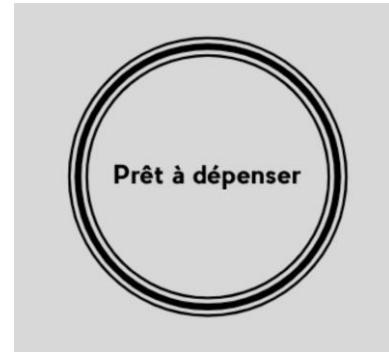


Construire un modèle de scoring

Prêt à dépenser





Sommaire

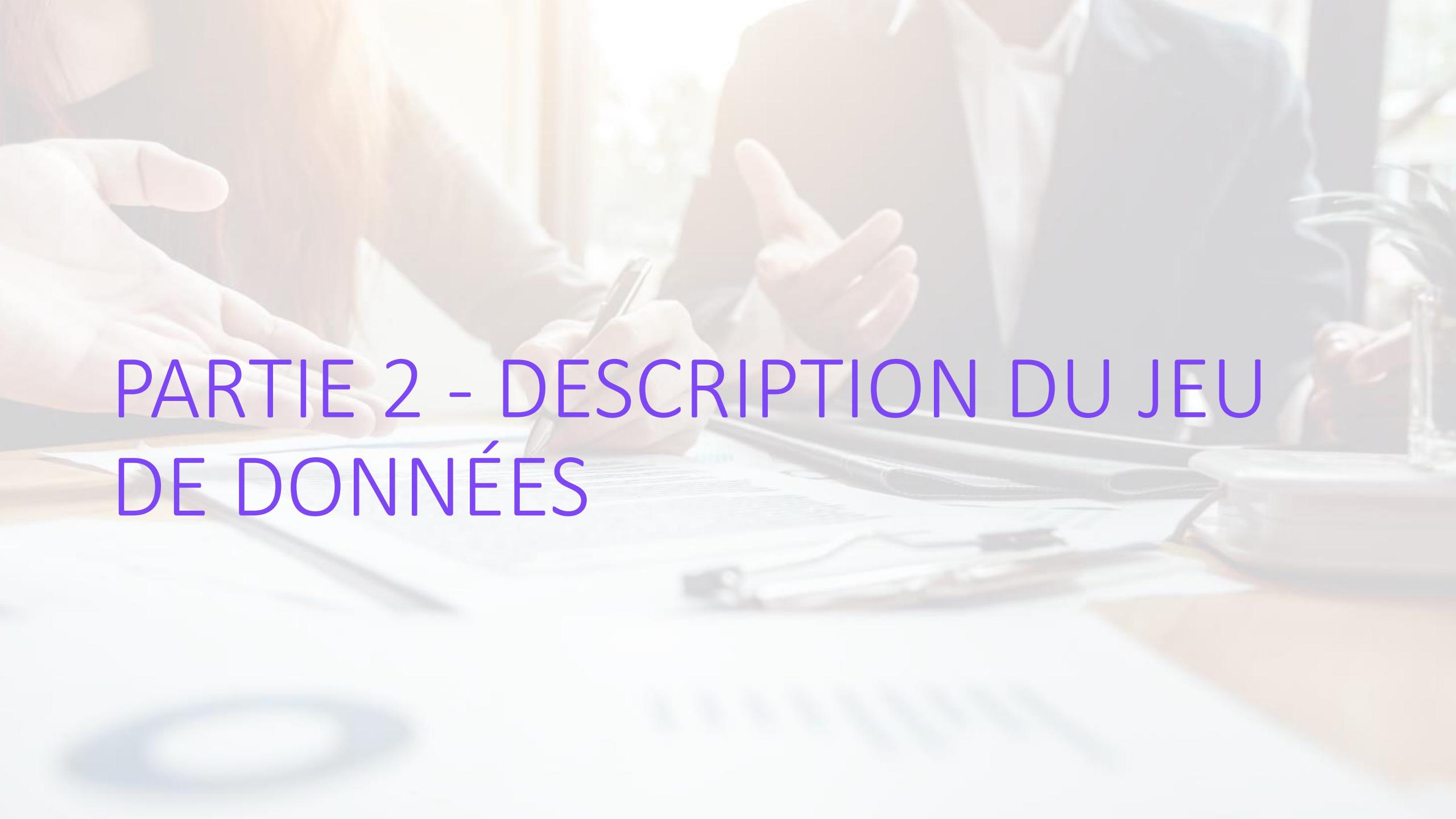
- PARTIE 1 – COMPRÉHENSION DE LA PROBLÉMATIQUE MÉTIER
- PARTIE 2 – DESCRIPTION DU JEU DE DONNÉES
- PARTIE 3 – TRANSFORMATION DU JEU DE DONNÉES
- PARTIE 4 – COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS
- PARTIE 5 – MODÈLE SÉLECTIONNÉ : ESSAI SUR LE TEST SET
- PARTIE 6 – INTERPRÉTABILITÉ DU MODÈLE
- CONCLUSION



PARTIE 1 - COMPRÉHENSION DE LA PROBLÉMATIQUE MÉTIER

COMPRÉHENSION DE LA PROBLÉMATIQUE MÉTIER

- Enjeux pour la banque :
 - mieux évaluer le risque de défaut d'un prospect, **aide à la décision**
 - **gagner du temps** lors de l'analyse des demandes de prêt
 - **optimiser les ressources** allouées par dossier
- Enjeux pour les chargés de relation client :
 - Prédiction **interprétable**...
 - Mesure de l'**importance** des features ...
 - ... pour pouvoir **vérifier dans le dossier** les éléments suspects
 - ... pour pouvoir « critiquer » le rendu du modèle et **déceler les éventuelles erreurs**



PARTIE 2 - DESCRIPTION DU JEU DE DONNÉES

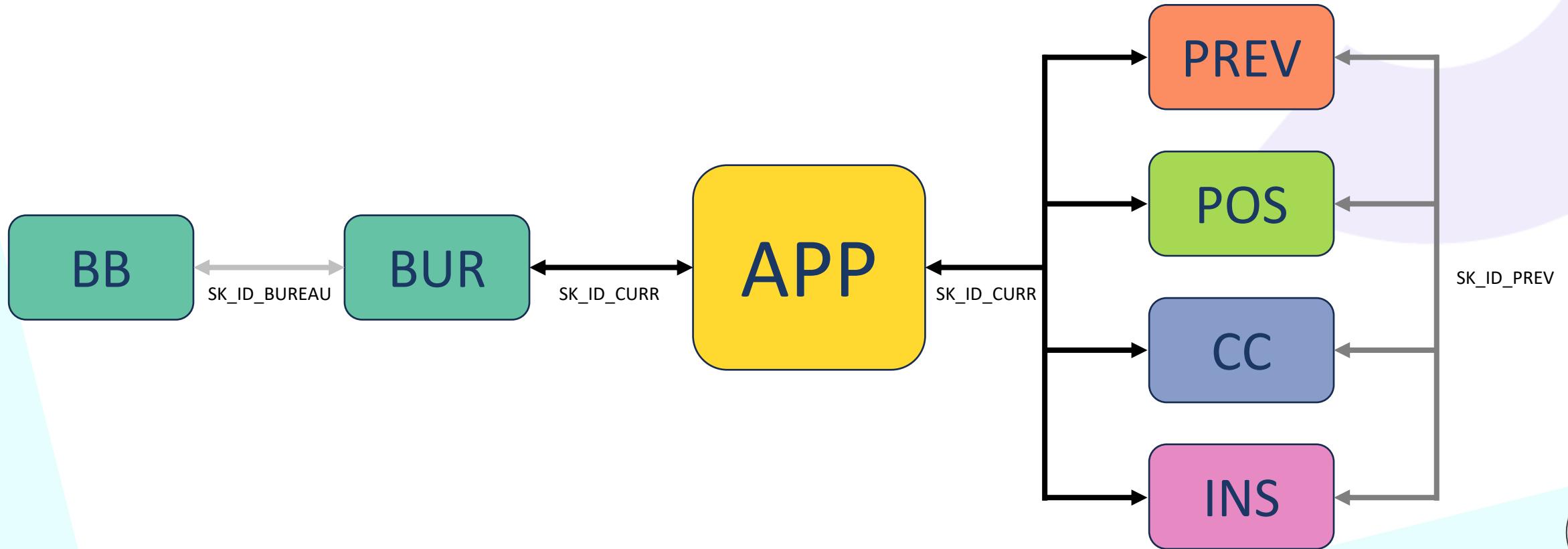
DESCRIPTION DU JEU DE DONNÉES

- Plusieurs jeux de données :

application_train.csv	Informations sur les clients actuels et leurs demandes de prêt Contient la target
bureau.csv	Informations sur d'anciennes demandes de prêt enregistrés au sein d'autres institutions financières (via l'organisme <i>Credit Bureau</i>)
bureau_balance.csv	Statuts mensuels des prêts <i>Credit Bureau</i> (<i>fermé, arriérés, etc.</i>)
previous_application.csv	Informations sur d'anciennes demandes de prêts
credit_card_balance.csv	Informations mensuelles sur d'anciens prêts de type « carte de crédit » (relevé, paiements, retraits, etc.)
installments_payments.csv	Informations mensuelles sur les versements d'anciens prêts
POS_CASH_balance.csv	Informations mensuelles sur d'anciens crédits de type « facilités de paiement »

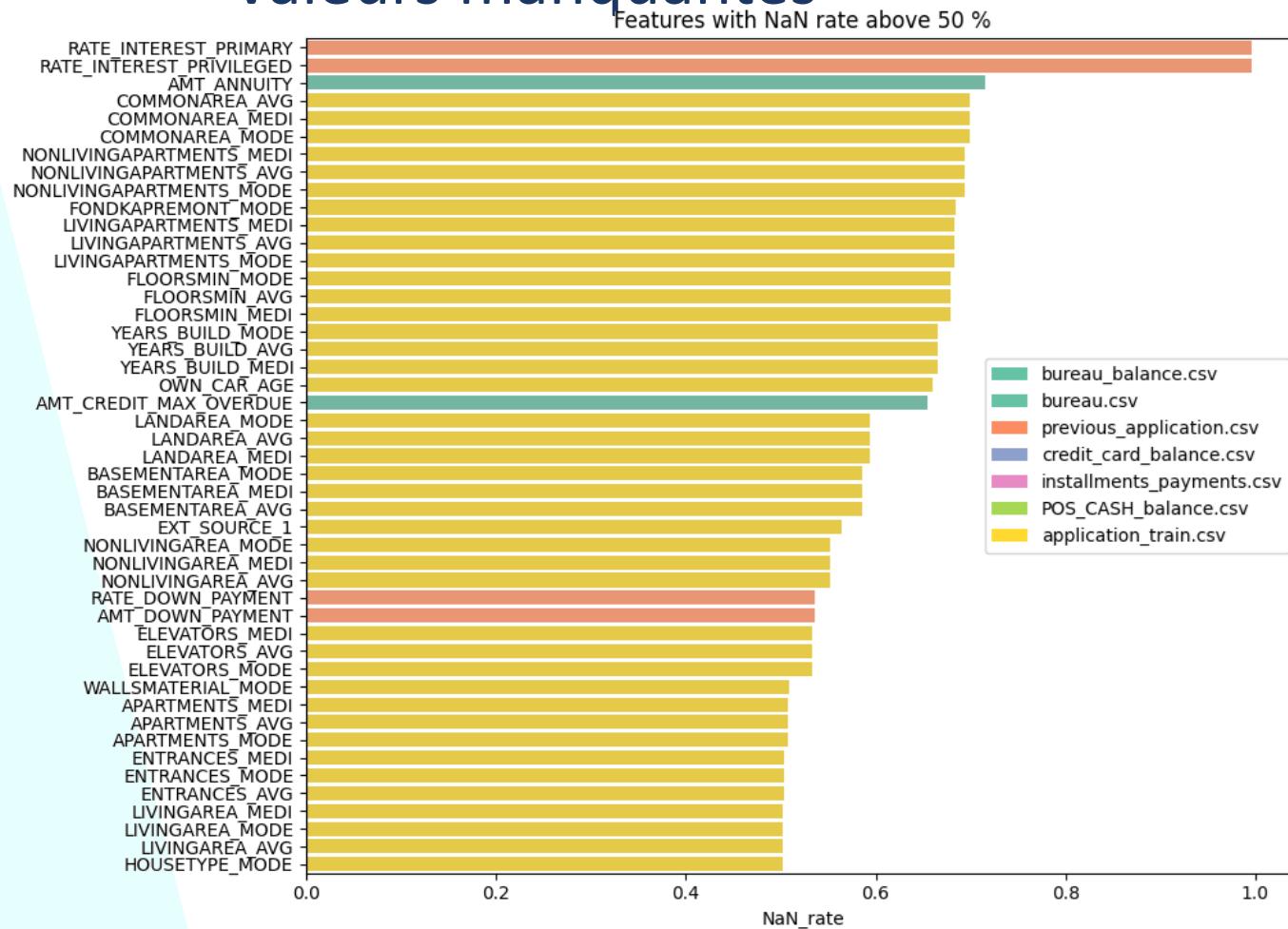
DESCRIPTION DU JEU DE DONNÉES

- Des jeux de données liés par des **identifiants** :



DESCRIPTION DU JEU DE DONNÉES

• Valeurs manquantes

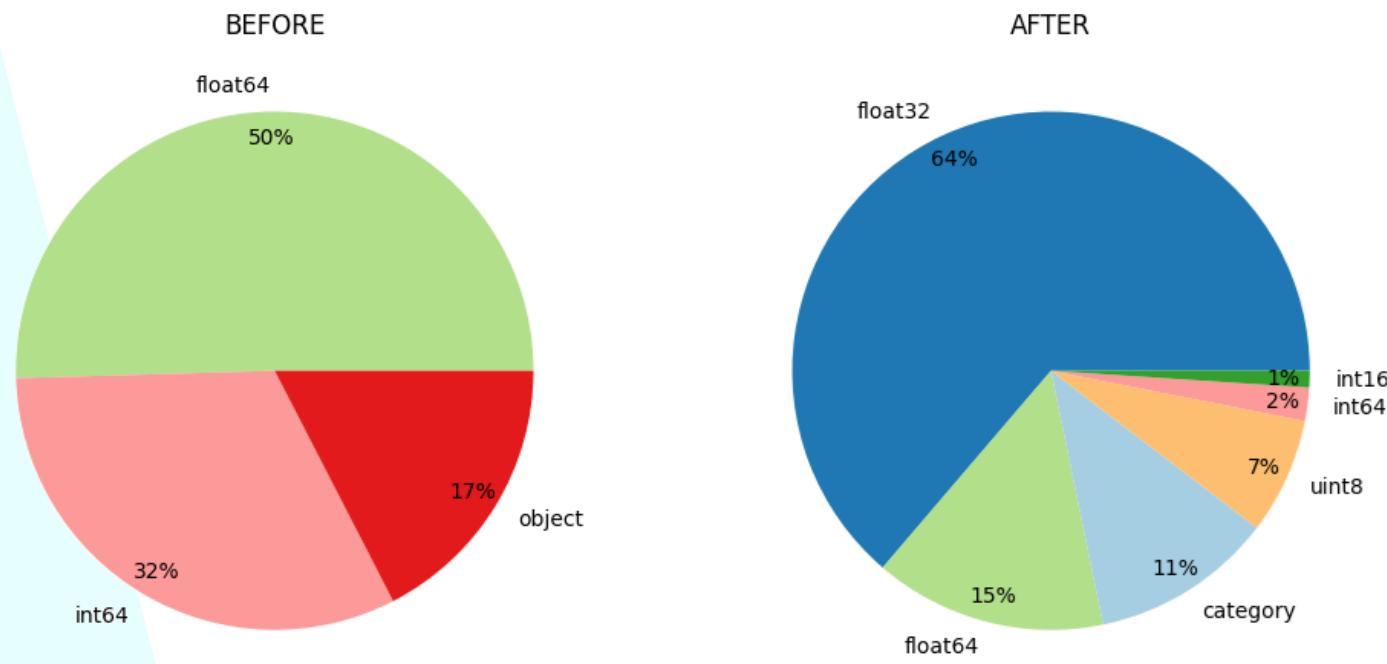


- Certaines features avec énormément de NaN
- Des valeurs à considérer comme valeurs manquantes :
 - 365243
 - « XNA »
 - « Unknown »
 - « not specified »
 - « Unknown type of loan »
 - « X »
 - etc.

