

BAILLET Benjamin

MILLOT Alexandra

OUAZZANI CHAHDI Ismail

Projet Master IREF : Prédiction du risque de crédit



BNP PARIBAS
PERSONAL FINANCE



03/02/2025

Université de Bordeaux
MASTER IREF 2024-2025

CONTEXTE ET PROBLÉMATIQUE

Contexte

Dans un contexte économique où le risque de crédit constitue un enjeu majeur pour les institutions bancaires, l'automatisation du processus d'octroi de crédit devient essentielle pour :

- **Réduire les pertes financières** liées aux défauts de paiement.
- **Optimiser la prise de décision** en s'appuyant sur des outils prédictifs fiables.
- **Accélérer le traitement des demandes** tout en améliorant l'expérience client.

Problématique

Comment concevoir un modèle automatisé capable de :

1. **Analyser les profils clients** en identifiant les facteurs de risque clés.
2. **Élaborer une grille de score fiable** pour évaluer le risque de défaut.
3. **Challenger cette grille** à l'aide de modèles de Machine Learning pour améliorer la précision des prédictions.

Ce projet s'inscrit dans une démarche visant à améliorer la gestion du risque tout en garantissant une approche équitable et transparente dans l'évaluation des demandes de crédit.

OBJECTIFS DU PROJET

Ce projet vise à développer un outil performant permettant d'automatiser et d'optimiser l'octroi de crédit pour une institution bancaire. Les objectifs principaux se déclinent en trois étapes clés :

- **Analyse descriptive des données**

- Explorer et analyser les données clients pour identifier les caractéristiques influençant le risque de crédit.
- Effectuer des analyses univariées, bivariées et multivariées pour mieux comprendre le portefeuille client et détecter les variables discriminantes.

- **Élaboration d'une grille de score**

- Construire un modèle de régression logistique pour établir une grille de score fiable.
- Normaliser le score, évaluer sa performance et déterminer un seuil optimal pour minimiser les pertes financières.

- **Challenge avec un modèle de Machine Learning**

- Comparer la grille de score avec des algorithmes de Machine Learning (Random Forest, Gradient Boosting, etc.).
- Optimiser les hyperparamètres, analyser la performance des modèles et assurer leur interprétabilité.

Résultat attendu

Proposer une solution complète et robuste qui combine analyse statistique, techniques de scoring et intelligence artificielle pour améliorer les décisions de crédit.

PRÉSENTATION DES DONNÉES

Les données utilisées pour ce projet proviennent de l'historique des crédits d'une banque. Elles contiennent des informations détaillées sur les clients comme nous pouvons le voir à droite.

Taille et Sources des Données

- **Volume des données :** Le jeu de données contient:

- 307511 clients (le nombre de lignes)
- 22 caractéristiques (le nombre de colonnes)

avec 3 types de données différentes (entiers, flottants et objets)

- **Origine des données :**

- Historique interne de la banque.
- Données externes fournies par des bureaux de crédit.

Ces données offrent une base riche pour analyser les comportements des emprunteurs et construire des modèles prédictifs robustes.

SK_ID_CURR : identifiant du client

GOOD_PAYER : 1 s'il n'y a pas eu de retard de remboursement, 0 sinon

CODE_GENDER : le genre du client

FLAG_OWN_CAR : le client possède une voiture

FLAG_OWN_REALTY : le client est propriétaire

CNT_CHILDREN : nombre d'enfant du client

AMT_INCOME_TOTAL : les revenus du client

AMT_CREDIT : le montant du crédit

AMT_GOODS_PRICE : le montant du produit pour lequel le crédit a été pris

NAME_INCOME_TYPE : le type de revenus du client

NAME_EDUCATION_TYPE : niveau académique du client

NAME_FAMILY_STATUS: status familial du client

NAME_CONTRACT_TYPE: crédit comptant ou en revolving

NAME_HOUSING_TYPE: situation habitat

TOTALAREA_MODE: surface normalisée d'habitation

DAYS_BIRTH: Age du client

DAYS_EMPLOYED : nombre d'année consécutif du dernier emploi du client

OCCUPATION_TYPE : profession du client

ORGANIZATION_TYPE : secteur d'emploi

EXT_SOURCE_1 : score de crédit bureau 1

EXT_SOURCE_2 : score de crédit bureau 2

EXT_SOURCE_3 : score de crédit bureau 3

AMT_REQ_CREDIT_BUREAU_YEARS: nombre de demande de crédit effectué par le client dans l'année précédente

TRAITEMENT DES DONNÉES



Avant l'analyse, les données ont été nettoyées pour une meilleure lisibilité et interprétation :

- Les variables [DAYS_BIRTH](#) et [DAYS_EMPLOYED](#) ont été examinées pour des valeurs peu intuitives.
- Une attention particulière a été portée à la variable [AMT_INCOME_TOTAL](#), dont une valeur maximale très éloignée des autres sera traitée dans la section Valeurs extrêmes.

Valeurs manquantes

7 variables contiennent des valeurs manquantes, représentant entre 0,09 % et 56,38 % des données et voici les stratégies:

- Les colonnes avec plus de 40 % de valeurs manquantes ([EXT_SOURCE_1](#) et [TOTALAREA_MODE](#)) ont été supprimées.
- Les autres variables ont été imputées selon leur type et leur contexte :
 - **Imputation conditionnelle** : Utilisation de variables connexes pour estimer les valeurs manquantes (ex: [OCCUPATION_TYPE](#) imputée avec [NAME_INCOME_TYPE](#)).
 - **KNN Imputation** : Pour les variables numériques, estimation basée sur les k plus proches voisins.
 - **Imputation catégorielle** : Remplacement des valeurs manquantes par une catégorie spéciale ou par la modalité la plus fréquente.

Valeurs extrêmes

La variable [AMT_INCOME_TOTAL](#) présente une valeur aberrante très éloignée des autres (117 millions). Ces valeurs extrêmes ont été identifiées et ajustées :

- **Suppression des observations** considérées comme aberrantes.
- **Transformation logarithmique** pour réduire l'influence des grandes valeurs sur les analyses.

ANALYSE UNIVARIÉE

Description des variables numériques

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	TOTALAREA_MODE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	AMT_REQ_CREDIT_BUREAU_YEAR	AGE_YEARS	YEARS_EMPLOYED
count	307511.000000	3.075110e+05	3.075110e+05	3.072330e+05	159080.000000	134133.000000	3.068510e+05	246546.000000	265992.000000	307511.000000	307511.000000
mean	0.416753	1.687979e+05	5.990260e+05	5.383962e+05	0.102547	0.502130	5.143927e-01	0.510853	1.899974	43.936992	5.355694
std	0.718240	2.371231e+05	4.024908e+05	3.694465e+05	0.107462	0.211062	1.910602e-01	0.194844	1.869295	11.956084	6.320809
min	0.000000	2.565000e+04	4.500000e+04	4.050000e+04	0.000000	0.014568	8.173617e-08	0.000527	0.000000	20.500000	0.000000
25%	0.000000	1.125000e+05	2.700000e+05	2.385000e+05	0.041200	0.334007	3.924574e-01	0.370650	0.000000	34.000000	0.800000
50%	0.000000	1.471500e+05	5.135310e+05	4.500000e+05	0.068800	0.505998	5.659614e-01	0.535276	1.000000	43.200000	3.300000
75%	1.000000	2.025000e+05	8.086500e+05	6.795000e+05	0.127600	0.675053	6.636171e-01	0.669057	3.000000	53.900000	7.600000
max	6.000000	1.170000e+08	4.050000e+06	4.050000e+06	1.000000	0.962693	8.549997e-01	0.896010	25.000000	69.100000	49.100000

Distribution des variables numériques

Normalité (Kolmogorov-Smirnov Test) : Aucune des variables analysées ne suit une distribution normale (p-value du test KS = 0.0 pour toutes les colonnes).

Kurtosis : La kurtosis permet de mesurer si une distribution présente des queues plus ou moins épaisses que la normale :

- Faible kurtosis (proche de -1) : [SK_ID_CURR](#), [AGE_YEARS](#), [AMT_CREDIT](#), [AMT_GOODS_PRICE](#), [EXT_SOURCE_2](#), [EXT_SOURCE_3](#).
- Kurtosis élevée (très positive) : [GOOD_PAYER](#), [Is_Outlier_IF](#).

Skewness : L'asymétrie est mesurée par la skewness :

- Skewness > 0 : distribution décalée à droite (queue vers les grandes valeurs).
- Skewness < 0 : distribution décalée à gauche (queue vers les petites valeurs).

Asymétrie positive (queue à droite) : [AMT_INCOME_TOTAL](#), [CNT_CHILDREN](#), [AMT_GOODS_PRICE](#), [YEARS_EMPLOYED](#).

Asymétrie négative (queue à gauche) : [GOOD_PAYER](#), [Is_Outlier_IF](#), [EXT_SOURCE_2](#), [EXT_SOURCE_3](#).

Nous allons donc transformer les variables numériques présentant une forte asymétrie en appliquant une transformation logarithmique ou une racine carrée, afin de réduire ces asymétries importantes (par exemple : [AMT_INCOME_TOTAL](#), [CNT_CHILDREN](#), [AMT_CREDIT](#)).

