

Outil de scoring credit

Évaluation des clients à risque

Mission

01	Cadre	<ul style="list-style-type: none">• Service de prêts d'une institution financière
02	Objectif	<ul style="list-style-type: none">• Identifier les clients à risque• Sans pénaliser les clients sans risque• Justifier le résultat.
03	Ressources	<ul style="list-style-type: none">• Historique de prêts;• Historique d'informations financières;• Comportement des emprunteurs.

Processus

Description du jeu de données

Nettoyage du jeu de données

Évaluation des modèles

Interprétabilité du modèle

Conclusion

Description du jeu de données

Historique des emprunteurs : application_train.csv & application_test.csv

307 507
applications

122
variables

1 Target

16
catégories

106
numériques

(307 507, 122)

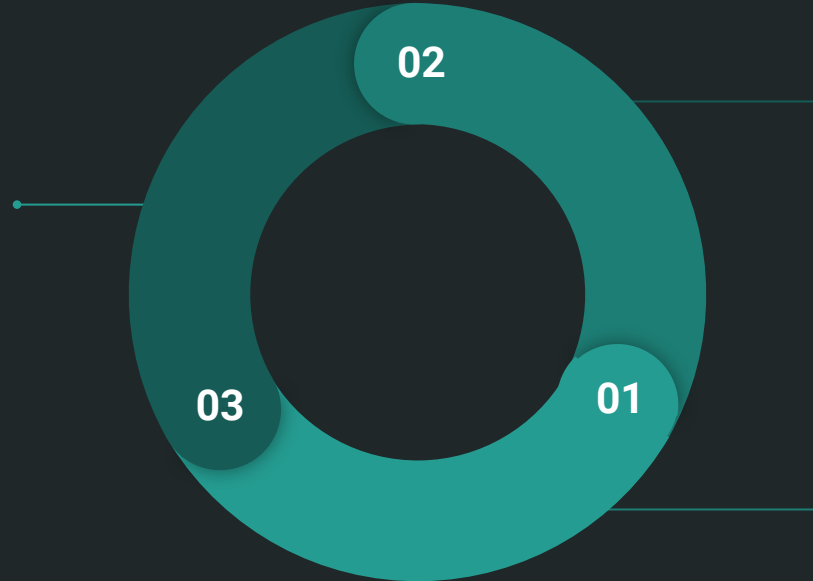
EDA

Features quantitatives

Valeurs aberrantes
Flags
Comptabilité du temps
Corrélation avec la target

Première sélection de features:

- Peu corrélées avec la target, plus de 50% de valeurs manquantes : 6
- Features très corrélées entre elles : 33



Features catégorielles

- Choix encodage
 - Ordinal Encoder
 - One Hot Encoder

Valeurs manquantes

41 features > 50%

EDA

corrélation avec la target $> 0,01$
valeurs manquantes $> 50\%$

6

corrélation $> 0,9$

30

(307 507, 87)

Échantillonnage

JEU DE BASE : (307 507, 87)

JEU RÉDUIT : (92 252, 87)

30%

JEU D'ENTRAÎNEMENT : (64 576, 85)

70%

JEU DE TEST : (64 576, 85)



Preprocessing

Quantitatives

Iterative Imputer

Qualitatives
bimodales

Simple Imputer (strategie : "most_frequent")

