Projet 7 - Mauyves NKONDO - Décembre 2023

Implémenter un modèle de Scoring

Parcours Data Scientist - OpenClassrooms

https://github.com/MauyvesNk

OPENCLASSROOMS

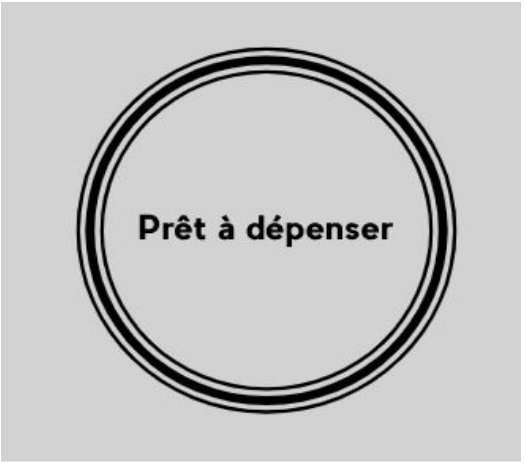
Sommaire :

Contexte, Objectifs et Set de Données

Partie I

Contexte et Objectifs :

La société financière **Prêt à Dépenser** offre des solutions de crédits à la consommation adaptées aux individus présentant un historique de prêt limité ou inexistant.



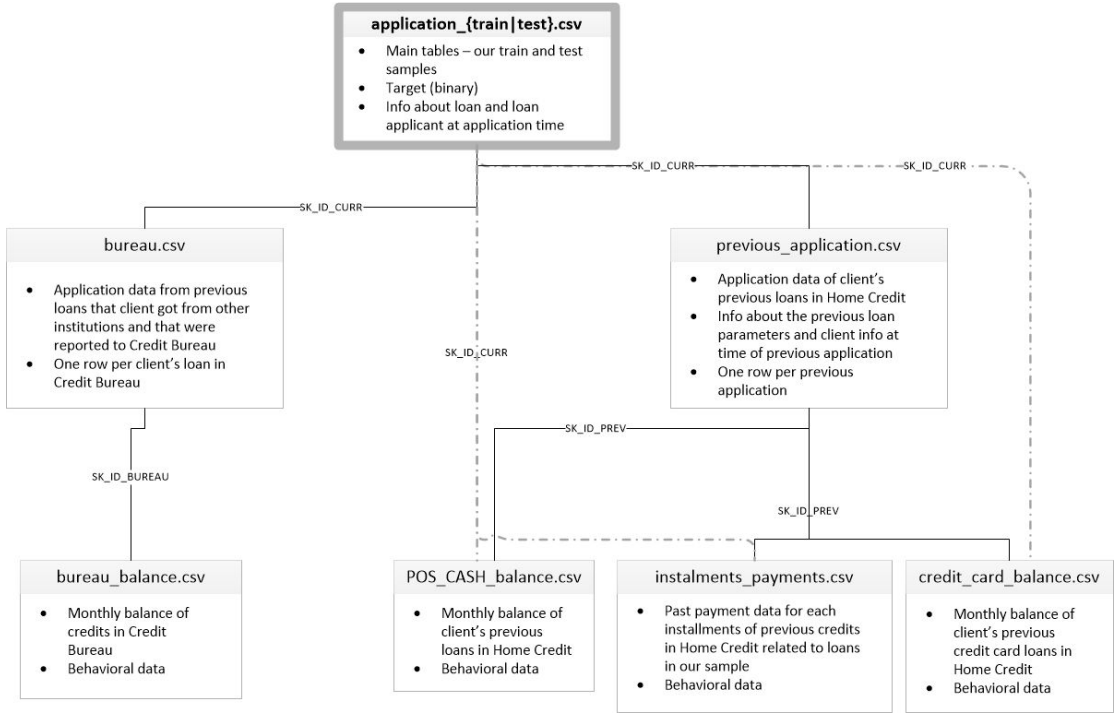
Dans le but d'optimiser son processus d'octroi de crédit, l'entreprise envisage la mise en place d'un outil de scoring crédit. Cet outil sera conçu pour évaluer la probabilité de remboursement d'un client et ainsi classifier les demandes en crédits accordés ou refusés.

Les données sources nécessaires sont disponibles en téléchargement sur Kaggle via le lien suivant.

