



Parcours Data Scientist - OpenClassrooms

Implémenter un modèle de scoring

Projet P7 – Isabelle Contant – 14/12/2022




CONTENTS

01 *Contexte, Objectifs et set de données*

02 *Modélisation*

03 *Dashboard*

04 *Limites et Améliorations Possibles*



Contexte, Objectifs et Set de données

PART.01

Contexte et 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.

Les données originales sont téléchargeables sur Kaggle à cette [adresse](#).

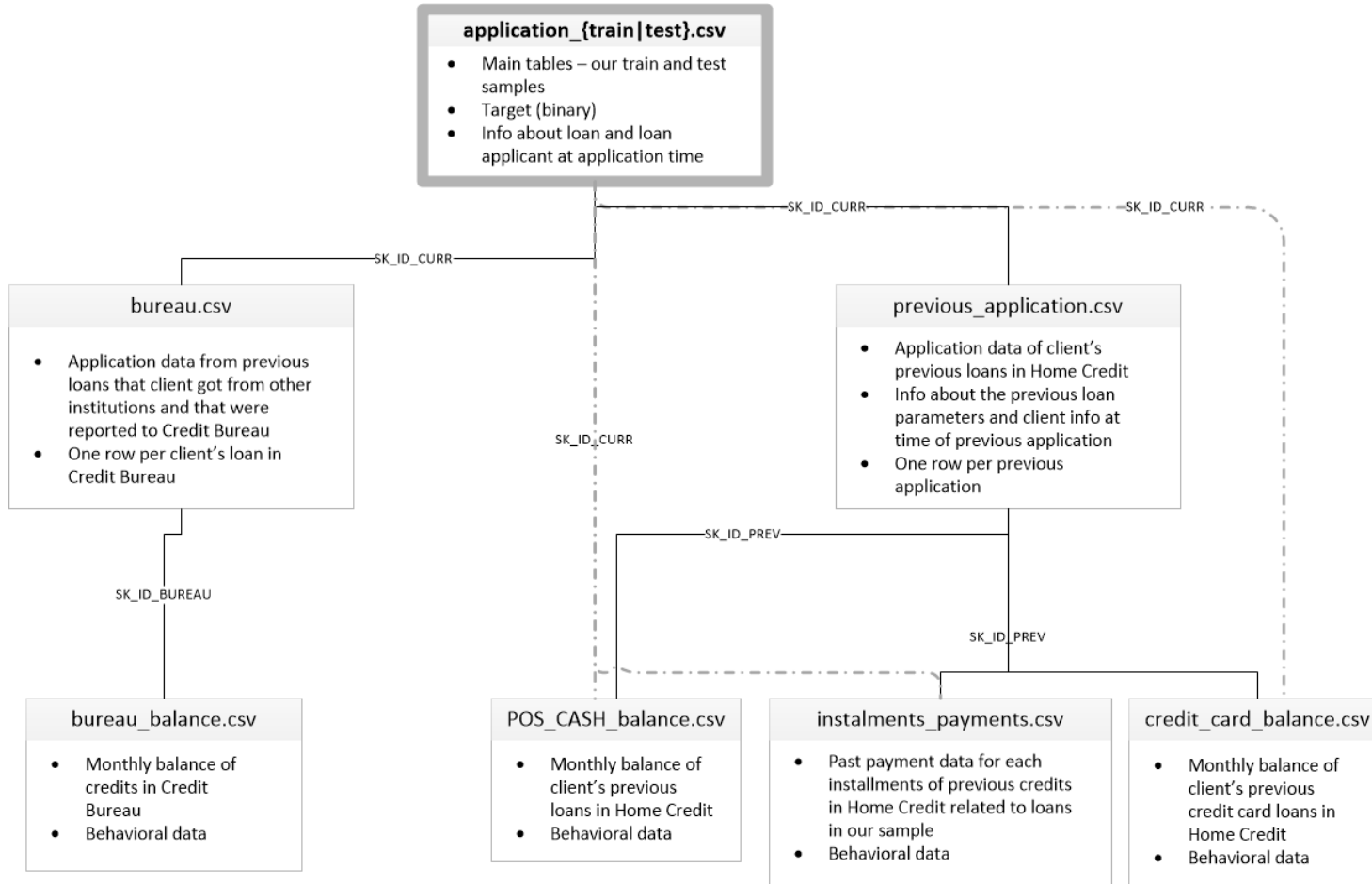
De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de transparence vis-à-vis des décisions d'octroi de crédit. Cette demande de transparence des clients va tout à fait dans le sens des valeurs que l'entreprise veut incarner. *Prêt à dépenser* décide donc de développer un dashboard interactif pour 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.



