Segmentation des comportements des clients

RAPPORT

Demande initiale

Nous travaillons pour une entreprise britannique leader dans la **vente en ligne** de toutes sortes d'objets.

Une des missions de l'équipe marketing est de mieux comprendre les comportements de ses clients pour **augmenter la fréquence d'achat et la valeur du panier moyen**.

Notre mission consiste à :

- Identifier **différents types de comportements**
- **Prédire le comportement d'un client** le plus rapidement possible dès le premier achat

- Nous utiliserons un **historique d'une année de transactions** sur ce site de vente en ligne

Interprétation

La demande exprimée consiste à réaliser 2 opérations :

Dans un premier temps il s'agit de réaliser un **clustering des clients**, sur la base de critères pertinents pour une utilisation marketing.

Le machine learning nous permettra de réaliser un **apprentissage non supervisé** et nous aidera à découvrir des profils de clients distincts entre eux et de contenu homogène.

Dans un deuxième temps un **apprentissage supervisé** nous permettra de **prédire le profil d'un client** sur la base des informations de sa première commande.

Pistes de recherche

Pour réaliser le **clustering des clients**, nous étudierons en priorité les features habituelles de ce métier, notamment :

- La fréquence d'achat
- Le volume d'achat

Pour réaliser la **prédiction de profil**, nous allons exploiter au maximum les informations de la première commande, soit :

- La date d'achat
- Le prix total de la commande
- Le nombre d'articles différents achetés
- La quantité d'articles identiques achetés
- Le prix des articles achetés
- Le type d'articles achetés
- Le pays

La performance de notre modèle de classification supervisée sera comparée à la performance d'un modèle « dummy » qui traite tous les clients de la même manière.

Plan de la présentation

01

Nettoyage et Exploration

02

Segmentation des comportements Clients

- Feature Engineering
- Modélisation

03

Prédiction des comportements clients

- Feature Engineering
- Modélisation

Nettoyage et Exploration

Variables présentes

- La base de données contient les informations suivantes :
  - InvoiceNo
  - StockCode
  - Description
  - Quantity
  - InvoiceDate
  - UnitPrice
  - CustomerID
  - Country
- Les factures couvent la période du 1/12/2010 au 9/12/2010
- 406829 lignes de facture
- 4372 clients
- 3684 codes produits
- 89% des clients sont britanniques, les autres dispersés dans 36 pays sans particularités notables.

Exploration et Cleaning

La base initiale est de bonne qualité : la plupart des informations sont complètes.

Nous avons réalisé les opérations de nettoyage suivantes :

- Sélection d'une fenêtre d'un an exactement pour éviter la sur-représentation du mois de décembre
- Suppression des transactions sans clients (n'apportent pas d'information sur les clients)
- Suppression des StockCode ne contenant que des lettres et ne correspondant pas à des achats de produits : ['POST', 'D', 'C2', 'M', 'BANK CHARGES', 'PADS', 'DOT', 'CRUK']
- Simplification de la variable Country : 1 pour UK, 0 sinon

Clustering des clients

FEATURE ENGINEERING

Features explorées

Fréquence des achats
- Nombre de commandes sur l'année
- Fréquence d'achat rapportée à la récence du client (nb d'achats/durée entre son premier achat et la fin de l'année)

Montant des achats
- Somme totale dépensée par le client sur l'année
- Somme totale dépensée par le client sur l'année rapportée à la récence du client (somme totale / durée entre son premier achat et la fin de l'année)
- Panier moyen
- Panier avec le montant le plus élevé pour ce client

Diversité des achats
- Nombre d'articles différents sur l'année
- Nombre maximum d'articles différents pour une seule facture

Quantité d'articles (particulier ? Grossiste ?)
- Quantité totale d'articles de toutes sortes commandés sur l'année
- Quantité minimale commandée pour un même article pour une même facture
- Quantité médiane commandée pour un même article pour une même facture
- Quantité maximale commandée pour un même article pour une même facture

Saisonnalité (client régulier ? Saisonnier ?) : variance du mois d'achat

Pour le calcul de ces features, par souci de simplification, nous n'avons conservé que les lignes de facture pour lesquelles la quantité était positive.
Les lignes de factures avec une quantité négative, correspondant à des retours ont été traitées séparément dans la feature « retours » , ci-dessous.

