IMPLÉMENTEZ UN MODÈLE DE SCORING

SOMMAIRE

I) MISSIONS

A) PROBLEMATIQUE

Contexte :

- Le risque de crédit résulte de l'incertitude quant à la possibilité ou la volonté des contreparties ou des clients de remplir leurs obligations. Très prosaïquement, il existe donc un risque pour la banque dès lors qu'elle se met en situation d'attendre une entrée de fonds de la part d'un client ou d'une contrepartie de marché. Par exemple, lors d'une demande de crédit, qu'il s'agisse d'une carte de crédit, d'un prêt automobile, d'un prêt personnel ou d'une hypothèque, le créancier voudra connaître le niveau de risque de crédit de chaque emprunteur. Ainsi, cela va aussi bien conscientiser chaque banque à identifier, surveiller et contrôler son risque de crédit aussi bien qu'à prévoir le capital suffisant contre ces risques. Grâce à cette analyse, chaque banque pourra se sortir indemne, de la manière la plus sûre, des risques à encourir. D'où l'utilité du crédit scoring.

Mission

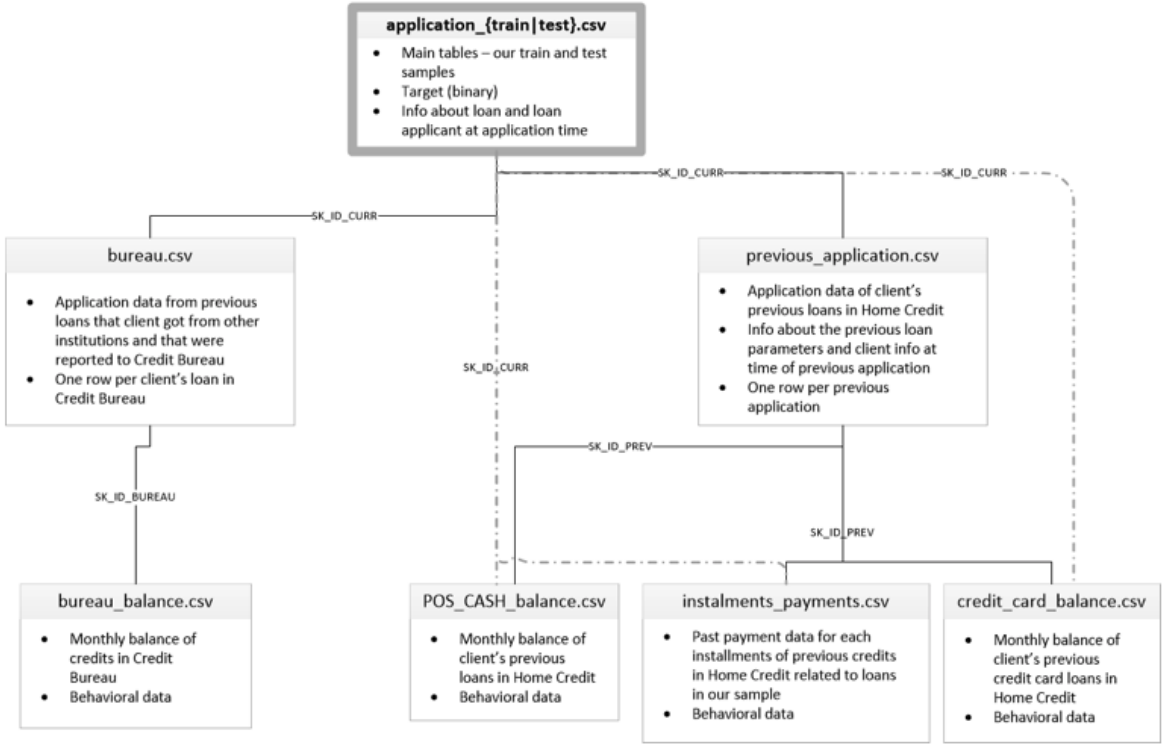
- Développer un modèle de scoring de la probabilité de défaut de paiement du client pour étayer la décision d'accorder ou non un prêt à un client potentiel.