Prêt à dépenser

Les Données

7 fichiers pour environ 200 variables



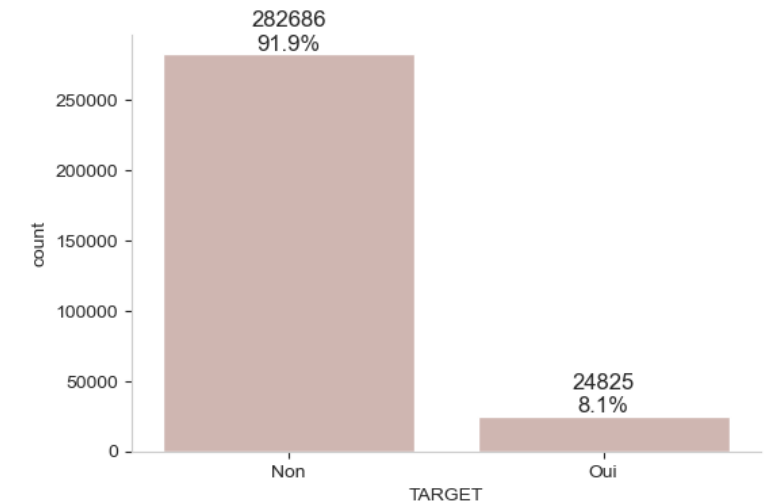
Historique de prêt du client dans d'autres institutions financières

Historique de prêt du client chez "Prêt à Dépenser"

- Application "train" regroupe 307 511 clients dont on connaît la décision de « Prêt à Dépenser » sur l'octroi du crédit (variable "Target")

Le client est-il en difficulté de paiement ?