Retours : Nombre d'articles retournés par le client

Aggrégation

- Lignes de facture
- Factures
- Clients

Outliers

Scaling

- Standardisation Min/Max

Shuffle

Pre-processing

Agrégation

Les informations de la base de données sont données sous la forme une ligne par ligne de panier

Nous avons construit 2 niveaux d'agrégation pour obtenir :

- **Une base aggrégée par facture** :

  Une ligne de la base regroupe tous les différentes lignes d'un même panier

- **Une base aggrégée par client** :

  Une ligne de la base par client

Outliers

Valeurs exceptionnelles

Ancienneté significative

Période de Noël

Valeurs exceptionnelles :
Une seule facture vraiment exceptionnelle a été identifiée. Avec un très grand nombre d'article et d'un montant élevé, elle a été remboursée quelques minutes après l'achat. Nous l'avons supprimé de la base.

Nous avons décidé de supprimer toutes les lignes pour lesquelles au moins une variable avait une **valeur supérieure au percentile 0,999.**

Nous avons choisi la valeur seuil de 0,999 parce que c'était celle qui permettait d'obtenir les meilleures performances de clustering (nombre et taille des clusters, que nous présenterons plus loin).

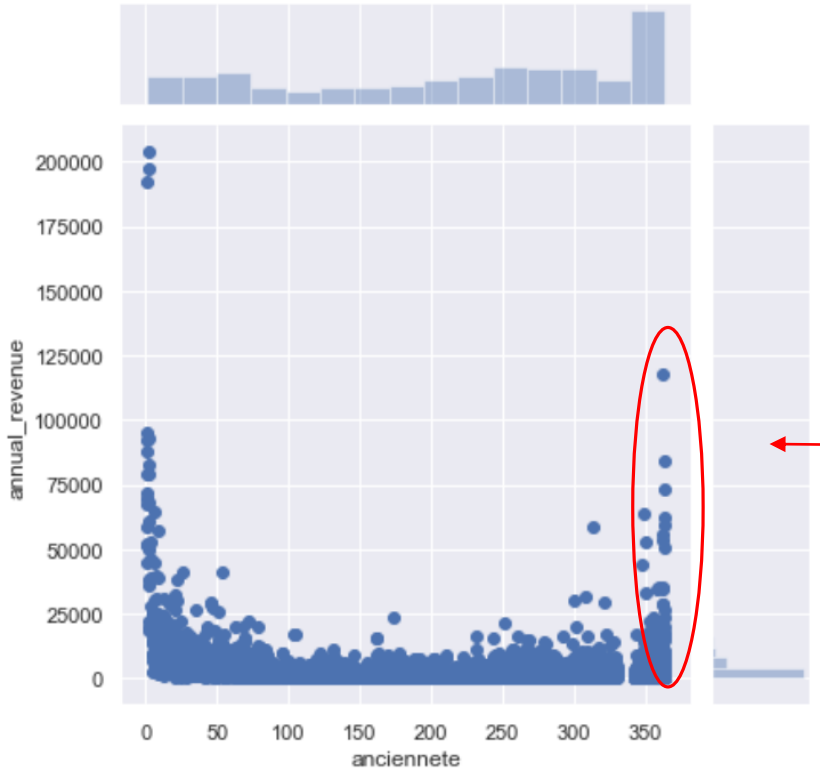
Sélection d'une période de temps significative
Pour les clients les plus récents, le nombre de commandes n'est pas significatif pour estimer le comportement en terme de fréquence d'achat. Nous avons donc exclus les clients avec une ancienneté inférieure à 200 jours. (à partir de cette valeur seuil, la fréquence se rapproche d'une distribution normale)

Exclusion du biais de Noël
Les achats effectués au mois de décembre sont nettement supérieurs à ceux achetés les autres mois. Cela induit un biais de différence entre les clients présents depuis fin 2010 et ceux ayant effectué un premier achat après cette date. Nous n'avons donc conservé que les clients ayant effectué leur premier achat après le 01/01/2011.

