Analyse des Résiliations de Cartes de Crédit

Chargement des données

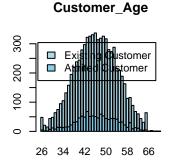
```
library(ggplot2)
library(skimr)
library(VIM)
library(outliers)
library(psych)
library(corrplot)

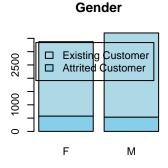
# Chargement des données
tab <- read.csv(file = "./data/credit-card-customers/cleaned_data.csv", sep = ',', dec = '.', header = '.'</pre>
```

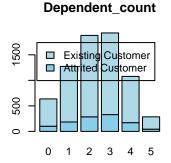
Objectif 1: Compréhension des caractéristiques démographiques

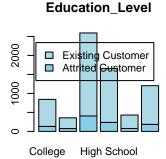
```
# Profils démographiques des clients résiliant leurs services de cartes de crédit
par(mfrow = c(2, 3))

# Analyse univariée pour chaque variable démographique
for (var in c("Customer_Age", "Gender", "Dependent_count", "Education_Level", "Marital_Status", "Income
   barplot(table(tab$Attrition_Flag, tab[[var]]), main = var, col = c("skyblue", "lightblue"), legend.te
}
```

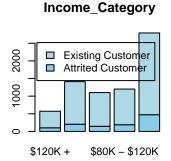












Objectif 2: Étude des relations entre les variables

Choix des variables

1. Représentativité des domaines clés :

- Customer_Age : L'âge du client est souvent un facteur important dans de nombreuses analyses démographiques.
- **Dependent_count :** Le nombre de personnes à charge peut influencer les habitudes de dépenses et la gestion financière.
- Credit_Limit : La limite de crédit est un indicateur financier clé qui peut être lié à d'autres comportements financiers.
- Total_Trans_Amt : Le montant total des transactions peut indiquer l'activité financière globale du client.
- Total_Trans_Ct : Le nombre total de transactions peut également fournir des informations sur l'utilisation des services.

2. Variables pertinentes pour la résiliation de carte de crédit :

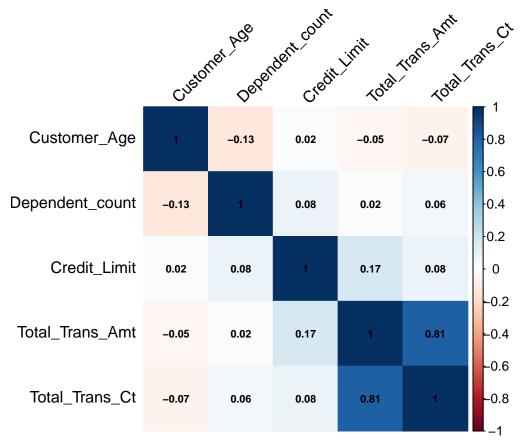
• Ces variables sont susceptibles d'influencer la décision de résiliation des services de carte de crédit. Par exemple, l'âge, la situation familiale, les habitudes de dépenses (indiquées par la limite de crédit et le montant total des transactions), et l'activité de transaction (nombre total de transactions) pourraient tous être liés à la résiliation.

3. Équilibre entre démographie et comportement financier :

• L'inclusion de variables démographiques (comme l'âge et le nombre de personnes à charge) en combinaison avec des variables transactionnelles permet d'obtenir une perspective équilibrée entre les caractéristiques personnelles et le comportement financier du client.

Matrice de corrélation

La matrice de corrélation montre les relations linéaires entre les variables démographiques et transactionnelles.



Interprétation des résultats :

• Customer_Age et Dependent_count :

- Corrélation = -0.13
- Interprétation : Il y a une corrélation négative faible entre l'âge du client et le nombre de personnes à charge. Cela suggère que, en général, les clients plus âgés ont tendance à avoir moins de personnes à charge.

• Customer_Age et Credit_Limit :

- Corrélation = 0.02
- Interprétation : La corrélation est très faible, indiquant une relation presque nulle entre l'âge du client et la limite de crédit.

• Customer_Age et Total_Trans_Amt :

- Corrélation = -0.05
- Interprétation : La corrélation est faible et négative, suggérant une tendance à une légère diminution du montant total des transactions avec l'âge.

• Customer_Age et Total_Trans_Ct :

- Corrélation = -0.07

- Interprétation : Il y a une corrélation négative faible entre l'âge du client et le nombre total de transactions. Cela suggère que les clients plus âgés ont tendance à effectuer moins de transactions.

• Dependent_count et Credit_Limit :

- Corrélation = 0.08
- Interprétation : Il y a une corrélation positive faible entre le nombre de personnes à charge et la limite de crédit. Cela pourrait signifier que les clients avec plus de personnes à charge ont tendance à avoir une limite de crédit légèrement plus élevée.

