PANTHÉON SORBONNE
UNIVERSITÉ PARIS 1

Projet Machine Learning

Prédiction d'un défaut de paiement bancaire

DATA ANALYTICS - SESSION 6

Présentées par

Patricia KOTO
Waï LEKONE ANTA

https://github.com/Patricia-Koto/sda_ML

PLAN

1. Présentation du sujet
Contexte et Problématique

2. Méthodologie
Objectifs et Process

3. Déploiement
Modèles et Résultats

4. Perspectives
Analyse et Axe d’amélioration

Présentation du sujet

CONTEXTE

Les défauts de paiement, ou le non-respect des obligations de paiement ont un grand impact très significative sur le système de fonctionnement des banques ou des institutions financières,

Ces derniers peuvent laisser les prêteurs dans une position précaire, affectant leur stabilité financière, leur rentabilité et l'ensemble de leurs opérations commerciales

Face à cette situation, il est crucial aux banques mettre en place un système permettant d'estimer la probabilité qu'un emprunteur ne rembourse pas son prêt.

PROBLÉMATIQUE

Quelles sont les caractéristiques qui peuvent influencer le défaut de paiement ?

Méthodologie

Objectifs

Notre démarche consiste à **garantir la viabilité économique des prêteurs** pour permettre aux banques de prendre des **décisions plus éclairées et d'éviter des prêts à risque.**

- **Créer un modèle prédictif : Prévoir si un demandeur de prêt sera en défaut de paiement.**
- **Identifier les facteurs clés : Déterminer les variables qui influencent le risque de défaut.**
- **Formuler des recommandations : Proposer des stratégies pour atténuer le risque financier.**

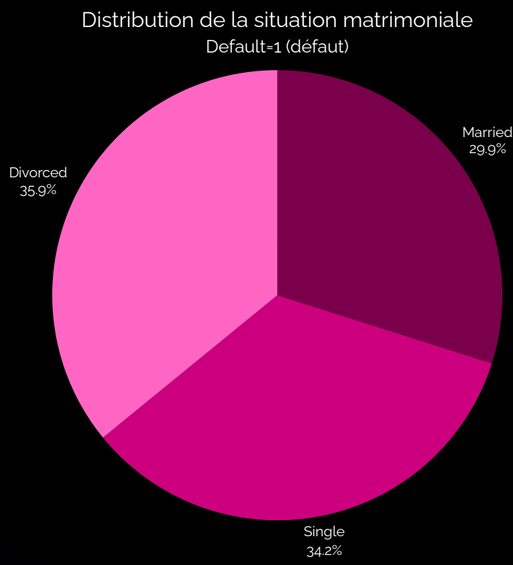
Présentation du sujet

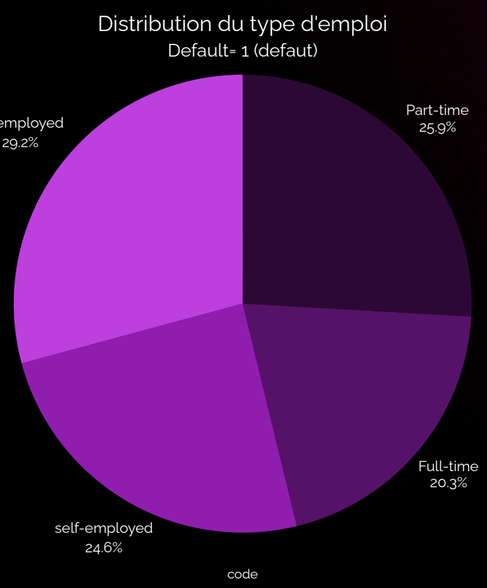
Variable	Explication
LoanID	Identifiant unique du prêt (chaîne de caractères).
Age	Âge de l'emprunteur (en années).
Income	Revenu annuel de l'emprunteur (en unités monétaires, ex. dollars).
LoanAmount	Montant du prêt demandé.
CreditScore	Score de crédit de l'emprunteur (ex. FICO, évalue la solvabilité).
MonthsEmployed	Nombre de mois travaillés chez l'employeur actuel.
NumCreditLines	Nombre total de lignes de crédit (cartes, prêts, etc.) que possède l'emprunteur.
InterestRate	Taux d'intérêt du prêt (en %).
LoanTerm	Durée du prêt (en mois).
DTIRatio	*Debt-to-Income Ratio* = ratio dette/revenu de l'emprunteur.
Education	Niveau d'éducation de l'emprunteur (ex. lycée, licence, master...).
EmploymentType	Type d'emploi (ex. salarié, indépendant, fonction publique...).
MaritalStatus	Statut matrimonial (ex. célibataire, marié, divorcé...).
HasMortgage	Indique si l'emprunteur a déjà un prêt hypothécaire (oui/non).
HasDependents	Indique si l'emprunteur a des personnes à charge (oui/non).
LoanPurpose	Raison du prêt (ex. achat voiture, études, consolidation de dettes...).
HasCoSigner	Indique si le prêt a un co-signataire/garant (oui/non).
Default	Variable cible : 1 si l'emprunteur est en défaut de paiement, 0 sinon.