DESCRIPTION DU JEU DE DONNÉES

- Types – optimisation de l'espace mémoire

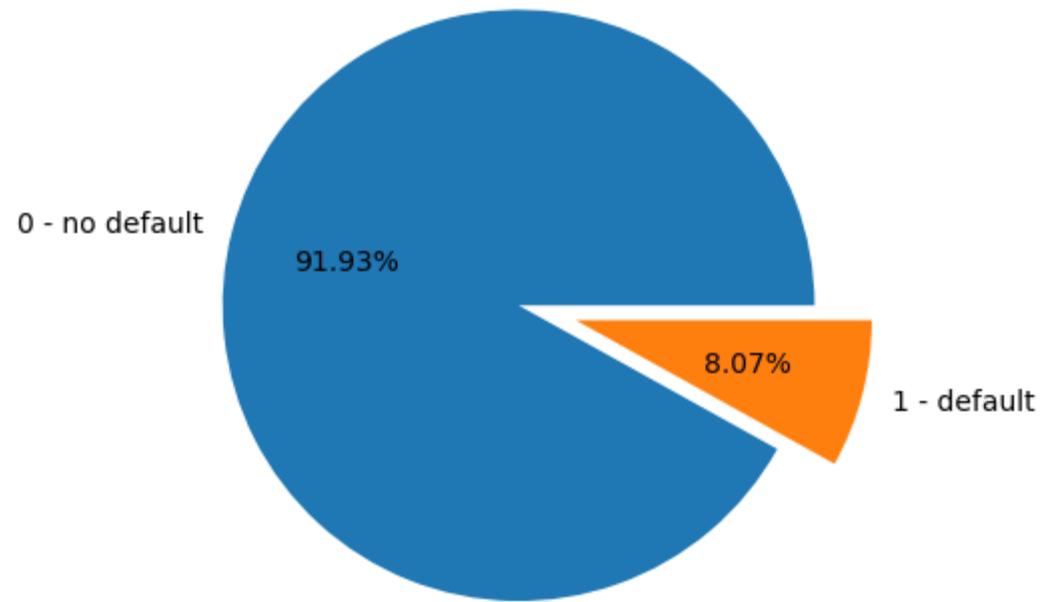
Memory management - dTypes



	rawImports	lowMemory	difference
application_train.csv	562761965	103029656	-81 %
bureau.csv	536987086	130451203	-75 %
bureau_balance.csv	2020194614	163800420	-91 %
credit_card_balance.csv	918225104	472359294	-48 %
installments_payments.csv	870745828	448978397	-48 %
POS_CASH_balance.csv	1192493276	220030874	-81 %
previous_application.csv	1992956783	192088203	-90 %

DESCRIPTION DU JEU DE DONNÉES

- Notre TARGET, quelle répartition des cas ?

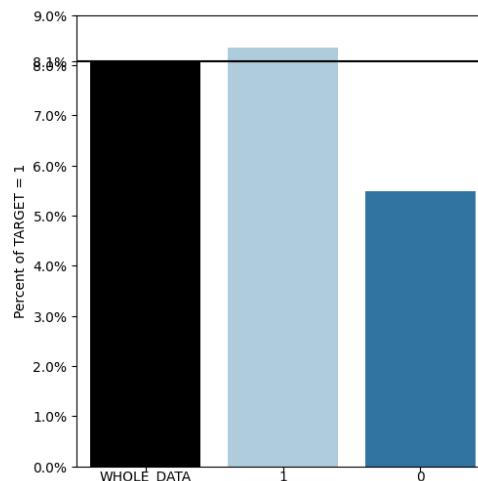
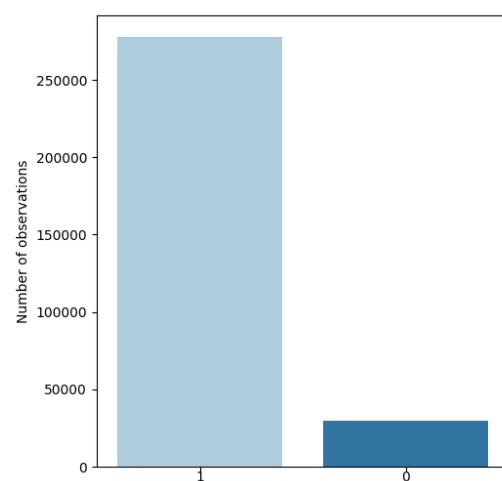


Problème de
Classification
déséquilibrée

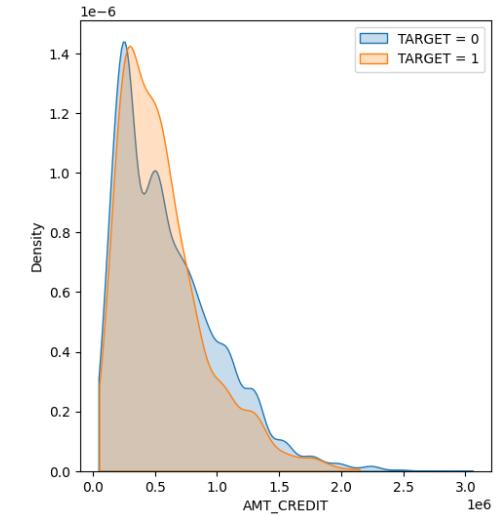
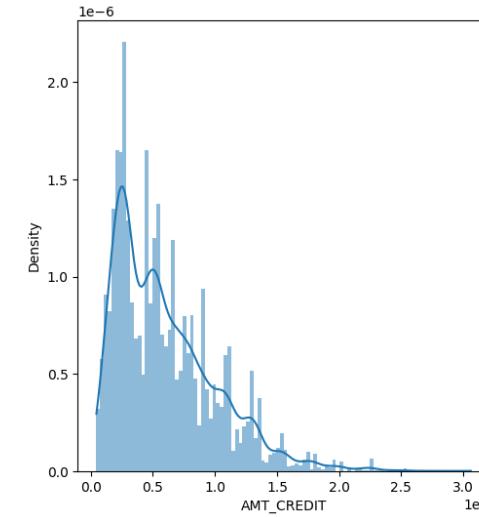
DESCRIPTION DU JEU DE DONNÉES

- Notre TARGET, des liens avec nos features ?

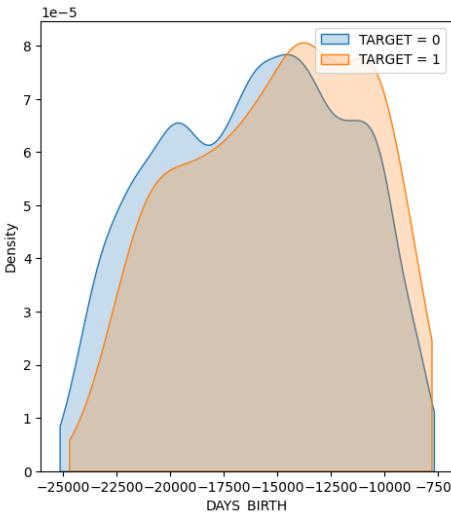
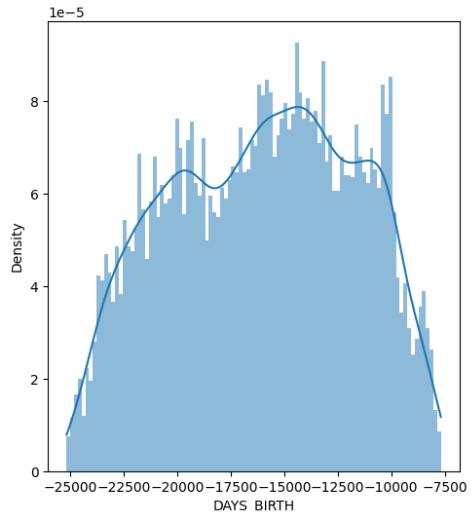
NAME_CONTRACT_TYPE (0.0% NaN) : distribution and relation with Target



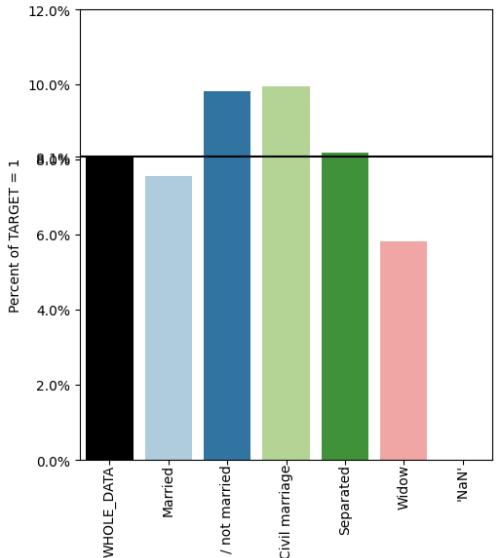
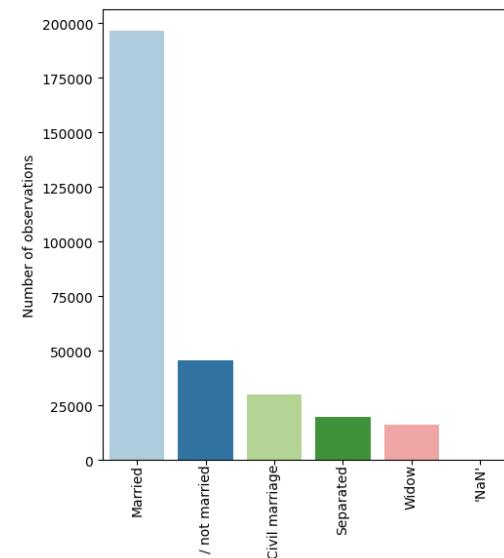
AMT_CREDIT (0.0% NaN) : distribution and relation with Target



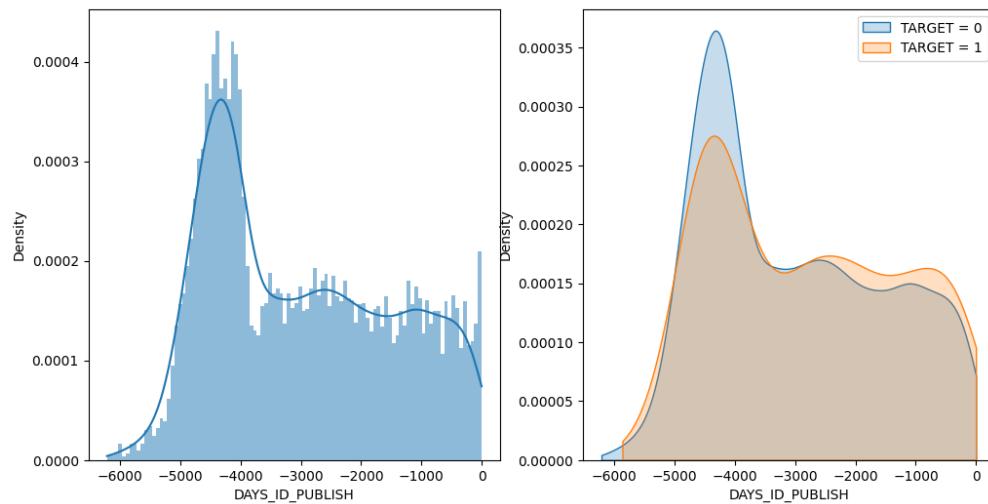
DAY_S_BIRTH (0.0% NaN) : distribution and relation with Target



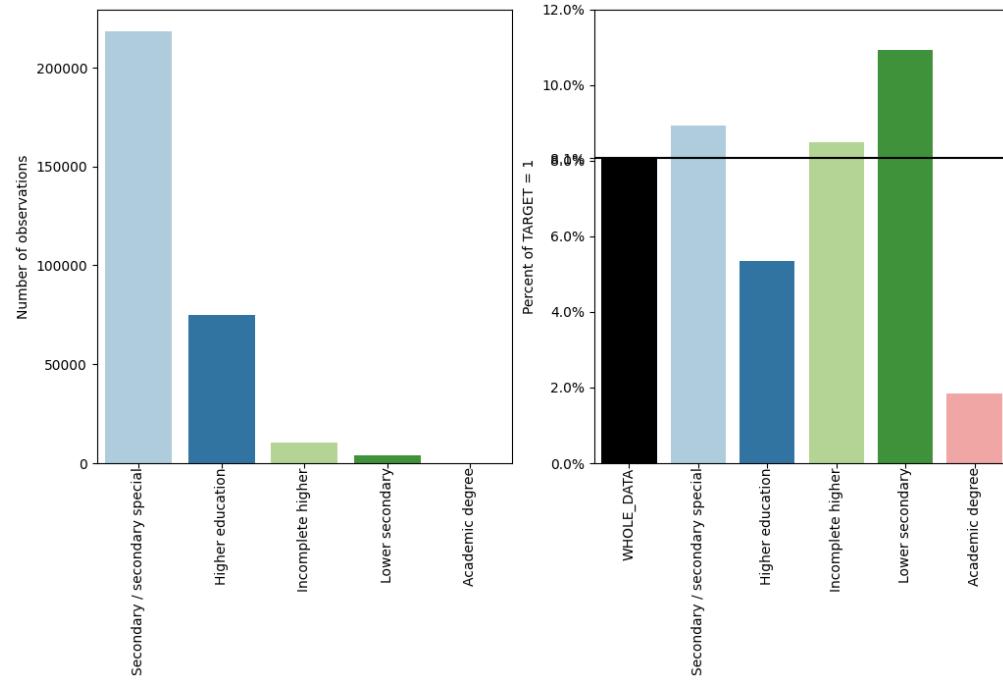
NAME_FAMILY_STATUS (0.0% NaN) : distribution and relation with Target



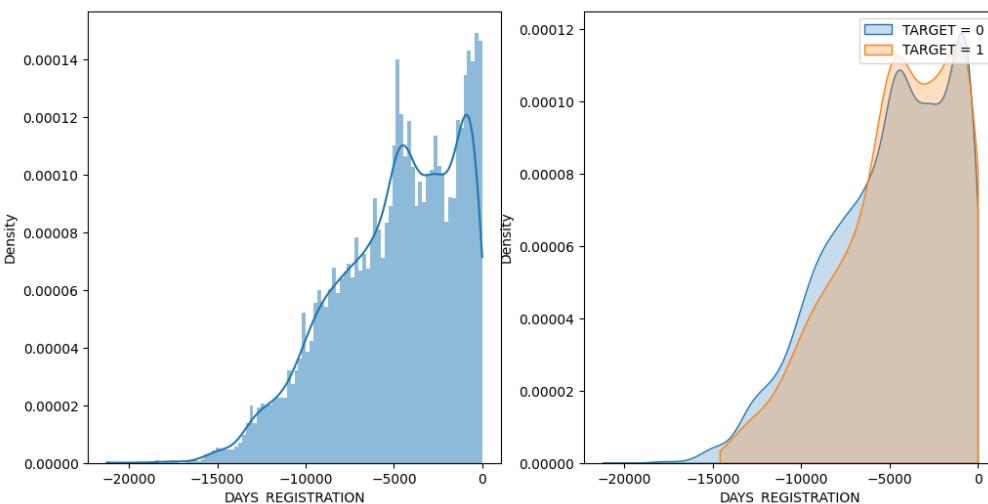
DAY_ID_PUBLISH (0.0% NaN) : distribution and relation with Target



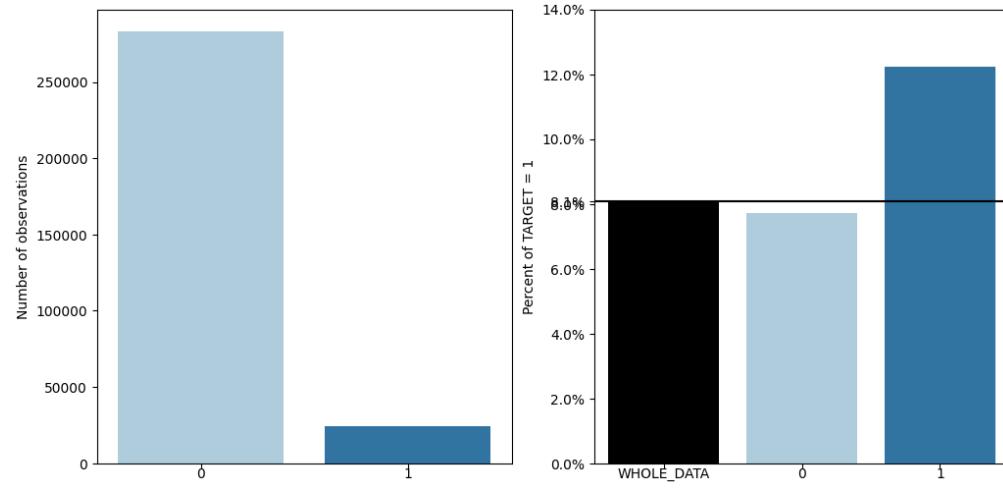
NAME_EDUCATION_TYPE (0.0% NaN) : distribution and relation with Target



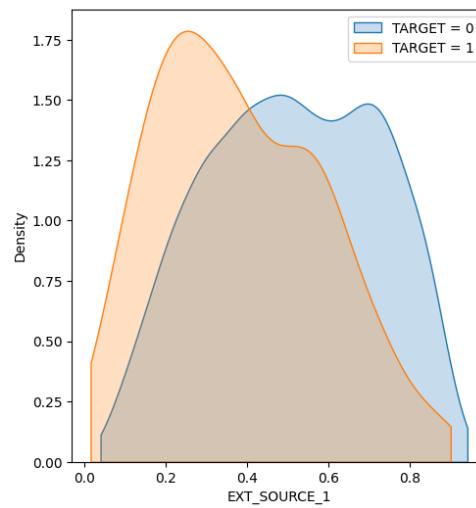
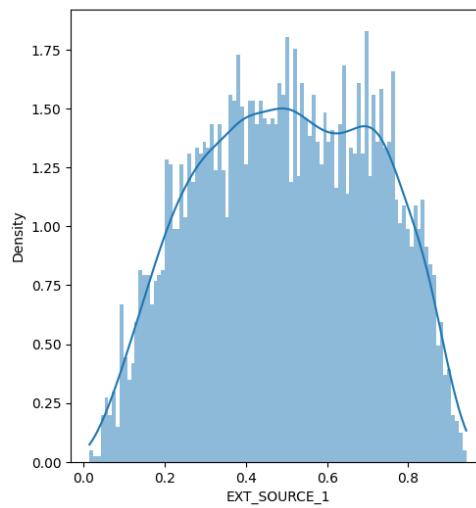
DAY_REGISTRATION (0.0% NaN) : distribution and relation with Target



