Étude d'une banque

Loïc Huang

Chadha Hassine

2025

Jeu de donnée

`https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn`

Table des Matières

1 Introduction

1.1 Contexte du projet

Le secteur bancaire fait aujourd'hui face à un enjeu stratégique majeur : la fidélisation de ses clients. Dans un environnement de plus en plus concurrentiel, les banques cherchent à anticiper les départs de clients, appelée *churn* ou attrition, en analysant leur profil et leur comportement. Ce projet vise à modéliser les facteurs expliquant l'attrition à partir d'un ensemble de variables socio-démographiques, comportementales et transactionnelles.

Pour cela, nous mobilisons différentes méthodes d'analyse de données étudiées dans le cadre du cours, telles que l'analyse en composantes principales (ACP), l'analyse factorielle des correspondances (AFC), la classification non supervisée (clustering) et la classification supervisée (modèles prédictifs).

1.2 Problématique

Peut-on regrouper les clients en profils homogènes afin de mieux comprendre leurs comportements bancaires ?

Peut-on prédire si un client va quitter la banque à partir de ses caractéristiques et de son comportement ?

1.3 Présentation du jeu de données

La base de données comporte les variables suivantes :

- *CLIENTNUM* : identifiant unique du client.
- *Attrition_Flag* : indique si le client a quitté la banque (*Attrited Customer*) ou s'il est encore client (*Existing Customer*).
- *Customer_Age* : âge du client en années.
- *Gender* : sexe du client (F ou M).
- *Dependent_count* : nombre de personnes à charge du client (enfants, conjoints, etc.).
- *Education_Level* : niveau d'éducation atteint par le client (ex. Graduate, High School, etc.).
- *Marital_Status* : situation maritale du client (ex. Married, Single, Divorced, Unknown).
- *Income_Category* : tranche de revenu annuel déclarée par le client (ex. Less than $40K).
- *Card_Category* : type de carte bancaire détenue par le client (Blue, Silver, Gold ou Platinum).
- *Months_on_book* : durée en mois depuis l'ouverture du compte du client.
- *Total_Relationship_Count* : nombre total de produits ou services bancaires utilisés par le client.
- *Months_Inactive_12_mon* : nombre de mois d'inactivité du client sur les 12 derniers mois.
- *Contacts_Count_12_mon* : nombre d'interactions avec le service client dans la dernière année.
- *Credit_Limit* : montant maximal de crédit autorisé pour le client.
- *Total_Revolving_Bal* : montant total du solde revolving, c'est-à-dire le montant de crédit reporté d'un mois à l'autre.

- *Avg_Open_To_Buy* : montant moyen disponible que le client peut encore dépenser sans dépasser sa limite de crédit.
- *Total_Amt_Chng_Q4_Q1* : ratio entre le montant total des transactions au quatrième trimestre et au premier trimestre.
- *Total_Trans_Amt* : montant total dépensé via des transactions par le client.
- *Total_Trans_Ct* : nombre total de transactions effectuées par le client.
- *Total_Ct_Chng_Q4_Q1* : ratio entre le nombre de transactions au quatrième trimestre et au premier trimestre.
- *Avg_Utilization_Ratio* : ratio moyen entre le montant utilisé et la limite de crédit autorisée.
- *Naive_Bayes_Classifier_..._1* : score de probabilité généré par un modèle Naive Bayes, à des fins de validation.
- *Naive_Bayes_Classifier_..._2* : deuxième score de probabilité généré par un modèle Naive Bayes, à des fins de validation.

2 Analyse exploratoire des données

2.1 Traitement des données

Nous avons effectué plusieurs étapes de traitement des données avant l'analyse :

- **Vérification des valeurs manquantes** : aucune valeur `NA` n'est présente .
- **Suppression de variables inutiles** : nous avons supprimé la variable `CLIENTNUM`, qui est un identifiant unique sans intérêt pour la prédiction. Nous avons également retiré les deux variables `Naive_Bayes_Classifier_...`, qui contiennent des scores issus du modèle Naive Bayes et qui pourraient biaiser l'entraînement de nos modèles.
- **Gestion des modalités inconnues** : certaines variables comme `Education_Level` et `Marital_Status` contiennent la modalité `"Unknown"`. Nous avons choisi de la conserver comme une modalité à part entière, car le fait de ne pas déclarer une information peut aussi révéler un comportement client spécifique utile à la modélisation.

2.2 Statistiques descriptives

Nous avons comparé la répartition des clients en fonction de leur statut d'attrition selon plusieurs variables numériques et qualitatives.

Globalement, les clients ayant quitté la banque présentent moins de transactions, des montants moins élevés et un taux d'utilisation du crédit plus faible. Pour les variables qualitatives, certaines modalités semblent plus liées à l'attrition, comme le type de carte ou la catégorie de revenu.

Les graphiques, 15 et 16, visibles en annexe, illustrent ces tendances. Le tableau détaillé issu de la commande `summary()` est disponible en annexe aussi (voir page 17).

Remarque : Pour des raisons de clarté et afin de respecter la limite imposée sur le nombre de pages, les graphiques liés à la statistiques descriptives et à la classification supervisée ont été placés en annexe.

Chaque figure mentionnée dans le corps du texte est accompagnée d'un numéro cliquable permettant d'accéder directement à l'annexe correspondante. De plus, sous chaque graphique en annexe, un numéro de section est également disponible pour revenir à la section d'analyse correspondante, facilitant ainsi la navigation sans perdre le fil de la lecture.

3 Méthode factorielle

3.1 ACP

On peut utiliser l'analyse par composante principale pour déterminer les liens entre les variables quantitatives.

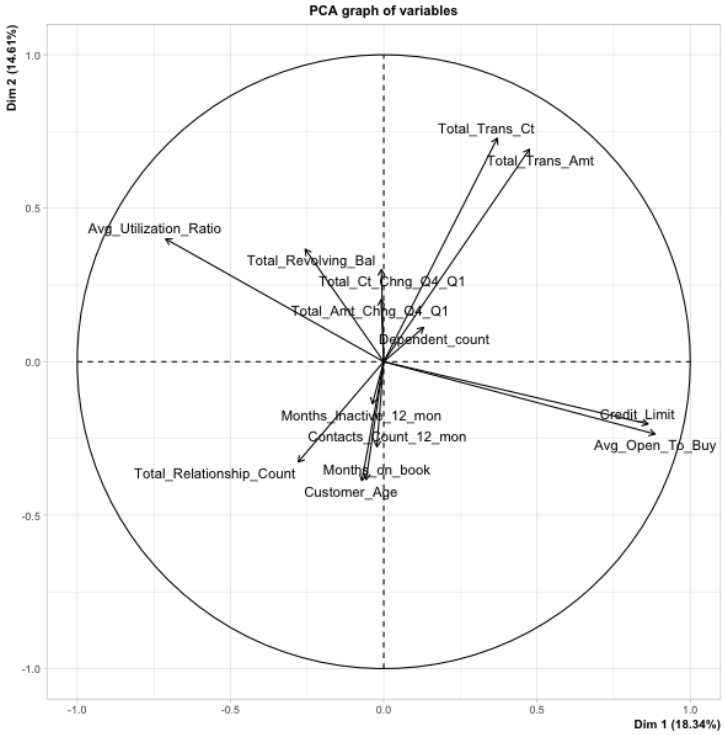

Figure 1: ACP avec toutes les variables quantitatives

Aucune combinaison d'axes ne permet d'expliquer plus de 50% de la variance donc on va chercher à supprimer des variables qui ont peu de contributions aux axes et ainsi réduire le bruit. On pourra en particulier supprimé des variables qui pourraient être difficiles à interpréter.

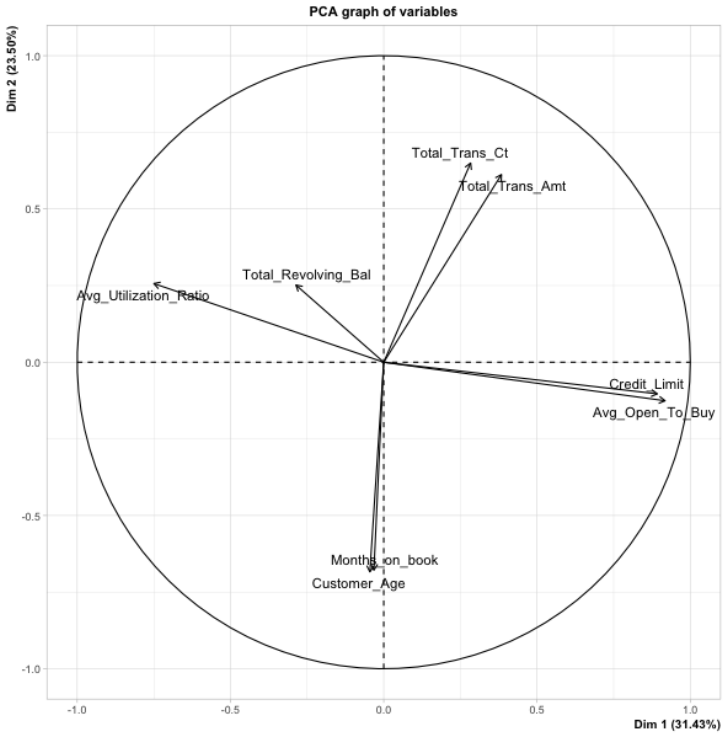

Figure 2: ACP avec les variables quantitatives importantes

Total_Trans_Ct et Total_Trans_Amt sont corrélés positivement. Les clients qui ont un nombre total de transactions élevés ont également un montant total des transactions élevé.

