BIG DATA ANALYTICS

PRÉDICTION DE L'OUVERTURE DE COMPTES APRÈS UNE CAMPAGNE MARKETING

RÉALISÉ PAR
SAIDI AMAL
MULTON THIBAUT
DIALLO BAILO

SUPERVISÉ PAR
SESSI TOKPAVI

MASTER ESA
Econométrie et Statistique Appliquée
Université d'Orléans

JANVIER 2025

Table des matières

1 Résumé

Ce projet vise à prédire l'ouverture de comptes suite à des campagnes marketing téléphoniques, en exploitant des méthodes avancées de machine learning et d'analyse statistique. En s'appuyant sur une base de données riche et variée, notre objectif est de comprendre les facteurs influençant les décisions des clients et d'optimiser les ressources allouées aux campagnes.

Deux approches principales ont été explorées : la pénalisation et l'agrégation. Les modèles de pénalisation, incluant Ridge, Lasso, Elastic Net et Adaptive Lasso, ont permis de gérer efficacement les corrélations entre variables et de sélectionner les caractéristiques les plus pertinentes. D'autre part, les méthodes d'agrégation, telles que Random Forest et XGBoost, se sont montrées particulièrement performantes pour capturer les interactions complexes et non linéaires entre les variables.

Les performances des modèles ont été évaluées selon des métriques clés comme l'AUC, l'accuracy, la précision, le rappel et le F1-score. L'analyse comparative révèle que XGBoost affiche les meilleures performances globales, tandis que les modèles de régression pénalisée se distinguent par leur interprétabilité.

En conclusion, ce projet offre des perspectives concrètes pour améliorer l'efficacité des stratégies marketing bancaires, tout en mettant en lumière l'intérêt des approches combinant puissance prédictive et simplicité.

2 Introduction

Dans le secteur bancaire, la compétitivité repose de plus en plus sur l'efficacité des stratégies de ciblage marketing, notamment pour maximiser la fidélité client et optimiser les ressources. En effet, la concurrence intense dans ce secteur oblige les institutions financières à exploiter au mieux les informations à leur disposition pour atteindre les clients susceptibles d'être intéressés par leurs offres. Historiquement, les campagnes marketing se fondaient principalement sur des approches classiques, telles que l'analyse descriptive et la segmentation statique des clients en fonction de critères simples comme l'âge, le revenu ou la localisation géographique. Ces approches étaient efficaces dans un contexte de données limitées et de technologies rudimentaires. Cependant, avec la diversification des comportements clients et la complexité croissante des marchés, ces méthodes montrent rapidement leurs limites. Elles produisent souvent des résultats biaisés ou incomplets, incapables de répondre avec précision aux attentes spécifiques des clients. Cela entraîne des coûts opérationnels élevés, des taux de réponse faibles et une insatisfaction générale.

Face à ces insuffisances, l'explosion des données disponibles, combinée aux avancées technologiques en analyse et traitement de données, a transformé le paysage du marketing bancaire. Aujourd'hui, les banques disposent d'un volume sans précédent de données transactionnelles, comportementales et contextuelles provenant de multiples sources : interactions numériques, historiques de transactions, réseaux sociaux ou encore données démographiques. Cependant, ce volume élevé de données n'a de valeur que s'il peut être exploité efficacement. C'est ici qu'interviennent les avancées technologiques, notamment le Big Data et les algorithmes d'apprentissage automatique. Ces outils permettent de traiter ces informations massives pour identifier des modèles complexes et non linéaires dans les comportements des clients. Contrairement aux méthodes traditionnelles, ces approches modernes offrent la possibilité d'anticiper avec précision les réponses des clients à une campagne marketing, tout en minimisant les inefficiences et les coûts. En optimisant la segmentation et le ciblage, elles permettent aux banques d'allouer leurs ressources de manière optimale et de maximiser les retours sur investissement.

C'est dans ce contexte que s'inscrit notre étude, qui a pour ambition d'évaluer l'efficacité des techniques modernes de machine learning pour prédire l'ouverture d'un compte bancaire suite à une campagne téléphonique. En utilisant une base de données provenant d'une institution bancaire portugaise, nous nous proposons de comparer les performances des modèles de pénalisation, tels que Ridge, Lasso et Elastic Net, avec celles des méthodes d'agrégation, comme Random Forest et XGBoost. L'objectif est d'identifier les approches les plus adaptées pour répondre à ce type de problématique, alliant précision prédictive et robustesse.

Ce rapport s'organise autour de plusieurs axes principaux : une description détaillée des données utilisées et de leur préparation, une présentation des méthodologies mises en œuvre, et une analyse comparative des résultats obtenus selon des métriques pertinentes. À travers cette étude, nous visons à fournir des recommandations concrètes pour optimiser les stratégies de ciblage marketing dans un environnement bancaire de plus en plus compétitif et complexe.

3 Description de la base des données

Dans le cadre de ce projet, les données utilisées proviennent du dépôt de données "UC Irvine Machine Learning Repository", une référence pour l'apprentissage statistique. Elles relèvent du domaine du marketing bancaire, et plus précisément des campagnes téléphoniques menées auprès de potentiels clients d'une institution bancaire portugaise.
La variable cible Y est qualitative et binaire, distinguant les clients ayant ouvert un compte de dépôt à l'issue de la campagne (yes) de ceux ne l'ayant pas fait (no). La base inclut 14 variables prédictives, dont 6 qualitatives (telles que job, marital, education, etc.) et 6 quantitatives (notamment age, balance, duration, etc.), permettant d'évaluer différents aspects du comportement et des caractéristiques des clients.

