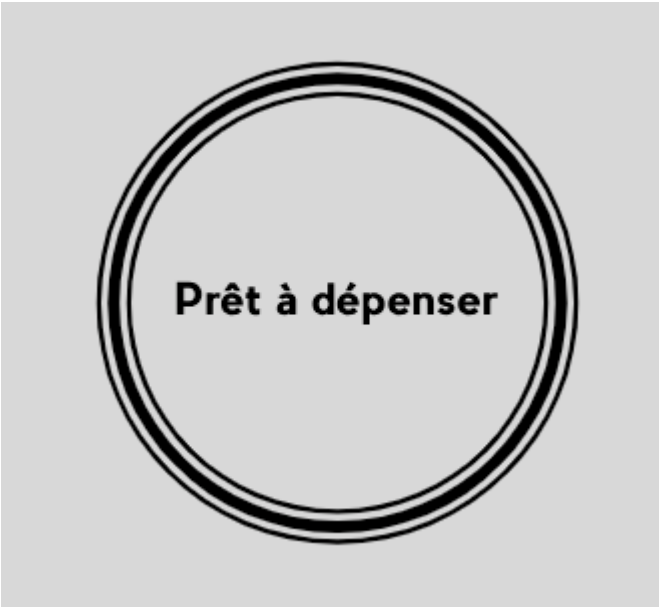
PROJET N°7

"IMPLEMENTEZ UN MODELE DE SCORING"

SOMMAIRE

AVANT PROPOS



CONTEXTE & OBJECTIFS :

La société financière Prêt à dépenser propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

L'entreprise souhaite

- ✓ mettre en œuvre un **outil de scoring crédit** qui calcule la probabilité qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé.

- ✓ - développer **un dashboard interactif** pour plus de transparence afin que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

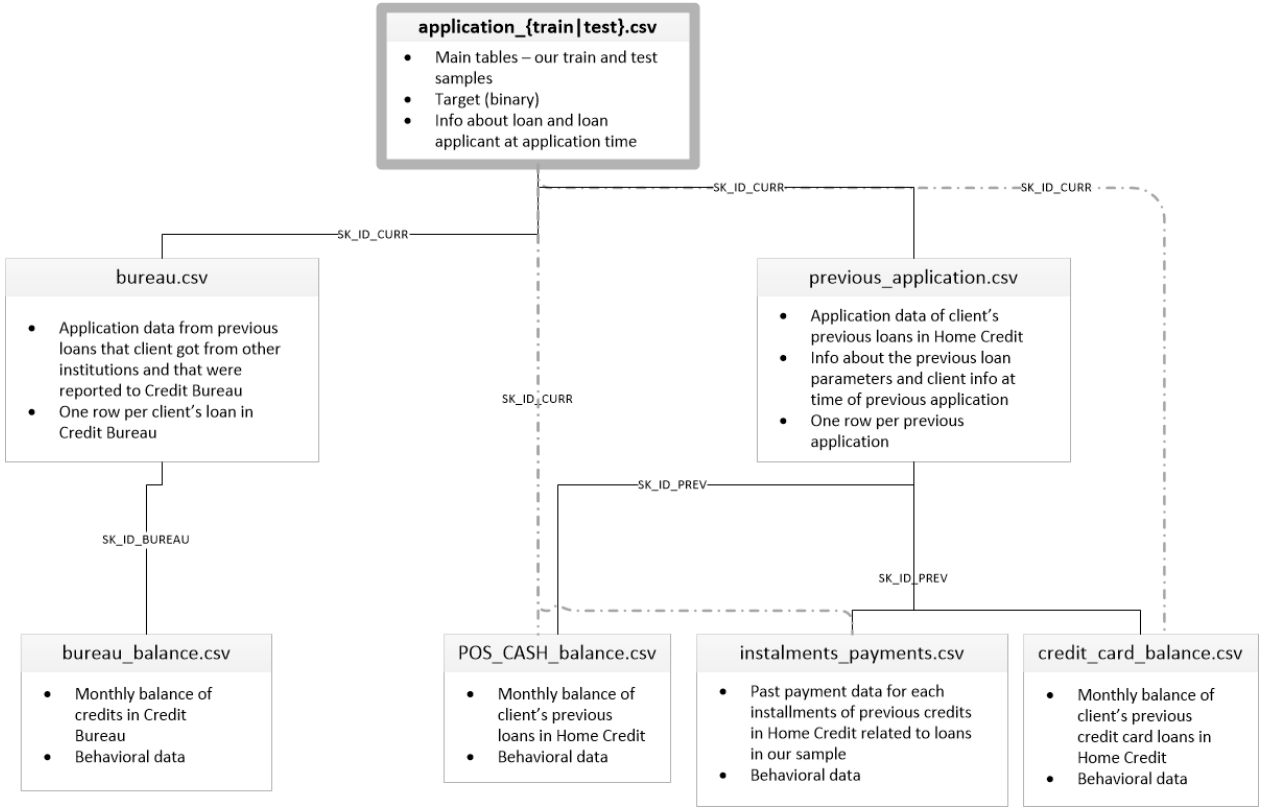
ETAPE 1 :

