Modèle de scoring

Sommaire

Rappel du projet

Développement d'un **algorithme de scoring** pour aider à décider si un prêt peut être accordé à un client.

L'algorithme devra calculer la probabilité qu'un client le rembourse ou non.

Problématique métier

Evaluer si un client sera à même de rembourser ou non est une tache complexe, il faut prendre en compte de nombreux paramètres.

Le modèle de scoring sera une aide à la décision. L'accord ou le refus de prêt par le modèle devra être compréhensible par les agents.

Jeu de données

7 sources de données (1 seule utilisée dans un premier temps)

307511 demandes de crédit

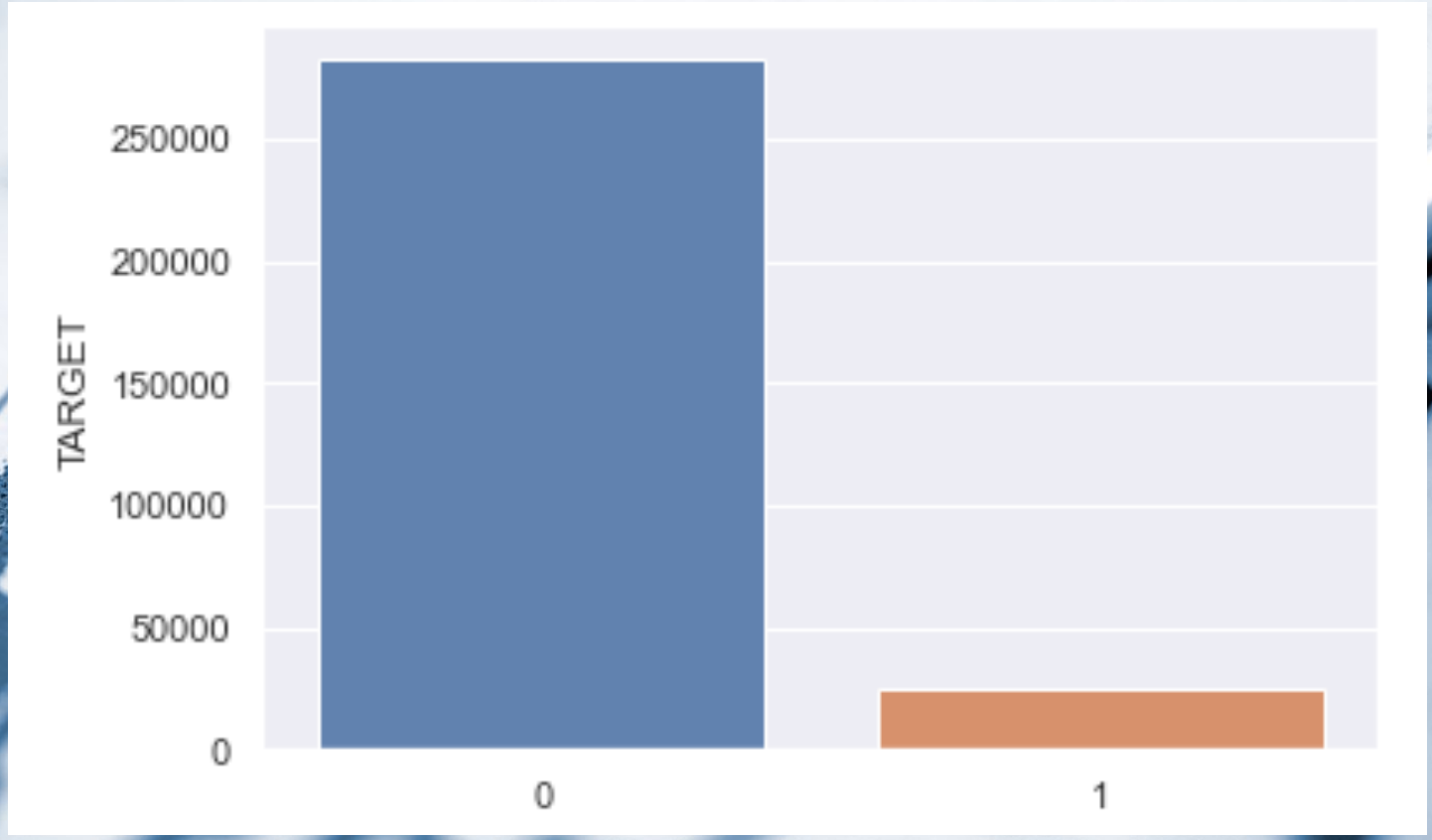
121 variables explicatives (emploi, revenus, montant du crédit, durée du crédit, etc.)

1 variable cible qui détermine si dans le passé le client a eu des difficultés de paiement ou non

Variable cible

Détermine si le client a eu des difficultés de paiement ou non

Les données sont déséquilibrées. Il y a plus de clients sans retard de paiement que de clients avec un retard de paiement.



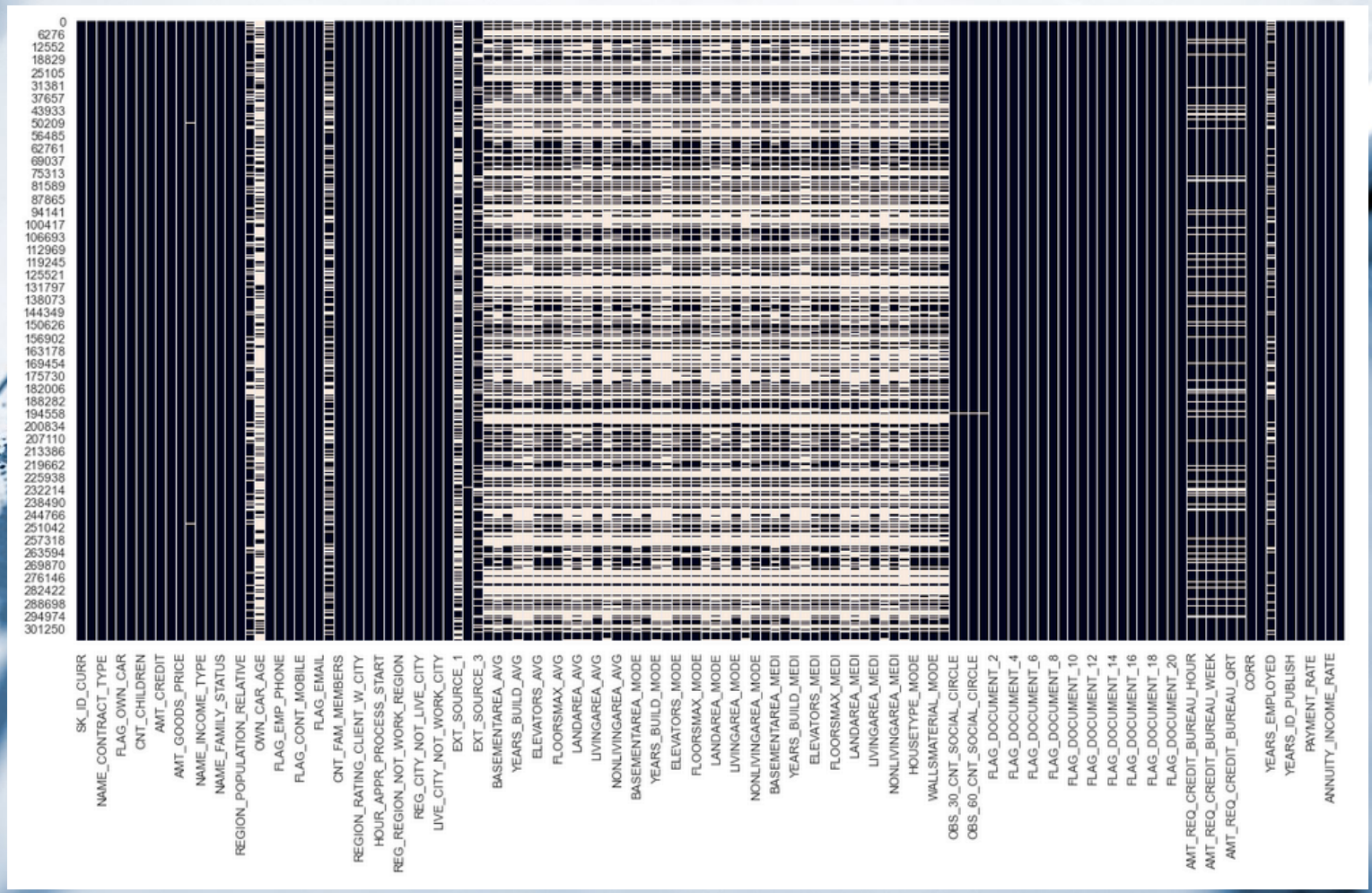
Under-sampling, réduction les données majoritaires à la taille des données minoritaires.