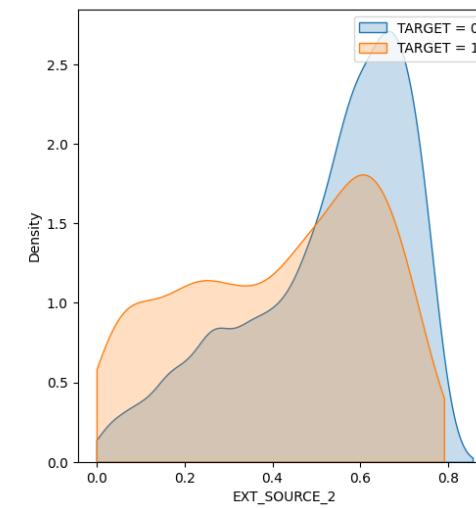
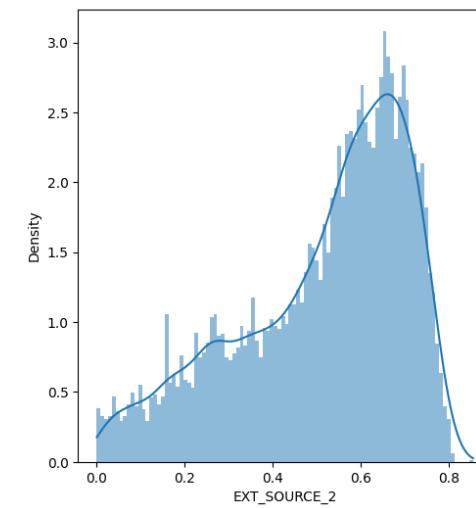
REG_CITY_NOT_LIVE_CITY (0.0% NaN) : distribution and relation with Target



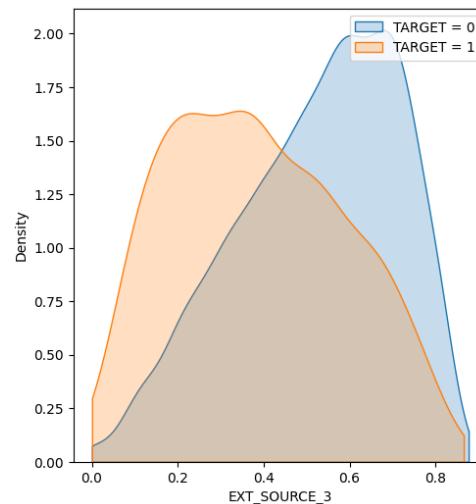
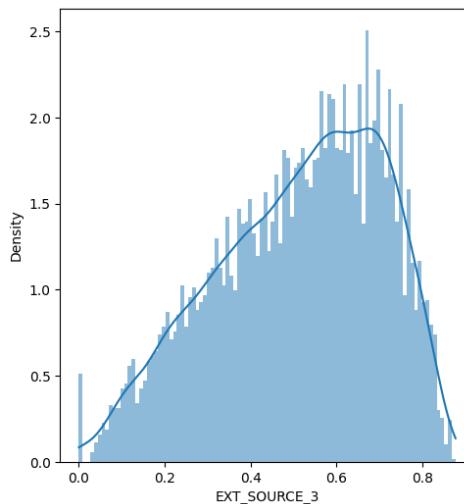
EXT_SOURCE_1 (56.4% NaN) : distribution and relation with Target



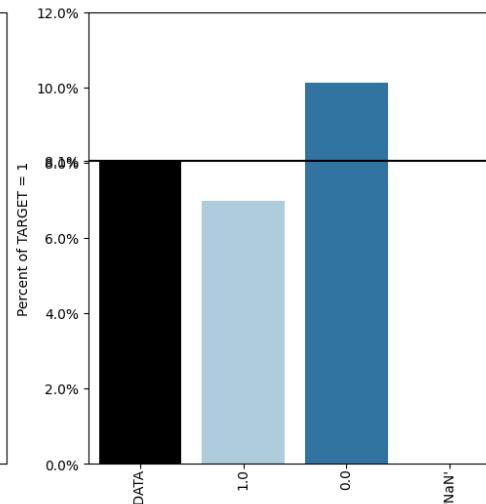
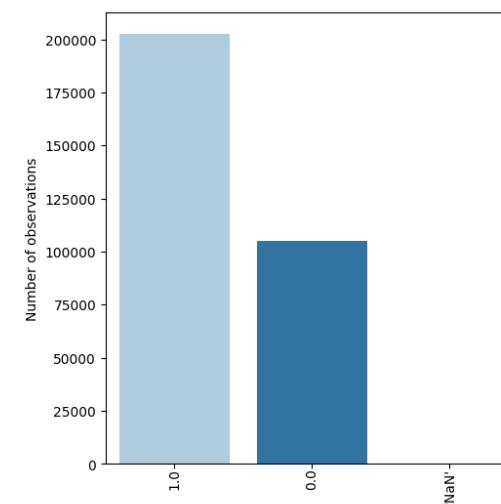
EXT_SOURCE_2 (0.2% NaN) : distribution and relation with Target



EXT_SOURCE_3 (20.0% NaN) : distribution and relation with Target



CODE_GENDER (0.0% NaN) : distribution and relation with Target



A blurred background image of a person sitting at a desk, looking down at a laptop screen with a thoughtful expression, with their hand near their chin.

PARTIE 3 - TRANSFORMATION DU JEU DE DONNÉES

TRANSFORMATION DU JEU DE DONNÉES

- Création de features - APP :

CAPITAL / MENSUALITÉ

APPORT NORMALISÉ

TAUX DE FINANCEMENT

MENSUALITÉ / REVENUS

DURÉE EMPLOI / AGE

Moyenne EXT_SOURCE s

Produit EXT_SOURCE s

Somme xx_NOT_xx s

Somme certains docs

- Création de features - PREV :

CAPITAL demandé – CAPITAL , normalisé

TX INTÉRÊT SIMPLE

TX INTÉRÊT SIMPLE

X

APPORT NORMALISÉ

TRANSFORMATION DU JEU DE DONNÉES

- Crédit de features - BUR :

CAPITAL RESTANT / DURÉE RESTANTE

- Crédit de features - CC :

Somme types RETRAIT

- Crédit de features - POS :

NB ÉCHÉANCES RESTANTES, au moment de la demande

- Crédit de features - INS :

PAIEMENT EFFECTUÉ – PAIEMENT ATTENDU, normalisé

DATE PAIEMENT – DATE ATTENDUE, normalisé

TRANSFORMATION DU JEU DE DONNÉES

- Crédit de features - BB CC POS INS :
 - S'intéresser aux derniers mois avant la demande de prêt !
 - Trier les datasets en fonction de la feature temporelle (par exemple *MONTH_BALANCE*)
 - Regrouper les observations par identifiant de prêt (ex *SK_ID_PREV*)
 - Et ne garder que les n plus récents

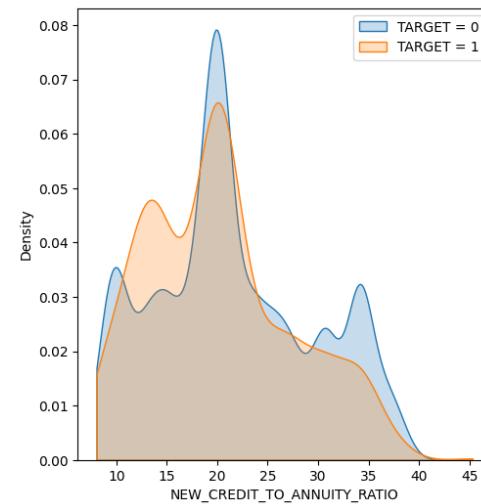
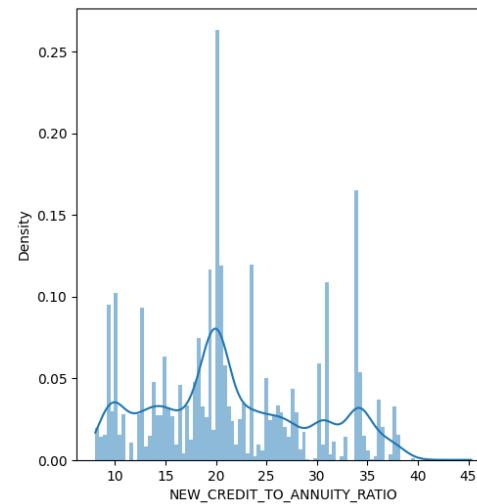
Last10_xxx

Last10_xxx

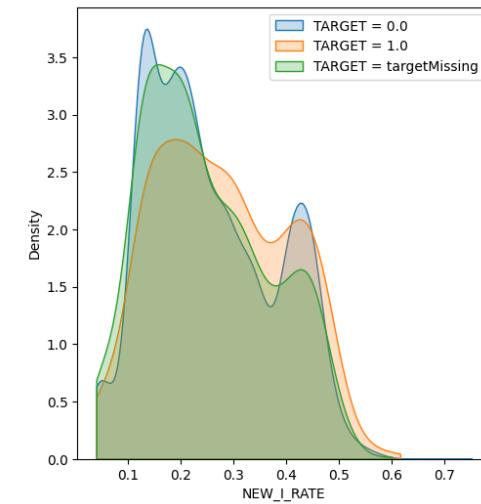
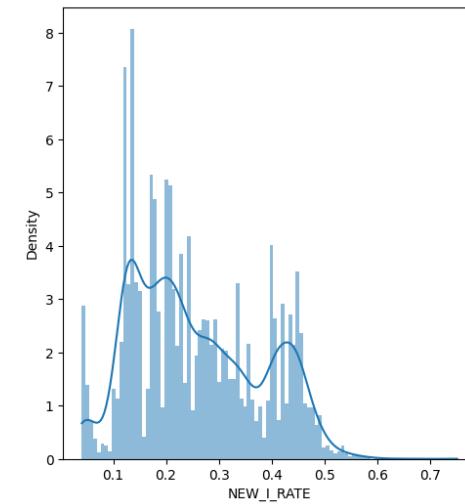
Last10_xxx

Last10_xxx

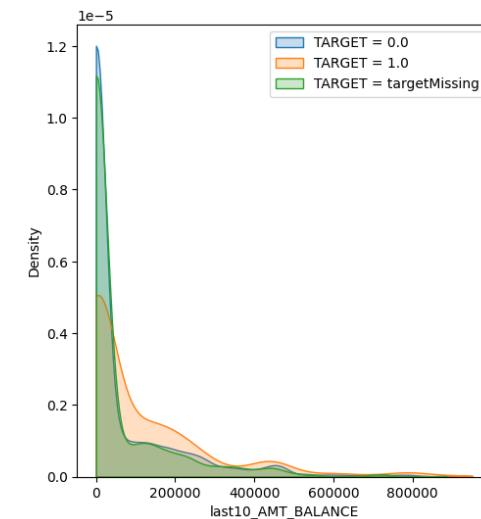
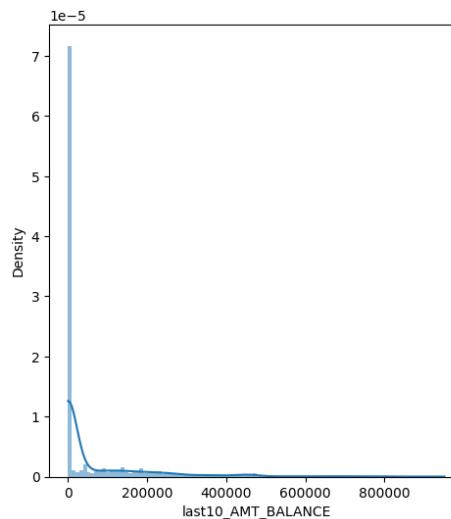
NEW_CREDIT_TO_ANNUITY_RATIO (0.0% NaN) : distribution and relation with Target



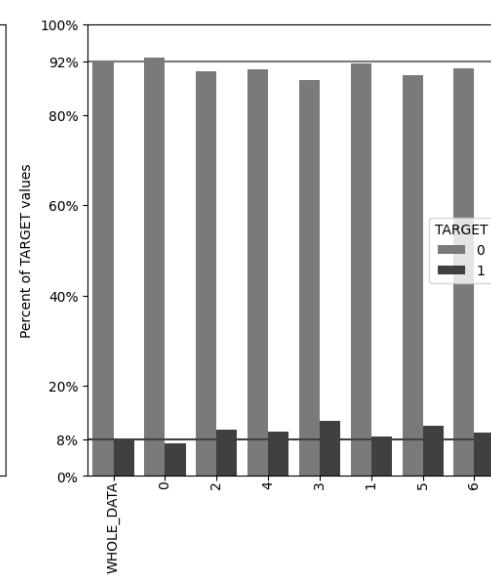
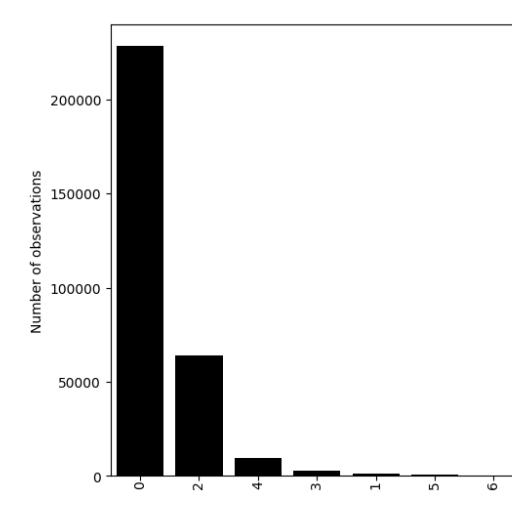
NEW_I_RATE (31.9% NaN) : distribution and relation with Target



last10_AMT_BALANCE (0.0% NaN) : distribution and relation with Target

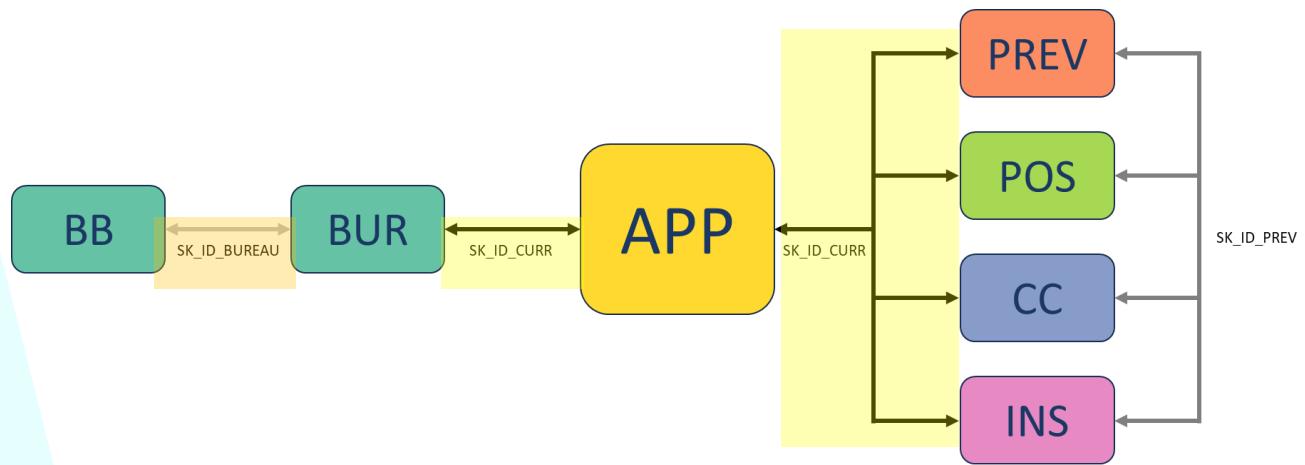


NEW_is_NOT_sth (0.0% NaN) : distribution and relation with Target



TRANSFORMATION DU JEU DE DONNÉES

- Unifier les datasets :
 1. Faire des **jointures**, grâce à nos **identifiants** !



	%InterWithAppli	%InterWithBur	%InterWithPrev
lowMemAppliDf	100%	NaN	NaN
lowMemBurDf	86%	100%	NaN
lowMemBurBalDf	NaN	45%	NaN
lowMemCcBalDf	28%	NaN	6%
lowMemInstalPaymDf	95%	NaN	57%
lowMemPosDf	94%	NaN	54%
lowMemPrevAppDf	95%	NaN	100%

- BB → BUR grâce à *SK_ID_BUREAU*
- BUR, PREV, POS, CC, INS → APP grâce à *SK_ID_CURR*

TRANSFORMATION DU JEU DE DONNÉES

- Unifier les datasets :
 2. Comment traiter les doublons de *SK_ID_BUREAU* et *SK_ID_CURR* ?
 - 1 prêt dans BUR a plusieurs lignes dans BB
 - 1 client peut avoir plusieurs anciens prêts
 - etc.

→ Faire des **agrégations** avant les jointures
 3. Comment traiter les features catégorielles ?
 - **encodage**
 - Pour les features ordinaires : Ordinal encoding
 - Pour les features nominales : One Hot encoding
 - Pour les features nominales avec beaucoup de catégories : Target encoding

TRANSFORMATION DU JEU DE DONNÉES

- Unifier les datasets :

3. Comment traiter les features catégorielles ?

target	feature
0	1 A
1	0 A
2	0 B
3	1 C
4	0 B
5	0 B
6	1 A
7	0 A
8	0 C
9	1 A

target	feature
0	1 0
1	0 0
2	0 1
3	1 2
4	0 1
5	0 1
6	1 0
7	0 0
8	0 2
9	1 0

Ordinal encoding

target	feature_A	feature_B	feature_C
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1
5	0	0	1
6	1	1	0
7	0	1	0
8	0	0	1
9	1	1	0

One Hot encoding

target	feature
0	1 0.350000
1	0 0.800000
2	0 0.200000
3	1 0.200000
4	0 0.133333
5	0 0.200000
6	1 0.350000
7	0 0.800000
8	0 0.700000
9	1 0.800000

Target encoding

TRANSFORMATION DU JEU DE DONNÉES

- Unifier les datasets :

4. Quels types d'agrégations ?

- *Moyenne*
- *Min*
- *Max*
- *Somme*
- Etc.