Biais de Noël : illustration

Annual_revue : estimation du CA annuel pour ce client, par extrapolation des commandes effectuées depuis son premier achat.

Figure : revenu annuel estimé en fonction de l'ancienneté du client



Clients présents depuis Noël 2010 : la projection de leur CA annuel est vraisemblablement faussée par leurs paniers de Noël. **Nous excluons ces clients de la modélisation.**

Scaling

Pour la segmentation des comportements clients, nous avons choisi d'utiliser un algorithme de clustering généraliste : Kmeans.

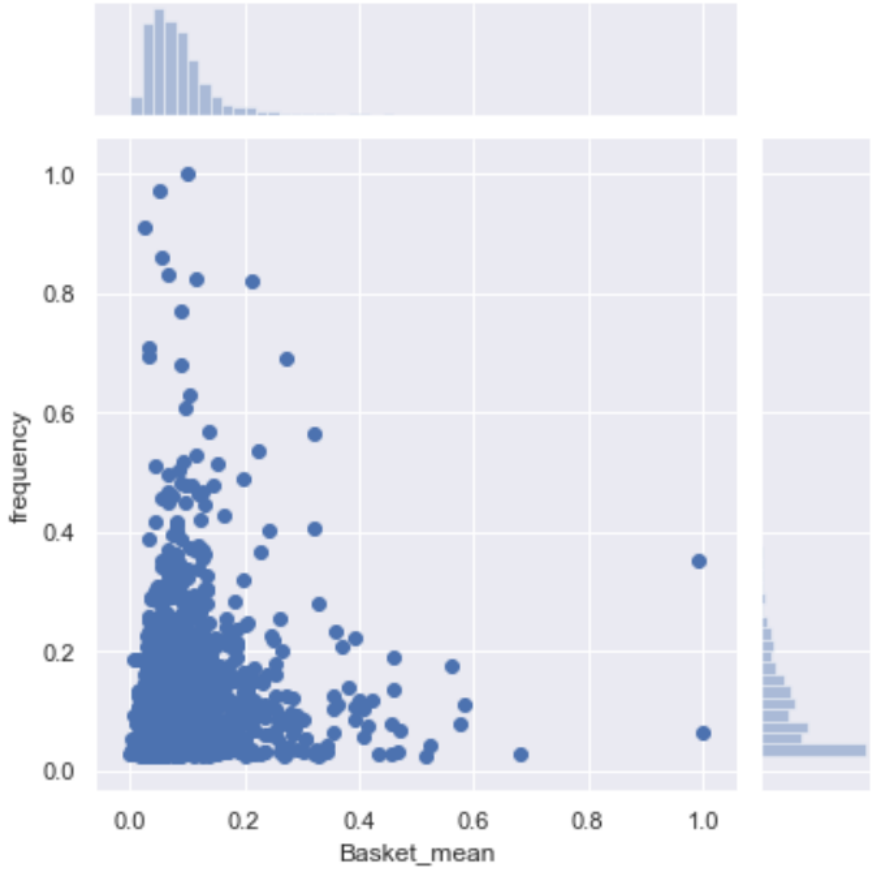
Si des variables ont une variance différente, Kmeans donnera plus d'importance aux variables à la variance plus élevée. Si nous voulons choisir nous même le poids de nos variables, ou leur donner un poids égal, nous devons d'abord rendre leur variance égale.

Nous avons donc standardisé les données.

Exploration visuelle des profils

aucun groupe de client évidnt n’apparaît visuellement

Distribution des clients selon leur fréquence d’achat et leur panier moyen



Clustering des clients

MODÉLISATION

Démarche

Nous avons choisi le modèle en optimisant les critères suivants :

Taille des clusters : nous souhaitons que chaque groupe ait une taille significative, pour permettre de mettre en œuvre des actions marketing collective sur chaque segment.

Différence entre groupes (distance des clusters) : nous souhaitons que les différents groupes se différencient nettement pour que l'on puisse comprendre les caractéristiques de chaque groupe

Choix du nombre de clusters :

Les clusters à 4 groupes donnait soit au moins un groupe trop petit, soit des groupes mal différenciés. Nous avons choisi un clustering en **3 groupes**.

