Projet 4 - Data Scientist

Jean-Baptiste Vayssade

Implémentez un modèle de scoring

2025

Introduction

Modèle de scoring crédit

Prêt à dépenser

- Projet 4 Parcours Data Scientist
- Python | Jupyter Notebook | Machine Learning
- Scikit-learn, SHAP, Pandas, NumPy
- OpenClassrooms - CentraleSupélec

Plan de la présentation

Sommaire

1. Contexte et problématique métier
2. Présentation et exploration des données
3. Préparation des données et feature engineering
4. Modélisation et score métier
5. Comparaison des modèles et résultats
6. Interprétabilité du modèle
7. Conclusion et perspectives

1. Contexte

Contexte et problématique

Entreprise : “Prêt à dépenser”	• Crédits à la consommation • Clientèle : peu ou pas d’historique de prêt
Besoin	• Outil de **scoring crédit** • Probabilité de remboursement • Aide à la décision : accord/refus
Utilisateurs	• Chargés de relation client • Besoin d’**interprétabilité**

Enjeux métier spécifiques

1. Déséquilibre des classes

- 92% bons clients vs 8% défauts
- Techniques de rééquilibrage nécessaires

2. Déséquilibre du coût

- **Faux Négatif (FN)** : Perte en capital
- **Faux Positif (FP)** : Manque à gagner
- **Hypothèse : Coût FN = 10 × Coût FP**

3. Score métier personnalisé

- Métrique adaptée aux coûts réels
- Optimisation du seuil de décision

2. Données

Source des données

Kaggle : “Home Credit Default Risk”

- Table principale : application_train.csv
  - 307 511 clients
  - 122 variables
  - TARGET : 0 = remboursé, 1 = défaut

Tables secondaires

- bureau.csv : Historique tous organismes
- previous_application.csv : Emprunts Home Credit
- POS_CASH_balance.csv : Crédits cash
- credit_card_balance.csv : Revolving
- installments_payments.csv : Historique paiements

Distribution de la cible

Déséquilibre important

Classe 0 (bons) : 91 907 (91,9%)

Classe 1 (défauts) : 8 093 (8,1%)

Total : 100 000 observations

Conséquences

Risque de modèle biaisé

Nécessité de techniques de rééquilibrage

SMOTE, pondération, sous-échantillonnage

Variables les plus corrélées

Top 5 corrélations négatives (bon signe)

- EXT_SOURCE_3 : -0.178
- EXT_SOURCE_2 : -0.162
- EXT_SOURCE_1 : -0.153
- NAME_EDUCATION_TYPE_Higher : -0.058
- CODE_GENDER_F : -0.055

Top 5 corrélations positives (risque)

- DAYS_BIRTH : 0.077
- DAYS_EMPLOYED : 0.073
- REGION_RATING_CLIENT_W_CITY : 0.063
- REGION_RATING_CLIENT : 0.062
- NAME_INCOME_TYPE_Working : 0.057

Analyse par âge

Observations

- Distribution : 21 à 70 ans
- Moyenne : 43,9 ans

Tendance

- Jeunes (20-30 ans) : ~12% défaut
- Seniors (65-70 ans) : ~3,6% défaut
- **L’âge est un facteur prédictif important**

Anomalies détectées

DAYS_EMPLOYED : 365 243 jours	• 17 865 observations (17,9%) • Probablement retraités/chômeurs
Traitement	• Variable DAYS_EMPLOYED_ANOM (flag) • Remplacement par NaN • Conservation de l'information
Impact	• Anomalies : 5,5% défaut • Non-anomalies : 8,7% défaut

3. Préparation

Gestion des valeurs manquantes

Constat

- 67 colonnes avec NaN (sur 122)
- Jusqu'à 70% de NaN pour certaines

Stratégie

- Seuil : 20% maximum de NaN
- Variables numériques : **KNN imputation (K=5)**
- Variables catégorielles : **Mode**
- Suppression des trop lacunaires

Feature Engineering

Variables créées	
CREDIT_DEBT_RATIO	•AMT_CREDIT / AMT_INCOME_TOTAL
PART_SOCIALE	•AMT_INCOME_TOTAL / (CNT_CHILDREN + 1)
YEARS_BIRTH	•DAYS_BIRTH / 365
Résultat	•122 → 240 features (après encoding) •Sélection → 27 features finales

Preprocessing

Encodage

1. **Label Encoding** : variables binaires (3 cols)
2. **One-Hot Encoding** : catégorielles multi-classes
   - NAME_INCOME_TYPE → 8 variables
   - NAME_EDUCATION_TYPE → 5 variables
   - OCCUPATION_TYPE → 18 variables

