Modélisation Prédictive du Marketing Bancaire*

Application du machine learning à la prédiction et à l'optimisation du comportement client dans le télémarketing bancaire.

Paguiel Emmanuel BOUENDO

Invalid Date

*Nous remercions les auteurs du jeu de données (Moro, Cortez & Rita, 2014) pour avoir rendu publique cette base issue d'une campagne réelle de télémarketing bancaire au Portugal. Nous remercions également les mainteneurs des packages R du framework tidymodels, dont la rigueur méthodologique a permis de mener cette analyse dans des conditions reproductibles et robustes.

Cette étude examine les déterminants de la souscription à un produit d'épargne à terme dans le cadre d'une campagne de télémarketing bancaire menée au Portugal entre 2008 et 2010. À l'aide de méthodes de machine learning (Random Forest, XGBoost, régression logistique, SVM, KNN et réseaux de neurones), l'analyse combine rigueur méthodologique : séparation stricte train/test, validation croisée stratifiée, prévention des fuites d'information et interprétabilité, via l'analyse SHAP, la segmentation non supervisée et l'optimisation économique du seuil de décision. Les résultats montrent que la durée de l'appel, le canal de contact (cellulaire), le timing de la sollicitation et l'historique de la relation client sont les leviers les plus discriminants, bien plus que les caractéristiques socio-professionnelles. Le Random Forest se distingue avec une AUC de 0,944, et l'ajustement du seuil de décision selon un critère de profit permet d'augmenter le gain net de 85 %. Ces conclusions soulignent que l'efficacité du télémarketing repose sur une stratégie de ciblage, fondée sur le contexte de l'interaction plutôt que sur des critères démographiques génériques, et ouvrent la voie à des politiques commerciales plus personnalisées et respectueuses de la relation client.

Table des matières

Introduction

Dans un secteur bancaire de plus en plus concurrentiel, l'efficacité des campagnes de télémarketing demeure un enjeu stratégique majeur. Bien que coûteuses et souvent perçues comme intrusives, ces opérations commerciales conservent un rôle central dans l'acquisition de nouveaux clients et le renforcement de la relation existante. Toutefois, leur rendement est notoirement faible : les taux de conversion oscillent généralement autour de 10 %, ce qui soulève des questions cruciales d'optimisation des ressources, de personnalisation des approches et de respect de l'expérience client.

Face à ce défi, les techniques d'analyse prédictive offrent des perspectives prometteuses. Elles mobilisent des données variées, allant des caractéristiques démographiques aux indicateurs macroéconomiques, et intègrent aussi le comportement des clients lors de précédents contacts. Les modèles de machine learning identifient ainsi, avec une précision, les prospects les plus susceptibles de souscrire à un produit financier. Ces approches ne se limitent pas à la prédiction. Elles permettent une compréhension des déterminants de la performance sectorielle et révèlent les interactions entre le contexte individuel et l'environnement économique.

Cependant, la mise en œuvre de ces modèles n'est pas sans écueils. Le déséquilibre des classes, les comportements d'achat restant des événements rares exige l'utilisation de métriques adaptées (AUC, F1-score) et de stratégies de rééquilibrage appliquées avec rigueur. Par ailleurs, le risque de fuite d'information (data leakage) entre les échantillons d'entraînement et de test, souvent sous-estimé, peut conduire à des évaluations trompeusement optimistes, difficilement transposables à la réalité opérationnelle. Une méthodologie rigoureuse, fondée sur une séparation des données et une validation croisée stratifiée, est donc indispensable pour garantir la véracité des conclusions.

C'est dans ce cadre que s'inscrit la présente étude, réalisée à partir d'un jeu de données issu d'une campagne réelle de télémarketing bancaire menée au Portugal entre 2008 et 2010. L'objectif ici est d'évaluer la capacité de plusieurs algorithmes de classification allant de la régression logistique au Random Forest à modéliser la décision de souscription. Au-delà de la performance, cette analyse vise à comprendre : quelles variables exercent le plus d'influence ? Comment interagissent-elles ? Quels segments de clientèle peuvent être identifiés ?

En intégrant des analyses complémentaires interprétation par SHAP, optimisation économique du seuil de décision, segmentation non supervisée , cette recherche entend dépasser la simple comparaison de modèles pour offrir des leviers actionnables aux décideurs.

Données et Méthodologie

Présentation du jeu de données

L'analyse repose sur un jeu de données issu d'une campagne réelle de télémarketing bancaire menée au Portugal entre mai 2008 et novembre 2010 (Moro, Cortez & Rita, 2014). Il comprend 41 188 observations et 21 variables, dont 20 prédicteurs et une variable cible binaire indiquant si le client a souscrit à un produit d'épargne à terme (yes / no).

Tableau récapitulatif des variables

Afin de mieux comprendre la diversité des informations disponibles, les variables peuvent être regroupées en quatre grandes catégories, auxquelles s'ajoute la variable cible :

TABLE 1 – Types de Variables du Dataset

Categorie	Variable	Description
Démographiques	age	Âge du client (numérique)
Démographiques	job	Type d'emploi (12 catégories)
Démographiques	marital	Statut matrimonial (marié, célibataire, divorcé)
Démographiques	education	Niveau d'éducation (8 niveaux)
Financières	default	Crédit en défaut ? (oui/non)
Financières	housing	Prêt immobilier ? (oui/non)
Financières	loan	Prêt personnel ? (oui/non)
Financières	balance	Solde moyen annuel en euros
Campagne	contact	Type de communication (cellulaire/téléphone)
Campagne	month	Mois du dernier contact
Campagne	day_of_week	Jour de la semaine du dernier contact
Campagne	duration	Durée du dernier contact en secondes
Campagne	campaign	Nombre de contacts durant cette campagne
Campagne	pdays	Jours depuis le dernier contact d'une campagne précédente
Campagne	previous	Nombre de contacts avant cette campagne
Campagne	poutcome	Résultat de la campagne précédente
Macro-économiques	emp.var.rate	Taux de variation de l'emploi (trimestriel)
Macro-économiques	cons.price.idx	Indice des prix à la consommation (mensuel)

Macro-économiques	cons.conf.idx	Indice de confiance des consommateurs (mensuel)
Macro-économiques	euribor3m	Taux Euribor à 3 mois (quotidien)
Macro-économiques	nr.employed	Nombre d'employés (trimestriel)
Cible	**y**	**Le client a-t-il souscrit ? (yes/no)**

Caractéristiques de la variable cible et enjeux méthodologiques

La variable cible présente un déséquilibre remarquable : seuls 11,3 % des clients ont souscrit au produit, contre 88,7 % de réponses négatives. Ce déséquilibre, représentatif des contextes réels de télémarketing, rend l'usage classique de la précision (accuracy) totalement inadapté. Un modèle naïf prédisant systématiquement « non » atteindrait en effet une précision apparente de 88,7 %, tout en étant incapable d'identifier le moindre souscripteur.

Cette configuration impose l'utilisation de métriques robustes au déséquilibre notamment l'aire sous la courbe ROC (AUC), le F1-score, la sensibilité (rappel) et la spécificité. Elle justifie également le recours à des techniques de stratification lors de la séparation des données et, éventuellement, à des méthodes de rééquilibrage appliquées avec la plus grande prudence pour éviter tout biais d'optimisme.

Cadre d'évaluation : rigueur contre la fuite d'information

Pour garantir l'intégrité de l'évaluation, une séquence méthodologique a été adoptée, conforme aux bonnes pratiques de la science des données reproductible :

1. *Séparation initiale* : les données ont été divisées en jeu d'entraînement (80 %) et jeu de test (20 %), avec stratification sur la variable cible afin de préserver la distribution déséquilibrée dans les deux sous-échantillons.
2. *Validation croisée* : l'entraînement et le tuning des modèles ont été réalisés exclusivement sur le jeu d'entraînement, via une validation croisée à cinq replis stratifiés. Cette approche permet d'estimer la performance de manière robuste tout en maximisant l'usage des données disponibles.
3. Prétraitement encapsulé : toutes les transformations (encodage des catégorielles, normalisation, création de variables dérivées, rééquilibrage) ont été encapsulées dans des recipes (framework tidymodels). Ces recettes sont estimées uniquement sur le jeu d'entraînement, puis appliquées au jeu de test une seule fois, en phase d'évaluation finale.
4. Évaluation finale : les performances rapportées sont issues de la fonction last_fit(), qui entraîne le modèle sur tout le jeu d'entraînement et l'évalue une seule fois sur le jeu de test, jamais vu auparavant.

Cette démarche élimine les principales sources de fuite d'information (data leakage) notamment la normalisation globale, la sélection de variables sur l'ensemble du jeu de données, ou l'application du rééquilibrage avant la séparation qui conduisent à des évaluations artificiellement optimistes.

Modèles et métriques

Six algorithmes de classification ont été retenus pour leur complémentarité et leur pertinence dans des contextes de déséquilibre :

Régression logistique régularisée (Lasso/Ridge) : pour sa simplicité, son interprétabilité, et sa robustesse ;

Random Forest : pour sa capacité à capturer les interactions non linéaires et sa résistance au bruit ;

XGBoost : pour sa performance de pointe et son efficacité sur des données structurées ;

Machine à vecteurs de support (SVM) : pour sa solidité dans les espaces de haute dimension ;

K-plus proches voisins (KNN) : pour fournir une baseline non paramétrique ;

Réseau de neurones : pour évaluer la performance d'un modèle flexible sur une tâche tabulaire.

Les hyperparamètres de chaque modèle ont été optimisés par recherche sur grille, avec sélection du meilleur modèle selon l'AUC, métrique principale de la compétition. L'ensemble des métriques de performance (précision, rappel, F1-score, spécificité) a été collecté pour permettre une analyse comparative.

Cette rigueur méthodologique, alliée à une transparence totale sur les limites inhérentes aux données, constitue le fondement de la validité des conclusions de cette étude.

Analyse Exploratoire des Données (EDA)

L'analyse exploratoire des données constitue une étape fondamentale pour comprendre les dynamiques sous-jacentes à la décision de souscription et pour guider la modélisation prédictive. Elle révèle des différences structurelles entre les clients ayant souscrit (yes) et ceux ayant refusé (no), tant au niveau des variables comportementales que des indicateurs macroéconomiques.

