Projet 5 : Segmentation du comportement clients > SOMMAIRE

> Choix des caractéristiques

- Objectifs et enjeux
- Présentation des variables et nettoyage des données
- Protection des données
- Piste de modélisation et choix des caractéristiques
- Présentation du modèle collaboratif et choix des hyper-paramètres adaptés

> Segmentation / prédiction

- Apprentissage non supervisé : méthode
- Segmentation et interprétation (modèle collaboratif)
- Segmentation > 3 scénarios, choix du nombre de clusters / interprétation des clusters / visualisation (modèle manuel)
- Apprentissage supervisé : méthode
- Présentation des algorithmes de classifications
- Performances de différents scénarios
- Prédiction finale

Problématique client > Segmenter les utilisateurs ? Pourquoi ?

Postulat : Objectif d'une entreprise > Faire du profit

Méthodes directes ou indirectes : améliorer la visibilité ou l'image de l'entreprise investir pour créer les nouveaux besoins, améliorer l'environnement de travail pour augmenter la productivité etc...)

- **Coûts structurels (entretient fonctionnement, équipements etc...)**
- **Investissement (développement de nouvelles gammes de produits, recherche et développement...)**
- **Marketing(publicité, démarchage, réductions etc...)**

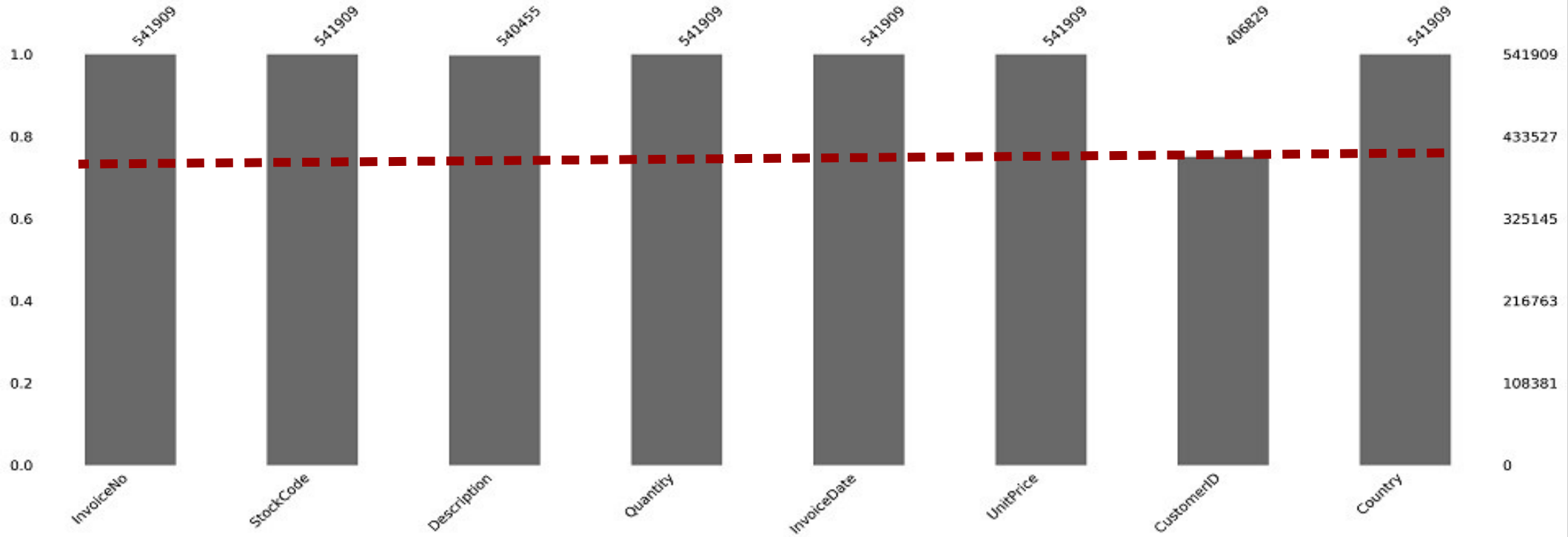
- Ces coûts sont très rarement indépendants des items proposés et de la nature des consommateurs.

Connaître les utilisateurs > Prédire leurs comportements futurs > établir des groupes d'utilisateurs caractéristiques > adapter la gouvernance, les stratégies de l'entreprise pour satisfaire les objectifs.

Présentation des variables

- Variables peu nombreuses et de deux genres : qualitatives et transactionnelles
- Variables qualitatives difficilement utilisables
- Variables transactionnelles : prix par unité, nombre d'unités, origine et date.
- Plus de 4250 items pour prés de 4000 utilisateurs
- La variable « utilisateur » n'a de sens que lorsque l'utilisateur est identifié

fig-1: Occurence en valeurs non manquantes des variables du jeux de donnée

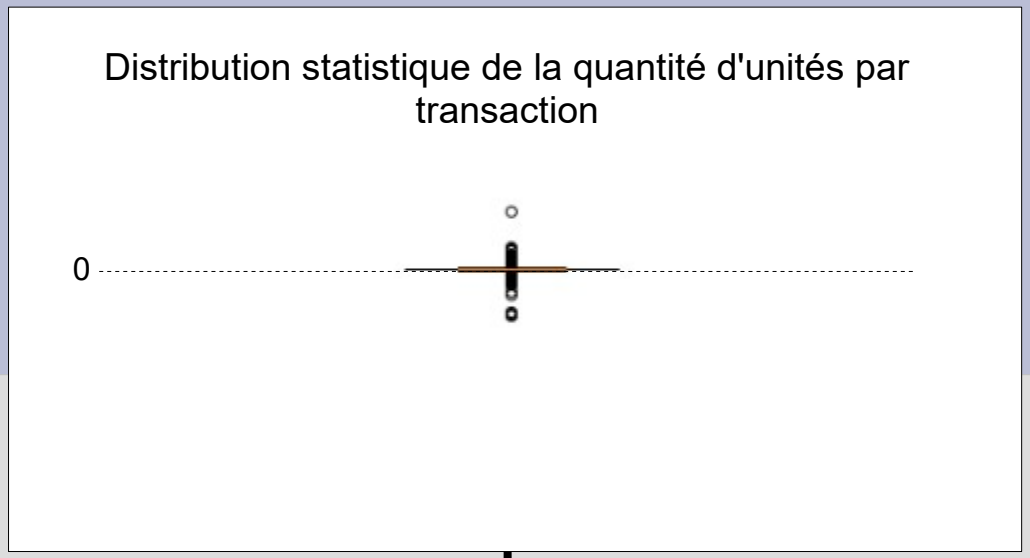


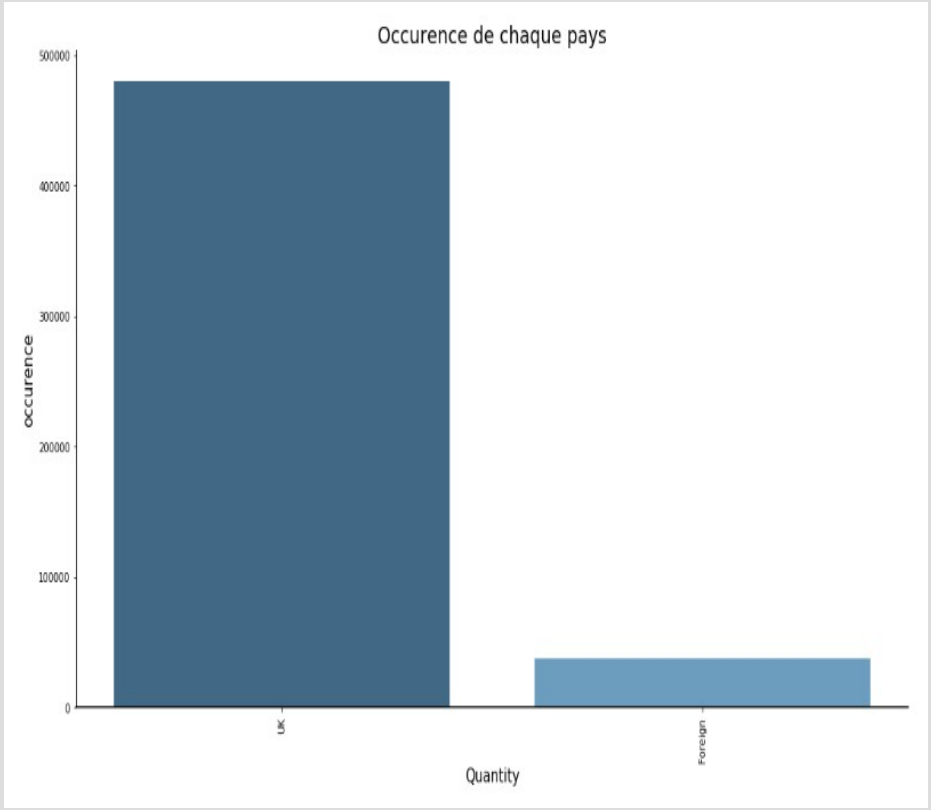
Nettoyage des données

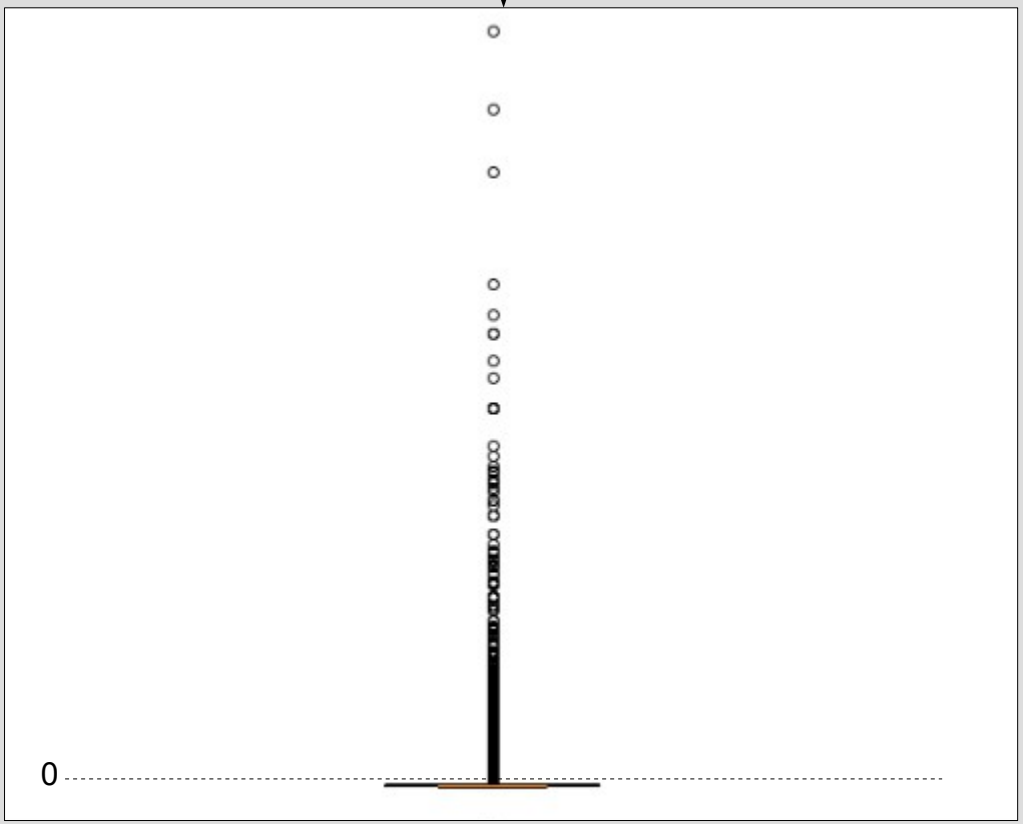
Présence de quantités négatives

Présence de données aberrantes

Traitement de la variable origine (prise en compte des clients anglais et des clients « étrangers »)

