Yann LECERF

P4
“Construction d’un modèle de scoring”

Prêt à dépenser

OPENCLASSROOMS / AI Engineer

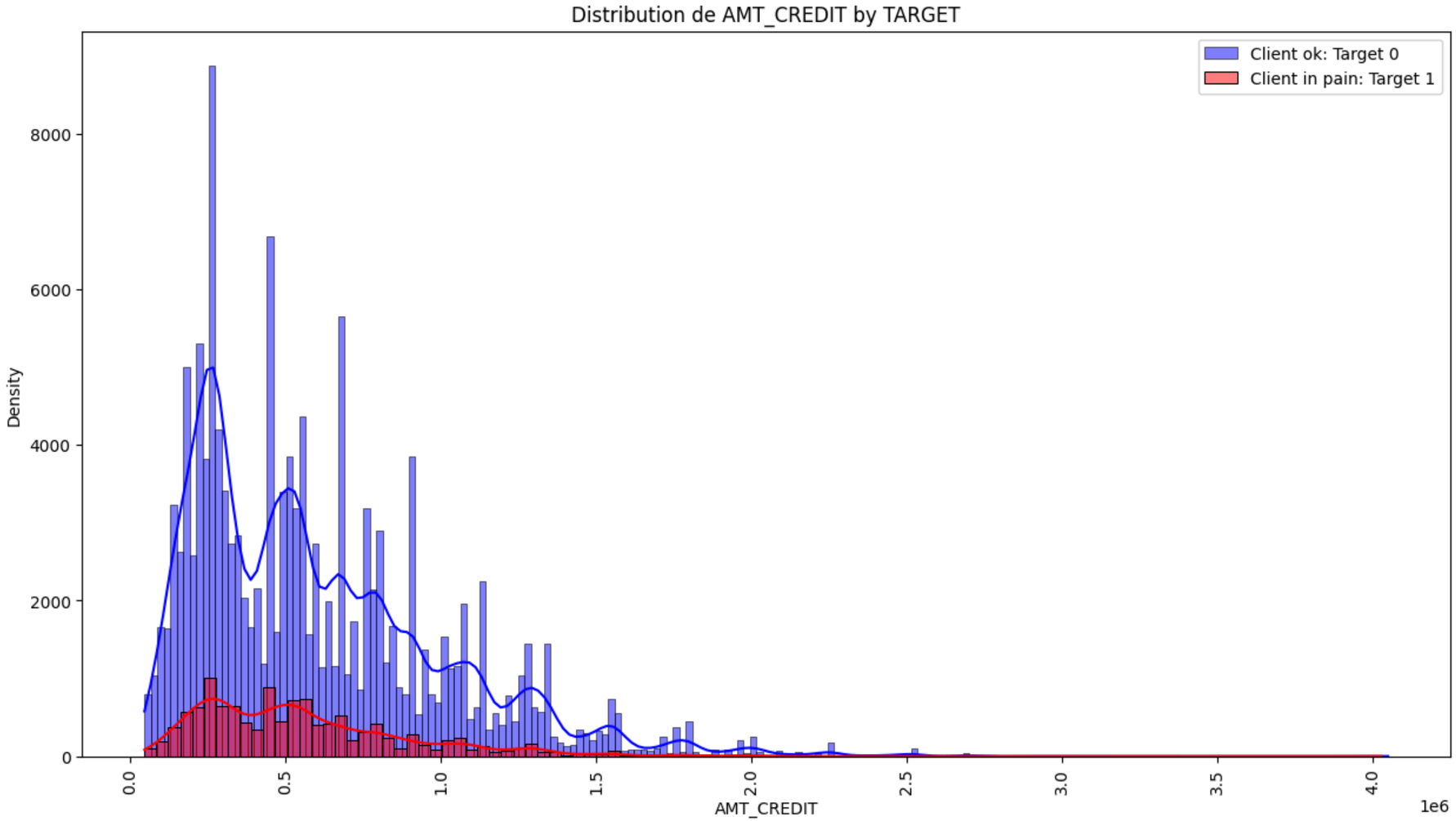
Yann LECERF

Sommaire

Compréhension de la problématique métier

RETOUR SOMMAIRE

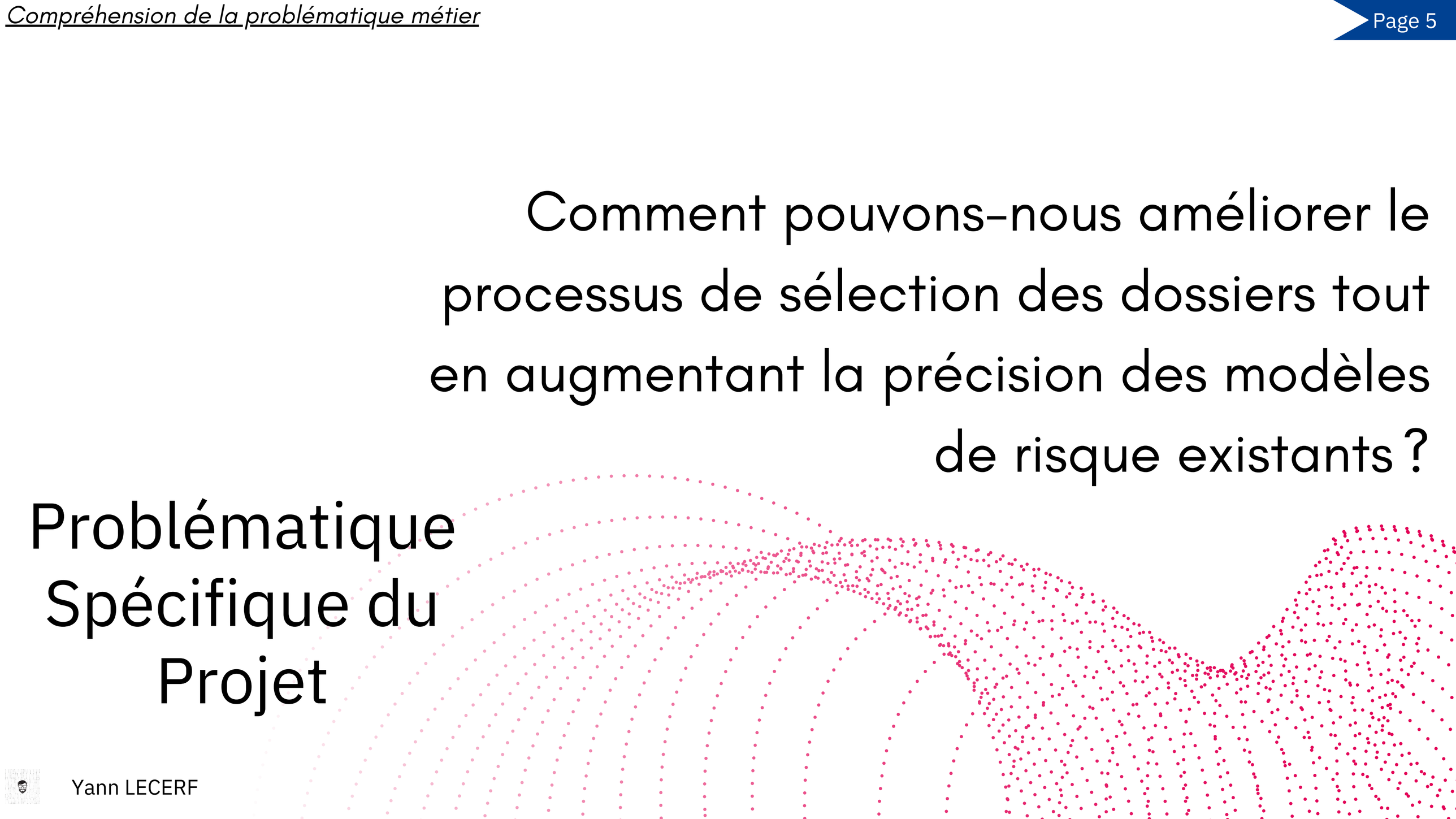
Contexte et Enjeux de la Gestion des Prêts



Le secteur bancaire fait face à des défis constants pour évaluer la solvabilité des clients et réduire le risque de défaut. Un modèle de scoring efficace permet de mieux gérer ce risque.

Problématique Spécifique du Projet

Comment pouvons-nous améliorer le processus de sélection des dossiers tout en augmentant la précision des modèles de risque existants ?



Impact pour les Chargés de Clientèle

Un modèle plus précis aide les chargés de relation client à prendre des décisions plus éclairées, améliorant ainsi la satisfaction client et réduisant les risques financiers.

