UQAC

TP4 - 8INF867 FONDEMENTAUX DE L'APPRENTISSAGE AUTOMATIQUE

Clustering

December 1, 2023

Université du Québec à Chicoutimi

Ali Ghammaz GHAA27070100
Atoumane Gaye GAYA19039305
Eric Moussinga

Contents

List of Figures

1 Introduction

L'objectif central de ce TP est de résoudre un problème spécifique de *Clustering* lié à l'industrie des cartes de crédit. Dans le domaine financier, la segmentation précise des clients en fonction de leur comportement d'utilisation des cartes de crédit revêt une importance stratégique. Ce problème de clustering permet d'identifier des groupes de clients similaires en termes de dépenses, de habitudes d'utilisation et de comportements financiers, offrant ainsi aux entreprises émettrices de cartes de crédit une meilleure compréhension de leur clientèle.

Notre approche méthodologique se concentre sur la création d'un *pipeline* complet, englobant des étapes cruciales telles que la préparation des données, le choix et l'entraînement de modèles de *Clustering*, ainsi que l'évaluation comparative de différentes méthodes telles que la k-moyenne, et le *DBSCAN*, parmi d'autres. Tout au long de ce processus, notre objectif est de développer des compétences pratiques en matière d'apprentissage non supervisé, spécifiquement adaptées à la segmentation des clients dans le secteur des cartes de crédit. Le résultat final de ce projet sera présenté sous la forme d'un rapport analytique détaillé, mettant en évidence nos choix, observations et conclusions, tout en démontrant notre engagement dans une approche scientifique rigoureuse.

2 Description du jeu de données

Le jeu de données sélectionné est disponible sur kaggle Ce jeu de données a été recueilli dans le cadre d'une enquête visant à élaborer une segmentation de la clientèle afin de définir une stratégie marketing. Il résume le comportement d'utilisation d'environ 9000 détenteurs de cartes de crédit actifs au cours des 6 derniers mois. Les données sont organisées au niveau du client et comprennent 18 variables comportementales.

Le dictionnaire des données pour le jeu de données sur les cartes de crédit est le suivant : *CUST ID*, *BALANCE*, *BALANCE FREQUENCY*, *PURCHASES*, *ONEOFF PURCHASES*, *INSTALLMENTS PURCHASES*, *CASH ADVANCE*, *PURCHASES FREQUENCY*, *ONEOFF PURCHASES FREQUENCY*, *PURCHASES INSTALLMENTS FREQUENCY*, *CASH ADVANCE FREQUENCY*, *CASH ADVANCE TRX*, *PURCHASES TRX*, *CREDIT LIMIT*, *PAYMENTS*, *MINIMUM PAYMENTS*, *PRC FULL PAYMENT*, *TENURE*. Pour des descriptions plus détaillées de ces variables, veuillez vous référer à l'annexe.

# Age	Gender	# Income	Education	Marital Sta...	# Number of...
25	Female	50000	Bachelor's Degree	Single	0
30	Male	100000	Master's Degree	Married	2
35	Female	75000	Doctorate	Married	1
40	Male	125000	High School Diploma	Single	0
45	Female	100000	Bachelor's Degree	Married	3
50	Male	150000	Master's Degree	Married	0

Figure 1: Aperçu de la *dataframe*

2.1 Gestion des valeurs manquantes

Initialement, une vérification a révélé que l'ensemble de données comportait au total 313 valeurs manquantes, dont 312 étaient présentes dans la colonne 'MINIMUM PAYMENTS' et une seule dans la colonne *'CREDIT LIMIT'*. Pour remédier à cela, la stratégie suivante a été mise en œuvre : la ligne contenant la valeur manquante dans la colonne *'CREDIT LIMIT'* a été supprimée, car il s'agissait d'une seule observation. En ce qui concerne la colonne *'MINIMUM PAYMENTS'*, les valeurs manquantes ont été remplacées par la médiane de cette colonne. Cette approche a permis de conserver l'intégrité des données tout en traitant efficacement les valeurs manquantes, fournissant ainsi un ensemble de données prêt pour une analyse plus approfondie.