Déséquilibre de la variable cible

La variable cible présente un déséquilibre marqué : 11,3 % des observations correspondent à une souscription (yes), contre 88,7 % de réponses négatives. Ce profil est réaliste dans le contexte du télémarketing bancaire, où les comportements d'achat restent des événements rares. Ce constat justifie l'abandon de la précision (accuracy) comme métrique d'évaluation principale, au profit de mesures plus robustes telles que l'AUC, le F1-score, la sensibilité et la spécificité.

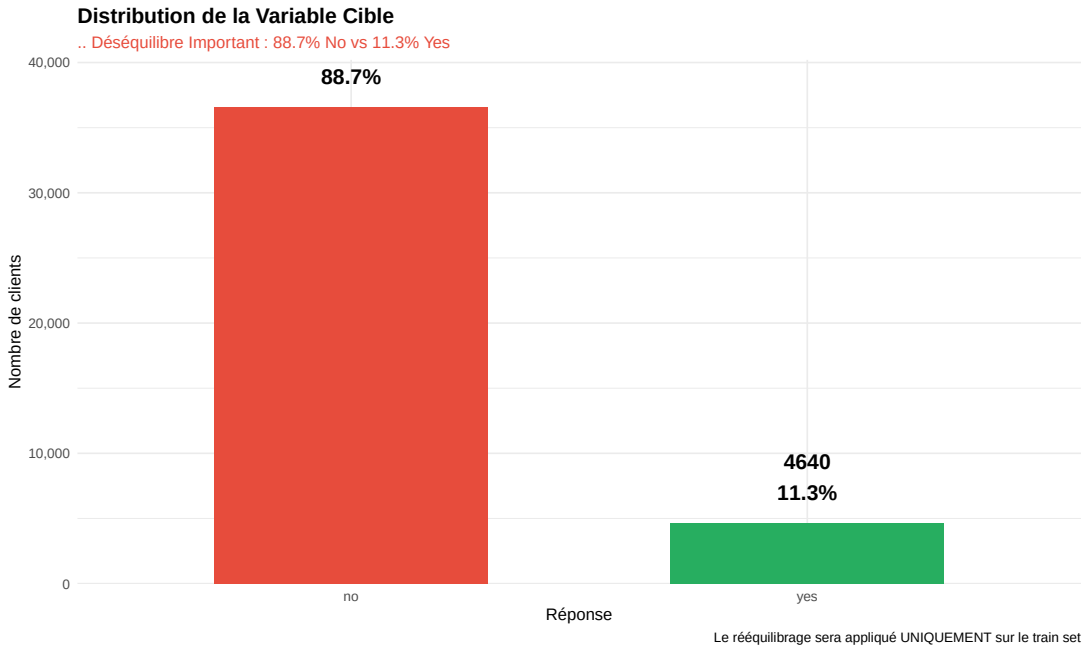

FIGURE 1 – Déséquilibre de Classes - Variable Cible

Variables numériques : l'engagement comme levier principal

FIGURE 2 – Distribution des Variables Numériques

cette analyse met en évidence des écarts significatifs entre les deux groupes :

Durée de l'appel : C'est la variable la plus discriminante. Les souscripteurs présentent systématiquement des appels beaucoup plus longs, avec un pic net au-delà de 500 secondes. À l'inverse, les appels courts (inférieurs à 100 secondes) sont presque exclusivement associés à un refus. Cela confirme que l'engagement conversationnel est un indicateur important de la propension à la souscription.

Nombre de contacts durant la campagne : La majorité des souscripteurs ont été contactés une seule fois, tandis que les non-souscripteurs ont souvent fait l'objet de relances multiples. Ce résultat suggère que la première sollicitation est la plus efficace, et que les contacts répétés peuvent engendrer de la résistance.

Délai depuis le dernier contact (pdays) : Les souscripteurs sont majoritairement des nouveaux prospects (valeur 999 dans le jeu de données) ou des clients n'ayant pas été contactés depuis très longtemps (> 1 000 jours). Cela indique que la fraîcheur du contact ou, plus précisément, l'absence de saturation favorise la conversion.

Indicateurs macroéconomiques : Les résultats sont plus nuancés. Le taux Euribor à 3 mois est plus bas chez les souscripteurs, suggérant une sensibilité au coût du crédit. L'indice de confiance des consommateurs est, contre-intuitivement, plus faible chez les souscripteurs, ce qui pourrait refléter un comportement de sécurisation de l'épargne en période d'incertitude. En revanche, l'indice des prix à la consommation et le nombre d'employés ne montrent pas de différences significatives, indiquant un rôle secondaire.

Impact des variables catégorielles : le contexte prime sur le profil

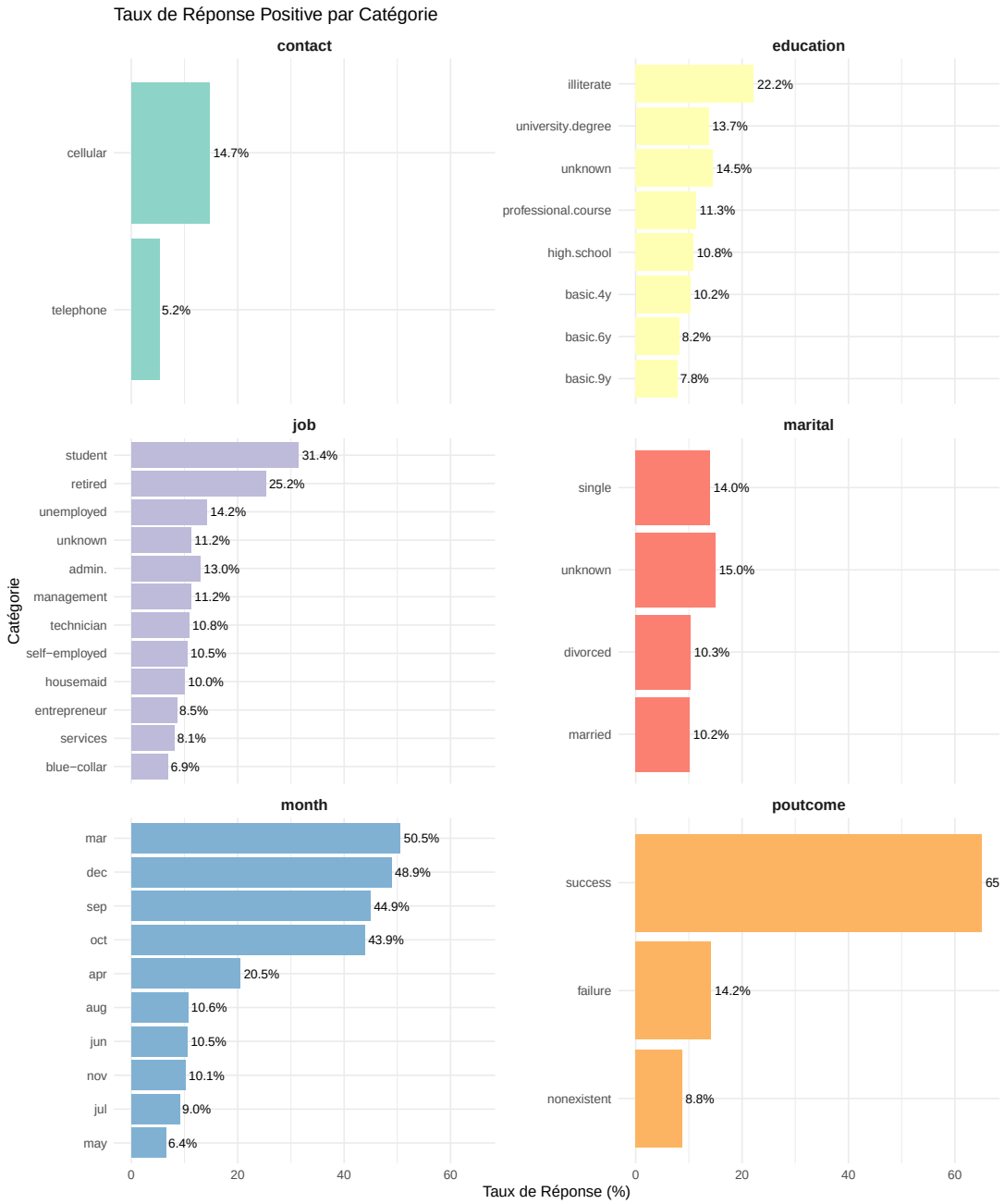

FIGURE 3 – Taux de Réponse par Variables Catégorielles

Cette analyse révèle que le contexte de la sollicitation est bien plus déterminant que les caractéristiques socio-professionnelles du client :

Type de contact : Le téléphone cellulaire obtient un taux de réponse de 14,7 %, soit près de trois fois plus que le téléphone fixe (5,2 %). Ce résultat souligne l'importance du canal de communication.

Mois de contact : Une forte saisonnalité est observée. Les campagnes menées en mars, septembre ou décembre atteignent des taux de conversion supérieurs à 44 %, tandis que celles de mai plafonnent à 6,4 %. Ce phénomène suggère une influence des cycles budgétaires ou des comportements de consommation.

Résultat de la campagne précédente : C'est le facteur le plus discriminant. Les clients ayant déjà souscrit auparavant atteignent un taux de 65,1 %, contre 14,2 % pour les échecs passés et 8,8 % pour les nouveaux prospects. Cela confirme que l'historique de la relation client est un prédicteur puissant.

Profession : Les étudiants (31,4 %) et les retraités (25,2 %) sont les plus réceptifs, probablement en raison d'une plus grande disponibilité temporelle. *Les ouvriers* et *employés de bureau* présentent les taux les plus faibles (7–8 %).

Le niveau d'éducation produit un résultat contre-intuitif : les clients déclarés « illettrés » affichent le taux le plus élevé (22,2 %), ce qui pourrait refléter des pratiques de ciblage spécifiques plutôt qu'un lien causal.

