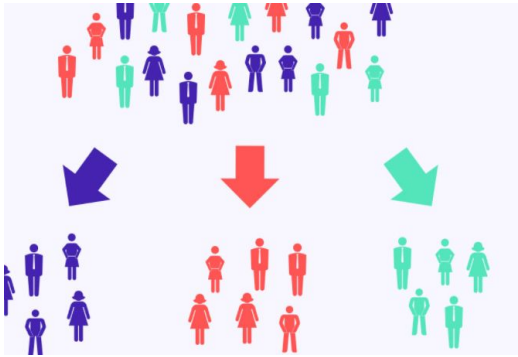
Formation DATA SCIENTIST PROJET

Segmenter les comportements des clients
Soutenance du 18/11/2021

Sommaire

Problématique



Vous êtes en poste chez Datazon, et travaillez au sein de l'équipe marketing, avec d'autres Data Scientists. L'entreprise britannique, leader dans la vente en ligne de toute sorte d'objets, a tout intérêt à mieux comprendre les comportements de ses clients pour augmenter la fréquence d'achat et la valeur du panier moyen.

Votre objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement dans la durée, afin de détecter les plus susceptibles de passer à l'achat.

Les livrables sont :

- Un notebook de nettoyage et d'exploration mis au propre
- Un notebook de test
- Le code Python final
- Un rapport complet

Description des données

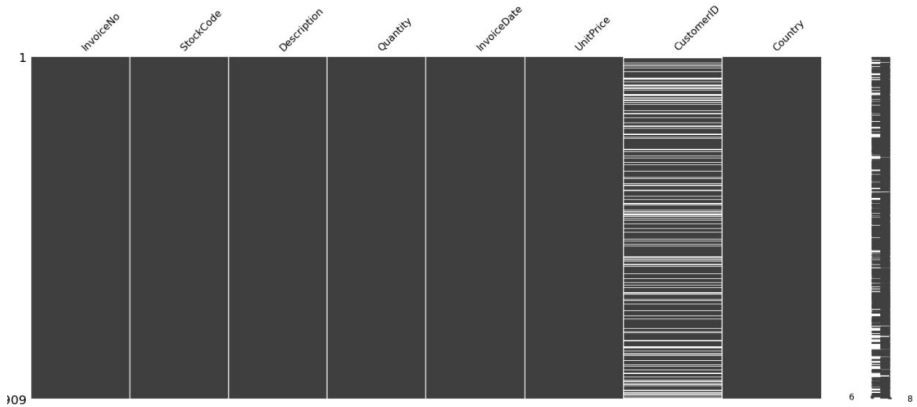
Il y 8 dimensions pour environ 500 000 lignes

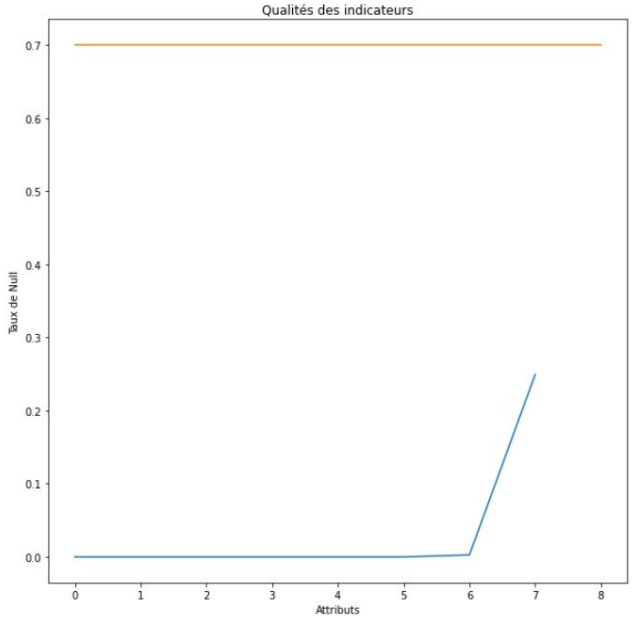
- Les informations de type quantitatives : Il s'agit d'information numérique caractérisant la commande comme par exemple
  a. Le nombre de produit acheté
  b. Le prix du produit acheté
- Les informations de type qualitatives ou catégorielles : Il s'agit d'information de catégorie sur commande
  a. Le pays d'origine du client
  b. Description du produit
- Les informations de date
  a. La Date de la commande
- Les informations d'identification
  a. Identifiant de la commande
  b. Identifiant du client
  c. Identifiant du produit

