CHURN PREDICTION

Junior Entreprise Centrale Casablanca

PROBLÉMATIQUE

Notre client, **Attijari Wafabank**, une des plus grandes banques de détail de la région MENA et du continent africain, souhaite utiliser son historique de données afin de comprendre le phénomène d'attrition de ses clients (churn) après avoir constaté que 30% de ses clients quitte la banque pour une autre.

Pour cela, il nous a fourni un jeu de données comportant un certain nombre d'informations clients ainsi que le constat du départ de la banque d'un client dans les 3 mois concernant son attrition. Certaines variables sont clairement exprimées car non sensibles, d'autres sont totalement anonymisées.

Objectifs

- Produire un système permettant de générer un score de churn afin de fournir aux conseillers une liste de clients à traiter en priorité.
- Les conseillers bancaires, lorsqu'ils reçoivent un client à traiter souhaiterait également comprendre les raisons qui l'amène à quitter la banque afin d'adapter leur discours commercial.

ANALYSES STATISTIQUES

Inspection des données de la banque

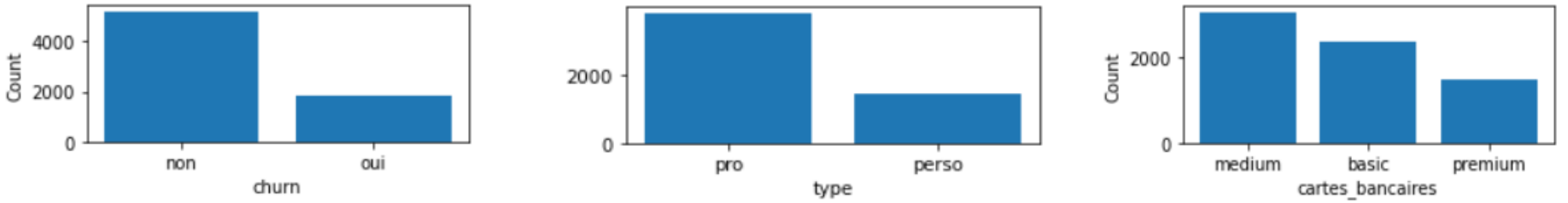

Figure 1. Le dataset n'est pas équilibré. Il y a environ **2000 churners contre plus que 4000 non-churners**. Ce qui est anormal dans le cas de l'activité d'une banque saine (30 % du churn); Les clients professionnels sont les plus majoritaires, il y a presque plus que **50%** de comptes professionnels par rapport aux comptes personnels; La majorité des clients possède une carte medium.