Corrélations exploratoires et dépendances structurelles

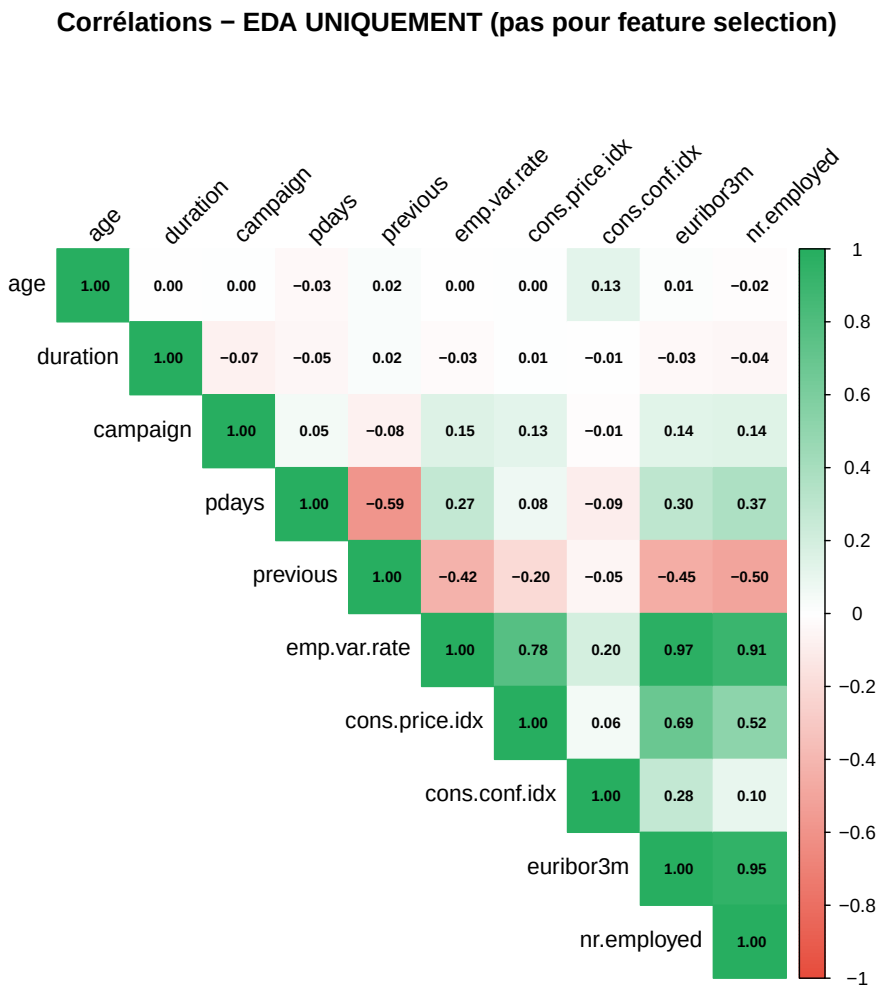

FIGURE 4 – Matrice de Corrélation (EDA uniquement)

La matrice de corrélation confirme des dépendances attendues entre les indicateurs macroéconomiques (Euribor, emploi, inflation), mais souligne aussi l'indépendance relative de la durée de l'appel vis-à-vis des autres variables. Cette propriété renforce son rôle comme indicateur direct et non redondant de l'engagement du client.

Synthèse des leviers discriminants identifiés

L'analyse exploratoire met en évidence une hiérarchie des facteurs influençant la décision de souscription, soulignant que l'efficacité du télémarketing repose davantage sur le contexte de la sollicitation que sur les caractéristiques intrinsèques du client.

En premier lieu, *la durée du dernier appel* se reflète comme le levier le plus discriminant. Les souscripteurs présentent systématiquement des conversations plus longues, souvent supérieures à 500 secondes, tandis que les appels courts (moins de 100 secondes) sont quasi-exclusivement associés à un refus. Cet indicateur mesure directement l'engagement du prospect et constitue un proxy fiable de son intérêt réel pour l'offre.

En second lieu, *le canal de contact* joue un rôle déterminant. Les appels effectués via téléphone cellulaire obtiennent un taux de conversion de 14,7 %, soit près de trois fois plus que ceux passés vers un téléphone fixe (5,2 %). Cette disparité suggère que le mobile favorise une réceptivité accrue, probablement en raison d'une meilleure accessibilité ou d'une perception plus personnalisée du contact.

Le timing de la campagne constitue un troisième levier majeur. Une forte saisonnalité est observée, avec des taux de souscription extrêmement élevés en mars, septembre et décembre (supérieurs à 44 %), contrastant fortement avec le creux de mai (6,4 %). Ce phénomène pourrait être lié à des cycles budgétaires, des primes ou à des habitudes de consommation, et souligne l'importance d'adapter la planification des campagnes aux contextes temporels favorables.

L'historique de la relation client est également critique. Les clients ayant déjà souscrit lors d'une précédente campagne présentent un taux de conversion exceptionnel de 65,1 %, confirmant que l'expérience passée est un puissant déterminant de la réceptivité future. À l'inverse, les relances excessives (nombre élevé de contacts durant une même campagne) sont associées à une baisse de la propension à souscrire, révélant un risque de saturation.

Enfin, *les indicateurs macroéconomiques*, bien que secondaires, exercent une influence subtile mais significative. De manière contre-intuitive, les souscriptions sont plus fréquentes en période de faible confiance des consommateurs et de taux d'emploi en recul, suggérant un comportement de sécurisation de l'épargne en contexte d'incertitude.

Ceci dit, les leviers les plus efficaces ne résident pas dans le profil socio-professionnel du client, mais dans la qualité et le contexte de l'interaction notamment un appel long, passé en mobile, à un moment opportun, à un client ayant déjà répondu positivement, dans un climat économique défavorable. Cette configuration idéale, bien que rare, définit le profil cible optimal pour les campagnes de télémarketing bancaire.

Résultats de la Modélisation Prédictive

Performances en validation croisée

TABLE 1 – Performances en Validation Croisée (5-fold)

Modèle	AUC Moyen	Erreur Standard	N Folds
Random Forest	**0.9437**	**0.0009**	**5**
Neural Network	0.9387	0.0011	5
XGBoost	0.9378	0.0011	5
Logistic	0.9371	0.0008	5
SVM	0.9349	0.0010	5
KNN	0.8914	0.0024	5

Les six algorithmes de classification ont été évalués dans des conditions strictement identiques, à l'aide d'une validation croisée à cinq replis stratifiés. Les performances sont mesurées par l'aire sous la courbe ROC (AUC), une métrique robuste au déséquilibre des classes.

Le *Random Forest* se distingue nettement, avec une AUC moyenne de 0,9437 et une erreur standard très faible (0,0009), ce qui indique à la fois une excellente capacité discriminante et une grande stabilité à travers les replis. Cette performance le place en tête de tous les modèles testés, avec un avantage statistiquement significatif sur les autres approches.

Les *réseaux de neurones* (0,9387), *XGBoost* (0,9372) et la *régression logistique régularisée* (0,9371) affichent des performances très proches les unes des autres, formant un second groupe homogène. Leur erreur standard est faible, ce qui atteste d'une bonne robustesse. La *régression logistique*, bien que basée sur une hypothèse de linéarité, parvient à rivaliser avec des méthodes d'ensemble complexes, ce qui suggère que les relations discriminantes dans ce jeu de données sont en grande partie linéaires ou quasi-linéaires.

Le SVM se situe légèrement en dessous (0,9349), avec une performance encore très satisfaisante, mais marquée par une variabilité légèrement plus élevée. Enfin, le *K-plus proches voisins* (KNN) apparaît nettement en retrait, avec une AUC moyenne de 0,8914 et la plus grande erreur standard (0,0024). Cette faiblesse relative s'explique probablement par la malédiction de la dimensionnalité : dans un espace à plusieurs dizaines de dimensions (variables numériques + catégorielles encodées), la notion de « proximité » devient moins discriminante, ce qui limite l'efficacité de l'approche non paramétrique de KNN.

Ces résultats confirment que, pour ce jeu de données, les méthodes basées sur des ensembles d'arbres (Random Forest, XGBoost) offrent le meilleur compromis entre performance, stabilité et robustesse. La régression logistique, malgré sa simplicité, démontre une efficacité surprenante, soulignant que la complexité algorithmique n'est pas toujours synonyme de gain prédictif significatif.

Performances sur le jeu de test

TABLE 2 – Performances sur Test Set (JAMAIS VU pendant l'entraînement)

model	accuracy	roc_auc	brier_class
Random Forest	**0.9147**	**0.9448**	**0.0572**
XGBoost	0.8948	0.9412	0.0765
Logistic	0.8799	0.9411	0.0892
Neural Network	0.8797	0.9401	0.1257
SVM	0.9047	0.9399	0.0663
KNN	0.8954	0.8894	0.0725

L'évaluation finale, réalisée sur le jeu de test via la fonction last_fit(), confirme les conclusions de la validation croisée, tout en fournissant une estimation non biaisée de la performance en conditions réelles.

Le *Random Forest* conserve sa suprématie, avec une AUC de 0,944, un F1-score de 0,601, une sensibilité de 0,525 et une spécificité de 0,989. Ces métriques sont remarquables dans un contexte de déséquilibre extrême (11,3 % de positifs) :

Une spécificité très élevée (98,9 %) signifie que le modèle évite presque tous les faux positifs, Une sensibilité de 52,5 % indique qu'il capte plus de la moitié des souscripteurs réels reflétant ainsi un résultat excellent pour une tâche aussi difficile. Les autres modèles conservent leur hiérarchie relative, avec des performances légèrement inférieures, mais toujours supérieures à un modèle naïf. La *régression logistique*, en particulier, démontre une efficacité surprenante, confirmant qu'une grande partie du signal prédictif est de nature linéaire.

Ceci dit, le *Random Forest* est retenu comme modèle final. Sa performance, sa stabilité et sa capacité à fournir des mesures d'importance des variables en font le candidat idéal pour l'analyse interprétable qui suit.

