Projet de Machine learning2

ISE2 2023/2024

BANK CHURN SCORING

Réalisé par:

Marie Agathe SECK

Diakhou NDAO

Fallou BADJI

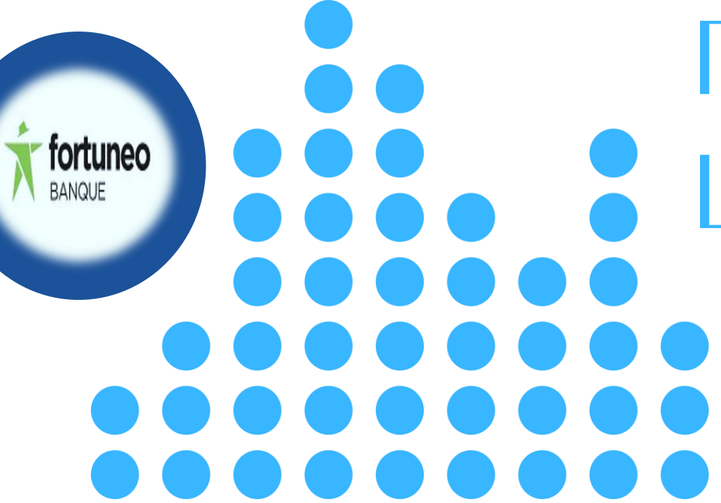
Elèves Ingénieurs Statisticiens Economistes

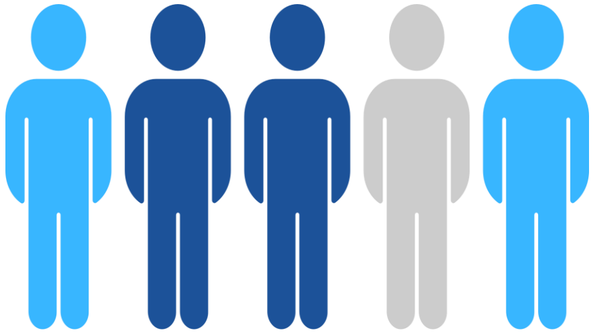
Formatrice :

Mme Mously DIAW

Machine Learning Engineer - Formatrice IA

Introduction





fortuneo est une banque en ligne qui propose une gamme variée de services financiers et bancaires pour des personnes ayant peu ou pas du tout d'historique de prêt.

Ses Services : Prêt personnel avec Younited Crédit, Compte bancaire ,Assurance vie ,Livret d'épargne, Crédit immobilier

Quelques Statistiques actuelles sur ses services bancaires

Compte bancaire

- **0€** de frais de tenue de compte
- **0€** cartes bancaires Mastercard gratuites, sous conditions

Livret d'épargne

- **4 livrets sans frais** pour une épargne 100% sécurisée et 100% disponible : Livret A, LDDS, Livret +, Livret Enfant.
- **0€** de frais de gestion, d'ouverture ou de clôture.

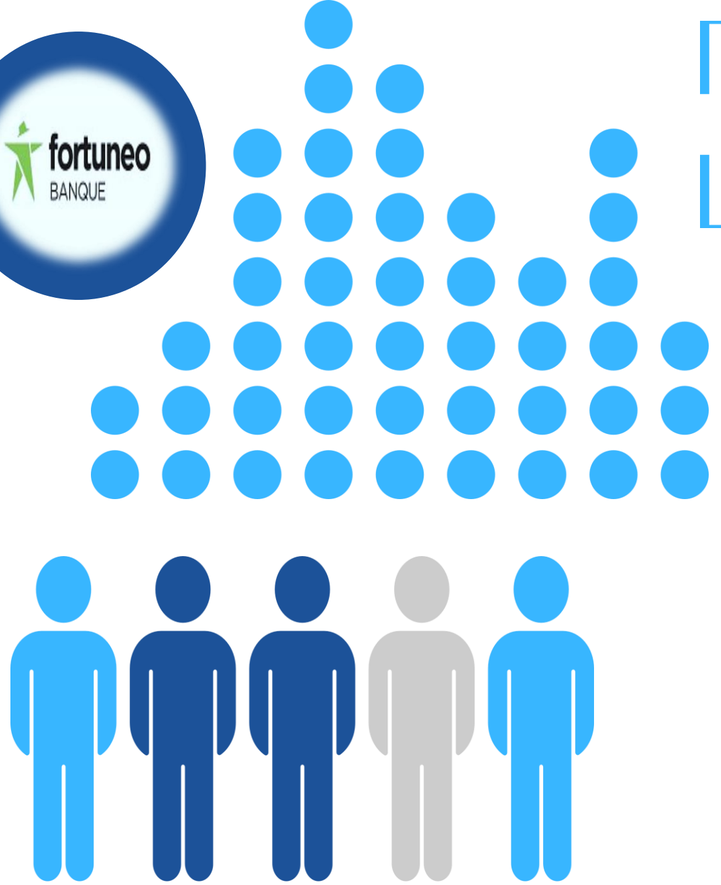
Assurance vie

- **0 €** de frais d'adhésion, de versements, d'arbitrages, de sortie.

Crédit immobilier

- **%** Taux compétitifs sans négociation
- **0€** de frais de dossier et de frais de remboursement anticipé

Introduction



Problématique

L’entreprise souhaite mettre en œuvre un outil de “churn scoring” pour calculer la probabilité qu’un client quitte la banque, puis classifie le client en ‘churn’ ou pas. Elle souhaite donc développer un algorithme de classification en s’appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

Objectif de la banque

Identifier les clients susceptibles de quitter la banque.

Objectif de l’étude

Concevoir un bon modèle prédictif à mettre à la disposition de fortuneo Banque afin qu’elle puisse réduire le **churn** en prenant des mesures préventives ciblées pour retenir les clients à haut risque.

Ses Services : Prêt personnel avec Younited Crédit, Compte bancaire ,Assurance vie ,Livret d’épargne, Crédit immobilier

PLAN

Introduction

01 Analyse exploratoire des données

02 Modélisation

03 Choix du modèle définitif

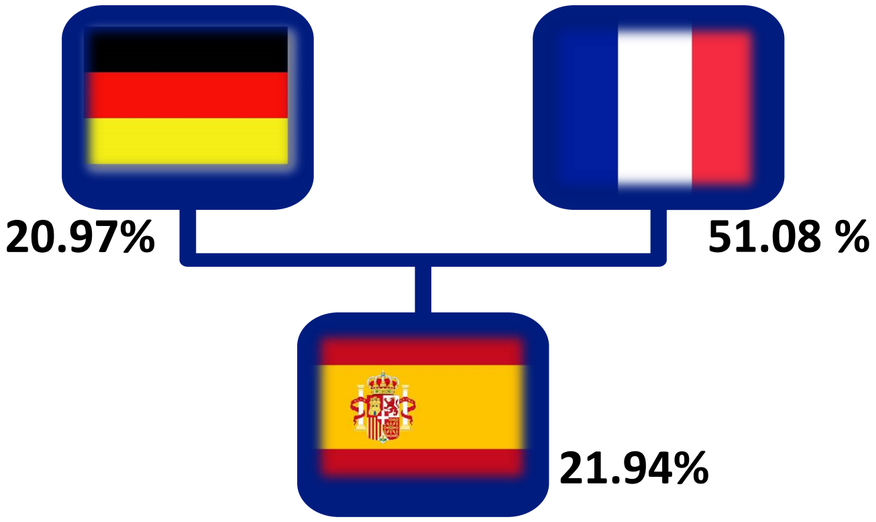
Déploiement et Mise en Production du Modèle