LES FACTEURS LIÉS AU CHURN

Exploratory Data Analysis - EDA

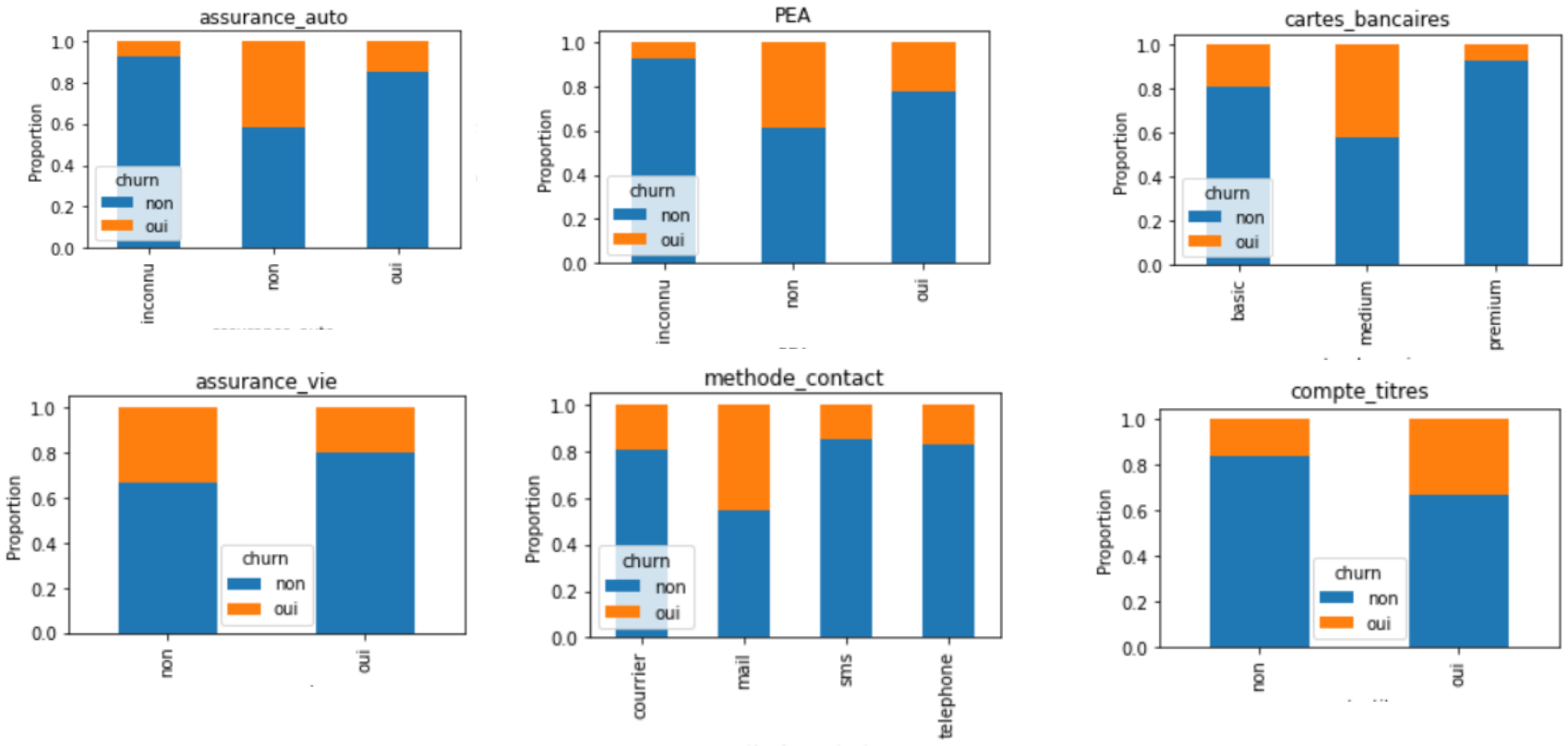

Figure 2 . Les clients **sans assurance auto, sans assurance vie, sans PEA** et **avec des comptes titres** et des cartes medium ont beaucoup plus tendance à churner. Les **clients contactés souvent par mail sont les plus susceptibles à manifester le churn.**

Ce sont, entre autres, les caractéristiques des clients professionnels !

