جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

Prédiction des Décisions d'Achat chez les Consommateurs en Ligne en Utilisant des Techniques de Data Mining

Professeur
Mohamed Sabiri

Réalisé par:
BEN OUAKRIM Ikrame

Module:
Data Mining et Statistique Décisionnelle

Plan

01
Introduction
et
problématique

02
Collection
des Données

03
Visualisation
des Données

04
Analyse
Exploratoire
des Données

05
Prétraitement
Des Données

06
Modélisation

07
Interprétation
des Résultats

08
Conclusion

01

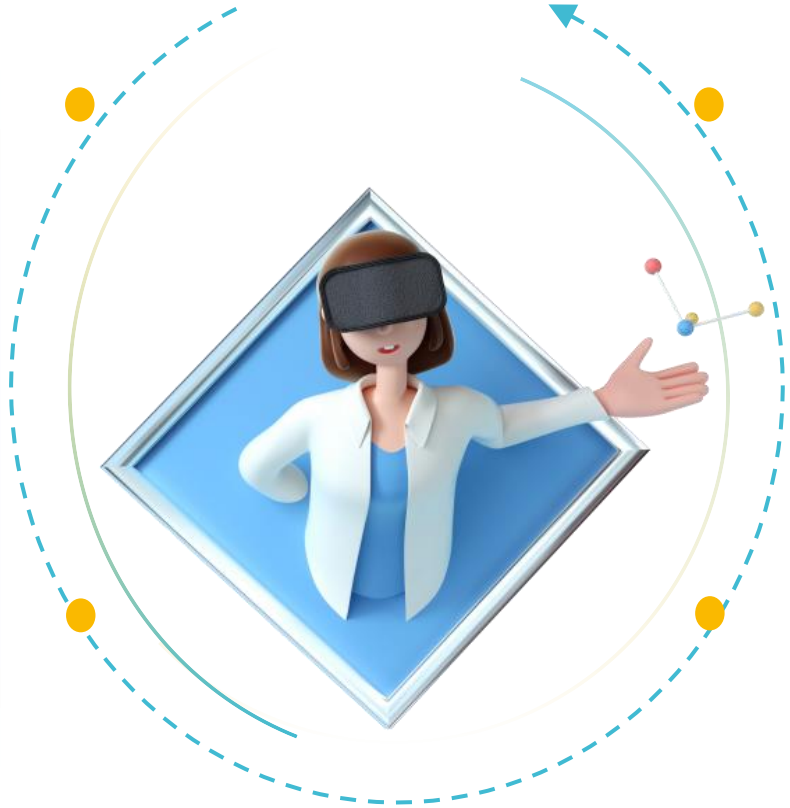
Introduction & Problématique

Introduction & Problématique

Contexte

L'évolution rapide du commerce électronique a profondément modifié la manière dont les consommateurs interagissent avec les produits et services.

À mesure que les transactions en ligne deviennent de plus en plus courantes, la capacité des entreprises à comprendre et à anticiper les comportements d'achat des visiteurs de leurs sites web devient cruciale.



Objectif du Projet

L'objectif principal de ce projet est de développer un modèle de classification utilisant des techniques de data mining pour prédire l'intention d'achat des visiteurs en ligne.

L'objectif ultime est d'offrir aux entreprises un outil précieux pour personnaliser leurs interactions avec les utilisateurs, améliorer l'efficacité des campagnes marketing, et optimiser la conversion des visiteurs en clients.

Introduction & Problématique

Enjeux et Défis

La complexité réside dans la gestion d'un ensemble diversifié de données comportementales, englobant des aspects tels que les pages consultées, la durée passée sur le site, et d'autres variables liées au comportement des utilisateurs. Tous cela pose des défis substantiels pour prédire de manière précise les intentions d'achat.

Signification et Implications

En surmontant ces défis, ce projet a des implications significatives pour les entreprises en ligne.

La capacité à anticiper les décisions d'achat des consommateurs permettra aux entreprises d'améliorer leur expérience client, d'ajuster leurs stratégies marketing en temps réel, et de maximiser leur potentiel de conversion.

02

Collection des données

Collection des données

Le jeu de données utilisé est le "Online Shoppers Purchasing Intention Dataset", disponible sur Kaggle. Il comprend diverses caractéristiques qui peuvent potentiellement influencer les décisions d'achat.
L'ensemble de données contient 12 330 enregistrements et 18 caractéristiques (17 variables d'entrée et 1 variable de sortie) avec des détails comme indiqué ci-dessous :

```
df.shape

(12330, 18)
```

Variables d'Entrée :

1. **Administrative :** Il s'agit du nombre de pages de ce type (administratif) que l'utilisateur a visitées.
2. **Administrative_Duration :** Il s'agit de la durée passée dans cette catégorie de pages.
3. **Informationnel :** Il s'agit du nombre de pages de ce type (informatif) que l'utilisateur a visitées.
4. **Informationnel_Duration :** Il s'agit de la durée passée dans cette catégorie de pages.
5. **ProductRelated :** Il s'agit du nombre de pages de ce type (liées au produit) que l'utilisateur a visitées.
6. **ProductRelated_Duration :** Il s'agit de la durée passée dans cette catégorie de pages.
7. **BounceRates :** Le pourcentage de visiteurs qui entrent sur le site web par cette page et sortent sans déclencher de tâches supplémentaires.
8. **ExitRates :** Le pourcentage de pages vues sur le site web qui se terminent sur cette page spécifique.
9. **PageValues :** La valeur moyenne de la page, calculée sur la valeur de la page cible et/ou l'achèvement d'une transaction de commerce électronique.

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.00	0.10	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.05	0.14	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.02	0.05	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

Collection des données

L'ensemble de données contient 12 330 enregistrements et 18 caractéristiques (17 variables d'entrée et 1 variable de sortie) avec des détails comme indiqué ci-dessous :

Variables d'Entrée :

10. **SpecialDay :** Cette valeur représente la proximité de la date de navigation par rapport aux jours spéciaux ou aux vacances (par exemple, la fête des mères ou la Saint-Valentin) où la transaction est plus susceptible d'être finalisée.
11. **Month :** Contient le mois où la page a été consultée, sous forme de chaîne de caractères.
12. **OperatingSystems :** Une valeur entière représentant le système d'exploitation sur lequel l'utilisateur était lorsqu'il consultait la page.
13. **Browser :** Une valeur entière représentant le navigateur que l'utilisateur utilisait pour consulter la page.
14. **Region :** Une valeur entière représentant la région dans laquelle l'utilisateur se trouve.
15. **TrafficType :** Une valeur entière représentant le type de trafic auquel l'utilisateur est catégorisé.
16. **VisitorType :** Une chaîne de caractères représentant si un visiteur est un Nouveau Visiteur, un Visiteur Régulier ou Autre.
17. **Weekend :** Un booléen représentant si la session a lieu pendant le week-end.

Variable de Sortie :

18. **Revenue :** Un booléen représentant si l'utilisateur a effectué ou non un achat.

```
df.shape

(12330, 18)
```

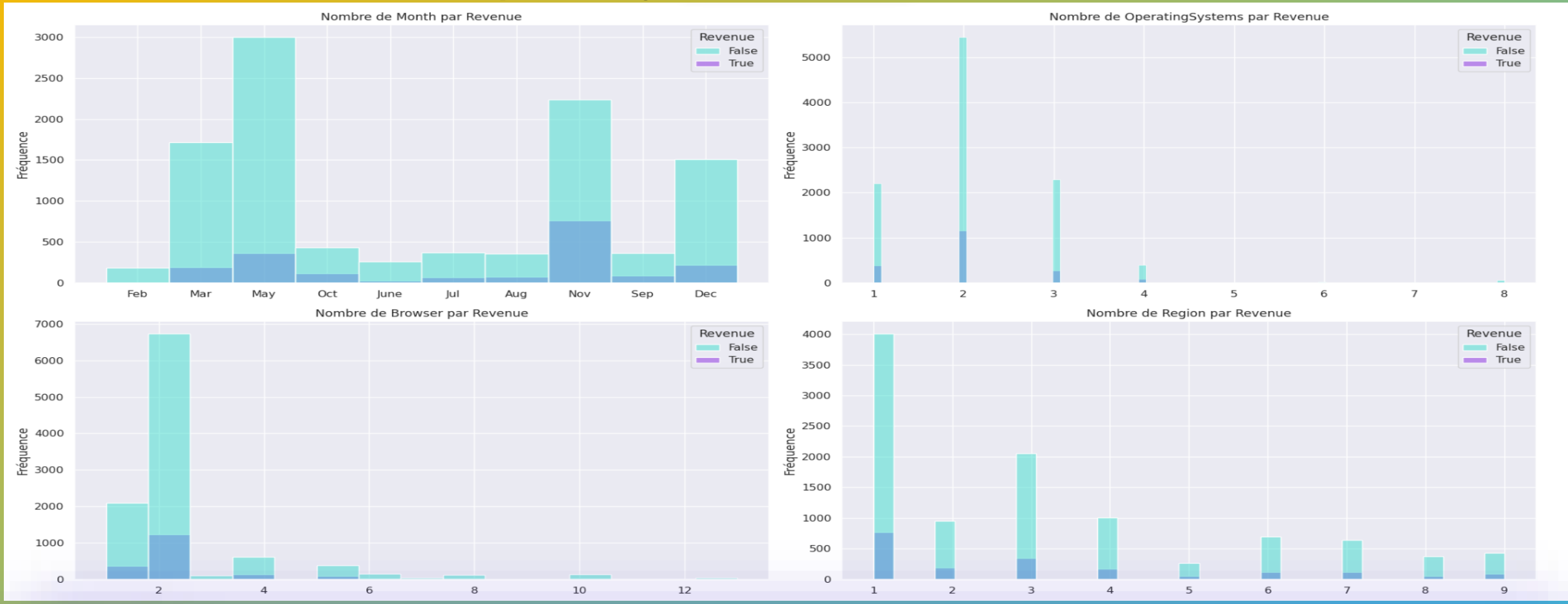
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.00	0.10	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.20	0.20	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.05	0.14	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.02	0.05	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

03

Visualisation des Données

Visualisation des Données

Visualisation des caractéristiques catégorielles



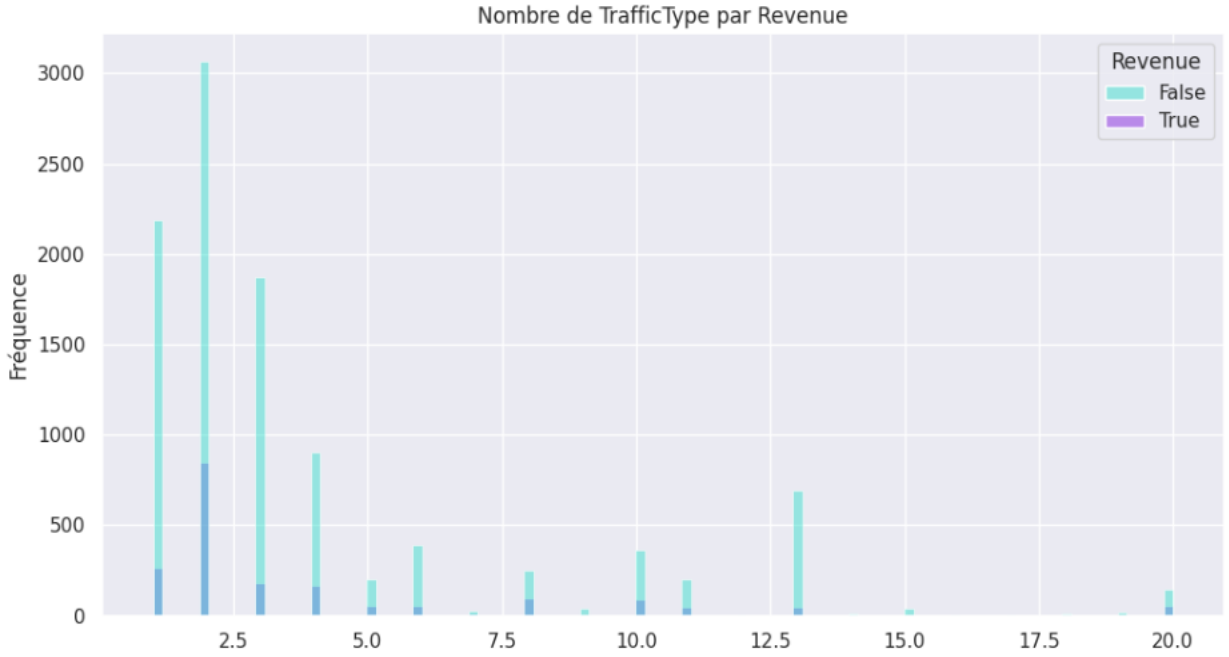
Conclusions:

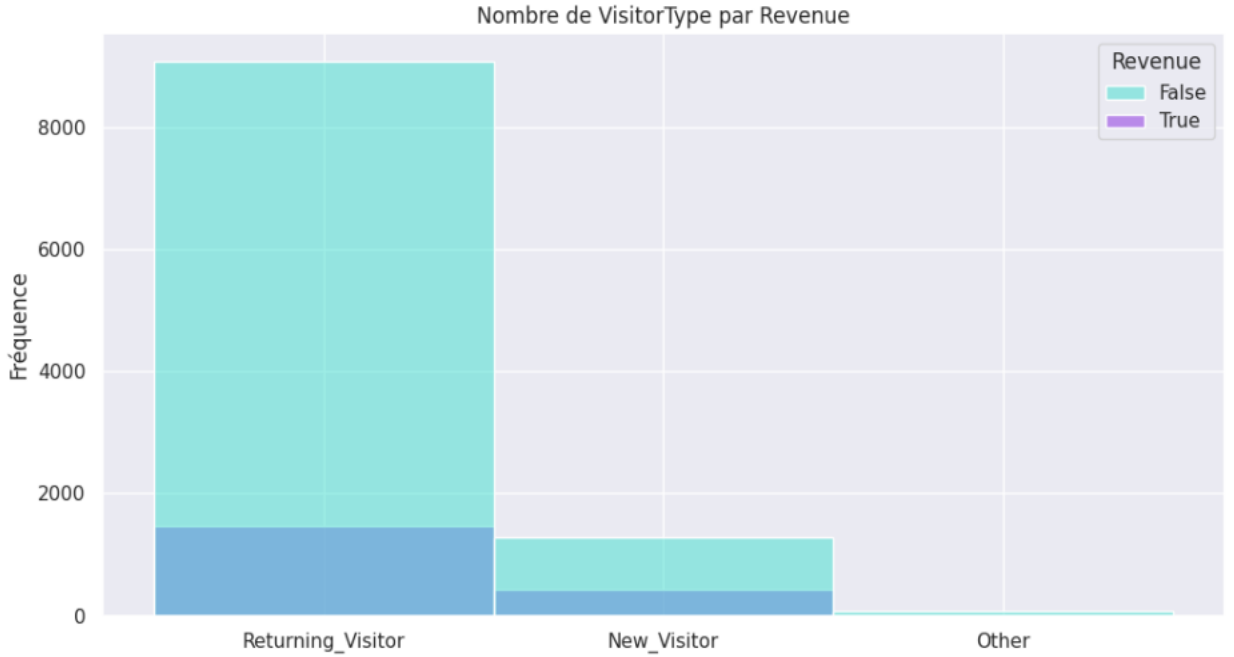
•**Month**: il semble y avoir une tendance saisonnière, certains mois comme novembre et mai ayant des fréquences d'achat plus élevées. Cela pourrait être dû aux périodes de magasinage des fêtes.

•**Operating System**: certains systèmes d'exploitation ont des fréquences d'achat plus élevées, ce qui peut indiquer une préférence de l'utilisateur ou une meilleure expérience utilisateur sur ces systèmes.

Visualisation des Données

Visualisation des caractéristiques catégorielles





Conclusions:

•**Browser**:Semblable au système d'exploitation, certains navigateurs affichent une fréquence d'achats plus élevée. Cela peut être dû à la compatibilité ou à la facilité de transaction sur certains navigateurs.