SMOTE (Synthetic Minority Oversampling Technique), création de nouvelles données

Variable explicatives

121 variables explicatives

Données manquantes



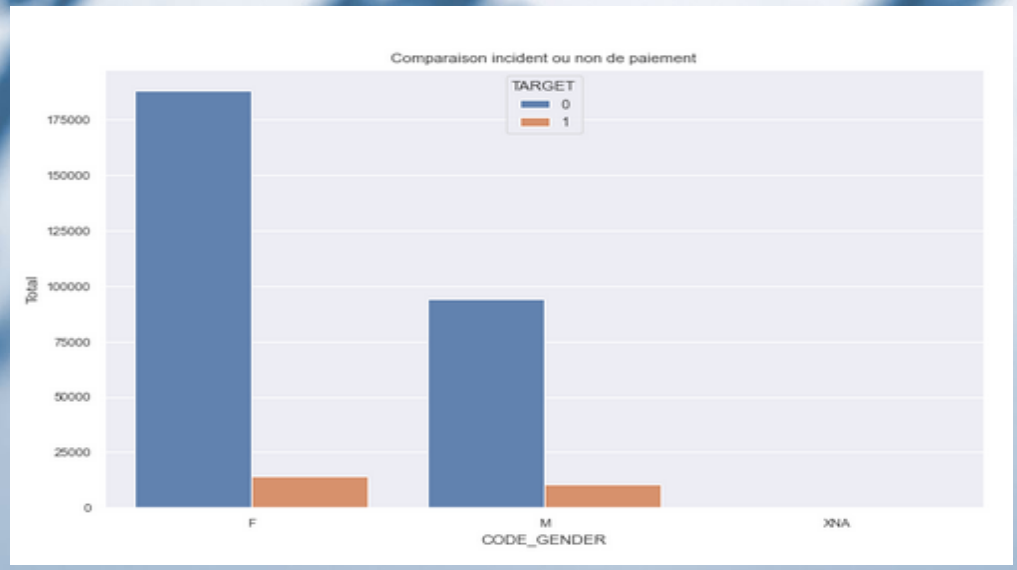
Il y a 41 variables qui ont plus de 50% de données manquantes

Variable explicatives

Genre du client (Homme, Femme, non défini)

Aucune valeur manquante

66% des femmes contractent un crédit



10% des hommes ne remboursent pas leur crédit

7% des femmes ne remboursent pas leur crédit

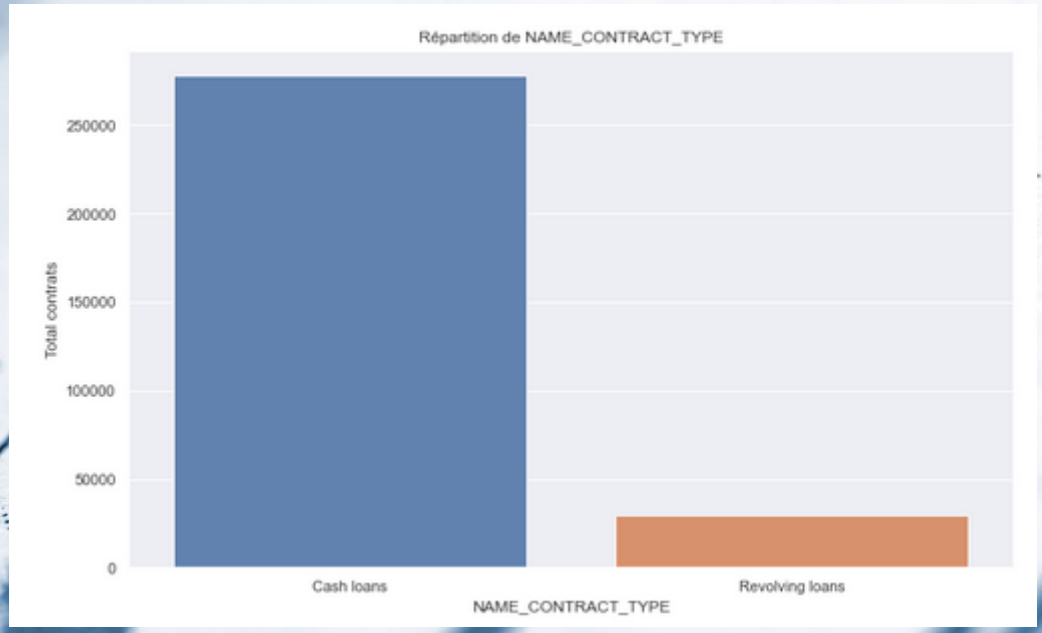
Variable explicatives

Type de contrat

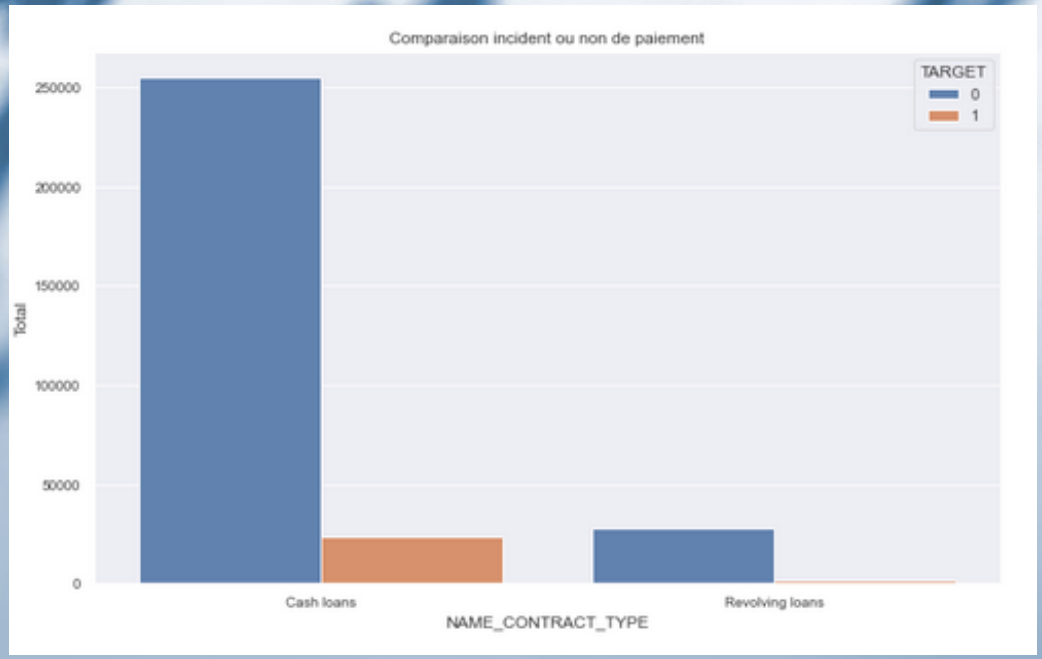
Crédit standard

Crédit renouvelable

Aucune valeur manquante



10% de crédits renouvelables



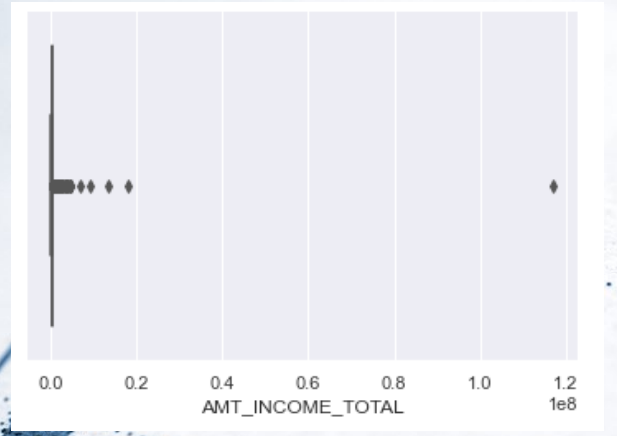
5% des crédits renouvelables ne sont pas remboursés