LES FACTEURS LIÉS AU CHURN

Exploratory Data Analysis - EDA

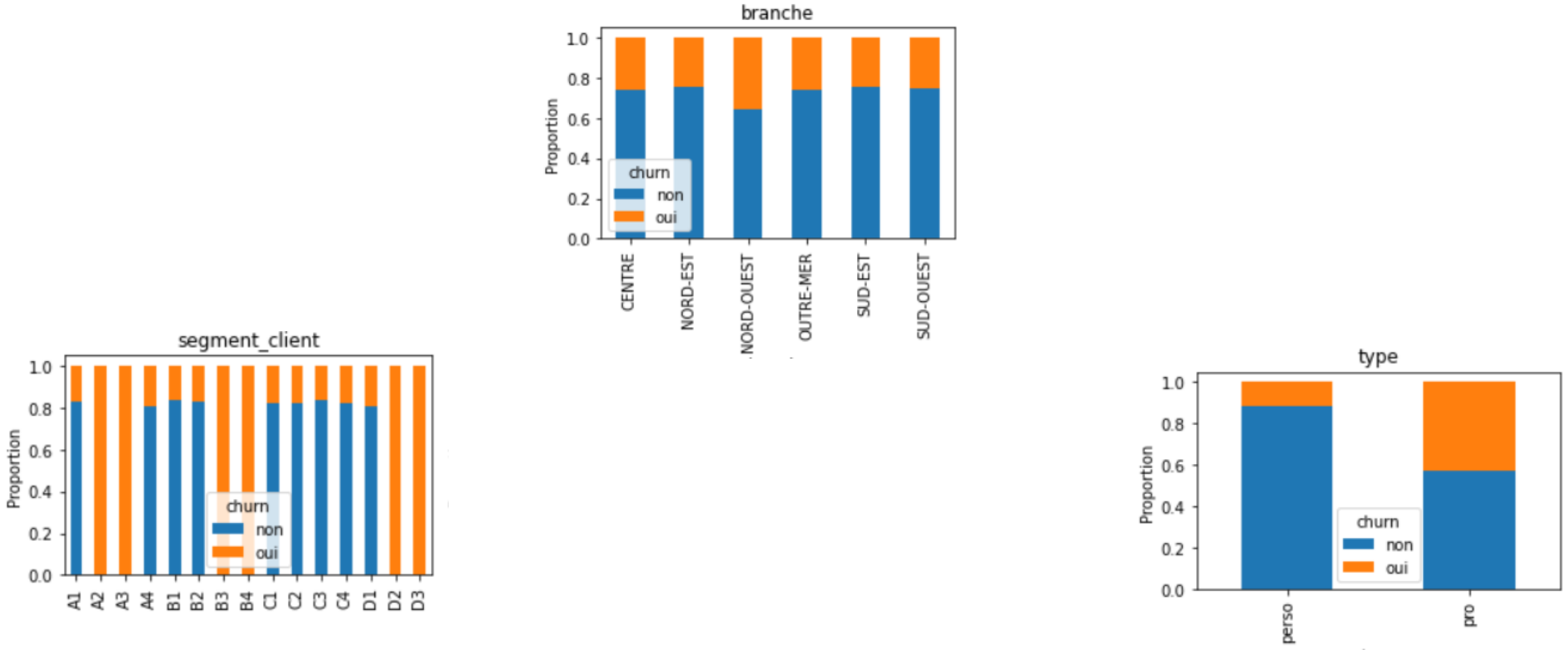

Figure 3. La **Branche Nord-Ouest de la banque possède le taux de churn le plus élevé**; Dans les segments client minoritaires de la banque (A2, A3, B3,B4, D2, D3),**100% des clients sont des churners**; **50% des clients professionnels sont des churners**.

POURQUOI LES PROFESSIONNELS CHURNENT LE PLUS ?

