

Projet Machine learning

BANK CHURN SCORING

Groupe 12

Emilia Laurine ADOGOUN

Ousseynou GUEYE

Isabelle Danielle MOSSE

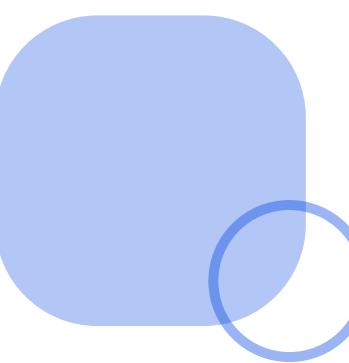
Famara SADIO



CONTEXTE

T

Fortuneo Banque, une banque en ligne innovante, cherche à développer un outil de "churn scoring" pour prédire la probabilité qu'un client quitte la banque. La rétention des clients est essentielle dans le secteur bancaire en ligne, où la concurrence est intense et les coûts d'acquisition de nouveaux clients sont élevés. Les outils de churn scoring jouent un rôle crucial en permettant aux entreprises de cibler efficacement leurs efforts de fidélisation.



01

Objectif

Identifier les clients susceptibles de quitter la banque.

02

Importance des outils churn

- Prévention proactive : Anticiper le départ des clients permet de prendre des mesures préventives avant qu'ils ne quittent.
- Réduction des coûts : Réduire le churn diminue les coûts liés à l'acquisition de nouveaux clients.
- Amélioration de la satisfaction client : En répondant aux besoins des clients à risque, la banque peut améliorer leur satisfaction et leur engagement.
- Optimisation des ressources : Cibler les clients à haut risque permet d'utiliser les ressources de manière plus efficace et efficiente.

03

Impact

Réduire le taux de churn en mettant en place des actions préventives ciblées pour retenir les clients à haut risque.



PLAN

- 01** Analyse exploratoire
- 02** Résultats des modèles
- 03** Choix du modèle final

ANALYSE EXPLORATOIRE



Description

- 14 variables et 165034 observations
- Pas de doublons
- Pas de valeurs manquantes



En résumé

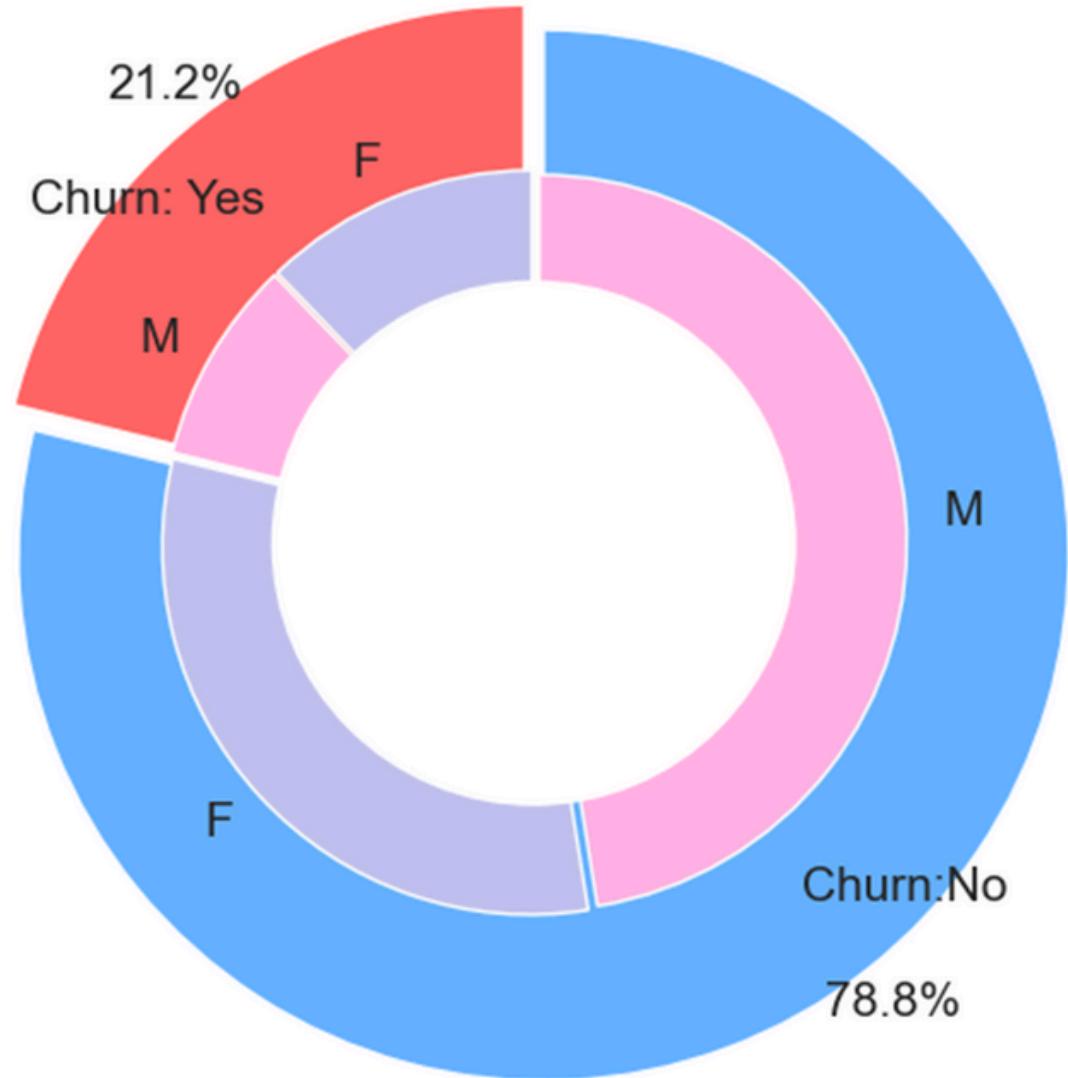
Le score de crédit moyen est de 656, avec la majorité des scores entre 597 et 710. L'âge moyen des clients est de 38 ans, avec un large éventail allant de 18 à 92 ans. En termes de fidélité, les clients restent en moyenne 5 ans avec la banque, bien que la durée varie de 0 à 10 ans.

La plupart des clients ont un solde nul, mais ceux qui ont un solde, en ont des très élevés, avec une moyenne de 55478 et un maximum de 250898. Les clients utilisent généralement entre 1 et 2 produits bancaires, avec une moyenne de 1,55. Environ 75 % des clients possèdent une carte de crédit, et seulement la moitié des clients sont des membres actifs.

Le salaire estimé moyen des clients est de 112574, avec des salaires allant de 11,58 à près de 200000. Enfin, environ 21 % des clients ont quitté la banque, ce qui indique que la majorité (79 %) des clients restent.

