Compétition Kaggle : Bank Churn Prediction

Chahla TARMOUN
Aya MOKHTAR

Table des Matières

I. Contexte général du challenge

II. Analyse exploratoire et visualisation des données

III. Préparation des données

IV. Entraînement et évaluation des modèles

V. Conclusion et Axes d’amélioration

I. Contexte Général du Challenge

1. Contexte de la compétition

Objectifs

- **Identifier** les clients susceptibles de quitter la banque.
- Permettre à la banque de **cibler ces clients** avec des actions de fidélisation adaptées.
- **Analyser les attributs** les plus influents dans la décision de churn, comme le solde, l'âge, ou l'engagement avec la banque.
- **Construire un modèle** prédictif fiable pour prédire si un client quittera la banque.

1. Contexte de la compétition

Métrique d'évaluation

- AUC (Area Under the Curve) à maximiser

Target

- Problème de classification. La cible: "Exited"

Données

Deux jeux de données:

- train.csv avec la cible
- test.csv sans la cible que nous devons prédire

2. Description du dataset

Les datasets utilisés pour ce projet contiennent des données réelles de clients bancaires :

- **Customer ID :** Identifiant unique pour chaque client (exclu de la modélisation).
- **Surname :** Nom de famille du client (exclu car sans valeur prédictive).
- **Geography :** Pays de résidence (France, Espagne ou Allemagne).
- **Gender :** Sexe du client (Homme ou Femme).
- **Age :** Âge du client.
- **Credit Score :** Indicateur numérique de la solvabilité.

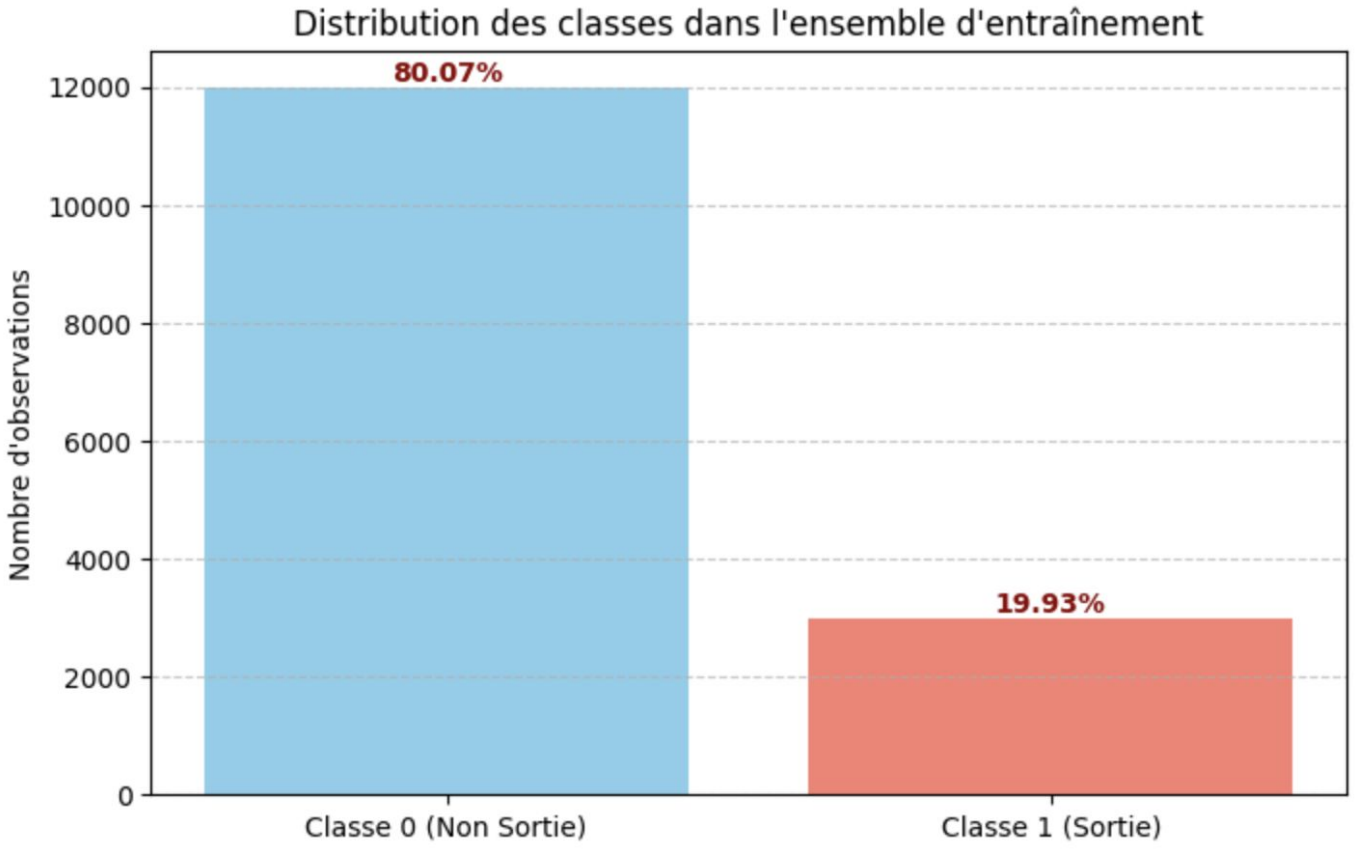
2. Description du dataset

- **Tenure :** Nombre d'années avec la banque.
- **Balance :** Solde du compte bancaire.
- **NumOfProducts :** Nombre de produits bancaires utilisés.
- **HasCrCard :** Possession d'une carte de crédit (1 = oui, 0 = non).
- **IsActiveMember :** Statut d'activité (1 = actif, 0 = inactif).
- **EstimatedSalary :** Salaire estimé du client.
- **Exited :** Indique le statut de churn (1 = client parti, 0 = client resté).

Les colonnes comme **Customer ID** et **Surname** ont été **exclues** de la modélisation car elles n'ont pas de pertinence prédictive.

