PROJET DE SCORING

Jean-Philippe KIENNER

PRESENTATION DU PROJET

On se place dans le cadre d'un opérateur de téléphonie mobile.

Nous sommes début avril 2023 et depuis quelques mois de nombreux clients résilient leur contrat et partent à la concurrence, ce qui a logiquement un impact très négatif sur les résultats de l'entreprise.

Afin de réduire ce problème, l'opérateur de téléphonie mobile souhaite contacter les clients encore actifs et leur proposer une offre de fidélisation afin qu'ils ne résilient pas leur contrat dans le futur, par exemple un nouveau téléphone gratuit ou une réduction sur leur abonnement mensuel ; il s'agit d'une campagne marketing de rétention (souvent appelée « anti-churn »).

Cependant cela coûterait beaucoup trop cher de faire cette proposition à l'ensemble des clients : le budget alloué à la campagne marketing permet de contacter uniquement 2000 personnes.

L'idée est donc de solliciter ceux dont on pense qu'ils ont le plus de risque de résilier dans les 3 prochains mois.

Notre objectif est d'identifier ces 2000 clients à contacter en priorité.

Le principe du projet sera de construire plusieurs ciblages de 2000 clients au moyen de méthodes statistiques plus ou moins complexes afin d'améliorer les performances de la campagne marketing :

- Ciblage aléatoire
- Ciblage métier
- Ciblage profilé
- Ciblage scoré V1
- Ciblage scoré V2

La notion de performance d'un ciblage sera définie de la manière suivante :

- Je connais la liste des clients qui ont réellement résilié entre le 01/04/2023 et le 30/06/2023
- Je pourrai donc comparer chacun de vos ciblages avec cette liste
- Votre ciblage sera d'autant meilleur que vous aurez réussi à identifier le plus de futurs résiliés

Remarque d'organisation :

A ce stade il est souvent utile de mettre en place une structure de répertoires pertinente, par exemple :

- Un répertoire « Scoring »
- Des sous-répertoires qui stockeront vos différents fichiers selon leur type / finalité
    - Documents
    - Programmes
    - Tables
    - Sources
    - Sorties
    - Ciblages

CIBLAGE ALEATOIRE

1 Introduction

L'objectif du projet est d'identifier les clients ayant le plus de risque de résilier afin de leur proposer une offre réengageante.

Un ciblage simpliste peut être fait en tirant aléatoirement 2000 clients dans la base.

Il est évident que ce ciblage ne donnera pas de résultat satisfaisant, il ne rentrera d'ailleurs logiquement pas dans la notation, mais il va permettre :

- De se fixer une référence à dépasser par la suite en utilisant des techniques de ciblage de plus en plus perfectionnées
- Et de bien maîtriser le processus de construction et de livraison du fichier au bon format afin de pouvoir en tester la performance

2 Travail à réaliser

L'objectif est de tirer un échantillon aléatoire de 2000 clients au sein de la base « BASE_TELECOM_2023_03 ».

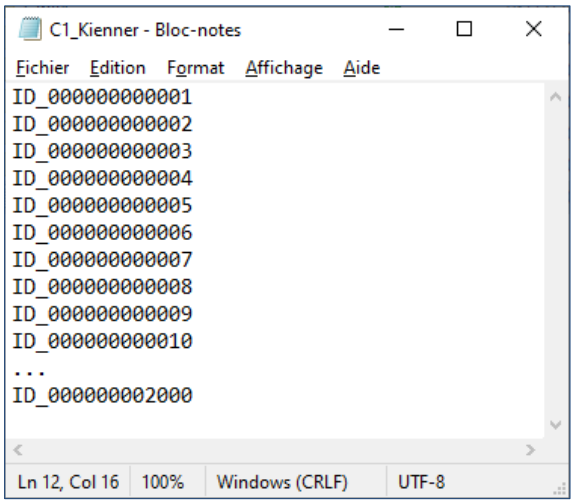
Afin d'avoir une liste de clients différente pour chaque participant, vous utiliserez une graine d'initialisation différente, par exemple votre date de naissance (exemple : 28042000).

3 Fichier de ciblage

Ce premier fichier de ciblage, à déposer sur Moodle, devra respecter le format suivant :

- Fichier texte nommé « C1_Nom1_Nom2.txt » au format Windows
- Une seule colonne sans en-tête (pas de nom de variable)
- 2000 lignes correspondant aux 2000 identifiants des clients sélectionnés (variable ID_CLIENT) sans numéro de ligne et sans cote ou guillemet

Exemple de fichier à envoyer (les numéros d'identifiants sont factices) :



Ce fichier sera directement intégré « tel quel » dans une moulinette qui permet de comparer votre liste à l'ensemble des résiliés réels et donc de connaître le nombre de résiliés que vous aurez réussi à identifier.

Il est donc de votre responsabilité de respecter les critères de format et de contenu du fichier de manière à ce qu'il ne soit pas rejeté.

