

Projet 7 : Implémentez un modèle de scoring

Nom : TRABIS

Prénom : Mohamed

Intitulé de formation : Data Scientist

Mentor: Mr. Christian NOUMSI



Table des matières

- Introduction
- Évaluation et découverte des données
- Analyse exploratoire des données
- Modèles de prédiction
- Dashboard
- Conclusion

Introduction



Introduction

□ Contexte :

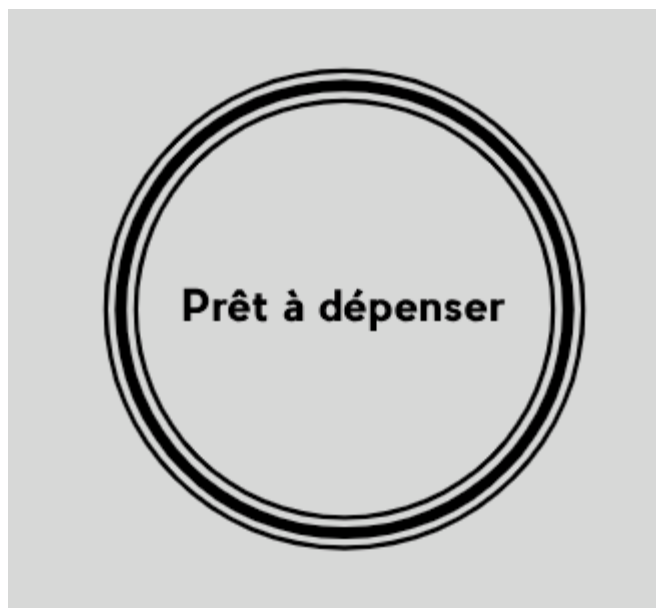
- L'entreprise «**Prêt à dépenser**» souhaite mettre en œuvre un outil de «scoring crédit» pour calculer la probabilité qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Elle souhaite donc développer un algorithme de classification en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).
- Prêt à dépenser décide de développer un **Dashboard** interactif pour que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.



Introduction

❑ Mission :

1. Construire un modèle de **scoring** qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.
2. Construire un **Dashboard** interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client des chargés de relation client.



Évaluation et découverte des données



Évaluation et découverte des données

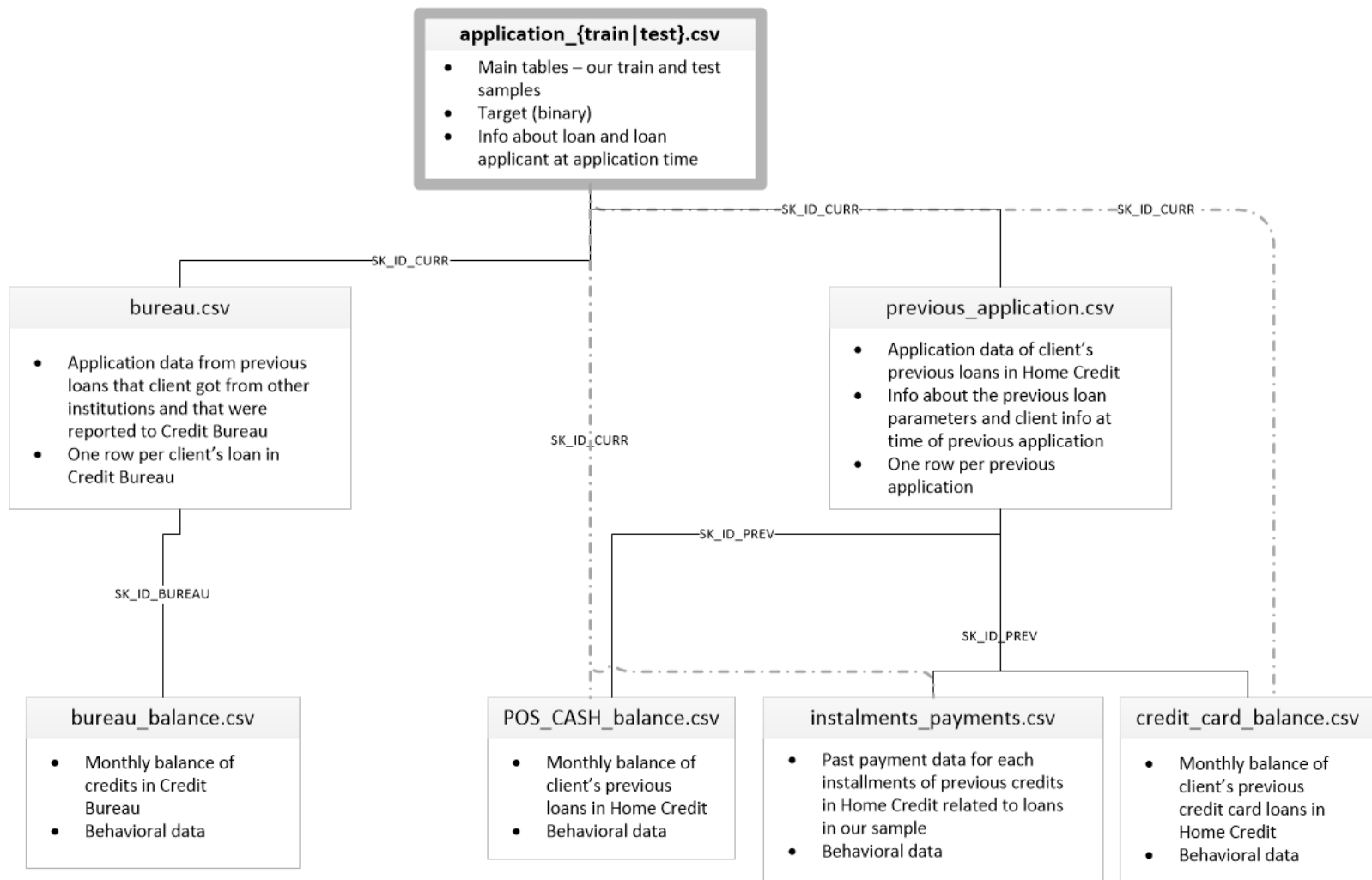
- ❑ Le jeu de données est composé de Plusieurs fichiers CSV :

Fichier CSV	Description
application_train.csv	Les principales données de formation avec des informations sur chaque demande de prêt chez Prêt à dépenser.
bureau.csv	Données concernant les crédits antérieurs du client auprès d'autres institutions financières.
bureau_balance.csv	Données mensuelles détaillées sur les crédits précédents dans le fichier bureau.csv
credit_card_balance.csv	Données mensuelles sur les cartes de crédit précédentes que les clients ont eues avec Prêt à dépenser.
installments_payments.csv	Historique de paiement pour les prêts précédents chez Prêt à dépenser.
previous_application.csv	Demandes précédentes de prêts chez Prêt à dépenser des clients qui ont des prêts dans le fichier application_train.csv
POS_CASH_balance.csv	Données mensuelles sur les clients précédents.



Évaluation et découverte des données

❑ Modèle de base de données:





Évaluation et découverte des données

- ❑ Les étapes effectuées pour le nettoyage et la validation des données :
 - Importer les fichiers CSV.
 - Effectuer des agrégations pour avoir une ligne par client (ex : moyenne).
 - Fusionner toutes les données dans une seule DataFrame.
 - Supprimer les colonnes avec un taux de valeurs manquantes supérieur à 30%.
 - Encoder les colonnes catégorielle (***get_dummies***).
 - Traiter les valeurs manquantes avec « ***SimpleImputer*** ».
 - Réduire l'utilisation de la mémoire de notre dataframe pour optimiser les performances.

- ❑ Remarque :

Suite à cette fusion nous avons 307505 lignes et 288 colonnes.

