Projet Python for Data Analysis

Les intentions des acheteurs en ligne

Aymeric Stheme de Jubecourt

DIA 5

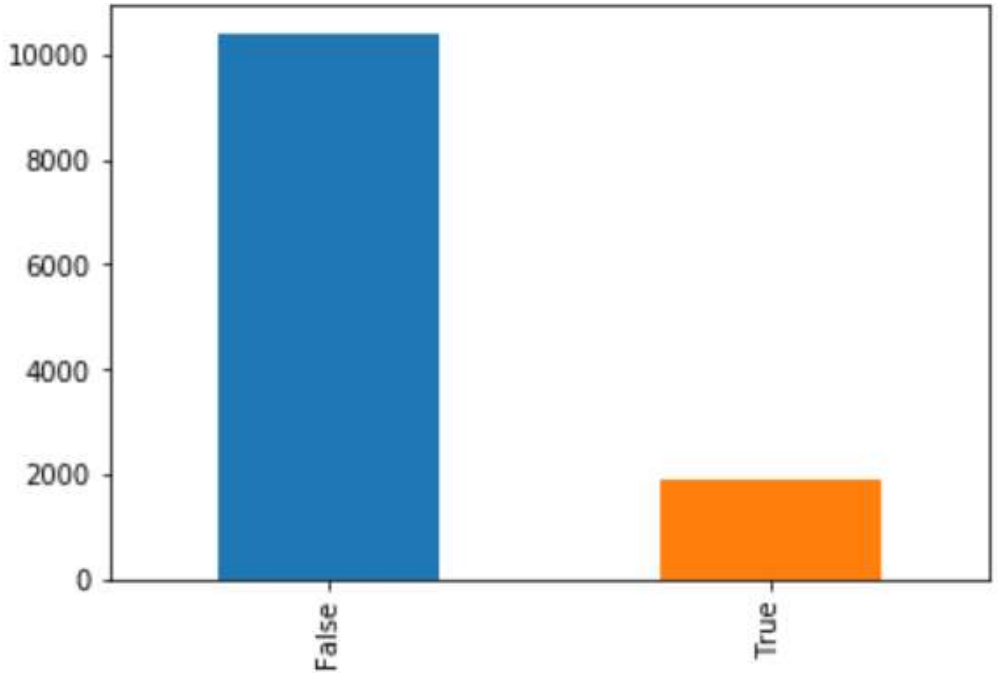
I/ Dataset

- Mon dataset présente l'intention qu'ont les acheteurs en ligne sur les sites webs.
- La première chose que j'ai regardé sur le dataset, ceux sont les variables. En effet, j'ai voulu savoir de quoi parlait mon sujet pour bien comprendre les données.
- Dans un premier temps mon dataset est constitué de 18 colonnes et 12330 lignes
- Le nombre de variables numériques et de 10 et de 8 pour les attributs catégoriels. La variable « Revenue » sera considéré comme une variable de classe

Administrative	int64
Administrative_Duration	float64
Informational	int64
Informational_Duration	float64
ProductRelated	int64
ProductRelated_Duration	float64
BounceRates	float64
ExitRates	float64
PageValues	float64
SpecialDay	float64
Month	object
OperatingSystems	int64
Browser	int64
Region	int64
TrafficType	int64
VisitorType	object
Weekend	bool
Revenue	bool

- Les principales variables que j'ai voulu mettre en avant sont les suivants: Administrative,Informationnal,ProductRelated
- J'ai sélectionné ces variables car ils me semblaient être les plus important et les plus impactants pour notre étude.