4 Exemple de programme R

```
# Affectation du répertoire d'étude (chemin à adapter)

setwd ( "C:/.../Scoring")


# Import de la base

base_telecom_2023_03 <- read.table
      (
            file       = "Sources/base_telecom_2023_03.txt",
            encoding   = "UTF-8",
            sep        = ";",
            header     = TRUE,
            na.strings = ""
      )


# Sélection aléatoire de 2000 clients (graine à adapter)

set.seed ( 18111977 )

ciblage_aleatoire <-
 base_telecom_2023_03[sample  (  1:nrow(base_telecom_2023_03)  ,  2000  ),
"id_client"]


# Export pour le test de campagne (chemin, nom et options à adapter)

write.table
      (
            x             = ciblage_aleatoire,
            file          = "Ciblages/C1_Kienner.txt",
            file_encoding = "UTF-8",
            row.names     = FALSE,
            col.names     = FALSE,
            quote         = FALSE,
            na            = ""
      )
```

Remarques :

Le code ci-dessus est à modifier afin d'intégrer les chemins de votre propre répertoire, votre graine d'initialisation du tirage aléatoire, et votre nom du fichier final.
Il est agencé afin de le rendre lisible dans ce document, il sera à réécrire correctement dans RStudio.

De plus si vous n'êtes pas dans un environnement Windows (ex. : Unix, Mac), les caractéristiques de votre environnement doivent être prises en compte afin de gérer la fin des lignes et que soit indiqué le marqueur « CRLF » (retour chariot et saut de ligne sous Windows) et non « CR » (Mac) ou « LF » (Unix).
L'option « eol » de la fonction « write.table » permet de gérer cela :

- `eol = "\n"` (génère un saut de ligne, utile sous Mac par exemple)
- `eol = "\r"` (génère un retour chariot, utile sous Unix par exemple)

CIBLAGE METIER

1 Introduction

Après avoir construit un 1er ciblage basé sur un tirage aléatoire, l'objectif est d'améliorer la performance de la campagne marketing en utilisant des critères de ciblages pertinents.

Cela peut être fait de manière simple en combinant des indicateurs. Par exemple, on cible les clients :

- Ayant entre 18 et 25 ans
- Détenteurs d'un forfait 4H
- Appelant entre 3H et 5H par mois

Cette technique sera utilisée pour les 2 prochains ciblages : ciblage métier et ciblage profilé.
Nous verrons par la suite qu'un ciblage peut être nettement optimisé par l'utilisation de méthodes statistiques avancées telles que le scoring.

2 Travail à réaliser

Dans le cadre d'un ciblage métier, l'identification des indicateurs (ex : l'âge des clients) et des seuils (ex : entre 18 et 25 ans) va se faire sur la connaissance du produit, du secteur, et des comportements clients et marchés.

Ici la difficulté réside dans le fait que vous ne connaissez ni l'entreprise ni ses clients.
En revanche votre bon sens, votre propre expérience et des recherches sur internet vous aideront à définir des critères de risques de résiliation pertinents.

Vous appliquerez ces critères sur la base « BASE_TELECOM_2023_03 » afin de construire un ensemble de 2000 clients à contacter.

Remarques :

- Vous ne trouverez probablement pas du premier coup les critères, il vous faudra probablement plusieurs itérations afin de constituer la cible de 2000 clients
- Si vous n'arrivez pas pile aux 2000 clients avec vos critères, vous pouvez
    - Faire un tirage aléatoire de 2000 clients parmi votre cible
    - Trier votre table en fonction d'un ou plusieurs indicateurs que vous estimez important et prendre les 2000 premiers clients

3 Fichier de ciblage

Ce deuxième fichier de ciblage, à déposer sur Moodle, devra respecter le même format que pour le 1er fichier (seul le nom diffère) :

- Fichier texte nommé « C2_Nom1_Nom2.txt » au format Windows
- Une seule colonne sans en-tête (pas de nom de variable)
- 2000 lignes correspondant aux 2000 identifiants des clients sélectionnés (variable ID_CLIENT) sans numéro de ligne et sans cote ou guillemet

Je pourrai alors comparer votre liste à l'ensemble des résiliés réels et vous indiquerai le nombre de résiliés que vous aurez réussi à identifier, ce qui donnera lieu à une première note.

Il est donc de votre responsabilité de parfaitement respecter les consignes sur le nom et le format du fichier car je l'intégrerai tel quel : si l'intégration ne fonctionne pas, vous n'aurez aucun résilié identifié et donc la note de 0 !

CIBLAGE PROFILE

1 Introduction

Après avoir construit des ciblages basés sur un tirage aléatoire puis sur des règles métiers, un 3ème ciblage va être expérimenté en commençant à utiliser des analyses statistiques simples.

2 Travail à réaliser

Ici les critères de ciblage vont être définis en identifiant les caractéristiques des clients qui ont résilié dans le passé : la logique voudrait que si on retrouve des clients avec ces mêmes caractéristiques dans la population actuelle, il y a de fortes chances qu'ils présentent eux aussi un fort risque de résiliation.

On dispose pour faire cela d'une 2ème table, « BASE_TELECOM_2022_12 », présentant la même structure que « BASE_TELECOM_2023_03 », avec une variable en plus indiquant pour chaque client s'il a résilié ou non au 1er trimestre 2023 (variable « flag_resiliation »).

Vous devez donc analyser le profil des clients résiliés sur la base « BASE_TELECOM_2022_12 », ce qui va vous permettre d'identifier les indicateurs principaux liés à la résiliation des clients, et donc les critères de ciblages.

Vous appliquerez ces critères sur la base « BASE_TELECOM_2023_03 » afin de construire un ensemble de 2000 clients à contacter.