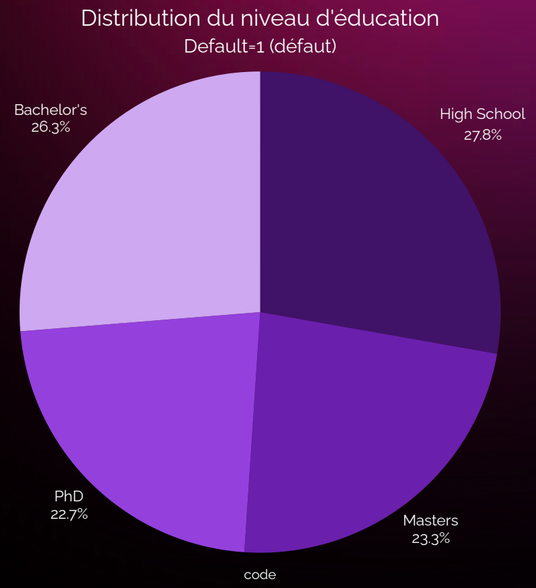
Jeu de données

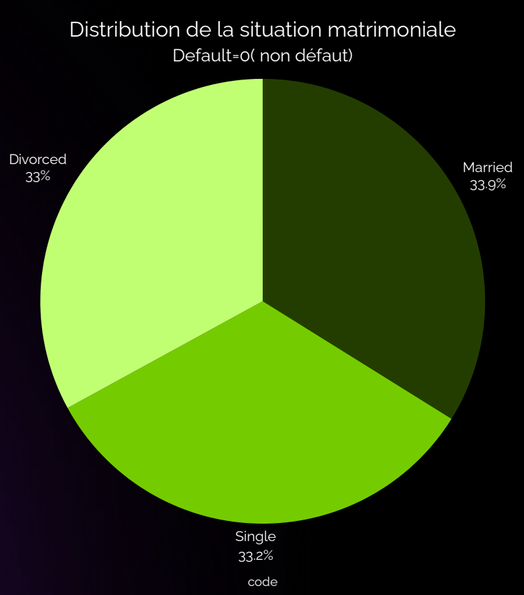
- "**loan_default_data.csv",** disponible sur kaggle
- **Plus de 250 000 observations**
- Informations personnelles **: âge, revenu, situation familiale**
- Informations financières **: montant du prêt, côte de crédit, taux d'intérêt**
- Variable cible binaire : **Default (1** défaut et **0** sinon**)**

