




CONSTRUISEZ UN MODÈLE DE SCORING

Xiaofan LEI

TABLE DES MATIÈRES



Contexte
Description et exploration du jeu de données
Feature Engineering et Equilibrage des datasets d'entrainement
Comparaison et synthèse des résultats pour les datasets utilisés
Comparaison et synthèse des résultats des algorithmes
Interprétation du modèle Catboost
Prédiction et conclusion

CONTEXTE

Contexte

Ce projet s'inscrit dans le cadre de la mise en œuvre d'un outil de "scoring crédit", permettant de calculer la probabilité qu'un client le rembourse ou non

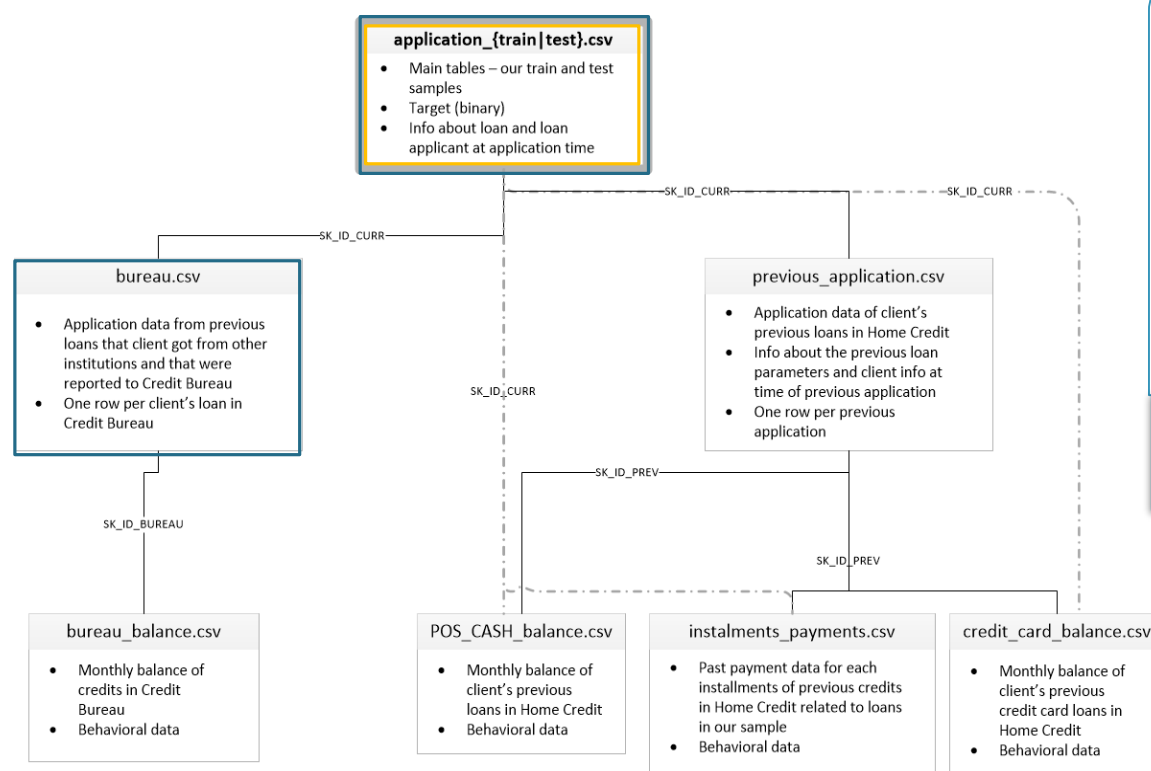
Objectif du projet

Entraîner un algorithme de classification pour effectuer des prédictions binaires
Sur lequel les chargés de clientèle peuvent s'appuyer pour classifier la demande : crédit accordé ou refusé.

Enjeux

L'algorithme doit accorder une attention particulière à la classe minoritaire, à laquelle les clients ne remboursent pas appartiennent.