Taux de remplissage

Constats :

Les données sont très bien renseignées sauf l'identifiant client qui est la donné la plus importante ⇒ Suppression des lignes sans identifiant client





Transformation des types

Dans le but de faciliter les manipulations par la suite, j'ai réalisé les transformations de type suivantes :

- Transformation de la date de la commande en champs DATE avec tri sur cette même dimension
- Transformation des champs quantitatifs en type numérique avec suppression des lignes impossibles à convertir
- Transformation de l'identifiant client en entier avec suppression des lignes impossibles à convertir

De plus j'ai supprimé les enregistrements suivants :

- Certain prix de produit était à 0. Je les ai remplacés par la moyenne d'achat de ce même produit

Création de plusieurs variables

Afin de catégoriser les clients je vais créer trois variables. Ces trois variables sont largement utilisées dans ce domaine d'analyse.

Données	*Traitement*
Récence : Date de la dernière commande (Ou nombre de jours depuis la dernière commande	Création d'une fonction qui calcul le nombre de jours depuis la dernière commande
Fréquence = Fréquence du nombre de commande	Nombre de commande sur la période
Montant = Montant des achats réalisés	Sommes des achats réalisés sur la période
Nombre de produits achetés en moyenne **(Ajout personnel)**	Nombre de produits achetés en moyenne par commande

Données Catégorielles

J'ai créé aussi des variables suivant le moment de la semaine et de la journée ou la commande était réalisée

Données	*Traitement*
Moment de la journée	Création d'une fonction qui calcul le moment de la journée de passage de la commande : Matin (5h-11h), Midi(11h-14h), Après-midi (14h-19h), soir(19h-23h)
Moment de la semaine	Création d'une fonction qui calcul le moment de la semaine de passage de la commande : Semaine(Lundi, Mardi, Mercredi, Jeudi, Vendredi), Week-end(Samedi, Dimanche)
Pays d'origine	One Hot encoding ⇒ **Finalement pas utilisé car non significatif avec la majorité des commandes en Angleterre**.

Agrégation au niveau client

Les informations que nous avons récoltées doivent maintenant être agrégées au niveau client afin de pouvoir créer nos catégories

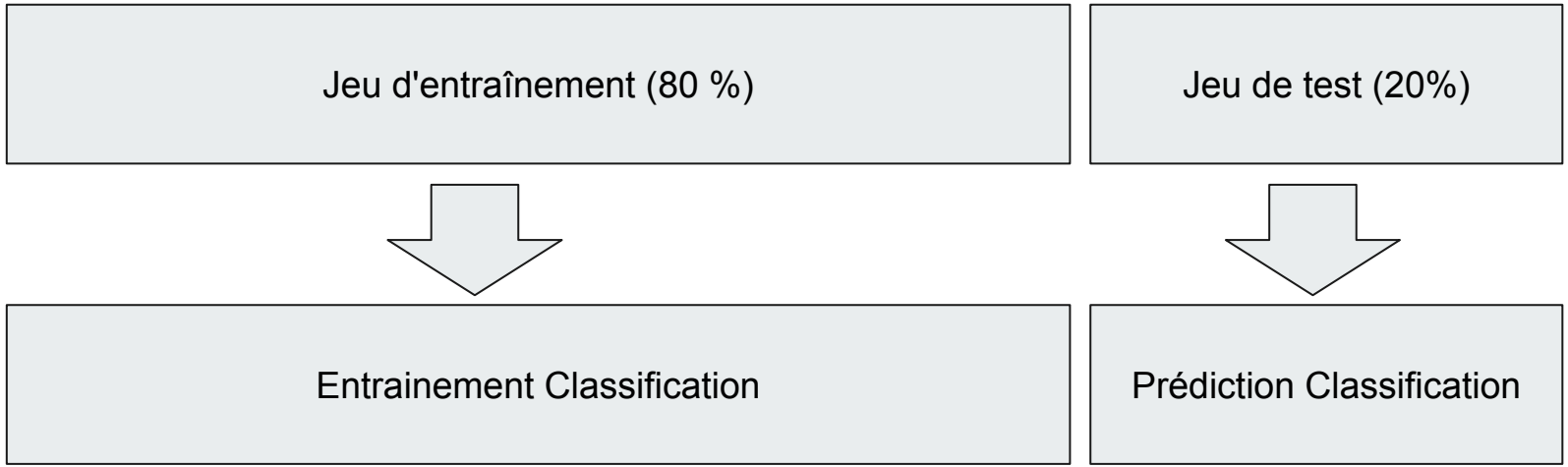
Données	*Traitement*
Nombre de commande en semaine	Calcul le nombre de commandes effectuées la semaine
Nombre de commande en week-end	Calcul le nombre de commandes effectuées le week-end
Nombre de commande le matin	Calcul le nombre de commandes effectuées le matin
Nombre de commande le midi	Calcul le nombre de commandes effectuées le midi
Nombre de commande l'après midi	Calcul le nombre de commandes effectuées l'après midi
Nombre de commande le soir	Calcul le nombre de commandes effectuées le soir

