

Description du Dataset - Projet d'Analyse de Données

Tchenang Lilyves - Rakotovao Johanna - Sonfack Samira
Plays Octave - Smiri Bechir
Groupe 5

3 novembre 2024

Nous avons choisi de travailler avec un dataset issu de la plateforme Kaggle, accessible via le lien suivant : <https://www.kaggle.com/datasets/arjunbhasin2013/ccdata/data>. Kaggle est une plateforme largement reconnue dans le domaine de la data science, reconnue pour la qualité et la rigueur de ses ressources. Elle est largement utilisée par les chercheurs et professionnels pour partager des bases de données fiables, développer des modèles de machine learning et organiser des compétitions de haut niveau. En choisissant ce dataset, nous nous assurons d'avoir accès à des données complètes, vérifiées et conformes aux standards de la communauté scientifique.

Description du dataset

Notre dataset porte sur des informations financières relatives à des clients payant par cartes de crédit, avec 8 950 entrées et 18 colonnes. Les variables comprennent des données sur les soldes, les montants et fréquences des achats, les paiements en avance, les transactions, les limites de crédit, et les paiements. Par exemple, les colonnes incluent :

- **BALANCE** : Solde actuel de la carte,
- **PURCHASES** et **ONEOFF_PURCHASES** : Montants des achats totaux et ponctuels,
- **CASH_ADVANCE** : Avances de fonds,
- **CREDIT_LIMIT** : Limite de crédit attribuée,
- **PAYMENTS** et **MINIMUM_PAYMENTS** : Montants des paiements et des paiements minimums.

Ce dataset présente également l'avantage d'être exhaustif, sans valeurs manquantes (NaN). Cela simplifie la préparation des données et nous permet de concentrer nos efforts sur l'analyse et l'interprétation des résultats. Le volume important de données permet aussi une représentativité renforcée, essentielle pour mener une analyse de segmentation client robuste.

Objectifs

Nous souhaitons comprendre les comportements financiers des utilisateurs de cartes de crédit à partir de ces données et identifier des segments de clients ayant des comportements similaires. Voici quelques problématiques potentielles :

1. **Segmentation des clients** : Quels groupes de clients peut-on identifier selon leurs habitudes de dépenses et de gestion de crédit ? Pourrait-on classifier les clients selon leurs fréquences d'achats, leurs habitudes de paiement, ou leur recours aux avances de fonds ?
2. **Étude des comportements d'achat** : Quels sont les clients qui utilisent majoritairement des achats en un paiement ou au comptant contre ceux qui préfèrent les achats en plusieurs paiements (installments) ? Existe-t-il des corrélations entre le montant de leurs achats, leur limite de crédit et leur propension à payer le solde complet ?
3. **Analyse du risque de crédit** : Peut-on identifier des indicateurs de risque basés sur la fréquence d'achats, l'utilisation de la limite de crédit, ou les retards dans les paiements minimums ? Quels sont les profils de clients présentant le plus de risques de défaut ?

Méthodes envisagées

Pour répondre à ces questions, nous envisageons d'utiliser plusieurs techniques d'analyse et d'exploration de données vues en cours, notamment :

- **Analyse de clustering** : Nous appliquerons des méthodes telles que le k-means ou le clustering hiérarchique pour regrouper les clients en clusters selon leurs caractéristiques financières.
- **Réduction de dimension** : En utilisant la Principal component analysis (PCA), nous pourrions identifier les facteurs principaux influençant les comportements.
- **Estimation paramétrique et test d'hypothèse** : Ces méthodes nous permettront de vérifier des hypothèses sur la variabilité des comportements selon des segments ou des groupes de clients.

En conclusion, notre projet vise à identifier et à caractériser des comportements types au sein de la clientèle payant par cartes de crédit, afin de mieux comprendre les différents profils d'utilisation. Nous fournirons des visualisations et des analyses qui illustrent ces profils pour aider à formuler des recommandations.