II. Analyse exploratoire et visualisation des données

1. Distribution de la variable cible: Exited



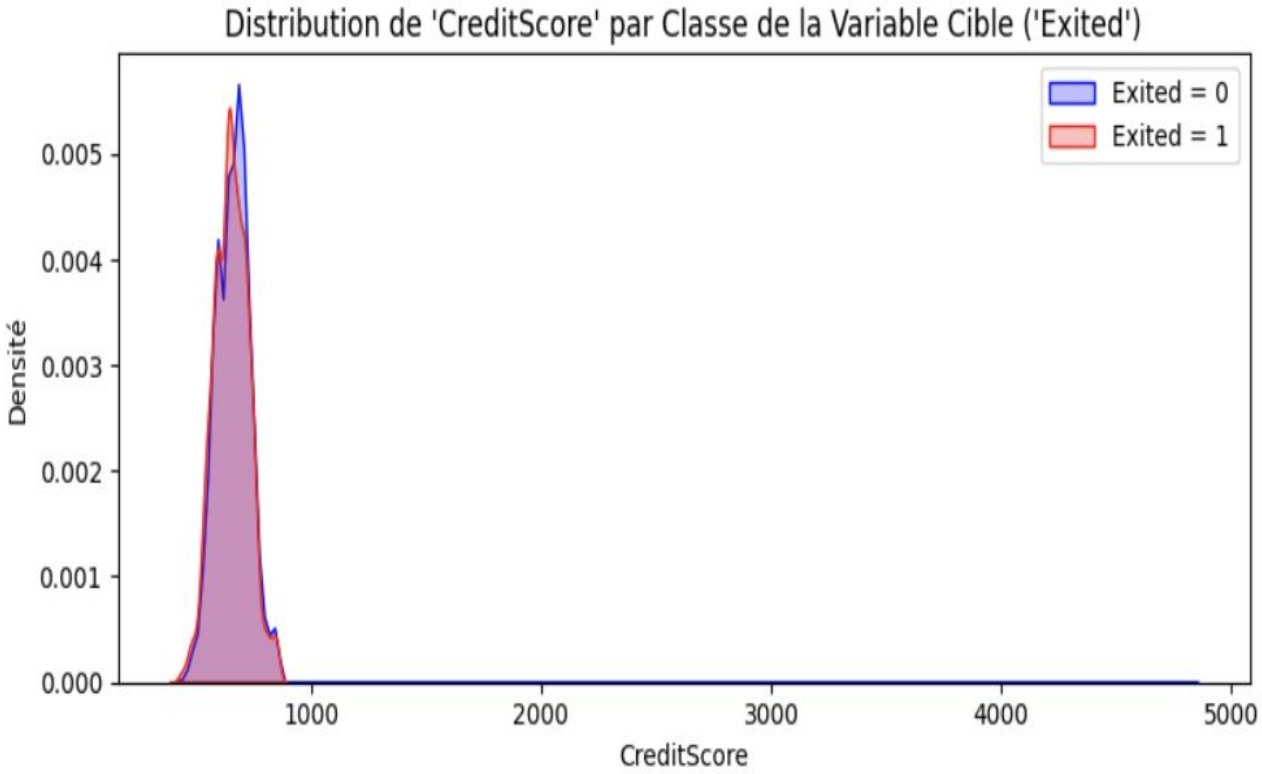
- Sortie : 80.07%
- Non Sortie : 19.93%

⇒ Un léger déséquilibre entre les deux classes, avec une prédominance de la classe "Sortie".

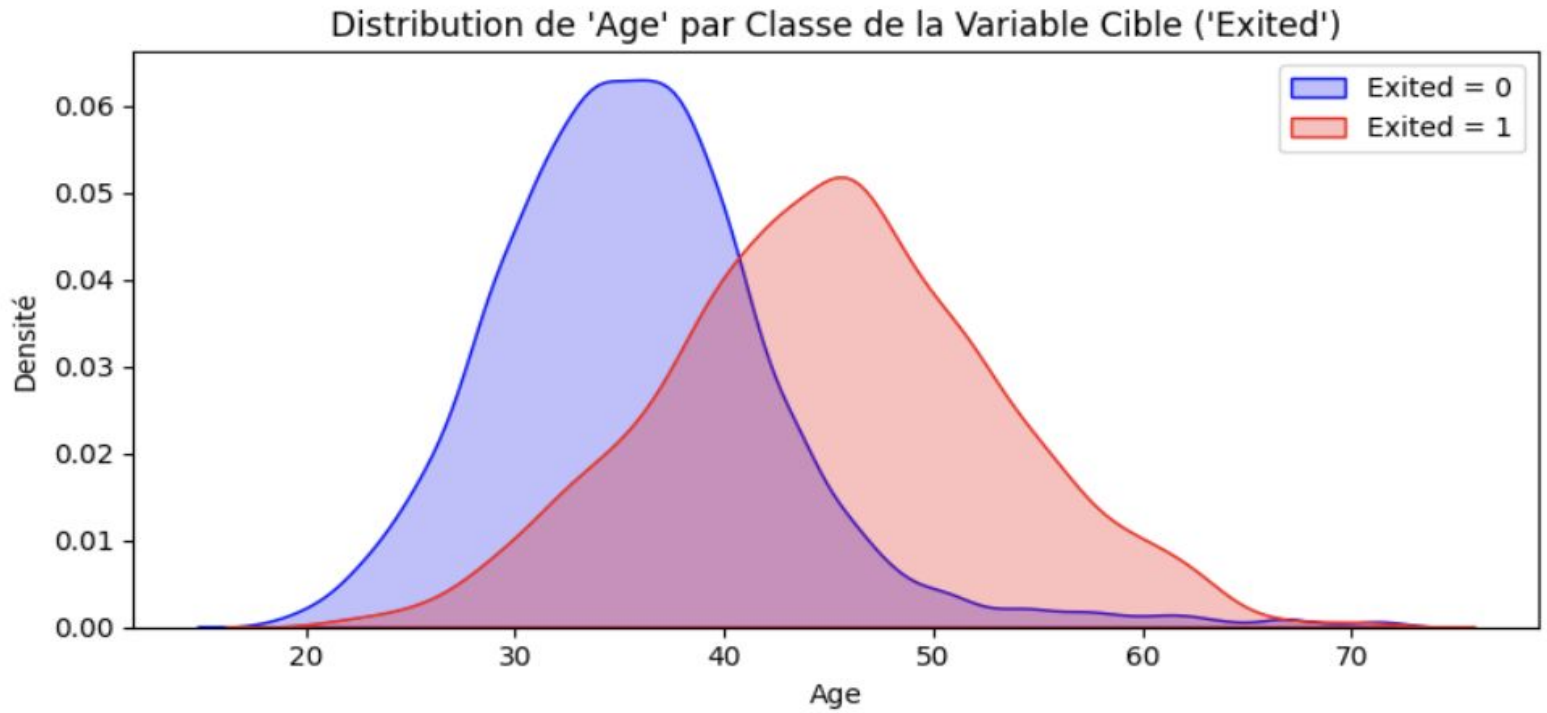
2. Variables Numériques : Score de crédit bancaire

- Distribution du score de crédit très similaires entre les deux groupes.

- Le score semble avoir un impact distinguant les clients qui quittent la banque de ceux qui ne la quittent pas, bien que cet impact soit relativement mineur.