Synthèse

Nous avons donc constitué un jeu de données avec

- Avec **4372** clients
- 11 variables : Identifiant client, R, F, M, Moyenne des quantités, nbre commande en semaine, nbre commande le week end, nbre commande la matin, nbre commande le midi, nbre commande l'après midi, nbre commande le soir

Création jeu d'entraînement et jeu de test



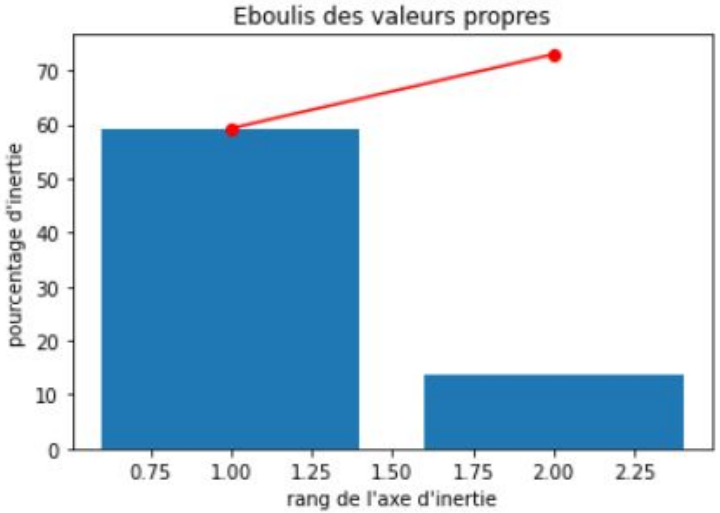
Standardisation

- Réalisation d’une standardisation à partir du jeu d'entraînement uniquement
- Utilisation de la librairie Sklearn **preprocessing.StandardScaler**

ACP

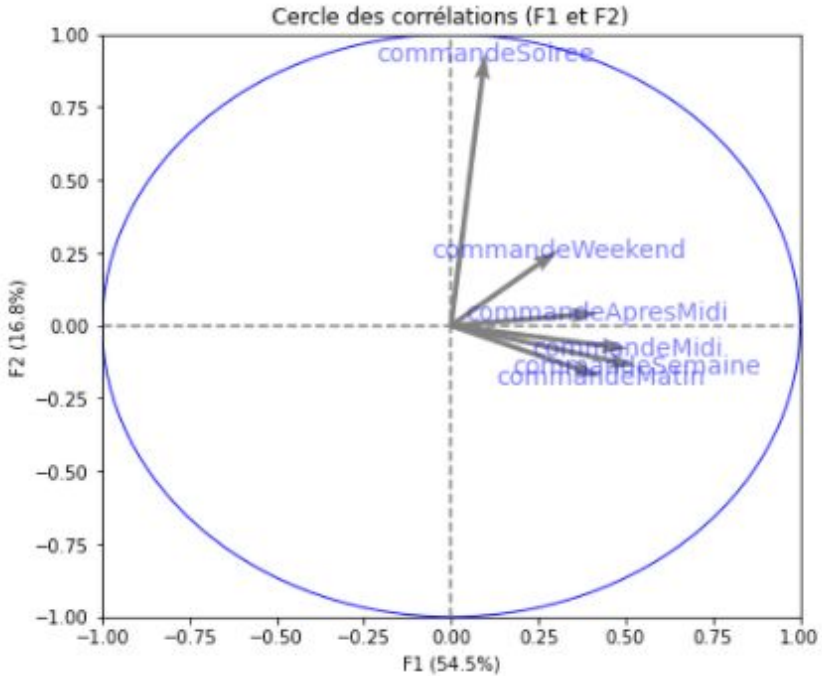
Afin de limiter l'influence des données catégorielles sur les indicateurs RFM. J'ai réalisé une ACP avec les données suivantes :