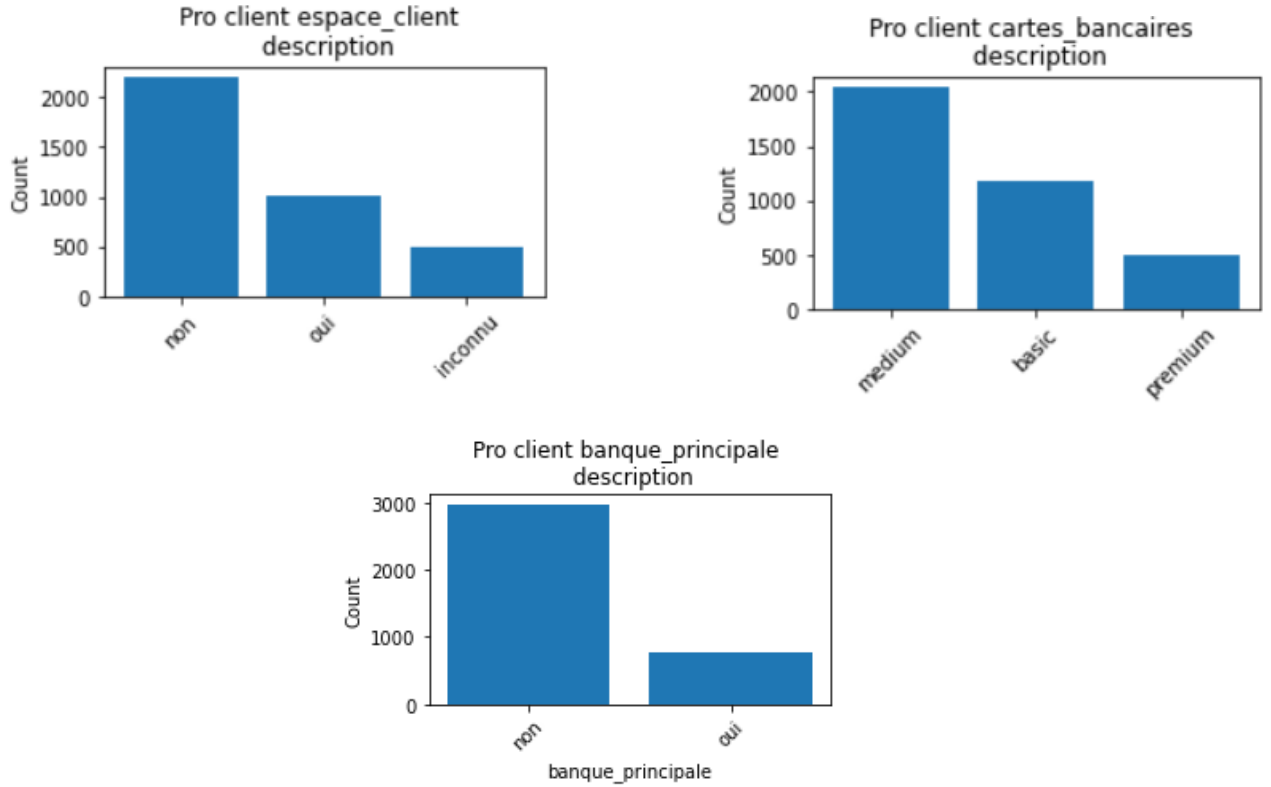

Figure 4. Les clients professionnels disposent généralement d'une carte medium. **Peut-être les services offerts ne sont pas compatibles avec ce segment de clients?** Les clients professionnels sont les clients majoritaires de la **banque bien que la banque étudiée ne soit pas leur banque principale**. Ainsi, **la plupart des clients professionnel n'ont pas un espace client**. Cela peut dégrader le relation client et causer le churn.

COMMENT EXPLIQUER LE TAUX DE CHURN ÉLEVÉ DANS LA BRANCHE NORD-OUEST ?