3. Variables Numériques : Âge

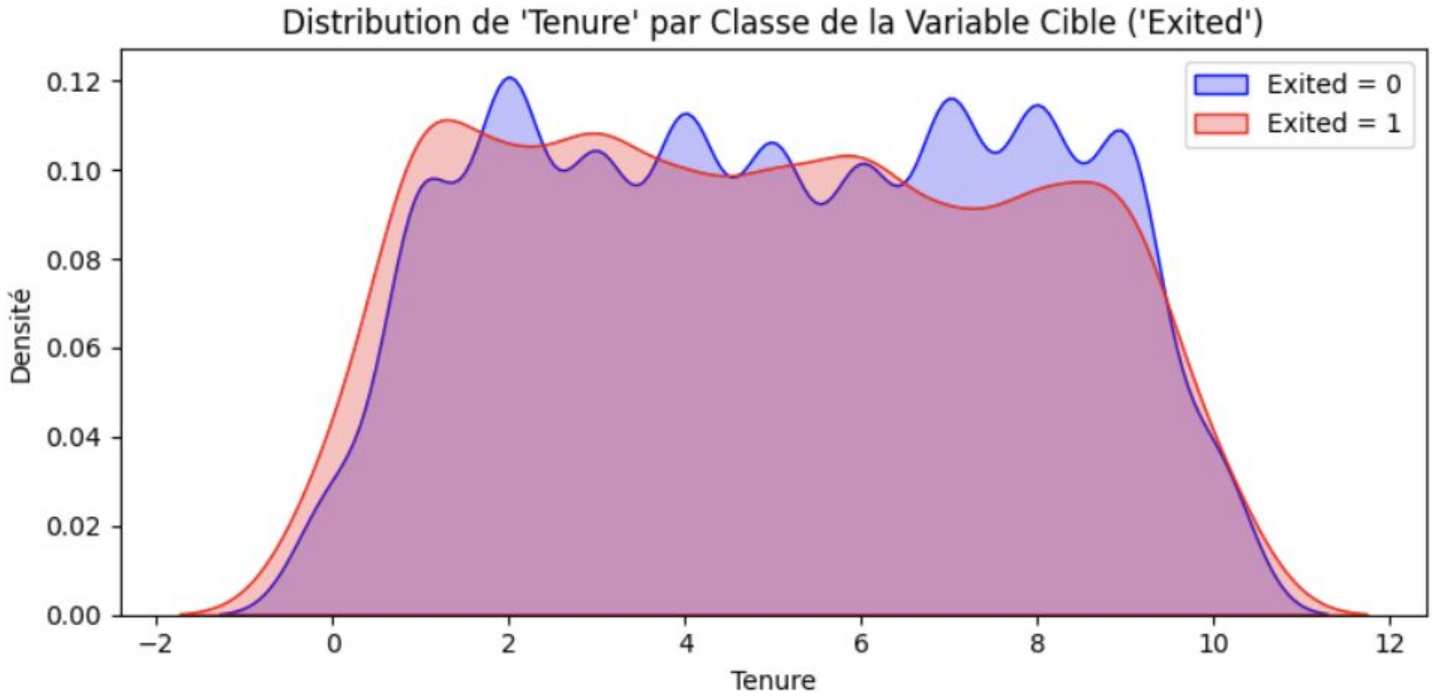


- Distribution des âges est clairement différente entre les deux groupes:

  Les clients plus âgés sont plus susceptibles de quitter la banque.

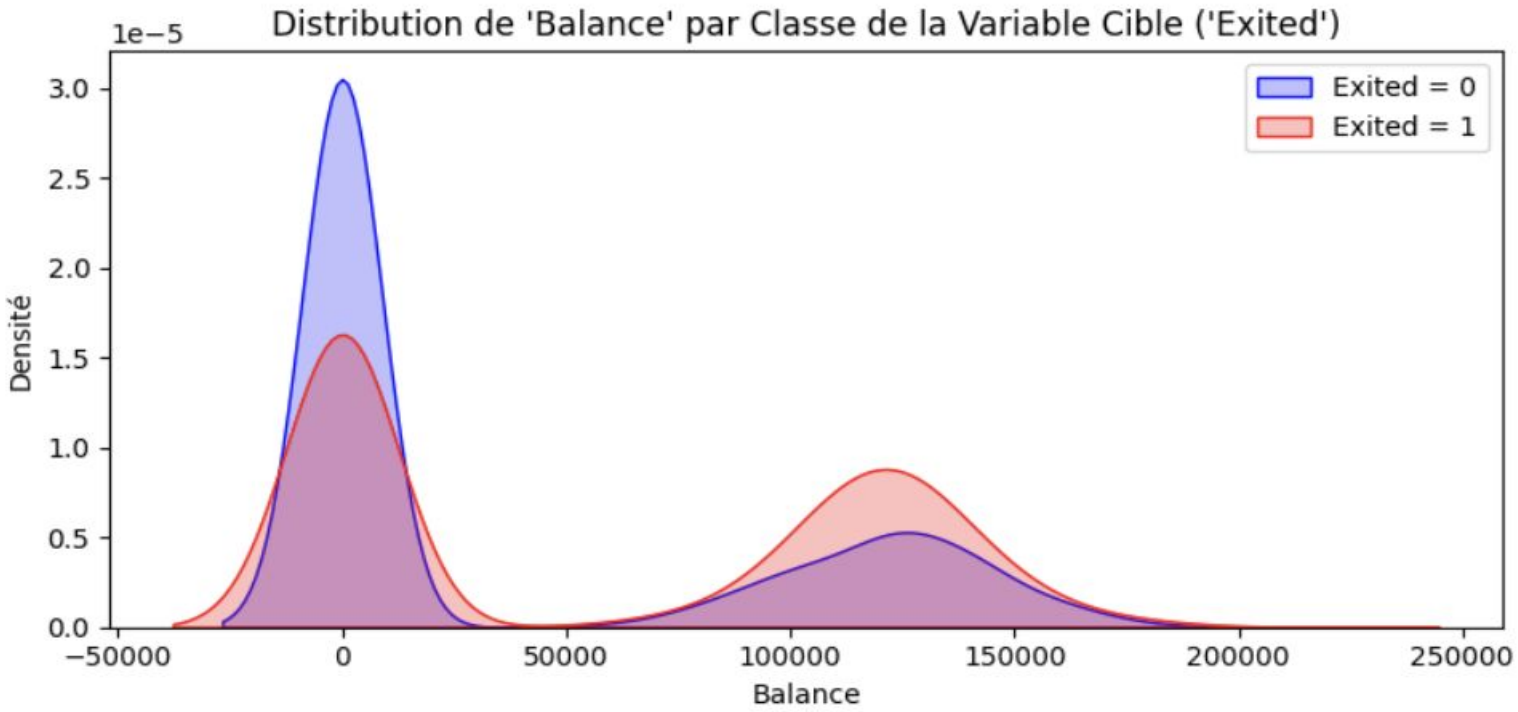
  → L'âge est une variable discriminante.

