Réalisé par :
Anis LOUISSI
Anisoara ABABII
Baptiste GAUTIER
Oussama LAAUMARI

Décembre 2022, Paris

Projet Modélisation du risque de crédit

Master 2 MoSEF, Université Paris 1 Panthéon Sorbonne

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

LCL
BANQUE ET ASSURANCE

Table des matières

Introduction 2

Source des données et cible à prédire 3

Analyse exploratoire des données 3

1.1.Valeurs manquantes et doublons 3

1.2. Distributions des variables 4

1.3. Analyse de la variable Age 5

1.4. Analyse de la variable contrat de travail 6

1.5. Analyse de la variable Métier 7

Découpage du jeu de données 7

Sélection des variables par tests statistiques 8

2.1. Le cas des variables continues 8

2.2. Le cas des variables discrètes 9

Sélection des variables explicatives et modélisation par régression logistique 10

3.1. Regroupement, Discrétisation 10

Modélisation 12

4.1. Hypothèse de régression logistique 12

Création d'une grille de score 13

Création de classes de risque 14

6.1. Clustering par K-means 14

6.2. Stabilité en risque des modalités du modèle : 15

6.3. Stabilité en effectif des modalités du modèle : 15

Analyse des performances 16

Tester des algorithmes de machine Learning et comparer les performances avec la régression logistique 17

Conclusion 19

Résumé

L'objectif de ce projet est de modéliser la Probabilité de tomber en défaut à 36 mois sur des dossiers immobiliers. Il s'agit également de construire une grille de score permettant d'évaluer pour un client donné un score de risque calculé selon certaines modalités. Ce score permettant en dernier lieu d'aider les conseillers sur leur décision d'acceptation d'un dossier de crédit immobilier.

Les spécifications du projet sont les suivantes (fournies par LCL) :

- Prendre en compte un profil emprunteur actualisé avec l'utilisation d'un historique de données plus récent ;
- Explorer et étudier l'apport de nouvelles variables discriminantes ;
- Tester l'apport de données externes (données socio-démographiques par exemple) ;
- Prendre en compte la Nouvelle Définition du Défaut Bâlois (NDB) comme variable cible (implémentée en août 2020 chez LCL)

Introduction

La société financière "LCL Banque et assurance" en collaboration avec "Crédit Agricole S.A. ", propose des crédits aux clients en ayant peu ou pas du tout d'historique de prêt. Afin de répondre aux demandes de transparence de ses clients vis-à-vis des décisions d'octroi de crédit, l'entreprise souhaite développer un modèle de scoring de la probabilité de défaut de paiement du client, réalisé à partir de sources de données variées.

Ce document présente la démarche mise en œuvre pour développer une approche quantitative, utile pour les banques de détail : le scoring. Ce modèle sera pour la suit un outil à l'aide de décision pour l'entreprise. De plus, ce modèle de scoring est calculé à partir des techniques statistiques sur des données historiques qui datent d'avril 2014 – avril 2018 (avec un taux de défaut observé à 36 mois).

Source des données et cible à prédire

Le jeu de données proposé pour l'élaboration du modèle est fourni par "LCL Banque et assurance" et n'est pas disponible en open source. Les données « Bases Récentes » se décomposent en 114416 observations et 124 variables. Celle-ci proposent une variété d'informations anonymisées sur la situation personnelle et les activités bancaires des clients.

La valeur à prédire, le risque de défaut du client, se situe dans la feature ***« defaut_36mois »***. Il prend la valeur 0 lorsque le client a remboursé son crédit, 1 s'il ne l'a pas fait. Il s'agira donc des 2 classes de notre problème.

Analyse exploratoire des données

Afin de se familiariser avec le jeu de données et de faire ressortir des caractéristiques pertinentes pour la modélisation, le jeu de données a été soumis à un processus d'analyse exploratoire en 3 parties :

- Étude des valeurs manquantes
- Étude de la distribution des valeurs
- Étude des corrélations

1.1. Valeurs manquantes et doublons

Dans la base de données il n'y a pas de doublons. Concernant les valeurs non-renseignées, environ 15% des colonnes présentent dans leur effectif plus de 50% de valeurs manquantes. Cela sera une information à considérer lors de l'étape de modélisation.

En effet, une étape d'imputation des valeurs manquantes est nécessaire pour rendre le jeu de données exploitable par certains algorithmes. Par précision les valeurs manquantes seront imputées par la mention « NR » (non-renseigné). Les colonnes qui d'un point de vue instinctif semblent les plus intéressantes semblent toutefois avoir un meilleur taux de remplissage *(Figure 1 : Valeurs manquantes dans la base Bases_Recentes.csv* (la figure est plus lisible dans 1_Dara_Description.ipynb)).

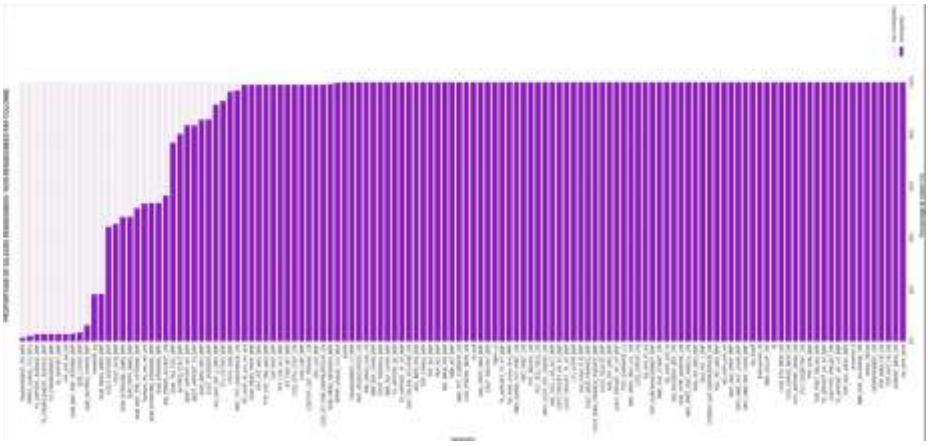


Figure 1 : Valeurs manquantes dans la base Bases_Recentes.csv

Les variables avec plus de 80% de valeurs manquantes ont été enlevées de la base de données uniquement pour l'étape de visualisation des données. Cela car elles ne contiennent pas assez de données renseignées pour apporter une information importante d'un point de vue analytique et visuel. En revanche, pour l'étape de modélisation nous avons procédé à une imputation des variables manquantes par la mention «NR » et nous les avons classés d'après la discrétisation de la librairie OptBinning dans une modalité différente des autres.

