Projet Machine Learning

PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Prédiction du Défaut de Paiement

Groupe d'étudiants:
SAWADOGO Issa, KPOGNON Koffi, FOMBA Abdou,

Nom de l'Enseignante: Lina FAIK

Du Sorbonne Data Analytics - Session 6 (2025-2026)

Sommaire

- I-Contexte et Problématique
- II-Les variables
- III-Prétraitement des données
- IV-Méthodologie
- V-Résultats et discussions

I-Contexte et Problématique

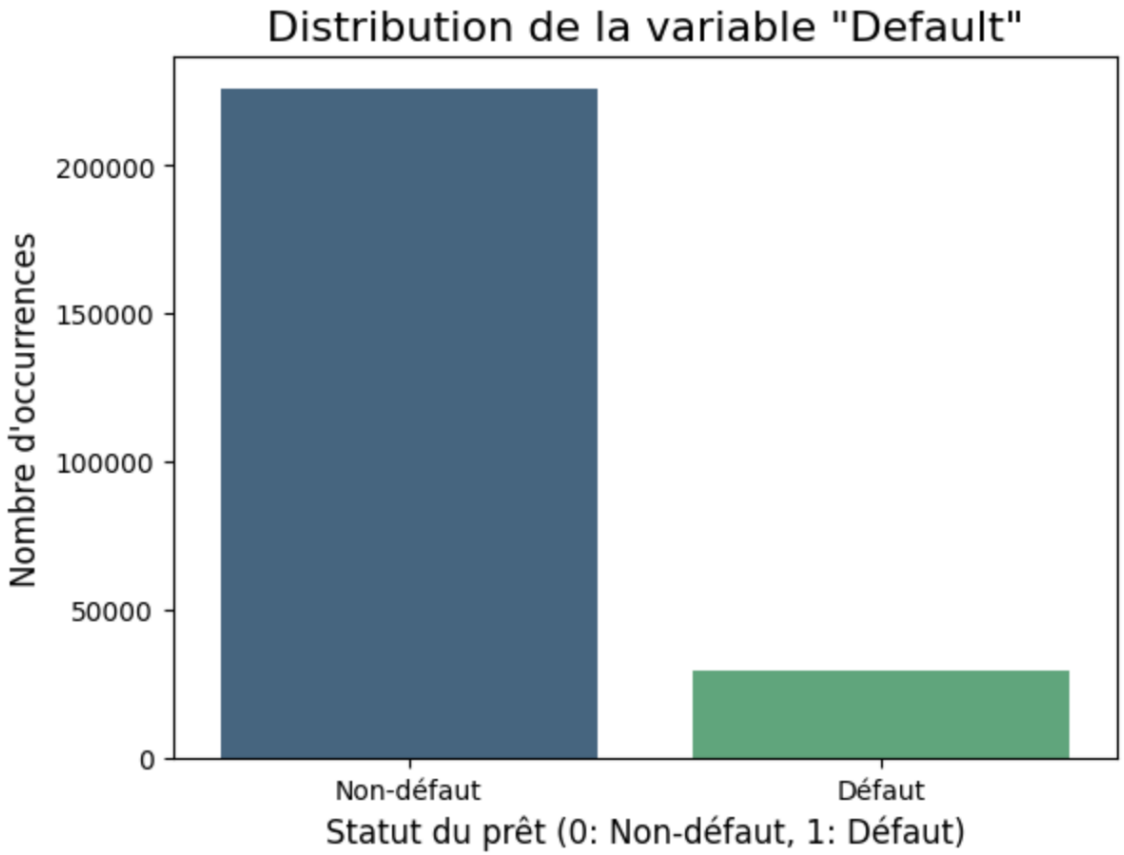
- Le défaut de paiement est l'incapacité d'un emprunteur à honorer ses engagements financiers.

Enjeux :

- Réduction des pertes pour les institutions financières.
- Amélioration de la gestion du risque.
- L'objectif de notre étude vise à prédire le risque de défaut de paiement à l'aide de plusieurs facteurs explicatifs et en usant des modèles connus dans la littérature et à l'aide des récentes techniques du machine learning.