Nous avons effectué des analyses graphiques ainsi que des statistiques descriptives afin de mieux comprendre et analyser la variable cible.

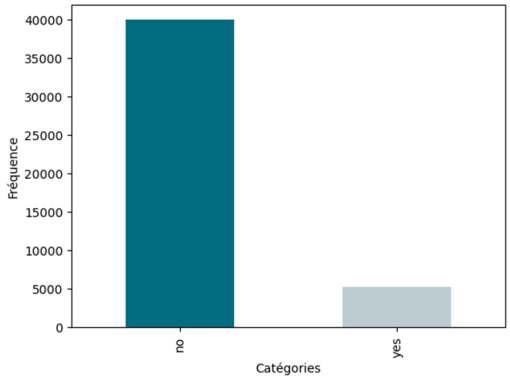

FIGURE 1 – Enter Caption

L'analyse de cette distribution met en évidence un fort déséquilibre entre les deux classes. La catégorie no (clients n'ayant pas ouvert de compte) est largement dominante avec une fréquence nettement supérieure à celle de la catégorie yes (clients ayant ouvert un compte). Cela reflète une faible proportion de succès des campagnes, ce qui est souvent le cas dans le domaine du marketing bancaire. Ce déséquilibre peut avoir des implications importantes pour la modélisation prédictive. Les modèles pourraient avoir tendance à favoriser la classe majoritaire (no), entraînant un risque de mauvaise classification des clients de la classe minoritaire (yes). L'utilisation de métriques adaptées, telles que le Recall, la Precision, et le F1-Score, sera cruciale pour évaluer correctement les performances des modèles.

Explorons à présent les variables explicatives présentes dans notre ensemble de données. Cette étape vise à analyser en détail les caractéristiques des variables indépendantes, afin de mieux comprendre leur nature, leurs distributions et leurs relations avec la variable cible Y.

Pour les variables quantitatives :

	Age	**Balance**	**Duration**	**Campaign**	**Pdays**	**Previous**	**Day**
Minimum	18.0	-8019.0	0.0	1.0	-1.0	0.0	1.0
Maximum	95.0	102127.0	4918.0	63.0	871.0	275.0	31.0
Moyenne	40.9362	1362.2721	258.1630	2.7638	40.1978	0.5803	15.8064
Médiane	39.0	448.0	180.0	2.0	-1.0	0.0	16.0

TABLE 1 – Statistiques descriptives des variables explicatives

On remarque que l'âge des clients varie de 18 à 95 ans, avec une moyenne de 40.93 ans, reflétant une population relativement jeune. Le solde annuel moyen (balance) montre une forte asymétrie positive, allant de -8019 (découverts) à 102127, avec une médiane de 448.

La durée des appels (duration) varie de 0 à 4918 secondes, avec une moyenne de 258 secondes, les valeurs nulles représentant probablement des contacts non réussis. La majorité des clients ont été contactés peu de fois lors de la campagne actuelle (campaign), avec une médiane de 2, tandis que les sollicitations lors des campagnes précédentes (previous) sont rares, avec une médiane de 0. La variable previous, représentant le nombre total de contacts lors de la campagne précédente, présente une valeur maximale de 275. Cette valeur, semble excessive et peu réaliste dans un contexte marketing. Par conséquent, cette observation a été supprimée afin de ne pas introduire de biais dans l'analyse et la modélisation. Enfin, la variable pdays, indiquant le nombre de jours depuis la dernière campagne, révèle que plus de 80 pourcent des clients n'ont pas été contactés auparavant (pdays = -1).

Quant aux variables qualitatives nous avons décidé de les représenter :

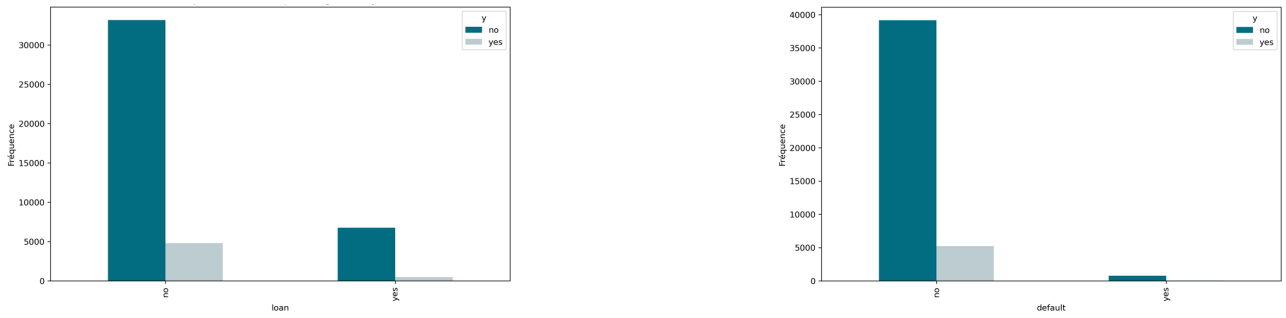

FIGURE 2 – Répartition des variables : Loan (gauche) et Default (droite) en fonction de Y

Le graphique de gauche montre que la majorité des clients n'ont pas de crédit personnel (loan = no), et parmi eux, on observe davantage de personnes ayant ouvert un compte (Y = yes) par rapport à celles ayant un crédit personnel (loan = yes), qui sont globalement moins nombreuses et moins susceptibles d'ouvrir un compte. Quant à la variable "default" indique que la majorité des clients n'ont pas de défaut de paiement (default = no), et parmi eux, on observe davantage de personnes ayant ouvert un compte (Y = yes) par rapport à celles ayant un défaut de paiement (default = yes), qui sont peu nombreuses et encore moins susceptibles d'ouvrir un compte.

