



Parcours AI Engineer

SOUTENANCE PROJET 4

“Construisez un modèle de scoring”

Stéphanie Duhem - Avril 2024



- 
- 
01. Contexte et orientation du projet
 02. Présentation générale du jeu de données
 03. Analyse exploratoire
 04. Approche méthodologique
 05. Synthèse des résultats

01. CONTEXTE & DÉROULÉ DU PROJET

LE CONTEXTE

Prêt à dépenser, société financière de crédit à la consommation, souhaite mettre en place un outil “scoring crédit” pour évaluer la probabilité qu’un client rembourse effectivement son prêt.

Il s’agit de développer un algorithme de classification qui permet d’accompagner la décision des chargés de clientèle pour déterminer si un prêt peut être accordé ou non. Le futur outil devra donc être aussi facilement interprétable.

LES 3 GRANDES ÉTAPES DU PROJET

- Explorer, analyser, traiter et bonifier les données à notre disposition
- Tester et comparer les différents modèles de classification
- Donner les clés d’interprétation du modèle



02. PRÉSENTATION DU JEU DE DONNÉES

LES FICHIERS

- 'application_train.csv' :
 - 307511 observations et 122 variables.
 - 1 observation = 1 crédit accordé
 - colonne 'TARGET' : 1 = crédit non remboursé, 0 = crédit remboursé
- 'HomeCredit_columns_description.csv' :
 - description de toutes les variables
- 'application_test.csv' et les autres fichiers ne seront pas utiles au projet

03. ANALYSES EXPLORATOIRES

LA COLONNE 'TARGET'

- Déséquilibre important dans la représentation des différents cas :
 - 92% des crédits accordés sont en 0 = crédit remboursé
 - 8% des crédits accordés sont en 1 = crédit non remboursé,

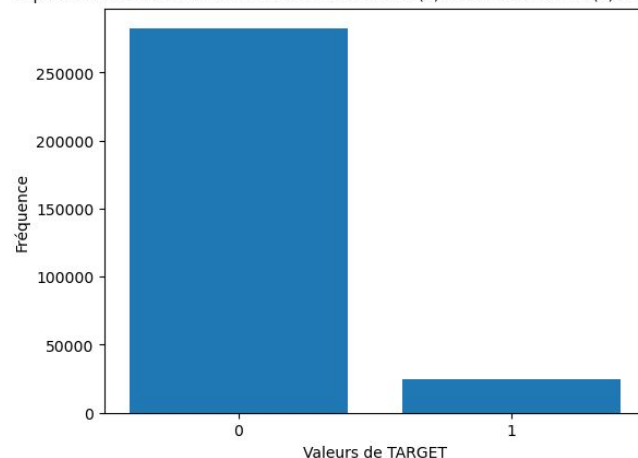
LES VALEURS MANQUANTES / NaN

- Taux global de NaN = 24,4%
- Sur 122 colonnes :
 - 67 colonnes contenant des NaN, dont 41 avec au moins 50% de valeurs manquantes.

LES TYPES DE DONNÉES

- 16 variables catégorielles
- 106 variables quantitatives

Répartition du nombre des cas de crédit remboursés (0) et non-remboursés (1) dans le dataset



03. LES VARIABLES CATÉGORIELLES

LA COLONNE 'TARGET'

- Déséquilibre important dans la représentation des différents cas :
 - 92% des crédits accordés sont en 0 = crédit remboursé
 - 8% des crédits accordés sont en 1 = crédit non remboursé,