Description du jeu de données

RETOUR SOMMAIRE

Structure et Provenance des Données

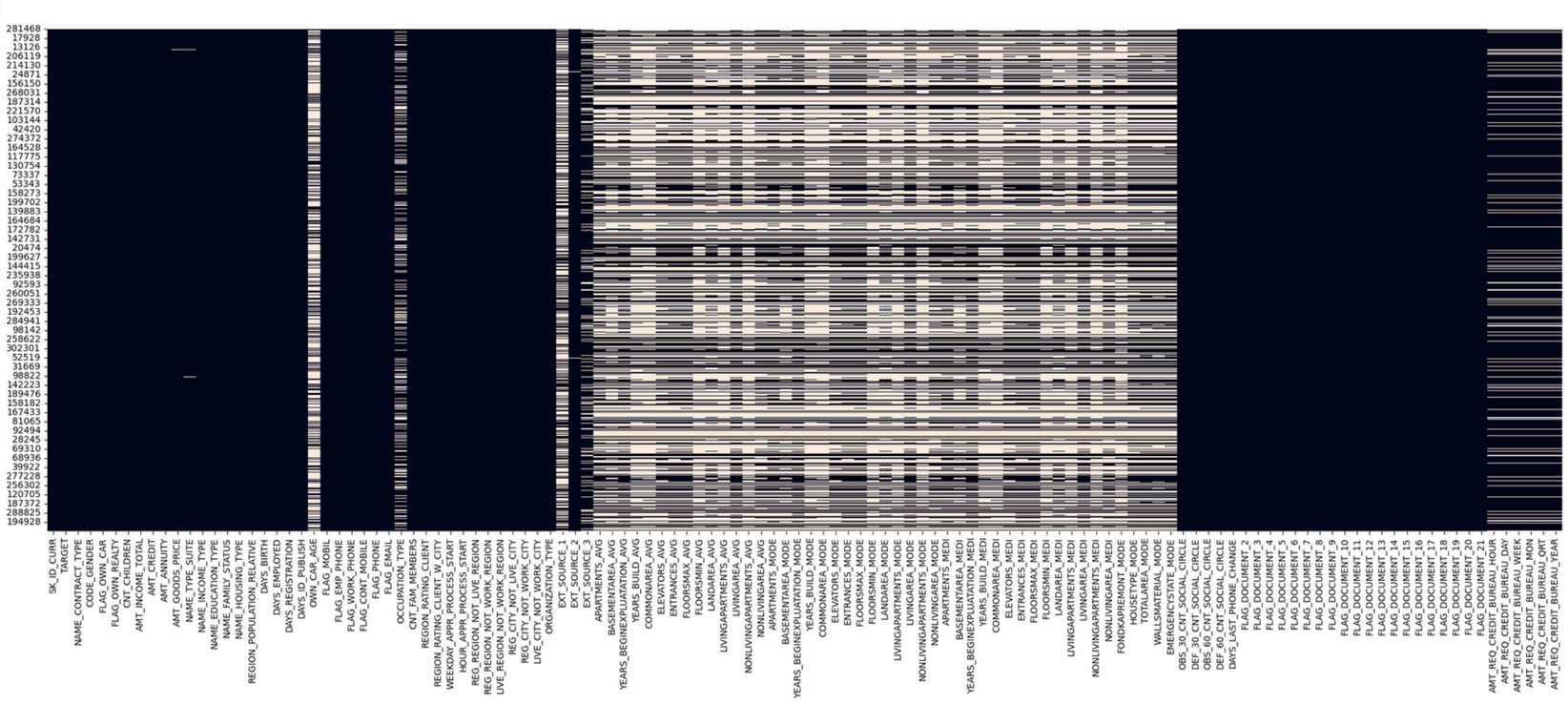
Dataset	nbrs_of_columns	columns_name	nbrs_of_rows
app_train	122	SK_ID_CURR, TARGET, NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, REGION_POPULATION_RELATIVE, DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, OWN_CAR_AGE, FLAG_MOBIL, FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL, OCCUPATION_TYPE, CNT_FAM_MEMBERS, REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY, WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START, REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY, ORGANIZATION_TYPE, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG, YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE, LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE, NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI, ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, DAYS_LAST_PHONE_CHANGE, FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12, FLAG_DOCUMENT_13, FLAG_DOCUMENT_14, FLAG_DOCUMENT_15, FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR	307511

Le dataset provient de sources financières et contient des détails sur les emprunteurs, tels que leurs caractéristiques démographiques, leurs revenus et leurs antécédents de crédit.

RETOUR SOMMAIRE

Analyse des Valeurs Manquantes

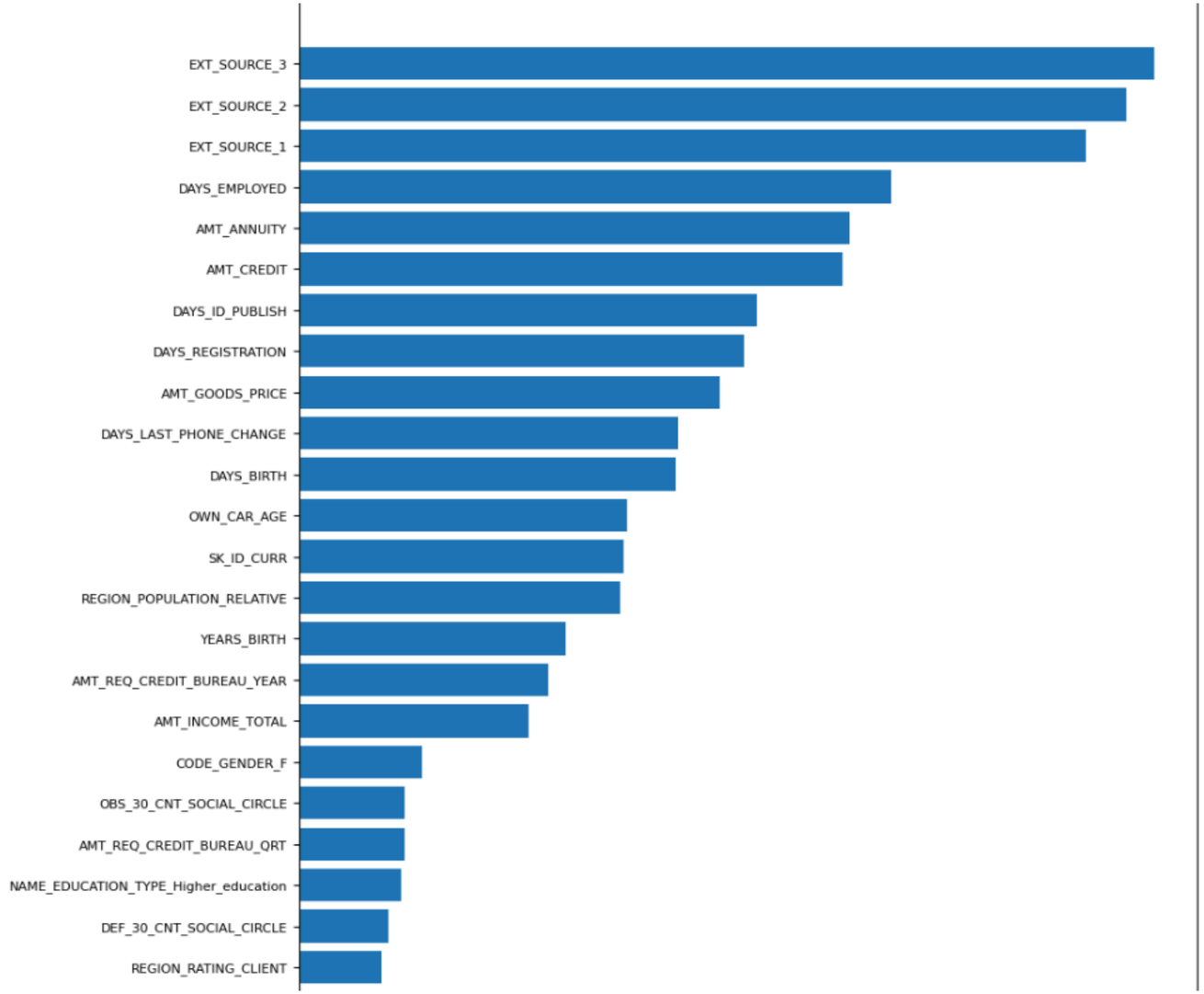
Un aspect clé de la préparation des données est la gestion des valeurs manquantes. Nous avons constaté que certaines variables avaient jusqu'à 60 % de valeurs manquantes.