Months_on_book et Customer_Age sont corrélés positivement. L'ancienneté d'un client est lié à son âge.

Credit_Limit et Avg_Open_Tol_Buy sont corrélés positivement. Les clients avec une limite de crédit élevée ont également un montant disponible moyen élevé.

Total_Trans_Ct et Total_Trans_Amt sont corrélés négativement avec Months_on_book et Customer_Age. Les clients qui un nombre total de transactions élevés et qui ont un montant total de transactions élevé sont des nouveaux clients avec un jeune âge.

Credit_Limit et Avg_Open_Tol_Buy sont corrélés négativement avec Avg_Utilization_Ratio. Les clients qui ont un taux d'utilisation élevé ont tendance à avoir un crédit limite ou un montant disponible moyen faible.

Total_Revolving_Bal est mal représenté.

Toutes ces corrélations semblent cohérentes par rapport à l'usage bancaire d'un client.

3.2 AFC

On peut utiliser l'analyse factorielle des composantes pour déterminer les liens entre les variables qualitatives.

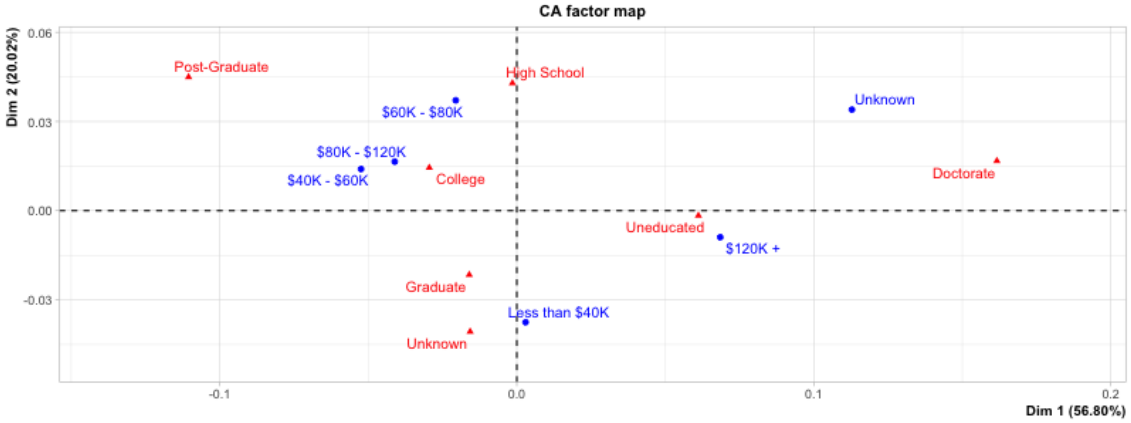

Figure 3: AFC entre le niveau d'etude et le revenu des clients

Les personnes qui n'ont pas fait d'études réussissent le mieu avec les revenues les plus haut.
Les doctorants sont aussi proche de ce revenu mais ont pour la plupart un revenu encore inconnu.
Les college et les post-graduate ont des revenues entre 40k et 120k Les niveaux lycées sont autour de 60k et 80k.
Les graduate et les personnes qui n'ont pas indiquées leurs études ont les revenus les plus faibles.

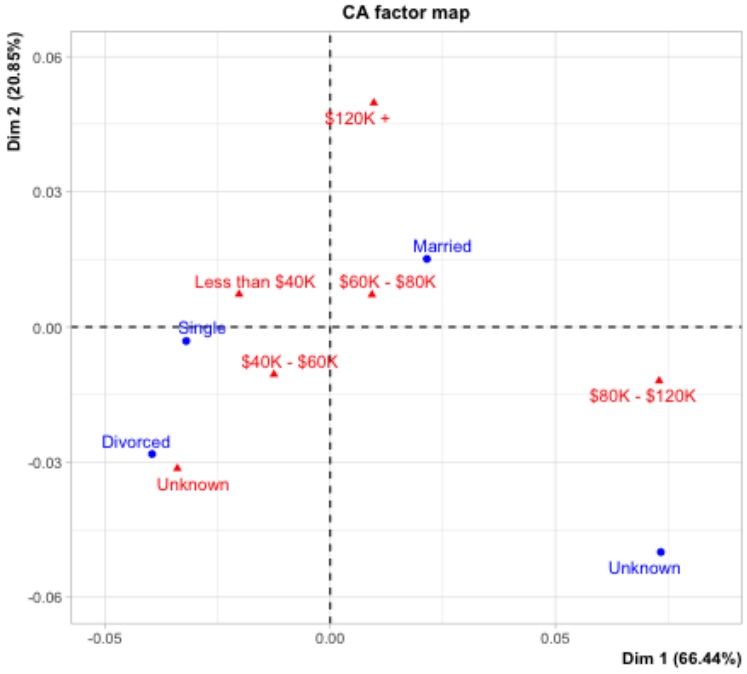

Figure 4: AFC entre le statut marital et le revenu des clients

Les clients célibataires ont des revenus de moins de 60k tandis que les personnes mariées ont des revenues plus élevés allant au dela de 60k.

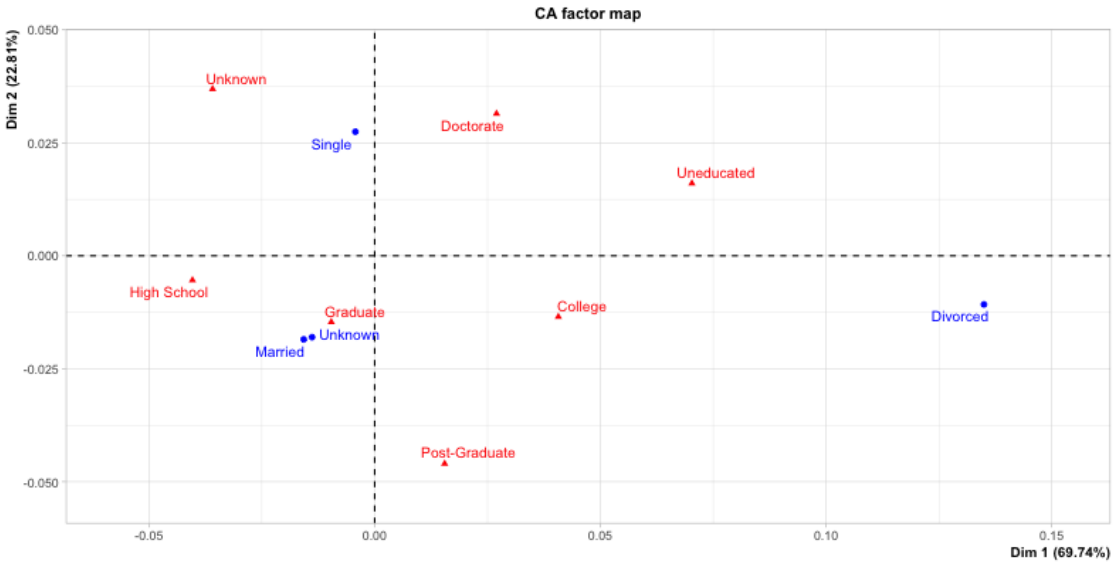

Figure 5: AFC entre le niveau d'etude et le statut marital des clients

Les clients célibataires sont en général des personnes qui sont en doctorate tandis que les personnes mariées ont un niveau d'éducation de graduate

4 Méthode de classification non supervisée

On peut utiliser le clustering pour déterminer les regroupements des clients selon certaines caractéristiques.

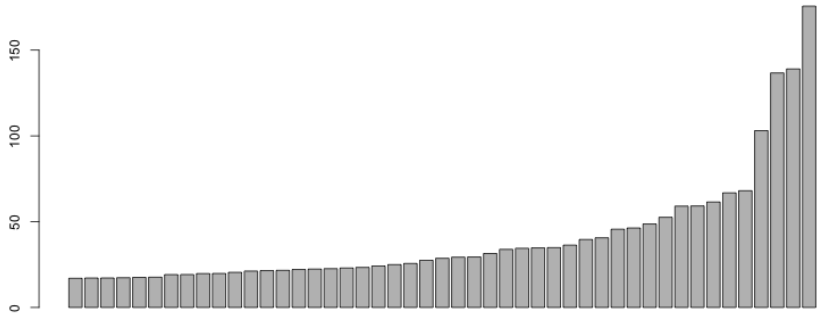

Figure 6: Diagramme de barres sur les resultats du dendogramme

La perte d'inertie semble faible jusqu'au passage de 5 à 4 classes qui est significative donc on prend k=5, donc on garde 5 classes.