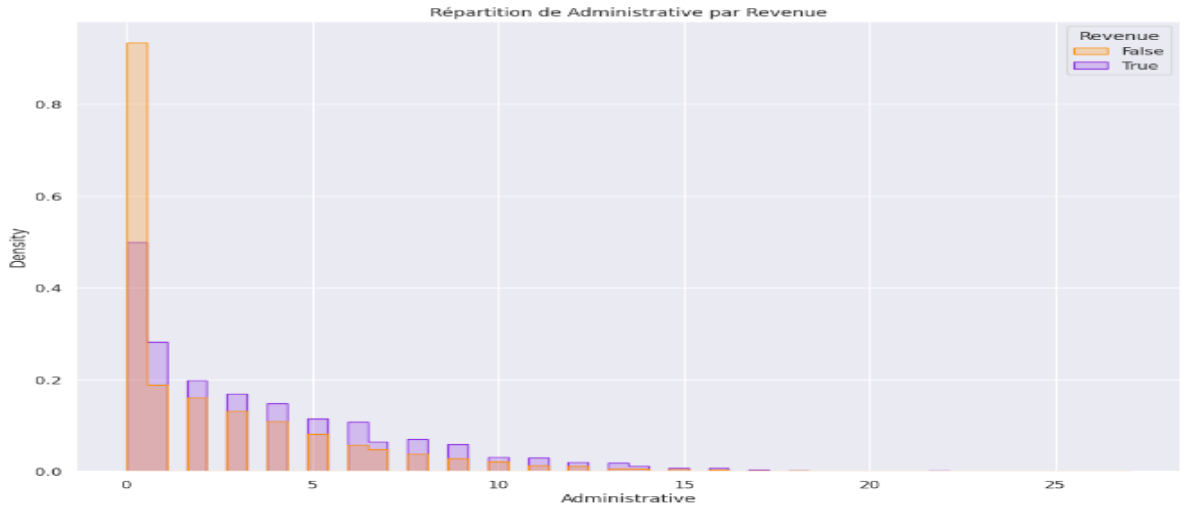
•**Region**: certaines régions affichent un comportement d'achat plus élevé. Cela pourrait être influencé par les promotions régionales ou les options d'expédition.

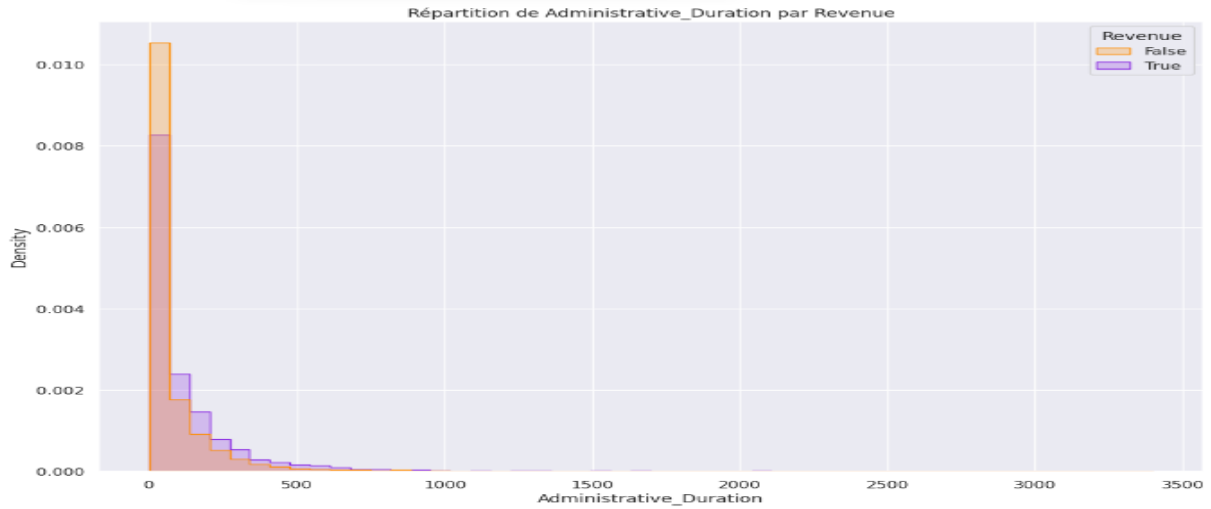
•**Traffic Type**: certains types de trafic vers le site sont plus susceptibles d'être convertis en achats. Par exemple, le trafic direct ou le trafic provenant de certaines campagnes marketing peuvent entraîner des conversions plus élevées.

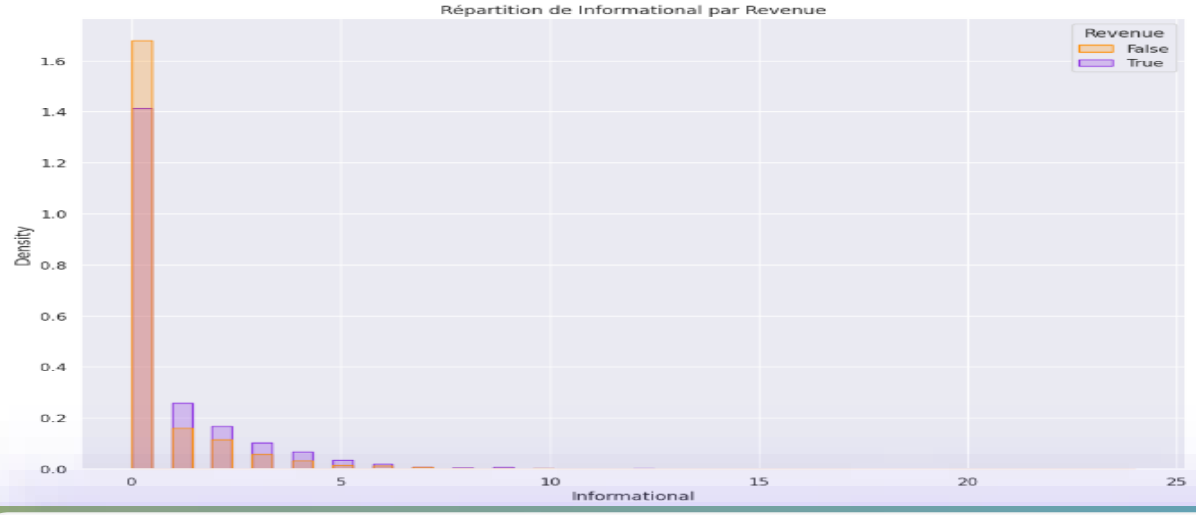
•**Visitor Type**: les visiteurs connus sont plus susceptibles d'effectuer un achat que les nouveaux visiteurs, ce qui indique l'importance de la fidélisation de la clientèle et la valeur potentielle du ciblage des clients réguliers.

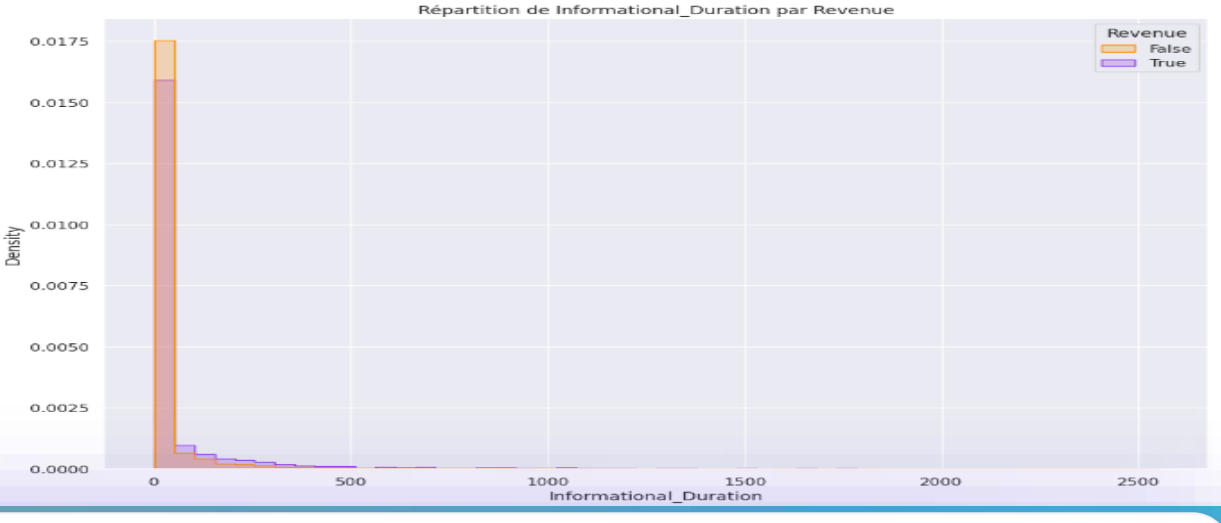
Visualisation des Données

Visualisation des caractéristiques numériques







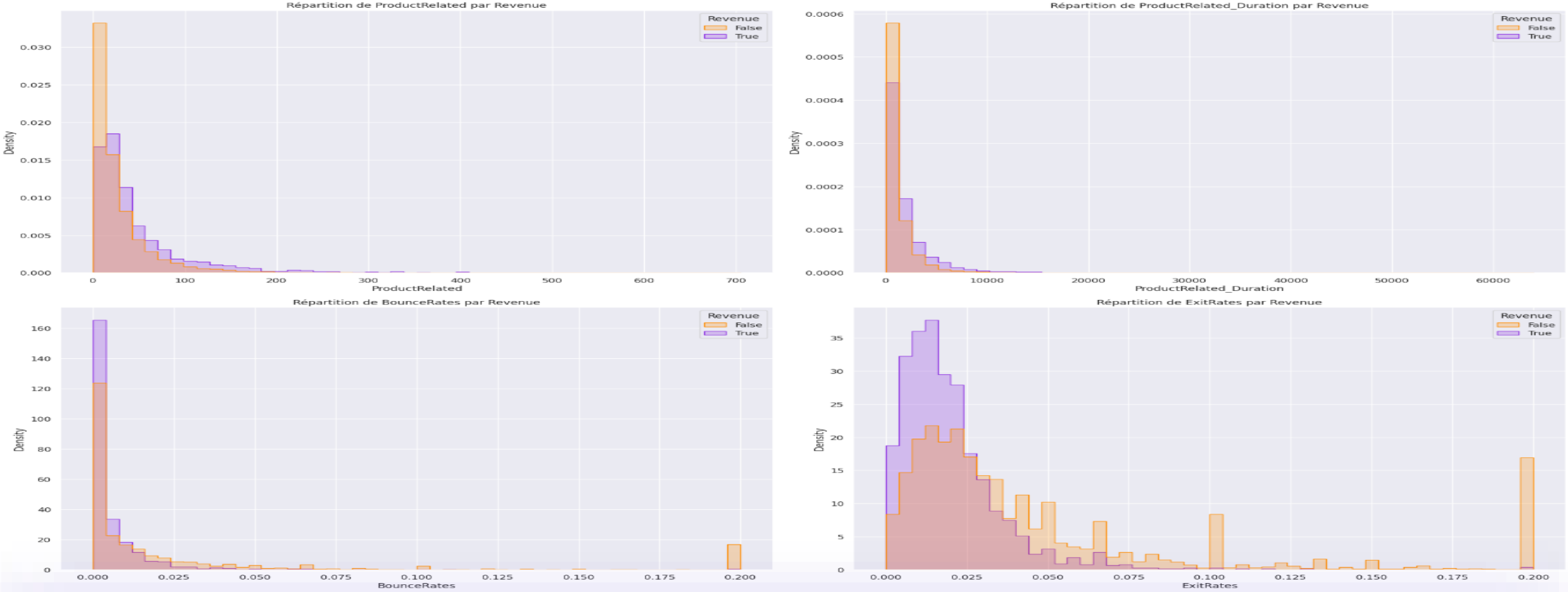


Les histogrammes divisés par la variable cible Revenue nous donnent un aperçu de la manière dont les différents comportements des utilisateurs et interactions sur le site sont associés à la probabilité qu'une session aboutisse à une transaction.

•**Administrative, Informational, Product Related**: le nombre de pages visitées dans ces catégories est généralement plus élevé pour les sessions qui génèrent des revenus. Cela suggère que l'interaction avec le contenu de ces différents types de pages est positivement associée à la probabilité d'une transaction.

Visualisation des Données

Visualisation des caractéristiques numériques

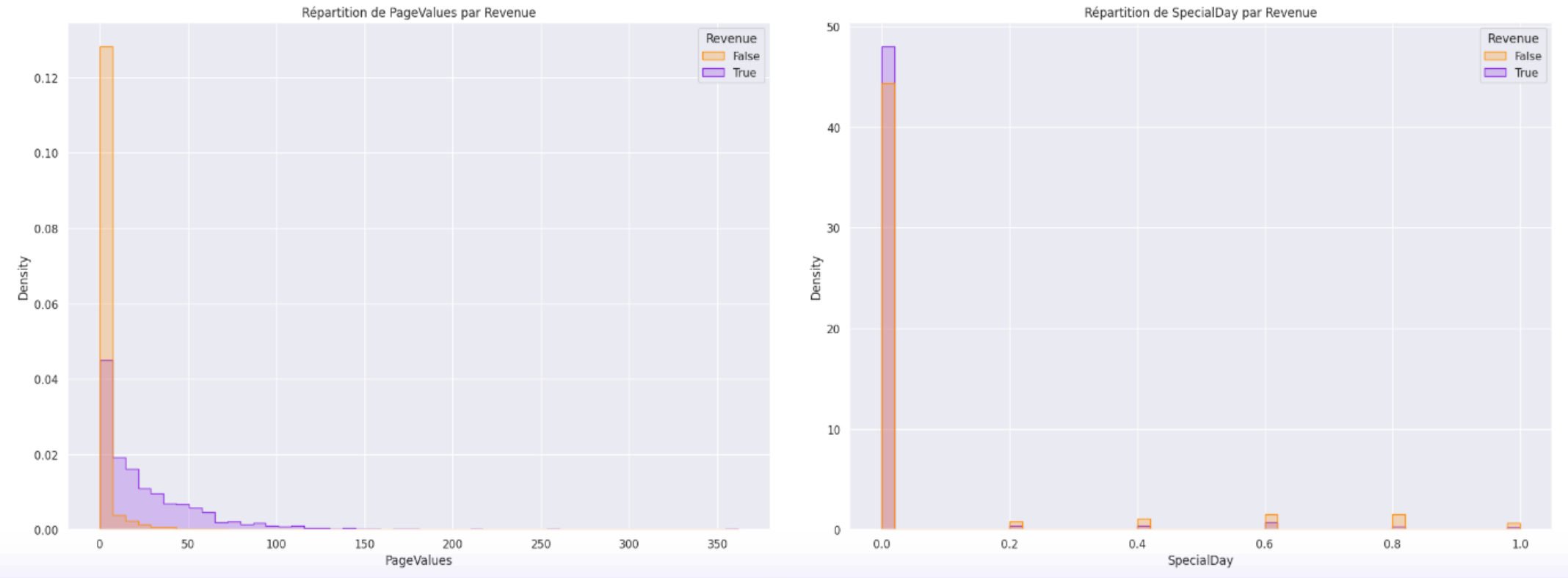


•L**Administrative Duration, Informational Duration, Product Related Duration**: de même, le temps passé sur ces types de pages est plus élevé pour les sessions génératrices de revenus. Cela implique que non seulement le nombre de pages, mais également la profondeur de l'engagement (mesurée par le temps) sont en corrélation avec la probabilité de transaction.

•**Bounce Rate**: les sessions qui n'ont pas généré de revenus ont tendance à avoir un taux de rebond plus élevé. Un taux de rebond élevé implique que les utilisateurs quittent le site après avoir consulté une seule page, ce qui est associé négativement aux transactions.

