CHALLENGE DATA MINING

VO Nguyen Thao Nhi
LAMOUR Samanta

SOMMAIRE

- Présentation du challenge (objectif)
- Présentation des données
- Analyse exploratoire des données
- Modélisation
- Conclusion

PRÉSENTATION DU CHALLENGE

- L'objectif principal de ce challenge est de développer un ou plusieurs modèle des classification capable de prédire la probabilité de churn (désengagement) des clients d'une banque.

- Les soumissions doivent respecter le format demandé (submission en format csv avec id et les probabilités de se désengager ou pas) et évaluer les prédictions sur les données de test en termes de probabilité, ce qui exige un focus sur des métriques comme l’AUC.

PRÉSENTATION DES DONNÉES

- Nous avions à notre disposition deux fichier csv , l’un représente les données d’entrainement et l’autre , l’un les données de test
- La base de données train a 15000 observations et 14 variables
- La base de données test a 10000 observations et 13 variables, ie elle ne contient pas la variable cible
- La variable cible Exited binaire représente si les clients ayant quitté la banque (1) ou non (0)
- Le dataset fourni contient des informations détaillées sur les clients, incluant des caractéristiques sociodémographiques (âge, genre, localisation), des comportements financiers (score de crédit, solde, nombre de produits utilisés), et des attributs relationnels (ancienneté avec la banque, statut de membre actif).

ANALYSE EXPLORATOIRE DES DONNÉES

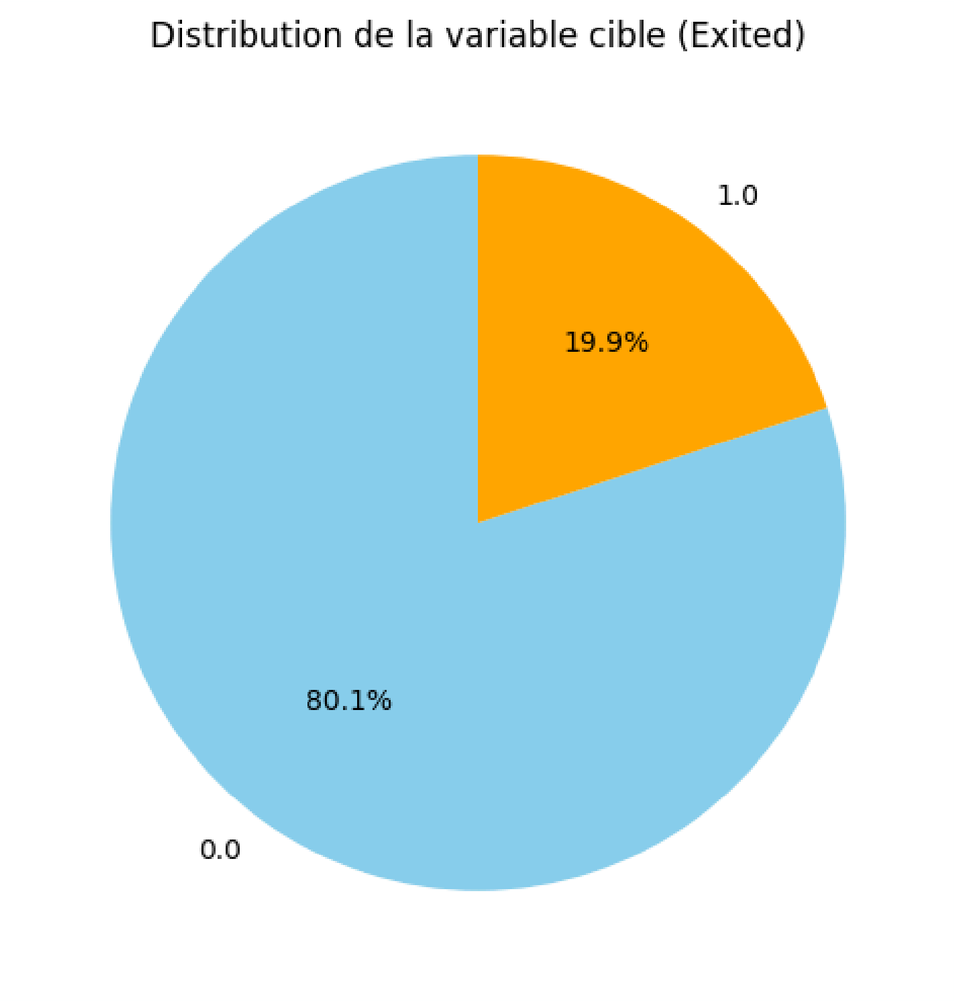
ANALYSE EXPLORATOIRE DE DONNÉES

- Les jeux de données train.csv et test.csv ont été chargés et inspectés. Les deux fichiers sont complets, sans valeurs manquantes, ce qui facilite leur traitement.