Remarques :

- L'analyse réalisée dans cette partie s'appuiera uniquement sur des méthodes statistiques descriptives univariées et bivariées. Aucune méthode mutivariée ne sera donc employée (arbre de décision, régression logistique, …).
- Vous ne trouverez probablement pas du premier coup les critères, il vous faudra probablement plusieurs itérations afin de constituer la cible de 2000 clients
- Si vous n'arrivez pas pile aux 2000 clients avec vos critères, vous pouvez
    - Faire un tirage aléatoire de 2000 clients parmi votre cible
    - Trier votre table en fonction d'un ou plusieurs indicateurs que vous estimez important et prendre les 2000 premiers clients

3 Fichier de ciblage

Ce troisième fichier de ciblage, à déposer sur Moodle, devra respecter le même format que pour les 1er et 2ème fichiers (seul le nom diffère) :

- Fichier texte nommé « C3_Nom1_Nom2.txt » au format Windows
- Une seule colonne sans en-tête (pas de nom de variable)
- 2000 lignes correspondant aux 2000 identifiants des clients sélectionnés (variable ID_CLIENT) sans numéro de ligne et sans cote ou guillemet

Je pourrai alors comparer votre liste à l'ensemble des résiliés réels et vous indiquerai le nombre de résiliés que vous aurez réussi à identifier, ce qui donnera lieu à une deuxième note.

Il est donc de votre responsabilité de parfaitement respecter les consignes sur le nom et le format du fichier car je l'intégrerai tel quel : si l'intégration ne fonctionne pas, vous n'aurez aucun résilié identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_TELECOM_2023_03 » et non ceux de « BASE_TELECOM_2022_12 » !

CIBLAGE SCORE V1

1 Introduction

Après avoir construit un 1er ciblage aléatoirement, puis un 2ème ciblage basé sur des règles métiers, suivi d'un 3ème ciblage utilisant des analyses statistiques simples, nous allons débuter un 4ème ciblage construit à partir d'un modèle de score.

Ce 1er score va être construit selon une méthodologie « traditionnelle », largement utilisée en entreprise et qu'il est donc nécessaire de maîtriser.
Une 2ème version du score permettra ensuite de tester d'autres éléments de méthodologie.

Les parties suivantes vont permettre de construire ce score progressivement en suivant les étapes définies dans le support de cours :

- Construction de la base d'étude
    - Identification de la population éligible
    - Définition de l'évènement à étudier
    - Détermination de la période d'étude
    - Construction des variables explicatives
    - Constitution des échantillons d'apprentissage et de validation(s)
    - Optimisation des variables explicatives
- Modélisation
    - Construction des modèles
    - Evaluation des modèles
    - Interprétation des modèles

Pour ce projet j'ai volontairement simplifié le processus en vous fournissant une base de données dans laquelle certaines étapes de la construction de la base d'étude ont déjà été faites :

- L'évènement à prédire est déjà matérialisé par la variable FLAG_RESILIATION
    - 0 si le client n'a pas résilié
    - 1 si le client a résilié
- La période d'étude a déjà été prise en compte et toutes les variables brutes ont été intégrées par rapport à une date de référence optimale

2 Travail à réaliser

2.1 Définition de l'évènement à étudier et de la population éligible

Y'a-t-il des exclusions de clients qui vous sembleraient pertinentes ?

- ⇨ Quantifier et préciser les traitements réalisés

Est-il nécessaire de stratifier le futur score ?

- ⇨ Discuter de la pertinence de cette mécanique sur ces données, cependant aucune stratification ne sera réalisée pour cette première version du score (ce pourra être un axe d'amélioration pour le score V2)

Est-il nécessaire de rééquilibrer les données ?

- ⇨ Discuter de la pertinence de cette mécanique sur ces données, et si vous le jugez nécessaire, réaliser les traitements

2.2 Nettoyage de la base de données

Y'a-t-il des valeurs manquantes, aberrantes ou extrêmes ?
Y'a-t-il des variables à supprimer ?
Y'a-t-il des incohérences entre variables ?

- ⇨ Quantifier et préciser les traitements réalisés

2.3 Construction des variables explicatives

Il est toujours pertinent de créer de nouvelles variables à partir des variables initiales, qui apporteraient une information supplémentaire ou bien une information plus synthétique.

- ⇨ Créer au moins 10 nouveaux indicateurs et expliquer leur intérêt et leur construction

Ce seuil de 10 nouvelles variables est purement indicatif, la base est suffisamment riche pour pouvoir créer plusieurs dizaines de nouveaux indicateurs.
Il n'est d'ailleurs pas gênant de conserver un grand nombre de variables ; et à ce stade il n'est pas utile de supprimer des variables, même si on soupçonne certaines d'être peu prédictives de la résiliation : c'est la phase de modélisation qui identifiera les variables pertinentes.

Attention, l'âge calculé à partir de la date de naissance, ou le volume d'appels exprimé en nombre d'heures au lieu d'un nombre de secondes, ne sont pas de nouveaux indicateurs (il n'y a pas d'information différente par rapport à la variable initiale).
De même un simple découpage en classes d'une variable quantitative ou un recodage en numérique d'une variable qualitative n'est pas un nouvel indicateur.

Il n'est d'ailleurs pas pertinent à ce stade de discrétiser les variables, par exemple avec des classes logiques / métier ou de même amplitude ou de même effectif : cet aspect sera traité ultérieurement grâce à l'optimisation des variables explicatives.

