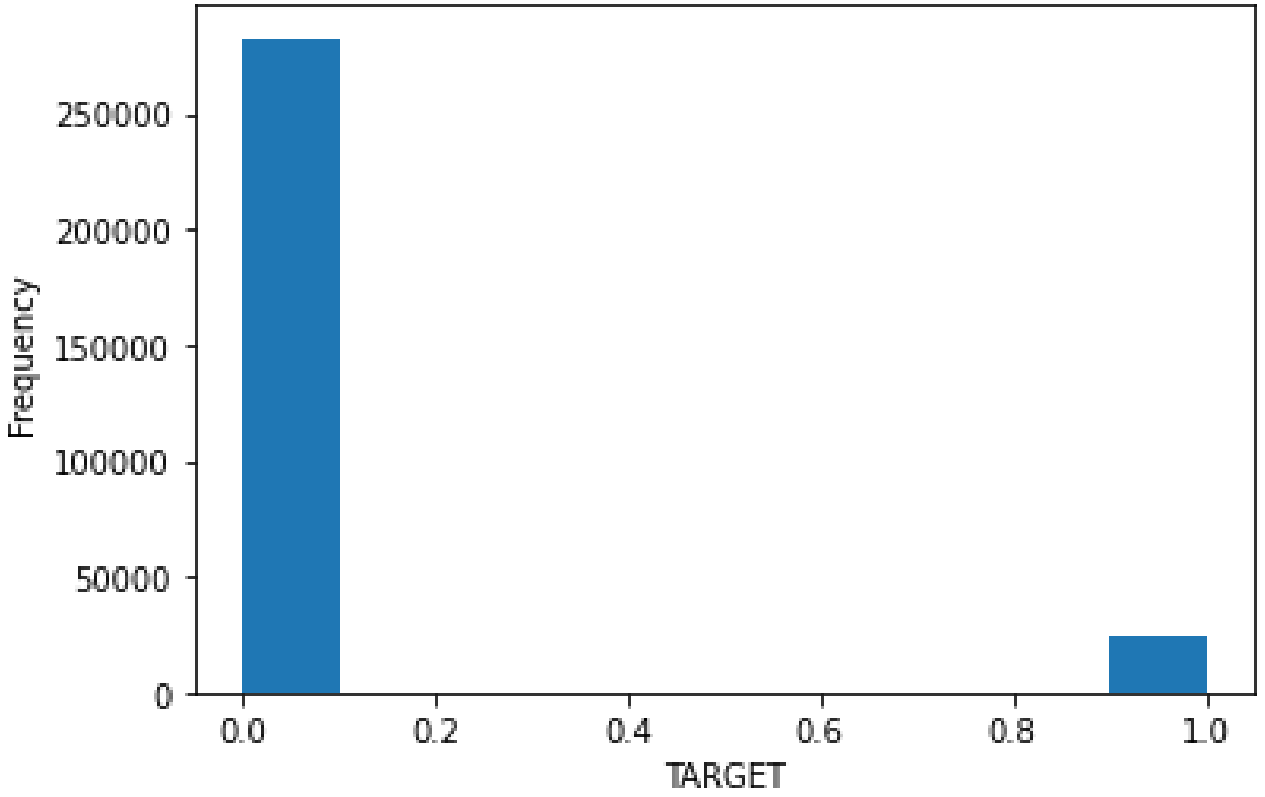
Contexte du projet

Présentation des données

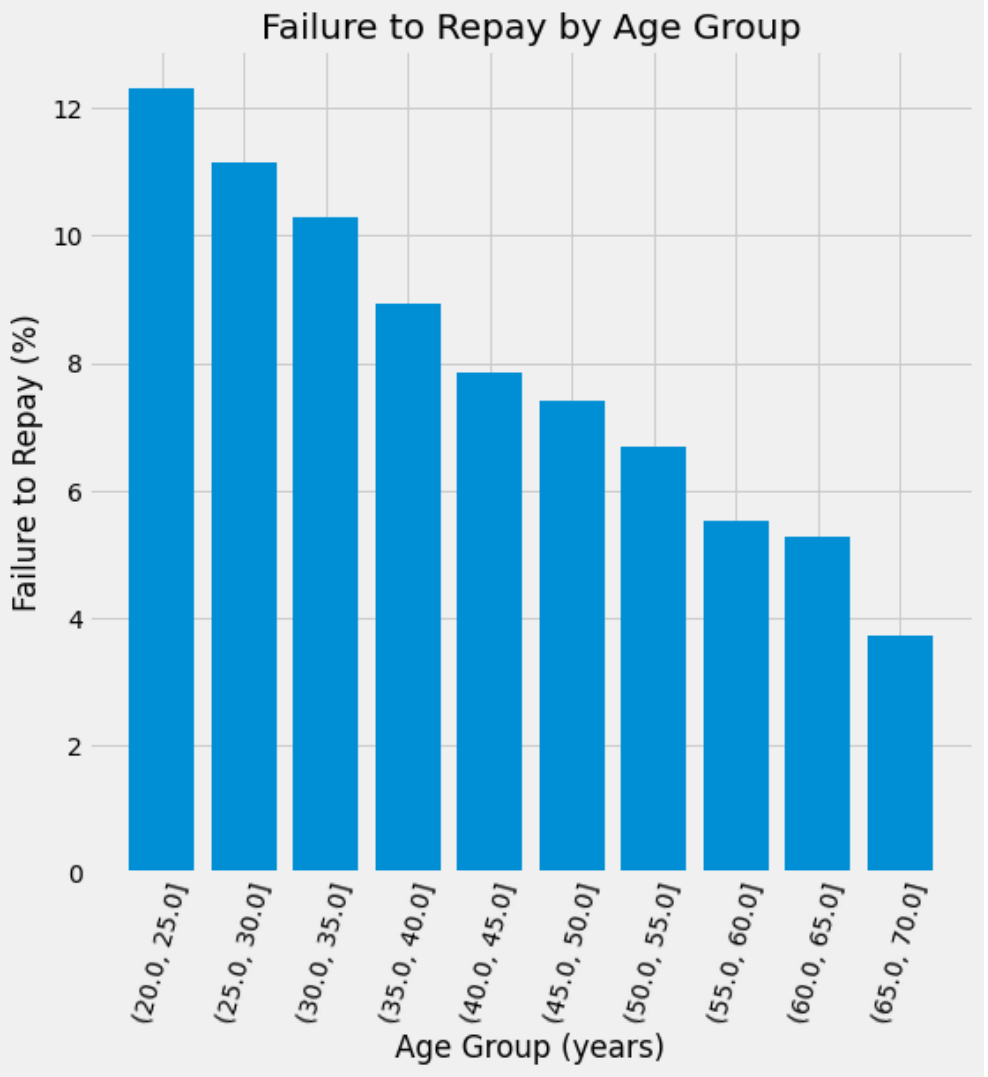
Training data shape: (307511, 122)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
0	100002	1	Cash loans	M	N	Y	0
1	100003	0	Cash loans	F	N	N	0
2	100004	0	Revolving loans	M	Y	Y	0
3	100006	0	Cash loans	F	N	Y	0
4	100007	0	Cash loans	M	N	Y	0