- Les données d'entraînement ont été divisées en deux ensembles :

1.-Entraînement : 80% des données.
2.-Validation : 20% des données, stratifiées selon la variable cible Exited pour maintenir une distribution équilibrée.

ANALYSE DE LA VARIABLE CIBLE



DÉSÉQUILIBRE DE CLASSE

La répartition de la variable Exited (clients ayant quitté la banque) est la suivante :
Ensemble d'entraînement : environ 80% de non-churn (0) et 20% de churn (1).
Ensemble de validation : répartition similaire, confirmant une bonne stratification.

ANALYSE DES AUTRES VARIABLES

- Population Stability Index (PSI)

Le PSI a été calculé pour évaluer la stabilité des distributions des variables quantitatives entre les ensembles d'entraînement et de test.

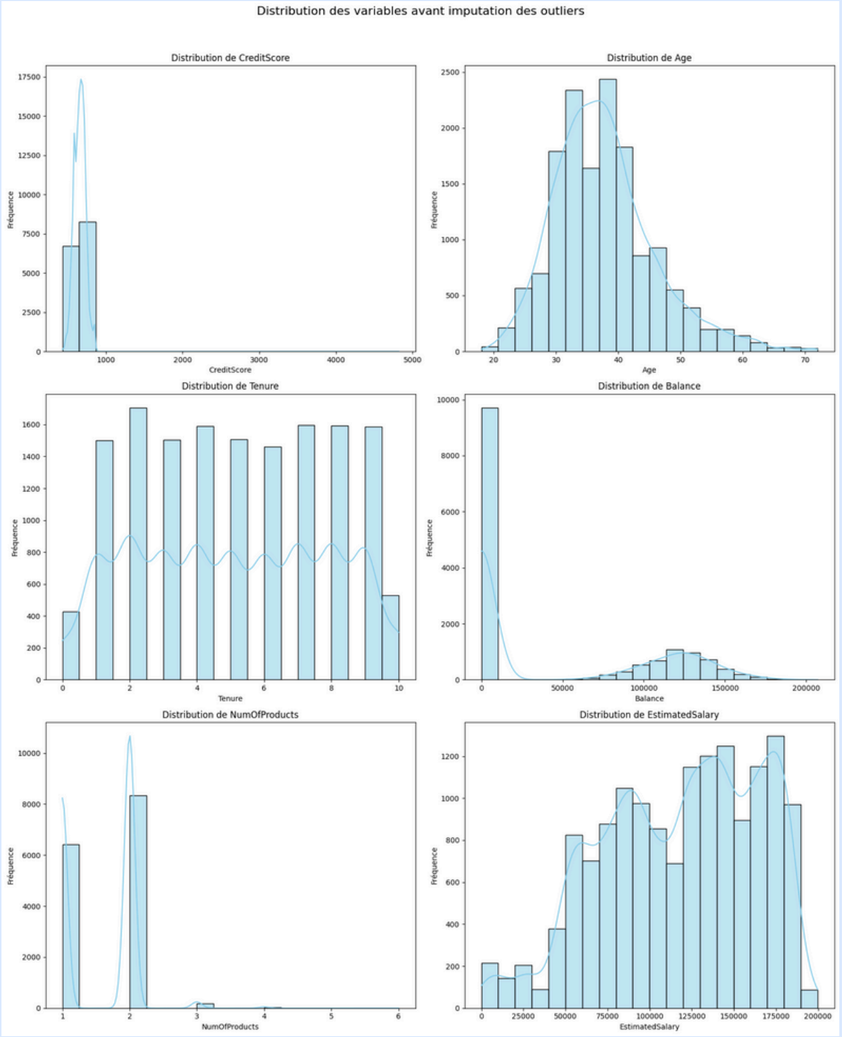
- Résultats :

les variables comme présentent des distributions similaires (PSI < 0.1).

