PROJET DE SCORING

Données Banque

Jean-Philippe KIENNER

2024 – 2025

PRÉSENTATION DU PROJET

1 Contexte métier

On se place dans le cadre d'un groupe bancaire proposant des crédits à la consommation.

Afin de maximiser le taux de souscription des crédits pour le prochain trimestre, les équipes marketing souhaitent contacter, parmi les clients ayant déjà souscrit un crédit, ceux ayant le plus d'appétence à en souscrire un de nouveau, et leur proposer une offre afin qu'ils concrétisent leur souscription.

Cependant cela coûterait beaucoup trop cher de faire cette campagne d'appels 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 chance de souscrire un crédit 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 :

- 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 :

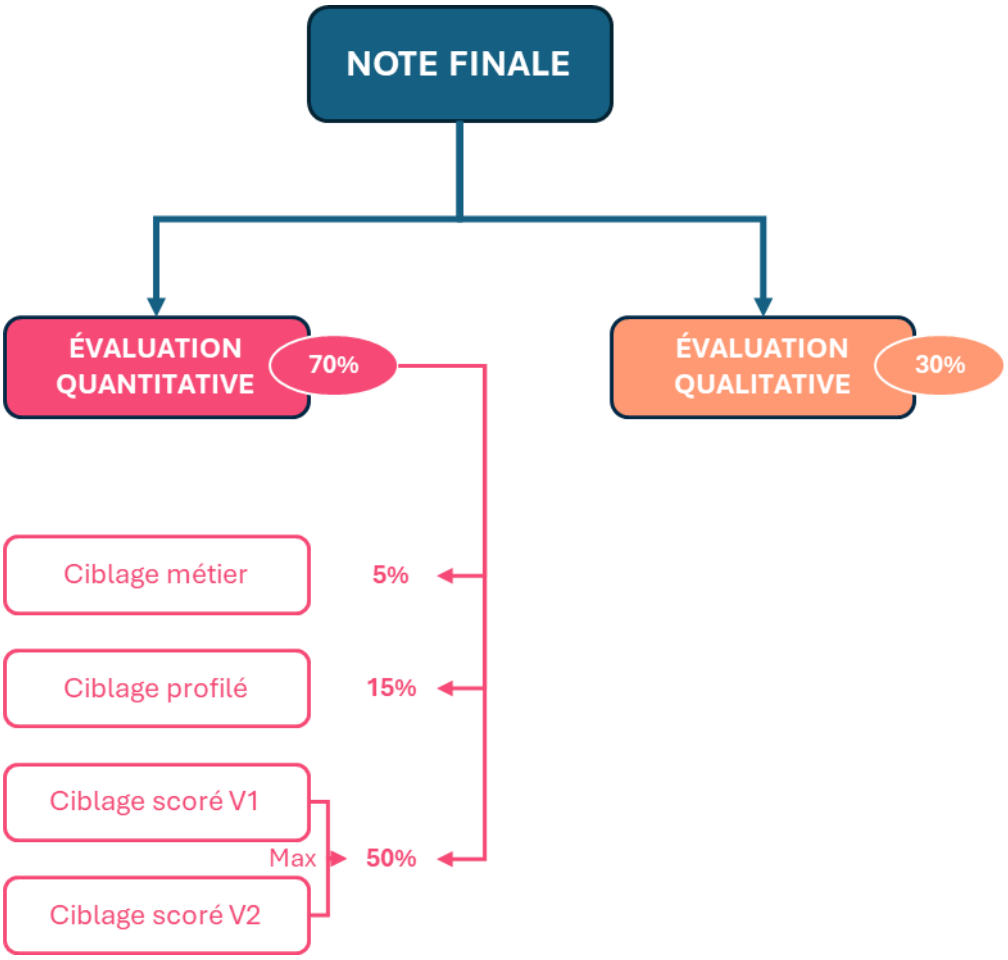
- Je connais la liste des clients qui ont réellement souscrit un crédit
- Je pourrai donc comparer chacun de vos ciblages avec cette liste
- Votre ciblage sera d'autant meilleur que vous aurez réussi à identifier le plus de futurs souscripteurs

2 Notation

2.1 Principe général

La note finale comprendra :

- Une évaluation quantitative, calculée à partir des performances des ciblages
- Une évaluation qualitative, calculée à partir d'une fiche de score



2.2 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 souscripteurs que vous aurez identifiés pour chaque ciblage.

À noter que la performance maximum sera prise en compte pour les 2 ciblages scorés.

2.3 Évaluation qualitative

L'idée est de présenter votre score en une 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 d'un rapport de projet académique 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 ?
- À 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 « Fiche_Nom1_Nom2.pdf » et déposé sur Moodle.