1.2. Distributions des variables

En étudiant le risque de défaut dans notre jeu de données, une large majorité des clients n'ont pas fait défaut (99,29%, *Figure 2 : Distribution des remboursements de crédit*).

Figure 2 : Distribution des remboursements de crédit

Il s'agit donc d'un problème présentant un fort déséquilibre de classes. Lors de l'étape de modélisation, il faudra donc réfléchir à la méthodologie à employer pour obtenir un modèle performant.

Parmi les variables continues significatives par rapport à la variable cible, nous remarquons des valeurs aberrantes. Le traitement sera mis en place dans la suite. *Figure 3 : Distribution des variables continues significatives au taux de défaut.*

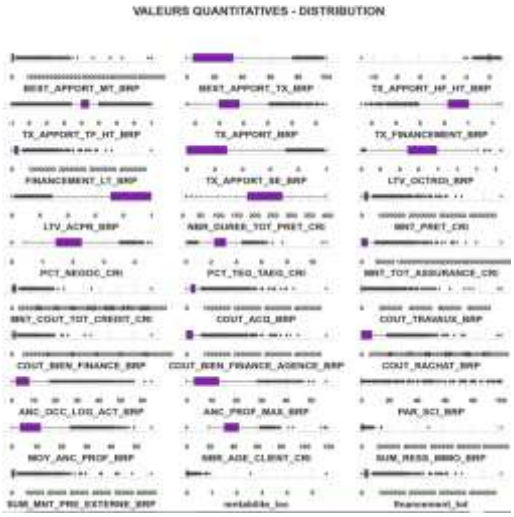


Figure 3 : Distribution des variables continues significatives au taux de défaut

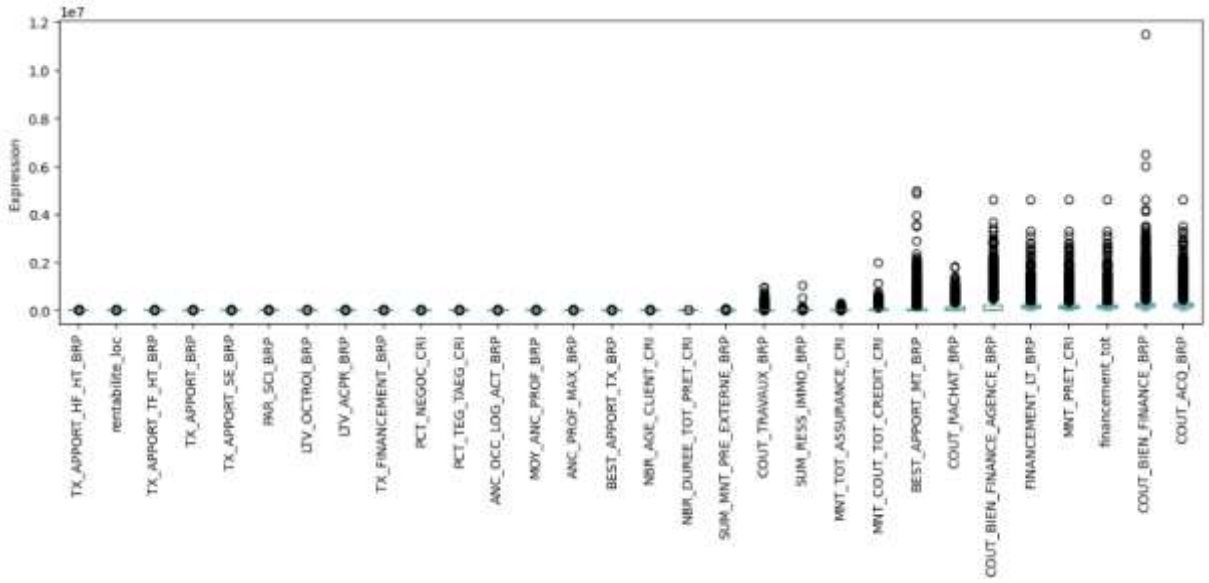


Figure 4 : Les niveaux d'expression des variables continues

Pour certaines variables continues les niveaux d'expression varient beaucoup, pour d'autres non. Les plages de valeurs sont différentes d'une variable à l'autre. Cette situation peut impacter le modèle que nous allons mettre en place. Lors de l'étape de discrétisation, les valeurs aberrantes seront classées dans une modalité particulière pour ne pas impacter le résultat final.

Le jeu de données présente nombre de variables corrélées entre elles : ex: Tx_Financement_BRP, LTV_OCTROI_BRP, LTV_ACPR_BRP (les variables corrélées seront soit regroupées soit supprimées, dépendamment du cas). 8 variables dichotomiques sont significatives au sens de Chi 2 parmi 14.

Concernant des anomalies, nous observons des taux négatifs, qui seront remplacés par des NR, car il ne peut pas à priori pas y avoir de taux négatif.

Ici, nous allons présenter des variables qui nous semblent pertinentes d'analyser avant de réaliser nos travaux sur la modélisation afin de comprendre de quelles informations nous disposons au sein et comment notre population est-elle répartie.

1.3. Analyse de la variable Age

Nous pensons que l'âge est un premier facteur d'accord d'un prêt pour un client donné. En effet, il est important d'avoir une idée sur la répartition de nos clients afin de pouvoir mieux les étudier et comprendre comment les segmenter par la suite. Nous pouvons remarquer sur la figure ci-dessous que l'on retrouve une population âgée de 20 à 70 ans (après nettoyage) ce qui nous semble déjà cohérent pour étudier le fait de faire défaut ou non car il s'agit d'une population susceptible de contracter un prêt bancaire à des fins immobilières.