ANALYSE EXPLORATOIRE

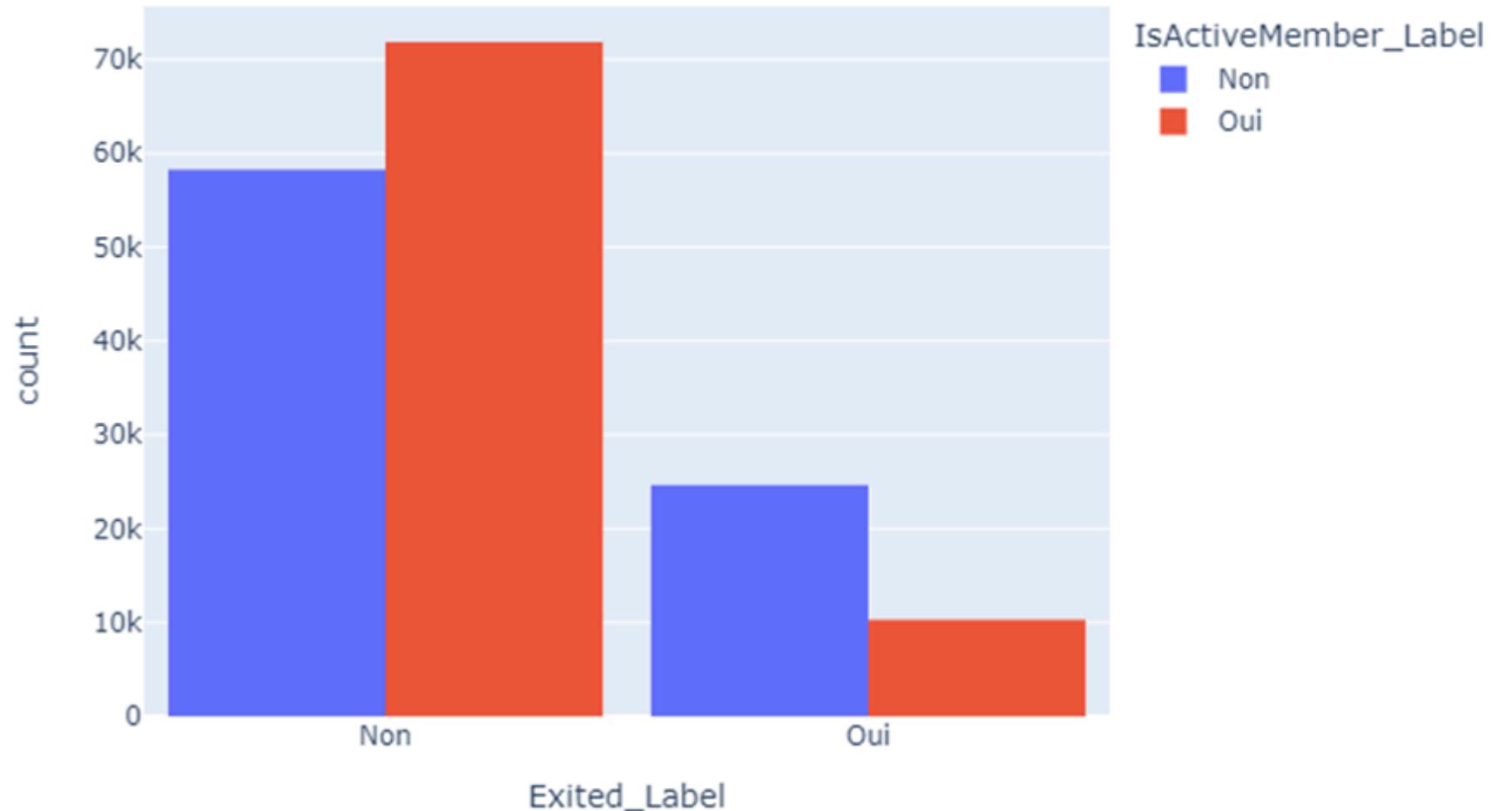
Churn Distribution w.r.t Gender: Male(M), Female(F)