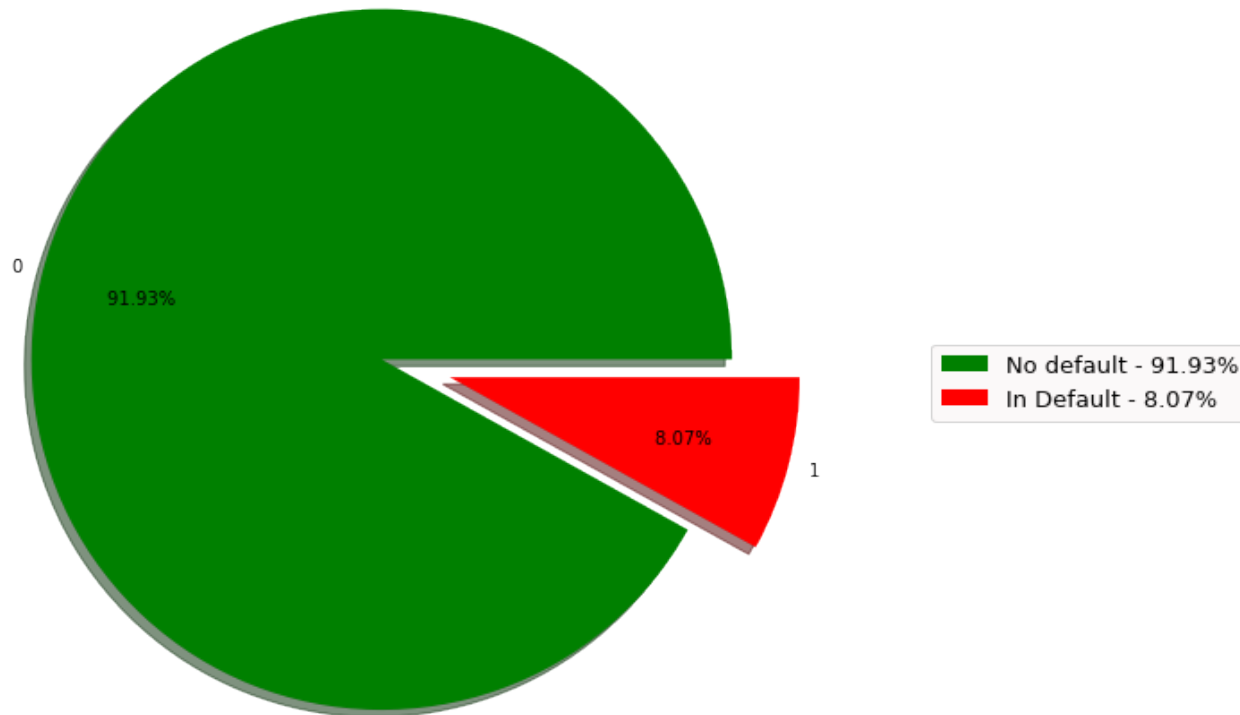
Analyse exploratoire des données



Analyse exploratoire des données

❑ Graphique de la distribution de la cible :

Distribution de la cible

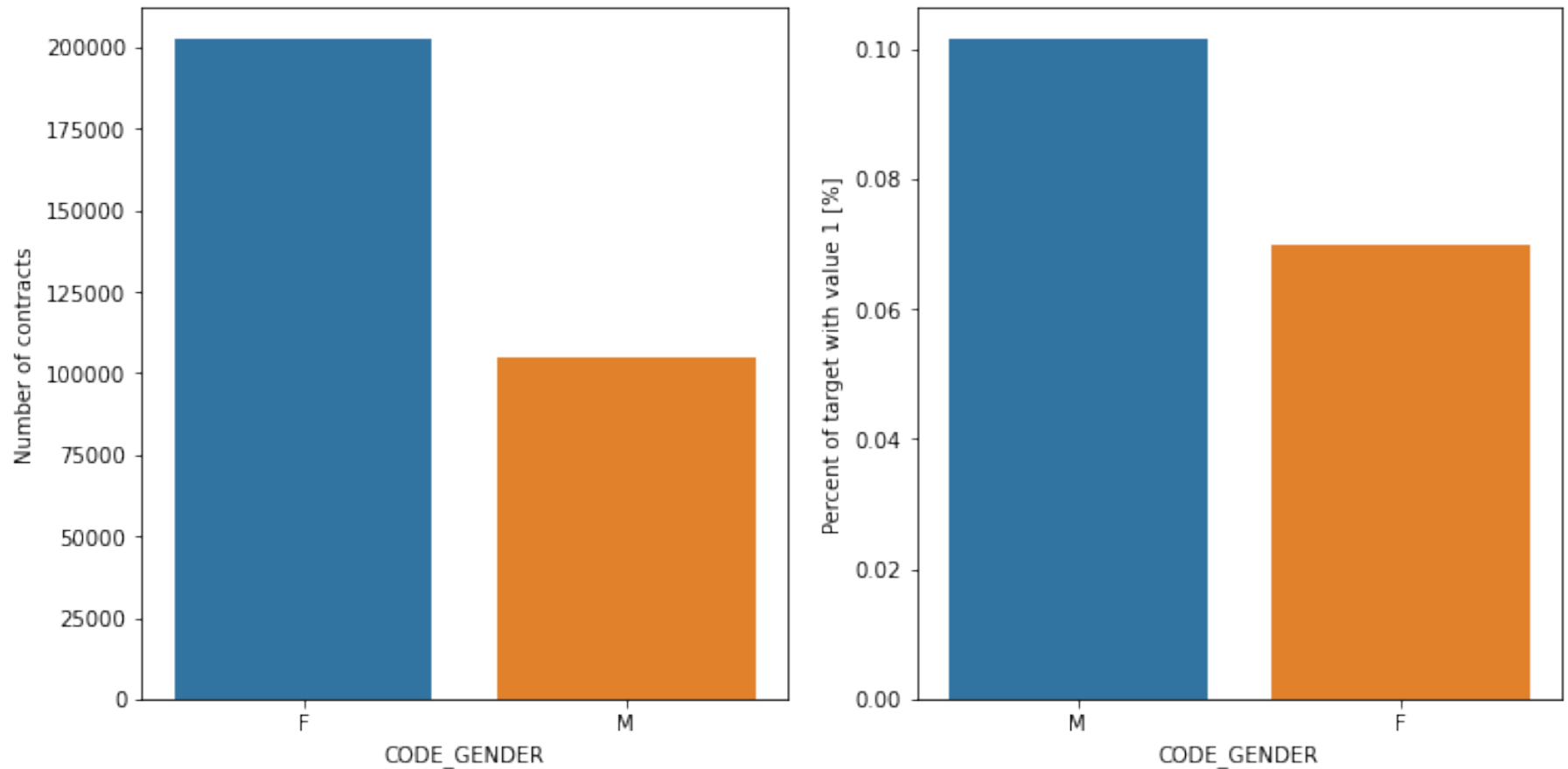


❑ Remarque : On constate un déséquilibre important entre les deux cibles.