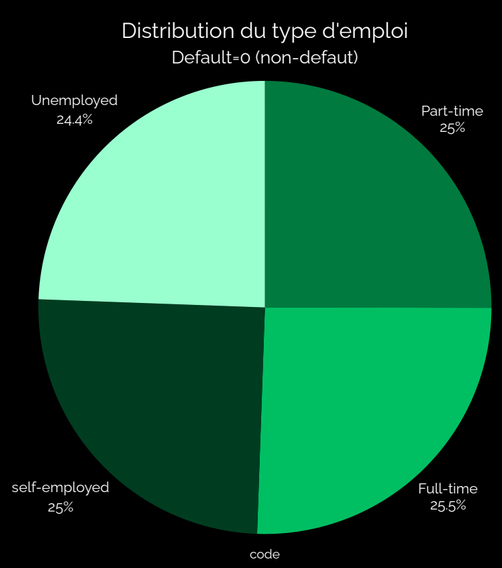
Présentation du sujet

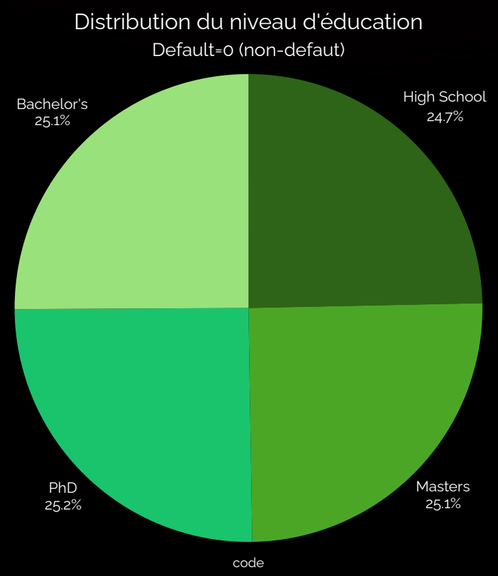




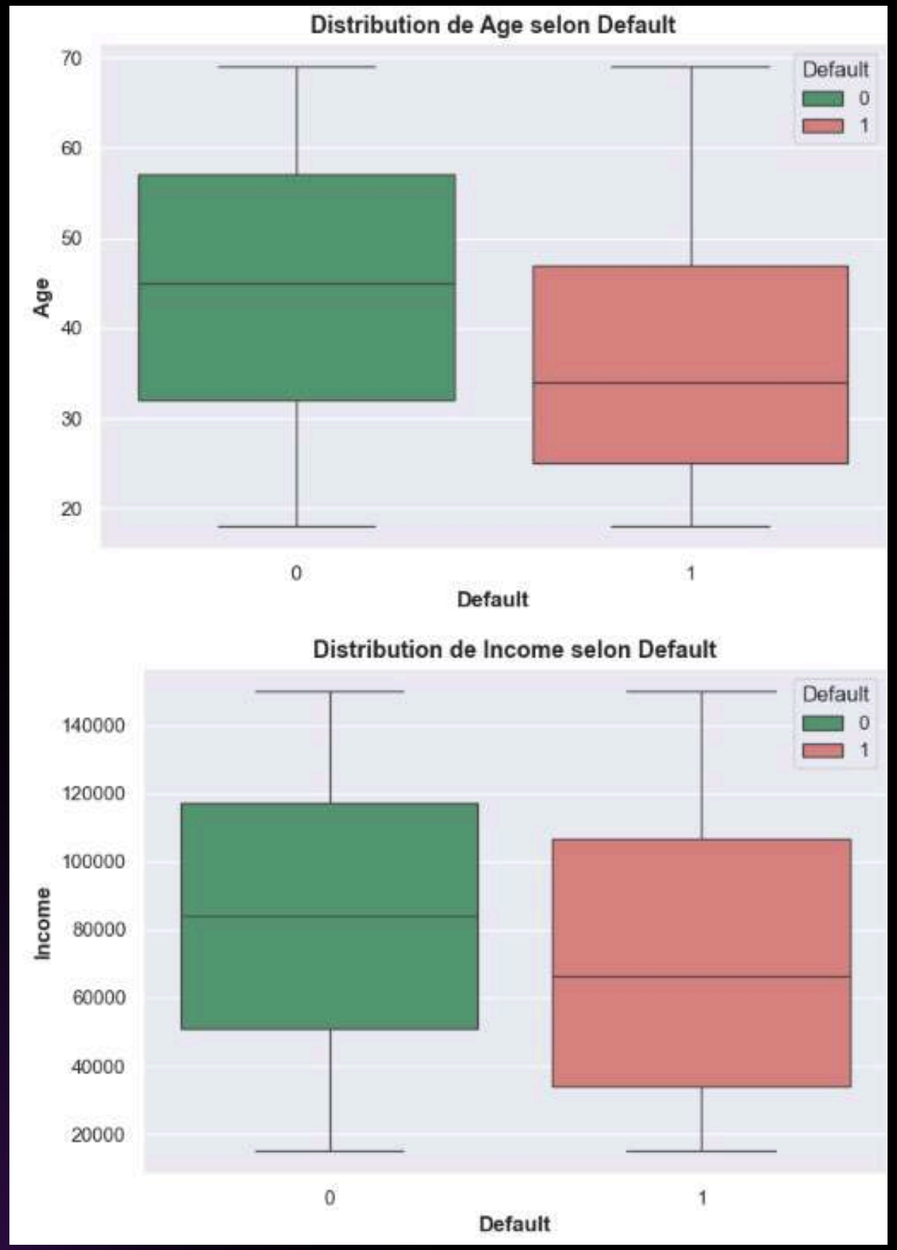


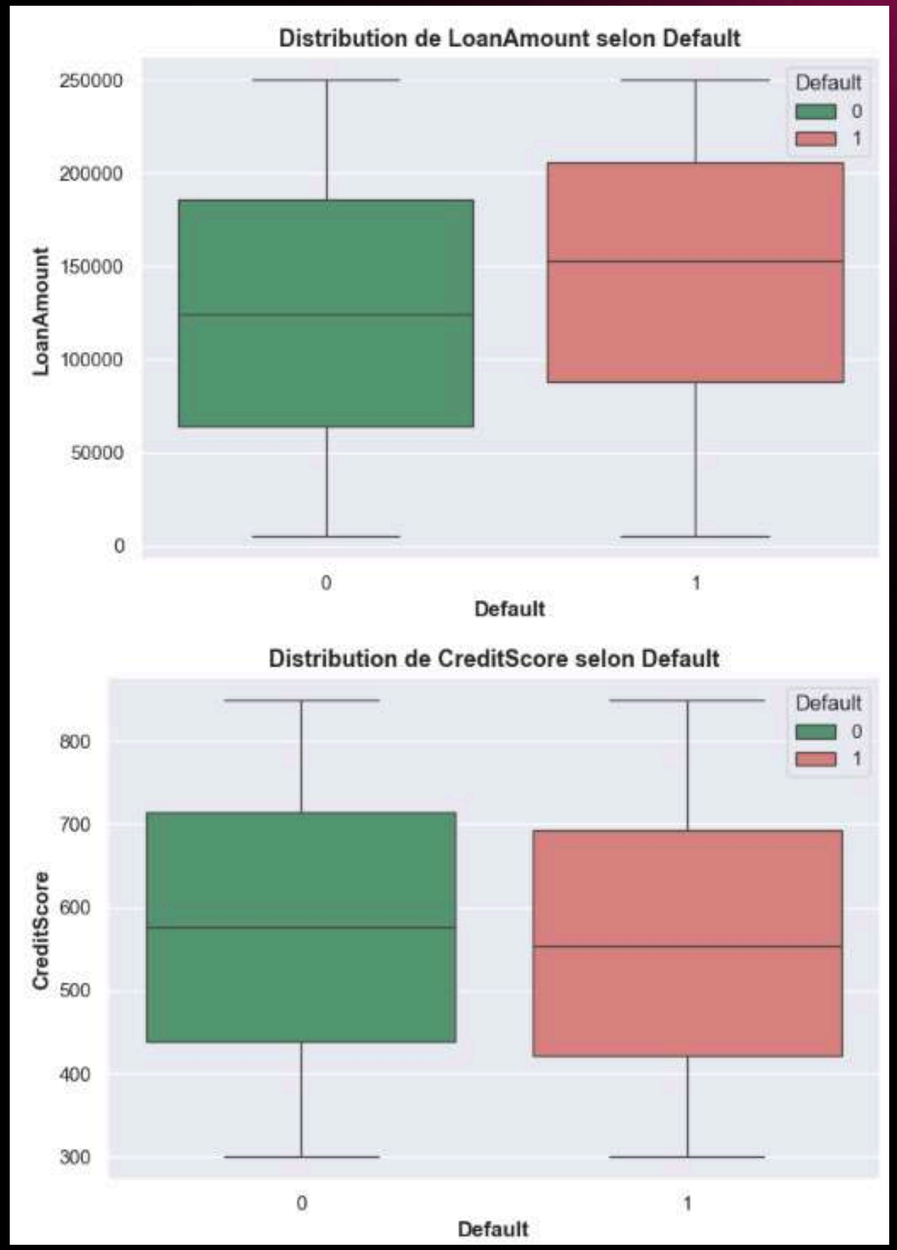