Transformation

- Log transformation sur AMT_INCOME_TOTAL
- Normalisation/Standardisation

Gestion du déséquilibre

Solutions testées

1. **SMOTE** : Over-sampling classe minoritaire
2. **Sous-échantillonnage** : Réduction classe majoritaire
3. **Pondération** : class_weight='balanced'

Approche retenue (combinée)

- Sur-échantillonnage modéré (50-100%)
- Sous-échantillonnage modéré (0-50%)
- Pondération pour équilibrer le reste

4. Modélisation

Score métier personnalisé

Matrice de coûts

	Prédit : 0	Prédit : 1
Réel : 0	TN : +1	FP : -1
Réel : 1	FN : **-10**	TP : +2

Formule

```
Score = (1 × TN) + (2 × TP) + (-1 × FP) + (-10 × FN)
```

Normalisation : [0, 1]

Modèles testés

Baseline

- Régression Logistique
- AUC : ~0.72

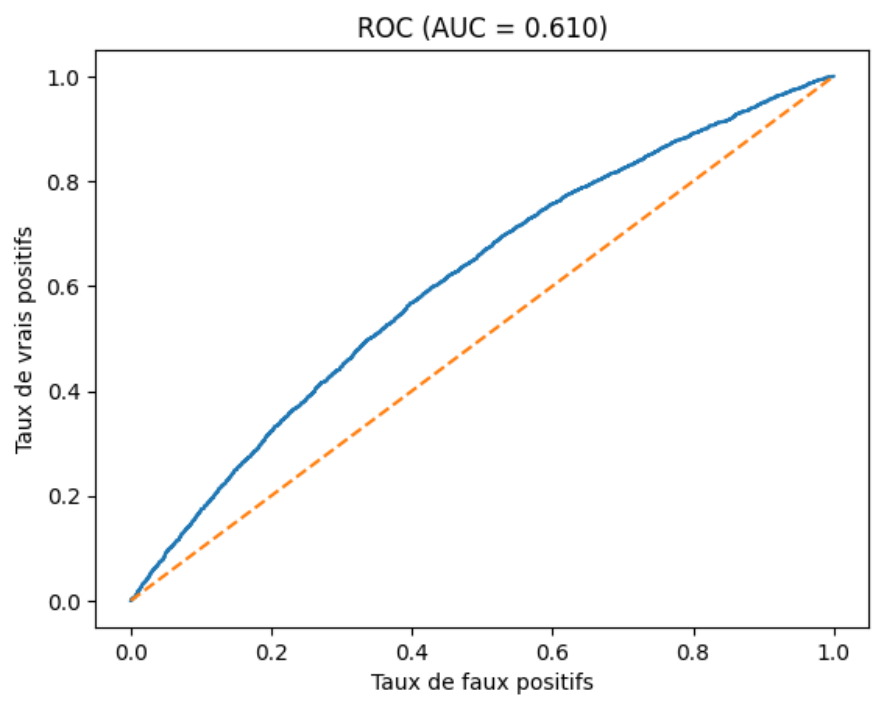
Modèles avancés

1. **Random Forest Classifier**
2. **LightGBM Classifier** ✓
3. XGBoost Classifier