Spécification

- Développement d'un dashboard viable et interactif permettant d'améliorer la connaissance client des chargés de relation client.

2) NETTOYAGE DU JEU DE DONNEES

A) PRESENTATION DU JEU DE DONNEES



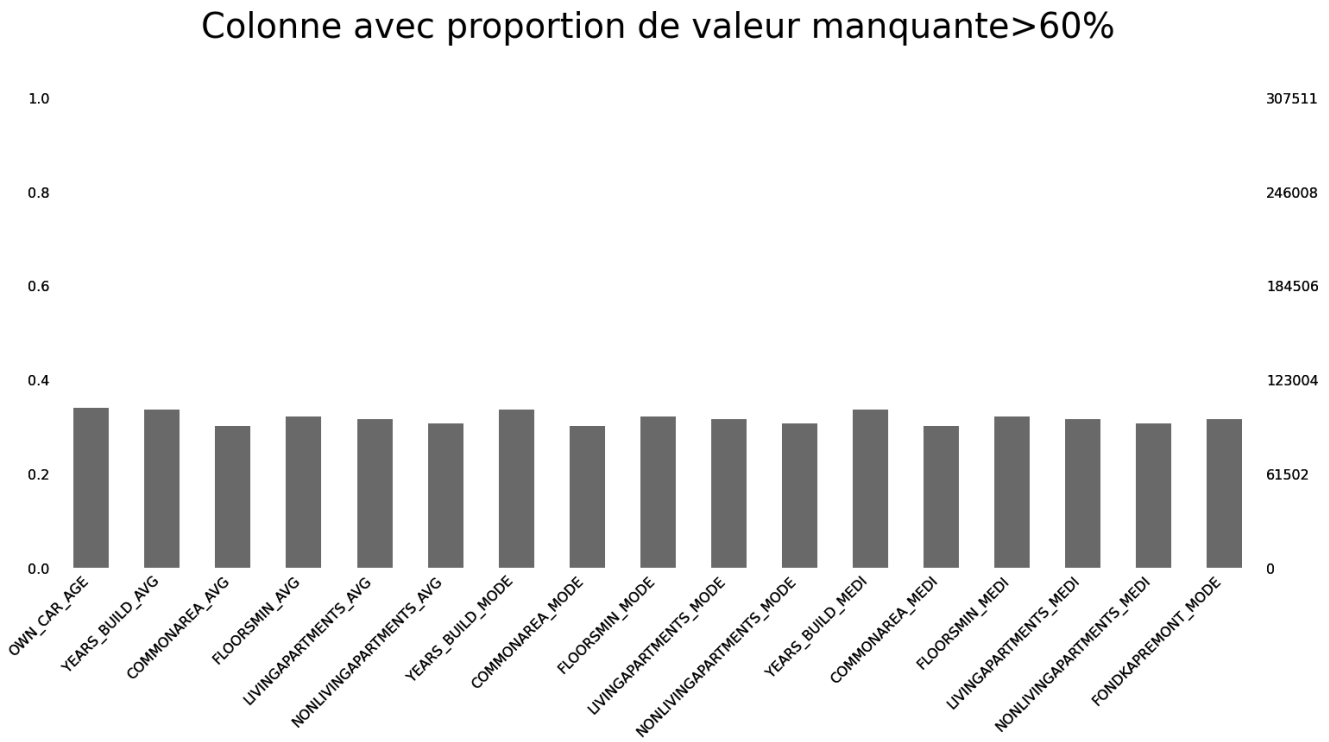
Sept jeux de données

- Différentes informations personnelles pour chacun des clients bancaires existants (sexe, revenu,emploi, défaut de paiement etc.)