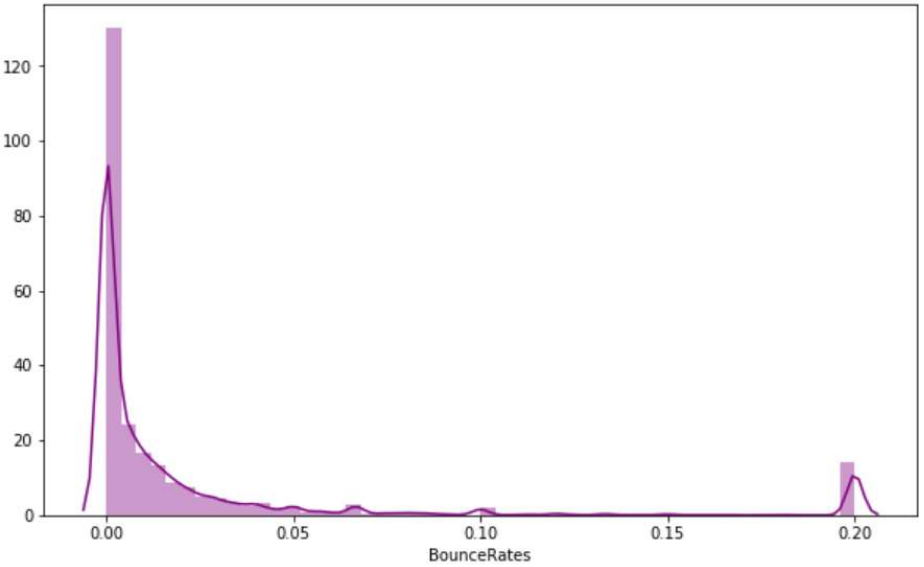
- «Administratve», «Administrative_duration», «Informational», «Informational_Duration», «ProductRelated» et «Product_Duration» représentent le nombre de différents types de pages visitées par le visiteur au cours de cette session et le temps total passé dans chacune des ces catégories de page.
- Les variables "Bounce Rate", "Exit Rate" et "Page Value"représentent les statistiques mesurées par "Google Analytics" pour chaque page du site de commerce électronique.
- Après avoir regardé le détail de la variable « Revenue » on obtient cela:



- On peut donc constater qu'environ 1% des personnes visitant les sites web quittent ces derniers sans acheter de produits.
- On voit aussi qu'il n'y a aucune valeur nulle dans notre dataset

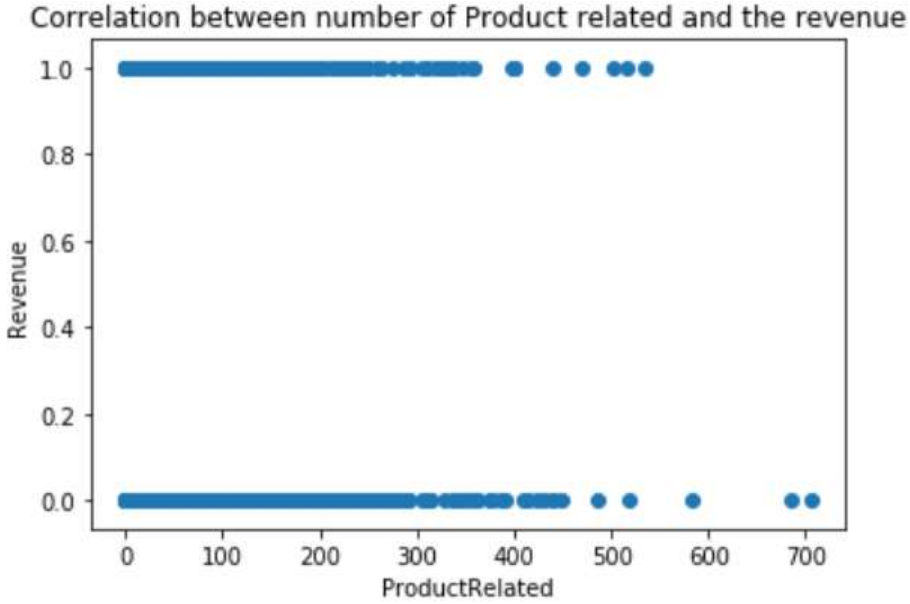
II/ Visualisation

- Pour mieux visualiser mon dataset et pour mieux le comprendre j'ai fait des graphiques afin de voir quels étaient les différents impact des variables.
- J'ai donc regardé le « Bouncerates » pour une page web qui représente le pourcentage de visiteurs qui entrent sur le site à partir de cette page, puis le quittent sans avoir d'autre demandes.