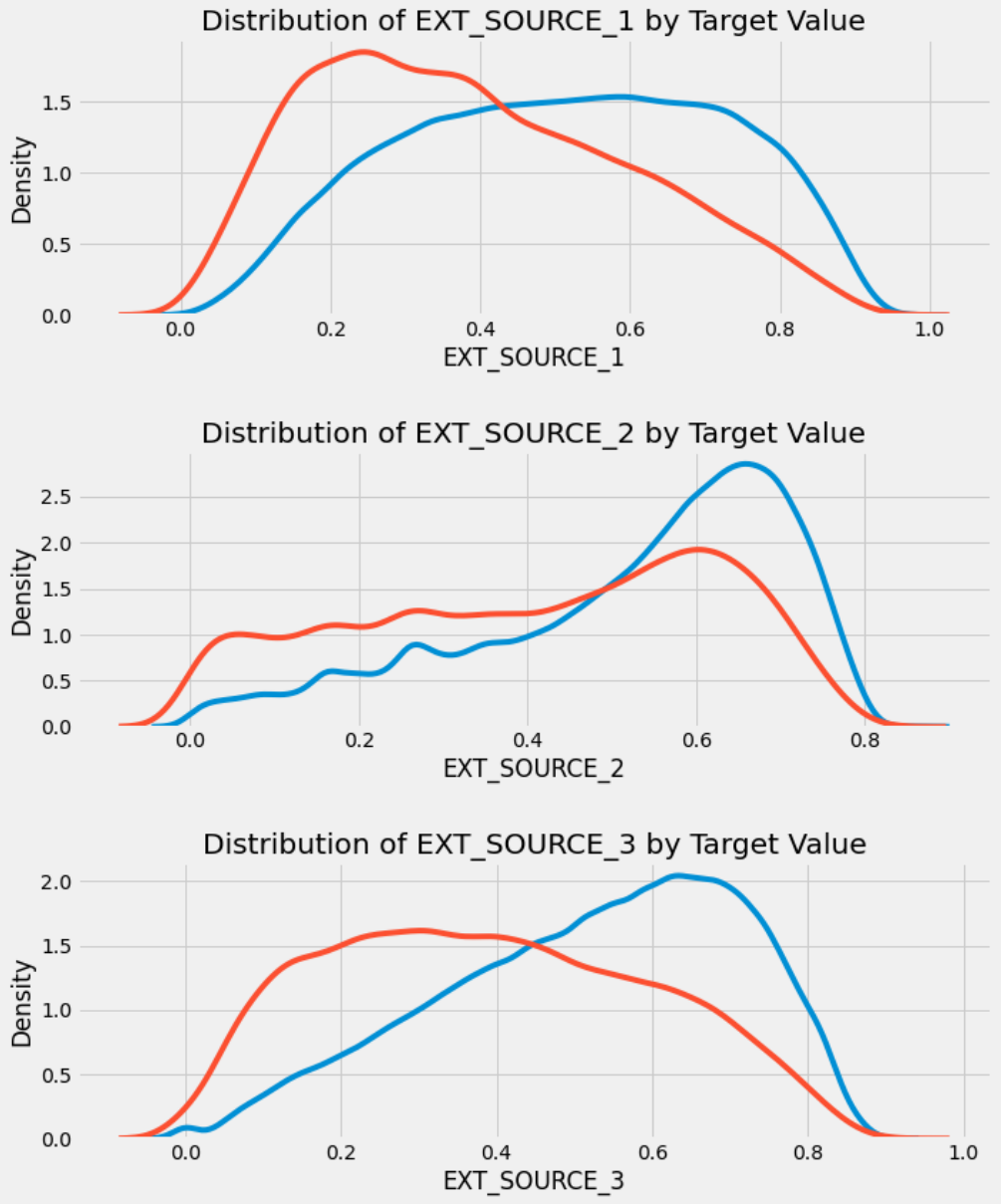
Analyse des données



Analyse des données

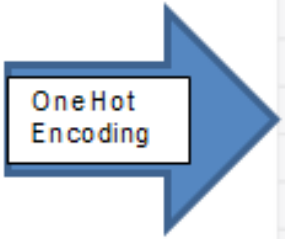


Analyse des données



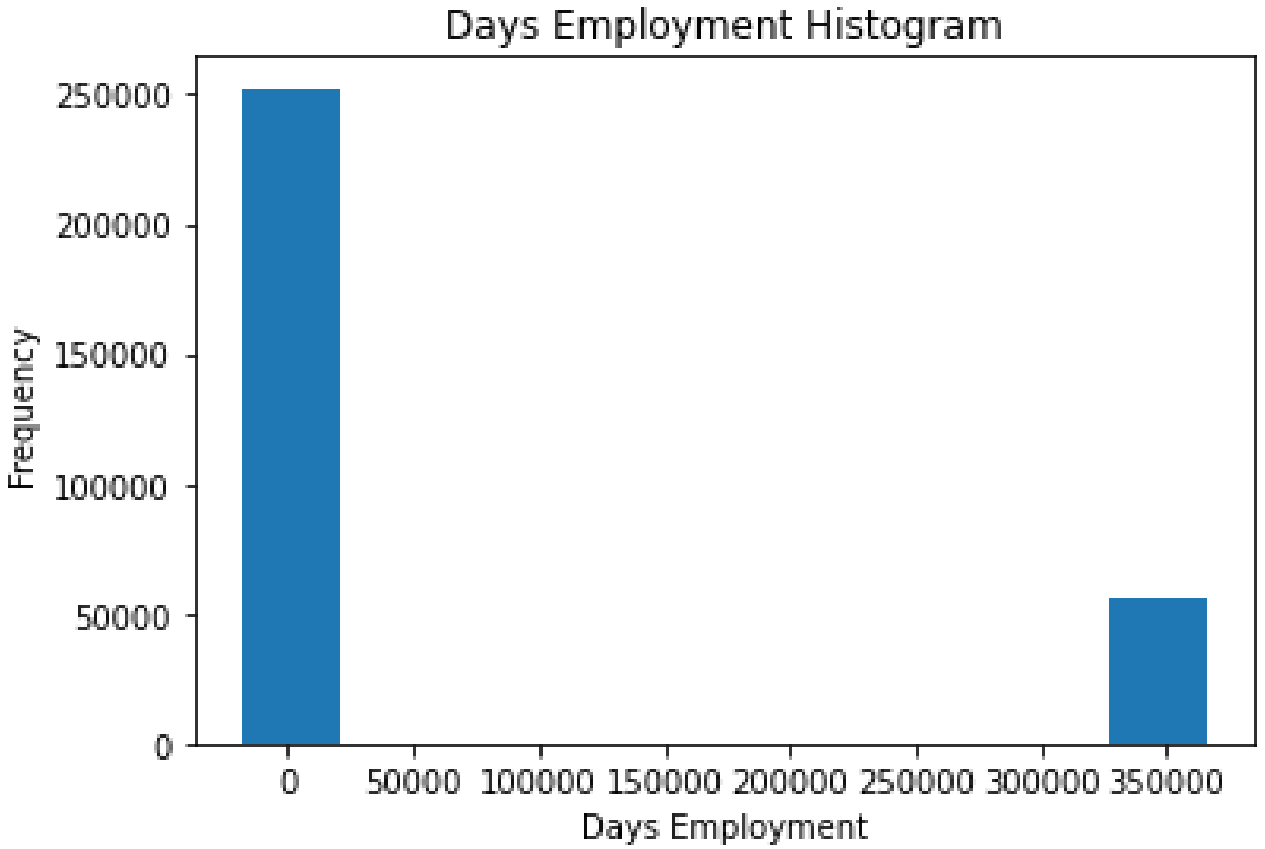
Nettoyage des données

	occupation
0	programmer
1	data scientist
2	engineer
3	manager
4	ceo



	occupation_ceo	occupation_data scientist	occupation_engineer	occupation_manager	occupation_programmer
0	0	0	0	0	1
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	1	0	0	0	0