Analyse exploratoire des données

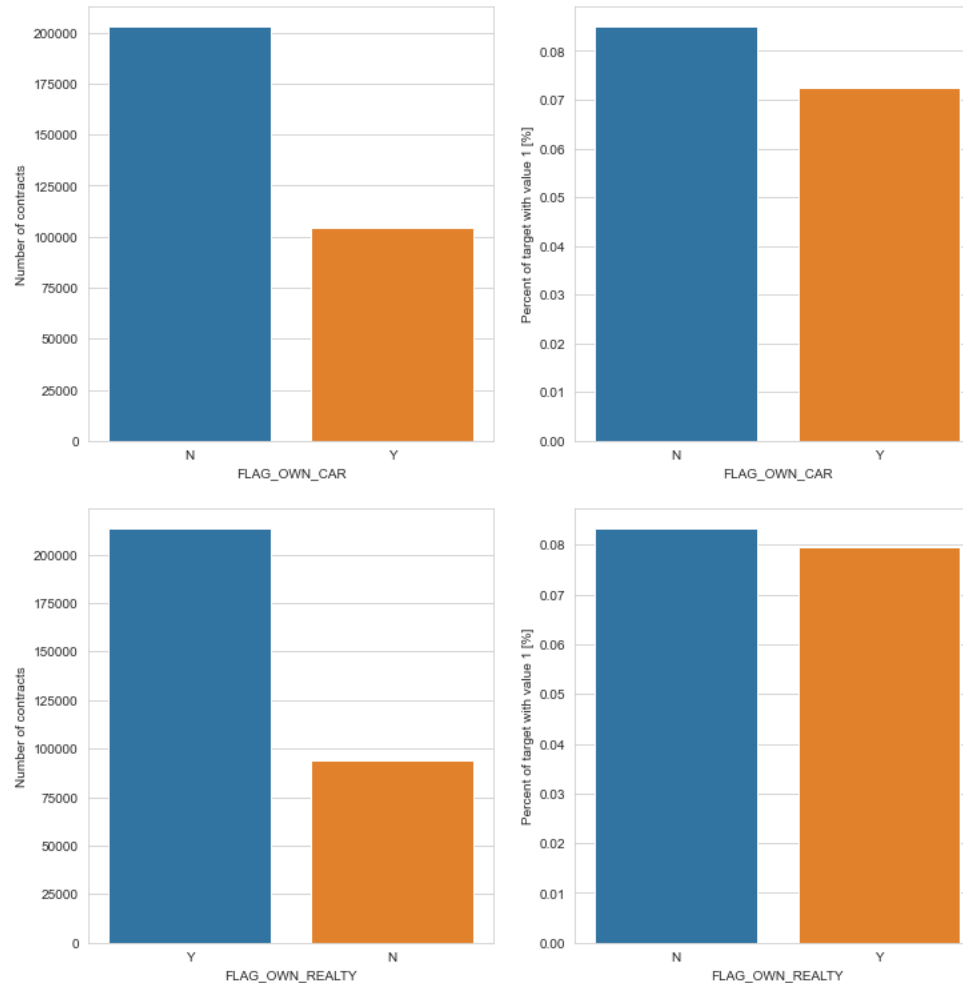
- ❑ Graphique de la distribution par rapport au sexe du client:





Analyse exploratoire des données

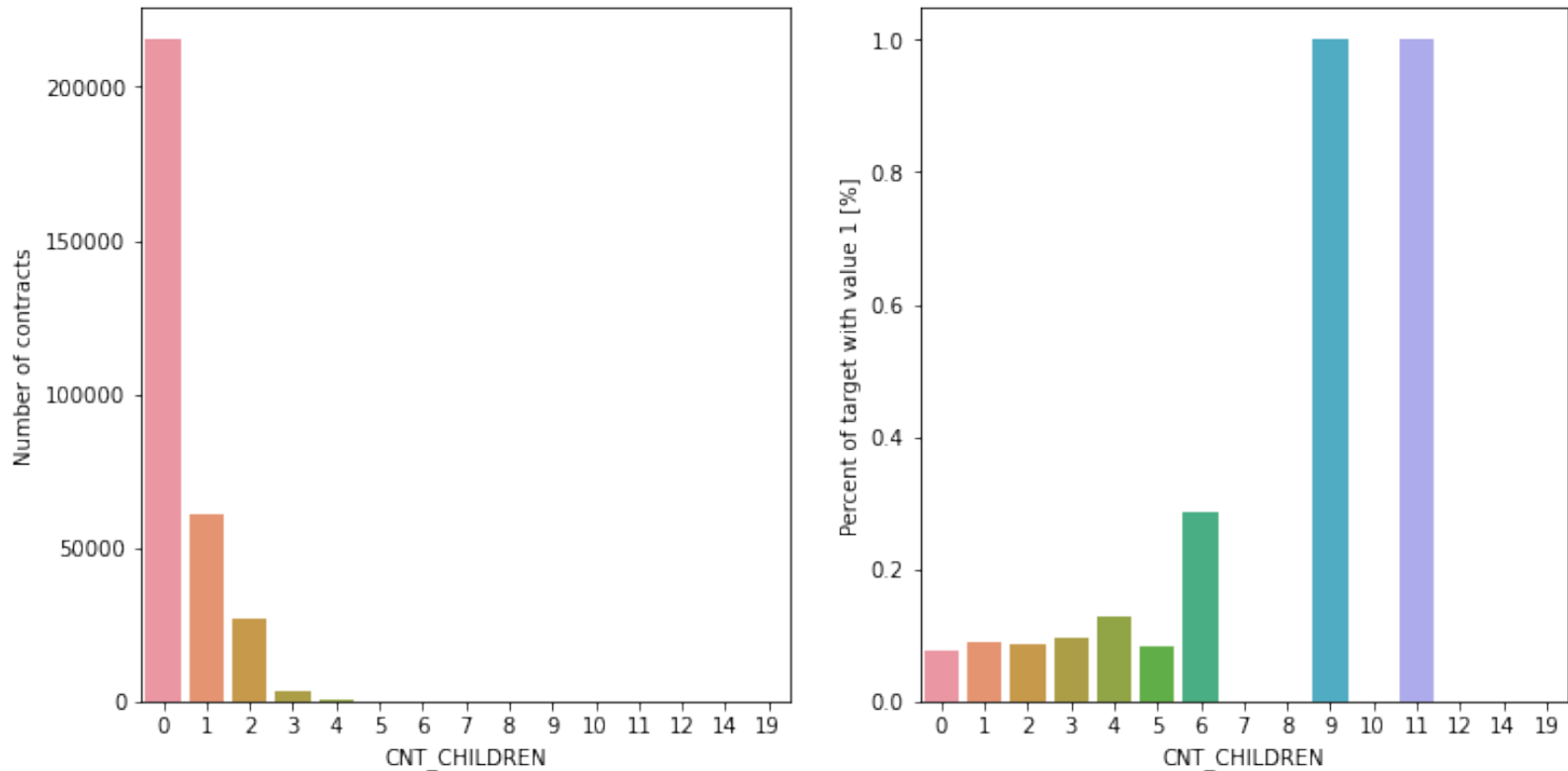
- Graphiques qui indiquent la distribution des clients qui possèdent ou non une voiture ou un bien immobilier :





Analyse exploratoire des données

- Distribution du nombre d'enfants des clients:

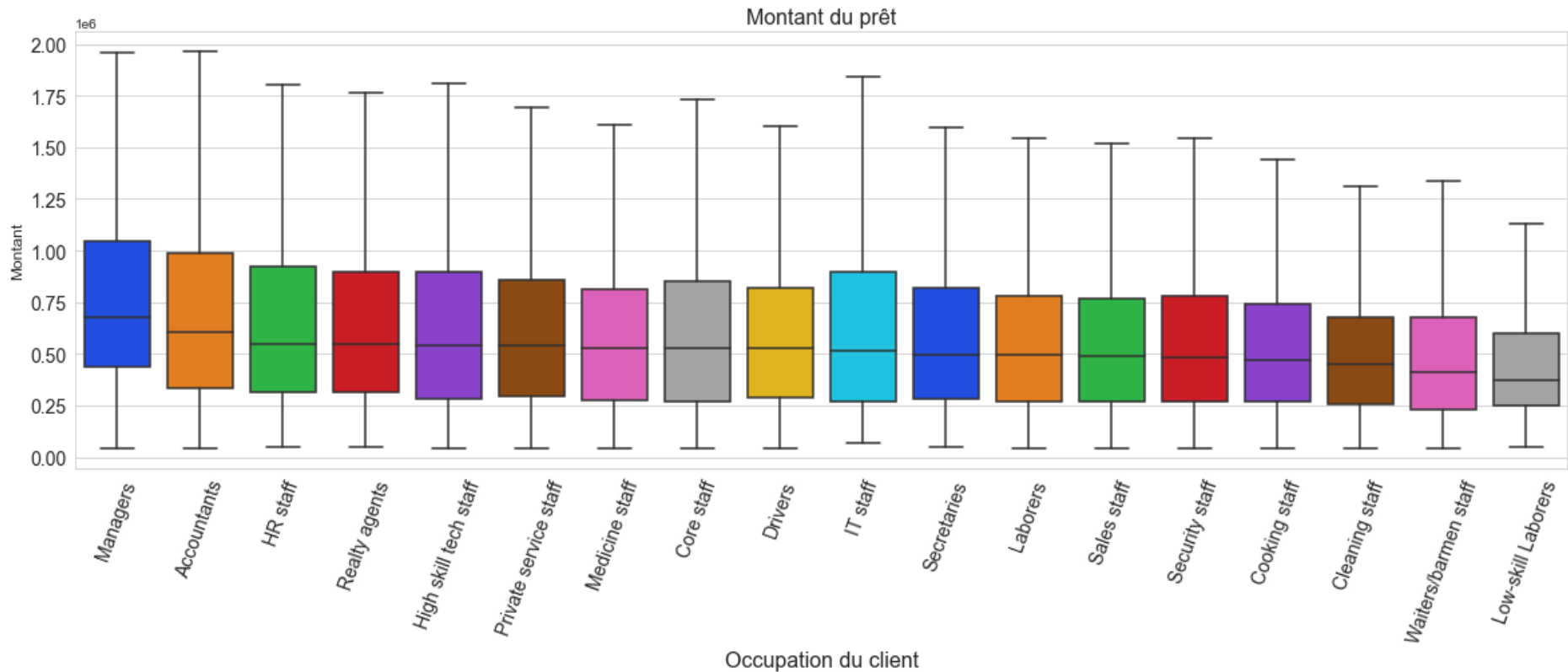


- Remarque : Les clients avec 9 et 11 enfants représentent un risque très élevé pour la banque.



Analyse exploratoire des données

❑ Montant du prêt par type d'occupation :



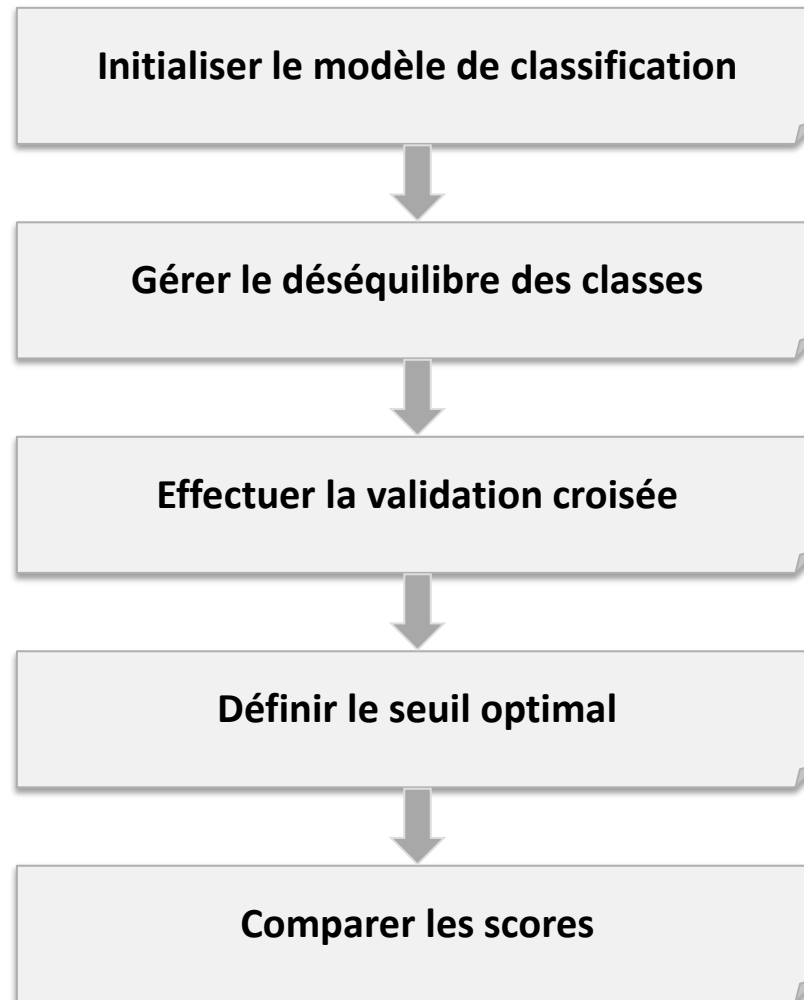
❑ Remarque : Les managers empruntent les montants les plus élevés par rapport aux autres types d'occupations.

Modèles de prédiction



Modèles de prédiction - Processus

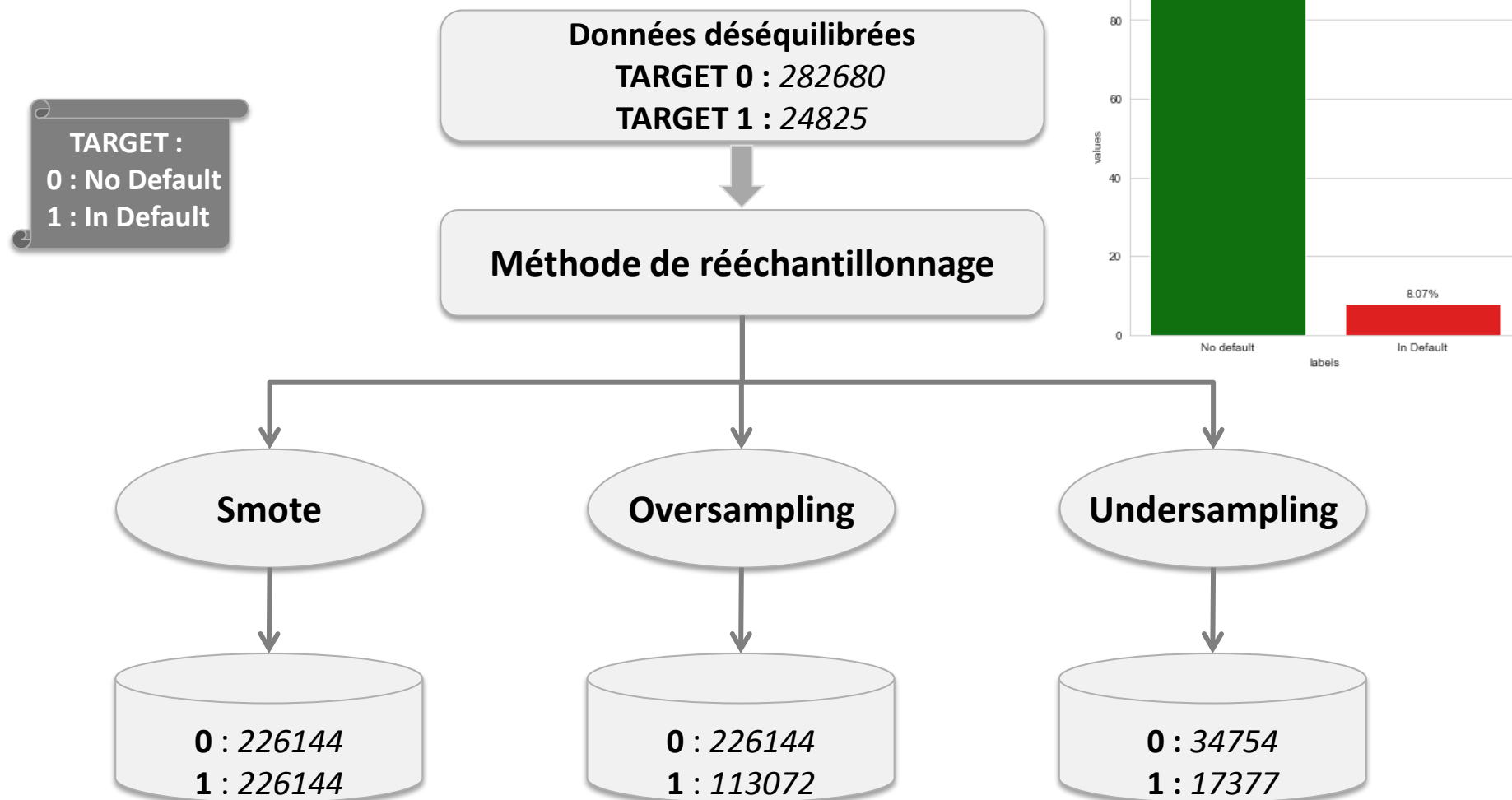
- ❑ La méthodologie de classification pour prédire notre cible :