Variable explicatives

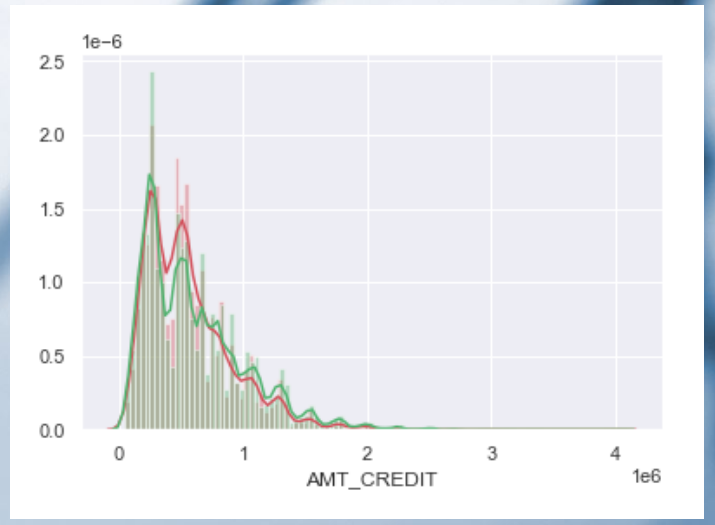
Revenu du client

Les données contiennent une valeur aberrante

Aucune valeur manquante



Cette variable comporte une valeur aberrante (117000000) qui fausse l'analyse. Il sera donc nécessaire de la supprimer.



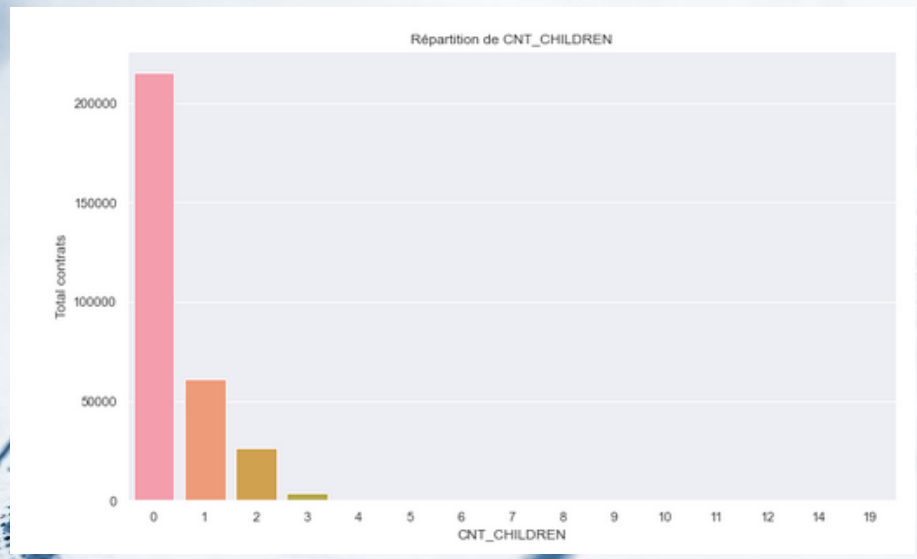
Les crédits autour de 270 000 en monnaie locale sont moins remboursés que les autres

Variable explicatives

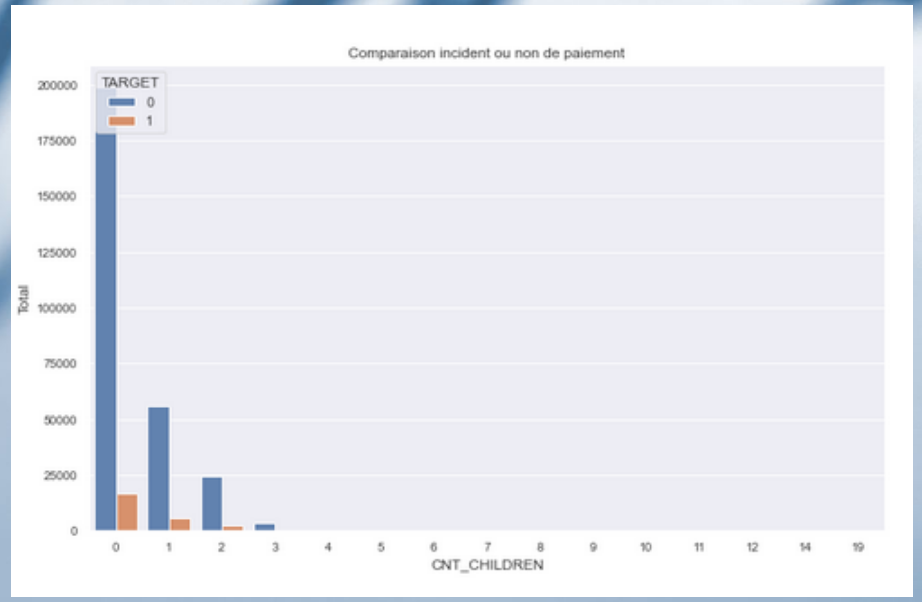
Nombre d'enfants

De 0 à 19 enfants

Aucune valeur manquante



70% des clients n'ont pas d'enfant



100% des clients qui ont 9 enfants ne remboursent pas leur crédit

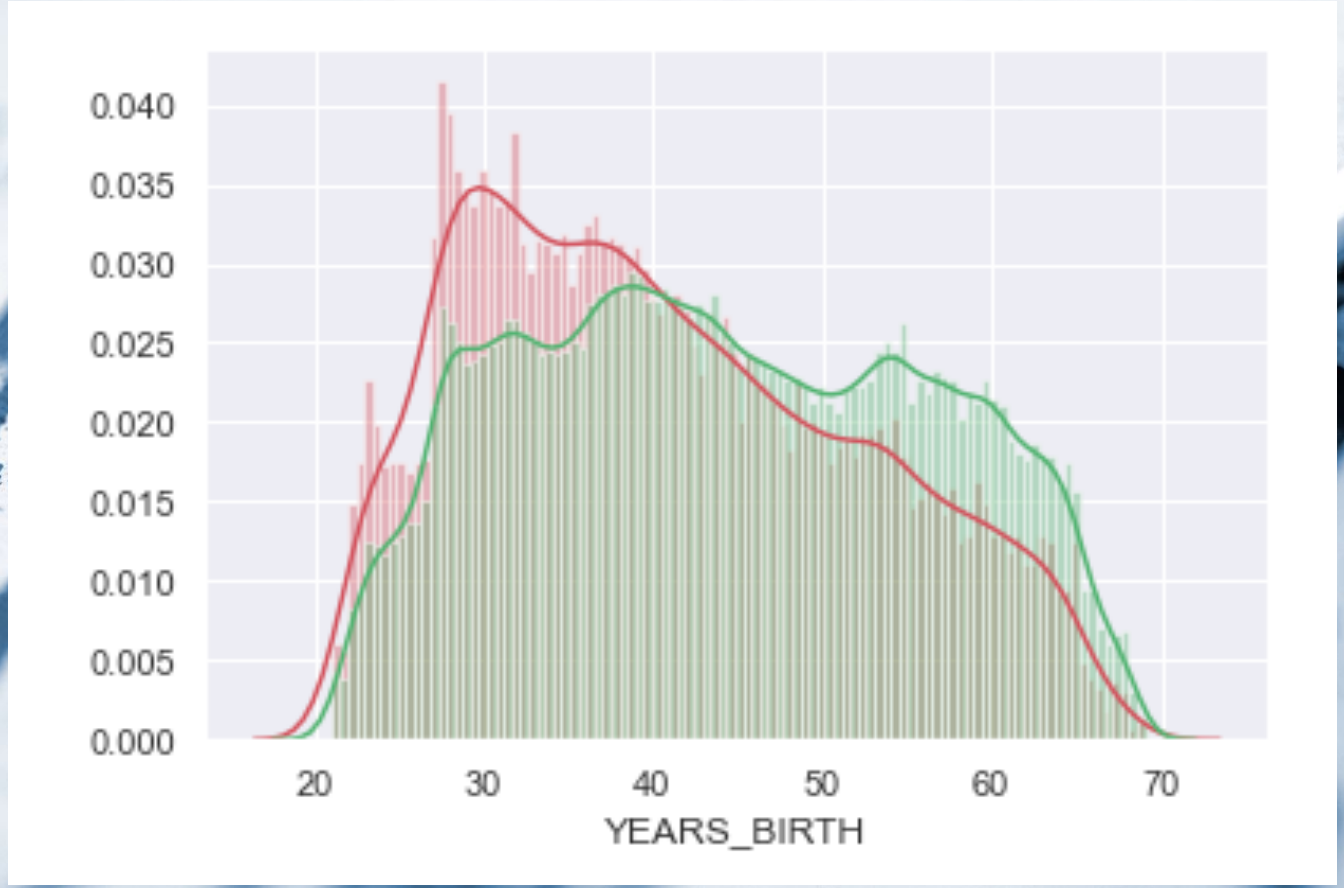
100% des clients qui ont 11 enfants ne remboursent pas leur crédit