ANALYSE BIVARIÉE ET CORRÉLATIONS

Nous allons analyser les corrélations entre les variables explicatives et la variable cible **GOOD_PAYER**, en utilisant les méthodes **ANOVA** et **Pearson**

Test Anova

	Variable	F-Stat	P-Value
0	CODE_GENDER	611.970774	6.048417e-135
1	FLAG_OWN_CAR	256.724260	9.487443e-58
2	FLAG_OWN_REALTY	1.882249	1.700800e-01
3	NAME_INCOME_TYPE	90.113413	1.916857e-113
4	NAME_CONTRACT_TYPE	330.967596	6.598701e-74
5	NAME_EDUCATION_TYPE	304.656706	6.301838e-262
6	NAME_FAMILY_STATUS	102.065917	5.366674e-87
7	NAME_HOUSING_TYPE	61.464686	2.923227e-64
8	OCCUPATION_TYPE	74.305359	2.077591e-272

Test Pearson

EXT_SOURCE_2	0.170337
EXT_SOURCE_3	0.158269
YEARS_EMPLOYED	0.079341
AGE_YEARS	0.065757
AMT_GOODS_PRICE	0.049886
AMT_CREDIT	0.040224
AMT_INCOME_TOTAL	0.036194
SK_ID_CURR	0.001413
CNT_CHILDREN	-0.007713
AMT_REQ_CREDIT_BUREAU_YEAR	-0.022462
dtype:	float64

Variables les plus importantes:

- [CODE_GENDER](#) (F-Stat : 828.80)
- [NAME_EDUCATION_TYPE](#) (F-Stat : 185.85) et [NAME_FAMILY_STATUS](#) (F-Stat : 175.23)
- [NAME_INCOME_TYPE](#) (F-Stat : 118.28) :

Variables d'impact modéré : [FLAG_OWN_CAR](#) et [NAME_CONTRACT_TYPE](#)

Variables à impact limité : [FLAG_OWN_REALTY](#) (F-Stat : 21.58)

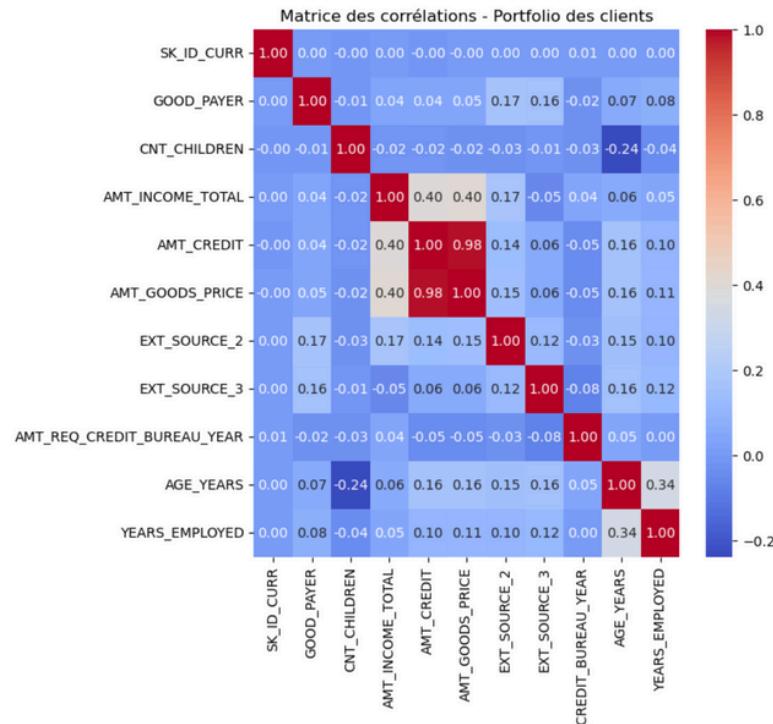
Variables avec une corrélation positive significative :

- [EXT_SOURCE_2](#) (0.0956) et [EXT_SOURCE_3](#) (0.0937)
- [AMT_GOODS_PRICE](#) (0.0617), [AGE_YEARS](#) (0.0610), et [AMT_CREDIT](#) (0.0540)
- [YEARS_EMPLOYED](#) (0.0539)
- [CNT_CHILDREN](#) (0.0439) et [AMT_INCOME_TOTAL](#) (0.0409)

Variables avec une corrélation négative ou neutre :

- [Is_Outlier_IF](#) (-0.0074) : Les anomalies détectées par l'algorithme **Isolation Forest** semblent légèrement associées à une probabilité plus faible d'être un bon payeur.
- Les autres variables comme [AMT_REQ_CREDIT_BUREAU_YEAR](#) (0.0188) et [SK_ID_CURR](#) (0.0082) montrent une corrélation presque nulle avec la cible, indiquant un faible rôle explicatif.

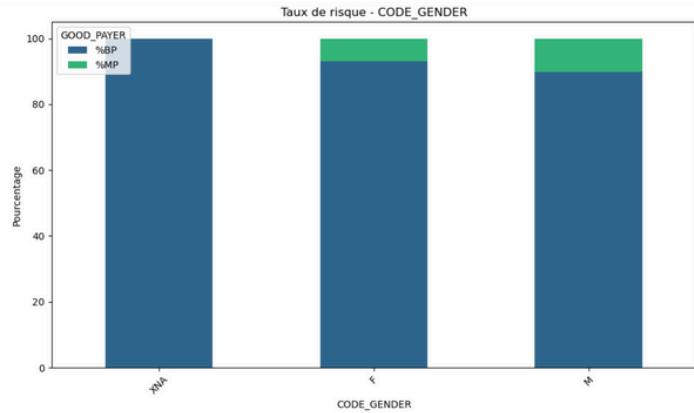
Matrice des corrélations



La **matrice de corrélation** met en relation 11 variables numériques entre elles, ce qui exclue les autres variables non numériques : **CODE_GENDER**, **FLAG_OWN_CAR**, **FLAG_OWN_REALTY**, **NAME_INCOME_TYPE**, **NAME_CONTRACT_TYPE**, **NAME_EDUCATION_TYPE**, **NAME_FAMILY_STATUS**, **NAME_HOUSING_TYPE**, **OCCUPATION_TYPE**, **ORGANIZATION_TYPE**

TAUX DE RISQUE, IV, GINI, TEST DE SIGNIFICATIVITÉ

Taux de risque



IV

Échelle d'interprétation de l'IV :

- 0 - 0.02 : Faiblement prédictive.
- 0.02 - 0.1 : Moyennement prédictive.
- 0.1 - 0.3 : Modérément prédictive.
- 0.3+ : Fortement prédictive.

Gini

Indice de Gini pour SK_ID_CURR : 0.0312
Indice de Gini pour CNT_CHILDREN : 0.0059
Indice de Gini pour AMT_INCOME_TOTAL : 0.0942
Indice de Gini pour AMT_CREDIT : 0.0795
Indice de Gini pour AMT_GOODS_PRICE : 0.1196
Indice de Gini pour EXT_SOURCE_2 : 0.3045
Indice de Gini pour EXT_SOURCE_3 : 0.3523
Indice de Gini pour AMT_REQ_CREDIT_BUREAU_YEAR : -0.0509
Indice de Gini pour AGE_YEARS : 0.0795
Indice de Gini pour YEARS_EMPLOYED : 0.1042

Test de significativité

Variables significatives ($p < 0.05$) :

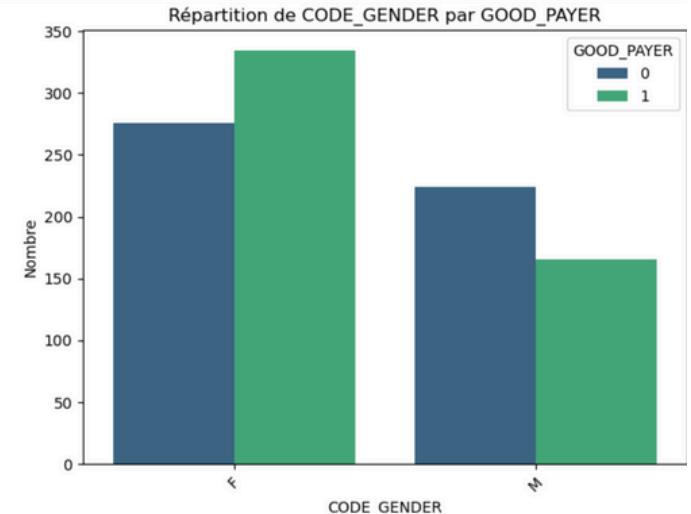
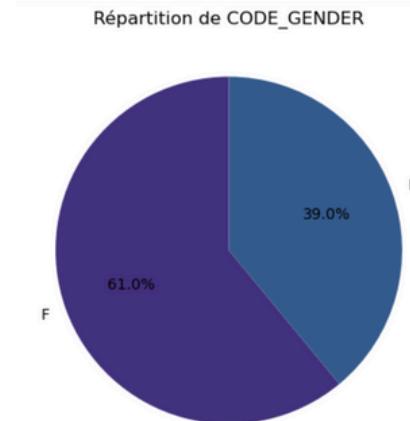
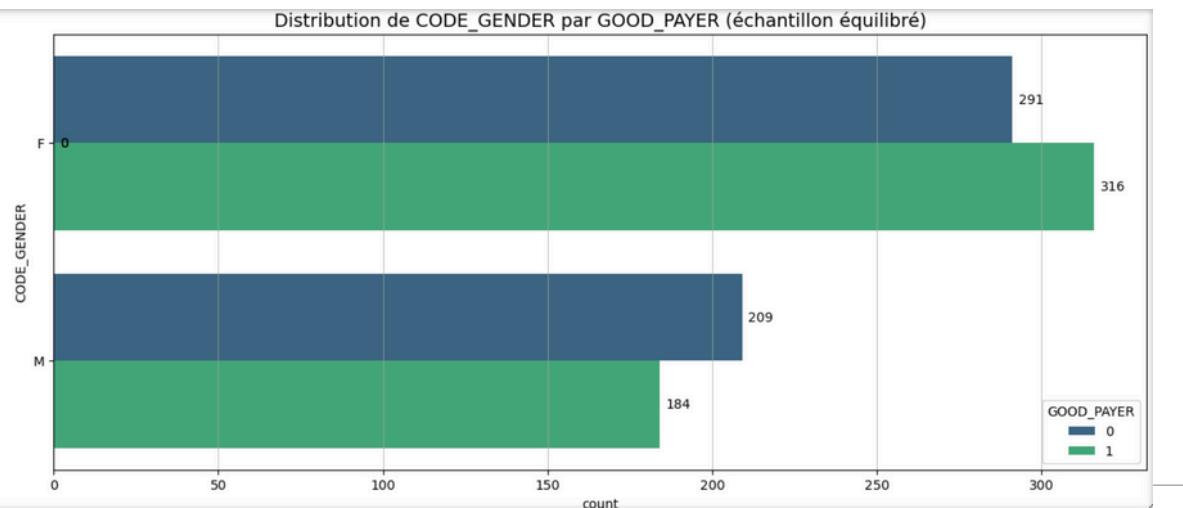
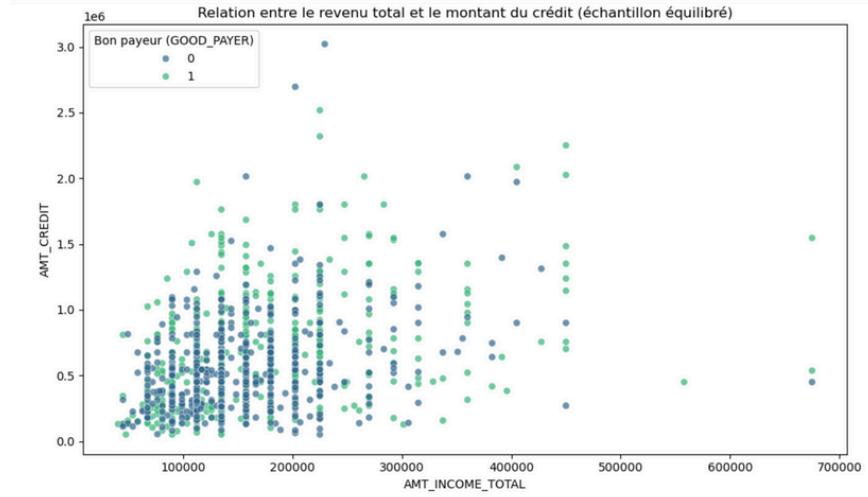
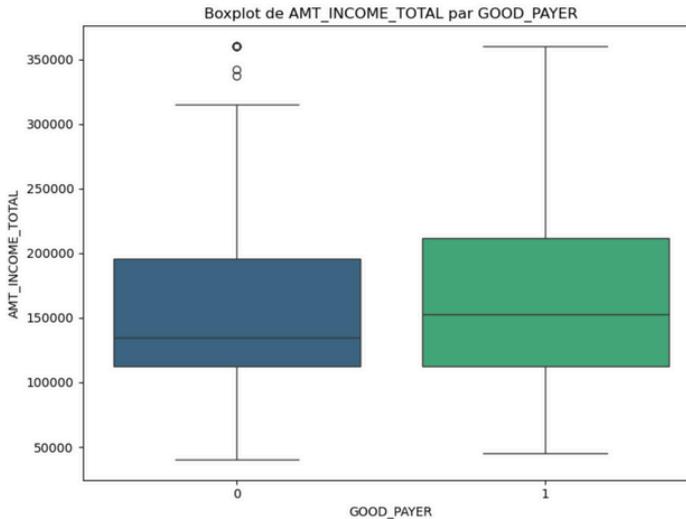
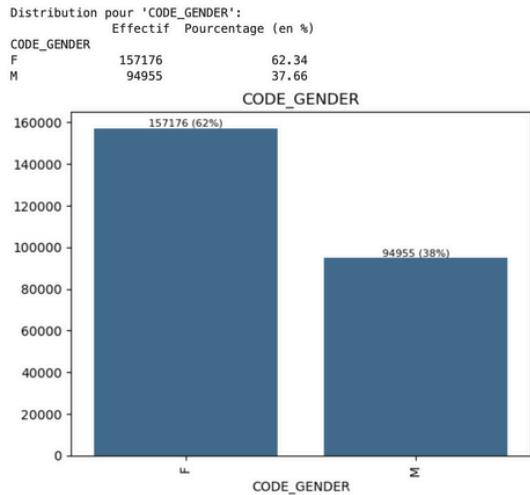
- AMT_CREDIT ($p = 0.0342$), CNT_CHILDREN ($p = 0.0159$), AMT_GOODS_PRICE ($p = 0.0008$), EXT_SOURCE_2 ($p = 0.0000$), EXT_SOURCE_3 ($p = 0.0000$), AGE_YEARS ($p = 0.0000$), YEARS_EMPLOYED ($p = 0.0007$) : Ces variables ont un indice de Gini significativement différent de 0, indiquant un pouvoir discriminant sur la variable cible. Elles peuvent être utilisées efficacement dans un modèle de scoring ou d'analyse prédictive.
- EXT_SOURCE_2 et EXT_SOURCE_3 : Les plus discriminantes (indices de Gini élevés)
- AGE_YEARS et YEARS_EMPLOYED
- AMT_GOODS_PRICE
- CNT_CHILDREN et AMT_CREDIT