Modèles de prédiction – Déséquilibre des classes

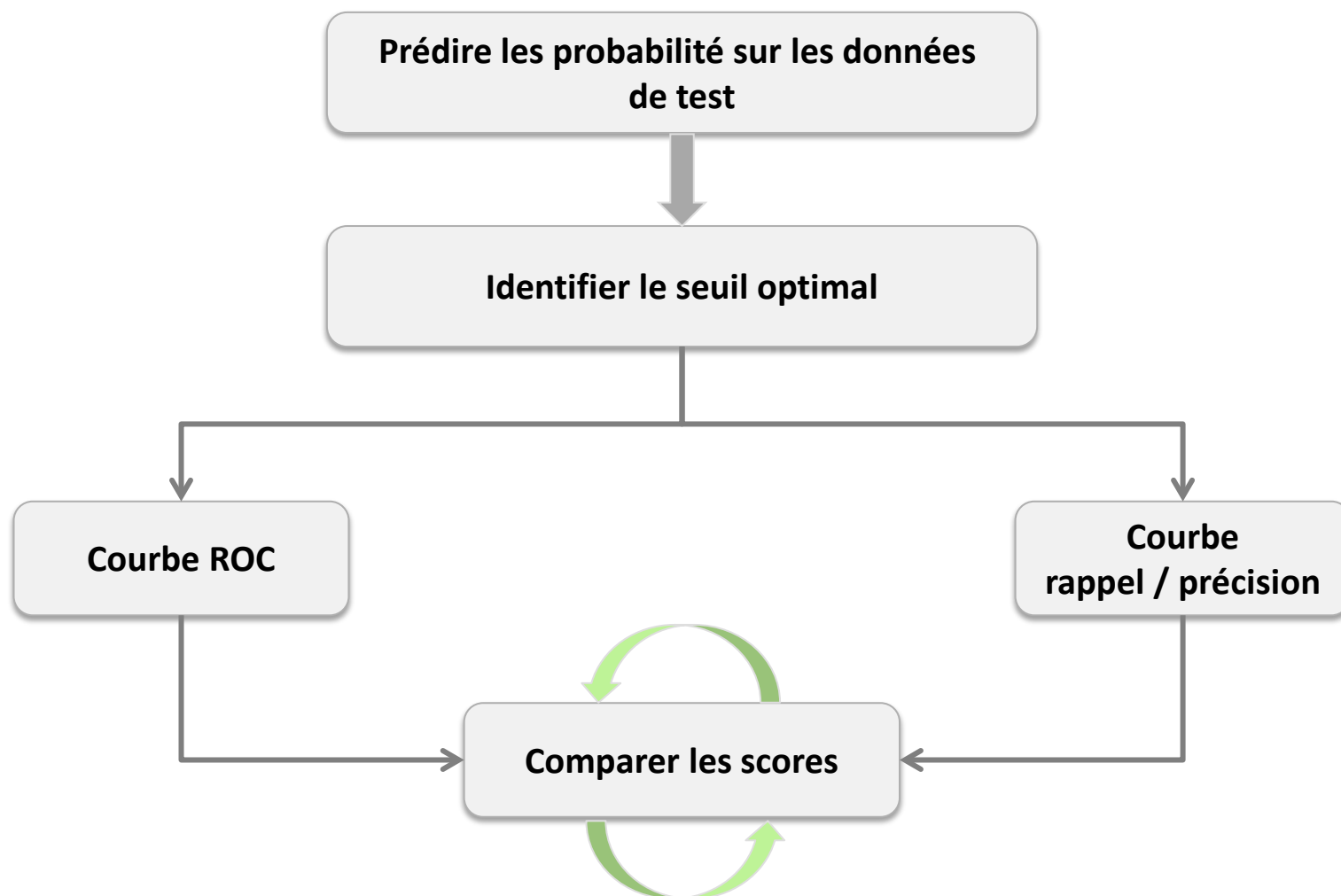
❑ La Gestion du déséquilibre des classes :





Modèles de prédiction – Seuil optimal

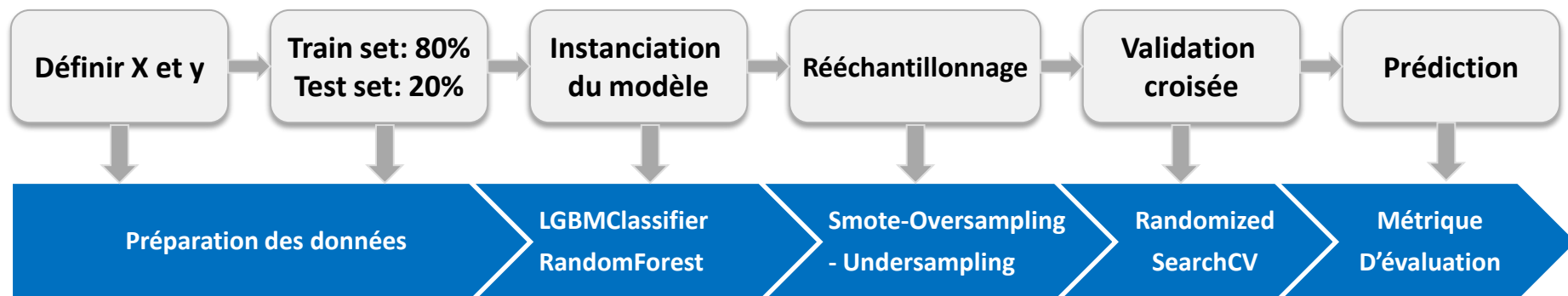
- ❑ Pour identifier le seuil optimal il existe plusieurs méthodes :





Modèles de prédiction – LGBMClassifier - RandomForestClassifier

❑ La méthodologie d'exploration des modèles de classification :



Modèle	Rééchantillonnage	Hyperparamètres
LGBMClassifier	Smote	<code>{'lgbm__scale_pos_weight': 6, 'lgbm__reg_lambda': 10, 'lgbm__reg_alpha': 20, 'lgbm__num_leaves': 15, 'lgbm__n_estimators': 400, 'lgbm__min_child_weight': 10, 'lgbm__min_child_samples': 50, 'lgbm__max_depth': 20, 'lgbm__learning_rate': 0.1, 'lgbm__colsample_bytree': 0.3}</code>
	Oversampling	<code>{'lgbm__scale_pos_weight': 6, 'lgbm__reg_lambda': 100, 'lgbm__reg_alpha': 20, 'lgbm__num_leaves': 20, 'lgbm__n_estimators': 1000, 'lgbm__min_child_weight': 5, 'lgbm__min_child_samples': 10, 'lgbm__max_depth': 5, 'lgbm__learning_rate': 0.09, 'lgbm__colsample_bytree': 0.1}</code>
	Undersampling	<code>{'lgbm__scale_pos_weight': 6, 'lgbm__reg_lambda': 100, 'lgbm__reg_alpha': 7, 'lgbm__num_leaves': 20, 'lgbm__n_estimators': 800, 'lgbm__min_child_weight': 10, 'lgbm__min_child_samples': 50, 'lgbm__max_depth': 20, 'lgbm__learning_rate': 0.09, 'lgbm__colsample_bytree': 0.3}</code>
RandomForestClassifier	Smote	<code>{'rfc__n_jobs': -1, 'rfc__n_estimators': 400, 'rfc__min_samples_split': 2, 'rfc__min_samples_leaf': 5, 'rfc__max_features': 'auto', 'rfc__max_depth': 20}</code>
	Oversampling	<code>{'rfc__n_jobs': -1, 'rfc__n_estimators': 400, 'rfc__min_samples_split': 2, 'rfc__min_samples_leaf': 5, 'rfc__max_features': 'auto', 'rfc__max_depth': 20}</code>
	Undersampling	<code>{'rfc__n_jobs': -1, 'rfc__n_estimators': 400, 'rfc__min_samples_split': 2, 'rfc__min_samples_leaf': 5, 'rfc__max_features': 'auto', 'rfc__max_depth': 20}</code>