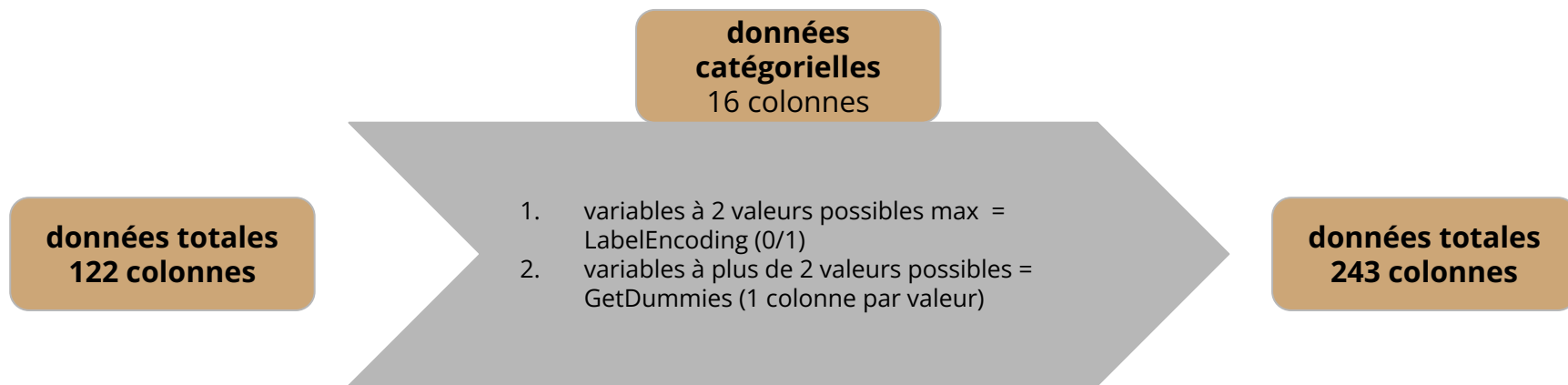
LES VALEURS MANQUANTES ('NaN')

- Taux global de NaN = 24,4%
- Sur 122 colonnes :
 - 67 colonnes contenant des NaN, dont 41 avec au moins 50% de valeurs manquantes.

LES TYPES DE DONNÉES

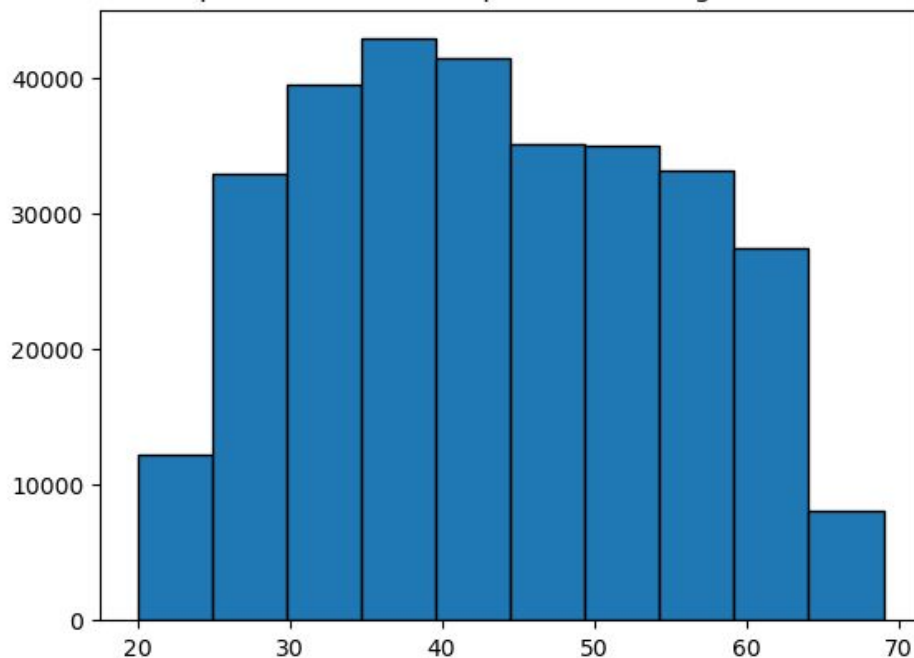
- 16 variables catégorielles
- 106 variables quantitatives