Par ailleurs, les chargés de relation client ont signalé une tendance croissante chez les clients à réclamer davantage de transparence concernant les décisions d'octroi de crédit. Cette exigence de transparence s'aligne parfaitement avec les valeurs prônées par l'entreprise. Afin de répondre à cette demande, **Prêt à Dépenser** a pris la décision de développer un **dashboard** interactif. Celui-ci permettra aux chargés de relation client d'expliquer de manière transparente les décisions d'octroi de crédit. De plus, il offrira la possibilité aux clients d'accéder facilement à leurs informations personnelles et de les explorer en toute simplicité.

Les Données :

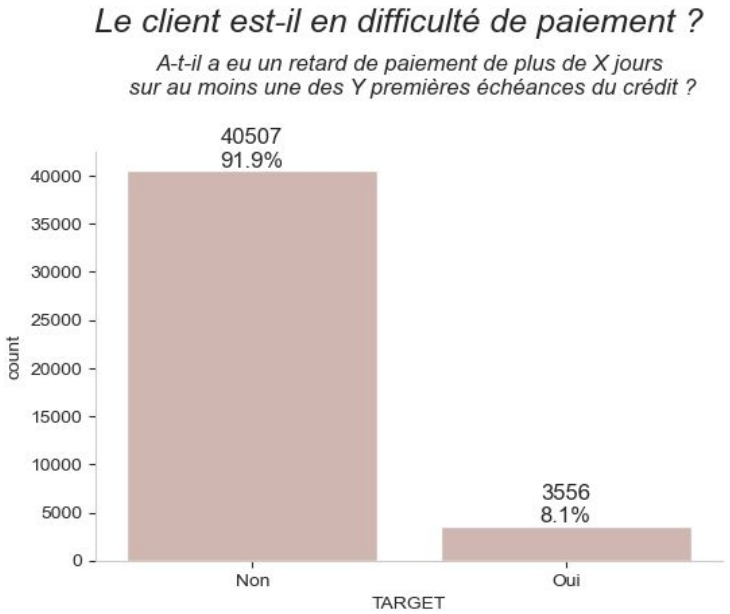
Sept fichiers contenant environ 200 variables au total.



Historique de crédit du client auprès d'autres institutions financières.

Historique de crédit du client chez "**Prêt à Dépenser**".

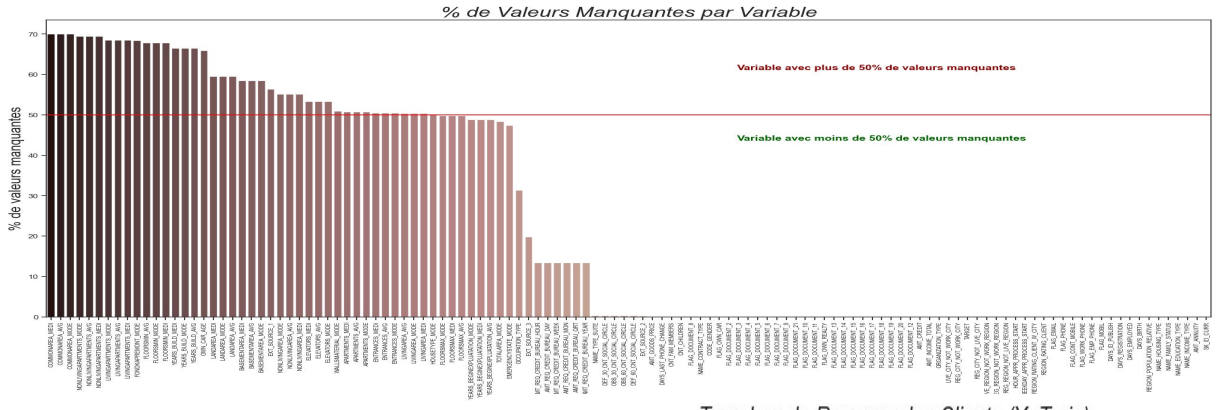
➢ L'application "train" rassemble 307 511 clients pour lesquels la décision d'octroi de crédit de « Prêt à Dépenser » est connue, avec la variable "Target".