Et pour les features devant subir un Target Encoding : *Mode*

TRANSFORMATION DU JEU DE DONNÉES

- Séparer nos données

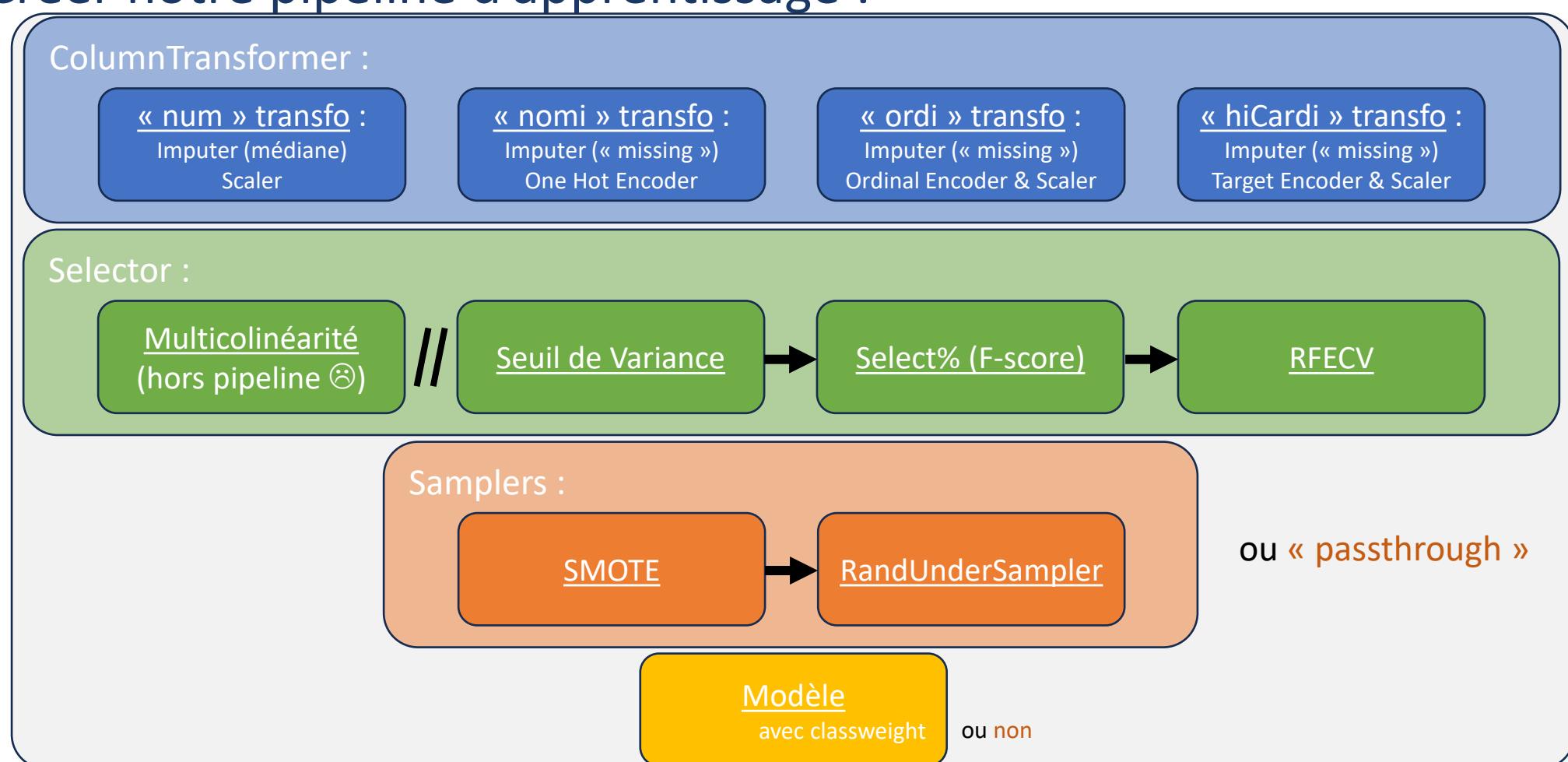
- Un jeu d'entraînement
- Un jeu de test

En conservant le déséquilibre de classes dans les 2 jeux :

	y	yTrain	yTest
0, no default	91.93%	91.93%	91.93%
1, default	8.07%	8.07%	8.07%

TRANSFORMATION DU JEU DE DONNÉES

- Créer notre pipeline d'apprentissage :





PARTIE 4 - COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Métrique :

- Une approche coût métier !
- Comment fonctionne le modèle ?

yProb : score/probabilité $\in [0,1]$

Th : seuil de classification $\in [0,1]$

yPred : prédiction

Si $yProb \geq Th$, alors $yPred = 1$

- Combien coûte une erreur ?

Faux-positif : demande de prêt refusée alors que le client serait solvable → manque à gagner (intérêts, etc.)

Faux-négatif : demande de prêt acceptée alors que le client fera défaut → coûts directs (impayés, etc.)

- Hypothèse : coût FN = 10 coût FP

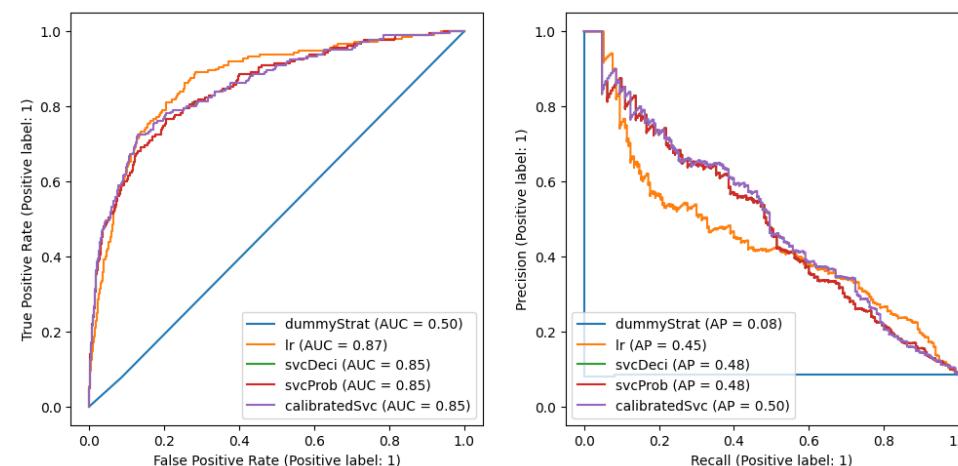
À discuter avec le business (quid des recettes ? Apports d'un TP ? Apports d'un TN ? etc.)

COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Métrique :
 - Minimiser les FN → Maximiser le Recall
 - Minimiser les FP → Maximiser la Precision
→ Precision / Recall Curve

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <i>Type II Error</i>	Sensitivity $\frac{TP}{TP + FN}$
	Negative	False Positive (FP) <i>Type I Error</i>	True Negative (TN)	Specificity $\frac{TN}{TN + FP}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TN + FP + FN}$

(Pourquoi ne pas minimiser la Spécificité et utiliser une ROC curve? Car FP « noyés » dans les TN :

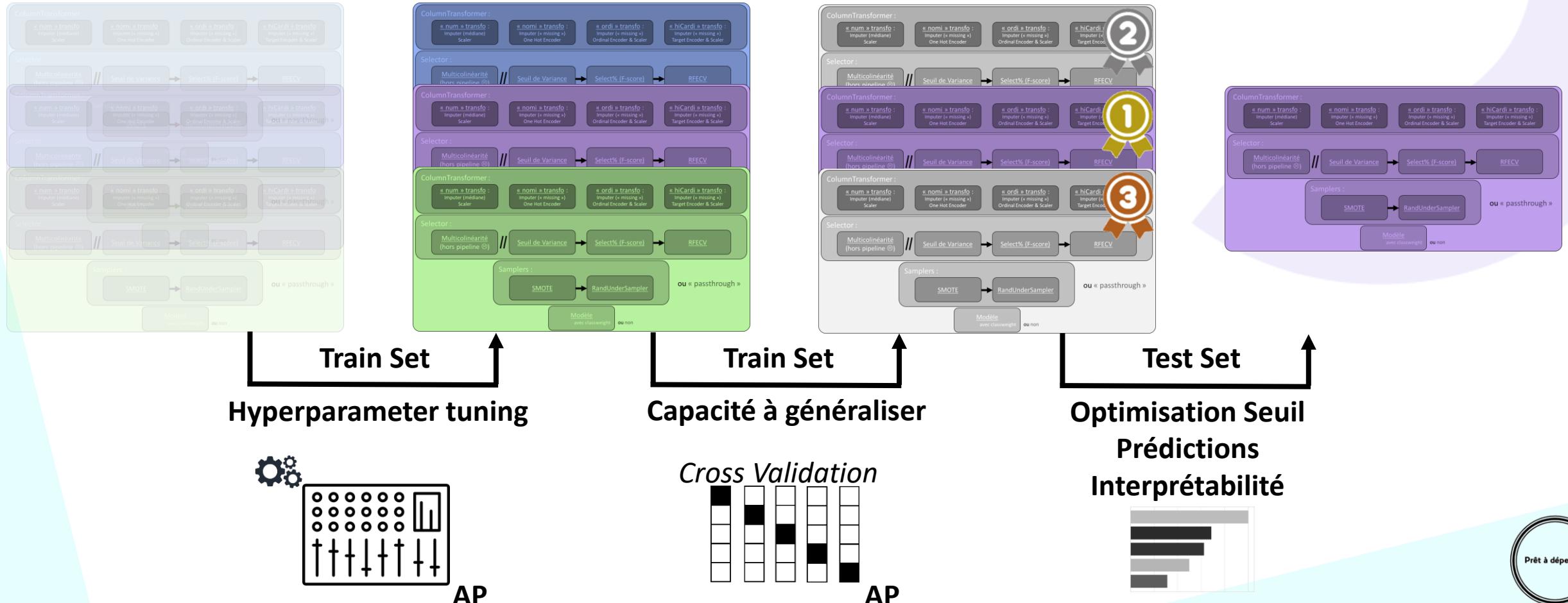


COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Métrique :
 - Métrique d'**évaluation** indépendante du seuil : **Average Precision (AP)**
 - **Une fois** le modèle sélectionné :
 - utiliser notre **hypothèse** ($cFN = 10 \times cFP$) !
 - Calculer coût moyen prédiction pour chaque seuil
 - Choisir le seuil avec le coût le plus bas

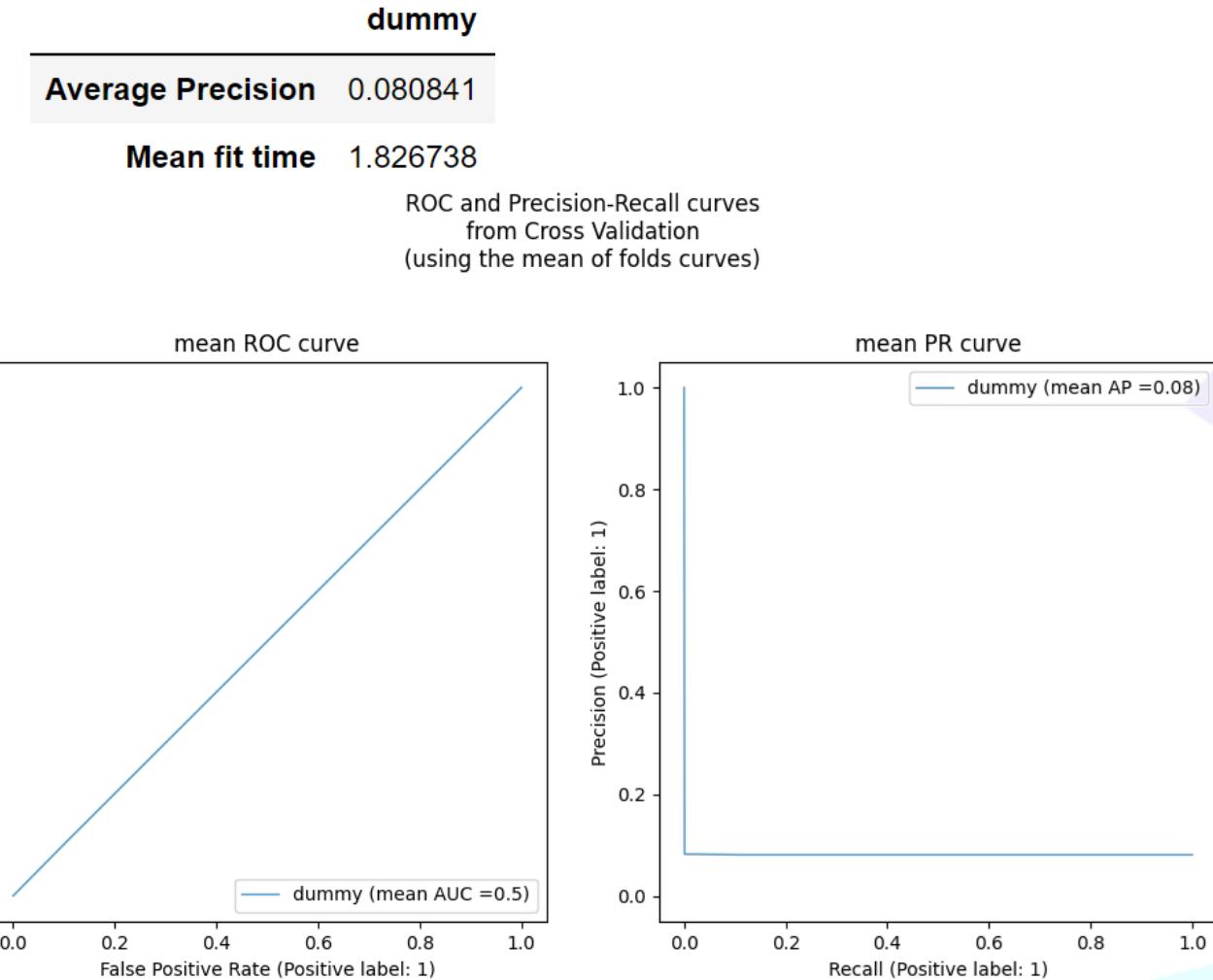
COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Méthode :



COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Dummy :

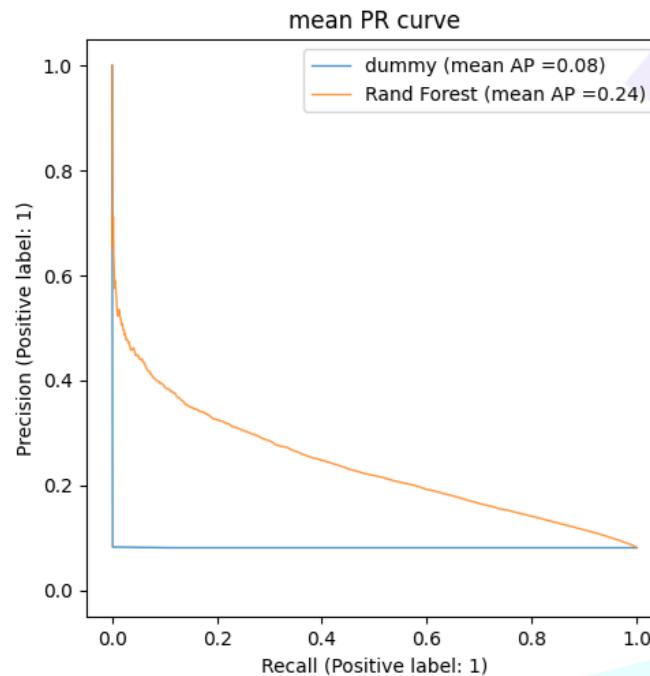
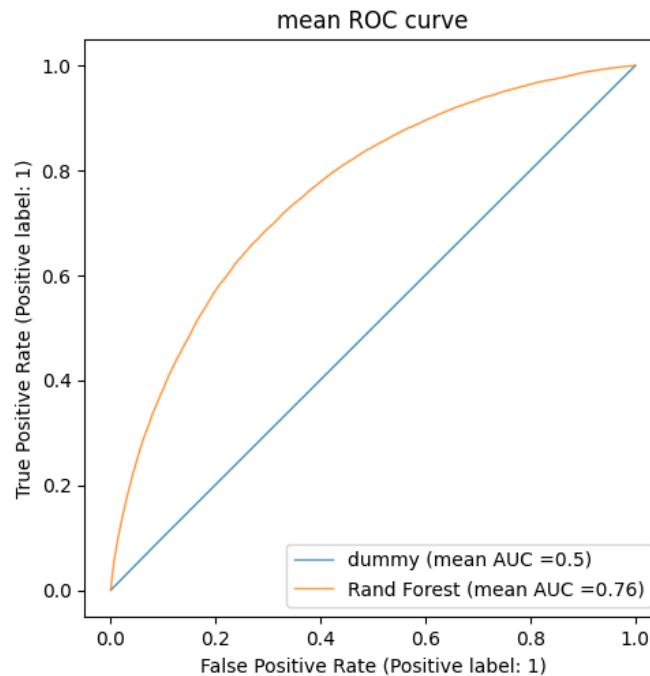


COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Random Forest :

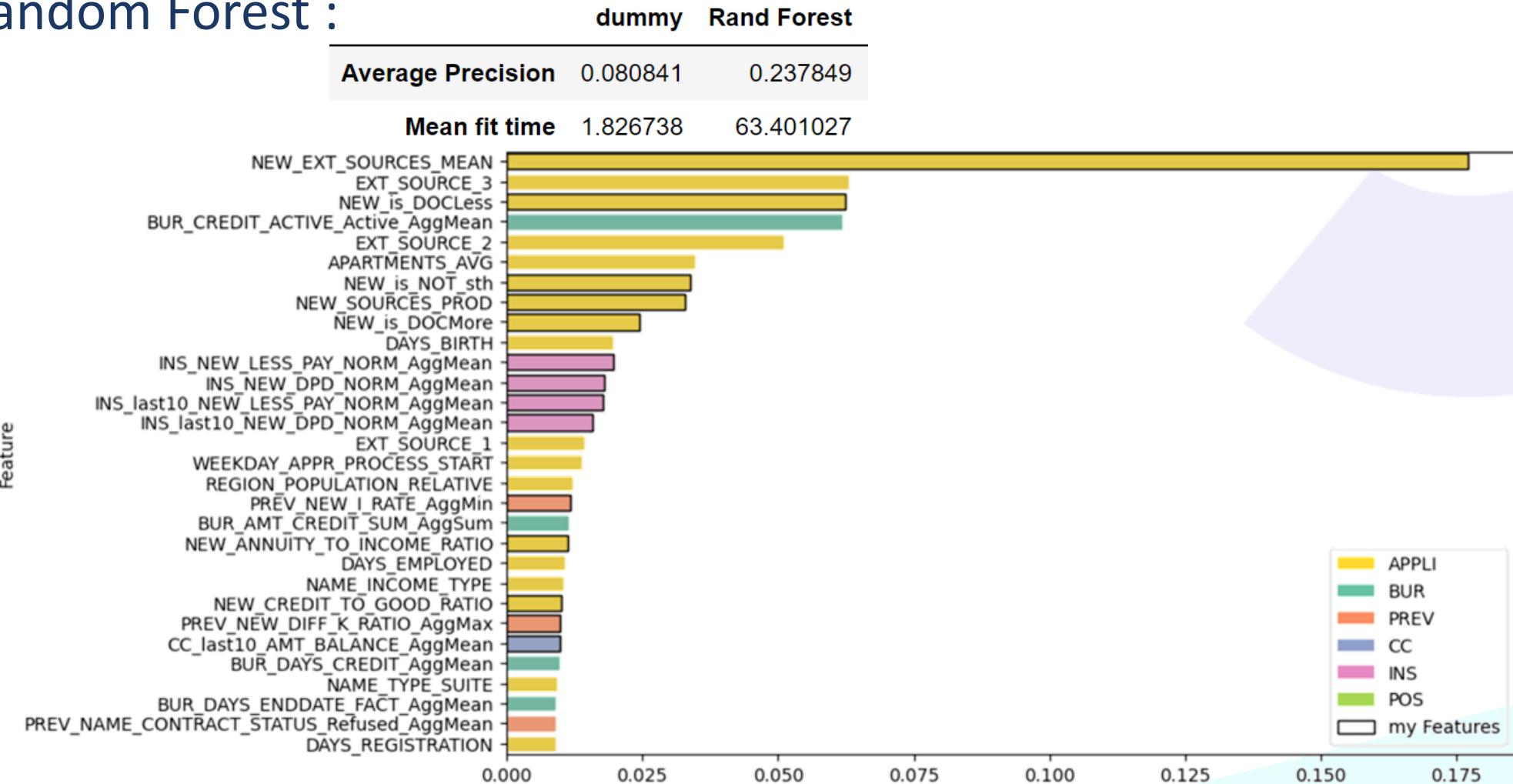
	dummy	Rand Forest
Average Precision	0.080841	0.237849
Mean fit time	1.826738	63.401027

ROC and Precision-Recall curves
from Cross Validation
(using the mean of folds curves)



COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Random Forest :

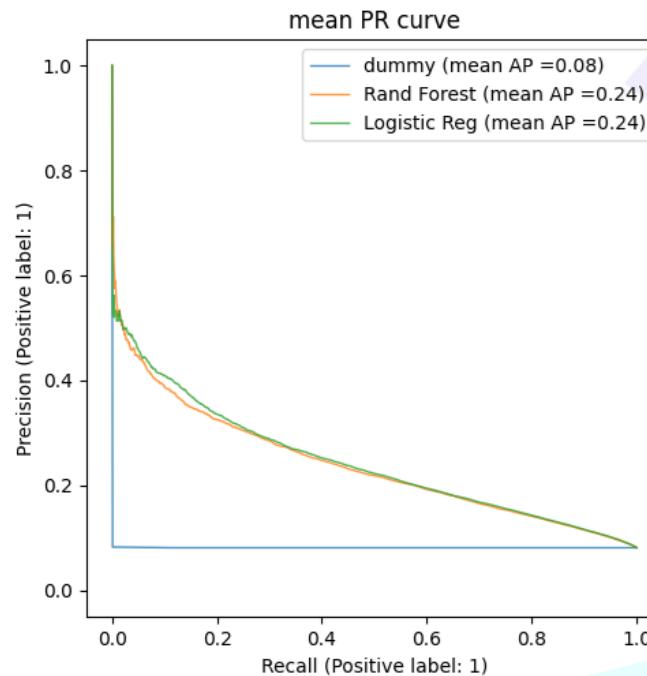
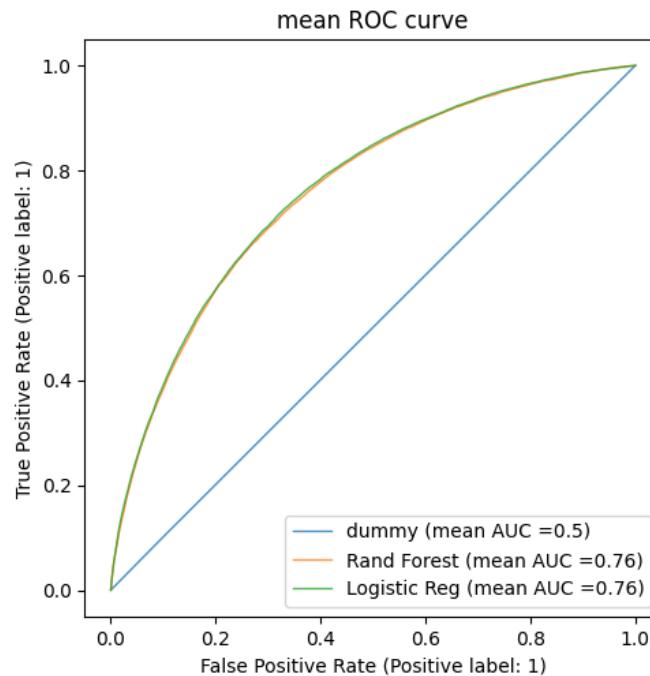


COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Régression logistique :

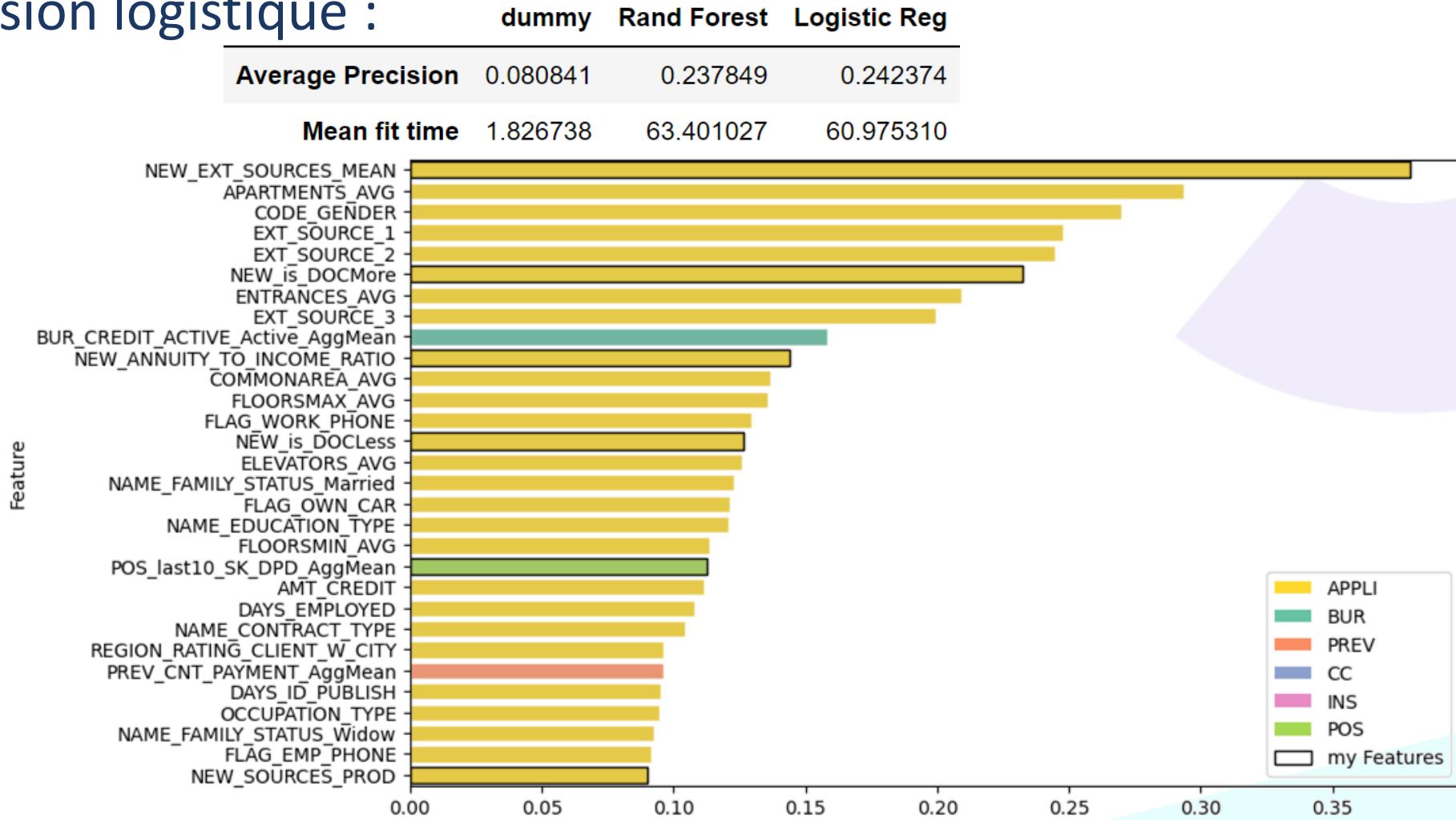
	dummy	Rand Forest	Logistic Reg
Average Precision	0.080841	0.237849	0.242374
Mean fit time	1.826738	63.401027	60.975310

ROC and Precision-Recall curves
from Cross Validation
(using the mean of folds curves)



COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- Régression logistique :

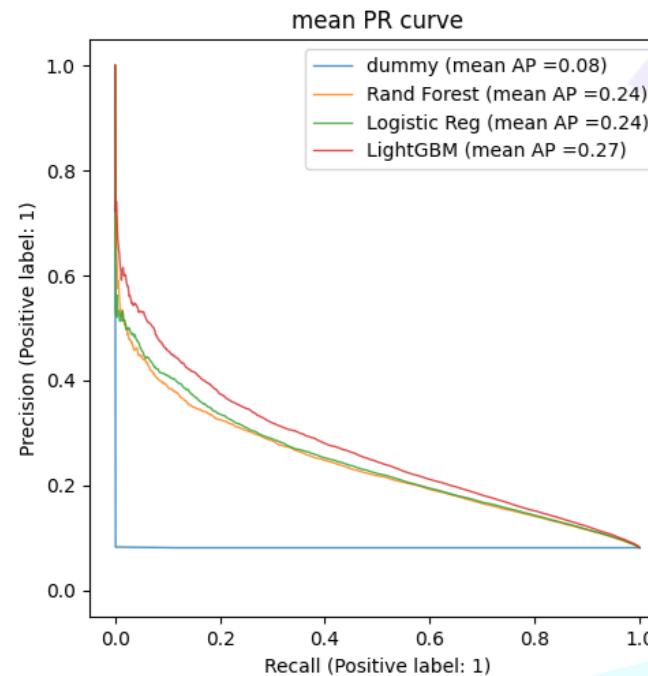
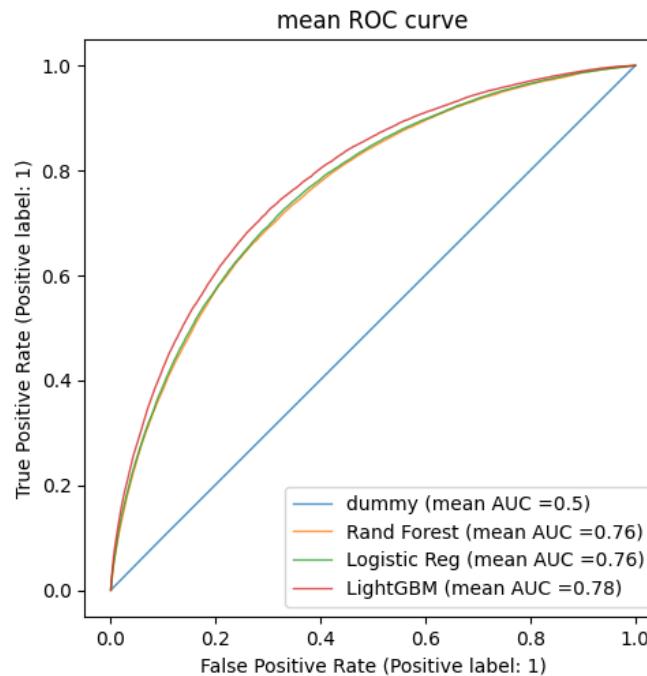


COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- LightGBM :

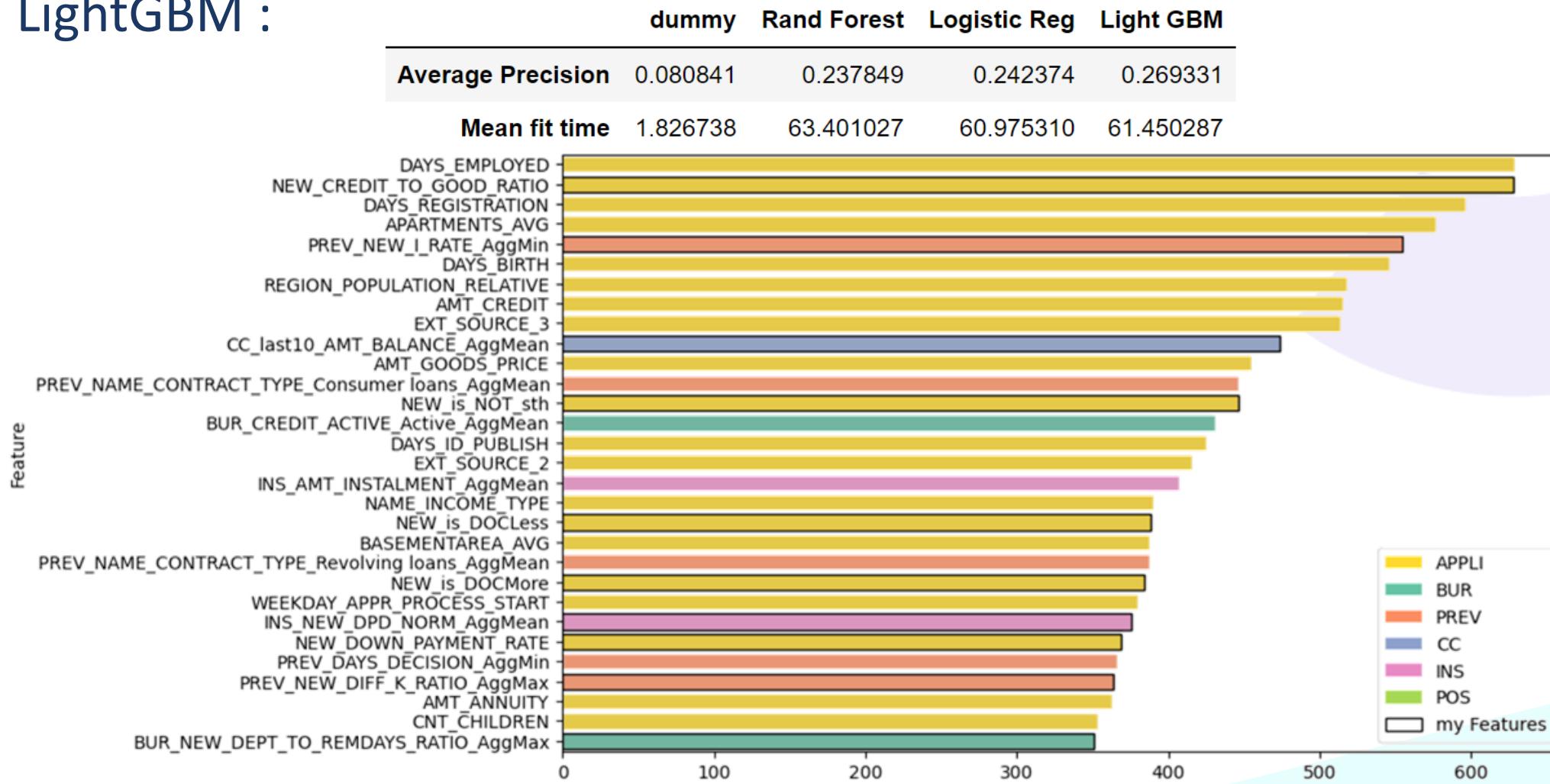
	dummy	Rand Forest	Logistic Reg	Light GBM
Average Precision	0.080841	0.237849	0.242374	0.269331
Mean fit time	1.826738	63.401027	60.975310	61.450287

ROC and Precision-Recall curves
from Cross Validation
(using the mean of folds curves)



COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

- LightGBM :



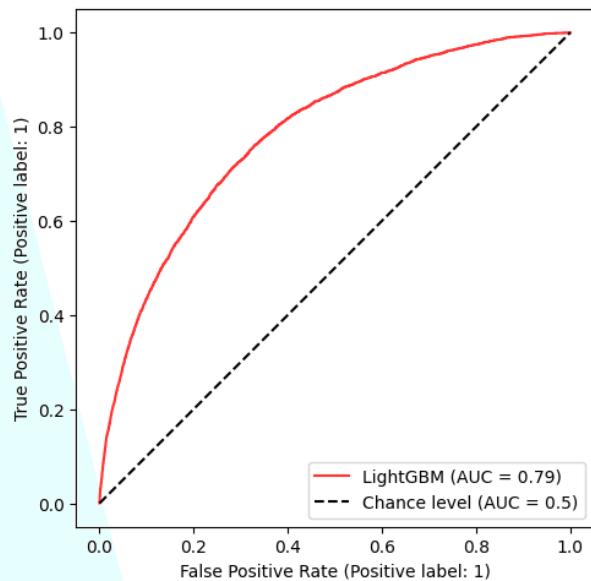
A blurred background image of a person sitting at a desk, looking down at a laptop screen with a thoughtful expression, with their hand near their chin.