Etude de la corrélation entre les variables quantitatives :

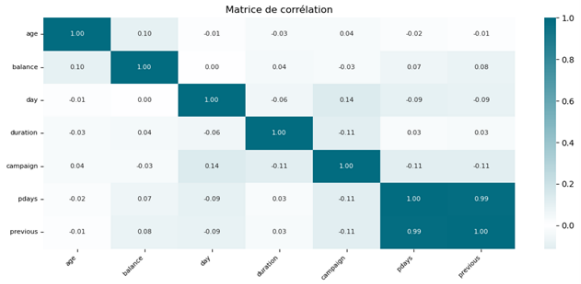

FIGURE 3 – Matrice de corrélation entre les variables quantitatives

Ces résultats suggèrent que la multicolinéarité n'est pas un problème majeur, à l'exception de la forte liaison entre pdays et previous.

Etude de la corrélation entre les variables qualitatives et la variable cible :

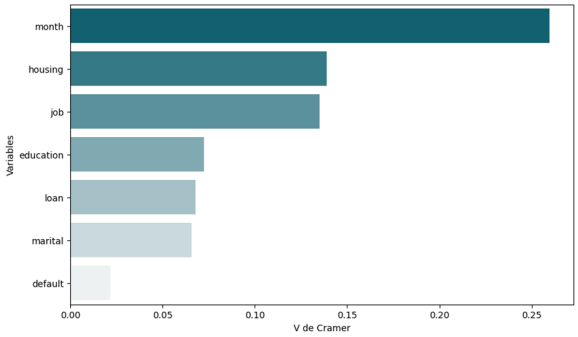

FIGURE 4 – V de cramer pour les variables qualitatives par rapport à la variable cible

Le graphique montre les valeurs du V de Cramer pour les variables qualitatives par rapport à la variable cible Y. La variable month est la plus corrélée à Y, suivie de housing et job, tandis que default présente la corrélation la plus faible, indiquant une influence limitée sur l'ouverture de compte. Ces résultats permettent de prioriser les variables pour la modélisation.

Etude de la corrélation entre les variables quantitatives et la variable cible :

Variable	**Statistique**	**P-Value**
age	3.461381	6.281783×10^{-2}
balance	454.774182	$6.593767 \times 10^{-101}$
day	39.473118	3.326067×10^{-10}
duration	5302.445386	0.000000
campaign	319.408727	1.948470×10^{-71}
pdays	1072.963591	$2.484050 \times 10^{-235}$
previous	1293.144416	$3.491720 \times 10^{-283}$

TABLE 2 – Statistiques descriptives et P-Values des variables quantitatives

Les variables balance, day, duration, campaign, pdays, et previous montrent des différences statistiquement significatives avec la variable cible, car leurs p-values sont inférieures à 0.05. La variable age ne montre pas de différence significative à un seuil de 5%.

4 Présentation des méthodes utilisées

4.1 Régression logistique

La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

Elle est également connue sous le nom de log-odds ou logarithme naturel de l'odds ratio, et cette fonction logistique est représentée par les formules suivantes :

$$Logit(\pi) = \frac{1}{1 + \exp(-\pi)}$$

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Dans cette équation de régression logistique, logit(pi) est la variable dépendante ou de réponse et x est la variable indépendante. Le paramètre bêta, ou coefficient, de ce modèle est généralement estimé via l'estimateur du maximum de vraisemblance (MLE).

4.2 Régression Ridge

La régression Ridge est une méthode avancée de régression linéaire intégrant la régularisation L2 pour limiter le sur-apprentissage (overfitting) en pénalisant la complexité du modèle. Développée pour résoudre les limitations de la régression classique, elle est particulièrement efficace lorsque le nombre de variables explicatives est élevé par rapport au nombre d'observations ou lorsque des corrélations entre les variables compromettent la stabilité numérique du modèle. Bien qu'elle soit principalement utilisée pour les problèmes de régression, elle peut être adaptée indirectement aux problèmes de classification binaire.

Fonction objectif de la régression Ridge :

$$PRSS = (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_2^2$$

Le paramètre $\lambda \geq 0$ contrôle l'intensité de la régularisation : plus λ est grand, plus la pénalité $\lambda\|\beta\|_2^2$ réduit la magnitude des coefficients, limitant ainsi le sur-apprentissage.

La régression Ridge se distingue par son efficacité à gérer la multicollinéarité des variables explicatives. La régularisation L2 favorise des coefficients plus petits sans les forcer à atteindre zéro, préservant ainsi toutes les variables dans le modèle.

4.3 Régression Lasso

La régression Lasso (Least Absolute Shrinkage and Selection Operator) est une méthode de régression pénalisée qui s'appuie sur la régularisation L1. Contrairement à la régression Ridge, qui réduit les coefficients sans les annuler, le Lasso introduit une pénalité basée sur la somme des valeurs absolues des coefficients, ce qui peut conduire certains d'entre eux à être exactement nuls. Cette caractéristique permet au Lasso de combiner régularisation et sélection automatique des prédicteurs, en identifiant ceux qui ont le plus d'impact sur la variable cible.