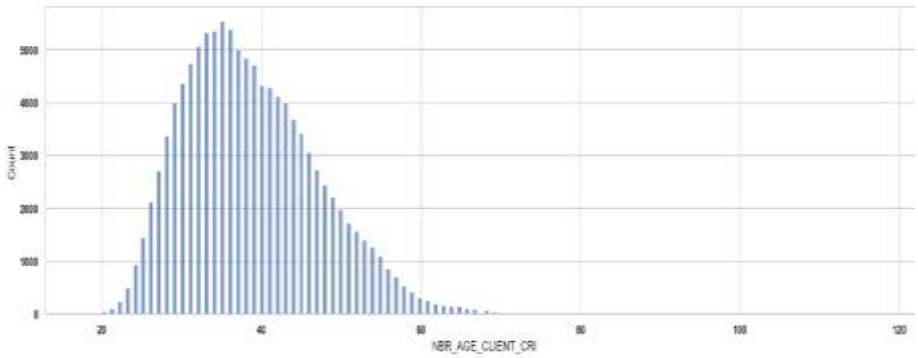


Figure 5 : Distribution de la variable Age

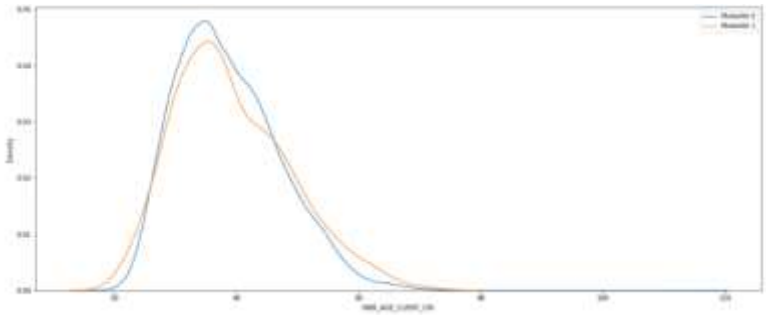


Figure 6 : Comparaison des densités de faire défaut ou non en fonction de l'âge

Ce 2ème graphique nous présente le nombre de personnes faisant défaut par âge. On remarque que logiquement ce sont les personnes les plus jeunes et les plus vieilles de notre échantillon qui ont plus de chances de faire défaut. Alors qu'à l'inverse une personne âgée de la quarantaine a plus de chances de rembourser son prêt par rapport aux extremums. De surcroît, que notre population est beaucoup plus dense en termes de volume autour de la quarantaine car ce type de personnes sont plus susceptibles de contracter un prêt immobilier.

1.4. Analyse de la variable contrat de travail

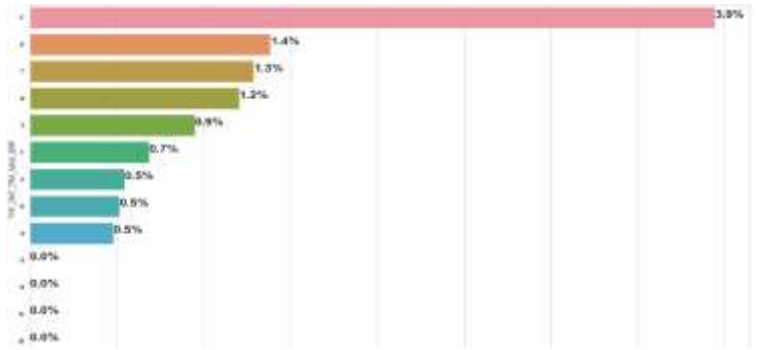


Figure 7 : Visualisation du taux de défaut en fonction du contrat de travail

Pour nous, un facteur important de l'accord de la banque suite à une demande de prêt est la situation professionnelle de la personne donc cela correspond à son type de contrat de travail ainsi qu'à son type de métier.

Cela nous confirme que lorsque nous nous trouvons en situation qui n'est pas stable nous avons plus de chances de tomber en défaut, par exemple on voit qu'il y a peu de personnes en CDI qui ne remboursent pas leur prêt contracté tandis que les étudiants ou les individus classés dans autres (artisans...) ont plus fortement tendance à ne pas rembourser leur prêt.

1.5. Analyse de la variable Métier

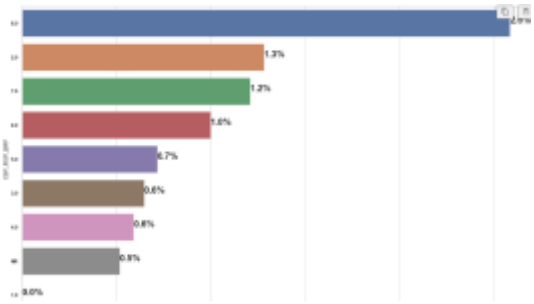


Figure 8 : Visualisation du taux de défaut en fonction du métier

Après avoir présenté le taux de défaut en fonction du contrat de travail il nous faut désormais présenter le taux de défaut en fonction du métier de l'individu. En effet, un cadre par exemple aura plus de chances de pouvoir rembourser un prêt qu'un ouvrier c'est ce que nous allons essayer de confirmer ici avec notre graphique.

On remarque qu'il y a une catégorie avec un bien plus fort taux de défaut que les autres catégories, il s'agit de la catégorie 8 qui comprend les personnes sans activité professionnelle. Ensuite, nous pouvons retrouver aussi les commerçants et les agriculteurs tandis que les métiers ayant la moins grande probabilité de faire défaut dans notre échantillon sont les personnes du secteur privé en l'occurrence les cadres.

Ces différents graphiques ont pu nous présenter des variables clés dans la discrimination de nos personnes par rapport au taux de défaut.