Ordinal Encoder

Qualitatives
multimodales

One Hot Encoder

MinMaxScaler

Évaluation du modèle : Classification des clients

0 = Bon client

1 = Mauvais client

		Réalité	
		Négatif : 0	Positif : 1
Prédictions	Négatif : 0		
	Positif : 1		

Recall

$$\frac{TP}{TP + \text{FN}} \longrightarrow 1$$

Precision

$$\frac{TP}{TP + FP}$$

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

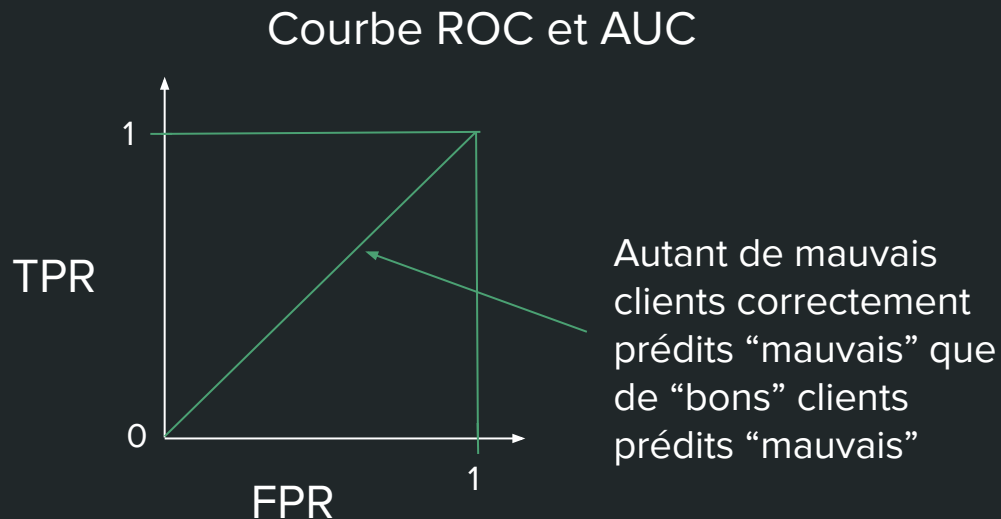
Évaluation du modèle : Autres métriques

Custom Recall

$$\frac{TP}{TP + 10 \times FN}$$

F1-score

$$\frac{Precision \times Recall}{Precision + Recall}$$



Méthodologie

Tests de models

Comparer
plusieurs modèles
entre eux

Resampling strategies

RandomUnderSampling
RandomOverSampling
SMOTE
NearMiss

Features engineering

Créer de nouvelles
features pour enrichir
le modèle

Hypertuning

Rechercher les
meilleurs
hyperparamètres du
modèle choisi

Test sur différents
seuils de décision.

Explicabilité

Comprendre les
variables les plus
influençantes

Évaluation du modèle : Choix de modèles

Classifieur Naïf

Simple
Classement aléatoire

Régression Logistique

Linéaire

Random Forest Classifieur

Arbre ensembliste

XG Boost

Gradient boosting

Light GBM






Gradient boosting

Évaluation du modèle : Sans strategies

Model	Stratégie	Accuracy	Recall	F2	ROC_AUC	Métier
Classifieur Naïf	“Stratified ”	0,85	0,08	0.08	0,5	0,008
Régression Logistique	“Balance d”	0,69	0,67	0.41	0,74	0,17
Random Forest Classifier	“Balance d”	0,92	0,0006	0.002	0,71	0,00006
XGBoost	Ratio	0,80	0,41	0.33	0,70	0,07
LGBM	Ratio	0,74	0,60	0.41	0.74	0.13

Entraînement : (64 576, 85)

Évaluation du modèle : Resampling strategies

Model	Resampling	Accuracy	Recall	F2	ROC_AUC	Métier
Régression Logistique	Random Undersampling	0,68	0,68	0.41	0,74	0,17
Random Forest Classifier	Random Oversampling	0,68 	0,65 	0.39	0,72	0,15 
XGBoost	RUS + ratio	0,76 	0,46 	0.35	0,69	0,08
LGBM	SMOTE + ratio	0,65	0,69	0.40	0,73	0,18

Entraînement : (64 576, 85)

Évaluation du modèle : Features engineering

Charge du crédit par rapport au revenu du demandeur : $\text{CREDIT} / \text{INCOME}$

Charge de paiement périodique par rapport au revenu du demandeur : $\text{ANNUITY} / \text{INCOME}$

Durée du crédit en termes de nombre de paiements d'annuité : $\text{CREDIT} / \text{ANNUITY}$

Proportion de la vie du demandeur passée en emploi : $\text{DAYS_EMPLOYED} / \text{DAYS_BIRTH}$

(64 576, 89)

Évaluation du modèle : Features engineering

Model	Resampling	Accuracy	Recall	F2	ROC_AUC	Métier
Régression Logistique	Random Undersampling	0,68	0,67	0.40	0,74	0,17
Random Forest Classifier	Random Undersampling	0,69	0,65	0.40	0,73	0,16
XGBoost	ratio	0,81	0,41	0.33	0,71	0,01
LGBM	SMOTE + ratio	0,66	0,69	0.40	0,73	0,18
LGBM	ratio	0,75	0,59	0.41	0,74	0,13

Entraînement : (64 576, 89)