PARTIE 5 – MODÈLE SÉLECTIONNÉ ESSAI SUR LE TEST SET

MODÈLE SÉLECTIONNÉ : ESSAI SUR LE TEST SET

- Quelle Average Precision ?

ROC and Precision-Recall curves
from preds on test set



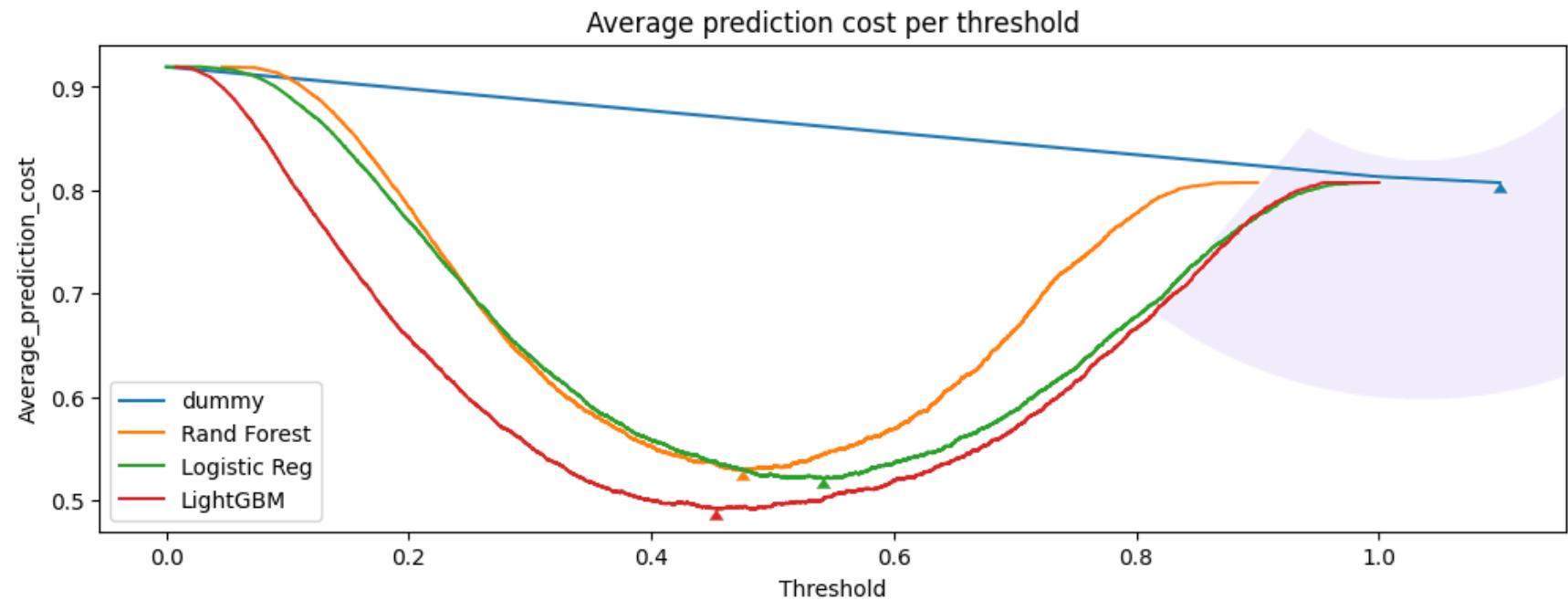
- Impact des Features composites :

LGBM Average Precision	diff%
with My and Agg features	0.278420 0.0 %
without My features	0.265955 -4.5 %
without My or Agg features	0.247446 -11.1 %

MODÈLE SÉLECTIONNÉ : ESSAI SUR LE TEST SET

- Optimisation du seuil de classification :

```
# build a cost matrix
# CTN  CFP
# CFN  CTP
PretADepenserCostMatrix=[  
    [0 , 1],  
    [10, 0]  
]
```



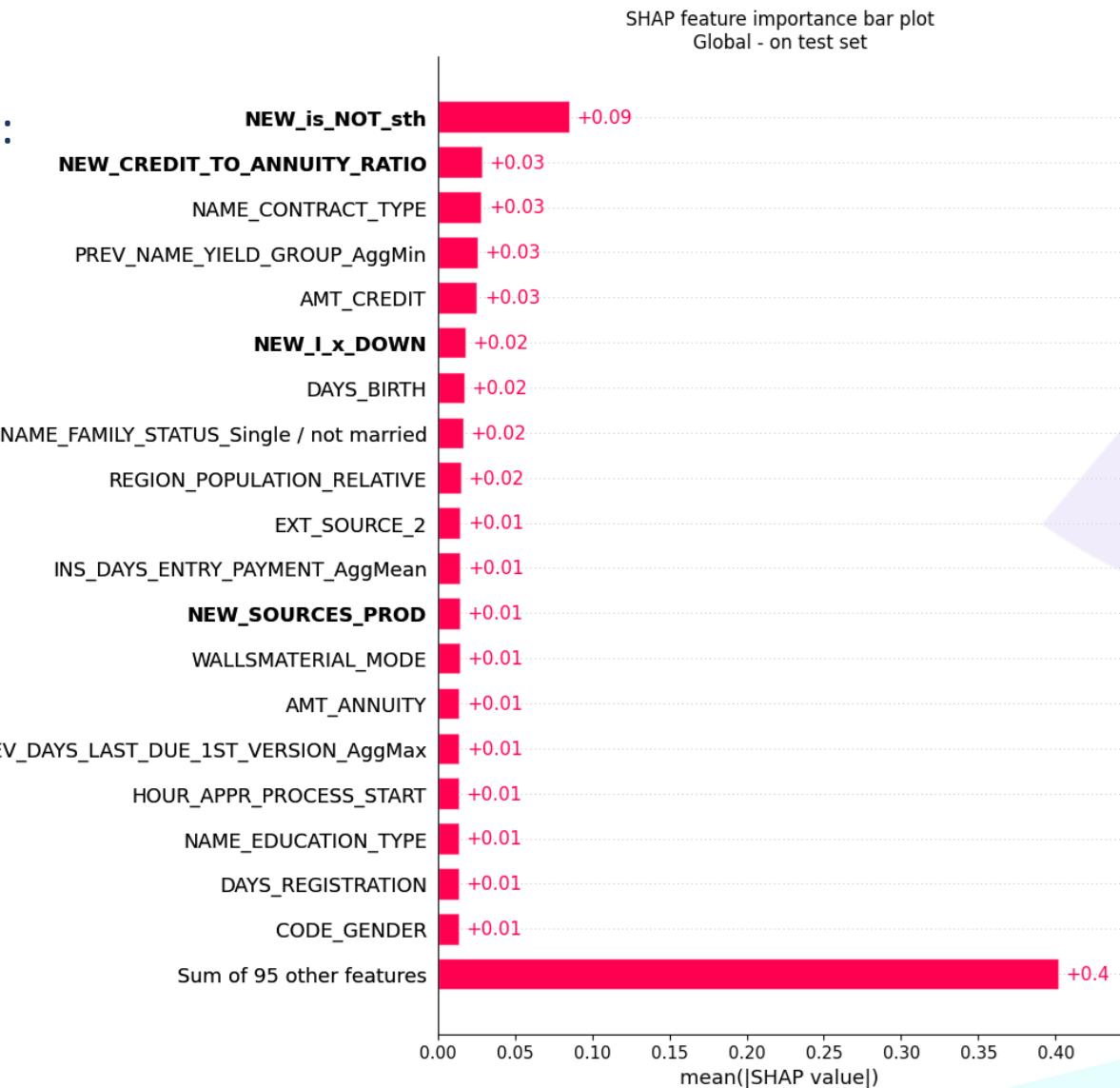
Nota : cela n'est pas indispensable, car ils sont déjà rejétés , mais nous avons regardé aussi pour les autres modèles.

A person is sitting at a desk, looking down at an open book or document. Their hands are clasped together in front of them. The background is slightly blurred.

PARTIE 6 – INTERPRÉTABILITÉ DU MODÈLE

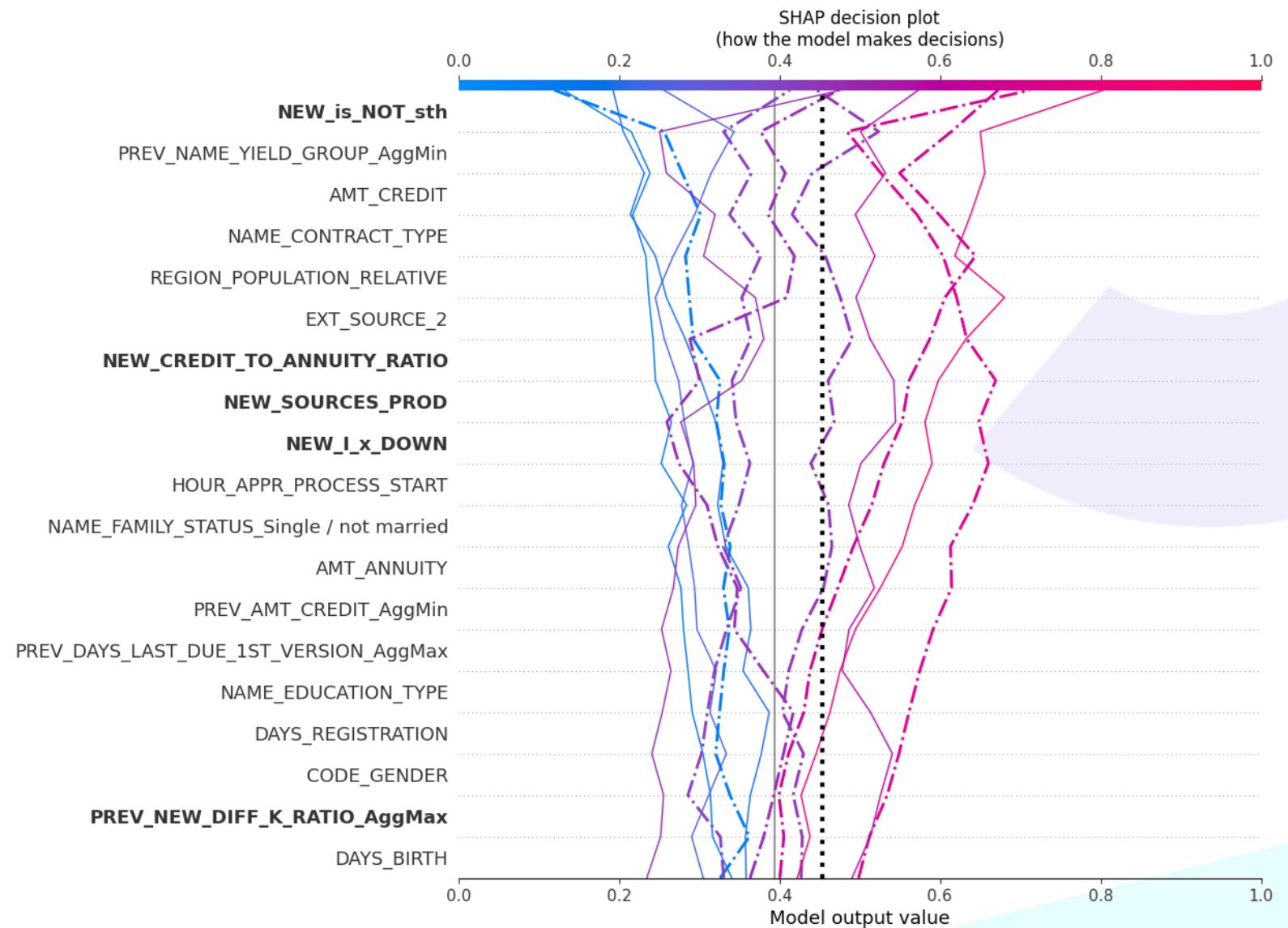
INTERPRÉTABILITÉ DU MODÈLE

- SHAP Importance Bar plot :



INTERPRÉTABILITÉ DU MODÈLE

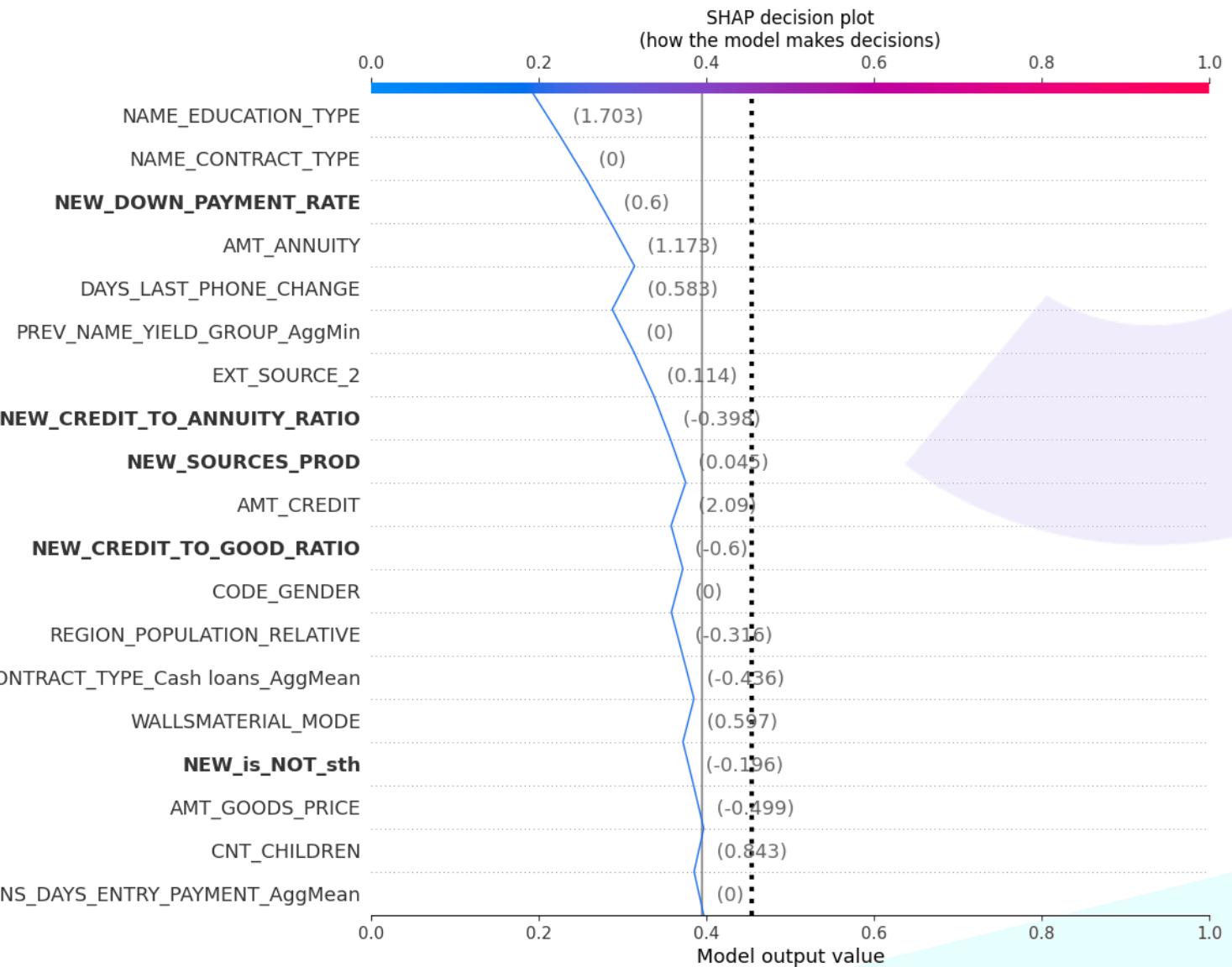
- SHAP Decision plot :