DONNÉES FOURNIES & DATA PREPROCESSING

DONNEES FOURNIES

7 Datasets:

- ✓ Reprenant des informations personnelles du client et sur le crédit souhaité
- ✓ Des historiques des crédits dans l'établissement bancaire
- ✓ Des historiques des crédits demandés dans d'autres institutions financières

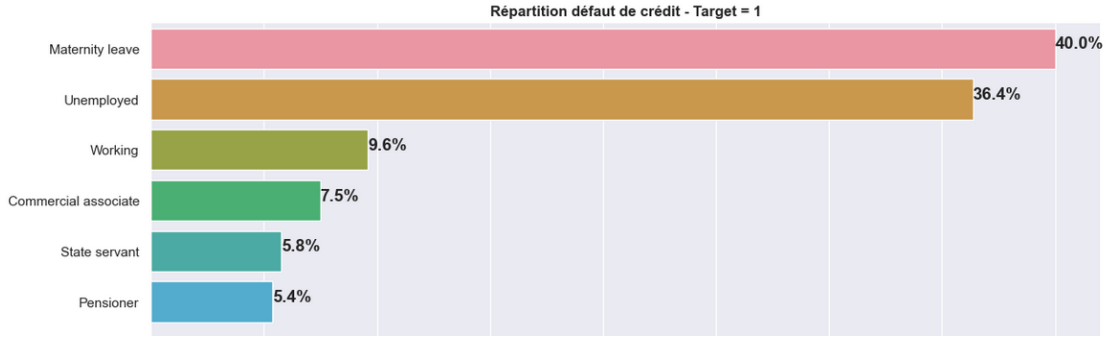


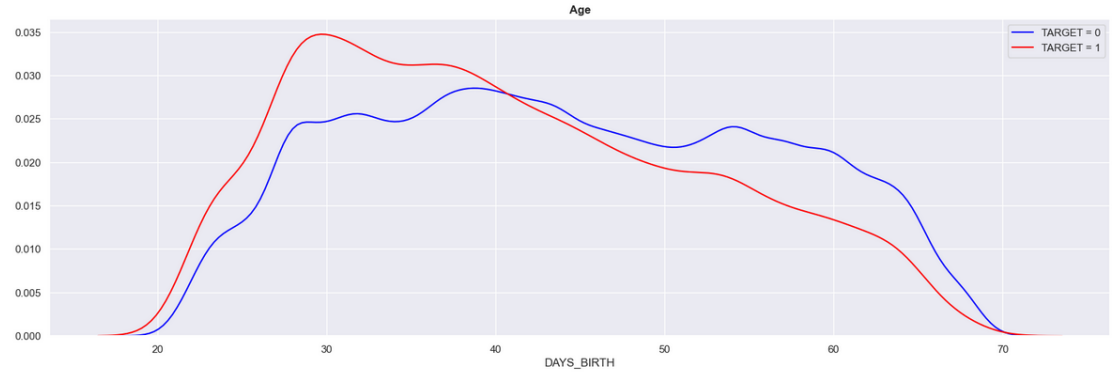
DATA PREPROCESSING

1 Analyse des 7 datasets séparement

- ✓ suppression des features non pertinentes
- ✓ Analyse des variables manquantes ou aberrantes
- ✓ Suppression de features fortement correlées

2 Analyse exploratoire





3 Agrégation des data

- ✓ Sur une data globale (train + test), intégration features brut ou agrégé des autres fichiers
- ✓ Traitement des valeurs manquantes (Avec simpleimputer – Médiane)