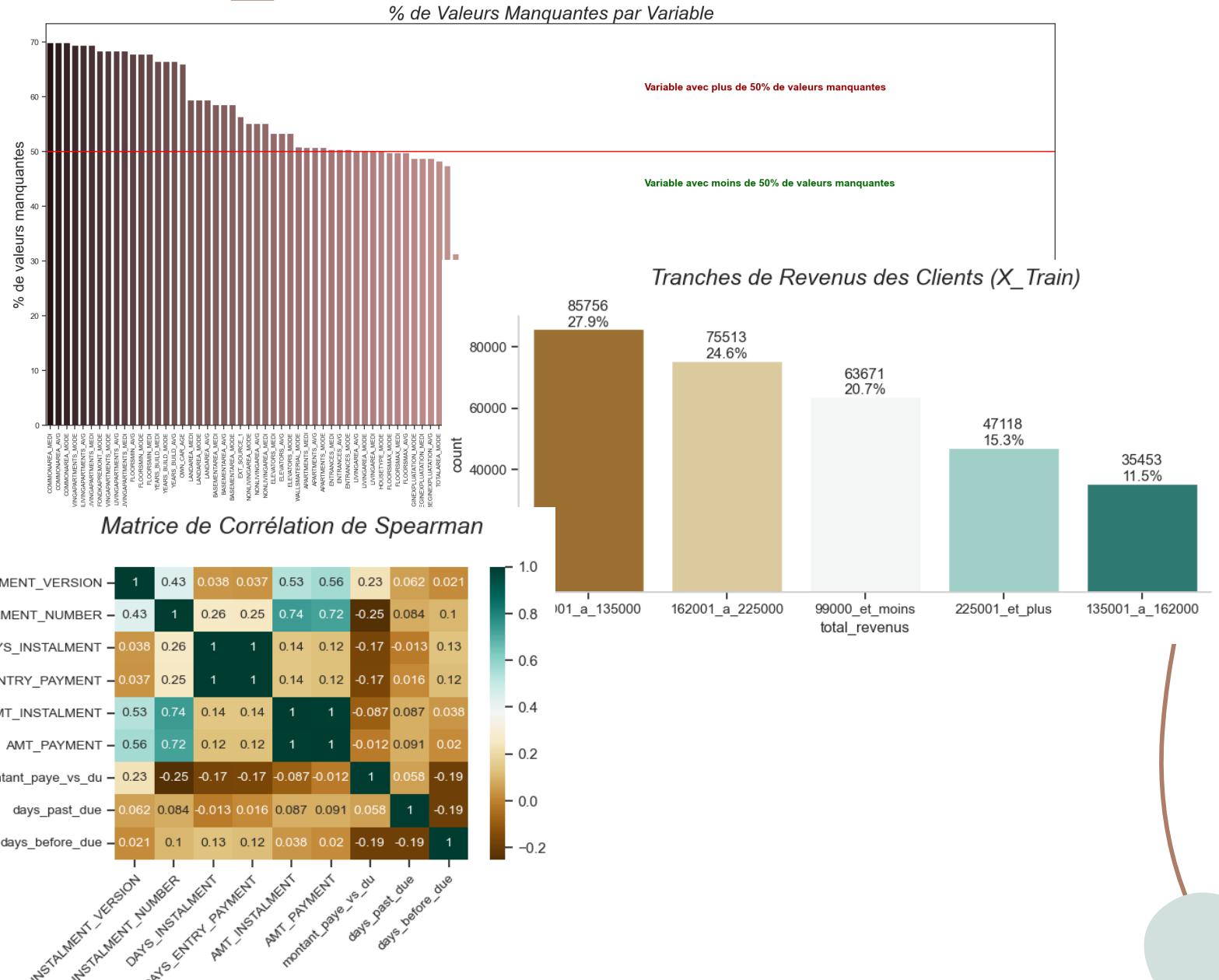
A-t-il eu un retard de paiement de plus de X jours sur au moins une des Y premières échéances du crédit ?



- Application "test" regroupe 48 744 clients dont on ne connaît pas cette décision.

Exploration et Nettoyage des données

- Suppression des variables avec plus de 1% de données manquantes
- Remplacement des valeurs manquantes par la médiane
- Discretisation des variables numériques pour neutraliser les outliers
- Regroupement des modalités à effectifs trop faibles des variables qualitatives
- Création de features (ratio)
- Agrégation (somme, moyenne)
- Suppression des variables corrélées
- Sélection des variables « pertinentes »



Modélisation

PART.02

Choix Métrique d'évaluation et Modèle

Problématique : Classification sur des données déséquilibrées

La difficulté ici réside dans le fait que les faux négatifs sont plus dommageables que les faux positifs. Les faux négatifs sur cet ensemble de données sont des cas où un **mauvais client est marqué comme un bon client et se voit accorder un prêt** alors que les faux positifs sont des cas où un bon client est marqué comme un mauvais client et la société de crédit ne lui accorde pas de prêt. Les faux négatifs lui coûtent donc plus cher : $\text{Cost}(\text{FalseNegatives}) > \text{Cost}(\text{FalsePositives})$