La fonction objectif de la régression Lasso est définie comme suit :

$$PRSS = (y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Cette approche est particulièrement adaptée aux situations où les données contiennent un grand nombre de prédicteurs, parmi lesquels certains peuvent avoir une contribution marginale. En ajustant la contrainte à l'aide du paramètre , le Lasso favorise la création de modèles plus simples et interprétables, tout en maîtrisant le sur-apprentissage et en sélectionnant les caractéristiques les plus pertinentes.

4.4 Régression Elastic Net

La régression Elastic Net est une méthode de régression pénalisée qui combine les régularisations L1 (utilisée dans le Lasso) et L2 (utilisée dans la régression Ridge). Proposée par Zou et Hastie en 2005, cette approche vise à tirer parti des forces des deux types de régularisation : la sélection de variables de Lasso et la gestion des prédicteurs corrélés de Ridge. L'Elastic Net est particulièrement utile lorsque les données contiennent de nombreuses variables corrélées, où le Lasso seul pourrait échouer à sélectionner correctement les prédicteurs pertinents.

La fonction objectif de la régression Elastic Net est donnée par :

$$PRSS = (y - X\beta)^T(y - X\beta) + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

où λ est le paramètre de régularisation et $\alpha \in [0, 1]$ contrôle l'équilibre entre les pénalités L1 et L2. Lorsque $\alpha = 1$, l'Elastic Net devient équivalent au Lasso, et lorsque $\alpha = 0$, il devient équivalent à la régression Ridge.

Cette approche permet de régulariser efficacement les modèles tout en sélectionnant automatiquement les prédicteurs les plus importants. En ajustant les paramètres λ et α, l'Elastic Net peut produire des modèles à la fois robustes et interprétables, particulièrement adaptés aux bases de données de haute dimension ou lorsque les prédicteurs présentent une multicolinéarité significative.

4.5 Régression Adaptive Lasso

La régression Adaptive Lasso est une extension de la méthode Lasso introduite par Zou en 2006. Elle améliore la sélection des variables en ajustant dynamiquement les pénalités appliquées aux coefficients des prédicteurs, en fonction de leur importance estimée lors d'une première étape. Contrairement au Lasso classique, qui applique la même pénalité à tous les prédicteurs, l'Adaptive Lasso attribue des poids différents à chaque coefficient, permettant une sélection plus précise des variables pertinentes.

La fonction objectif de la régression Adaptive Lasso est définie comme suit :

$$PRSS = (y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{p} w_j|\beta_j|$$

où $w_j = \frac{1}{|\hat{\beta}_j|^\gamma}$ représente le poids attribué à chaque prédicteur j, calculé à partir des coefficients $\hat{\beta}_j$ estimés lors d'une première étape (généralement avec une régression Ridge ou Lasso standard), et $\gamma > 0$ est un paramètre qui contrôle la sensibilité des poids.

L'Adaptive Lasso combine les avantages de la régularisation L1 avec une pondération adaptative qui favorise les prédicteurs les plus influents tout en pénalisant davantage ceux qui ont une contribution moindre. Cette approche est particulièrement efficace dans les scénarios où certaines variables importantes pourraient être négligées par le Lasso en raison de la corrélation entre les prédicteurs. En utilisant les poids adaptatifs, l'Adaptive Lasso offre une meilleure précision dans la sélection des variables et améliore la généralisation du modèle.

4.6 Random Forest

Le Random Forest, formalisé par Breiman (2001), est une méthode ensembliste qui constitue une amélioration significative du bagging dans le cadre des arbres de décision, en particulier de l'algorithme CART. Cette technique est principalement utilisée pour des problèmes de classification et de régression, avec un objectif clair : réduire la variance tout en augmentant la robustesse et la précision des prédictions.

Cependant, la performance reste limitée si une forte corrélation existe entre les arbres. C'est ici que le Random Forest innove, en cherchant à diminuer cette corrélation grâce à un double échantillonnage : d'une part sur les exemples de données (échantillonnage bootstrap), et d'autre part sur les prédicteurs disponibles à chaque division binaire.

Le processus peut être décrit en trois grandes étapes :

- Construction d'un échantillon bootstrap : À chaque itération, un sous-ensemble de données est tiré aléatoirement avec remplacement à partir du jeu de données initial. Cela génère une diversité au niveau des échantillons d'apprentissage.

- Sélection des prédicteurs candidats : Lors de chaque division d'un arbre, un sous-ensemble de prédicteurs est choisi de manière aléatoire parmi l'ensemble des caractéristiques disponibles. En pratique, la valeur optimale du nombre de prédicteurs sélectionnés (q) est souvent $\sqrt{p}$ pour

la classification et $\log_2(p)$ pour la régression, où p représente le nombre total de prédicteurs.

- Agrégation des prédictions : Une fois tous les arbres construits, les prédictions sont combinées pour fournir un résultat final. Dans le cas d'une classification, la classe majoritaire parmi les arbres détermine la prédiction, tandis que pour la régression, c'est la moyenne des prédictions qui est retenue.
Le Random Forest permet d'améliorer les performances globales et d'obtenir des modèles capables de généraliser efficacement sur de nouvelles données. En outre, l'utilisation de multiples arbres construits à partir d'échantillons aléatoires réduit considérablement le risque de surajustement.