Variable explicatives

Âge du client en jours au moment de la demande

Cette donnée était en jours et a été transformée en années afin d'être plus parlante

Aucune valeur manquante



Les clients ont entre 20 ans et 70 ans et la moyenne se situe autour de 43 ans.

Les clients jeunes (25-30 ans) auront plus tendance à ne pas rembourser leur crédit au contraire des plus âgés (55-65 ans)

Modification de variables

Transformation de l'âge du client en jours au moment de la demande en années

Transformation en années du nombre jours avant la demande où la personne à commencé son emploi actuel

Transformation en années du nombre de jours avant la demande où le client a modifié son inscription

Transformation en années du nombre de jours avant la demande le client où il a modifié le document d'identité avec lequel il avait demandé le prêt

Transformation logarithmique du revenu

Nouvelles variables

Pourcentage entre les annuités et le montant du crédit

(AMT_ANNUITY / AMT_CREDIT) * 100

Pourcentage entre le revenu annuel et le montant emprunté

(AMT_INCOME_TOTAL / AMT_CREDIT) * 100

Pourcentage entre les annuités et le revenu annuel

(AMT_ANNUITY / AMT_INCOME_TOTAL) * 100

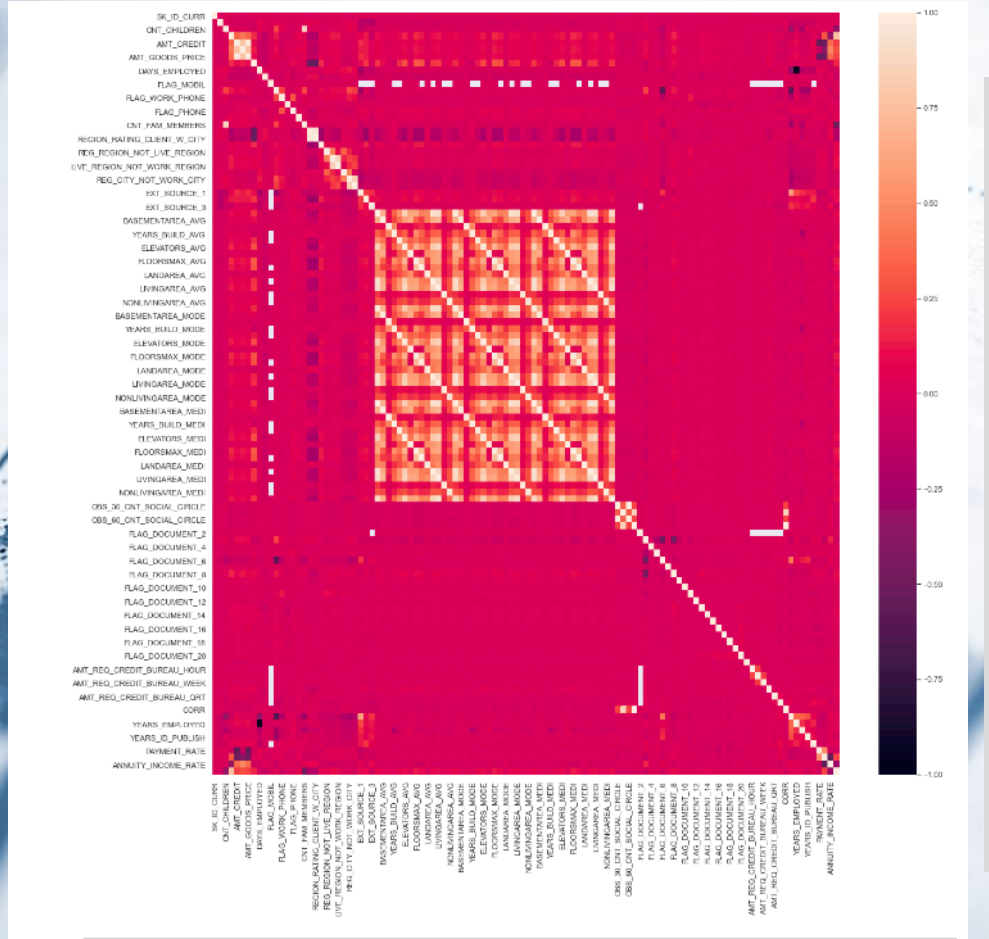
Revenu annuel par personne au foyer

(AMT_INCOME_TOTAL / CNT_FAM_MEMBERS) * 100

Corrélations

Les corrélations entre les variables vont nous permettre de détecter les variables qui seraient semblables.

Les corrélations étant à plus de 90% seront éliminées. Une seule des deux variables sera conservée.



Corrélation de 99% entre le montant du crédit et le montant de l'achat.

Corrélation de 99% entre les clients qui ont 30 jours de retard et ceux qui ont 60 jours de retard.

Corrélation de 99% pour l'appréciation de la région et l'appréciation de la ville

61 variables sont corrélées à plus de 90%

Préparation des variables

Suppression des variables ayant plus de 90% de corrélation

61 variables corrélées entre elles

Imputation des variables binaires

Transformation de Y/N par 1/0

Imputation des variable qualitative

Transformation des variables manquantes

Transformation en indicateurs

Imputation des variables quantitatives

Utilisation de la médiane pour les valeurs manquantes

Normalisation des données

Classement des variables

Classement des variables en fonction de leur importance (selectKBest)

1. Score normalisé 2 provenant d'une source de données externe
2. Score normalisé 3 provenant d'une source de données externe
3. Score normalisé 1 provenant d'une source de données externe
4. Année de naissance
5. Nombre d'années de travail

...