Choix des features et de leur poids. Nous avons fait varier les features et leur poids pour optimiser les critères ci-dessus.

Modèle retenu

Interprétation

Basic customers

Le groupe orange, majoritaire, représente des clients avec des paniers moyens bas et une fréquence d'achat basse

Best frequency

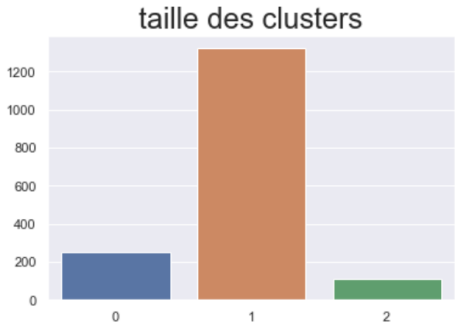
Le groupe bleu représente les clients qui rapportent le meilleur CA essentiellement grâce à une fréquence d'achat 2 fois supérieure à la moyenne.

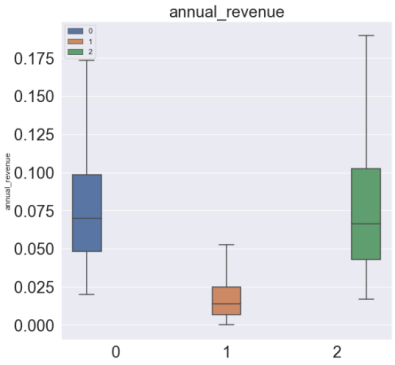
Best Basket

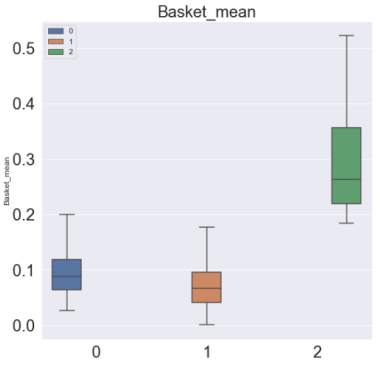
Le groupe vert représente des clients apportant un meilleur revenu grâce à un panier moyen 2 fois plus élevé et une fréquence plus élevée.

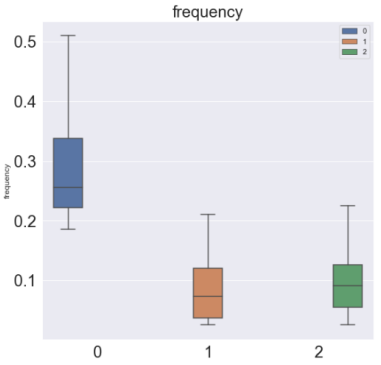
Features et poids

coeff = {'annual_revenue' : 1, 'Basket_mean' : 1,'Quantity_mean' :0,'StockCode_size' : 0, 'StockCode_mean' : 0,'frequency' : 1, 'Invoice_count': 0, 'return_rate' : 0, 'seasonality' : 0}









Prédiction du profil client

FEATURE ENGINEERING

Features

La demande consistant à prédire le comportement futur d'un client **dès son premier achat**, nous avons recherché **une modélisation qui prend en entrée une facture**, avec ses différentes lignes et produit en sortie un groupe.

Les features utilisées pour réaliser la prédiction sont les suivantes :

Variables quantitatives :