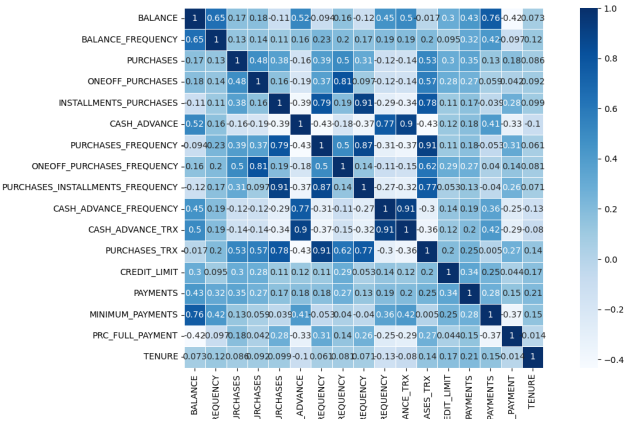
2.2 Gestion des variables catégorique

La seule variable catégorique qui se présente dans notre jeu de donnée est la *CUST ID* .Nous avons décidé de supprimer cette colonne vu qu'elle ne porte pas d'informations pertinentes pour le processus de clustering. elle est utilisée pour identifier de manière unique chaque Client, mais ils ne contribuent pas à la similarité ou à la distance entre les observations.

2.3 Distribution et valeurs aberrantes

Afin d'améliorer la distribution des données et de réduire l'impact des valeurs aberrantes, nous avons appliqué une transformation logarithmique sur l'ensemble de jeu de données. Cette transformation facilite à notre modèle l'interprétation des résultats du *clustering* en rendant les distributions plus symétriques.

2.4 Analyse bivariable



L'analyse de la matrice de corrélation révèle des relations intéressantes entre les variables du jeu de données.

La corrélation la plus forte, de 0.92, entre PURCHASES et ONEOFF PURCHASES, suggère une relation presque parfaite entre le montant total des achats effectués et le montant maximum d'achat réalisé en une seule fois. Cela indique que les clients qui effectuent des achats plus importants ont tendance à réaliser des achats significatifs en une seule fois.

D'autre part, la corrélation négative faible entre BALANCE et PRC FULL PAYMENT indique une tendance où les clients ayant un solde plus élevé sont moins enclins à effectuer des paiements intégraux. Cette relation pourrait être attribuée au fait que les clients qui maintiennent un solde plus élevé préfèrent peut-être payer partiellement plutôt que de régler intégralement leur solde.

La corrélation de -0.16 entre BALANCE FREQUENCY et PRC FULL PAYMENT indique une faible relation négative entre la fréquence de mise à jour du solde et le pourcentage de paiement intégral. Cela suggère que les clients qui mettent fréquemment à jour leur solde ont légèrement moins tendance à effectuer des paiements intégraux.

L'analyse de ces relations met en lumière des tendances générales, mais il est important de souligner que ces corrélations ne capturent pas nécessairement toute la complexité des comportements financiers des clients. Les relations entre ces variables sont multifactorielles, et d'autres facteurs non pris en compte dans l'analyse de corrélation peuvent influencer les comportements financiers individuels.

C'est précisément pour cette raison que l'utilisation d'algorithmes de clustering devient pertinente. Les algorithmes de clustering permettent d'identifier des caractéristiques similaires au sein du jeu de données, re-

groupant les clients ayant des comportements financiers plus similaires. En utilisant ces techniques de regroupement, nous pourrions obtenir des segments de clients plus homogènes en termes de comportements financiers, offrant ainsi une compréhension plus approfondie et nuancée des différents profils de clients.