ANALYSE UNIVARIÉE

- **Variables quantitatives**

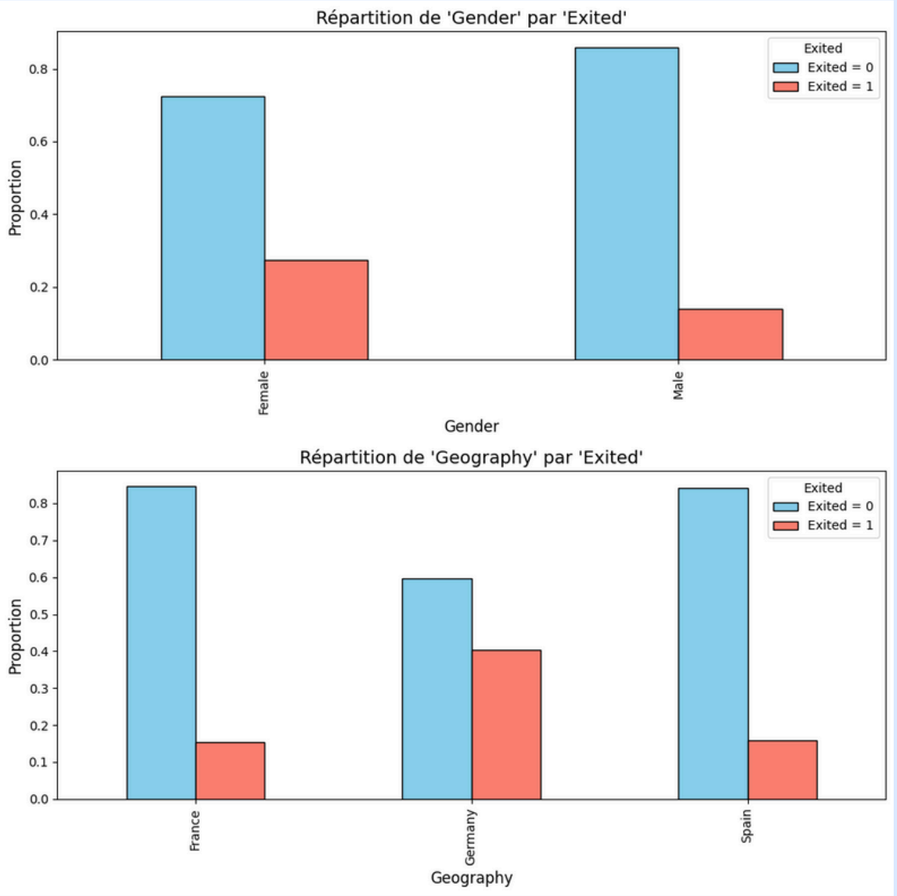
Les variables montrent une certaine asymétrie



ANALYSE UNIVARIÉE

- **Variables catégorielles**

La répartition des genres (Gender) et des régions (Geography) a été examinée par rapport à Exited, révélant des corrélations potentielles.



ANALYSE UNIVARIÉE

- **Détection des doublons et outliers**

Aucun doublon n’a été trouvé dans les données.
Les outliers ont été détectés et traités à l'aide de la méthode IQR, et certaines valeurs extrêmes (ex. : CreditScore > 1000) ont été supprimées.

ANALYSE UNIVARIÉE

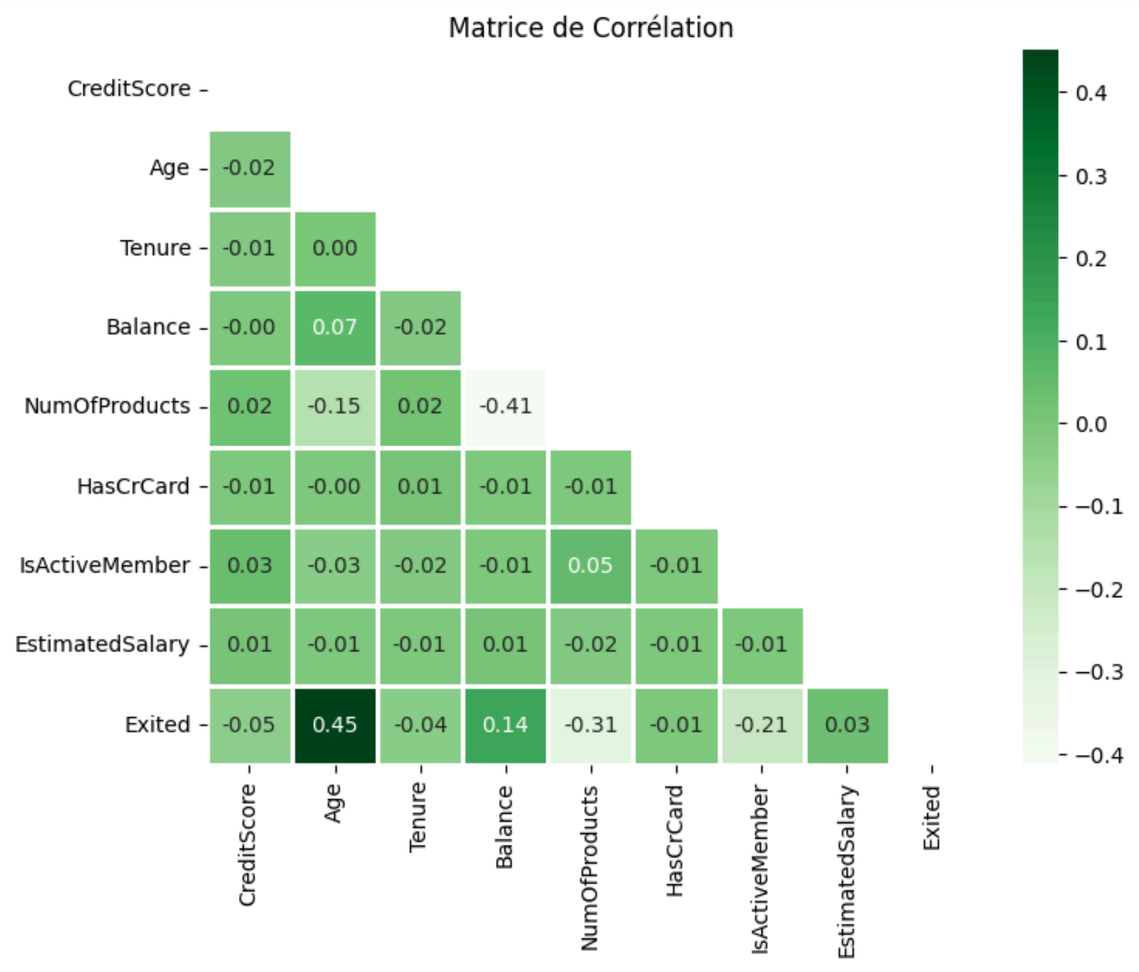
- **Encodage et transformation des variables**