Un jeu de donnée principal

- 307 511 clients et 122 features

B) TRAITEMENTS DES VARIABLES



Types de variable

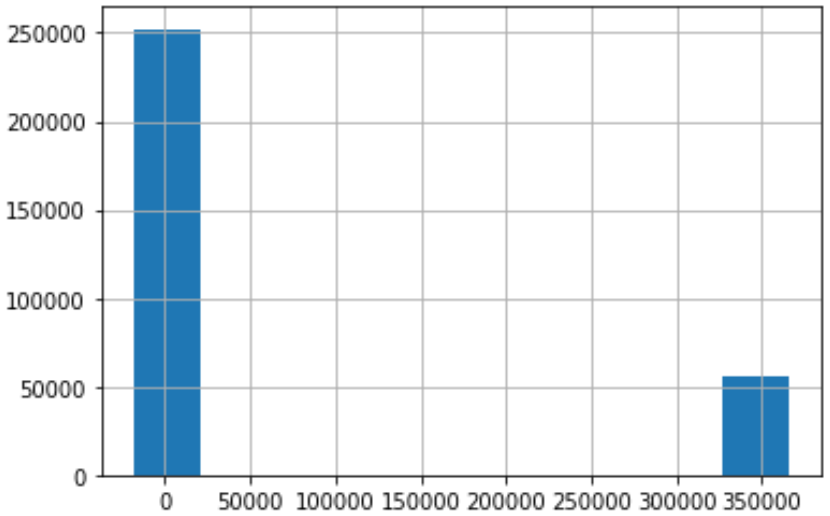
- 16 variables de type catégorique, et 105 variables numérique dont 65 variables de type flottants et 41 de type entier

Valeurs manquantes

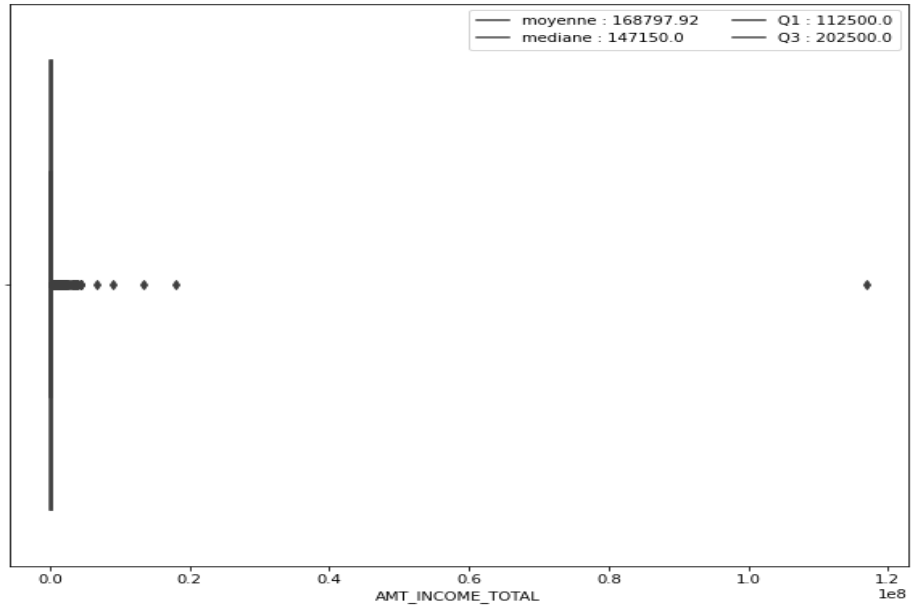
- 17 colonnes soit près de 14 % des variables dont la proportion de valeurs manquantes est supérieure à 60%

