RÉSUMÉ

L'objectif de ce rapport est de prédire la détention d'une carte Visa Premier à partir des comportements bancaires de 1 063 clients, et d'identifier les profils les plus susceptibles d'être intéressés par cette offre premium. Trois approches complémentaires ont été mises en œuvre : un arbre de décision (CART), un classificateur naïf de Bayes, et un réseau de neurones artificiels.

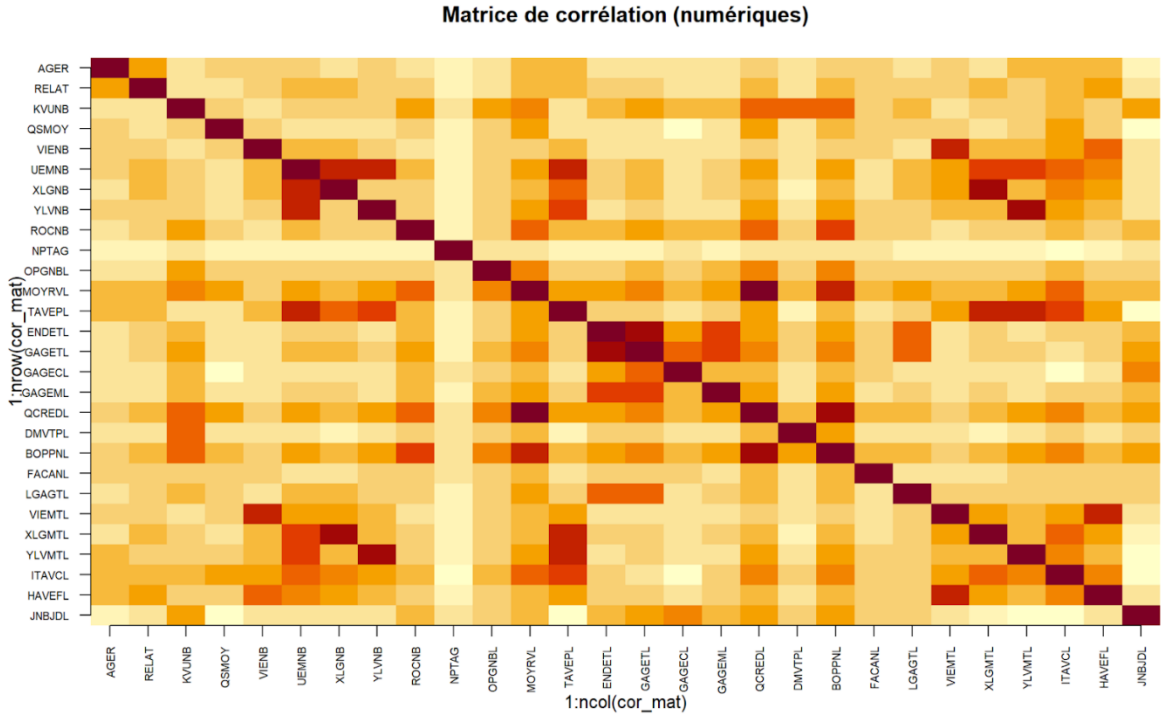
La stratégie fondée sur l'arbre de décision permet d'obtenir un modèle interprétable sous forme de règles de segmentation explicites. Après validation croisée et application de la règle du 1-SE, l'arbre retenu comporte 11 feuilles. Sur l'échantillon test, il atteint une AUC de 0,9229 et un taux d'erreur de 9,86 %, avec une sensibilité de 84,0 % et une spécificité de 92,8 %. Il offre ainsi le meilleur compromis entre capacité discriminante et lisibilité opérationnelle.

Le classificateur naïf de Bayes constitue un modèle de référence probabiliste plus simple. Après discrétisation des variables numériques et sélection basée sur le Khi² et le V de Cramer, le modèle réduit obtient une AUC de 0,8477 sur l'échantillon test, avec un taux d'erreur de 23,94 %. Bien que performant, il demeure structurellement limité par l'hypothèse d'indépendance conditionnelle, partiellement violée en présence de corrélations fortes entre variables financières.

