ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Rapport:

Advanced Machine Learning

DATA & AI

HAOUAS Lies FADILI Yanis

Professeur :

Denis OBLIN

Introduction 3
Présentation du dataset 3
Data pre-processing 7
Feature engineering 9
Modèles **9**
Conclusion **10**

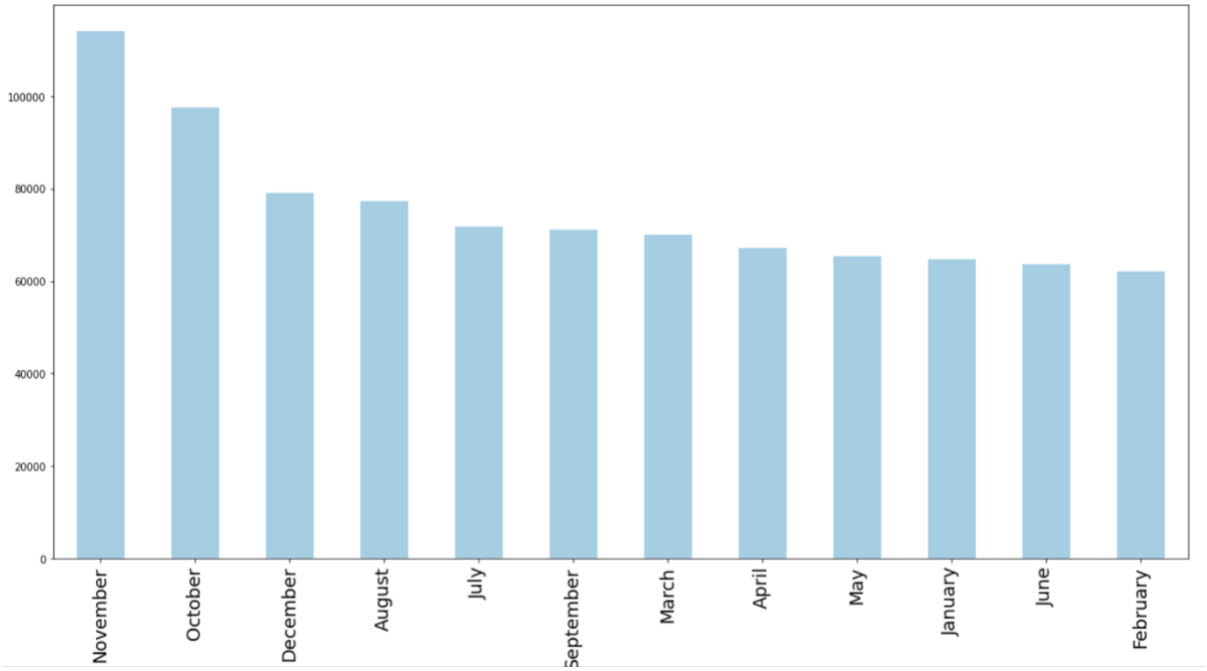
1. Introduction

1.1. Présentation du dataset

Le dataset contient 55 colonnes et 903 653 lignes. Ce dataset contient les différentes opérations de ventes, dans les 55 colonnes nous avons des informations sur la date, l'heure, le navigateur et le système d'opération utilisé.