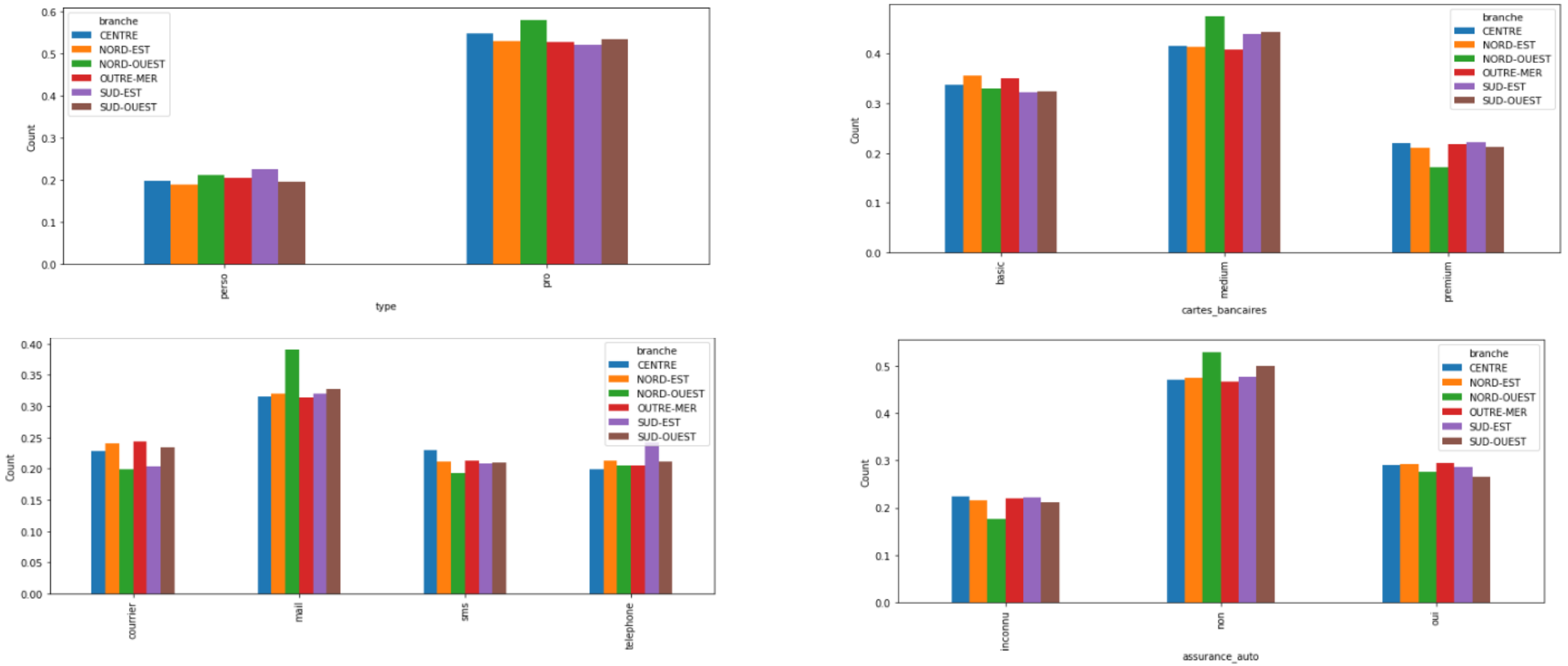

Figure 5. Le taux de churn élevé dans la branche nord-ouest s'explique par la présence de beaucoup plus de clients professionnels par rapport aux autres branches.

D’AUTRES FACTEURS LIÉS AU CHURN

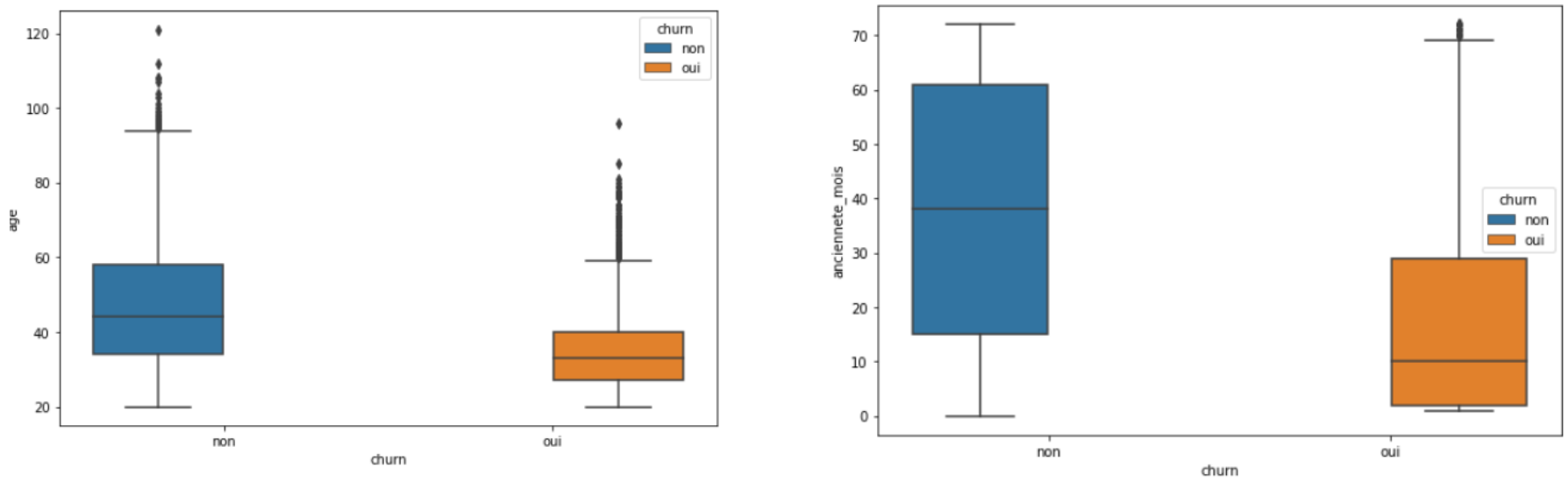

Figure 6. Plus l'ancienneté des clients est élevée, plus il est moins probable qu’ils churnent. **50% des churners ont une ancienneté de moins de 10 mois**. L'âge est facteur de churn. **Les clients au-dessous de 35 ans ont une probabilité de churn presque sûr.**