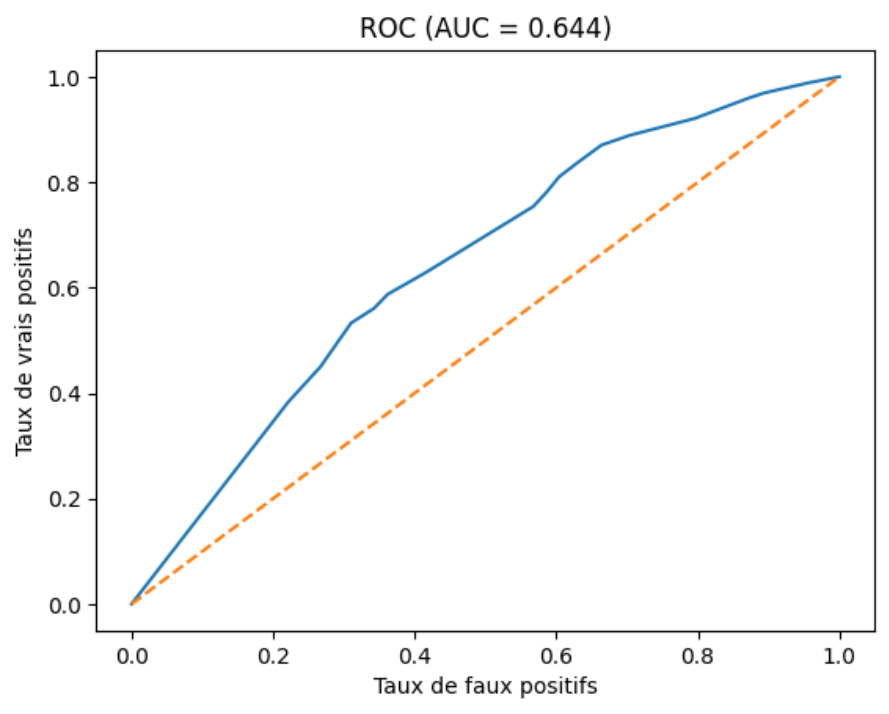
Méthodologie

- Cross-validation : 5 StratifiedKFolds
- GridSearch / RandomizedSearch
- Métriques : Score métier, AUC, Accuracy