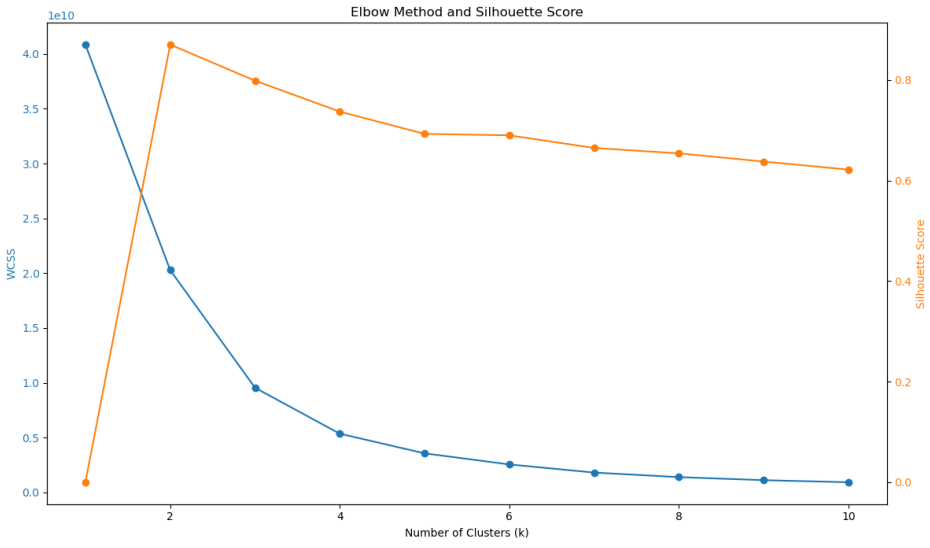
3 Implémentation de l'algorithmes de Clustering

3.1 Réduction de dimensionnalité PCA

La première étape de cette démarche consiste à appliquer une réduction de dimensionnalité sur la dataset à l'aide de l'analyse en composantes principales (PCA). Cela est particulièrement bénéfique dans le contexte du clustering, où la recherche de patterns significatifs est plus efficace dans un espace de dimension réduite.

3.2 Nombre optimal de clusters (k)

Ensuite, la démarche se poursuit avec la détermination du nombre optimal de clusters (k), en utilisant la méthode du coude (Elbow Method) et l'analyse du silhouette score. Ces deux métriques sont des outils cruciaux pour évaluer la cohérence et la compacité des clusters. La méthode du coude vise à identifier le point où l'ajout de clusters supplémentaires n'apporte plus une réduction significative de la variance intra-cluster (WCSS), tandis que le silhouette score mesure à quel point chaque point de données est similaire à son propre cluster par rapport aux clusters voisins.

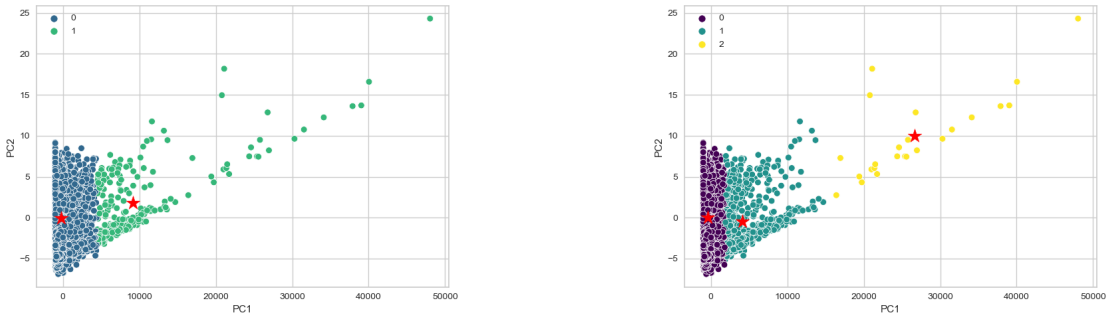


les valeurs de k=2 et k=3 sont deux options qui présentent des scores de

silhouette élevés et des WCSS relativement bas, offrant une indication que la structure des clusters pourrait être bien capturée avec l'un ou l'autre de ces deux choix de k. Nous allons donc établir une analyse plus approfondie afin de déterminer la meilleure option en fonction de notre contexte spécifique .