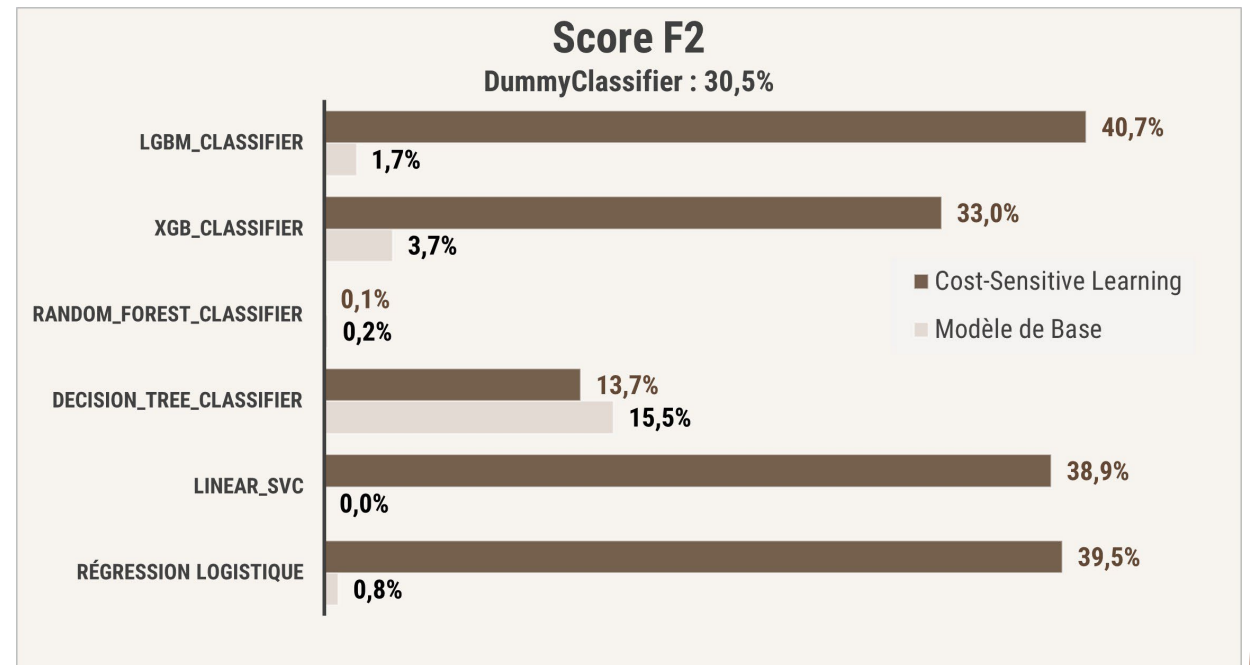
Train/Test Split : 70% / 30%.

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.

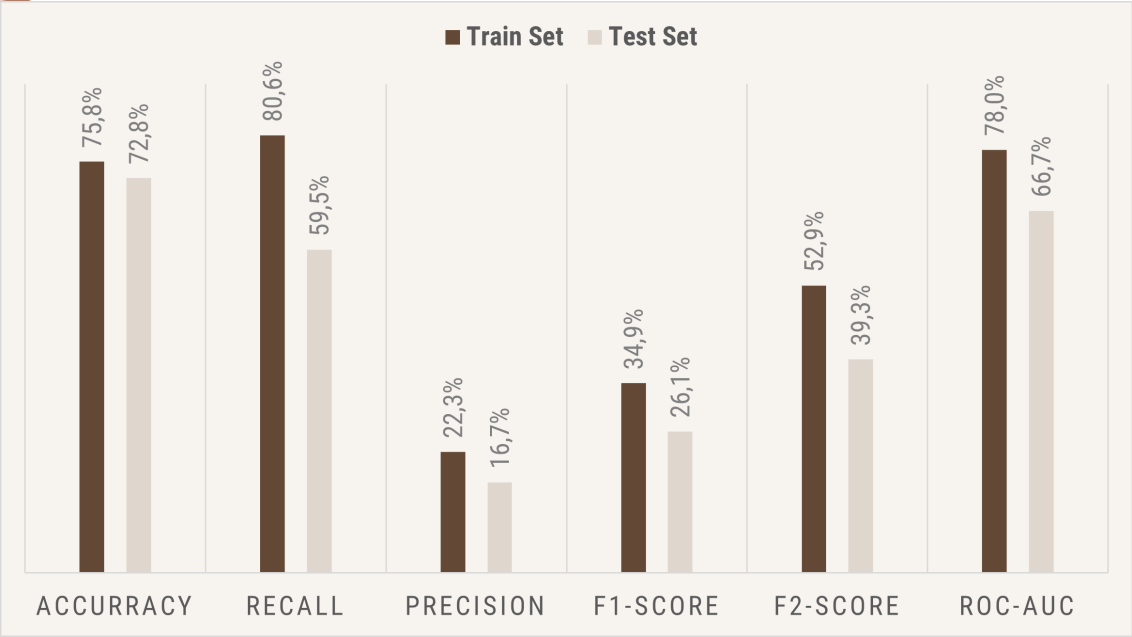
Évaluation des modèles candidats à l'aide d'une validation croisée stratifiée répétée k-fold.