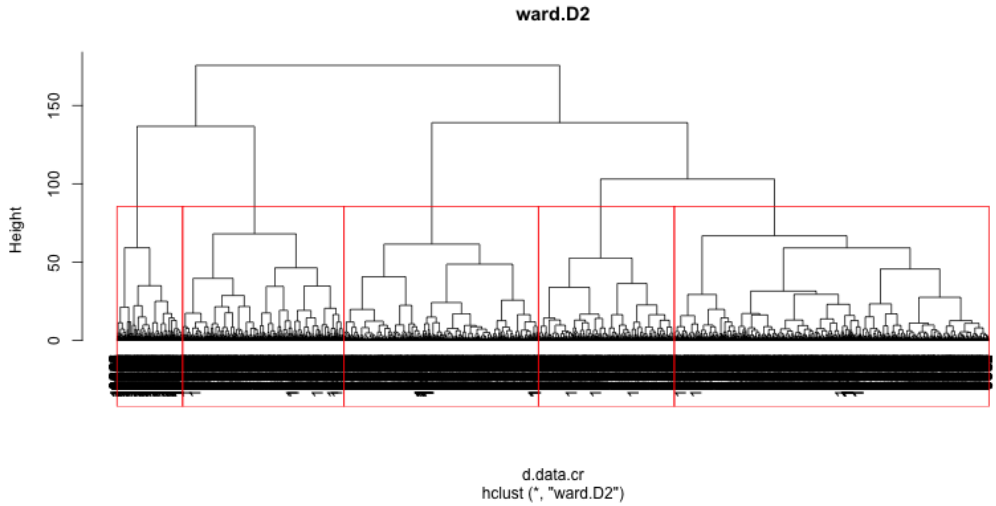

Figure 7: Dendogramme avec les classes

Le dendogramme représente la répartition des 5 classes sur le jeu de donnée.

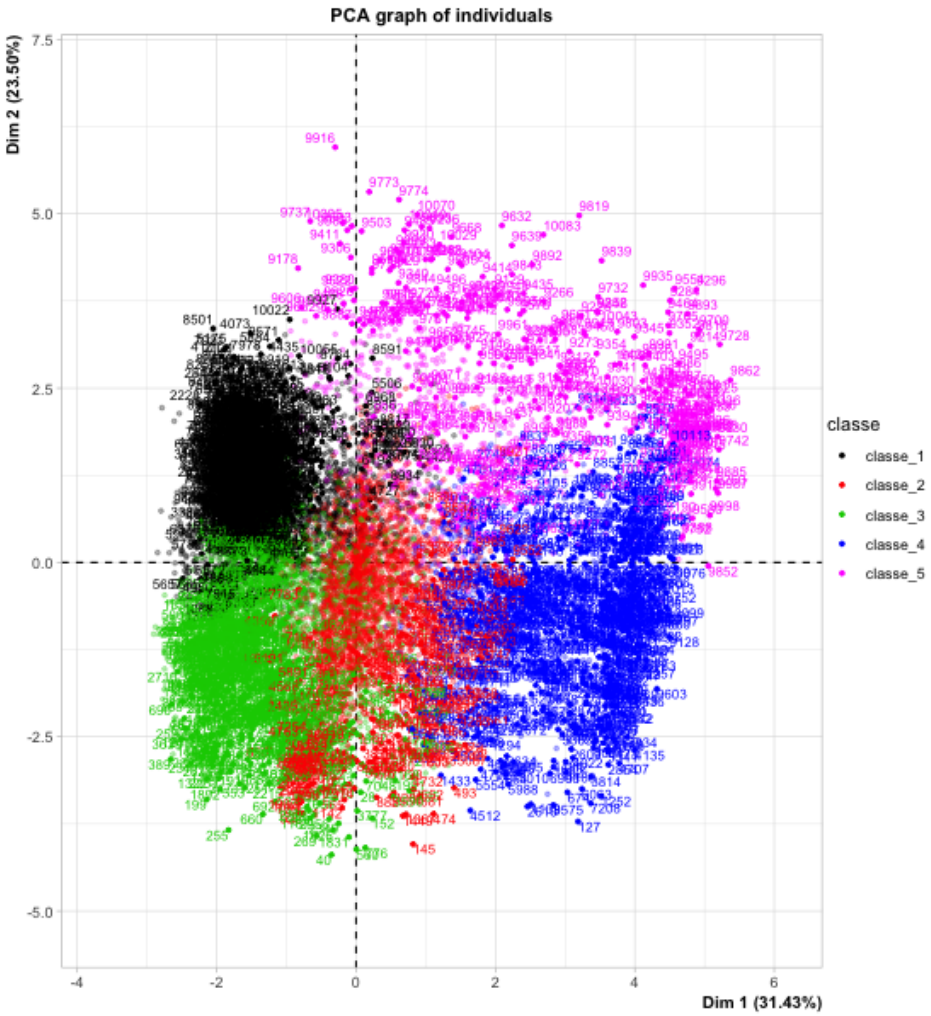

Figure 8: Representation des clients dans leurs classes respectives

Le classe 2 correspond aux clients avec Credit_Limit et Avg_Open_Tol_Buy élevés. Ce sont donc des personnes qui une limite de crédit et un montant disponible moyen élevé.

Le classe 3 correspond aux clients avec Total_Trans_Ct et Total_Trans_Amt élevé. Ce sont des personnes qui effectuent beaucoup de transactions et avec une utilisation monétaire importante.

Le classe 4 correspond aux clients avec Avg_Utilization_Ratio élevé. Ce sont des personnes qui ont tendance à beaucoup utiliser leur compte bancaire pour des achats avec montant élevé ou non.

Le classe 1 est l'opposé du classe 3 avec un nombre de transaction faible et à montant faible.

Le classe 1 et 5 correpondent aux client avec Months_on_book et Customer_Age élevé. Ce sont des clients avec une ancienneté importante dans la banque.

Group.1 <int>	Customer_Age <dbl>	Months_on_book <dbl>	Credit_Limit <dbl>	Total_Revolving_Bal <dbl>	Avg_Open_To_Buy <dbl>	Total_Trans_Amt <dbl>	Total_Trans_Ct <dbl>	Avg_Utilization_Ratio <dbl>
1	46.12262	35.93803	4745.304	103.8358	4641.468	3437.315	58.04073	0.02165516
2	55.56119	44.52505	4904.043	1658.9803	3245.063	3115.566	56.74509	0.46262460
3	43.08233	32.82823	4215.472	1557.5878	2657.885	3911.489	66.32795	0.47962199
4	45.67182	35.18517	22724.998	1178.5027	21546.496	3484.044	58.86553	0.06269637
5	44.98555	34.80946	14407.442	1342.9514	13064.490	14576.263	109.61104	0.17658081

Figure 9: Tableau des donnees pour chaque groupe

On peut déduire à partir de ce tableau que le groupe 1 correspond à la classe 5, le groupe 2 correspond à la classe 4, le groupe 3 correspond à la classe 3, le groupe 4 correspond à la classe 2 et le groupe 5 correspond à la classe 1,

5 Méthode de classification supervisée

5.1 Équilibrage des données (SMOTE)

La variable cible `Statut_attrition` est fortement déséquilibrée, avec environ 84 % de clients fidèles (*Existing Customer*) et seulement 16 % de clients partis (*Attrited Customer*). Ce déséquilibre peut biaiser les modèles de classification en les incitant à favoriser la classe majoritaire.

Pour illustrer cela, nous avons d'abord entraîné un modèle *Random Forest* sur les données d'origine, sans rééquilibrage. Les performances obtenues paraissent bonnes au premier abord, avec une erreur globale faible.

Cependant, une analyse plus fine montre que l'erreur sur la classe minoritaire (clients churnés) reste très élevée (17 %), tandis qu'elle est très faible pour la classe majoritaire. Cela indique un fort biais du modèle en faveur des clients fidèles, limitant sa capacité à détecter efficacement les churners.

Les résultats détaillés (courbe d'erreur et matrice de confusion) sont présentés en annexe, figures 17 et 18. La courbe d'erreur comprend trois courbes distinctes : la courbe noire représente l'erreur globale du modèle (out-of-bag), la courbe rouge correspond à l'erreur spécifique à la classe *Attrited Customer* (clients churnés), tandis que la courbe verte montre l'erreur pour la classe *Existing Customer* (clients fidèles).

Pour remédier à ce problème de déséquilibre et améliorer la performance du modèle, nous appliquons la méthode SMOTE. Cette technique (Synthetic Minority Over-sampling Technique) permet de générer artificiellement de nouveaux exemples pour la classe Attrited Customer, en interpolant entre les observations existantes. Elle permet ainsi de rééquilibrer le jeu de données sans perte d'information.

5.2 Modèles testés

Dans cette section, nous testons différents algorithmes de classification supervisée sur la base de données rééquilibrée (`data.train.balanced`). L'objectif est de comparer leurs performances pour prédire le statut d'attrition des clients.

5.2.1 Régression logistique

Dans cette partie, nous cherchons à identifier les variables les plus significatives dans la prédiction du churn. Pour cela, nous avons entraîné un modèle de régression logistique sur les données équilibrées, puis calculé les *odds ratios* (valeurs exponentiées des coefficients) afin de faciliter l'interprétation.

La variable *Total_Ct_Chng_Q4_Q1* (OR $\approx$ 22,6) apparaît comme la plus influente. Le tableau des odds ratios est présenté en annexe (voir figure 19).

Ensuite, une sélection automatique des variables a été réalisée via la méthode `stepAIC`, permettant de ne conserver que les variables les plus significatives.

L'analyse des coefficients montre que plusieurs variables ont un impact important sur la probabilité de churn : le genre (les hommes ont une probabilité plus élevée), le nombre de personnes à charge (effet négatif), le niveau d'éducation (notamment les titulaires d'un doctorat ou d'un diplôme post-graduate), ainsi que le statut marital. D'autres variables comme le nombre de produits bancaires détenus, l'inactivité, les contacts avec le service client, le solde revolving, le nombre total de transactions ou encore la variation d'activité entre les trimestres (Q4 vs Q1) sont également fortement discriminantes.