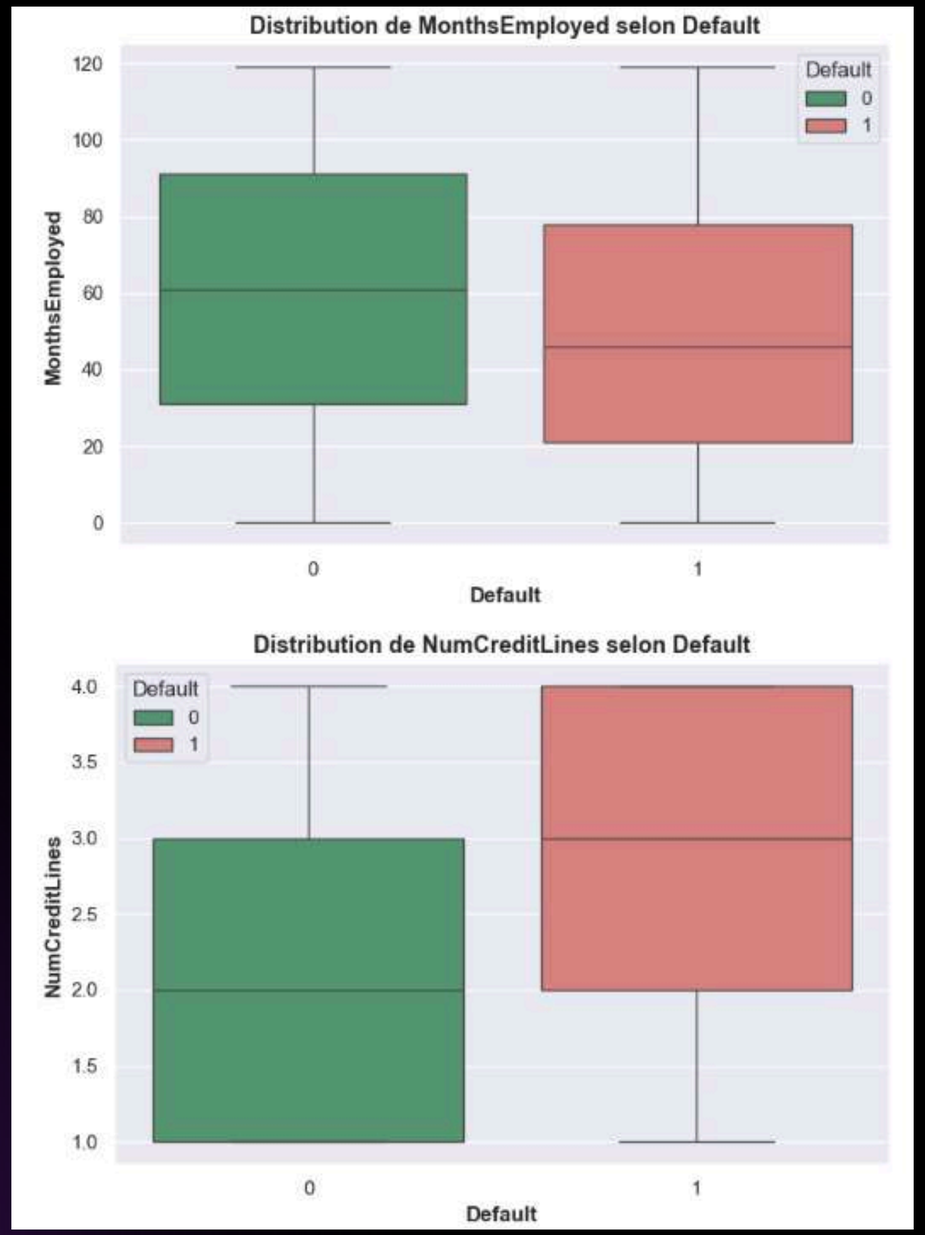


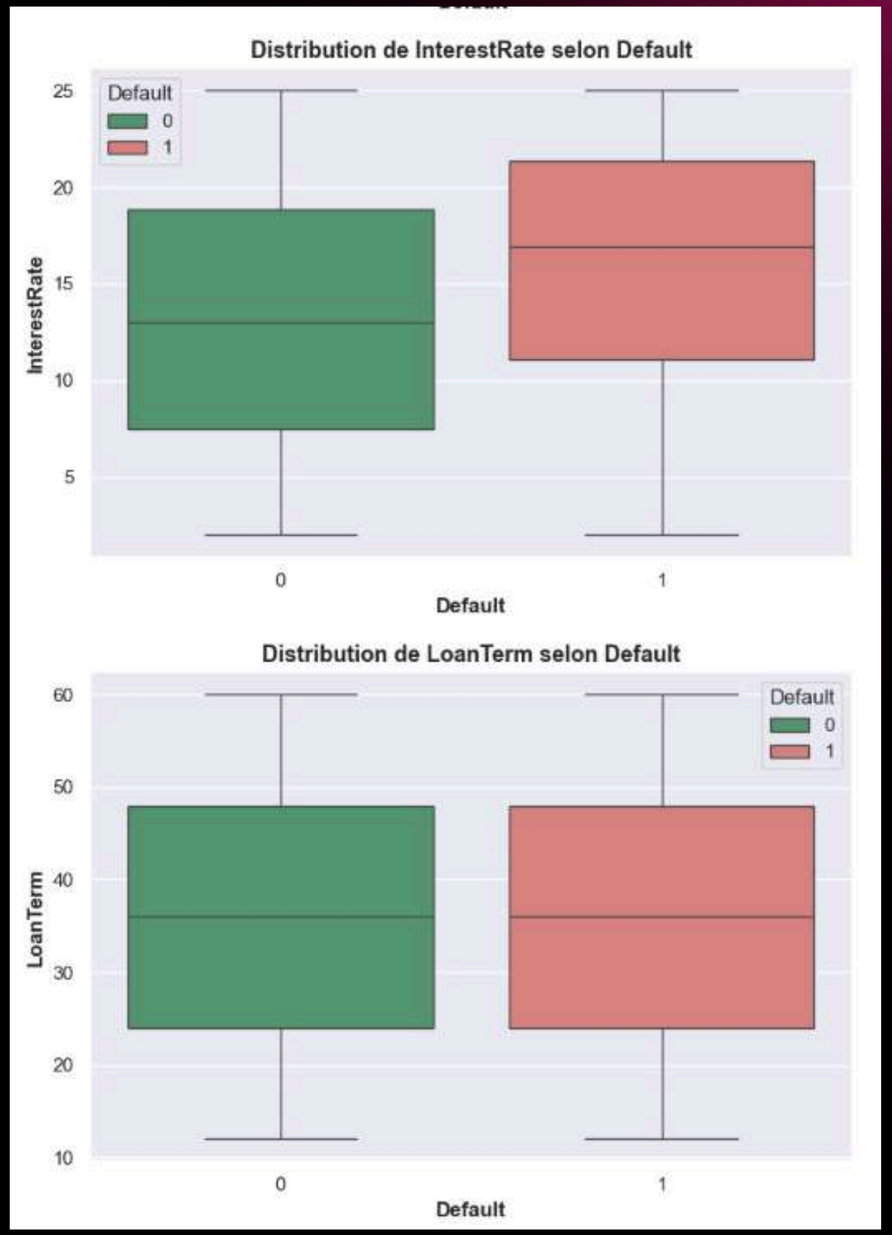
Présentation du sujet



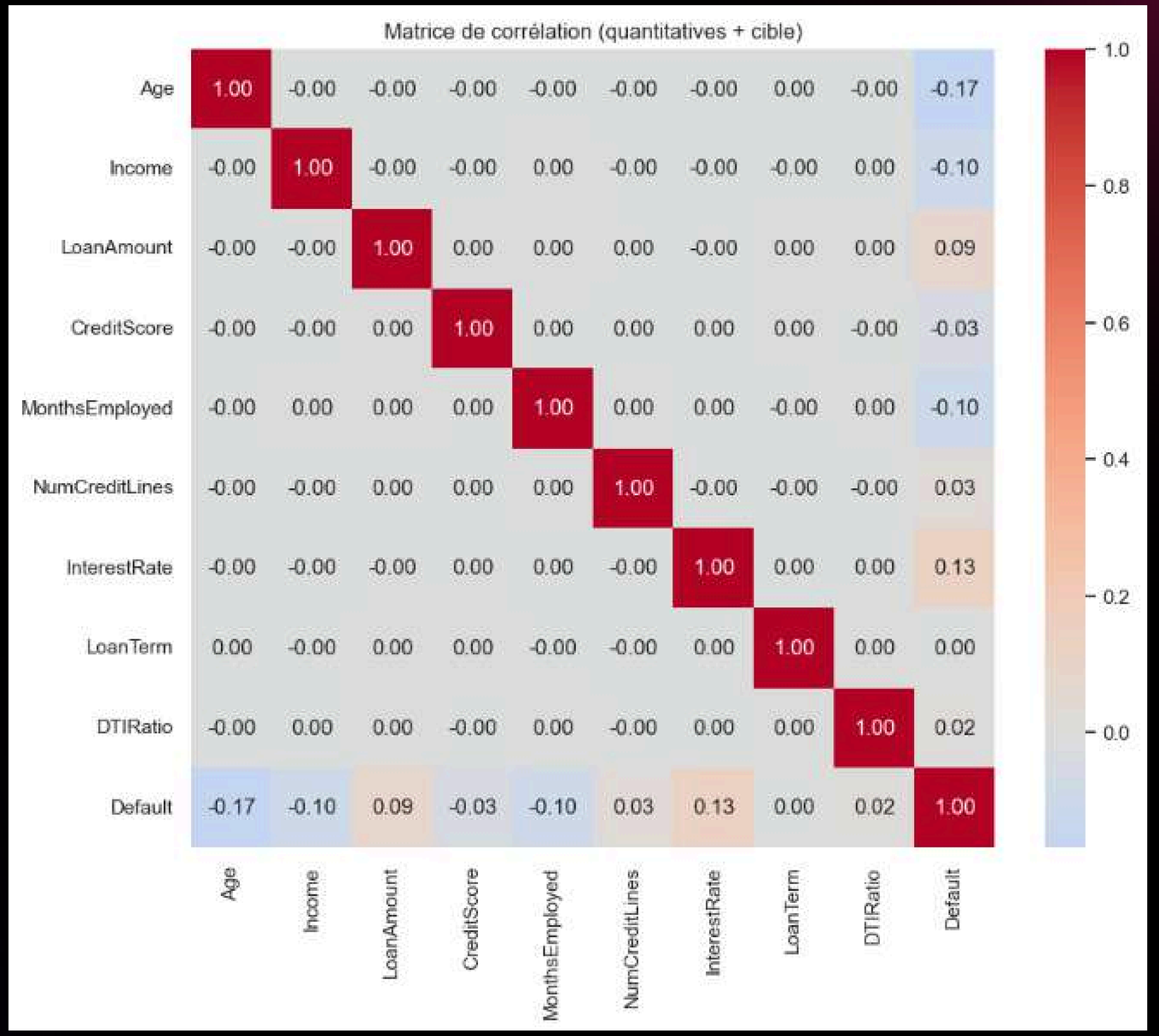


Présentation du sujet





Présentation du sujet



Méthodologie

Gestion de l'équilibre

Trois (3) méthodes d'équilibrage

Echantillonnage aléatoire

Consiste à supprimer aléatoirement certaines données de la classe majoritaire

Tomek Link

Consiste à supprimer les points de la classe majoritaire qui sont les plus proches voisins de la classe minoritaire