- Commande Semaine
- Commande Week end
- Commande Matin
- Commande Midi
- Commande Après midi
- Commande Soir



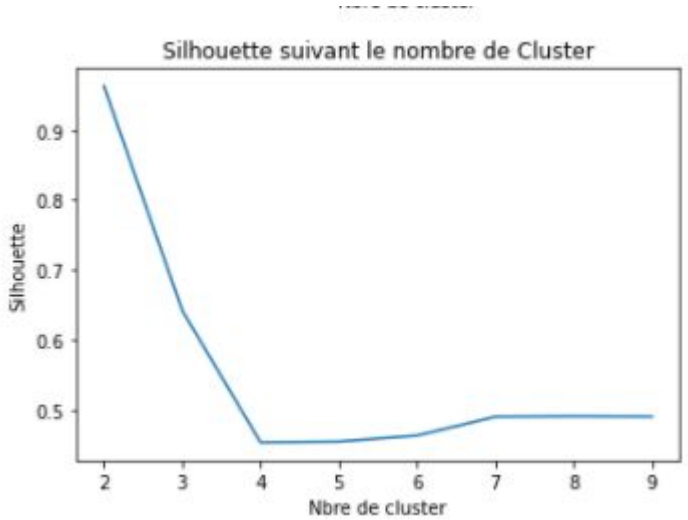
Réalisation d'un Éboulis pour identifier le nombre de composantes à conserver⇒ **Conservation de deux composantes**

ACP


Cercle des corrélations (F1 et F2)
commandeSoiree
commandeWeekend
commandeApresMidi
commandeMidi
commandeSemaine
commandeMatin
F1 (54.5%)
F2 (16.8%)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Catégorisation

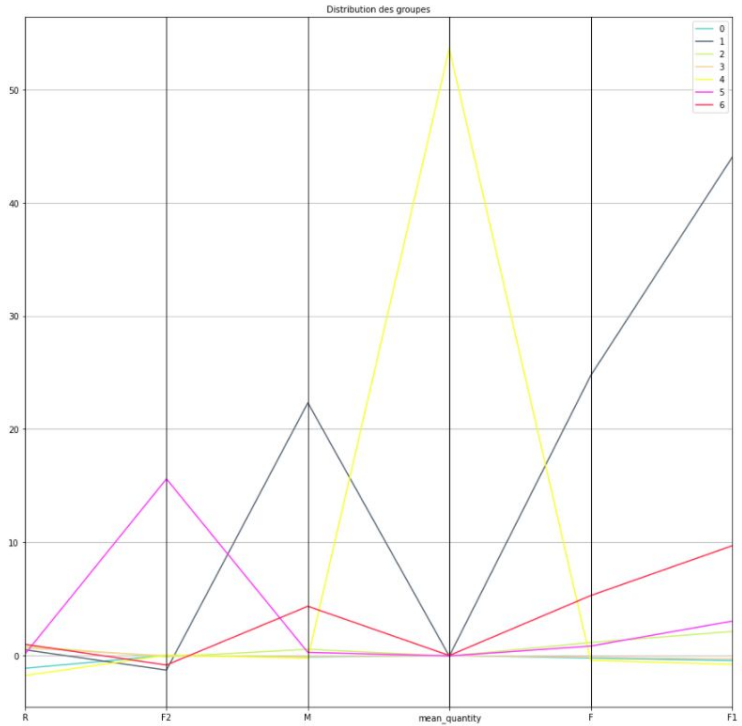
Avant de commencer à lancer les différents modèles il a fallu catégoriser les clients.
Pour cela j'ai utilisé un K_Means.
Afin d'identifier mon nombre de cluster j'ai réalisé un graphique afin de le détecter ⇒ **J'ai conservé 7 clusters** et fusionné après coup les 4 plus petits