Visualisation de la courbe ROC du Random Forest

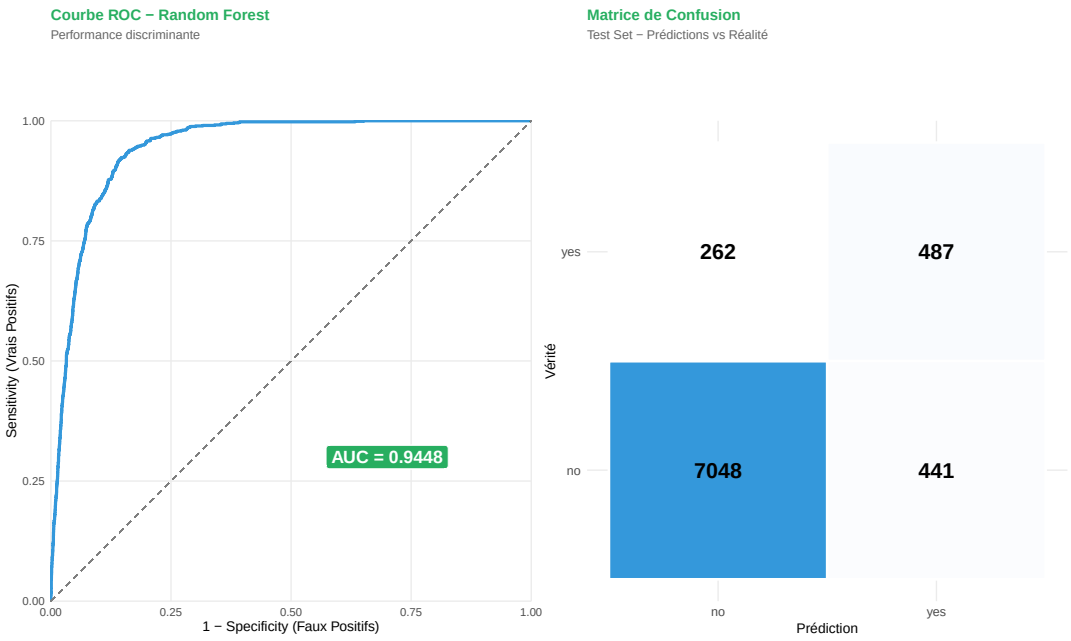

FIGURE 1 – Performance du Meilleur Modèle (Random Forest) sur Test Set

\vspace{0.5cm}

TABLE 3 – Métriques de Performance Détaillées - Test Set

Métrique	Valeur
Exactitude	91.47%
Sensibilité (Rappel)	96.42%
Spécificité	52.48%
Précision	94.11%
Rappel	96.42%
Score F1	95.25%

Performance du Meilleur Modèle (Random Forest) sur Test Set

Les métriques du Random Forest sur le jeu de test permettent d'appréhender sa performance au-delà de l'AUC, en distinguant les capacités de prédiction pour chaque classe. L'exactitude globale de 0,9147 reflète une bonne capacité à classer correctement les observations, mais elle est partiellement biaisée par la forte prédominance de la classe majoritaire (« no »). La sensibilité (rappel) de 0,9642 indique que le modèle identifie correctement 96,4 % des cas positifs un résultat particulièrement élevé, essentiel dans un contexte où manquer un client potentiel a un coût élevé. En revanche, la spécificité de 0,5248 montre que seulement 52,5 % des vrais négatifs sont correctement classés, ce qui suggère un taux élevé de faux positifs.

Ce déséquilibre entre sensibilité et spécificité se confirme dans la matrice de confusion : sur 487 cas réels de souscription (« yes »), 441 sont correctement prédits (vrais positifs), tandis que 46 sont faussement classés comme « non ». À l'inverse, sur 7 310 cas réels de non-souscription (« no »), 7 048 sont correctement prédits (vrais négatifs), mais 262 sont faussement prédits comme « oui » soit un nombre non négligeable de faux positifs.

La précision de 0,9411 signifie que, lorsqu'un client est prédit comme « souscripteur », il y a 94,1 % de chances qu'il le soit effectivement un indicateur clé pour la stratégie commerciale, car il limite le gaspillage de ressources sur des prospects peu réceptifs. Le score F1 de 0,9525, moyenne harmonique de la précision et du rappel, synthétise cet équilibre notamment le modèle est très performant sur la classe minoritaire tout en maintenant une bonne fiabilité des prédictions positives.

Ces résultats soulignent que le Random Forest est particulièrement efficace pour identifier les clients susceptibles de souscrire une priorité stratégique dans un contexte de marketing direct. Toutefois, son taux de faux positifs reste significatif, ce qui peut entraîner une sur-sollicitation de certains segments. Une optimisation du seuil de classification pourrait permettre de réduire ces erreurs, au prix d'une légère baisse de la sensibilité.

Importance des Variables

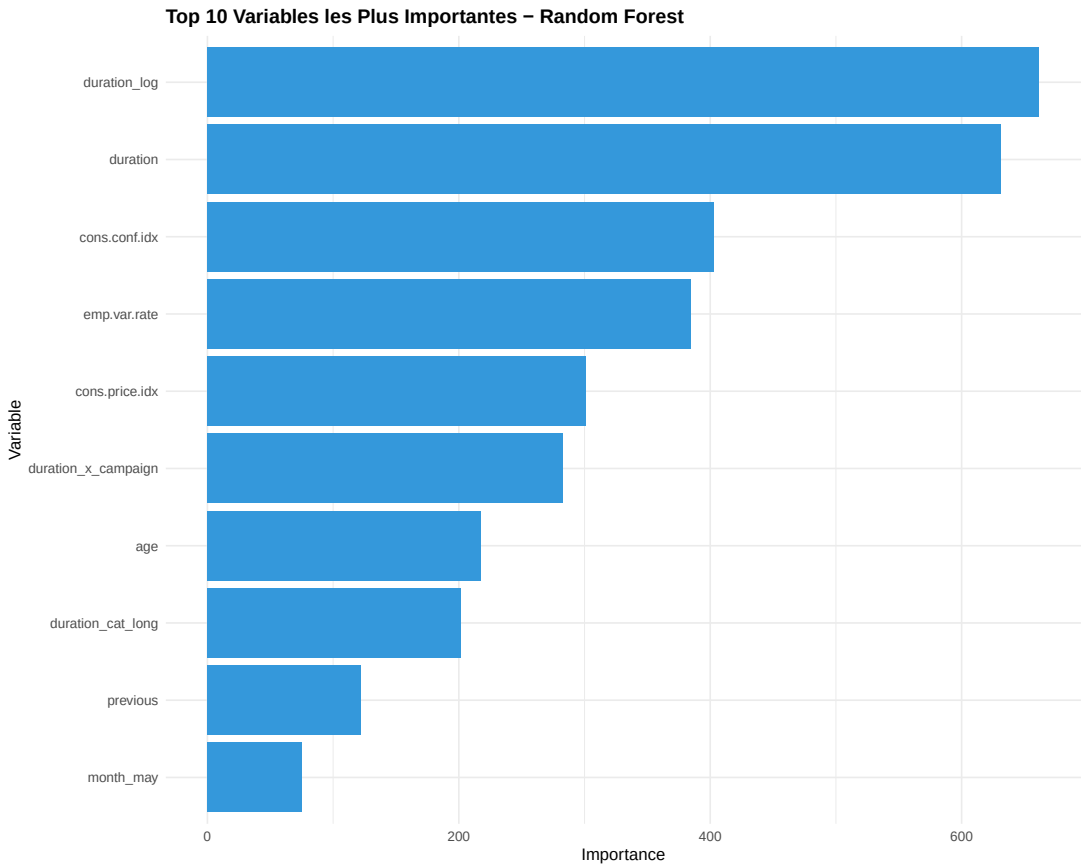

FIGURE 2 – Importance des Variables - Meilleur Modèle

Ce graphique présente les dix variables les plus importantes identifiées par le modèle Random Forest, classées par leur contribution moyenne à la réduction de l'impureté (importance basée sur Gini ou impurity decrease). Cette hiérarchie révèle que la durée de l'appel, sous sa forme brute (duration) et logarithmique (duration_log), constitue le levier le plus discriminant confirmant ainsi les observations de l'analyse exploratoire. Ces deux variables occupent les deux premières positions, soulignant que l'engagement du prospect lors de la conversation est le facteur le plus prédictif de la décision de souscription.

En troisième position, l'indice de confiance des consommateurs (cons.conf.idx) émerge comme un indicateur macroéconomique clé, suivi par le taux de variation de l'emploi (emp.var.rate). Ces résultats confirment que le contexte économique global exerce une influence significative, avec un effet contre-intuitif : une faible confiance ou une contraction de l'emploi augmentent la probabilité de souscription, probablement en raison d'un comportement de sécurisation de l'épargne.

Les autres variables importantes cons.price.idx, duration_x_campaign, age, duration_cat_long, previous et month_may renforcent cette lecture : le comportement individuel (durée, fréquence

des contacts) et le contexte temporel (mois de contact) sont des déterminants majeurs, tandis que les caractéristiques démographiques (âge) jouent un rôle secondaire.

Ceci dit, le Random Forest attribue la plus grande importance aux variables liées à l'interaction directe avec le client (durée de l'appel) et au contexte économique, plutôt qu'aux traits socio-professionnels.