L'ensemble des coefficients estimés est disponible en annexe (voir tableau 1).

5.2.2 Régression logistique Lasso

La régression Lasso permet la sélection automatique des variables les plus pertinentes, tout en réduisant le surapprentissage.

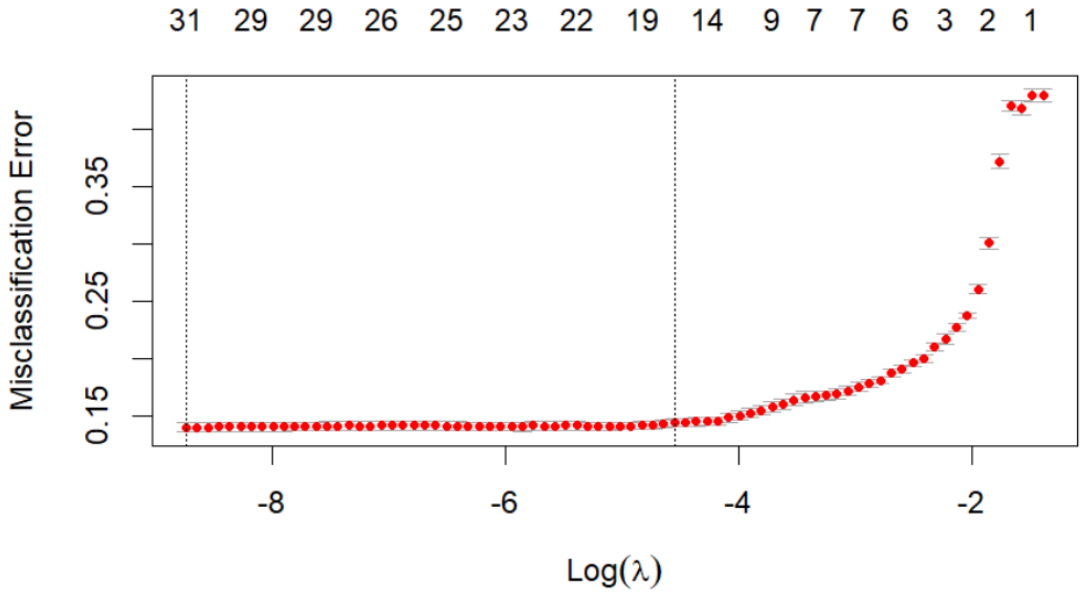

Figure 10: Erreur de classification selon la valeur de λ dans la régression Lasso

Nous avons utilisé une validation croisée pour sélectionner la valeur optimale du paramètre de régularisation λ. Une fois ce paramètre choisi, le modèle a été entraîné sur les données équilibrées, puis utilisé pour faire des prédictions sur l'échantillon test.

5.2.3 Arbre de classification (CART)

Le modèle CART permet de visualiser les règles de décision utilisées pour prédire le statut d'attrition. L'arbre généré met en évidence que le nombre total de transactions (`Total_Trans_Ct`), le solde revolving (`Total_Revolving_Bal`) et le montant total des transactions (`Total_Trans_Amt`) jouent un rôle déterminant dans la séparation entre clients fidèles et clients partis.

On observe par exemple que les clients ayant effectué moins de 58 transactions sont davantage susceptibles de quitter la banque. D'autres variables comme le nombre de produits détenus (`Total_Relationship_Count`) ou le changement de volume de transactions (`Total_Ct_Chng_Q4_Q1`) interviennent aussi dans les décisions.

Pour éviter le surapprentissage, nous avons optimisé la taille de l'arbre à l'aide de la méthode de *pruning*. Le paramètre de complexité (`cp`) optimal a été choisi automatiquement à partir du tableau des erreurs croisée (`cptable`). La valeur minimisant l'erreur de classification est `cp = 0.01`. L'arbre correspondant est plus simple et généralise mieux sur de nouvelles données.

L'arbre final, obtenu après élagage, est présenté ci-dessous (figure 11).

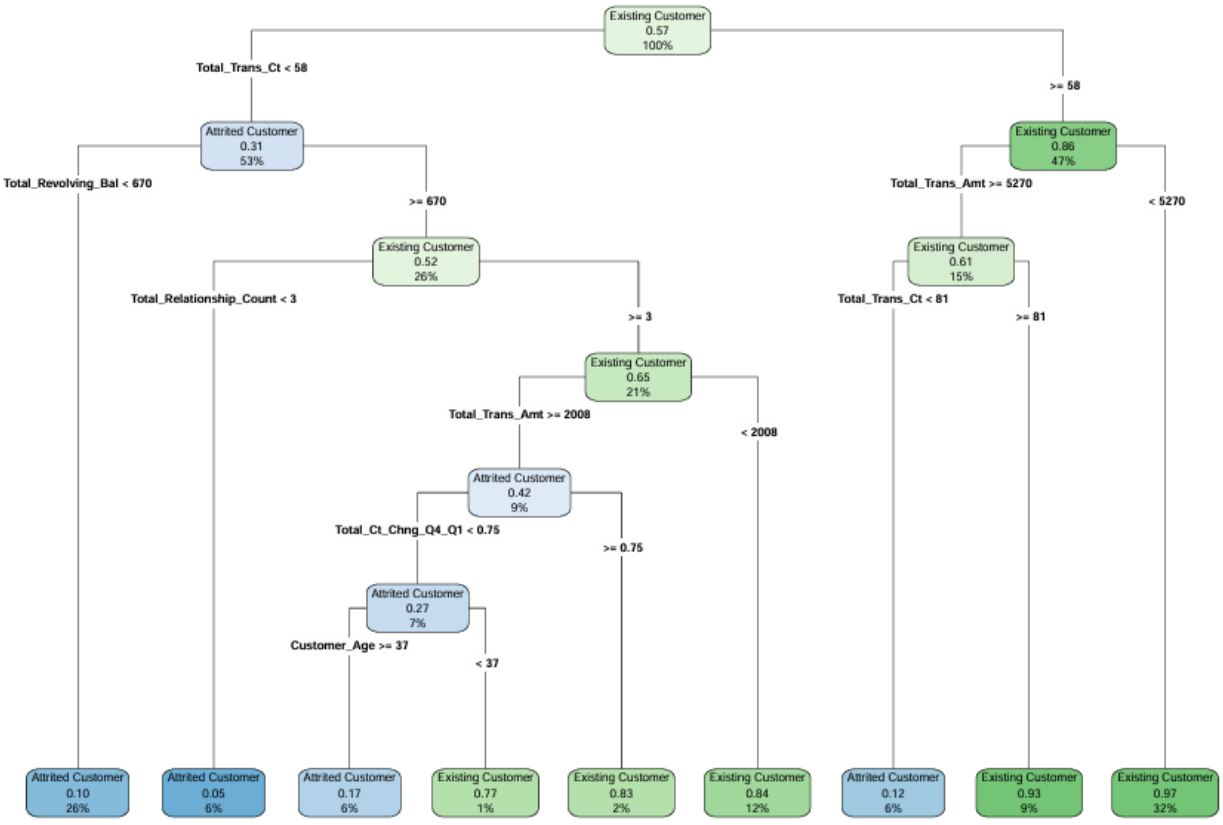

Figure 11: Arbre optimal de décision généré par le modèle CART

5.2.4 Random Forest (après équilibrage)

Le modèle Random Forest a été entraîné sur les données équilibrées issues de SMOTE. Il affiche une excellente performance avec un taux d'erreur global OOB de seulement 2,04 %. Les erreurs de classification sont faibles et bien réparties entre les deux classes : environ 2,3 % pour les clients churners et 1,85 % pour les clients fidèles. Les détails de l'apprentissage (structure du modèle, taux d'erreur et répartition des prédictions) sont fournis en annexe (Figures 20 et 21).

5.2.5 Adaboost

Le modèle Adaboost a d'abord été entraîné avec les paramètres par défaut sur 100 arbres. Afin d'améliorer les performances, nous avons ensuite cherché à déterminer le nombre optimal d'itérations (B) en utilisant une validation croisée à 5 folds, avec différentes valeurs maximales (3000, 8000, 12000 arbres).

À chaque fois, la fonction `gbm.perf` a affiché une courbe où la perte pour le jeu d'entraînement (courbe noire) et la perte estimée par validation croisée (courbe verte) se superposent presque parfaitement. Ce comportement indique que le modèle ne montre pas de signe de sur-apprentissage : la performance sur les données d'entraînement est très proche de celle en validation.

Nous avons également testé avec `cv.folds = 10`, mais l'affichage reste identique. Cela peut s'expliquer par une structure de données relativement simple après équilibrage (SMOTE), où la séparation entre les classes est suffisamment nette pour que le modèle généralise bien, quel que soit le nombre d'itérations.

La figure illustrant la courbe de sélection de B est présentée en annexe (Figure 22).

5.3 Comparaison des performances

Pour évaluer les performances des modèles testés, nous avons comparé à la fois l'accuracy et la courbe ROC (Receiver Operating Characteristic), en mettant en évidence la surface sous la courbe (AUC) pour chaque méthode.

Les résultats montrent que le modèle Random Forest est celui qui obtient les meilleures performances globales, avec une accuracy de 95,8 % et une AUC de 0,985. Il est suivi de près par AdaBoost (accuracy : 93,8 %, AUC : 0,974) et la régression logistique avec pénalisation Lasso (accuracy : 86,7 %, AUC : 0,927).