4.7 XGBoost

L'Extreme Gradient Boosting (XGBoost), introduit par Chen et Guestrin en 2016, est une amélioration sophistiquée du Gradient Boosting, lui-même conçu par Friedman en 2001. Bien que le Gradient Boosting et XGBoost partagent une structure de base similaire, XGBoost se distingue par ses optimisations avancées qui le rendent plus performant en pratique.
L'objectif principal de XGBoost est de pallier les limitations du Gradient Boosting traditionnel, telles qu'une lenteur d'exécution due à son entraînement séquentiel complexe et une faible tolérance aux données manquantes.

XGBoost surmonte plusieurs défis grâce à ses innovations clés : une régularisation avancée qui pénalise le nombre de feuilles pour éviter le surajustement, une optimisation basée sur une approximation de second ordre de Taylor pour améliorer la précision et accélérer les calculs, et une gestion efficace des données manquantes qui renforce la robustesse tout en optimisant la vitesse d'exécution et l'utilisation de la mémoire.

Le modèle repose sur l'idée d'un ensemble d'arbres de décision, où chaque arbre successif corrige les erreurs des prédictions précédentes. La prédiction finale pour une observation donnée est calculée comme la somme des prédictions individuelles des arbres :

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \quad \forall i = 1, \ldots, n, \quad f_t \in \mathcal{F}$$

où f_t représente une fonction dans l'ensemble des arbres de décision. Pour les problèmes de classification, la décision finale est obtenue via la fonction signe appliquée à : $\hat{y}_i$.

Dans le cas des problèmes de classification binaire, XGBoost s'appuie sur une fonction objective régularisée pour apprendre les différentes fonctions f_t. Cette fonction a pour rôle d'optimiser les prédictions en équilibrant l'ajustement au jeu de données et la complexité du modèle. La fonction objective peut être exprimée ainsi :

$$\mathcal{L} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{t=1}^{T} \Omega(f_t)$$

Où $l(\hat{y}_i, y_i)$ représente une fonction de perte qui quantifie l'écart entre les prédictions $\hat{y}_i$ et les valeurs réelles y_i.
Le terme $\Omega(f_t)$ est un terme de régularisation visant à contrôler la complexité des arbres, souvent proportionnel au nombre de feuilles ou à d'autres métriques structurelles.

5 Critère d'évaluation

5.1 F1-Score

Le F1-score est une mesure d'évaluation utilisée pour évaluer la performance des modèles de classification, particulièrement utile lorsque les classes sont déséquilibrées. Il combine la précision et le rappel dans une moyenne harmonique, offrant ainsi un équilibre entre ces deux métriques.

La formule du F1-score est la suivante :

$$F1 = 2 \cdot \frac{Précision \cdot Rappel}{Précision + Rappel}$$

où :

Précision : Proportion des prédictions positives correctes parmi l'ensemble des prédictions positives.
Rappel : Proportion des vrais positifs parmi l'ensemble des instances positives.

Le F1-score varie entre 0 et 1, où 1 représente une performance parfaite. Il est particulièrement adapté lorsque les faux négatifs et les faux positifs ont des coûts similaires dans l'analyse.

5.2 Recall

Le recall (ou sensibilité) est une mesure d'évaluation utilisée pour évaluer la capacité d'un modèle de classification à identifier correctement les instances positives. Il est particulièrement utile lorsque l'objectif est de minimiser les faux négatifs.

La formule du Recall est la suivante :

$$Recall = \frac{VraisPositifs(TP)}{VraisPositifs(TP) + FauxNégatifs(FN)}$$

où :

TP : Nombre de vrais positifs (instances positives correctement classées).
FN : Nombre de faux négatifs (instances positives incorrectement classées comme négatives).

Le Recall varie entre 0 et 1, où 1 indique que toutes les instances positives ont été correctement identifiées. Il est crucial dans les cas où il est plus important de ne pas manquer de cas positifs (par exemple, dans les diagnostics médicaux ou les fraudes).

5.3 Précision

La précision est une mesure d'évaluation utilisée pour évaluer la proportion des prédictions positives correctes parmi toutes les prédictions positives effectuées par le modèle. Elle est particulièrement importante lorsque le coût des faux positifs est élevé.

La formule de la précision est la suivante :

$$Précision = \frac{VraisPositifs(TP)}{VraisPositifs(TP) + FauxPositifs(FP)}$$

où :

TP : Nombre de vrais positifs (instances positives correctement classées).
FP : Nombre de faux positifs (instances négatives incorrectement classées comme positives).

La Précision varie entre 0 et 1, où 1 indique que toutes les prédictions positives effectuées par le modèle sont correctes. Elle est cruciale dans les cas où les faux positifs ont des conséquences importantes (par exemple, dans la détection de spams ou de fraudes).

5.4 ROC-AUC

La courbe ROC (Receiver Operating Characteristic) est une représentation graphique qui illustre la performance d'un modèle de classification et permet de visualiser le compromis entre ces deux métriques à différents seuils de décision. L'AUC (Area Under the Curve) est la surface sous la courbe ROC. Elle fournit une mesure quantitative de la performance globale du modèle.

$$AUC = \int_0^1 ROC(FPR)\, d(FPR)$$

où :

TPR : Le taux de vrais positifs.
FPR : Le taux de faux positifs.