MODÉLISATION

Feature Selection

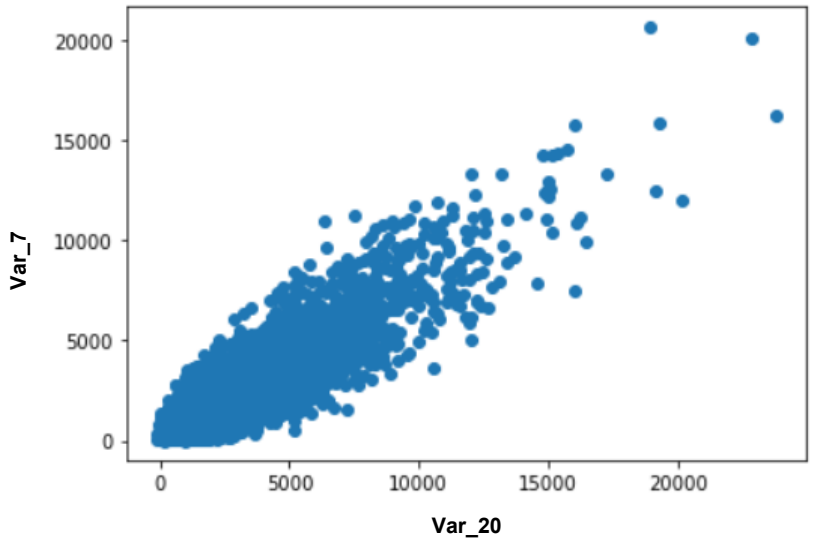

Figure 7. La corrélation entre les variables var_7 et var_20 est linéaire et fort.

Figure 8. Matrice de corrélation

La plupart des variables continues sont corrélées. Par exemple la variable sur **l'intérêt total du compte d'épargne** est corrélée à 83 % avec **l'ancienneté du client**. Seules les variables dont la corrélation est inférieure à 0,7 ont été conservée. Cela va simplifier le traitement des valeurs manquantes et réduire la dimensionnalité du dataset.

MODÉLISATION

Analyse des Valeurs Manquantes

- Pour les variables catégorielles, les valeurs manquantes ont été considérées comme une classe supplémentaire.
- Pour les variables continues, on a utilisé l'imputation par KNN.
- Pour chaque variables continues, on a entrainé un KNN pour prédire ses valeurs manquantes à partir des variables catégorielles.
- On n'a réussi à obtenir un bon modèle KNN d'imputation que pour une seule variable qui est « agios_6mois ».
- var_7 a été imputée en utilisant une régression linéaire à partir de la variable var_20 étant donné la corrélation forte et linéaire entre les deux variables.
- Après imputation, on a supprimé les valeurs manquantes restantes qui sont en quantité très négligeables par rapport à la taille des données traitées; 6627 contre 7043 du dataset de départ.

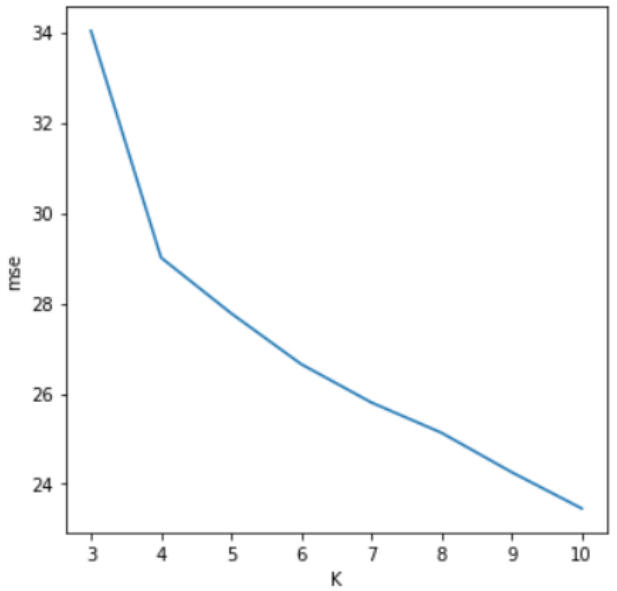

Figure 9. Méthode « Elbow » pour le choix de k lors de l'entrainement du KNN pour imputer les valeurs manquantes de la variable « agios_6mois ». Le modèle a une performance de R2 = 0,96. ce qui témoigne de la qualité d'imputation.

MODÉLISATION

Entrainement du Modèle de Scoring du Churn

Modèles	Précision	Recall	Roc Auc	F1 Score
Régression Logistique	0.83	0.69	**0.93**	0.75
Arbre de Décision	0.69	0.46	**0.80**	0.55
Forêt Aléatoire	0.90	0.28	**0.89**	0.43
Gradient Boosting	0.85	0.63	**0.92**	0.72

Tableau 1. Métriques d'évaluation sur les données d'entrainement.