Évaluation du modèle : Top features

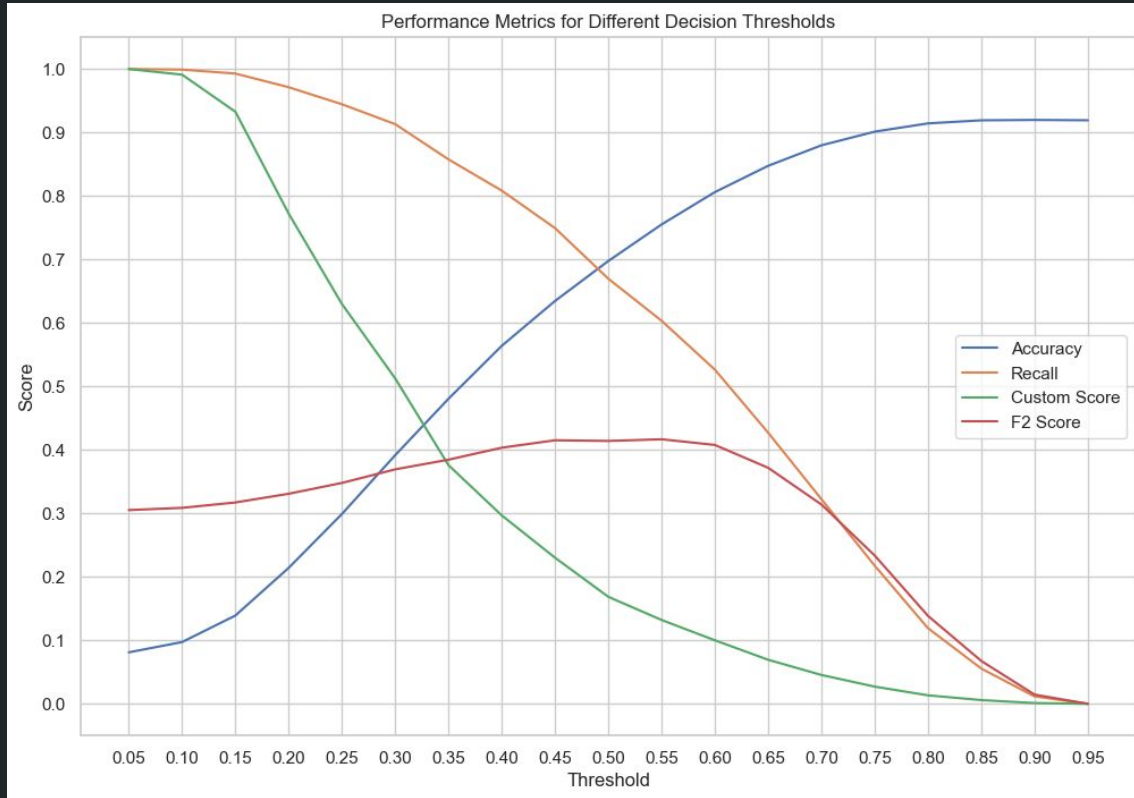
Model	Resampling	Accuracy	Recall	F2	ROC_AUC	Métier
Régression Logistique	Random Undersampling	0,68	0,67	0.40	0,74	0,17
Random Forest Classifier	Random Undersampling	0,68	0,65	0.39	0,72	0,15
XGBoost	ratio	0,80	0,42	0.33	0,70	0,06
LGBM	SMOTE + ratio	0,42	0,85	0.36	0,69	0,33
LGBM	ratio	0,74	0,60	0.41	0,74	0,13

Entraînement : (64 576, 25)