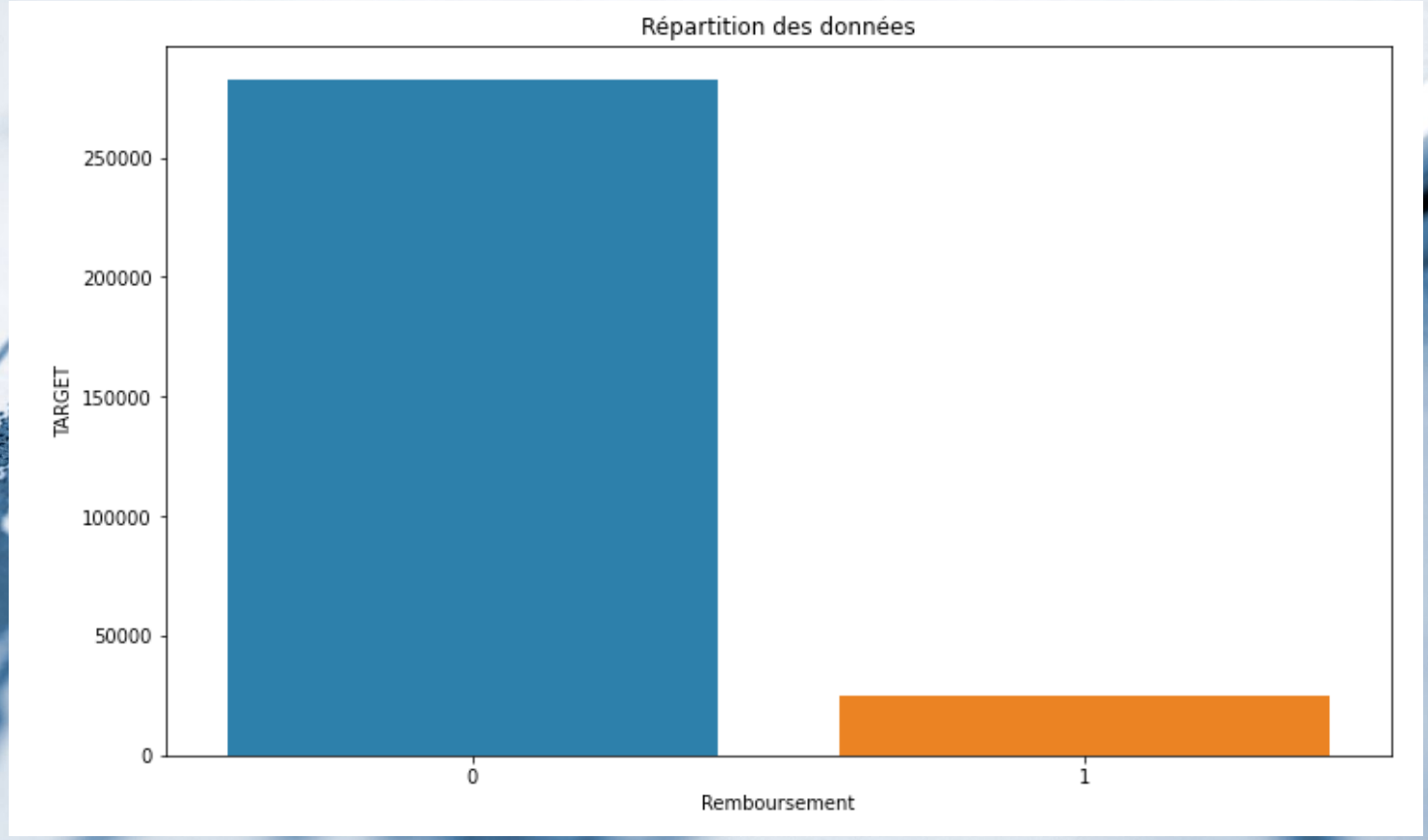
87. Le client a-t-il fourni le document 5 ?
88. Situation du client en matière de logement
89. Le client a-t-il fourni le document 20 ?
90. Type d'organisation où le client travaille
91. Nombre de demandes de renseignements au bureau de crédit concernant le client une heure avant la demande

Variable à prédire

Déséquilibre des données

Under-Sampling

Over-Sampling (SMOTE)



24 824 clients qui ne remboursent pas leur crédit
282 686 clients qui remboursent leur crédit

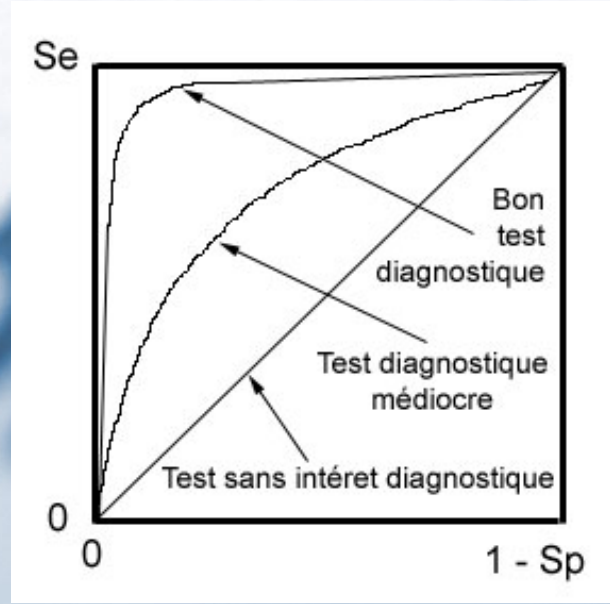
Évaluation scoring

Identifier les clients ayant un fort risque de ne pas rembourser

	Prédit non solvable	Prédit solvable
Non solvable	Vrai positif	faux négatif
Solvable	faux positif	vrai négatif

Éviter les **faux négatifs**, c'est-à-dire prédire que le client est solvable alors qu'il ne l'est pas.

Éviter les **faux positif**, c'est-à-dire que le client n'est pas solvable alors qu'il est.



Une courbe ROC trace les valeurs TVP et TFP pour différents seuils de classification

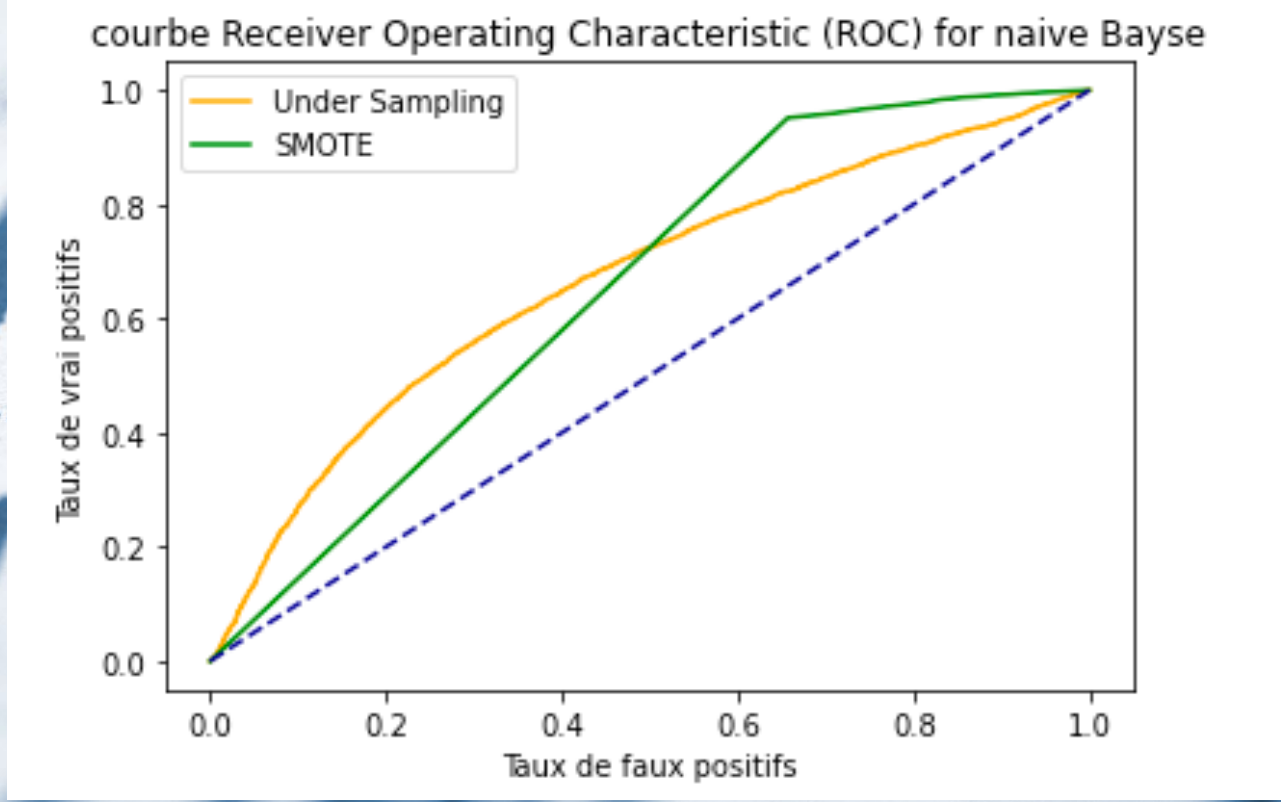
Test des modèles

Naive Bayse

Recherche du nombre de variables optimales

Under-Sampling

Over-Sampling (SMOTE)