Les courbes ROC comparées permettent de visualiser la capacité de chaque modèle à distinguer les churners des clients fidèles. Le tableau des accuracy et AUC est également présenté ci-dessous.

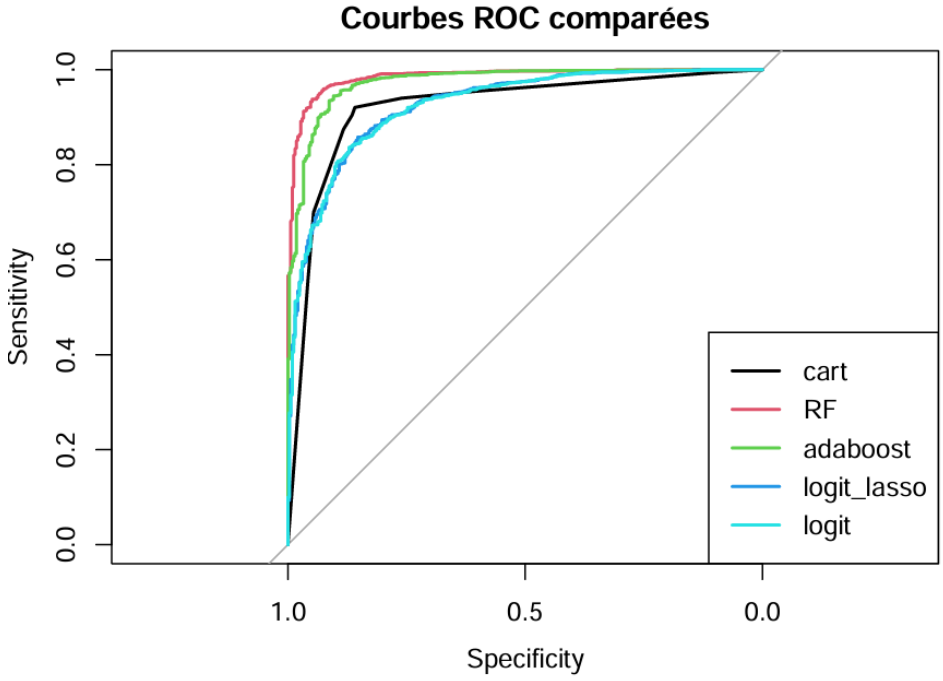

Figure 12: Courbes ROC comparées des modèles

```
##              cart        RF  adaboost logit_lasso     logit
## accuracy 0.9106173 0.9580247 0.9377778   0.8671605 0.8656790
## AUC      0.9210460 0.9851234 0.9742189   0.9265968 0.9258159
```

Figure 13: Accuracy et AUC pour chaque modèle

Remarque : Les tables de confusion des différents modèles sont disponibles en annexe, aux figures 23, 24, 25 et 26.

5.4 Analyse de l'importance des variables

Pour mieux comprendre les facteurs influençant le *churn*, nous avons analysé l'importance des variables selon trois méthodes d'apprentissage : **CART**, **Random Forest** et **Adaboost**. Ces modèles permettent tous d'estimer l'impact relatif de chaque variable dans la classification, mais chacun le fait avec un critère spécifique.

Avec **CART**, les variables les plus importantes sont le nombre total de transactions effectuées (*Total_Trans_Ct*) et le montant total (*Total_Trans_Amt*). Ces deux variables sont également les premières utilisées dans l'arbre optimal, ce qui confirme leur rôle central dans la prédiction.

Le modèle **Random Forest**, qui agrège plusieurs arbres, confirme l'importance de ces deux mêmes variables. Il met aussi en avant la variable *Total_Revolving_Bal* (solde renouvelable) et *Total_Ct_Chng_Q4_Q1* (changement de fréquence des transactions entre les trimestres), ce qui suggère que la dynamique de comportement joue un rôle fort dans le churn.

Enfin, **Adaboost**, en se concentrant sur les erreurs de classification successives, fait ressortir de manière marquée *Total_Relationship_Count* comme la variable la plus influente, suivie de *Contacts_Count_12_mon* et *Months_Inactive_12_mon*, soulignant l'importance des interactions avec la banque.

Les graphiques comparant l'importance des variables sont disponibles en annexe (Figures 27, 28 et 29).

6 Lien entre l'analyse factorielle et la classification supervisée

Avec l'ACP, il n'a pas été possible d'obtenir un graphique des individus suffisamment lisible permettant de distinguer les deux classes *Existing Customer* et *Attrited Customer*. Nous avons donc appliqué la méthode FAMD sur notre base de données, ce qui a permis une meilleure visualisation. Les résultats obtenus sont globalement similaires à ceux de l'ACP sur le plan des axes factoriels.

Le graphique des individus (Figure 14) montre une séparation partielle entre les deux classes de la variable cible *Attrition_Flag*. Les clients fidèles (*Existing Customer*) sont majoritairement regroupés dans la partie droite du plan factoriel, tandis que les clients perdus (*Attrited Customer*) apparaissent davantage dans la partie inférieure gauche.

On remarque que les clients fidèles présentent des valeurs élevées de *Total_Trans_Ct* et *Total_Trans_Amt*, ce qui confirme les résultats obtenus par les méthodes supervisées.

En effet, lors de l'analyse de l'importance des variables dans les modèles supervisés (CART, Random Forest et Adaboost), ces deux variables ont été identifiées comme étant parmi les plus influentes pour prédire le churn, aux côtés de *Total_Revolving_Bal*, *Total_Ct_Chng_Q4_Q1* et *Total_Relationship_Count*. La FAMD met également en évidence certaines de ces variables, notamment

Total_Revolving_Bal et *Avg_Utilization_Ratio*, en cohérence avec les résultats issus des modèles de régression logistique.

Ainsi, l'analyse factorielle non supervisée offre une vision exploratoire cohérente avec les résultats prédictifs issus de la classification supervisée. Elle permet de confirmer visuellement et structurellement les variables clés associées au churn.

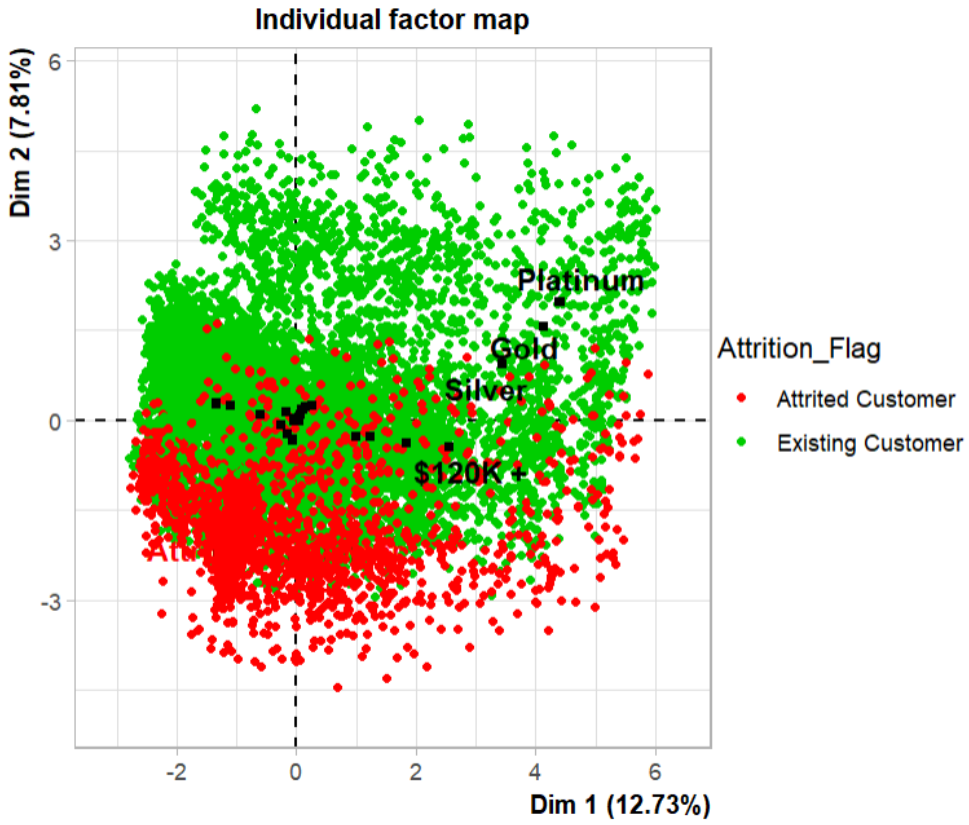

Figure 14: Carte factorielle des individus selon la FAMD (en fonction de la variable *Attrition_Flag*)

7 Conclusion

Ce projet nous a permis d'explorer et de modéliser les comportements clients d'une banque, dans le but de mieux comprendre les facteurs associés à l'attrition.

L'ACP nous a permit de comprendre l'influence de certains paramètres d'un compte bancaire sur les autres. La consommation des clients est plus élevé à un jeune âge. Un compte avec un montant disponible faible suggère que le client à un ratio d'utilisation de la carte élevé.

L'AFC nous a permit de savoir les liens entre les différents profils des clients avec le niveau d'étude, le revenu annuel et le statut marital.

La classification non supervisée nous a permit de répartir tous les clients dans des différents groupes qui ont été formés selon les profils de consommation bancaire relevés.

La classification supervisée nous a permit de prédire si un client peut quitter la banque selon des caractéristiques liées à ses dépenses.