Nettoyage des données

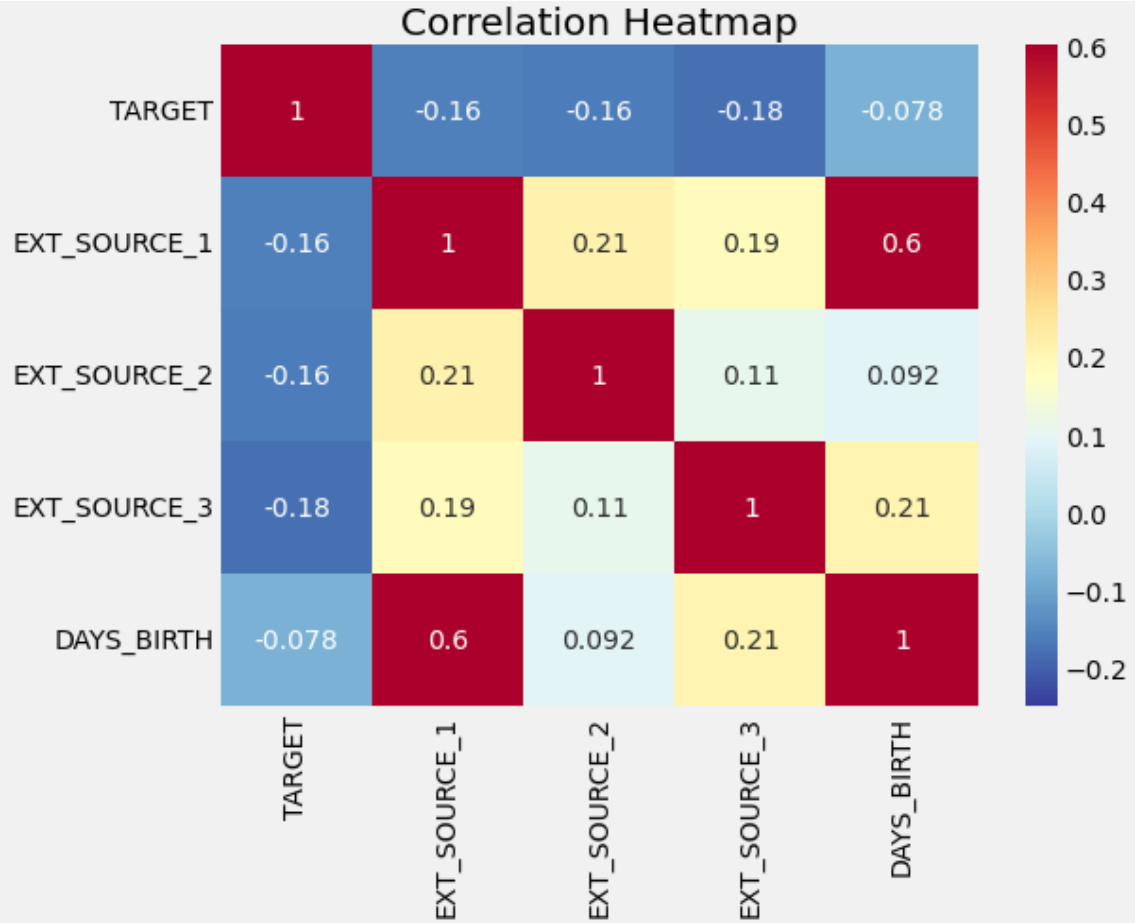


Imputation des données

	Missing Values	% Missing
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4

Feature engineering

Polynomiale

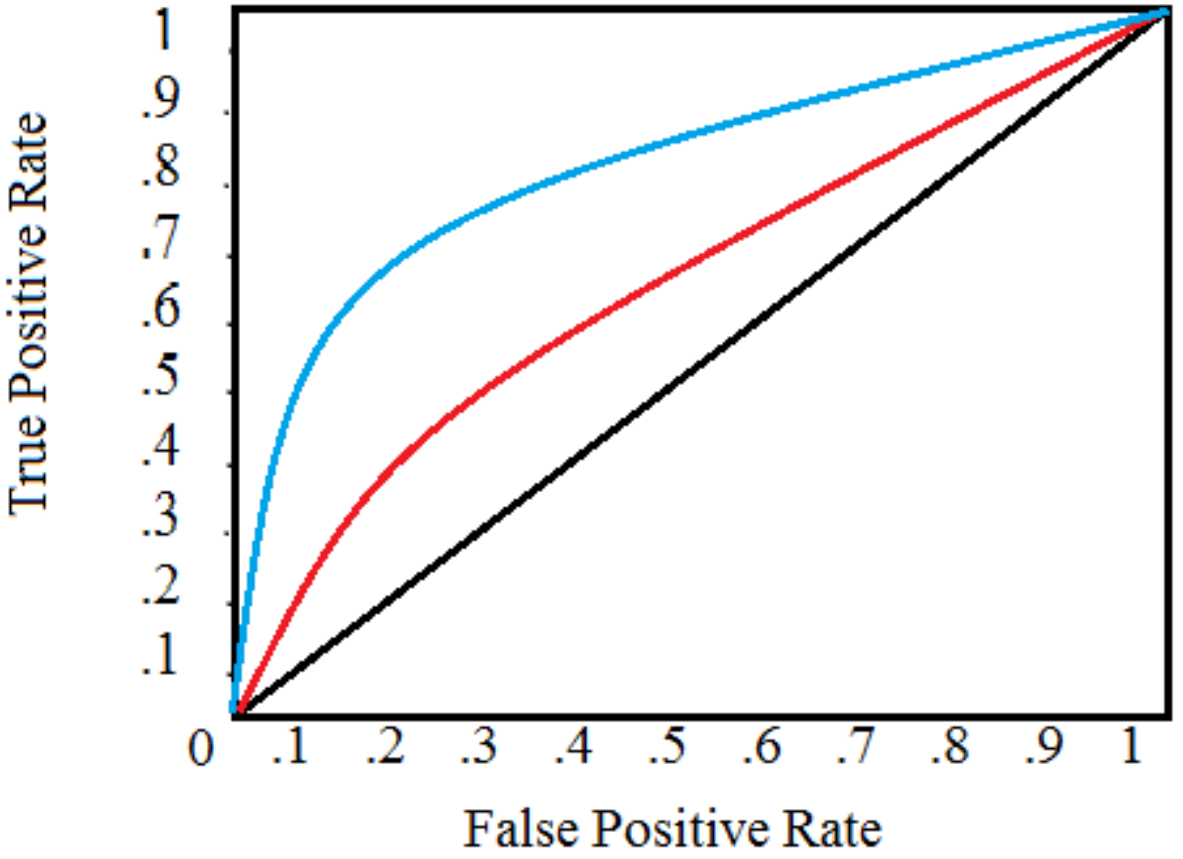


Feature engineering

Métier

- days employed percent