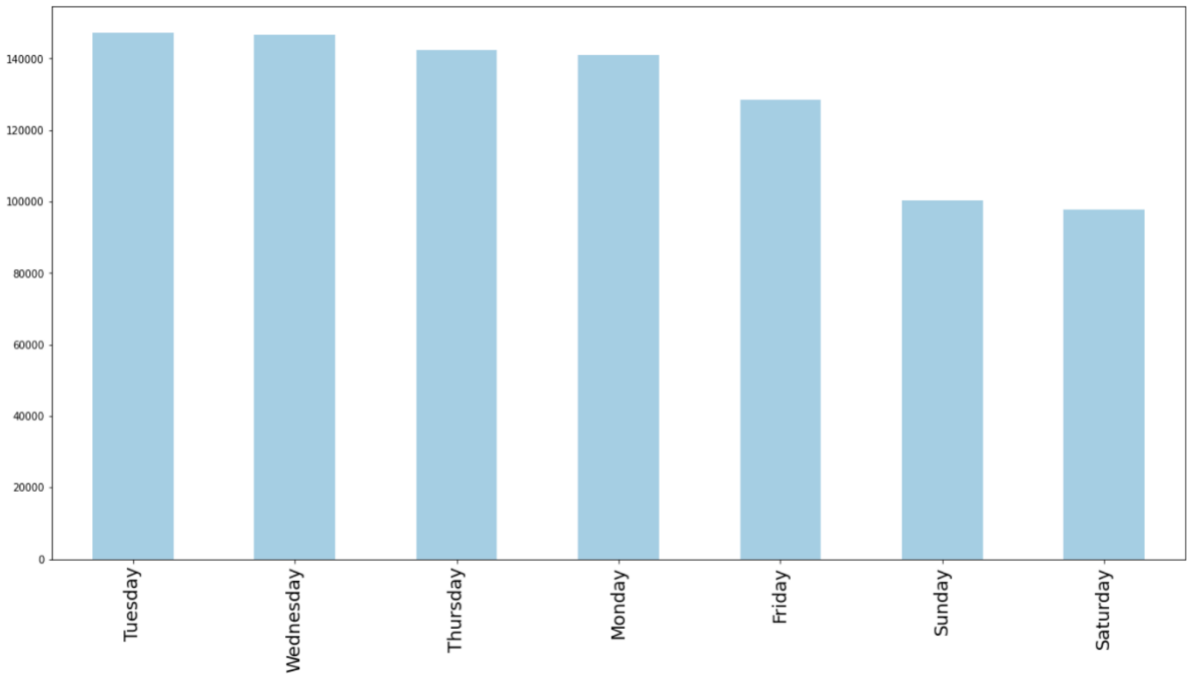
L'objectif du projet est ici de travailler sur ce jeu de données afin de prédire le revenu qu'un utilisateur génère au totale. Le projet est développé en Python 3 sur Jupyter Notebook, majoritairement avec la bibliothèque Pandas, Matplotlib.

Diagramme en bar du nombre de transaction par mois :



On peut voir que le mois de Novembre contient le plus d'achat.

Diagramme en bar du nombre de transaction par jour :



On peut voir le nombre d'achat diminue vraiment le week-end comparé aux autres jours de la semaine.

Pour vous donner plus d'information, sur la répartition des achats sur les différents pays et continents.

Voici, un tableau représentant le classement des 5 continents avec le plus transaction :

index	count
Americas	450377
Asia	223698
Europe	198311
Oceania	15054
Africa	14745

Diagramme en bar du nombre de transaction par continent :

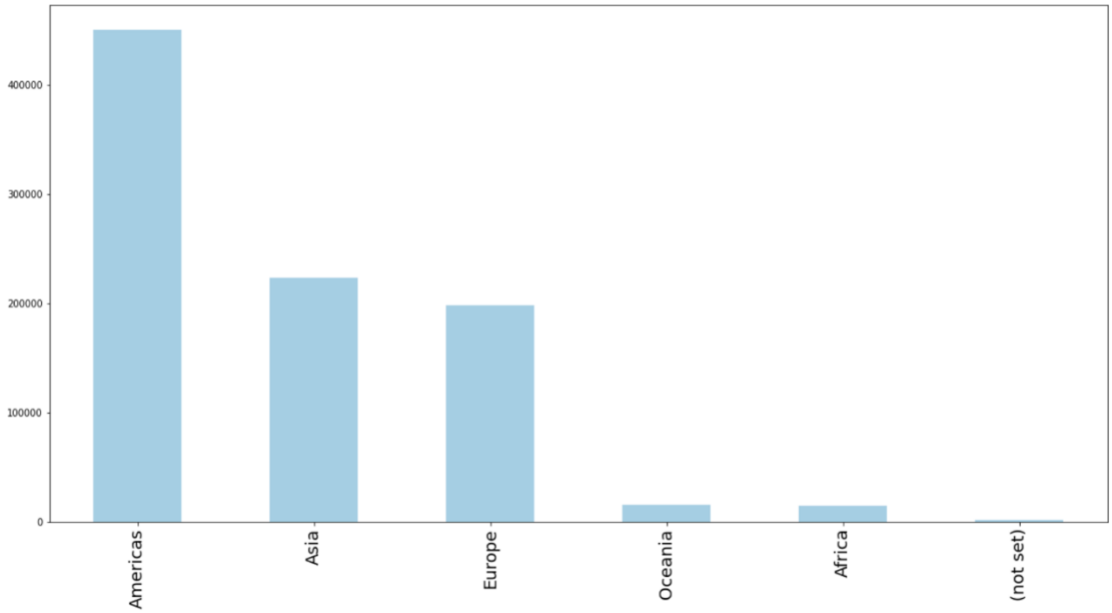


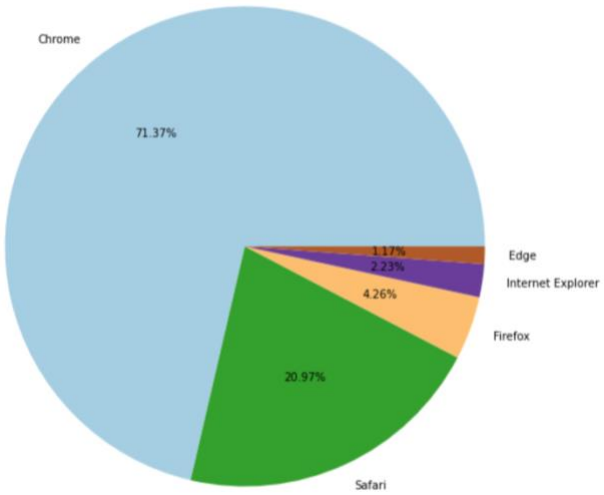
Tableau représentant le classement des 10 pays avec le plus transaction :