• Dependent_count et Total_Trans_Amt :

- Corrélation = 0.02
- Interprétation : La corrélation est très faible, indiquant une relation presque nulle entre le nombre de personnes à charge et le montant total des transactions.

• Dependent_count et Total_Trans_Ct :

- Corrélation = 0.06
- Interprétation : Une corrélation positive faible suggère qu'il existe une tendance à une légère augmentation du nombre total de transactions avec le nombre de personnes à charge.

• Credit_Limit et Total_Trans_Amt :

- Corrélation = 0.17
- Interprétation : Il y a une corrélation positive modérée entre la limite de crédit et le montant total des transactions. Cela indique que les clients avec une limite de crédit plus élevée ont tendance à effectuer des transactions de montant plus élevé.

• Credit_Limit et Total_Trans_Ct :

- Corrélation = 0.08
- Interprétation : La corrélation est positive faible, suggérant une relation modérée entre la limite de crédit et le nombre total de transactions.

• Total_Trans_Amt et Total_Trans_Ct :

- Corrélation = 0.81
- Interprétation : Il y a une forte corrélation positive entre le montant total des transactions et le nombre total de transactions. Cela indique que les clients effectuant un plus grand nombre de transactions ont tendance à avoir un montant total de transactions plus élevé.

Test de corrélation

Les tests utilisent la statistique de corrélation de Pearson ou corrélation linéaire.

```
## Corrélation entre Customer_Age et Dependent_count :
## P-value = 2.98963e-27
##
## Corrélation entre Customer_Age et Credit_Limit :
## P-value = 0.03722771
##
## Corrélation entre Customer_Age et Total_Trans_Amt :
## P-value = 0.0001127818
##
## Corrélation entre Customer_Age et Total_Trans_Ct :
## P-value = 3.981838e-09
##
## Corrélation entre Dependent_count et Credit_Limit :
```

```
## P-value = 7.178391e-12
##
## Corrélation entre Dependent_count et Total_Trans_Amt :
## P-value = 0.05111694
##
## Corrélation entre Dependent_count et Total_Trans_Ct :
## P-value = 1.649984e-06
##
## Corrélation entre Credit_Limit et Total_Trans_Amt :
## P-value = 4.594966e-48
##
## Corrélation entre Credit_Limit et Total_Trans_Ct :
## P-value = 7.134977e-12
```

Interprétation des résultats :

• Customer Age et Dependent count :

- Corrélation significative (p-value = 2.98963e-27).
- Interprétation : Il y a une corrélation statistiquement significative entre l'âge du client et le nombre de personnes à charge.

• Customer_Age et Credit_Limit :

- Corrélation significative (p-value = 0.03722771).
- Interprétation : Il existe une corrélation statistiquement significative entre l'âge du client et la limite de crédit, bien que la corrélation soit faible.

• Customer_Age et Total_Trans_Amt :

- Corrélation significative (p-value = 0.0001127818).
- Interprétation : Une corrélation négative significative existe entre l'âge du client et le montant total des transactions. Cela peut indiquer que les clients plus jeunes ont tendance à effectuer des transactions plus importantes.

• Customer_Age et Total_Trans_Ct :

- Corrélation significative (p-value = 3.981838e-09).
- Interprétation : Il y a une corrélation statistiquement significative entre l'âge du client et le nombre total de transactions. Les clients plus jeunes ont tendance à effectuer un nombre plus élevé de transactions.

• Dependent_count et Credit_Limit :

- Corrélation significative (p-value = 7.178391e-12).
- Interprétation : Il existe une corrélation significative entre le nombre de personnes à charge et la limite de crédit.

• Dependent_count et Total_Trans_Amt :

- Corrélation non significative (p-value = 0.05111694).
- Interprétation : Aucune corrélation statistiquement significative entre le nombre de personnes à charge et le montant total des transactions.

• Dependent count et Total Trans Ct:

- Corrélation significative (p-value = 1.649984e-06).
- Interprétation : Une corrélation significative existe entre le nombre de personnes à charge et le nombre total de transactions.

• Credit_Limit et Total_Trans_Amt :

- Corrélation significative (p-value = 4.594966e-48).
- Interprétation : Une corrélation positive significative entre la limite de crédit et le montant total des transactions. Les clients avec une limite de crédit plus élevée ont tendance à effectuer des transactions plus importantes.

• Credit_Limit et Total_Trans_Ct :

- Corrélation significative (p-value = 7.134977e-12).
- Interprétation : Corrélation significative entre la limite de crédit et le nombre total de transactions.
 Les clients avec une limite de crédit plus élevée ont tendance à effectuer un plus grand nombre de transactions.