Sur les données de test afin de les catégoriser j'ai fait une prédiction

Interprétation des groupes

Groupe	*Interprétation*
3 : 1809 Individus	Client moyen n'ayant pas commandé récemment
0 : 1298 Individus	Clients moyen ayant commandé récemment
2 : 355 individus	Clients qui commande fréquemment et pour un bon montant. **Bon client**
3 : 35 Individus. Fusion des groupes 5, 1, 4 et 3	Clients qui commandent beaucoup de produit ou très chère ou le week end. **Très bon client**



Modélisation

- Une fois la catégorisation obtenue j'ai pu tester différents modèles afin d'évaluer le plus pertinent
  - DummyRegressor pour la baseline
  - Régression linéaire avec coefficient Ridge
  - Arbre de décision

Modèles

Synthèse des résultats obtenus

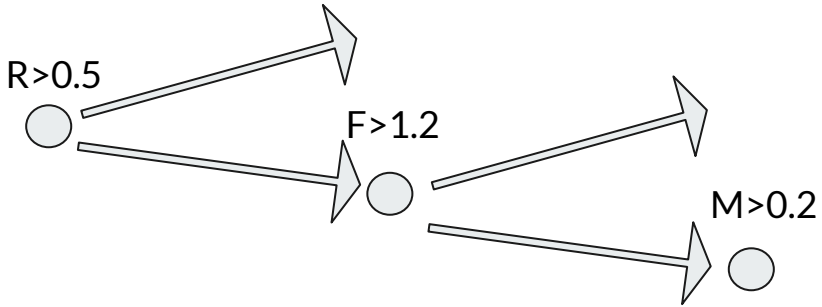
Modèle	Précision
Dummy Regressor	-0.007
Régression linéaire avec coefficient Ridge	0.882
Arbre de décision	0.982

Modèles

Focus sur l'arbre de décision

Avantages	Inconvénients
• Algorithme facile à comprendre • Résultats interprétables • Temps d'exécution raisonnable	• Faible performance • Risque de sur apprentissage

Un arbre de décision permet d'expliquer une variable cible à partir d'autres variables dites explicatives. Dans notre cas présent il s'agit de prédire la classification en fonction des autres variables (RFM, F1, F2)



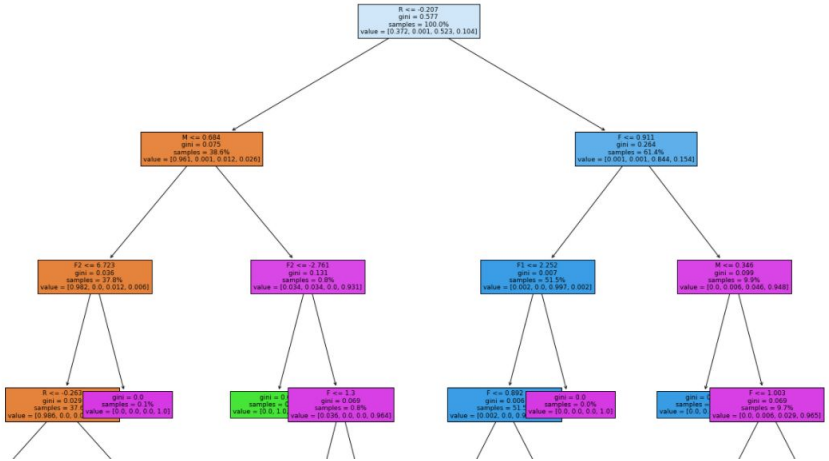
Hyperparamètre	Explication
Profondeur de l'arbre	Profondeur maximum de l'arbre
Nombre min d'observation dans un noeud pour le diviser	Le nombre minimum pour diviser un noeud
Nombre maximum de noeud terminaux	Le nombre maximum de noeud terminaux
Nombre min d'observation dans un noeud final	Le nombre minimum que l'on souhaite dans un noeud final

Graphique de l'arbre de décision

Comme nous pouvons le voir un arbre de décision est facilement interprétable.