Montant de la facture

Nombre d'articles différents achetés

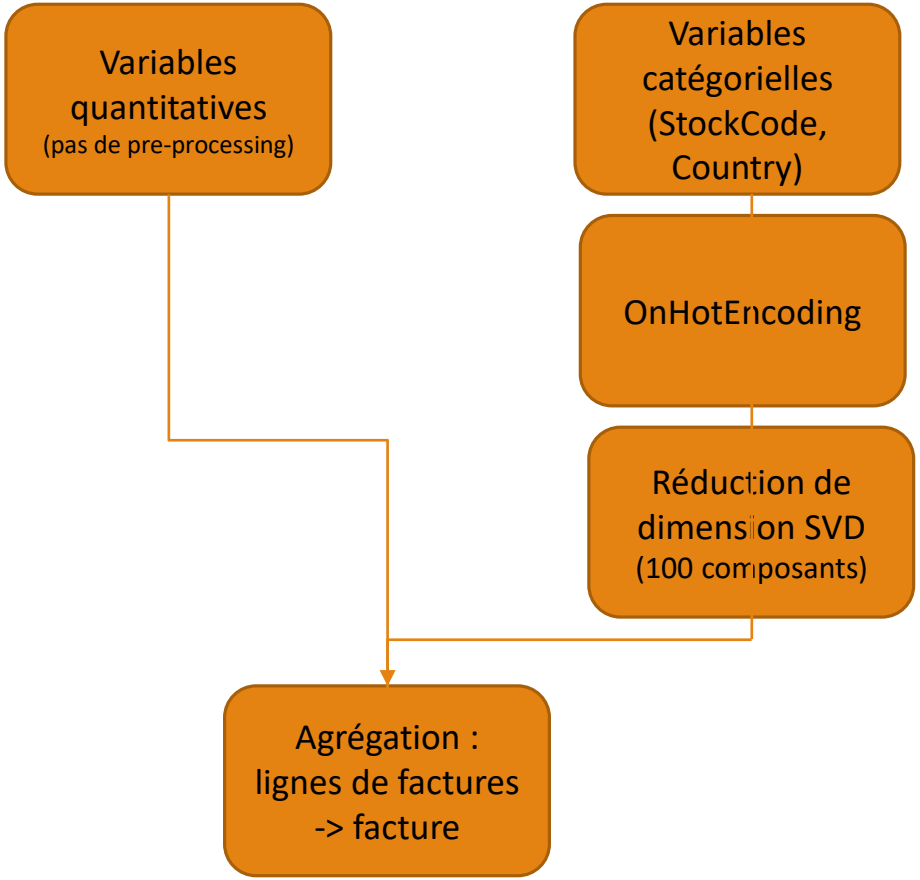
Prix moyen des articles achetés

Quantité moyenne pour un même article

Variables qualitatives :

- **Codes Articles**
- **Pays**

Pre-processing



Prédiction du profil Client

MODÉLISATION

Démarche

Baseline : Modélisation avec un modèle **Dummy**, avec l'option « most_frequent » : ce modèle traite tous les clients de la même manière, comme un client du groupe majoritaire.

Modélisation avec les algorithmes **Random Forest** et **XGB**, avec les paramètres par défaut de ces algorithmes (un travail ultérieur pourra être réalisé pour rechercher les meilleures hyper-paramètres par validation croisée grâce à GridSearchCV).

Remarque : lors du clustering des clients, nous avons exclus de la modélisation les clients dont l'ancienneté était insuffisante pour que leur comportement soit significatif. En conséquence, ces clients n'ont pas été classés dans des groupes. Nous avons donc exclus ces clients également pour la modélisation de la prédiction de comportement.

Modèle final

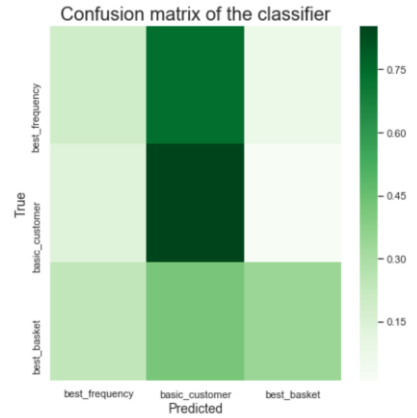
Le Recall Score mesure notre capacité à « ne pas rater » un client.