Découpage du jeu de données

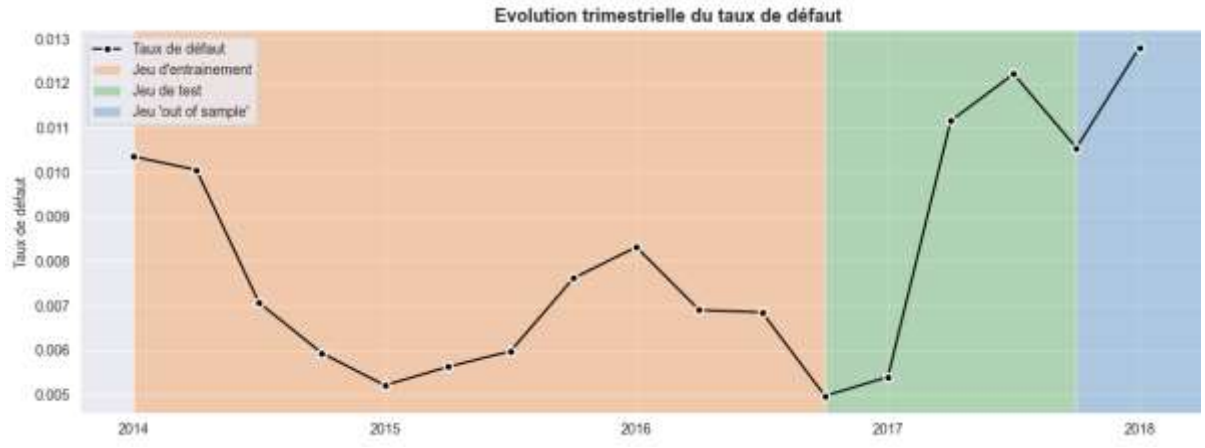
Pour la stratégie de sélection des variables présentée ci-après, nous avons au préalable réalisé un découpage de notre jeu de donnée. Ce découpage nous permet d'éviter les problématiques de *data leakage*, ou de fuite d'information entre notre base de données d'entrainement et notre base d'évaluation. Par exemple, si l'on observe statistiquement qu'une variable est significativement liée à la cible sur l'ensemble du jeu de donnée. Nous choisirons de ce fait de considérer cette variable comme pertinente pour la modélisation. Dès lors, même si l'on réalise un découpage du jeu de donnée pour évaluer notre modélisation, une information présente dans le jeu de test aura été tacitement prise en compte pour l'étape de *feature selection*. C'est pourquoi nous avons décidé d'effectuer le découpage avant même l'étape de sélection des variables.

Concernant la stratégie de découpage, l'approche traditionnelle est de diviser le jeu de données en un jeu d'entrainement et en un jeu de test en subdivisant éventuellement le jeu d'entrainement en un jeu d'entrainement et en un jeu de validation. Cela tout en gardant la

même proportion de clients tombés en défaut dans chaque jeu (surtout dans notre cas ou la répartition est fortement déséquilibrée). Cependant ici, la consigne nous a été donnée de raisonner plutôt comme sur un cas de série temporelle. Ainsi, notre échantillonnage se fit de la façon suivante :

- Toutes les données qui précèdent à 2017 constituent le jeu d'entrainement.
- Les données de 2017 constituent le jeu de test.
- Les données correspondant à l'année 2018 (nous ne disposons que deux trimestres) seront constitutives de l'échantillon *out-of-sample.*

Voci, un graphique, illustratif de notre découpage :



Ainsi, toute la partie qui suit sur la sélection des variables a été réalisée sur le jeu d'entrainement.

Sélection des variables par tests statistiques

Les variables explicatives de la variable cible correspondent aux facteurs qui ont conduit à une situation de défaut. Plutôt que de les sélectionner manuellement (un travail long et fastidieux compte tenu du nombre de variables), nous avons choisi d'avoir recours à des tests statistiques. En effet, on ne veut garder que des variables qui sont significatives statistiquement. Cela nous a permis d'effectuer un premier tri dans le choix des variables à garder pour la modélisation. Le choix du test adéquat dépend du type de la variable (continu ou discret) analysée, comme nous allons le voir ci-après.

2.1. Le cas des variables continues

Afin de détailler notre procédure concernant les tests sur les variables continues, il faut rappeler que notre variable indicatrice du taux de défaut est binaire et est définie selon le schéma suivant :

$$defaut = \begin{cases} 1 \; si \; le \; client \; est \; tombé \; en \; défaut \\ 0 \; sinon \end{cases}$$

A partir de cette observation, nous pouvons scinder chaque variable continue en 2 vecteurs distincts :

- Un premier vecteur v_1 contenant les observations de la variable pour les clients tombés en défaut ($defaut = 1$)
- Un second vecteur v_2 contenant les observations de la variable pour les clients non-défaillants ($defaut = 0$)

On veut alors comparer ces 2 vecteurs, afin de savoir si leur distribution est significativement différente ou non. Dans le premier cas, on considère que la variable peut être sélectionnée comme candidate à notre modèle car elle permet de segmenter dans une certaine mesure notre variable cible. Dans le second cas, on considère que l'on peut jeter la variable car elle n'apporte aucune information sur la tombée en défaut.

Reste alors à déterminer le choix du test entre les deux vecteurs v_1 et v_2. La première approche lorsque l'on veut comparer deux variables continues est souvent de recourir au test de Student. Cependant, ce dernier pose l'hypothèse de la normalité de nos deux variables. Or, il s'agit là d'une hypothèse forte que nombre de nos variables ne vérifient pas comme montré dans la première partie. Dès lors, il faut alors se tourner vers l'horizon des tests non-paramétriques. Ce qui nous amène au test de Wilcoxon-Mann-Whitney, ou de sa généralisation à k échantillons : Le test de **Kruskal-Wallis**.

L'hypothèse nulle et l'hypothèse alternative du test sont les suivantes (pour 2 échantillons) :

- $H_0 : P(v_1 > v_2) = P(v_2 > v_1)$, c'est-à-dire que la probabilité qu'une observation de v_1 soit supérieure à une observation de v_2 est la même que la probabilité qu'une observation de v_2 soit supérieure à une observation de v_1. On considère alors que les deux distributions sont identiques.
- $H_A : P(v_1 > v_2) \neq P(v_2 > v_1)$, où les distributions sont différentes.

Si l'on ordonne le vecteur v_1 U v_2 et que l'on considère S_{v1} la somme des rangs des éléments de v_1 au sein du vecteur v_1 U v_2. On peut alors montrer que, sous H_0, $S_{v1} = t$ suit une loi de distribution connue, que l'on peut approcher par une loi normale de paramètres $\mu = \frac{\#v_1 \#v_2}{2}$ et $\sigma^2 = \frac{\#v_1 \#v_2 (\#v_1 + \#v_2 + 1)}{12}$.

$\#v_1$ en faisant référence au cardinal du vecteur v_1.

A partir de cette loi, on peut ainsi calculer la valeur $\epsilon = \frac{|S_{v1} - \mu|}{\sigma^2}$, que l'on compare avec un seuil de risque à 5% (ce seuil est ajustable, mais nous avons pris 5% par défaut). Ainsi, si la valeur de ϵ est supérieure à 1.96, on peut rejeter l'hypothèse nulle stipulant que nos échantillons sont égaux. Dans ce cas, on peut garder notre variable pour la modélisation.