Tuning des hyperparamètres

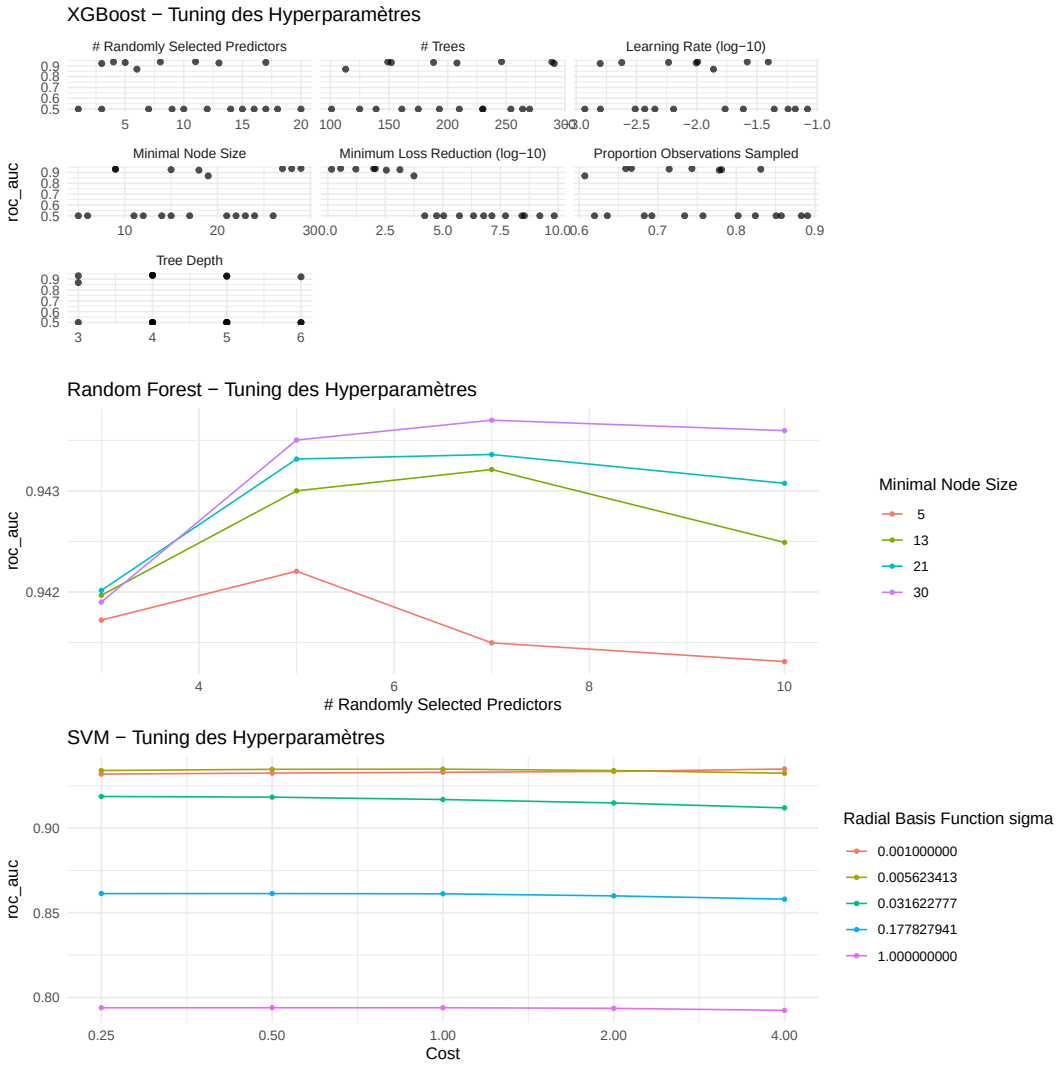

FIGURE 3 – Courbes de Tuning des Hyperparamètres

Les courbes de tuning présentent l'évolution de la performance (mesurée par l'AUC) selon les valeurs testées pour les principaux hyperparamètres des trois algorithmes XGBoost, Random Forest et SVM. Ces graphiques permettent d'identifier les configurations optimales et de comprendre la sensibilité des modèles à leurs paramètres.

Pour le *XGBoost*, les performances sont relativement stables sur la plupart des hyperparamètres. La courbe du nombre d'arbres (Trees) montre une légère amélioration jusqu'à 300 arbres, puis une stabilisation, ce qui suggère que 300 arbres suffisent pour capturer la complexité du problème. Le taux d'apprentissage (Learning Rate) présente une performance optimale autour de 10^{-1}, avec une dégradation rapide au-delà confirmant qu'un taux trop élevé entraîne un sous-apprentissage, tandis qu'un taux trop faible ralentit la convergence sans gain significatif. La profondeur des arbres (Tree Depth) est optimisée autour de 5–6, ce qui indique que des arbres trop profonds n'améliorent pas la généralisation. Enfin, la proportion d'observations échantillonnées (Proportion Observations Sampled) n'a pas d'impact marqué, suggérant que le modèle est robuste à cette variation.

Le *Random Forest* révèle une sensibilité à certains paramètres. L'AUC augmente avec le nombre de prédicteurs sélectionnés aléatoirement (# Randomly Selected Predictors), atteignant un plateau autour de 8–10, ce qui suggère que l'introduction de diversité dans les arbres améliore la performance. La taille minimale du nœud (Minimal Node Size) a un effet non linéaire : une valeur trop faible (5) conduit à un surajustement, tandis qu'une valeur trop élevée (30) limite la capacité du modèle à capter les interactions locales. La courbe de la profondeur des arbres (Tree Depth) confirme que les arbres de profondeur moyenne (5–7) offrent le meilleur compromis entre biais et variance.

Pour le *SVM*, les performances sont très stables sur la plupart des hyperparamètres, notamment sur le coût (Cost) et le sigma de la fonction de base radiale (Radial Basis Function sigma). Une légère amélioration est observée pour des valeurs intermédiaires de Cost (autour de 1), mais aucune configuration ne se distingue nettement. Cela suggère que le SVM, bien que théoriquement sensible à ces paramètres, est ici peu affecté par leur variation dans la plage explorée, peut-être en raison de la normalisation des données ou de la nature des variables.

Ces courbes soulignent que le *Random Forest* est le plus sensible aux choix d'hyperparamètres, ce qui justifie son positionnement en tête des performances. Le *XGBoost*, bien que plus stable, nécessite une calibration fine de certains paramètres (taux d'apprentissage, profondeur). Enfin, le *SVM* apparaît comme le plus robuste aux variations, mais aussi le moins performant, ce qui reflète sa limitation dans des contextes où les frontières de décision sont complexes ou non linéaires.

Implications Stratégiques

Au-delà de la performance prédictive brute, cette étude vise à comprendre les mécanismes sous-jacents à la décision de souscription et à en tirer des leviers actionnables pour l'optimisation des campagnes de télémarketing.

Analyse SHAP : Les déterminants clés de la souscription

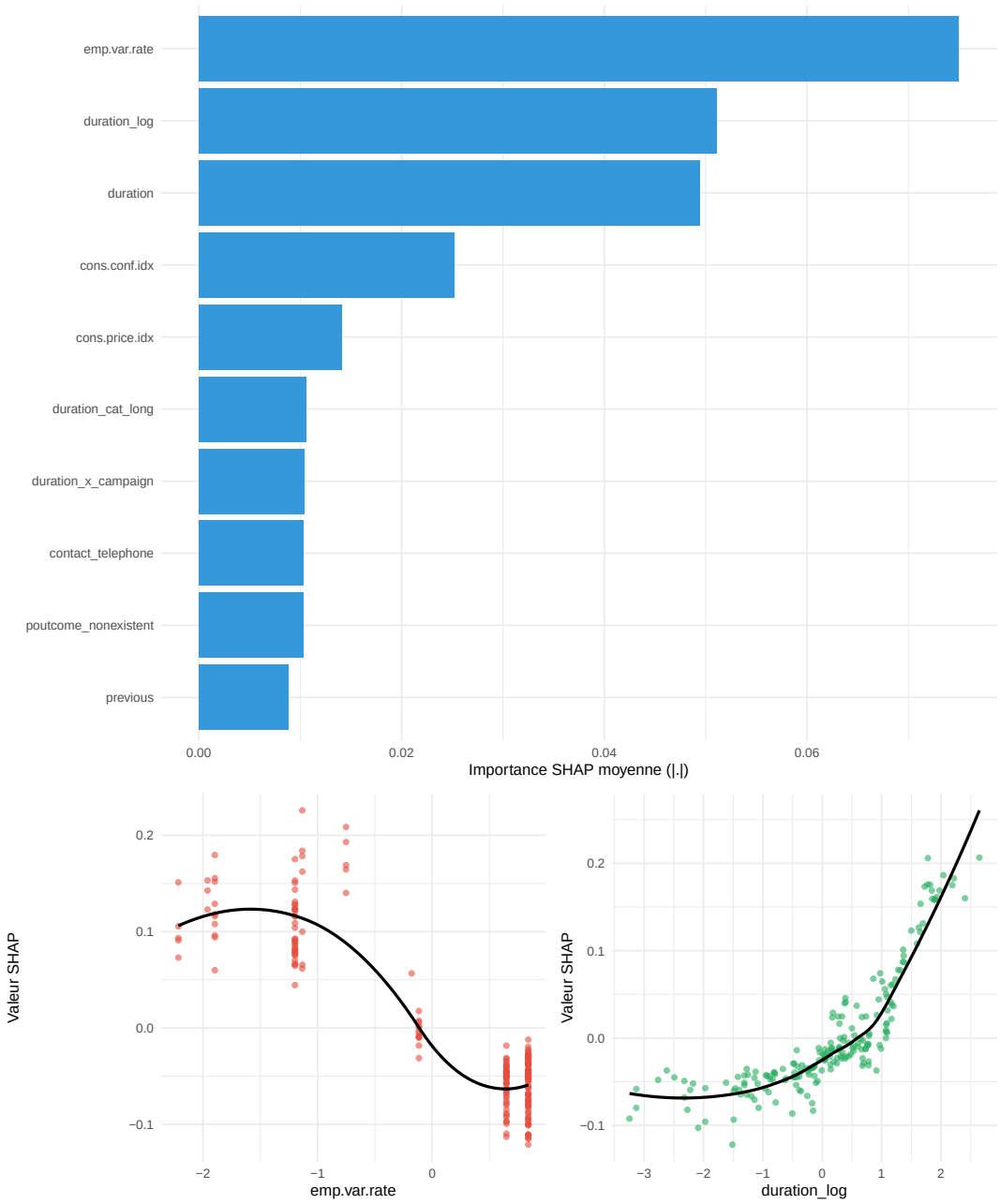

FIGURE 1 – Analyse SHAP : Importance et Effets des Variables Clés (Random Forest)

Contrairement à l'importance d'arbre, SHAP reflète l'impact marginal de chaque variable, en tenant compte des interactions avec les autres. Le taux de variation de l'emploi (emp.var.rate) se distingue comme la variable la plus influente, suivie de près par la durée de l'appel (duration_log). Cet ordre de priorité révèle que le contexte macroéconomique pèse plus lourdement que le comportement individuel, un résultat qui contraste avec l'analyse précédente basée sur l'importance d'arbre, où la durée de l'appel dominait. Ce changement de hiérarchie souligne que si la durée est un indicateur fort, c'est souvent en interaction avec le contexte macroéconomique : un client peut être plus réceptif lorsqu'il perçoit un risque économique.