Visualisation des Données

Visualisation des caractéristiques numériques



•**Exit Rate**: les sessions sans revenus ont des taux de sortie plus élevés, ce qui indique que les utilisateurs sont plus susceptibles de quitter le site à partir d'une page donnée. Un taux de sortie plus faible lors des sessions génératrices de revenus suggère que les utilisateurs parcourent davantage de pages avant de terminer leur session, ce qui est un signe positif d'engagement.

•**Page Value**: il existe une distinction claire dans la répartition de la valeur des pages entre les sessions payantes et non payantes. Les pages visitées avant une transaction ont des valeurs de page plus élevées, ce qui montre leur importance dans le processus de conversion.

•**Special Day**: l'histogramme montre que la présence de revenus est plus probable lorsque la valeur du « Jour spécial » est plus proche de zéro, ce qui implique que les jours normaux peuvent entraîner plus de transactions que les jours spéciaux, ce qui pourrait être contre-intuitif et justifier une enquête plus approfondie.

04

Analyse Exploratoire des Données

Analyse Exploratoire des Données

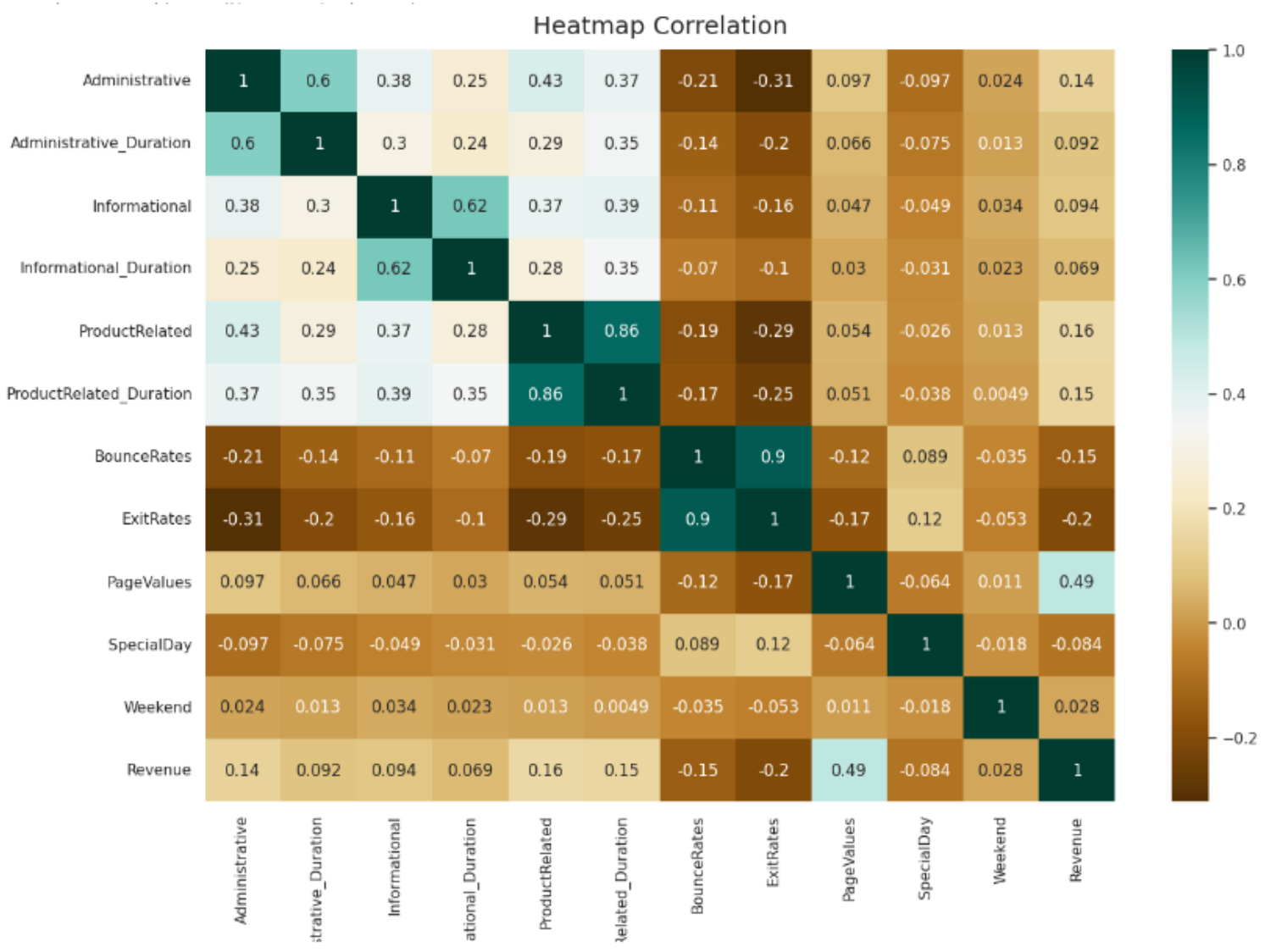
La matrice de corrélation montre qu'il existe peu de corrélation entre les différentes fonctionnalités, à l'exception des éléments suivants :

- Forte corrélation entre :

- BounceRates & ExitRates (0.91).
- ProductRelated & ProductRelated_Duration (0.86).

- Corrélations modérées:

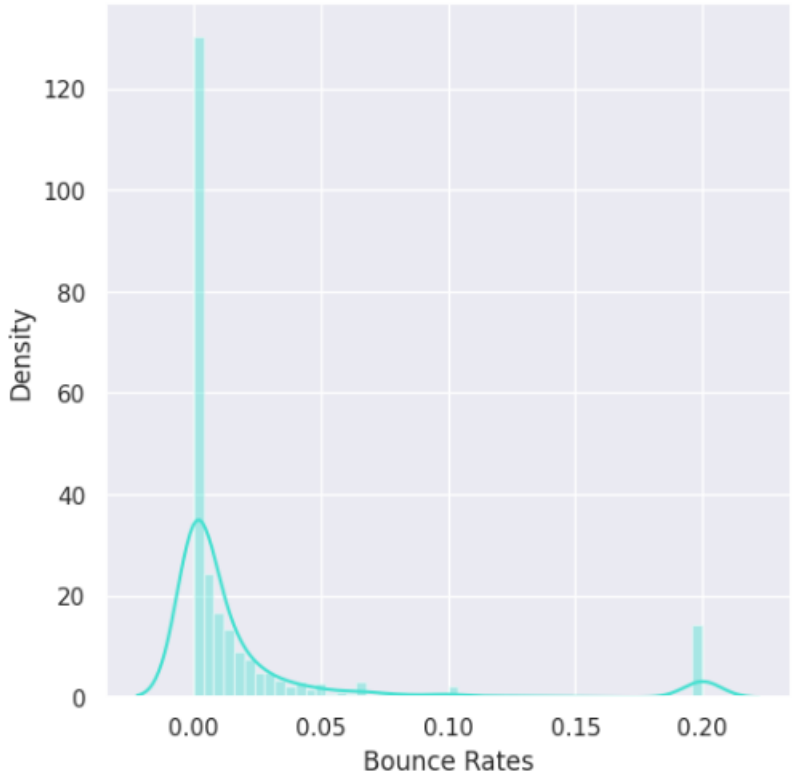
- Administrative & Administrative DUration (0.6)
- Informational & Informational Duration (0.62)
- Page Values & Revenue (0.49)

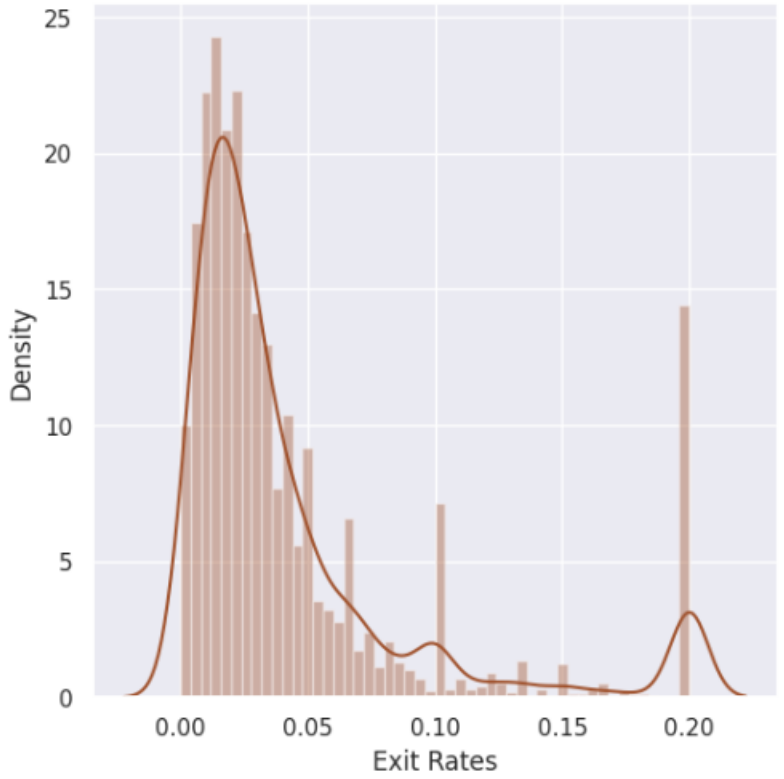


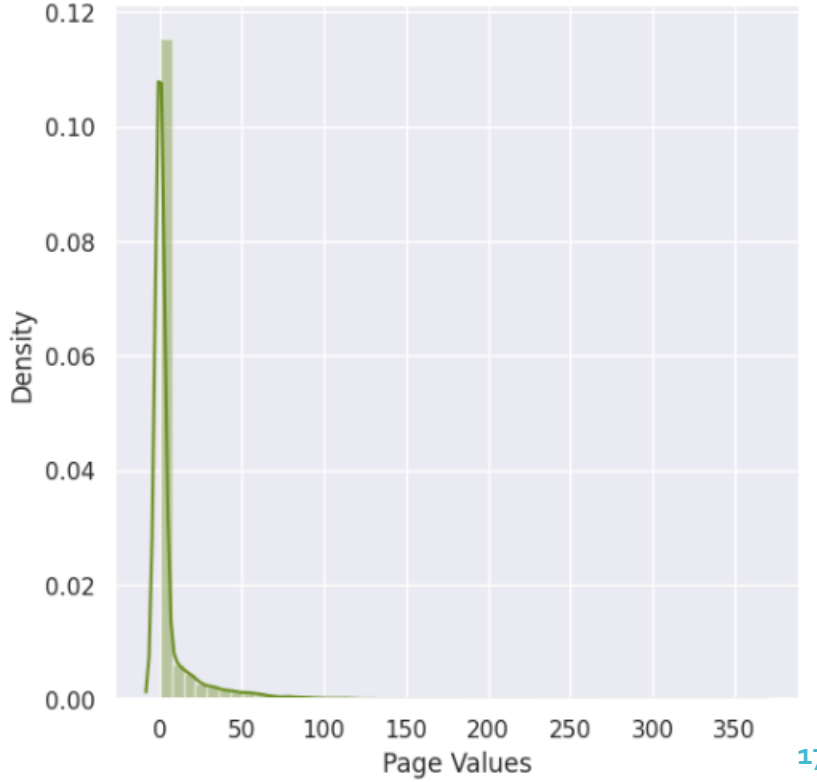
Analyse Exploratoire des Données

Les diagrammes de distribution ci-dessus des métriques de page montrent les éléments suivants :

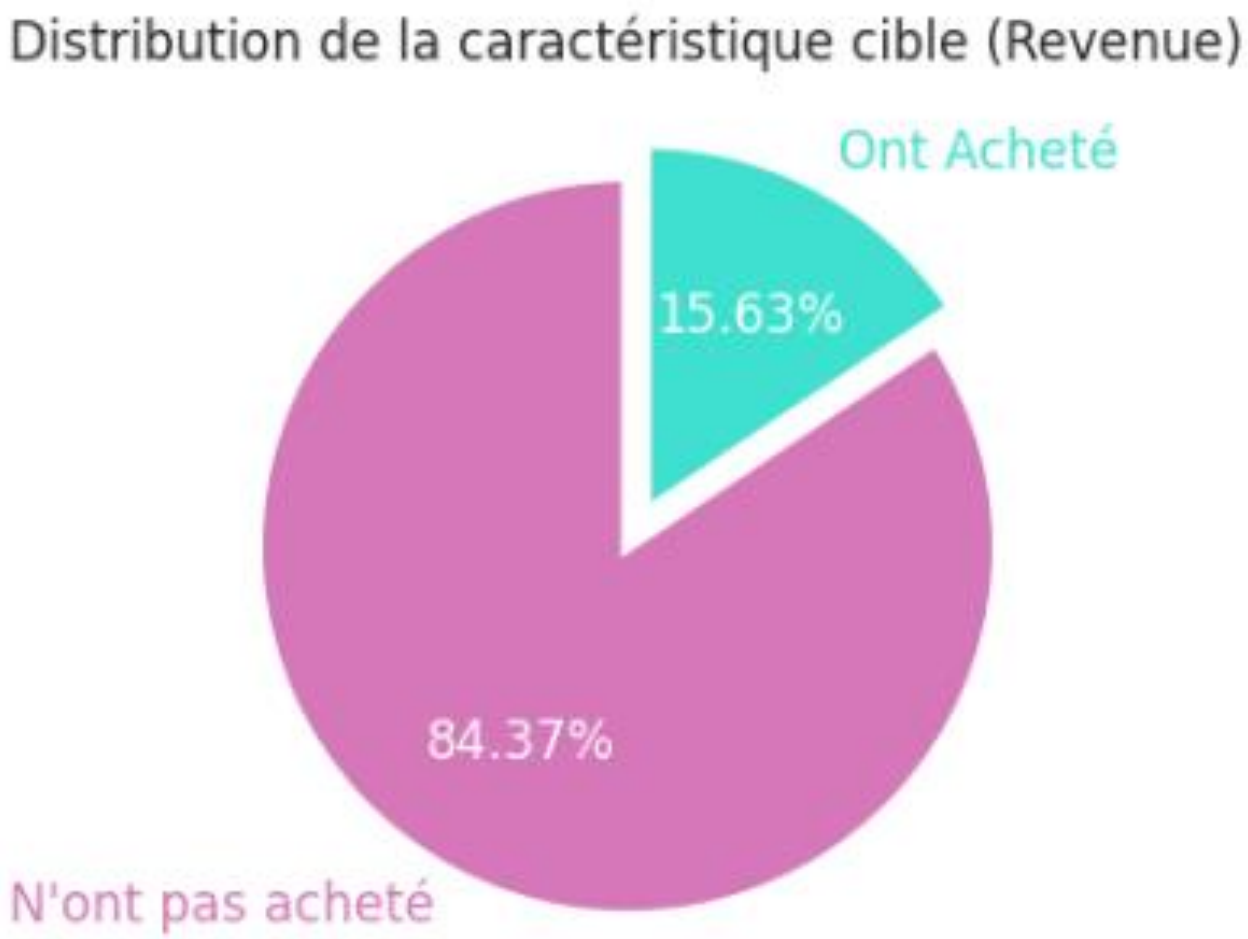
- Les 3 caractéristiques ont des distributions très asymétriques avec de nombreuses valeurs aberrantes.
- Le taux de **Bounce Rates** dans la plupart de nos points de données est faible. Il s'agit d'une observation positive, car des taux élevés indiqueraient que les visiteurs n'interagissent pas avec le site Web.
- **Exit rates** sont plus élevés en valeurs que **Bounce Rates**. Ceci est attendu car nous pouvons supposer que les pages de confirmation de transaction entraîneront une augmentation du taux de sortie moyen.