3.3 Comparaison des résultats avec différents valeur de k

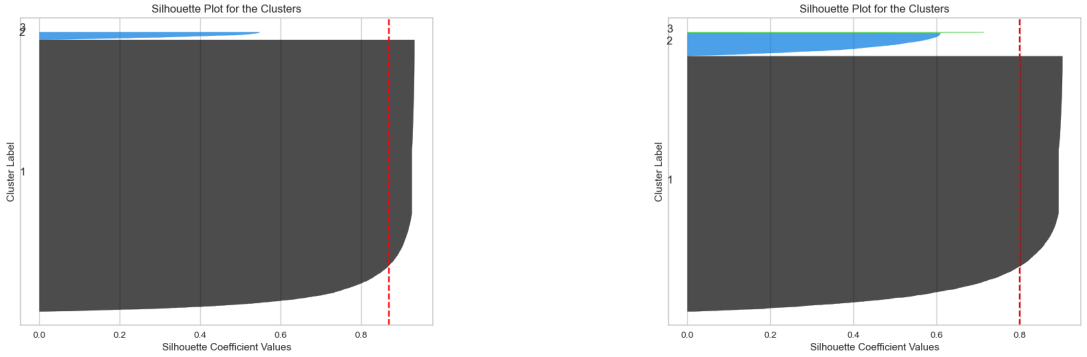
2D clustering:



(a) k=2 (b) k=3

Figure 2: Comparison of 2D clustering

Les graphiques 2D pour les clusters k=2 et k=3 permettent de visualiser la répartition des points de données dans l'espace réduit par PCA. Visuellement k=3 permet une meilleur séparation Silouhette Score:



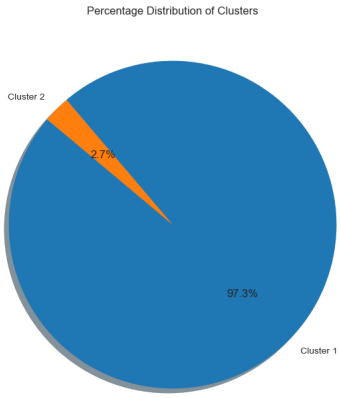
(a) k=2 (b) k=3

Figure 3: Comparison of Silouhette Score

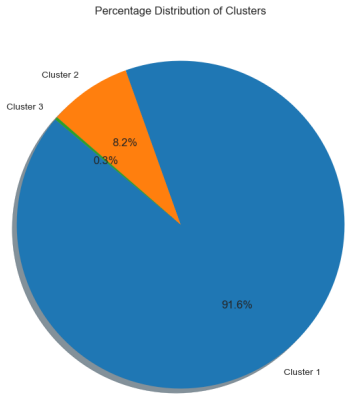
Les scores de silhouette mesurent la cohésion et la séparation des clusters. Un score de silhouette élevé indique que les points d'un cluster sont

similaires entre eux et différents des points des autres clusters.pour k=2 le score de silhouette est légèrement supérieur à celui obtenu pour k=3.

Clustering percentage:



(a) k=2



(b) k=3

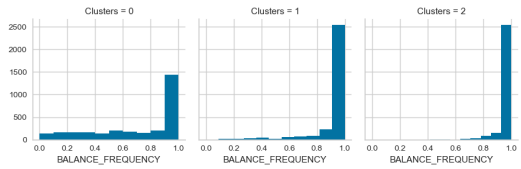
Figure 4: Comparison of Clustering percentage

En examinant la répartition des clients dans les clusters, pour k=2, le premier cluster représente 97.3 des clients, tandis que le deuxième cluster en représente seulement 2.7. Pour k=3, le premier cluster englobe la majorité des clients avec 91.6, suivi du deuxième cluster avec 8.2, et enfin, le troisième cluster représente une proportion marginale de 0.3.

Cette observation suggère que le modèle avec k=2 offre une segmentation plus concentrée, avec un cluster dominant et un autre moins fréquent, tandis que le modèle avec k=3 introduit une complexité supplémentaire avec un troisième cluster.