2.4 Échantillonnage

Afin de ne pas biaiser l'estimation des indicateurs de qualité des modèles, on les calcule à la fois sur l'échantillon qui a servi à construire le modèle, mais aussi sur un échantillon « indépendant ».

- ⇨ Séparer la base en échantillons d'apprentissage et de validation

2.5 Optimisation des variables explicatives

Il est fréquent de ne travailler qu'avec des variables qualitatives dont les modalités sont rendues les plus discriminantes par rapport à la variable à expliquer (incluant les variables quantitatives discrétisées et les variables qualitatives dont les modalités peuvent être regroupées).
Ce n'est pas obligatoire mais nous allons utiliser cette technique dans le cadre de ce 1er score.

- ⇨ Discrétiser l'ensemble des variables explicatives en optimisant le découpage en fonction de la variable à expliquer

Attention, une discrétisation en fonction des effectifs, par exemple en quartiles, n'est pas pertinente !

2.6 Modélisation

La construction du score se fait dans cette première version du score à partir d'un modèle de régression logistique (vous pourrez utiliser d'autres méthodes de modélisation dans le score V2).

- ⇨ Construire au moins 10 modèles différents
- ⇨ Comparer ces modèles au moyen d'indicateurs de qualité et choisir le meilleur modèle
- ⇨ Interpréter le modèle final, par exemple avec les odds-ratios ou les coefficients normalisés

Attention, les « N » itérations d'une sélection forward ne comptent pas pour « N » modèles !
L'idée est de construire des modèles différents (c'est-à-dire incluant des variables différentes) et de les comparer.

Pour rappel, réduire le nombre de variables n'est pas le principal objectif d'un score, le meilleur modèle n'est pas systématiquement celui qui contient le moins de variables …

2.7 Application du score

Une fois votre score construit, vous devez appliquer le modèle.

- ⇨ Reproduire à l'identique sur la table « BASE_TELECOM_2023_03 » les traitements réalisés à partir des décisions prises précédemment pour tous les éléments en « entrée » de votre modèle : population éligible, nettoyage des données, variables explicatives
- ⇨ Appliquer le modèle que vous avez choisi
- ⇨ Sélectionner les 2000 clients ayant les plus fortes probabilités de résiliation

3 Fichier de ciblage

Ce quatrième fichier de ciblage, à déposer sur Moodle, devra respecter le même format que pour les 1er, 2ème et 3ème fichiers (seul le nom diffère) :

- Fichier texte nommé « C4_Nom1_Nom2.txt » au format Windows
- Une seule colonne sans en-tête (pas de nom de variable)
- 2000 lignes correspondant aux 2000 identifiants des clients sélectionnés (variable ID_CLIENT) sans numéro de ligne et sans cote ou guillemet

Je pourrai alors comparer votre liste à l'ensemble des résiliés réels et vous indiquerai le nombre de résiliés que vous aurez réussi à identifier, ce qui donnera lieu à une troisième note.

Il est donc de votre responsabilité de parfaitement respecter les consignes sur le nom et le format du fichier car je l'intégrerai tel quel : si l'intégration ne fonctionne pas, vous n'aurez aucun résilié identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_TELECOM_2023_03 » et non ceux de « BASE_TELECOM_2022_12 » !

CIBLAGE SCORE V2

1 Introduction

L'objectif est ici d'améliorer le score précédent en construisant un deuxième modèle.

2 Travail à réaliser

2.1 Construction du score V2

Les étapes présentées précédemment constituent la trame classique d'un projet de scoring, néanmoins chacun peut y apporter des modifications en fonction de son expérience et de sa sensibilité.

- **Construction de la base d'étude :**
  - Modification de la population éligible
  - Stratification de la population d'étude : pas de stratification / stratification (et dans ce cas comment agréger des probabilités issues de modèles stratifiés)
  - Rééquilibrage de la variable à expliquer : pas de rééquilibrage / over-sampling / under-sampling
  - Nouvelles variables explicatives
  - Calibrage des variables explicatives : pas de discrétisation / discrétisation manuelle / discrétisation automatique / dichotomisation des variables qualitatives
  - Analyse factorielle des variables explicatives
  - Echantillonnage : apprentissage – validation classique / plusieurs échantillons de validation / validation croisée
- **Méthodes de modélisation :**
  - Tests de plusieurs méthodes de machine learning
  - Stacking de modèles (et dans ce cas comment agréger des probabilités issues de modèles différents)
- **Interprétation des modèles :**
  - Importance des variables
  - Compréhension des valeurs des variables entrainant une forte / faible probabilité : PDP, ICE, LIME, SHAP, …

Chaque élément pourra être étudié au travers de son impact :

- Sur le calcul de la probabilité
- Sur la performance du modèle

2.2 Application du score

Une fois votre score construit, vous devez appliquer le modèle.