3 Présentation de la base de données

3.1 Principe général

Les clients de la banque sont regroupés dans une base de données arrêtée au 31/03/2024 et décrivant les clients au travers de leurs caractéristiques sociaux-démographiques et des crédits déjà souscrits.

Cette base de données s'appelle « BASE_BANQUE_2024_03 » et les variables sont détaillées ci-dessous.

C'est à partir de cette base que devra être extraite la cible de 2000 clients qu'on souhaite contacter afin qu'ils souscrivent un nouveau crédit au 2ème trimestre 2024.

Je vous fournirai à partir du ciblage profilé une deuxième base « BASE_BANQUE_2023_12 » de structure équivalente mais arrêtée au 31/12/2023, et qui contiendra une colonne en plus « FLAG_CREDIT » indiquant si le client a souscrit un crédit entre le 01/01/2024 et le 31/03/2024.

Cette base vous permettra de construire les modèles statistiques.

3.2 Description des variables

Base_banque_2024_03 :

VARIABLE	LIBELLÉ
id_client	Identifiant du client
age	Age du client (en années)
sexe	Sexe du client
situation_familiale	Situation familiale
statut_logement	Statut de la résidence principale
revenu	Revenu déclaré par le client
flag_tel	Le téléphone est-il connu
flag_email	L'e-mail est-il connu
nb_credits_total	Nombre total de crédits
mt_credits_total	Montant total des crédits
nb_credits_actuel	Nombre de crédits en cours
mt_credits_actuel	Montant des crédits en cours
mt_echeances_actuel	Montant de l'échéance mensuelle des crédits en cours
duree_remboursement_actuel	Durée des crédits en cours (en années)
mt_premier_credit	Montant du premier crédit
anc_premier_credit	Ancienneté du premier crédit (en jours)
canal_premier_credit	Canal de souscription du premier crédit
mt_dernier_credit	Montant du dernier crédit
anc_dernier_credit	Ancienneté du dernier crédit (en jours)
canal_dernier_credit	Canal de souscription du dernier crédit

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

Base_banque_2023_12 (disponible à partir du ciblage profilé) :

VARIABLE	LIBELLÉ
id_client	Identifiant du client
flag_credit	Le client a-t-il souscrit un crédit ?
age	Age du client (en années)
sexe	Sexe du client
situation_familiale	Situation familiale
statut_logement	Statut de la résidence principale
revenu	Revenu déclaré par le client
flag_tel	Le téléphone est-il connu
flag_email	L'e-mail est-il connu
nb_credits_total	Nombre total de crédits
mt_credits_total	Montant total des crédits
nb_credits_actuel	Nombre de crédits en cours
mt_credits_actuel	Montant des crédits en cours
mt_echeances_actuel	Montant de l'échéance mensuelle des crédits en cours
duree_remboursement_actuel	Durée des crédits en cours (en années)
mt_premier_credit	Montant du premier crédit
anc_premier_credit	Ancienneté du premier crédit (en jours)
canal_premier_credit	Canal de souscription du premier crédit
mt_dernier_credit	Montant du dernier crédit
anc_dernier_credit	Ancienneté du dernier crédit (en jours)
canal_dernier_credit	Canal de souscription du dernier crédit

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

La variable « FLAG_CREDIT » n'est évidemment présente que dans cette base.

CIBLAGE ALÉATOIRE

1 Introduction

L'objectif du projet est d'identifier les clients ayant le plus d'appétence de souscrire à nouveau un crédit, et leur proposer une offre afin qu'ils concrétisent cette souscription.

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_BANQUE_2024_03 ».

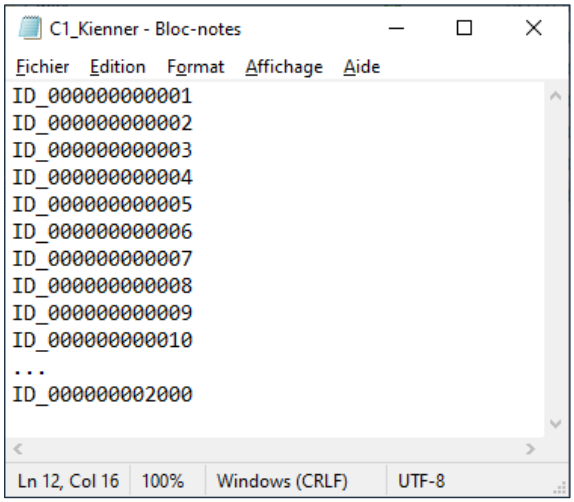
Afin d'avoir une liste de clients différente pour chaque participant, vous utiliserez une graine d'initialisation différente, par exemple votre date de naissance (exemple : 28042000).