4. Variables Numériques : Ancienneté



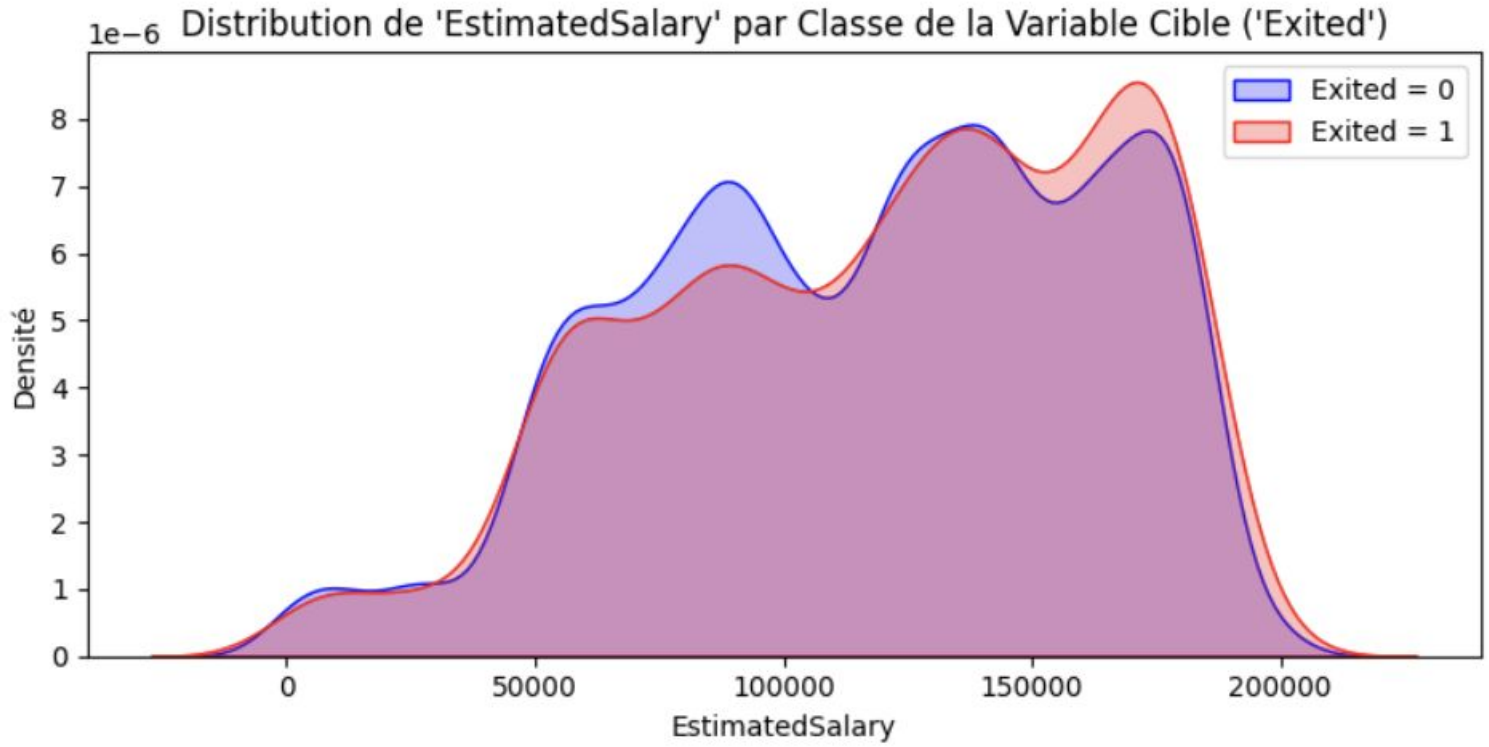
Les distributions sont très proches pour les deux groupes.

5. Variables Numériques : Solde Bancaire



- Les clients qui quittent ont tendance à avoir des soldes plus élevés, tandis que ceux qui restent ont une densité significative autour de zéro (peu ou pas de solde).
- **Raisons possibles :** un mécontentement ou une recherche de meilleures opportunités

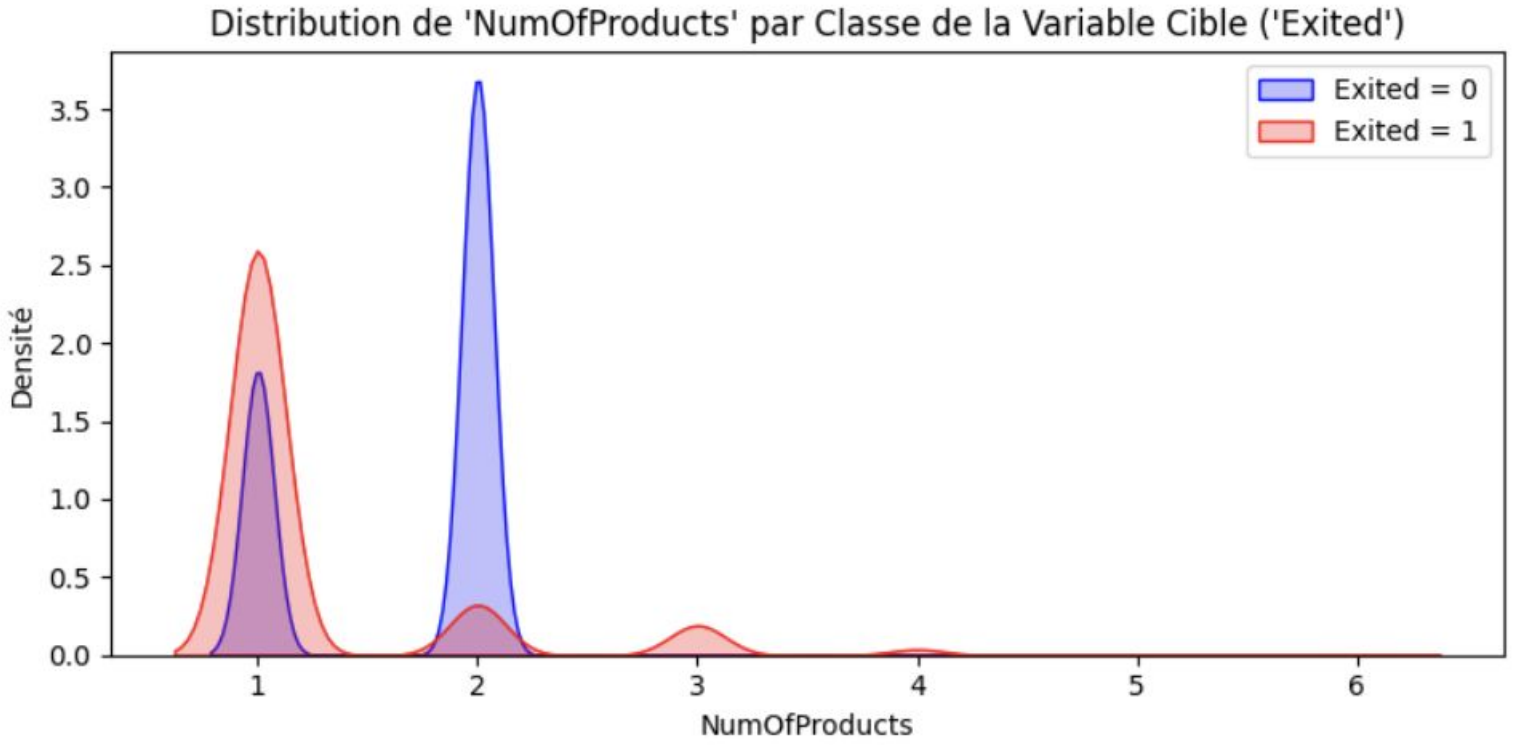
6. Variables Numériques : Salaire Estimé



- Les distributions sont presque identiques pour les deux groupes.

  → Le salaire estimé ne semble pas avoir d'impact significatif sur la probabilité de quitter ou rester