DATA PREPROCESSING

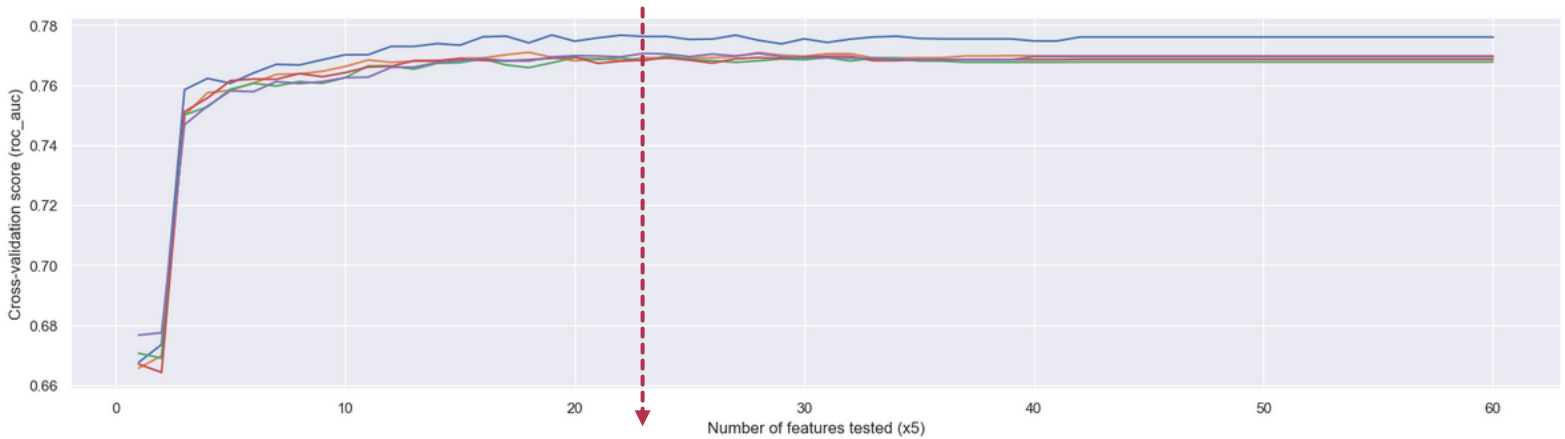
4 Préprocessing des données

- ✓ Création de 4 nouvelles variables métier
- ✓ Encodage des variables catégorielles (OneHotEncoder)
- ✓ Standardisation des données (StandardScaler)

----→ **A ce stade : 174 features**

5 Réduction de dimension

- ✓ Utilisation de la technique d'élimination des caractéristiques récursives avec validation croisée (RFECV).



Seuil Optimal AUC/Features : 92

ETAPE 2 :

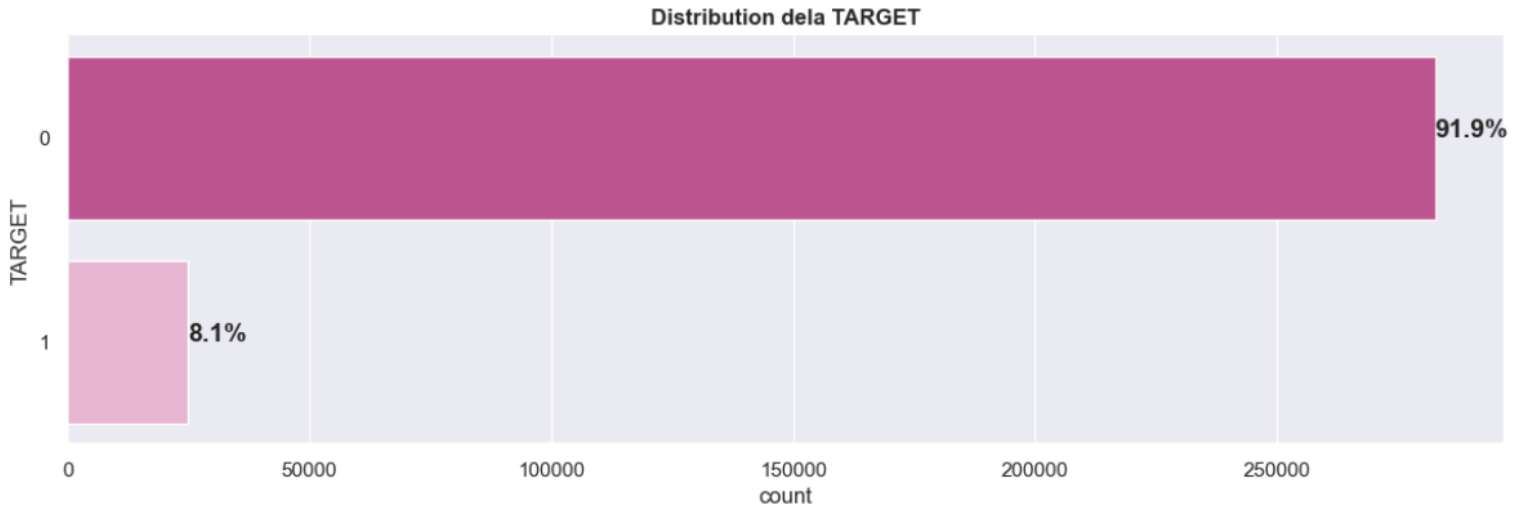
MODELISATION

MODELISATION

TRAITEMENT DU DESEQUILIBRE DES CLASSES

PROBLEME : Les clients en difficulté de paiement sont largement sous-représentés (8.1%) dans les données d'entraînement