Analyse Exploratoire des Données



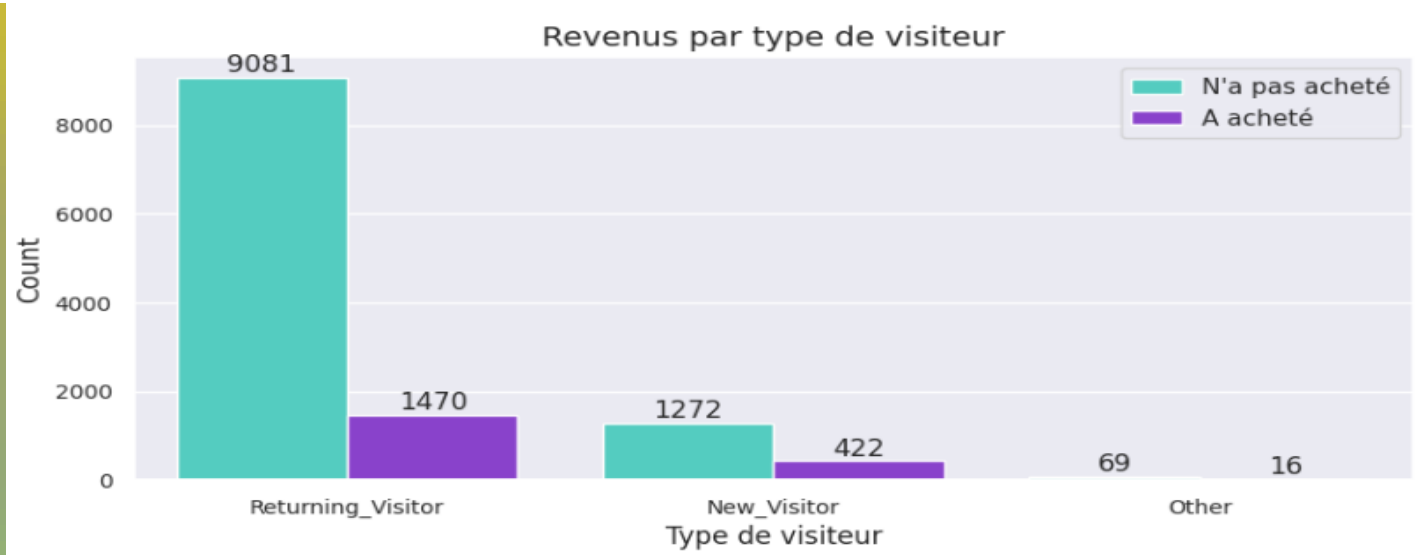
Distribution de la caractéristique Target (Revenue)

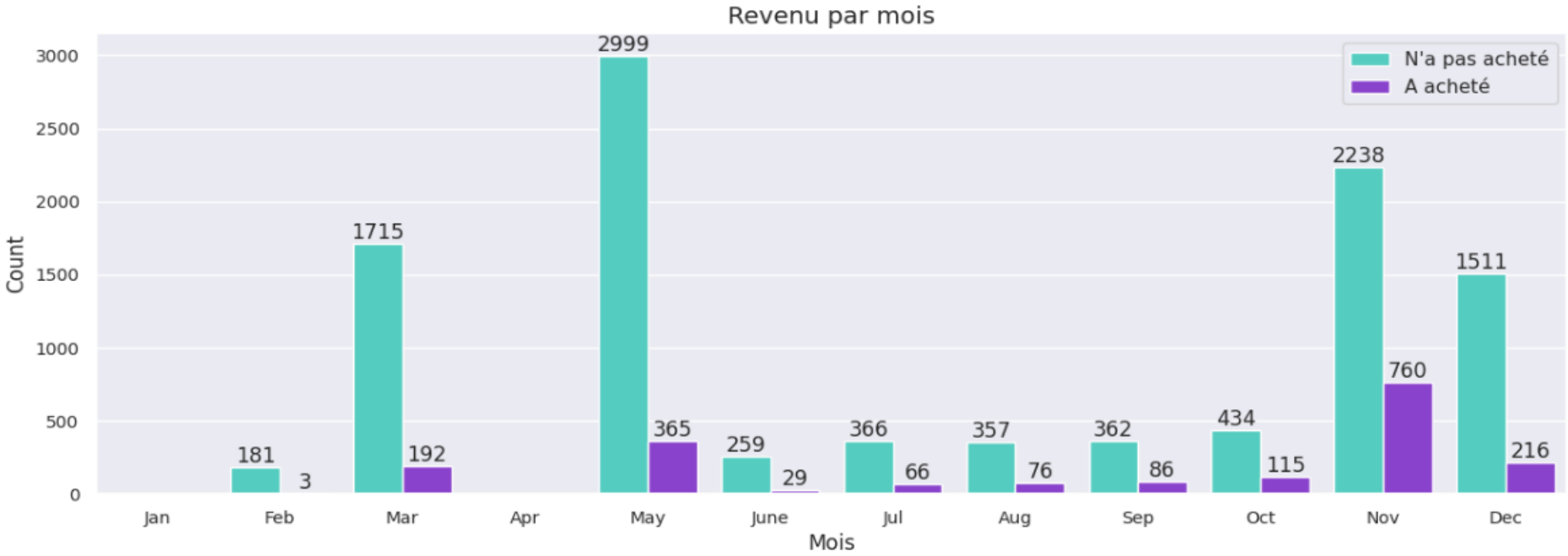
Il existe un déséquilibre dans la variable de sortie (« Revenue »), où la proportion de visiteurs qui n'ont pas effectué d'achat par rapport à ceux qui ont effectué un achat est respectivement de 84,37 % et 15,63 %.

Analyse Exploratoire des Données

Revenue par Mois

•En janvier et avril, aucune visite du site Web n'a été enregistrée.
•De nombreuses transactions ont lieu vers la fin de l'année, novembre et décembre enregistrant les revenus générés les plus élevés et les troisièmes. Cependant, ces mois ne connaissent pas les visites de sites Web les plus élevées enregistrées.
•Les quatre mois avec le plus de visites sont mai, novembre, mars et décembre.

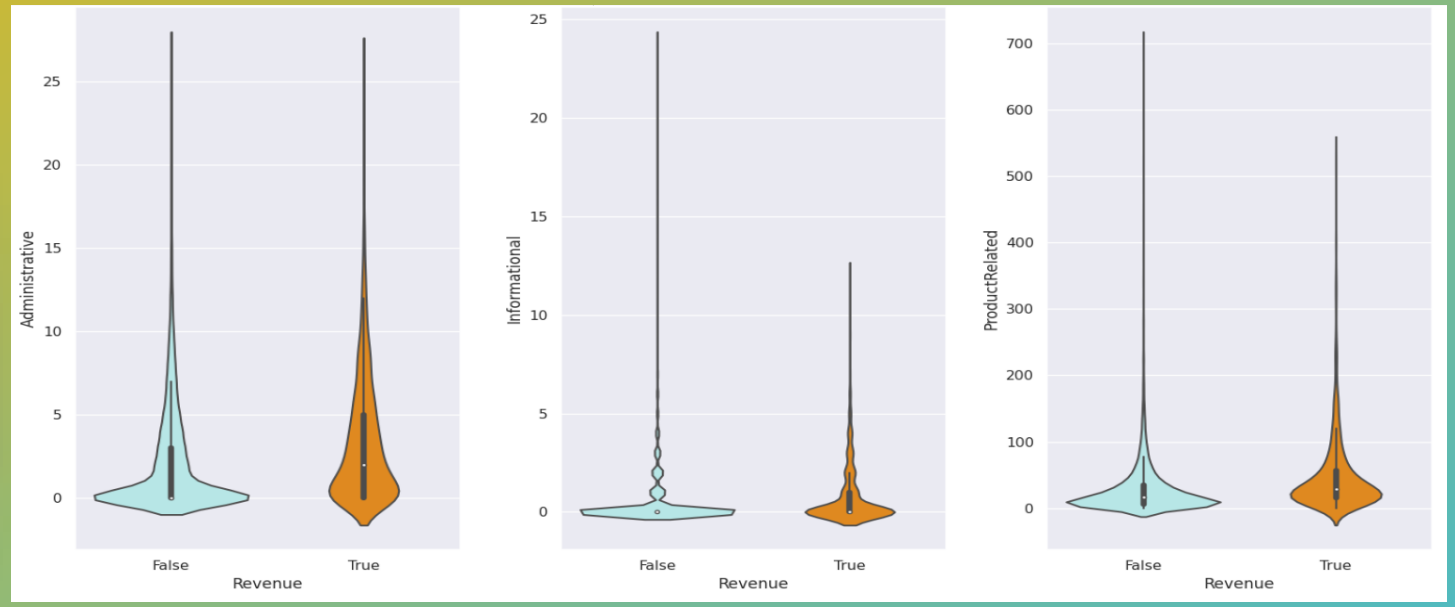


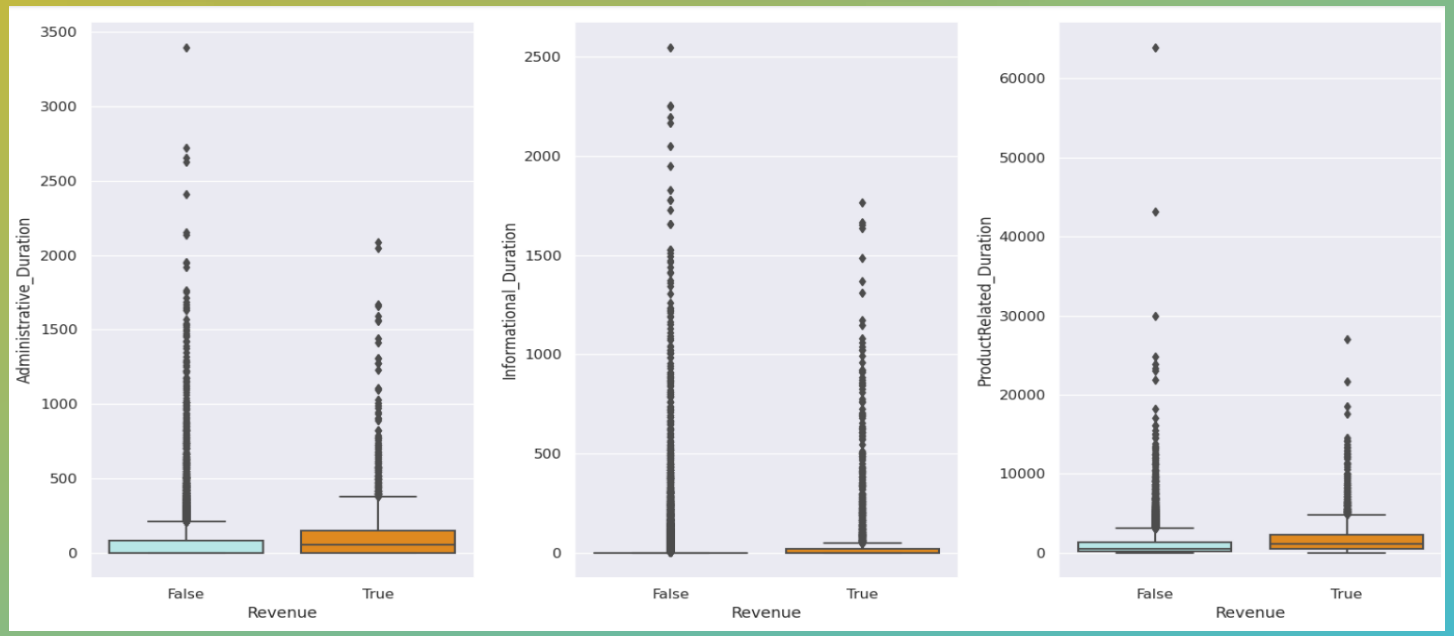


Analyse Exploratoire des Données

Revenue par type de page/durée

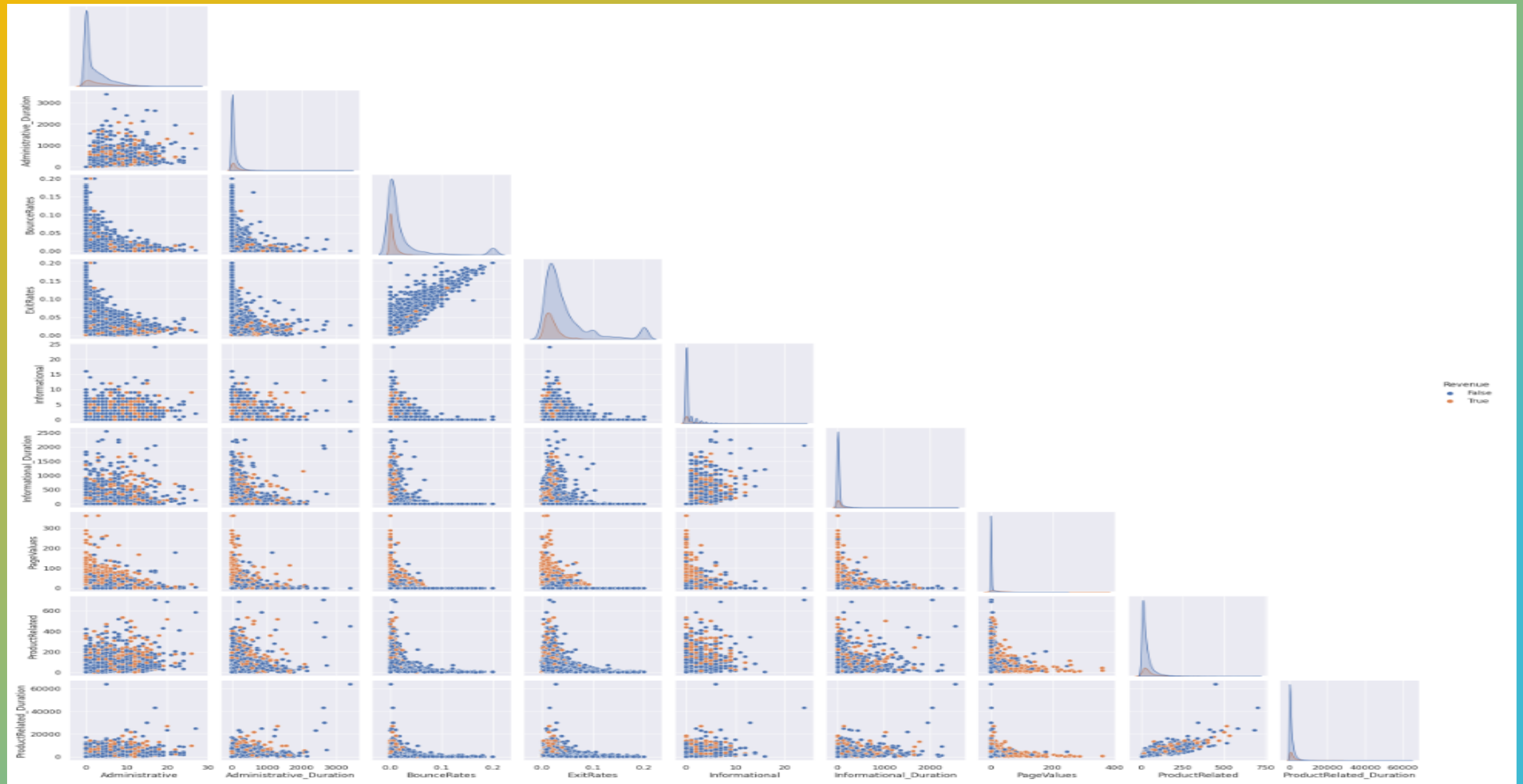
•Les visiteurs ont tendance à visiter moins de pages et à passer moins de temps s'ils ne souhaitent pas effectuer d'achat.
•Le nombre de pages liées aux produits visitées et le temps passé dessus sont supérieurs à ceux des pages liées au compte ou aux informations.