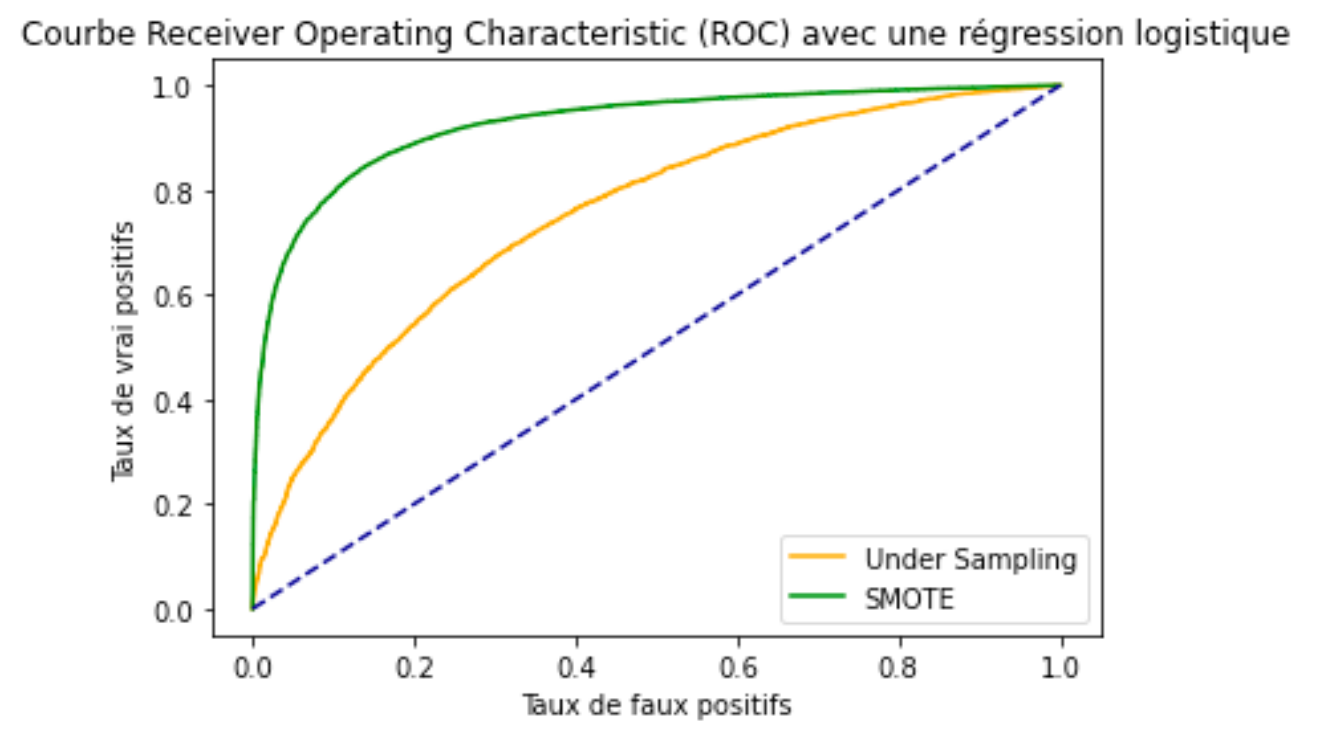


Under-Sampling
181 est le nombre de variables optimales avec un score de 0.64

Over-Sampling
81 est le nombre de variables optimales avec un score de 0.76

Test des modèles

Régression Logistique



Under-Sampling

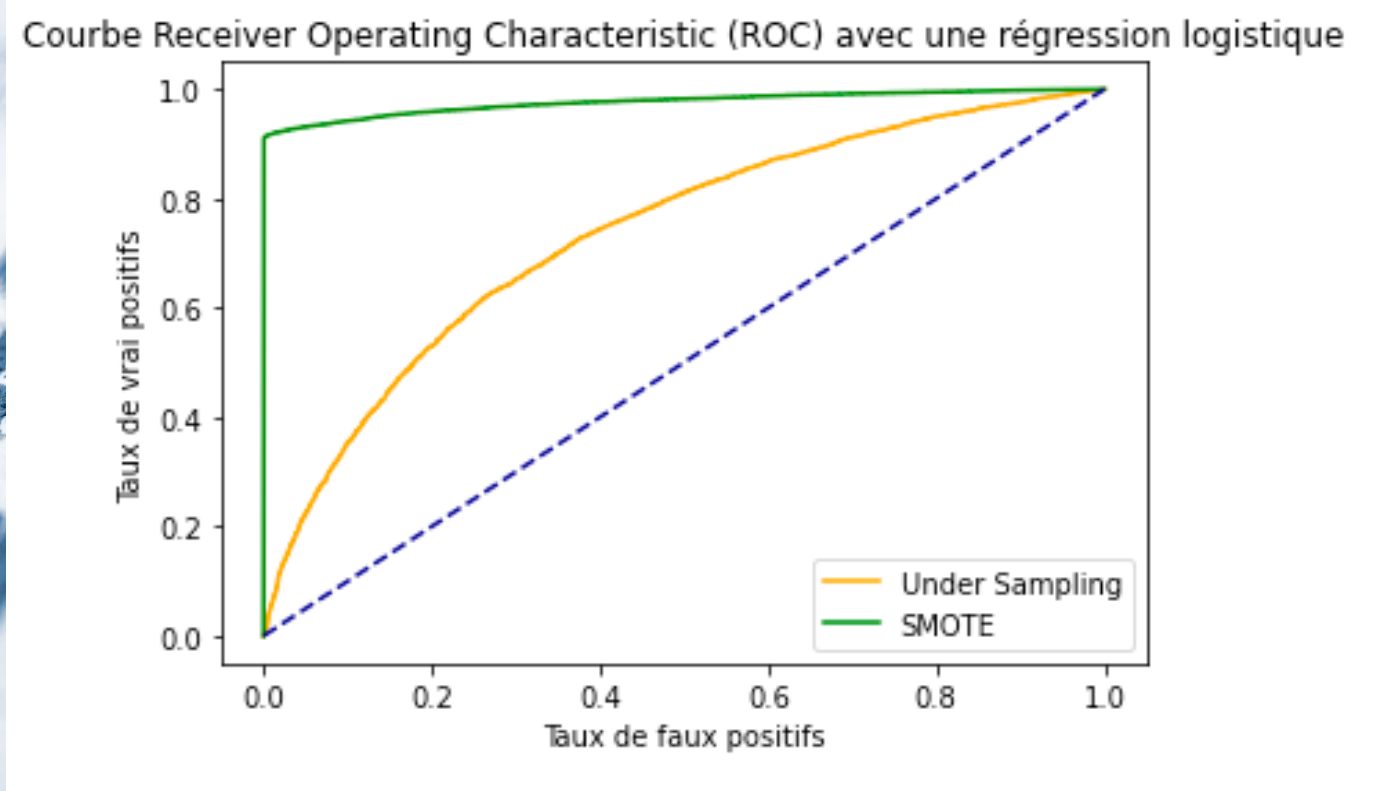
116 est le nombre de variables optimales avec un score de 0,75