→ Mauvais résultat ou induire en erreur avec des scores trop optimistes



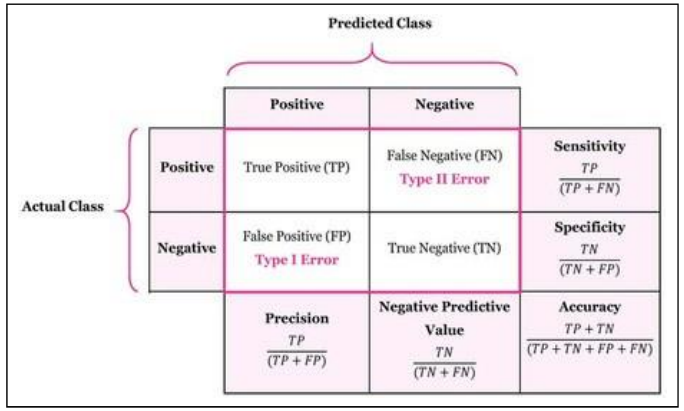
SOLUTIONS :

- ✓ **Méthodes de Data Level** qui consistent en une modification des données d'entraînement (comme over sampling, under sampling ou SMOTE (Suréchantillonnage)
- ✓ **Les Méthodes Algorithm-level** qui ne modifient pas les données, mais qui pénalisent les mauvaises prédictions sur la classe minoritaire en donnant plus de poids à la fonction de perte

→ La méthode RandomUnderSampling sera mise en avant dans ce rapport, celle-ci étant suffisante et les autres méthodes n'ayant pas donné de meilleurs résultats

MODELISATION
METRIQUES & SCORE METIER

La matrice de confusion



Les métriques classiques

Accuracy	Rappel	Précision	F1-score
% de bonnes prédictions	% de la classe positive détectée	% des vraies positifs dans les positifs détectés	Moyenne harmonique de Précision / Rappel
→ Affecté par le déséquilibre des classes → Ne distingue pas les erreurs commises	→ Reflète la capacité à éviter les faux positifs	→ Reflète la capacité à détecter tous les cas positifs	→ Combinaison du rappel et de la précision → A privilégier si déséquilibre des classes

Le score métier : d'un point de vue métier, le risque de refuser un crédit à tort (faux négatifs) n'a évidemment pas le même poids que d'accorder un crédit à un client n'ayant pas honoré son engagement (faux positifs).

→ Création d'un **score métier normalisé** en prenant en compte un risque 10 x plus élevé de prédire un faux positif qu'un faux négatif. Un poids de 1 est attribué aux bonnes prédictions :

$$J = 1 * TP + 1 * TN - 1 * FP - 10 * FN$$

$$SCORE_{METIER} = \frac{(J - Jmin)}{(Jmax - Jmin)} \quad \text{Où} \quad Jmax = (FP + TN) * 1 + (FN + TP) * 1$$

$$Jmin = (FP + TN) * (-1) + (FN + TP) * (-10)$$

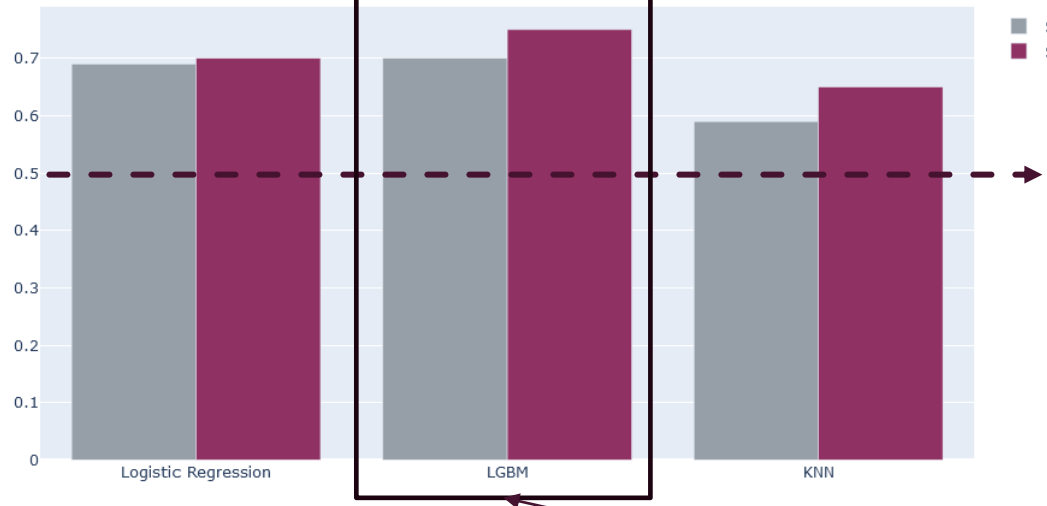
MODELISATION

MODELISATION

La modélisation a été élaborée :