Yann LECERF

Analyse des Variables Clés

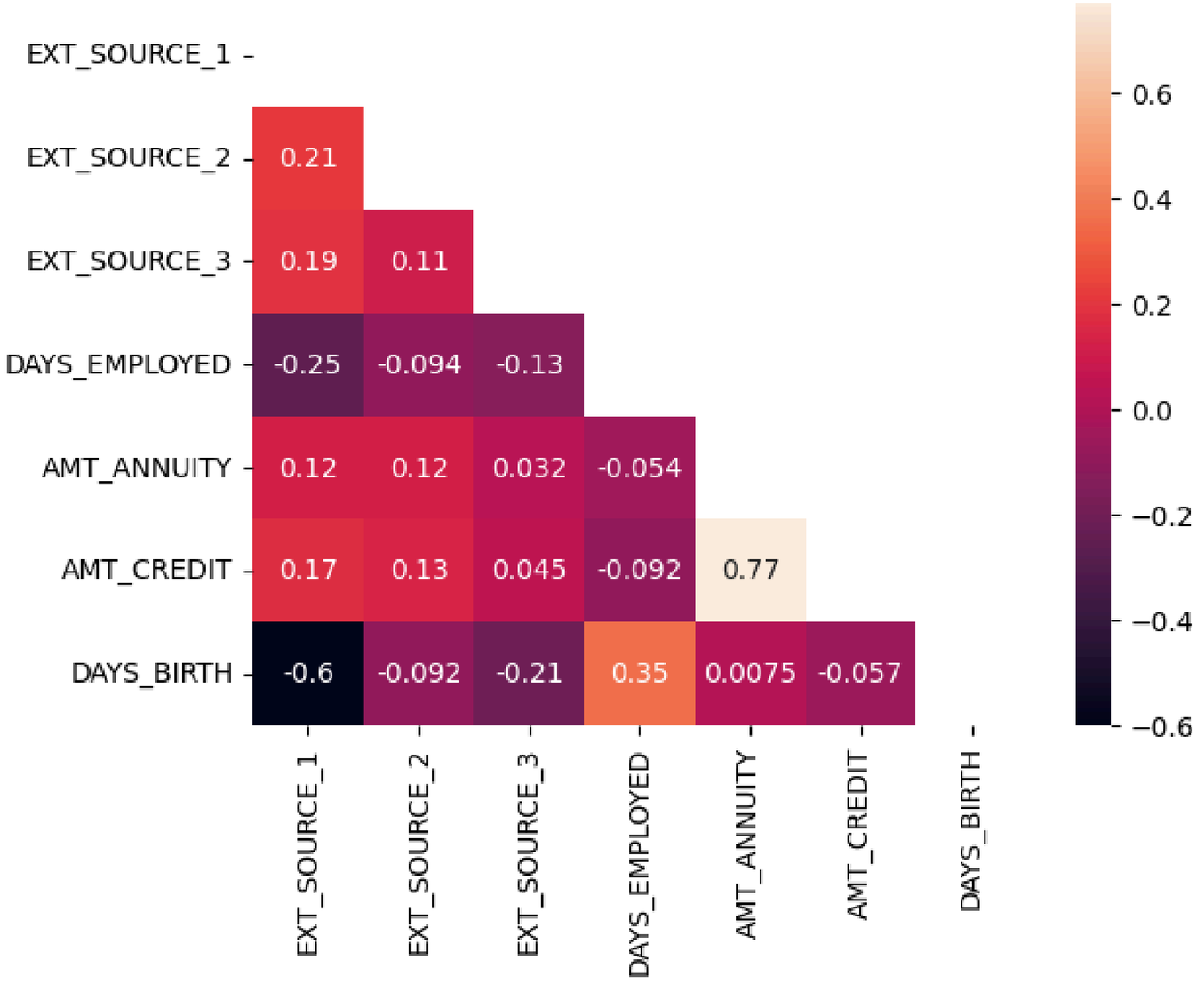
Certaines variables, ont montré une forte corrélation entre elle.



Yann LECERF

Analyse des Variables Clés

Certaines variables, ont montré une forte corrélation entre elle.



Analyse des Variables Clés

- **Taux d’endettement**

  “DEBT RATE”

- **Reste à vivre**

  “KEEP FOR LIVING”

- **Revenu par nombre de personne dans le foyer**

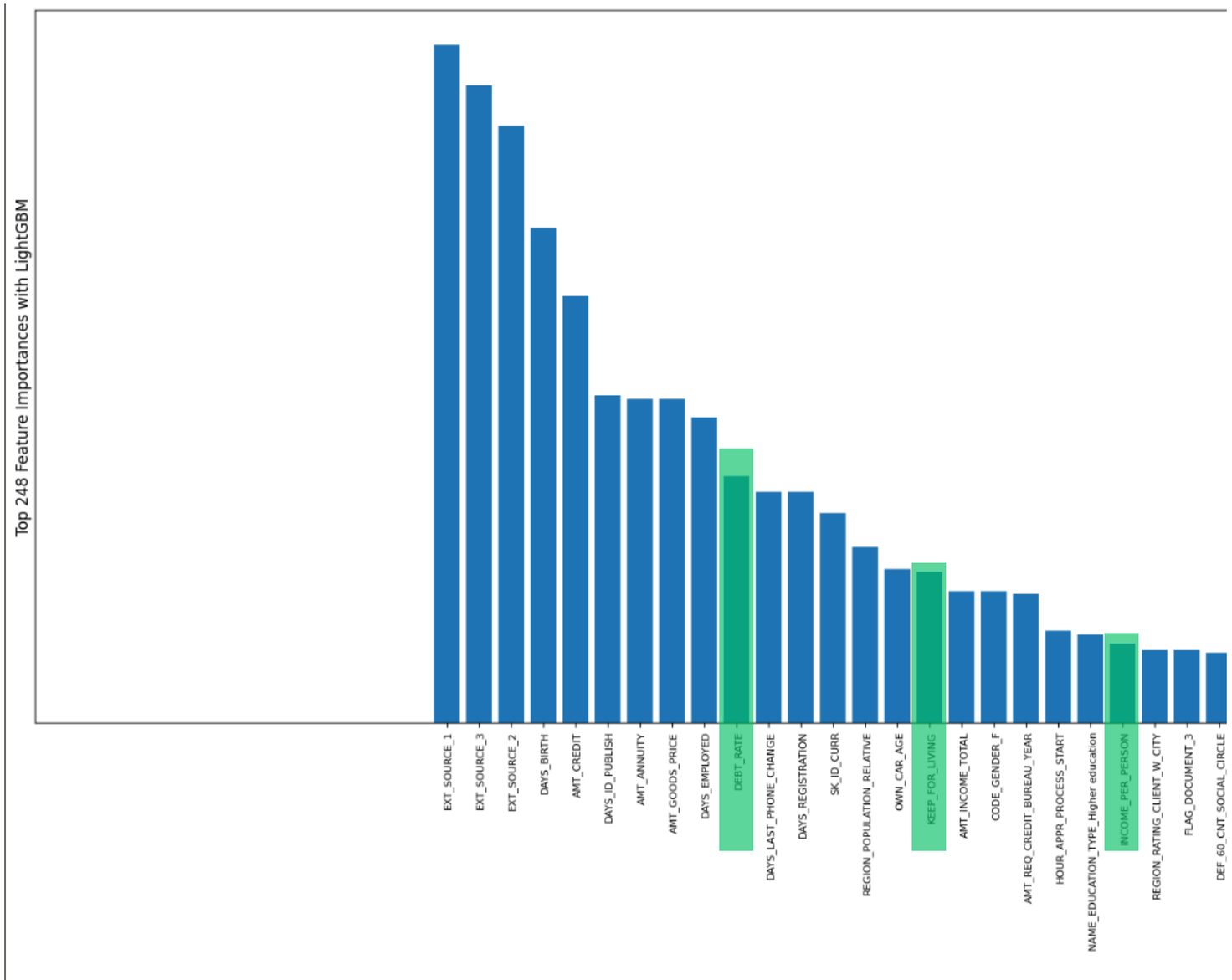
  “”INCOME PER PERSON”

Transformation du jeu de données (nettoyage et feature engineering)