Variables non significatives ($p \geq 0.05$) :

- AMT_INCOME_TOTAL ($p = 0.0502$) : Très proche du seuil de significativité.
- AMT_REQ_CREDIT_BUREAU_YEAR ($p = 0.4009$) : Pas significative, avec un indice de Gini négatif. Cette variable n'a pas de pouvoir discriminant et pourrait être exclue des modèles.

VISUALISATION DES DONNÉES

Voici différentes visualisations permettant de mettre en avant des relations spécifiques entre les variables explicatives et la variable cible:



PRÉPARATION DES DONNÉES

Pour construire une grille de score efficace à partir d'une régression logistique, deux étapes essentielles ont été réalisées : la division des données en sous-ensembles (split) et l'équilibrage des classes par la **méthode des poids**.

Étape 1 : Division des données en sous-ensembles

Les données ont été divisées en trois ensembles distincts pour garantir une modélisation rigoureuse et une évaluation fiable :

- **Ensemble d'entraînement (60%)** : utilisé pour ajuster les coefficients du modèle.
- **Ensemble de validation (20%)** : utilisé pour tester les hyperparamètres et ajuster le modèle.
- **Ensemble de test (20%)** : réservé à une évaluation finale sur des données indépendantes.

La méthode `train_test_split` avec stratification a été appliquée pour conserver les proportions de la variable cible ([GOOD_PAYER](#)) dans chaque sous-ensemble.

Étape 2 : Traitement du déséquilibre avec la méthode des poids

Pour corriger ce déséquilibre sans perdre de données, nous avons choisi **d'attribuer des poids aux observations plutôt que d'utiliser une méthode d'oversampling**.

La **méthode des poids** permet de rééquilibrer l'influence des classes sans introduire d'échantillons artificiels. Chaque observation se voit attribuer un poids inversement proportionnel à la taille de sa classe. Ainsi, les mauvais payeurs, sous-représentés, auront un poids plus élevé que les bons payeurs, garantissant un apprentissage robuste sans perte d'information.

Avant la méthode des poids

```
Train size: 151278, Validation size: 50426, Test size: 50427
```

Distribution dans chaque set:
 Train: GOOD_PAYER
 1 0.913398
 0 0.086602
 Name: proportion, dtype: float64
 Validation: GOOD_PAYER
 1 0.913398
 0 0.086602
 Name: proportion, dtype: float64
 Test: GOOD_PAYER
 1 0.9134
 0 0.0866
 Name: proportion, dtype: float64

Après la méthode des poids

Effectif pondéré après le split :
 Train : GOOD_PAYER
 0 75639.300000
 1 75638.971556
 Name: WEIGHT, dtype: float64
 Validation : GOOD_PAYER
 0 25213.100000
 1 25212.990519
 Name: WEIGHT, dtype: float64
 Test : GOOD_PAYER
 0 25213.100000
 1 25213.537925
 Name: WEIGHT, dtype: float64

TRANSFORMATION DES VARIABLES

Dans le cadre de l'objectif 2, une attention particulière a été portée à la structuration des données afin d'assurer la robustesse du modèle de scoring basé sur la régression logistique. Il a été constaté un fort déséquilibre dans la base de données, avec **91 % de bons payeurs contre seulement 9 % de mauvais payeurs**. Une correction de ce déséquilibre a été nécessaire pour éviter que le modèle ne soit biaisé en faveur de la classe majoritaire.

1. Transformation des variables catégorielles

- Les variables catégorielles ont été encodées via **l'encodage one-hot ou l'encodage ordinal appliqué avec la méthode indicatrice imbriquée**.
- Pour certaines variables présentant de nombreuses modalités, un **groupement par fréquence ou taux de risque** a été réalisé afin de limiter la création de dimensions inutiles.
- Des tests de significativité ont été effectués afin de **sélectionner les modalités les plus pertinentes** et éviter d'introduire du bruit dans le modèle (χ^2 , ANOVA).

2. Transformation des variables numériques

- Scaling (normalisation/standardisation)** pour harmoniser les échelles et éviter un effet disproportionné sur le modèle.
- Binning optimisé par Gini ou expertise** pour améliorer la segmentation des variables numériques.
- Suppression des variables avec IV < 0.02** pour ne garder que les prédicteurs pertinents

IV avant transformation

IV AVANT transformation :
 CODE_GENDER: 0.0281
 FLAG_OWN_CAR: 0.0121
 FLAG_OWN_REALTY: 0.0001
 NAME_INCOME_TYPE: 0.0294
 NAME_CONTRACT_TYPE: 0.0170
 NAME_EDUCATION_TYPE: 0.0643
 NAME_FAMILY_STATUS: 0.0215
 NAME_HOUSING_TYPE: 0.0141
 OCCUPATION_TYPE: 0.0709
 ORGANIZATION_TYPE: 0.0491

IV après transformation

IV APRÈS transformation :
 CODE_GENDER: 0.0281
 NAME_INCOME_TYPE: 0.0289
 NAME_EDUCATION_TYPE: 0.0618
 NAME_FAMILY_STATUS: 0.0199
 OCCUPATION_TYPE: 0.0663
 ORGANIZATION_TYPE: 0.0462

FLAG_OWN_CAR,
 FLAG_OWN_REALTY
 NAME_CONTRACT_TYPE
 NAME_HOUSING_TYPE
 ont un IV trop faible et sont supprimées

IV avant transformation

Variable	IV
EXT_SOURCE_2	0.322925
EXT_SOURCE_3	0.291345
AMT_GOODS_PRICE	0.110009
YEARS_EMPLOYED	0.100022
AGE_YEARS	0.060609
AMT_CREDIT	0.050859
AMT_INCOME_TOTAL	0.020417
AMT_REQ_CREDIT_BUREAU_YEAR	0.009556
SK_ID_CURR	0.001046
CNT_CHILDREN	0.000508

IV après transformation

Somme des IV par variable :	
Variable	
CNT_CHILDREN	0.000097
SK_ID_CURR	0.001046
AMT_REQ_CREDIT_BUREAU_YEAR	0.007977
AMT_INCOME_TOTAL	0.020123
AMT_CREDIT	0.050859
AGE_YEARS	0.060609
YEARS_EMPLOYED	0.100022
AMT_GOODS_PRICE	0.110009
EXT_SOURCE_3	0.291673
EXT_SOURCE_2	0.322869