II-Les variables

Target: Default



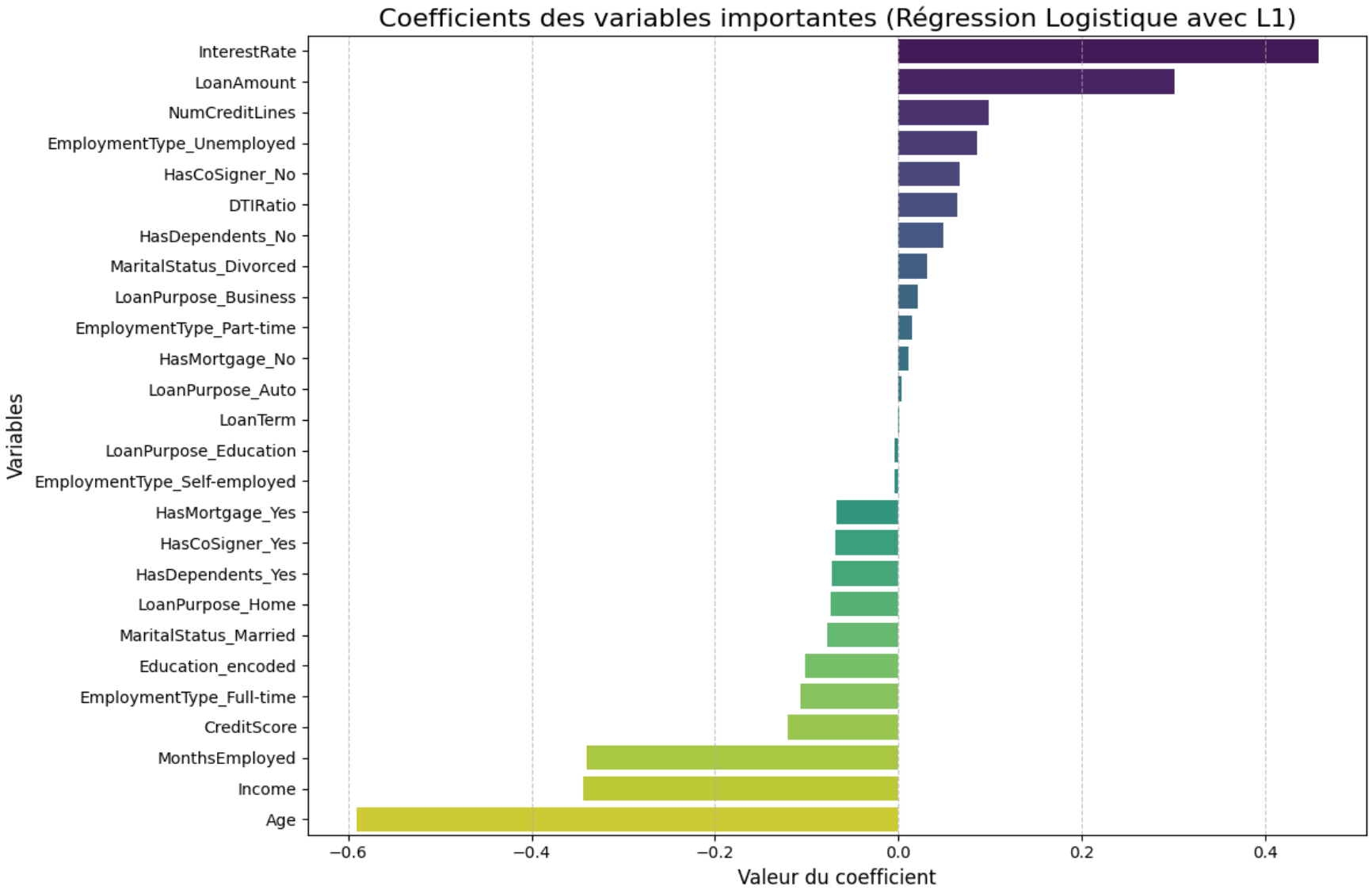
II-Les variables

Variable	Définition	Valeurs non-nulles	Type de données
LoanID	Identifiant unique du prêt	255347	object
Age	Âge du demandeur	255347	int64
Income	Revenu annuel du demandeur	255347	int64
LoanAmount	Montant du prêt demandé	255347	int64
CreditScore	Cote de crédit du demandeur	255347	int64
MonthsEmployed	Nombre de mois d'emploi du demandeur	255347	int64
NumCreditLines	Nombre de lignes de crédit ouvertes	255347	int64
InterestRate	Taux d'intérêt du prêt	255347	float64
LoanTerm	Durée du prêt en mois	255347	int64
DTIRatio	Ratio dette sur revenu	255347	float64
Education	Niveau d'éducation du demandeur	255347	object
EmploymentType	Type d'emploi du demandeur	255347	object
MaritalStatus	État civil du demandeur	255347	object
HasMortgage	Indicateur de l'existence d'un prêt hypothécaire	255347	object
HasDependents	Indicateur de l'existence de personnes à charge	255347	object
LoanPurpose	Raison de la demande de prêt	255347	object
HasCoSigner	Indicateur de l'existence d'un cosignataire	255347	object