En fonction de ces résultats, le modèle avec k=3 semble mieux représenter la structure sous-jacente des données, offrant une segmentation plus nette et concise de la clientèle.

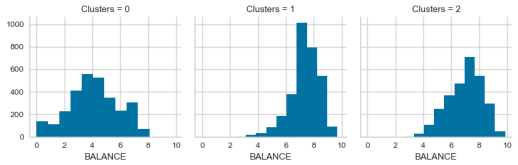
4 Caractéristiques des Clusters



(a) BALANCE FREQUENCY

(b) CASH ADVANCE FREQUENCY

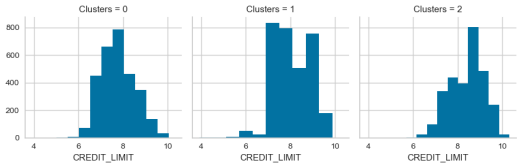
(a) MINIMUM PAYMENTS



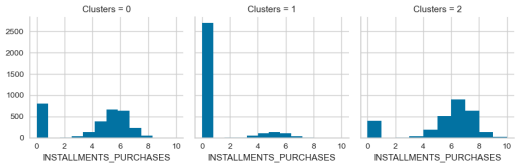
(b) BALANCE

(a) CASH ADVANCE

(b) CASH ADVANCE TRX



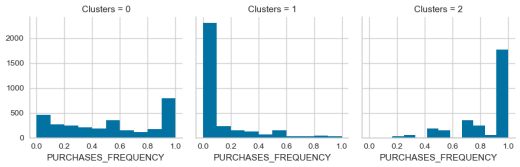
(a) CREDIT LIMIT



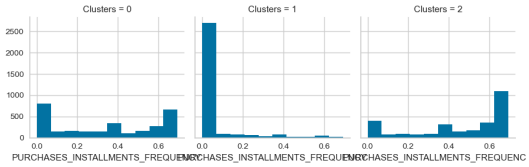
(b) INSTALLMENTS PURCHASES

(a) ONEOFF PURCHASES FREQUENCY

(b) PAYMENTS



(a) PF



(b) PIF

4.1 Cluster 1:

Cash Advance Frequency : Les clients de ce cluster ont une fréquence d'utilisation faible des avances de trésorerie, indiquant une utilisation peu fréquente de ce service.

Balance: Les clients de ce cluster ont tendance à avoir un solde faible sur leur compte. **Cash Advance** : Le montant des avances de trésorerie pris par les clients de ce cluster est relativement élevé.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence faible.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre élevé de transactions d'avance de trésorerie.

4.2 Cluster 2:

Cash Advance Frequency : Les clients de ce cluster utilisent fréquemment les services d'avance de trésorerie.

Balance : Les clients de ce cluster maintiennent un solde élevé sur leur compte.

Cash Advance : Le montant des avances de trésorerie pris par les clients de ce cluster est faible.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence élevée.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre faible de transactions d'avance de trésorerie.

4.3 Cluster 3:

Cash Advance Frequency : Les clients de ce cluster ont une fréquence d'utilisation faible des avances de trésorerie.

Balance : Les clients de ce cluster ont tendance à avoir un solde faible sur leur compte.

Cash Advance : Le montant des avances de trésorerie pris par les clients de ce cluster est faible.

Installments Purchases : Les clients de ce cluster effectuent des achats en versements avec une fréquence faible.

Purchase Frequency : La fréquence globale d'achats pour ce cluster est non élevée.

Cash Advance TRX : Les clients de ce cluster ont un nombre élevé de transactions d'avance de trésorerie.

4.4 Résultats:

Cluster 1 : Ce cluster représente des clients qui utilisent peu fréquemment les avances de trésorerie, ont un solde faible, mais effectuent des achats relativement fréquents.

Cluster 2 : Ce cluster regroupe des clients qui utilisent fréquemment les avances de trésorerie, ont un solde élevé, et effectuent des achats réguliers, souvent en versements.

Cluster 3 : Ce cluster concerne des clients avec une faible utilisation des avances de trésorerie, un solde faible, et des achats peu fréquents.