Nous avons donc réalisé cette procédure pour l'ensemble des variables continues.

2.2. Le cas des variables discrètes

Dans le cas des variables discrètes, la méthode définie pour les variables continues n'a plus de sens. Il faut avoir recours à un autre test statistique : Le test du χ^2. Ce dernier peut être utilisé à diverses fins, nous l'utiliserons ici pour tester l'indépendance entre notre cible (la tombée en défaut) et les différentes variables discrètes dont nous disposons.

Pour chaque variable var, l'hypothèse nulle (H_0) du test de χ^2 stipule que la variable var et la variable cible $defaut$ sont indépendantes. Autrement dit, le fait d'avoir une

information sur la variable var ne nous donne aucune information sur la variable $defaut$. La procédure du test est la suivante :

On pose N le nombre d'observations. var et $defaut$ prennent un nombre fini de valeurs ; respectivement K et 2. On note O_{kj} l'effectif observé des données pour lequel var prend la valeur k et $defaut$ la valeur j. Sous H_0, dans une telle configuration, la valeur espérée est donnée par :

$$E_{kj} = \frac{O_{k+} * O_{+j}}{N}$$

Avec $O_{k+} = \sum_{j=1}^{2} O_{kj}$

Et $O_{+j} = \sum_{k=1}^{K} O_{kj}$

On calcule la statistique de test T comme suit :

$$T = \sum_{k,j} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}$$

A noter que T suit une loi de χ^2 à $(K-1)(2-1) = (K-1)$ degrés de liberté.

On peut alors estimer la probabilité de l'erreur de première espèce avec un seuil là encore fixé par défaut à 5%. Si la p-valeur se situe en deçà des 5%, on ne peut pas accepter H_0. On considère que notre variable n'apporte pas d'information sur notre cible. On peut donc en faire fi. Dans le cas contraire, on garde la variable comme significative vis-à-vis du test de χ^2, elle nous apporte donc de l'information sur notre cible.

En appliquant cette procédure à nos 14 variables qualitatives, on en garde 6 comme statistiquement significatives.

Sélection des variables explicatives et modélisation par régression logistique

Après avoir réalisé nos différents tests statistiques afin de pouvoir sélectionner nos variables comme expliqué ci-dessus nous avons par la suite réalisé une étape de discrétisation sur ces variables. Cette étape est très importante pour établir un modèle. En effet, si nous prenons par exemple les variables qualitatives, il y en a avec énormément de modalités et donc au moment de l'encodage de ces variables on obtiendrait un modèle très complexe nous exposant directement à la menace de *l'overfitting*.

3.1. Regroupement, Discrétisation

C'est pourquoi dans notre projet, nous avons utilisé une librairie du nom d'OptBinning. Le but étant de regrouper des valeurs (d'une certaine variable qu'elle soit quantitative ou qualitative) ayant des comportements similaires par rapport à notre target ici `'defaut_36mois'`. Cette librairie nous permet d'établir un regroupement le plus optimal possible en nous donnant le contrôle sur les paramètres et les contraintes.

A noter, que plus tôt, lors de l'étape de nettoyage des données, nous n'avons pas traité les valeurs manquantes et les valeurs extrêmes car cette technique peut résoudre ces problèmes en créant par exemple une modalité comprenant les valeurs manquantes (NR).

De surcroît, les valeurs extrêmes vont se retrouver groupées dans des classes, leur influence sera alors limitée dans nos estimations et cette méthodologie va par ailleurs accroître la stabilité du modèle que nous allons construire.

Pour obtenir un regroupement optimal nous utilisons la classe Binning_Process. Nous obtiendrons à la suite de cela plusieurs résultats dont notamment :

- Un score de qualité
- L'information value
- Nombre de groupes

L'information value va jouer un rôle important ici. En effet, il s'agit d'une des techniques les plus utiles pour sélectionner des variables dans un modèle prédictif. Il permet de classer les variables en fonction de leur importance sur l'estimation de la variable cible.

On calcule l'IV de la manière suivante : IV = $\sum(\%\ de\ non\ événements - \%\ d'événements) * WOE$

Avec le WOE qui correspond au poids de la preuve. Il s'agit d'une mesure de séparation des bons et des « mauvais clients » c'est-à-dire les clients qui ont fait défaut pour les « mauvais clients » et ceux qui ont remboursé leur prêt.

Ci-dessous, nous pouvons trouver un tableau évoquant les règles relatives à l'information value :

Information Value (IV)	Predictive Power of the Variable
<0.02	Useless for prediction
0.02 to 0.10	Weak predictor
0.10 to 0.30	Medium predictor
0.30 to 0.50	Strong predictor
>0.50	Suspicious behaviour

Note de lecture : Règles relatives de l'Information Value

C'est pourquoi nous allons surtout nous intéresser aux variables avec plus de 0.10 en IV afin d'éviter les prédicteurs faibles et les prédicteurs inutiles.

De surcroît, l'intérêt d'utiliser la librairie OptBinning est de pouvoir - comme nous l'avons mentionné - utiliser un regroupement optimal. Cela car on pourrait seulement agir en fonction de l'IV et de ce fait on pourrait créer beaucoup plus de groupes et ainsi il pourrait y avoir des groupes ayant un très petit nombre d'événements et de non-événements. Tandis que l'IV serait élevé.

C'est pourquoi il nous a semblé pertinent d'allier ces différents méthodologies et outils.

Ainsi, nous retrouvons nos variables gardées lors de la phase de sélection de variables à l'aide de tests statistiques. Nous les discrétisons de façon optimale afin d'obtenir des variables avec plusieurs modalités, au minimum 2, que ce soit pour les variables continues ou les variables

qualitatives. Tout d'abord, pour les variables qualitatives, nous discriminons nos modalités en fonction du taux de défaut afin de réunir celles ayant un comportement semblable au niveau du défaut. Tandis que pour les variables quantitatives, nous créons des intervalles à partir desquelles si un individu se trouve dans cet intervalle il a environ la même probabilité de tomber en défaut qu'un autre individu présent dans cette modalité créée (en l'occurrence l'intervalle ici). Le tout ici est réalisée en maximisant aussi l'information value afin de pouvoir obtenir les meilleurs prédicteurs possibles et par ailleurs en gardant une certaine homogénéité de nos populations dans chaque modalité, en effet, les modalités sont dites stables en termes d'effectif.