Modèles de prédiction – Résultats des scores

LGBMClassifier

Méthode	Score AUC
Smote	0.781
Oversampling	0.782
Undersampling	0.781

RandomForestClassifier

Méthode	Score AUC
Smote	0.731
Oversampling	0.759
Undersampling	0.762

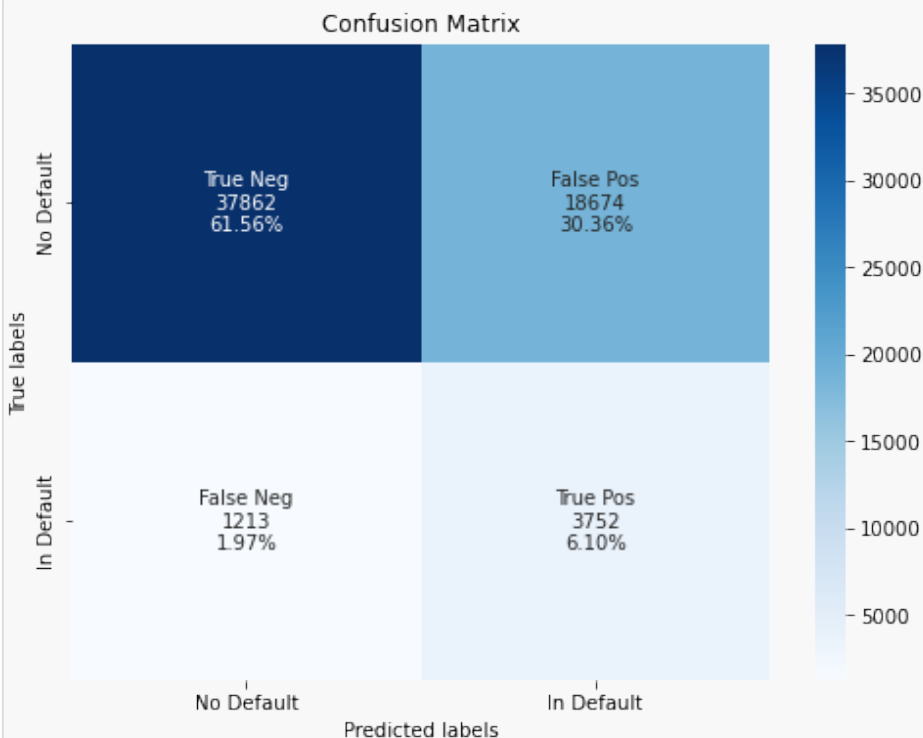
Choix du modèle : LGBMClassifier
Rééchantillonnage : Oversampling



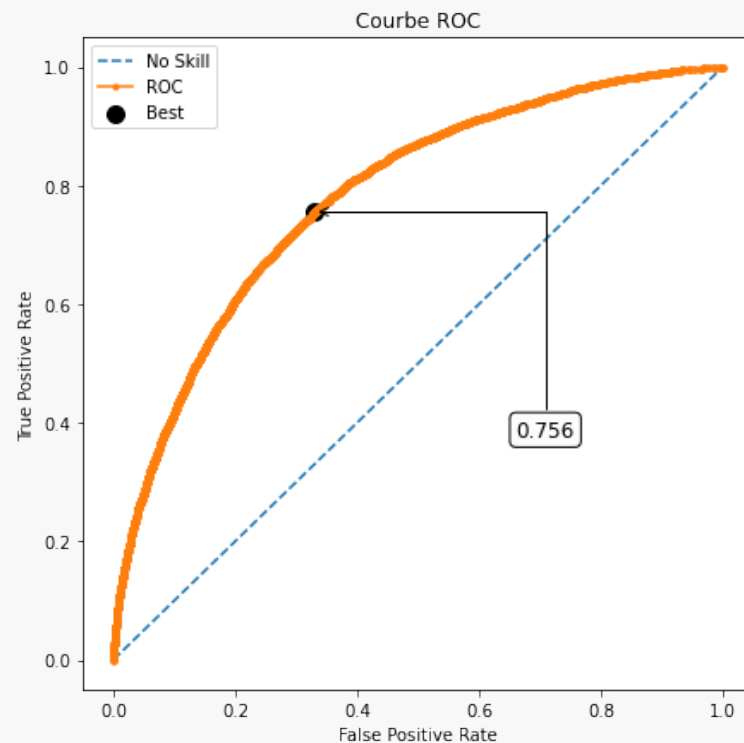
Modèles de prédiction - Métrique d'évaluation

- ❑ La matrice de confusion et la courbe ROC du modèle LightGBM :

Matrice de confusion



Courbe ROC





Modèles de prédiction – Seuil métier

- ❑ Le seuil optimal consiste à calculer le gain obtenu pour l'ensemble des individus du jeu de données.
- ❑ Pour cela nous avons fixé un poids pour chacune des prédictions par rapport à leurs valeurs réelles :

→ fn_value = -10

→ tp_value = 0

→ tn_value = 0

→ fp_value = -1

Confusion matrix		Reality	
		Negative : 0	Positive : 1
Prediction	Negative : 0	True Negative : TN	False Negative : FN
	Positive : 1	False Positive : FP	True Positive : TP

❑ Remarque :

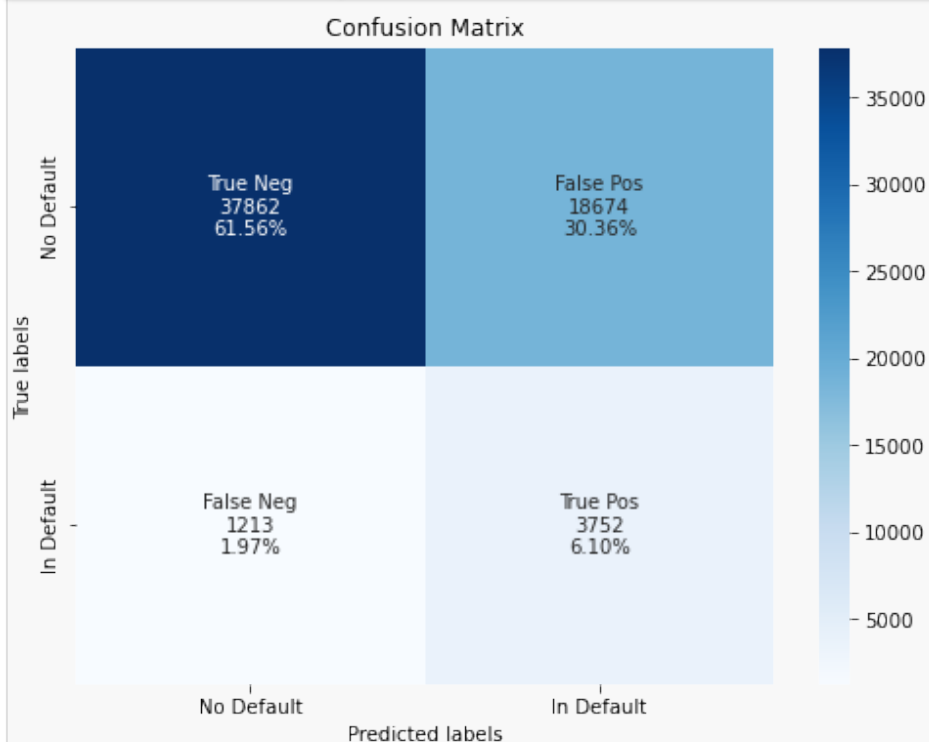
Les prêts accordés aux individus qui ne sont pas solvables sont dotés d'une pénalisation négative de -10. Cette notation peut être modifiée selon les besoins métier.