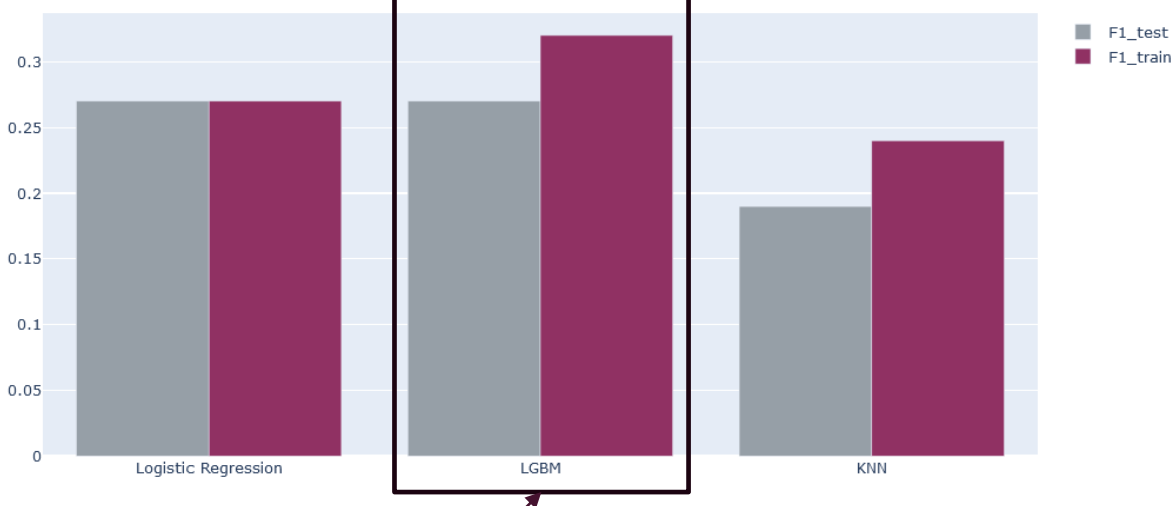
- Avec un pipeline de transformation
- Une optimisation des paramètres
- et avec GridsearchCV et 5 cross validation, ce qui signifie que le modèle est tester 5 fois pour chaque ensemble d'hyperparamètres sélectionnés.