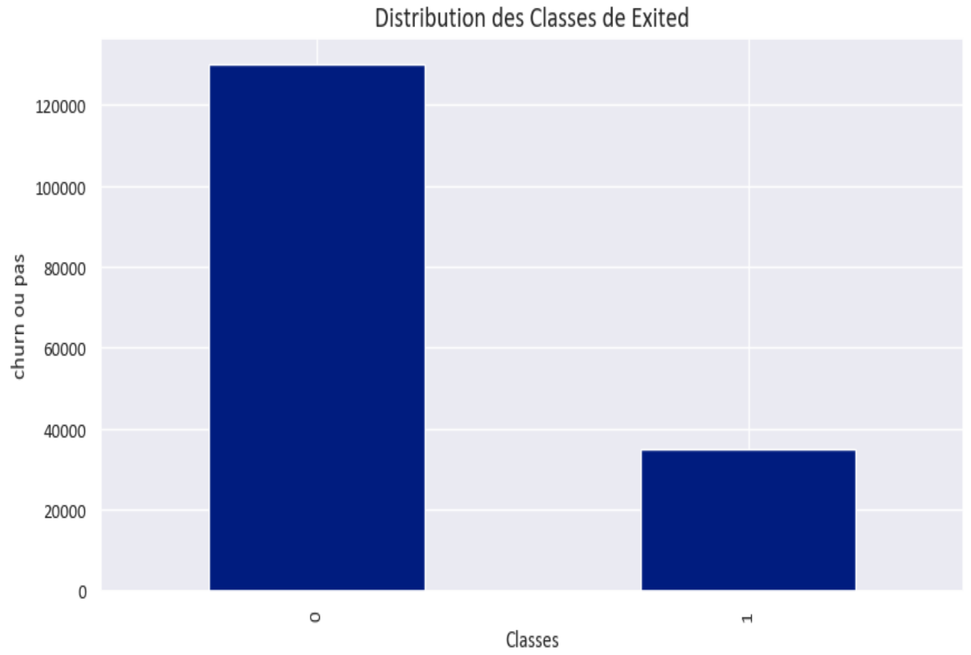
01. Analyse Exploratoire

Statistiques Globales

Le jeu de données dont on dispose comporte des informations sur les clients de la banque qui ont quitté ou qui continuent d'être clients. Le jeu de données nous renseigne sur leurs caractéristiques telles que L'âge, le sexe, le pays, ainsi que d'autres variables comme le nombre de produits utilisés par les clients: s'ils sont des membres actifs ou non, leurs salaires estimés, s'ils détiennent des cartes de crédit…

Répartition des Clients



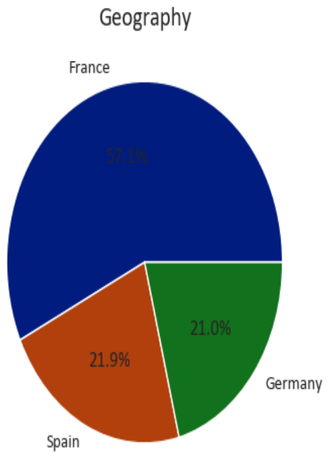


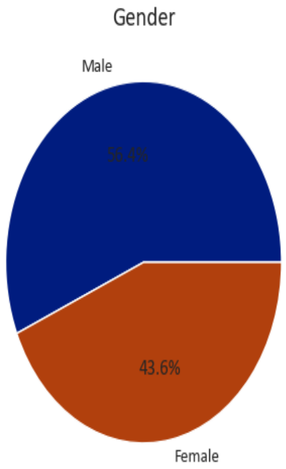
Autres Tendances

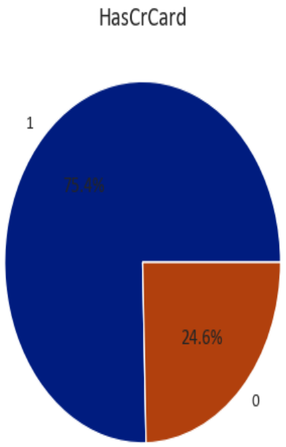
- Clients détenteurs de carte bancaire : **75.39 %**
- Clients utilisant au moins deux types de produits : **50%**
- Age moyen des clients : **38 ans** dont **75%** ont moins de **42 ans**
- Salaire moyen: **112574.82 €**