Modèles	Précision	Recall	Roc Auc	F1 Score
Régression Logistique	0.84	0.67	**0.93**	0.75
Arbre de Décision	0.71	0.46	**0.81**	0.56
Forêt Aléatoire	0.93	0.32	**0.89**	0.47
Gradient Boosting	0.87	0.62	**0.92**	0.73

Tableau 2. Métriques d'évaluation sur les données de test. Le roc auc est la métrique la plus fiable car il ne dépend d'aucun treshold de classification.

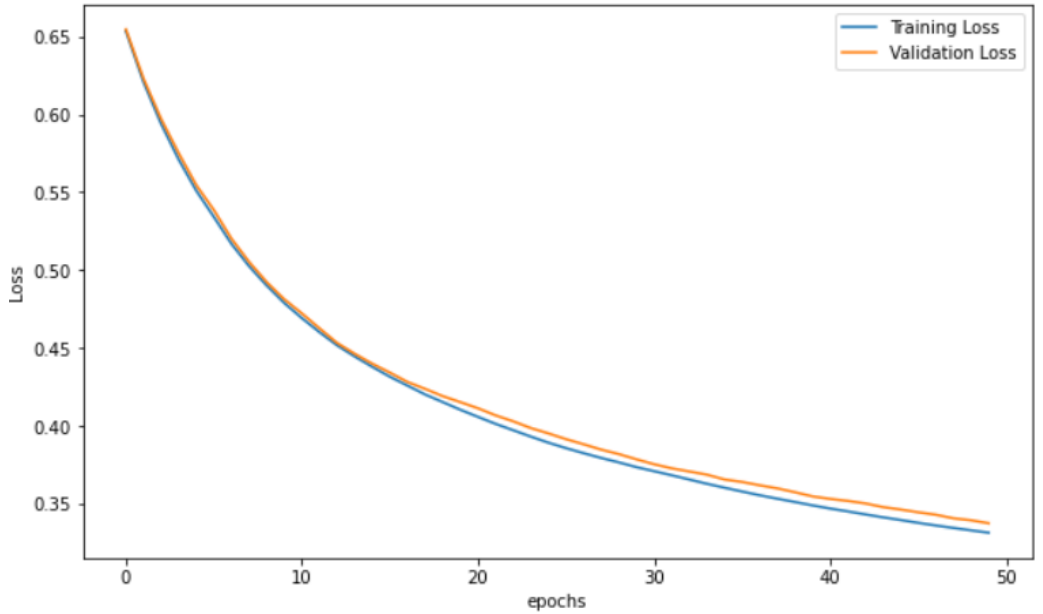

Figure 10. Courbe d'apprentissage du Gradient Boosting. Le modèle converge et est stable. L'Overfitting est presque inexistant.