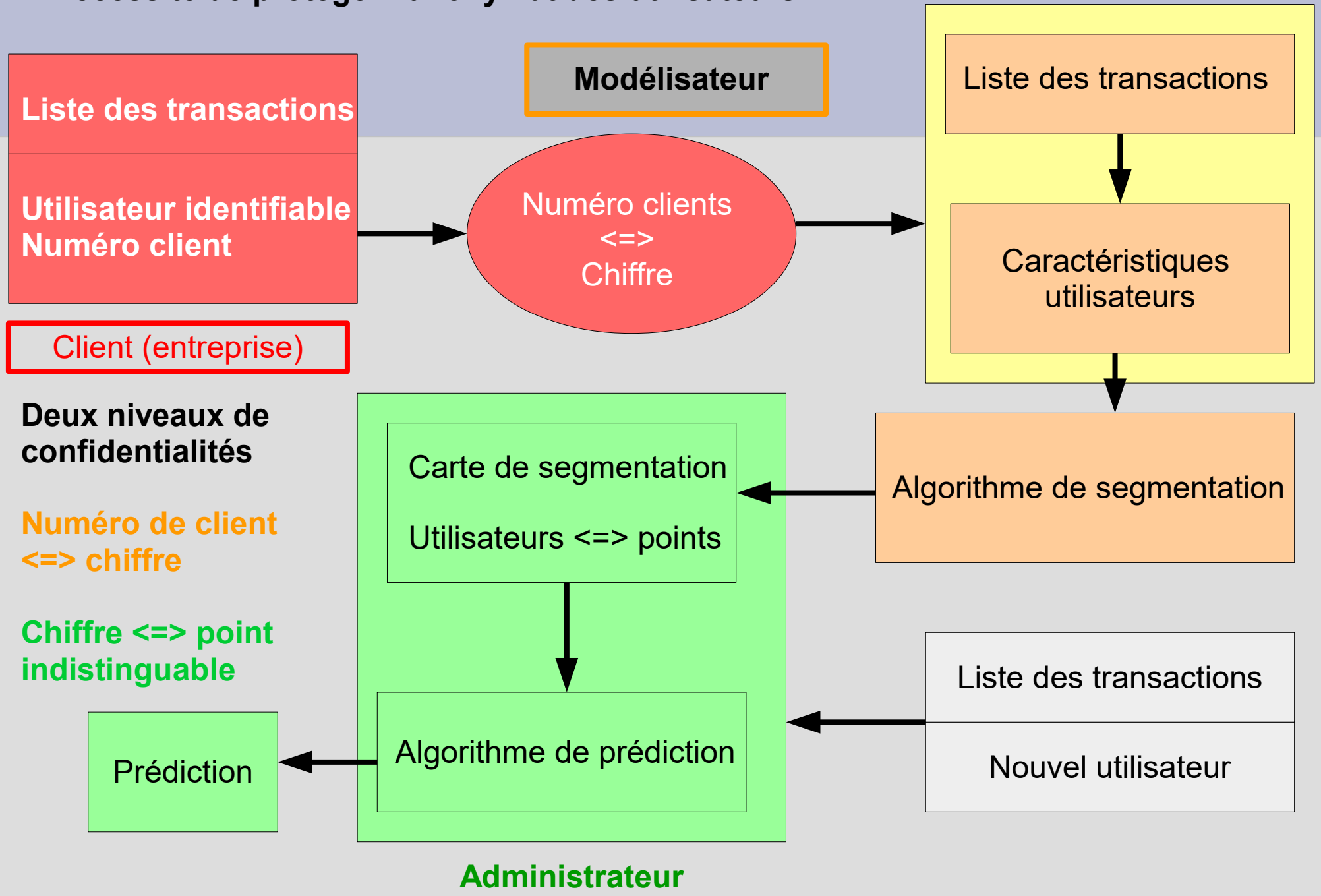




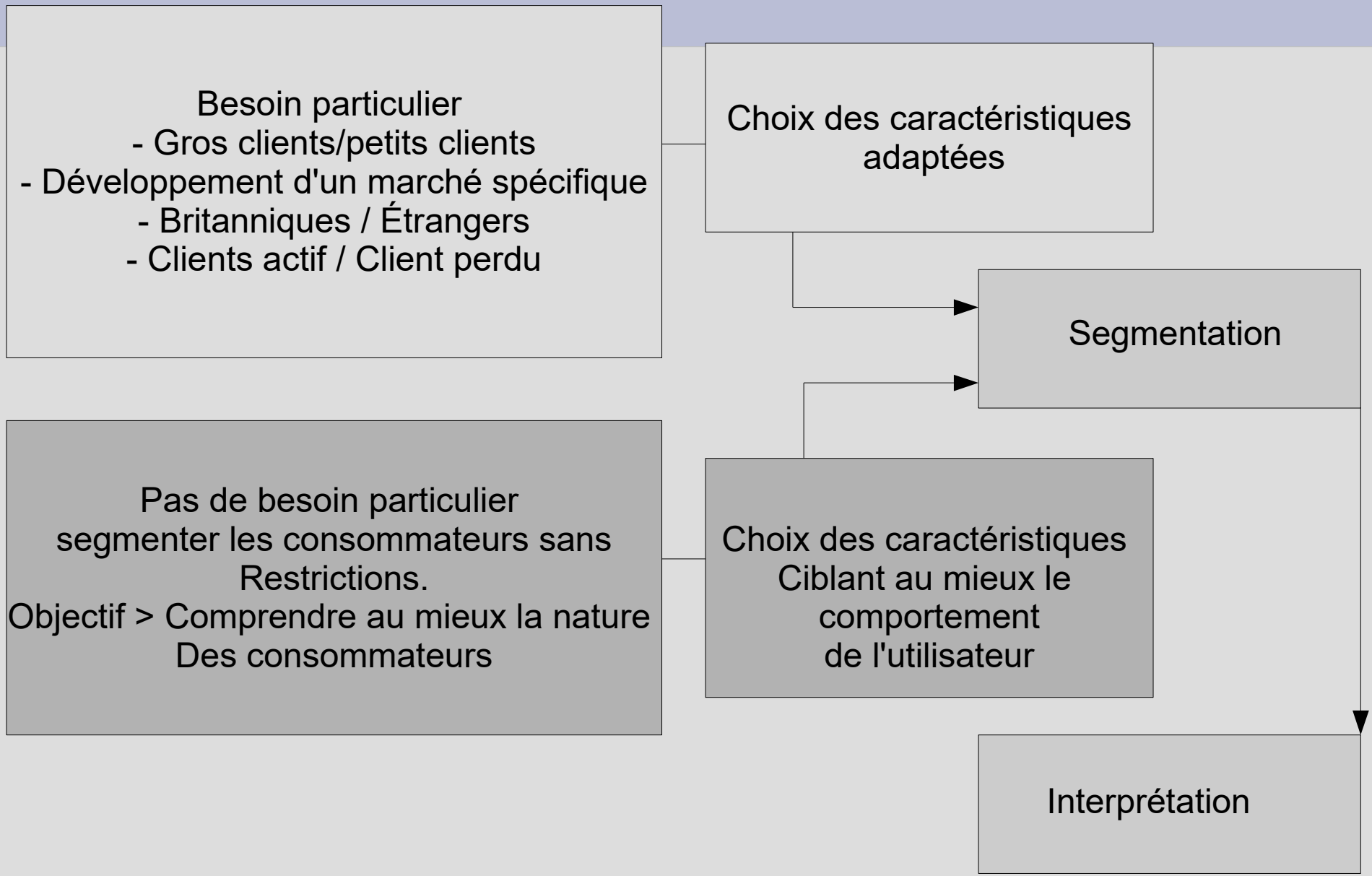


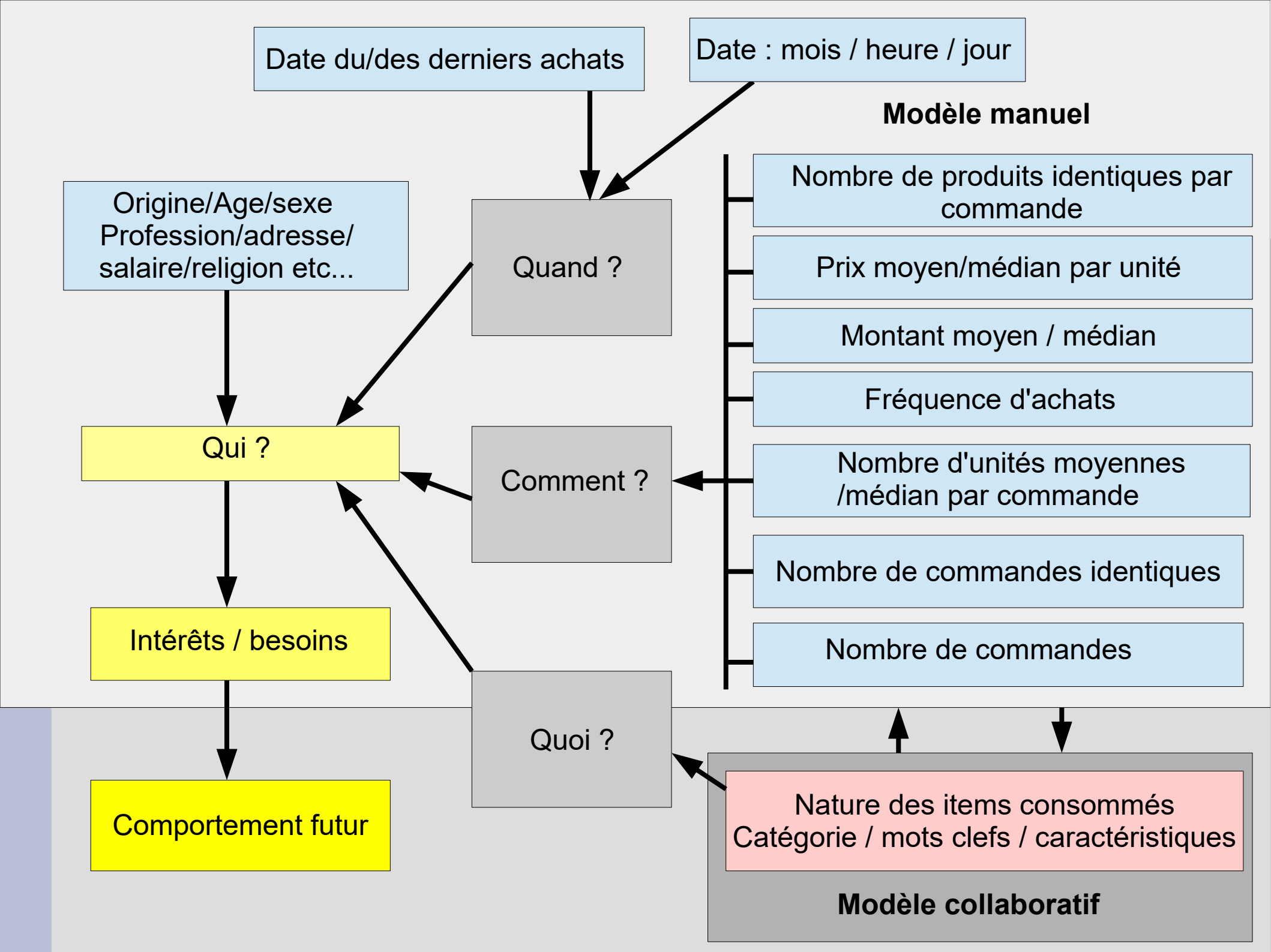
Confidentialité

- **Nécessité de protéger l'anonymat des utilisateurs**

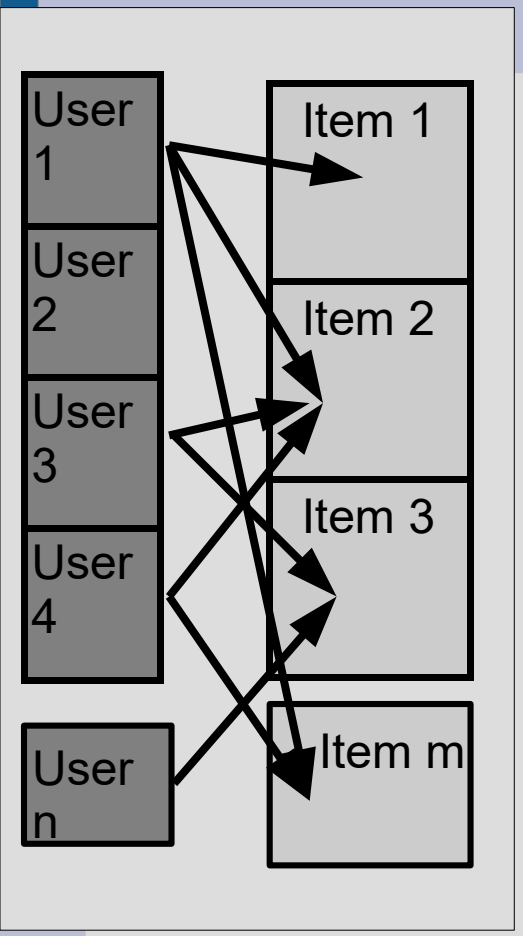


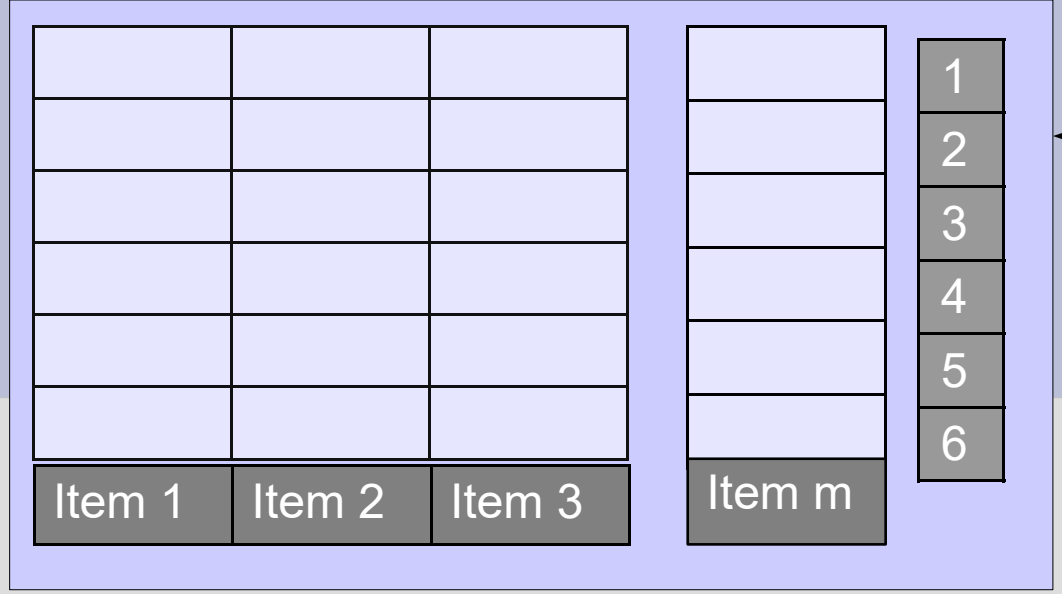
Besoins de l'entreprise > Problématique spécifique / Comprendre les dynamiques d'achats




Date du/des derniers achats
Date : mois / heure / jour
Modèle manuel
Origine/Age/sexe
Profession/adresse/
salaire/religion etc...
Quand ?
Nombre de produits identiques par commande
Prix moyen/médian par unité
Montant moyen / médian
Fréquence d'achats
Qui ?
Comment ?
Nombre d'unités moyennes /médian par commande
Nombre de commandes identiques
Intérêts / besoins
Nombre de commandes
Quoi ?
Comportement futur
Nature des items consommés
Catégorie / mots clefs / caractéristiques
Modèle collaboratif

- **Caractéristiques Modèle collaboratif**

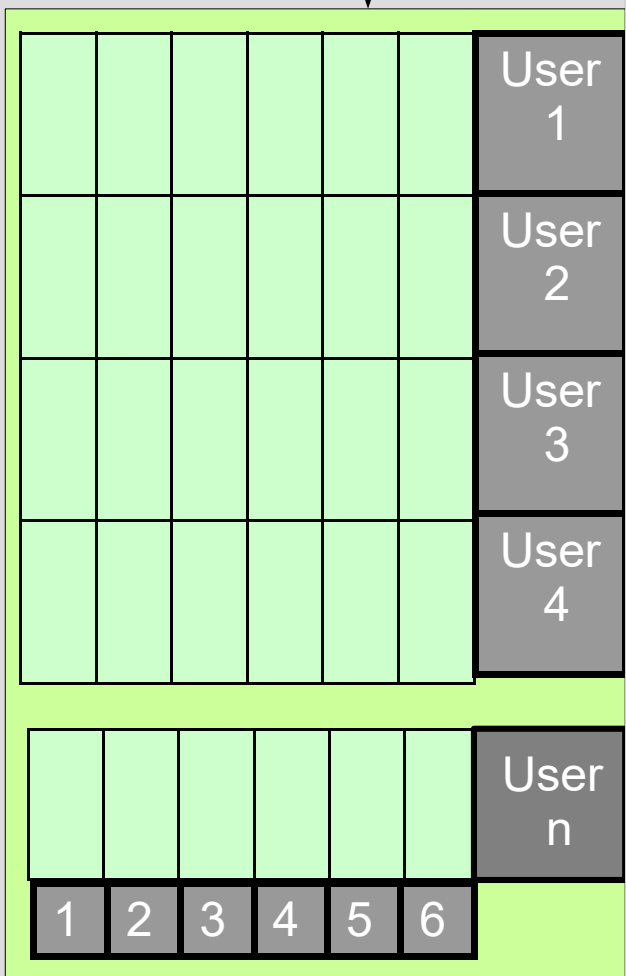




- Vecteurs Caractéristiques Items

- Vecteurs Caractéristiques utilisateurs

$\hat{R}\_i1u1$ R_i1u1	$\hat{R}\_i2u1$ R_i2u1	$\hat{R}\_i3u1$ R_i3u1	$\hat{R}\_imu1$ R_imu1
$\hat{R}\_i1u2$ R_i1u2	$\hat{R}\_i2u2$ R_i2u2	$\hat{R}\_i3u2$ R_i3u2	$\hat{R}\_imu2$ R_imu2