RETOUR SOMMAIRE

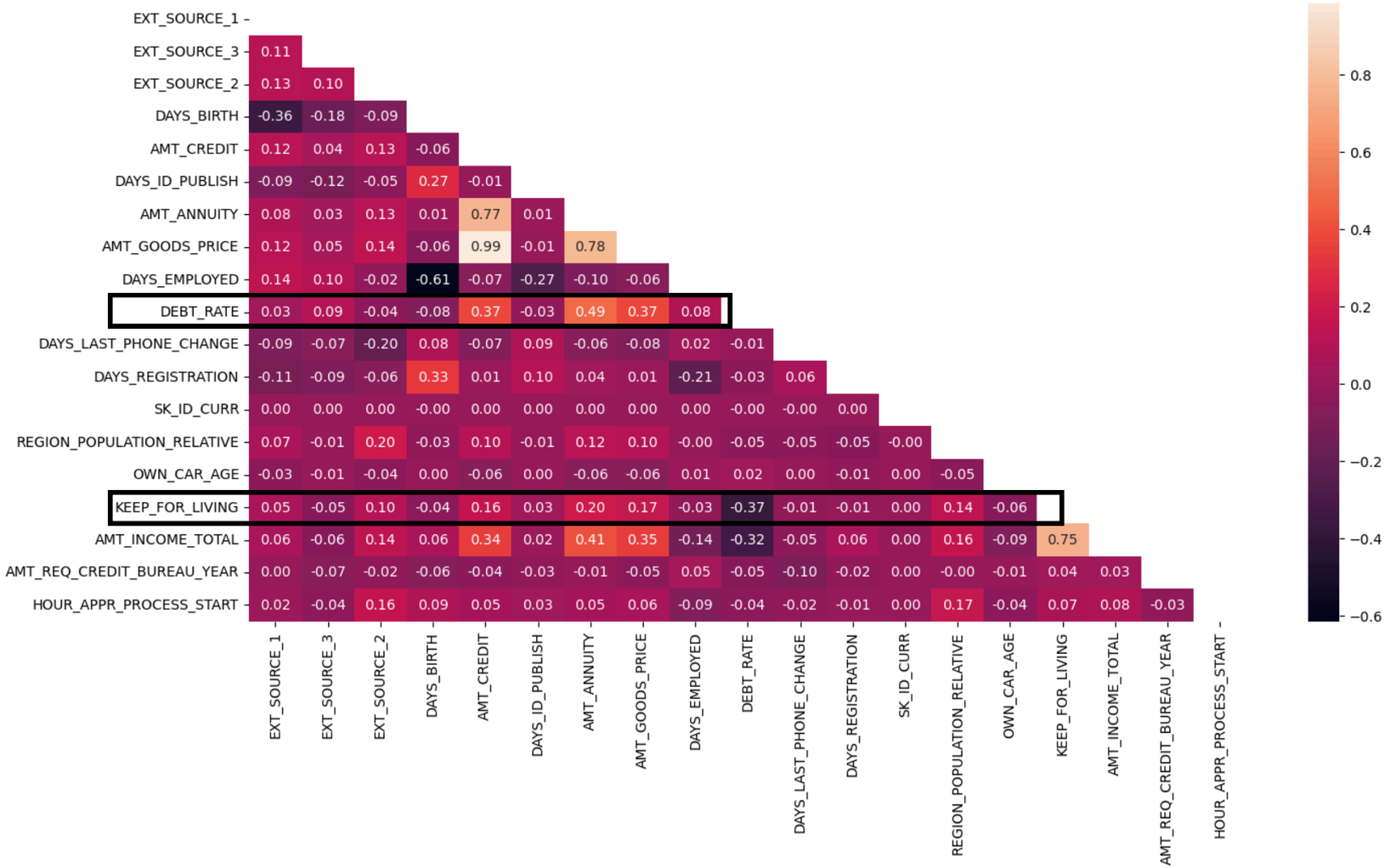
Sélection de Features de Base et personnelles