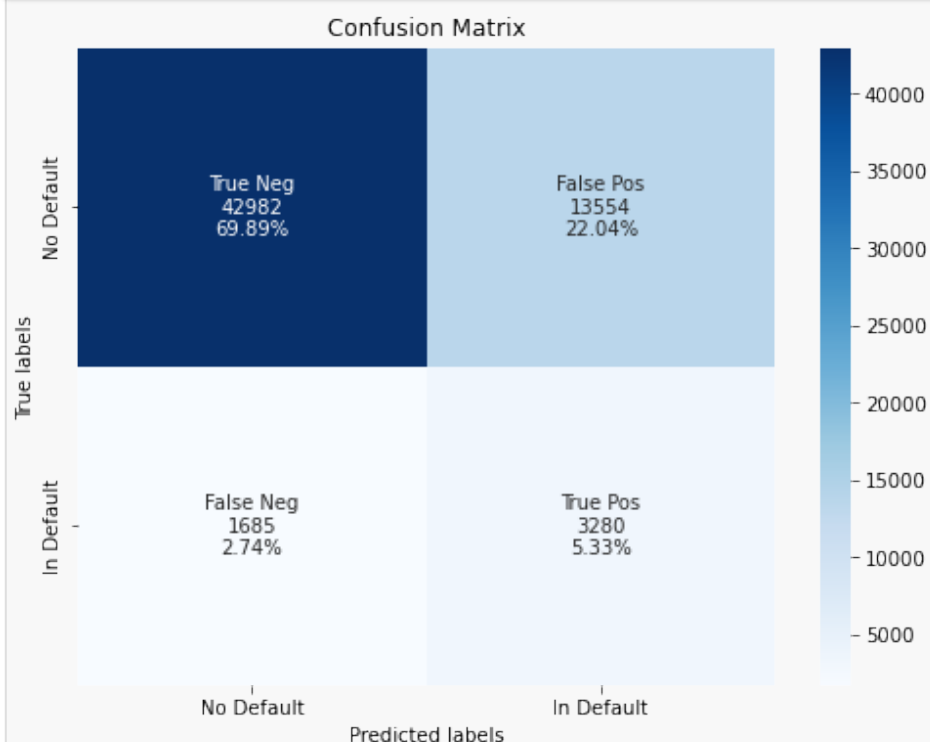
Modèles de prédiction – Seuil métier

- ❑ Comparaison entre le seuil ROC et le seuil métier:

Matrice de confusion – ROC
Seuil optimal ROC = 0.679



Matrice de confusion - seuil métier
Seuil métier = 0.750





Modèles de prédiction – Analyse Shapley

- ❑ **SHAP** (SHapley Additive exPlanations) : Est une approche théorique pour expliquer la sortie de tout modèle d'apprentissage automatique, ci-dessous un exemple (source) :



SHAP



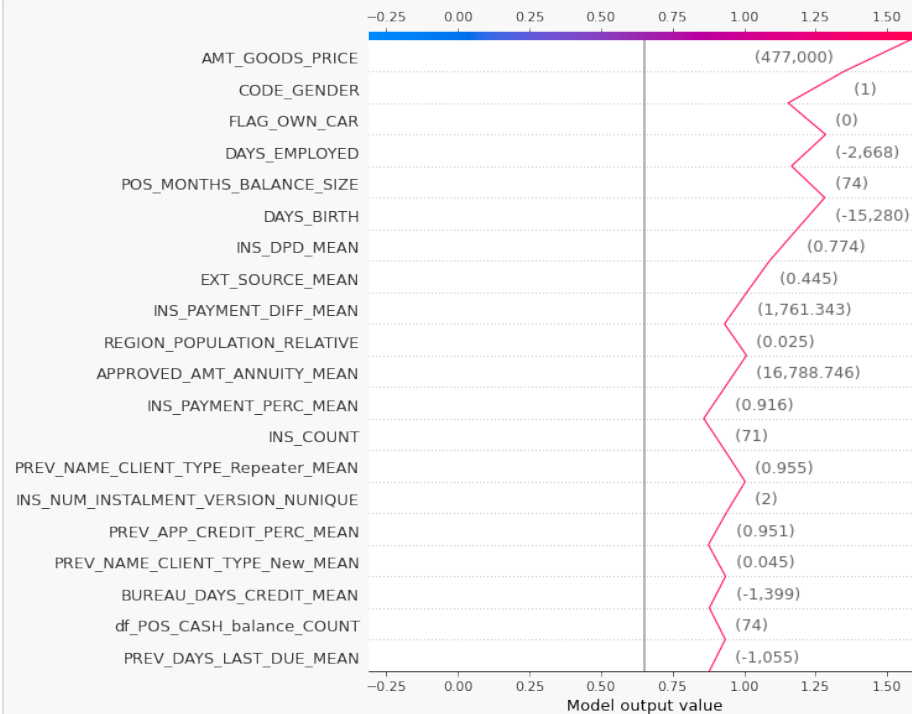
- ❑ **Remarque** : L'approche SHAP va permettre aux chargés de relation client d'expliquer de façon la plus transparente possible les décisions de la banque.



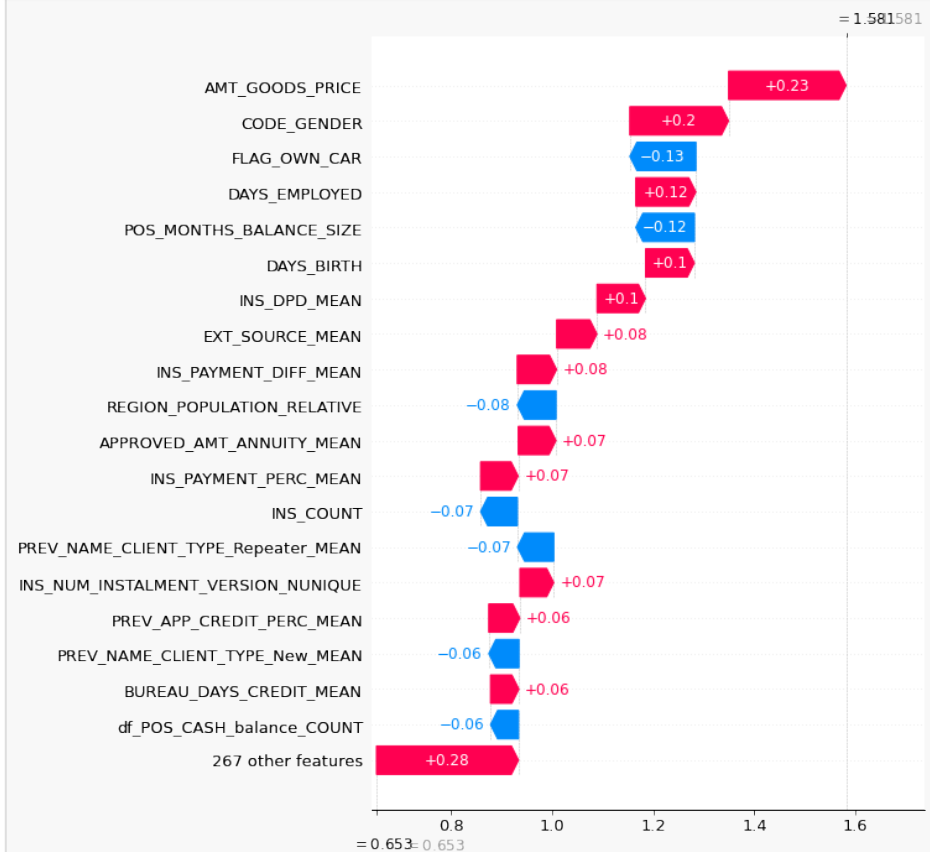
Modèles de prédiction – Analyse Shapley

❑ Exemple de Graphiques SHAP pour le client **100031** :

