Segmentez les comportements de clients

Projet OpenClassrooms présenté par F. Kossi

Livrables déployés sur GitHub

Introduction

L'objet de ce projet est de pouvoir catégoriser les clients du site de e-commerce Datazon dès leur premier achat.

Je dispose à cet effet du fichier des commandes d'environ 4000 clients, du 12/01/2010 au 10/12/2011, caractérisées ainsi :

- InvoiceNo : Identifiant unique de chaque transaction sur 6 chiffres. Nominal, s'il commence par 'c', il s'agit d'une annulaton.
- StockCode : Code produit unique sur 5 chiffres. Nominal.
- Description : Nom du produit. Nominal
- Quantity : Quantité de produit pour chaque transaction. Numérique.
- InvoiceDate : Date et heure de la commande. Numérique.
- UnitPrice : Prix unitaire en livre sterling. Numérique.
- CustomerID : Identifiant unique client sur 5 chiffres. Nominal.
- Country: Nom du pays de résidence des clients. Nominal.

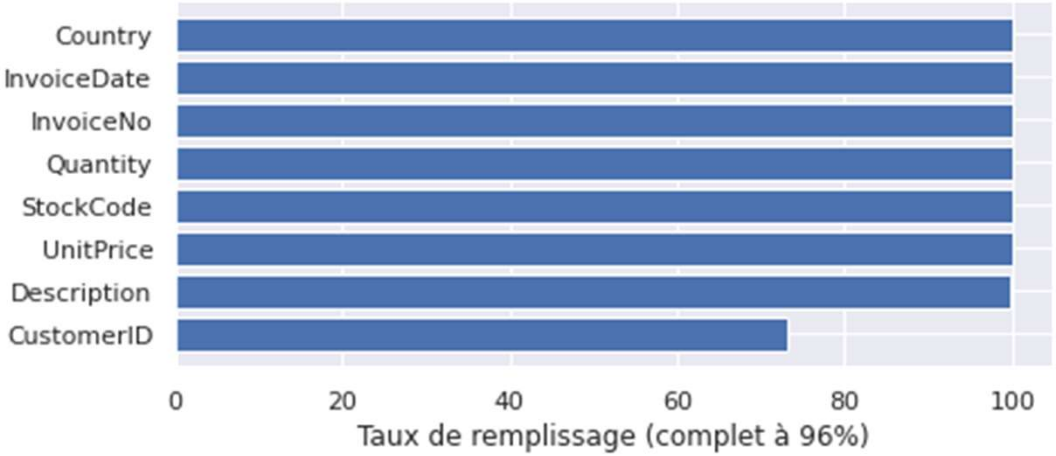
Préparation

❑ Contenu du jeu de données

produits	transactions	clients
4070	25900	4372

❑ Jeu de données rempli à 96%

Je supprime les 135k enregistrements dont les CustomerID sont absents car c'est la clé de ce projet



❑ 5152 doublons (toutes les variables sauf Country et Description) supprimés

Préparation

❑ Pays

Je conserve les clients britanniques, majoritaires dans le jeu de données. Par ailleurs, les pratiques d'achat sont propres à chaque pays/culture.

Pays	Nombre
United Kingdom	495478
Germany	9495
France	8557
EIRE	8196
Spain	2533

❑ 7218 commandes annulées

Plusieurs itérations me permettent de diminuer les commandes correspondantes (client, produit, prix unitaire) si la quantité annulée est plus grande que la commande. Sinon, je diminue la commande annulée de la quantité de la commande associée. In fine, je supprime les 2417 commandes annulées non réconciliées.

Préparation

❑ Codes produit

Des codes produits ne sont pas sur 6 chiffres (cf. caractéristiques en introduction). Il s'agit visiblement de transaction spéciales (rabais, charges, …) à l'initiative de DataZon.