F-mesure avec une valeur bêta 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.



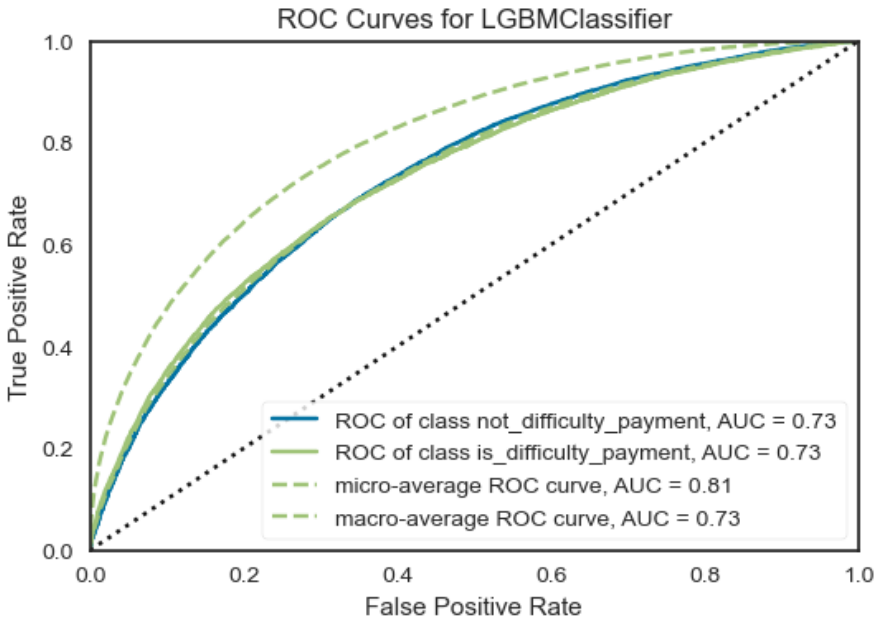
Performances du modèle après optimisation des hyperparamètres



Le F2-score a augmenté de 40,7% à 52,9% sur notre jeu d'entraînement, mais n'est que de 39% sur le set de test.

Taux d'erreurs de classification : 26%

		Classes Prédites		Recall
		is_difficulty_payment	not_difficulty_payment	
Classes réelles	is_difficulty_payment	TP 4 875	FN 3 317	59,5%
	not_difficulty_payment	FP 24 334	TN 68 953	73,9%
Precision		16,7%	95,4%	