Le réseau de neurones artificiels permet de modéliser des relations non linéaires plus complexes. Après optimisation des hyperparamètres par validation interne, le modèle retenu (RN5) atteint une AUC de 0,8736 sur l'échantillon test, avec un taux d'erreur de 20,56 %, une sensibilité de 71,7 % et une spécificité de 82,7 %. Malgré une bonne capacité de discrimination, ses performances restent inférieures à celles de l'arbre de décision, et son interprétabilité demeure plus limitée.

La comparaison des trois approches met en évidence la supériorité du modèle CART, tant en termes de discrimination globale que de performance de classification. Au-delà des indicateurs chiffrés, son avantage principal réside dans la transparence des règles produites, directement exploitables dans une perspective de ciblage commercial. L'arbre de décision à 11 feuilles apparaît ainsi comme la solution la plus pertinente dans ce contexte.

Figure 7 - Matrice de corrélation des variables quantitatives



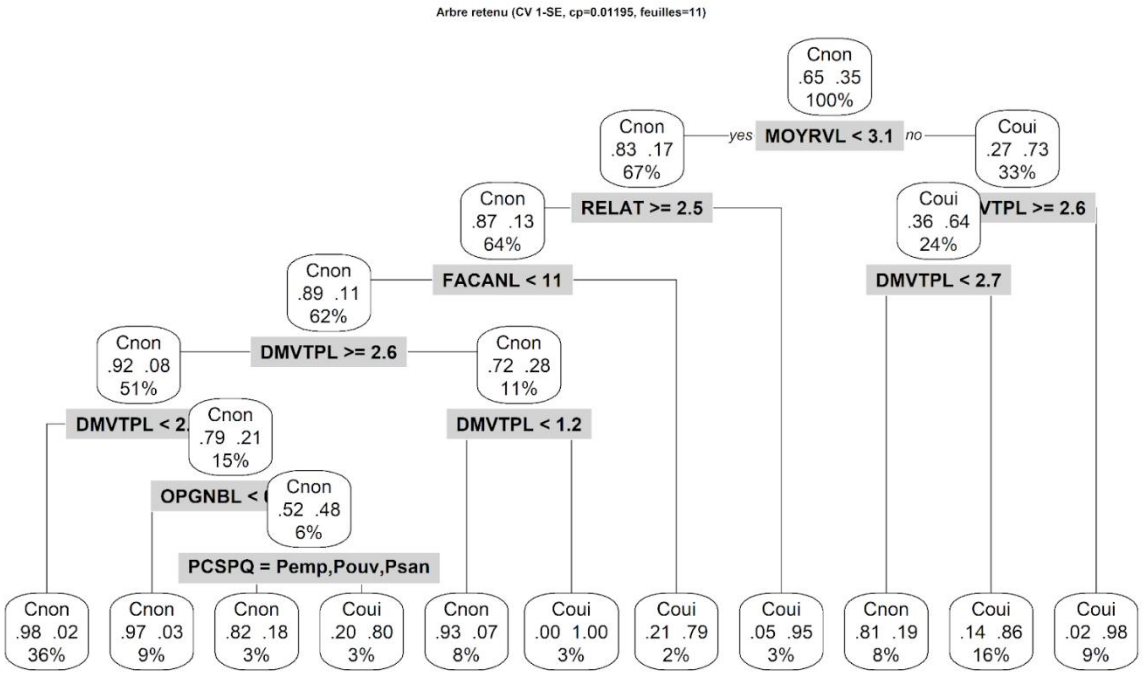
Au niveau des variables quantitatives, les corrélations entre les variables numériques et CARVPr montrent des associations globalement faibles à modérées, les plus fortes concernant des indicateurs liés au crédit/solvabilité (MOYRVL : r = 0,492; QCREDL : r = 0,457).