SMOTE

Consiste à augmenter la taille de la classe minoritaire en générant des échantillons synthétiques

Algorithmes
modèle de classification

Après l'étape de la préparation des données et encodage des variables catégorielle, notre cas d'étude est de type **classification binaire**

Régression Logistique

Son approche statistique, ce modèle s'adapte à notre cas d'usage car il permet de créer une frontière de séparation entre les deux classes différentes

La régularisation par défaut **Ridge** et en utilisation la pondération des classes avec l'attribut **balanced.**

Arbre de Décision

Par sa simplicité, ce modèle adapté à ce cas d'usage permet d'établir des règles pour aider à la prise de décision.

Utiliser **l'indice de GINI,** comme critère de séparation et une **profondeur de 4**

Évaluation : Utilisation de la validation croisée (K-Fold) pour garantir la robustesse du modèle.

Métriques : Performance mesurée par le score d'exactitude, Accuracy, Précision, F1-Score et Recall

Déploiement

Interprétation Modèles

VS

Régression Logistique

```
=== Top-10% (thr=0.7184) ===
ROC AUC: 0.7529299252190208
PR  AUC: 0.31161387107803795
TN=62810 FP=4899 | FN=6135 TP=2761
              precision    recall  f1-score   support

           0      0.911     0.928     0.919     67709
           1      0.360     0.310     0.334      8896

    accuracy                          0.856     76605
   macro avg      0.636     0.619     0.626     76605
weighted avg      0.847     0.856     0.851     76605

Coût(c_FN=5, c_FP=1) = 35574
```