Les variables catégorielles (Geography, Gender) ont été encodées avec un OneHotEncoder.

- **Nettoyage des données**

- Les colonnes inutiles (Surname, CustomerId, etc.) ont été supprimées.
- Les jeux de données nettoyés et transformés ont été exportés pour les étapes suivantes.

CORRELATION

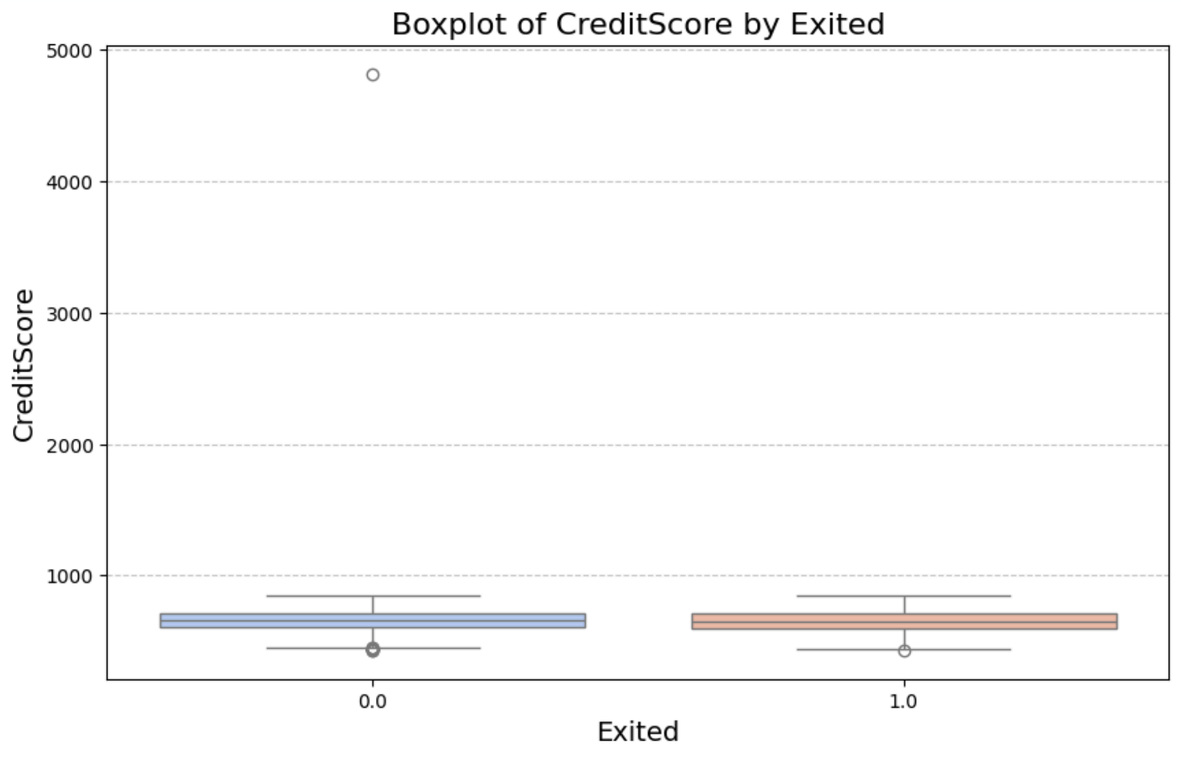


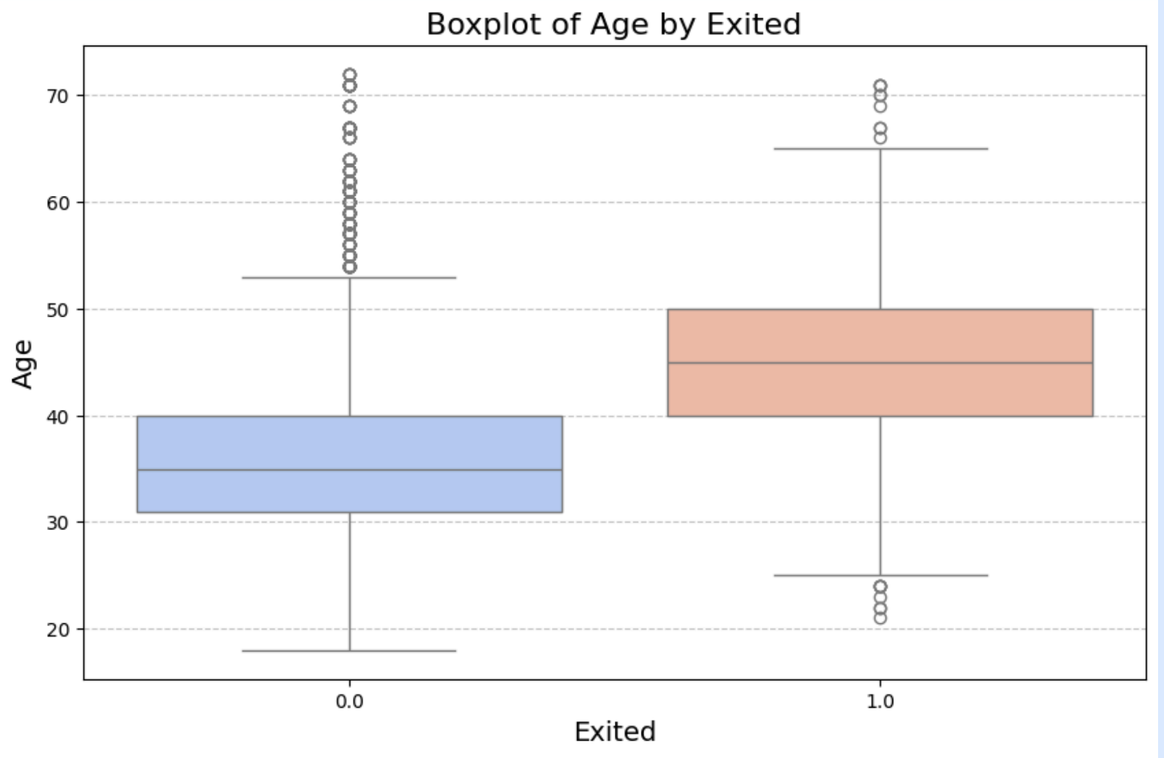
PAS DE FORTE CORRÉLATION ENTRE LES VARIABLES

PRE PROCESSING

- Division des données pour l'entraînement (70%) et la validation (30%) en conservant les proportions des classes dans les deux sous-ensembles.
- Détection des outliers avec la méthode IQR: Age et CreditScore

OUTLIERS


Boxplot of CreditScore by Exited
CreditScore
5000
4000
3000
2000
1000
0.0
1.0
Exited