Une AUC égale à 1 indique un modèle parfait, capable de différencier sans erreur les classes positives et négatives. À l'inverse, une AUC égale à 0.5 signifie que le modèle ne fait pas mieux qu'un tirage aléatoire. Plus l'AUC se rapproche de 1, meilleure est la capacité du modèle à distinguer les classes. Cette métrique est particulièrement utile pour comparer les performances de différents modèles ou pour ajuster le seuil de classification afin d'optimiser les résultats.

5.5 Accuracy

L'accuracy est une mesure d'évaluation utilisée pour quantifier la proportion totale des prédictions correctes effectuées par un modèle de classification. Elle est particulièrement utile lorsque les classes positives et négatives sont équilibrées.
La formule de l'accuracy est donnée par :

$$Accuracy = \frac{VraisPositifs(TP) + VraisNégatifs(TN)}{Totaldesinstances(TP + TN + FP + FN)}$$

L'accuracy varie entre 0 et 1, où une valeur proche de 1 indique que le modèle effectue un nombre élevé de prédictions correctes. Cependant, cette métrique peut être trompeuse lorsque les classes sont déséquilibrées, car elle ne distingue pas entre les erreurs pour les classes positives et négatives. Dans ce cas, d'autres métriques comme le recall ou la précision peuvent être plus appropriées.

6 Analyse des résultats

Régression Logistique

AUC	*Accuracy*	*Précision*	*Recall*	*F*1*Score*
0.878	0.888	0.572	0.212	0.309

Table 3 – Indicateurs de performance - Régression Logistique

Le modèle de régression logistique présente une AUC de 0.878 et une accuracy de 88,8 %, ce qui reflète une bonne capacité discriminante et une proportion élevée de prédictions correctes. Cependant, la précision (57,2 %) montre que seulement un peu plus de la moitié des prédictions positives sont correctes, tandis que le recall (21,2 %) révèle une faible capacité à identifier les cas positifs. Le F1-score, à 30,9 %, illustre un équilibre limité entre précision et recall, mettant en évidence une performance globalement moyenne pour capturer les cas positifs.

Régression Ridge

AUC	*Accuracy*	*Précision*	*Recall*	*F*1*Score*
0.887	0.890	0.605	0.174	0.270

Table 4 – Indicateurs de performance - Régression Ridge

La régression Ridge présente une AUC de 0.887 et une accuracy de 89 %, indiquant une bonne capacité discriminante et une proportion élevée de prédictions correctes. Cependant, la précision (60,5 %) est modérée, et le recall (17,4 %) est faible, ce qui montre une difficulté à identifier les cas positifs. Le F1-score, à 27 %, reflète un déséquilibre important entre précision et recall, limitant l'efficacité du modèle pour capturer les cas positifs.

Variables	**duration**	**housing**	**campaign**	**pdays**	**loan**
Coefficients	1.0650	-0.8370	-0.3312	0.0279	-0.5320

TABLE 5 – Extrait des variables et leurs coefficients - Ridge

Les coefficients du modèle Ridge montrent que la duration (1.0650) a un effet positif similaire à celui du Lasso, augmentant la probabilité de succès avec la durée de la campagne. Housing (-0.8370) reste négatif, indiquant que l'association avec le secteur du logement diminue cette probabilité. Campaign (-0.3312) a également un impact négatif, bien que modéré. Pdays (0.0279) montre un effet positif plus faible que dans le modèle Lasso, suggérant un impact plus limité du contact préalable. Enfin, loan (-0.5320) a un effet négatif similaire à celui du Lasso, diminuant la probabilité de réponse positive chez les emprunteurs.

Régression Lasso

AUC	*Accuracy*	*Précision*	*Recall*	*F1Score*
0.888	0.892	0.589	0.256	0.357

TABLE 6 – Indicateurs de performance - Régression Lasso

La régression Lasso présente une AUC de 0.888 et une accuracy de 89,2 %, témoignant d'une bonne capacité discriminante et d'une proportion élevée de prédictions correctes. La précision (58,9 %) reste modérée, tandis que le recall (25,6 %) est relativement faible, indiquant une difficulté à identifier les cas positifs. Le F1-score, à 35,7 %, reflète un équilibre limité entre précision et recall, mais légèrement supérieur à celui de la régression Ridge, suggérant une amélioration dans la gestion des compromis entre ces deux métriques.

Variables	**duration**	**housing**	**campaign**	**pdays**	**loan**
Coefficients	1.0646	-0.8343	-0.3317	0.2379	-0.5302

TABLE 7 – Extrait des variables et leurs coefficients - Lasso

Le coefficient positif de duration (1.0646) suggère qu'une augmentation de la durée de la campagne augmente la probabilité de succès. En revanche, le coefficient négatif de housing (-0.8343) indique que l'association avec le secteur du logement diminue cette probabilité. Campaign (-0.3317) a un effet légèrement négatif, tandis que pdays (0.2379) montre que le contact préalable améliore nos chances de succès. Enfin, loan (-0.5302) indique que les personnes ayant un prêt sont moins susceptibles de répondre positivement.