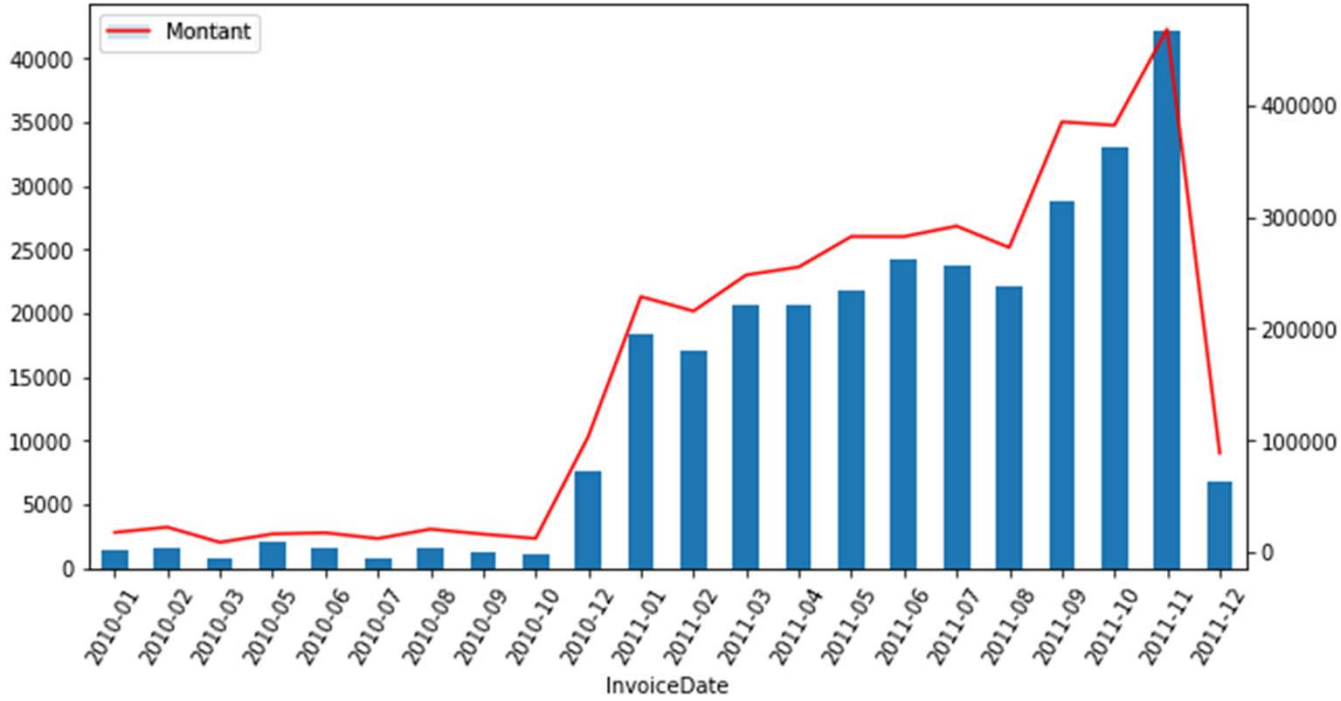
Je supprime ces enregistrements car ce ne sont donc pas de réelles transactions client.

StockCode	Description	Nb
M	Manual	380
POST	POSTAGE	86
D	Discount	74
C2	CARRIAGE	30
DOT	DOTCOM POSTAGE	16
CRUK	CRUK Commission	16
BANK CHARGES	Bank Charges	12
PADS	PADS TO MATCH ALL CUSHIONS	4

❑ Calcul des montants de transactions : TotalPrice = Quantity x UnitPrice

Exploration

❑ Réduction temporelle



Les données antérieures à Janv. 2010 ou postérieures à Nov. 2011 ne semblent pas pertinentes ou complètes. Je réduis donc mon jeu de données à la période de Janv. 2010 à Nov. 2011.

Exploration

❑ Feature engineering

Pour chaque client, j'agrège les informations suivantes en m'inspirant de la démarche marketing RFM pour agréger les données par client, soit :

- sa récence (R), durée depuis sa dernière commande (recency)
- sa fréquence (F), nombre de commandes passées sur la période observée (frequency)
- le montant (M) global de ses transactions (monetary_value)
- la durée entre le 1° et le dernier achat (duration)
- le nombre de commandes distinctes (orders)
- le nombre de produits total achetés par un utilisateur (pdt_mean)
- la quantité moyenne par transaction (line_mean)
- le nb moyen d'articles différents par achat (pdt_mean)
- le prix moyen des produits achetés par transaction (price_mean)
- le nombre de produits différents achetés

Notons qu'à ce date, **les commandes uniques représentent 36.22%** des enregistrements (panier moyen de 324.60£).

Exploration

❑ Etiquetage RFM

L'idée est de pouvoir comparer la segmentation proposée par mon étude avec le « standard » R/F/M.