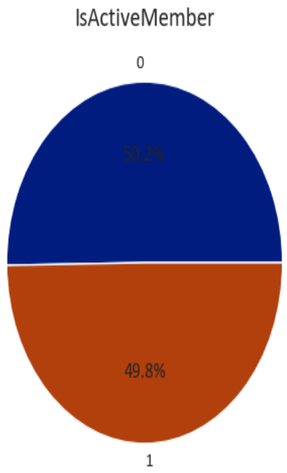
Statistiques Exploratoires

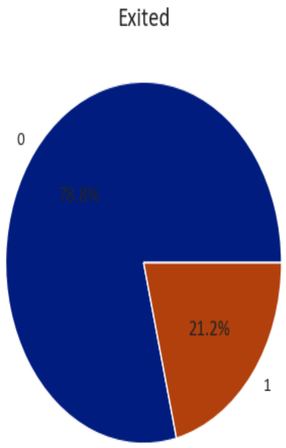
Variables Catégorielles



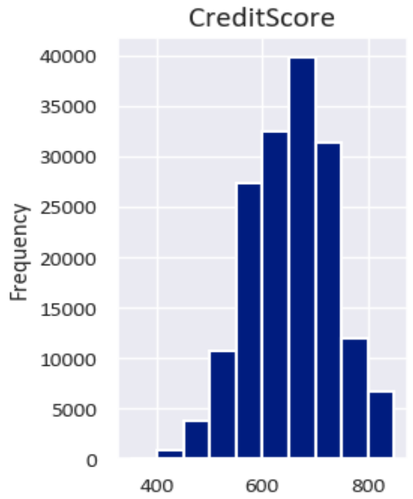


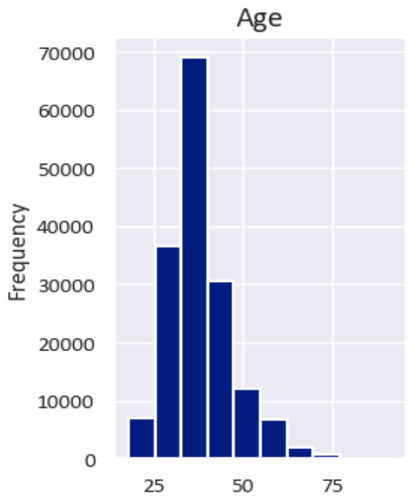


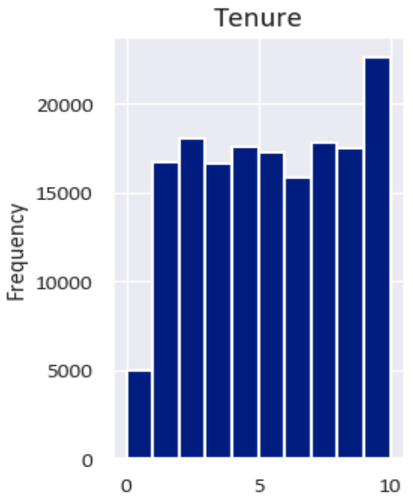


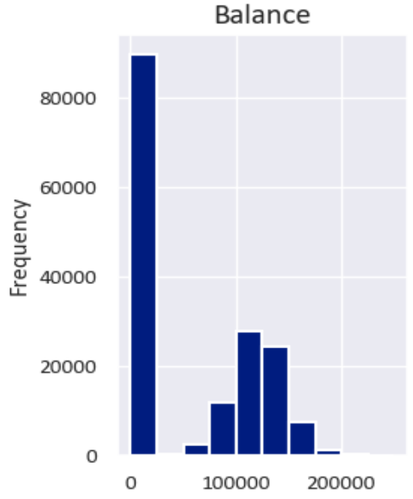


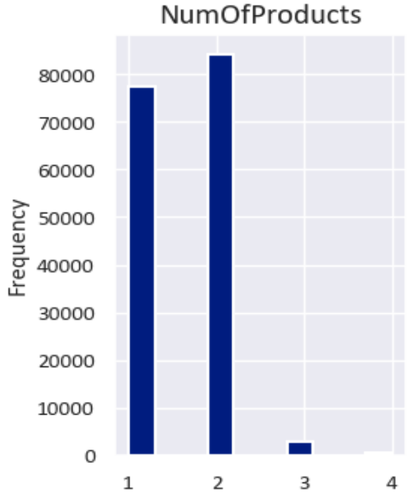
Variables Numériques

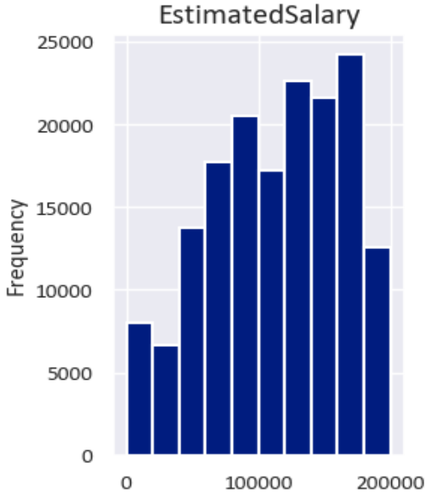






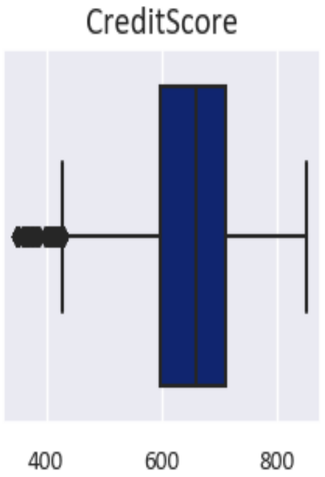


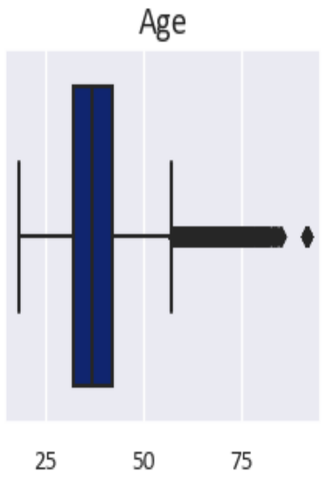


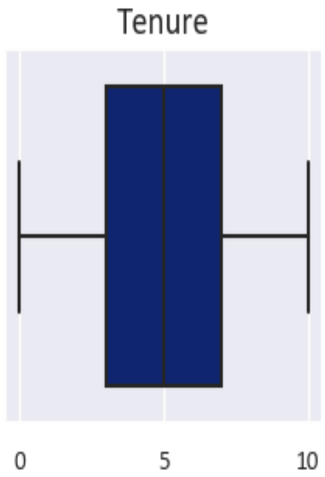


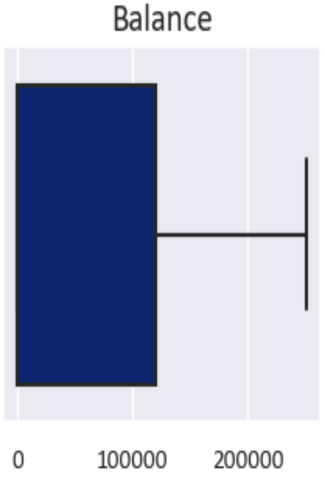
Statistiques Exploratoires

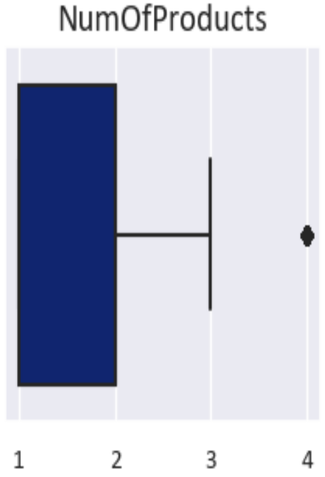
Boxplot des variables numériques

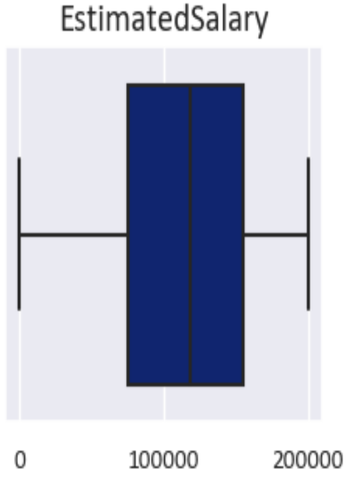




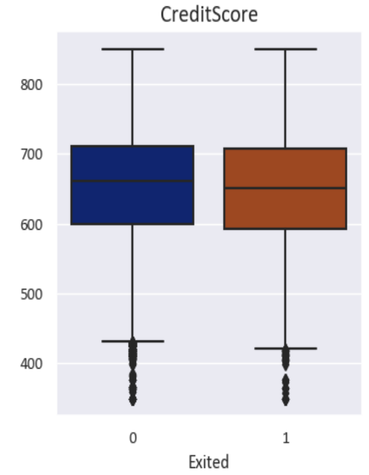


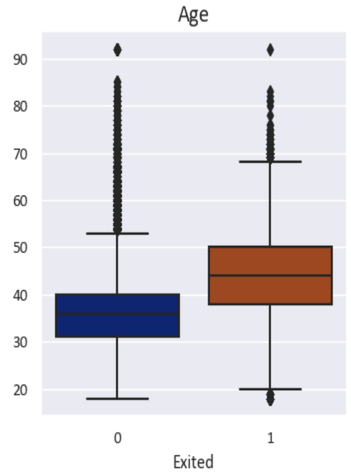


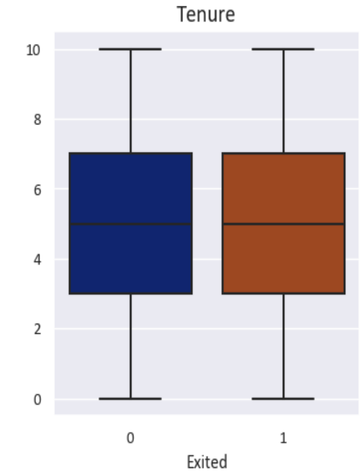


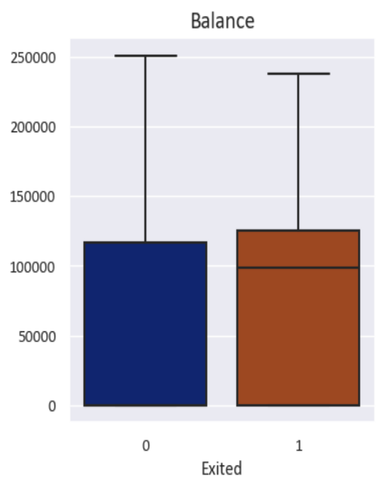


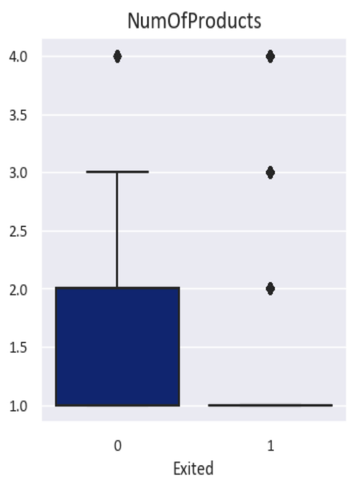
Boxplot de ces variables en fonction des classes