Ensuite, nous ne gardons que les variables qui sont au minimum considéré comme des prédicteurs moyens en l'occurrence des variables ayant une IV supérieur à 0.10 (= le seuil qu'on choisit pour sélectionner les variables explicatives), à noter que nous n'avons aucune variable avec une IV supérieur à 0.50 témoignant d'un prédicteur dit 'suspect'.

De ce fait, nous obtenons 9 variables significatives en fonction de notre seuil. Ces variables vont comprendre comme valeur le numéro de leur classe correspondant à la modalité à laquelle l'individu appartient.

Ainsi, pour que la régression logistique puisse lire nos données nous allons créer une variable qui correspondra à une modalité de la variable. Donc, nous allons retrouver nos modalités en tant que colonnes et en tant que variables explicatives dans notre modèle. Les individus auront pour valeur 0 s'ils n'appartiennent pas à la modalité et 1 s'ils y sont présents.

Modélisation

Nous voulons modéliser la probabilité à 36 mois qu'un client soit en défaut donc nous utilisons un modèle de classification binaire. Le modèle utilisé dans toutes les banques est la régression logistique car elle offre une meilleure interprétabilité que les autres modèles de classification et dispose de coefficient permettant de réaliser une grille score. Le modèle logistique permet une expression non linéaire, variant de façon monotone entre 0 et 1, cette expression de la probabilité à calculer en fonction des variables explicatives (Xj) est la suivante :

$$\ln\left(\frac{p(1|X)}{1-p(1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

En effet, après transformation de l'équation ci-dessus, nous obtenons :

$$p(1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

4.1. Hypothèse de régression logistique

- La régression logistique nécessite que la variable dépendante soit binaire et la régression logistique ordinale exige que la variable dépendante soit ordinale.
- La régression logistique nécessite que les observations soient indépendantes les unes des autres. En d'autres termes, les observations ne doivent pas provenir de mesures répétées ou de données appariées.
- La régression logistique exige qu'il y ait peu ou pas de multi-colinéarité parmi les variables indépendantes. Cela signifie que les variables indépendantes ne doivent pas être trop fortement corrélées entre elles.

- Contrairement au Moindre Carré Ordinaire (MCO), l'homoscédasticité n'est pas requise et les termes d'erreur n'ont pas besoin d'être distribués normalement.
- La variable cible est catégorielle (Oui / Non).
- Les termes d'erreur sont indépendants.

Pour entraîner notre modèle nous avons utilisé une méthode de validation croisée adapté à nos données déséquilibrées. Cette méthode est le RepeatedStratifiedKFold qui coupe le jeu de donnée en 5 par exemple et entraine le modèle sur les 4 premiers découpage et test sur le dernier découpage qui fait office de jeu de validation. De plus cette méthode conserve la part d'observation en défaut et non défaut à chaque étape de découpage. Nous avons intégré notre méthode de validation à une méthode de recherche d'hyperparamètres optimaux appelé GridSearchCV pour être sûr que notre modèle ne colle pas trop aux données d'entraînement (overfitting). Dans la librairie scikit-learn nous ne pouvions pas spécifier dans la même colonne la variable de référence et obtenir les coefficients associés aux modalités. Nous avons donc créé des variables dummy pour chaque modalité en enlevant la première pour éviter le problème de colinéarité. Ensuite nous avons pu estimer les coefficients de modalités. Or les p-values associés aux modalités n'étant pas toutes significative, nous avons gardé seulement celles significatives avec un test du χ^2.

Création d'une grille de score

Dans la réalisation de la grille de score nous avons recouru surtout à un modèle logit qu'à un modèle linéaire, car le modèle logit défit plus précisément la probabilité cumulative de défaut des clients. La raison pour laquelle le modèle linéaire reste assez faible en robustesse dans l'étape de construction du score est qu'il conduit à des probabilités qui sortent de l'intervalle [0 ;1].

Pour la construction de la grille nous nous sommes résumés à une échelle de score allant de 0 à 1000 :

- 0 → degré élevé de risque de défaut
- 1000 → degré bas de risque de défaut

Un mot maintenant sur les résultats obtenus au sein de grille de score. En effet, il convient de les aborder car il s'agit là du rendu final du projet. Vous trouverez d'ailleurs cette dernière dans le fichier archivé que nous vous avons joint. Nous avons suivi la méthode donnée pour calculer nos coefficients associés à chaque modalité. De fait, nous observons des résultats semblant incohérents d'un point de vue métier. Nous souhaitons ici discuter de la cause de ces incohérences.

Tout d'abord, nous avions initialement retenus 11 variables par le biais des différentes méthodes de sélection de nos variables se concluant par l'utilisation d'OtpBining et de sa méthode de discrétisation par l'Information Value. Or nous avons remarqué en construisant la grille que deux de nos variables présentaient la même description et quasiment les mêmes modalités que deux autres variables présentes dans notre sélection. Nous avons donc fait le choix d'enlever manuellement ces variables, pour gagner en interprétabilité métier. De facto, nos résultats sont restés quasiment identiques, bien que nous ayons perdus en performance de prédiction sur l'échantillon out of sample.

Le second point qu'il nous faut aborder ici concerne la technique de sélection des modalités. L'approche courante est d'adopter une démarche stepwise, enlevant progressivement les modalités si ces dernières sont considérées comme non-significatives statistiquement. Ici nous avons choisi une approche différente. Nous avons décidé de sélectionner les modalités significatives par rapport à la tombée en défaut via un test du χ^2 avec un seuil de confiance fixé à 15%. Cela car nous rencontrions des p-valeurs en NaN lors de notre régression stepwise avec un modèle Logit. Ainsi, notre approche astucieuse nous a permis de valider statistiquement la pertinence de nos modalités. Nous pensons que les résultats singuliers obtenus sont peut-être la conséquence de cette sélection des modalités.