RÉGRÉSSION LOGISTIQUE

Régression logistique avec 23 variables

Variables finales sélectionnées :	'EXT_SOURCE_2', 'EXT_SOURCE_3', 'YEARS_EMPLOYED', 'NAME_EDUCATION_TYPE_Secondary / secondary special', 'CODE_GENDER_M', 'ORGANIZATION_TYPE_Military', 'ORGANIZATION_TYPE_Self-employed', 'NAME_FAMILY_STATUS_Married', 'NAME_INCOME_TYPE_Working', 'AMT_INCOME_TOTAL', 'ORGANIZATION_TYPE_Medicine', 'OCCUPATION_TYPE_Drivers', 'AGE_YEARS', 'OCCUPATION_TYPE_Laborers', 'OCCUPATION_TYPE_Sales staff', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'ORGANIZATION_TYPE_Other', 'NAME_FAMILY_STATUS_Single / not married', 'OCCUPATION_TYPE_Managers', 'OCCUPATION_TYPE_High skill tech staff', 'NAME_INCOME_TYPE_State servant', 'ORGANIZATION_TYPE_Government'				
Logit Regression Results					
Dep. Variable:	GOOD_PAYER				
No. Observations:	151278				
Model:	Logit				
Method:	MLE				
Date:	Sat, 01 Feb 2025				
Pseudo R-squ.:	0.09772				
Time:	17:49:33				
Log-Likelihood:	-40212.				
converged:	True				
LL-Null:	-44567.				
Covariance Type:	nonrobust				
LR p-value:	0.000				
coef	std err	z	P> z	[0.025	0.975]
const	0.0431	0.067	0.647	0.518	-0.088
EXT_SOURCE_2	2.2396	0.047	47.751	0.000	2.148
EXT_SOURCE_3	2.6271	0.052	50.754	0.000	2.526
YEARS_EMPLOYED	0.0334	0.002	14.382	0.000	0.029
NAME_EDUCATION_TYPE_Secondary / secondary special	-0.3333	0.026	-12.627	0.000	-0.385
CODE_GENDER_M	-0.2460	0.023	-10.678	0.000	-0.291
ORGANIZATION_TYPE_Military	0.4118	0.059	6.992	0.000	0.296
ORGANIZATION_TYPE_Self-employed	-0.1194	0.023	-5.164	0.000	-0.165
NAME_FAMILY_STATUS_Married	0.1718	0.025	6.833	0.000	0.123
NAME_INCOME_TYPE_Working	-0.1316	0.023	-5.805	0.000	-0.176
AMT_INCOME_TOTAL	7.824e-07	1.57e-07	4.978	0.000	4.74e-07
ORGANIZATION_TYPE_Medicine	0.8378	0.001	3.798	0.000	0.865
OCCUPATION_TYPE_Drivers	-0.2941	0.043	-6.755	0.000	-0.379
AGE_YEARS	0.0854	0.001	5.113	0.000	0.063
OCCUPATION_TYPE_Laborers	-0.2003	0.036	-5.614	0.000	-0.270
OCCUPATION_TYPE_Sales staff	-0.1981	0.039	-5.146	0.000	-0.274
AMT_CREDIT	-2.466e-06	1.49e-07	-16.636	0.000	-2.76e-06
AMT_GOODS_PRICE	2.796e-06	1.71e-07	16.322	0.000	2.46e-06
ORGANIZATION_TYPE_Other	0.0929	0.033	2.779	0.005	0.027
NAME_FAMILY_STATUS_Single / not married	0.0693	0.032	2.188	0.029	0.007
OCCUPATION_TYPE_Managers	-0.1210	0.049	-2.489	0.013	-0.216
OCCUPATION_TYPE_High skill tech staff	-0.0867	0.045	-1.927	0.054	-0.175
NAME_INCOME_TYPE_State servant	-0.0940	0.047	-1.983	0.047	-0.187
ORGANIZATION_TYPE_Government	0.0818	0.043	1.986	0.057	-0.002

La première étape clé pour construire un modèle de régression logistique est la sélection des variables explicatives afin d'améliorer la précision et l'interprétabilité du modèle tout en évitant le sur-ajustement. Pour cela, la **méthode ascendante (Forward Selection)** est utilisée : elle commence avec un modèle vide et ajoute progressivement les variables les plus pertinentes selon des critères statistiques comme l'**AIC**, le **test de vraisemblance et la significativité des coefficients**. Le processus s'arrête lorsque l'ajout d'une nouvelle variable n'améliore plus significativement le modèle. Cette approche permet d'obtenir un modèle optimisé, ne conservant que les variables ayant un impact significatif sur la prédiction du statut de bon payeur (**GOOD_PAYER**). Notre modèle a sélectionné **23 variables** après la méthode ascendante basée sur l'AIC.

Régression logistique avec 20 variables

Logit Regression Results						
Dep. Variable:	GOOD_PAYER	No. Observations:	151257	Model:	Logit	Df Residuals:
Method:	MLE	Df Model:	20	Date:	Sat, 01 Feb 2025	Pseudo R-squ.:
Time:	17:49:33	Log-Likelihood:	-40217.	converged:	True	LL-Null:
Covariance Type:	nonrobust	LLR p-value:	0.000	const	0.0092	0.063
EXT_SOURCE_2	2.2396	0.047	47.751	0.000	0.145	0.884
EXT_SOURCE_3	2.6267	0.052	50.750	0.000	2.525	2.728
YEARS_EMPLOYED	0.0334	0.002	14.382	0.000	0.029	0.038
NAME_EDUCATION_TYPE_Secondary / secondary special	-0.3378	0.026	-12.849	0.000	-0.389	-0.286
CODE_GENDER_M	-0.2494	0.023	-10.839	0.000	-0.294	-0.204
ORGANIZATION_TYPE_Military	0.3879	0.057	6.788	0.000	0.276	0.508
ORGANIZATION_TYPE_Self-employed	-0.1268	0.022	-5.645	0.000	-0.171	-0.083
NAME_FAMILY_STATUS_Married	0.1724	0.025	6.858	0.000	0.123	0.222
NAME_INCOME_TYPE_Working	-0.1178	0.021	-5.573	0.000	-0.159	-0.076
AMT_INCOME_TOTAL	7.798e-07	1.57e-07	4.969	0.000	4.72e-07	1.09e-06
ORGANIZATION_TYPE_Medicine	0.1077	0.036	3.032	0.002	0.038	0.177
OCCUPATION_TYPE_Drivers	-0.2559	0.039	-6.637	0.000	-0.331	-0.180
AGE_YEARS	0.0054	0.001	5.115	0.000	0.003	0.007
OCCUPATION_TYPE_Laborers	-0.1618	0.029	-5.528	0.000	-0.219	-0.104
OCCUPATION_TYPE_Sales staff	-0.1629	0.033	-4.989	0.000	-0.227	-0.099
AMT_CREDIT	-2.466e-06	1.49e-07	-16.644	0.000	-2.76e-06	-2.18e-06
AMT_GOODS_PRICE	2.796e-06	1.71e-07	16.327	0.000	2.46e-06	3.13e-06
ORGANIZATION_TYPE_Other	0.0773	0.033	2.357	0.018	0.013	0.142
NAME_FAMILY_STATUS_Single / not married	0.0697	0.032	2.204	0.028	0.008	0.132
OCCUPATION_TYPE_Managers	-0.0836	0.045	-1.863	0.062	-0.172	0.004

✓ AIC après suppression des variables inutiles : 80476.22

Pour éviter le sur-ajustement, il est possible de simplifier le modèle en **supprimant les variables dont la p-value est supérieure à 0.05**. Cela renforce la robustesse du modèle en conservant uniquement les variables significatives.

Ainsi, les variables:

- OCCUPATION_TYPE_High skill tech staff**
- ORGANIZATION_TYPE_Government**
- NAME_INCOME_TYPE_State**

sont supprimées pour améliorer la performance du modèle.

Nous avons donc sélectionné **20 variables explicatives**, ce qui reste relativement élevé et pourrait entraîner un risque de **sur-ajustement (overfitting)**.



BNP PARIBAS
PERSONAL FINANCE



Régression logistique avec 14 variables

Entrainement du modèle...						
Optimization terminated successfully.						
Current function value: 0.266904						
Iterations 7						
Logit Regression Results						
Dep. Variable:	GOOD_PAYER	No. Observations:	151278	Model:	Logit	Df Residuals:
Method:	MLE	Df Model:	13	Date:	Sat, 01 Feb 2025	Pseudo R-squ.:
Time:	20:27:00	Log-Likelihood:	-40377.	converged:	True	LL-Null:
Covariance Type:	nonrobust	LLR p-value:	0.000	const	0.2145	0.052
EXT_SOURCE_2	2.2972	0.046	49.596	0.000	4.114	0.000
EXT_SOURCE_3	2.6092	0.051	50.864	0.000	2.206	0.317
YEARS_EMPLOYED	0.0351	0.002	15.275	0.000	0.031	0.040
NAME_EDUCATION_TYPE_Secondary / secondary special	-0.3730	0.026	-14.432	0.000	-0.424	-0.322
CODE_GENDER_M	-0.2243	0.022	-10.186	0.000	-0.267	-0.181
ORGANIZATION_TYPE_Military	0.3659	0.056	6.596	0.000	0.256	0.476
ORGANIZATION_TYPE_Self-employed	-0.1521	0.021	-7.175	0.000	-0.194	-0.111
NAME_FAMILY_STATUS_Married	0.1488	0.020	7.470	0.000	0.110	0.188
NAME_INCOME_TYPE_Working	-0.1260	0.021	-6.025	0.000	-0.167	-0.085
OCCUPATION_TYPE_Drivers	-0.2719	0.036	-7.636	0.000	-0.342	-0.202
AGE_YEARS	0.0048	0.001	4.660	0.000	0.003	0.007
OCCUPATION_TYPE_Laborers	-0.1751	0.026	-6.811	0.000	-0.225	-0.125
OCCUPATION_TYPE_Sales staff	-0.1781	0.030	-5.947	0.000	-0.237	-0.119

Précision du modèle : 0.9131
UIC-ROC : 0.7239
Matrice de confusion :

[14 4353]
[27 46032]

Nous avons précédemment sélectionné **20 variables explicatives**, ce qui peut entraîner un risque de sur-ajustement et réduire la capacité de généralisation du modèle. Pour y remédier, nous appliquons plusieurs techniques de sélection :

- Suppression des variables avec un IV < 0.02** : élimine les variables peu informatives.
- Filtrage des variables avec un VIF > 10** : identifie et supprime celles présentant une multi-colinéarité élevée.
- Sélection via Lasso (L1)** : applique une pénalisation pour conserver uniquement les variables les plus pertinentes.

Nous avons donc sélectionné **14 variables explicatives**

CONSTRUCTION DE LA GRILLE DE SCORE

La grille de score a été construite en utilisant une méthodologie de **régression logistique ascendante**, permettant de sélectionner progressivement les variables les plus pertinentes pour le modèle.

La méthode ascendante (Forward Selection) ajoute progressivement les variables les plus pertinentes selon des critères statistiques :

- **AIC (Akaike Information Criterion)** : Plus l'AIC diminue, meilleur est le modèle.
- **Test de vraisemblance (LLR p-value)** : Vérifie si une variable améliore significativement le modèle.
- **Significativité des coefficients (p-value)** : Une variable est conservée si $p < 0.05$.