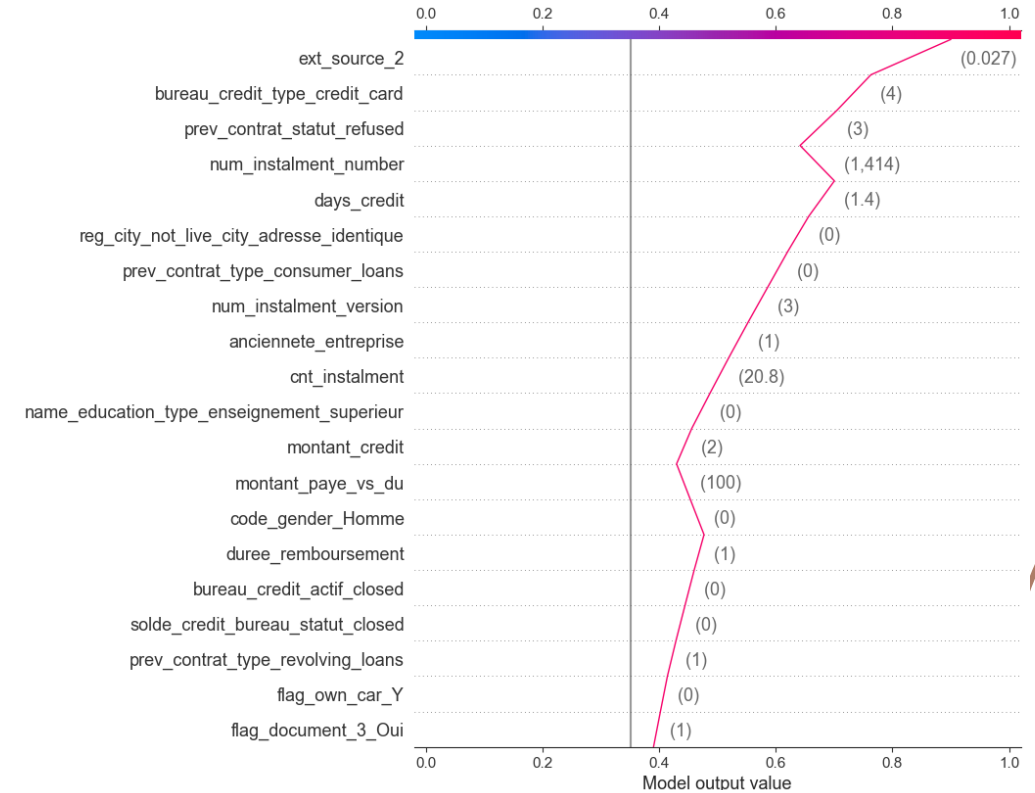
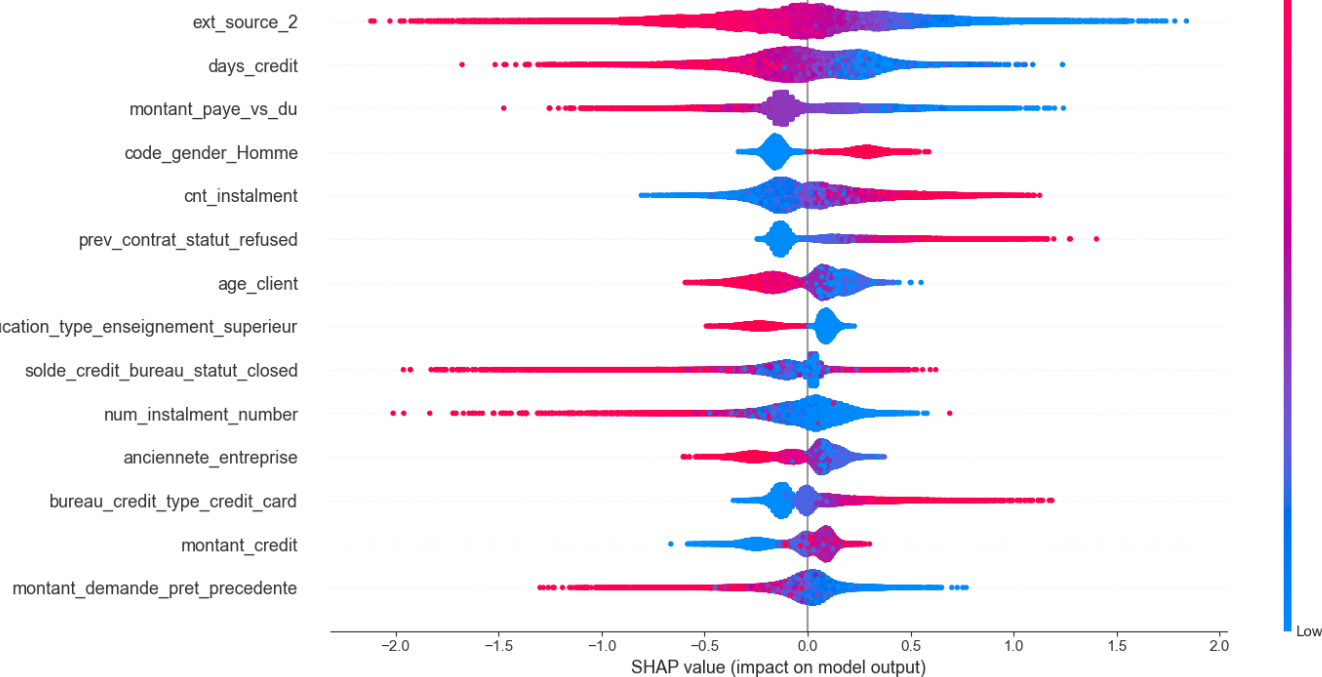