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

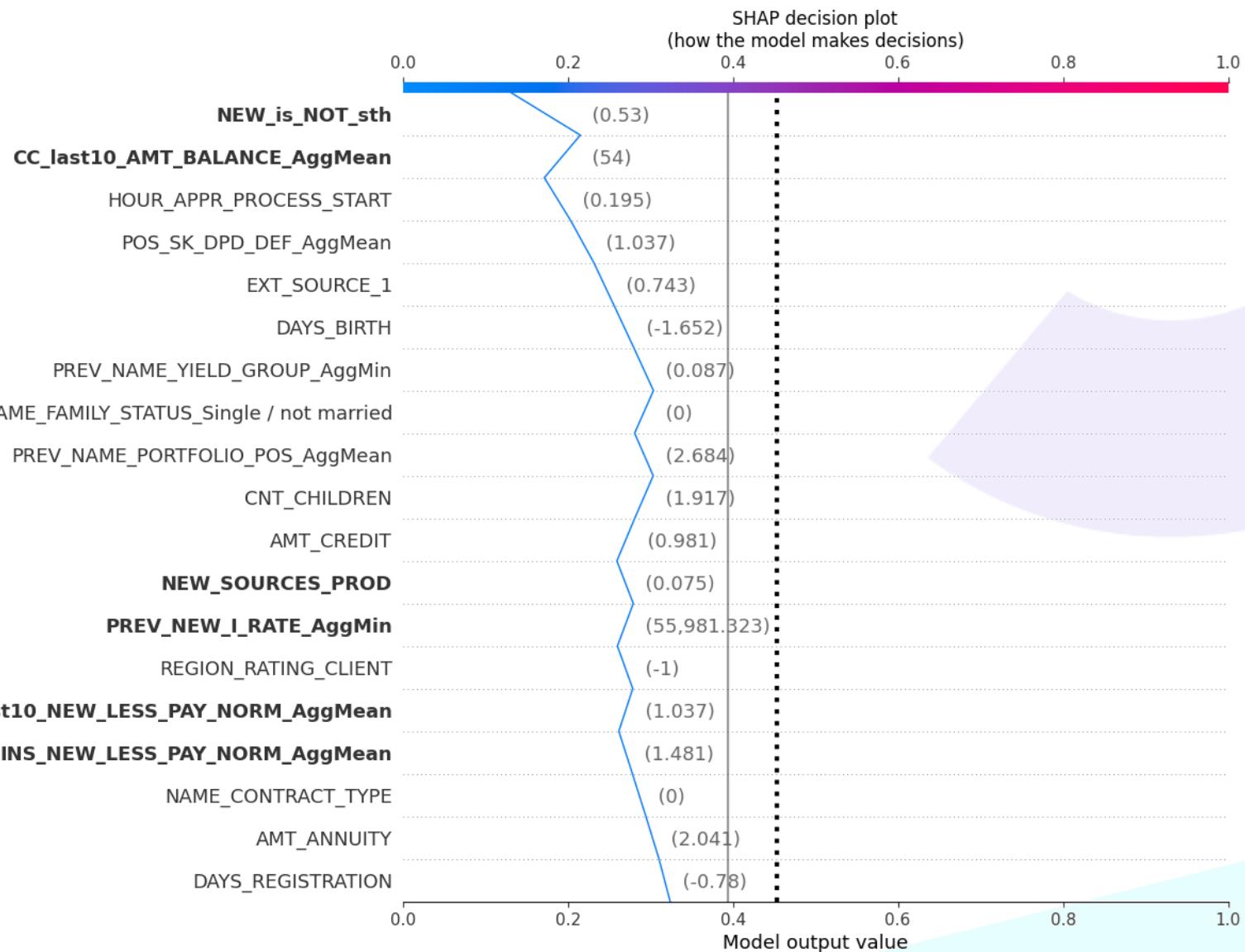
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

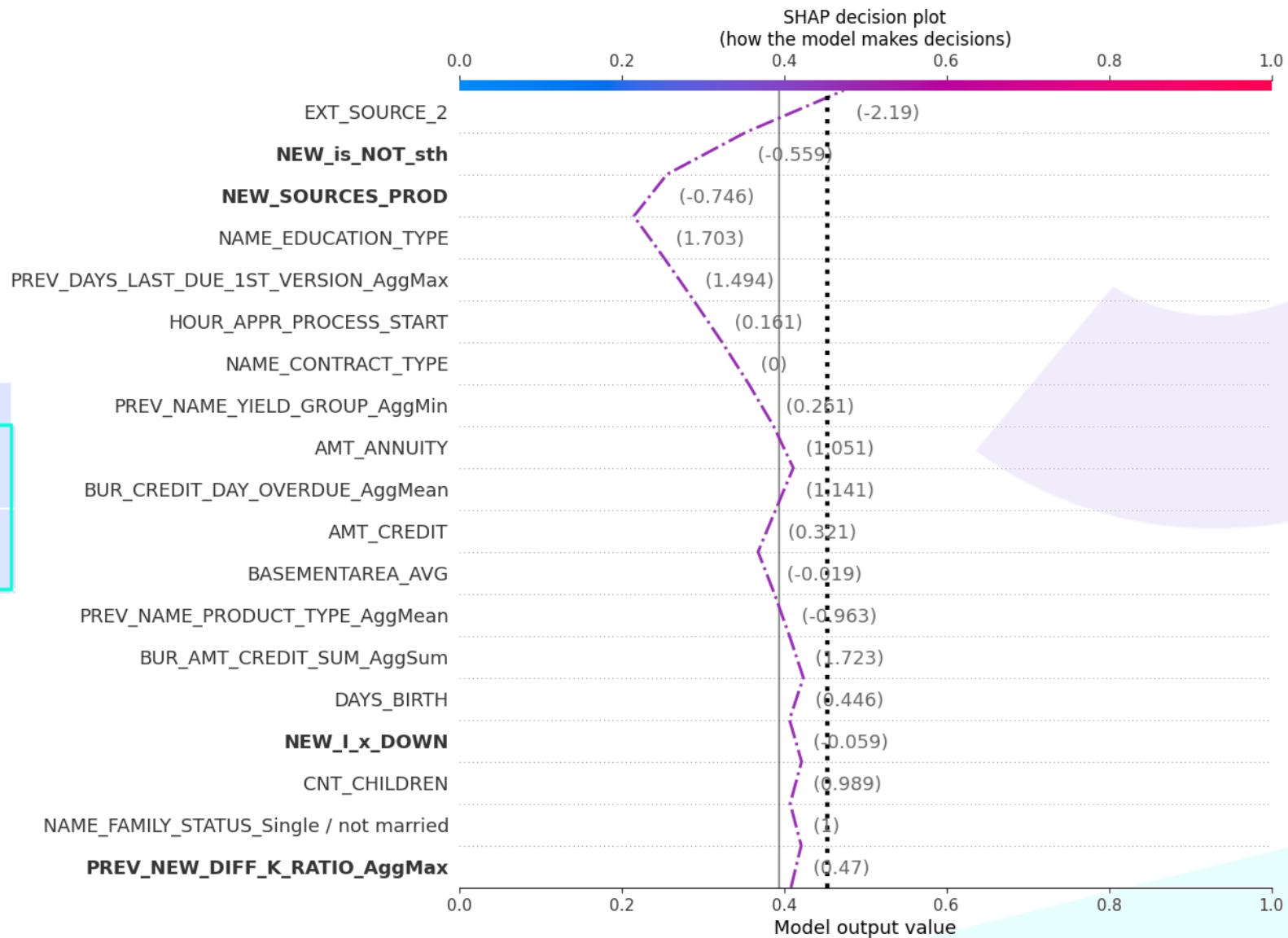
		Predicted Class
		Positive Negative
Actual Class	Positive	True Positive (TP)
	Negative	False Positive (FP) Type I Error
	Positive	False Negative (FN) Type II Error
	Negative	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

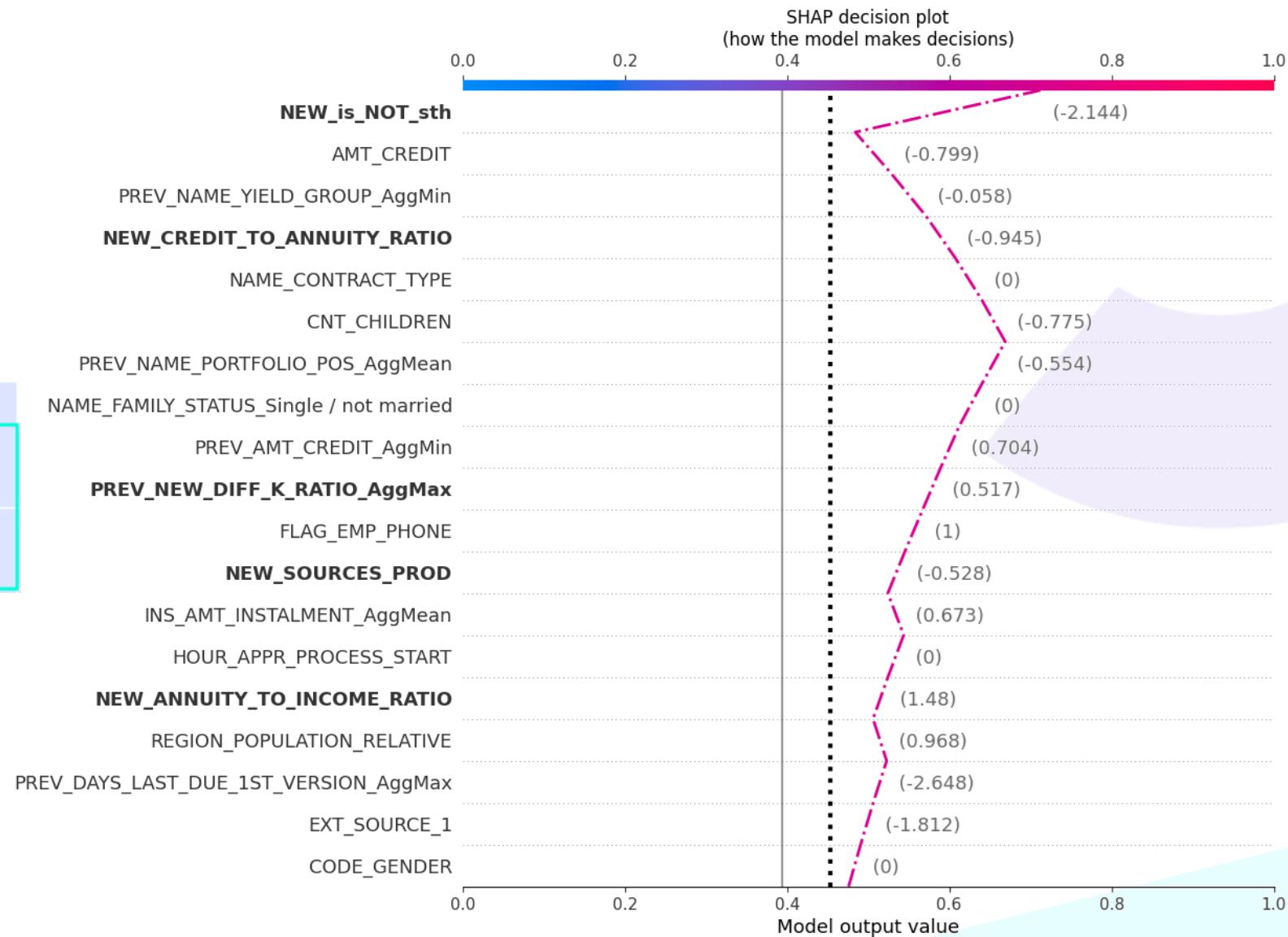
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

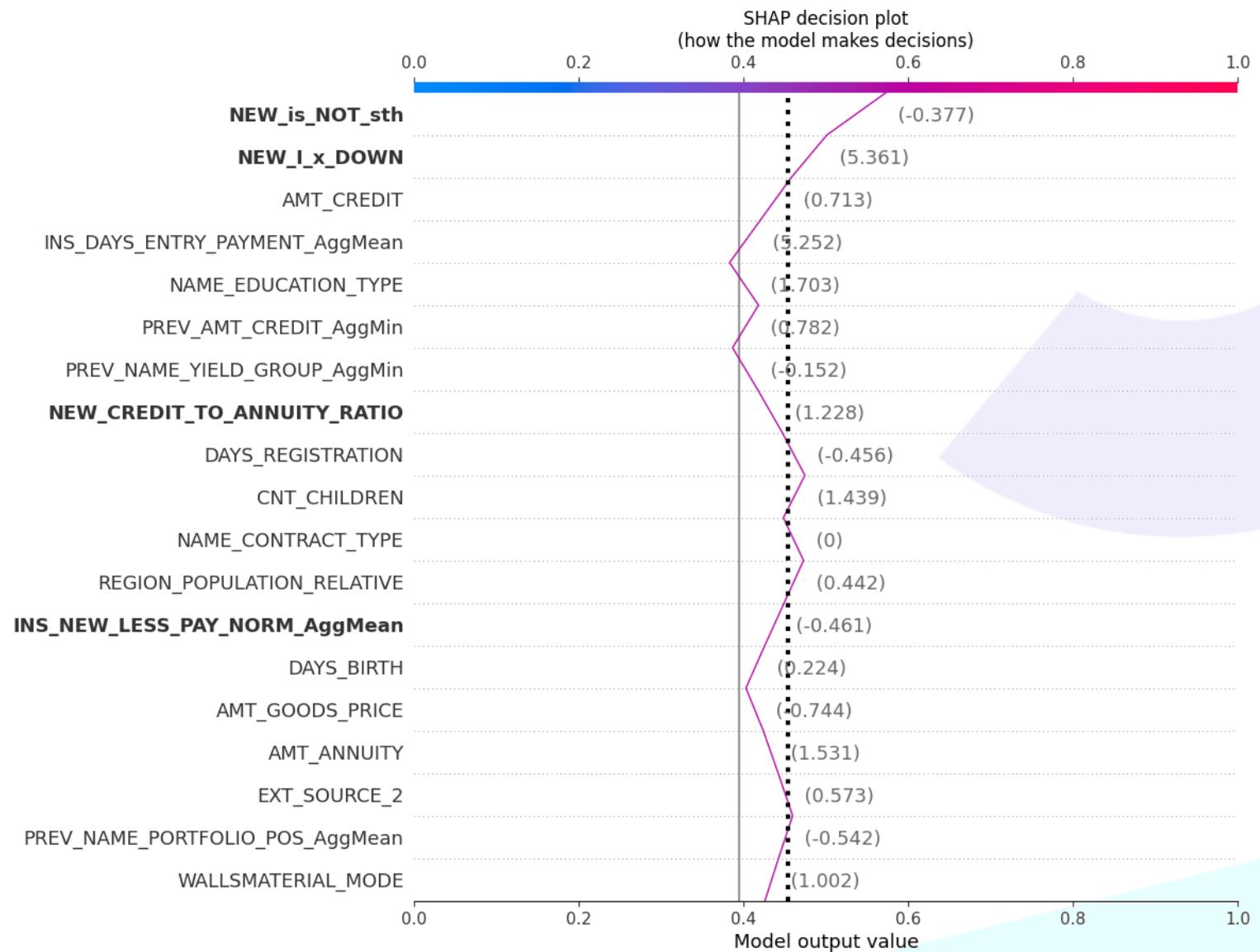
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