Notre modèle est meilleur :

pour détecter les clients « best_basket »

que pour détecter les clients « best_frequency »

Matrice de confusion normalisée pour le modèle XGB



	precision	recall	f1-score	support
best_frequency	0.47	0.19	0.27	652
basic_customer	0.59	0.85	0.70	868
best_basket	0.32	0.34	0.33	76
micro avg	0.56	0.56	0.56	1596
macro avg	0.46	0.46	0.43	1596
weighted avg	0.53	0.56	0.51	1596

Comparaison des performances

Overfitting

Scores aggrégés

Score par groupe

```
DummyClassifier(constant=None, random_state=None, strategy='most_frequent')
dummy
estimator.score on train : 0.5460224906289046
estimator.score on test : 0.543859649122807
precision_score  : 0.29578331794398277
recall_score  : 0.543859649122807
f1_score : 0.38317384370015956
                precision    recall  f1-score   support

best_frequency       0.00      0.00      0.00       652
basic_customer       0.54      1.00      0.70       868
   best_basket       0.00      0.00      0.00        76

     micro avg       0.54      0.54      0.54      1596
     macro avg       0.18      0.33      0.23      1596
  weighted avg       0.30      0.54      0.38      1596
```

```
RandomForestClassifier(bootstrap=True,
            class_weight={0.0: 1907, 1.0: 2622, 2.0: 273},
            criterion='gini', max_depth=None, max_features='auto',
            max_leaf_nodes=None, min_impurity_decrease=0.0,
            min_impurity_split=None, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=400, n_jobs=None, oob_score=False,
            random_state=None, verbose=0, warm_start=False)
rf
estimator.score on train : 1.0
estimator.score on test : 0.5494987468671679
precision_score  : 0.540564858104558
recall_score  : 0.5494987468671679
f1_score : 0.5096725942325111
                precision    recall  f1-score   support

best_frequency       0.46      0.27      0.34       652
basic_customer       0.58      0.80      0.67       868
   best_basket       0.83      0.07      0.12        76

     micro avg       0.55      0.55      0.55      1596
     macro avg       0.62      0.38      0.38      1596
  weighted avg       0.54      0.55      0.51      1596
```

overfitting

Meilleure détection des clients « best_frequency »

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0,
       max_depth=4, min_child_weight=1, missing=None, n_estimators=100,
       n_jobs=1, nthread=None, objective='multi:softprob', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=2622, seed=None,
       silent=True, subsample=1)
xgb
estimator.score on train : 0.8134110787172012
estimator.score on test : 0.5576441102756893
precision_score  : 0.529662665401581
recall_score  : 0.5576441102756893
f1_score : 0.5056165967809199
                precision    recall  f1-score   support

best_frequency       0.47      0.19      0.27       652
basic_customer       0.59      0.85      0.70       868
   best_basket       0.32      0.34      0.33        76