Enfin, la convergence entre les variables importantes issues de la FAMD et celles sélectionnées dans les modèles prédictifs renforce la robustesse de nos conclusions..

A Annexes

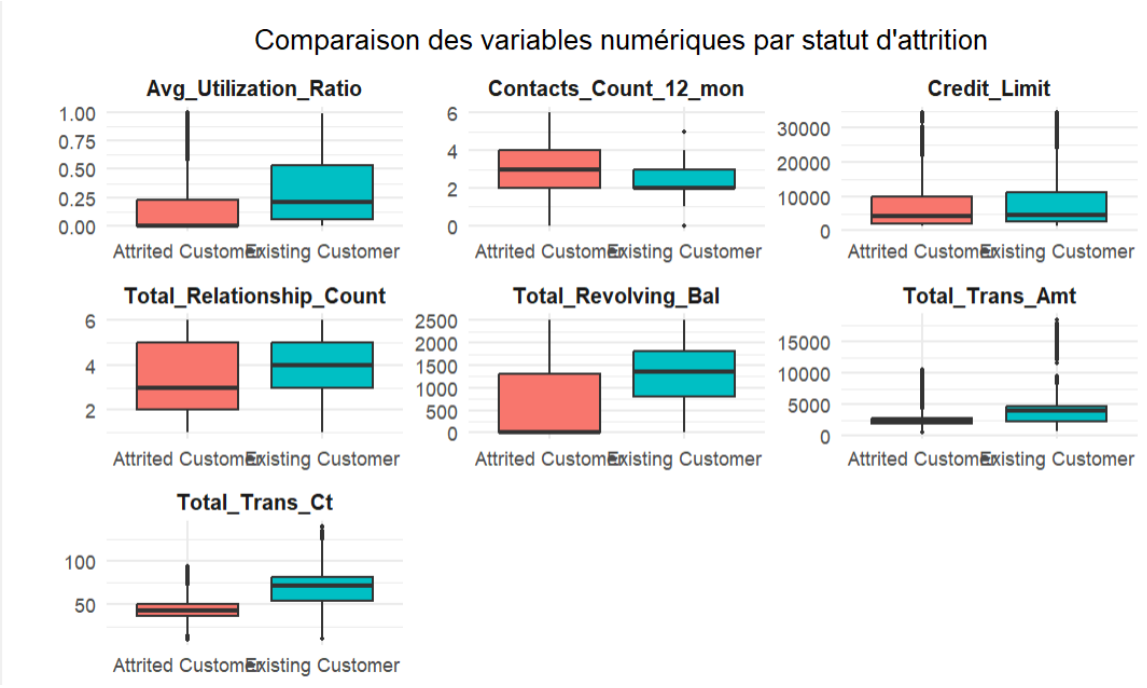

Figure 15: Comparaison des variables numériques selon le statut d'attrition

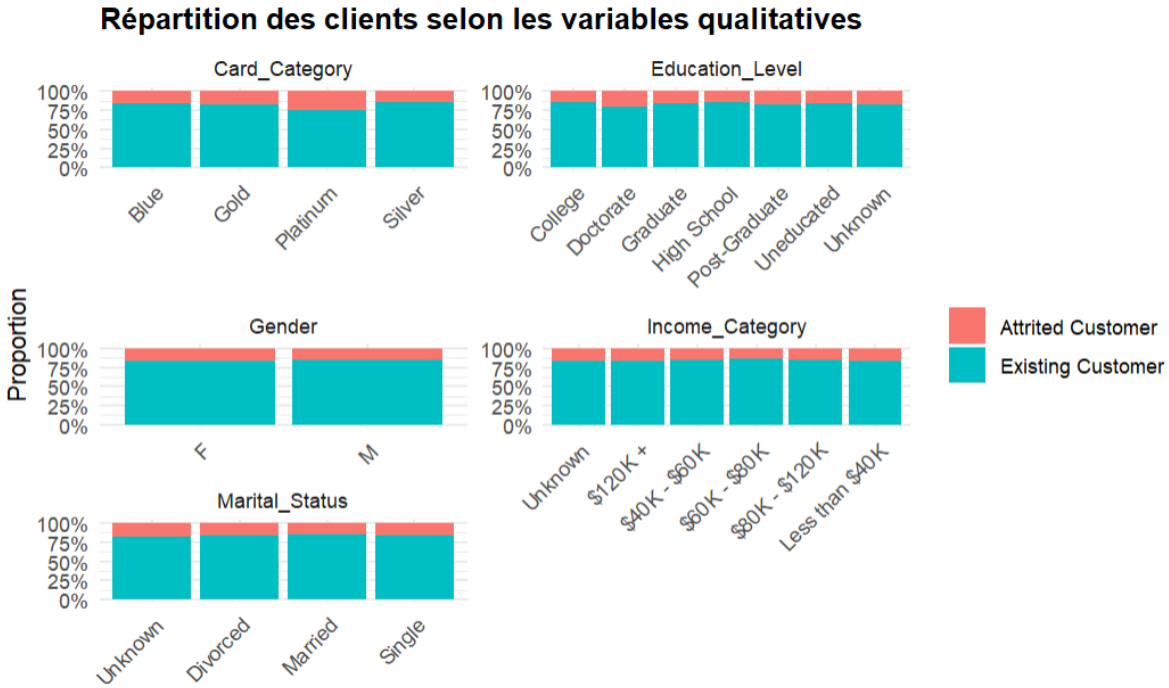

Figure 16: Répartition des clients selon les variables qualitatives

Voir interprétation dans la section 2.2.

Résumé statistique des données

```
          Attrition_Flag  Customer_Age   Gender   Dependent_count      Education_Level
Attrited Customer:1627   Min.   :26.00   F:5358   Min.   :0.000   College      :1013
Existing Customer:8500   1st Qu.:41.00   M:4769   1st Qu.:1.000   Doctorate    : 451
                         Median :46.00            Median :2.000   Graduate     :3128
                         Mean   :46.33            Mean   :2.346   High School  :2013
                         3rd Qu.:52.00            3rd Qu.:3.000   Post-Graduate: 516
                         Max.   :73.00            Max.   :5.000   Uneducated   :1487
                                                                  Unknown      :1519
 Marital_Status       Income_Category  Card_Category  Months_on_book
Divorced: 748   $120K +       : 727   Blue    :9436   Min.   :13.00
Married :4687   $40K - $60K   :1790   Gold    : 116   1st Qu.:31.00
Single  :3943   $60K - $80K   :1402   Platinum:  20   Median :36.00
Unknown : 749   $80K - $120K  :1535   Silver  : 555   Mean   :35.93
                Less than $40K:3561                   3rd Qu.:40.00
                Unknown       :1112                   Max.   :56.00

Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon  Credit_Limit
Min.   :1.000            Min.   :0.000          Min.   :0.000         Min.   : 1438
1st Qu.:3.000            1st Qu.:2.000          1st Qu.:2.000         1st Qu.: 2555
Median :4.000            Median :2.000          Median :2.000         Median : 4549
Mean   :3.813            Mean   :2.341          Mean   :2.455         Mean   : 8632
3rd Qu.:5.000            3rd Qu.:3.000          3rd Qu.:3.000         3rd Qu.:11068
Max.   :6.000            Max.   :6.000          Max.   :6.000         Max.   :34516
 Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt Total_Trans_Ct
Min.   :   0        Min.   :    3   Min.   :0.0000       Min.   :  510   Min.   : 10.00
1st Qu.: 359        1st Qu.: 1324   1st Qu.:0.6310       1st Qu.: 2156   1st Qu.: 45.00
Median :1276        Median : 3474   Median :0.7360       Median : 3899   Median : 67.00
Mean   :1163        Mean   : 7469   Mean   :0.7599       Mean   : 4404   Mean   : 64.86
3rd Qu.:1784        3rd Qu.: 9859   3rd Qu.:0.8590       3rd Qu.: 4741   3rd Qu.: 81.00
Max.   :2517        Max.   :34516   Max.   :3.3970       Max.   :18484   Max.   :139.00

Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
Min.   :0.0000      Min.   :0.0000
1st Qu.:0.5820      1st Qu.:0.0230
Median :0.7020      Median :0.1760
Mean   :0.7122      Mean   :0.2749
3rd Qu.:0.8180      3rd Qu.:0.5030
Max.   :3.7140      Max.   :0.9990
```

Revenir à la section 2.2.