Étapes du Processus

- **Démarrage** avec un modèle contenant uniquement l'intercept.
- **Ajout progressif des variables les plus impactantes** à chaque itération.
- **Arrêt du processus** lorsqu'aucune variable n'améliore significativement l'AIC.

Grille de score

La grille de score est construite à partir de **6 variables maximum**, en respectant les règles suivantes : **Valeur de référence** : 500, **Facteur multiplicatif** : 2, **Nombre de points** : 20, **Ratio de référence** : 1. Lors de la construction de la grille de score, nous avons identifié des **incohérences de monotonie** : certaines catégories à plus haut risque se voyaient attribuer un score plus élevé que des catégories à faible risque, ce qui pouvait altérer la qualité des prédictions. Pour corriger cela, nous avons appliqué une **méthode de lissage et de regroupement adaptatif**.

Grilles de score

Grille de score pour EXT_SOURCE_2 :			
	EXT_SOURCE_2	WOE	Score
0	(-0.0009999183, 0.222]	-0.973336	500.779788
1	(0.222, 0.345]	-0.453674	522.805347
2	(0.345, 0.445]	-0.257315	539.167710
3	(0.445, 0.514]	0.019349	386.170002
4	(0.514, 0.566]	0.097364	432.791064
5	(0.566, 0.609]	0.210434	455.028935
6	(0.609, 0.646]	0.363841	470.827658
7	(0.646, 0.682]	0.556871	483.108347
8	(0.682, 0.721]	0.695275	489.513167
9	(0.721, 0.855]	1.050422	501.419413

Grille de score pour EXT_SOURCE_3 :			
	EXT_SOURCE_3	WOE	Score
0	(-0.000473, 0.245]	-0.958481	501.223532
1	(0.245, 0.349]	-0.428722	524.437628
2	(0.349, 0.422]	-0.162167	552.488767
3	(0.422, 0.478]	-0.062310	580.087489
4	(0.478, 0.524]	0.138801	443.022027
5	(0.524, 0.57]	0.233582	458.040109
6	(0.57, 0.616]	0.296916	464.962657
7	(0.616, 0.669]	0.459410	477.557177
8	(0.669, 0.731]	0.681097	488.918706
9	(0.731, 0.888]	0.967613	499.050078

Grille de score pour AGE_YEARS :			
	AGE_YEARS	WOE	Score
0	(20.499, 27.5]	-0.331494	531.858811
1	(27.5, 30.7]	-0.319137	532.954991
2	(30.7, 33.9]	-0.182703	549.048475
3	(33.9, 37.1]	-0.094040	568.211392
4	(37.1, 39.9]	-0.009382	634.715435
5	(39.9, 43.0]	0.122633	439.448613
6	(43.0, 46.3]	0.118107	438.363662
7	(46.3, 50.0]	0.214023	455.516839
8	(50.0, 54.2]	0.271958	462.429176
9	(54.2, 69.0]	0.472482	478.366683

Grille de score pour NAME_EDUCATION_TYPE_Secondary / secondary special :			
	NAME_EDUCATION_TYPE_Secondary / secondary special	WOE	Score
0	False	0.469961	478.21232
1	True	-0.132159	558.39277

Grille de score pour CODE_GENDER_M :			
	CODE_GENDER_M	WOE	Score
0	False	0.139771	443.222884
1	True	-0.201397	546.237546

Scores et monotonie

Vérification de la monotonie des scores et du risque moyen					
	Score	mean_risk	monotonic_increasing	monotonic_decreasing	monotonicity_respected
0	1676.069921	0.842105	True	True	0 True
1	1677.154872	1.000000	True	False	1 True
2	1691.088002	0.888889	False	True	2 True
3	1692.172954	0.882353	False	True	3 True
4	1693.223097	0.866667	False	True	4 True
... True
9618	2289.776480	0.916667	True	False	9618 True
9619	2291.957933	0.900000	False	True	9619 True
9620	2301.914756	0.962963	True	False	9620 True
9621	2303.194291	0.947368	False	True	9621 True
9622	2319.556654	0.850000	False	True	9622 True

ÉVALUATION DE LA GRILLE

L'évaluation de la grille repose sur des indicateurs de performance et une analyse de l'impact économique pour valider son efficacité.

Indicateurs de performance

1. Indice de GINI : Évalué sur les ensembles d'entraînement,

validation et test pour mesurer la capacité discriminante de la grille.

2. Courbe KS : Identifie le point où la différence entre les pourcentages cumulés de "bons" et "mauvais" clients est maximale, optimisant les décisions d'acceptation/refus.

3. Déciles de score : Analyse des risques par tranche pour mieux segmenter les clients.

Impact économique

Les erreurs de classification ont été quantifiées :

- Faux positifs** : Perte de la totalité du crédit financé.
- Faux négatifs** : Perte de 8% liée aux intérêts non perçus.

Une **courbe d'équilibre financier** a permis de déterminer le seuil de score optimal, maximisant le gain ou minimisant les pertes, rendant la grille performante et rentable pour la banque.

Tableau des déciles

Table des déciles - Entraînement						
	count	Bad	Good	Bad Rate	Cumulative Bad Rate	Cumulative Good Rate
Decile						
10	15130	14413	717	0.952611	0.104308	0.054729
9	15130	14317	813	0.946266	0.207922	0.116785
8	15125	14225	900	0.940496	0.310869	0.185482
7	15127	14079	1048	0.930720	0.412760	0.265476
6	15127	14009	1118	0.926092	0.514145	0.350813
5	15130	13805	1325	0.912426	0.614053	0.451950
4	15127	13789	1338	0.911549	0.713845	0.554080
3	15130	13403	1727	0.885856	0.810844	0.685902
2	15128	13336	1792	0.881544	0.907358	0.822685
1	15124	12801	2323	0.846403	1.000000	1.000000

Table des déciles - Validation						
	count	Bad	Good	Bad Rate	Cumulative Bad Rate	Cumulative Good Rate
Decile						
10	5044	4767	277	0.945083	0.103498	0.063430
9	5044	4798	246	0.951229	0.207668	0.119762
8	5040	4733	307	0.939087	0.310428	0.190062
7	5044	4698	346	0.931404	0.412428	0.269292
6	5043	4666	377	0.925243	0.513732	0.355622
5	5041	4624	417	0.917278	0.614125	0.451111
4	5042	4548	494	0.902023	0.712868	0.564232
3	5045	4500	545	0.891972	0.810569	0.689031
2	5040	4418	622	0.876587	0.906490	0.831463
1	5043	4307	736	0.854055	1.000000	1.000000

Table des déciles - Test						
	count	Bad	Good	Bad Rate	Cumulative Bad Rate	Cumulative Good Rate
Decile						
10	5043	4816	227	0.954987	0.104559	0.051981
9	5044	4755	289	0.942704	0.207794	0.118159
8	5041	4734	307	0.939099	0.310573	0.188459
7	5044	4688	356	0.929421	0.412353	0.269979
6	5043	4621	422	0.916320	0.512679	0.366613
5	5043	4595	448	0.911164	0.612440	0.469201
4	5045	4576	469	0.907037	0.711789	0.576597
3	5038	4462	576	0.885669	0.808663	0.708496
2	5045	4477	568	0.887413	0.905862	0.838562
1	5041	4336	705	0.860147	1.000000	1.000000

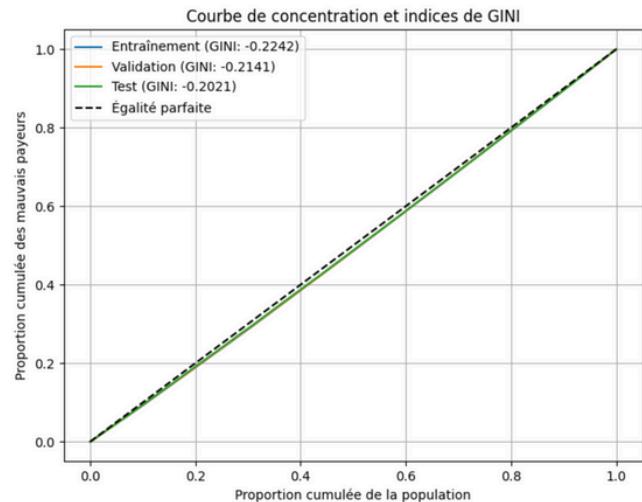
Résultats clés

Seuil optimal ajusté pour le gain économique : 2597.51
Train - Gain: 72672.00, TP: 138177, FP: 13101, FN: 0, TN: 0
Validation - Gain: 24222.92, TP: 46058, FP: 4367, FN: 1, TN: 0
Test - Gain: 24225.00, TP: 46060, FP: 4367, FN: 0, TN: 0

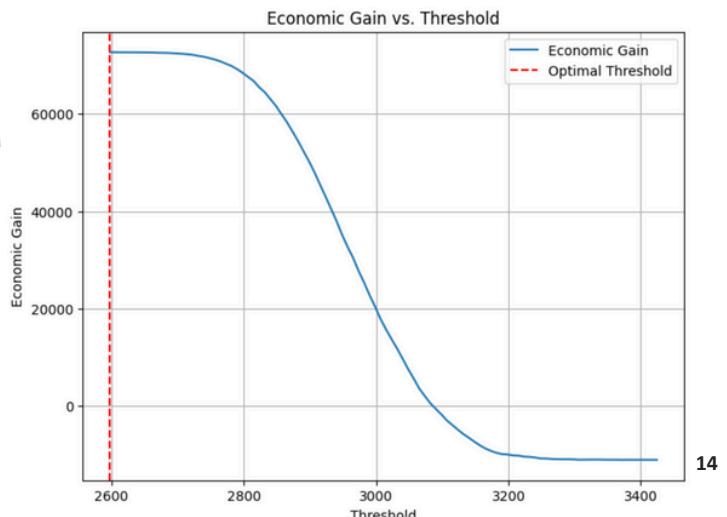
Évaluation du Modèle :
 R^2 Ajusté (Train) : 0.0939
AIC : 80781.35
BIC : 80920.32