C) TRAITEMENTS DES OUTLIERS

Outliers sur la variable
days_employed



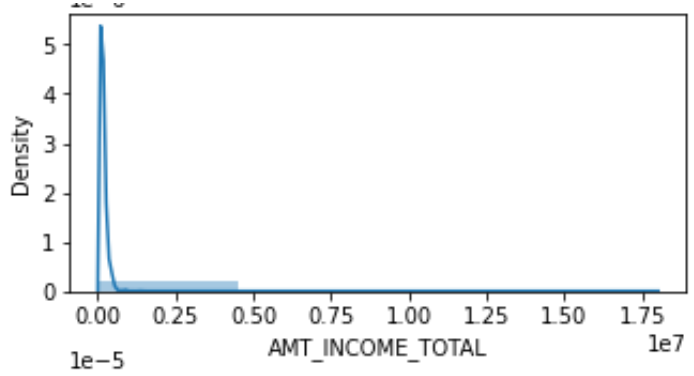
DAYS_EMPLOYED
moy: 63815.04
med: -1213.0
mod: 0 365243
dtype: int64

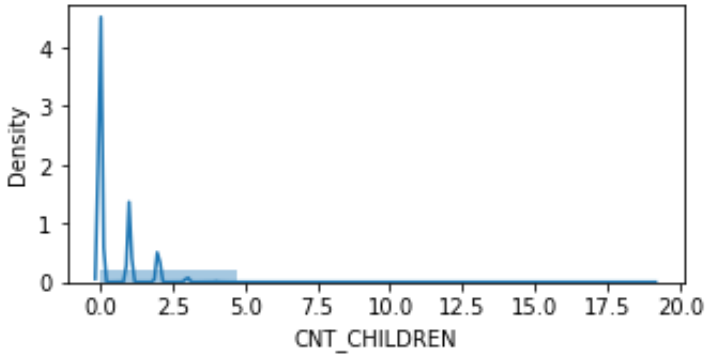


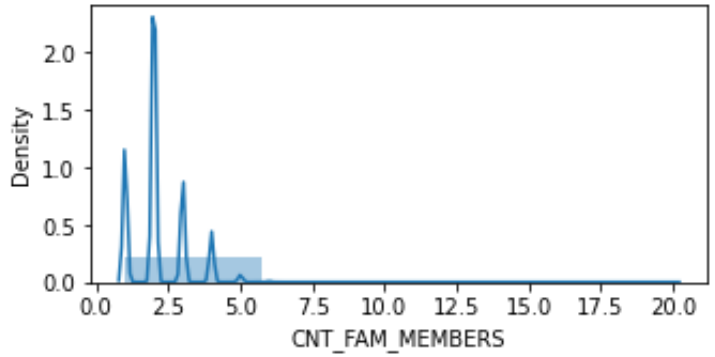
Outliers sur la variable
amt_income_total