	precision	recall	f1-score	support
0	0.95	0.74	0.83	93287
1	0.17	0.60	0.26	8192
accuracy			0.73	101479
macro avg	0.56	0.67	0.55	101479
weighted avg	0.89	0.73	0.79	101479

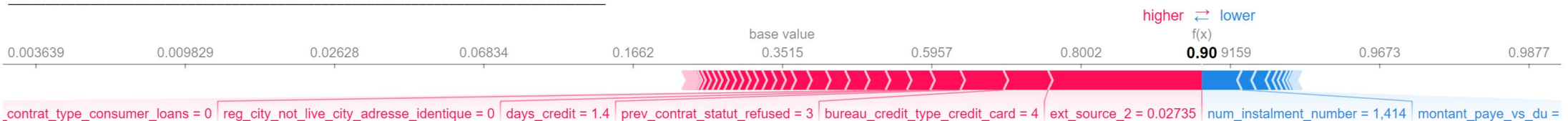
Interprétation des résultats du modèle

Les valeurs de Shapley 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.

Interprétation Globale : Diagramme des valeurs SHAP
Moyenne des valeurs absolues des valeurs de Shap ≥ 0.1



Client numero : 106854
Model Prediction : Classe 1
Il y a 90.0% de risques que le client ait des difficultés de paiement



Dashboard

PART.03

<https://isabellecontant-p7-dashboard-streamlit-01--homepage-ubegdo.streamlit.app/>

Limites et Améliorations possibles

PART.04

Limites et Améliorations possibles

- **Méconnaissance du milieu bancaire** → vérifier la cohérence du pré-process
- Définir plus finement la **métrique d'évaluation** et la **fonction de coût** en collaboration avec les équipes métier
- Améliorer les performances de la modélisation

Il serait intéressant de développer un dashboard avec une page « banque » et une page « client ». Cela permettrait à la personne de « *Prêt à dépenser* » qui explique la décision à un client d'avoir accès à certaines données permettant d'expliquer la réponse, sans nécessairement pouvoir les montrer au client. Par ailleurs, il serait intéressant de rajouter une partie interactive qui permettrait au client de voir quelle valeur sur quelle variable aurait pu lui permettre d'obtenir son crédit. En ce sens, on pourrait envisager une page « scenario » où l'on pourrait changer une ou plusieurs valeurs du profil du client et voir l'impact sur la réponse de la banque.



Merci