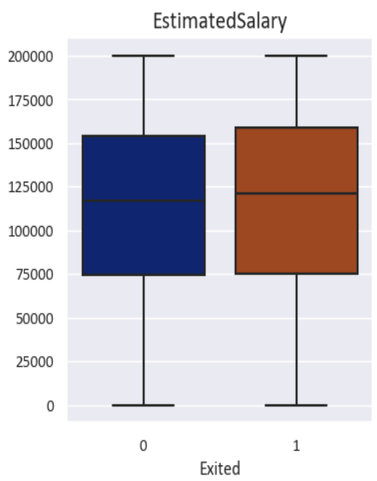






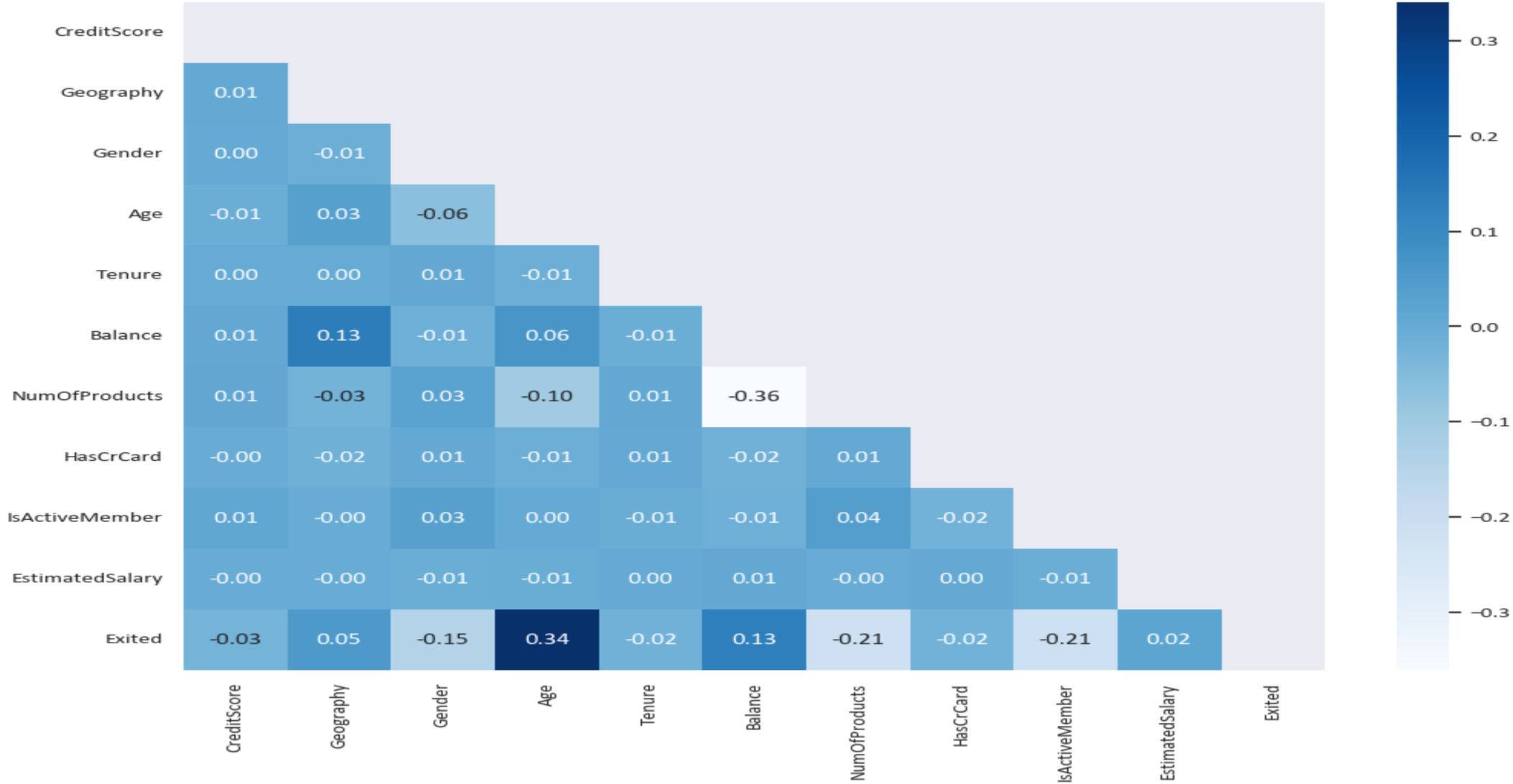






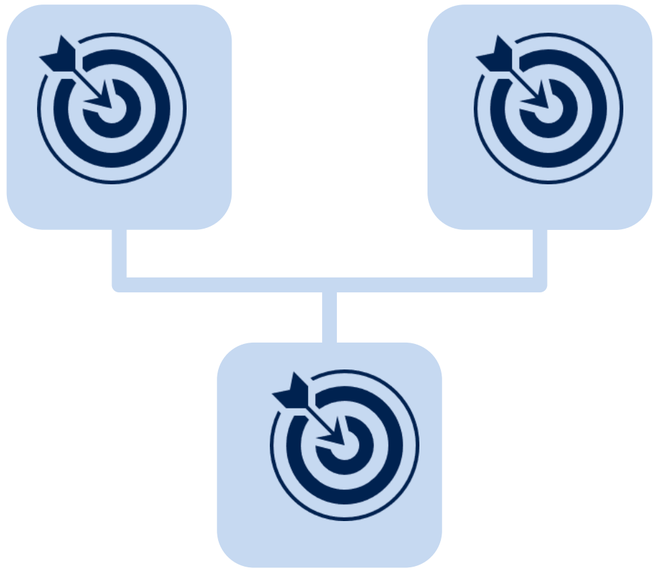
Statistiques Exploratoires

Corrélations



02.Modélisation

Démarche



Le modèle de prédiction que nous voulons mettre en place sera un outil d'aide à la prise de décision pour l 'entreprise fortuneo banque. Etant conscient de cet enjeu, il est important de prendre en compte le contexte métier afin d'élaborer un contexte de travail en conformité avec l'objectif attendu de cette étude.

L 'idée est d'explorer plusieurs modèles , paramétriques comme ensemblistes, afin d 'en choisir le meilleur suivant un ensemble de métriques d'évaluation .

- D'une part, nous allons optimiser selon l'AUC pour tenir compte du déséquilibre entre le classes (métrique très utilisé dans le cas des classifications binaires avec déséquilibre).

- D'autre part, en raison du domaine (bancaire), nous testerons l'optimisation des modèles de sorte à minimiser l'erreur que l'on commet en prédisant qu'un client va rester alors qu'il est sur le point de se désabonner (minimiser les faux négatif). Et nous utilisons à cet effet un f_β_ score avec un $\beta > 1$ ($\beta = 2$) basé sur la littérature dans le domaine de la banque et de la finance.

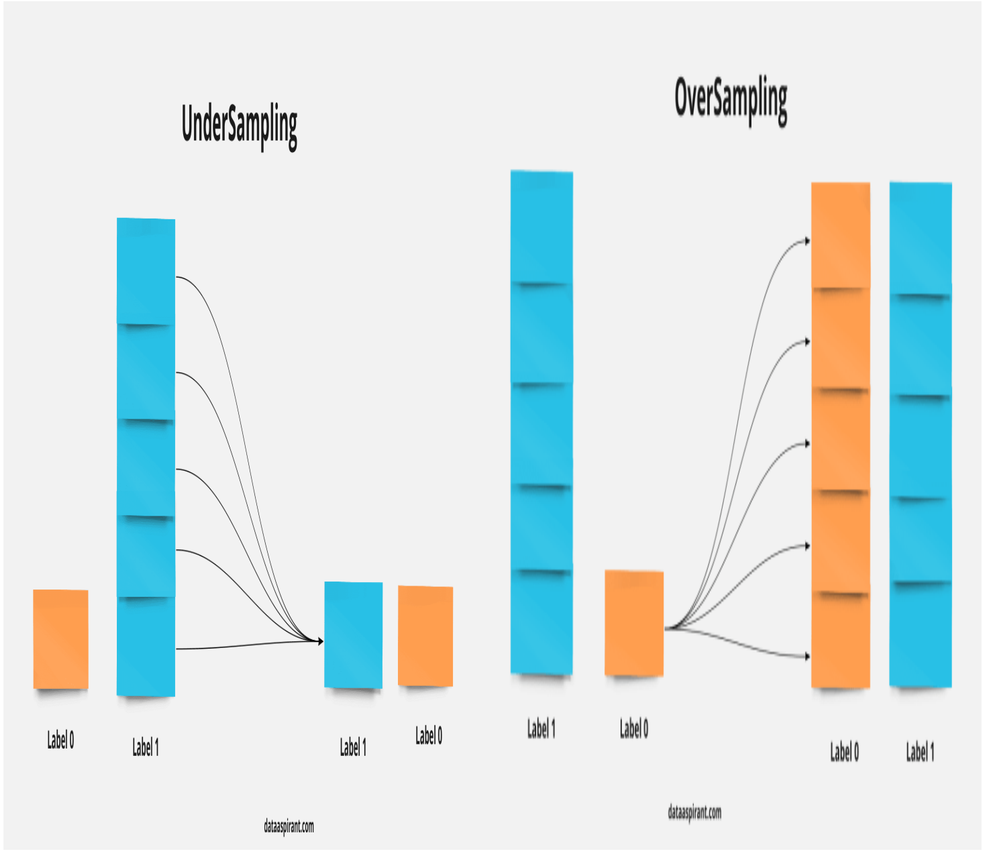
Le déséquilibre des classes sera prise en compte dans le pipeline de la modélisation à l'aide de certaines méthodes de réechantillonnage que nous présenterons par la suite.

Gestion du problème de déséquilibre

La classification sur données déséquilibrées est un problème où l'échantillon d'apprentissage contient une forte disparité entre les classes à prédire ,c'est le cas de nos données avec 21.1% pour la classe positive. Cette situation peut conduire à des modèles biaisés. Si nous entraînons un classificateur sur des données déséquilibrées, il pourrait apprendre à favoriser la classe majoritaire et à ignorer la classe minoritaire.
Pour apporter une solution à ce problème, plusieurs méthodes sont envisageables :

- ✦ **Collecter davantage de données;**
- ✦ **Utiliser un modèle pénalisé;**
- ✦ **Utiliser des méthodes de rééchantillonnage;**
- ✦ **Utiliser sa créativité**

Dans le cadre de ses travaux ,nous avons opté pour la démarche suivant:
Pour chaque modèle à entrainer, nous faisons un premier test en utilisant tous les types de méthodes de rééchantillonnage notamment le **oversampling** ,le **undersampling** et la méthode **mixte** .Et à l'issu de cette première étape, on récupère les trois (3) meilleurs samplers pour ensuite optimiser les hyperparamètres du modèle avec ces samplers afin d'en sortir le meilleur modèle suivant l'AUC ou le f_2_ score .



Spécification

La matrice X des variables explicatives

- CreditScore
- Geography
- Gender
- Age
- Tenure
- Balance
- NumOfProducts
- HasCrCard
- IsActiveMember
- EstimatedSalary

La variable Cible

- Exited

Métriques d'évaluation

- Recall
- Precision
- f1_score
- f2_score
- Accuracy
- AUC
- macro avg (du f1, du recall, etc.)
- Weighted avg (du f1, du recall, etc.)