$\hat{R}\_i1u3$ R_i1u3	$\hat{R}\_i2u3$ R_i2u3	$\hat{R}\_i3u3$ R_i3u3	$\hat{R}\_imu3$ R_imu3
$\hat{R}\_i1u4$ R_i1u4	$\hat{R}\_i2u4$ R_i2u4	$\hat{R}\_i3u4$ R_i3u4	$\hat{R}\_imu4$ R_imu4
$\hat{R}\_i1un$ R_i1un	$\hat{R}\_i2un$ R_i2un	$\hat{R}\_i3un$ R_i3un	$\hat{R}\_imun$ R_imun



Présentation (mathématique) de l'algorithme et choix des hyper paramètres

Nombre de transactions utilisateur > item

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases}$$

$$\min_{x_\star, y_\star} \sum_{u,i} c_{ui} \left(p_{ui} - x_u^T y_i \right)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

$$c_{ui} = 1 + \alpha r_{ui}$$

Degrés de confiance

Régularisation

hyper-paramètres

vecteurs utilisateurs

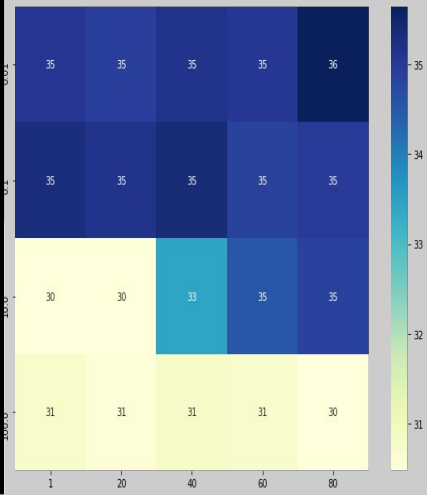
$$x_u = (Y^T C^u Y + \lambda I)^{-1} Y^T C^u p(u)$$

vecteurs items

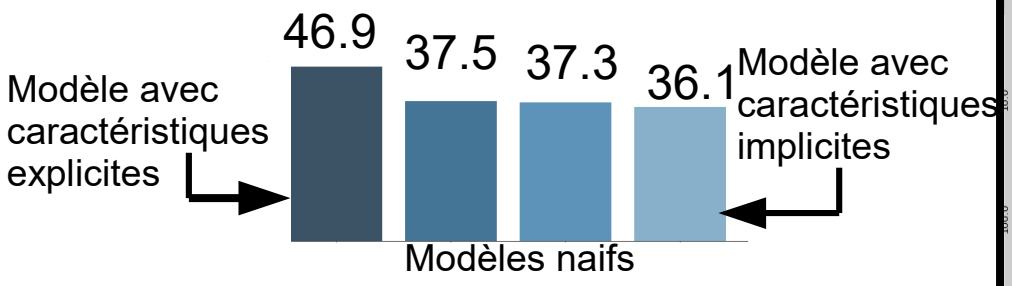
$$y_i = (X^T C^i X + \lambda I)^{-1} X^T C^i p(i)$$

- Meilleurs vecteurs utilisateurs > Meilleure prédiction
- Nécessité de définir un score de prédiction
- Définition des jeux de test et d'entraînement

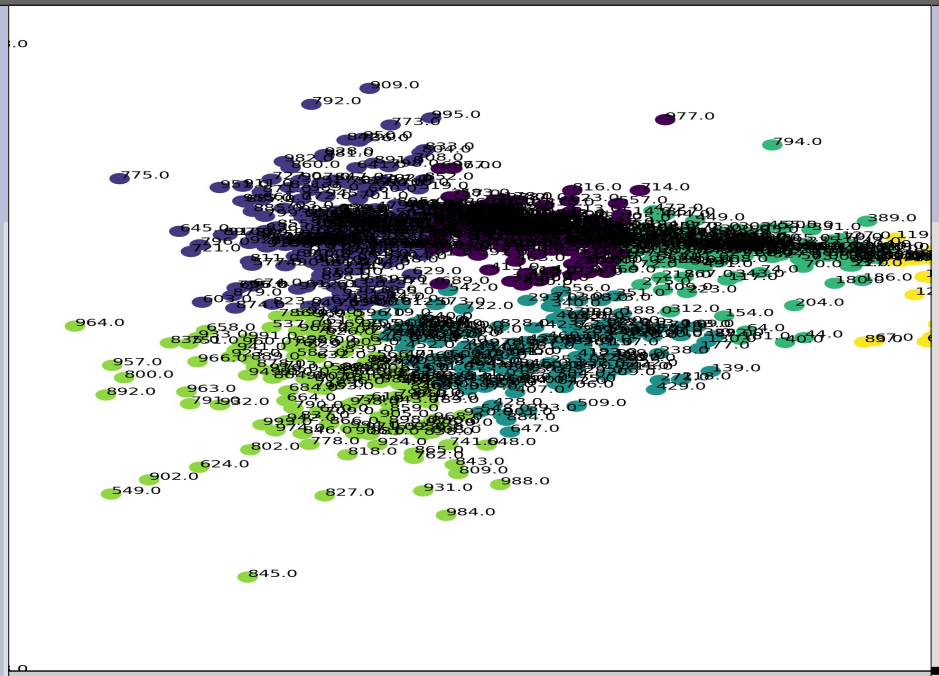
Détermination des hyper-paramètres par validation croisée sur grille



> Comparaison des modèles
Algorithme de prédiction manuel
distance item/barycentre des transactions dans l'espace items