Decision plot



Waterfall Plot

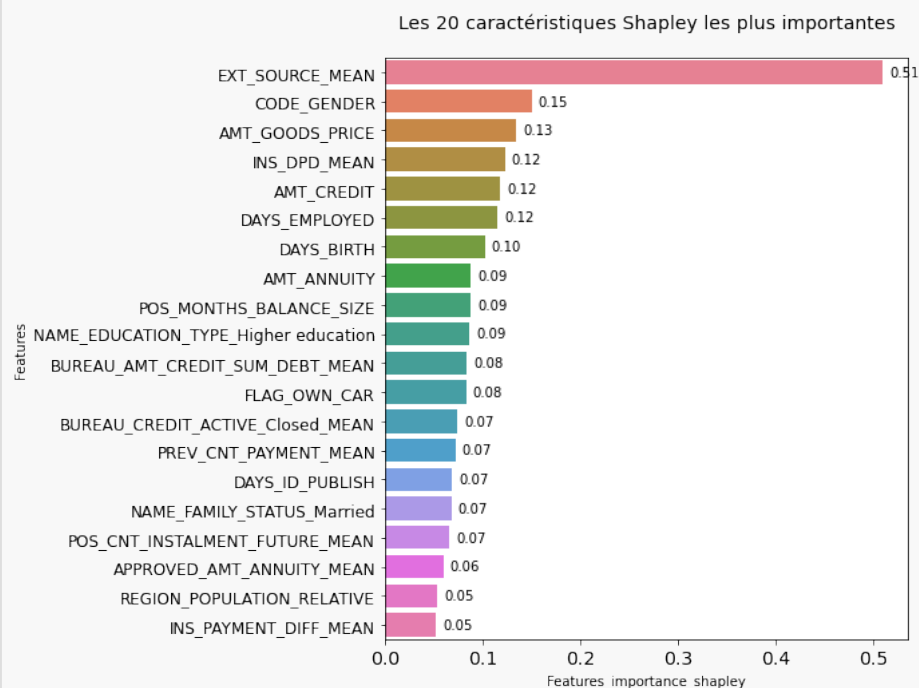




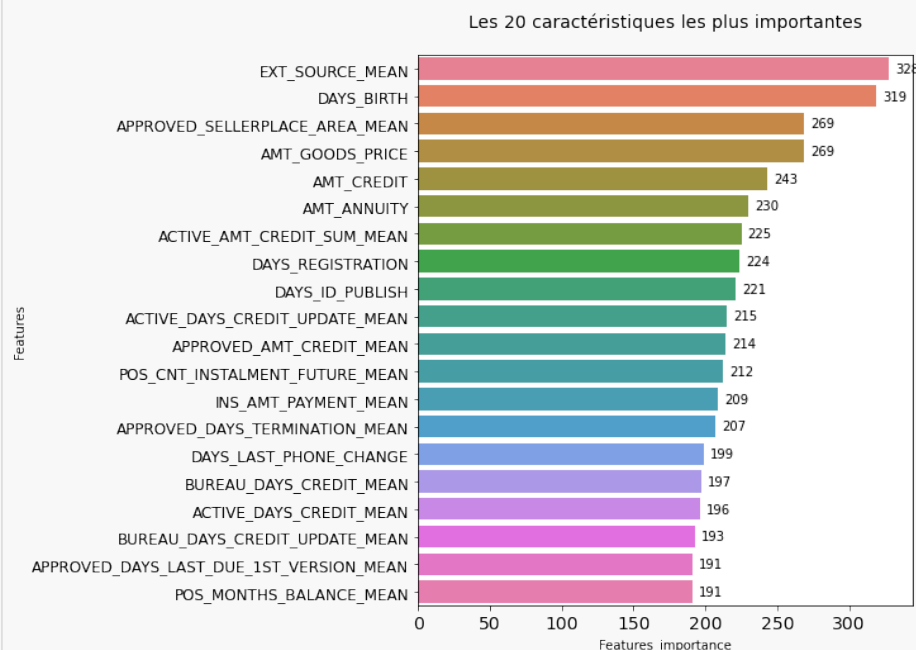
Modèles de prédiction – Features importances

- ❑ Les 20 caractéristiques les plus importantes :

Features importances - Shapley



Features importances - Model






Dashboard



Dashboard – Présentation

- ❑ Les outils utilisés pour développer le Dashboard interactif :

Outil	Description
	Framework open-source qui facilite la création et le partage des applications Web personnalisées pour l'apprentissage automatique (front end).
	Framework Web pour la création d'API, qui permet de faire la prédiction de notre modèle (back end).
	Gestionnaire de versions qui permet aux développeurs de conserver un historique des modifications et des versions de tous leurs fichiers dans un dépôt distant.



Dashboard – Page d'accueil

Sélectionner un client

100030

Seuil de solvabilité en %

71,40

Prêt à dépenser

Data Client

- Analyse
- Shapley
- Description

Informations Client : 100030

	Data
SK_ID_CURR	100030
CODE_GENDER	F
AGE	52.97
FLAG_OWN_CAR	N
FLAG_OWN_REALTY	Y
CNT_CHILDREN	0
AMT_INCOME_TOTAL	90,000.00
AMT_CREDIT	225,000.00
AMT_ANNUITY	11,074.50
AMT_GOODS_PRICE	225,000.00
YEARS_EMPLOYED	9.57

Threshold = 71.4 %

35.2
▼-36.2

✓ Crédit accordé

Allez vers
mon Dashboard

Sélectionner un client

100030

Seuil de solvabilité en %

71,40

Home Credit

- Data Client
- Analyse**
- Shapley
- Description

Prêt à dépenser

Dashboard

Informations Client : 100030

	Data
SK_ID_CURR	100030
CODE_GENDER	F
AGE	52.97
FLAG_OWN_CAR	N
FLAG_OWN_REALTY	Y
CNT_CHILDREN	0
AMT_INCOME_TOTAL	90,000.00
AMT_CREDIT	225,000.00
AMT_ANNUITY	11,074.50
AMT_GOODS_PRICE	225,000.00
YEARS_EMPLOYED	9.57

Threshold = 71.4 %

35.2
▼-36.2

✓ Crédit accordé

Analyse Client : 100030

Sélectionner une ou plusieurs variables

DATA_SMT1

Submit

Client's age in days at the time of application

Variable Description

DATA_SMT1 Client's age in days at the time of application

Allez vers
mon Github

Conclusion



Conclusion

❑ Approches supplémentaires pour :

■ Améliorer le modèle:

- ✓ Améliorer le data engineering avec une analyse comptable plus poussée.
- ✓ Affiner l'analyse avec l'équipe métier pour déterminer le seuil optimal.
- ✓ Optimiser la performance du modèle (par ex : réglage des hyperparamètres).
- ✓ Essayer d'autres techniques d'encodage de données.

■ Améliorer le Dashboard :

- ✓ Optimiser les performances pour un chargement de données plus rapide.
- ✓ Ajouter d'autres onglets pour une analyse complète.
- ✓ Intégrer des graphiques interactifs.
- ✓ Créer une API pour la consultation de la base de données.

OPENCLASSROOMS

Merci pour votre attention
Fin de la présentation