Les résultats de notre analyse démontrent qu'il existe des relations statistiquement significatives entre les caractéristiques démographiques, telles que l'âge du client, le nombre de personnes à charge, et les transactions financières, notamment la limite de crédit, le montant total des transactions, et le nombre total de transactions. Ces liens mettent en lumière l'importance de prendre en considération ces facteurs lors de l'évaluation des motifs de résiliation des services de cartes de crédit.

Objectif 3: Facteurs sous-jacents à la résiliation

```
# Sélection des variables pertinentes pour l'analyse factorielle
selected_vars <- c("Customer_Age", "Credit_Limit", "Total_Trans_Amt")</pre>
# Résumé statistique des colonnes sélectionnées
summary(tab[, selected_vars])
                    Credit Limit
                                    Total Trans Amt
##
    Customer_Age
                         : 1438
##
          :26.00
                                    Min. : 510
  {	t Min.}
                   Min.
  1st Qu.:41.00
                  1st Qu.: 2496
                                    1st Qu.: 2090
## Median: 46.00 Median: 4287
                                    Median: 3831
## Mean
          :46.35
                  Mean : 8490
                                    Mean : 4395
## 3rd Qu.:52.00
                   3rd Qu.:10708
                                    3rd Qu.: 4740
## Max.
          :73.00
                          :34516
                                           :17995
                   Max.
                                    Max.
# Analyse factorielle pour réduire la dimensionnalité des données
nb_facteurs <- 2 # Nombre de facteurs à extraire
fact_model <- fa(tab[, selected_vars], nfactors = nb_facteurs, rotate = "varimax")</pre>
# Affichage des résultats de l'analyse factorielle
print(fact_model)
## Factor Analysis using method = minres
## Call: fa(r = tab[, selected_vars], nfactors = nb_facteurs, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##
                    MR1
                          MR2
                                  h2
                                       u2 com
## Customer Age
                   -0.01 0.30 0.088 0.91 1.0
                  0.43 0.09 0.196 0.80 1.1
## Credit Limit
## Total Trans Amt 0.43 -0.14 0.205 0.80 1.2
##
##
                         MR1 MR2
## SS loadings
                        0.37 0.12
```

```
## Proportion Var
                        0.12 0.04
## Cumulative Var
                        0.12 0.16
## Proportion Explained 0.76 0.24
## Cumulative Proportion 0.76 1.00
## Mean item complexity = 1.1
## Test of the hypothesis that 2 factors are sufficient.
## df null model = 3 with the objective function = 0.03 with Chi Square = 234.87
## df of the model are -2 and the objective function was 0
## The root mean square of the residuals (RMSR) is 0
## The df corrected root mean square of the residuals is NA
## The harmonic n.obs is 7079 with the empirical chi square 0 with prob < NA
## The total n.obs was 7079 with Likelihood Chi Square = 0 with prob < NA
## Tucker Lewis Index of factoring reliability = 1.013
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
                                                      MR1
                                                            MR2
## Correlation of (regression) scores with factors
                                                     0.56 0.34
## Multiple R square of scores with factors
                                                     0.32 0.12
## Minimum correlation of possible factor scores
                                                    -0.37 - 0.77
```

Objectif 4: Modélisation prédictive

##

(Intercept)

Customer Age

Credit_Limit

9.104 < 2e-16 ***

3.835 0.000125 ***

Estimate Std. Error z value Pr(>|z|)

-9.097e-03 4.462e-03 -2.039 0.041458 *

-1.164e-05 4.300e-06 -2.707 0.006790 **

2.360e+00 2.592e-01

Total_Trans_Amt 4.004e-04 2.112e-05 18.958 < 2e-16 ***
Total_Trans_Ct -9.493e-02 3.295e-03 -28.814 < 2e-16 ***

Dependent_count 1.108e-01 2.889e-02

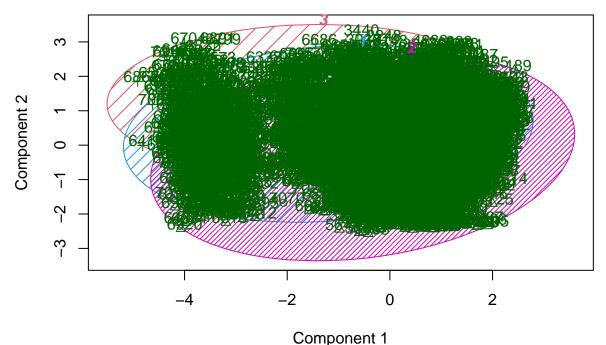
```
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 6159.3 on 7078 degrees of freedom
## Residual deviance: 4804.0 on 7073 degrees of freedom
## AIC: 4816
##
## Number of Fisher Scoring iterations: 6
```