03. TRAITEMENT SUR LES VARIABLES CATÉGORIELLES

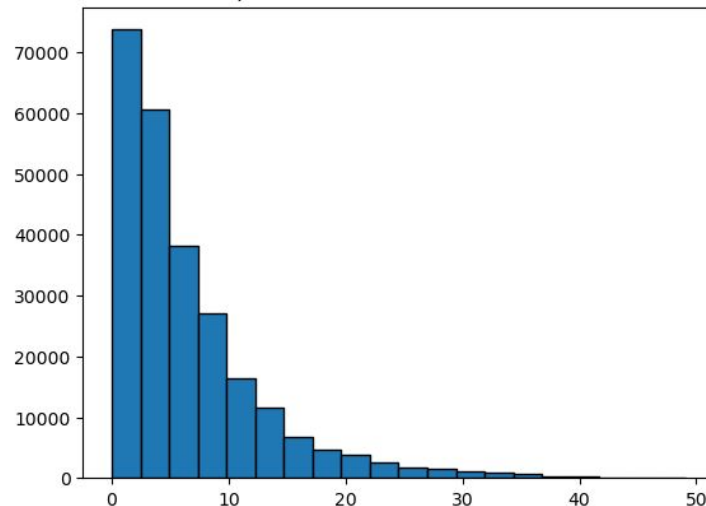


03. ANALYSES SUR LES VARIABLES QUANTITATIVES

Répartition des clients par tranche d'âge de 5 ans



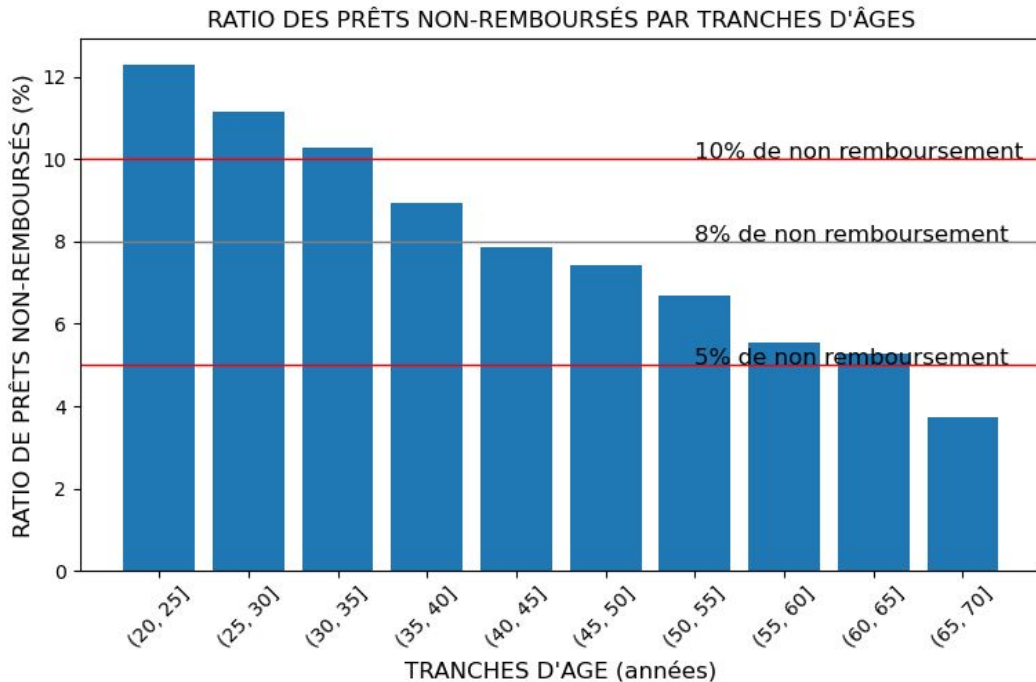
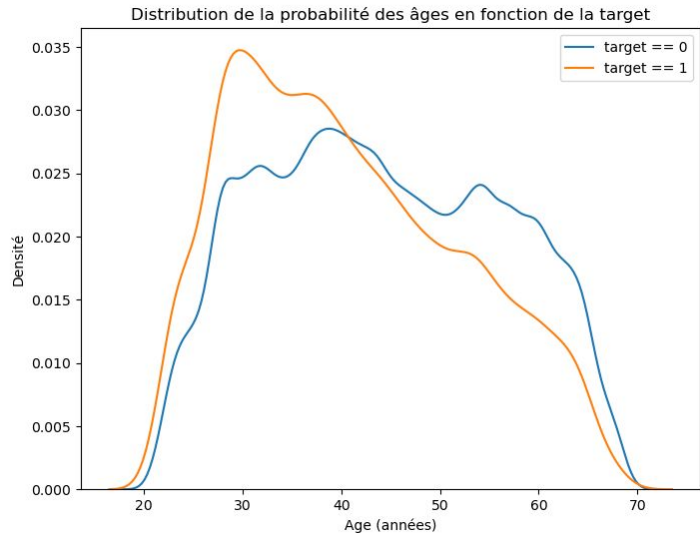
nombre d'années où l'emprunteur a travaillé avant la date de demande du crédit