3) ANALYSE DES DONNEES

A) ANALYSE UNIVARIEE





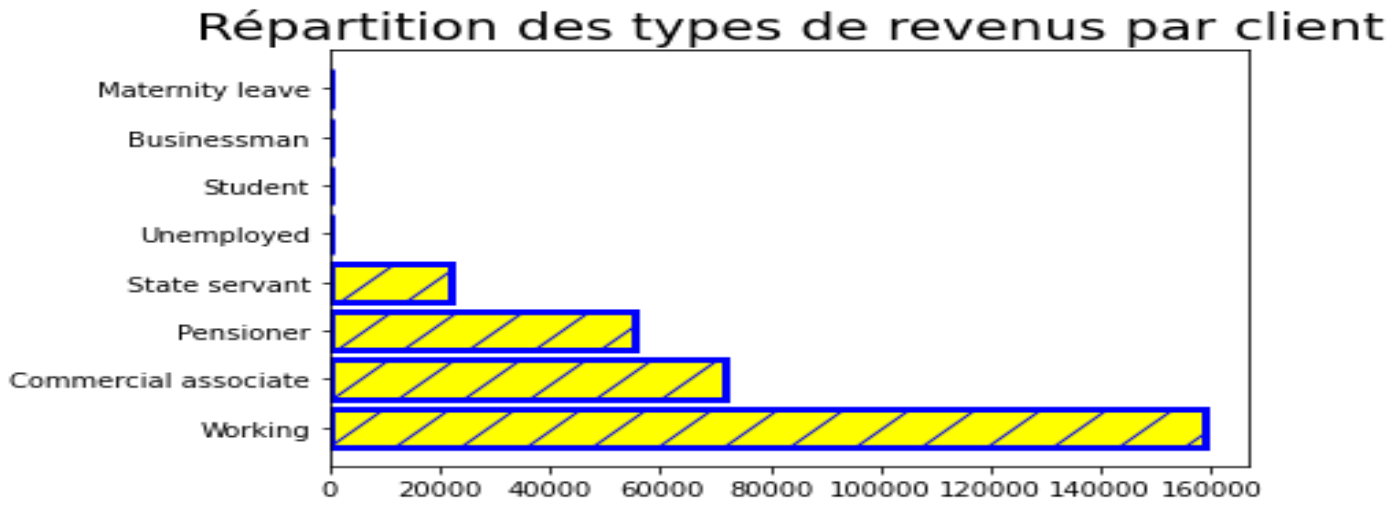


Etat matrimonial et familial

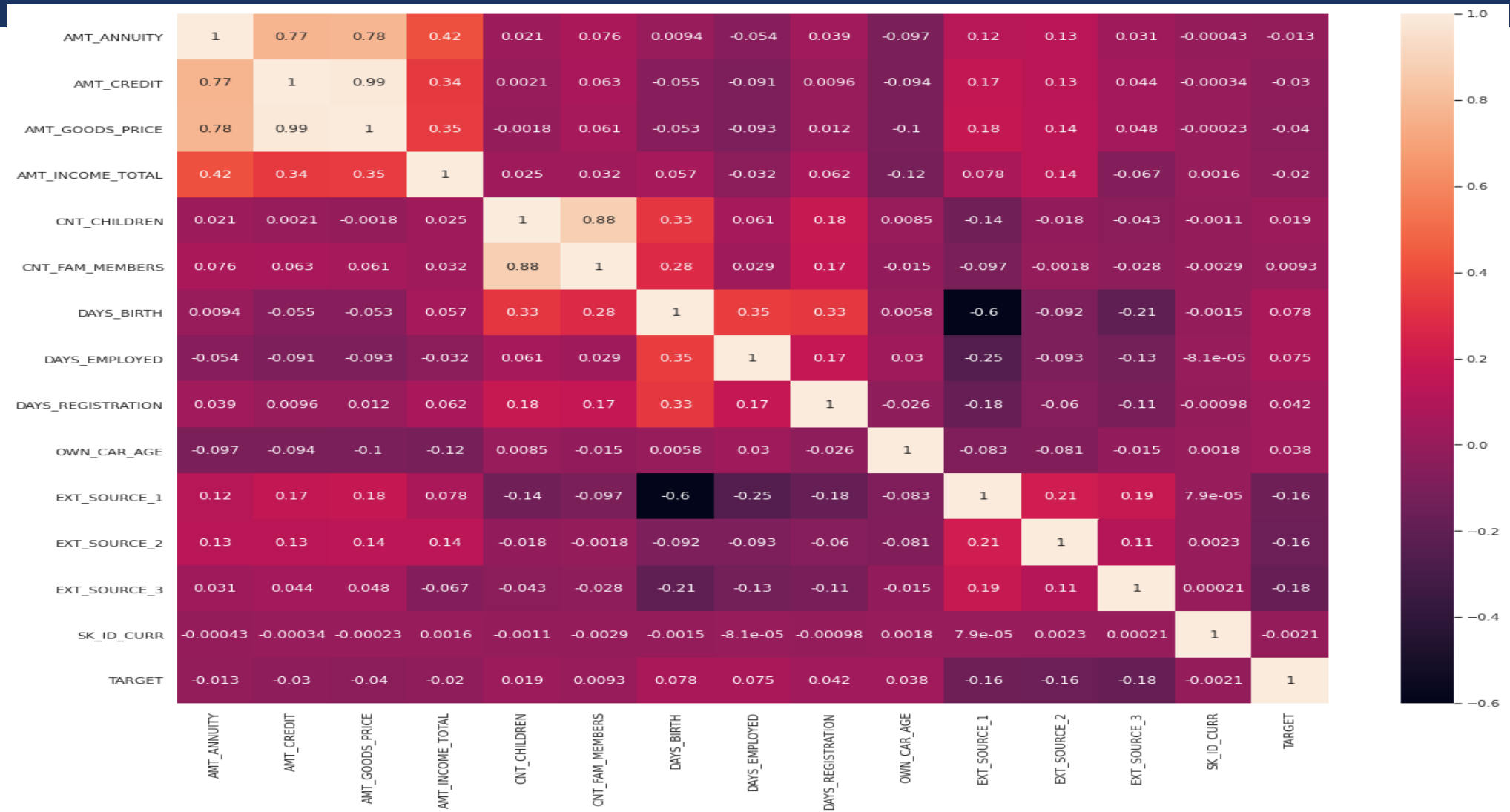
- La plupart des clients qui contractent un prêt n'ont pas d'enfants. Les clients en couples sont les plus nombreux, suivis des personnes seule, et des familles avec un enfant ou deux.