Boxplot of Age by Exited
Age
70
60
50
40
30
20
0.0
1.0
Exited

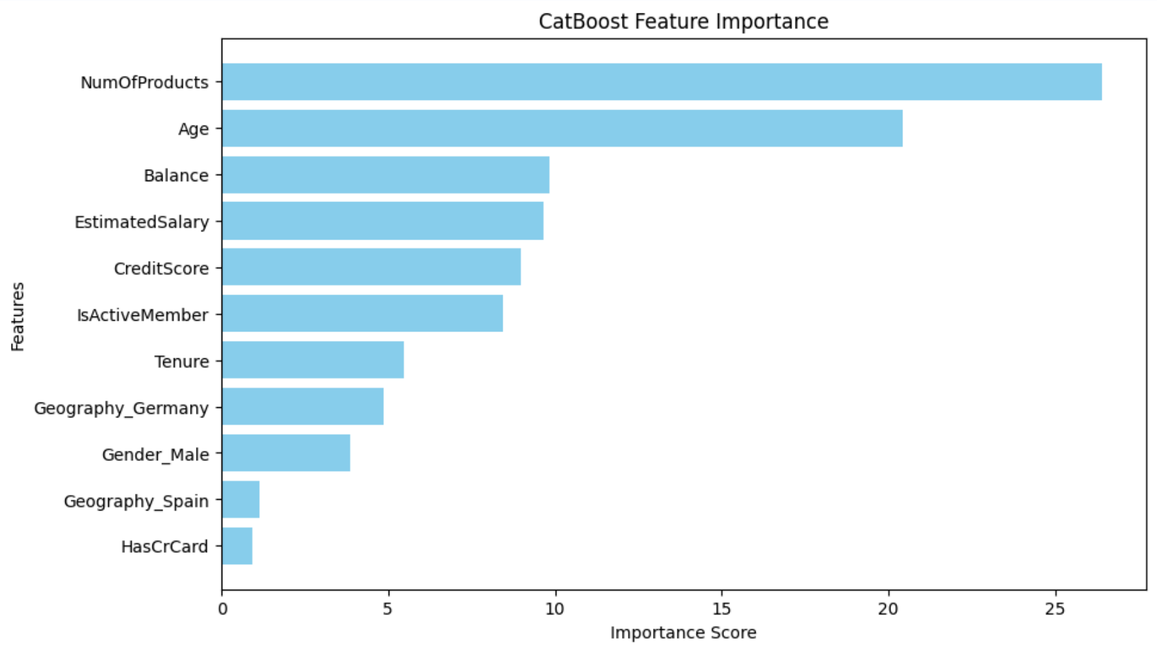
MODELES

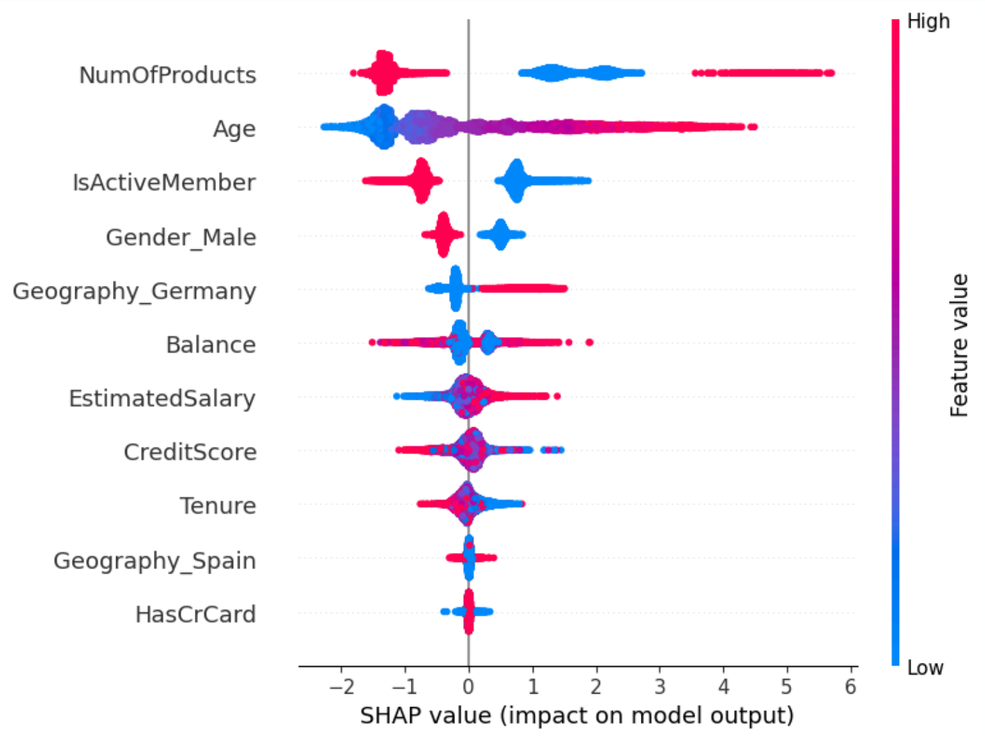
MODÈLE 1

- Modèle: Catboost
- Division des données pour l'entraînement (70%) et la validation (30%) en conservant les proportions des classes dans les deux sous-ensembles.
- Fine tuner les hyperparamètres: GridSearchCV, StratifiedKFold (5 splits)
- Les hyperparamètres finaux:

```
params = {
    'depth': 6, 'iterations': 500, 'learning_rate': 0.01,
    'verbose': 0,
    "loss_function": "Logloss",
    'cat_features': categorical_features,
    'random_state':42
}
```