Preprocessor

- StandardScaler: **numeric_transformer**
- OneHotEncoder: **categorical_transformer**
- Sampling **: Over, under ou combine (à optimiser)**

Spécification

Modèle Paramétrique

Régression Logistique

La régression logistique modélise la probabilité qu'une observation appartienne à la classe positive comme une transformation logistique d'une combinaison linéaire des variables. Elle permet de modéliser un problème de classification binaire . Les données sont considérées comme n points (observations) en p dimensions, représentés par la matrice X ∈ ℝ_n×p. Leurs étiquettes, représentées par un vecteur y ∈ {0,1}, représentent l'appartenance (1) ou non (0) à une classe.

$$p(Y = 1|x) = \frac{1}{1 + \exp - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}$$

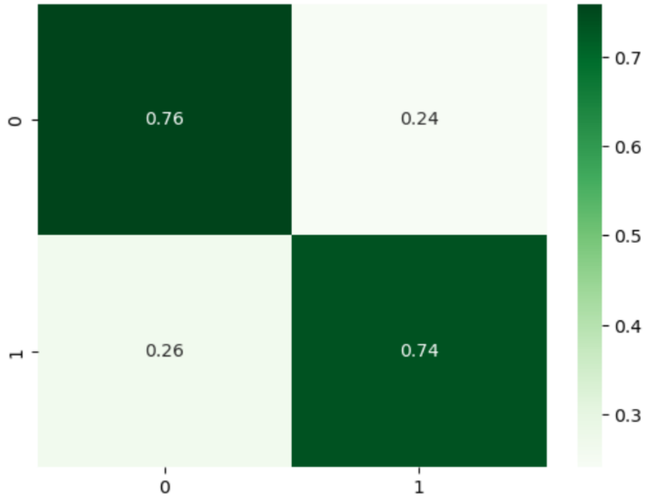
Les coefficients s'apprennent par maximisation de vraisemblance. Les coefficients peuvent être contrôlés avec régularisation ℓ2 pour éviter le sur-apprentissage et avec régularisation ℓ1 pour obtenir un modèle parcimonieux.

Présentation des résultats

R. Logistique Optimisé selon AUC

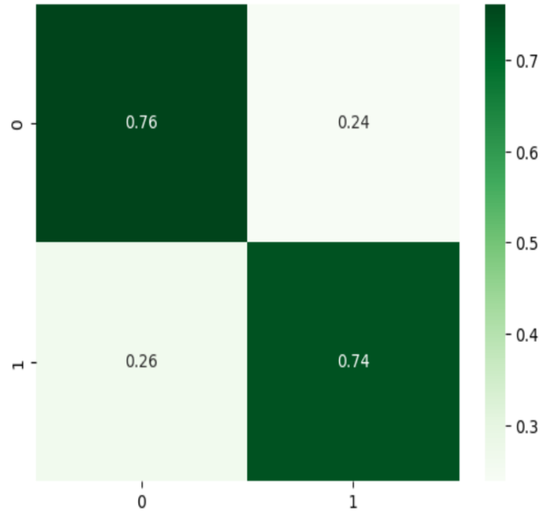
AUC: 0.8184

TRAIN



	precision	recall	f1-score	support
0	0.91	0.76	0.83	104023
1	0.45	0.74	0.56	27905
accuracy			0.75	131928
macro avg	0.68	0.75	0.69	131928
weighted avg	0.82	0.75	0.77	131928

TEST

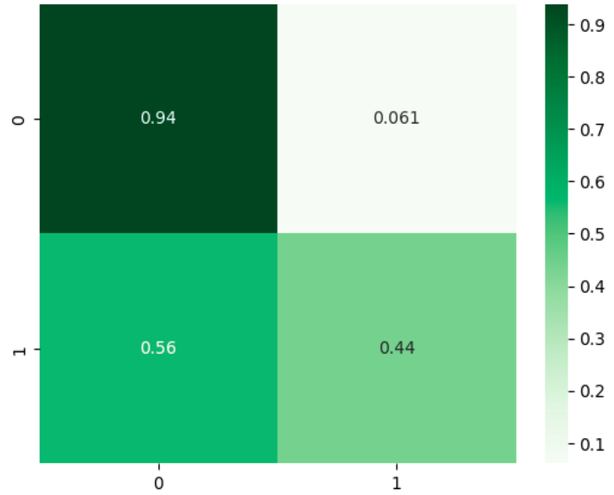


	precision	recall	f1-score	support
0	0.91	0.76	0.83	25979
1	0.45	0.74	0.56	7004
accuracy			0.76	32983
macro avg	0.68	0.75	0.70	32983
weighted avg	0.82	0.76	0.77	32983

R. Logistique optimisé selon f2_weighted

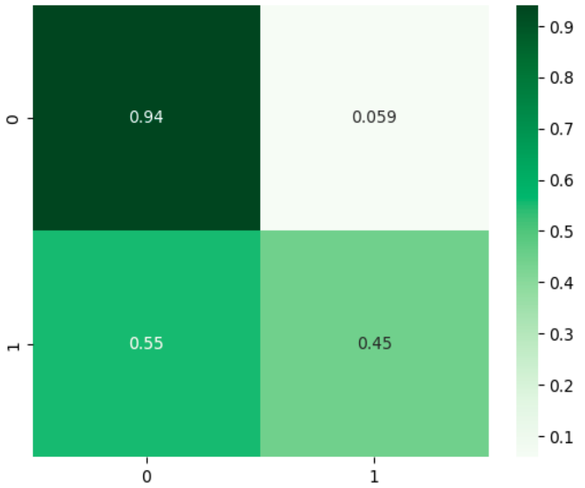
F2_weighted : 0.83

TRAIN



	precision	recall	f1-score	support
0	0.86	0.94	0.90	104023
1	0.66	0.44	0.53	27905
accuracy			0.83	131928
macro avg	0.76	0.69	0.71	131928
weighted avg	0.82	0.83	0.82	131928

TEST

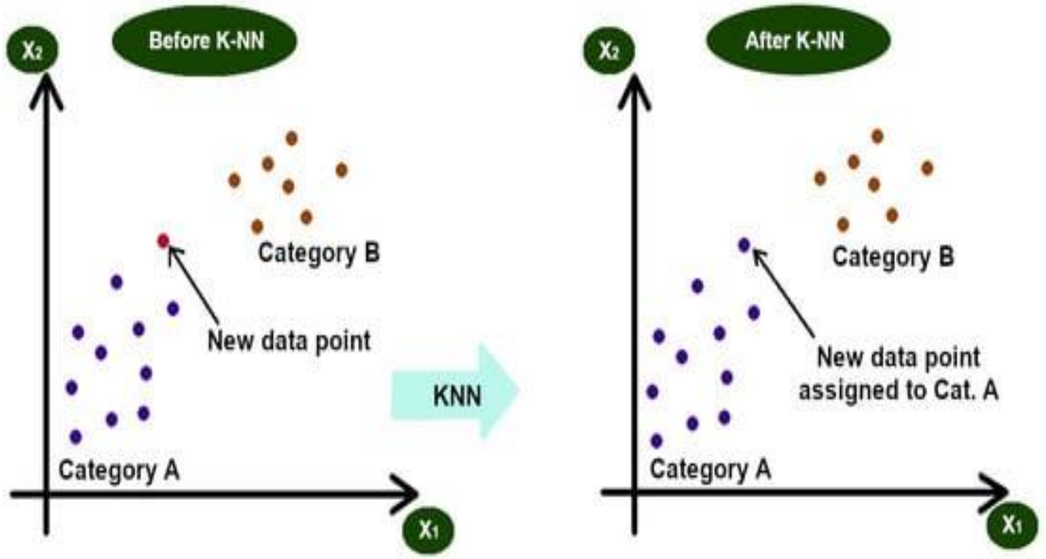


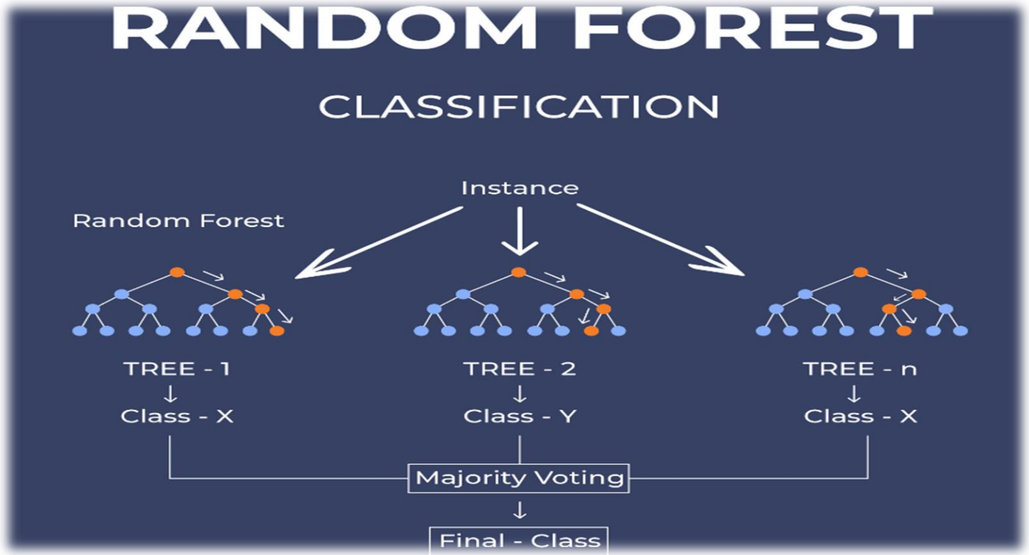
	precision	recall	f1-score	support
0	0.86	0.94	0.90	25979
1	0.67	0.45	0.53	7004
accuracy			0.84	32983
macro avg	0.77	0.69	0.72	32983
weighted avg	0.82	0.84	0.82	32983

Méthodes Ensembliste

K plus proches voisins (KNN)

C'est une méthode non paramétrique dans laquelle le modèle mémorise les observations de l'ensemble d'apprentissage pour la classification(ou même régression) des données de l'ensemble de test. Toutefois ,avant qu'une classification puisse être effectuée, la distance doit être définie. Il existe plusieurs distances: la distance euclidienne, la distance de Minkowski , la distance de Manhattan, la distance de Hamming … Et la distance euclidienne est la plus couramment utilisée.