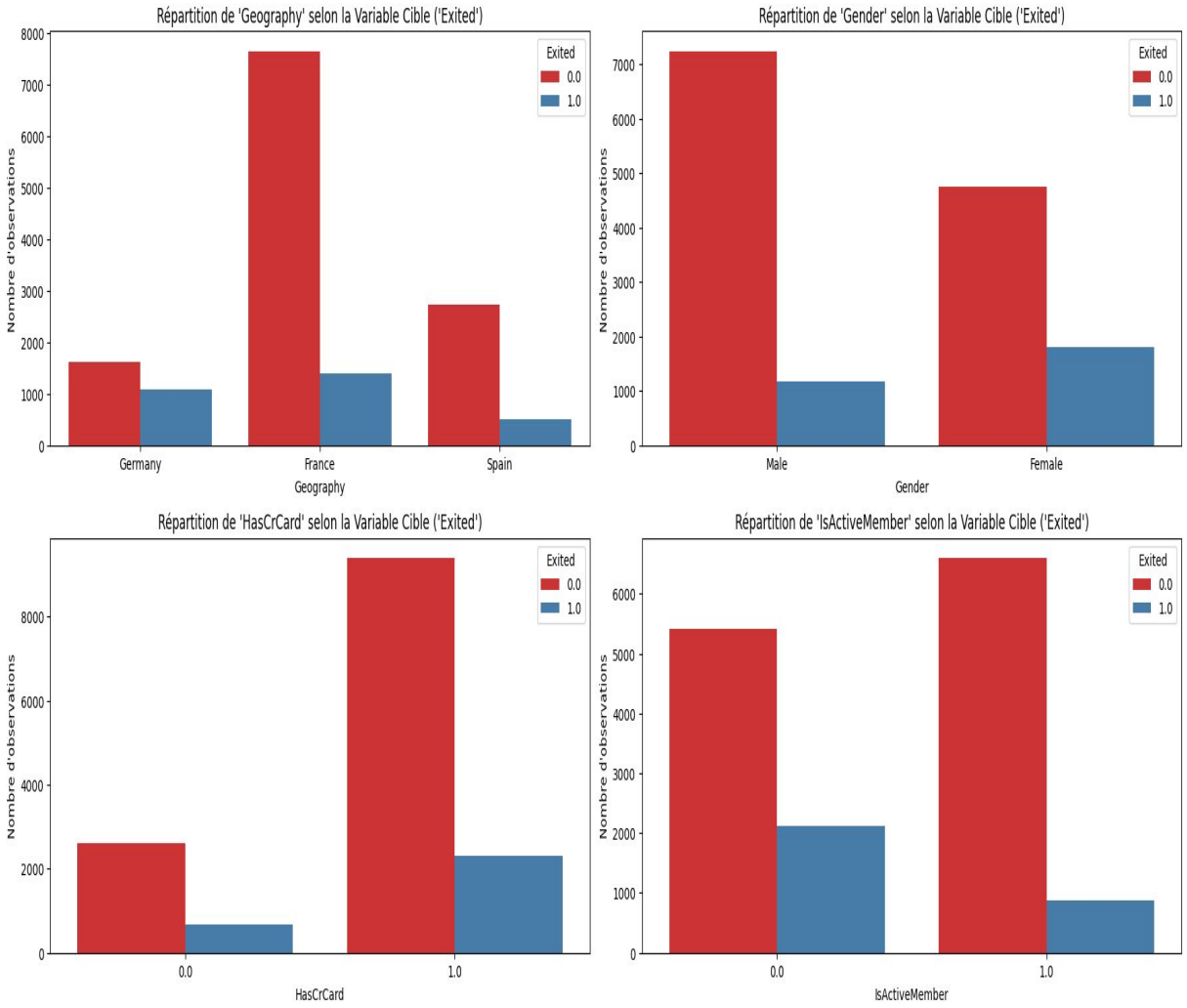
7. Variables Numériques : Nombre de produits bancaires



- Les clients avec **2** produits restent majoritairement.

→ Les clients ayant plus de produits sont plus fidèles, tandis que ceux avec un seul produit sont plus susceptibles de quitter.

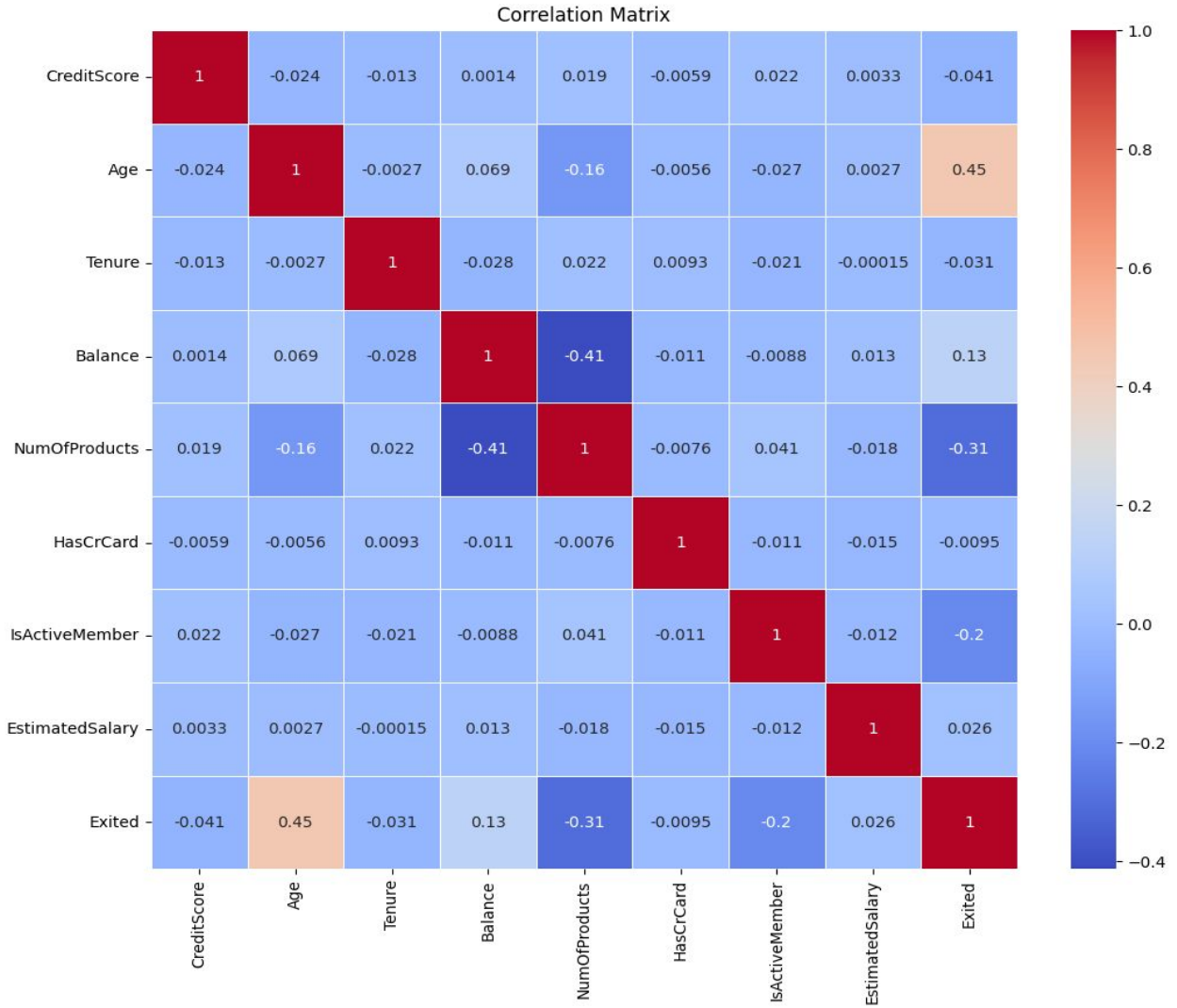
8. Variables Catégorielles



- Les femmes semblent avoir un taux de départ légèrement plus élevé proportionnellement.
- La proportion de clients quittant la banque semble relativement similaire entre les pays.
- Avoir une carte de crédit ne semble pas fortement influencer la décision de partir.
- Les membres inactifs ont tendance à quitter la banque plus fréquemment que les membres actifs.

9. Corrélation

- La plupart des variables ont des corrélations faibles entre elles.
- Profil type du client à risque serait : une personne âgée, inactive, avec peu de produits mais un solde élevé.