Sexe des clients

- Plus de femmes ayant quitté la banque que d'hommes

Membre actif

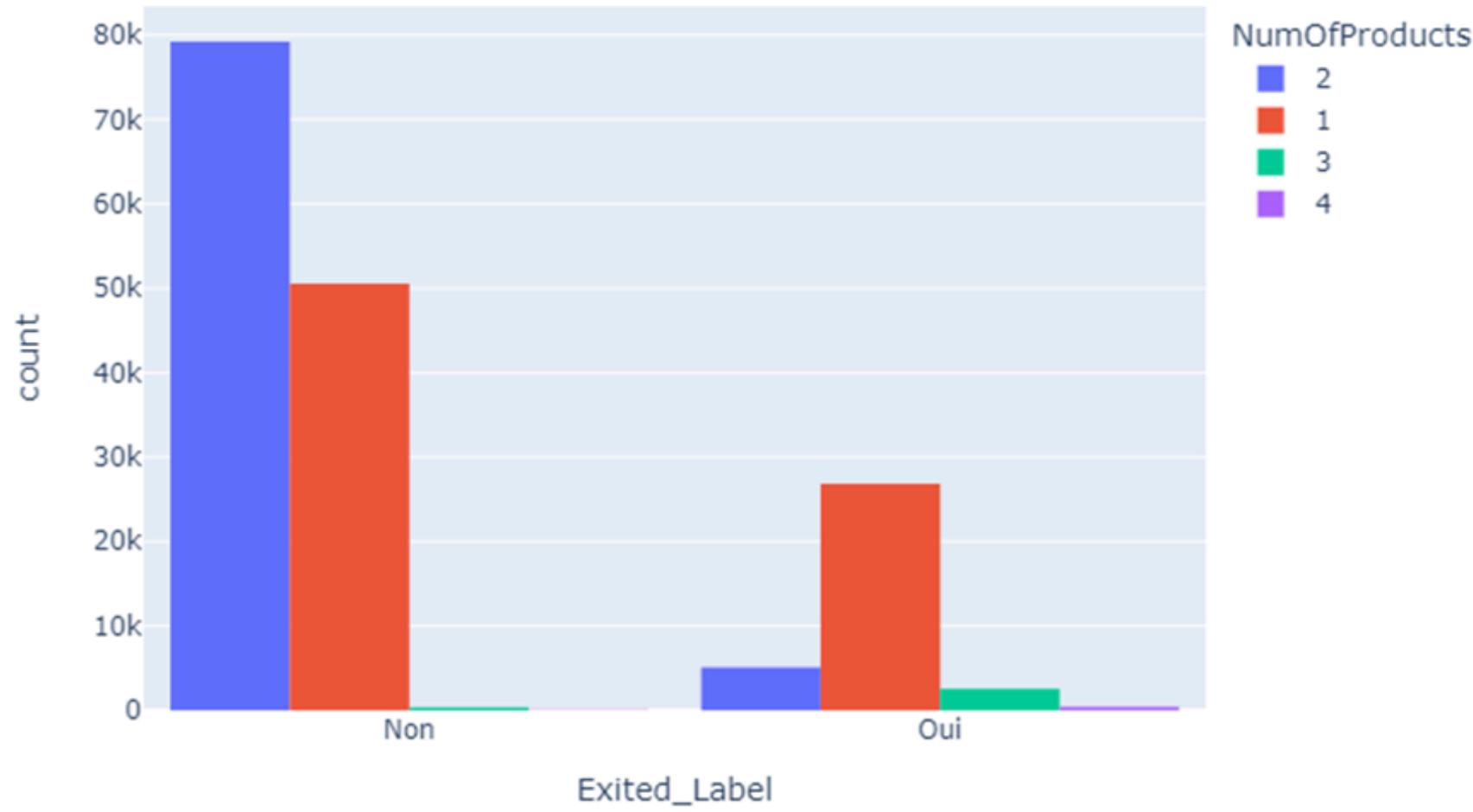


Membre actif

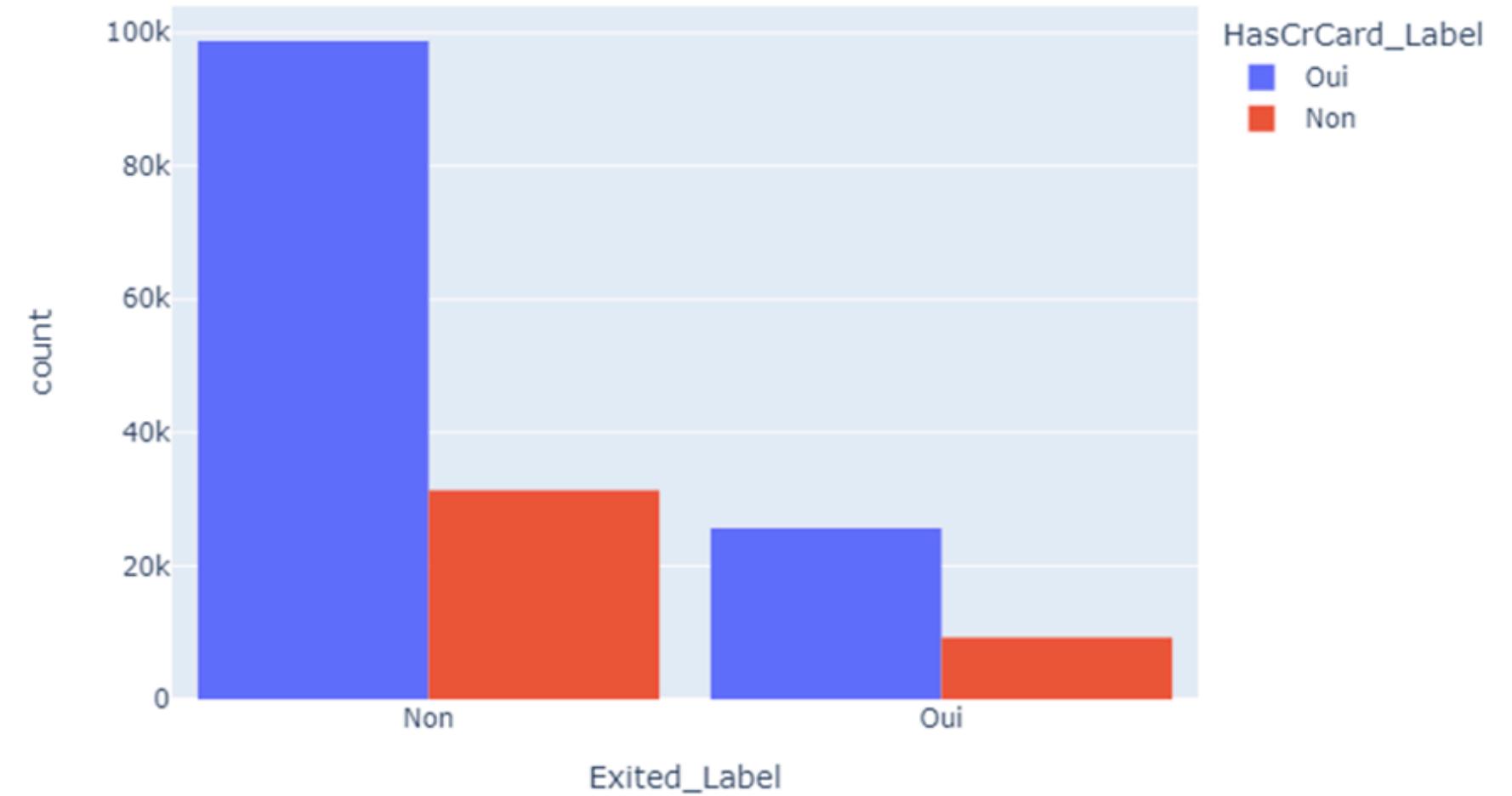
- La plupart de ceux qui quittent la banque n'était pas des clients actifs
- La majorité des membres restants sont des clients actifs

ANALYSE EXPLORATOIRE

Nombre de produits souscrits



Carte de crédit



Produits souscrits

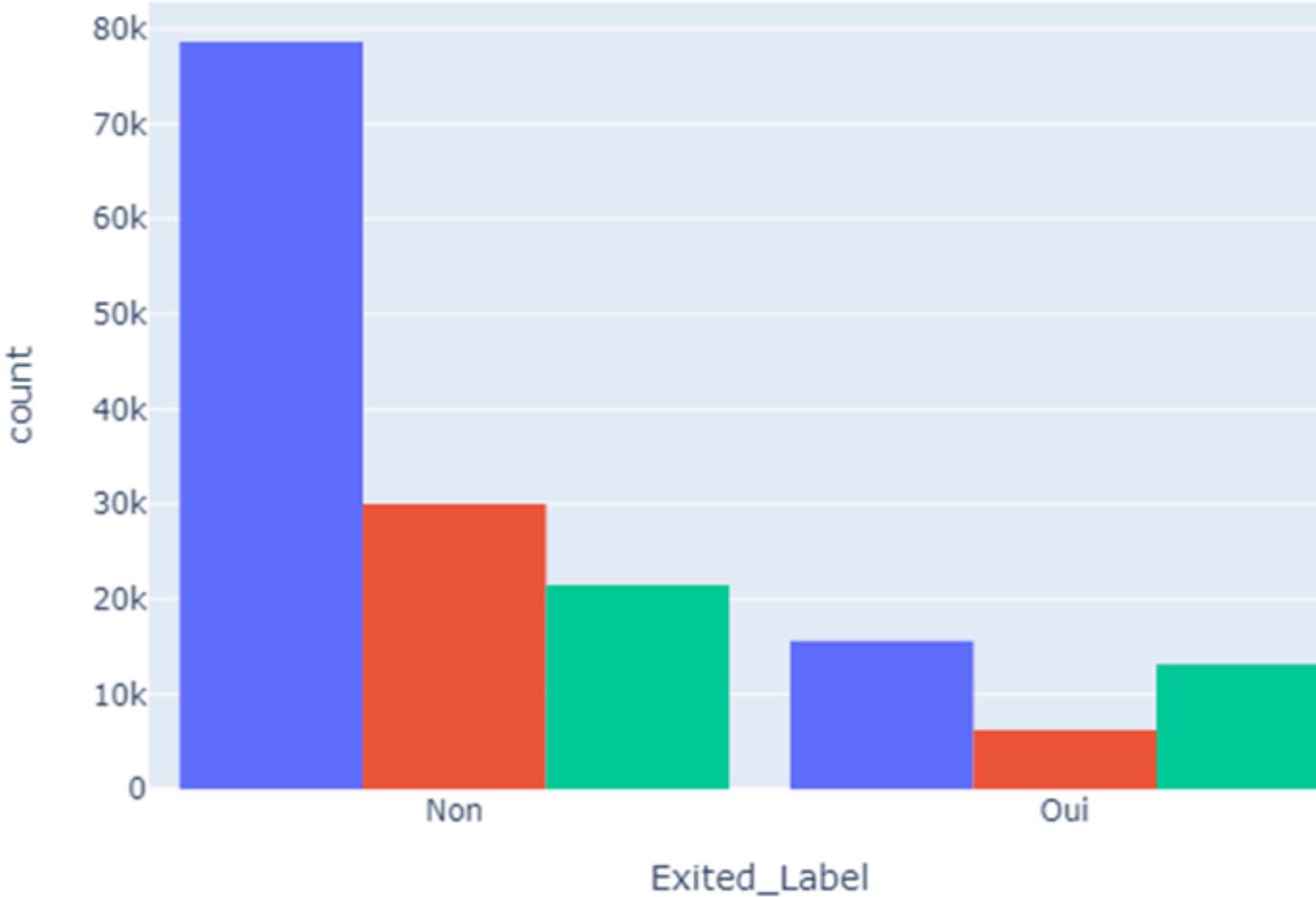
- La plupart de ceux qui ont quitté la banque avaient un seul produit