Régression Elastic Net

AUC	*Accuracy*	*Précision*	*Recall*	*F1Score*
0.855	0.89	0.59	0.21	0.31

TABLE 8 – Indicateurs de performance - Régression Elastic Net

La régression Elastic Net présente une AUC de 0.855 et une accuracy de 88.9 %, indiquant une bonne capacité discriminante et une proportion élevée de prédictions correctes. Avec une précision de 59 %, le modèle identifie efficacement les vrais positifs parmi les prédictions positives, bien que le recall de 21 % montre une faible capacité à capturer les cas positifs réels. Le F1-score, à 31 %, reflète un équilibre limité entre précision et rappel, suggérant que le modèle privilégie les prédictions positives correctes au détriment de la couverture des cas positifs.

Variables	**duration**	**housing**	**campaign**	**pdays**	**loan**
Coefficients	0.1085	-0.0337	0.0019	0.0210	-0.0072

TABLE 9 – Extrait des variables et leurs coefficients - Elastic Net

Le tableau présente les coefficients d'un modèle Elastic Net, qui combine les régularisations L1 et L2 pour sélectionner les variables pertinentes et gérer la multicolinéarité. Le coefficient positif de duration (0.1085) suggère qu'une augmentation de la durée est associée à une légère augmentation de la probabilité de l'événement cible, tandis que le coefficient négatif de housing (-0.0337) indique qu'un lien avec le logement diminue cette probabilité. Le coefficient très faible de campaign (0.0019) montre que son impact sur la probabilité est négligeable, tandis que pdays (0.0210) a un léger effet positif, suggérant qu'un contact préalable augmente légèrement la probabilité de l'événement. Enfin, le coefficient négatif de loan (-0.0072) indique que les personnes ayant un prêt sont légèrement moins susceptibles de répondre positivement à la campagne. L'Elastic Net a régularisé ces coefficients, réduisant l'impact des variables moins significatives tout en maintenant les plus influentes, avec des effets généralement modérés.

Régression Adaptive Lasso

AUC	*Accuracy*	*Précision*	*Recall*	*F1Score*
0.86	0.89	0.60	0.212	0.31

TABLE 10 – Indicateurs de performance - Régression Adaptive Lasso

La régression Adaptive Lasso présente une AUC de 0.86 et une accuracy de 78 %, indiquant une capacité discriminante correcte, bien qu'inférieure à celle d'autres modèles. Avec une précision de 60 %, le modèle identifie efficacement les vrais positifs parmi les prédictions positives. Cependant, un recall de 21,2 % souligne une difficulté à capturer une majorité de cas positifs. Le F1-score de 31 % reflète cet équilibre modéré entre précision et rappel, limitant son efficacité globale dans un contexte où la détection des cas positifs est cruciale.

Variables	**duration**	**housing**	**campaign**	**pdays**	**loan**
Coefficients	1.0437	-0.3268	-0.1624	0.0916	-0.0575

TABLE 11 – Extrait des variables et leurs coefficients - Adaptive Lasso

Les coefficients issus de l'Adaptive Lasso montrent que la duration a un impact positif important sur la probabilité de succès de l'événement cible, tandis que des variables comme housing et loan ont des effets négatifs, suggérant que les personnes avec un prêt ou un logement stable

sont moins susceptibles de répondre favorablement. Le coefficient négatif de campaign indique qu'une sur-exposition à la campagne pourrait réduire son efficacité, tandis que pdays a un effet positif mais faible, suggérant qu'un contact antérieur peut légèrement améliorer la réponse. L'Adaptive Lasso permet ici de sélectionner efficacement les variables les plus pertinentes tout en régularisant celles de moindre importance.

Random Forest

AUC	*Accuracy*	*Précision*	*Recall*	*F1Score*
0.919	0.900	0.650	0.380	0.480

TABLE 12 – Indicateurs de performance - Random Forest

Le modèle Random Forest affiche une excellente capacité discriminante avec une AUC élevée de 0.919 et une accuracy de 90 %, indiquant une forte proportion de prédictions correctes. Cependant, la précision (65 %) et le recall (38 %) révèlent un compromis où le modèle favorise les prédictions positives correctes au détriment de la détection de toutes les observations positives. Le F1-score, à 48 %, reflète cet équilibre modéré.

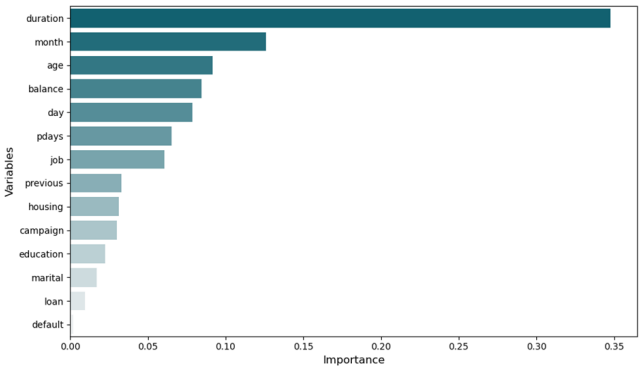

FIGURE 5 – Importance des variables - Random Forest

Dans la modélisation par Random Forest, toutes les variables contribuent à la prédiction, mais leur importance varie significativement. Comme illustré dans la figure, la variable duration se démarque nettement avec une contribution dominante, indiquant qu'elle joue un rôle clé dans la prédiction de la variable cible. Cette importance s'explique probablement par sa capacité à capturer des informations discriminantes essentielles.

D'autres variables, comme month, age, et balance, montrent également des contributions notables, reflétant leur pertinence dans le modèle. En revanche, des variables telles que marital ou loan ont des importances plus faibles, suggérant qu'elles apportent une valeur limitée à la prédiction.