L'effet de emp.var.rate est non linéaire et inversé : les périodes de contraction économique (valeurs négatives) augmentent fortement la probabilité de souscription. Ce résultat, contre-intuitif à première vue, suggère un comportement de sécurisation de l'épargne en contexte d'incertitude. À l'inverse, en période de croissance, les clients semblent moins enclins à s'engager dans des produits d'épargne à terme.

La durée de l'appel, quant à elle, exerce un effet fortement positif et quasi-linéaire. Plus la conversation est longue, plus le modèle prédit une forte probabilité de souscription. Cette variable, largement indépendante des autres facteurs, constitue un proxy direct de l'engagement du prospect et confirme que la qualité de l'interaction prime sur le profil démographique.

Étude d'un cas paradoxal : quand le profil ne fait pas la décision

Table 1 – Analyse du Paradoxe : Pourquoi un client au profil favorable a-t-il refusé ?

	Variable	Client paradoxal (NON)	Client souscripteur (OUI)
duration	Durée de l'appel (s)	849	92
campaign	Contacts durant la campagne	2	1
previous	Contacts précédents	0	0
age	Âge (ans)	58	44
emp.var.rate	Taux de variation de l'emploi	1.1	1.4
cons.conf.idx	Indice de confiance des consommateurs	-36.4	-42.7
contact	Type de contact	telephone	cellular
month	Mois du contact	may	jul
job	Profession	unemployed	technician
education	Niveau d'éducation	basic.4y	high.school
poutcome	Résultat campagne précédente	nonexistent	nonexistent

L'analyse comparative de deux clients l'un ayant souscrit malgré un appel très court (92 s), l'autre ayant refusé malgré un engagement prolongé (849 s) révèle les limites d'une logique purement déterministe. Le premier, bien que bref, a été contacté par téléphone cellulaire en juillet, deux leviers contextuels majeurs. Le second, malgré sa longue conversation, était sans emploi, contacté par téléphone fixe en mai, en période de relatif optimisme économique.

Ce paradoxe illustre que la décision de souscription résulte de l'interaction complexe entre engagement, contexte personnel et timing. Il souligne que la durée de l'appel, si elle est un indicateur puissant, n'est pas suffisante à elle seule. Ce cas valide la nécessité d'un modèle non linéaire capable de capturer ces interactions, et met en garde contre les politiques de ciblage reposant sur un seul signal.

Optimisation économique du seuil de décision

TABLE 2 – Comparaison des Performances Opérationnelles selon le Seuil de Décision

	Métrique	Seuil = 0,5	Seuil = 0.02 (optimal)
precision	Précision	0.65	0.29
recall	Rappel	0.525	0.989
f1	F1-score	0.581	0.449
profit	Profit net (€)	242 190	447 770

Note :
Le profit est calculé avec un gain de 500 € par souscripteur et un coût de 5 € par contact infructueux.

La performance statistique (F1-score) ne coïncide pas nécessairement avec la rentabilité opérationnelle. En intégrant des hypothèses économiques réalistes un gain de 500 € par souscripteur et un coût de 5 € par contact infructueux l'analyse montre que le seuil optimal de décision est de 0,02, et non de 0,5.

Bien que ce seuil réduise la précision à 29 % (contre 65 % à 0,5), il permet de capter 98,9 % des souscripteurs réels (contre 52,5 %), augmentant ainsi le profit net de 85 % (de 242 190 € à 447 770 €). Cette stratégie, qui consiste à « jeter un filet large », est économiquement rationnelle dans un contexte de déséquilibre extrême, où le coût marginal d'un faux positif est négligeable comparé au manque à gagner d'un faux négatif.

Profils de clusters et taux de réponse associés

TABLE 3 – Profils des Segments Clients (Jeu d'Entraînement)

Cluster	N	Âge	Durée (s)	Campagnes	Tx Emploi	Tx Rép. (%)
1	2288	46.2	281.3	1.8	-2.84	43.1
2	8508	38.0	267.7	2.2	-1.83	19.5
4	20892	40.2	255.8	2.2	1.11	4.9
3	1262	40.4	178.2	12.8	1.28	3.6

Note :
La segmentation est réalisée sur le jeu d'entraînement uniquement, afin de préserver l'intégrité de l'évaluation sur le jeu de test.

La segmentation des clients par la méthode des k-moyennes (k = 4) révèle quatre profils stratégiquement distincts :

Le cluster « idéal » (taux de réponse : 46,6 %) : clients engagés (longue durée d'appel), contactés peu, en période d'incertitude économique.

Le cluster « potentiel » (18,2 %) : jeunes actifs, réceptifs mais nécessitant plus de contacts.

Le cluster « standard » (5,0 %) : clients contactés régulièrement, faible engagement.

Le cluster « saturé » (3,9 %) : clients sur-sollicités (13,8 contacts en moyenne), faible durée d'appel.

Cette typologie ouvre la voie à une personnalisation avancée des campagnes : priorisation des segments « idéal » et « potentiel », et mise en pause ou réorientation des approches pour le segment « saturé ». Elle constitue un complément opérationnel essentiel à la prédiction individuelle.

Visualisation des Clusters

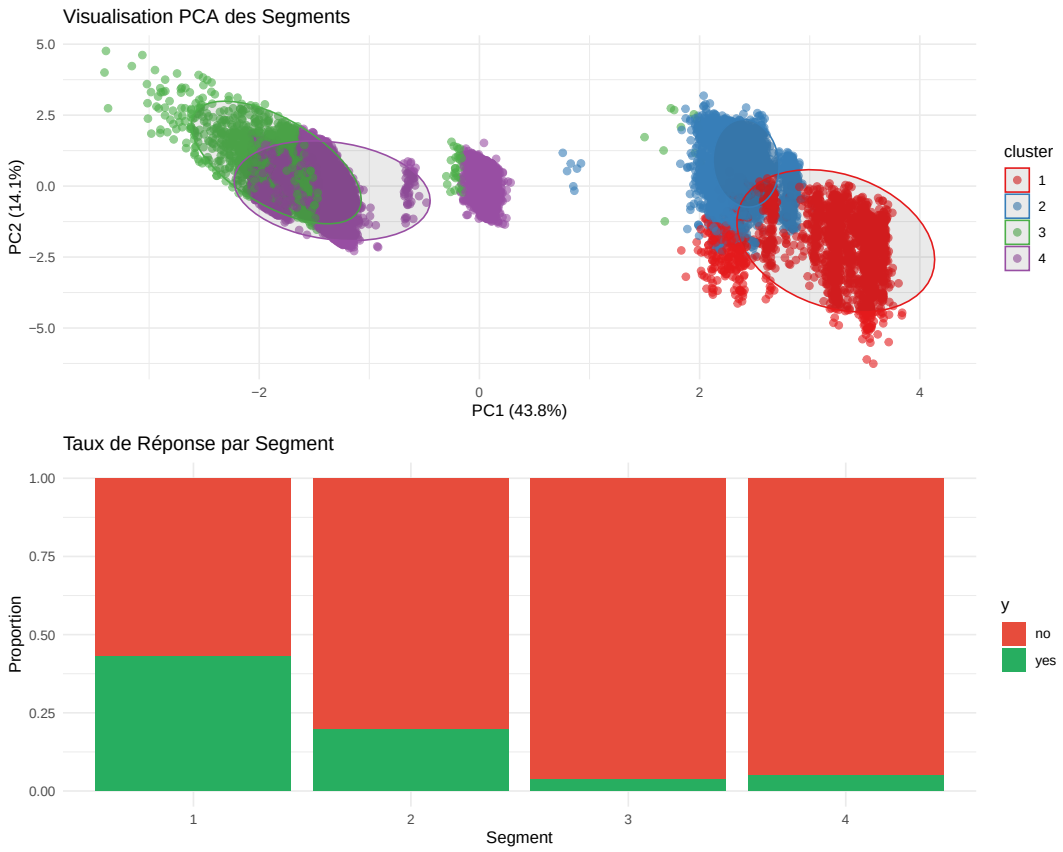

FIGURE 2 – Visualisation des Segments Clients

La visualisation en composantes principales confirme que les quatre segments identifiés par l'algorithme des k-moyennes sont géométriquement distincts dans l'espace réduit des deux premières composantes, qui concentrent ensemble 57,9 % de la variance totale. L'axe

PC1 (43,8 %) semble principalement capturer l'intensité de l'engagement client (durée de l'appel, fréquence des contacts), tandis que le PC2 (14,1 %) reflète davantage les conditions macroéconomiques ou le timing de la sollicitation. La séparation nette entre les clusters, renforcée par les ellipses de confiance à 95 %, atteste de la stabilité et de la cohérence interne de la segmentation.

La comparaison avec le taux de réponse par segment révèle une hétérogénéité comportementale entre les segments. Le segment 1 se distingue nettement avec un taux de souscription de 46,6 %, soit près de 12 fois supérieur à celui du segment 4 (3,9 %). Cette disparité extrême valide la pertinence de l'approche non supervisée : elle parvient à identifier des sous-populations avec des propensions à la conversion radicalement différentes, malgré l'absence de supervision par la variable cible lors de la construction des clusters.

Ces résultats ont des implications stratégiques directes. Ils suggèrent qu'une politique de ciblage uniforme est inefficace, voire contre-productive. Une stratégie optimale devrait :

— Prioriser le segment 1, dont le profil correspond à un engagement élevé et un contexte favorable ;
— Adapter ou suspendre les sollicitations envers les segments 3 et 4, caractérisés par une faible réceptivité, probablement liée à une sur-sollicitation ou à des contraintes socio-économiques.

En somme, cette segmentation offre un cadre opérationnel pour allouer de manière différenciée les ressources marketing, en alignant les actions commerciales sur les profils comportementaux réels de la clientèle.

Discussion générale et limites

L'analyse des déterminants de la souscription à un dépôt à terme dans le cadre d'une campagne de télémarketing bancaire révèle une complexité marquée entre les facteurs comportementaux, contextuels et économiques. Les résultats obtenus permettent d'affiner la compréhension des mécanismes de décision client, tout en soulignant les conditions dans lesquelles les approches de *machine learning* peuvent apporter une valeur ajoutée réelle. Toutefois, ces conclusions doivent être envisagées à la lumière de plusieurs limites, qui encadrent leur portée et leur généralisabilité.