Membre actif

- Que le client ait une carte de crédit ou non ne semble pas influencer sa décision de quitter ou pas

ANALYSE EXPLORATOIRE

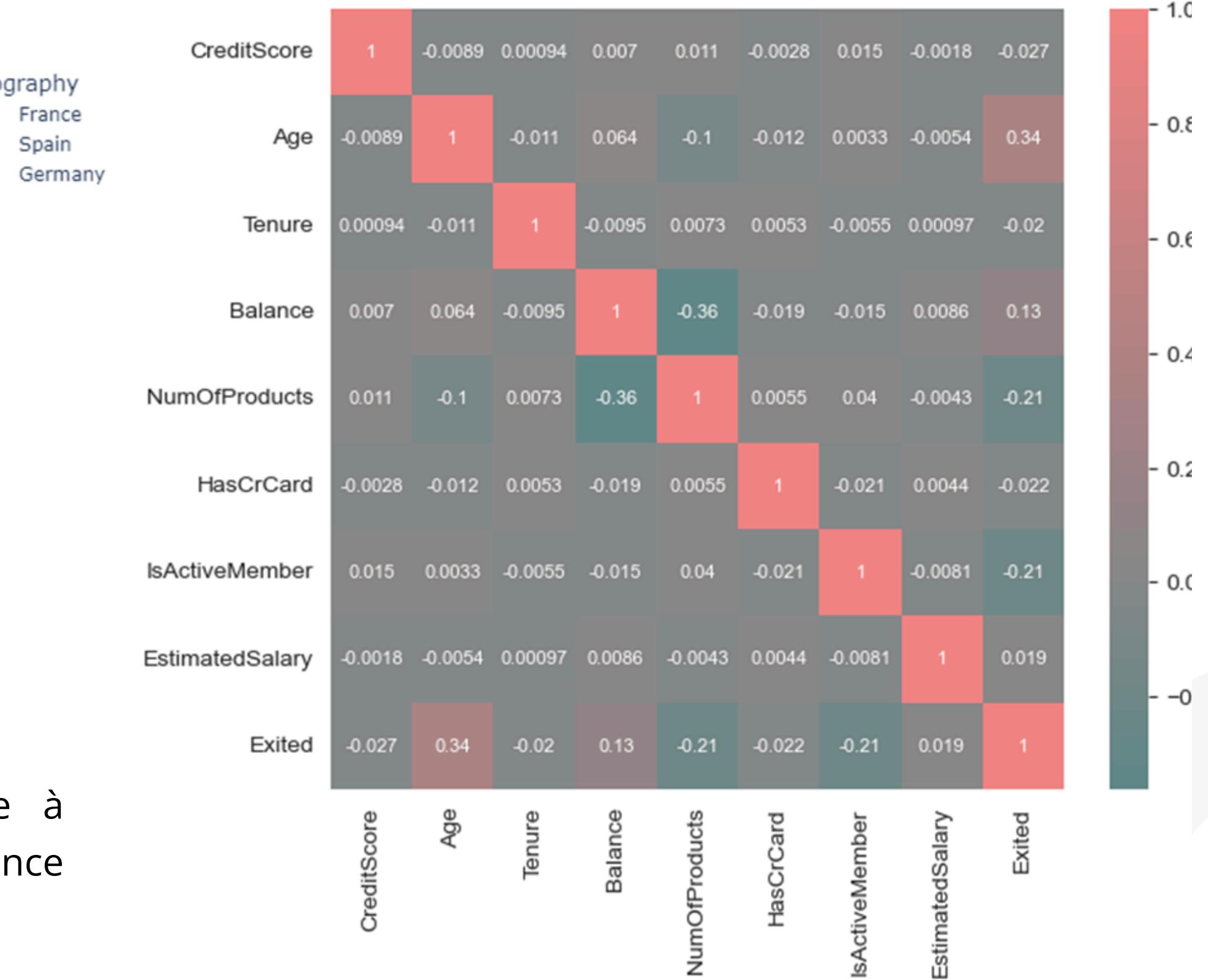
Geography



Pays d'origine des clients

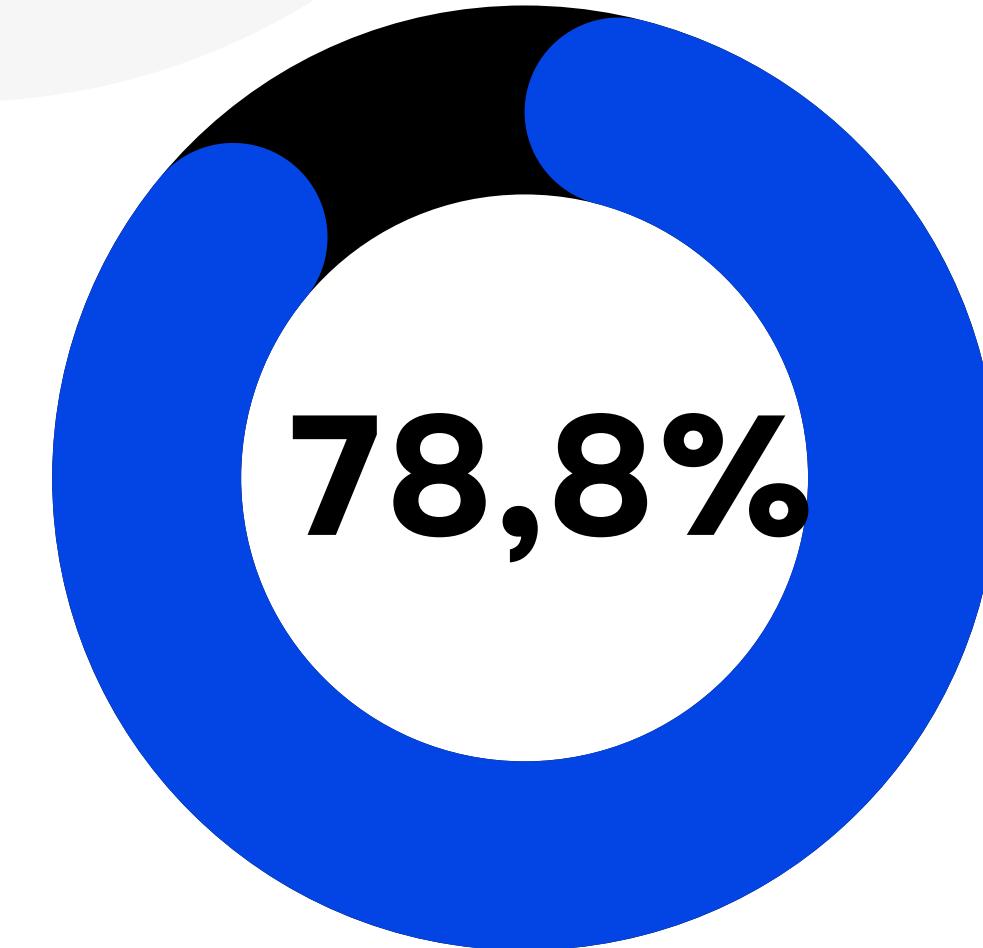
- Les clients d'Allemagne ont plus tendance à quitter que ceux d'Espagne, qui ont plus tendance à rester