Random forest

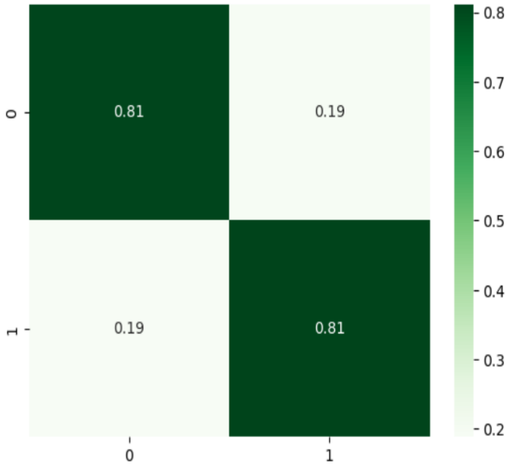
Le Random Forest combine plusieurs arbres de décision, formés aléatoirement sur des sous-ensembles de données et de caractéristiques, pour améliorer la robustesse du modèle par le vote majoritaire ou la moyenne.

Présentation des résultats

KNN optimisé selon AUC

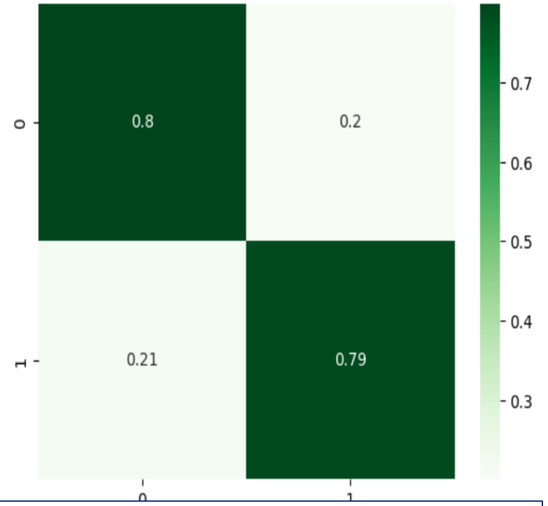
AUC: 0.8668

TRAIN



	precision	recall	f1-score	support
0	0.94	0.81	0.87	104023
1	0.53	0.81	0.64	27905
accuracy			0.81	131928
macro avg	0.74	0.81	0.76	131928
weighted avg	0.85	0.81	0.82	131928

TEST

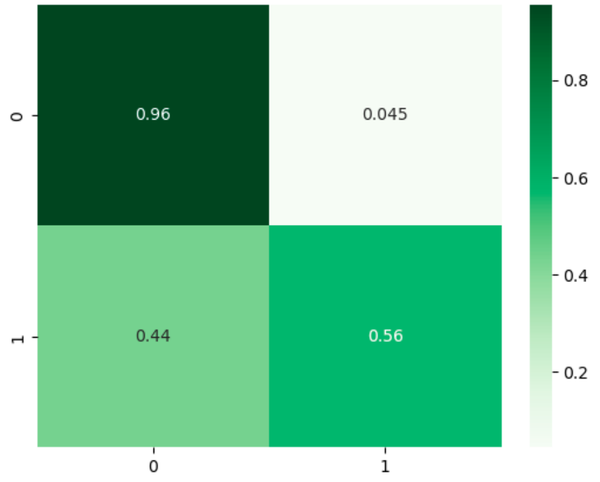


	precision	recall	f1-score	support
0	0.93	0.80	0.86	25979
1	0.51	0.79	0.62	7004
accuracy			0.80	32983
macro avg	0.72	0.79	0.74	32983
weighted avg	0.84	0.80	0.81	32983

KNN optimisé selon f2_weighted

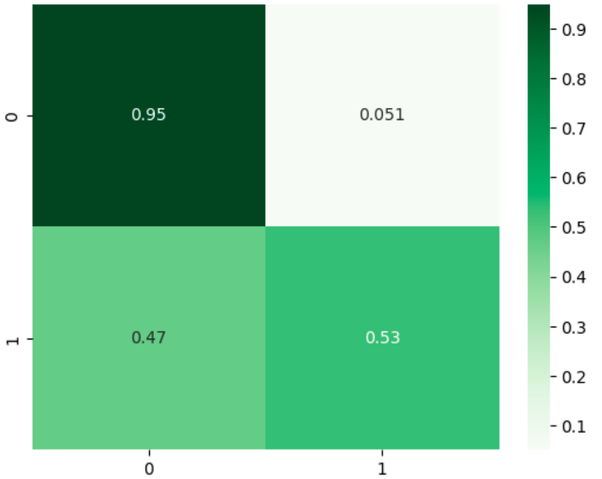
f2_weighted : 0.8545

TRAIN



	precision	recall	f1-score	support
0	0.89	0.96	0.92	104023
1	0.77	0.56	0.65	27905
accuracy			0.87	131928
macro avg	0.83	0.76	0.79	131928
weighted avg	0.87	0.87	0.86	131928

TEST



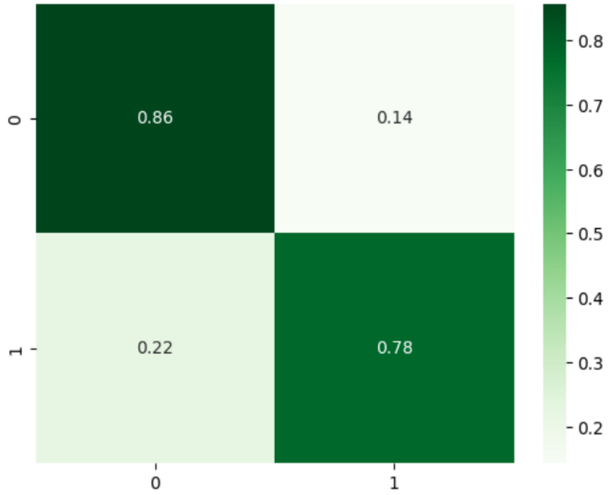
	precision	recall	f1-score	support
0	0.88	0.95	0.91	25979
1	0.74	0.53	0.62	7004
accuracy			0.86	32983
macro avg	0.81	0.74	0.77	32983
weighted avg	0.85	0.86	0.85	32983

Présentation des résultats

RF optimisé selon AUC

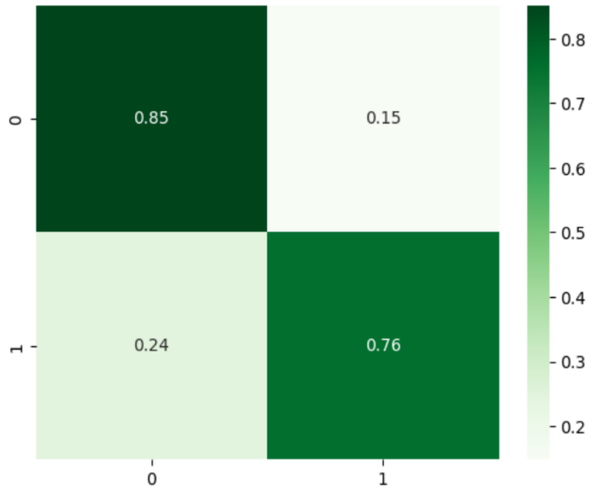
AUC: 0.8865

TRAIN



	precision	recall	f1-score	support
0	0.93	0.86	0.89	104023
1	0.59	0.78	0.67	27905
accuracy			0.84	131928
macro avg	0.76	0.82	0.78	131928
weighted avg	0.86	0.84	0.85	131928