Courbe de concentration

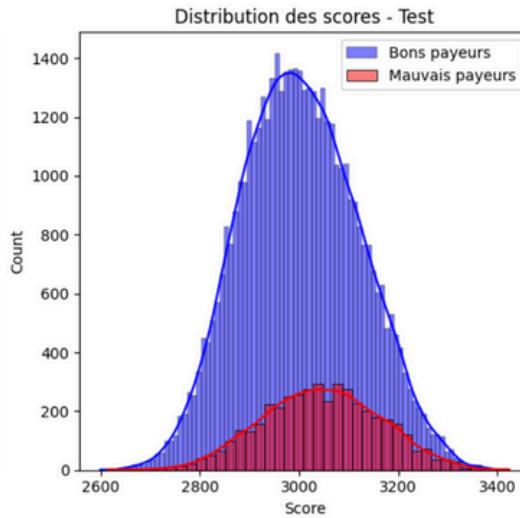
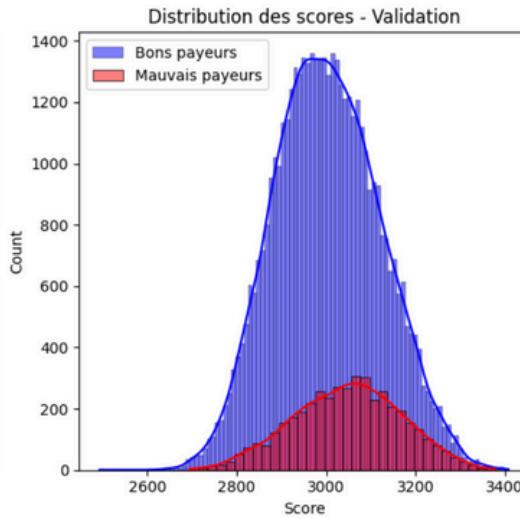
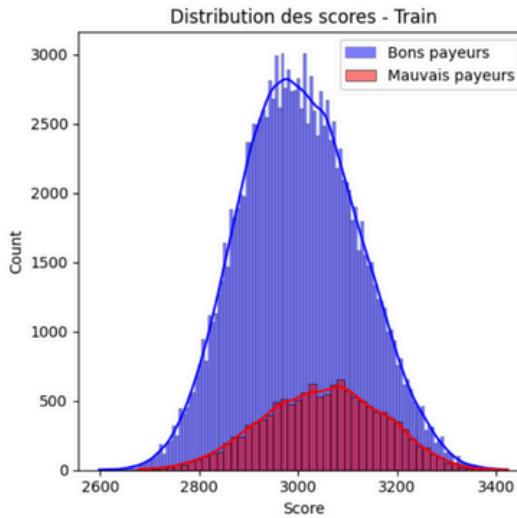


Courbe d'équilibre financier

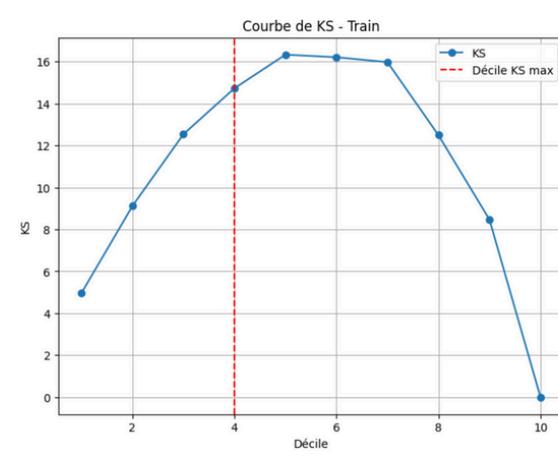
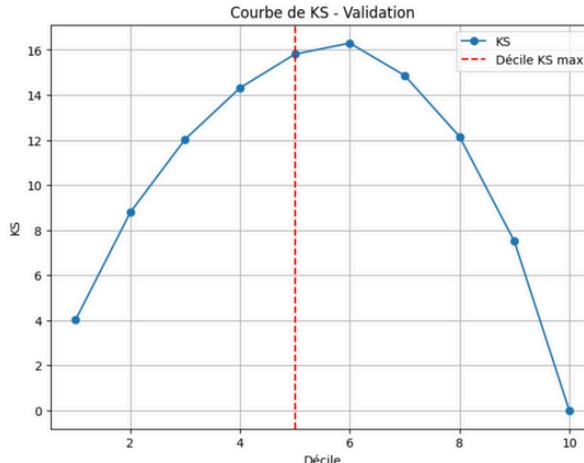
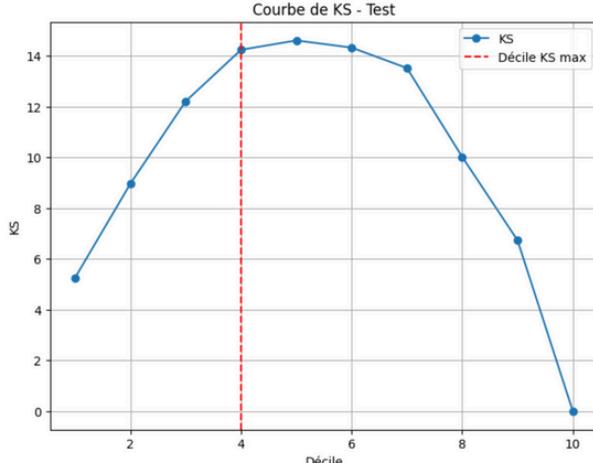


ÉVALUATION DE LA GRILLE

Distribution des scores

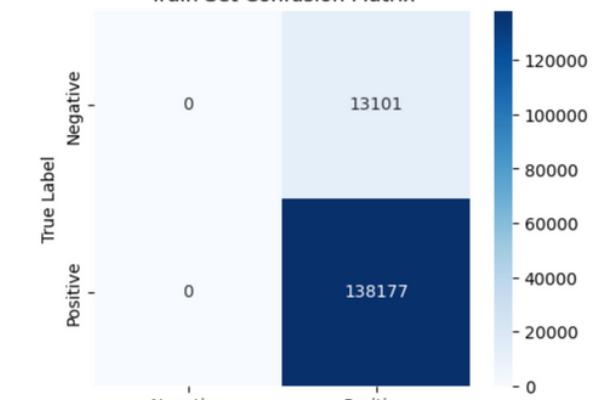


Courbes KS



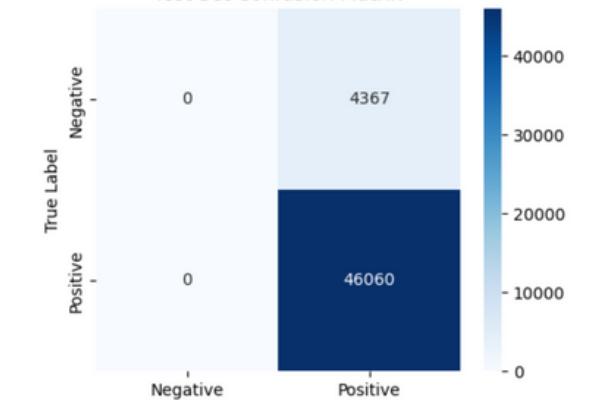
Matrices de confusion

Train Set Confusion Matrix



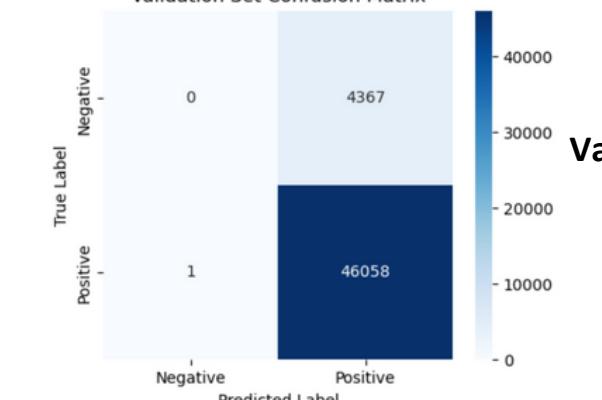
Train

Test Set Confusion Matrix



Test

Validation Set Confusion Matrix



Validation



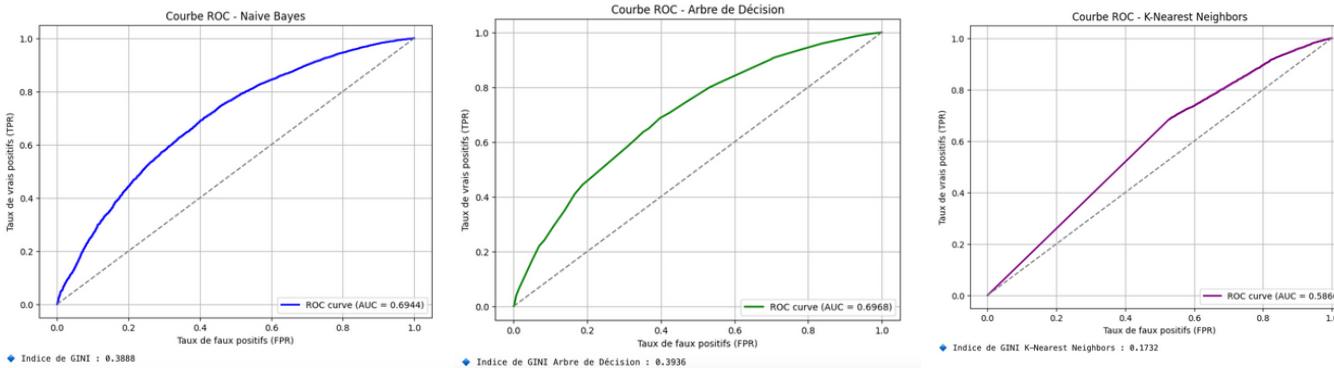
INTRODUCTION AUX MODÈLES DE MACHINE LEARNING

Pour challenger la grille de score basée sur la régression logistique, nous avons testé plusieurs modèles de Machine Learning, simples et complexes, afin d'améliorer la prédiction du risque de crédit.

Modèles Simples

- Naïve Bayes (NB)** : Basé sur le théorème de Bayes, efficace pour des données textuelles et bien adaptées aux hypothèses d'indépendance.
- Arbre de Décision** : Segmente les données en fonction des caractéristiques les plus discriminantes, simple et interprétable.
- K-Nearest Neighbors (KNN)** : Classe une observation selon ses voisins les plus proches, efficace sur de petites bases.