- credit income percent
- annuity income percent
- credit term

Sélection du modèle



Sélection du modèle

Pipeline :

- équilibrage des classes
- normalisation
- grid search / validation croisée
- test des performances

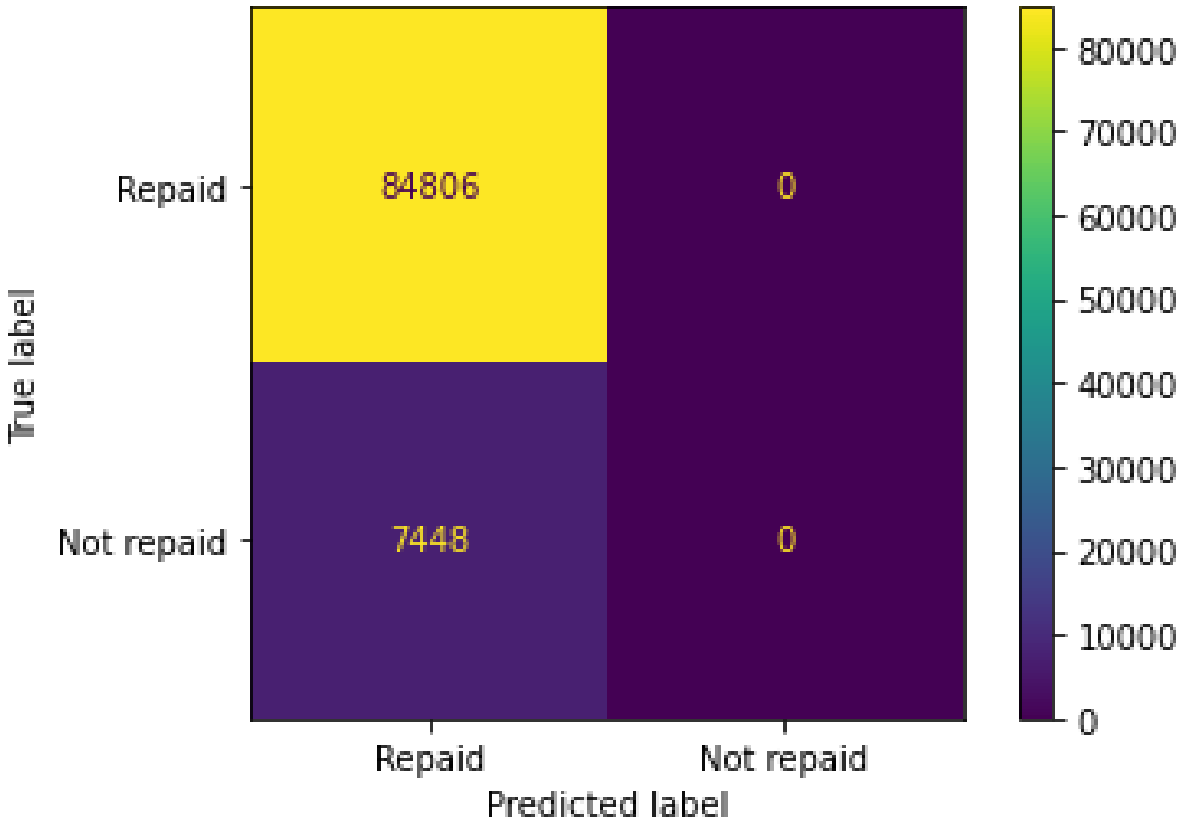
Sélection du modèle

modèles :