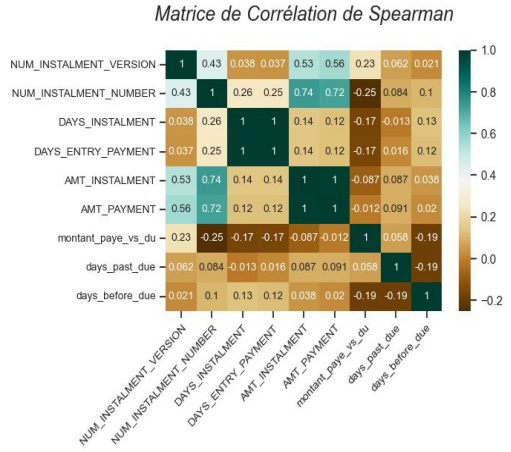


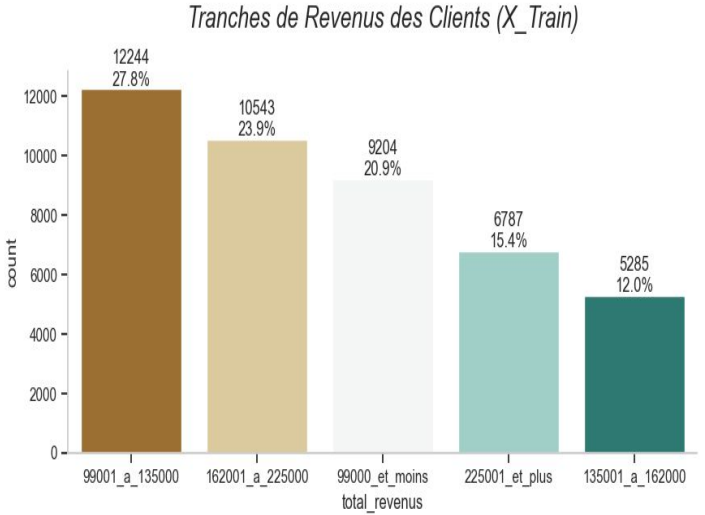
➢ L'application "test" rassemble 48 744 clients pour lesquels la décision n'est pas connue.

Exploration et Nettoyage des données:

- Élimination des variables comportant plus de 1 % de données manquantes.
- Substitution des valeurs manquantes par la médiane.
- Discrétisation des variables numériques pour atténuer les valeurs aberrantes.
- Regroupement des catégories avec des effectifs trop faibles pour les variables qualitatives.
- Création de nouvelles caractéristiques basées sur des ratios.
- Agrégation (somme, moyenne).
- Élimination des variables présentant une corrélation.
- Sélection des variables pertinentes.







Modélisation

Partie II

Choix Métrique d'évaluation et Modèle:

Problématique : ***Classification dans un contexte de données déséquilibrées.***

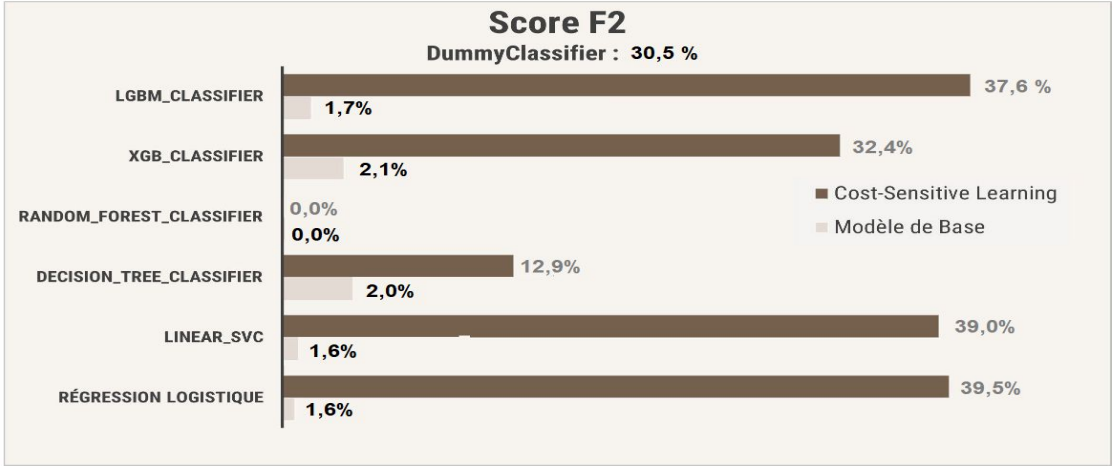
La complexité de la situation réside dans le fait que les faux négatifs ont des conséquences plus préjudiciables que les faux positifs. Les faux négatifs dans cet ensemble de données se produisent lorsqu'un client à risque est identifié à tort comme étant sans risque, obtenant ainsi un prêt qui pourrait ne pas être remboursé. En revanche, les faux positifs correspondent à des cas où **un client fiable est erronément considéré comme risqué,** conduisant la société de crédit à refuser à tort le prêt. Les faux négatifs entraînent donc des coûts plus importants : Coût(Faux Négatifs) > Coût(Faux Positifs).