- ⇨ Reproduire à l'identique sur la table « BASE_TELECOM_2023_03 » les traitements réalisés à partir des décisions prises précédemment pour tous les éléments en « entrée » de votre modèle : population éligible, nettoyage des données, variables explicatives
- ⇨ Appliquer le modèle que vous avez choisi
- ⇨ Sélectionner les 2000 clients ayant les plus fortes probabilités de résiliation

3 Fichier de ciblage

Ce cinquième fichier de ciblage, à déposer sur Moodle, devra respecter le même format que pour les 1er, 2ème, 3ème et 4ème fichiers (seul le nom diffère) :

- Fichier texte nommé « C5_Nom1_Nom2.txt » au format Windows
- Une seule colonne sans en-tête (pas de nom de variable)
- 2000 lignes correspondant aux 2000 identifiants des clients sélectionnés (variable ID_CLIENT) sans numéro de ligne et sans cote ou guillemet

Je pourrai alors comparer votre liste à l'ensemble des résiliés réels et vous indiquerai le nombre de résiliés que vous aurez réussi à identifier, ce qui donnera lieu à une troisième note.

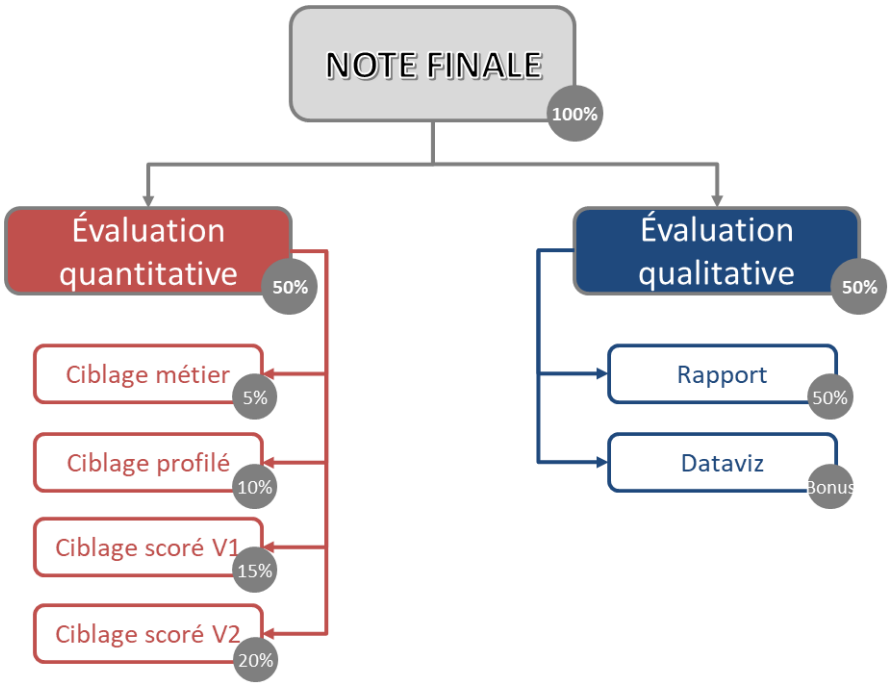
Il est donc de votre responsabilité de parfaitement respecter les consignes sur le nom et le format du fichier car je l'intégrerai tel quel : si l'intégration ne fonctionne pas, vous n'aurez aucun résilié identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_TELECOM_2023_03 » et non ceux de « BASE_TELECOM_2022_12 » !

NOTATION

La note finale comprendra

- Une évaluation quantitative, calculée à partir des performances constatées sur les 4 ciblages
- Une évaluation qualitative **optionnelle**, calculée à partir d'un rapport et d'une « dataviz »



1 Évaluation quantitative

L'évaluation de la performance des ciblages permet de déterminer si votre travail est performant.

Elle se fait en calculant le nombre de résiliés que vous aurez identifiés pour chaque ciblage.

	Seuil pour une note de 0	Seuil pour une note de 20	1 point tous les ... résiliés
Ciblage métier	300	1200	47
Ciblage profilé	300	1300	53
Ciblage scoré V1	300	1600	68
Ciblage scoré V2	300	1700	74

Le seuil de 300 résiliés identifiés, en deçà duquel vous aurez la note de « 0 », correspond à la pire performance d'un ciblage aléatoire (parmi 1000 échantillons tirés).
Les seuils hauts ont été définis selon les performances réellement constatées sur ces ciblages.

Exemples pour le ciblage métier :

- 258 résiliés identifiés sur les 2000 clients ciblés : note de 0 (pas de points négatifs)
- 937 résiliés identifiés sur les 2000 clients ciblés : note de 14
- 1315 résiliés identifiés sur les 2000 clients ciblés : note de 22 (points au-delà de 20 conservés)

2 Évaluation qualitative

Le rapport permet d'expliquer votre travail.

Le nombre de pages importe peu (le remplissage est proscrit), l'idée est d'expliquer et justifier tous les choix que vous avez fait :

- Synthétisez les ciblages métiers et profilés en justifiant les choix que vous avez faits, les résultats et apportez un regard critique
- Exposez votre réflexion sur la population éligible, la stratification et le rééquilibrage des données
- Expliquez votre démarche globale pour identifier les problèmes et nettoyer les données ; pour chaque type d'anomalie prenez une variable en exemple et détaillez précisément son traitement ; quantifiez les impacts sur la base de données
- Présentez précisément les nouvelles variables que vous créez : intérêt et mode de calcul
- Prenez un ou deux exemples de discrétisation pour illustrer votre méthodologie
- Détaillez très précisément votre stratégie de modélisation : quels modèles ont été construits et pourquoi, synthétisez les indicateurs de performance des modèles dans un tableau récapitulatif, justifiez le choix du modèle final
- Rendez l'interprétation du modèle compréhensible pour un non-statisticien

Il doit naturellement être correctement rédigé et illustré :

- Pas de code informatique
- Pas de sortie logicielle
- Bien rédigé, sans faute d'orthographe
- Tableaux et graphiques bienvenus
- Un tableau ou un graphique ne sont pertinents que s'ils sont commentés
- Explications précises et compréhensibles

Ce rapport sera nommé « Rapport_Nom1_Nom2.pdf » et déposé sur Moodle.