Pour cette donnée :

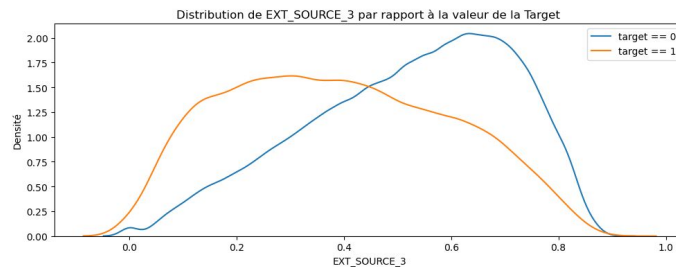
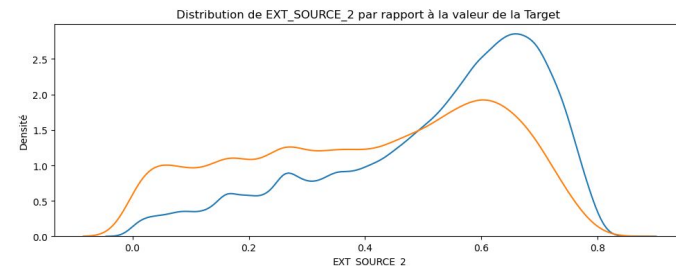
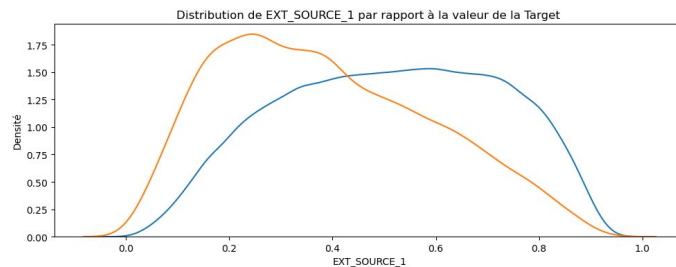
Traitement des outliers ("-1000ans" d'emploi)

03. CORRÉLATIONS ENTRE L'ÂGE ET LA CIBLE



03. CORRÉLATIONS ENTRE 'EXT_SOURCE_' ET LA CIBLE

- Ces données correspondent à des scores normalisés depuis des données extérieures.
- Ces 3 variables ont les corrélations (négatives) les plus fortes en valeur absolue avec la cible (lorsqu'elles augmentent, le taux de non remboursement diminue)



03. FEATURE ENGINEERING

**données
243 colonnes**

1. processus pour les features 'métier'
2. processus pour les features polynomiales

**données
243 colonnes**

Réduction du nombre de colonnes en retenant les features dont le taux de corrélation est supérieur ou égale à la médiane de toutes les corréaltions avec la cible

**dataset final
139 colonnes
307511 lignes**

Dernier traitement :
imputation par la médiane des valeurs nulles dans les colonnes

03. MODÉLISATION

préparation des données

1. séparation du jeu de données en 80/20, pour isoler le jeu de test de l'entraînement
2. Mise à l'échelle des données pour éviter des biais

score métier

pour rappel

- quand la target = 0 : le prêt est remboursé
- quand la target = 1 : le prêt est non remboursé

Supposition que le coût d'un faux négatif est 10 fois supérieur à un faux positif