La mention Train/Test Split : 70% /30%.
Chaque sous-échantillon contient le même mélange d'exemples par classe, c'est-à-dire environ 92% de classe 0 et 8% de classe 1.

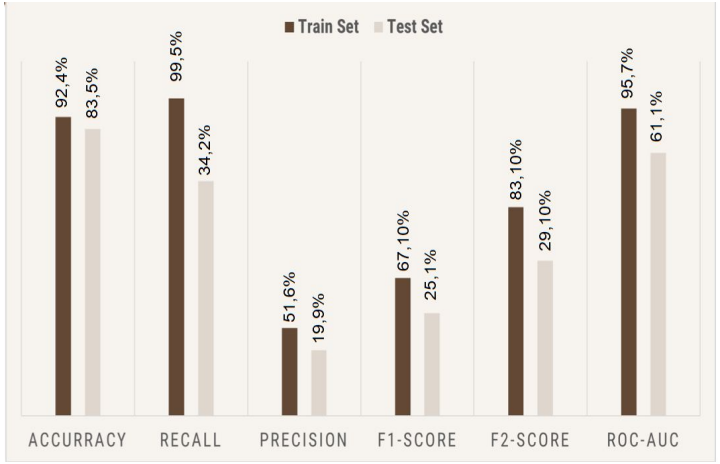
Évaluation des modèles candidats à l'aide d'une validation croisée stratifiée répétée k-fold.

F-mesure avec une valeur béta de 2 qui accorde plus d'attention au rappel qu'à la précision.

Cost Sensitive Learning : attribution des coûts basés sur la distribution inverse des classes.



Performance du modèle après optimisation des hyperparamètres :



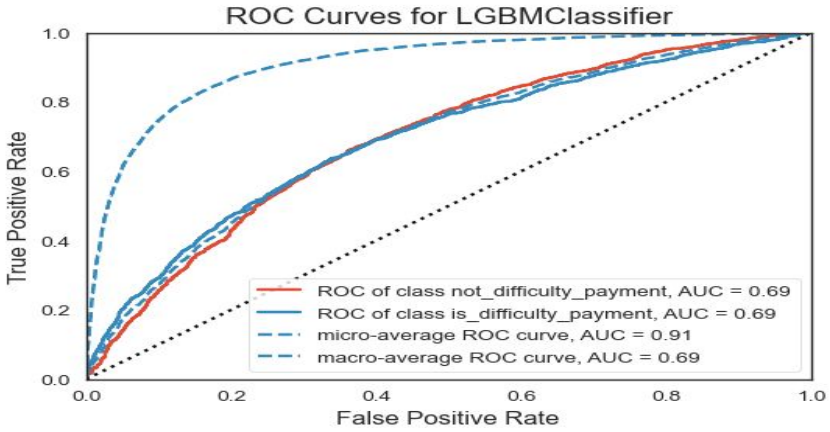
Le modèle optimal pour notre situation est le **classificateur LGBM.**

Le score F2 a connu une amélioration, passant de 37,6 % à 83,10 % sur notre ensemble d'entraînement, cependant, il ne s'élève qu'à 29,10 % sur l'ensemble de test.

	precision	recall	f1-score	support
0	0.94	0.88	0.91	13368
1	0.20	0.34	0.25	1173
accuracy			0.84	14541
macro avg	0.57	0.61	0.58	14541
weighted avg	0.88	0.84	0.85	14541

Taux d'erreurs de Classification : 25%

		Classes Prédites		
		is_difficulty_payment	not_difficulty_payment	Recall
Classes réelles	is_difficulty_payment	TP 402	FN 771	34,3%
	not_difficulty_payment	FP 1628	TN 11740	87,8%
	Precision	19,8%	93,8%	