Métrique	Définition	Interprétation
AUC (Area Under Curve)	Aire sous la courbe ROC	Plus proche de 1 = meilleur modèle
Courbe ROC	Graphique du taux de vrais positifs contre le taux de faux positifs	Plus la courbe est proche du coin supérieur gauche, mieux c'est
Indice de GINI	$GINI = 2 \times AUC - 1$	Plus proche de 1 = modèle performant

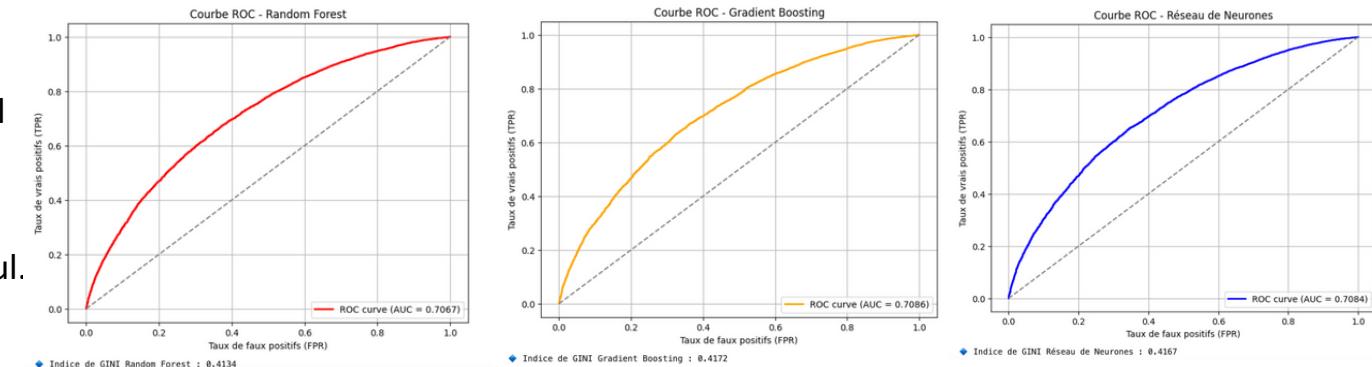


Modèles Complexes

- Random Forest (RF)** : Agrégation d'arbres de décision, robuste et réduit le surapprentissage.
- Gradient Boosting (XGBoost, LightGBM)** : Entraînement séquentiel pour améliorer progressivement les erreurs, très performant.
- Réseau de Neurones (MLP)** : Capture des relations non linéaires complexes mais nécessite plus de données et de puissance de calcul.

Avant l'entraînement, nous avons utilisé l'**encodage One-Hot** pour convertir les variables catégorielles en valeurs numériques.

Les modèles ont été comparés selon des métriques comme l'**AUC-ROC** et le **F1-score** afin d'optimiser la prédiction du risque de crédit.



PRÉPARATION DES DONNÉES POUR LE MACHINE LEARNING

1. Encodage des variables catégorielles

Les variables catégorielles sont transformées via **One-Hot Encoding**, permettant de convertir chaque catégorie en colonnes binaires exploitables par les modèles.

2. Optimisation des hyperparamètres

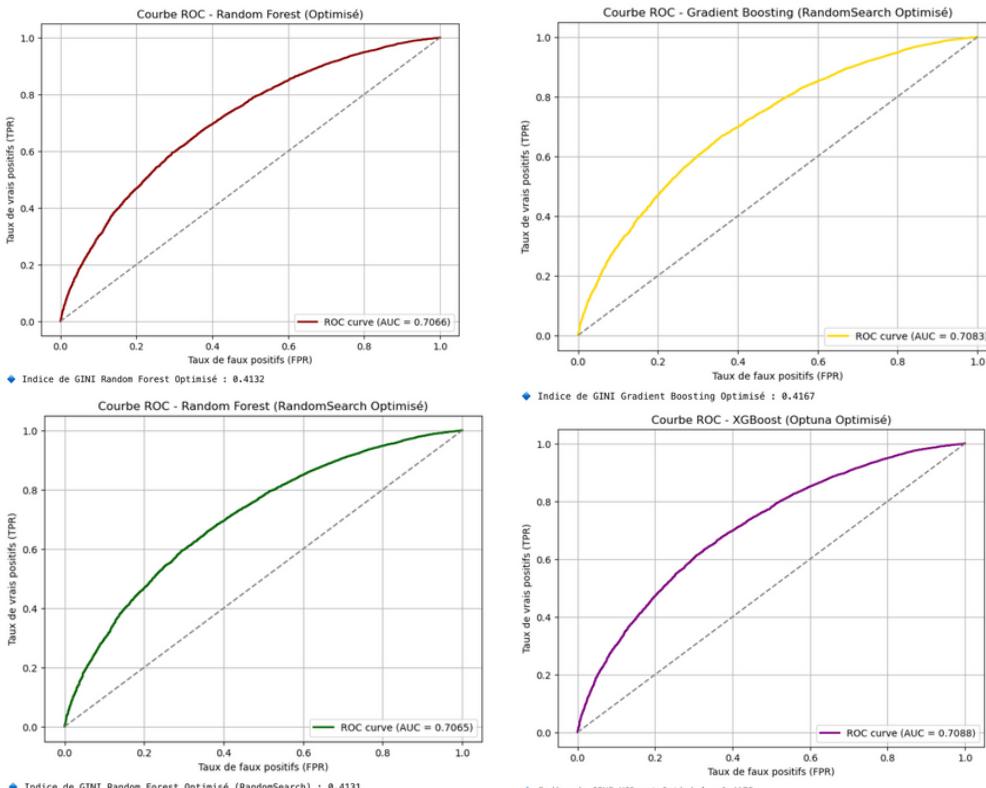
L'optimisation améliore la performance des modèles en ajustant leurs paramètres :

- **GridSearchCV:** Utilisé pour optimiser Random Forest, il explore différentes configurations en effectuant une validation croisée et sélectionne la meilleure combinaison selon l'AUC.
- **Optuna:** Employé pour XGBoost, il optimise les hyperparamètres comme `n_estimators`, `learning_rate`, et `max_depth`, avec une approche bayésienne accélérant la recherche des meilleures valeurs.

Comparaison des performances avant optimisation des hyperparamètres

Comparaison des performances :
 AUC Naïve Bayes : 0.6944
 AUC Arbre de Décision : 0.6968
 AUC K-Nearest Neighbors : 0.5866
 AUC Random Forest : 0.7067
 AUC Gradient Boosting : 0.7086
 AUC Réseau de Neurones : 0.7084

Indice de GINI Naïve Bayes : 0.3888
 Indice de GINI Arbre de Décision : 0.3936
 Indice de GINI K-Nearest Neighbors : 0.1732
 Indice de GINI Random Forest : 0.4134
 Indice de GINI Gradient Boosting : 0.4172
 Indice de GINI Réseau de Neurones : 0.4167



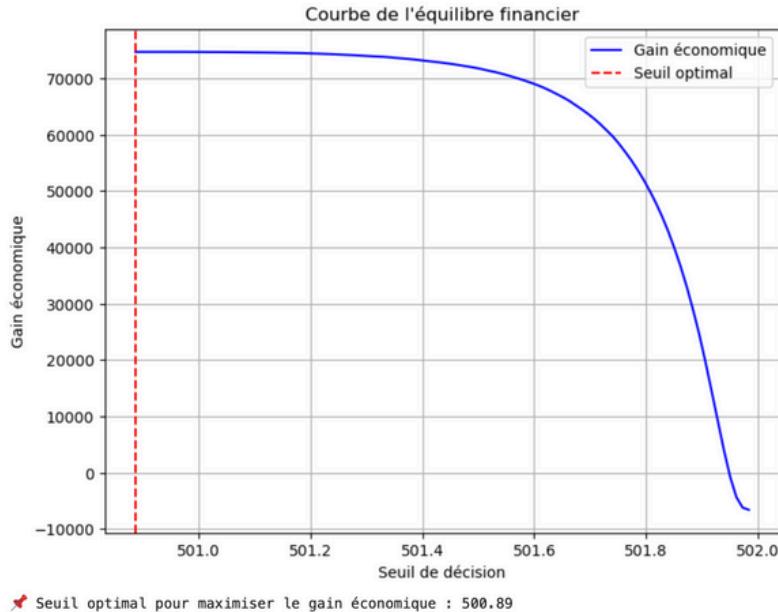
Comparaison des performances après optimisation des hyperparamètres

Comparaison des AUC après optimisation :
 AUC Naïve Bayes : 0.6944
 AUC Arbre de Décision : 0.6968
 AUC K-Nearest Neighbors : 0.5866
 AUC Random Forest (GridSearchCV) : 0.7066
 AUC Random Forest (RandomizedSearchCV) : 0.7065
 AUC Gradient Boosting (RandomizedSearchCV) : 0.7083
 AUC XGBoost (Optuna) : 0.7088
 AUC Réseau de Neurones : 0.7084

Indice de GINI Naïve Bayes : 0.3888
 Indice de GINI Arbre de Décision : 0.3936
 Indice de GINI K-Nearest Neighbors : 0.1732
 Indice de GINI Random Forest Optimisé : 0.4132
 Indice de GINI Random Forest Optimisé (RandomSearch) : 0.4131
 Indice de GINI Gradient Boosting Optimisé : 0.4167
 Indice de GINI XGBoost Optimisé : 0.4175

- **Meilleurs modèles :** XGBoost (AUC = 0.7088, GINI = 0.4175) et Gradient Boosting (AUC = 0.7083, GINI = 0.4167).
- **Random Forest** est aussi performant (AUC = 0.7066, GINI = 0.4132).
- **Recommandation :** Privilégier XGBoost et Gradient Boosting.

IMPACT ÉCONOMIQUE



Objectif du graphique

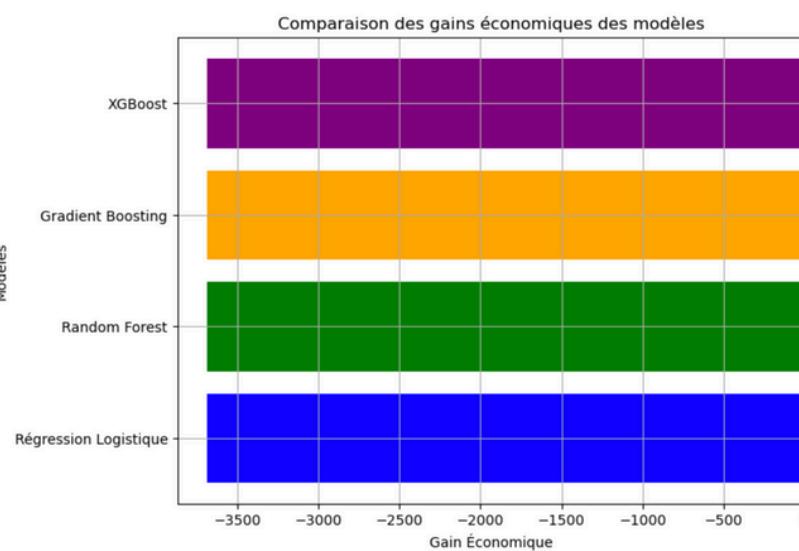
- Visualiser l'**impact du choix du seuil de décision sur les pertes et gains financiers**.
- Identifier un seuil optimal qui maximise la rentabilité en trouvant un bon compromis entre taux d'acceptation des crédits et taux de défaut.