Bien que KMeans ait permis d'obtenir les résultats de clustering souhaités, nous avons également exploré d'autres algorithmes tels que DBSCAN et OPTICS pour évaluer leur performance. Cependant, ces méthodes n'ont pas produit des résultats aussi satisfaisants que KMeans dans le contexte spécifique de notre ensemble de données sur les cartes de crédit.

DBSCAN et OPTICS sont des algorithmes de clustering basés sur la densité qui peuvent être plus adaptés à des distributions de données spécifiques. Dans notre cas, les caractéristiques de l'ensemble de données et la structure des clusters semblaient mieux correspondre au modèle de KMeans, qui divise les données en clusters de forme sphérique.

5 Conclusion

En conclusion, ce travail a été consacré à l'exploration approfondie des comportements financiers des titulaires de cartes de crédit à travers l'application de techniques avancées de clustering. Nous avons débuté par une gestion judicieuse des valeurs manquantes, la suppression des variables catégoriques non informatives, et l'atténuation des effets des valeurs aberrantes par une transformation logarithmique. L'analyse de la matrice de corrélation a permis de mettre en lumière des tendances générales entre différentes variables, soulignant l'importance du clustering pour dévoiler des relations plus complexes.

La réduction de dimensionnalité par PCA a constitué une étape cruciale, permettant de simplifier la complexité des données tout en préservant les informations essentielles. Le choix du nombre optimal de clusters (k) a été abordé à travers la méthode du coude et l'analyse du silhouette score, indiquant que k=2 et k=3 étaient des options viables.

L'implémentation de l'algorithme KMeans a abouti à des clusters distincts, mettant en lumière des profils financiers spécifiques. Une comparaison avec d'autres algorithmes tels que DBSCAN et OPTICS a souligné l'efficacité de KMeans dans notre contexte. L'analyse approfondie des clusters pour k=2 et k=3 a révélé des comportements financiers différenciés entre les groupes, apportant ainsi des insights pertinents pour des stratégies de segmentation de la clientèle.

En définitive, ce travail démontre l'importance de l'apprentissage non supervisé pour dévoiler des schémas complexes au sein de données financières. Les résultats obtenus ouvrent la voie à des applications pra-

tiques telles que la personnalisation des offres, l'optimisation du service client, et la prise de décision stratégique dans le domaine des services financiers. Il est également essentiel de souligner que la compréhension approfondie des caractéristiques des algorithmes de clustering et le choix judicieux des paramètres sont cruciaux pour obtenir des résultats pertinents et exploitables.

6 Annexe

CUST ID : Identification du titulaire de la carte de crédit (Catégorique)
BALANCE : Montant du solde restant sur le compte pour effectuer des achats
BALANCE FREQUENCY :Fréquence de mise à jour du solde, score entre 0 et 1
PURCHASES : Montant des achats effectués depuis le compte
ONEOFF PURCHASES : Montant maximum d'achat effectué en une seule fois
INSTALLMENTS PURCHASES : Montant des achats effectués en versements
CASH ADVANCE : Avance de trésorerie accordée par l'utilisateur
PURCHASES FREQUENCY:Fréquence des achats, score entre 0 et 1
ONEOFF PURCHASES FREQUENCY : Fréquence des achats
PURCHASES INSTALLMENTS FREQUENCY : Fréquence des achats en versements
CASH ADVANCE FREQUENCY : Fréquence du paiement de l'avance de trésorerie
CASH ADVANCE TRX : Nombre de transactions effectuées avec "Cash in Advanced"
PURCHASES TRX : Nombre de transactions d'achat effectuées
CREDIT LIMIT : Limite de la carte de crédit pour l'utilisateur
PAYMENTS : Montant du paiement effectué par l'utilisateur
MINIMUM PAYMENTS : Montant minimum des paiements effectués par l'utilisateur
PRC FULL PAYMENT : Pourcentage du paiement intégral effectué par l'utilisateur
TENURE : Durée du service de la carte de crédit pour l'utilisateur