Clusters et interprétation : modèle collaboratif



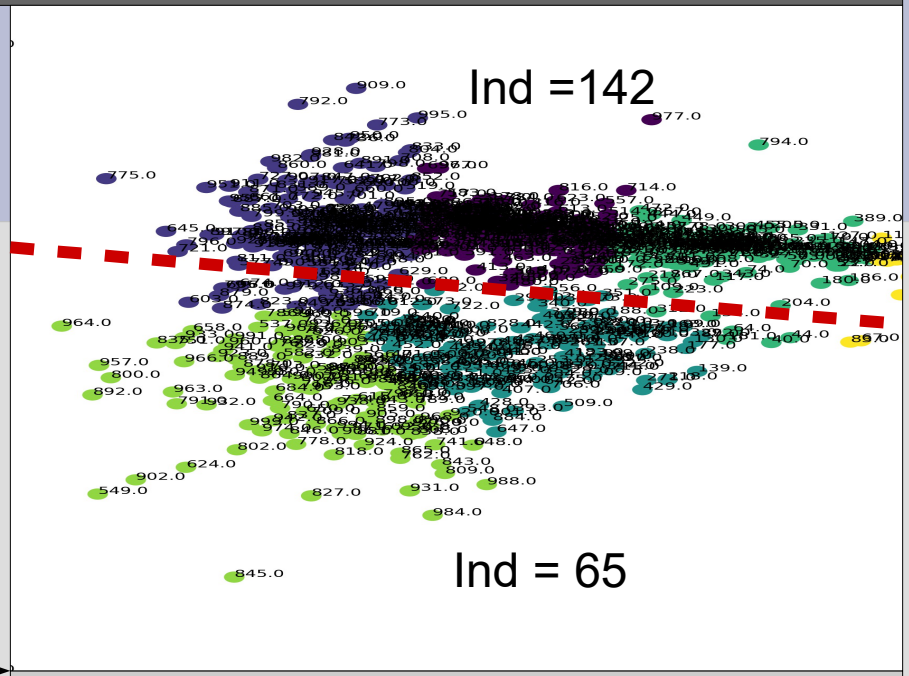
Visualisation de l'historique des utilisateurs dans un cluster donné

↓

Interprétation qualitative périlleuse
Clients <=> entreprises concurrentes

↓

Composante caractérisant le nombre de transactions effectuées



Définition d'un indice de popularité > « Ind »

Le modèle est peu performant pour prédire les comportements utilisateurs

Le modèle est peu performant pour prédire les comportements utilisateurs > mais fait émerger une caractéristique item basée sur la popularité

Choix des variables et features engineering

Prendre en compte la distribution statistique des transactions

Transactions :

- ~~**Nombre de transactions**~~
- **Fréquence de transaction**
- **Montant moyen par transaction**
- **Nombre d'unités moyennes par transaction**
- **Prix moyen unitaire par transaction**

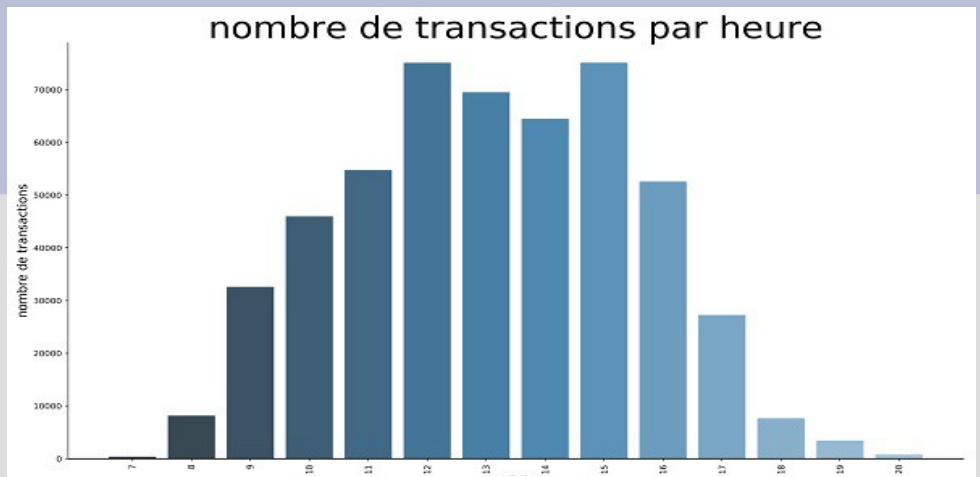
- Nombre de produit différent par commandes
- Nombre de transaction identiques

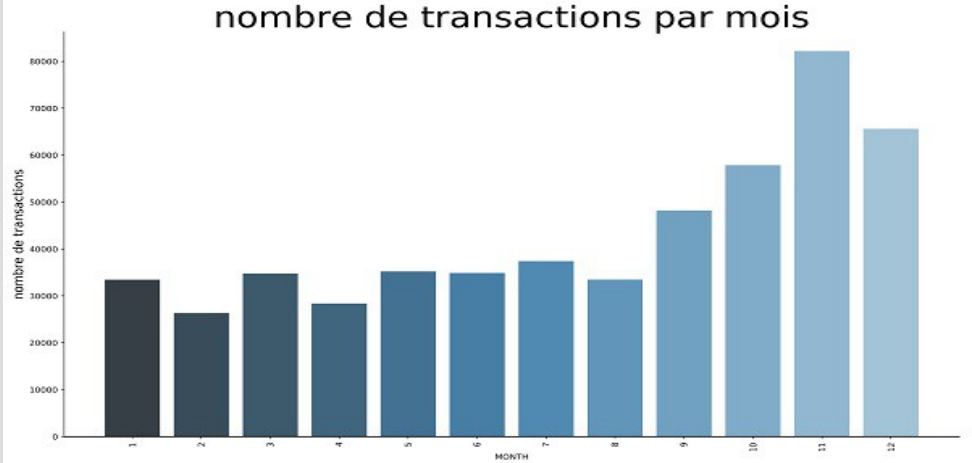
Prendre en compte les différences de consommation liées à l'origine

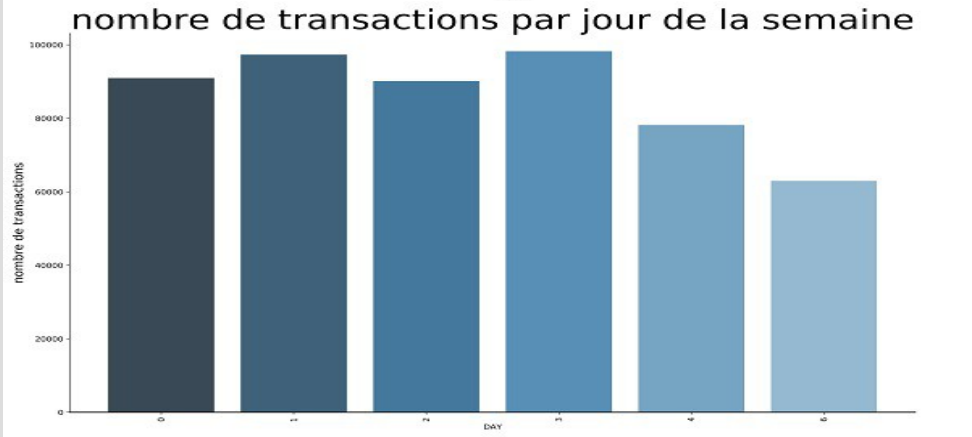
Pays d'origine > 0 ou 1 (UK ou étranger

Prendre en compte les nouveaux consommateurs / les clients perdus

- Date de la dernière commande







Choix des variables et features engineering

Prendre en compte l'impact du temps.

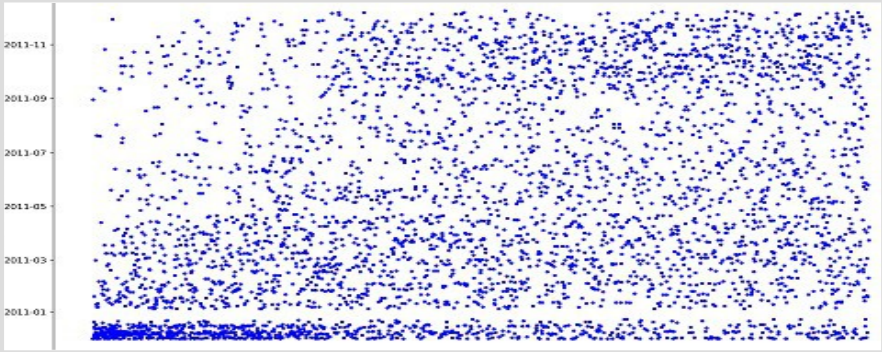
Jour de la semaine et heure
> particulier / entreprise
> grande ou petite entreprise

- Jour (heure) de maximum de consommation
- Probabilité de transaction le jour du maximum de consommation

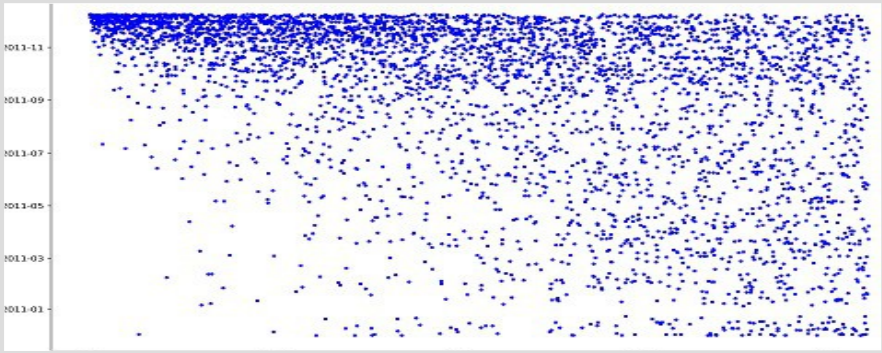
Mois (périodes de vacances/nature des produits etc...)
> Entreprise (fonctionnement interne/catégorie etc...)
> particulier (religion/préférences etc...)

- nombre de commandes par mois
- nombre de transaction au sein d'une commande par mois
- quantité moyen d'unités par transaction par mois
- prix moyen par transaction par mois

Date de la première commande par utilisateur



Date de la dernière commande par utilisateur

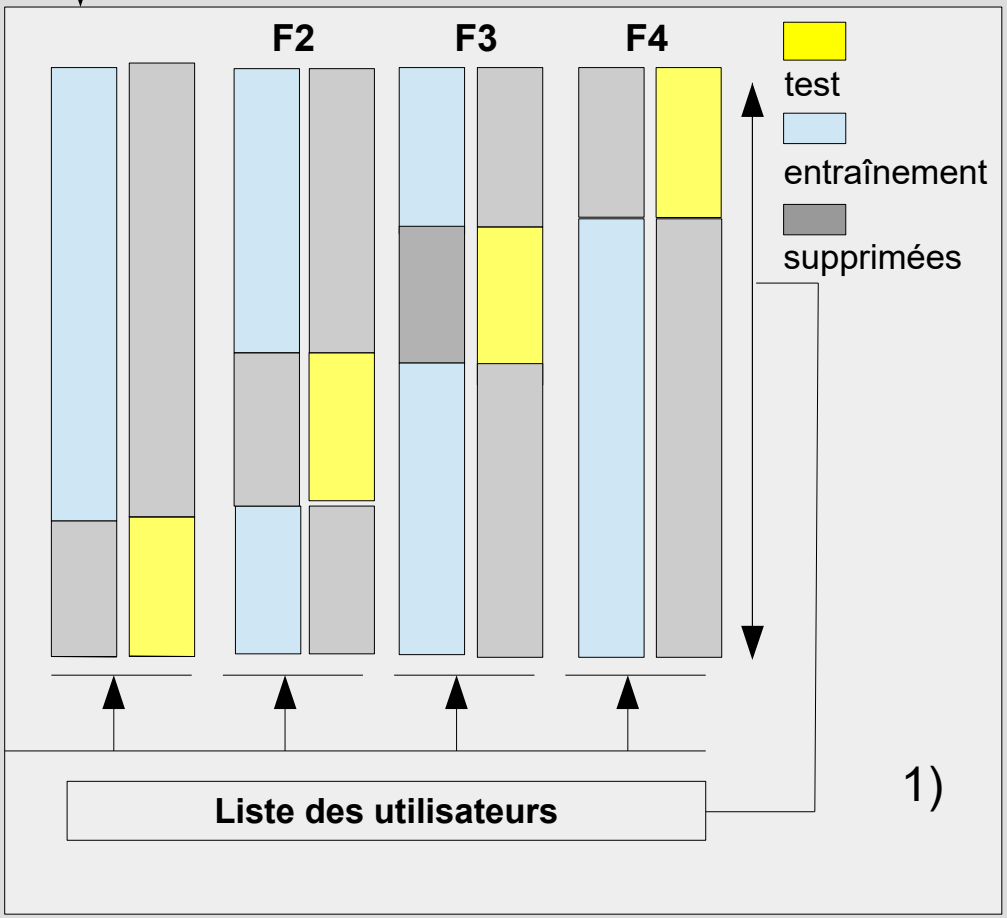


Imputation par la moyenne
Imputation par régression
-Technique > effectuer une validation croisée sur grille pour trouver automatiquement la constante de régularisation > temps de calcul
-Conceptuel > comment connaître réellement qu'elle est la première commande ?