SCORE METIER



Baseline

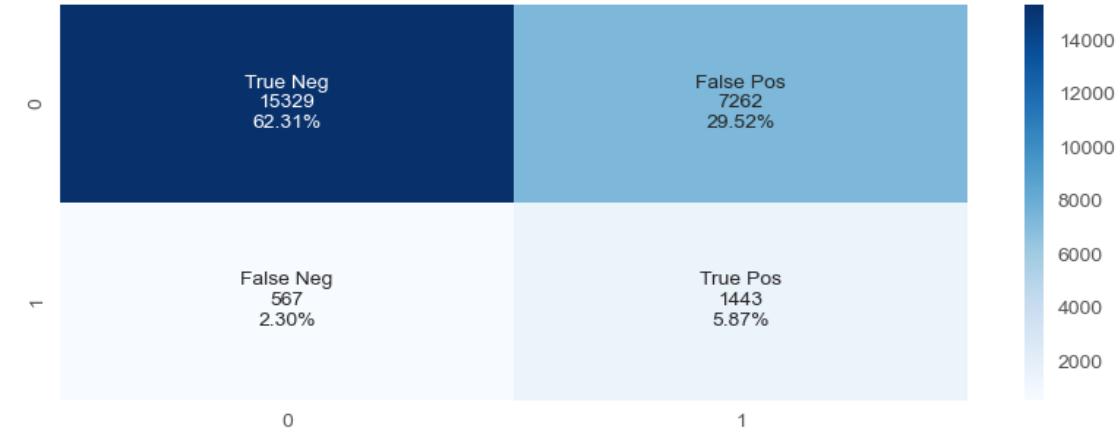
F1-SCORE

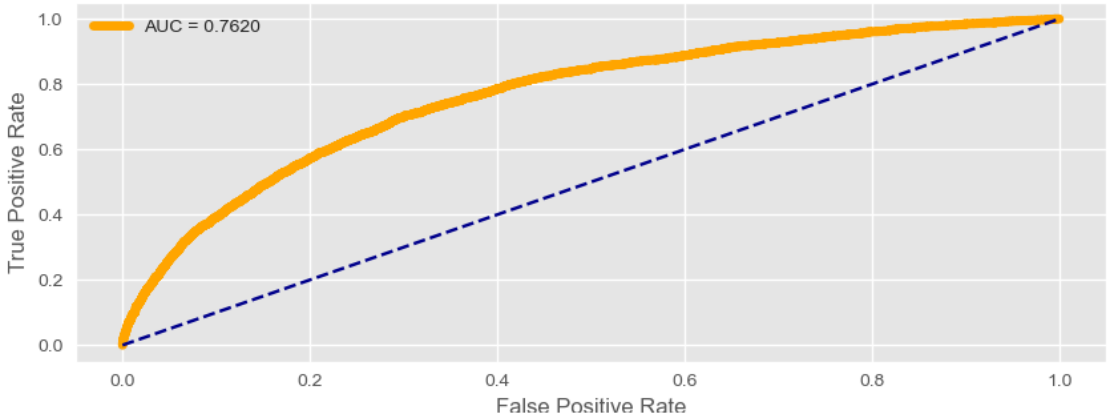


Meilleurs résultats pour le Score métier comme pour le F1_Score avec LGBM

MODELISATION

MODELISATION





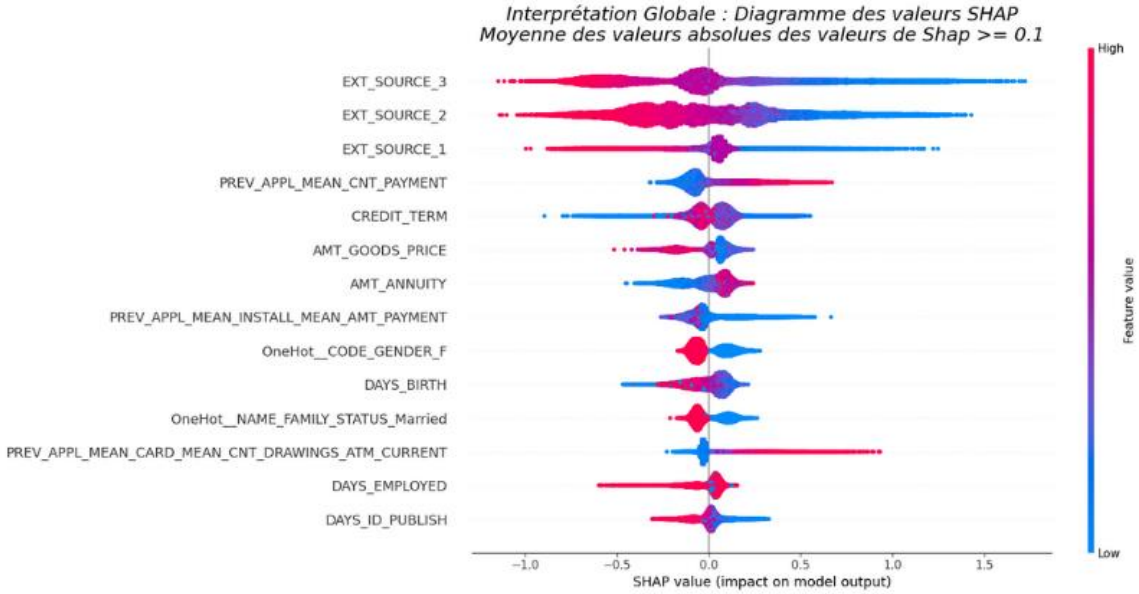
Ci-contre la matrice de confusion et la courbe ROC, qui montre un AUC de 0,762

On peut également noter que le seuil optimal du seuil de probabilité est aux alentours de 0,48

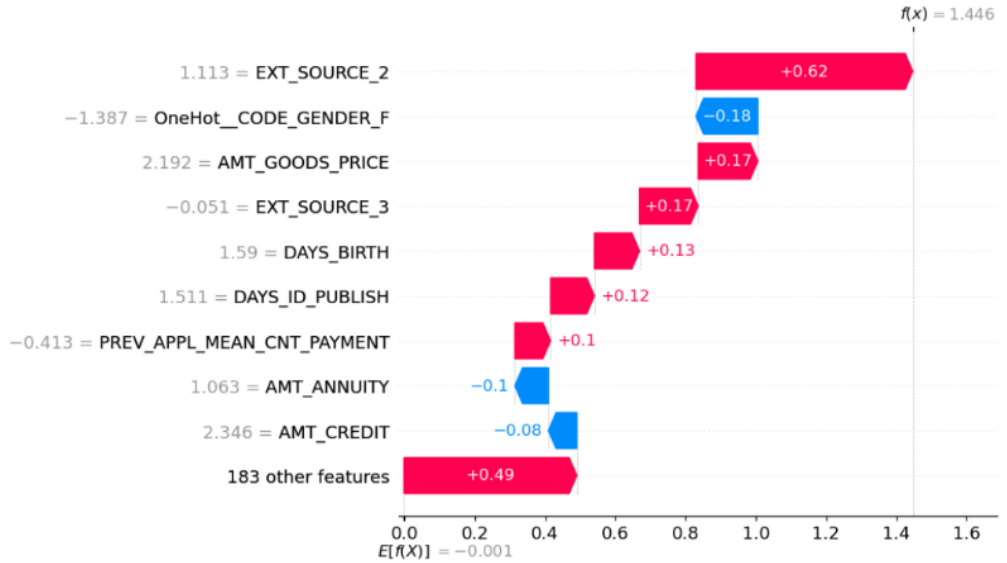
MODELISATION
INTERPRETATION DU MODELE

LES VALEURS DE SHAP : elles calculent l'importance d'une variable en comparant ce qu'un modèle prédit avec et sans cette variable. Cependant, étant donné que l'ordre dans lequel un modèle voit les variables peut affecter ses prédictions, cela se fait dans tous les ordres possibles, afin que les fonctionnalités soient comparées équitablement.