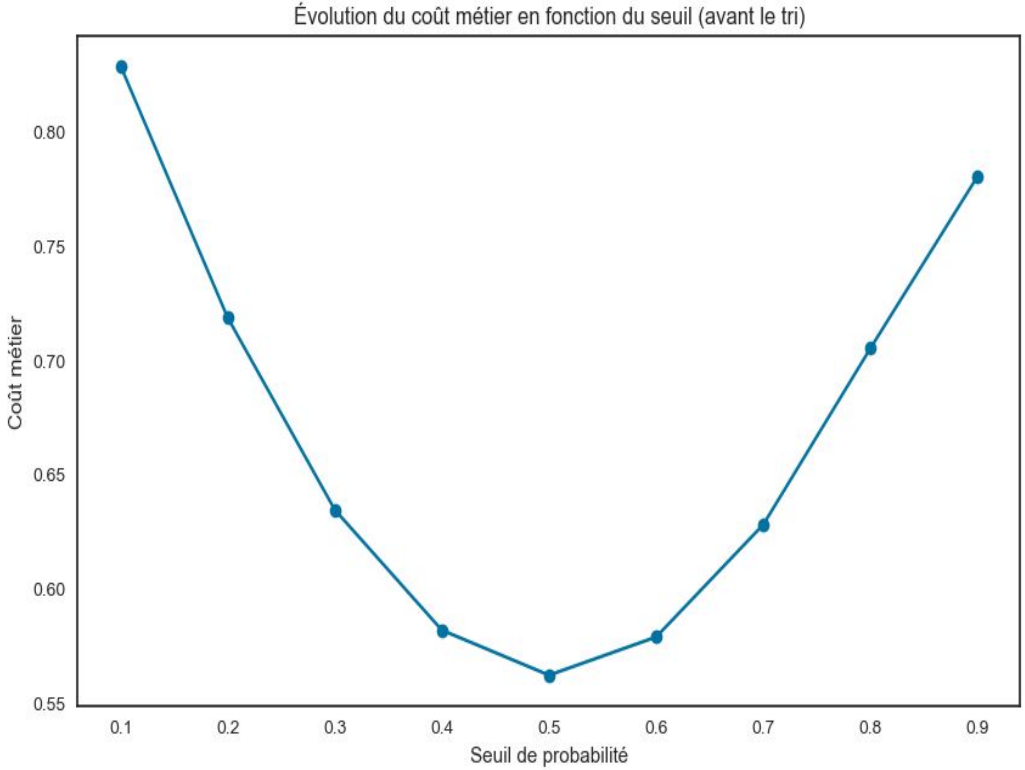


Coût Métier :

Définition du seuil dans notre problématique

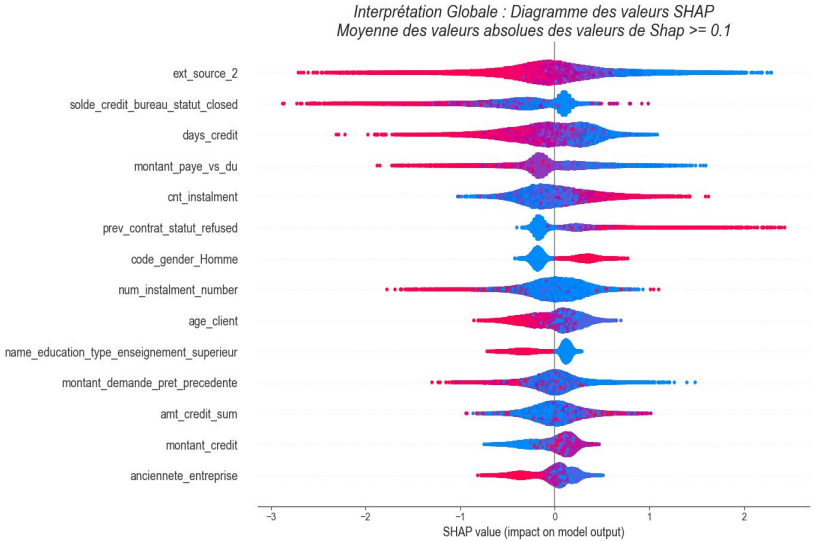
Un prêt non remboursé engendre un fardeau financier bien plus conséquent pour la banque que le bénéfice potentiel qu'elle aurait pu en tirer.