Intégration de ce score métier comme métrique de sélection du modèle

Pipeline d'entraînement

1. Oversampling pour diminuer le déséquilibre de classe (92% de négatifs, 8% de positifs)
2. Hyperparamètres propres à chaque modèle

03. MODÉLISATION

TESTS SUR 2 MODÈLES DE CLASSIFICATION

- RANDOMFORESTCLASSIFIER (modèle non linéaire)
- RÉGRESSION LOGISTIQUE (modèle linéaire)

Recherche par grille des hyperparamètres de chaque modèle :

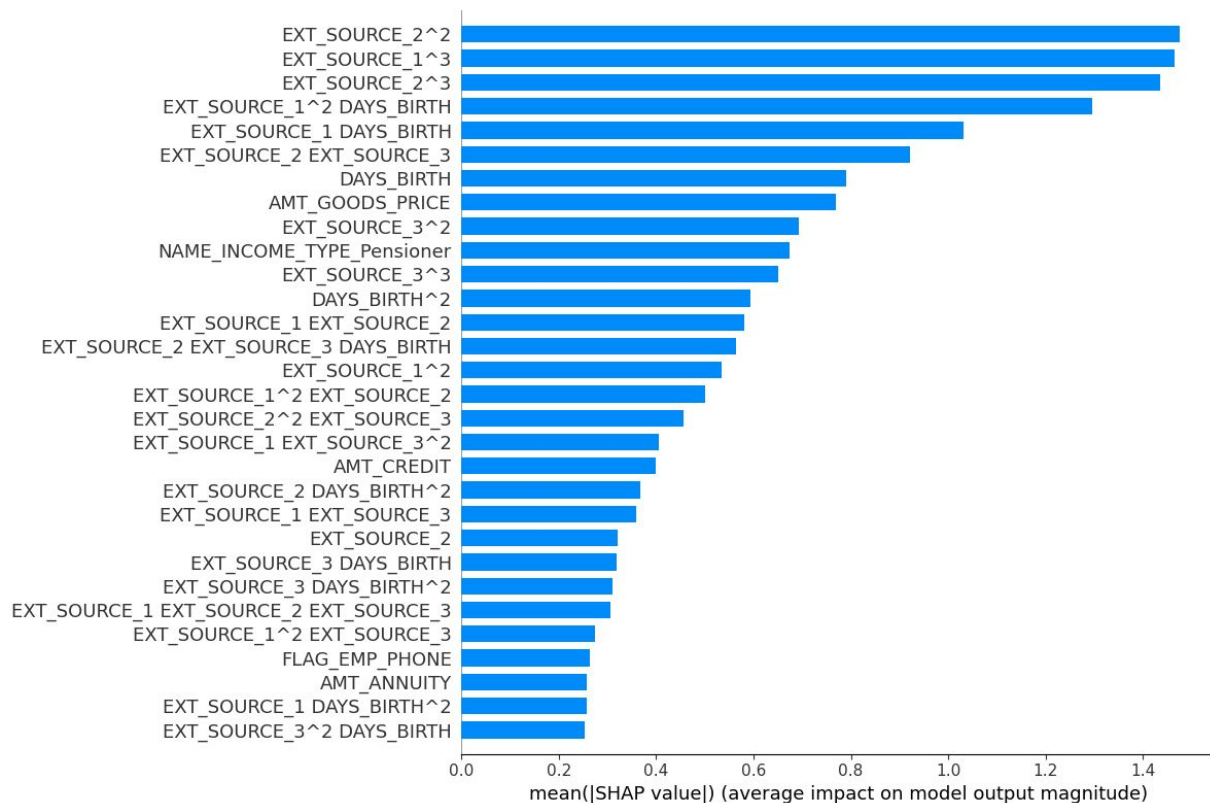
- sélection des meilleurs modèles en fonction du seuil le plus bas de coût métier

Comparaison des résultats

Sélection du meilleur modèle

-
-

03. ANALYSE DE L'IMPORTANCE GLOBALE (modèle)



03. ANALYSE DE L'IMPORTANCE DES VARIABLES LOCALE (client donné)

15