Barème indicatif :

- Ciblage métier 1 pt
- Ciblage profilé 2 pts
- Ciblage scoré
    - Construction de la base d'étude
        - Définition de l'évènement à étudier et de la population éligible 2 pts
        - Nettoyage de la base de données 2 pts
        - Construction des variables explicatives 3 pts
        - Échantillonnage 1 pt
        - Optimisation des variables explicatives 2 pts
    - Modélisation
        - Construction de plusieurs modèles 3 pts
        - Comparaison des modèles 2 pts
        - Interprétation du modèle choisi 2 pts

La dataviz ou « fiche de score » a pour objectif d'illustrer votre score en une page.

L'idée est de présenter votre score en 1 page (type slide Powerpoint) à un décideur, par exemple un directeur marketing, qui en parcourant votre fiche doit être convaincu qu'il va pouvoir lancer une campagne basée sur cet outil.

Ce n'est donc pas une simple transposition de la structure de votre rapport de projet ni un outil informatique type R-Shiny.
Ce n'est pas non plus un document technique présentant du code, des noms de packages ou de méthodes de machine learning.

Pensez « métier », réfléchissez à ce dont a besoin cet interlocuteur pour comprendre votre travail, le valoriser, décider de son utilisation et savoir comment l'utiliser.

Sans que ce soit exhaustif voici quelques questions auxquelles votre interlocuteur doit trouver une réponse dans votre dataviz :

- Qu'est-ce que cet outil ?
- A quoi cela va-t-il servir ?
- Est-ce que ça marche ?
- Qu'est-ce qu'il y a dedans ?
- Comment l'utiliser ?

Cette dataviz sera nommée « Dataviz_Nom1_Nom2.pdf » et déposé sur Moodle.

Le programme (non noté) permettra de comprendre en détail votre travail.

Vous regrouperez dans un seul fichier l'ensemble de votre code ayant permis de construire les 5 ciblages.
Bien que la qualité de programmation ne soit pas évaluée, il est important d'avoir un programme clair, structuré et commenté.
Le programme ne vous exonère pas d'être exhaustif et précis dans votre rapport.

Ce programme sera nommé « Programme_Nom1_Nom2.sas » et/ou « Programme_Nom1_Nom2.R » et déposé sur Moodle.

PRESENTATION DE LA BASE DE DONNEES

Les clients de l'opérateur de téléphonie mobile sont regroupés dans une base de données arrêtée au 31/03/2023 et décrivant les clients au travers de leurs caractéristiques sociaux-démographiques, les produits qu'ils détiennent, et l'usage qu'il font de leurs produits.

Cette base de données s'appelle « BASE_TELECOM_2023_03 » et les variables sont détaillées page suivante.

C'est à partir de cette base que devra être extraite la cible de 2000 clients.

Je vous fournirai à partir du ciblage profilé une deuxième base de structure équivalente avec deux éléments qui diffèrent :

- La base est arrêtée au 31/12/2022 et s'appellera « BASE_TELECOM_2022_12 »
- Elle contiendra une colonne en plus « FLAG_RESILIATION » indiquant si le client a résilié entre le 01/01/2023 et le 31/03/2023

Cette base vous permettra de construire les modèles statistiques.

Afin d'utiliser correctement ces bases, voici les principales règles :

- Lorsqu'une personne souscrit un forfait de téléphonie mobile, on dit qu'il s'active.
    - Il peut souscrire dans plusieurs enseignes différentes : boutique de l'opérateur, grande distribution, internet
    - Il souscrit un forfait avec une certaine durée : ½H, 1H, 2H, 3H, 4H, 6H, 8H, 10H
    - Il peut souscrire à des services additionnels
    - Il choisit un téléphone qui peut être bas de gamme, milieu de gamme ou haut de gamme ; s'il détient déjà un téléphone, il peut prendre une carte SIM seule
    - Il s'engage pour une période initiale de 1 an ou 2 ans, ce qui définit sa date de fin d'engagement

- Dans sa vie de client
    - Il peut modifier son forfait (« migration ») en augmentant ou diminuant sa durée
    - Il peut se réengager pour une période de 1 an ou 2 ans
        - Sa date de fin d'engagement est repoussée d'autant
        - S'il ne le fait pas, il est « libre d'engagement » (mais reste client)

- L'usage du client est caractérisé par des appels qu'il passe et des SMS qu'il envoie
    - On connait le volume de ces appels (en nombre de secondes) et le nombre de SMS sur les 6 derniers mois
    - On sait s'il a passé des appels surtaxés : international, numéros spéciaux

- Enfin le client peut décider de résilier son contrat
    - Il est censé être libre d'engagement lors de la résiliation mais peut faire le choix de payer des pénalités s'il est encore sous engagement au moment de sa résiliation

Remarque :
La date de fin d'engagement ne correspond pas à une date de résiliation !
La date de fin d'engagement d'un client peut donc être dépassée depuis plusieurs années mais il continue à être un client « normal » : il est simplement libre de résilier à tout moment, sans pénalité.

