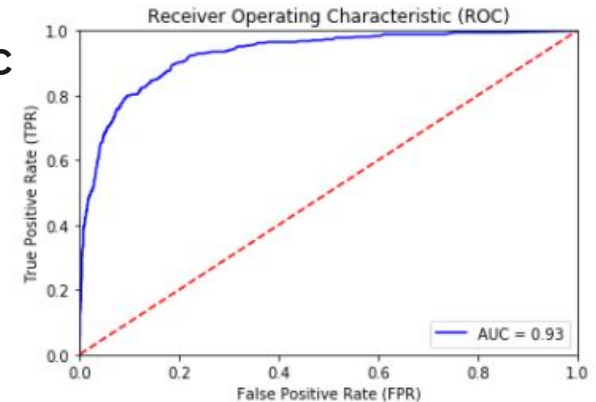
**Grid Search : Optimisation des paramètres du modèle**

**Accuracy : 0.9067**

**Confusion Matrix**



**Courbe ROC**

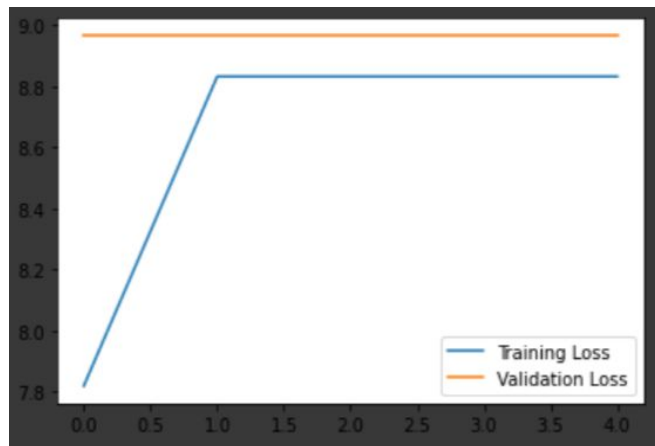
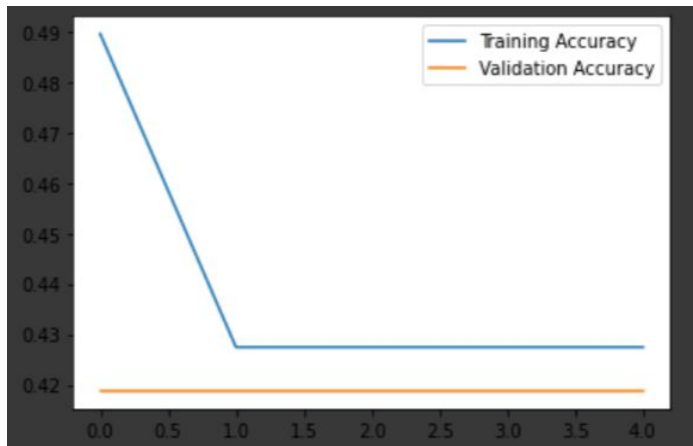