Néanmoins, la matrice de corrélation met en évidence une redondance entre plusieurs variables numériques, qui peuvent être regroupées en blocs :

- Bloc « solvabilité » : les associations les plus fortes concernent MOYRVL–QCREDL (r = 0,931) et QCREDL–BOPPNL (r = 0,845), suggérant une information largement partagée entre ces indicateurs
- Bloc « intensité d'usage » : des corrélations élevées apparaissent entre mesures exprimées en comptages et en montants, notamment YLVNB–YLVMTL (r = 0,847) et XLGNB–XLGMTL (r = 0,836) ; par ailleurs, UEMNB présente des liens avec plusieurs variables d'activité (par ex. UEMNB–TAVEPL : r = 0,772).
- « Dépendances ponctuelles » : certaines dépendances, plus localisées, sont observées entre ENDETL–GAGETL (r = 0,822) et VIENB–VIEMTL (r = 0,727), en cohérence avec des indicateurs décrivant un même processus sous des formulations proches.

Dans l'ensemble, la structure de corrélations indique une duplication partielle de l'information. Elle est susceptible de réduire l'interprétabilité des modèles paramétriques (via des effets de multi colinéarité) et de renforcer la non-indépendance entre prédicteurs dans le cadre du

Figure 10 - Arbre de décision final sélectionné par validation croisée



3. Performance de généralisation du modèle

Après avoir sélectionné le modèle final sur l'échantillon d'apprentissage à l'aide de la validation croisée (et des critères associés de choix de complexité), nous évaluons sa capacité de généralisation en l'appliquant à l'échantillon test, resté à l'écart de la phase d'entraînement. Les performances sont alors synthétisées à partir de la matrice de confusion et d'indicateurs usuels (taux d'erreur, sensibilité, spécificité), complétés par l'AUC afin de mesurer la capacité de séparation des deux classes indépendamment du seuil de décision.

Tableau 3 - Matrice de confusion du modèle final sur l'échantillon test

Réel \ Prédit	Positif	Négatif	Total
Positif	TP = 89	FN = 17	106
Négatif	FP = 18	TN = 231	249
Total	107	248	355

4. Performance sur l'échantillon test

Tableau 12 - Matrice de confusion du réeau de neurones RN5 sur l'échantillon test

Réel \ Prédit	Positif	Négatif	Total
Positif	TP = 89	FN = 17	106
Négatif	FP = 18	TN = 231	249
Total	107	248	355

Tableau 13 - Indicateurs de performance du modèle RN5 sur l'échantillon test (AUC, erreur, sensibilité, spécificité)

Indicateur	Valeur
AUC	0.8736
Erreur	20.56 %
Sensibilité (TPR)	71.7 %
Spécificité (TNR)	82.7 %