XGBoost

AUC	*Accuracy*	*Précision*	*Recall*	*F1Score*
0.928	0.910	0.630	0.450	0.520

TABLE 13 – Indicateurs de performance - XGBoost

Le modèle XGBoost se distingue par des performances globales élevées, avec une AUC de 0.928 et une accuracy de 91 %, témoignant d'une excellente capacité discriminante et d'une forte proportion de prédictions correctes. La précision (63 %) et le recall (45 %) montrent un meilleur équilibre que le Random Forest, reflété par un F1-score de 52 %, qui indique une gestion plus efficace des compromis entre faux positifs et détection des cas positifs.

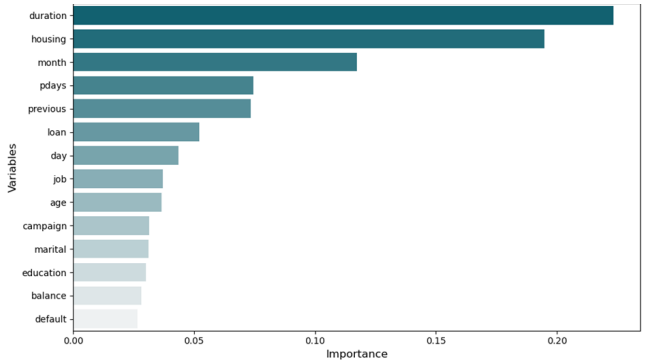

FIGURE 6 – Importance des variables - XGBoost

Dans le modèle XGBoost, certaines variables se distinguent par leur contribution prédominante. Comme illustré dans la figure, la variable duration se positionne comme la plus influente, jouant un rôle central dans la prédiction de la variable cible grâce à sa forte capacité discriminante. Cette importance traduit son impact significatif sur les performances globales du modèle.

D'autres variables telles que month, campaign, et age affichent également des contributions notables, soulignant leur pertinence dans le processus de décision. En revanche, des variables comme loan ou marital montrent une importance relativement faible, suggérant qu'elles ont un impact limité sur les prédictions finales.

6.1 Comparaison des résultats

Dans cette section, nous présentons un tableau récapitulatif des performances des différentes méthodes utilisées, évaluées selon plusieurs métriques. Ces métriques incluent notamment la précision, le rappel, le score F1, l'AUC et l'accuracy. Chaque méthode a été évaluée sur un ensemble de données de test afin de fournir une comparaison objective de leurs performances respectives. Le tableau ci-dessous résume les résultats obtenus pour chaque méthode.

Modèle	**AUC**	**Accuracy**	**Précision**	**Recall**	**F1 Score**
Logistique	0.878	0.888	0.572	0.212	0.309
Ridge	0.887	0.890	0.605	0.174	0.270
Lasso	0.888	0.892	0.589	0.256	0.357
Elastic Net	0.855	0.88	0.59	0.212	0.31
Adaptive Lasso	0.86	0.89	0.60	0.210	0.310
Random Forest	0.919	0.900	**0.650**	0.380	0.480
XG Boost	**0.928**	**0.910**	0.630	**0.450**	**0.520**

TABLE 14 – Tableau résumant les performances des modèles sur l'échantillon Test

Parmi les modèles évalués, XGBoost se distingue avec une AUC de 0.928 et une accuracy de 91%, témoignant d'une excellente capacité discriminante et d'un bon équilibre entre précision (63%) et rappel (45%). Random Forest suit avec une AUC de 0.919 et une précision légèrement supérieure (65%), mais un rappel plus faible (38%).
Les modèles de régression, comme Lasso, Ridge et Elastic Net, affichent des performances correctes (AUC entre 0.855 et 0.888) mais inférieures à XGBoost et Random Forest, notamment en termes de rappel et de F1-score. Enfin, Adaptive Lasso et la régression logistique présentent des résultats plus limités, avec des F1-scores respectivement de 31% et 30.9%.

7 Conclusion

En conclusion, cette étude a exploré diverses approches de machine learning pour prédire l'ouverture de comptes suite à des campagnes marketing, en mettant l'accent sur les performances des modèles de régression et d'agrégation. À travers l'évaluation de métriques clés telles que l'AUC, l'accuracy, la précision, le recall et le F1-score, nous avons identifié des tendances significatives et des compromis entre les différents modèles.
Les méthodes de pénalisation, comme la régression Ridge, Lasso et Elastic Net, ont montré une capacité à gérer efficacement la complexité des données et à sélectionner les variables pertinentes. Parmi elles, Elastic Net s'est distingué par un bon équilibre entre précision et recall, tout en maintenant une bonne capacité discriminante.

Dans le domaine de l'agrégation, le modèle XGBoost s'est affirmé comme la méthode la plus performante, combinant une AUC élevée de 0.928 et une accuracy de 91%. Ce modèle a su capturer les relations complexes entre les variables explicatives et la variable cible, tout en offrant une gestion équilibrée des erreurs.

Ces résultats illustrent l'importance d'adapter le choix des modèles aux spécificités des données et des objectifs du projet. En combinant les points forts des différentes approches, cette étude apporte des orientations pratiques pour optimiser les stratégies marketing, en identifiant les clients les plus susceptibles d'ouvrir un compte, et en améliorant ainsi l'efficacité globale des campagnes.