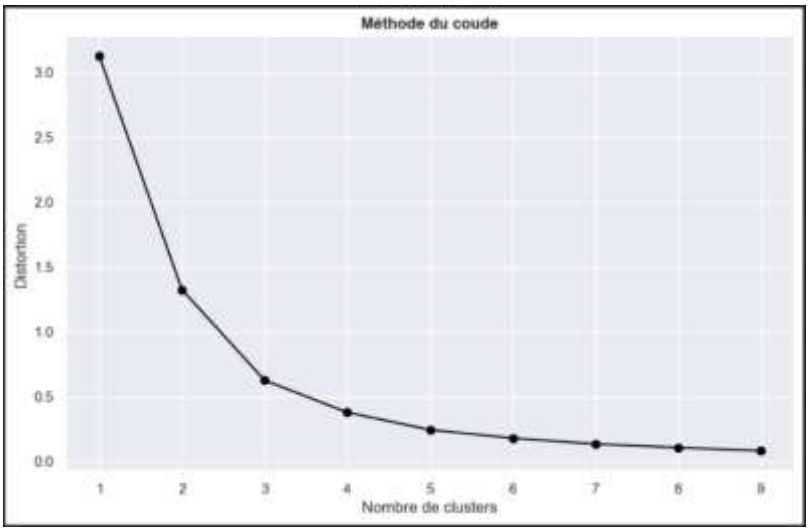
Création de classes de risque

6.1. Clustering par K-means

Afin d'orienter la décision du conseiller, on regroupe les individus dans différentes classes en fonction de la probabilité de défaut estimée par le score en l'occurrence on veut obtenir des classes regroupant des personnes ayant un risque faible, un risque modéré et enfin un risque très élevé.

Pour déterminer les classes, nous avons utilisé un algorithme de Machine Learning : Le clustering par les K-moyennes.

Tout d'abord, afin de classifier les données, nous devons choisir le nombre de clusters pertinents à priori. Pour ce faire, nous utilisons la méthode 'Elbow' afin d'identifier le nombre de clusters pour lequel l'ajout d'un autre cluster ne donne pas une meilleure modélisation des données. En effet, utiliser trop peu de clusters va nous conduire à des classes avec des caractéristiques hétérogènes. A l'inverse, si l'on utilise trop de clusters, on aura des classes qui ne se différencie pas suffisamment.



On se rend compte avec ce graphique que la courbe est plutôt coudée pour une valeur de k=3. Il semble donc y avoir 3 clusters pertinents. Ainsi, nous avons segmentés notre population en 3 classes. Une fois nos 3 classes créées nous avons pu constater que ces classes de risques sont assez différenciées en regardant la moyenne des taux de défauts de chaque classe. La première

a un taux de défaut moyen de 0.0035, la deuxième de 0.0320 et enfin la dernière de 0.0138. Ainsi, nous avons bien une différenciation appropriée entre les classes.

Après avoir réalisé notre clustering, nous avons analysé la stabilité trimestrielle de nos taux de défaut par classe sur l'ensemble de la période donné. On remarque une certaine stabilité des classes de risque faibles et des classes de risques moyen. C'est au niveau des classes des risques élevés où là la stabilité est moins importante. Mais on peut tout de même noter que la classe à risque élevé reste tout de même plus élevé en termes de taux de défaut que les autres classes au fil du temps hormis au début de l'année 2017.

6.2. Stabilité en risque des modalités du modèle :

La construction des classes de risque et évaluation des probabilités de défaillance sont des étapes fortement liées entre elles. Les intervalles qui délimitent ces classes sont calculés de façon à garantir la stabilité temporelle des probabilités, quel que soit l'échantillon temporel étudié.

En particulier, à l'horizon 2014/04/ -2017/04, pour les modalités des variables retenues pendant l'étape de modélisation nous avons constaté une certaine stabilité trimestrielle en risque pendant les 6 premiers trimestres. Les modalités des variables deviennent instables dans la dernière année 2017. Cela se traduit économiquement par le fait que les émissions de crédits auprès des clients ont souffert des épisodes d'instabilité autour de cette période. (Graphique Annexes : Stabilité temporelle (trimestre) en risque des variables). Ensuite, cette problématique liée à la temporalité en risque est probablement dû au nombre petit du taux de défaut présent à cette période. Les variables : « Taux d'apport par personne », « Top premier achat immo client » et « Nature du projet » sont les variables plus stables à l'horizon trimestriel, semestriel et annuel (voire Annexes).

6.3. Stabilité en effectif des modalités du modèle :

Graphiquement, la stabilité en effectif du taux de défaut reste meilleure que celle en risque pour la majorité des variables du modèle. Une seule modalité de la variable « Montant total de l'assurance » reste in stable par rapport à l'effectif pendant l'année 2014, tandis que les autres modalités des variables ont enregistré une stabilité temporelle à l'horizon trimestriel/semestriel/ annuel.

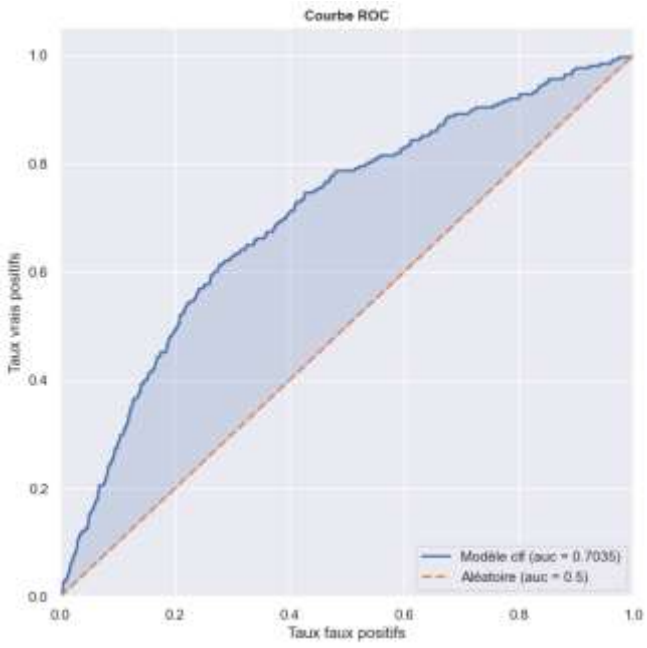
Cette stabilité temporelle observée sur nôtres variables nous permet de conclure que les variables ont été sélectionnées et discrétisées de manière optimale par rapport à la variable d'intérêt qui est le taux de faire un défaut auprès d'un client.