III-Prétraitement

- Vérification des valeurs manquantes, des données dupliquées de la base et de la typologie des variables.
- Analyse graphique des donnes.
- Encodage des variables catégorielles.
- Choix des variables les plus importantes à partir d'une régression Logistique avec une régularisation L1

Régression Logistique avec une régularisation L1 et choix des variables les plus importantes



IV-Méthodologie

- Modèles à tester :

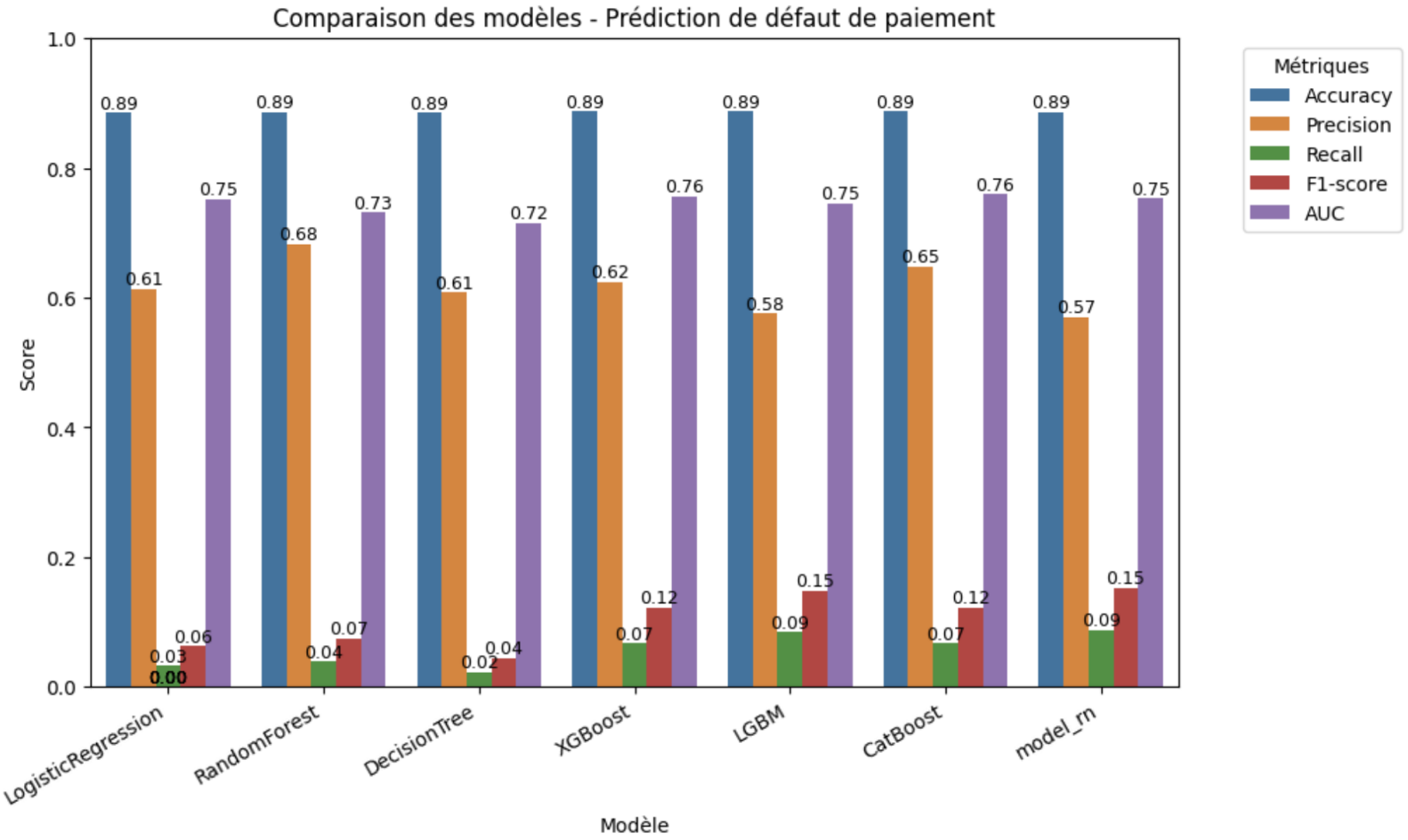
Modèle	Performance	Interprétabilité	Robustesse	Complexité	Cas d'Usage Recommandé
Régression Logistique	▲▲△	★★★	★★★	△△△	Base de référence, exigence réglementaire forte
Arbres de Décision	▲△△	★★★	▲△△	▲△△	Prototypage, règles métier simples
Random Forest	▲▲▲	▲▲△	★★★	▲▲△	Benchmark solide, équilibre performance/interprétation
XGBoost/LightGBM/ CatBoost	★★★	▲▲△	▲▲▲	▲▲▲	Meilleure performance pure
Réseaux de Neurones	★★★	△△△	▲▲△	★★★	Très grands volumes de données (big data)
Stacking Ensemble	★★★	△△△	▲▲▲	★★★	Projets avancés où chaque gain de performance compte