Features importances



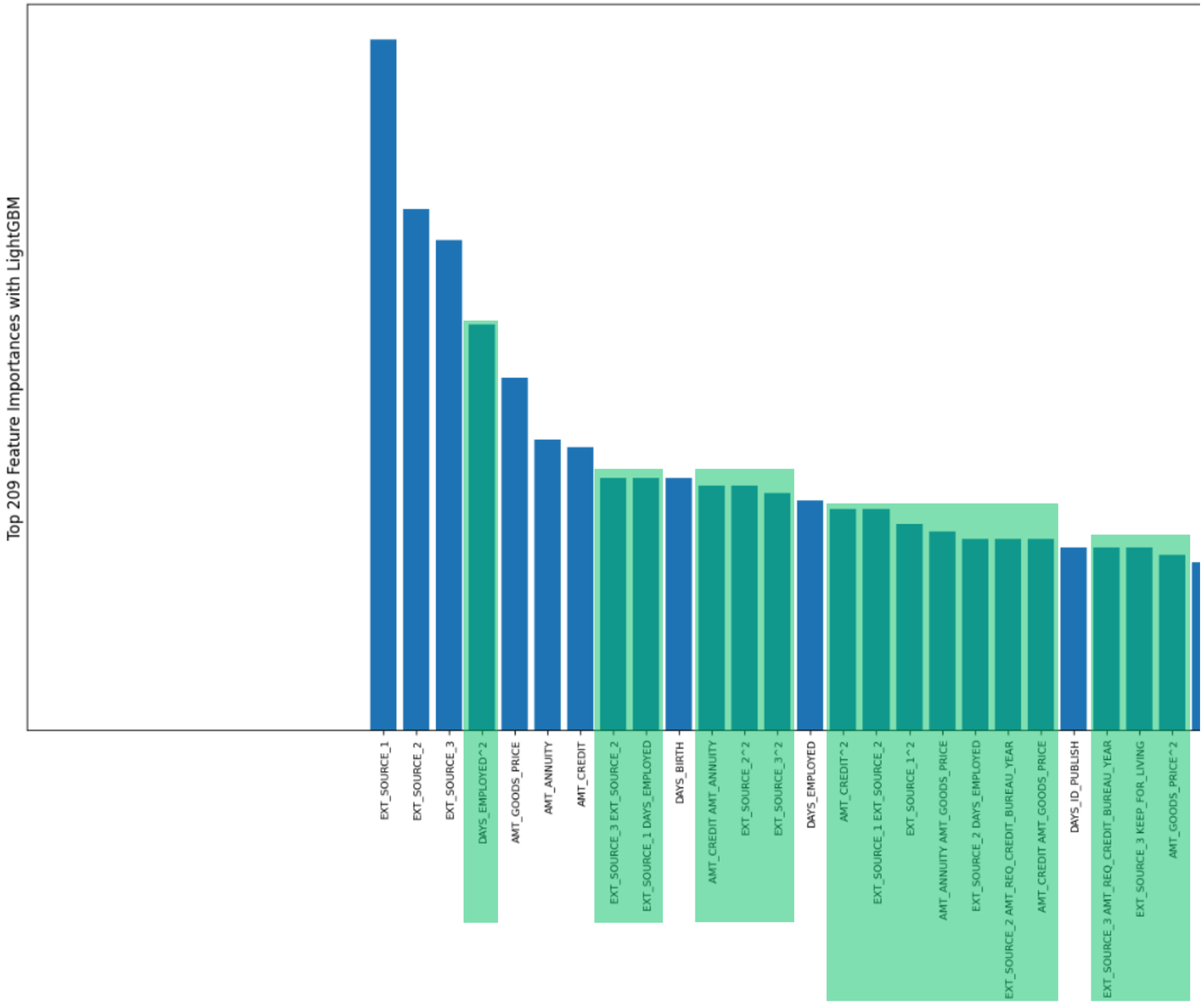
Yann LECERF

Sélection avec Variables Personnalisées



Sélection avec Features Polynomiales

Extrait de polynomiales features

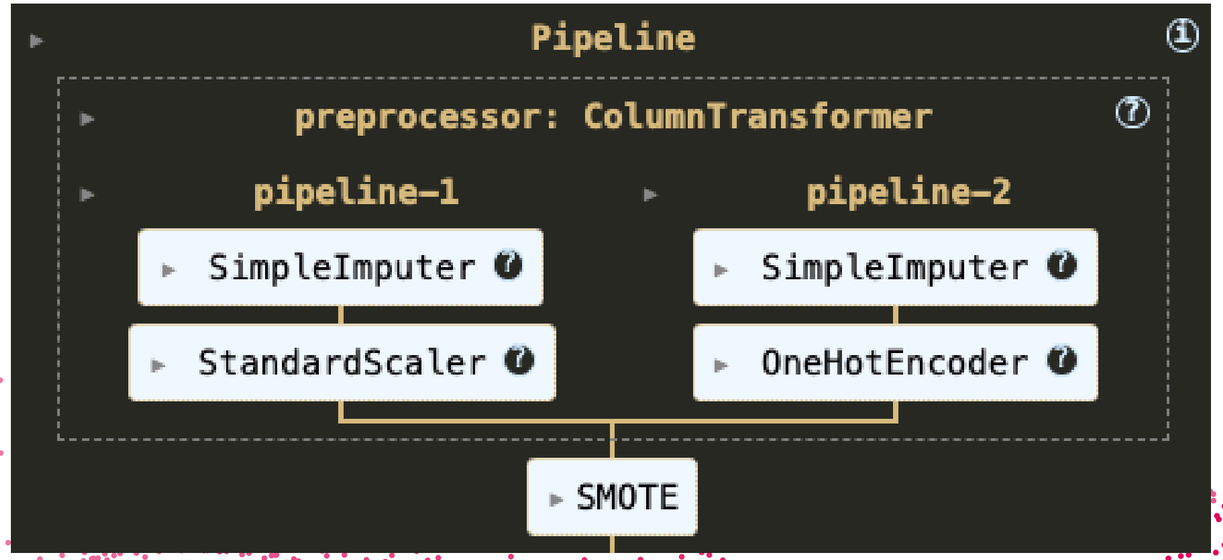


Comparaison et synthèse des résultats pour les modèles utilisés.

RETOUR SOMMAIRE

La création du modèle a été pensée pour refléter les besoins métiers, avec une approche axée sur la précision et l'interprétation des résultats.

Conception du Modèle Alignée à la Fonction Métier



Comparaison des différents modèles avec les paramètres par defauts.

	Modéle	Moyenne AUC	Moyenne Business Cost	Temps d'exécution (s)
0	RandomForest	0.700935	29872.6	2792.452951
1	LogisticRegression	0.723513	22784.0	33.592061
2	LGBM	0.722291	30996.6	41.267289
3	XGBOOST	0.712043	30506.0	33.870847
4	DummyClassifier	0.500000	31776.0	12.947418

Choix des Algorithmes et Paramètres

Comparaison et synthèse des résultats pour les modèles utilisés.

```
Meilleurs paramètres pour LogisticRegression : {'classifier__C': 0.01, 'classifier__solver': 'saga'}
Score moyen pour LogisticRegression : 21934.2

ROC AUC du meilleur modèle : 0.7373384704904477
```

```
LogisticRegression
LogisticRegression(C=0.01, random_state=42, solver='saga')
```

Choix des Algorithmes et Paramètres

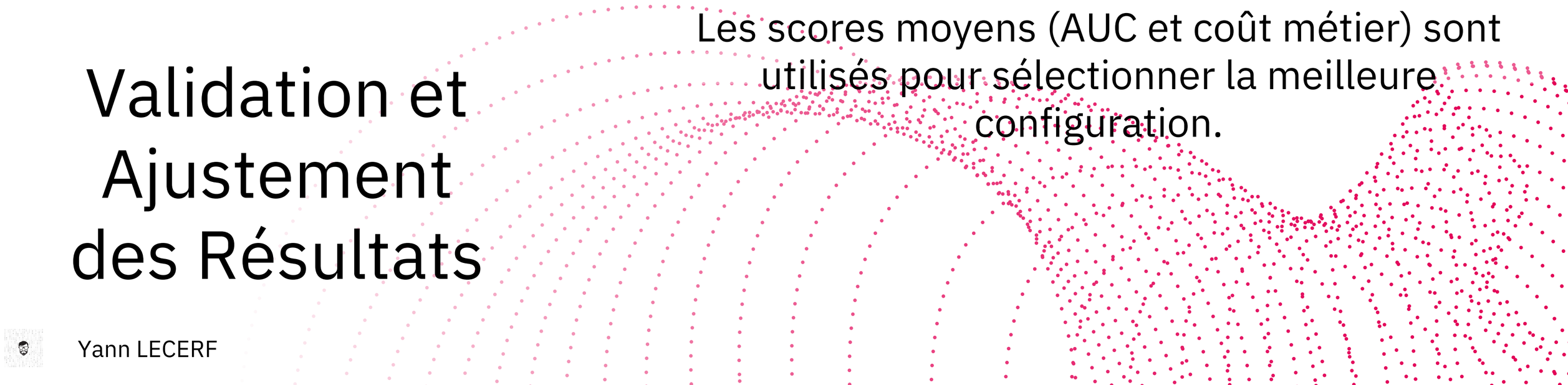
Yann LECERF

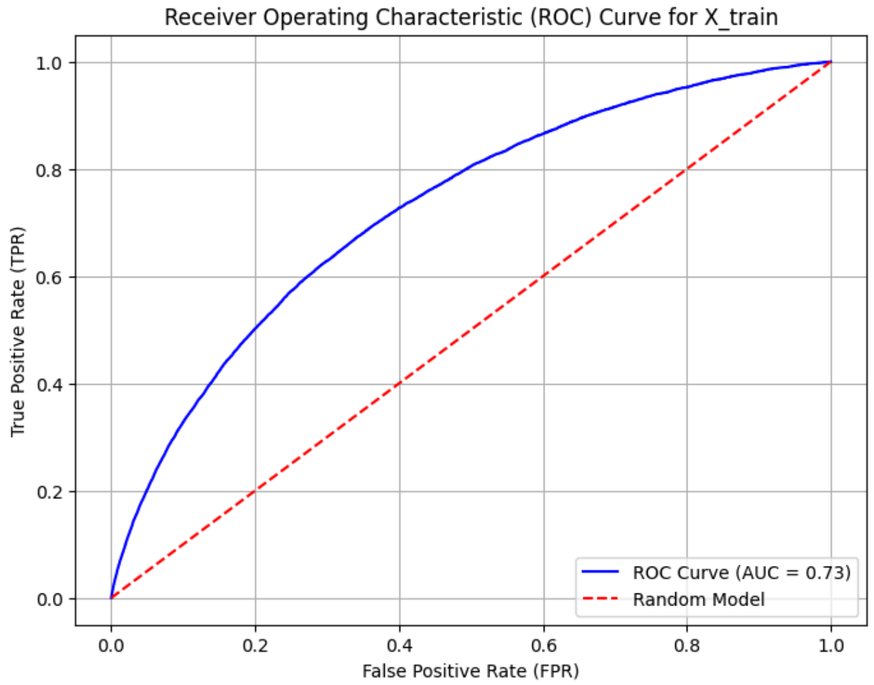
Validation et Ajustement des Résultats

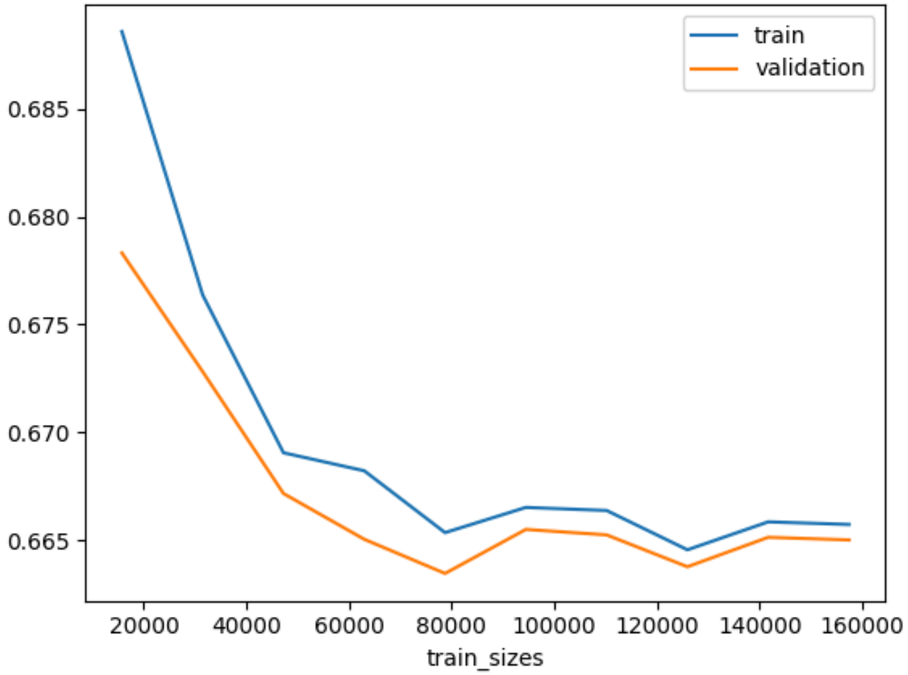
GridSearchCV est une technique d’optimisation des modèles qui permet de tester de manière exhaustive toutes les combinaisons possibles des hyperparamètres spécifiés.