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 souscripteurs réels et donc de connaître le nombre de souscripteurs 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 qu'il ne soit pas rejeté.

CIBLAGE MÉTIER

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 :

- Les propriétaires
- Ayant un total de crédit supérieur à 10000 €

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 : les propriétaires) et des seuils (ex : plus de 10000 €) 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 pertinents qui qualifient les clients les plus enclins à souscrire un crédit.

Vous appliquerez ces critères sur la base « BASE_BANQUE_2024_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 souscripteurs réels et vous indiquerai le nombre de souscripteurs 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 souscripteur identifié et donc la note de 0 !

CIBLAGE PROFILÉ

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 souscrit un crédit 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 une forte chance de souscrire un nouveau crédit.

On dispose pour faire cela d'une 2ème table, « BASE_BANQUE_2023_12 », présentant la même structure que « BASE_BANQUE_2024_03 », avec une variable en plus indiquant pour chaque client s'il a souscrit un crédit ou non (variable « flag_credit »).

Vous devez donc analyser le profil des clients souscripteurs sur la base « BASE_BANQUE_2023_12 », ce qui va vous permettre d'identifier les indicateurs principaux liés à la souscription des clients, et donc les critères de ciblages.

Vous appliquerez ces critères sur la base « BASE_BANQUE_2024_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 souscripteurs réels et vous indiquerai le nombre de souscripteurs 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 souscripteur identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_BANQUE_2024_03 » et non ceux de « BASE_BANQUE_2023_12 » !

CIBLAGE SCORÉ 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 la variable à expliquer
  - Détermination de la période d'étude
  - Nettoyage de la base de données
  - Construction des variables explicatives
  - Constitution des échantillons d'apprentissage, de validation(s) et de test(s)

- Modélisation
  - Construction des modèles
  - Évaluation des modèles
  - Interprétation des modèles

2 Travail à réaliser

2.1 Identification de la population éligible

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

- ⇨ Quantifier et préciser les traitements réalisés

2.2 Définition de la variable à expliquer

Pour ce projet j'ai volontairement simplifié le processus en vous fournissant une base de données dans laquelle l'évènement à prédire est déjà matérialisé par la variable FLAG_CREDIT :

- ⇨ 0 si le client n'a pas souscrit un crédit
- ⇨ 1 si le client a effectivement souscrit un crédit

Il n'y a donc rien à faire ici (contrairement à un projet de scoring réel).

2.3 Détermination de la période d'étude

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.

Il n'y a donc rien à faire ici (contrairement à un projet de scoring réel).

2.4 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

2.5 Création de nouvelles variables

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 de nouveaux indicateurs (au moins une dizaine) 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 souscription : c'est la phase de modélisation qui identifiera les variables pertinentes.

Attention, une durée calculée à partir d'une date n'est pas un nouvel indicateur (il n'y a pas d'information différente par rapport à la variable initiale).
De même un simple recodage en numérique d'une variable qualitative n'est pas un nouvel indicateur.

2.6 Séparation en plusieurs échantillons

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 test

2.7 Construction du modèle

La construction du score se fait au moyen d'un modèle statistique laissé au libre choix.

- ⇨ Construire plusieurs modèles différents (au moins une dizaine)
- ⇨ Comparer ces modèles au moyen d'indicateurs de qualité et choisir le meilleur modèle
- ⇨ Interpréter le modèle final

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 …

2.8 Application du score

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

- ⇨ Reproduire à l'identique sur la table « BASE_BANQUE_2024_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 souscription

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 souscripteurs réels et vous indiquerai le nombre de souscripteurs 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 souscripteur identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_BANQUE_2024_03 » et non ceux de « BASE_BANQUE_2023_12 » !

CIBLAGE SCORÉ V2

1 Introduction

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

2 Travail à réaliser

2.1 Amélioration du score

Les étapes présentées précédemment constituent la trame classique d'un projet de scoring, néanmoins plusieurs pistes d'amélioration peuvent être testées :

- ⇨ Modification de la population éligible
- ⇨ 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
- ⇨ Stratification de la population d'étude : pas de stratification / stratification (et dans ce cas comment agréger des probabilités issues de modèles stratifiés)
- ⇨ Échantillonnage : apprentissage – validation classique / plusieurs échantillons de validation / validation croisée
- ⇨ Tests de plusieurs méthodes de machine learning

2.2 Application du score

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

- ⇨ Reproduire à l'identique sur la table « BASE_BANQUE_2024_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 souscription

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 souscripteurs réels et vous indiquerai le nombre de souscripteurs 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 souscripteur identifié et donc la note de 0 !

En particulier, pensez à mettre les ID_CLIENT de la table « BASE_BANQUE_2024_03 » et non ceux de « BASE_BANQUE_2023_12 » !