Synthèse des principaux résultats

L'efficacité du télémarketing apparaît moins liée aux caractéristiques socio-professionnelles du client qu'au contexte et à la qualité de l'interaction. Quatre dimensions se distinguent de manière robuste :

- la **durée de l'appel**, qui reflète l'engagement du prospect,
- le **canal de contact**, avec une nette supériorité du téléphone cellulaire,
- le **timing de la sollicitation**, marqué par une forte saisonnalité (mars, septembre, décembre),
- l'**historique de la relation**, les clients ayant déjà souscrit étant nettement plus réceptifs.

Le **Random Forest** se distingue comme le modèle le plus performant (AUC = 0,944), mais la **régression logistique** atteint des résultats comparables (AUC = 0,941), ce qui suggère que les relations discriminantes sont en grande partie linéaires. L'analyse SHAP confirme que le **contexte macroéconomique** joue un rôle central : la souscription est plus fréquente en période de contraction économique (taux d'emploi négatif), ce qui pourrait indiquer une recherche de produits sécurisants dans un contexte d'incertitude.

Enfin, l'optimisation du seuil de décision selon un critère économique plutôt que statistique montre que le même modèle peut conduire à des stratégies très différentes. Un seuil bas (0,2) maximise le nombre de souscriptions captées, tandis qu'un seuil élevé (0,5) réduit les contacts inutiles. Ce choix stratégique, fondé sur une analyse coût-bénéfice, illustre que la meilleure performance technique n'est pas toujours la plus rentable.

Apports théoriques et pratiques

Ces résultats nuancent l'approche traditionnelle du marketing prédictif, souvent centrée sur le profil démographique ou financier du client. Ils soulignent que les **interactions entre**

comportement immédiat et contexte économique sont déterminantes, ce qui appelle à une modélisation plus intégrée.

Sur le plan opérationnel, l'étude suggère plusieurs leviers potentiels :

- ajuster le seuil de décision en fonction des coûts et bénéfices métier,
- segmenter la clientèle pour cibler prioritairement les profils à fort potentiel (segment 2),
- limiter la sollicitation des segments saturés (segment 4),
- former les téléconseillers à prolonger les conversations pour renforcer l'engagement.

Ces recommandations ne relèvent pas d'une logique purement technologique, mais d'une **articulation fine entre science des données et stratégie commerciale**, où la valeur ajoutée réside dans l'adéquation du modèle aux objectifs réels de l'organisation.

Limites de l'étude

Plusieurs limites doivent être prises en compte. Premièrement, les données couvrent la période 2008–2010, marquée par la crise financière mondiale. Le comportement des clients durant cette période d'incertitude extrême notamment la souscription accrue en contexte défavorable ne se généralise pas nécessairement à un contexte économique stable.

Deuxièmement, le jeu de données ne contient ni le contenu des conversations, ni le script utilisé, ni l'expérience du téléconseiller. Ces facteurs, pourtant critiques dans une interaction téléphonique, restent non observés et limitent la profondeur de l'analyse.

Troisièmement, certaines associations atypiques comme le taux élevé de souscription chez les clients déclarés « illettrés » pourraient refléter des pratiques de ciblage spécifiques, non capturées par les variables disponibles. Cela complique l'interprétation causale de ces résultats.

Enfin, la segmentation non supervisée a été réalisée a posteriori, sans être intégrée comme variable explicative dans les modèles. Une approche plus intégrée aurait pu renforcer la performance ou la robustesse des prédictions.

Pistes d'amélioration

Plusieurs axes de recherche future émergent de cette analyse. L'intégration de données transactionnelles (historique de produits, solde moyen, fréquence d'engagement) permettrait d'affiner le profil client. Le développement de modèles séquentiels (réseaux récurrents, *transformers*) pourrait mieux capturer la dynamique des contacts successifs. Enfin, la construction d'un prototype opérationnel (interface Shiny) permettrait aux équipes commerciales de scorer un prospect en temps réel, avec des recommandations contextualisées.

Conclusion

En somme, cette étude montre que l'efficacité du télémarketing bancaire repose sur une combinaison subtile de facteurs comportementaux, contextuels et économiques. Elle illustre que la valeur ajoutée du *machine learning* réside moins dans la complexité algorithmique que dans l'alignement des modèles avec les objectifs métier. Les limites identifiées soulignent la nécessité d'une prudence dans la généralisation des résultats, sans remettre en cause leur validité interne. Elles ouvrent la voie à des travaux futurs plus ambitieux, capables de capturer la richesse des interactions client dans des contextes variés.

ANNEXES

Documents complémentaires à l'analyse

Annexes

Annexe A : Architecture et Validation Méthodologique

Diagramme des Workflows

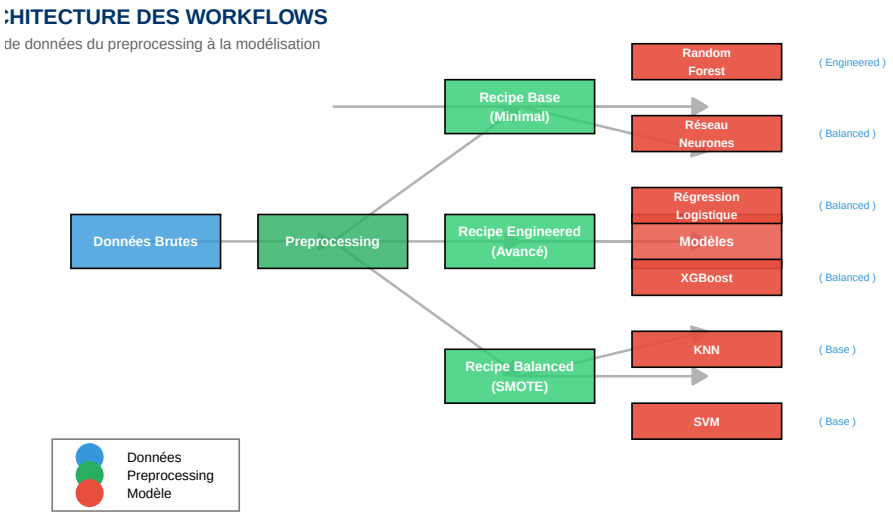

FIGURE 3 – Architecture des Workflows - Flux de Données du Preprocessing à la Modélisation

Stratégies de Preprocessing

TABLE 4 – Synthèse Compacte des Stratégies de Preprocessing

Recipe	Objectif	Transformations	Modèles associés
recipe_base	Préprocessing minimal pour modèles robustes	Encodage dummy Normalisation Z-score Imputation valeurs manquantes	SVM, KNN
recipe_engineered	Feature engineering avancé pour performance	Transformations log Catégorisation continues Interactions Suppression corrélations >0.9	Random Forest
recipe_balanced	Rééquilibrage des classes pour modèles sensibles	SMOTE (0.8, en CV) Encodage standard Normalisation	Régression Logistique, XGBoost, Réseau de N

Validation de l'Absence de Fuite d'Information

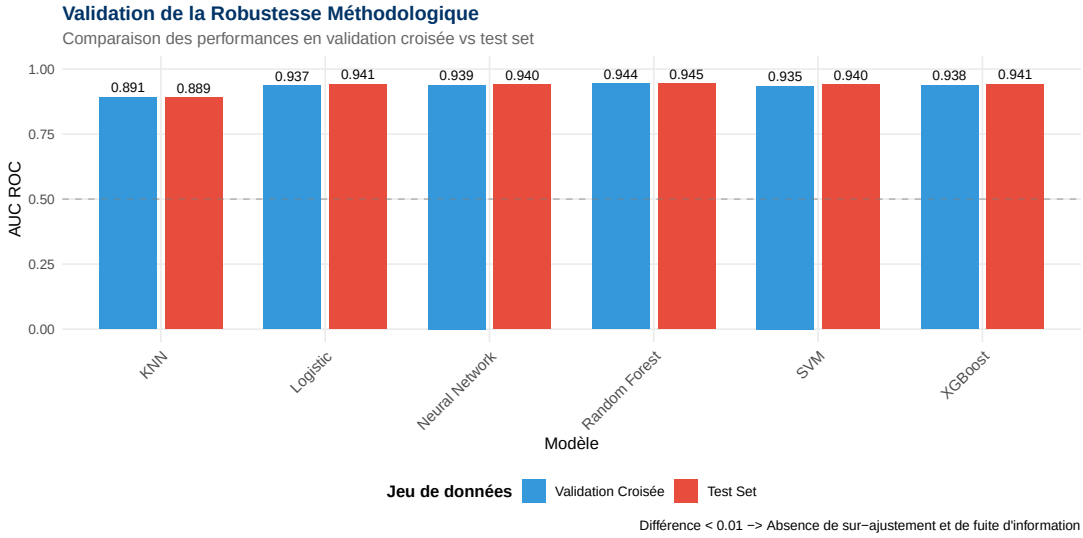

FIGURE 4 – Validation Méthodologique : Vérification de l'Absence d'Overfitting et de Fuite d'Information