FEATURE IMPORTANCE



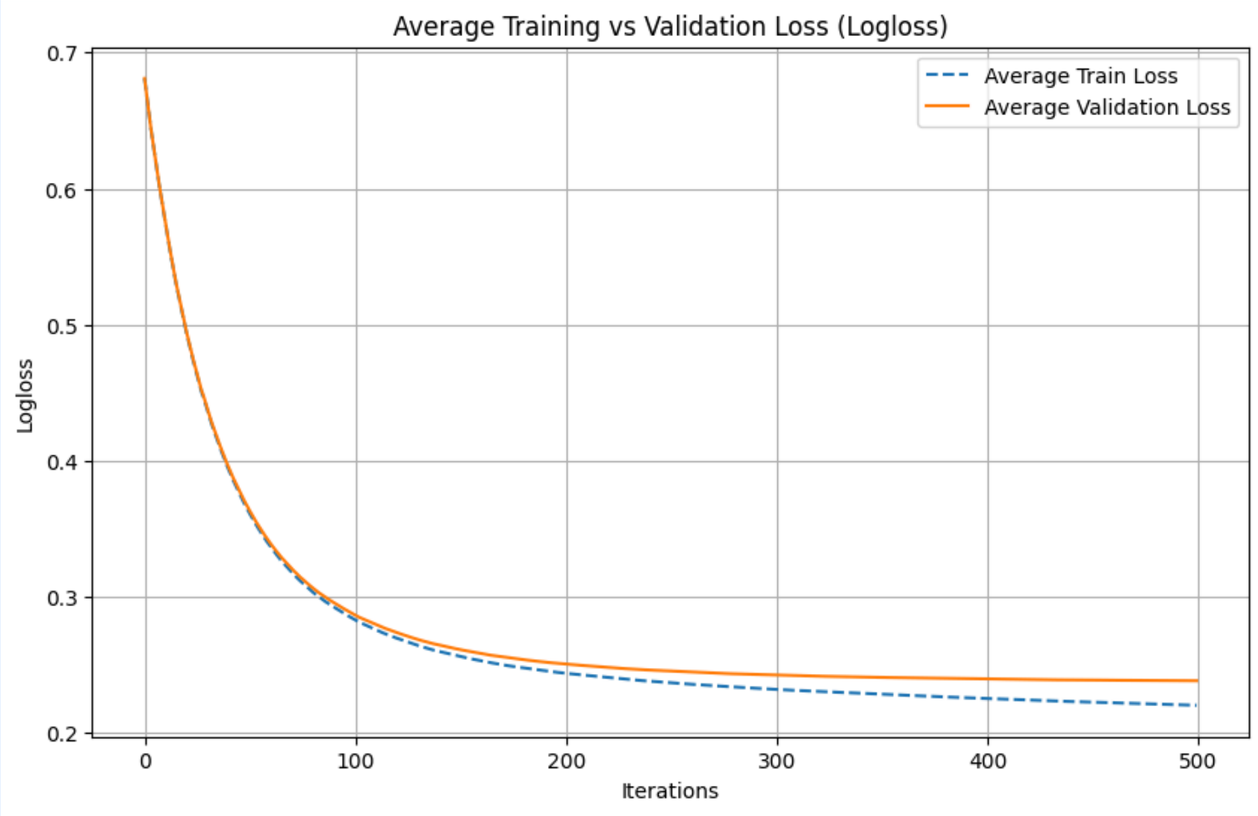


Feature Importance: Analyse de l'importance des variables avec CatBoost et SHAP

MODÈLE 1

EVALUATION

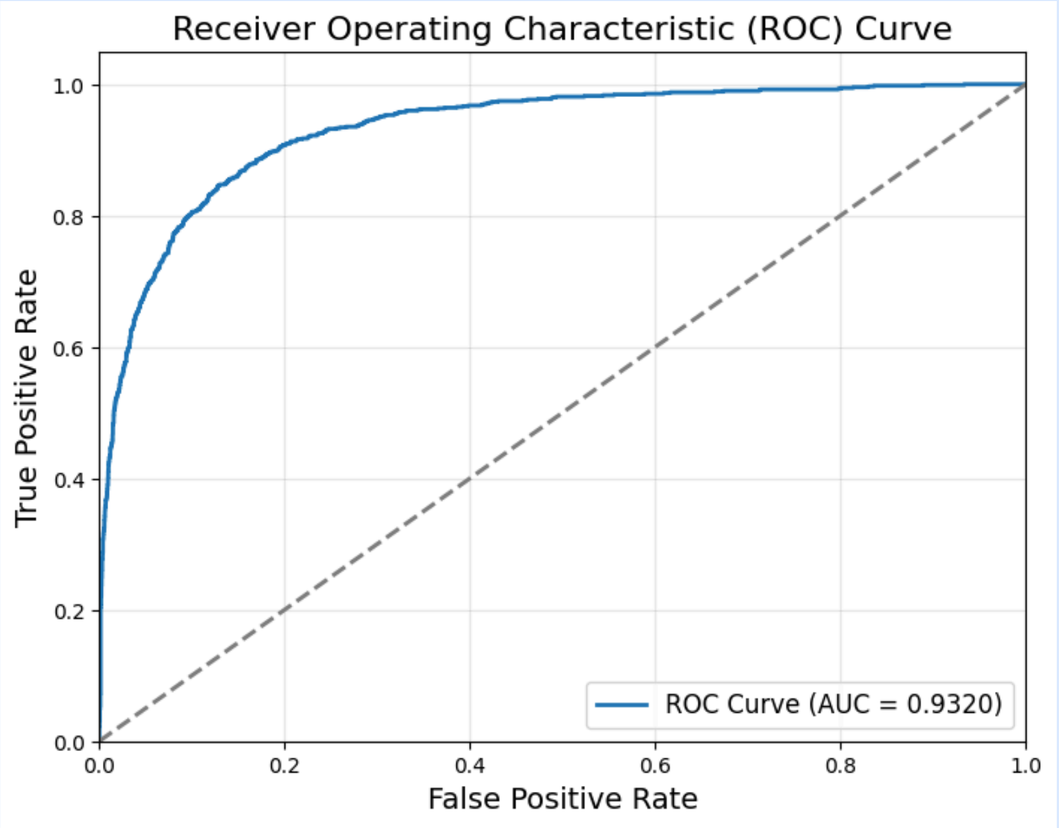
- Fold 1: Train AUC = 0.9446, Validation AUC = 0.9356
- Fold 2: Train AUC = 0.9438, Validation AUC = 0.9396
- Fold 3: Train AUC = 0.9429, Validation AUC = 0.9415
- Fold 4: Train AUC = 0.9439, Validation AUC = 0.9370
- Fold 5: Train AUC = 0.9425, Validation AUC = 0.9405
-
- Mean Train AUC: 0.9435
- Mean Validation AUC: 0.9388
- Validation AUC Standard Deviation: 0.0022