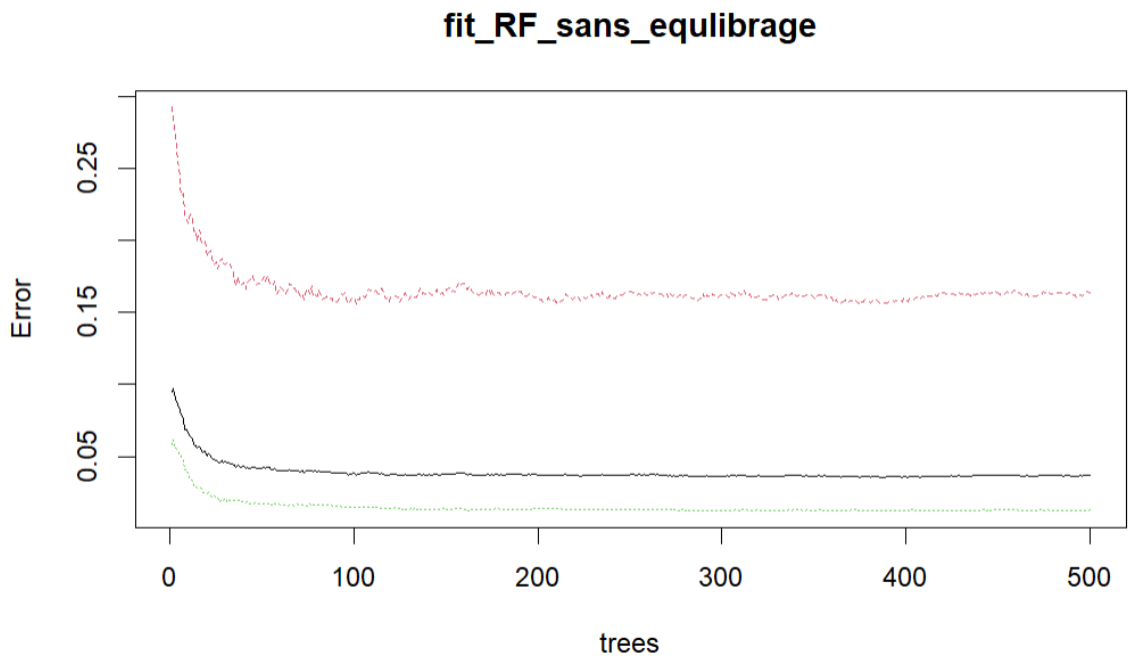

Figure 17: Courbe d'erreur du modèle Random Forest sans équilibrage

```r
confusion__sans_equlibrage = table(class_RF_sans_equlibrage, data.test$Attrition_Flag)
confusion__sans_equlibrage
```

class_RF_sans_equlibrage	Attrited Customer	Existing Customer
Attrited Customer	272	12
Existing Customer	61	1680

Figure 18: Table de confusion du modèle Random Forest sans équilibrage

Voir l'analyse dans la section 5.1.

```r
# OR
exp(logit.train$coefficients)
```

(Intercept)	Customer_Age	GenderM
0.000437792	1.000942263	1.440673664
Dependent_count	Education_LevelDoctorate	Education_LevelGraduate
0.839933114	0.496155359	0.764562055
Education_LevelHigh School	Education_LevelPost-Graduate	Education_LevelUneducated
0.818330667	0.443764041	0.659067773
Education_LevelUnknown	Marital_StatusMarried	Marital_StatusSingle
0.709753120	1.978120983	1.149022609
Marital_StatusUnknown	Income_Category$40K - $60K	Income_Category$60K - $80K
1.126891507	1.481013219	1.671695378
Income_Category$80K - $120K	Income_CategoryLess than $40K	Income_CategoryUnknown
1.092459127	1.300802460	1.155150093
Card_CategoryGold	Card_CategoryPlatinum	Card_CategorySilver
0.214694068	0.221765386	0.225498297
Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon
1.014316009	1.493119889	0.551350044
Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal
0.593067697	1.000029082	1.000889666
Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt
NA	2.108875987	0.999439178
Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
1.142549602	22.614872563	0.972059499

Figure 19: Valeurs des Odds Ratios (OR) estimés pour le modèle de régression logistique

Table 1: Résumé des coefficients du modèle de régression logistique

```
Call:
glm(formula = Attrition_Flag ~ Gender + Dependent_count + Education_Level +
    Marital_Status + Income_Category + Card_Category + Months_on_book +
    Total_Relationship_Count + Months_Inactive_12_mon + Contacts_Count_12_mon +
    Credit_Limit + Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 +
    Total_Trans_Amt + Total_Trans_Ct + Total_Ct_Chng_Q4_Q1, family = binomial,
    data = data.train.balanced)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -7.718e+00  3.877e-01 -19.906  < 2e-16 ***
GenderM                     3.654e-01  8.169e-02   4.473 7.72e-06 ***
Dependent_count            -1.745e-01  2.798e-02  -6.237 4.47e-10 ***
Education_LevelDoctorate   -7.013e-01  1.769e-01  -3.965 7.35e-05 ***
Education_LevelGraduate    -2.687e-01  1.227e-01  -2.190 0.028527 *
Education_LevelHigh School -2.010e-01  1.319e-01  -1.523 0.127767
Education_LevelPost-Graduate -8.128e-01  1.780e-01  -4.566 4.98e-06 ***
Education_LevelUneducated  -4.171e-01  1.364e-01  -3.057 0.002237 **
Education_LevelUnknown     -3.427e-01  1.375e-01  -2.493 0.012674 *
Marital_StatusMarried       6.820e-01  1.351e-01   5.047 4.50e-07 ***
Marital_StatusSingle        1.386e-01  1.360e-01   1.019 0.308073
Marital_StatusUnknown       1.199e-01  1.698e-01   0.706 0.479929
Income_Category$40K - $60K  3.916e-01  1.610e-01   2.432 0.015012 *
Income_Category$60K - $80K  5.129e-01  1.591e-01   3.223 0.001268 **
Income_Category$80K - $120K 8.734e-02  1.505e-01   0.580 0.561804
Income_CategoryLess than $40K 2.612e-01  1.577e-01   1.656 0.097818 .
Income_CategoryUnknown      1.439e-01  1.745e-01   0.824 0.409707
Card_CategoryGold          -1.540e+00  2.563e-01  -6.008 1.87e-09 ***
Card_CategoryPlatinum      -1.507e+00  4.244e-01  -3.550 0.000385 ***
Card_CategorySilver        -1.489e+00  1.332e-01 -11.180  < 2e-16 ***
Months_on_book              1.499e-02  4.471e-03   3.352 0.000801 ***
Total_Relationship_Count    4.009e-01  2.406e-02  16.664  < 2e-16 ***
Months_Inactive_12_mon     -5.956e-01  3.772e-02 -15.790  < 2e-16 ***
Contacts_Count_12_mon      -5.225e-01  3.370e-02 -15.505  < 2e-16 ***
Credit_Limit                2.940e-05  4.998e-06   5.882 4.05e-09 ***
Total_Revolving_Bal         8.832e-04  3.865e-05  22.852  < 2e-16 ***
Total_Amt_Chng_Q4_Q1        7.450e-01  1.734e-01   4.297 1.73e-05 ***
Total_Trans_Amt            -5.607e-04  2.224e-05 -25.206  < 2e-16 ***
Total_Trans_Ct              1.332e-01  3.646e-03  36.543  < 2e-16 ***
Total_Ct_Chng_Q4_Q1         3.120e+00  1.771e-01  17.612  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12371.6  on 9057  degrees of freedom
Residual deviance:  5824.9  on 9028  degrees of freedom
AIC: 5884.9

Number of Fisher Scoring iterations: 6
```

Revenir à la section 5.2.1.

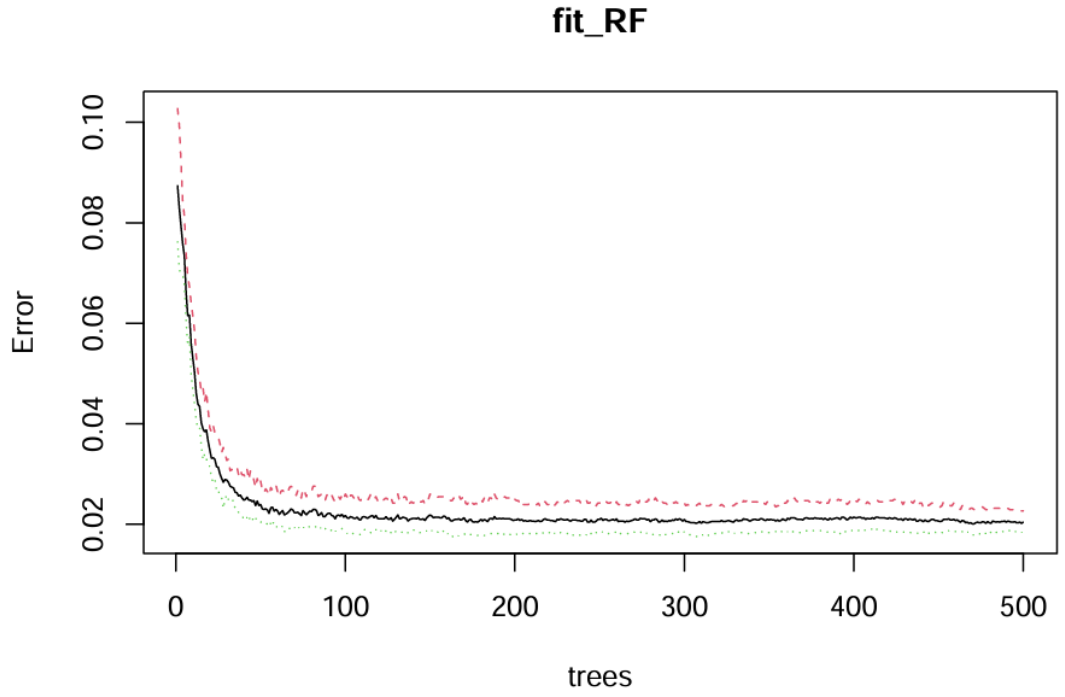

Figure 20: Courbe d'erreur du modèle Random Forest après équilibrage (SMOTE)

```
fit_RF <- randomForest(Attrition_Flag~.,data.train.balanced)
fit_RF
```

```
##
## Call:
##  randomForest(formula = Attrition_Flag ~ ., data = data.train.balanced)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 2.04%
## Confusion matrix:
##                    Attrited Customer Existing Customer class.error
## Attrited Customer              3794                88  0.02266873
## Existing Customer                97              5079  0.01874034
```

Figure 21: Sortie du modèle Random Forest sur données équilibrées (résumé du modèle)

Revenir à la section 5.2.4.