Chaque combinaison est évaluée à l’aide de la validation croisée stratifiée pour assurer la robustesse du modèle.

Les scores moyens (AUC et coût métier) sont utilisés pour sélectionner la meilleure configuration.

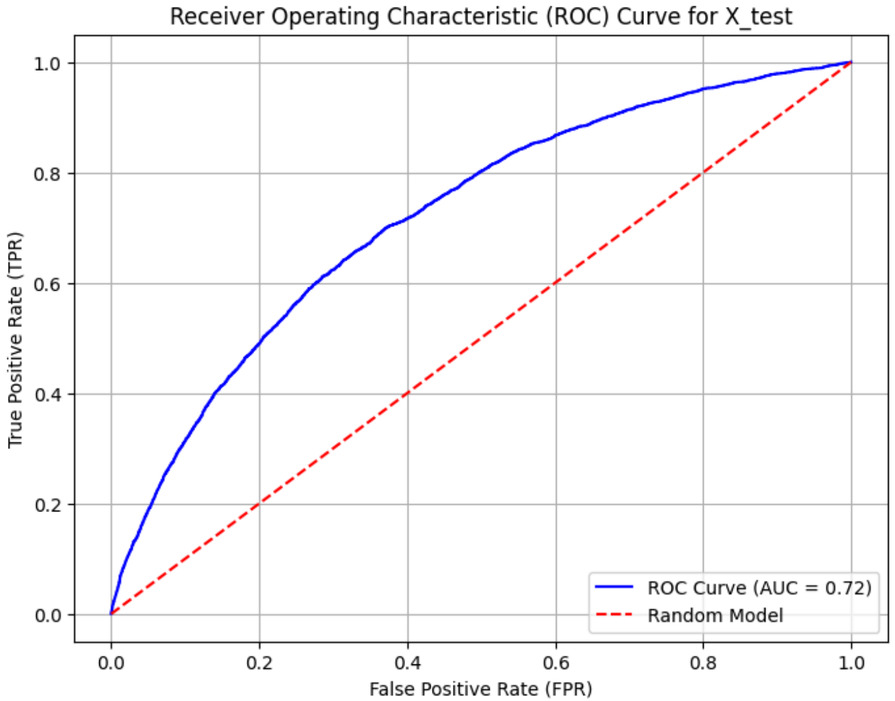


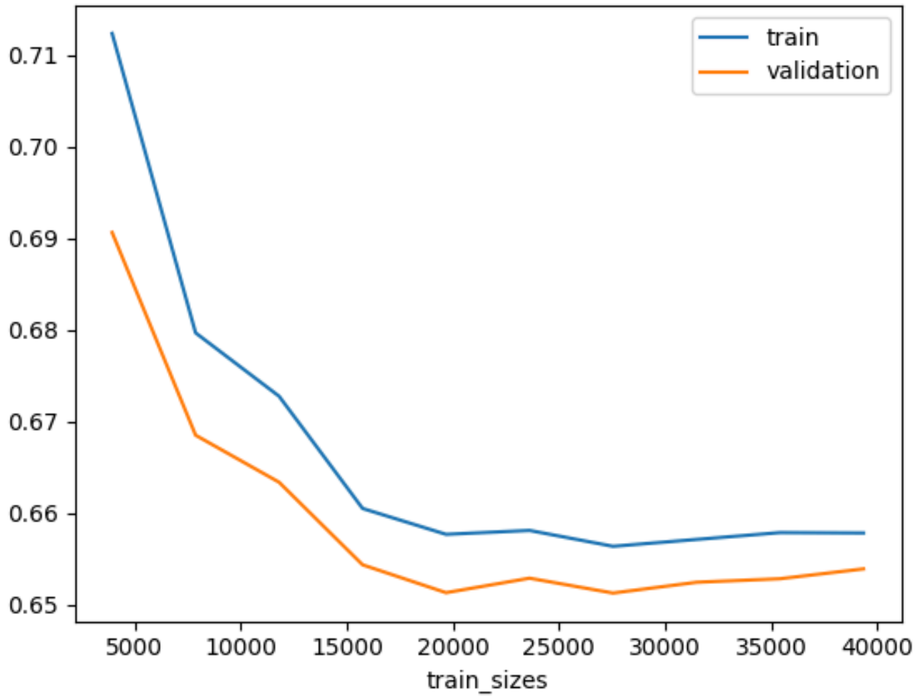




Validation et Ajustement des Résultats

Yann LECERF

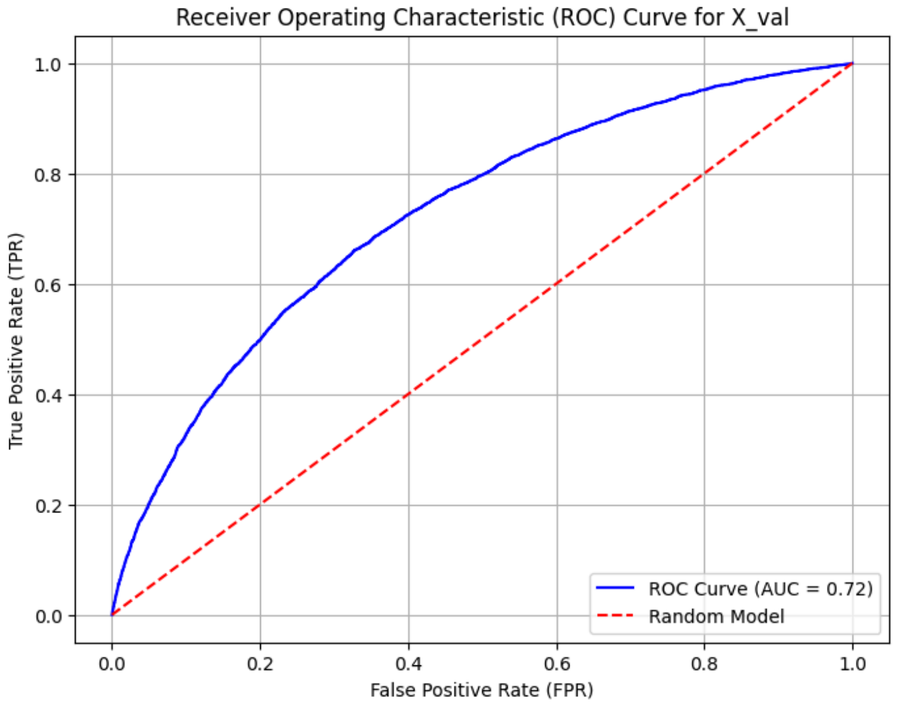


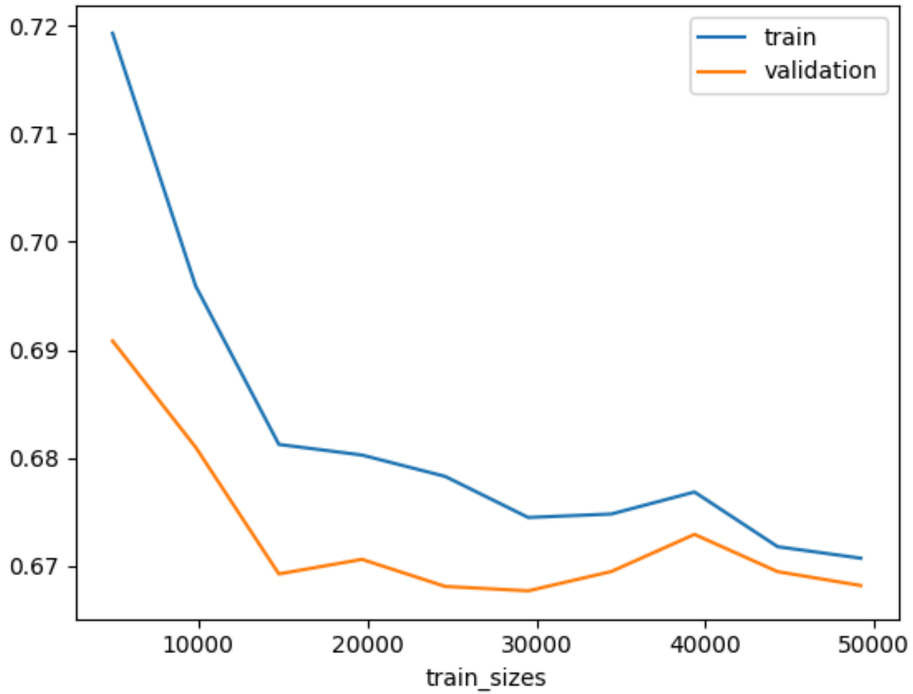


Validation et Ajustement des Résultats

Yann LECERF

Comparaison et synthèse des résultats pour les modèles utilisés.