Figure 12 - Courbe ROC du réseau de neurones RN5 sur l'échantillon test

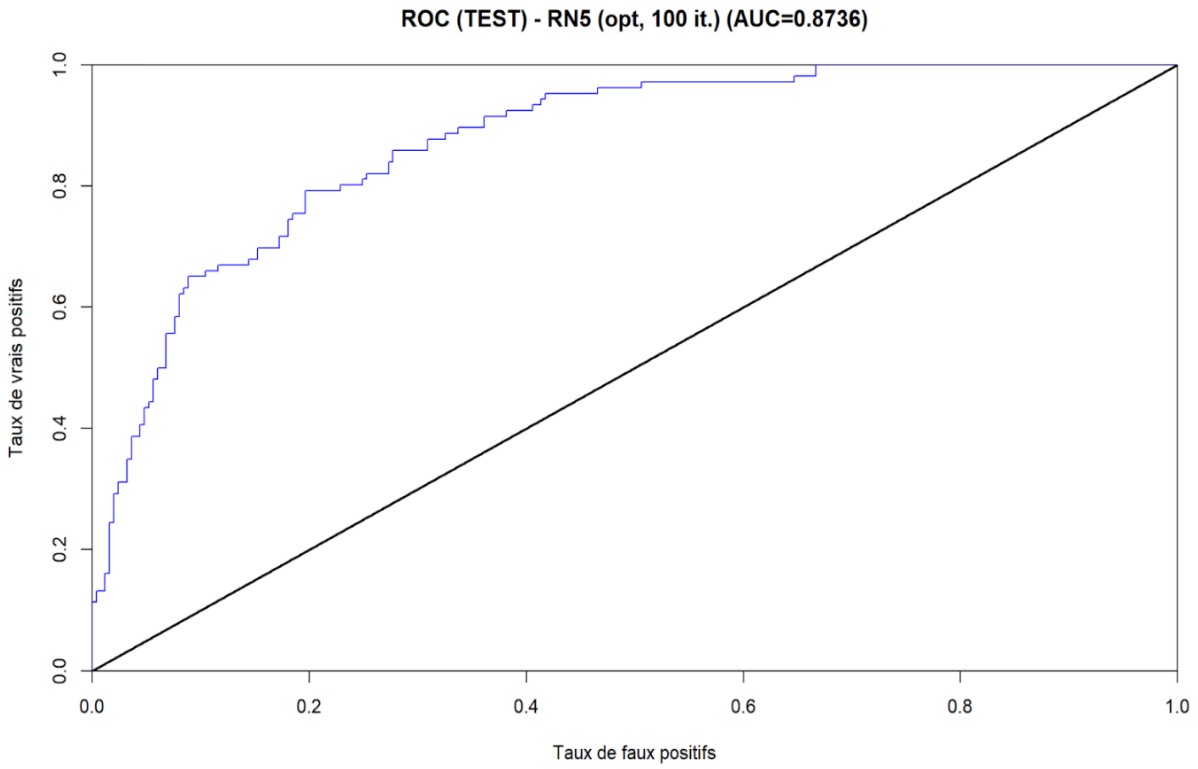


Tableau 14 - Comparaison des performances des modèles sur l'échantillon test

Modèle	AUC	Erreur	Sensibilité	Spécificité
Arbre de décision (11 feuilles)	0,9229	9,86 %	84,0 %	92,8 %
Réseau de neurones (RN5)	0,8736	20,56 %	71,7 %	82,7 %
BAYES Réduit	0,8477	23,94 %	83,0 %	73,1 %

La courbe ROC comparative confirme une hiérarchie nette des capacités de discrimination : CART domine l'ensemble de la plage des faux positifs, ce qui se traduit par l'AUC la plus élevée (0,9229). Cette supériorité se retrouve également sur les métriques à seuil fixe, avec un taux d'erreur minimal (9,86 %) et un profil équilibré, associant une sensibilité élevée (84,0 %) à une spécificité très forte (92,8 %). Autrement dit, l'arbre parvient simultanément à identifier une large part des détenteurs et à limiter efficacement les faux positifs, ce qui en fait le meilleur compromis dans ce cadre.

Le réseau de neurones (RN5) présente des performances intermédiaires : son AUC (0,8736) traduit une discrimination correcte mais inférieure à celle de CART, et son taux d'erreur (20,56 %) indique une généralisation moins favorable. Son profil (sensibilité 71,7 %, spécificité 82,7 %) suggère un compromis davantage orienté vers la réduction des faux positifs que vers la maximisation de la détection, au prix d'une proportion plus importante de détenteurs non identifiés.

Le modèle de Bayes réduit obtient l'AUC la plus faible (0,8477) et l'erreur la plus élevée (23,94 %), indiquant une capacité de discrimination plus limitée. Il se distingue toutefois par une sensibilité élevée (83,0 %), proche de CART, mais associée à une spécificité plus faible (73,1 %), ce qui traduit une stratégie plus « inclusive » générant davantage de faux positifs.

En synthèse, la comparaison sur test conduit à retenir CART comme modèle le plus performant, tant en discrimination globale (ROC/AUC) qu'en performance de classification (erreur,