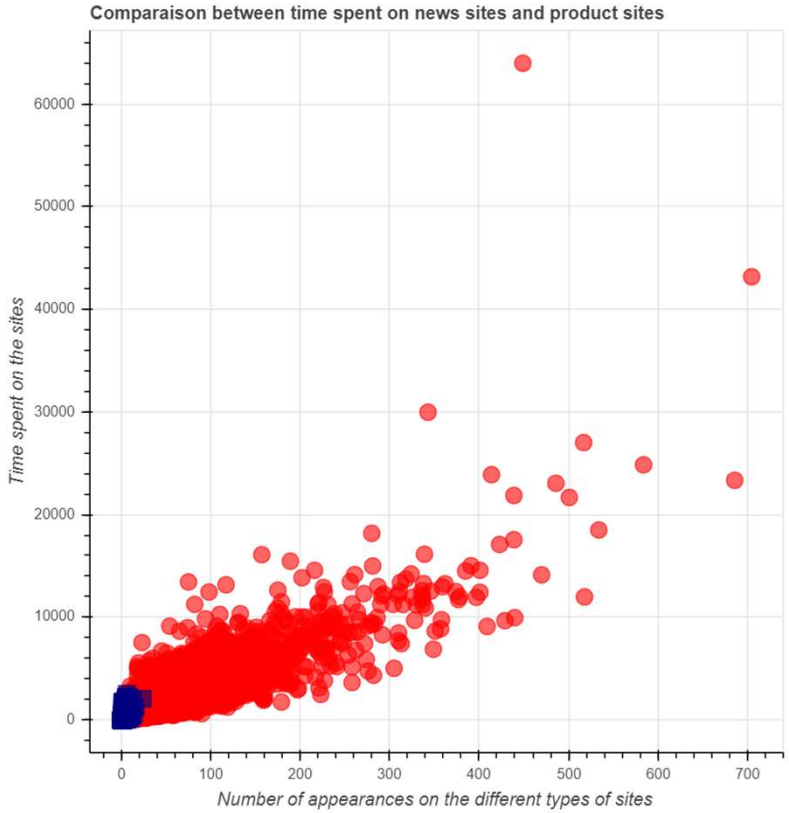
- On peut s'apercevoir que beacoup de personnes visitant une page web y restent pour effectuer d'autres recherches mais que la plupart restaient sur un site web pour qu'une seule recherche.

- Pour le deuxième graphique, je voulais savoir quels étaient la corrélation entre « Revenue » et « productRelated », c'est-à-dire si les acheteurs en lignes dépensaient leur argent dans des produits.