TEST

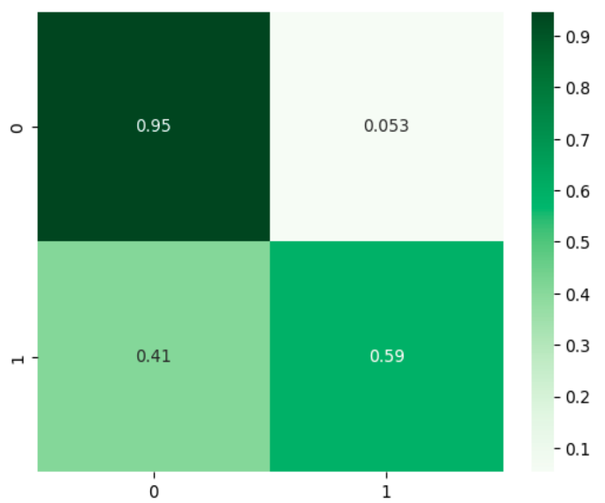


	precision	recall	f1-score	support
0	0.93	0.85	0.89	25979
1	0.58	0.76	0.66	7004
accuracy			0.83	32983
macro avg	0.75	0.80	0.77	32983
weighted avg	0.85	0.83	0.84	32983

RF optimisé selon f2_weighted

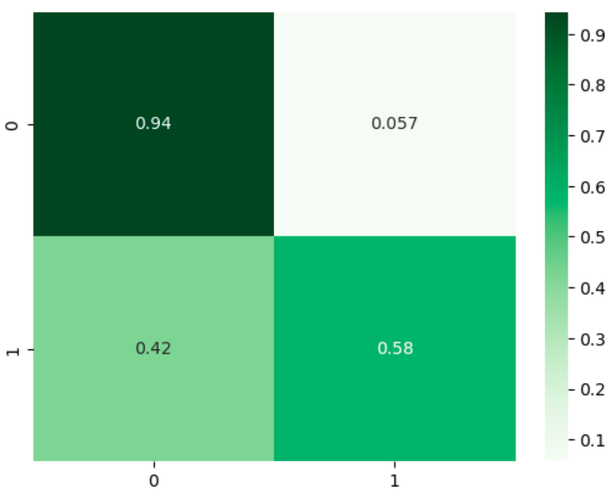
f2_weighted : 0.8612

TRAIN



	precision	recall	f1-score	support
0	0.90	0.95	0.92	104023
1	0.75	0.59	0.66	27905
accuracy			0.87	131928
macro avg	0.82	0.77	0.79	131928
weighted avg	0.87	0.87	0.87	131928

TEST



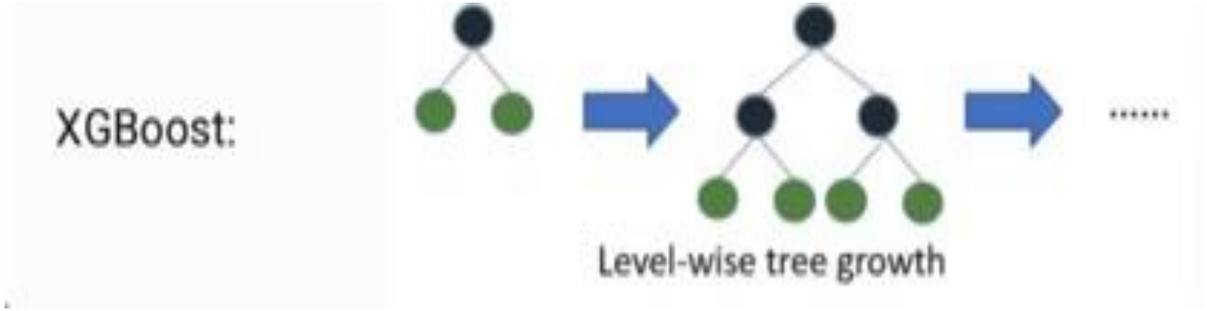
	precision	recall	f1-score	support
0	0.89	0.94	0.92	25979
1	0.73	0.58	0.65	7004
accuracy			0.87	32983
macro avg	0.81	0.76	0.78	32983
weighted avg	0.86	0.87	0.86	32983

Méthodes Ensembliste

XGBOOST Classifier

C’est une méthode de **Boosting** .Elle fait un assemblage d'arbres décisionnels (weak learners) qui prédisent les résidus et corrige les erreurs des arbres décisionnels précédents.

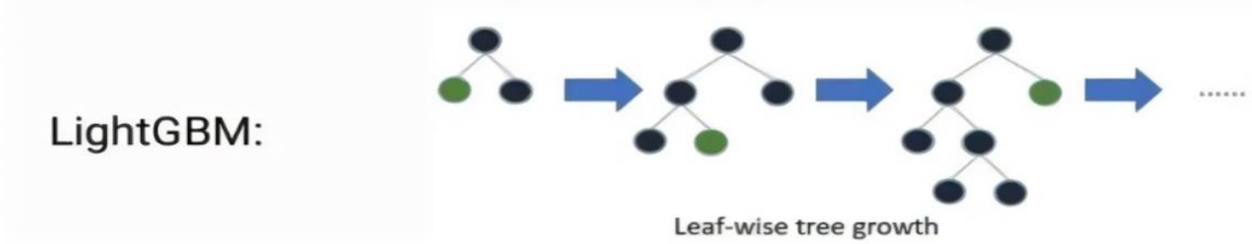
L’avantage est qu’elle permet de réduire les biais, et rend plus performent les prédictions.



LightGBM Classifier

Light GBM est un cadre de renforcement de gradient qui utilise un algorithme d'apprentissage basé sur des arbres.

Light GBM fait croître l'arbre verticalement tandis que d'autres algorithmes font pousser des arbres horizontalement, ce qui signifie que Light GBM fait pousser l'arbre par **feuille** tandis que l'autre algorithme se développe par niveau. Il choisira la feuille avec une perte de delta maximale pour se développer. Lors de la croissance de la même feuille, l'algorithme par feuille peut réduire plus de perte qu'un algorithme par niveau.

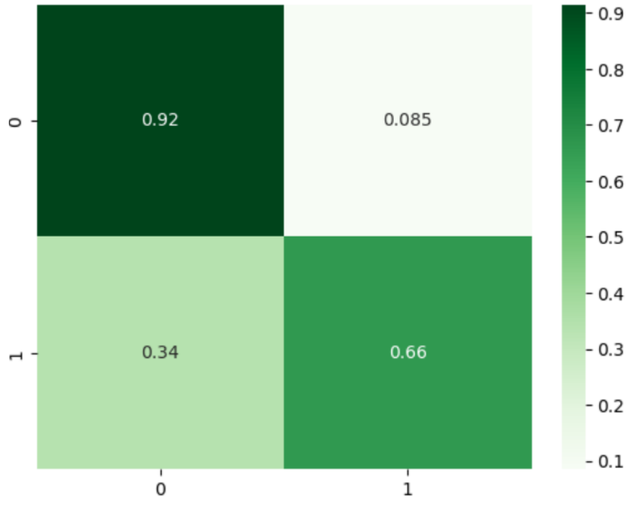


Présentation des résultats

XGB optimisé selon AUC

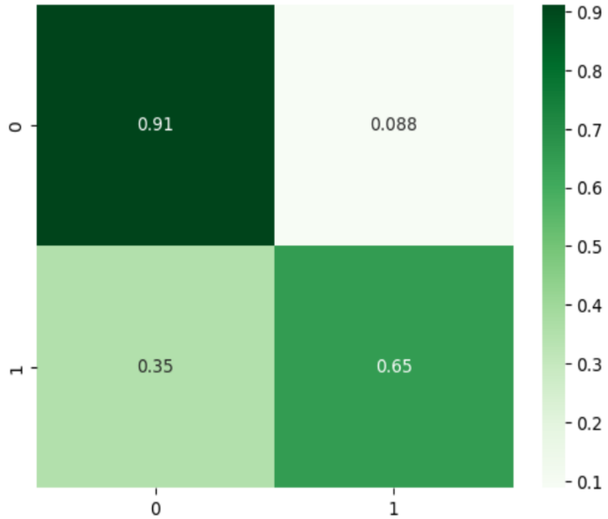
AUC: 0.8872

TRAIN



	precision	recall	f1-score	support
0	0.91	0.92	0.91	104023
1	0.68	0.66	0.67	27905
accuracy			0.86	131928
macro avg	0.79	0.79	0.79	131928
weighted avg	0.86	0.86	0.86	131928