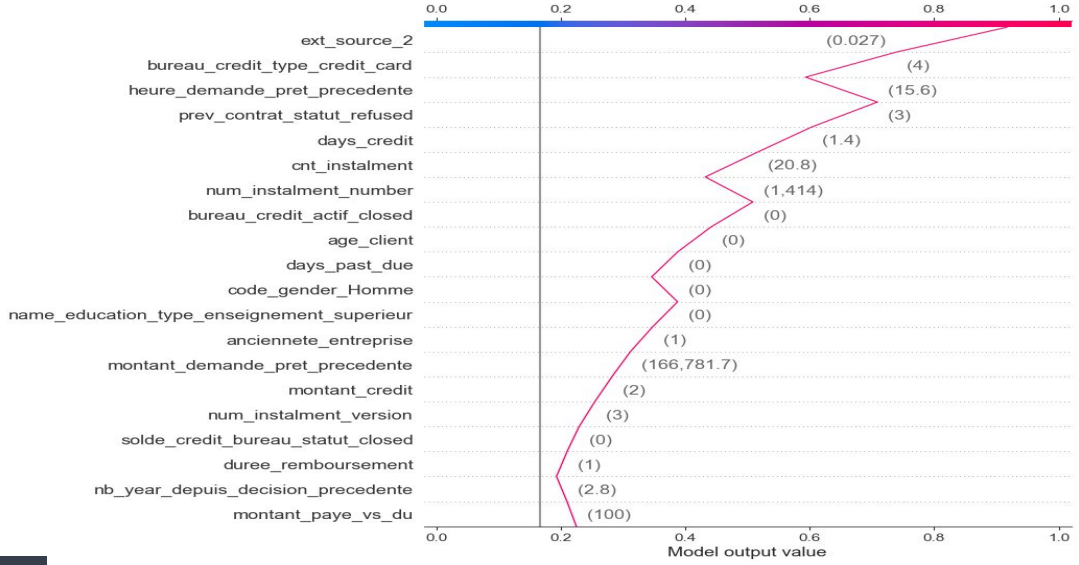
Le coût est optimal (minimal) pour la valeur de 0.5.
Pour chaque seuil, les faux négatifs (FN) et les faux positifs (FP) ont été calculés, fournissant ainsi le coût métier sur le jeu de validation. La valeur optimale du seuil de probabilité est celle qui minimise le coût métier.



Interprétation des résultats du modèle :

Les valeurs de Shapley mesurent l'importance d'une variable en évaluant la différence entre les prédictions d'un modèle avec et sans cette variable. Cependant, étant donné que l'ordre dans lequel un modèle rencontre les variables peut influencer ses prédictions, cette évaluation est effectuée pour tous les ordres possibles afin d'assurer une comparaison équitable des caractéristiques.