Principaux résultats

- Le modèle utilise la régression logistique avec des variables comme `EXT_SOURCE_2`, `EXT_SOURCE_3`, `YEARS_EMPLOYED`, et `AGE_YEARS` pour prédire un score de solvabilité.
- Les pertes et gains varient fortement selon le seuil choisi : un seuil trop bas entraîne trop de défauts, tandis qu'un seuil trop élevé rejette des clients solvables et réduit la rentabilité.

Enjeux économiques

- Trouver l'**équilibre idéal** entre minimisation des pertes et maximisation du nombre de crédits accordés.
- Ajuster le coût des erreurs pour mieux refléter l'impact économique réel des faux positifs et faux négatifs.



Ce graphique vise à comparer les **performances économiques** des différents modèles de Machine Learning testés, notamment : **Régression Logistique, Random Forest, Gradient Boosting, XGBoost**

Principaux résultats du graphique:

- Tous les modèles aboutissent à une **perte économique similaire d'environ -3684.72**.
- Malgré des performances variées en termes d'AUC et de GINI, aucun modèle ne génère un gain financier positif avec le seuil de décision actuel.

Explications possibles de ces résultats:

1. Mauvais choix du seuil de décision
2. Variables discriminantes insuffisantes
3. Coût des erreurs mal calibré



BNP PARIBAS
PERSONAL FINANCE

INTERPRÉTABILITÉ DES MODÈLES DE MACHINE LEARNING

L'interprétabilité du modèle est étudiée à travers trois approches :

1. Importance des variables pour les modèles à base d'arbres

- Utilisée dans Random Forest, Gradient Boosting et XGBoost, elle repose sur la réduction de l'impureté, évaluant la contribution de chaque variable à la séparation des classes.

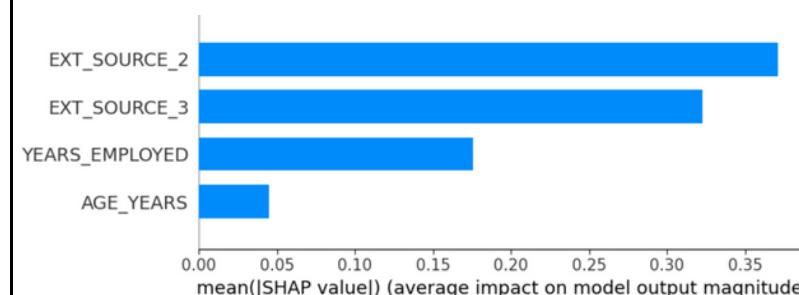
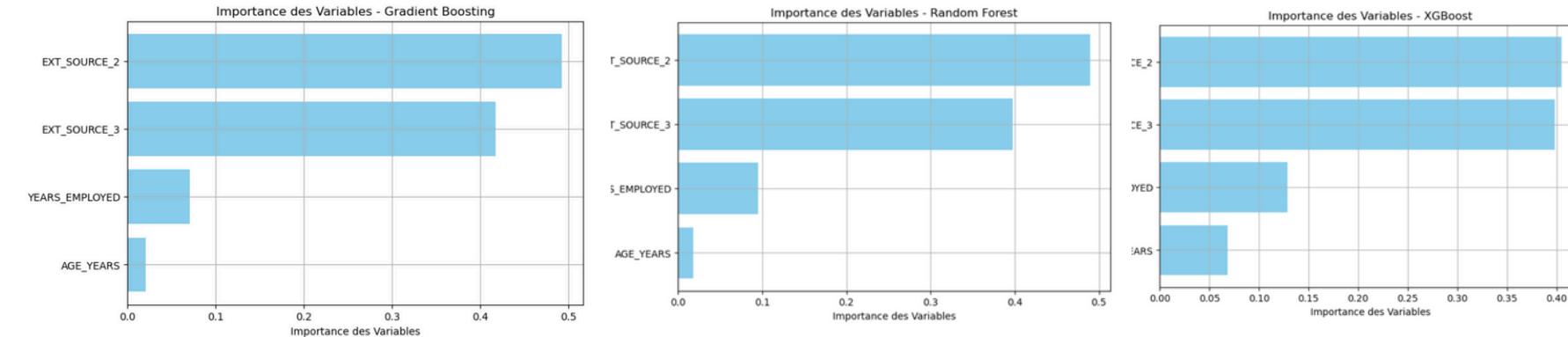
2. Analyse avec SHAP (Shapley Values)

- Fournit une explication détaillée de l'impact de chaque variable sur la prédiction.
- Permet d'identifier les facteurs déterminants pour chaque décision individuelle.

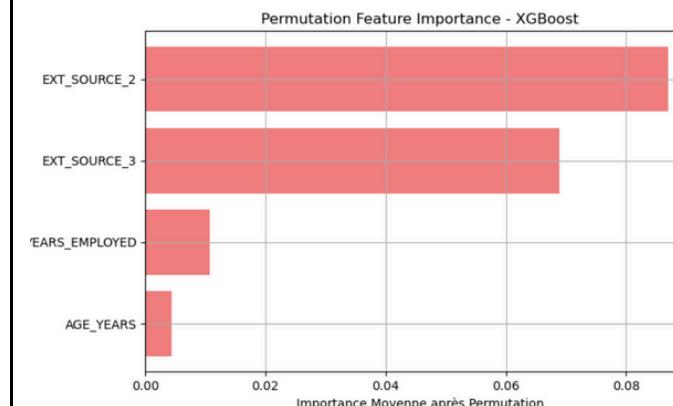
3. Permutation Feature Importance

- Mesure la dégradation des performances lorsque l'on perturbe une variable.
- Indique quelles variables influencent réellement les prédictions en dehors des biais du modèle.

Les modèles **Random Forest**, **Gradient Boosting**, et **XGBoost** montrent que **EXT_SOURCE_2** et **EXT_SOURCE_3** sont les variables les plus influentes dans l'octroi de crédit. **YEARS_EMPLOYED** a un impact modéré et **AGE_YEARS** est la moins significative.



L'analyse SHAP confirme que **EXT_SOURCE_2** et **EXT_SOURCE_3** sont les variables ayant le plus d'impact sur la décision du modèle **Gradient Boosting**, suivies par **YEARS_EMPLOYED** et **AGE_YEARS**. La cohérence avec l'importance des features des modèles d'arbres renforce la fiabilité de ces variables pour l'octroi de crédit.

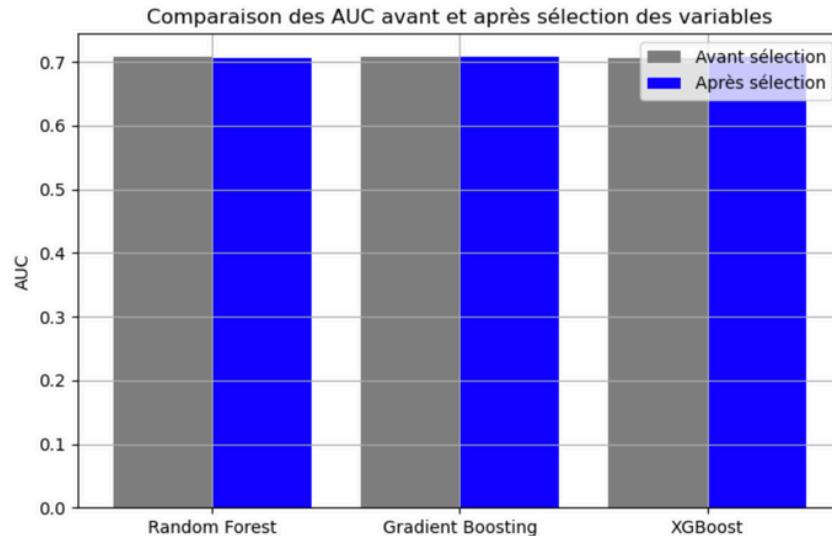


L'analyse de l'importance des variables par **Permutation Feature Importance** sur le modèle **XGBoost** confirme les tendances observées précédemment :

- EXT_SOURCE_2** et **EXT_SOURCE_3** restent les variables les plus influentes.
- YEARS_EMPLOYED** et **AGE_YEARS** ont un impact significatif mais moindre sur la prédiction.

RÉENTRAINEMENT DU MODÈLE ET CONCLUSION

RÉENTRAINEMENT DU MODÈLE



Après l'analyse de l'importance des variables, les modèles ont été réentraînés en utilisant uniquement les variables les plus influentes : **EXT_SOURCE_2**, **EXT_SOURCE_3**, **YEARS_EMPLOYED** et **AGE_YEARS**. Les performances en termes d'AUC et d'indice de GINI restent stables malgré la simplification :

- **Random Forest** : AUC légèrement réduit (0.7060 vs. 0.7067), sans impact majeur.
- **Gradient Boosting** : AUC inchangé (0.7083), confirmant que la réduction n'altère pas sa performance.
- **XGBoost** : AUC quasi identique (0.7086 vs. 0.7088), prouvant l'efficacité du modèle avec moins de variables.

CONCLUSION

Ce projet a permis d'explorer une **approche complète et rigoureuse** du **scoring de crédit**, en intégrant des techniques avancées de **Machine Learning**, une **analyse économique** et des **méthodes d'interprétabilité**.

- **Objectif 1** : Une **préparation et une exploration approfondies** des données ont permis de comprendre la structure des variables et d'anticiper les défis liés à la modélisation.
- **Objectif 2** : La **construction et l'optimisation des modèles** ont mis en avant **XGBoost** et **Gradient Boosting** comme les plus performants, en s'appuyant sur des **critères robustes** comme l'**AUC** et le **GINI**.
- **Objectif 3** : L'**impact économique**, l'**interprétabilité** et le **réentraînement avec les meilleures variables** ont permis d'optimiser la prise de décision en équilibrant **performance et compréhension métier**.

Perspectives :

Les résultats obtenus ouvrent la voie à des **améliorations futures**, notamment en intégrant **d'autres sources de données**, en explorant des **techniques plus avancées** (Deep Learning, Feature Engineering plus poussé) et en mettant en place une **approche plus dynamique** de gestion du risque de crédit.

 Ce projet constitue ainsi une **base solide** pour une **meilleure prise de décision** dans le domaine du **scoring de crédit**.