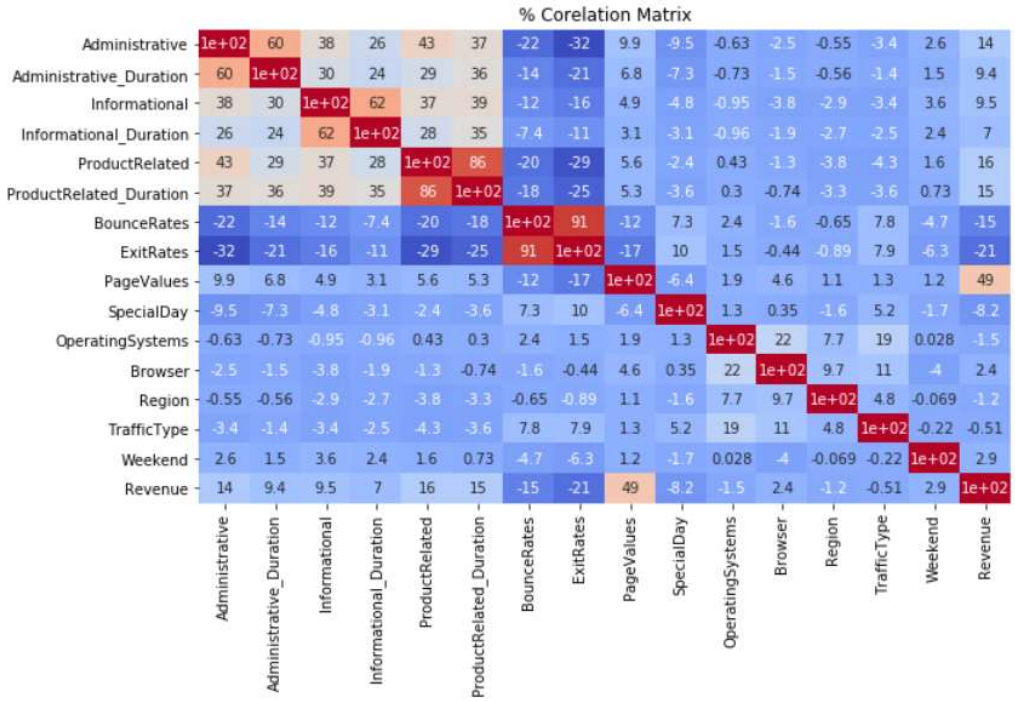
- On peut se rendre compte avec ce graphique, que la plupart des acheteurs en ligne s'intéressent surtout à des pages web sur des produits plutôt qu'à celles sur l'Admistratif ou l'information.
- C'est pour cela aussi que le nombre de produits achetées est très élevés, il est proportionnel aux nombres de personnes visitant des pages web sur des produits.
- Avec cette observation, nous pouvons en déduire que la variable « ProductRelated » peut avoir une influence sur notre variable de classe « Revenue

- Pour continuer les observations, j'ai voulu voir la différence entre le temps passé sur des sites d'information et sur des sites vendant des produits.



- Avec cette observation, on peut en conclure que les acheteurs en ligne passent plus de temps sur les sites de produits et y vont plus souvent.

- Pour finir, j'ai voulu montrer la matrice de corrélation entre les variables car elle m'a permis de voir quelles étaient les variables les plus influentes sur la variable de classe « Revenue ».



- Avec cette matrice de corrélation, on peut voir la corrélation qu'ont les variables entre elles, je me suis intéressais à la corrélation des variables avec notre variable de classe.
- Sans surprise, j'ai vu que les variables « Administrative », « ProductRelated » et « Informational » avait une forte corrélation mais j'ai aussi découvert la variable « PageValues » qui a la plus forte corrélation.

III/ Modélisation

- La première chose que j'ai fait lorsque je suis arrivé au moment de la modélisation c'est le choix des variables.
- En effet, lorsque j'ai voulu modéliser un arbre avec 12330 lignes et 18 colonnes cela a pris beaucoup de temps et en pur perte malheureusement.
- J'ai donc sélectionné les variables les plus impactantes sur notre variable de classe et les plus intéressantes pour les utilisateurs qui vont utiliser ma Web App.
- Pour tous les modèles suivants j'ai donc utilisé ces variables pour X:

  « ProductRelated »

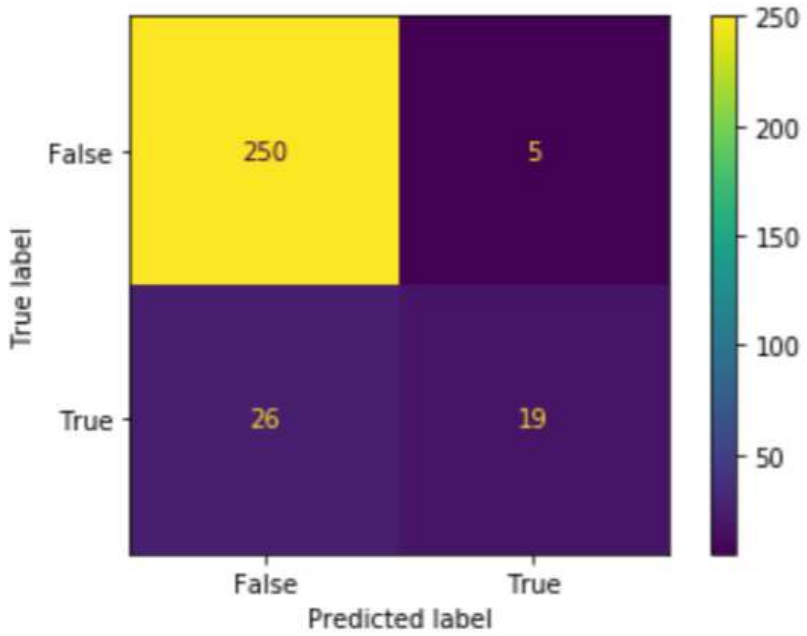
  « Informational »