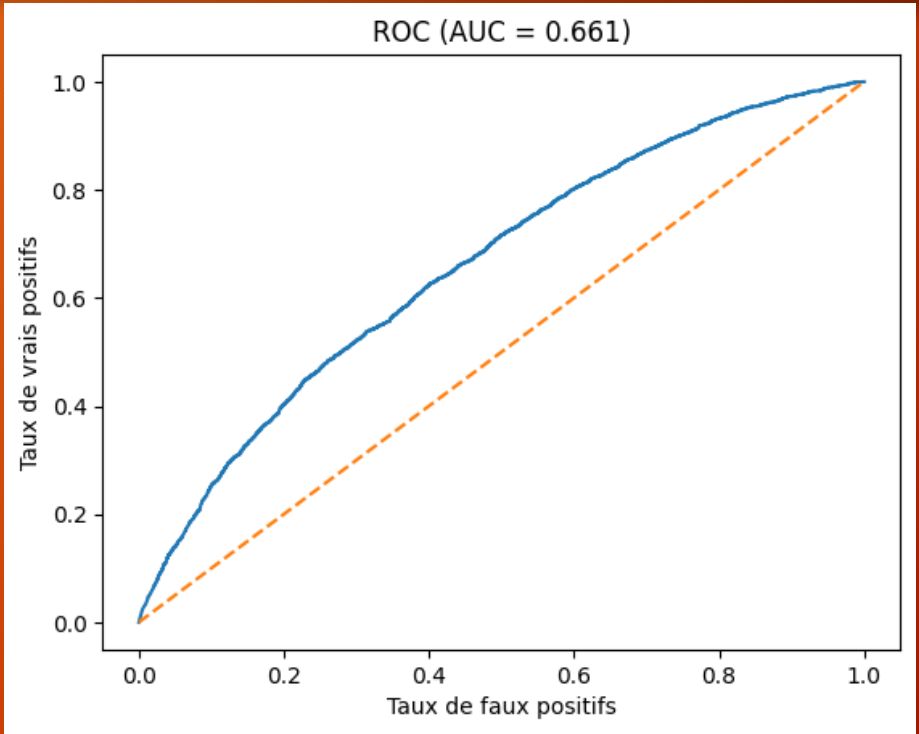
AUC de la régression logistique



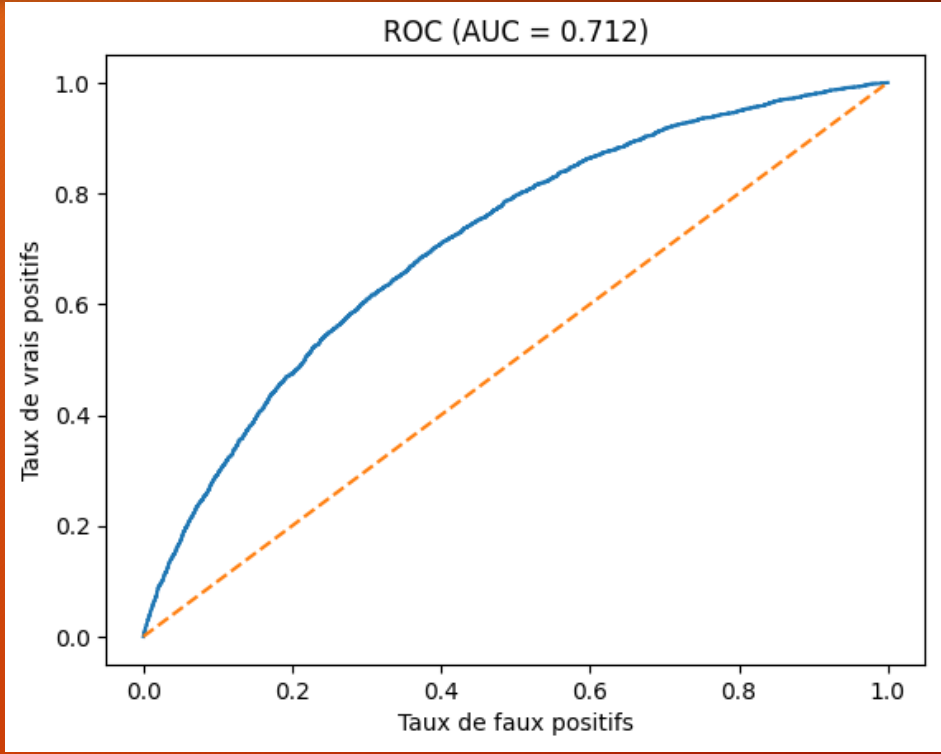
AUC Decision Tree



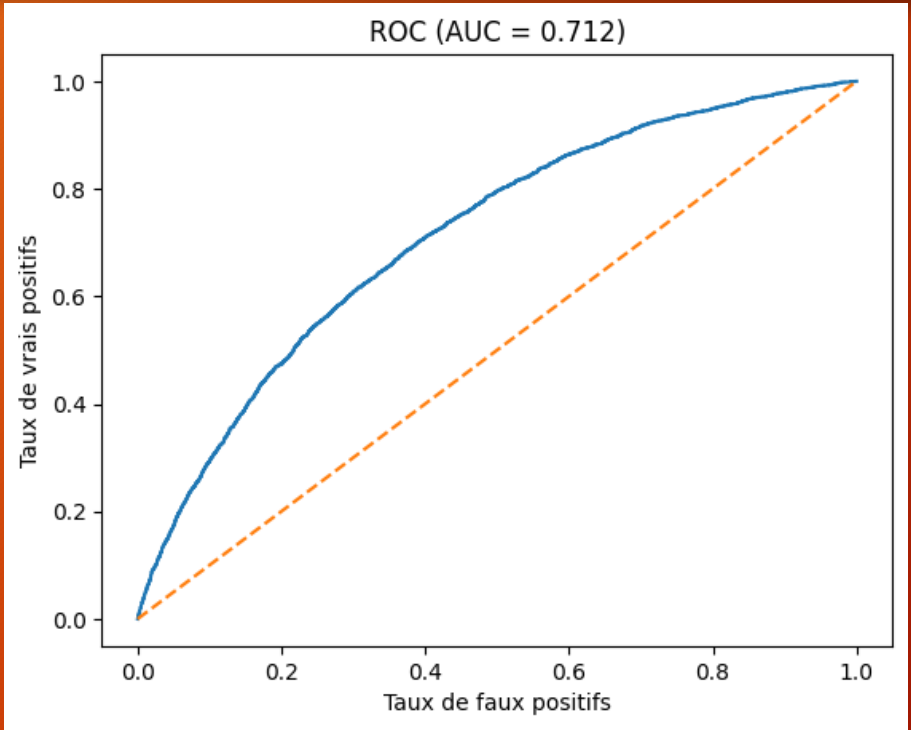
AUC Random Forest



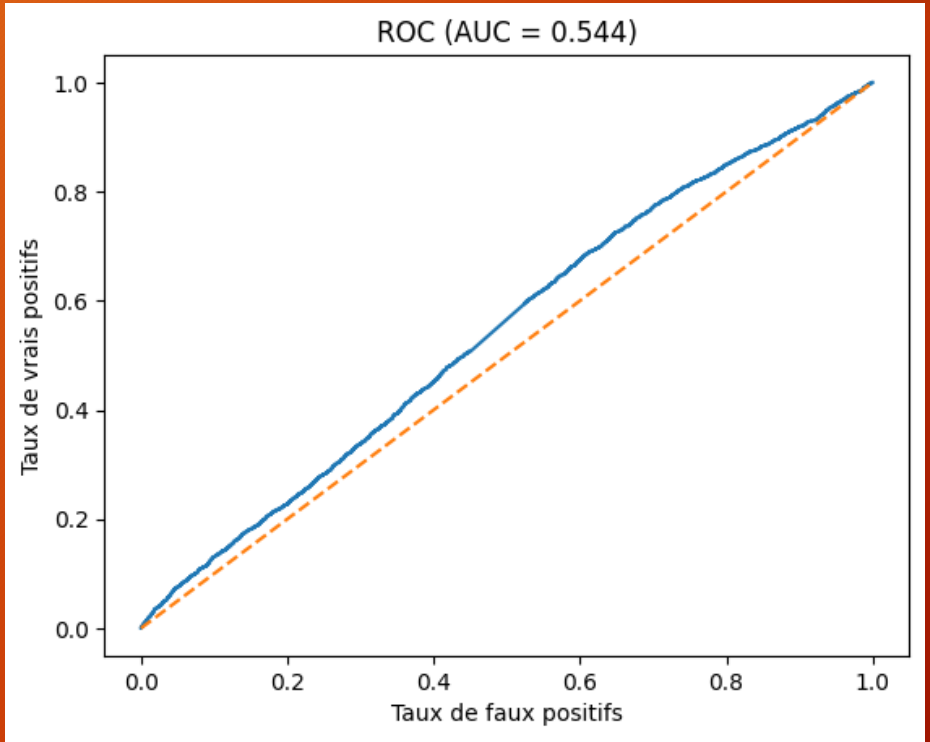
AUC Gradient Boosting



AUC KNN Classifier



AUC Support Vector Machine



Optimisation des hyperparamètres

Pipeline testé

1. **Préparation données**
   - Over/under-sampling
   - Scalers variés
2. **Optimisation modèle**
   - GridSearchCV
   - Métrique : Score métier
3. **Paramètres LightGBM**
   - n_estimators
   - max_depth
   - learning_rate

5. Résultats

Comparaison des modèles

Modèle	Score Métier	AUC	Accuracy
Régression Log.	0.675	0.720	0.68
Random Forest	0.679	0.732	0.71
LightGBM	**0.693**	**0.741**	**0.72**

Matrice LightGBM

	Prédit : 0	Prédit : 1
Réel : 0	22 036	8 580
Réel : 1	966	1 652