III. Préparation des données

1. Vérification des valeurs uniques et des valeurs manquantes

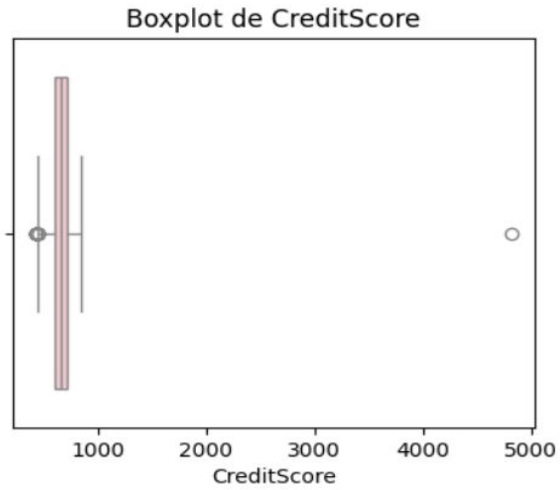
Valeurs uniques

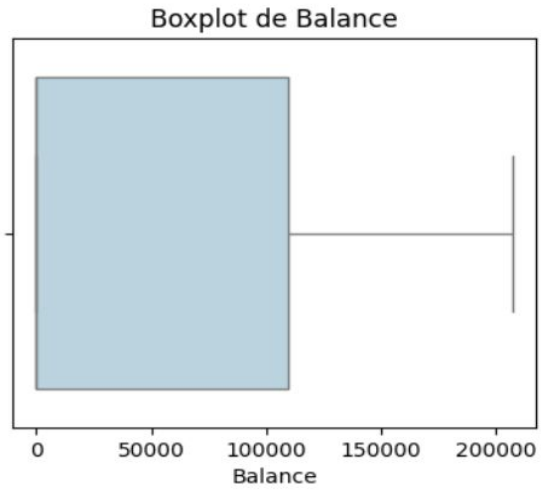
	Nombres de valeurs uniques
Balance	3307
Surname	755
CreditScore	376
Age	55
Tenure	11
NumOfProducts	5
Geography	3
Gender	2
HasCrCard	2
IsActiveMember	2

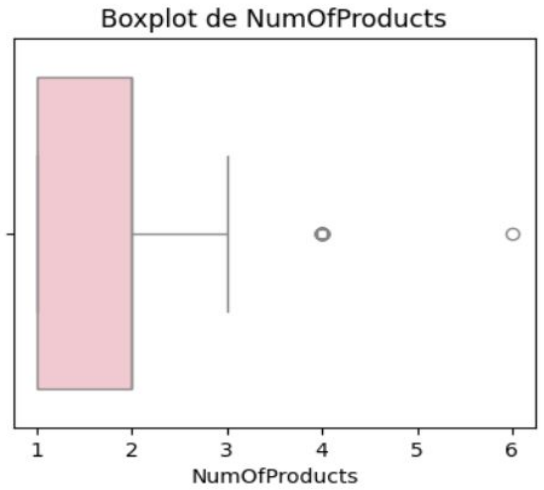
Valeurs manquantes

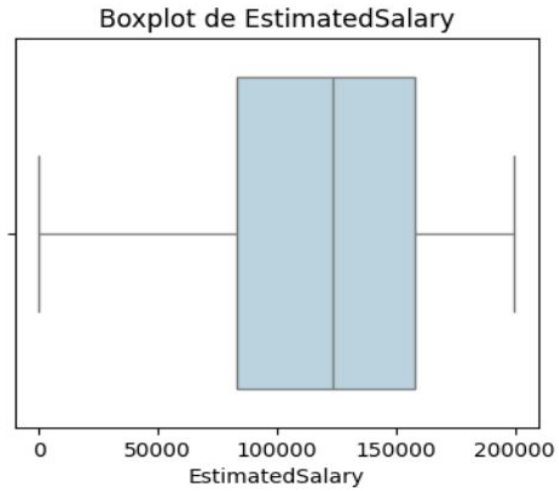
	Colonne	pourcentage manquant	nombre
0	id	0.0	0
1	CustomerId	0.0	0
2	Surname	0.0	0
3	CreditScore	0.0	0
4	Geography	0.0	0
5	Gender	0.0	0
6	Age	0.0	0
7	Tenure	0.0	0
8	Balance	0.0	0
9	NumOfProducts	0.0	0
10	HasCrCard	0.0	0
11	IsActiveMember	0.0	0
12	EstimatedSalary	0.0	0
13	Exited	0.0	0

2. Outliers









- Les valeurs semblent être logiques, à l'exception du CreditScore, qui nécessite un traitement spécifique.

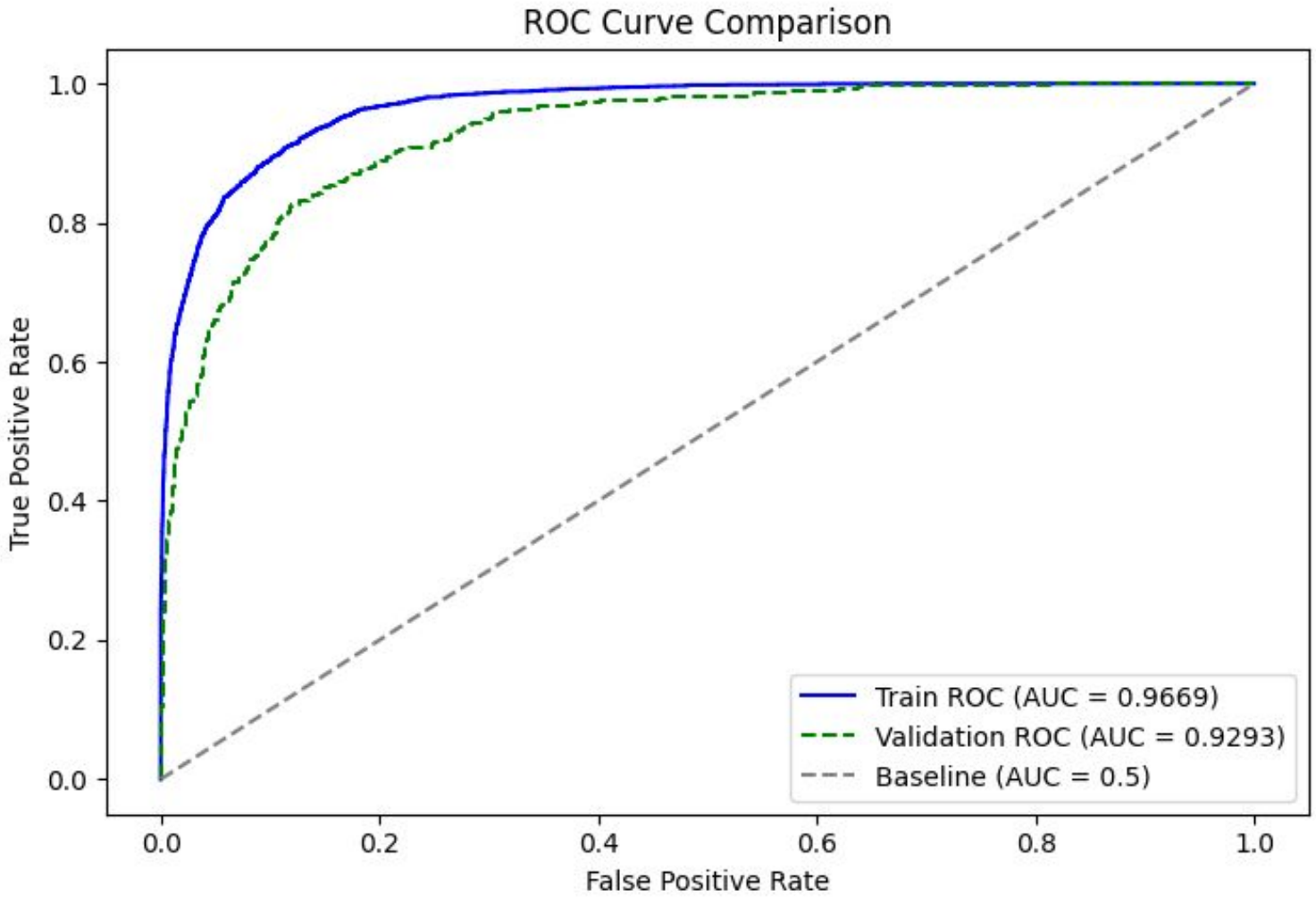
❗ Après avoir testé la suppression de cette valeur aberrante, nous avons constaté une baisse significative des performances des modèles → Nous avons décidé de la conserver.