INTERPRETATION GLOBALE



INTERPRETATION LOCALE



ETAPE 3 :

DASHBOARD

DASHBOARD

DEPLOIEMENT DU MODELE – PLATEFORME / BIBLIOTHEQUE

Site de partage de code, sur lequel on peut publier des projets dont le code est géré avec le système de gestion de version Git.

Framework permettant de faire des tests et de vérifier si les différentes conditions sont juste ou fausse. Il permet de tester les éléments un à un mais on peut aussi lui demander de faire une série de tests.

Plateforme permettant de créer, déployer et faire évoluer vos applications (site statique rapide et gratuit). En associant le site à un dépôt Github, Render met automatiquement à jour votre site à chaque poussée vers la branche spécifiée.

Plateforme d'intégration continue et livraison continue (CI/CD) qui permet d'automatiser le pipeline de génération, de test et de déploiement grâce à un workflow (processus automatisé - un fichier YAML qui est déclenché par un événement dans votre dépôt

Bibliothèque open-source qui permet aux data scientists de créer des applications web pour la visualisation de données de manière rapide et efficace

DASHBOARD

PRINCIPE DE DEPLOIEMENT & INTEGRATION DES MODIFICATIONS

BACK

Dépôt GITHUB : https://github.com/helene1219/PROJET_7_BIS_BACK

Dépôt Github API / TEST

CI / CD

GitHub Actions

pytest

Déploiement sur Render – connecté au dépôt GITHUB

Render

https://scoring-p7.onrender.com

FRONT

Dépôt GITHUB : https://github.com/helene1219/PROJET_7_BIS_FRONT

Dépôt Github DASHBOARD – URL Render

Déploiement Web APP et Modification si new commit
- Connecté au dépôt GITHUB

Streamlit

https://scoring-p7.streamlit.app/

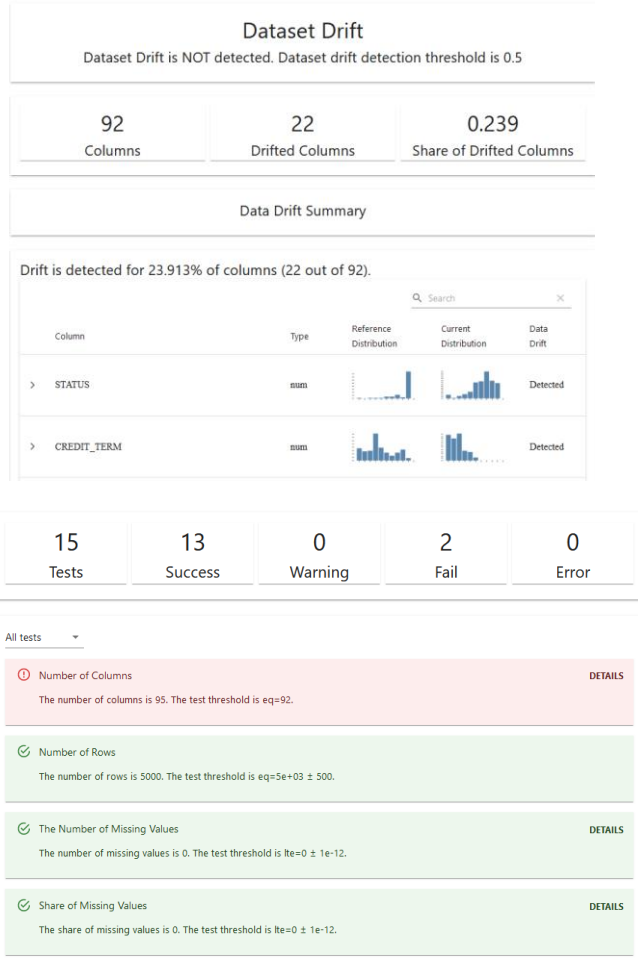
DASHBOARD
SUIVI DU MODELE

EVIDENTLY AI

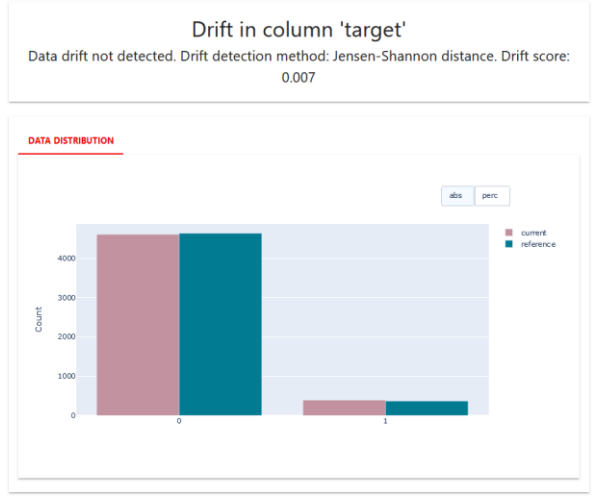
Il est essentiel de suivre un modèle en production et notamment :

- ✓ La dérive et qualité des données .
- ✓ La dérive de la cible et la perfomance du modèle