Courbe ROC et AUC

Performance

- Régression Logistique : AUC = 0.720
- Random Forest : AUC = 0.732
- **LightGBM : AUC = 0.741 ✓**

Validation

- AUC < 0.82 → Pas d'overfitting ✓
- Bonne capacité de discrimination

Seuil optimal

- Seuil standard : 0.5
- **Seuil optimisé : 0.45**
- Score métier maximisé

Optimisation du seuil

Impact du seuil

Seuil	Score Métier	FN	FP
0.30	0.650	450	15000
0.45	**0.693**	**966**	**8580**
0.50	0.680	1200	6500
0.60	0.620	1800	4000

Seuil optimal : 0.45

- Balance FN/FP optimale
- Maximise le score métier

6. Interprétabilité

Importance des variables

Top 10 features (LightGBM)

1. EXT_SOURCE_2 : 15,3%
2. EXT_SOURCE_3 : 14,7%
3. EXT_SOURCE_1 : 12,1%
4. DAYS_BIRTH : 8,9%
5. CREDIT_DEBT_RATIO : 7,2% *
6. DAYS_EMPLOYED : 6,5%
7. AMT_CREDIT : 5,8%
8. AMT_INCOME_TOTAL : 4,9%
9. PART_SOCIALE : 3,7% *
10. OWN_CAR_AGE : 3,2%