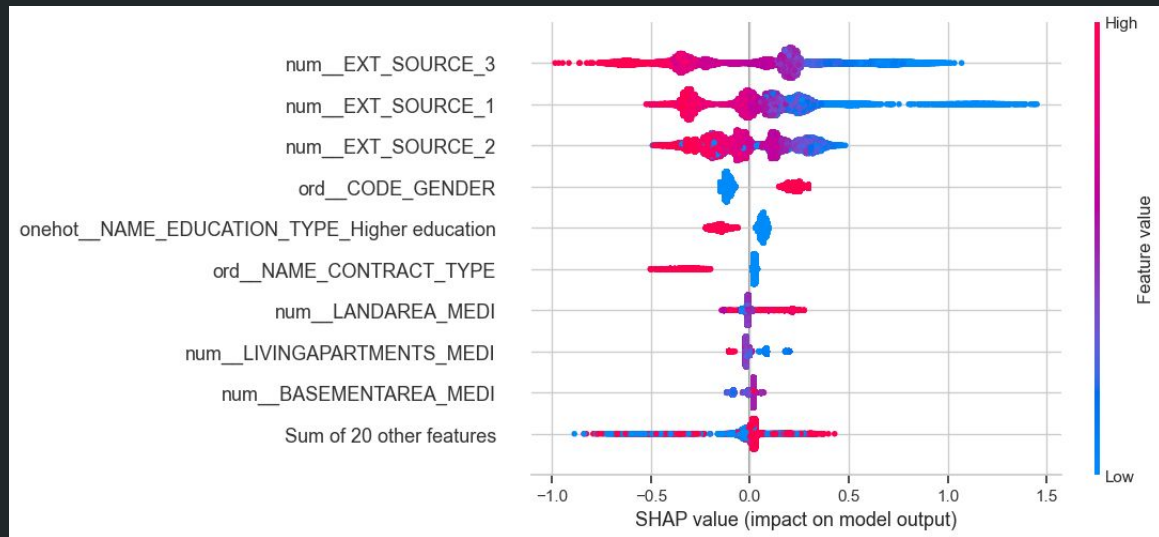
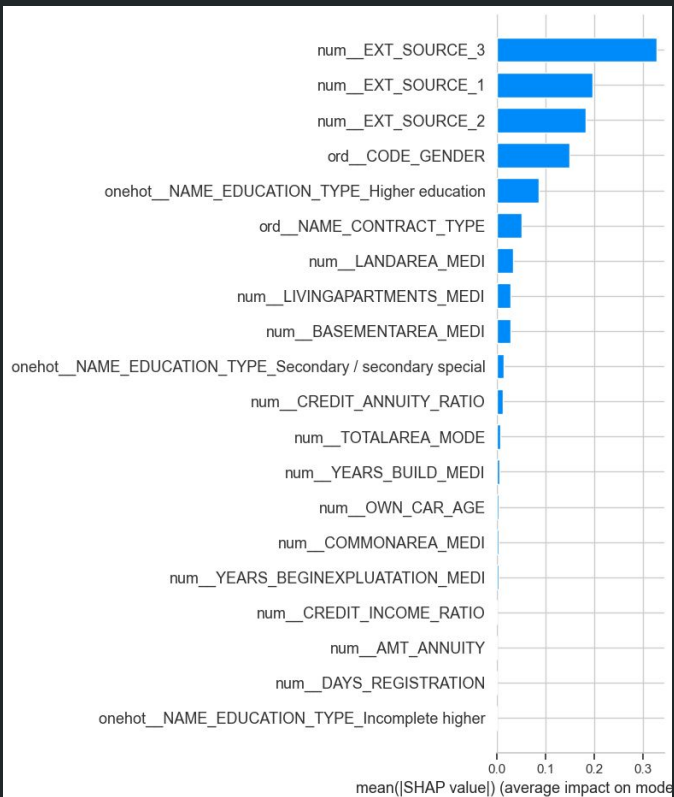
Évaluation du modèle : Hypertuning

Model	Resampling	Accuracy	Recall	F2	ROC_AUC	Métier
XGBoost	Ratio	0,70	0,66	0.41	0,75	0,16