Analyse Exploratoire des Données

Analyse bivariée



05

Prétraitements des Données

Prétraitements des Données

Encodage de la fonctionnalité Mois à l'aide de l'encodage cyclique

```
# Attribuer une valeur numérique à chaque mois
month_to_num = {
    'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5,
    'June': 6, 'Jul': 7, 'Aug': 8, 'Sep': 9,
    'Oct': 10, 'Nov': 11, 'Dec': 12
}
df['month_num'] = df['Month'].map(month_to_num)

# Encoder avec sinus et cosinus
df['month_sin'] = np.sin((df['month_num'] - 1) * (2. * np.pi / 12))
df['month_cos'] = np.cos((df['month_num'] - 1) * (2. * np.pi / 12))

df.drop(columns=['Month', 'month_num'], inplace=True)
df.head()
```

ctRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue	month_sin	month_cos
1	0.000000	0.20	0.20	0.0	0.0	1	1	1	1	Returning_Visitor	False	False	0.5	0.866025
2	64.000000	0.00	0.10	0.0	0.0	2	2	1	2	Returning_Visitor	False	False	0.5	0.866025
1	0.000000	0.20	0.20	0.0	0.0	4	1	9	3	Returning_Visitor	False	False	0.5	0.866025
2	2.666667	0.05	0.14	0.0	0.0	3	2	2	4	Returning_Visitor	False	False	0.5	0.866025
10	627.500000	0.02	0.05	0.0	0.0	3	3	1	4	Returning_Visitor	True	False	0.5	0.866025

Cette étape est cruciale pour garantir une représentation adéquate des informations temporelles dans notre modèle. Voici une brève explication du processus :

- attribution d'une valeur numérique à chaque mois, simplifiant ainsi le traitement de cette caractéristique dans nos analyses subséquentes.
- L'encodage cyclique a été réalisé en utilisant les fonctions sinus et cosinus. Cela permet de transformer la caractéristique "**Mois**" en deux nouvelles caractéristiques, **'month_sin'** et **'month_cos',** qui capturent efficacement la périodicité de l'année.
- les colonnes redondantes telles que **'Month'** et **'month_num'** ont été supprimées de notre ensemble de données pour maintenir la propreté de notre modèle.

Prétraitements des Données

Encodage de la Caractéristique "VisitorType"

```
df['VisitorType_Returning_Visitor'] = 0
df['VisitorType_New_Visitor'] = 0
df['VisitorType_Other'] = 0

df.loc[df['VisitorType']=='Returning_Visitor', 'VisitorType_Returning_Visitor'] = 1
df.loc[df['VisitorType']=='New_Visitor', 'VisitorType_New_Visitor'] = 1
df.loc[df['VisitorType']=='Other', 'VisitorType_Other'] = 1

df.drop(columns=['VisitorType'], inplace=True)
```

L'objectif était de transformer cette caractéristique en plusieurs colonnes binaires, chacune indiquant la présence d'un type de visiteur spécifique. Voici une brève explication du processus :

- Créer trois nouvelles colonnes binaires : **'VisitorType_Returning_Visitor'**, **'VisitorType_New_Visitor'**, et **'VisitorType_Other'**. Chacune de ces colonnes indique la catégorie de visiteur correspondante et est codée en binaire (0 ou 1) en fonction de la présence de ce type de visiteur dans chaque observation.
- Utiliser la fonction '**loc**' pour assigner les valeurs binaires appropriées en fonction du type de visiteur dans la colonne d'origine « **VisitorType** » .
- Supprimer la colonne d'origine "**VisitorType**" pour éviter la redondance des informations.

Cet encodage de caractéristique garantit que les informations relatives au type de visiteur sont intégrées de manière efficace dans notre modèle d'apprentissage automatique, facilitant ainsi une analyse plus approfondie et des prédictions plus précises.

Prétraitements des Données

Encodage avec One-Hot Encoding des Caractéristiques Catégorielles

```
encoder = OneHotEncoder(sparse_output=False)

for col in df.select_dtypes(include='object'):
    transformed = encoder.fit_transform(df[[col]])

    encoded_df = pd.DataFrame(transformed, columns=[f"{col}_{category}" for category in encoder.categories_[0]]) # Create a DataFrame from the encoded columns

    df = df.join(encoded_df.set_index(df.index)) # Ajouter les nouvelles colonnes au DataFrame d'origine


df.drop(columns=df.select_dtypes(include='object').columns, inplace=True) # Supprimer les colonnes catégorielles d'origine
```

Cette approche permet de convertir les variables catégorielles en un format numérique adapté à l'analyse et à l'apprentissage automatique. Voici une brève explication du processus:

- Utiliser la classe **OneHotEncoder** pour encoder toutes les fonctionnalités catégorielles de notre ensemble de données. Cette étape consiste à créer de nouvelles colonnes binaires pour chaque catégorie possible dans chaque fonctionnalité catégorielle.
- Une fois l'encodage **One-Hot** effectué, nous avons supprimé les colonnes catégorielles d'origine pour éviter la redondance des informations.

06

Modélisation

Modélisation

Entrainement du modèle

Afin de préparer nos données pour le processus d'apprentissage automatique, une étape cruciale de séparation entre les étiquettes et les caractéristiques est effectuée. Voici une brève explication du processus:

Extraire la variable cible, représentée par la colonne « Revenue » Ces étiquettes, indiquant si une transaction a été effectuée (revenu généré) ou non, constituent la base sur laquelle notre modèle sera formé.

```
[66] y = df['Revenue'] # Étiquettes
     X = df[df.columns.difference(['Revenue']).to_list()] # Caractéristiques
```

```
[67] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Les caractéristiques, représentées par toutes les colonnes restantes dans notre ensemble de données, ont été extraites de manière à exclure la variable cible. Cela inclut toutes les variables numériques ainsi que les nouvelles colonnes créées par les processus d'encodage.

Les modèles prédictifs qui seront utilisés sont la régression logistique, le classificateur KNeighbours, le SVM, l'arbre de décision et le classificateur de forêt aléatoire.

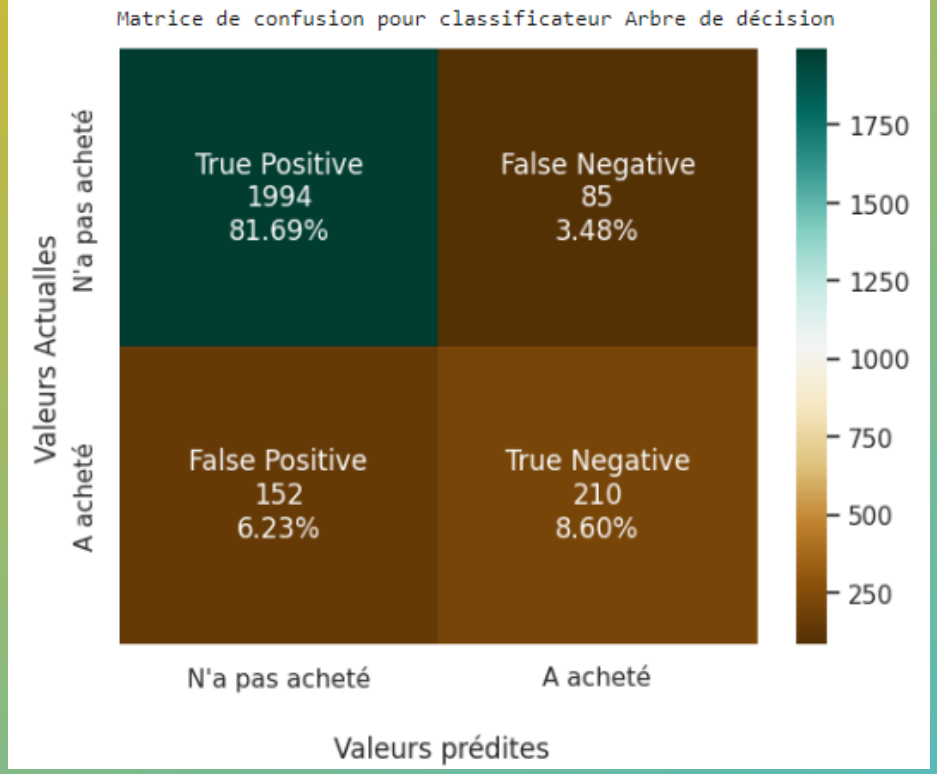
Modélisation

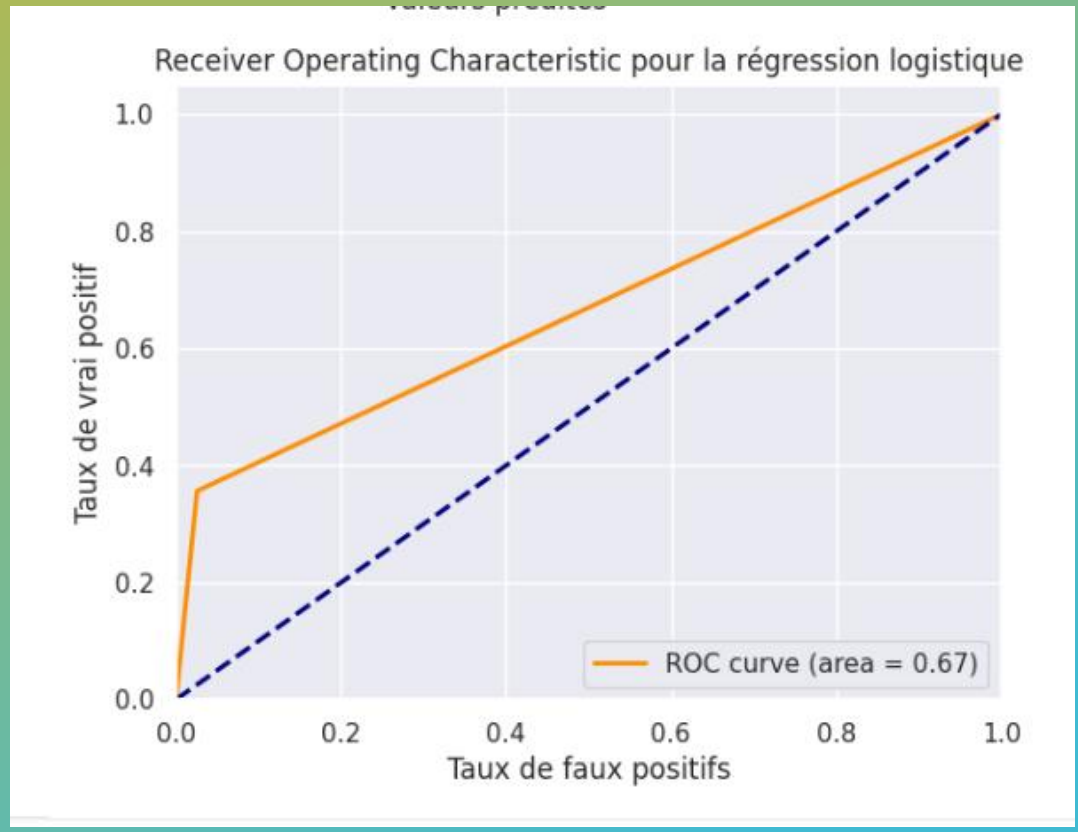
Modèle de la régression logistique ou Logistic Regression

Pour Arbre de décision, Accuracy est 0.9029086439983613

	precision	recall	f1-score	support
False	0.93	0.96	0.94	2079
True	0.71	0.58	0.64	362
accuracy			0.90	2441
macro avg	0.82	0.77	0.79	2441
weighted avg	0.90	0.90	0.90	2441

```
[[1994   85]
 [ 152  210]]
```

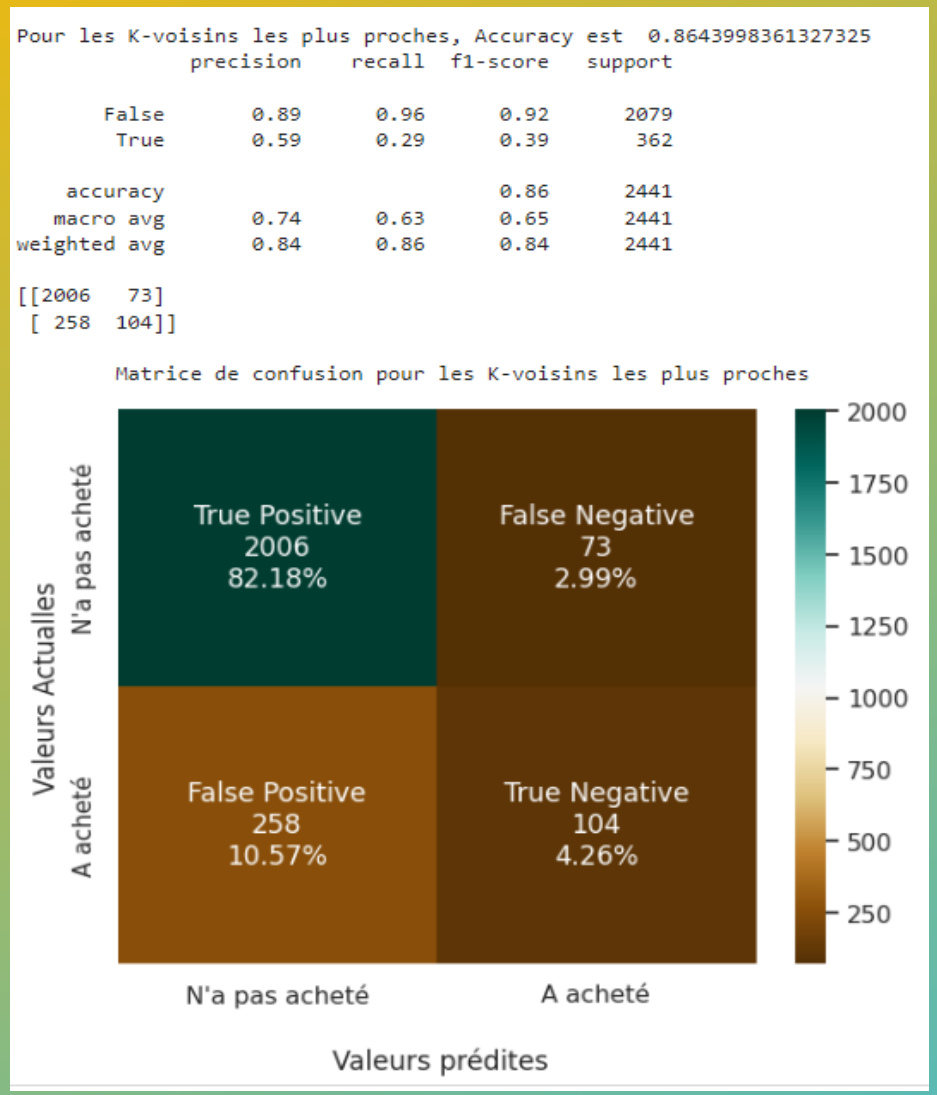


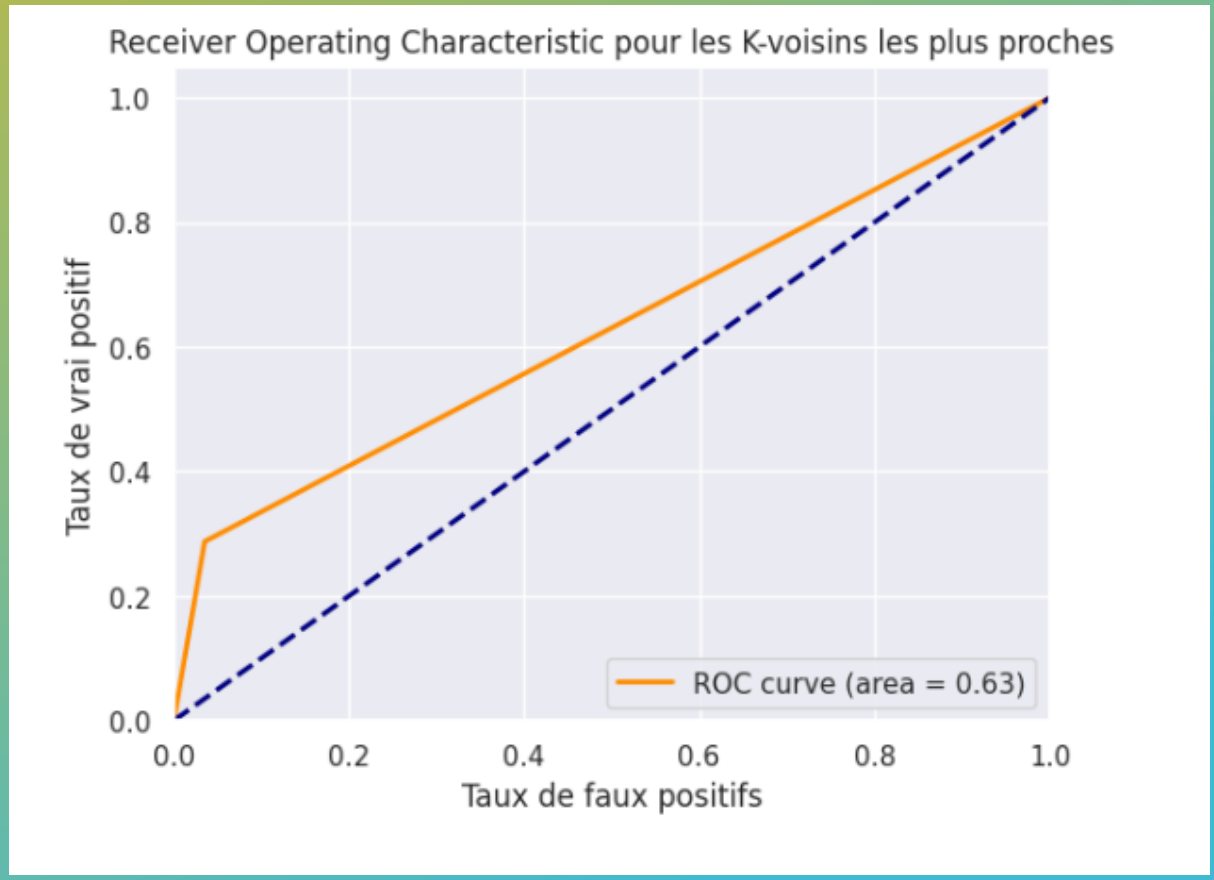


Modélisation

Modèle des K plus proches voisins ou K Neighbors