* Feature engineered

SHAP - Interprétabilité locale

Méthode SHAP

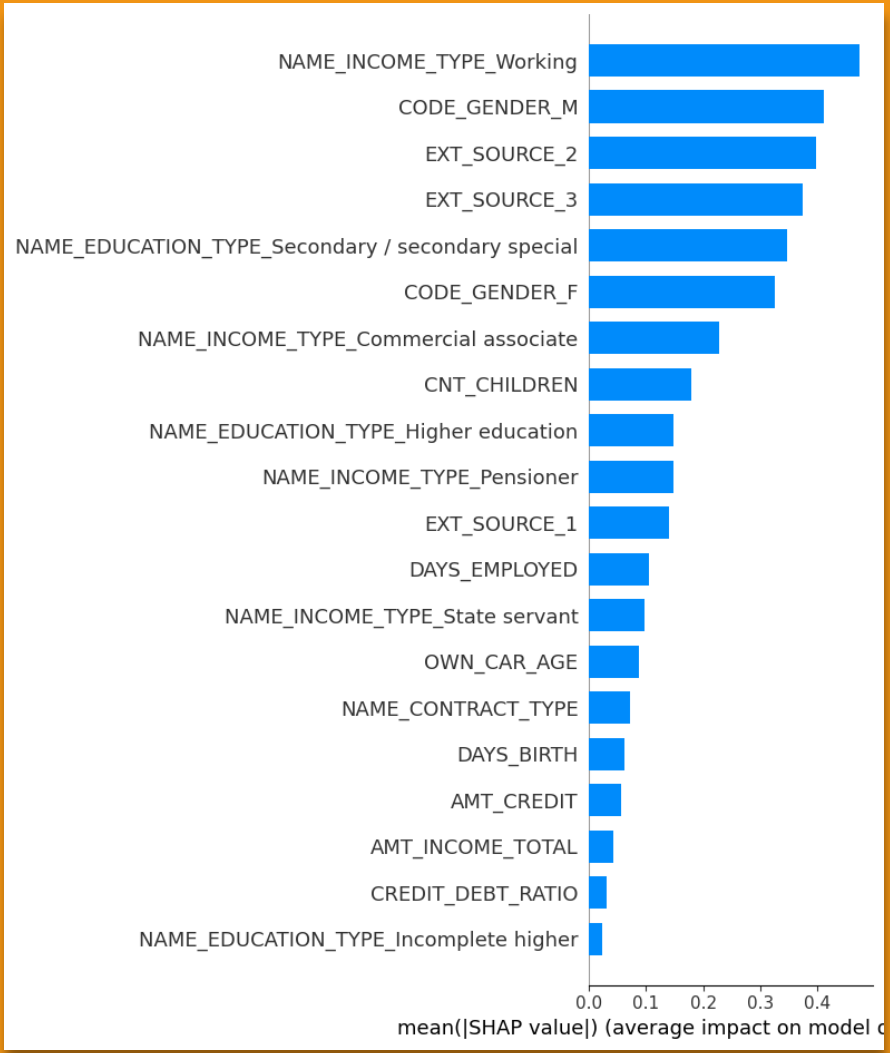
Explication individuelle des prédictions