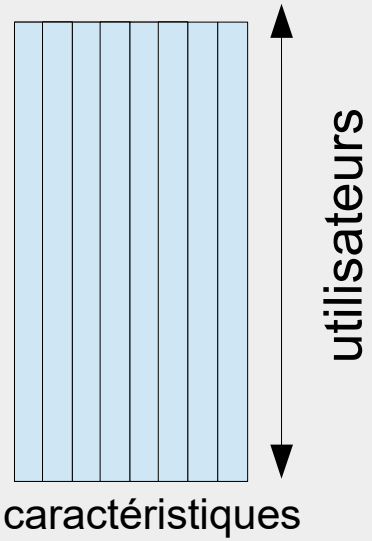
- Établir plusieurs modèles de prédiction en fonction de la commande requête

Apprentissage non supervisé

Entrée > Jeu Original : liste des transactions



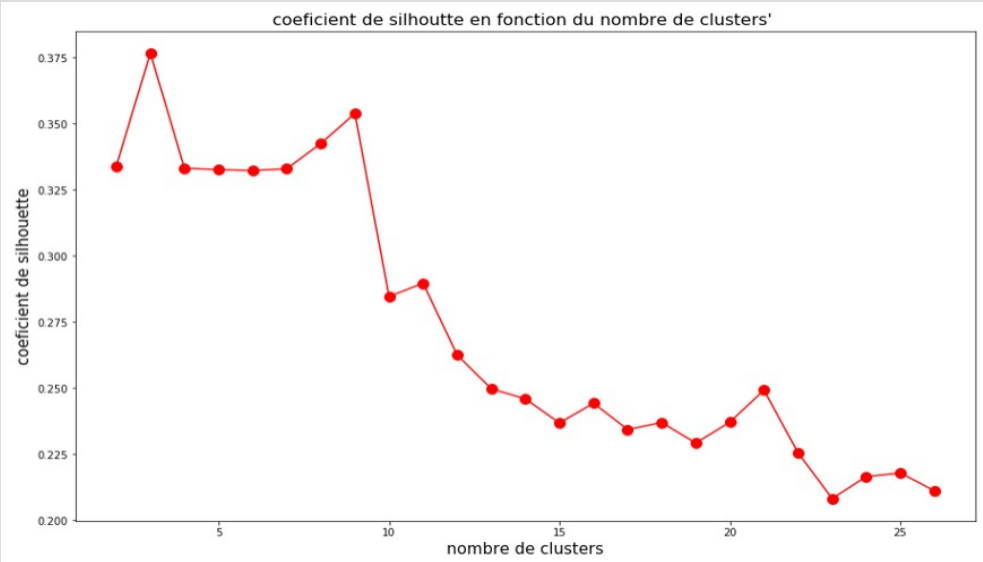
Génération des caractéristiques de clustering



Segmentation K-means pour plusieurs K

Évaluation > structure du cluster : Coefficient de silhouette

3)



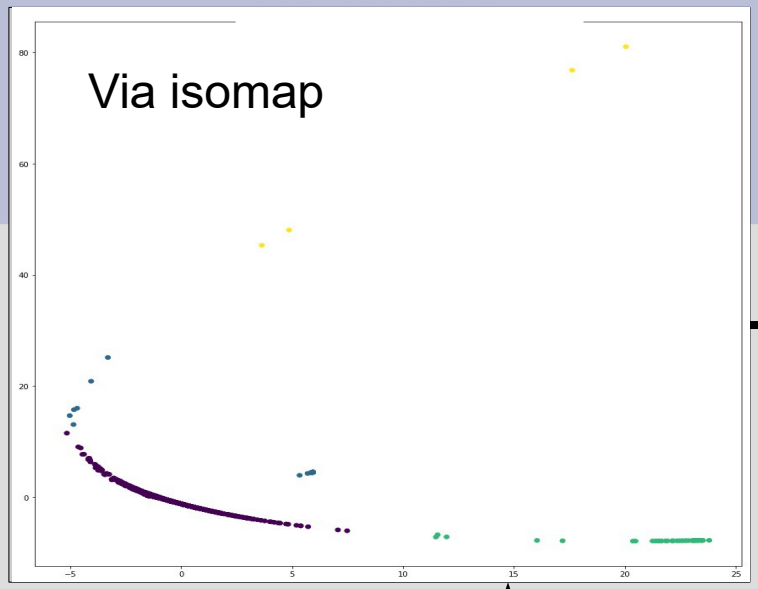
Étude des caractéristiques moyennes par cluster

Interprétation qualitative

Sortie : Labels par client

Segmentation > modèle manuel

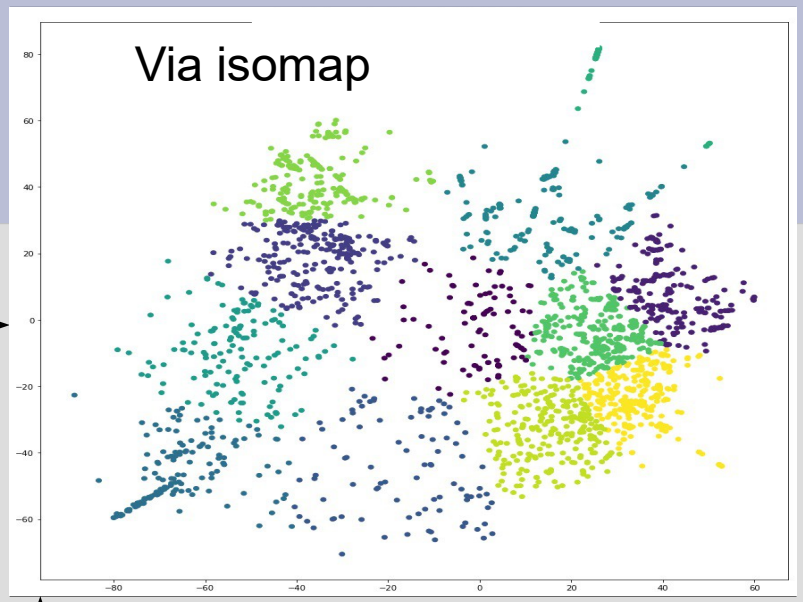
Quel comportement souhaite t-on capturer ?



Taille des clusters / Nombre de clusters minimum /

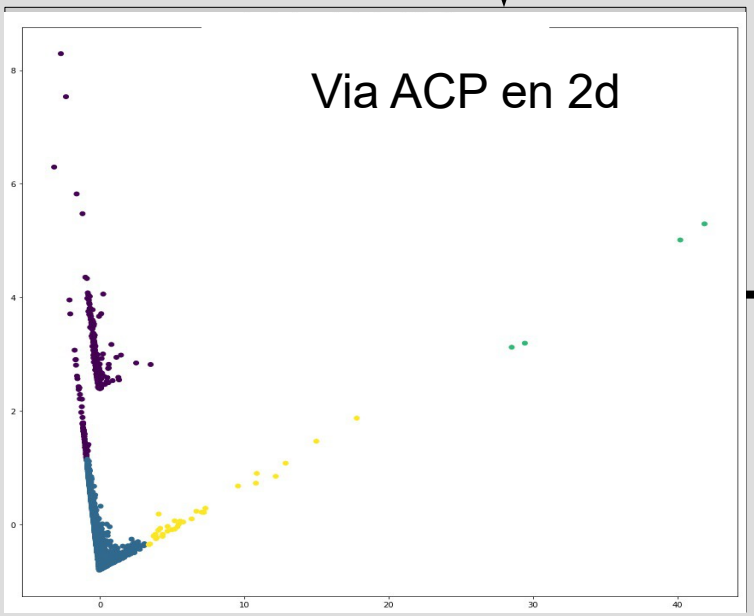
maximum

Dynamique linéaire / non linéaire



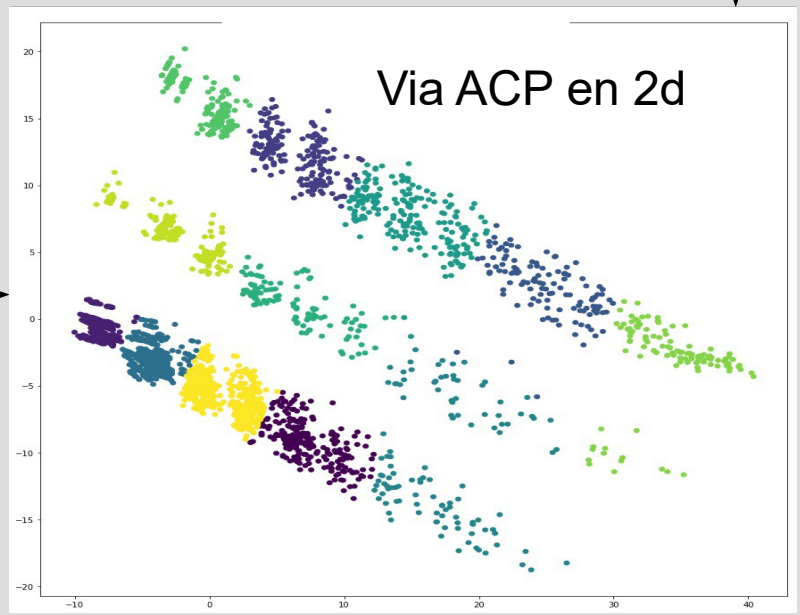
Standard Scaler

Normal Quantil Transformer



Données « écrasées » par les outliders

Utiliser un standardiser différent

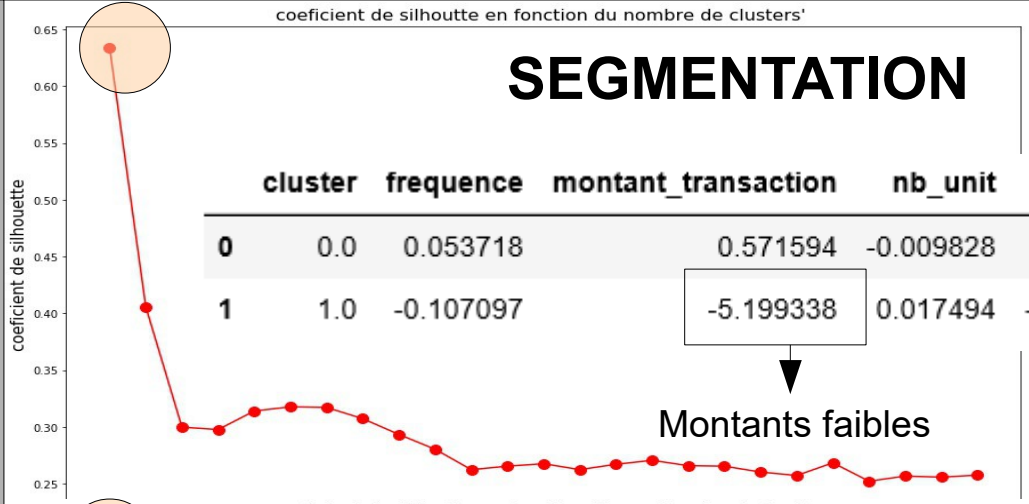


Scénario 1 : Objectif > Distinguer les petits clients des gros clients

Scénario 2 : Objectif > évaluer les différences de comportements des clients étrangers