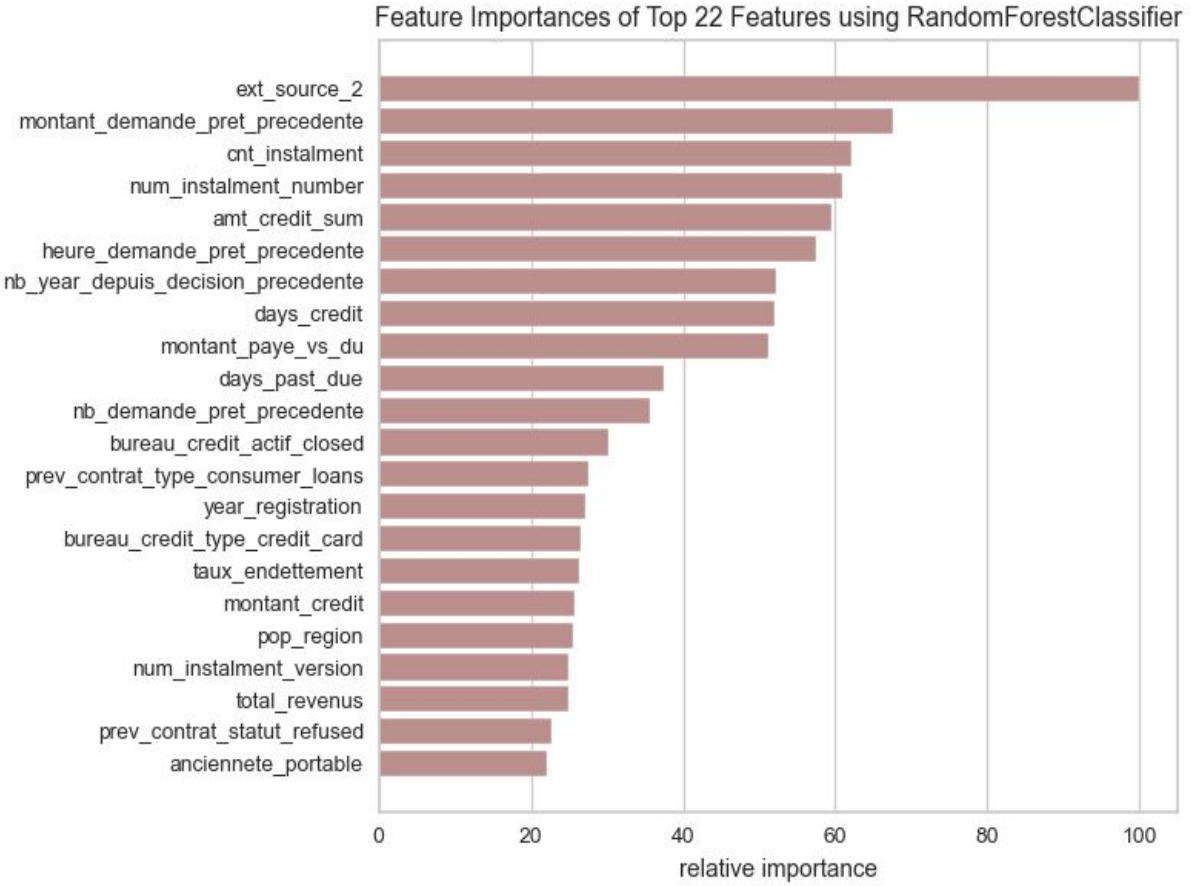




Ce graphique de décision SHAP fournit une compréhension visuelle de la contribution de chaque feature à la prédiction du modèle pour le client spécifique, en montrant comment chaque feature pousse la prédiction vers le haut (classe positive) ou vers le bas (classe négative) par rapport à la valeur d'espérance du modèle.

Features importances:

La visualisation des 22 importances features vise à quantifier l'influence de diverses variables sur la performance du modèle. Ces indications sont précieuses pour déterminer quelles variables exercent la plus grande influence dans le processus de prise de décision du modèle.



Dashboard

Le **Dashboard Home Credit Default Risk** offre une plateforme complète pour explorer, analyser et prédire le risque de défaut de paiement des clients. Il peut être utilisé pour prendre des décisions éclairées dans le domaine du crédit et de la gestion financière.

Ce **dashboard**, développé avec des technologies telles que **Streamlit, Plotly**, et **Shap**, offre une expérience utilisateur conviviale et interactive. Il peut être utilisé comme outil de référence pour les professionnels du crédit et les analystes financiers dans leurs activités quotidiennes.