Analyse des performances

Grace à l'analyse exploratoire faite plus haut, nous avons constaté un déséquilibre de classes existant au sein du jeu de données. Pour faire face à ce problème de classification déséquilibré en termes de performances du modèle, nous avons choisi un indicateur pertinent qui permettait de juger réellement la qualité de notre modèle.

La courbe ROC est utilisé pour évaluer les performances de ce modèle. Graphiquement, la mesure ROC est représentée sous la forme d'une courbe qui donne le taux de vrais positifs, la sensibilité, en fonction du taux de faux positifs, l'anti-spécificité (1 - spécificité). La courbe ROC, représente l'outil idéal pour résumer la performance de classifieur binaire de tomber en défaut ou pas en fonction de tous les seuils possibles car il résume parfaitement les 2 ratios de la matrice de confusion que l'on a cité. En l'occurrence, notre courbe va être construite en utilisant des matrices de confusion issues de seuils compris entre 0 et 1000 et pilotant leur taux de vrais positifs et leur taux de faux positifs. Sachant que pour notre modèle un vrai positif est que l'emprunteur a fait défaut et que le modèle le prédit aussi, un faux positif est lorsque l'individu n'a pas fait défaut mais que le modèle l'a prédit et inversement pour les individus classés en négatifs. En l'occurrence, le modèle n'a prédit aucun défaut alors que la personne a fait défaut il s'agit là d'un vrai négatif et enfin lorsque le modèle n'a prédit aucun défaut mais que l'emprunteur a fait défaut là on parle de faux négatif.

Pour une précision des résultats, nous avons calculé l'aire sous la courbe (AUC : Area Under the Curve), qui reste un indicateur très utile pour résumer la capacité du modèle à distinguer les classes des clients de tomber en défaut de la classe de ne pas avoir faire une incapacité de payement. Le score AUC obtenu est de 0,70 sur le test set. Ce score signifie que le modèle est meilleur qu'une classification aléatoire, c'est à dire un modèle assez bien prédictif.



La 2ème métrique à analyser est l'indice de Gini, elle nous permet d'évaluer la performance de notre modèle de risque de crédit. L'indice de Gini est directement lié à la valeur de l'AUC et utilisée comme métrique pour discriminer notre modèle. En effet, il teste l'efficacité du modèle à différencier les 'mauvais' emprunteurs des 'bons' emprunteurs. C'est-à-dire les personnes faisant défaut et les personnes qui ne font pas défauts. Comme on l'a dit, cet indice est lié directement à l'AUC car il existe une relation linéaire entre ces 2 métriques. On obtient pour notre part un indice de Gini de 0.40, cette métrique nous a servi pour comparer nos modèles réalisés afin de vous proposer le meilleur possible et celui qui est le plus significatif.

L'indice de Gini est donné par : Gini = 2 * AUC – 1

Tester des algorithmes de machine Learning et comparer les performances avec la régression logistique

Nous allons à présent aborder la modélisation de nos données par d'autres modèles que la régression logistique. Après avoir testé différents modèles, nous nous sommes rendus compte que notre jeu de données et notre pré-traitement nous donnent de bons résultats sur des algorithmes linéaires paramétriques. En revanche, quand il s'agit d'avoir recours à des modèles non paramétriques comme les algorithmes ensemblistes de *bagging* ou de *boosting*, nos performances se dégradent. A cela nous voyons deux explications.

La première explication, c'est que l'ensemble de notre pré-traitement a été réalisé de sorte à obtenir une bonne performance sur un modèle linéaire. Ainsi, nous avons choisi nos variables puis discrétisés ces dernières sur des critères statistiques. C'est surtout le premier point qui pose problème selon nous. En effet, nous avons sélectionné nos colonnes de sorte à éviter la multi colinéarité. On le sait, la multi colinéarité peut fausser les résultats de l'inférence, il faut donc la limiter au maximum dans un modèle linéaire. Les modèles non-paramétriques, en revanche ne sont pas sensibles à cette contrainte. Ajouter des variables corrélées peut s'avérer bénéfique, si tant est que ces dernières apportent de l'information sur la cible. Ainsi, cantonner le nombre de variables comme nous l'avons revient à tirer une balle dans le pied de nos modèles non paramétriques. C'est l'éternelle dualité entre le *machine learning* et la statistique ; significativité statistique et pouvoir de prédiction ne sont pas synonymes, et peuvent même devenir antagonistes.

La seconde explication concerne la distribution des données. On le sait, les algorithmes non paramétriques ont une tendance à overfitter rapidement. La condition *sine qua none* pour que cet overfitting diminue lors du processus ensembliste, est que les observations soient suffisamment différentes les unes aux autres (voire idéalement i.i.d). Ici, ce n'est pas le cas, nos lignes sont toutes similaires (la similarité cosinus minimum est de 0.96). Combiné au faible nombre de défauts dans nos données, cela donne une bonne piste pour expliquer l'overfitting des modèles plus complexes implémentés.

Voici un tableau récapitulatif des résultats obtenus (avec en bleu les modèles linéaires et en vert les modèles non linéaires) :

Modèle	AUC sur le train set	AUC sur le test set
Régression Logit	0,75	0,70
Linear Discriminant Analysis	0,76	0,70
Light GBM	0,94	0,63
Random Forest	0,94	0,63

AUC : Aire sous la courbe ROC

Conclusion

Vous trouverez ainsi au sein de ce rapport, de la grille fournie, et de nos fichiers de code l'ensemble de notre démarche vis-à-vis du projet de modélisation des probabilités de défaut sur les dossiers immobiliers fournis par LCL. Bien que les résultats obtenus au niveau de la grille de score soient parfois singuliers d'un point de vue métier, nous sommes quand même assez contents du travail réalisé. Il convient de noter que l'exercice était assez difficile compte tenu du fort déséquilibre entre les classes (0,07% de clients tombés en défaut). Nous estimons avoir réussi à proposer un modèle somme tout assez performant. Peut être serait-il intéressant de refaire une étude avec un jeu de données plus équilibré. Cela augmentera notamment fortement les performances des modèles challengers, pour qui le déséquilibre semble invariablement mener vers un overfitting dans la configuration actuelle.