- gradient boosting
- forêt aléatoire
- XGBoost
- régression logistique
- SGD

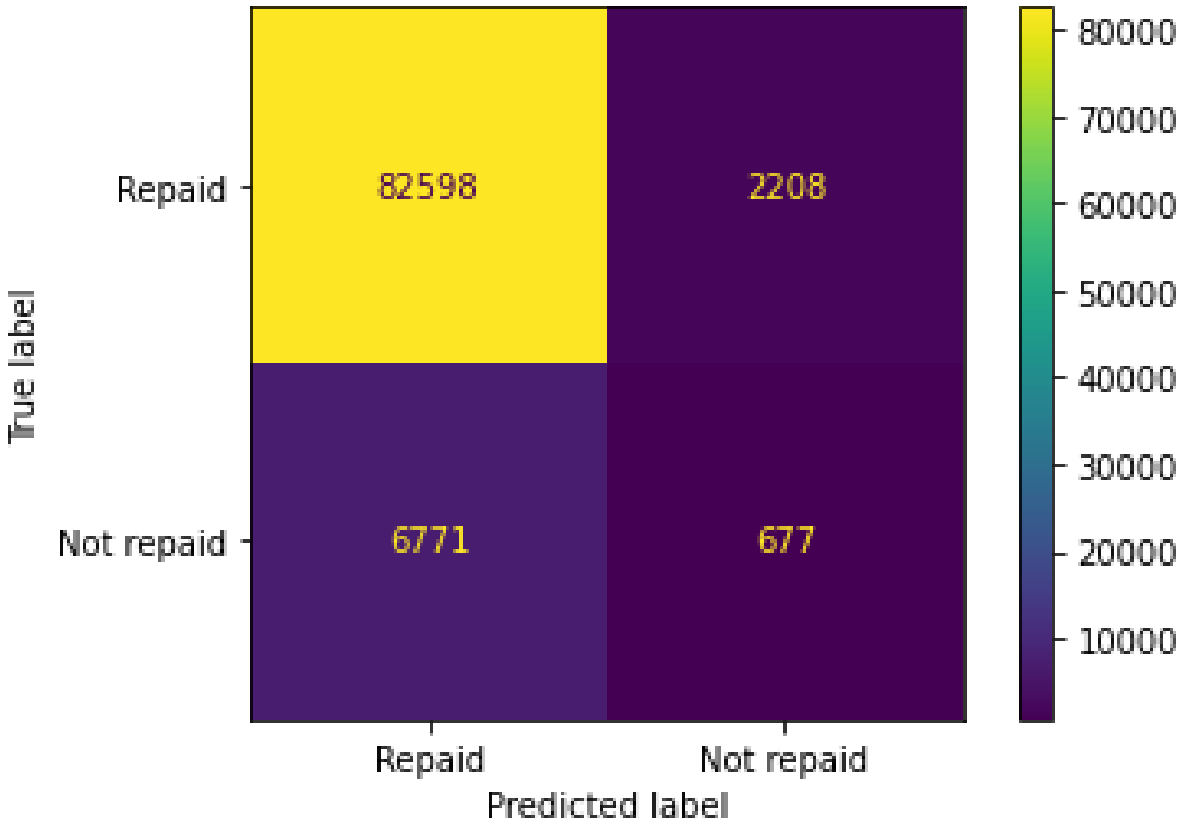
Sélection du modèle

Dummy : 0,500



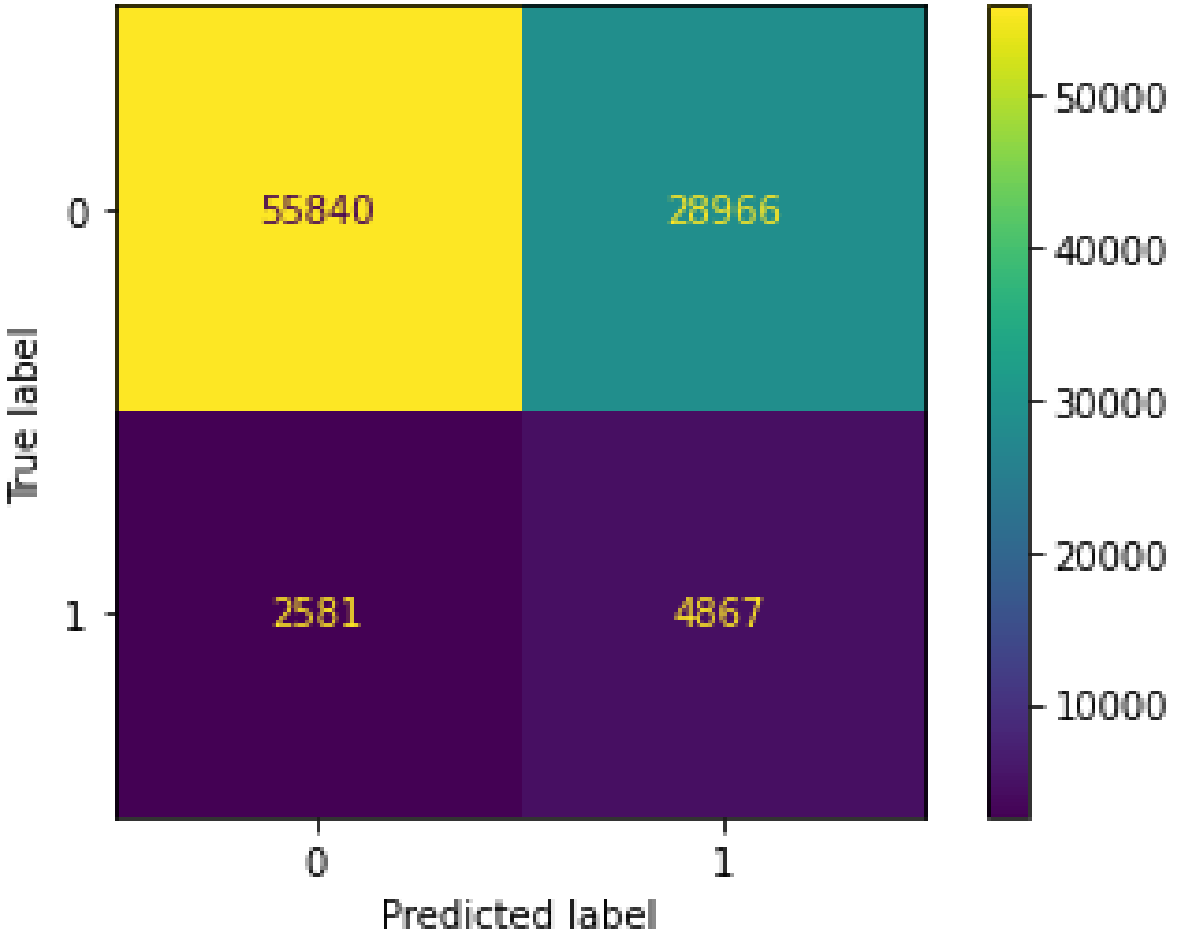
Sélection du modèle

Gradient boosting : 0,703



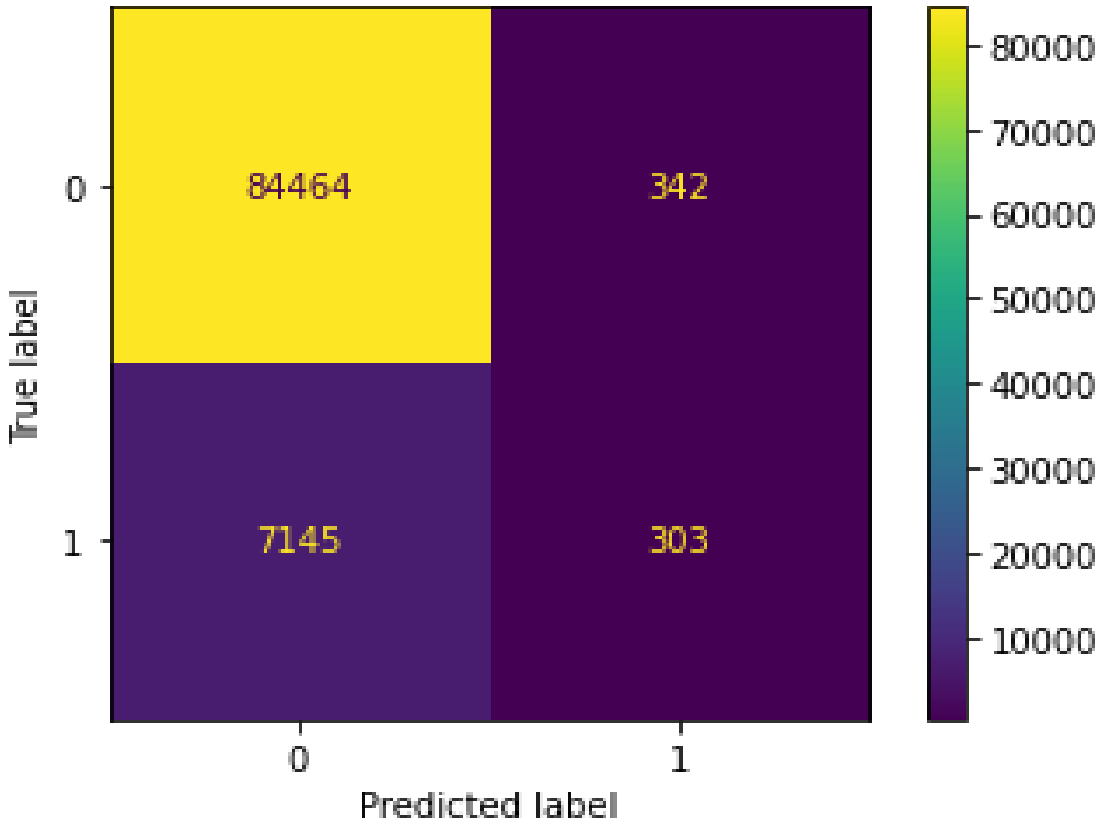