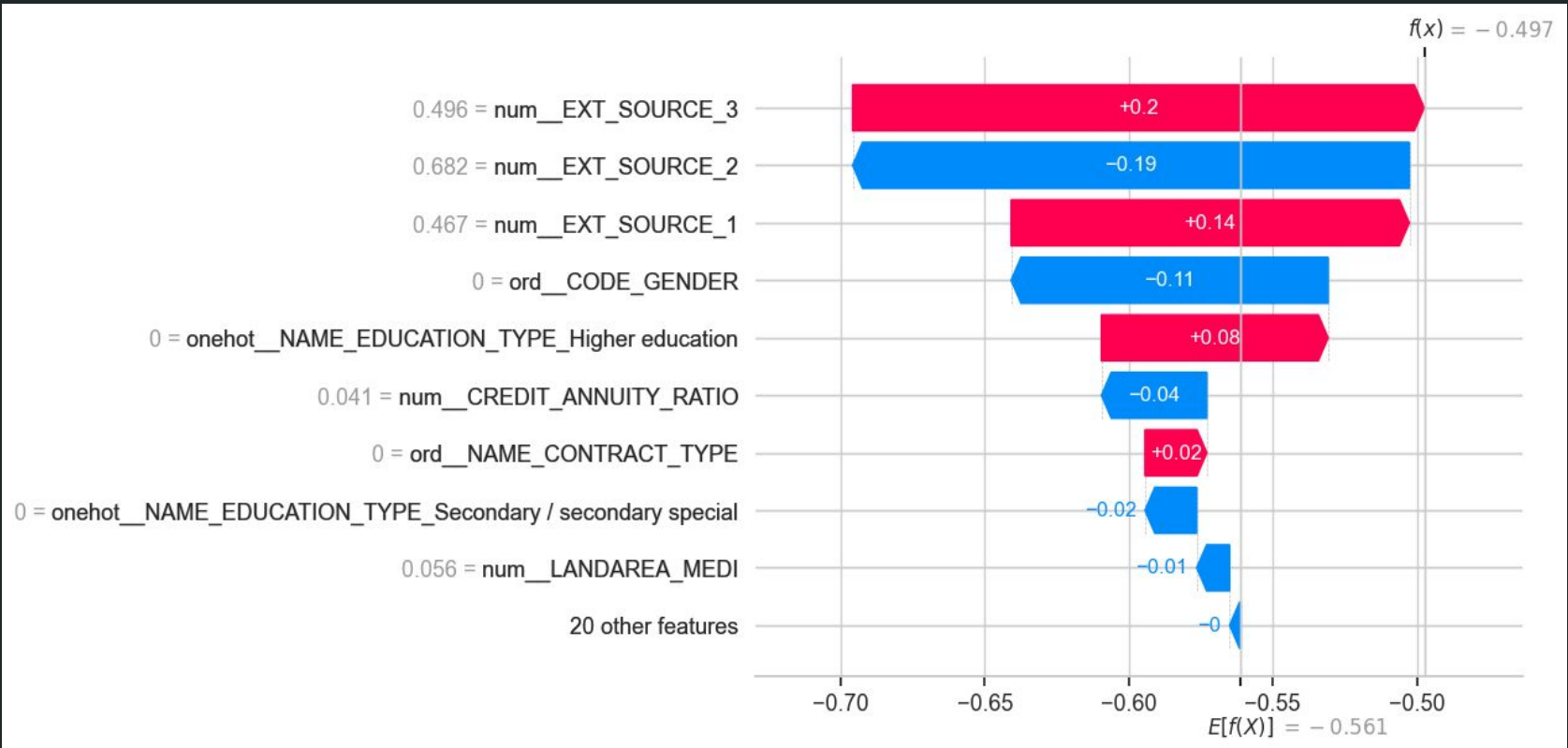
Évaluation du modèle : Seuil de décision