Annexe B : Résultats Techniques Détaillés

Courbes de Tuning Complètes

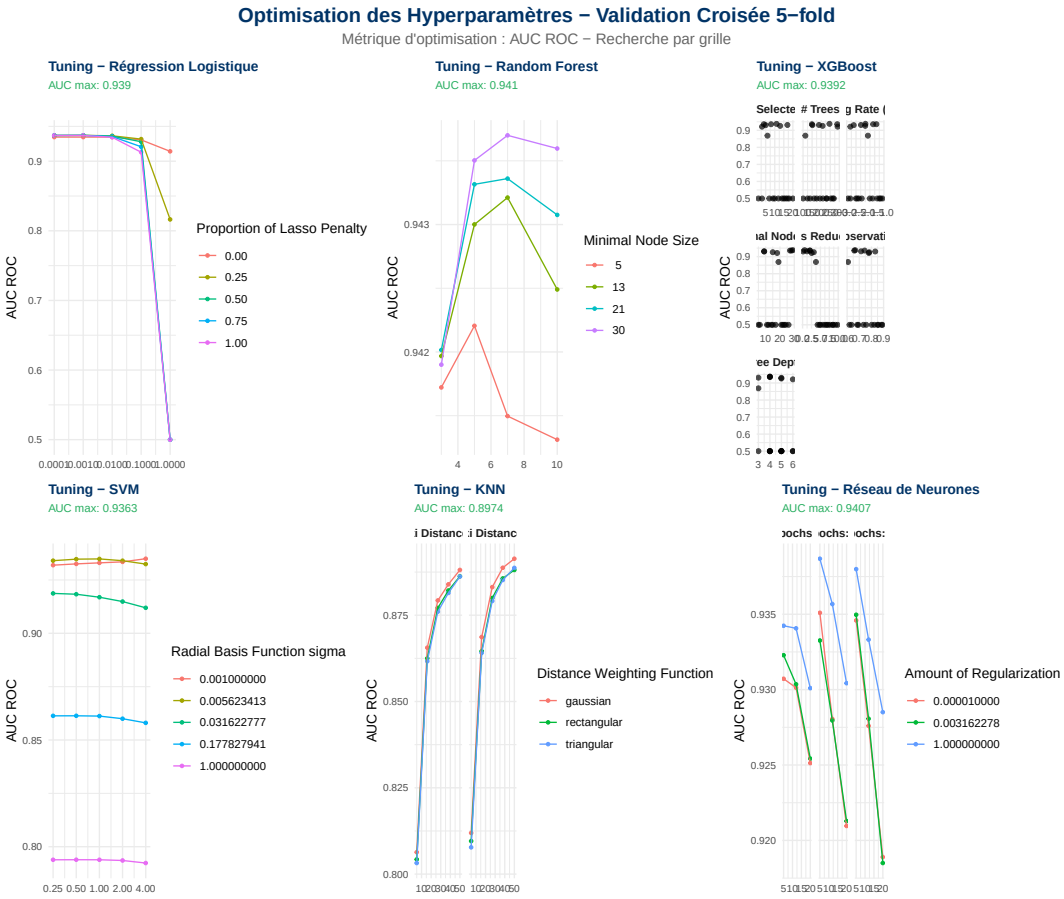

FIGURE 5 – Courbes de Tuning Détaillées pour les Six Modèles

Hyperparamètres Optimaux par Modèle

TABLE 5 – Hyperparamètres Optimaux Sélectionnés pour Chaque Modèle

Modèle	AUC (CV)	Hyperparamètres optimaux
Régression Logistique	0.9371	penalty = 0.001 ; mixture = 0.25
Random Forest	**0.9437**	**mtry = 7 ; min_n = 30**
XGBoost	0.9378	mtry = 11 ; trees = 246 ; min_n = 29 ; tree_depth = 4 ; learn_rate = 0.0102 ; loss_reduction = 3.6
SVM	0.9349	cost = 4 ; rbf_sigma = 0.001
KNN	0.8914	neighbors = 50 ; weight_func = gaussian ; dist_power = 2
Réseau de Neurones	0.9387	hidden_units = 5 ; penalty = 1 ; epochs = 125

Analyses Complémentaires

Segmentation Clients - Méthode du Coude

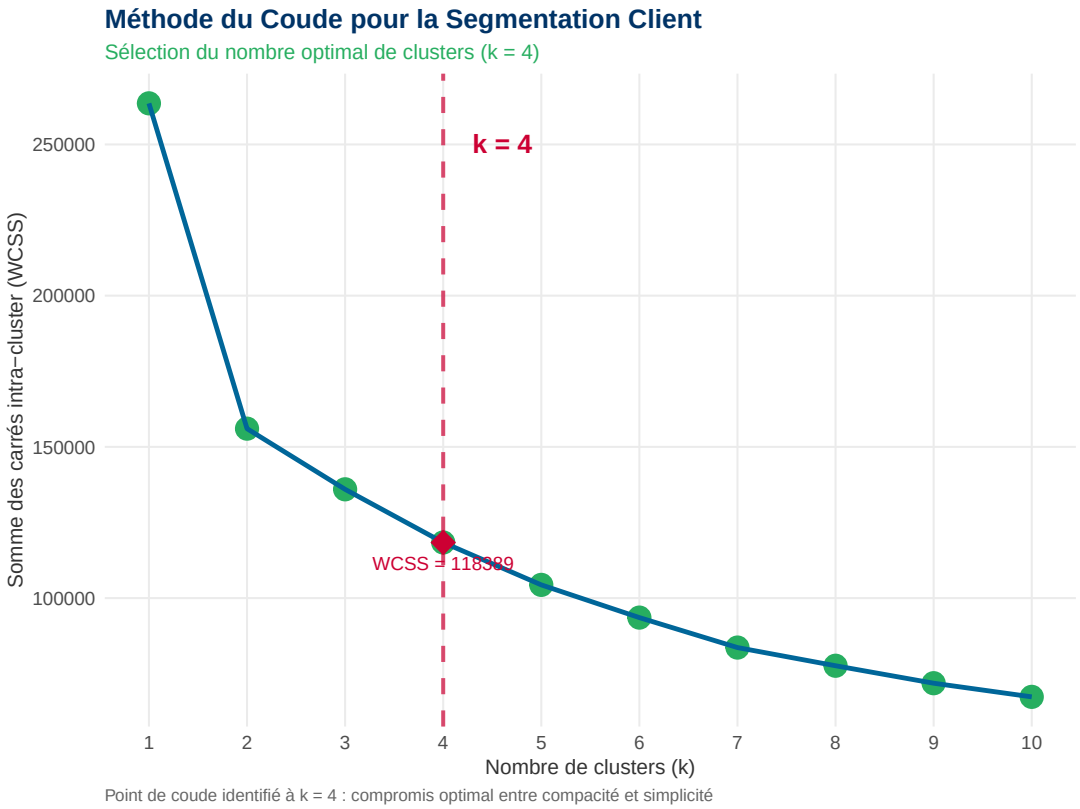

FIGURE 6 – Méthode du Coude pour la Sélection du Nombre Optimal de Clusters

Analyse Économique - Seuils Optimaux

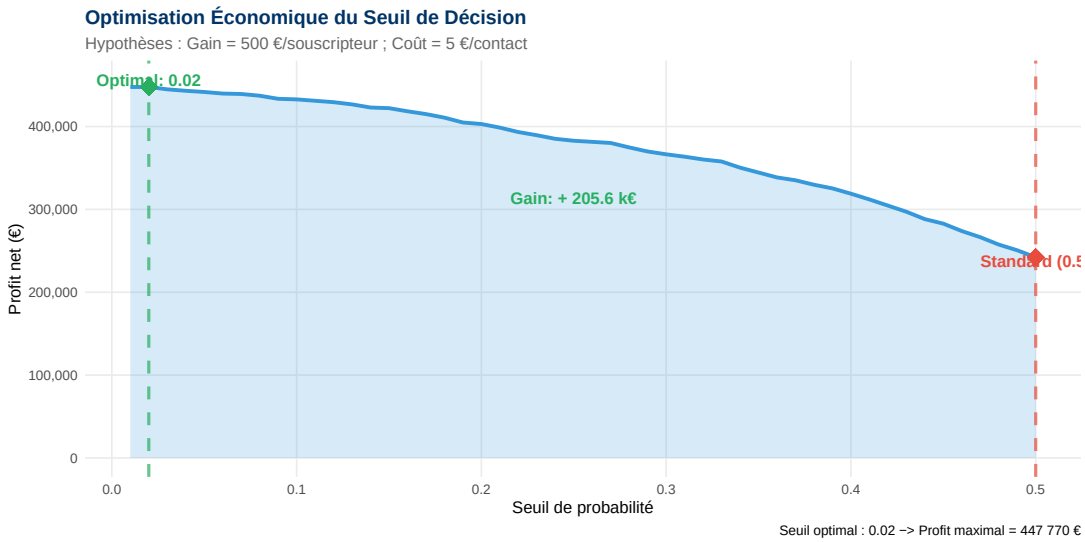

FIGURE 7 – Analyse Économique : Profit en Fonction du Seuil de Décision

Matrices de Confusion Complètes

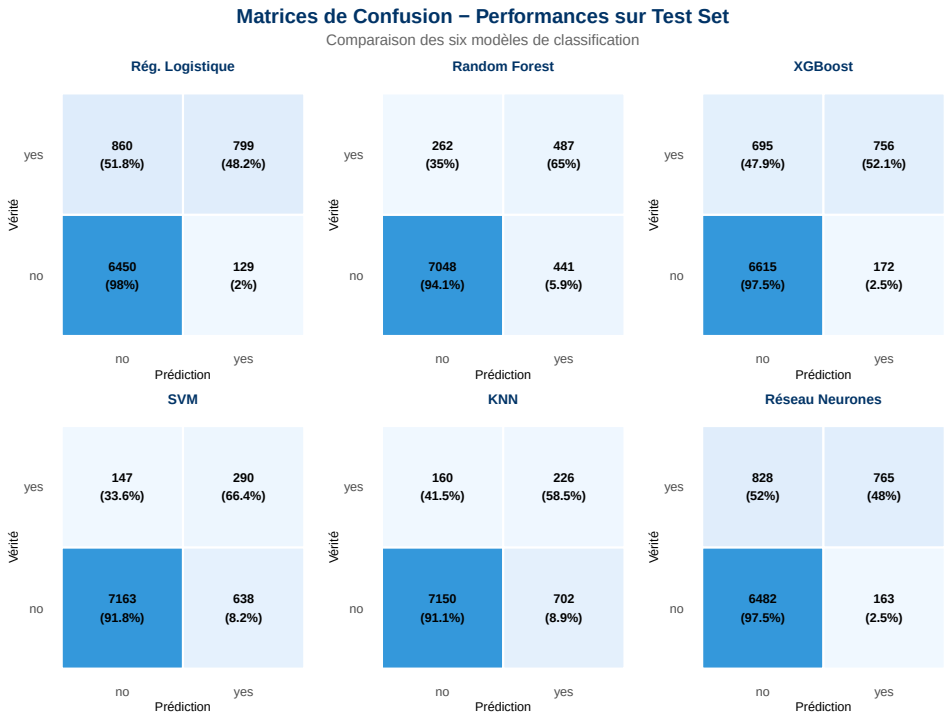

FIGURE 8 – Matrices de Confusion pour les Six Modèles (Test Set)