J'attribue donc une note de 1 (meilleur) à 4 (pire) selon les quartiles de chaque attribut. Puis, fonction de la somme de ces notes, je crée les 4 segments de clients suivants :

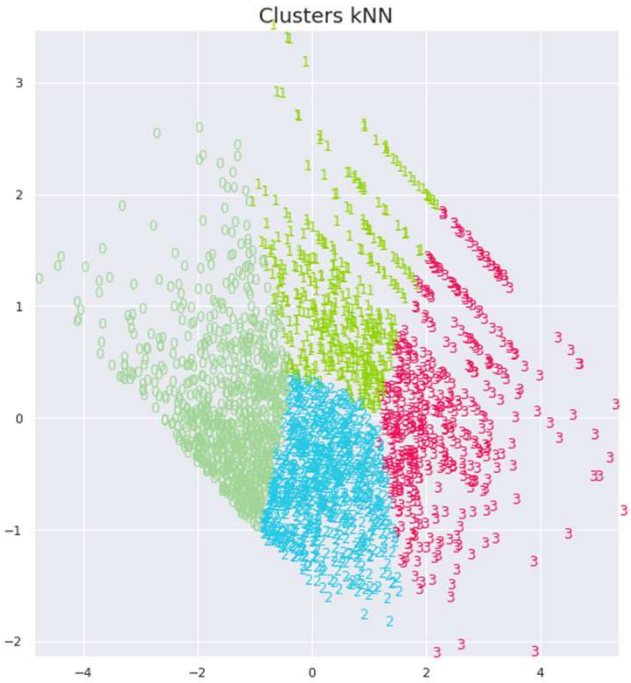
- Platinum si R+F+M = 3
- Gold si 4 <= R+F+M < 6
- Silver si 6 <= R+F+M < 10
- Bronze si R+F+M >= 10

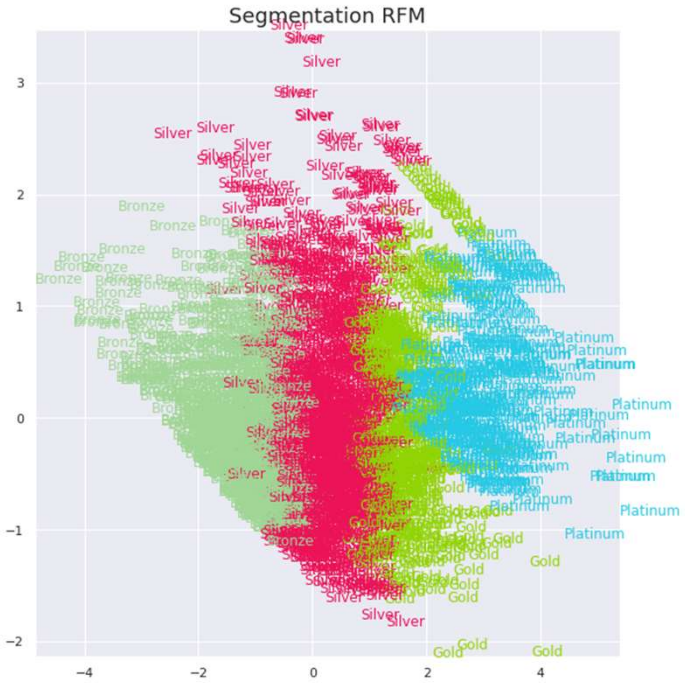
Clustering

❑ KMeans

Après PCA en 2 dimensions des variables, je visualise un Knn de 4 clusters (cf. catégories de clients créées). Les clusters calculés (à gauche) présentent des similitudes sur les catégories extrêmes (0/Bronze et 3/Platinium).

Les différences sur les valeurs centrales sont probablement liées à un défaut de granularité de mon étiquetage RFM. Par exemple, la note de 6 (Silver) pourrait être composée de 2+2+2 (OK) mais aussi de 1+1+4 qui avec 2 « top » attributs tendrait vers Gold.





Clustering

❑ Etiquetage clustering

J'affecte les étiquettes suivantes en fonction des moyennes des 3 variables, importance donnée aux montants des commandes

cluster	nombre	recency	frequency	monetary_value
3	461	16.34	280.98	6683.72
2	841	108.49	68.65	1165.83
1	460	12.00	48.96	729.35
0	877	156.51	13.50	273.43

Label
Platinum
Gold
Silver
Bronze

Classification

- Comparaison des classifieurs avec des jeux d'entrainement et de test définis aléatoirement (resp. 70% et 30% du total) ou temporellement (resp. Janv.-Août 2011 et Sept.-Nov.2011)

Classifieur	Precision	
	Random split	Time split
k-Nearest Neighbors	97.44 %	96.98 %
Support Vector Machine Classifier (SVC)	97.00 %	94.30 %
Random Forrest	97.08 %	95.85 %

- Je choisis donc de baser mon modèle définitif sur le k-NN, visiblement plus performant quelle que soit la clé de répartition train/test

Conclusion

Ce projet permet de catégoriser les clients Datazon en 4 familles : Platinum, Gold, Silver et Bronze en se basant sur leurs récences, leur nombres et montants d'achats.

Le modèle implémenté est le k-Nearest Neighbors, visiblement plus performant (précision > 97%).

Une segmentation des produits achetés améliorerait probablement les résultats obtenus ici.