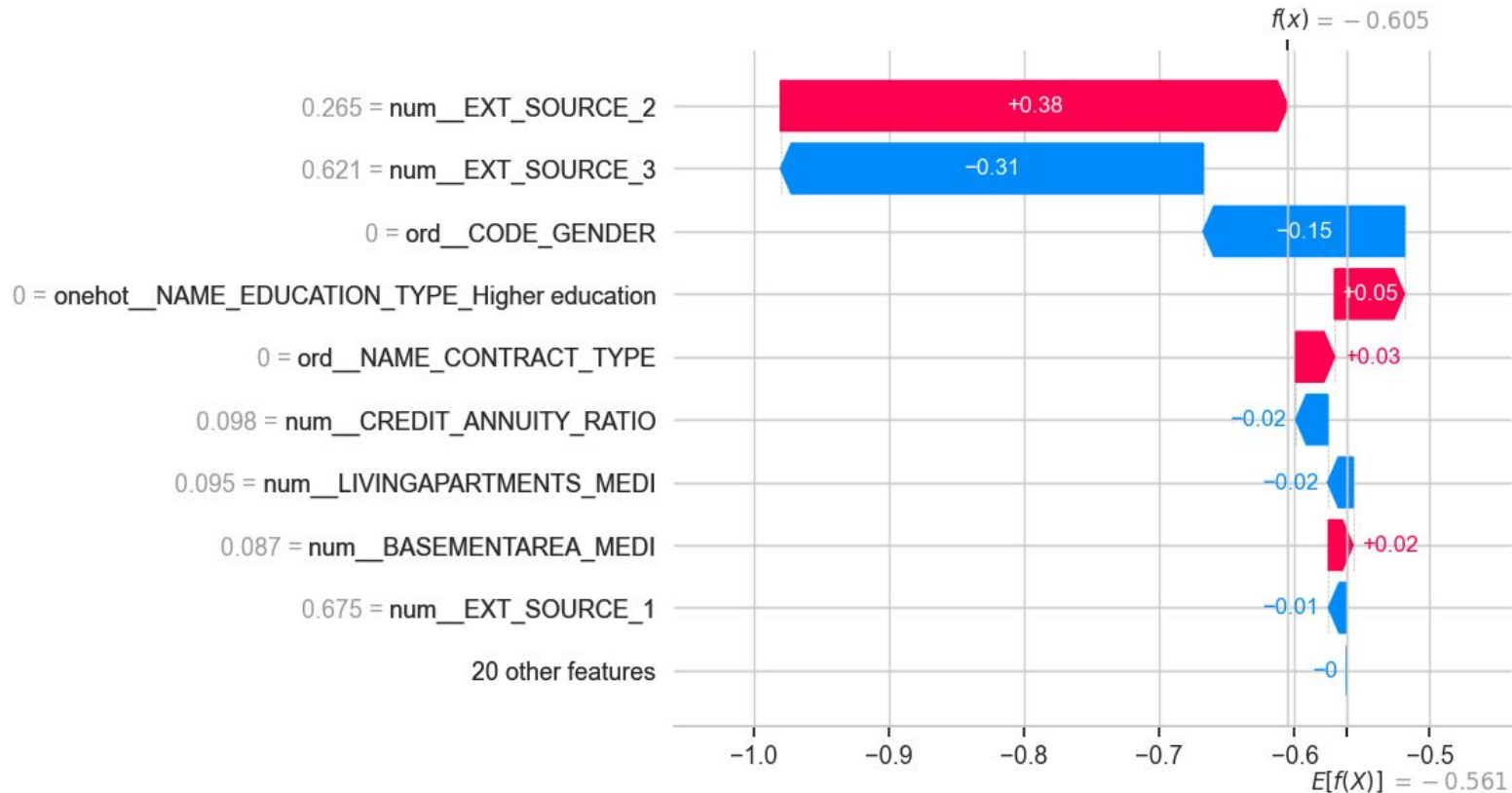
Explicabilité du modèle : Variables globales



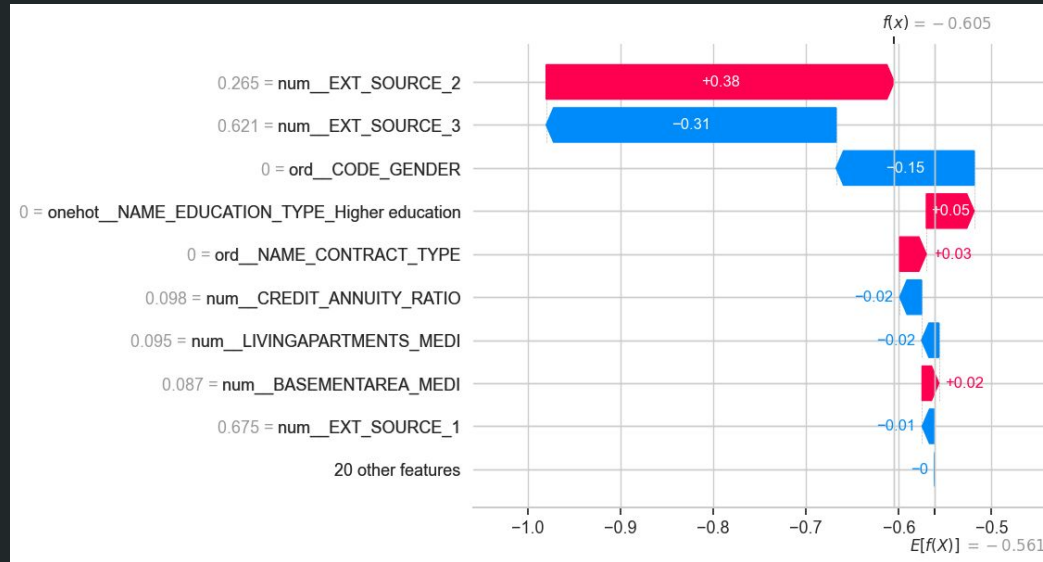
Explicabilité du modèle : Variables locales



Explicabilité du modèle : Variables locales



Explicabilité du modèle : Variables locales



Pistes d'amélioration

Travail sur les différents fichiers disponibles.

Valider la sélection de features avec l'équipe métier.

Travail sur l'imputation des valeurs manquantes.