TEST

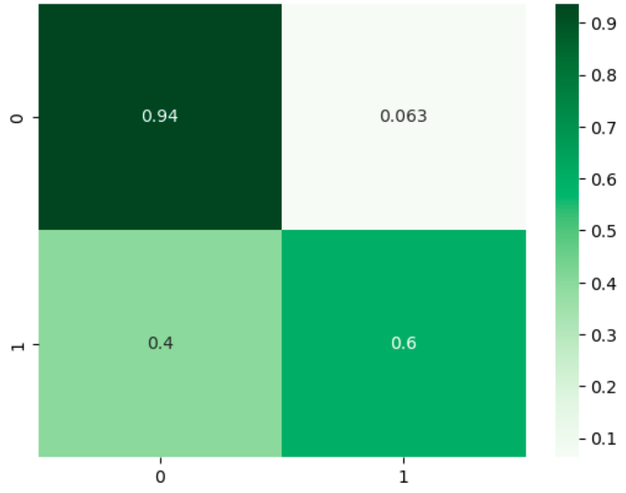


	precision	recall	f1-score	support
0	0.91	0.91	0.91	25979
1	0.67	0.65	0.66	7004
accuracy			0.86	32983
macro avg	0.79	0.78	0.78	32983
weighted avg	0.86	0.86	0.86	32983

XGB optimisé selon f2_weighted

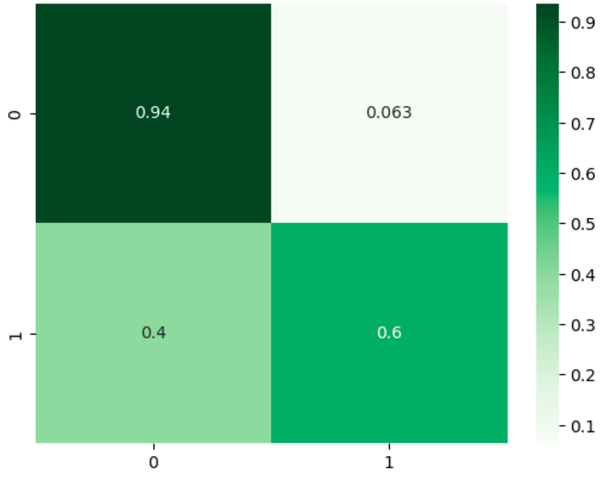
f2_weighted : 0.8622

TRAIN



	precision	recall	f1-score	support
0	0.90	0.94	0.92	104023
1	0.72	0.60	0.66	27905
accuracy			0.87	131928
macro avg	0.81	0.77	0.79	131928
weighted avg	0.86	0.87	0.86	131928

TEST



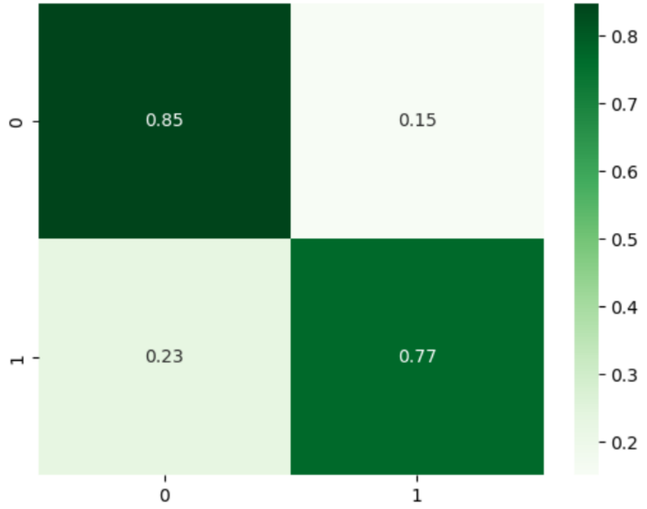
	precision	recall	f1-score	support
0	0.90	0.94	0.92	25979
1	0.72	0.60	0.66	7004
accuracy			0.87	32983
macro avg	0.81	0.77	0.79	32983
weighted avg	0.86	0.87	0.86	32983

Présentation des résultats

LGBM optimisé selon AUC

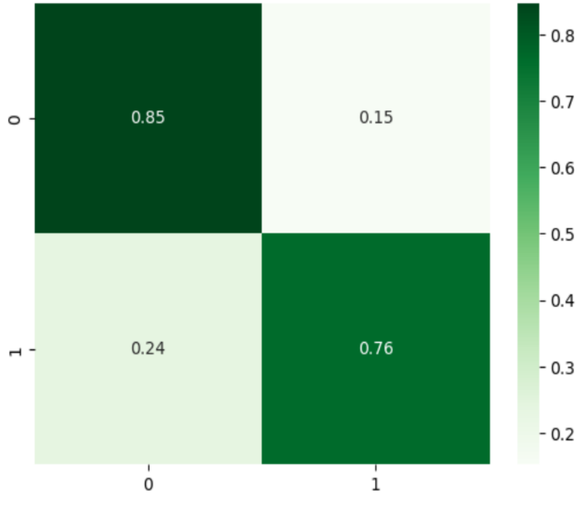
AUC: 0.8887

TRAIN



	precision	recall	f1-score	support
0	0.93	0.85	0.89	104023
1	0.58	0.77	0.66	27905
accuracy			0.83	131928
macro avg	0.75	0.81	0.77	131928
weighted avg	0.86	0.83	0.84	131928

TEST

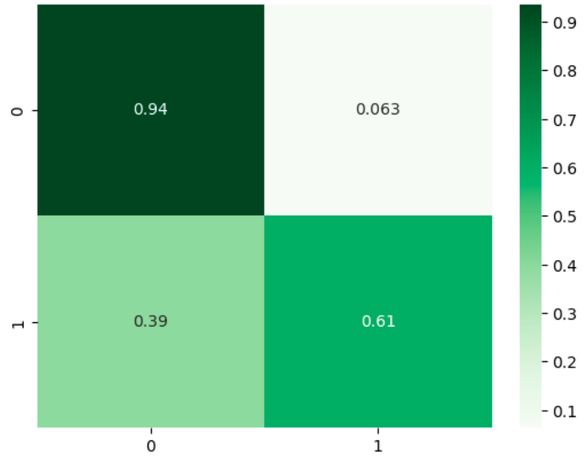


	precision	recall	f1-score	support
0	0.93	0.85	0.89	25979
1	0.58	0.76	0.66	7004
accuracy			0.83	32983
macro avg	0.75	0.81	0.77	32983
weighted avg	0.86	0.83	0.84	32983

LGBM optimisé selon f2_weighted

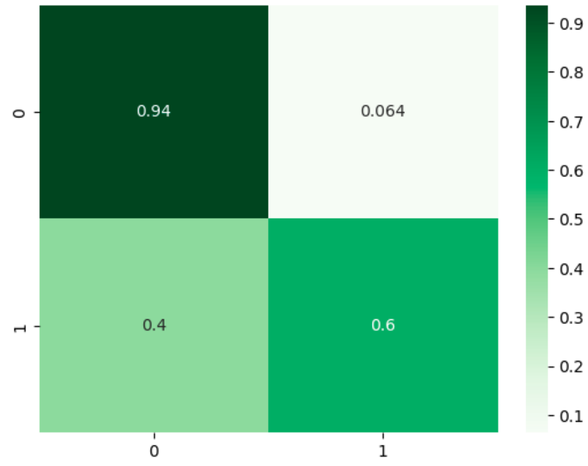
f2_weighted : 0.8625

TRAIN



	precision	recall	f1-score	support
0	0.90	0.94	0.92	104023
1	0.72	0.61	0.66	27905
accuracy			0.87	131928
macro avg	0.81	0.77	0.79	131928
weighted avg	0.86	0.87	0.86	131928

TEST



	precision	recall	f1-score	support
0	0.90	0.94	0.92	25979
1	0.72	0.60	0.66	7004
accuracy			0.87	32983
macro avg	0.81	0.77	0.79	32983
weighted avg	0.86	0.87	0.86	32983

03.Choix du Meilleur Modèle

XGBOOST Classifier

Déploiement et mise en production du modèle

fortuneo BANQUE

Prédiction de désabonnement des clients de la banque Fortuneo

Bienvenue dans notre outil de prédiction de résiliation de comptes clients. Cet outil a été conçu pour vous aider à anticiper et à gérer les désabonnements de vos clients, vous permettant ainsi d'améliorer leur fidélité et satisfaction. Ce modèle de scoring prédit la probabilité de résiliation pour chaque client et le classe dans une catégorie (churn ou pas). Veuillez renseigner les informations suivantes pour voir les prédictions du modèle.

Sélectionnez la zone

France

Sélectionnez le genre du client

Male

Le client a t-il une carte de crédit

Oui

Le client est il membre actif

Oui

Score de Crédit du client

0

Sélectionnez la zone

France

Sélectionnez le genre du client

Female

Le client a t-il une carte de crédit

Oui

Le client est il membre actif

Oui

21

Le nombre d'année en tant que client de la banque

1

Solde actuelle du compte

13908

Nombre de produits bancaires utilisé par le client

1

Salaire estimé du client

17300

Prédiction

✅ Client non désabonné avec une probabilité de 85.03%

MERCI POUR VOTRE ATTENTION