3. Autres prétraitements

- **One hot Encoding:** Geography et Gender
- **Normalisation :** CreditScore, Age, Balance, EstimatedSalary, Tenure, NumOfProducts
- **Séparation du dataset en :** Train 80% et Validation 20%
- Les colonnes comme **Customer ID** et **Surname** ont été exclues de la modélisation car elles n'ont pas de pertinence prédictive.

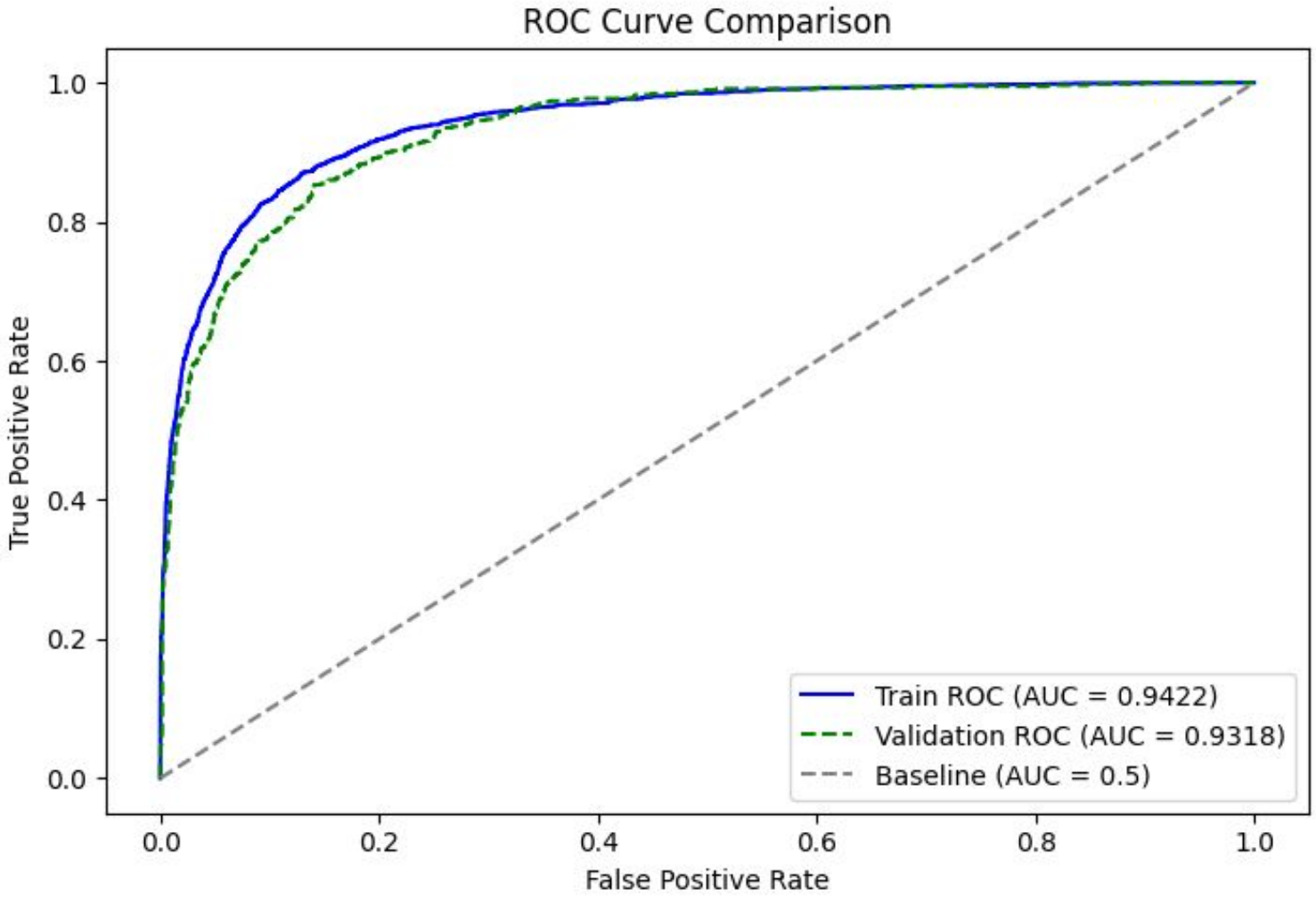
IV. Entraînement et évaluation des modèles