```
Pour les K-voisins les plus proches, Accuracy est  0.8643998361327325
              precision    recall  f1-score   support

       False       0.89      0.96      0.92      2079
        True       0.59      0.29      0.39       362

    accuracy                           0.86      2441
   macro avg       0.74      0.63      0.65      2441
weighted avg       0.84      0.86      0.84      2441

[[2006   73]
 [ 258  104]]
```

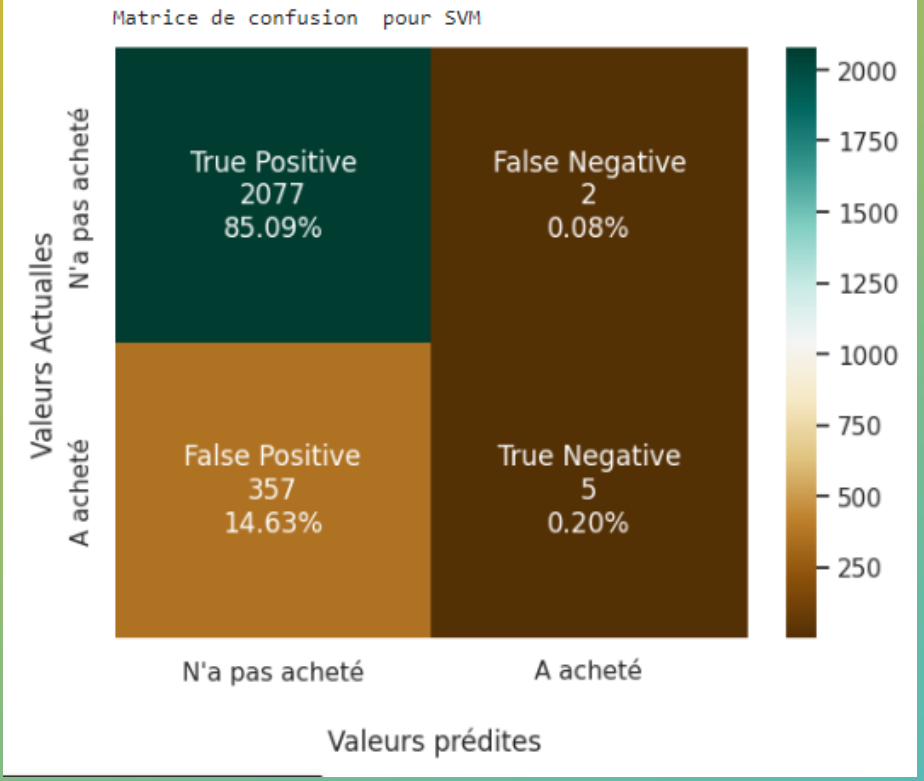


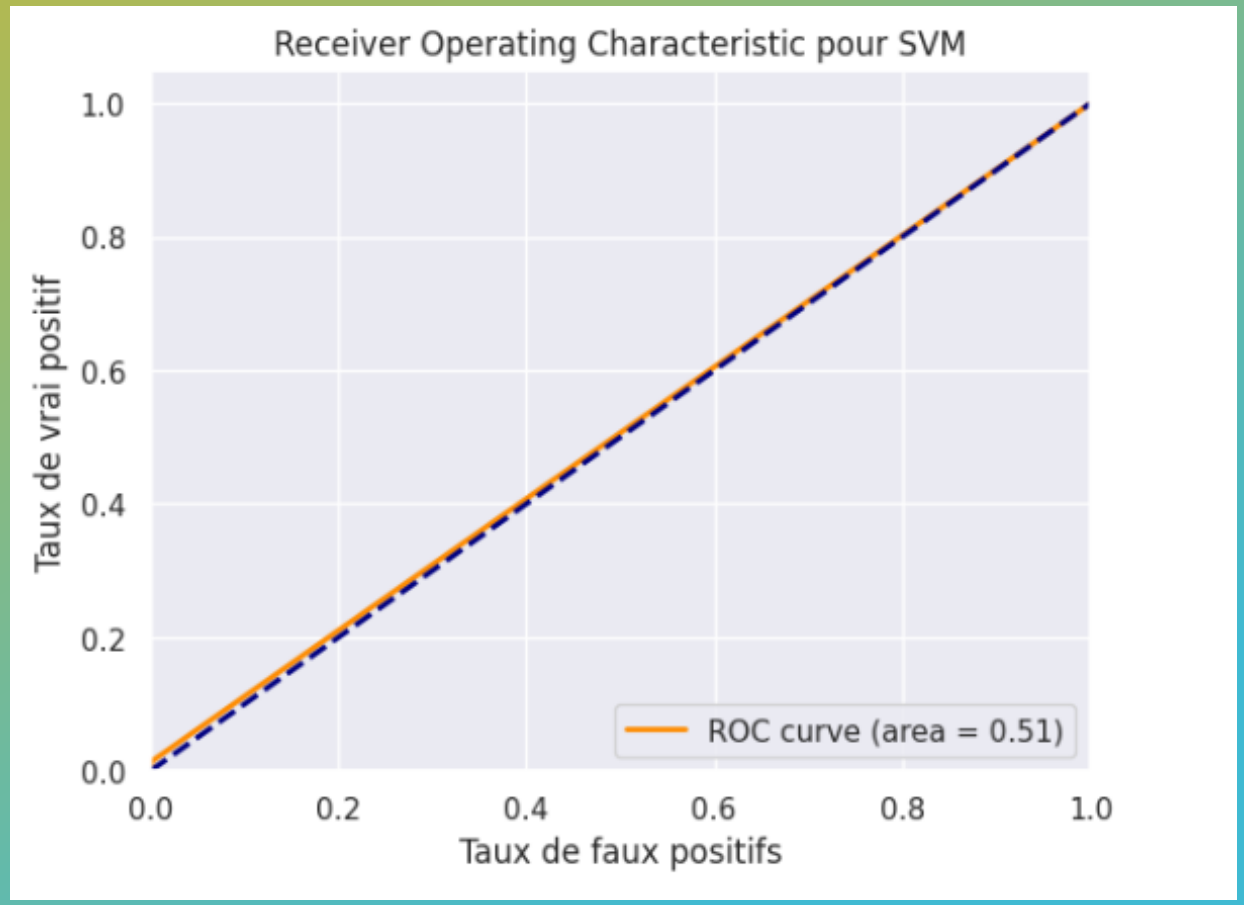


Modélisation

Modèle Support Vector Machine ou Machine à vecteurs de support(SVM)

```
Pour SVM, Accuracy est  0.8529291274068005
              precision    recall  f1-score   support

       False       0.85      1.00      0.92      2079
        True       0.71      0.01      0.03       362

    accuracy                           0.85      2441
   macro avg       0.78      0.51      0.47      2441
weighted avg       0.83      0.85      0.79      2441

[[2077    2]
 [ 357    5]]
```

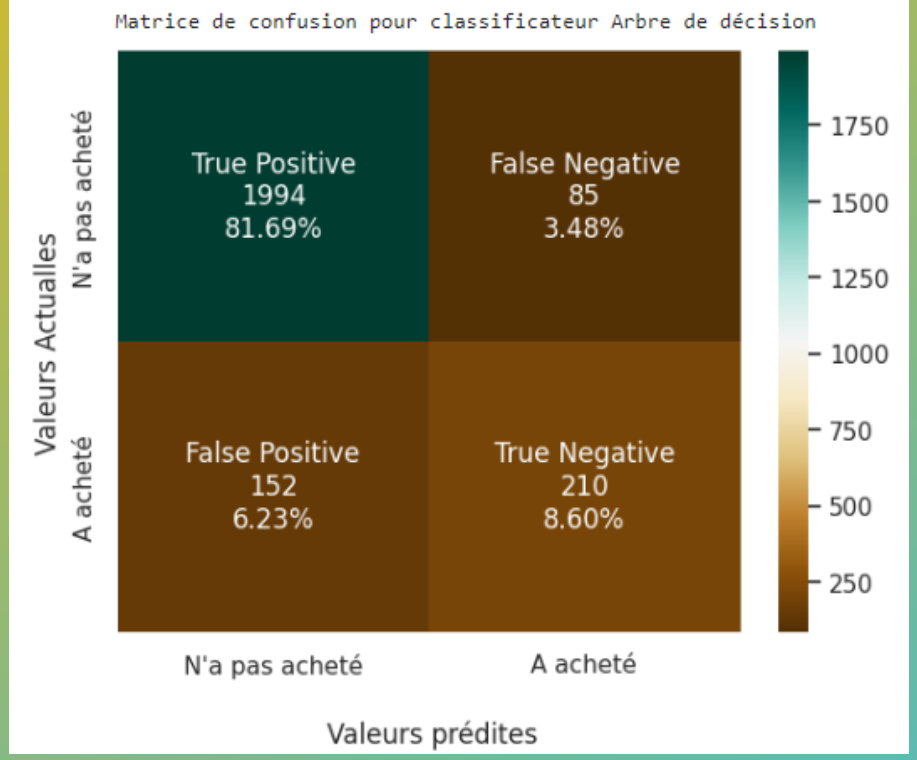


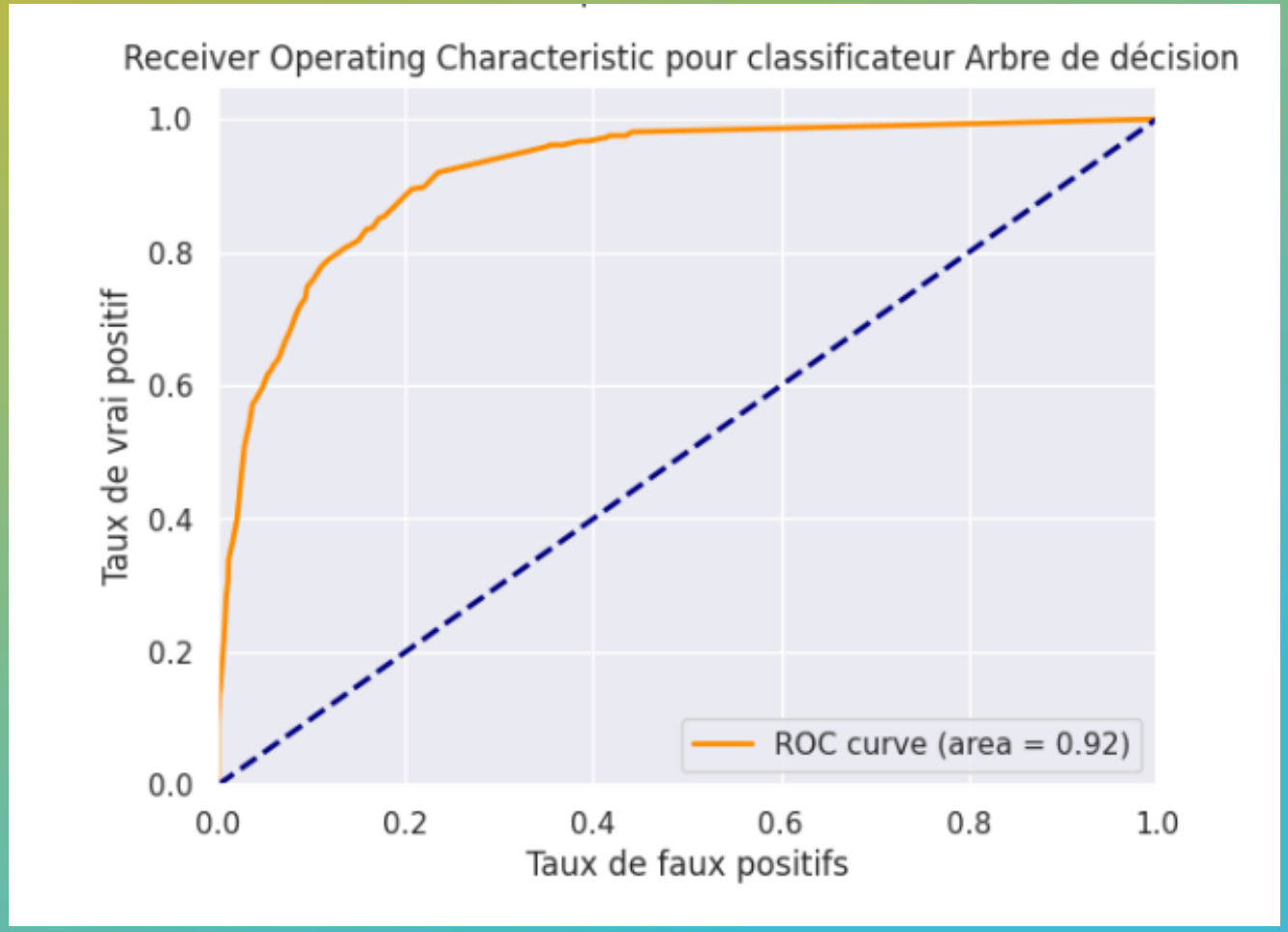


Modélisation

Modèle d'arbre de décision ou Decision Tree

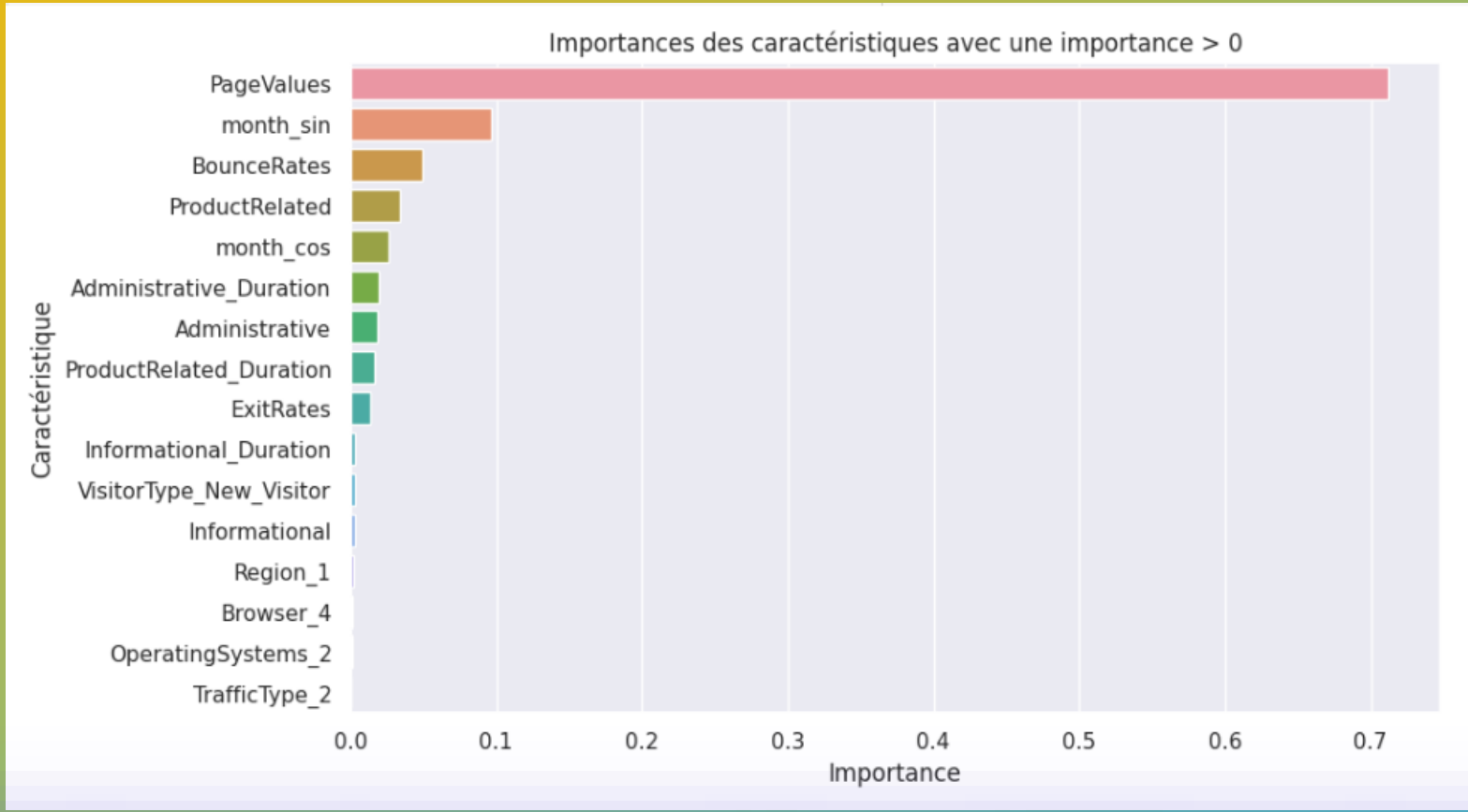
```
Pour Arbre de décision, Accuracy est  0.9029086439983613
              precision    recall  f1-score   support