Attribution de l'impact de chaque feature

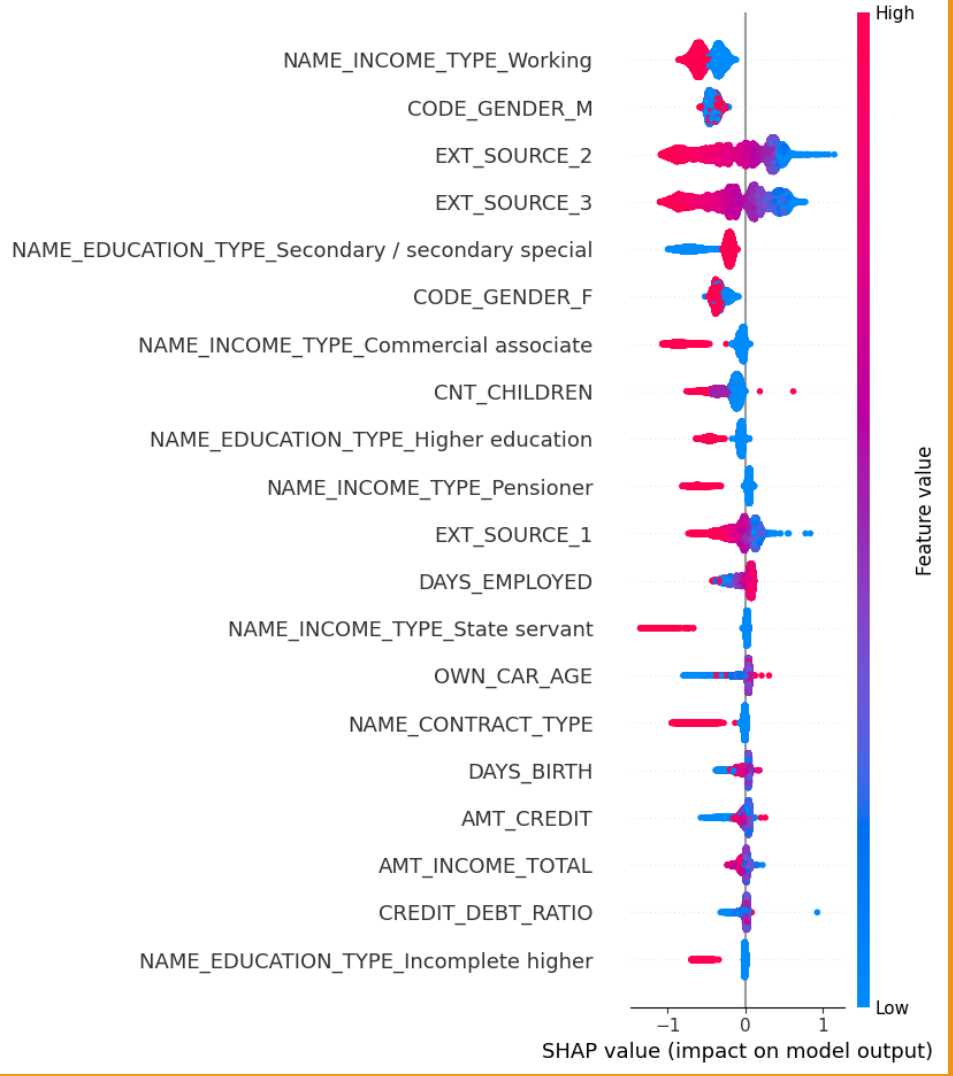
Exemple client faible risque

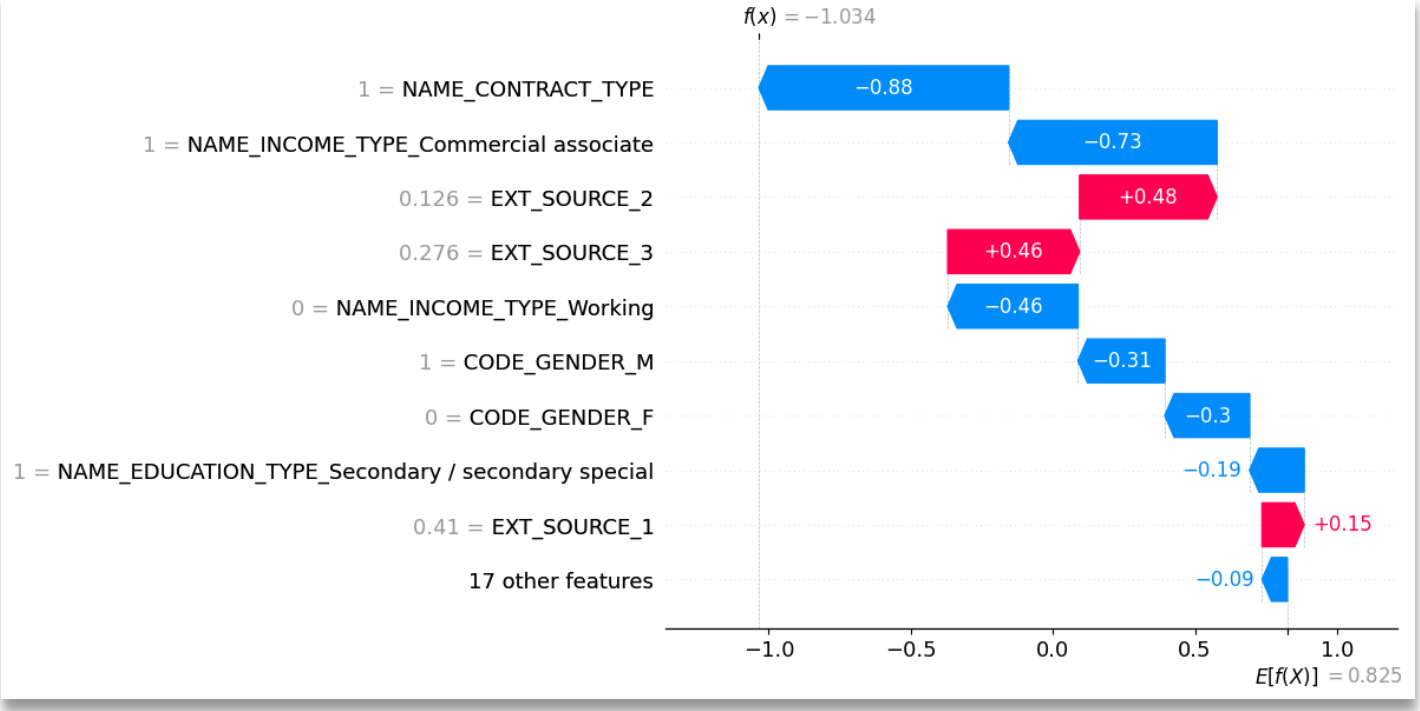
Feature	Impact SHAP
EXT_SOURCE_2: 0.85	-0.15 ✓
DAYS_BIRTH: -15000	-0.08 ✓
CREDIT_DEBT: 1.8	-0.06 ✓