Légende

★★★ Excellent | ▲▲▲ Très Bon | ▲▲△ Bon | ▲△△ Moyen | △△△ Faible

- Évaluation : F1-score,Recall,Precision, AUC-ROC, (Accuracy)

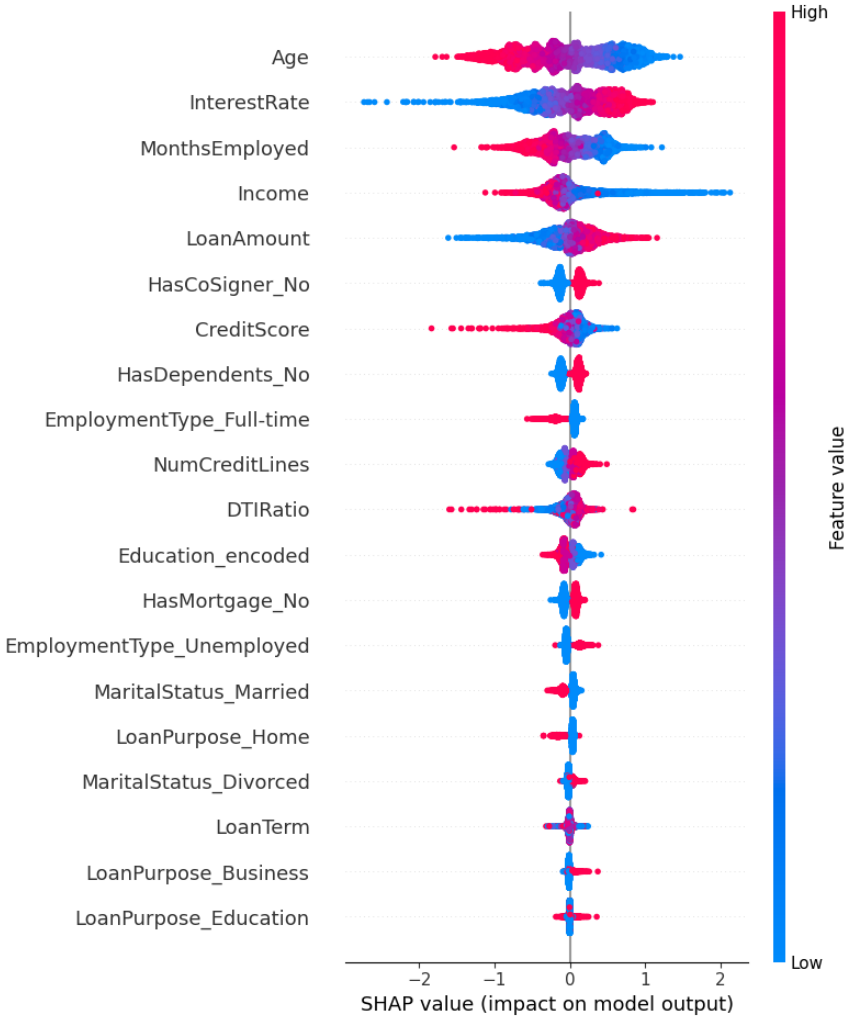
V-Résultats et Discussions



V-Résultats et Discussions

Le choix en terme de performance a porté sur les modèles de réseaux de neurones et le LightGBM grâce aux métriques F1-score et AUC qui sont les plus élevés pour les deux modèles.

Modèle LightGBM + SHAP



Conclusion et Perspectives

- Apres avoir implémenté les différents modèles, le choix en terme de performance a porté sur les modèles de réseaux de neurones et LightGBM.
- Mais en termes de perspectives, on pourrait faire un stacking ensemble pour avoir des métriques plus intéressantes.

Conclusion et Perspectives

- Tous les modèles ont un problème de rappel trop faible: il faudra ajuster :

- Seuil de décision (par défaut on utilise 0.5; on peut essayer 0.3 ou 0.25 pour augmenter le rappel).

- Rééchantillonnage (SMOTE, oversampling) pour corriger le déséquilibre de classes.

- Optimisation coût-sensible (pondérer davantage les défauts dans la loss function).

PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Merci pour votre attention