Over-Sampling

113 est le nombre de variables optimales avec un score de 0,98

Test des modèles

Descente de gradient stochastique



Under-Sampling

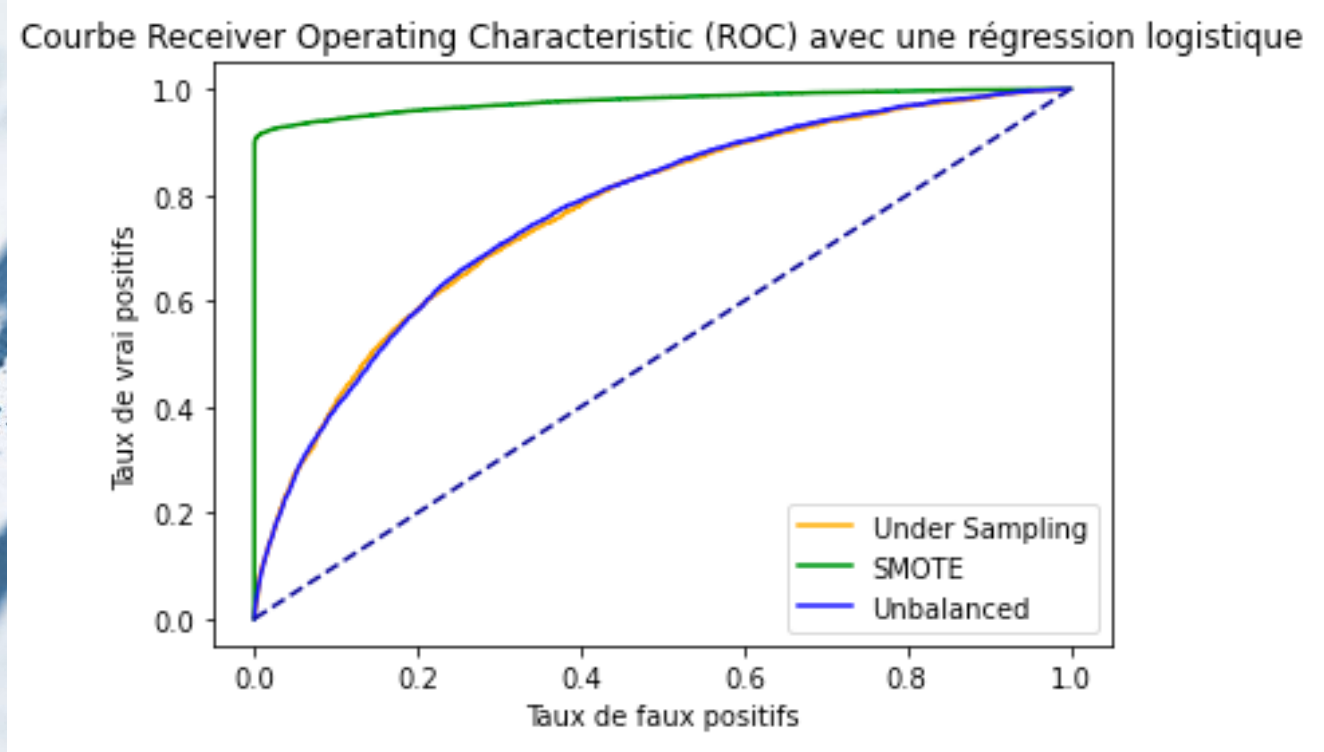
5 est le nombre de variables optimales avec un score de 0,73

Over-Sampling

188 est le nombre de variables optimales avec un score de 0,98

Test des modèles

Light Gradient Boosting



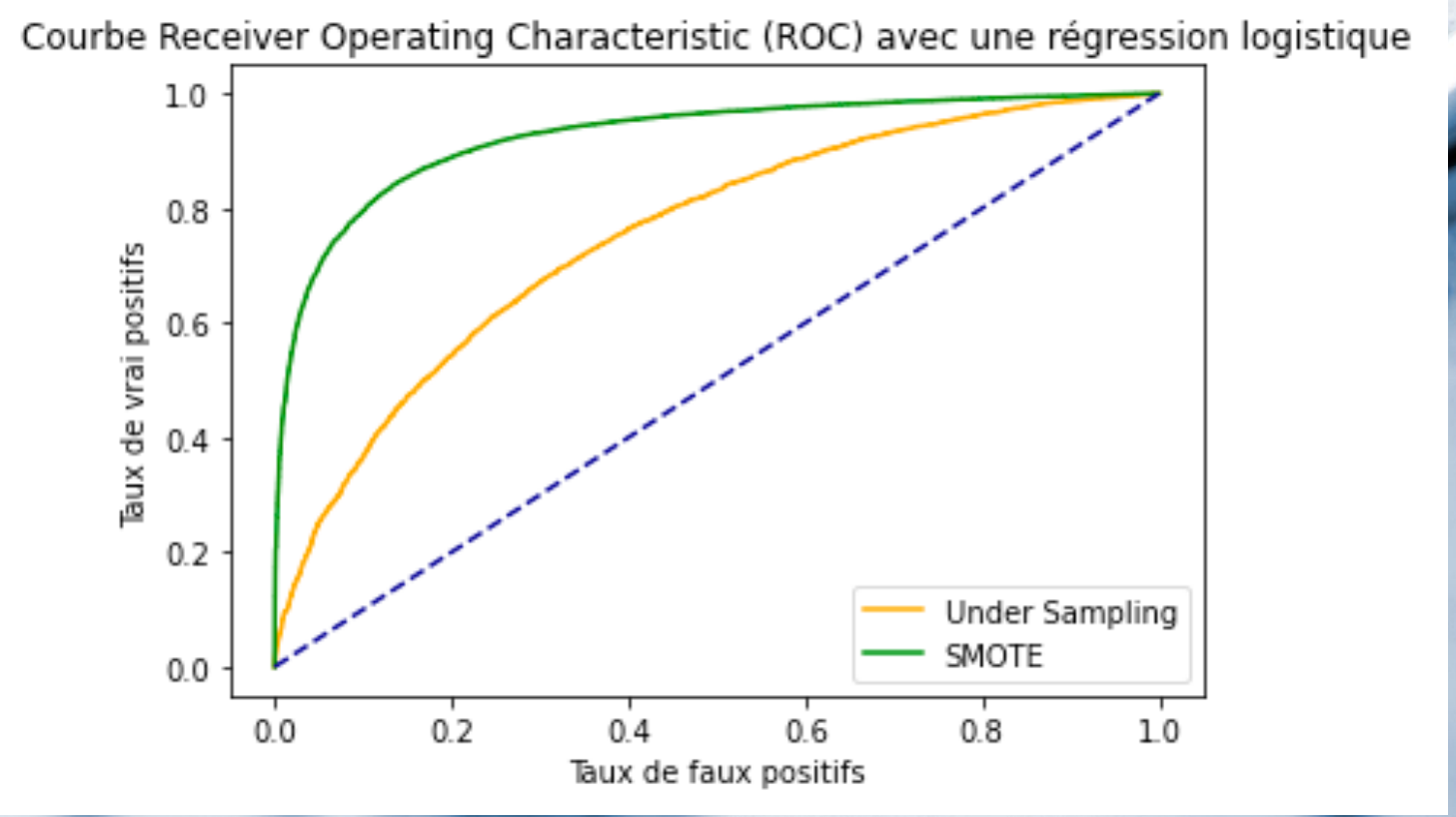
Under-Sampling
112 est le nombre de variables optimales avec un score de 0,77