Matrice de corrélation des variables

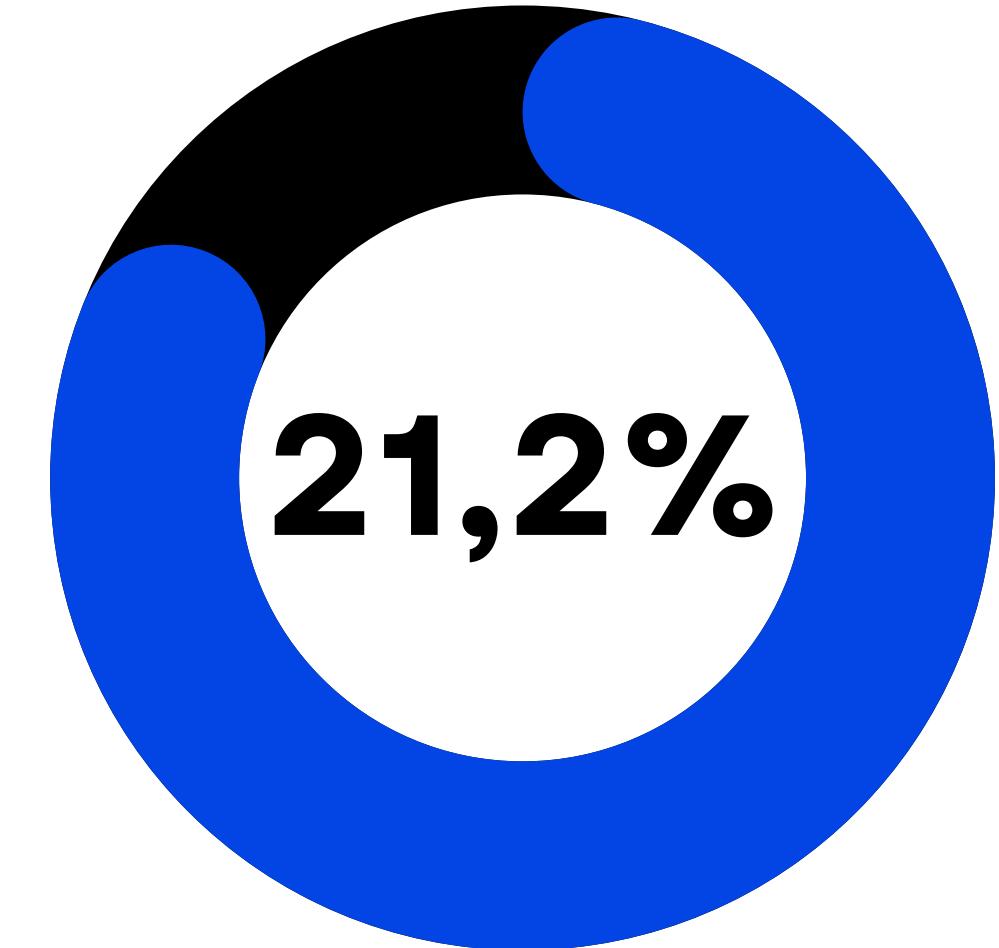


ANALYSE EXPLORATOIRE

Ceux qui n'ont pas quitté



Ceux qui ont quitté



- Un déséquilibre de la variable cible "Exited"



RESULTATS DES MODELES

DESCRIPTION DES MODÈLES

—

Afin de faire de meilleures prédictions nous avons testés plusieurs modèles d'apprentissage.

Gaussian Naive Bayes

Il s'agit d'un type d'algorithme de classification probabiliste basé sur le théorème de Bayes.

Logistic Regression

un modèle d'apprentissage automatique utilisé pour la classification binaire ou multiclasse. l'algorithme apprend à prédire la probabilité qu'une observation appartienne à une classe spécifique en utilisant une fonction logistique

KNN

Il s'agit d'un algorithme de classification qui attribue une étiquette de classe en fonction des étiquettes des K plus proches voisins dans l'espace des caractéristiques.

DESCRIPTION DES MODÈLES



Random forest

Il s'agit d'un ensemble de plusieurs arbres de décision formant une forêt. Les prédictions sont faites par agrégation des résultats de plusieurs arbres.

DecisionTreeclassifier

C'est un algorithme de machine learning supervisé utilisé pour les tâches de classification. C'est un schéma ayant la forme d'un arbre, qui présente les données possibles d'une série de choix interconnectés.

Perceptron

Algorithme d'apprentissage supervisé simple pour les tâches de classification binaire. Utilise une seule couche neuronale pour prendre des décisions basées sur une fonction linéaire.

XGBoost

Il s'agit d'un algorithme d'apprentissage automatique basé sur le boosting gradient pour améliorer la précision de la prédiction. Il fournit un boost d'arbre parallèles pour créer un modèle fort et précis.