Scénario 3 : Ajout de caractéristiques temporelles, popularité, date de dernière transaction

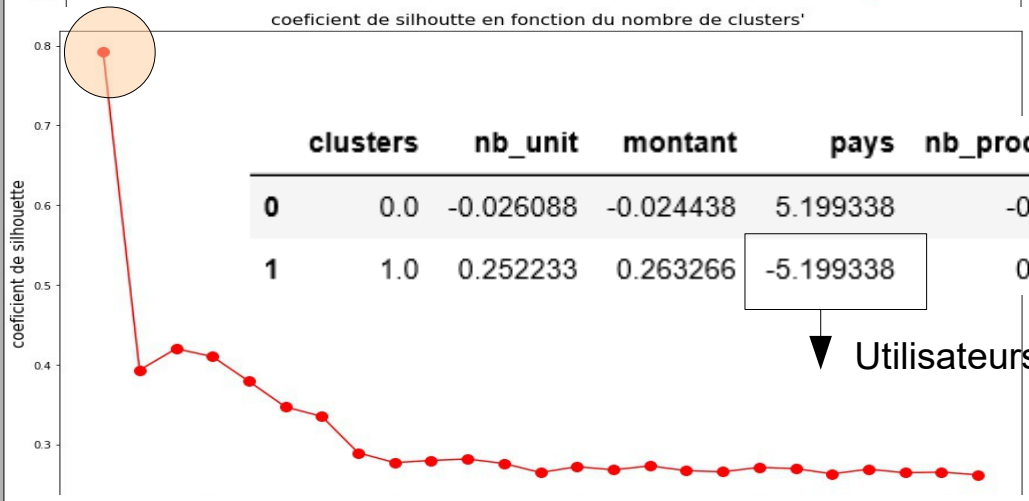
SEGMENTATION



	cluster	frequence	montant_transaction	nb_unit	prix_unit	dimension
0	0.0	0.053718	0.571594	-0.009828	0.034137	64.963961
1	1.0	-0.107097	-5.199338	0.017494	-0.063825	35.036039

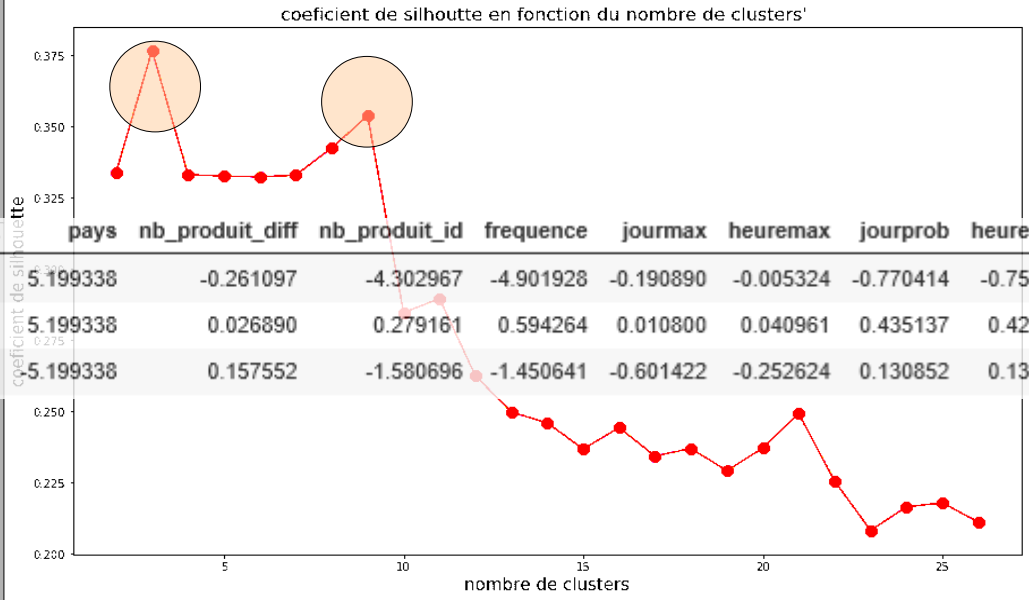
Montants faibles

Étude des caractéristiques moyennes par cluster

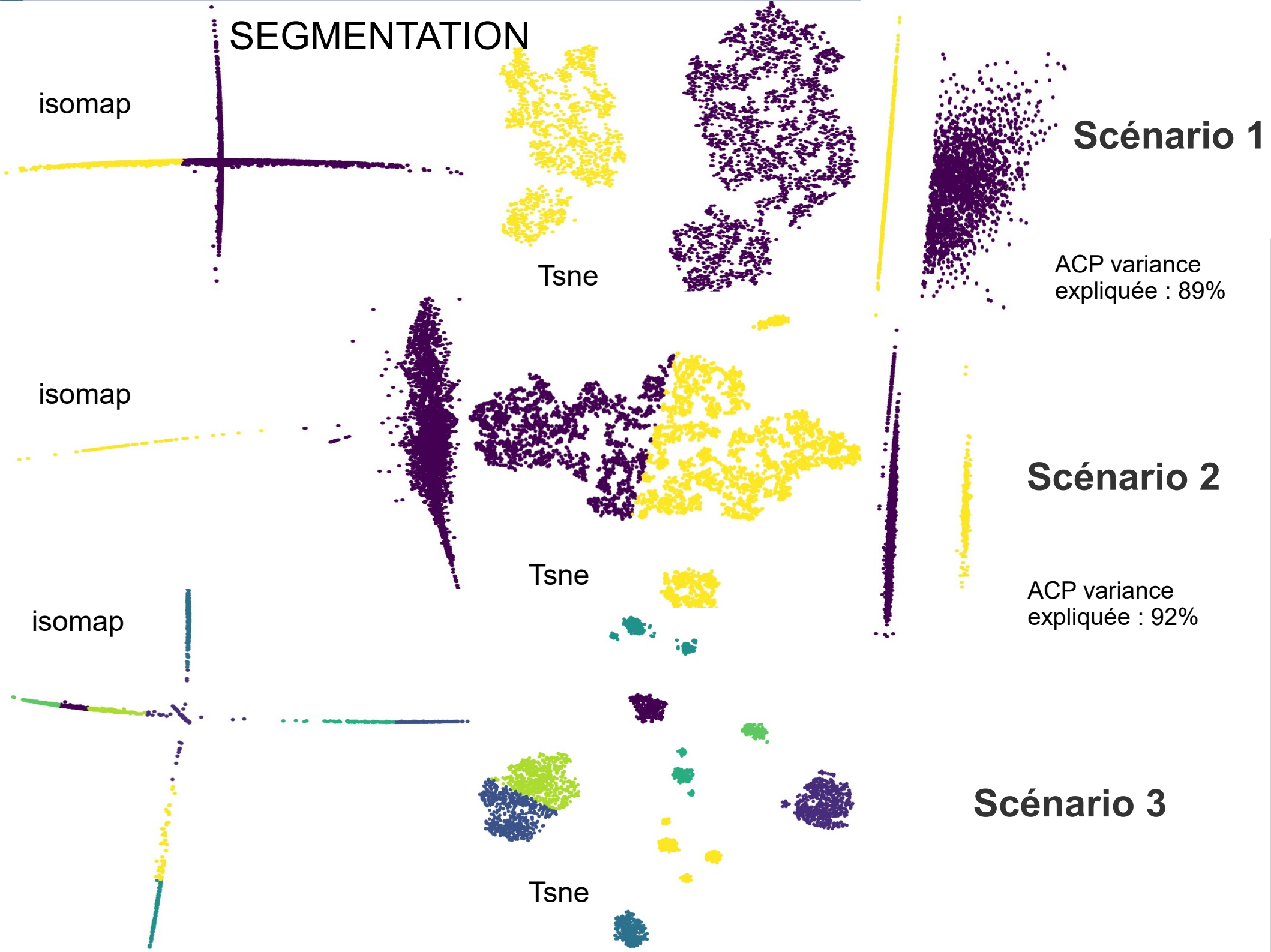


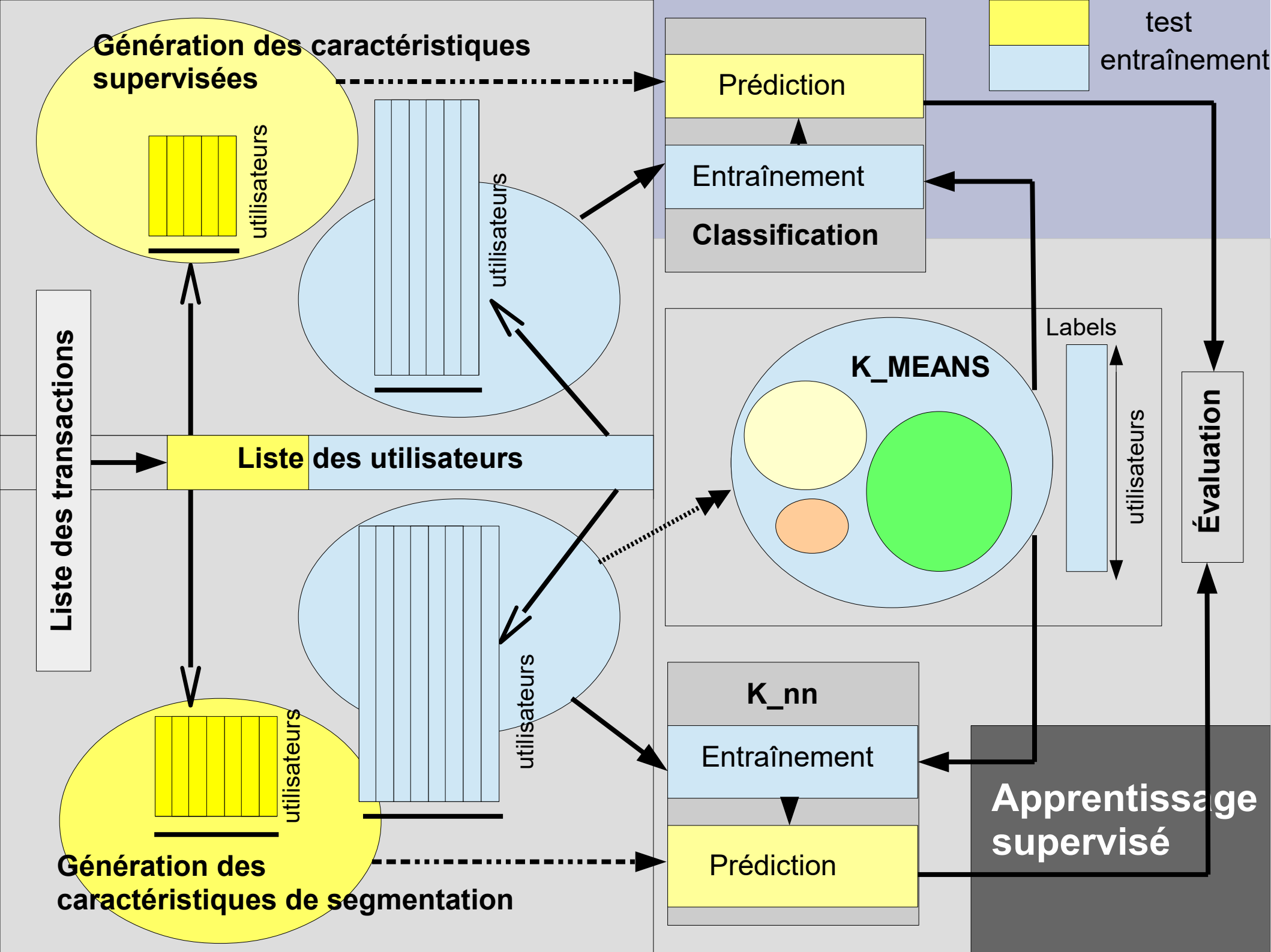
	clusters	nb_unit	montant	pays	nb_produit_diff	dimension
0	0.0	-0.026088	-0.024438	5.199338	-0.078112	91.570041
1	1.0	0.252233	0.263266	-5.199338	0.171465	8.429959