Type de revenus

- La plupart des clients sont des salariés, suivis par les statuts commercants, retraités et fonctionnaires.



B) ANALYSE BIVARIEE



4) INGENIERIE ET SELECTION DES VARIABLES PREDICTRICES

A) CREATION DE NOUVELLES VARIABLES

'TAUX_ENDETTEMENT'

- Pourcentage du montant du crédit par rapport au revenu du client

'CAPACITE_REMBOURSEMENT'

- Montant de l'échéance mensuelle que le client est en capacité de rembourser en tenant compte de ses ressources nettes.

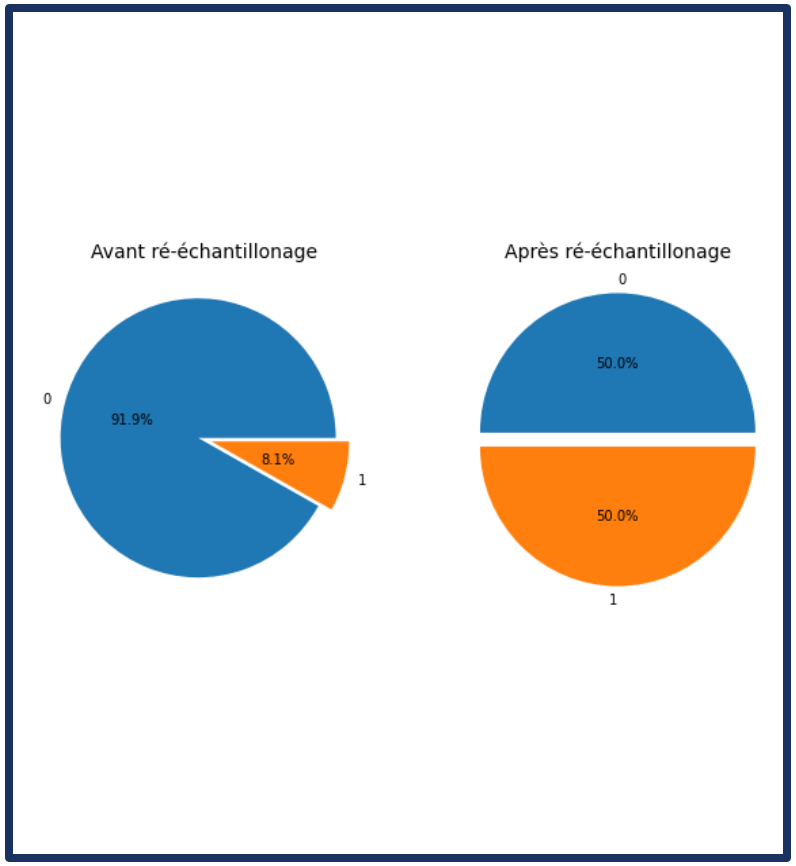
'DAYS_EMPLOYED_PERCENT'

