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 ».

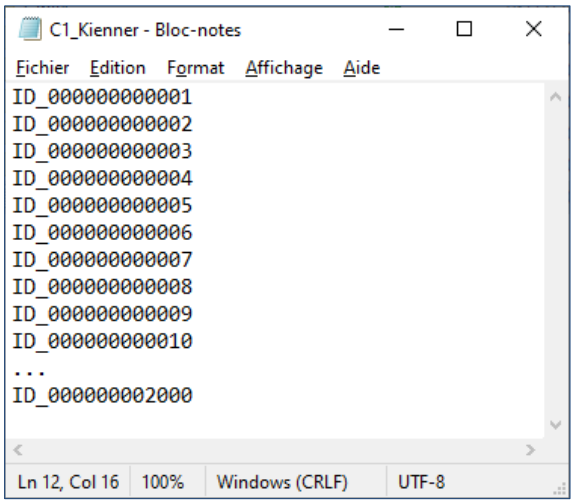
Afin d'avoir une liste de clients différente pour chaque participant, vous utiliserez une graine d'initialisation différente, par exemple votre date de naissance (exemple : 28042000).

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 » !