       False       0.93      0.96      0.94      2079
        True       0.71      0.58      0.64       362

    accuracy                           0.90      2441
   macro avg       0.82      0.77      0.79      2441
weighted avg       0.90      0.90      0.90      2441

[[1994   85]
 [ 152  210]]
```





Modélisation

Évaluation de l'Importance des caractéristiques



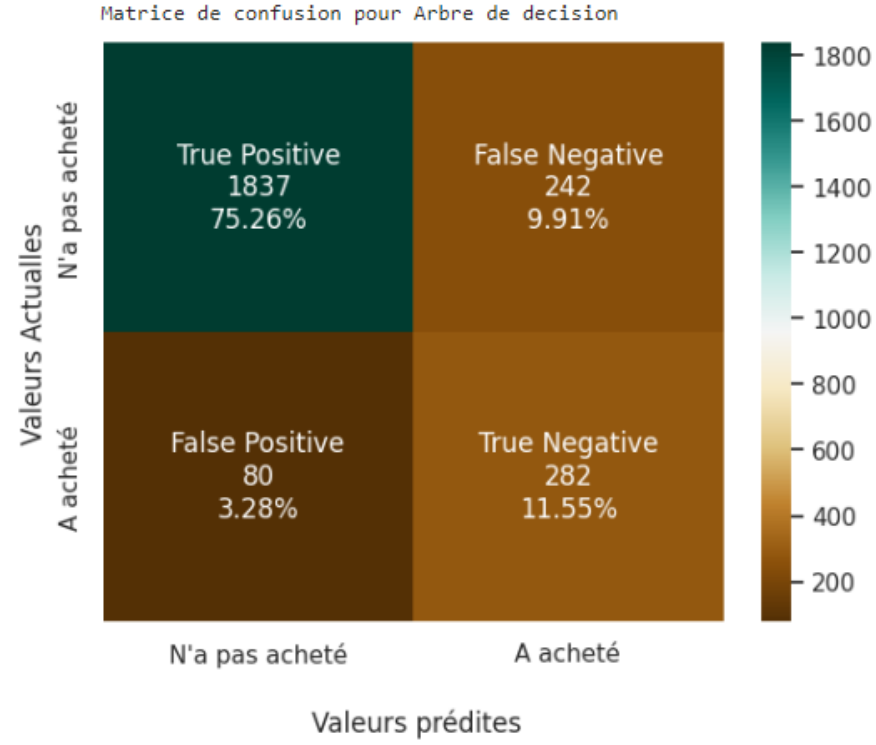
Cette analyse nous permet de cibler les aspects clés qui contribuent significativement à la capacité prédictive de notre modèle, en extractant les importances des fonctionnalités à partir de notre modèle, puis créé un tableau récapitulatif mettant en évidence les relations entre chaque fonctionnalité et son niveau d'importance. Ensuite , on utilise une visualisation graphique avec un diagramme à barres pour représenter graphiquement l'importance relative de chaque fonctionnalité.

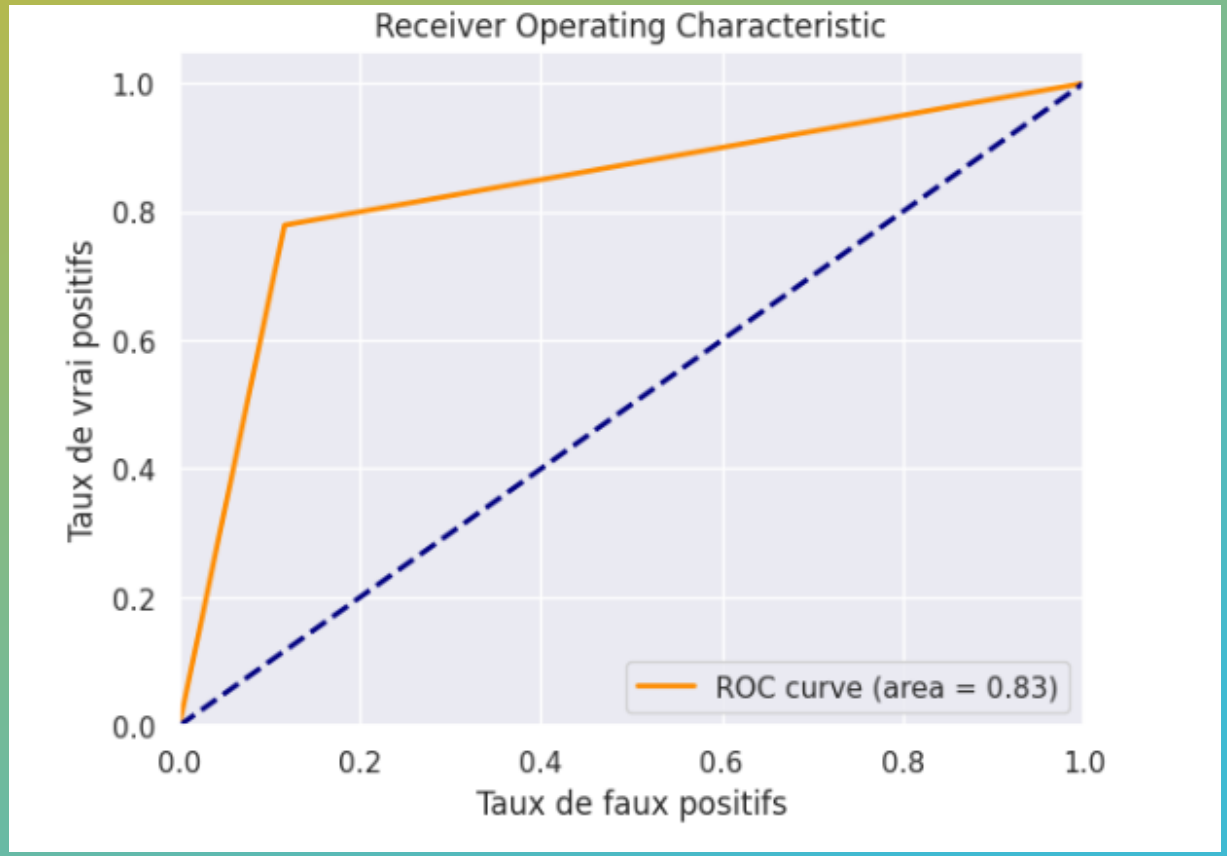
Il semble que « **PageValues** » soit considérablement plus important que toutes les autres caractéristiques. Essayons d'adapter un classificateur d'arbre de décision en utilisant uniquement la caractéristique « **PageValues** ».

Modélisation

Modèle d'arbre de décision ou Decision Tree

```
Pour Arbre de décision, Accuracy est  0.8680868496517821
              precision    recall  f1-score   support

       False       0.96      0.88      0.92      2079
        True       0.54      0.78      0.64       362

    accuracy                           0.87      2441
   macro avg       0.75      0.83      0.78      2441
weighted avg       0.90      0.87      0.88      2441

[[1837  242]
 [  80  282]]
```



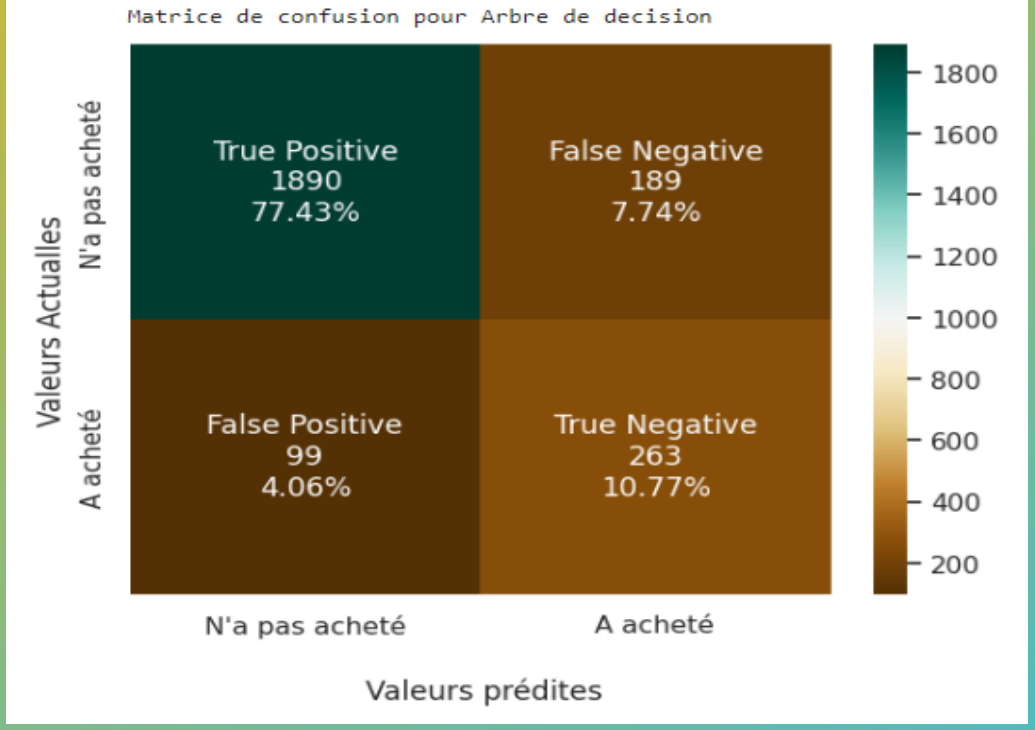


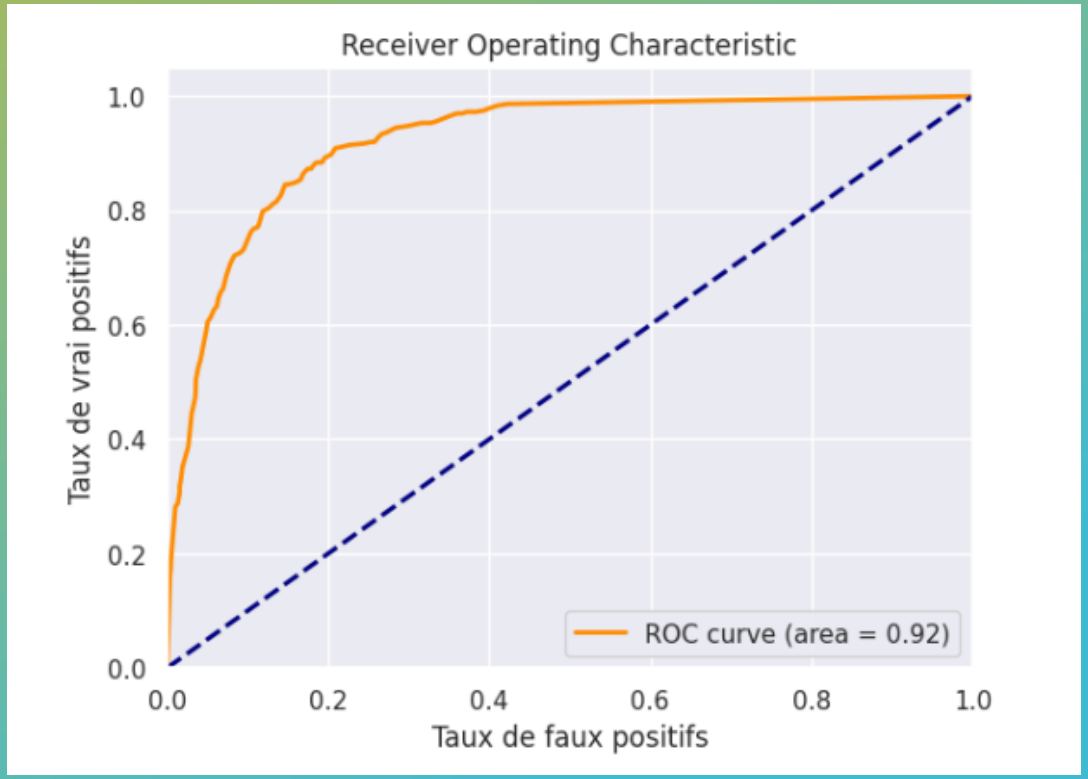
Modélisation

Modèle d'arbre de décision ou Decision Tree : Suréchantillonnage

Étant donné que les données sont très déséquilibrées dans l'étiquette, essayons d'utiliser la technique de suréchantillonnage SMOTE dans le pipeline

```
Pour Arbre de décision, Accuracy est  0.8820155673904138
              precision    recall  f1-score   support

       False       0.95      0.91      0.93      2079
        True       0.58      0.73      0.65       362

    accuracy                           0.88      2441
   macro avg       0.77      0.82      0.79      2441
weighted avg       0.90      0.88      0.89      2441