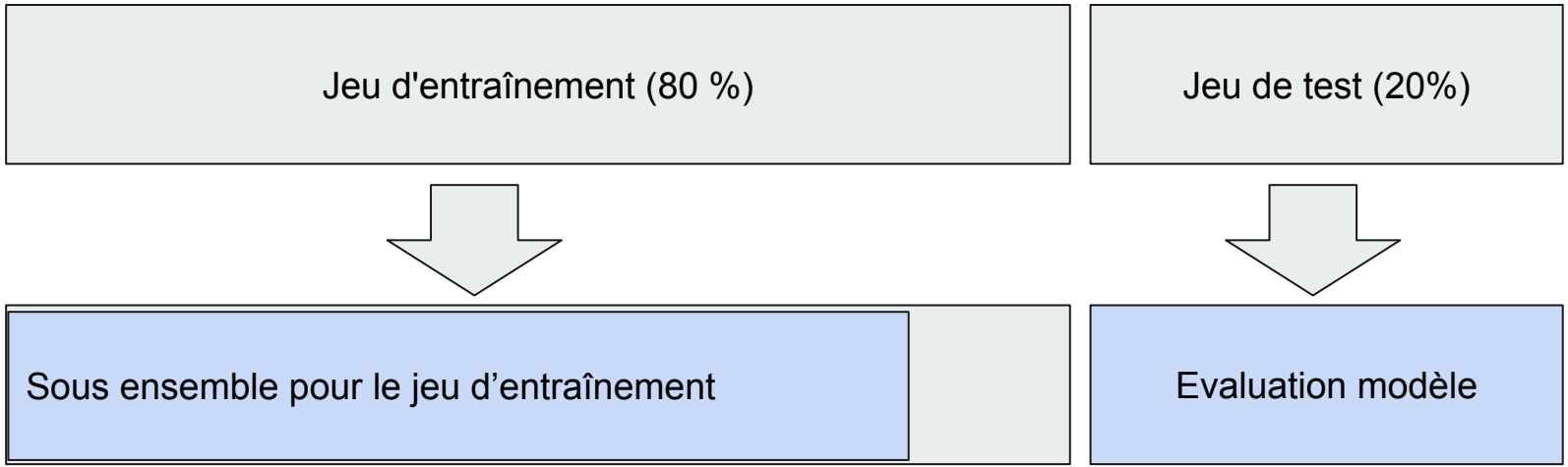
Le graphique suivant n'est qu'une sous partie de notre arbre complet.

un affichage moins graphique est aussi possible afin de mieux l'interpréter



Modèles : Découpage jeu d'entraînement

Un élément important du cahier des charges était d'évaluer l'importance de la création du jeu de test. En effet suivant ce découpage les résultats peuvent être très différents. Pour cela j'ai pris un sous ensemble de mon jeu d'entraînement afin de construire le modèle. La vérification se fait toujours sur l'ensemble du jeu de test qui lui n'a pas bougé



Modèles : Découpage jeu d'entraînement

Voici les résultats obtenus

Constitution du sous ensemble	Résultat
Utilisation de la dimension classification pour faire un découpage stratifié (Conservation des proportions de nos groupes de client)	Résultat identique
Prendre les 2800 première lignes avec tri sur la date de commande	Résultat identique
Prendre les 2800 premiers clients ordonnés par la variable date de dernière commande	0,66 ⇒ Très mauvais résultat
Prendre les 2800 premiers clients ordonnés par la variable montant commandé	0,96 ⇒ Moins bon résultat

Conclusion

Ce projet m'a permis de bien comprendre la notion de modèle supervisé ou non supervisé avec son approche, Classification puis prédiction.

En effet la première partie a permis de créer un label donc des cluster de clients. Et la deuxième partie de prédire ce label. Faire cela dans un même projet m'a aidé à faire cette distinction.

Le deuxième élément que je retiens de ce projet est le découpage du jeu de test. Celui-ci est primordiale et influence grandement le résultat final avec des risques de sur apprentissage. Il est donc indispensable de prendre le temps de découper son jeu d'entraînement et de test à travers une analyse rapide métier. Ceci afin de permettre d'avoir une représentation fidèle de la réalité. C'est aussi une grande difficulté.