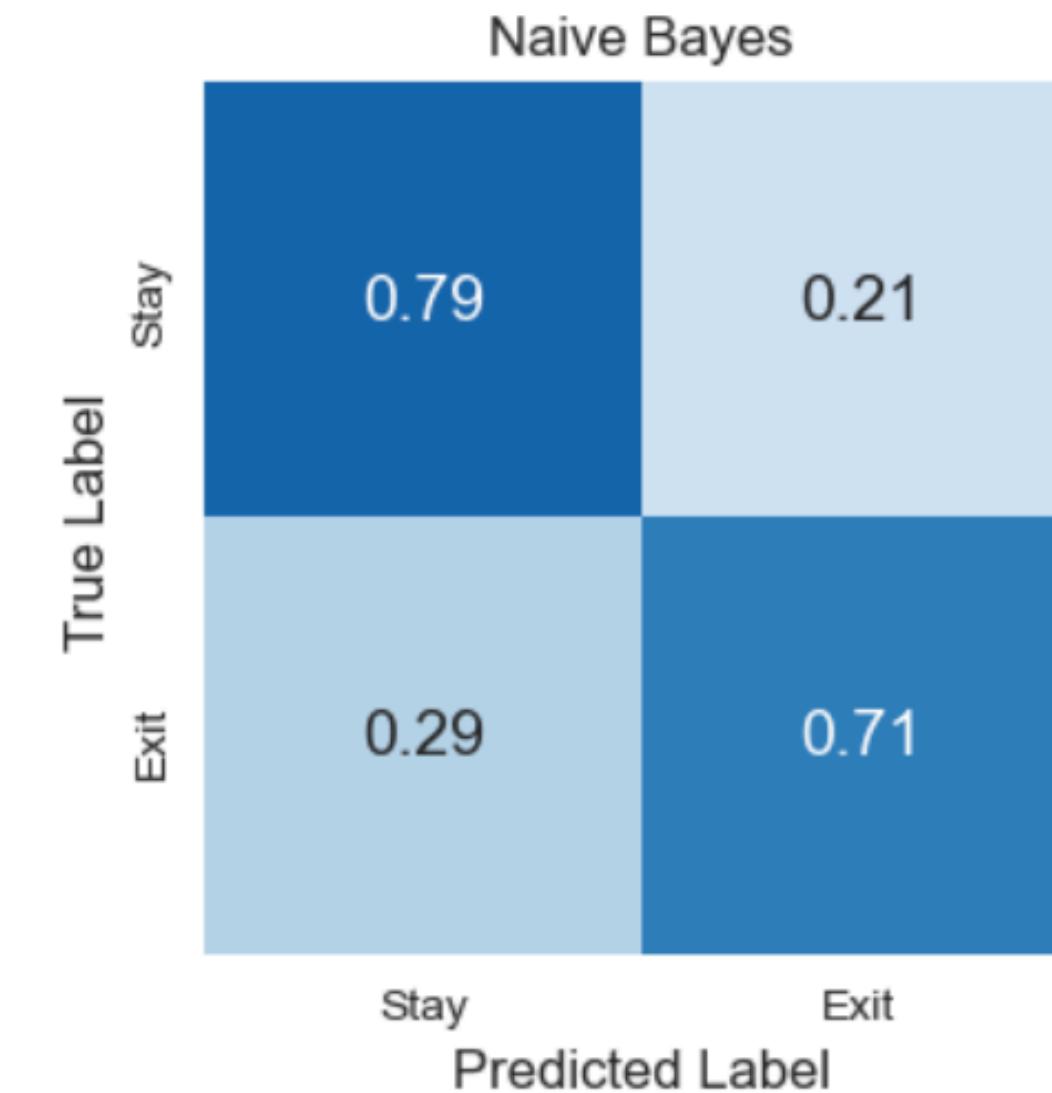
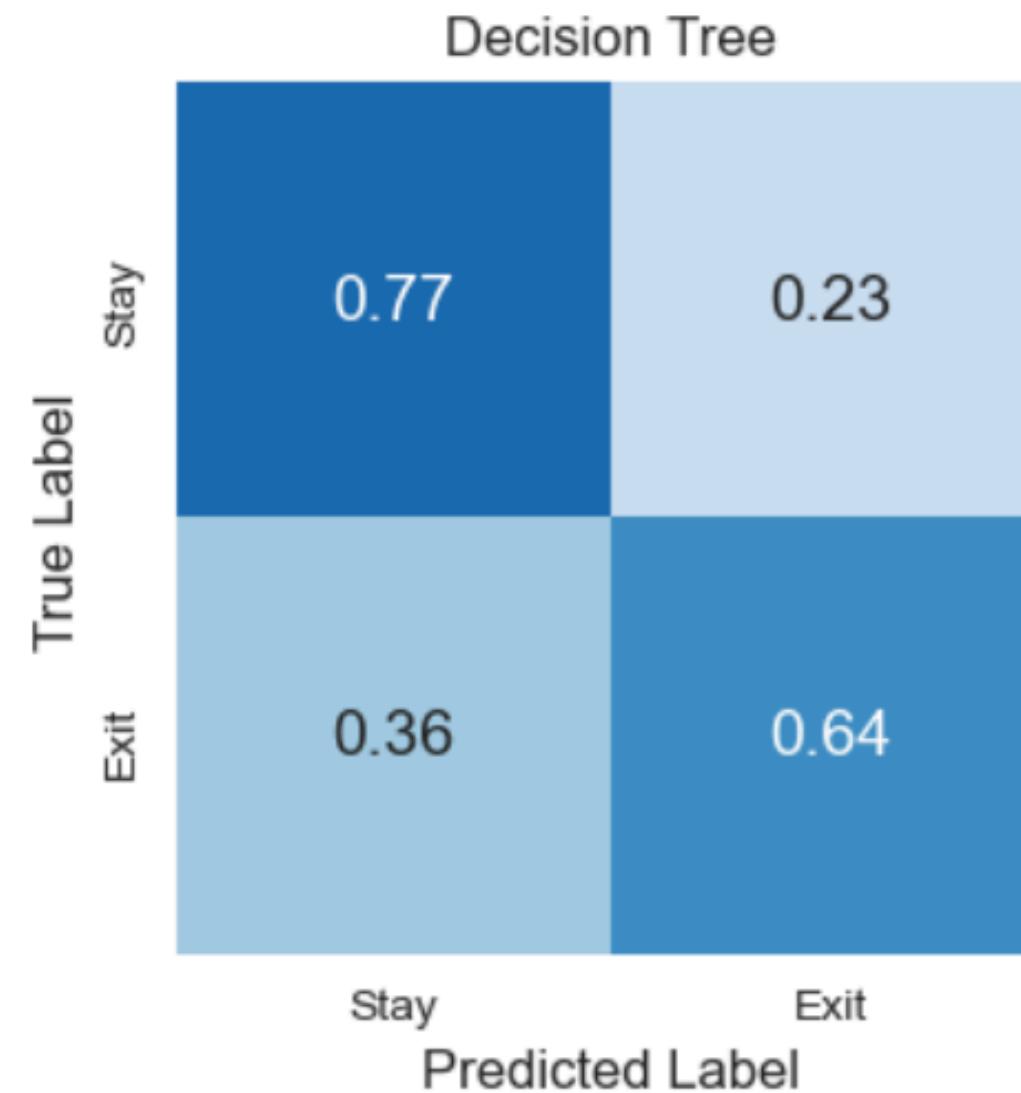
PERFORMANCE DES MODÈLES



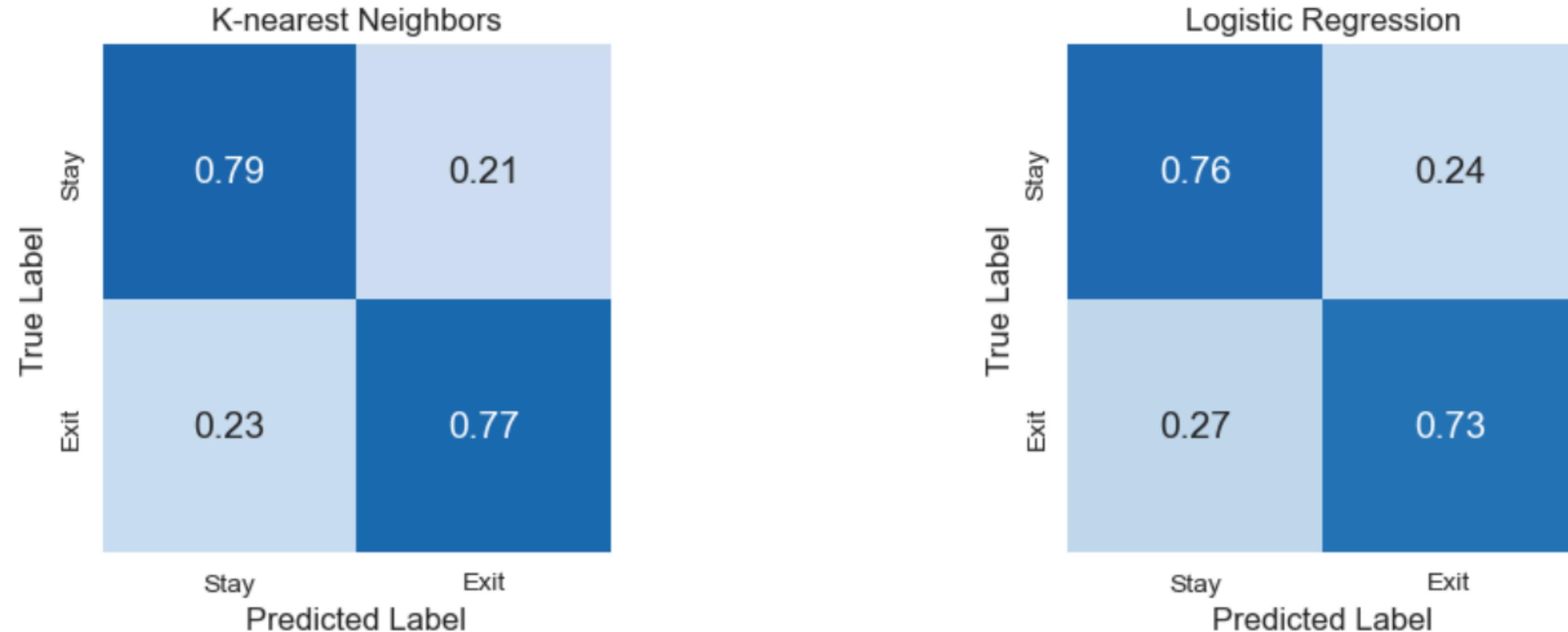
Dans notre contexte, il est généralement plus crucial de minimiser les faux négatifs (clients à haut risque de churn que le modèle ne détecte pas) que de minimiser les faux positifs (clients que le modèle prédit comme churn mais qui ne le sont pas en réalité). Cela est dû au fait que rater un client qui va partir peut entraîner une perte significative de revenus et d'opportunités de vente croisée ou de fidélisation. Dans le calcul du F2-Score, il est donc question d'accorder plus de poids au Recall.