Over-Sampling
81 est le nombre de variables optimales avec un score de 0,98

Tel quel
201 est le nombre de variables optimales avec un score de 0,77

Test des modèles

Random Forest



Under-Sampling
130 est le nombre de variables optimales avec un score de 0,75

Over-Sampling
79 est le nombre de variables optimales avec un score de 0,98

Sélection du modèle

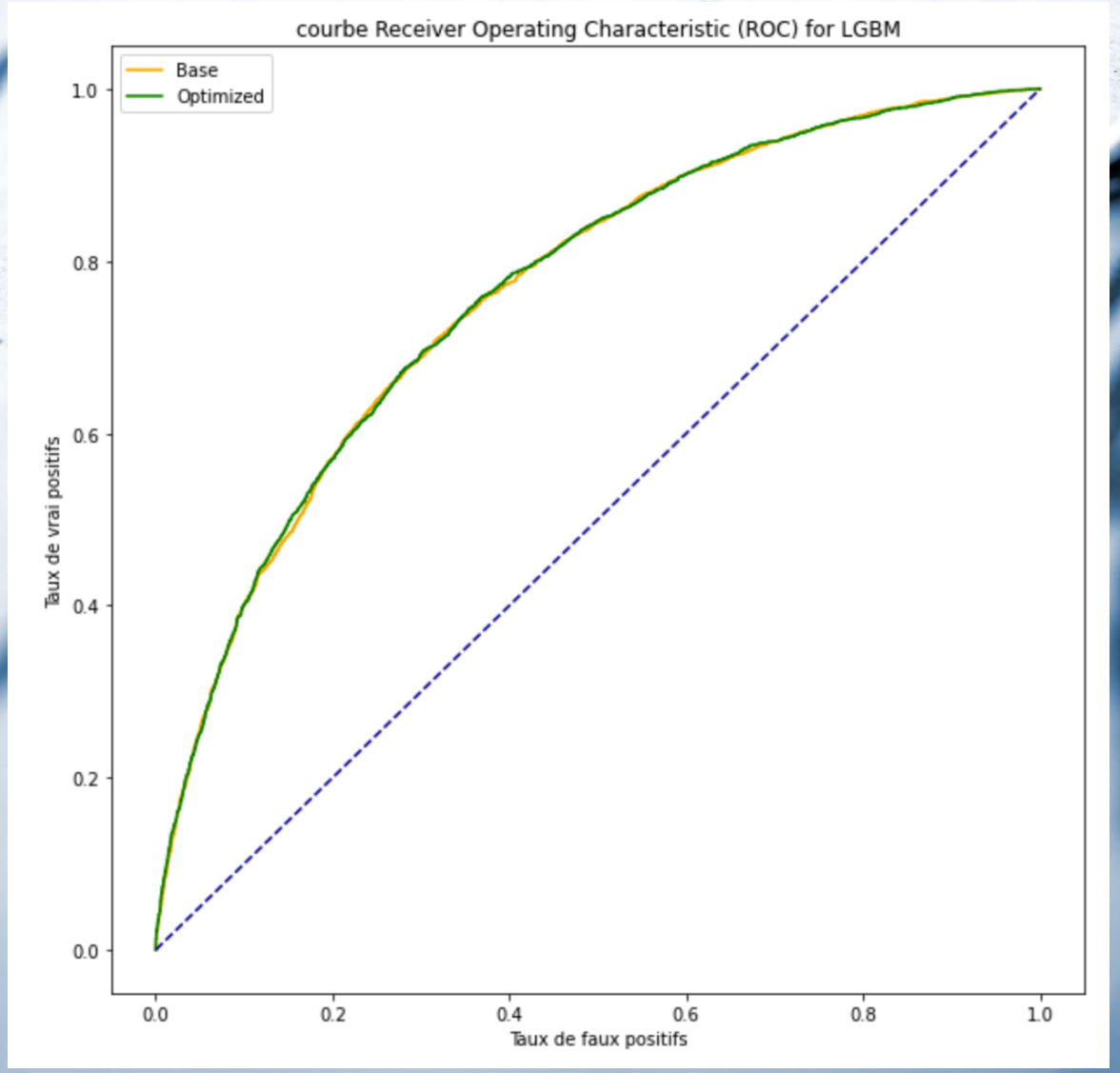
Light Gradient Boosting

Modèle	Données	Nb Variables	Score
Naive Bayse	Under-Sampling	181	0,64
	SMOTE	81	0,76
Régression Logistique	Under-Sampling	116	0,75
	SMOTE	113	0,98
Descente de gradient stochastique	Under-Sampling	5	0,73
	SMOTE	188	0,98
Light Gradient Boosting	**Under-Sampling**	**112**	**0,77**
	SMOTE	81	0,98
Random Forest	Under-Sampling	130	0,75
	SMOTE	79	0,98

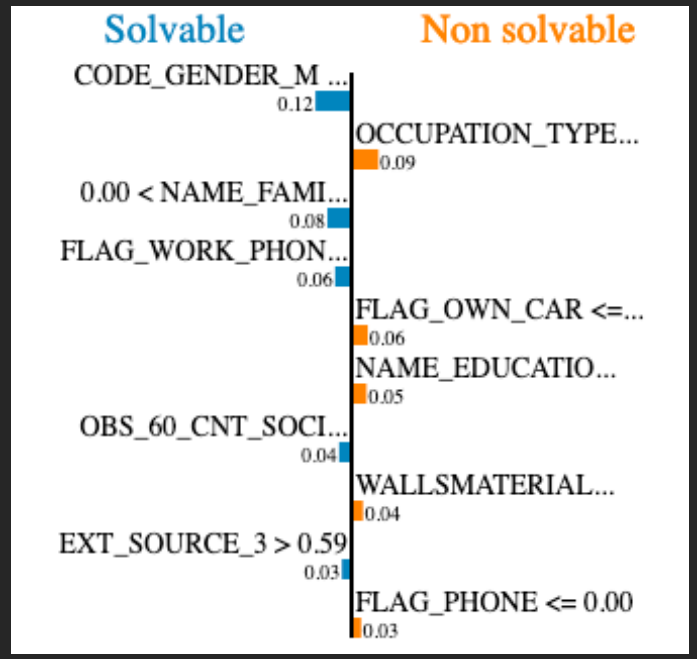
Optimisation du modèle

Light Gradient Boosting

Tests avec différentes variables du modèle



Explication du modèle



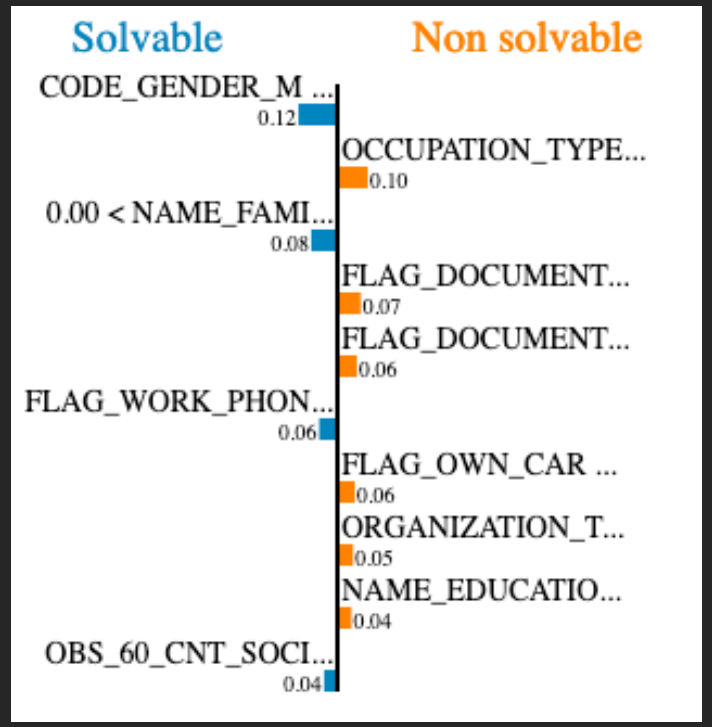
Librairie LIME
Local Interpretable Model-Agnostic Explanations

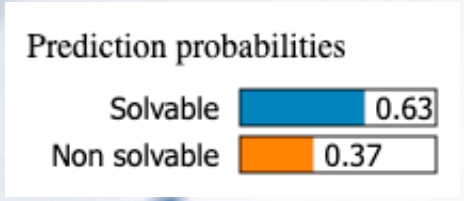
Solvable
C'est une femme
Elle est mariée
Elle n'a pas fourni de n° téléphone professionnel
Il n'y pas eu d'observation dans 60 derniers jours
Les données EXT_SOURCE_3 existent

Non solvable
Elle n'est pas un personnel de base
Elle ne possède pas de voiture
Elle n'a pas d'éducation supérieure
Elle habite une maison en pierre ou en brique
Elle n'a pas fourni de n° de téléphone

Exemple d'explication

Client potentiellement solvable





Solvable à 63%
C'est une femme
Elle est mariée
Elle n'a pas fourni de téléphone n° professionnel
Il n'y pas eu d'observation dans 60 derniers jours

Non solvable à 37%
Elle n'est pas un personnel de base
Elle n'a pas fourni le document 18
Elle n'a pas fourni le document 16
Elle ne possède pas de voiture
Elle n'a pas d'éducation supérieure
Elle ne travaille pas au ministère de la sécurité
Elle n'a pas fourni de n° de téléphone

Conclusion

Modèle avec un score de 77%

Axes d'améliorations

Ajouter les autres jeux de données fournis

Obtenir plus d'observations

Augmenter les performances avec d'autres modèles