- Pourcentage de jours employés par rapport à l'âge du client

NOUVELLES FEATURES A PARTIR DES FEATURES RETENUES

- Fonctions somme, minimum, maximum, moyenne

B) PREPROCESSING DES VARIABLES



Encodage des variables

- Utilisation de Label encoding pour les colonnes qui avaient un comportement ordinal
- Utilisation de l'encodage One-Hot pour les autres

Standardisation des variables

- Technique de mise à l'échelle avec StandardScaler()

Ré-échantillonage

- Sous-échantillonnage avec RandomUnderSampler
- Sur-échantillonage avec SMOTE

5) MODELISATION

A) MODELES UTILISEES

- Dummy Classifier
  - Baseline
- Logistic Regression
- Random Forest Classifier
- XGBoost
- LGBM

B) MESURE DE PERFORMANCE DES MODELES

Matrice de confusion

Valeur actuelle / Valeur prédite	Avec défaut de paiement	Sans défaut de paiement
Avec défaut de paiement	VP	FP=coût d'opportunité
Sans défaut de paiement	FN=coût important	VN

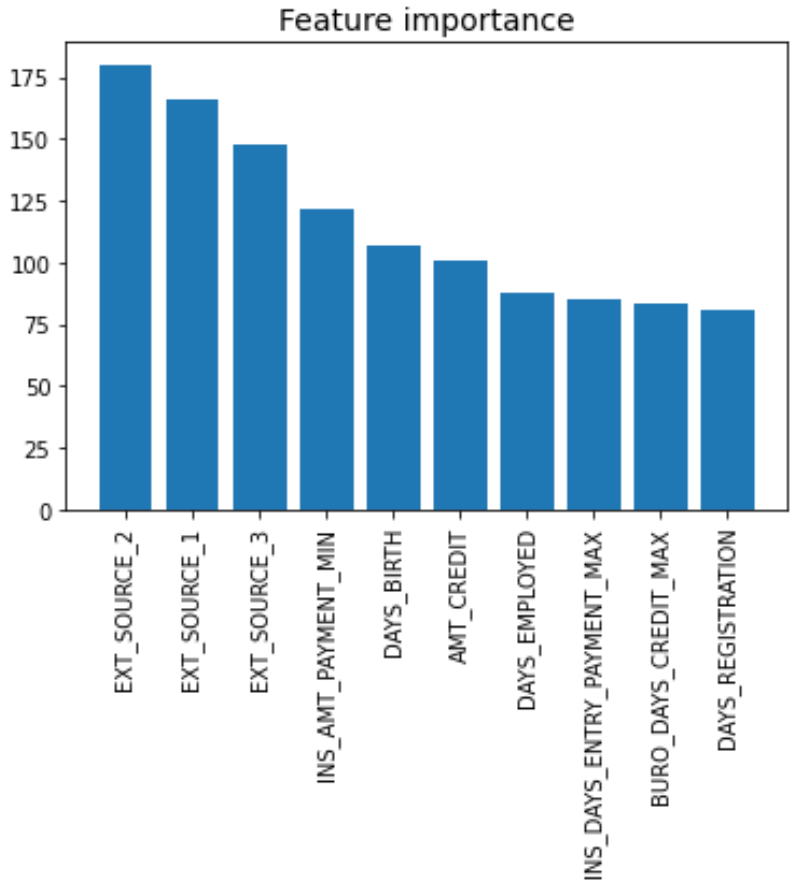
AUC ROC (aire sous la courbe ROC)

- La courbe ROC (Receiving Operator Characteristic) sert à comparer différents classifieurs. Plus une courbe a des valeurs élevées, plus l'aire sous la courbe est grande, moins le classifieur fait d'erreur.