Sélection du modèle

Random forest : 0,714



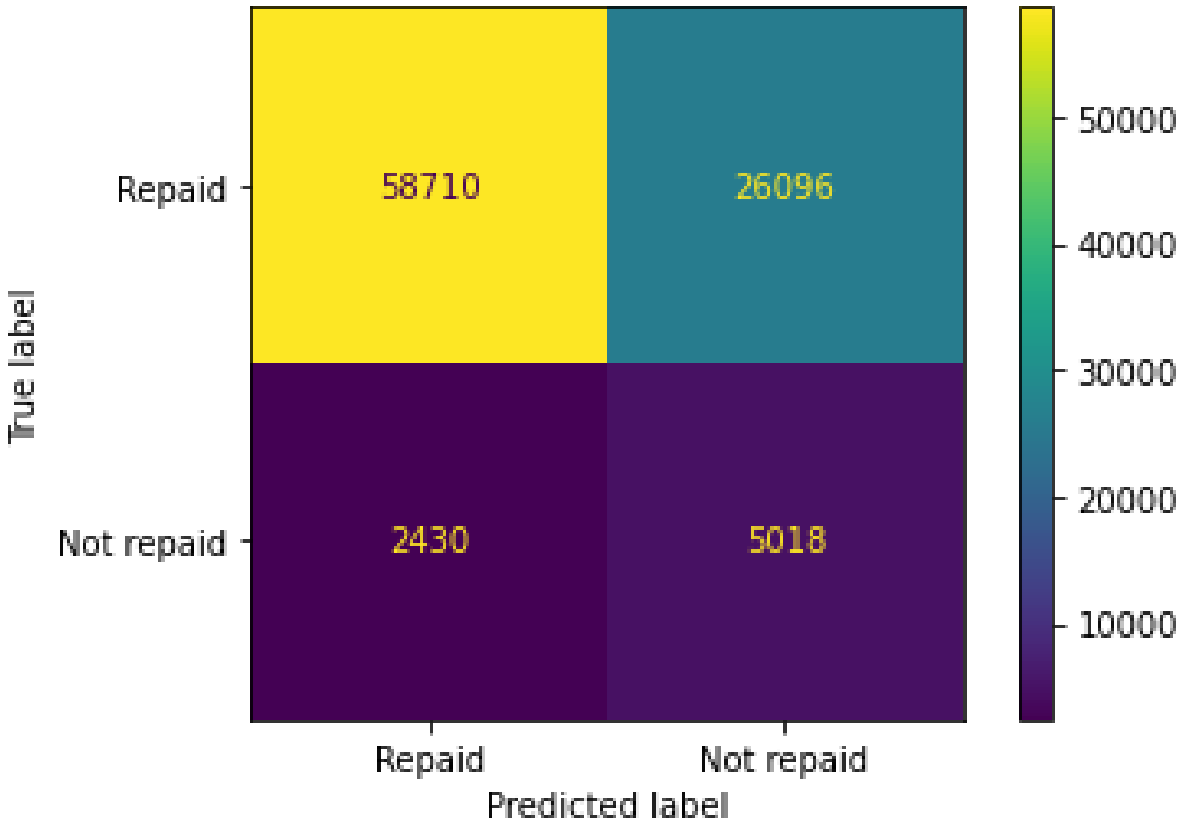
Sélection du modèle

XGBoost : 0,746



Sélection du modèle

Régression logistique : 0,745



Optimisation du modèle

- jeux de données

- grid search plus précise

- optimisation du seuil