Objectif 5: Classification des clients

```
# Utilisation de méthodes de classification (par exemple, k-means)
set.seed(123) # Pour la reproductibilité
num clusters <- 3
cluster_vars <- tab[, c("Customer_Age", "Credit_Limit", "Total_Trans_Amt", "Total_Trans_Ct")]</pre>
# Algorithme de k-means
kmeans_model <- kmeans(cluster_vars, centers = num_clusters, nstart = 20)</pre>
# Affichage des résultats de la classification
print(kmeans_model)
## K-means clustering with 3 clusters of sizes 1377, 4964, 738
##
## Cluster means:
  Customer_Age Credit_Limit Total_Trans_Amt Total_Trans_Ct
## 1
    46.45389
         13980.772
                 5173.005
                        66.97313
## 2
    46.23832
          3604.487
                 3989.158
                        63.13578
    46.88482
## 3
         31109.801
                 5674.447
                        69.21274
##
## Clustering vector:
##
  ##
  ##
 [408] \ 2\ 1\ 3\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 3\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3
## [445] 2 2 2 1 2 2 2 1 2 1 2 2 2 2 2 2 1 1 1 1 1 2 2 2 1 2 2 2 2 2 2 1 1 1 2 2 1 2 1
##
 [593] 2 3 2 2 3 2 2 2 2 1 1 2 2 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 1 2 1 3 2 3 3
##
```

```
## [6772] 2 1 2 2 1 3 3 2 3 1 1 1 1 1 1 2 2 1 2 3 2 1 2 2 1 1 1 3 2 2 1 1 3 2 2 2 3
## [7068] 2 3 1 2 2 1 1 1 2 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 47106841027 59161782918 27891462441
 (between_SS / total_SS = 80.1 %)
##
## Available components:
##
## [1] "cluster"
         "centers"
                            "tot.withinss"
               "totss"
                      "withinss"
## [6] "betweenss"
               "iter"
                      "ifault"
# Visualisation en 2D des clusters créés par k-means
library(cluster)
par(mfrow = c(1, 1))
clusplot(cluster_vars, kmeans_model$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```

CLUSPLOT(cluster_vars)



These two components explain 71.98 % of the point variability.

TODO:

Objectif 1: Compréhension des caractéristiques démographiques

Profils démographiques des clients résiliant leurs services de cartes de crédit L'analyse univariée des caractéristiques démographiques des clients permet de visualiser la distribution des résiliations pour chaque variable. Voici quelques observations :

- Customer Age: La répartition des résiliations semble similaire pour différentes tranches d'âge.
- Gender : Les résiliations semblent réparties de manière équivalente entre les genres.
- Dependent_count : La plupart des clients ayant résilié ont un faible nombre de personnes à charge.
- Education_Level : Les clients ayant résilié proviennent de divers niveaux d'éducation, mais les diplômés universitaires semblent être légèrement plus nombreux.
- Marital_Status : Les clients célibataires semblent plus enclins à résilier.
- Income Category: Les résiliations sont plus fréquentes chez les clients ayant des revenus modérés.

Objectif 3: Facteurs sous-jacents à la résiliation

L'analyse factorielle avec rotation varimax sur les variables sélectionnées (Customer_Age, Credit_Limit, Total_Trans_Amt) réduit la dimensionnalité des données. Les résultats montrent les facteurs sous-jacents aux caractéristiques sélectionnées, mais il est essentiel d'interpréter ces facteurs en fonction du contexte.

Objectif 4: Modélisation prédictive

Le modèle de régression logistique est utilisé pour estimer la probabilité de résiliation en fonction des variables disponibles. Les coefficients du modèle indiquent l'impact de chaque variable sur la probabilité de résiliation. Une analyse plus approfondie des résultats, notamment des p-values, est nécessaire pour évaluer la significativité des coefficients.

Objectif 5: Classification des clients

L'utilisation de l'algorithme de classification k-means regroupe les clients en clusters distincts en fonction de leurs caractéristiques. La visualisation en 2D des clusters montre comment les clients sont regroupés en fonction de l'âge, du montant de crédit, du total des transactions, et du total des transactions effectuées. Cette classification peut aider à identifier des segments de clients ayant des comportements similaires.

Ces résultats fournissent une base pour mieux comprendre les caractéristiques des clients associées à la résiliation des services de cartes de crédit et pour développer des stratégies visant à réduire la résiliation. Il est recommandé d'approfondir l'analyse, d'évaluer la performance du modèle prédictif sur un ensemble de test distinct et de vérifier la stabilité des clusters au fil du temps.

FAIT

Objectif 2