B) MESURE DE PERFORMANCE DES MODELES

	Méthode de rééchantillonage	Dummy Classifier	Logistic Regression	Random Forest	Xgboost	LGBM
Accuracy	Aucune	0.8519	0.919067	0.8795	0.919367	0.9193
ROC AUC	Aucune	0.500309	0.698229	0.699396	0.764194	0.764638
F1	Undersampling	0.136225	0.291718	0.258632	0.308565	0.312481
Score	SMOTE	0.501133	0.690867	0.8677	0.909068	0.913
ROC	SMOTE	0.488967	0.739008	0.663199	0.701821	0.715308

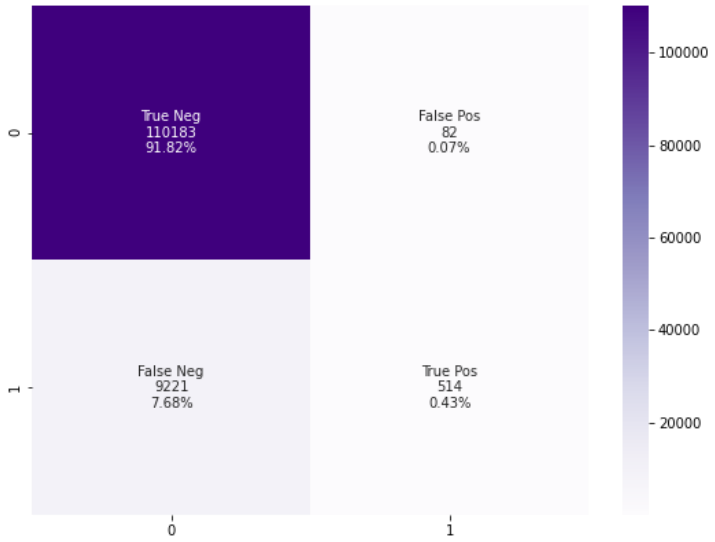
C) OPTIMISATION DU MODELE

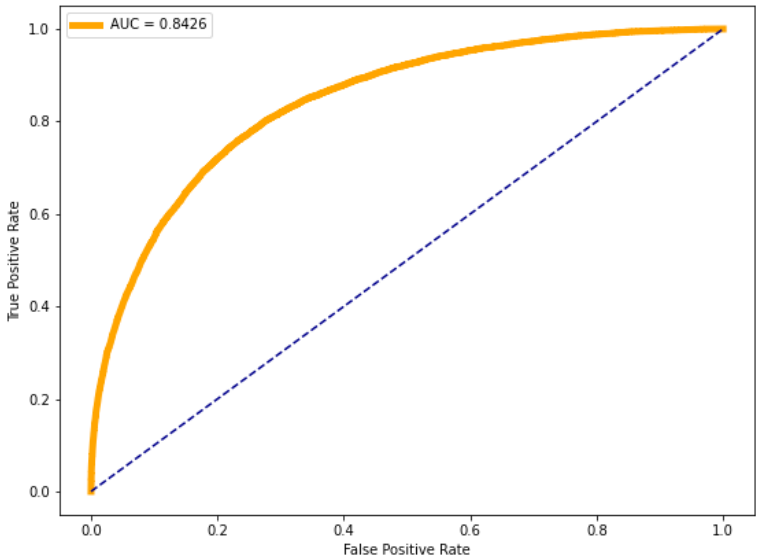


Optimisation des hyperparamètres avec gridsearchcv

- 'n_estimators': 10 000
- 'learning_rate': 0.05
- 'objective': binary
- 'subsample': 40 000
- 'colsample_bytree': 0.71
- 'reg_lambda':0.097959

C) OPTIMISATION DU MODELE





Score

'AUC': 0.8426497045197152

'accuracy': 0.9326

Score personnalisé

Avant optimisation : 0,02

Après optimisation : 0,11

6) REALISATION DU DASHBOARD

A) CREATION ET DEPLOIEMENT DU DASHBOARD

Streamlit

- Framework d'application open source développé pour créer des tableaux de bord ML