PERFORMANCE DES MODÈLES

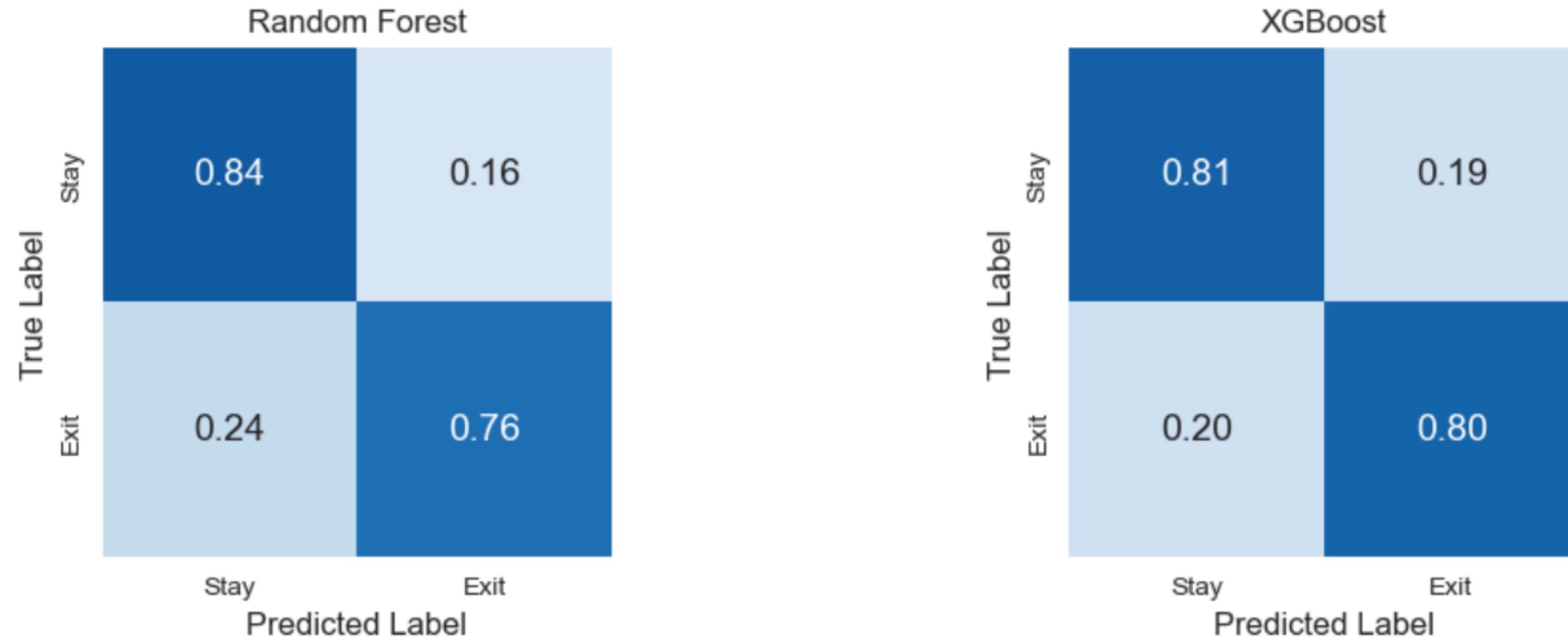
Normalized Confusion Matrices



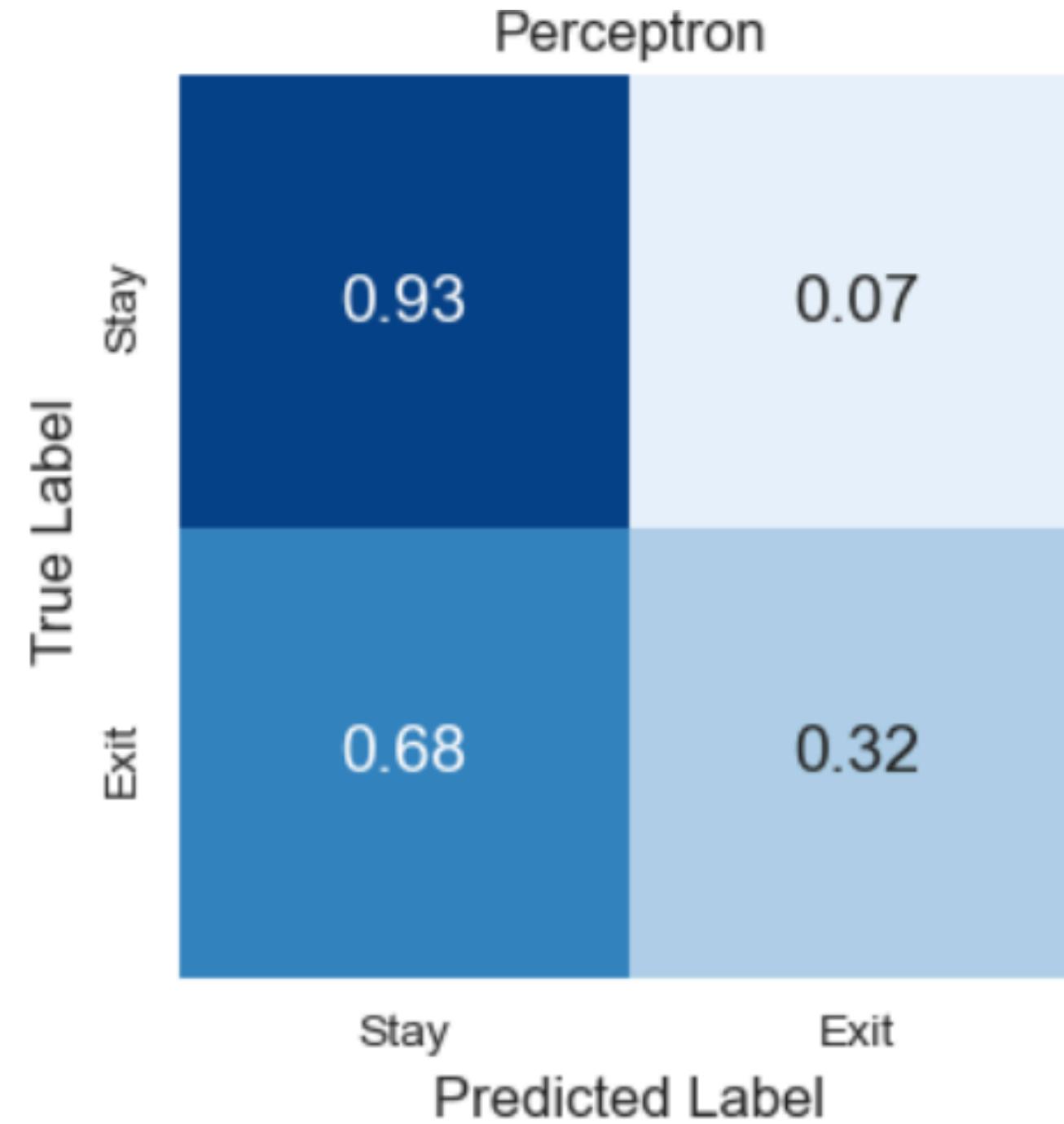
PERFORMANCE DES MODÈLES



PERFORMANCE DES MODÈLES



PERFORMANCE DES MODÈLES





CHOIX DU MEILLEUR MODELE

CHOIX DU MEILLEUR MODELE

	precision	recall	f1_macro	f2_macro	accuracy
--	-----------	--------	----------	----------	----------

model

dt	0.659657	0.706353	0.670729	0.688337	0.744663
----	----------	----------	----------	----------	----------

nb	0.694022	0.749867	0.708978	0.729532	0.774899
----	----------	----------	----------	----------	----------

knn	0.710264	0.779665	0.726431	0.752480	0.783341
-----	----------	----------	----------	----------	----------

lr	0.686356	0.746941	0.695166	0.719852	0.755913
----	----------	----------	----------	----------	----------