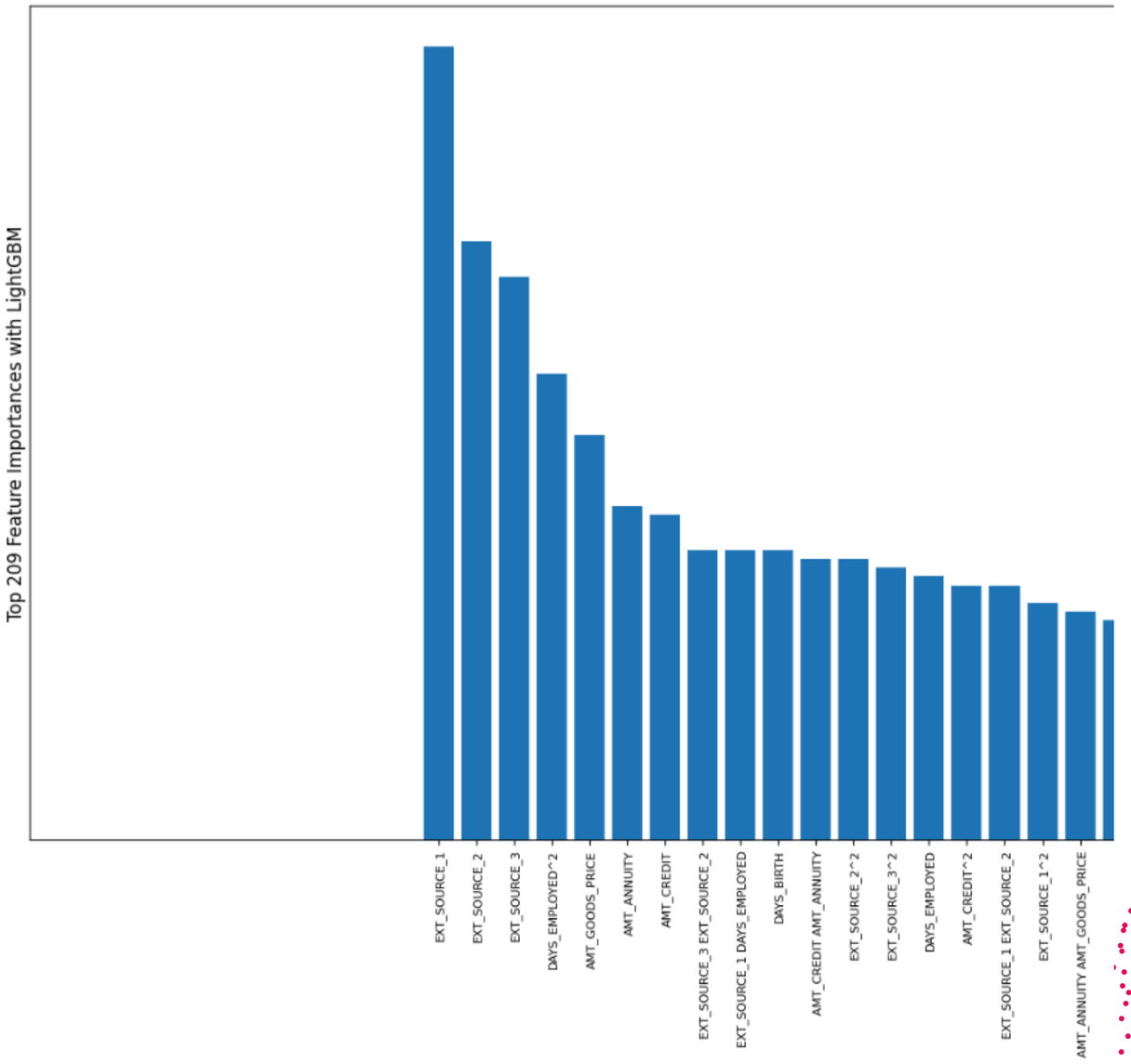
Validation et Ajustement des Résultats

Yann LECERF

Interprétabilité du modèle.

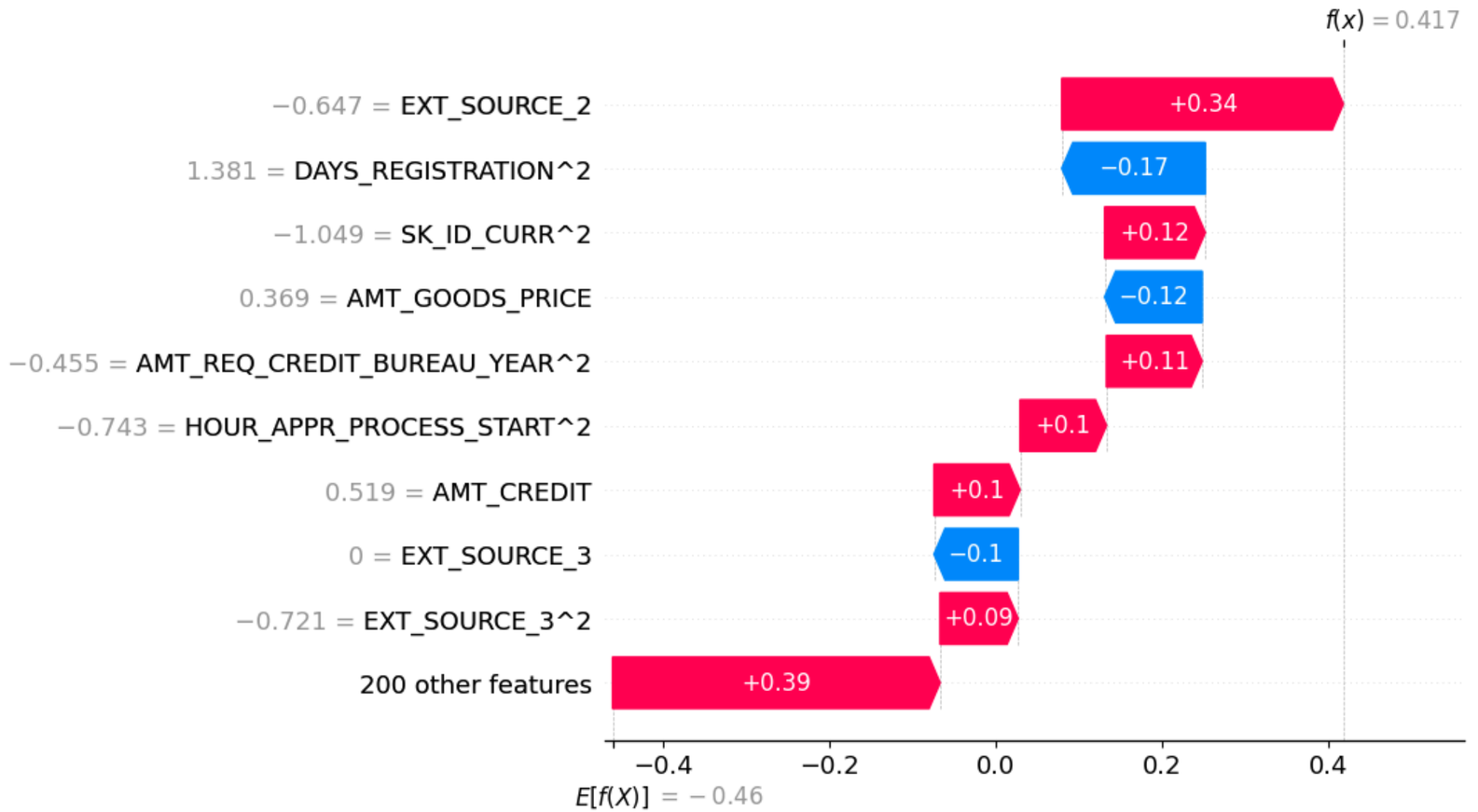
RETOUR SOMMAIRE

Pour comprendre comment le modèle prend ses décisions, nous avons analysé l’importance des variables. Certaines caractéristiques comme l’âge et le revenu se sont révélées cruciales.