     micro avg       0.56      0.56      0.56      1596
     macro avg       0.46      0.46      0.43      1596
  weighted avg       0.53      0.56      0.51      1596
```

Meilleure détection des clients « best_basket »

Modèle retenu et amélioration possibles

Les résultats agrégés des modèles Random Forest et Xgb sont très proches.

Pour les résultats par groupe, chacun est meilleur pour un groupe donné.

Nous retenons le modèle XgBoost, qui présente moins d'overfitting et sera plus généralisable.

Des optimisations pourront être recherchées ultérieurement sur les deux modèles en utilisant une recherche de recherche d'hyper-paramètres avec validation croisée grâce à GridSearchCV.

Conclusion

Les performances obtenues sont-elles bonnes ?

Intuitivement, on pouvait savoir que prédire le comportement futur d'un client à partir de sa seule première commande n'est pas une opération facile.

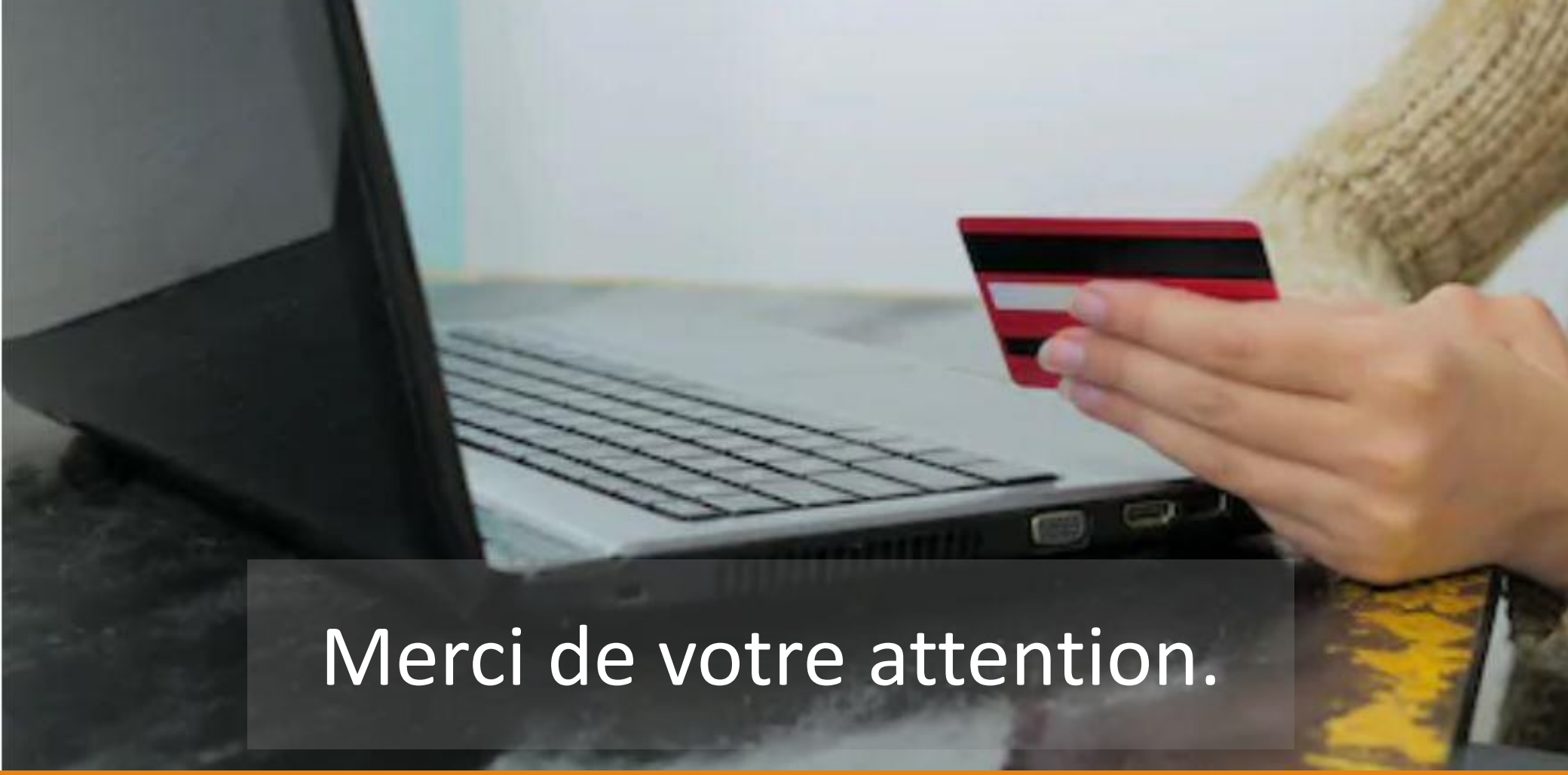
Notre modèle apporte une performance supérieure à celle d'un modèle baseline qui traite tous les clients de la même manière.

Si l'on considère que le site de vente en ligne n'applique encore aucune démarche marketing différenciée lors d'un premier achat, **notre modèle va déjà lui permettre de réaliser ses premières actions différenciées** :

Par exemple :

- Proposer aux clients prédits « best_basket » des offres les incitant à acheter plus fréquemment (livraison gratuite sur votre prochaine commande…)
- Proposer aux clients prédits « best_frequency » des offres incitant à augmenter leur panier (ajoutez un article à votre panier pour obtenir un cadeau, une réduction …)

Même si la prédiction est parfois fausse, elle sera plus souvent vraie que fausse et pourra conduire, globalement, à une augmentation du CA réalisé sur ces clients en attendant de mieux les connaître dans la durée.



Merci de votre attention.

contact : brigitte.maillere@Machine-Learn.it