

Projet d'Analyse de Données

2024-2025

Document de travail

1 Introduction

Le projet se déroule en binôme et compte pour la note finale (%) à définir) du module Analyse de Données Financières.

Le projet d'analyse de données est l'occasion de mettre en oeuvre les méthodes d'analyse exploratoire et de modélisation sur des données réelles.

Pour réaliser ces analyses, il sera nécessaire d'utiliser le logiciel libre R ou le langage Python. Ce projet constitue l'occasion de développer leur apprentissage. Le travail attendu consiste en un rapport d'étude correctement rédigé qui présente les résultats des analyses effectuées. Le rendu d'un rapport de projet est indispensable pour la bonne compréhension des concepts de l'analyse des données et des statistiques ; l'absence de rendu d'un rapport sera synonyme d'un zéro.

2 Partie 1

2.1 Présentation

L'objectif est d'examiner les données fournies par le service marketing d'une banque afin de comprendre et d'identifier les caractéristiques des clients qui souhaitent se désengager de la banque et ainsi de pouvoir réagir de manière proactive afin de leur fournir de meilleurs services et les amener à changer leur décision ; la finalité étant d'augmenter la fidélisation des clients.

2.2 Les axes thématiques abordées en cours

1. Chargement des Données Il s'agit de s'assurer de la mise en forme des données sous la forme d'un tableau $n \times p$, avec les individus en ligne et les variables en colonne. Il faut détecter la présence de données manquantes, de valeurs éventuellement aberrantes, s'assurer du type des différentes variables (quantitatives, qualitatives).
2. Compréhension des variables
1ère étape d'analyse descriptive. Elle fournit des informations de la manière la plus efficace possible par l'utilisation des outils statistiques classiques (moyenne, médiane, écart-type,).
3. Etude des relations entre les variables
(2 ième étape d'analyse descriptive.)
 - (a) Variables à tester
 - (b) Matrice de corrélation pour les variables démographiques et transactionnelles
4. Analyse factorielle : Facteurs sous-jacents à la résiliation
 - (a) Sélection des variables pertinentes pour l'analyse factorielle
 - (b) Conversion de variables en variables numériques
 - (c) Analyse factorielle pour réduire la dimensionnalité des données (ACP -AFD)
 - (d) Réalisation de l'analyse factorielle

- (e) Affichage des résultats de l'analyse factorielle
5. Modélisation prédictive
- (a) Modèle de régression
 - (b) Modèle de régression logistique, prédictive pour estimer la probabilité de résiliation
 - (c) Calculs d'incertitudes - Méthodes Monte-Carlo - Application aux paramètres du modèle supposés aléatoires
6. Classification des clients : k-means
- (a) La Méthode du coude Réduction de la somme de la variance intracusters grâce à l'augmentation du nombre de clusters. Plus il est élevé, plus il permet d'extraire des groupes plus fins à partir de l'analyse d'objets de données qui ont plus de similarité entre eux. On utilise le point de retournement de la courbe de la somme des variances pour choisir le bon nombre de clusters.
 - (b) Le Score de silhouette Evaluation de la qualité des clusters créés grâce aux algorithmes de clustering. Compris entre $[-1, 1]$, le score silhouette pourra être utilisé pour trouver la valeur optimale du nombre de clusters "k".
7. Visualisation Visualisation des résultats de la classification - Visualisation en 2D des clusters créés par k-means

3 Partie 2

Pour justifier ces bons résultats, la banque décide de fournir des actualités financières sur les cotations boursières qu'elle a plébicitées et sur lesquelles elle s'est appuyée pour proposer une qualité de service irréprochable.

L'objectif est ici de procéder à l'extraction des informations et à trouver des cotations d'entreprises à partir de symboles ou de noms.

Les actions à traiter sont "Tesla", "Stellantis", Une fois la cotation trouvée, il est proposé de représenter ses performances graphiquement en fonction de différents indicateurs, paramètres et durées.