VARIABLE	LIBELLE
ID_CLIENT	Identifiant du client
FLAG_RESILIATION	***Le client a-t-il résilié (0 = non / 1 = oui)***
DATE_NAISSANCE	Date de naissance
SEXE	Sexe
CSP	Catégorie socio-professionnelle
CODE_POSTAL	Code postal
TAILLE_VILLE	Nombre d'habitants dans la ville
TYPE_VILLE	Catégorie de la ville
REVENU_MOYEN_VILLE	Revenu moyen des habitants de la ville
DATE_ACTIVATION	Date d'activation du contrat
ENSEIGNE	Enseigne de souscription du contrat
MODE_PAIEMENT	Mode de paiement
DUREE_OFFRE_INIT	Nombre d'heures du forfait initial
DUREE_OFFRE	Nombre d'heures du forfait actuel
NB_MIGRATIONS	Nombre de changements de forfait
FLAG_MIGRATION_HAUSSE	Le client a-t-il augmenté la durée de son forfait
FLAG_MIGRATION_BAISSE	Le client a-t-il diminué la durée de son forfait
NB_SERVICES	Nombre de services additionnels détenus
FLAG_PERSONNALISATION_REPONDEUR	Le client a-t-il personnalisé son répondeur
FLAG_TELECHARGEMENT_SONNERIE	Le client a-t-il téléchargé une sonnerie
TELEPHONE_INIT	Catégorie de téléphone initial
TELEPHONE	Catégorie de téléphone actuel
DATE_FIN_ENGAGEMENT	Date de fin d'engagement
NB_REENGAGEMENTS	Nombre de réengagements
DATE_DERNIER_REENGAGEMENT	Date du dernier réengagement
SITUATION_IMPAYES	Situation d'impayés
VOL_APPELS_M6	Volume d'appels il y a 6 mois (en secondes)
VOL_APPELS_M5	Volume d'appels il y a 5 mois (en secondes)
VOL_APPELS_M4	Volume d'appels il y a 4 mois (en secondes)
VOL_APPELS_M3	Volume d'appels il y a 3 mois (en secondes)
VOL_APPELS_M2	Volume d'appels il y a 2 mois (en secondes)
VOL_APPELS_M1	Volume d'appels il y a 1 mois (en secondes)
FLAG_APPELS_VERS_INTERNATIONAL	Le client a-t-il appelé vers l'international
FLAG_APPELS_DEPUIS_INTERNATIONAL	Le client a-t-il appelé depuis l'international
FLAG_APPELS_NUMEROS_SPECIAUX	Le client a-t-il appelé des numéros spéciaux
NB_SMS_M6	Nombre de SMS envoyés il y a 6 mois
NB_SMS_M5	Nombre de SMS envoyés il y a 5 mois
NB_SMS_M4	Nombre de SMS envoyés il y a 4 mois
NB_SMS_M3	Nombre de SMS envoyés il y a 3 mois
NB_SMS_M2	Nombre de SMS envoyés il y a 2 mois
NB_SMS_M1	Nombre de SMS envoyés il y a 1 mois
SEGMENT	Segment du client (A = meilleurs clients)

Toutes les variables « FLAG » sont binaires : 0 = non / 1 = oui.

Attention, la variable « FLAG_RESILIATION » n'est évidemment pas présente dans la base au sein de laquelle sélectionner les 2000 clients à contacter !

Quelques notions méritent d'être approfondies pour bien comprendre la base de données, en particulier les dates, voici donc un exemple détaillé de la vie d'un client avec les valeurs des variables correspondantes.

Etape 1

Le client souscrit une offre de téléphonie mobile le 15/03/2010.
Il choisit un forfait 2H, 3 services additionnels et un téléphone « milieu de gamme ».
Il décide de s'engager pour une période de 2 ans : cela signifie qu'en théorie il ne peut pas résilier son contrat avant la fin de cette période d'engagement, sauf à payer des pénalités

BASE AU 15/03/2010	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	2
NB_MIGRATIONS	0
FLAG_MIGRATION_HAUSSE	0
FLAG_MIGRATION_BAISSE	0
NB_SERVICES	3
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Milieu de gamme
DATE_FIN_ENGAGEMENT	15/03/2012
NB_REENGAGEMENT	0
DATE_DERNIER_REENGAGEMENT	

Etape 2

Au bout d'un an et demi, le 20/09/2011, le client souhaite changer de téléphone.
Il appelle le service client qui lui propose un nouveau téléphone avec une belle réduction.
En échange de cet effort commercial, le client accepte de se réengager pour une durée d'un an

BASE AU 20/09/2011	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	2
NB_MIGRATIONS	0
FLAG_MIGRATION_HAUSSE	0
FLAG_MIGRATION_BAISSE	0
NB_SERVICES	3
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Haut de gamme
DATE_FIN_ENGAGEMENT	20/09/2012
NB_REENGAGEMENT	1
DATE_DERNIER_REENGAGEMENT	20/09/2011

Etape 3

Plusieurs années passent, le client étant satisfait de son équipement actuel.
A noter qu'il n'est plus engagé depuis le 20/09/2012, il aurait donc pu résilier son contrat sans frais, mais il ne l'a pas fait, il continue à être client tout à fait normalement.