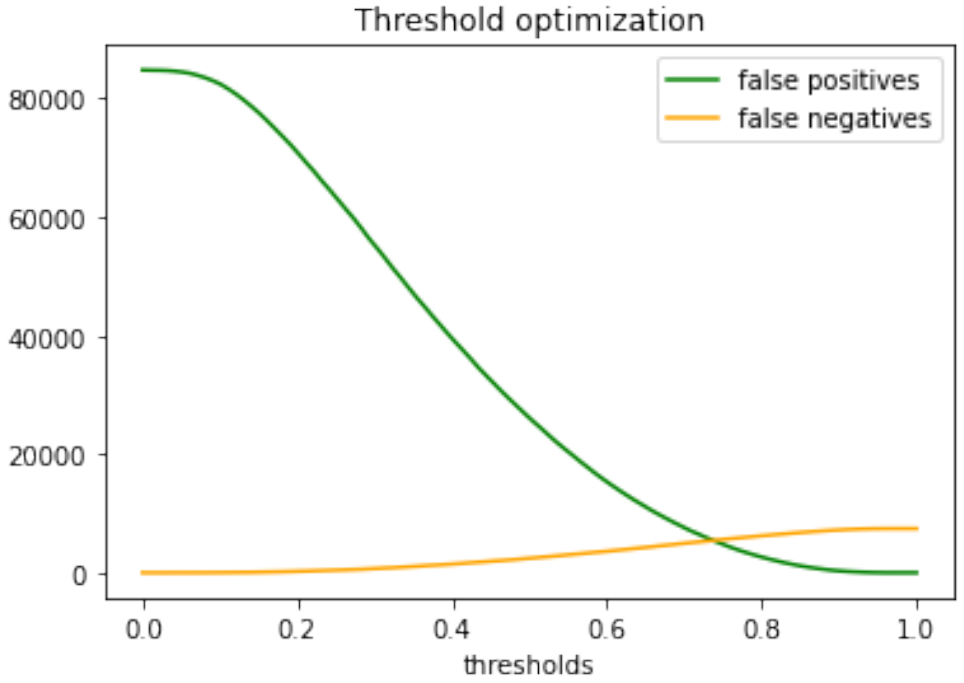
Optimisation du modèle

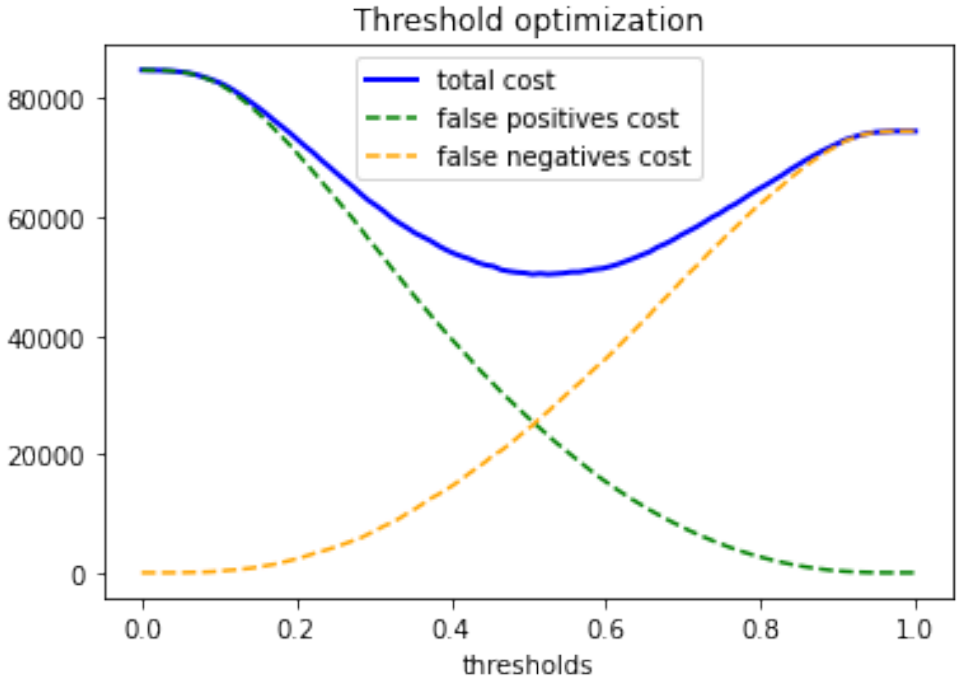
Fonction coût = FP + FN

Faux Positif = 1 Faux Négatif = 10

Optimisation du modèle

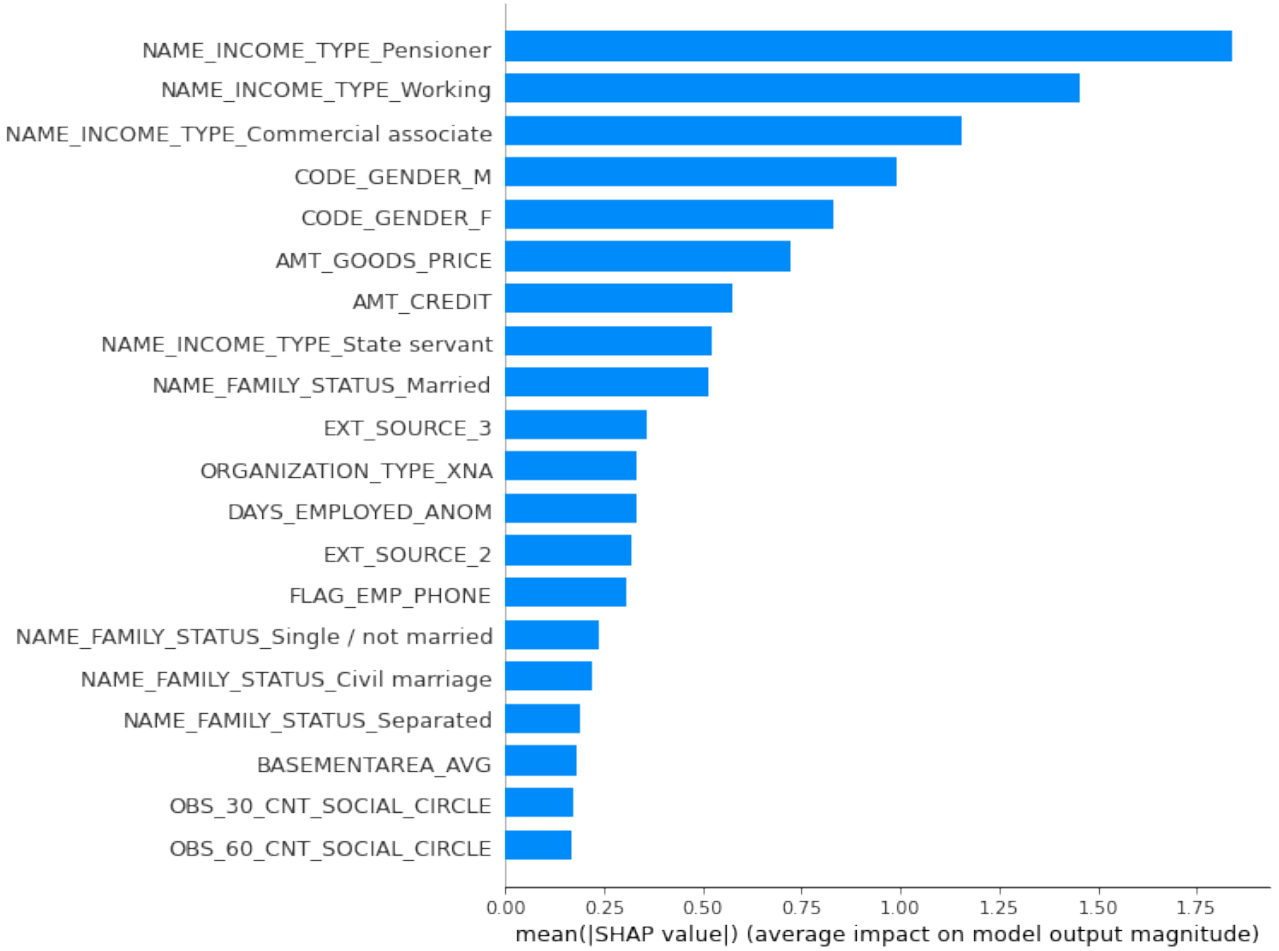
Régression logistique





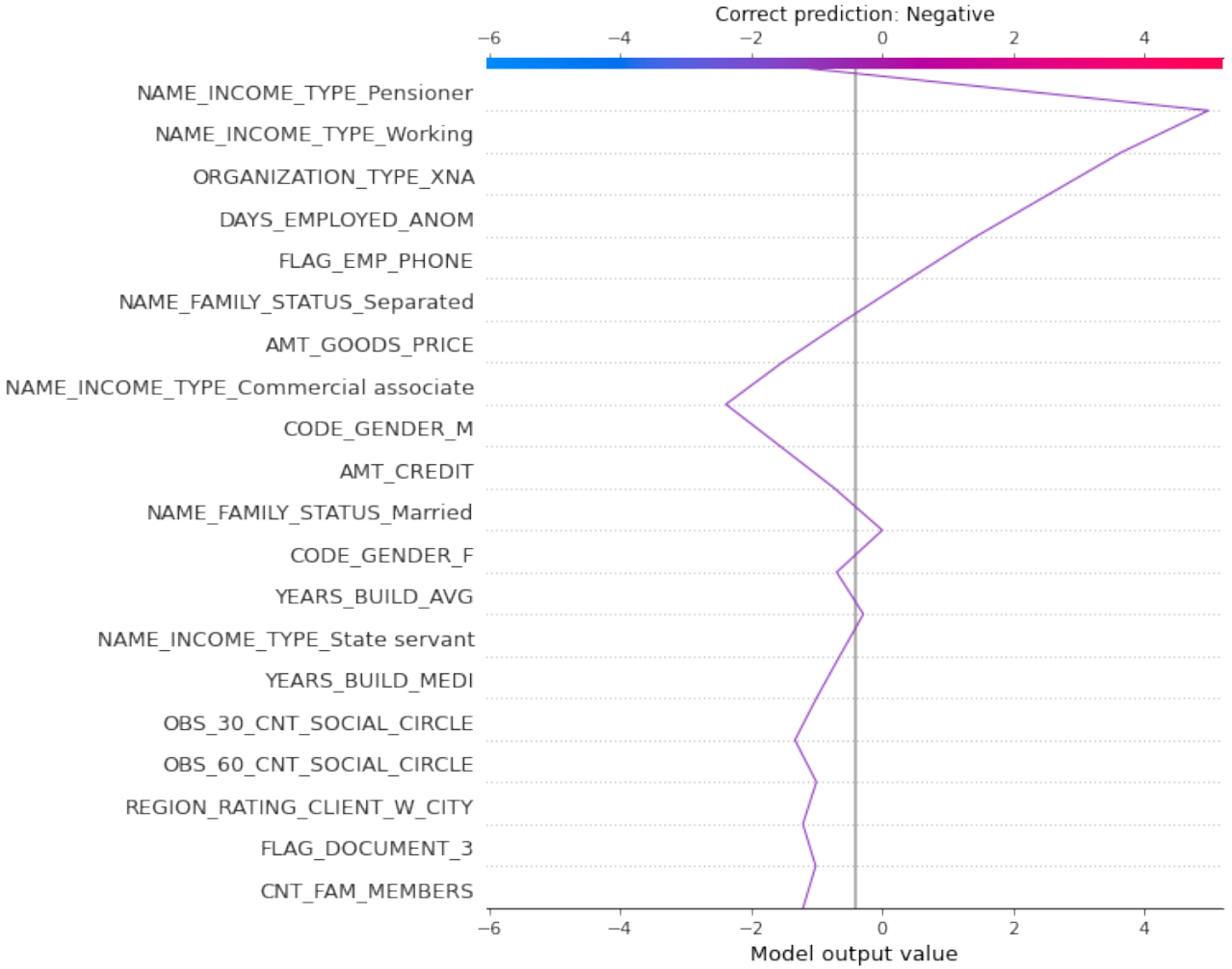