index	count
United States	364744
India	51140
United Kingdom	37393
Canada	25869
Vietnam	24598
Turkey	20522
Thailand	20123
Germany	19980
Brazil	19783
Japan	19731

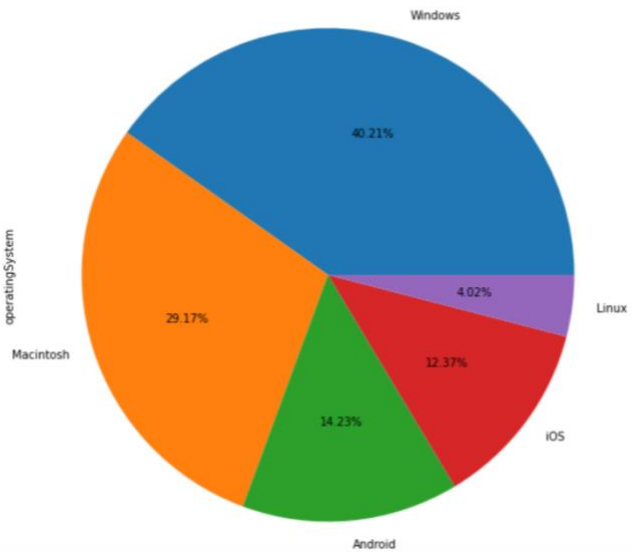
On peut voir que sur le continent Américain, il y a le plus de transaction et quasiment le double comparé au deuxième continent avec le plus de transaction.
L'Asie et l'Europe sont assez proche, avec en tête comme pays l'Inde et le Royaume-Uni, le deuxième et troisième pays avec le plus de transaction au total.
On peut maintenant avoir une meilleure idée de la répartition des transactions entre les différents continent et pays.
On peut faire l'hypothèse que le site est très bien implanté dans les pays anglophones tels que les États-Unis, l'Inde, et le Royaume-Uni.
Ainsi pour la suite, c'est une information pertinente pour savoir quels pays et segment de marché visée pour augmenter leur visibilité ou leurs parts de marché sur un pays et adapté leur stratégie marketing en conséquence.

Maintenant, nous allons regarder la répartition par rapport au navigateur, système d'exploitation, et la source :

Répartition des achats par navigateur :



Répartition des achats par système d'exploitation :



On peut ainsi observer que la plupart des utilisateurs utilisent un ordinateur le plus souvent le système d'exploitation Windows et le navigateur Chrome, ce qui nous donne un indice de quel type de plateforme les utilisateurs utilisent et ainsi permettre d'améliorer l'UX (user expérience) sur ce type de plateforme ou même à l'inverse comprendre pourquoi sur certains de plateforme il y a moins d'achat et ainsi améliorer l'UX spécifiquement sur ce type de plateforme.

Après avoir analysé les données, nous avons trouvé quelques problèmes comme les colonnes avec des valeurs nulles ou des colonnes porteuses de peu d'information ou que le dataset était trop volumineux. Ainsi, nous avons décidé de faire du data pre-processing.

1.2. Data pre-processing

Le nom des colonnes et leurs nombres de null values :

index	count	count in %
adContent	892707	98.8
transactionRevenue	892138	98.7
slot	882193	97.6
page	882193	97.6
isVideoAd	882193	97.6
adNetworkType	882193	97.6
gclId	882092	97.6
isTrueDirect	629648	69.7
referralPath	572712	63.4
keyword	502929	55.7
bounces	453023	50.1
newVisits	200593	22.2

On a remarqué que plusieurs colonnes contenaient énormément de valeurs nulles pour pallier ce problème, nous avons préféré supprimer les colonnes ayant plus de 50% de valeurs manquantes.
À part pour la colonne « *transactionRevenue* » qui est notre valeur à prédire et pour cela nous avons juste remplacer les valeurs manquantes par 0 étant donné que si l'utilisateur n'a pas procédé à un achat lors de sa visite, la valeur de la colonne « *transactionRevenue* » n'est pas renseigné. C'est l'hypothèse que nous avons faite.