Utilisateurs étrangers



	clusters	last_trans	nb_unit	prix_unit	montant	pays	nb_produit_diff	nb_produit_id	frequence	jourmax	heuremax	jourprob	heureprob	prop1	prop2	dimension
0	0.0	0.643874	-0.076650	0.004649	-0.043788	5.199338	-0.261097	-4.302967	-4.901928	-0.190890	-0.005324	-0.770414	-0.756257	-0.023475	-0.041427	33.970542
1	1.0	-0.375659	0.002486	0.006982	-0.014038	5.199338	0.026890	0.279161	0.594264	0.010800	0.040961	0.435137	0.423585	0.014055	0.035286	57.787527
2	2.0	-0.022331	0.249555	-0.066632	0.281910	-5.199338	0.157552	-1.580696	-1.450641	-0.601422	-0.252624	0.130852	0.130806	-0.002940	-0.078066	8.241930


SEGMENTATION
isomap
Tsne
Scénario 1
ACP variance
expliquée : 89%
isomap
Tsne
Scénario 2
ACP variance
expliquée : 92%
isomap
Tsne
Scénario 3


Génération des caractéristiques supervisées
utilisateurs
utilisateurs
Liste des transactions
Liste des utilisateurs
utilisateurs
utilisateurs
Génération des caractéristiques de segmentation
test
entraînement
Prédiction
Entraînement
Classification
Labels
K_MEANS
utilisateurs
Évaluation
K_nn
Entraînement
Prédiction
Apprentissage supervisé

Étude du la qualité de prédiction en fonction du scenario

entraînement
test

Liste des transactions

Liste des utilisateurs

Liste des commandes par utilisateur et par dates

t=1

t=2

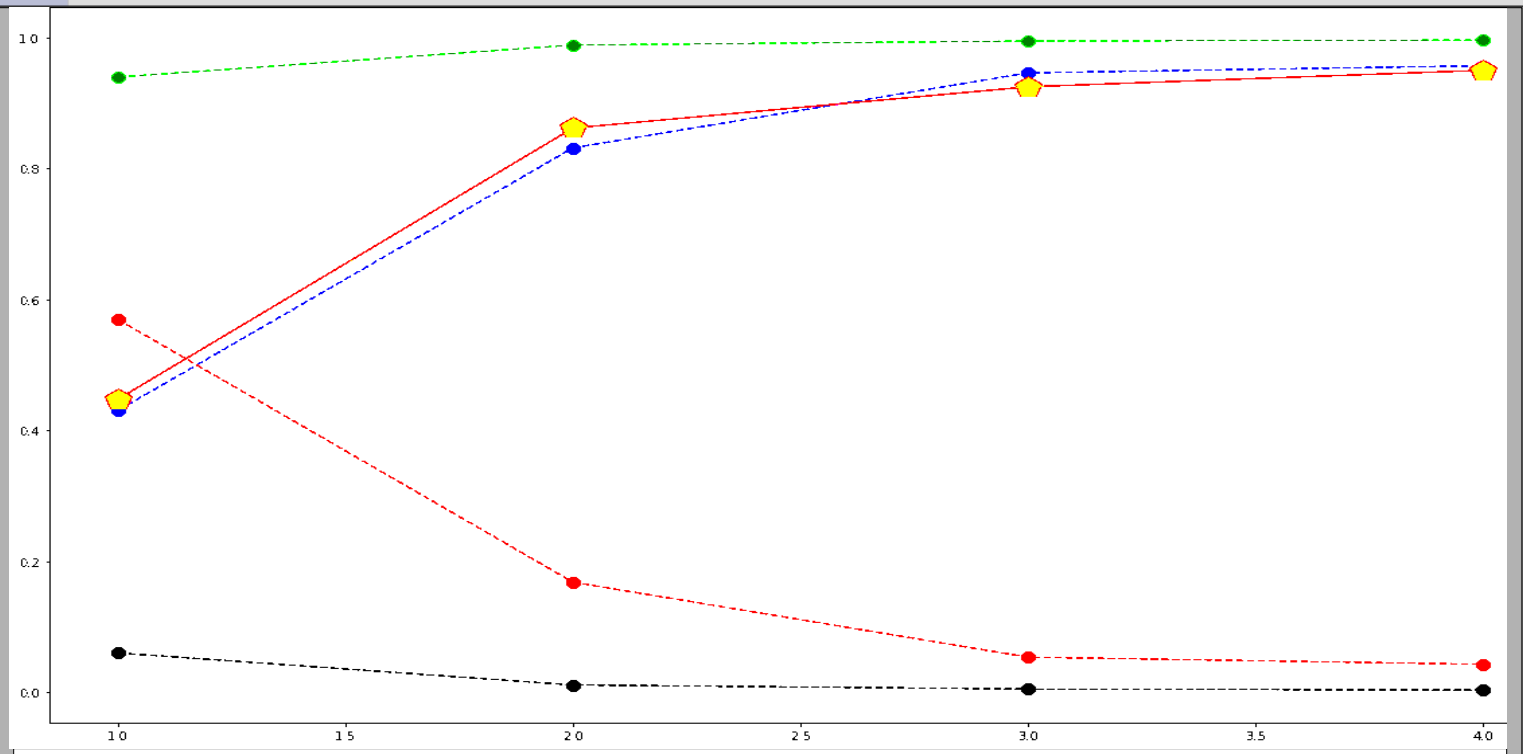
t=3

Liste des transactions

Mise en place d'un scénario spécifique
> Restriction sur le choix des données utilisées pour l'apprentissage supervisé