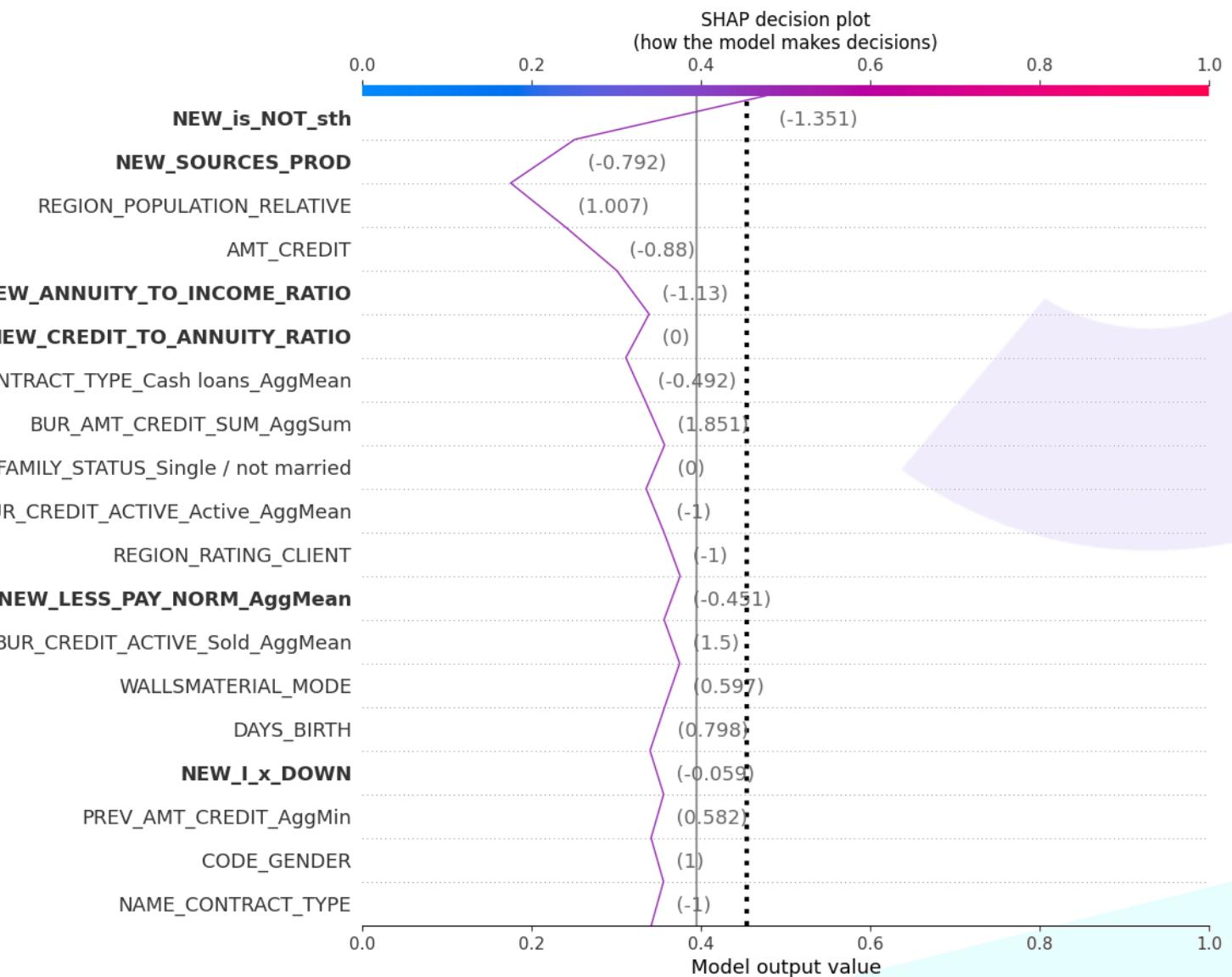
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

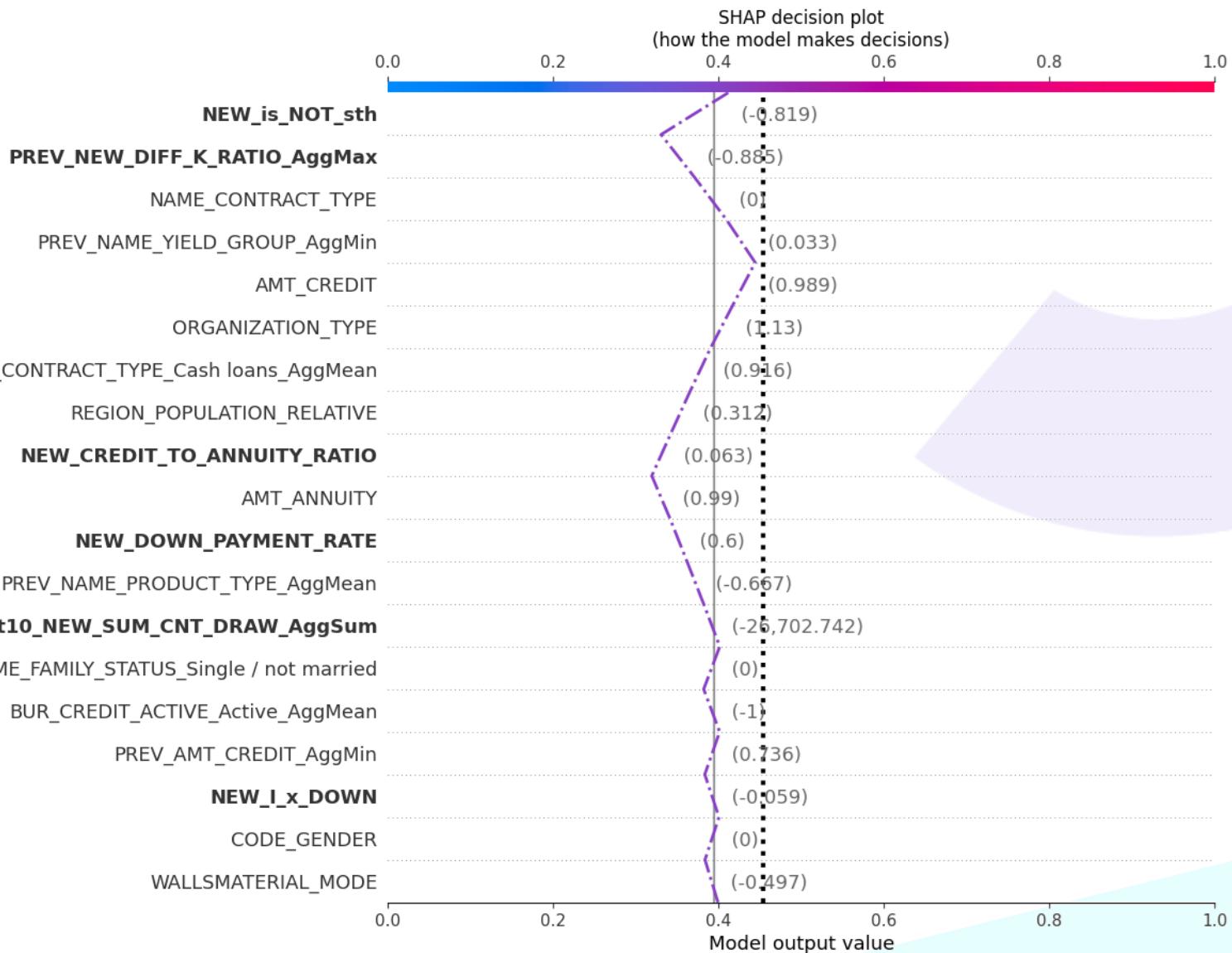
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

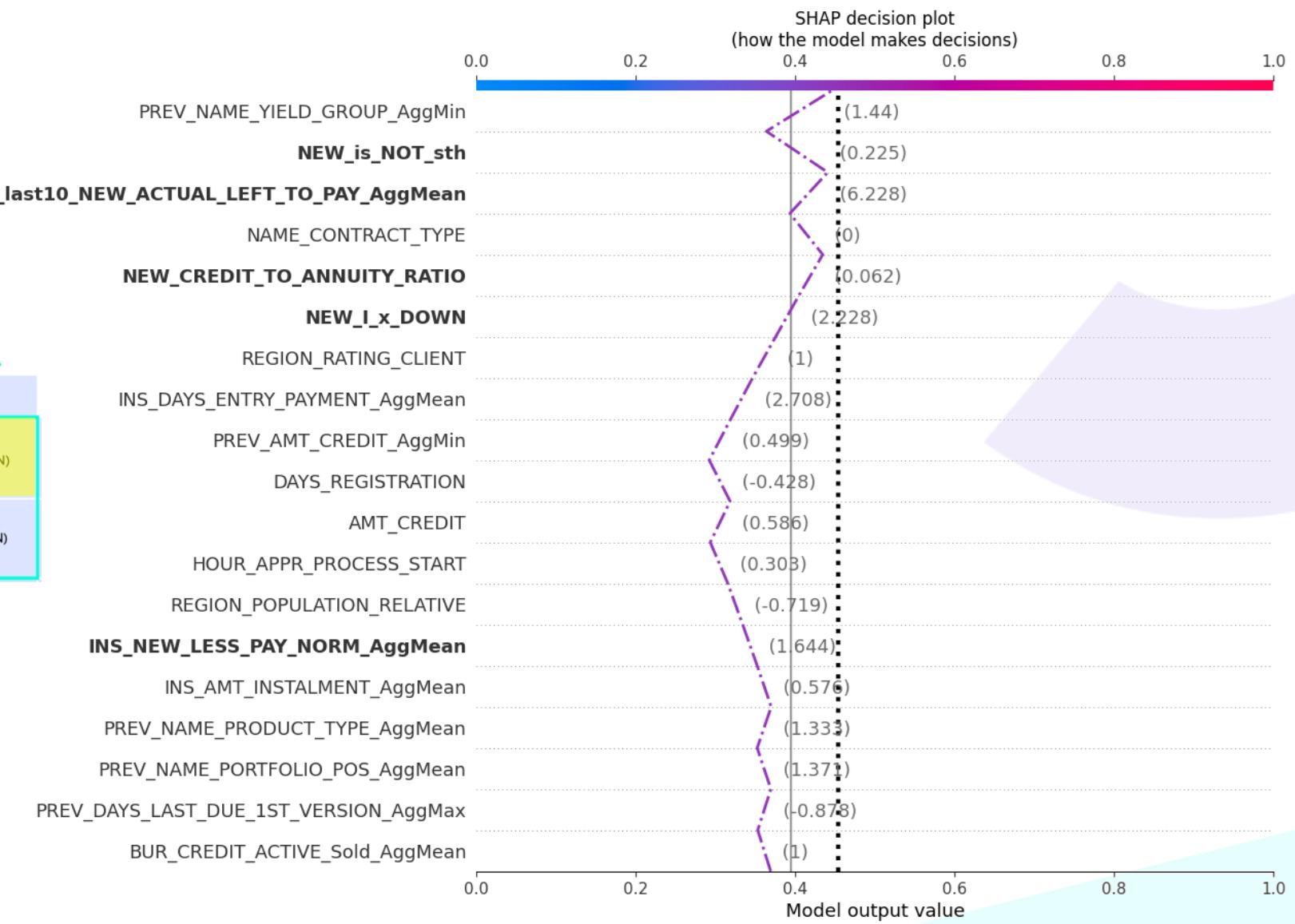
		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



INTERPRÉTABILITÉ DU MODÈLE

- SHAP Decision plot :

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)



A person is sitting at a desk, writing in a notebook with a pen. The background is slightly blurred.

CONCLUSION

CONCLUSION

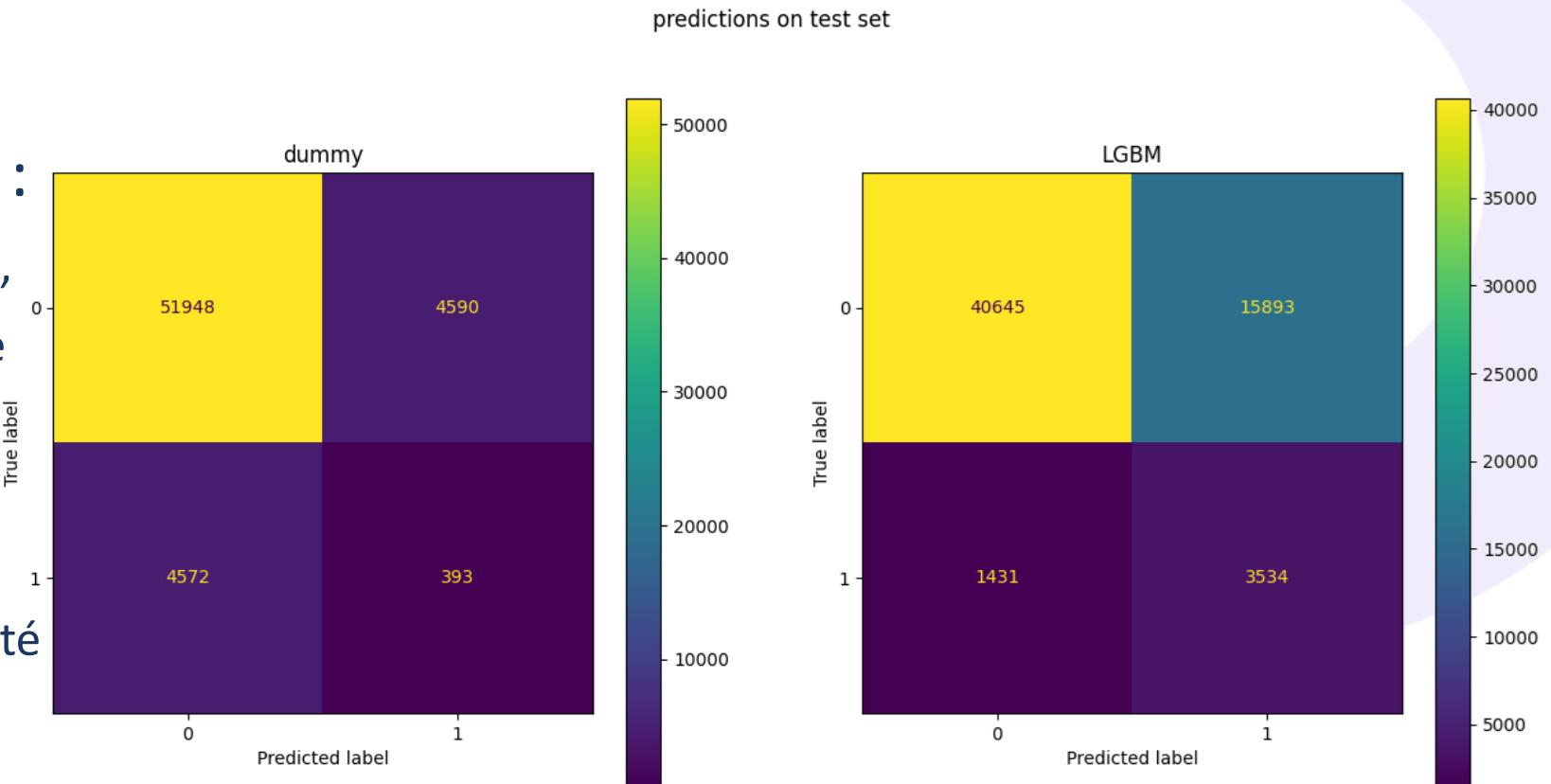
- **Un modèle imparfait :**

Effectivement peu de FN (2,3%),
mais au détriment d'un nombre
accru de FP (26%)

→ rôle conforté des chargés de
relation client 🔎

→ intérêt réel de l'interprétabilité

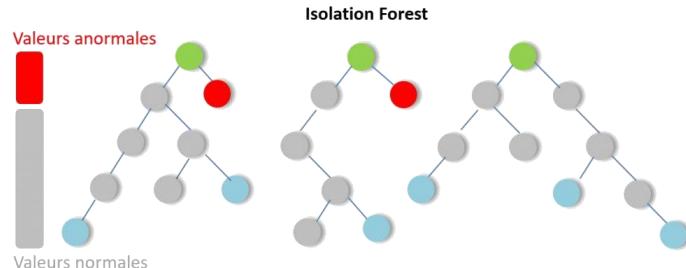
→ ne pas laisser passer trop de
clients



CONCLUSION

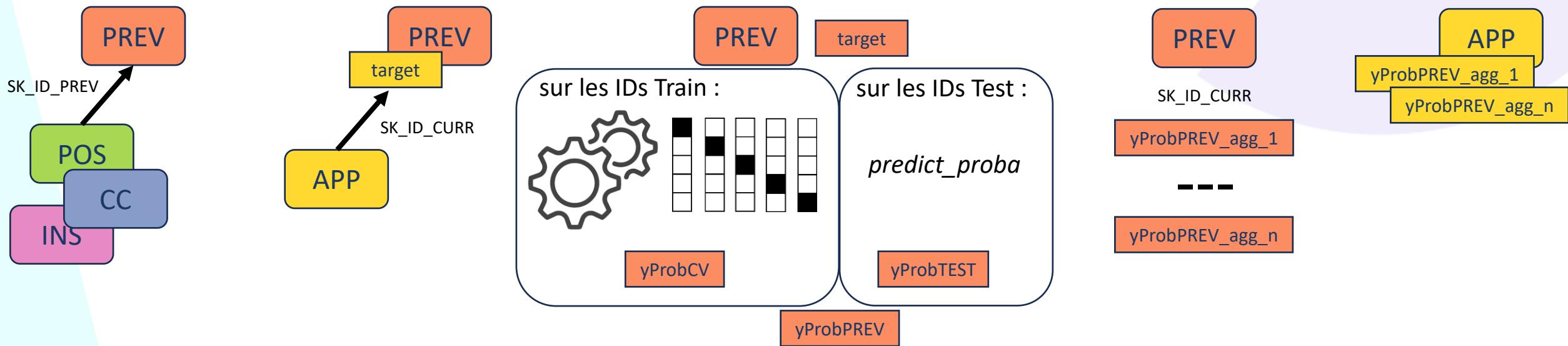
- Perspectives d'évolution ? :

- Donner sa chance au *SMOTE* :
 - plus de temps et plus de calcul → mieux « fine tuner » *k_neighbors* et *sampling_strategy*
 - transformer le pipeline et utiliser *SMOTENC* pour éviter d'utiliser *SMOTE* sur des variables catégorielles encodées...
 - remplacer le *RandomUnderSampling* par un algorithme plus habile comme *l'EditedNearestNeighbours*
- Gérer plus finement les imputations :
 - pour les features numériques → plus de calcul, iterative imputer, knn imputer ?
 - pour les features catégorielles → imputer basé sur le knn
 - pour les features intéressantes (ex : EXT_SOURCE1, 2, 3), développer un modèle de ML ?
- Gérer plus finement les outliers :
 - *FunctionSampler (func = IsolationForest ou autre ?)*



CONCLUSION

- Perspectives d'évolution ? :
 - D'autres features composites ? Nécessiterait du travail, mais intéressant :



A close-up photograph of a person's hands writing in a spiral-bound notebook with a black pen. The person is wearing a light-colored long-sleeved shirt. The background is slightly blurred, showing more of the notebook and some papers on a desk.

merci