SHAP


NAME_INCOME_TYPE_Working
CODE_GENDER_M
EXT_SOURCE_2
EXT_SOURCE_3
NAME_EDUCATION_TYPE_Secondary / secondary special
CODE_GENDER_F
NAME_INCOME_TYPE_Commercial associate
CNT_CHILDREN
NAME_EDUCATION_TYPE_Higher education
NAME_INCOME_TYPE_Pensioner
EXT_SOURCE_1
DAYS_EMPLOYED
NAME_INCOME_TYPE_State servant
OWN_CAR_AGE
NAME_CONTRACT_TYPE
DAYS_BIRTH
AMT_CREDIT
AMT_INCOME_TOTAL
CREDIT_DEBT_RATIO
NAME_EDUCATION_TYPE_Incomplete higher
0.0
0.1
0.2
0.3
0.4
mean(|SHAP value|) (average impact on model

SHAP


NAME_INCOME_TYPE_Working
CODE_GENDER_M
EXT_SOURCE_2
EXT_SOURCE_3
NAME_EDUCATION_TYPE_Secondary / secondary special
CODE_GENDER_F
NAME_INCOME_TYPE_Commercial associate
CNT_CHILDREN
NAME_EDUCATION_TYPE_Higher education
NAME_INCOME_TYPE_Pensioner
EXT_SOURCE_1
DAYS_EMPLOYED
NAME_INCOME_TYPE_State servant
OWN_CAR_AGE
NAME_CONTRACT_TYPE
DAYS_BIRTH
AMT_CREDIT
AMT_INCOME_TOTAL
CREDIT_DEBT_RATIO
NAME_EDUCATION_TYPE_Incomplete higher
High
Feature value
Low
−1
0
1
SHAP value (impact on model output)


f(x) = −1.034
1 = NAME_CONTRACT_TYPE
1 = NAME_INCOME_TYPE_Commercial associate
0.126 = EXT_SOURCE_2
0.276 = EXT_SOURCE_3
0 = NAME_INCOME_TYPE_Working
1 = CODE_GENDER_M
0 = CODE_GENDER_F
1 = NAME_EDUCATION_TYPE_Secondary / secondary special
0.41 = EXT_SOURCE_1
17 other features
−0.88
−0.73
+0.48
+0.46
−0.46
−0.31
−0.3
−0.19
+0.15
−0.09
−1.0
−0.5
0.0
0.5
1.0
E[f(X)] = 0.825

SHAP

Exemple client risque élevé

Feature	Impact SHAP
EXT_SOURCE_2: 0.15	+0.22 ⚠
DAYS_BIRTH: -8000	+0.12 ⚠
CREDIT_DEBT: 8.5	+0.18 ⚠

Validation et contrôle

Tests effectués

1. **Absence de data leakage**
   - AUC = 0.741 < 0.82 ✓
2. **Stabilité CV**
   - Écart-type : ±0.015
   - Résultats cohérents ✓
3. **Cohérence métier**
   - Features logiques ✓
4. **Généralisation**
   - Test : 0.693
   - Train : 0.698
   - Diff : 0.005 ✓

7. Conclusion

Limites identifiées

Limites techniques

- Taux FN encore élevé (36,9%)
- Taux FP : 28%
- Dépendance scores externes (42%)

Limites métier

- Hypothèse coût 10:1 à valider
- Modèle à ré-entraîner régulièrement
- Formation utilisateurs nécessaire

Perspectives d’amélioration

Améliorations possibles

1. **Feature engineering avancé**
   - Agrégations temporelles
   - Features d’interaction
2. **Algorithmes complémentaires**
   - Stacking/Blending
   - Neural Networks
3. **Optimisation coût métier**
   - Affiner ratios avec business
   - Optimisation multi-objectif
4. **Mise en production**
   - API de scoring
   - Interface utilisateur
   - Monitoring

Synthèse

Objectifs atteints

- Modèle de scoring développé
- Score métier personnalisé
- Gestion du déséquilibre
- Interprétabilité assurée
- Performance : AUC = 0.741

Modèle retenu : LightGBM

- Seuil optimal : 0.45
- 27 features finales
- Score métier : 0.693

Prochaines étapes :

- Validation métier
- Tests pilotes
- Déploiement progressif
- Formation utilisateurs

Merci !

Questions ?