[[1890  189]
 [  99  263]]
```





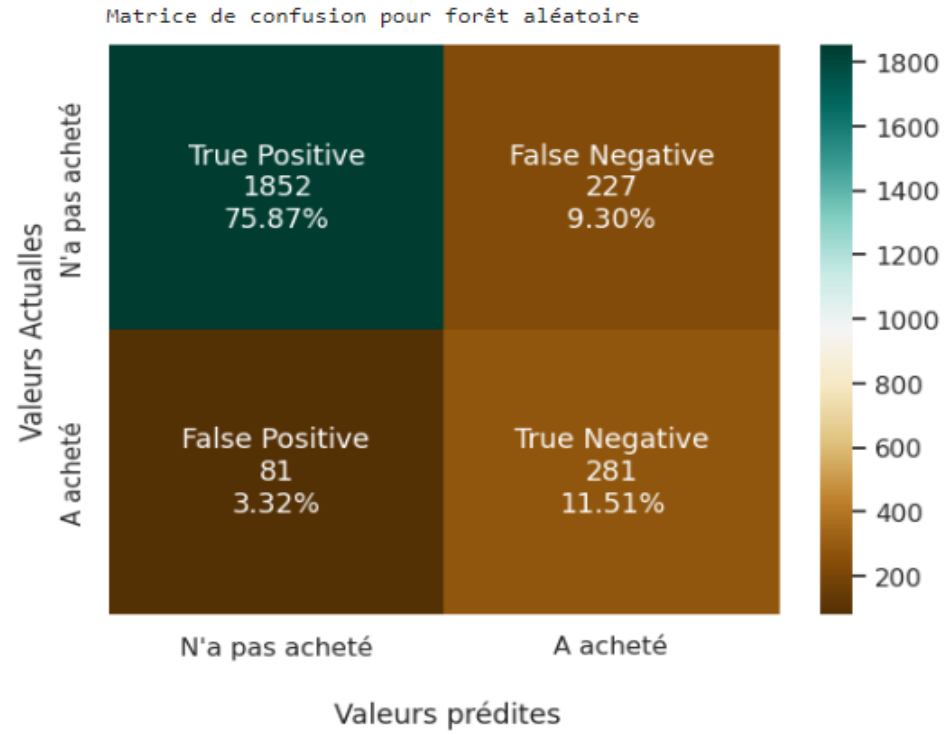
Modélisation

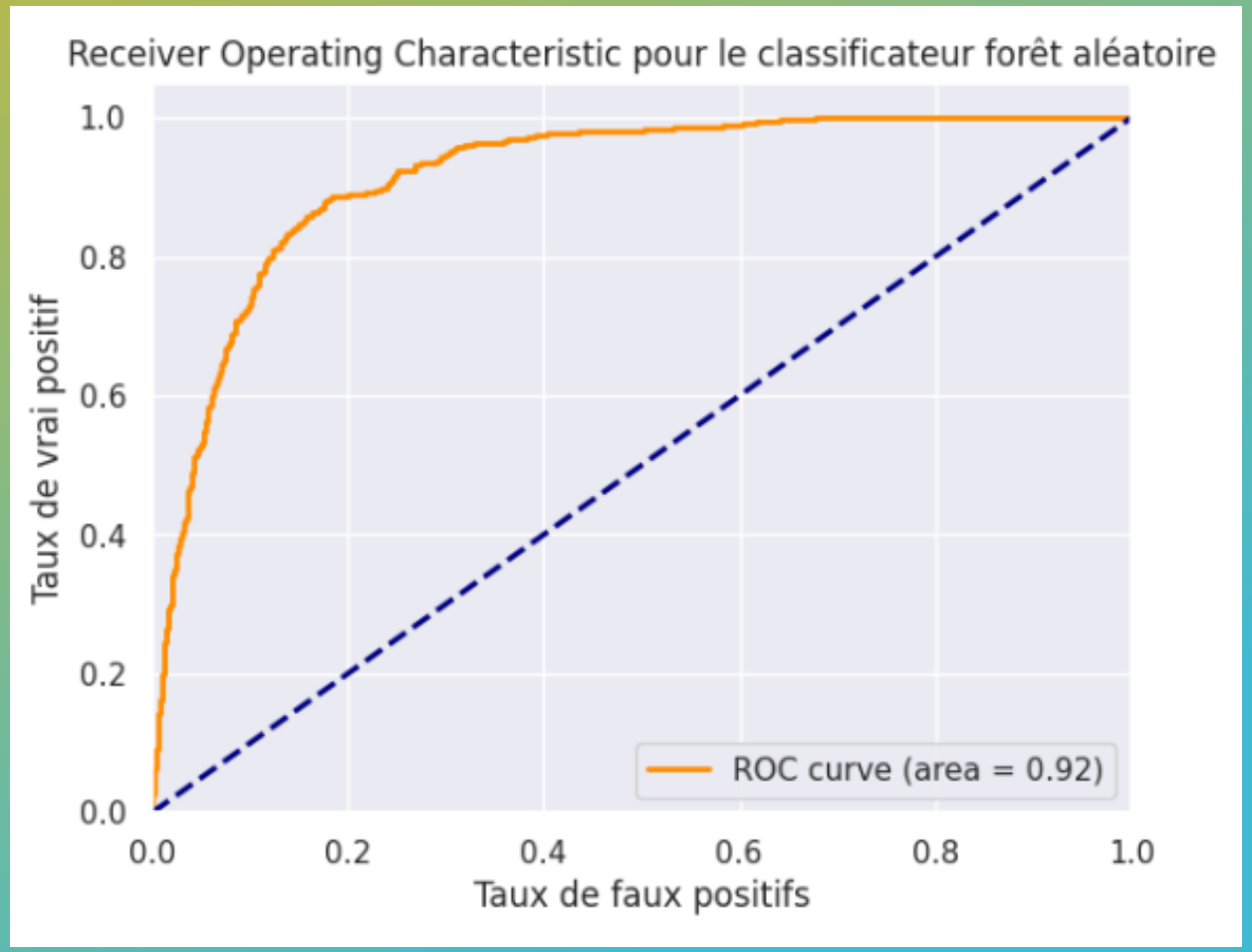
Modèle de forêt aléatoire ou Random Forest

	precision	recall	f1-score	support
False	0.96	0.89	0.92	2079
True	0.55	0.78	0.65	362
accuracy			0.87	2441
macro avg	0.76	0.83	0.78	2441
weighted avg	0.90	0.87	0.88	2441

```
[[1852  227]
 [  81  281]]
```

Pour forêt aléatoire, Accuracy est 0.8738222040147481





Modélisation

Modèle de forêt aléatoire ou Random Forest

Optimisation des Hyperparamètres avec Recherche Aléatoire de Moitié

- On adapte une démarche d'optimisation des hyperparamètres de notre modèle RandomForestClassifier en utilisant la technique de recherche aléatoire de moitié, accompagnée d'une validation croisée pour garantir la robustesse de notre processus d'optimisation.
- On construit un pipeline qui inclut la mise à l'échelle des données, l'application de la technique **SMOTE** pour gérer le déséquilibre de classe, et l'utilisation du modèle RandomForestClassifier.
- On configure la recherche aléatoire de moitié en définissant une grille d'hyperparamètres sur laquelle effectuer la recherche. Cette grille inclut des paramètres tels que le nombre d'estimateurs, la profondeur maximale, et d'autres paramètres spécifiques au **RandomForestClassifier**.
- On utilise la métrique du score F1 pondéré comme critère d'évaluation pour la recherche aléatoire de moitié..
- Après avoir exécuté la recherche aléatoire de moitié, nous avons identifié les meilleurs hyperparamètres pour notre modèle RandomForestClassifier.

```
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_search.py:952: UserWarning: One or more of the train scores are non-finite: [       nan       nan       nan ... 0.92157369 0.
  warnings.warn(
Best hyperparameters: {'classifier__n_estimators': 75, 'classifier__min_samples_leaf': 4, 'classifier__max_samples': 0.9, 'classifier__max_features': 'sqrt', 'classifier__max_depth': None}
```

Modélisation

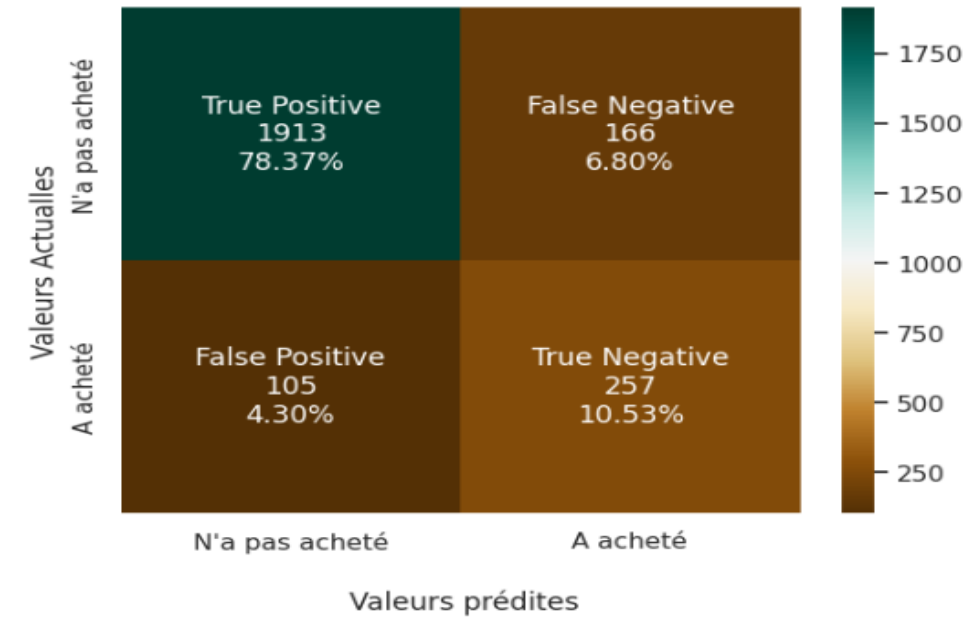
Modèle de forêt aléatoire ou Random Forest

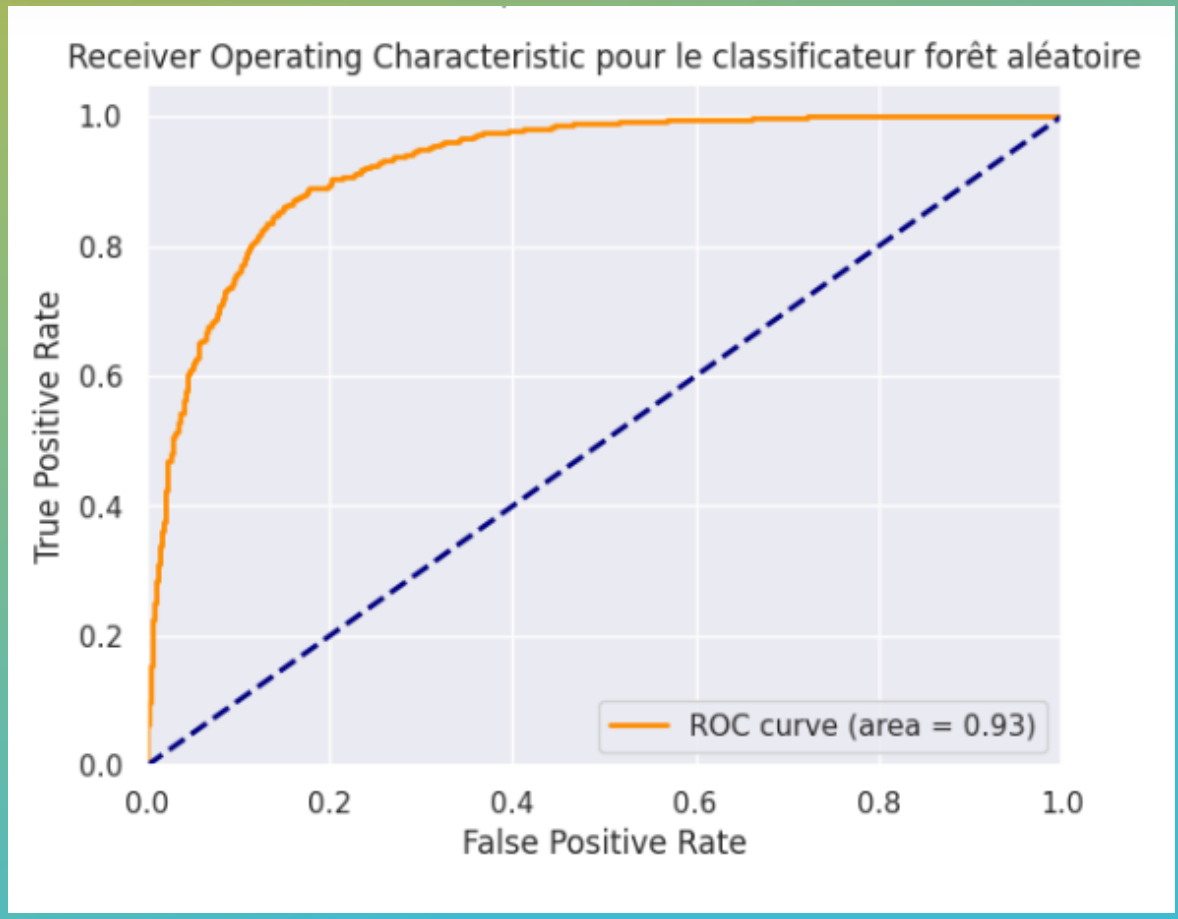
Mise en place d'un pipeline intégrant plusieurs étapes, notamment la mise à l'échelle des données, l'utilisation de la technique SMOTE pour gérer le déséquilibre de classe, et la formation du modèle RandomForestClassifier en ajustant les paramètres de notre modèle par les paramètres trouvés dans la diapositive précédente.

Pour forêt aléatoire, Accuracy est 0.8889799262597297

	precision	recall	f1-score	support
False	0.95	0.92	0.93	2079
True	0.61	0.71	0.65	362
accuracy			0.89	2441
macro avg	0.78	0.82	0.79	2441
weighted avg	0.90	0.89	0.89	2441

```
[[1913  166]
 [ 105  257]]
```





07

Interprétation des résultats

Interprétation des résultats

Importance des Caractéristiques

L'analyse de l'importance des caractéristiques a révélé les principaux contributeurs à la prédiction des intentions d'achat en ligne. Certains facteurs ont émergé comme des déterminants significatifs, guidant ainsi notre compréhension des aspects spécifiques qui influencent le comportement des acheteurs.

Optimisation des Hyperparamètres

La recherche aléatoire de moitié avec validation croisée a permis d'identifier les configurations optimales des hyperparamètres pour notre modèle. Ces paramètres ajustés ont significativement amélioré les performances du modèle, augmentant sa capacité à généraliser efficacement à de nouvelles données..

Interprétation des Métriques d'Évaluation

Les métriques d'évaluation telles que l'accuracy, le score F1, et la courbe ROC offrent une vision complète de la performance du modèle.
Ces indicateurs ont été cruciaux pour évaluer la capacité du modèle à discriminer entre les classes positives et négatives, ainsi que pour fournir des informations sur les taux de faux positifs et faux négatifs, contribuant ainsi à une interprétation approfondie des résultats du projet.

08

Conclusion

Conclusion

- Dans le cadre de ce projet, on a développé des modèles capables de classer les visiteurs d'un site web marchand et de prédire s'ils sont susceptibles de réaliser un achat sur le site ou non. Cinq classificateurs d'apprentissage (Régression Logistique, KNN, SVM, Arbre de Décision et Forêt Aléatoire) ont été évalués.
- L'analyse d'importance des fonctionnalités a révélé que la fonctionnalité "Page Values" est le déterminant le plus crucial de l'intention d'achat d'un visiteur.
- D'autres caractéristiques importantes comprennent le taux de sortie, le taux de rebond, le type de pages visitées, ainsi que la durée passée sur ces pages. Cette information fournit des indications essentielles pour optimiser l'expérience utilisateur et orienter les efforts de marketing.
- En conclusion, ce projet offre des perspectives significatives pour les entreprises en ligne en fournissant des outils efficaces pour anticiper les comportements des consommateurs. L'utilisation du Classifier de Forêt Aléatoire, associée à l'identification de caractéristiques clés, permet d'améliorer la prise de décision et d'optimiser les stratégies commerciales. L'aptitude à généraliser à de nouvelles données renforce la pertinence de notre modèle dans des contextes opérationnels réels.

M e r c i

F i n d e p r e s e n t a t i o n