Importance des Variables

L’approche SHAP (SHapley Additive exPlanations) a été utilisée pour évaluer l’impact de chaque variable sur les prédictions.



Analyse SHAP pour l’Interprétabilité

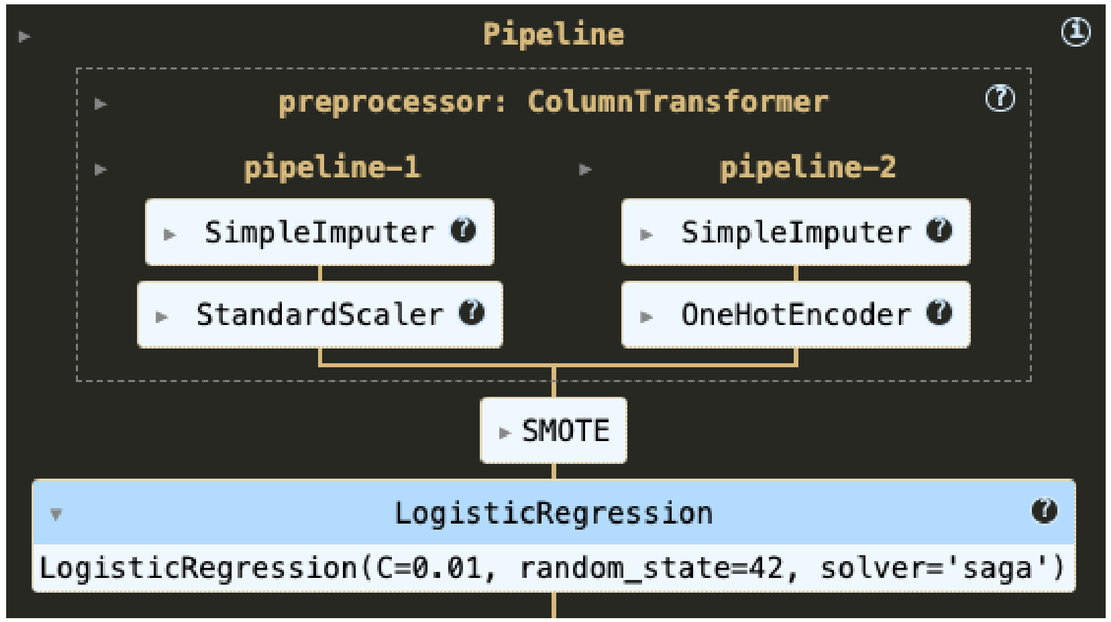
Yann LECERF

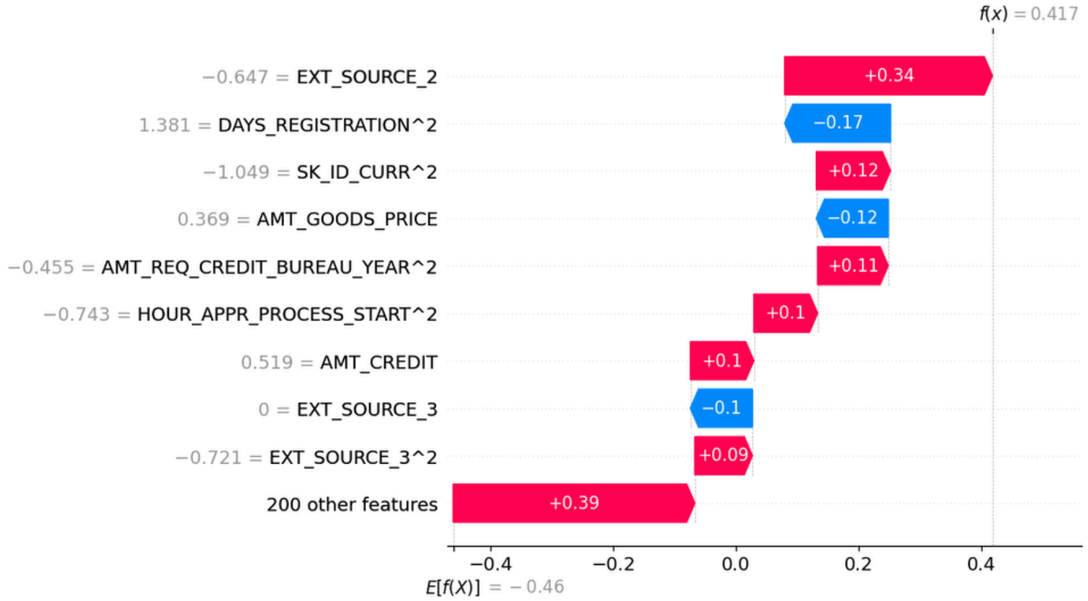
L’outil SHAP permet de visualiser l’impact des variables sur chaque prédiction individuelle, ce qui peut aider les conseillers à mieux comprendre les recommandations.

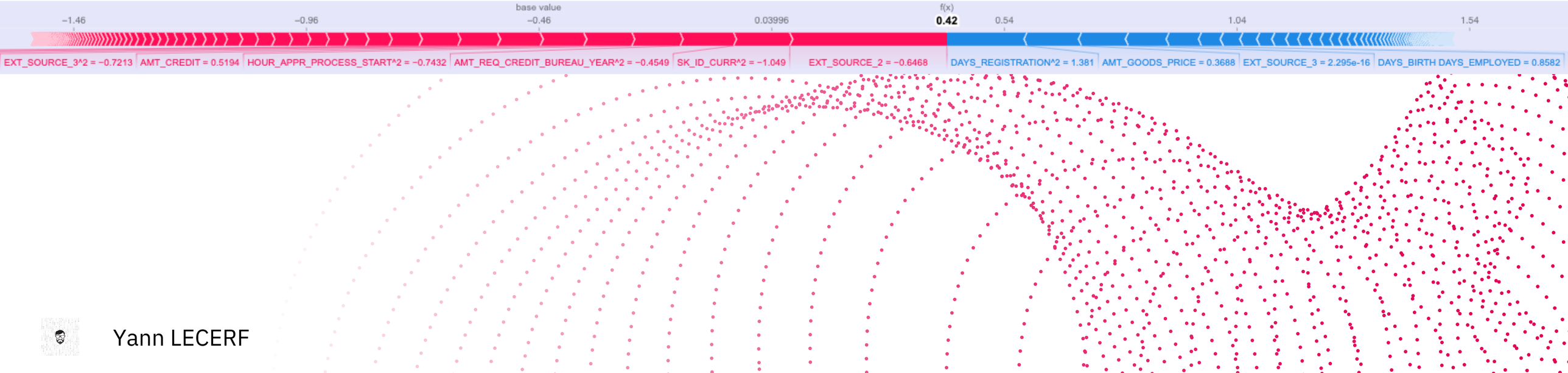
Interprétation des Prédictions Individuelles

CONCLUSIONS

RETOUR SOMMAIRE


Pipeline
preprocessor: ColumnTransformer
pipeline-1
pipeline-2
SimpleImputer
StandardScaler
SimpleImputer
OneHotEncoder
SMOTE
LogisticRegression
LogisticRegression(C=0.01, random_state=42, solver='saga')


f(x) = 0.417
−0.647 = EXT_SOURCE_2
1.381 = DAYS_REGISTRATION^2
−1.049 = SK_ID_CURR^2
0.369 = AMT_GOODS_PRICE
−0.455 = AMT_REQ_CREDIT_BUREAU_YEAR^2
−0.743 = HOUR_APPR_PROCESS_START^2
0.519 = AMT_CREDIT
0 = EXT_SOURCE_3
−0.721 = EXT_SOURCE_3^2
200 other features
+0.34
−0.17
+0.12
−0.12
+0.11
+0.1
+0.1
−0.1
+0.09
+0.39
−0.4
−0.2
0.0
0.2
0.4
E[f(X)] = −0.46


base value
f(x)
−1.46
−0.96
−0.46
0.03996
0.42
0.54
1.04
1.54
EXT_SOURCE_3^2 = −0.7213
AMT_CREDIT = 0.5194
HOUR_APPR_PROCESS_START^2 = −0.7432
AMT_REQ_CREDIT_BUREAU_YEAR^2 = −0.4549
SK_ID_CURR^2 = −1.049
EXT_SOURCE_2 = −0.6468
DAYS_REGISTRATION^2 = 1.381
AMT_GOODS_PRICE = 0.3688
EXT_SOURCE_3 = 2.295e-16
DAYS_BIRTH DAYS_EMPLOYED = 0.8582

RETOUR SOMMAIRE

Questions ?