Liste des utilisateurs

Entraînement sur 6 mois / prédiction sur 6 mois



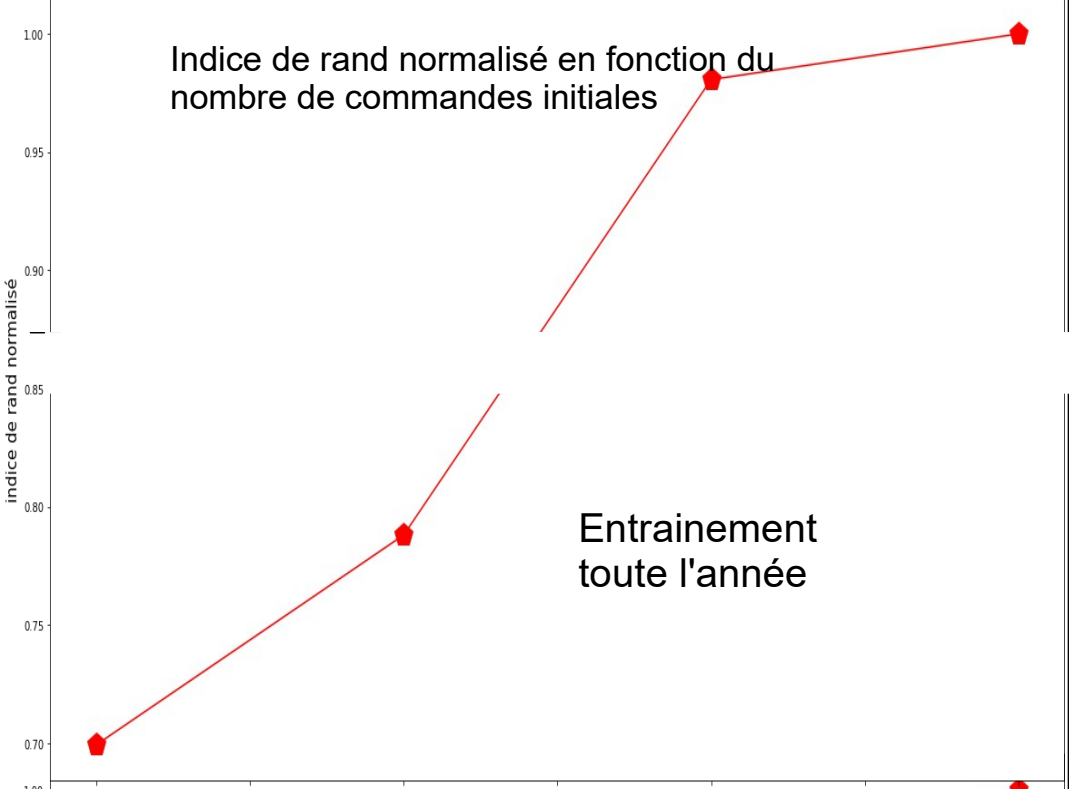
Vrais négatif

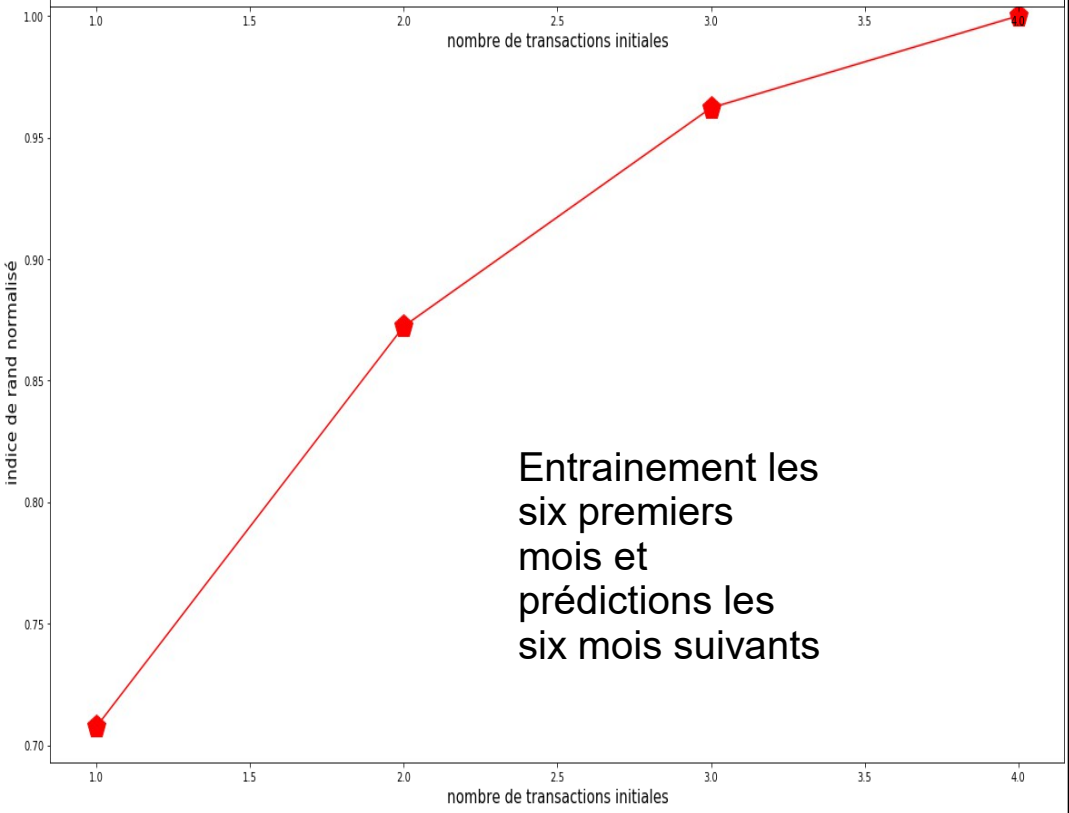
Indice de rand normalisé

Vrais positif

Faux negatif

Faux positif





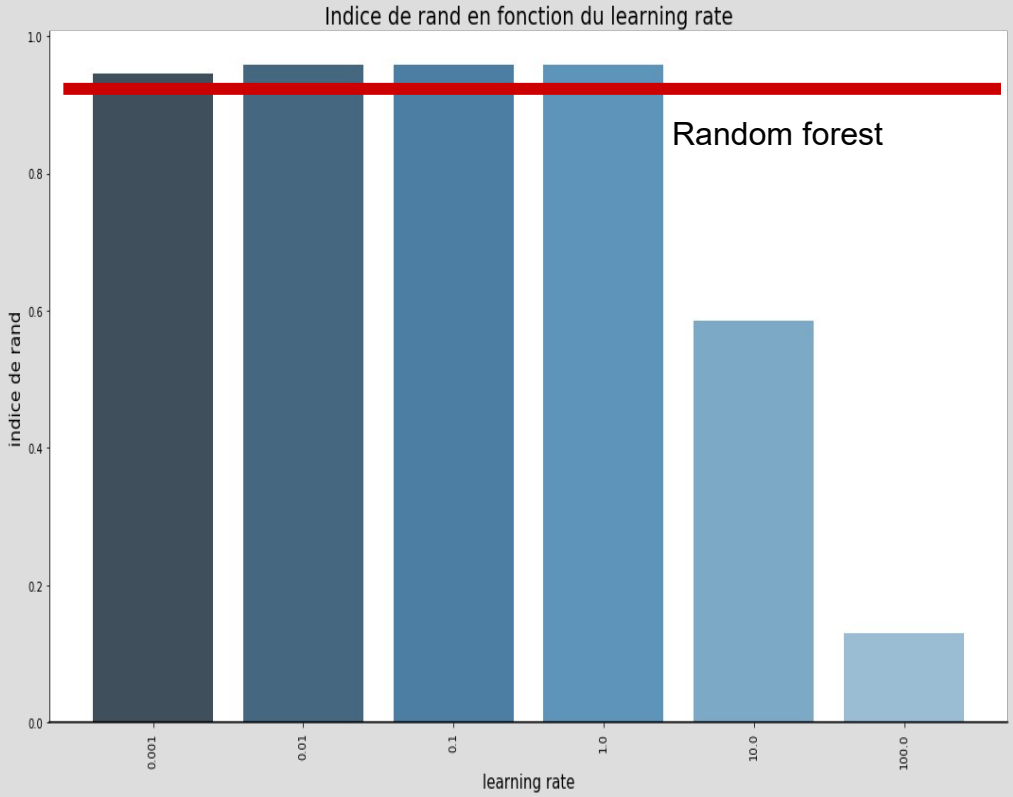
Apprentissage supervisé

Comparaison des modèles

Random forest classifier

Gradient boosting classifier

- Détermination de l'hyper-paramètre « alpha » par grille



Algorithme du bagging > présentation théorique

Méthode ensembliste > combinaison d'apprenants faibles pour obtenir des résultats plus performants

- Bootstrap > effectuer n échantillonnages avec remise

  effectuer la moyenne ou vote à la majorité des apprenants faibles appliqués aux n échantillons.

  Modèles individuels simples > forte variance / erreur répartie de manière normale

  conjugaison des modèles > réduction de la variance et donc prédiction plus fiable

Forêts aléatoires > présentation théorique

- Utilisation d'arbre de décisions :

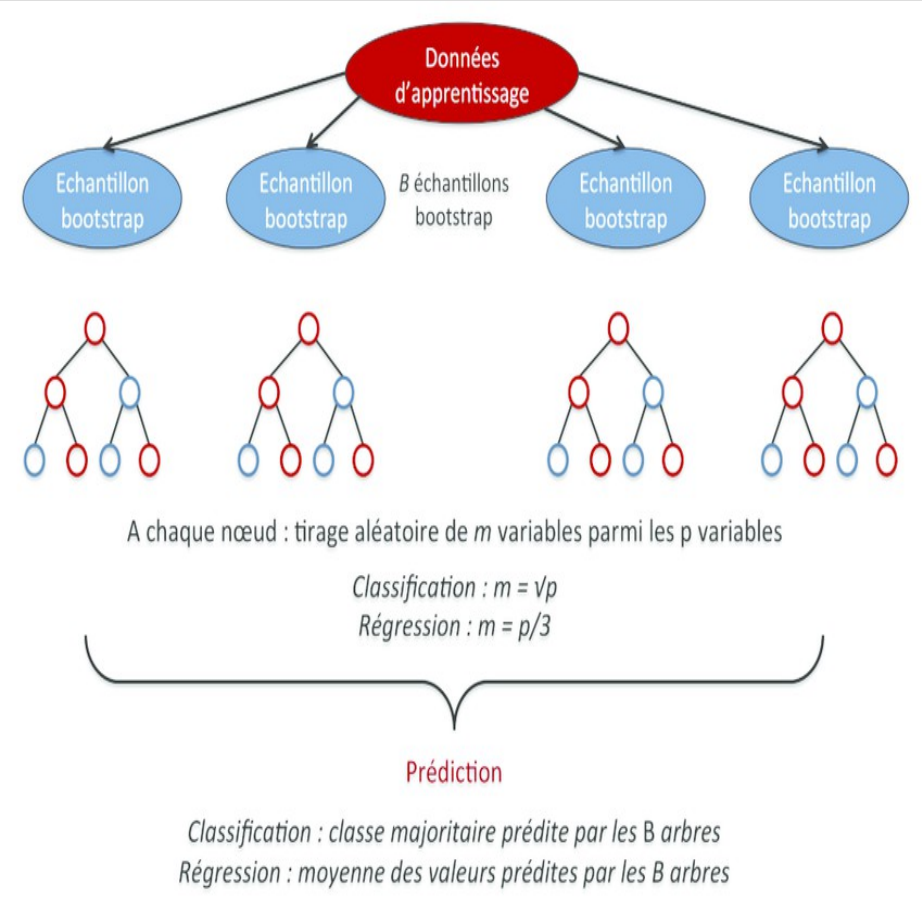
  chaque branche > détermination de la caractéristique et de la valeur de séparation > utilisation d'heuristique (ex : index de Gini)

  nombre de branches optimal > compromis biais / variance

- Forêt > Agrégation d'arbres qui sur-apprennent individuellement

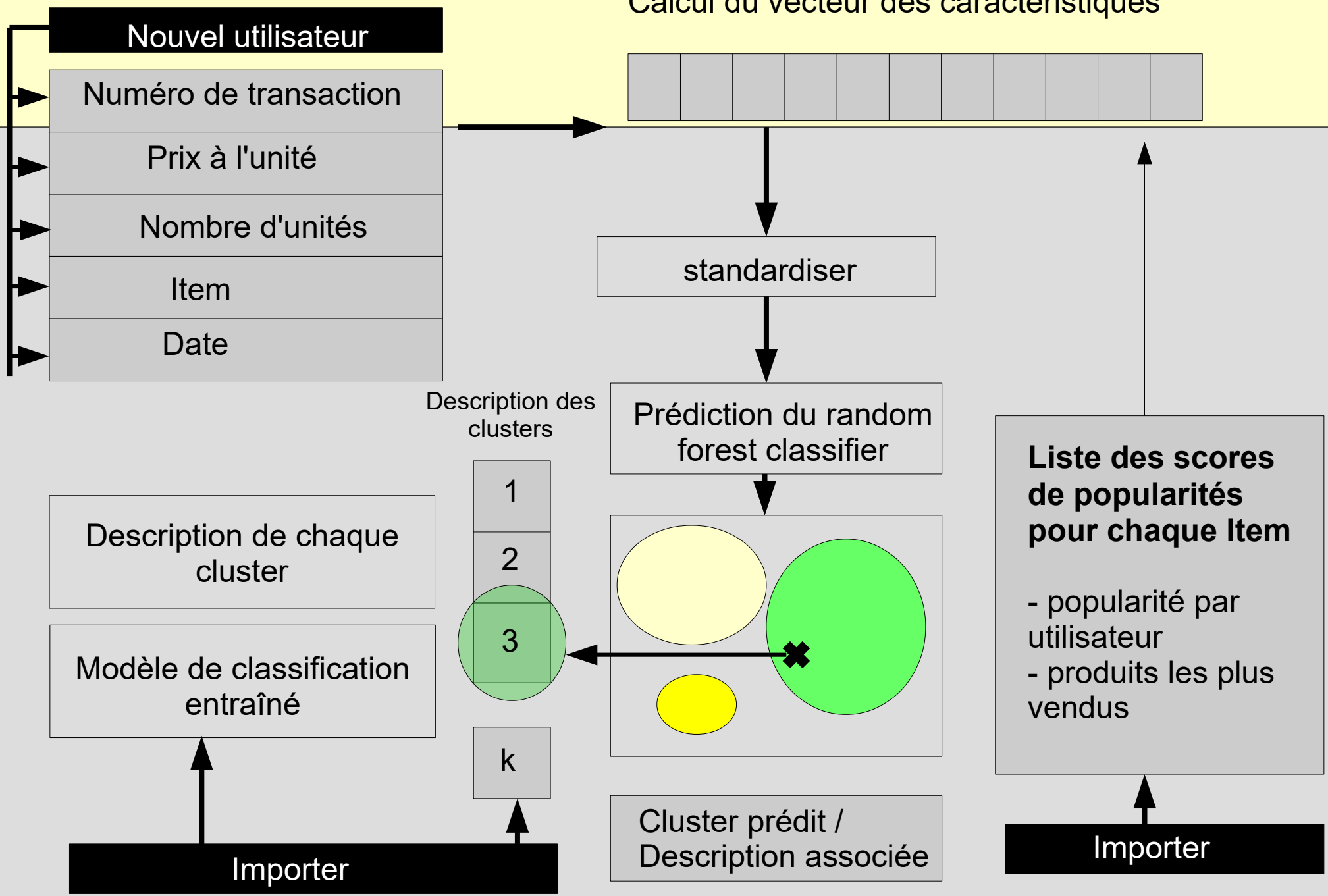
  - application du bootstraap pour n arbres différents

  - prise en compte d'un sous ensemble aléatoire de features à chaque nœud > réduction de la corrélation entre les arbres



- Avantages : mémoire > taille du dataset initial / pas d'overfitting / complexité raisonnable

Prédiction


Nouvel utilisateur
Numéro de transaction
Prix à l'unité
Nombre d'unités
Item
Date
Calcul du vecteur des caractéristiques
standardiser
Prédiction du random forest classifier
Description des clusters
1
2
3
k
Description de chaque cluster
Modèle de classification entraîné
Importer
Cluster prédit / Description associée
Liste des scores de popularités pour chaque Item
- popularité par utilisateur
- produits les plus vendus
Importer

Conclusion

	Modèle collaboratif	Modèle manuel
Détermination des objectifs	Comprendre au mieux la nature des comportements client	
Séparer les utilisateurs en jeux de test et d'entraînement → Étude des variables	Variables manuelles (transactionnelles/temporelles etc..) Variables descriptives des items	
Qualité de prédiction > Qualité des vecteurs utilisateurs générés - détermination des jeux de test et d'entraînement - détermination d'un score - validation croisée sur grille		- définition des caractéristiques supposées pertinentes - détermination d'un algorithme de prédiction pour la comparaison des modèles - définition des jeux de test et d'entraînement
Segmentation via K-means / Choix du nombre de clusters via l'indice de silhouette		
Interprétations qualitatives difficiles Deux caractéristiques interprétables émergent.		- Interprétation des scénarios via la moyenne des caractéristiques par cluster
Prédiction > modèles supervisés > différents scenarii (nombre de commandes initiales, échelle de temps...)	Évaluation > Indice de rand normalisé Nouveau client > fonction python	