1. Random Forest



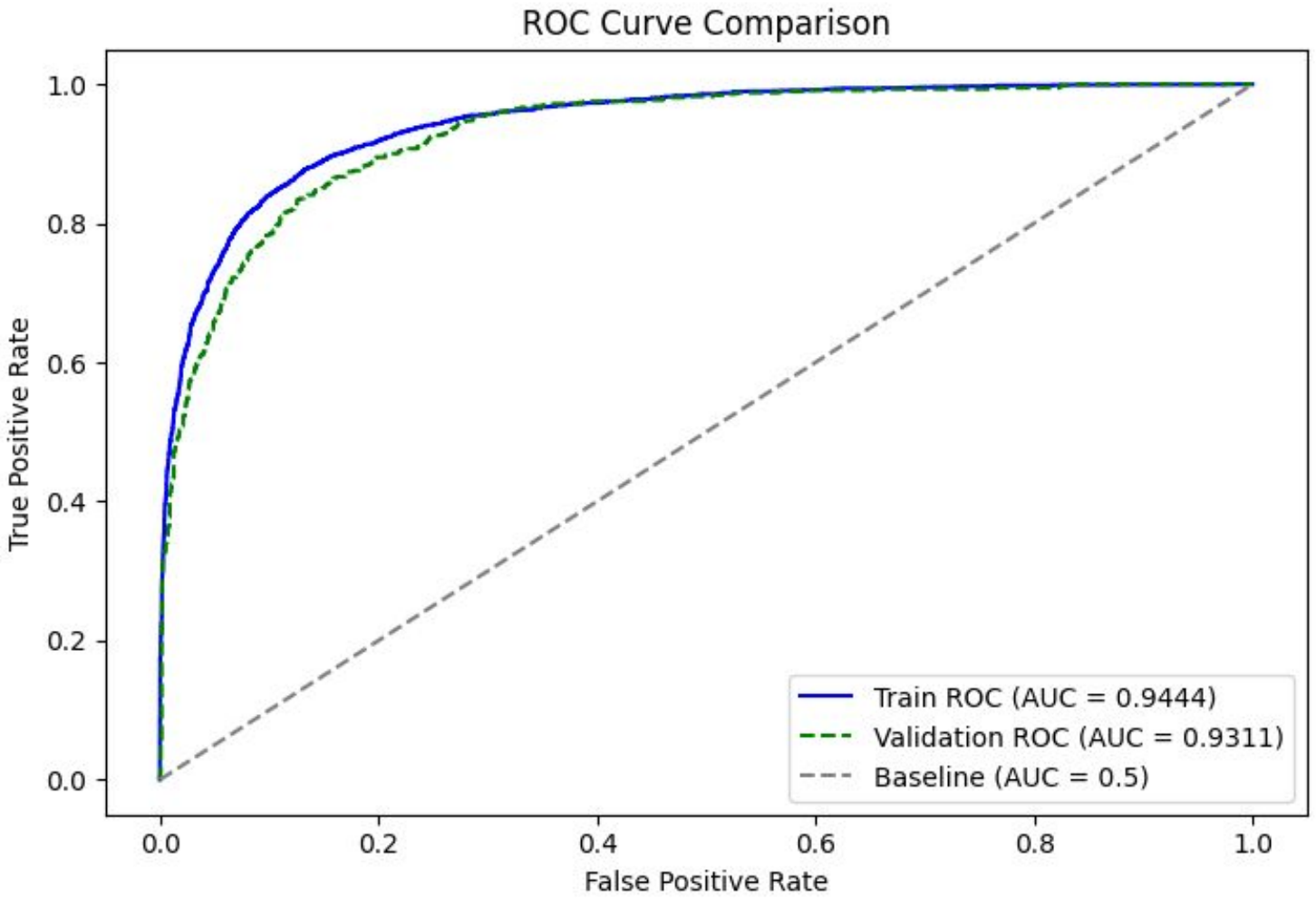
- **1000** essais avec Optuna
- AUC Score sur le dataset validation : **0.9293**
- Score sur Kaggle : **0.9313**

2. XGBoost



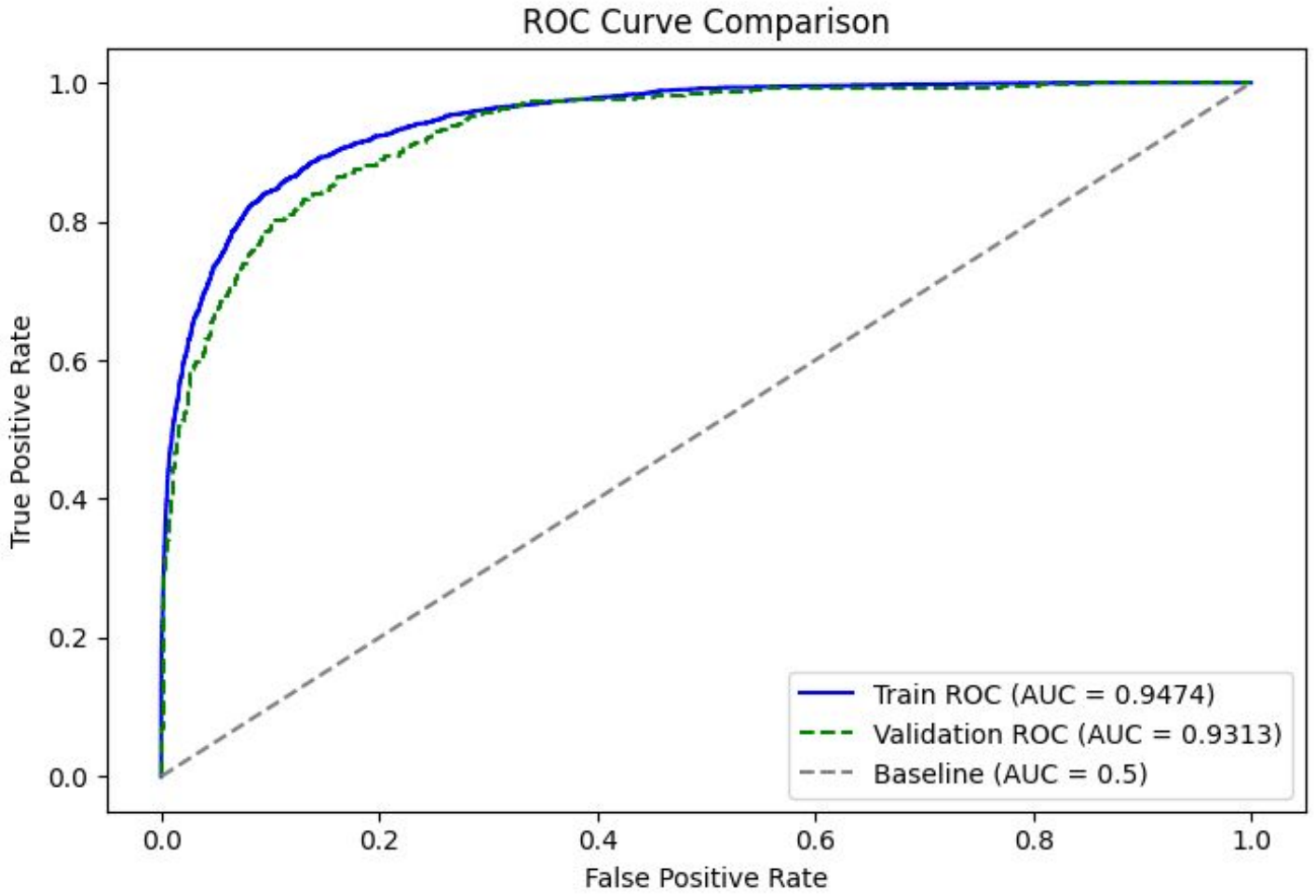
- **1000** essais avec Optuna
- AUC sur validation : **0.9318**
- Score sur Kaggle : **0.9309**

3. CatBoost



- AUC sur validation : **0.9311**
- Score sur Kaggle : **0.9324**

4. Blending



- Catboost, XGBoost, LightGBM, Random Forest
- AUC sur validation : **0.9313**
- Score sur Kaggle : **0.9323**

V. Conclusion

1. Conclusion

Les résultats obtenus pour prédire le churn bancaire ont mis en évidence la robustesse de différents modèles de machine learning: Le CatBoost obtient le meilleur score d'AUC, suivi de près par le Blending, mettant en évidence l'efficacité des approches d'ensemble.

	Random Forest	XGB	CATBoost	Blending
AUC SCORE	**0.9313**	**0.9309**	**0.9324**	**0.9323**

2. Axes d'amélioration

- Mieux hyperparamétriser nos modèles.

  → Limite: augmentation du temps d'exécution

- Trouver un Feature Engineering qui permet d'augmenter la performance de nos modèle.

  → Limite: nécessité de la connaissance métier.

- Tester des pipelines automatisés : utiliser des solutions AutoML (e.g., H2O, TPOT) pour générer rapidement des modèles optimisés.

MERCI POUR VOTRE ATTENTION