# Réseau de neurones

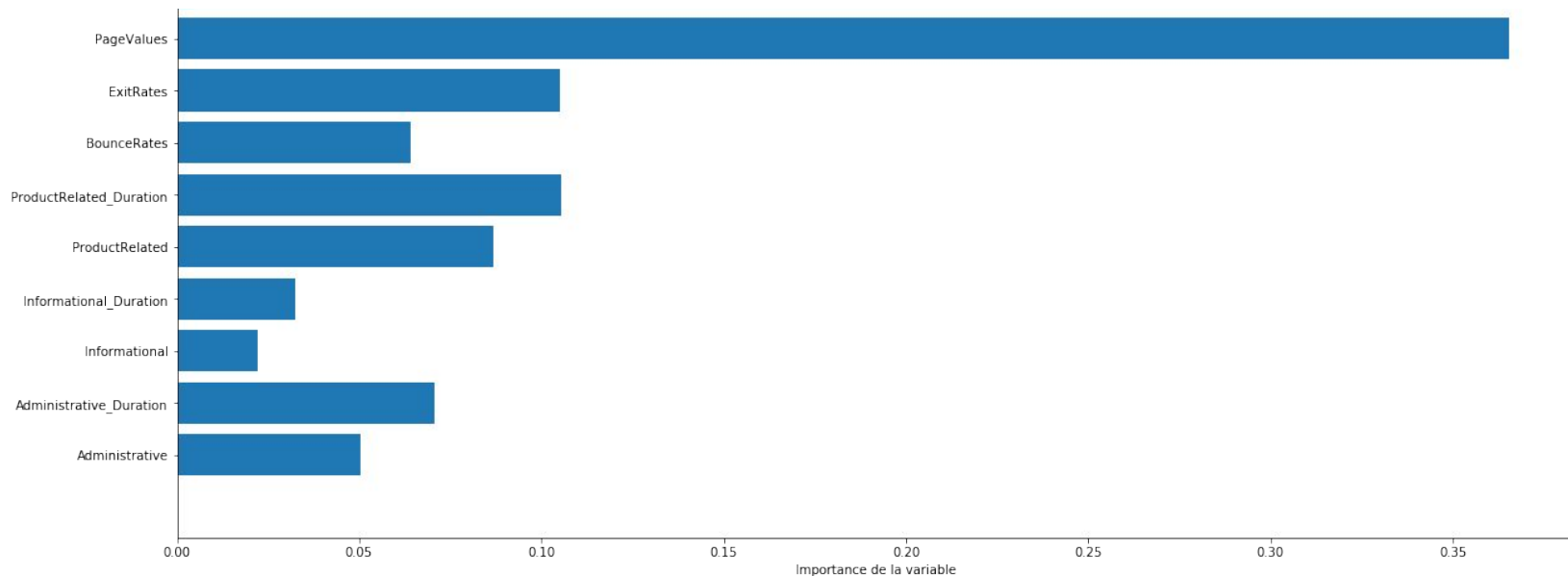
L'utilisation de deep learning n'est pas adaptée à ce problème, notamment à cause du fait de l'**imbalance data**. Les prédictions générées par le modèle vont être presque constantes (proches de 0, puisque Revenu 0 majoritaires dans le dataset) afin de minimiser la fonction d'erreur du modèle.

De ce fait, nous avons créé un dataframe contenant un nombre égal de lignes avec Revenu et sans Revenu. Mais alors, le modèle semble **manquer de données** pour pouvoir être entraîné correctement.

Nous n'avons **pas intégré** le réseau de neurones dans notre API.



# Most Important Features



# API



data test api.json







# Flask

Données supprimées (meilleures performances de modèles):

- OperatingSystems
- Browser
- Region
- TrafficType

```
1  {
2      "Administrative": 4,
3      "Administrative_Duration": 70,
4      "Informational": 0,
5      "Informational_Duration": 0,
6      "ProductRelated": 15,
7      "ProductRelated_Duration": 250,
8      "BounceRates": 0,
9      "ExitRates": 0.4,
10     "PageValues": 7,
11     "SpecialDay": 0,
12     "Weekend": false,
13     "Month_Feb": 0,
14     "Month_Mar": 0,
15     "Month_May": 0,
16     "Month_Jun": 0,
17     "Month_Jul": 1,
18     "Month_Aug": 0,
19     "Month_Sep": 1,
20     "Month_Oct": 0,
21     "Month_Nov": 0,
22     "Month_Dec": 0,
23     "VisitorType_New_Visitor": 1,
24     "VisitorType_Returning_Visitor": 0,
25     "VisitorType_Other": 0
26 }
```



 random_forest.pickle	1/8/2021 4:23 AM	PICKLE File	138,363 KB
 decision_tree.pickle	1/8/2021 4:45 AM	PICKLE File	125 KB
 naive_bayes.pickle	1/8/2021 4:45 AM	PICKLE File	2 KB
 logistic_regression.pickle	1/8/2021 4:42 AM	PICKLE File	1 KB

```
C:\Users\User\Documents\ESILV\A5\Python for data analysis\Project>curl localhost:5000/api -H 'Content-Type:application/json' -d @data_test_api.json
{"response": false}
```

*Modèle par défaut: random\_forest*

```
C:\Users\User\Documents\ESILV\A5\Python for data analysis\Project>curl localhost:5000/api/random_forest -H 'Content-Type:application/json' -d @data_test_api.json
{"response": false}
```

```
C:\Users\User\Documents\ESILV\A5\Python for data analysis\Project>curl localhost:5000/api/logistic_regression -H 'Content-Type:application/json' -d @data_test_api.json
{"response": false}
```

```
C:\Users\User\Documents\ESILV\A5\Python for data analysis\Project>curl localhost:5000/api/naive_bayes -H 'Content-Type:application/json' -d @data_test_api.json
{"response": false}
```