MODÈLE 1

EVALUATION

- AUC Score pour l'ensemble de validation: 0.9320
- AUC Score pour les données publiques sur Kaggle: 0.9326



MODÈLE 2

- Modèle: XGBoost
- Fine tuner les hyperparamètres: OPTUNA
- Les variables quantitatives ont été normalisées (Min-Max Scaling) pour homogénéiser les échelles.

PARAMS = {'N_ESTIMATORS': 343,
'MAX_DEPTH': 8,
'LEARNING_RATE': 0.08571583489125943,
'SUBSAMPLE': 0.7817344756772059,
'COLSAMPLE_BYTREE': 0.5393607402976373,
'COLSAMPLE_BYLEVEL': 0.9209678660886534,
'GAMMA': 2.382541298014462,
'REG_ALPHA': 4.064590577146683,
'REG_LAMBDA': 8.924597552521014,
'MIN_CHILD_WEIGHT': 5,
'RANDOM_STATE': 8687}

- AUC Score pour l'ensemble de validation: 0.9344
- AUC Score pour les données publiques sur Kaggle: 0.9321

MODÈLE 3

- Modèle: HistGradientBoostingClassifier
- Fine tuner les hyperparamètres: OPTUNA

LE MODÈLE HISTGRADIENTBOOSTINGCLASSIFIER FAIT PARTIE DES ALGORITHMES DE BOOSTING. IL EST PARTICULIÈREMENT OPTIMISÉ POUR DES DONNÉES DE GRANDE DIMENSION GRÂCE À SON IMPLÉMENTATION HISTOGRAMME.

MODÈLE 3

Le modèle HistGradientBoostingClassifier présente des limites potentielles, notamment un risque d'overfitting dû à une mauvaise calibration des hyperparamètres ou à la présence de bruit et de redondances dans les données, ce qui peut entraîner des performances élevées en validation croisée mais une dégradation sur les données de test. Il est également sensible aux biais dans les données, notamment en cas de sur- ou sous-représentation de certaines classes ou d'une gestion inadéquate des variables catégoriques.

- AUC SCORE POUR L'ENSEMBLE DE VALIDATION: 0.9321
- AUC SCORE POUR LES DONNÉES PUBLIQUES SUR KAGGLE: 0.9288

AUTRES MODÈLES

Tableau Résumé : Autres Modèles Testés

Modèle	Description
XGBoost Simple	XGBoost avec paramètres par défaut ou légèrement optimisés.
XGBoost avec SMOTE	XGBoost appliqué sur un dataset rééquilibré via SMOTE.
LightGBM Simple	LightGBM avec configuration par défaut ou légèrement optimisée.
LightGBM avec SMOTE	LightGBM appliqué sur des données rééquilibrées.
Random Forest Simple	Modèle de forêt aléatoire classique avec hyperparamètres de base.

AUTRES MODÈLES

Random Forest avec SMOTE	Forêt aléatoire appliquée sur un dataset rééquilibré.
Stacking avec Régression Logistique	Combinaison des prédictions de tous les modèles testés, pondérées via une régression logistique.
Modèle de Deep Learning	Réseau de neurones profond optimisé pour la classification.

CONCLUSION/CHOIX FINAL

FIN DU CHALLENGE