DERIVE ET QUALITE DES DONNEES

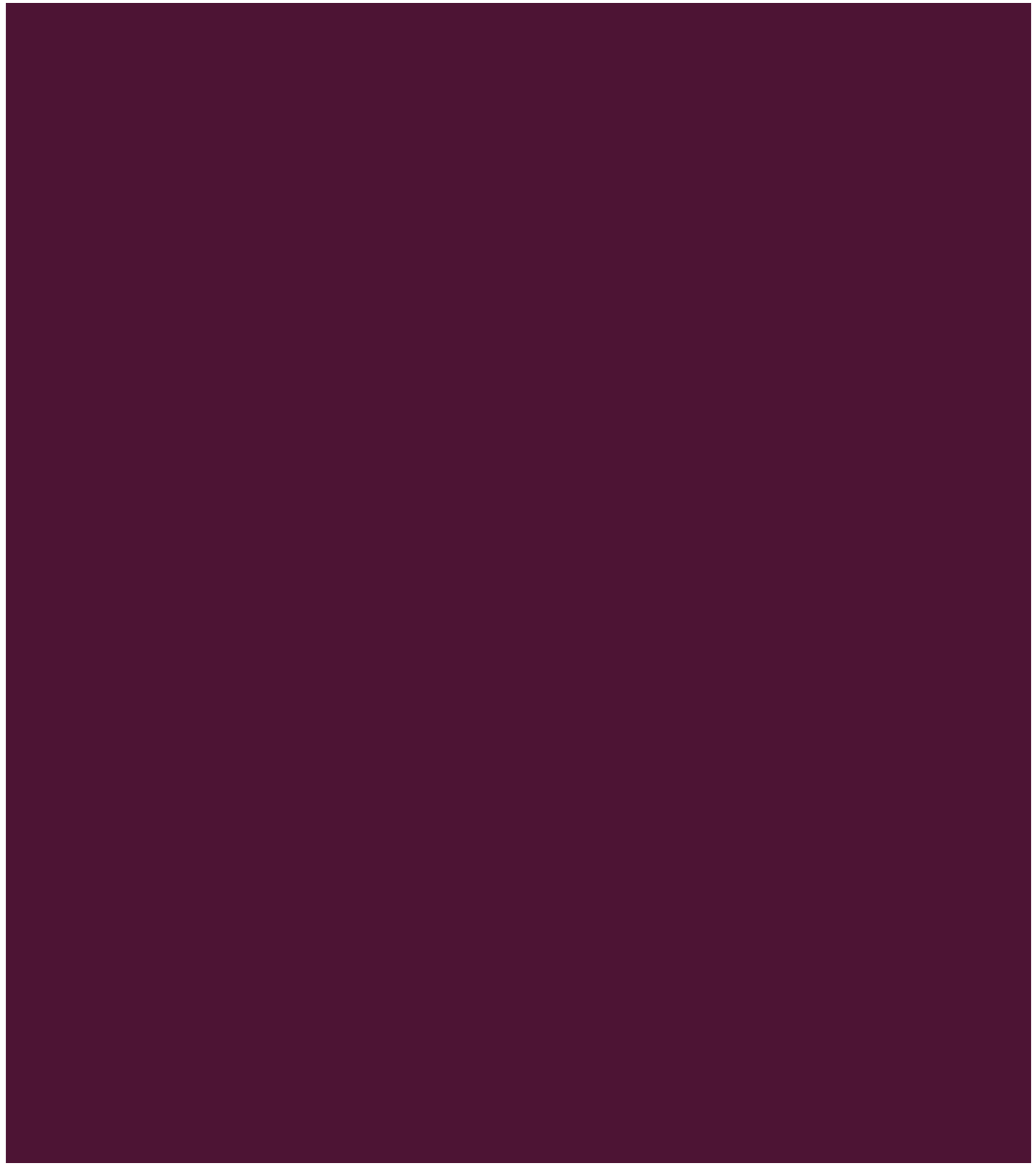


PERFORMANCE DU MODELE



ETAPE 4 :

CONCLUSION



CONCLUSION

Les données mise à disposition ont permis de mettre en place un algorithme de classification assez efficace avec une modélisation optimale avec l'algorithme LGBM. Cependant, afin de pouvoir améliorer ces résultats, il serait pertinent :

- ✓ **De disposer d'une meilleure méconnaissance du milieu bancaire** ce qui permettrait de vérifier / améliorer le processus de traitement des données
- ✓ D'avoir une meilleure compréhension des variables essentielles dans l'explication du modèle (EXT_SOURCE)
- ✓ De définir plus finement la **métrique d'évaluation** et la **fonction de coût** en collaboration avec les équipes métier
- ✓ D'améliorer les performances de la modélisation en intégrant de nouveaux hyperparamètres et / ou augmentant les valeurs testées
- ✓ D'avoir des connaissances de développement plus robuste afin d'optimiser le dashsbaord.