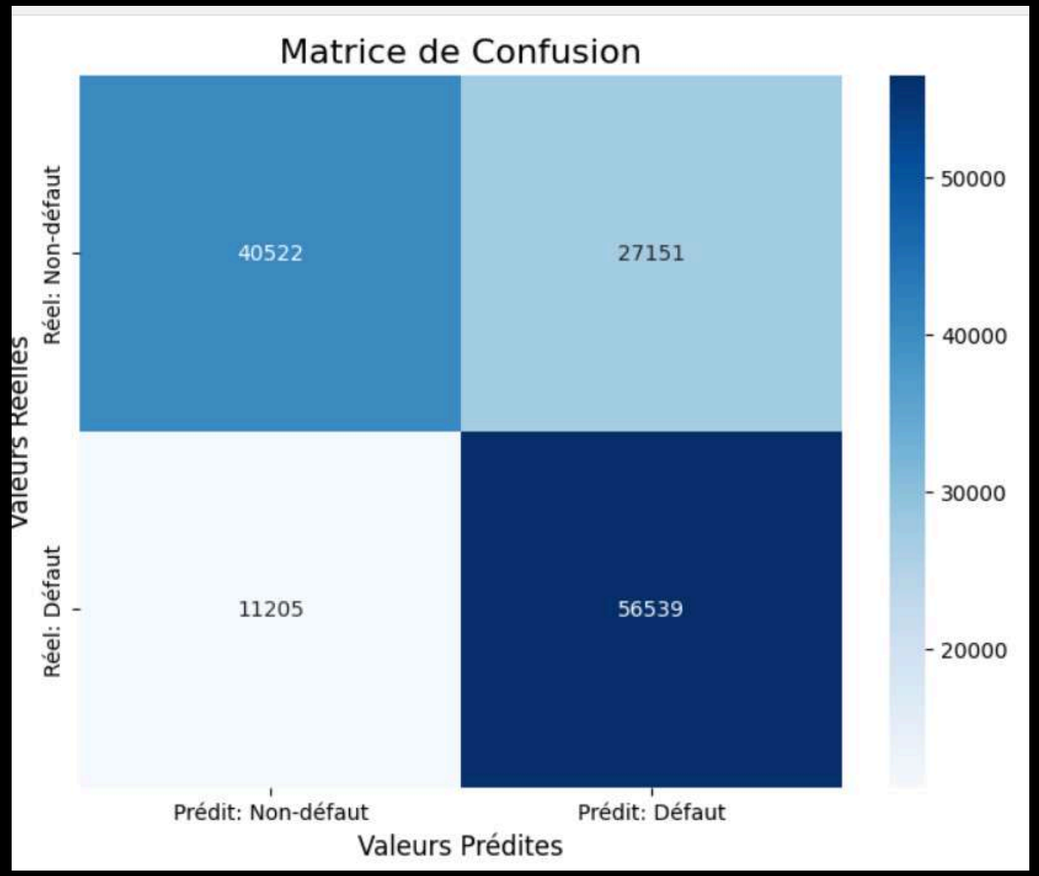
Arbre de décision

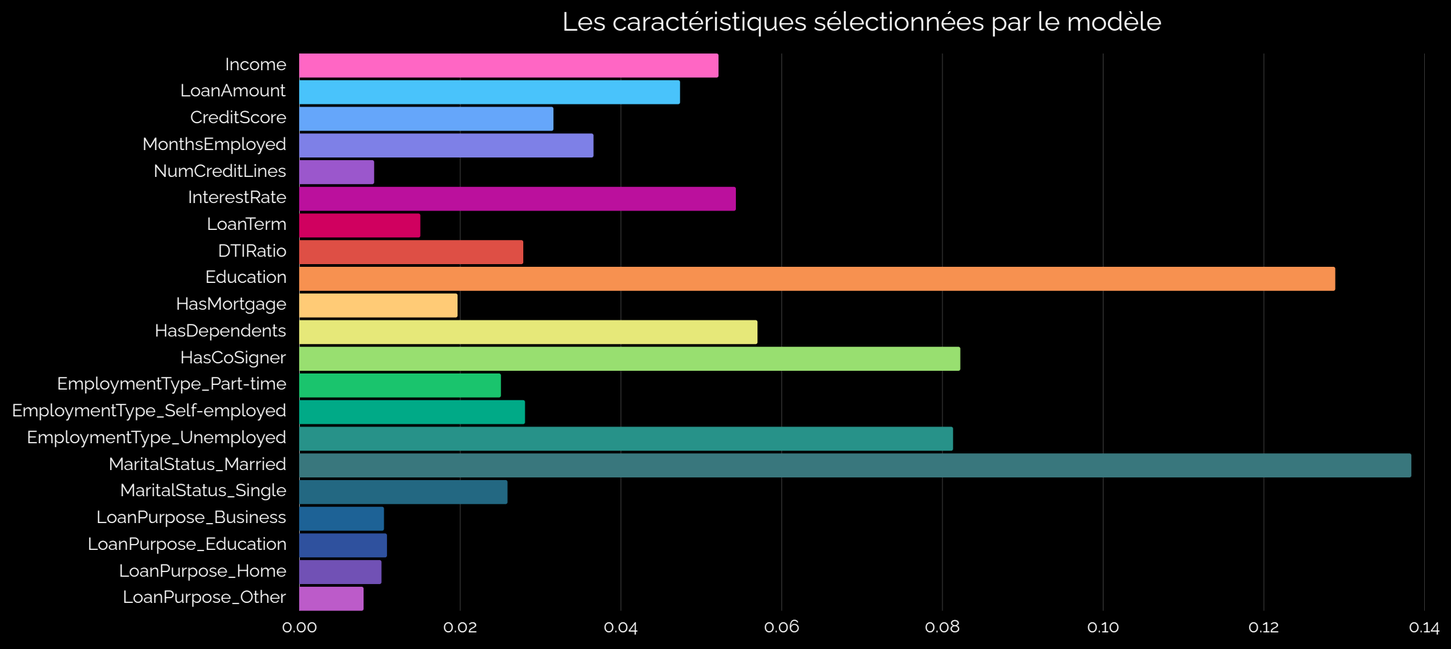
```
Accuracy :  0.7167563895227335
F1-score :  0.7467147404149662
Recall :  0.8345978979688238
Accuracy :                precision    recall  f1-score   support

           0       0.78      0.60      0.68     67673
           1       0.68      0.83      0.75     67744

    accuracy                           0.72    135417
   macro avg       0.73      0.72      0.71    135417
weighted avg       0.73      0.72      0.71    135417
```

Déploiement

Caractéristiques **importantes**

ARBRE
DE DÉCISION





Perspectives

Mise au point

Organisation

Echange en permanence sur le groupe de travail

Créer un répertoire de travail sur GitHub

Difficultés

Gestion de l'équilibrage

Effectuer l'ingénierie des caractéristiques

L'entraînement du modèle

Explication du modèle

Axe d'amélioration

Affiner les hyperparamètres

Interprétabilité du modèle

Ressources utilisées

- Cours de classe de ML
- https://www.3vfinance.com/infinance-le-blog/fr/risque-de-credit
- https:/shs.hal.science/halshs-04518248/file/Machine_Learning_et_mod%C3%A8les_IRB.pd
- https://fastercapital.com/fr/contenu/Defaut-de-pret---Comprendre-les-defauts-de-pret---l-impact-sur-les-prets-classes.html#L-impact-des-d-fauts-de-paiement-sur-les-pr-teurs
- https://dl.acm.org/doi/full/10.1145/3728725.3728813
- https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
- https://larevueia.fr/comment-gerer-le-desequilibre-des-classes-en-machine-learning//

Merci !

Des questions...