SELECTION DU MODÈLE

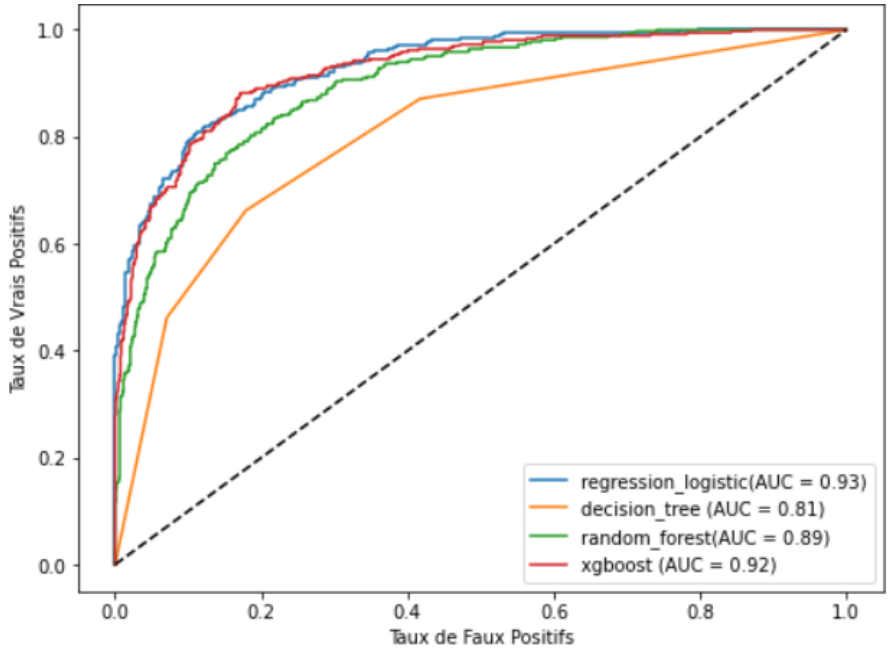

Figure 11. Courbe ROC des modèles entrainés1

- La régression logistique et le Gradient Boosting sont les modèles les plus robustes.
- La précision est la métrique qui nous intéresse car il reflète la capacité du modèle à prédire les churners (1). On peut choisir une treshold maximisant la précision. Néanmoins, le Recall est aussi important si le coût opérationnel de targetter un client, qui ne va pas churner, est élevée. Cela dépend des objectifs business et des mécanismes du métier.
- Le Gradient Boosting peut atteindre une meilleure précision que la régression logistique. Cela veut dire que le modèle ne se trompe presque pas sur les clients qui auront la tendance de churner.

INTERPRÉTABILITÉ GLOBALE DU MODÈLE XGBOOST

Shapley Values explainer

- Comme résultats, les clients professionnels ont plus tendance à churner, ainsi que les clients possédant une carte medium.
- Les clients de moins de 35 ans, ainsi que les clients ayant moins de 10 mois d'ancienneté ont une probabilité de churn élevée.
- Les clients des segments D2, B4, B3, D3, A3, A2 ainsi que les clients sans compte courant sont beaucoup plus susceptibles de manifester le churn.

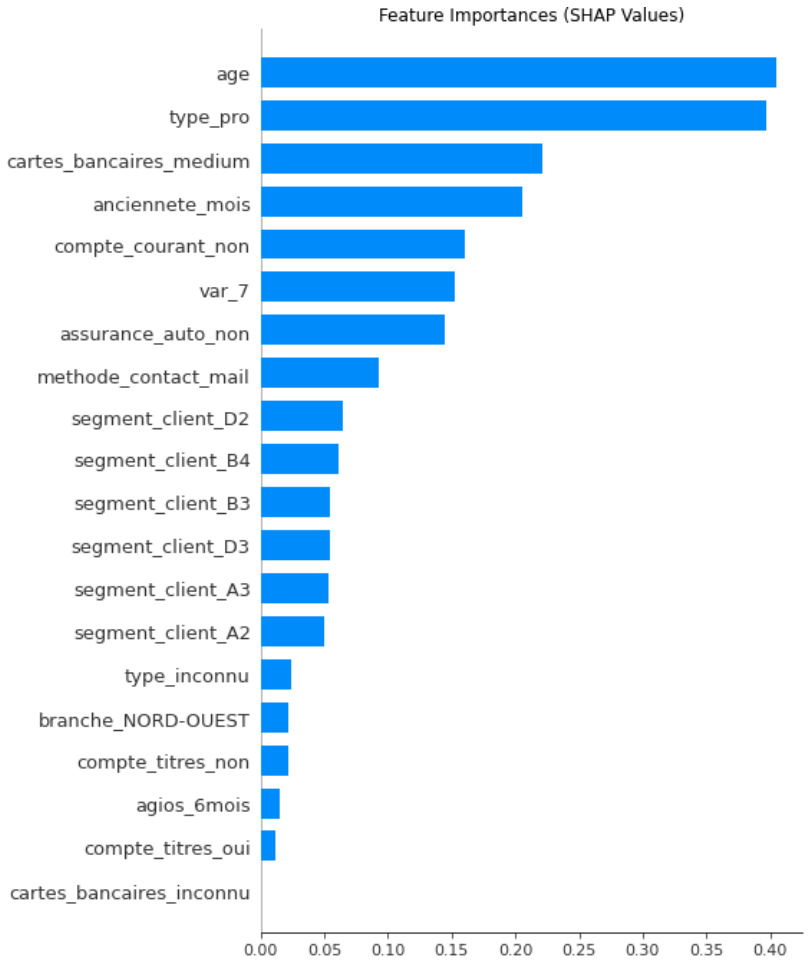

Figure 12. Shapley values global scores

THANK YOU

Omar Elghaffouli

+33 6 29 28 99 95

omar.elghaffouli@ensae.fr