Etape 4

Le 24/02/2017, le client décide de changer complètement son offre de téléphonie mobile.
Il augmente sa durée de forfait et passe à 4H, il souscrit à 2 options en plus et prend un nouveau téléphone « Haut de gamme ».
Ces changements, accompagnés d'une forte réduction, entrainent un réengagement de 2 ans.

BASE AU 24/02/2017	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	4
NB_MIGRATIONS	1
FLAG_MIGRATION_HAUSSE	1
FLAG_MIGRATION_BAISSE	0
NB_SERVICES	5
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Haut de gamme
DATE_FIN_ENGAGEMENT	24/02/2019
NB_REENGAGEMENT	2
DATE_DERNIER_REENGAGEMENT	24/02/2017

Etape 5

Le 03/04/2017, le client souhaite augmenter sa durée de forfait à 6H.
Aucune autre modification n'est faite, ce changement n'entraine aucun réengagement.

BASE AU 03/04/2017	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	6
NB_MIGRATIONS	2
FLAG_MIGRATION_HAUSSE	1
FLAG_MIGRATION_BAISSE	0
NB_SERVICES	5
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Haut de gamme
DATE_FIN_ENGAGEMENT	24/02/2019
NB_REENGAGEMENT	2
DATE_DERNIER_REENGAGEMENT	24/02/2017

Etape 6

Le 16/11/2018, le client souhaite diminuer sa durée de forfait à 2H et résilier 4 options.
Il a toujours le même téléphone mais beaucoup d'autres modèles sont sortis entre temps, il est donc désormais catégorisé en « Bas de gamme ».
Aucune autre modification n'est faite, ce changement n'entraine aucun réengagement.

BASE AU 16/11/2018	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	2
NB_MIGRATIONS	3
FLAG_MIGRATION_HAUSSE	1
FLAG_MIGRATION_BAISSE	1
NB_SERVICES	1
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Bas de gamme
DATE_FIN_ENGAGEMENT	24/02/2019
NB_REENGAGEMENT	2
DATE_DERNIER_REENGAGEMENT	24/02/2017

Etape 7

Le 04/03/2023, le client souhaite changer d'opérateur de téléphonie mobile, il résilie donc son contrat.
Sa date de fin d'engagement étant passée, il est libre de résilier sans frais.
Ses données restent donc en l'état.

BASE AU 04/03/2023	
DATE_ACTIVATION	15/03/2010
DUREE_OFFRE_INIT	2
DUREE_OFFRE	2
NB_MIGRATIONS	3
FLAG_MIGRATION_HAUSSE	1
FLAG_MIGRATION_BAISSE	1
NB_SERVICES	1
TELEPHONE_INIT	Milieu de gamme
TELEPHONE	Bas de gamme
DATE_FIN_ENGAGEMENT	24/02/2019
NB_REENGAGEMENT	2
DATE_DERNIER_REENGAGEMENT	24/02/2017

Résumé de la vie du client au travers de la base de données extraite aux différentes dates

	15/03/2010	20/09/2011	24/02/2017
DATE_ACTIVATION	15/03/2010	15/03/2010	15/03/2010
DUREE_OFFRE_INIT	2	2	2
DUREE_OFFRE	2	2	4
NB_MIGRATIONS	0	0	1
FLAG_MIGRATION_HAUSSE	0	0	1
FLAG_MIGRATION_BAISSE	0	0	0
NB_SERVICES	3	3	5
TELEPHONE_INIT	Milieu de gamme	Milieu de gamme	Milieu de gamme
TELEPHONE	Milieu de gamme	Haut de gamme	Haut de gamme
DATE_FIN_ENGAGEMENT	15/03/2012	20/09/2012	24/02/2019
NB_REENGAGEMENT	0	1	2
DATE_DERNIER_REENGAGEMENT		20/09/2011	24/02/2017

	03/04/2017	16/11/2018	04/03/2023
DATE_ACTIVATION	15/03/2010	15/03/2010	15/03/2010
DUREE_OFFRE_INIT	2	2	2
DUREE_OFFRE	6	2	2
NB_MIGRATIONS	2	3	3
FLAG_MIGRATION_HAUSSE	1	1	1
FLAG_MIGRATION_BAISSE	0	1	1
NB_SERVICES	5	1	1
TELEPHONE_INIT	Milieu de gamme	Milieu de gamme	Milieu de gamme
TELEPHONE	Haut de gamme	Bas de gamme	Bas de gamme
DATE_FIN_ENGAGEMENT	24/02/2019	24/02/2019	24/02/2019
NB_REENGAGEMENT	2	2	2
DATE_DERNIER_REENGAGEMENT	24/02/2017	24/02/2017	24/02/2017

Ce client fait partie de la base clients au 31/12/2022, et il aura un FLAG_RESILIATION = 1 car il a résilié au cours du 1er trimestre 2023.
Il ne fait en revanche pas partie de la base clients au 31/03/2023 puisqu'il a résilié entre-temps.
Un bon modèle de score aurait donc dû être capable de détecter ce client avant qu'il ne résilie.

Parmi les clients encore présents au 31/03/2023, il existe des personnes qui risquent de résilier dans les mois suivants.
L'objectif du projet est de les identifier avant qu'ils ne résilient.