Pour la colonne « *pageviews* », on a décidé de remplacer les valeurs manquantes par la valeur la plus fréquentes. Pour la colonne « newVisits », on a remplacé les valeurs « nan » par « 0 » pour faciliter le traitement.

Ensuite, nous avons supprimé toutes les colonnes qui n'était pas disponible dans la version démo du dataset, c'est-à-dire les colonnes qui contenait uniquement la valeur : « not available in demo dataset »

De plus, les colonnes contenant uniquement un seul type de valeur, tels que "*socialEngagementType*" qui ne contenait uniqument la valeur : "*Not Socially Engaged*", nous avons décidé de les supprimé car elles n'apportaient aucune information en plus à notre input, vu que c'était toujours la même valeur.

Le nom des colonnes qui ne sont pas disponible dans la version démo du dataset :

```
cityId
latitude
longitude
networkLocation
browserVersion
browserSize
operatingSystemVersion
mobileDeviceBranding
mobileDeviceModel
mobileInputSelector
mobileDeviceInfo
mobileDeviceMarketingName
flashVersion
language
screenColors
screenResolution
criteriaParameters
```

Le nom des colonnes qui ne contiennent qu'une seule valeur :

```
socialEngagementType ['Not Socially Engaged']
datasplit ['train']
visits [1]
```

Pour conclure, nous avons convertit les colonnes de type catégorie qui veut dire qui contient un nombre fini de valeur possible, comme la colonne « channelGrouping » qui ne contient que ces valeurs :

```
Organic Search
Referral
Paid Search
Affiliates
Direct
Display
Social
(Other)
```

1.3 Feature engenieering

Il est question dans cette section de créer des colonnes nous permettant par la suite de pouvoir les tester sur différents modèles de machine learning.

On a rajouté les nouvelles colonnes « *day* », « *month* », « *dayWeek* » à partir de la colonne « date » qu'on a ensuite supprimé. Il nous n'a pas semblé pertinent de rajouter la colonne année car le dataset comprend les données sur une année du 8 janvier 2016 au 8 janvier 2017.

2. Modèles

Voici les résultats obtenus avec les différents algorithmes :

	Scores
RandomForestClassifier	0.995431
GradientBoostingClassifier	0.970800
BaggingRegressor	0.087525
LinearRegression	0.073179

On peut conclure que le modèle Random Forest est celui qui nous donne la meilleure prédiction, suivi du modèle Gradient Boosting Classifier avec 97% de précision.

3. Conclusion

A l'issue de ce projet, nous avons atteint l'objectif final de proposer une analyse complète du dataset. Les différents modèles à programmer ont été réalisé avec succès. Ainsi, nous ne pouvons que nous satisfaire d'avoir répondu aux attentes du projet et de l'avoir achevé.

Nous sommes heureux d'avoir abouti à ce stade pour une première expérimentation d'un projet de Machine Learning d'environ un mois. Nous avons pu avoir un bref aperçu d'un projet professionnel dans le domaine de la data-science.

De plus, la réalisation de ce projet nous a été plus que bénéfique. Il nous a permis d'en tirer des leçons pour la suite et nous a apporté un réel apprentissage. Dans un premier temps, il nous a permis de développer, d'approfondir, de mieux comprendre et de mettre en pratique les connaissances que l'on a acquis au cours de ce début d'année.

Par ailleurs, ce projet nous a permis de développer nos capacités à gérer notre temps, collaborer et travailler en groupe pour adopter les bonnes méthodes de travail. Ainsi, il a été bénéfique dans la mesure où le travail sur des projets est indispensable dans la plupart des secteurs du monde professionnel.

Cependant, étant encore novice, de nombreuses difficultés ont été rencontrés et de nombreux points peuvent être encore améliorés. Cela nous a permis de nous rendre compte qu'il reste encore beaucoup à apprendre et qu'il me faut accumuler de l'expérience. Les premières difficultés rencontrées ont été en début de projet. Nous nous sommes sentis un peu perdu puisque nous ne savions pas vraiment par où commencer. Grâce aux enseignants et aux cours que l'on nous a donné, on a vite pu être mieux dirigé et débuter notre projet sur de bonnes bases.

Nous avons aussi eu des difficultés dans la compréhension des analyses attendues dans les différentes parties. Enfin, le dernier obstacle fut lors du codage d'un des modèles de prédiction, nous n'avions pas su trouver de solutions convenables à ce problème et qui justement reste encore à améliorer : le choix des features reste encore à améliorer. Concernant le codage le choix des features, nous avons surement dû avoir un manque d'analyse mais j'imagine que c'est un problème qui se résout surtout avec le temps et l'expérience.