  « Administrative »

  « PagesValues »
- Et pour y, j'ai donc utilisé notre variable de classe:

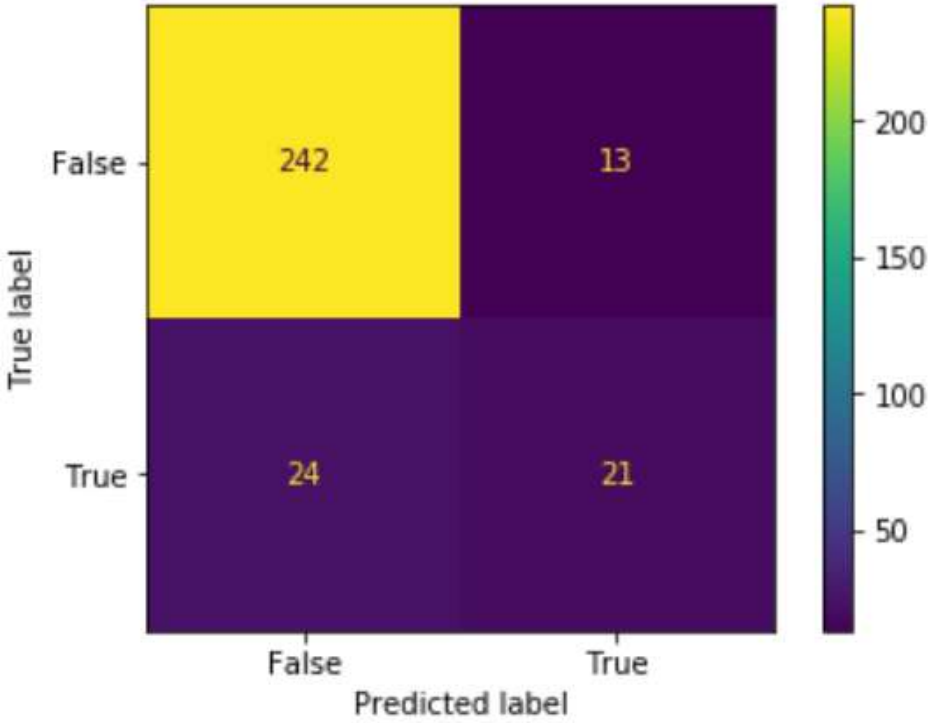
  « Revenue »
- J'ai donc modifié notre dataset pour permettre à l'ordinateur de « fitter » nos modèles.

- Le premier modèle que j'ai utilisé c'est donc une régression linéaire. La première régression linéaire que j'ai fait je n'avais pas mis la variable « PageValues » car je me disais qu'elle n'avait pas sa place pour les utilisateurs de mon API mais au vu de sa corrélation avec la variable « Revenue » je ne pouvais pas ne pas la mettre. Ce fut une bonne idée car l'accuracy score de mes modèles ont tous augmenté en rajoutant cette variable, celui de la régression est de : 0.896666. De plus, ce modèle a pour meilleur hyper-paramètre C=0.1
- Après le premier modèle j'ai décidé de montrer la matrice de confusion de ce modèle:



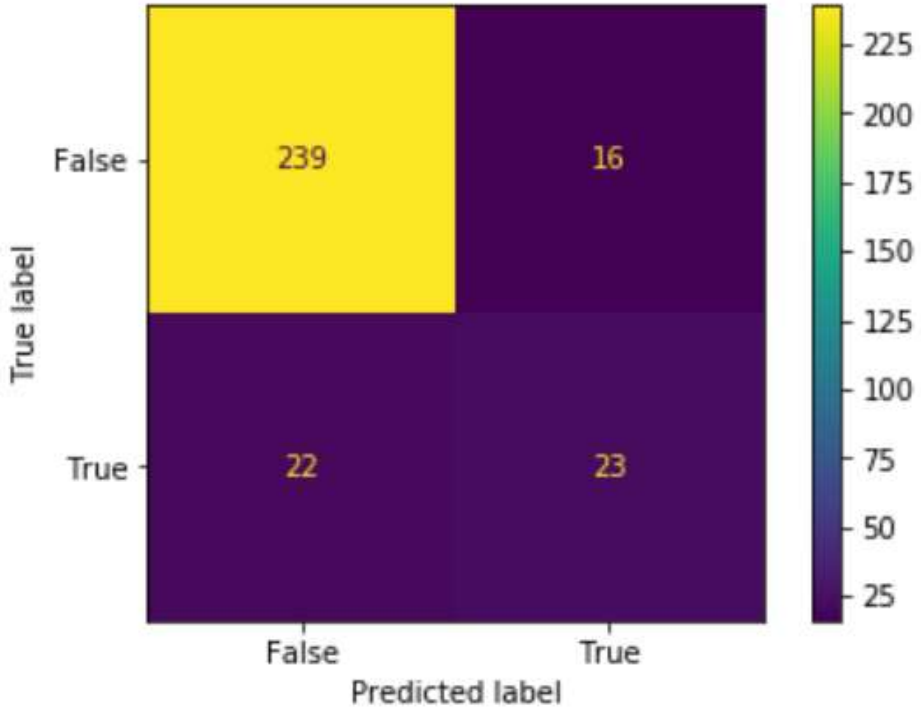
- On peut voir sur cette matrice il y a eu 250 bonnes prédictions « False » et 19 bonnes prédictions « True » et seulement 26 mauvaises prédictions « False » et 5 mauvaises prédictions « True ».

- Pour le deuxième modèle, j'ai utilisé un knn pour voir si son accuracy score sera plus élevé. Le knn a obtenu comme meilleur score 0.8851;
- J'ai pu constater que ces meilleurs paramètres étaient de 50 concernant la liste que j'ai implanté pour le nombres de voisins et « uniform » pour l'hyper paramètre « poids ».



- On peut voir que comme sur le modèle précédent, les prédictions sont plutôt positives car nous avons 263 bonnes prédictions et 37 mauvaises prédictions.

- J’ai aussi effectuer un modèle svm mais il m’a donné le même accuracy score que le modèle knn et, à 2 prédictions prêts, la même matrice de corrélation.
- Je me suis donc orienté vers une grille ou arbre de recherche, j’ai pu remarquer que son accuracy score était 0.873333 et que son meilleur accuracy score était de 0.89733. J’ai constaté que ce meilleur accuracy score est arrivé lorsque le « max_depth » donc la profondeur la plus efficace pour ce modèle est de 4 et que le n_estimators était égale à 100 (j’ai pu vérifier cette information grâce à la fonction best_param).
- Et ce modèle m’a donné comme matrice de confusion :



IV/ Web App

- Concernant ma Web App, j'ai voulu mettre comme variable la valeur des pages car elle avait une influence trop importante sur le modèle et donc sur la prédiction.
- C'est pour cela que je vous mets une explication de ce qu'est la valeur d'une page web: la valeur de la page correspond à la valeur moyenne d'une page qu'un utilisateur a visitée avant d'accéder à la page objectif ou d'effectuer une transaction e-commerce(ou les deux).
- Voila pourquoi cette variable a une grande importance, car cela change tout à votre prédiction, je vous mets un exemple de ma Web App :