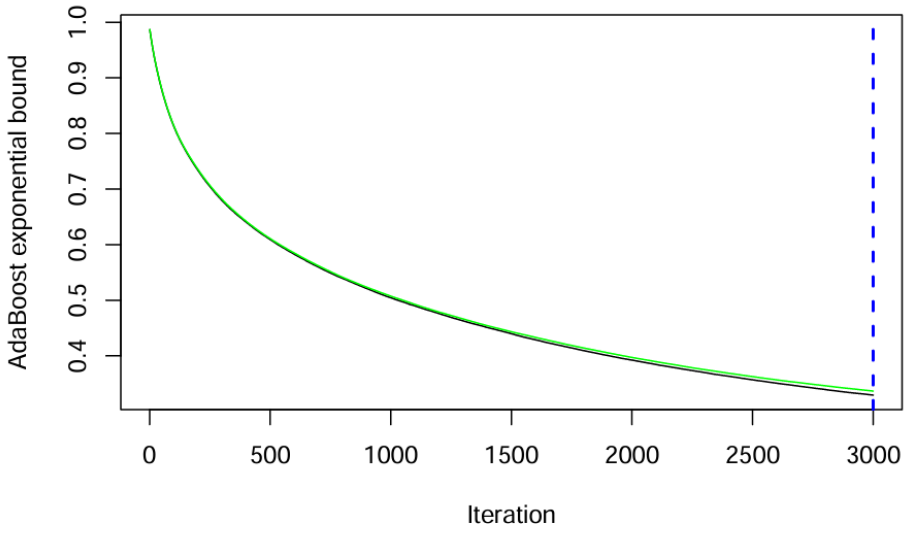

Figure 22: Évolution de la perte exponentielle selon le nombre d'itérations — Adaboost

Revenir à la section 5.2.5.

```
table (class_cart,data.test$Attrition_Flag)
```

```
##
## class_cart          Attrited Customer Existing Customer
##   Attrited Customer               286               134
##   Existing Customer                47              1558
```

Figure 23: Table de confusion – Modèle CART

```
table(class_RF, data.test$Attrition_Flag)
```

```
##
## class_RF            Attrited Customer Existing Customer
##   Attrited Customer               302                54
##   Existing Customer                31              1638
```

Figure 24: Table de confusion – Random Forest

```
table(class_adaboost, data.test$Attrition_Flag)
```

```
##
## class_adaboost      Attrited Customer Existing Customer
##   Attrited Customer               297                94
##   Existing Customer                36              1598
```

Figure 25: Table de confusion – Adaboost

```
table(class_logit_lasso,data.test$Attrition_Flag)
```

```
##
## class_logit_lasso   Attrited Customer Existing Customer
##   Attrited Customer               274               211
##   Existing Customer                59              1481
```

Figure 26: Table de confusion – Régression Lasso

Revenir à la section 5.3.

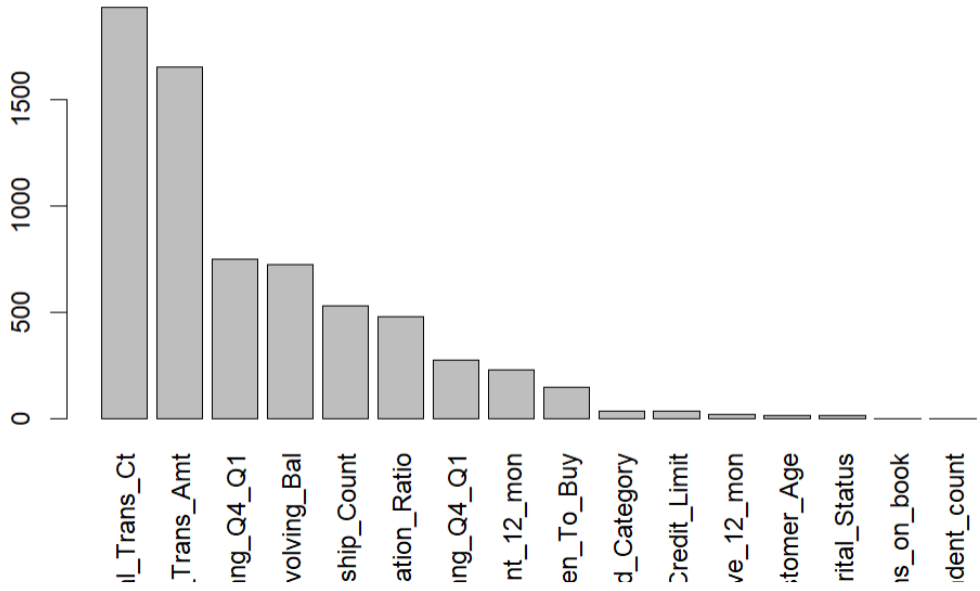

Figure 27: Importance des variables selon CART

Revenir à la section 5.4.

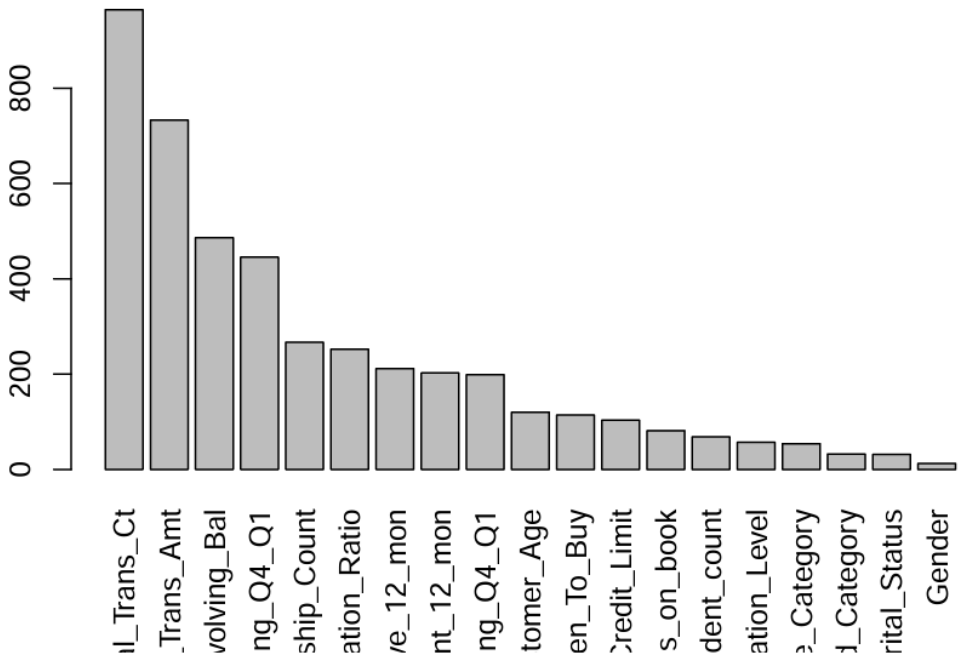

Figure 28: Importance des variables selon Random Forest

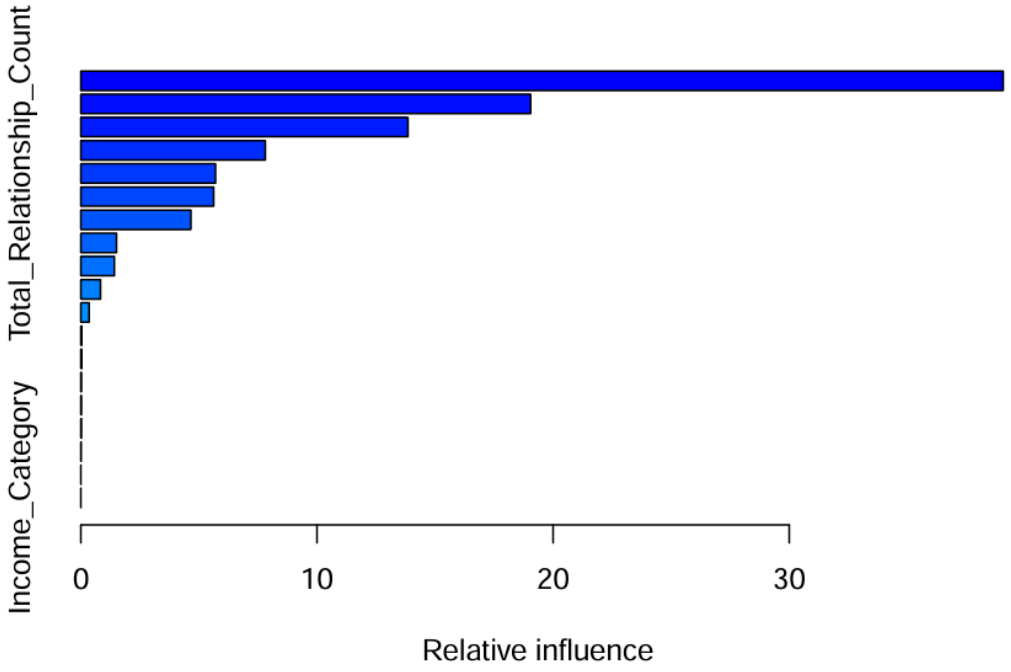

Figure 29: Importance des variables selon Adaboost

Revenir à la section 5.4.