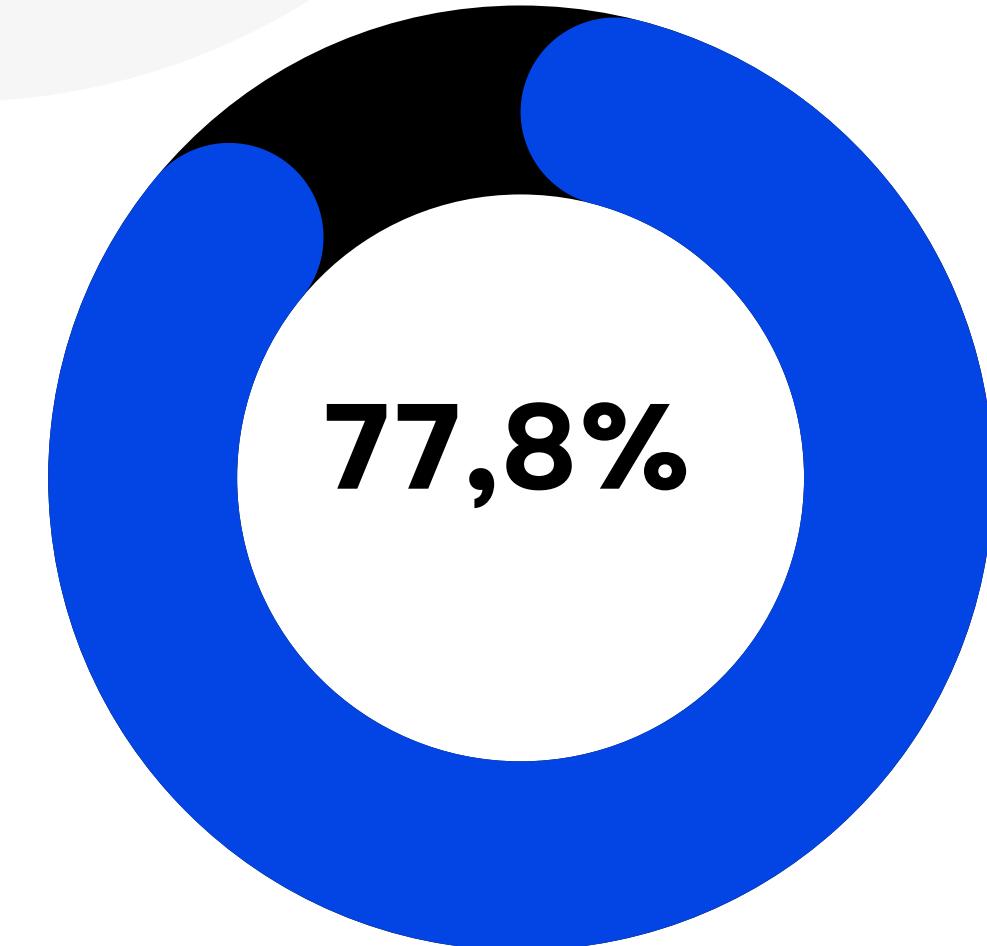
rf	0.740130	0.795943	0.758982	0.778492	0.819030
----	----------	----------	----------	----------	----------

xgb	0.732978	0.804156	0.752248	0.778295	0.806487
-----	----------	----------	----------	----------	----------

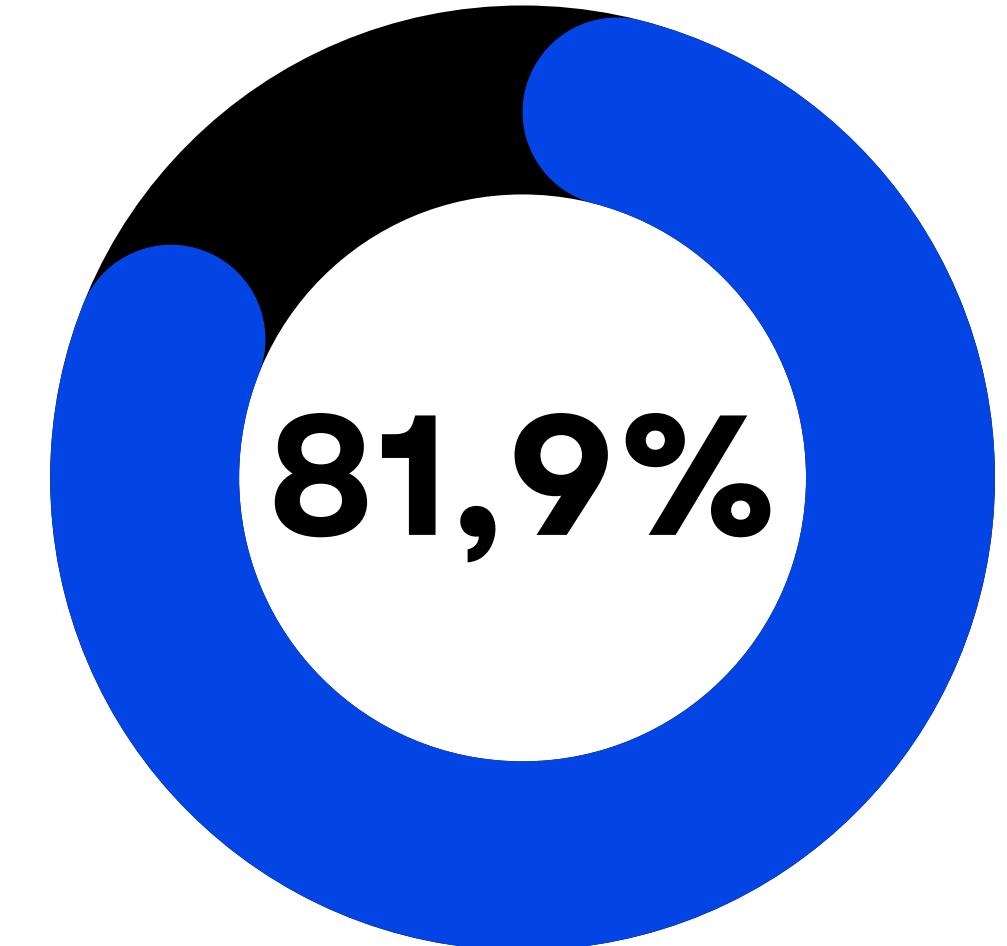
pcp	0.693766	0.626023	0.643716	0.630490	0.801075
-----	----------	----------	----------	----------	----------

CHOIX DU MEILLEUR MODELE

F2-Score



Accuracy



Random Forest

THANK YOU

FOR YOUR ATTENTION

GROUPE 12

