SEGMENTATION DE LA PROBABILITÉ DE DÉFAUT IFRS9

GROUPE 24 : GANIYU Isaac TOURE Abdoul-Aziz SANDJA Nathan TORRES Jason

ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE DE L'INFORMATION

Contexte

Objectif et méthode

OBJECTIF : Segmenter les clients en groupes homogènes par rapport au taux de défaut.

MÉTHODE : Mettre en exergue des variables segmentant au mieux la probabilité de défaut ainsi que les découpages adéquats de ces variables. Pour cela on utilise des modèles de segmentation. On construira ensuite les segments par croisement des variables retenues.

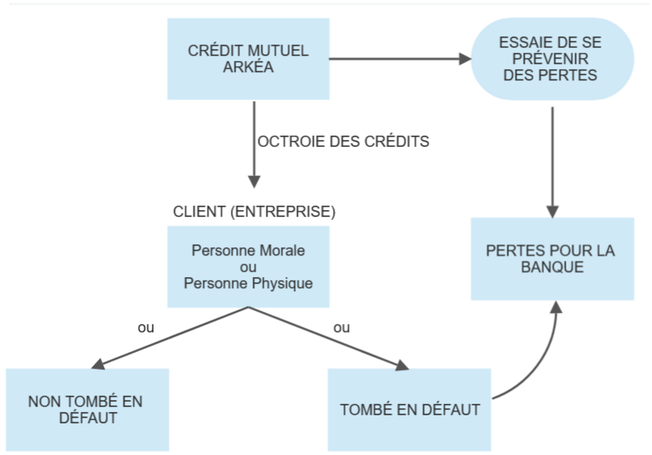

Figure 1. Structuration de la banque vis-à-vis des tombés en défaut

Données

Données : Base de données d'emprunteurs de Crédit Mutuel Arkéa avec plusieurs variables signalétiques du défaut.

Description de la base de données

Dans le cadre de notre projet, nous possédons un jeu de données renseignant les caractéristiques de différents emprunteurs. Le jeu de données de notre étude comporte 618 745 observations qui représentent les emprunteurs sains et tombés en défaut au cours des années 2017 à 2021. Il y a 22 variables d'étude, dont 1 variable quantitative et 21 variables qualitatives. La variable d'intérêt désigne la variable de mise en défaut, cette variable vaut :

- 0 : Non défaut
- 1 : Tombé en défaut

Modélisation

Présélection des variables

Régression logistique

Afin de pouvoir présélectionner un premier groupe de variables, nous avons réalisé une régression logistique :

$$\text{logit}(p) = \beta_0 + \beta_1 1_{\text{modalite1}} + \beta_2 1_{\text{modalite2}} + ... + \beta_n 1_{\text{modalite}_n}$$

Où n désigne le nombre de modalités totales traitées par l'ensemble des variables explicatives.
Afin de déterminer la significativité des variables indicatrices, on utilise le critère : $p\ valeur < 0.05$.

ANOVA

Afin de déterminer les variables en relation avec la probabilité de défaut nous avons réalisé une analyse de la variance (ANOVA) à plusieurs facteurs. Le but de l'analyse de variance est de comparer, pour une variable donnée, les moyennes de ses modalités en terme de défaut. On détermine ensuite si les moyennes sont significativement différentes les unes des autres. Si c'est le cas, alors la variable considérée a un effet significatif sur notre variable d'intérêt.

Modèles de prédiction du défaut

L'étude consiste à mettre en évidence un petit nombre de variables (3 au maximum) permettant d'aboutir à la meilleure segmentation de notre base de données avec des segments homogènes et distincts les uns des autres. Pour cela, on effectue les divisions successives de nos modèles de segmentation selon l'indice de Gini. Dans le cas de deux classes, l'indice de Gini s'exprime par :

$$I(h) = 1 - P_0(h)^2 - P_1(h)^2 = 2 * P_1(h) * (1 - P_1(h))$$

Avec $P_1(h)$ le taux de défaut pour un groupe de clients h donné.

CART avec des coûts d'erreurs

Dans ce modèle CART (Classification And Regression Trees), on définit des coûts d'erreur de classement. Pour cela, on utilise la matrice de coûts suivante:

$$C = \begin{bmatrix} 0 & 0.7 \\ 0.3 & 0 \end{bmatrix}$$

Avec $C_{ij} = c(i|j)$, i,j = 0 ou 1.
$c(i|j)$ le coût d'erreur d'affectation d'un individu à la classe i alors qu'il appartient à la classe j.

CART avec une base équilibrée

Au vu du déséquilibre de la base de données initiale par rapport à notre variable d'intérêt, Nous avons équilibré les proportions de chaque modalité de la variable d'intérêt dans la base qui servira à entraîner le modèle CART. L'apprentissage dans ce modèle s'est donc fait sur une base de données avec une proportion de tombés en défaut égale à celle des non tombés en défaut.

Forêts aléatoires

La méthode des forêts aléatoires (Random Forest) est une autre méthode d'apprentissage supervisé reposant sur la génération d'un grand nombre d'arbres aléatoires. Les règles de décision associées sont agrégées pour obtenir le classifieur final. À chaque construction d'un arbre, on effectue deux variations: La variation des observations et la variation des variables.

Résultats et Comparaisons

Choix du modèle final

Modèle	Accuracy	Balanced accuracy	F1-score
CART avec coûts d'erreur	0.945	0.73749	0.27210
CART avec base équilibrée	0.8258	0.82487	0.15635
Random Forest	0.8075459	0.81673461	0.14484740

Table 1. Indicateurs de nos modèles

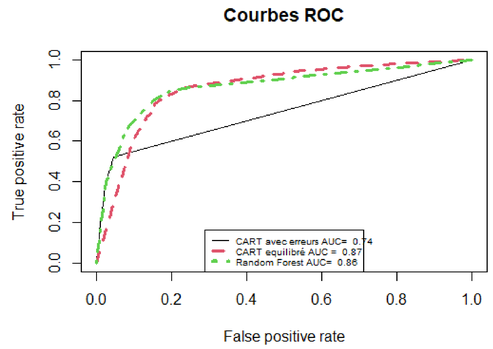

Figure 2. Courbes ROC des modèles

Compte tenu de ces indicateurs et pour des raisons d'interprétabilité, nous retenons le modèle CART avec coûts d'erreur.

Conclusion et interprétations

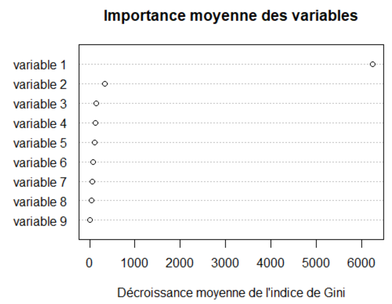

Figure 3. Importance moyenne des variables

- Les variables les plus importantes sont retenues en tenant compte des résultats des trois modèles
- On effectue la segmentation suivant les découpages retenus des variables retenues.
- La segmentation nous permet d'aboutir à plusieurs interprétations. Toutefois, pour des raisons de confidentialité, nous ne pouvons pas présenter de résultats plus détaillés