Interprétation du modèle

Globale



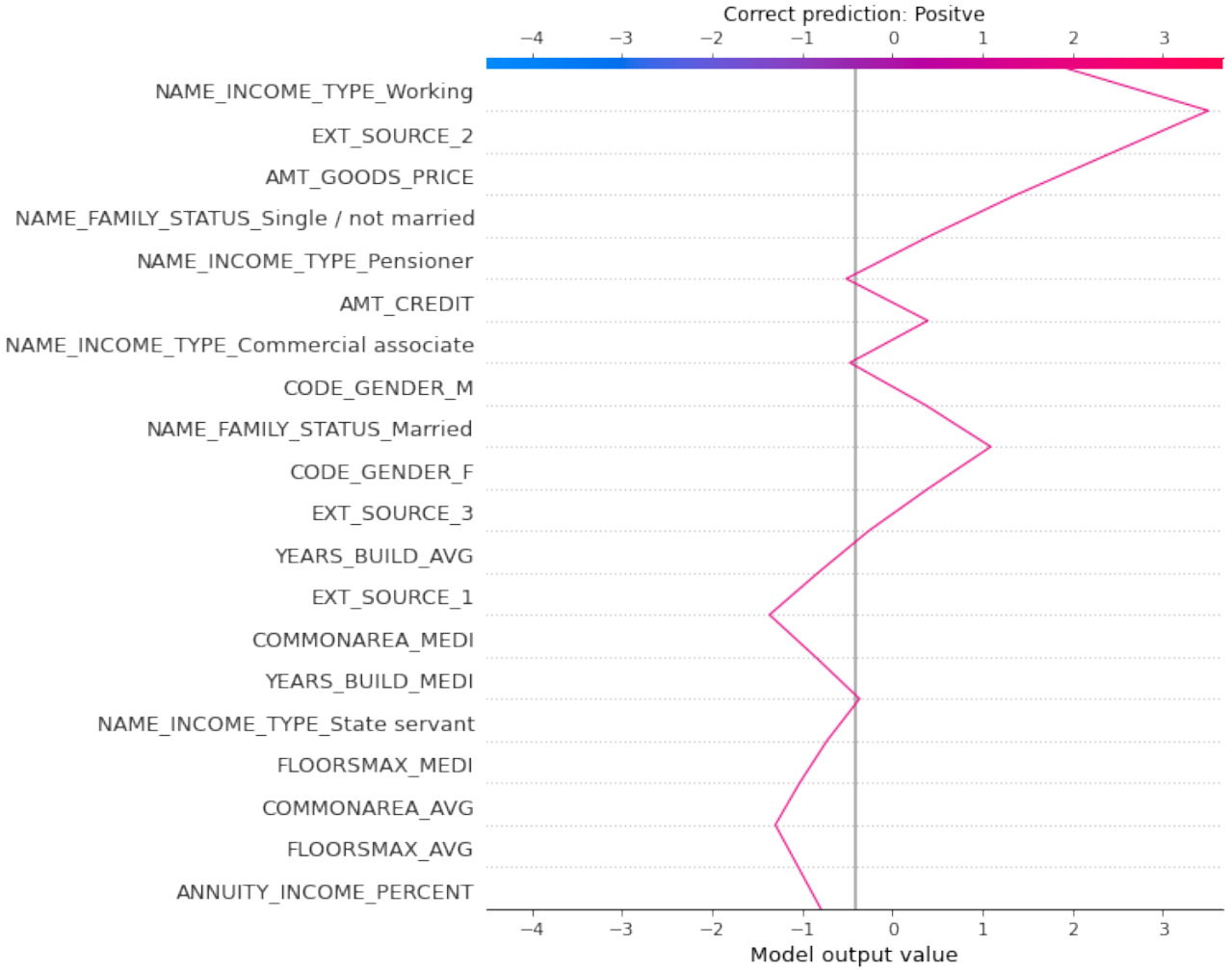
Interprétation du modèle

Locale



Interprétation du modèle

Locale



Axes d'améliorations

- données
- feature engineering
- imputation
- plus de modèles
- plus gros modèles
- fonction coût