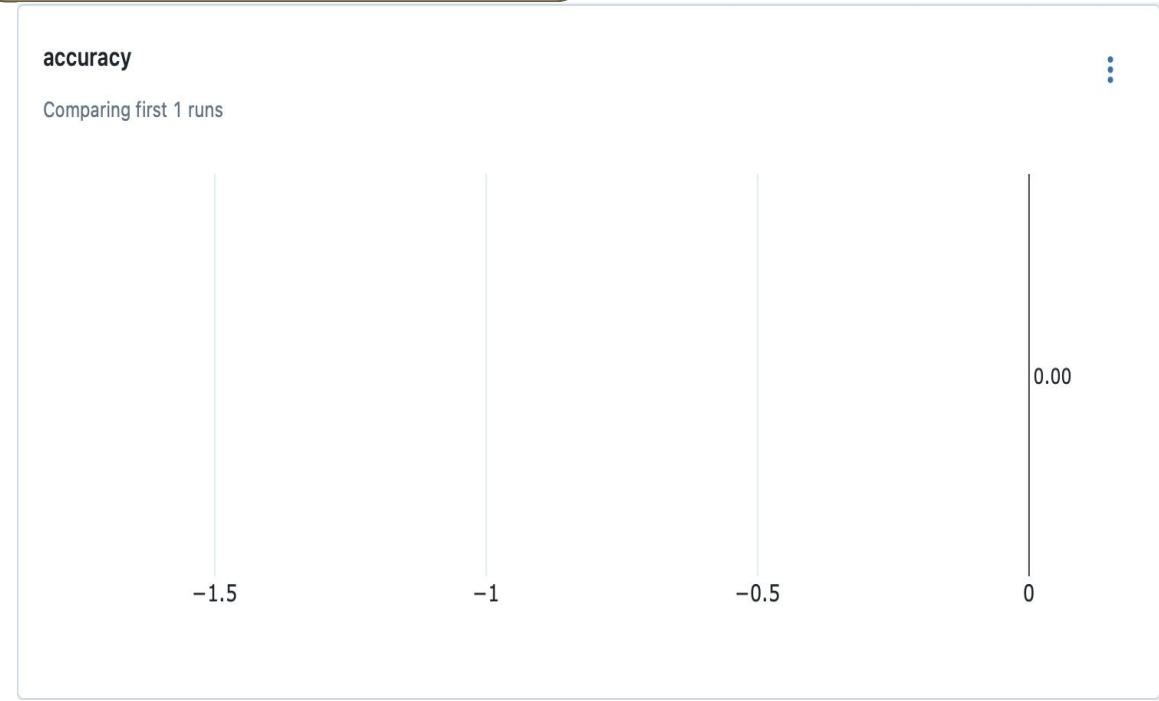
https://dashboard.heroku.com/apps/app-streamlit

https://app-streamlit-efb0ba034283.herokuapp.com

Mlflow

L'intégration de **MLflow** dans notre processus de développement de modèles machine learning a considérablement amélioré la gestion et la traçabilité de nos expériences. **MLflow** offre une solution complète pour suivre et organiser chaque étape du cycle de vie des modèles, de l'expérimentation initiale à la mise en production.

Nous avons démontré comment **MLflow** facilite le suivi des hyperparamètres, la sauvegarde des métriques de performance telles que accuracy, et l'enregistrement transparent des modèles avec toutes les informations nécessaires. La signature automatique du modèle et la possibilité de le charger ultérieurement ont simplifié la gestion des versions des modèles.



localhost:8085/

Data Drift

L'analyse complète effectuée de **Data Drift** entre les données de référence et actuelles, en utilisant diverses métriques et tests. Il offre une compréhension approfondie de la stabilité des données, met en évidence les changements potentiels, et évalue la qualité des données. Pour générer un rapport de présentation, nous pouvons agréger les résultats significatifs, fournir des explications claires sur les divergences détectées, et proposer des recommandations pour maintenir la qualité des données.

Visualisation de la page HTML

Partie III

Limites et Améliorations Possibles

Partie IV

Limites et Améliorations possibles :

AMÉLIORATIONS SUGGÉRÉES :

- **Correction des coquilles :**
  - "pré-process" devrait être "**prétraitement**".
  - "modélisation" peut être reformulé comme "**modèles prédictifs**" pour plus de clarté.
- **Réécriture :**
  - Comprendre le domaine bancaire est essentiel; il est recommandé de vérifier la cohérence du processus de prétraitement des données.
  - Affiner avec précision les métriques d'évaluation et les fonctions de coût en collaboration avec les équipes métier.
  - Améliorer les performances des modèles prédictifs.
  - Proposer le développement d'un dashboard avec deux pages distinctes, l'une dédiée à la "**banque**" et l'autre au "**client**". Cela permettrait au personnel de "**Prêt à Dépenser**" expliquant la décision à un client d'accéder à des données pertinentes sans nécessairement les partager intégralement avec le client.
  - Introduire une fonctionnalité interactive dans laquelle le client peut visualiser comment une modification de certaines variables aurait pu influencer la décision de crédit. Une page dédiée, telle que "**scénario**", pourrait être envisagée, où le client peut ajuster une ou plusieurs valeurs de son profil pour voir l'impact sur la réponse de la banque.

Merci

Parcours Data Scientist - OpenClassrooms

https://github.com/MauyvesNk