DESCRIPTION DU JEU DE DONNÉES



- Principales données d'entraînement
- Une ligne = un prêt
- Target = 1 : non remboursement
- Taille d'origine: (307511, 122)
- Taille après transformation : (307511, 199)

application_train.csv



- Données complémentaires
- Credits précédents dans d'autres établissements
- Plusieurs lignes = 1 prêt
- Taille d'origine : (1716428, 17)
- Taille après transformation : (305811, 71)

Bureau.csv:

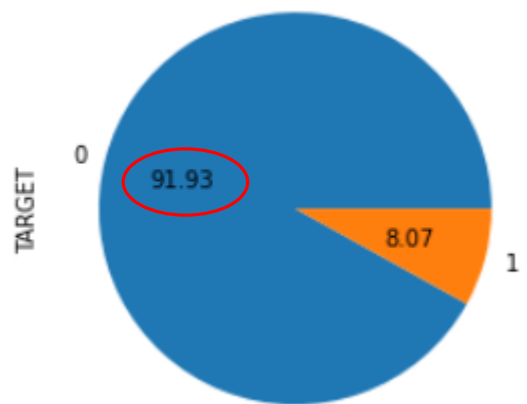


- Principales données de test (sans Target)
- Une ligne = un prêt
- Taille d'origine: (48744, 121)
- Taille après transformation & ajout des variables bureau : (48744, 235)

application_test.csv



EXPLORATION DES DONNÉES



Distribution de classes déséquilibrée:

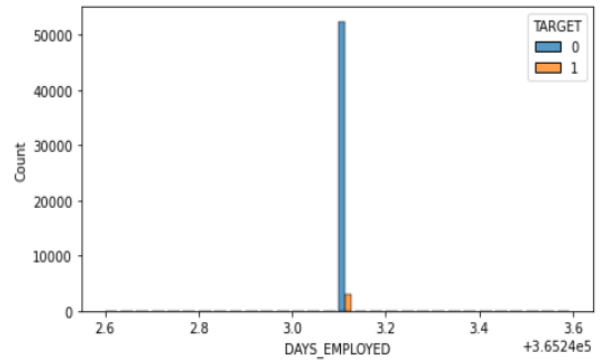
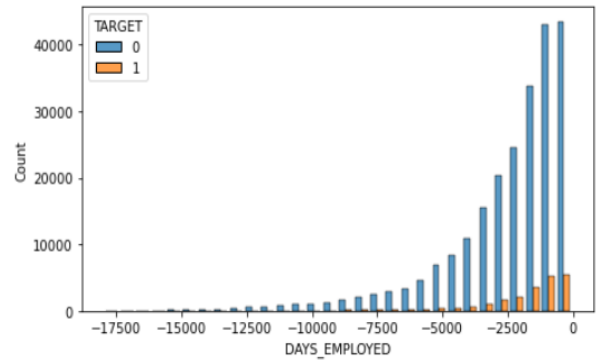
- 0 : prêts remboursés 92%
- 1 : prêts non remboursés 8%

	nb na	% na
OWN_CAR_AGE	202929	65.99%
YEARS_BUILD_AVG	204488	66.50%
COMMONAREA_AVG	214865	69.87%
LIVINGAPARTMENTS_AVG	210199	68.35%
YEARS_BUILD_MODE	204488	66.50%
COMMONAREA_MODE	214865	69.87%
LIVINGAPARTMENTS_MODE	210199	68.35%
YEARS_BUILD_MEDI	204488	66.50%
COMMONAREA_MEDI	214865	69.87%
LIVINGAPARTMENTS_MEDI	210199	68.35%
NONLIVINGAPARTMENTS_MEDI	213514	69.43%

Vérification des valeurs manquantes :

- Une dizaine de colonnes ont le taux de valeurs manquantes de plus de 60%.
- Les valeurs manquantes seront remplacées par 0.

	count	mean	std	min	25%	50%	75%	max
DAYS_EMPLOYED	307511.0	63815.045904	141275.766519	-17912.0	-2760.0	-1213.0	-289.0	365243.0



The non-anomalies default on 8.66% of loans
The anomalies default on 5.40% of loans
There are 55374 anomalous days of employment

DAYS_EMPLOYED:

- La valeur maximum très éloignée du reste
- La seule valeur positive est 365243
- Ce chiffre semble avoir une influence sur le taux de remboursement
- L'anomalie sera mise dans une nouvelle colonne

FEATURE ENGINEERING - ENCODAGE

Traitement des variables catégoriques

16 colonnes contiennent des catégories en texte

Encodage d'étiquettes(Label encoding) pour les catégories binaires (3 colonnes concernées)

Encodage à chaud(OneHotEncoder) pour les autres catégories

Taille de la base après transformation : (307511, 244)

Encodage d'étiquettes (Label Encoding)

color	color
red	0
green	1
blue	2
red	0

Encodage à chaud(OneHot Encoding)

color	color_red	color_blue	color_green
red	1	0	0
green	0	0	1
blue	0	1	0
red	1	0	0

FEATURE ENGINEERING - MATRICE DE CORRÉLATION

Vérification de la colinéarité des variables

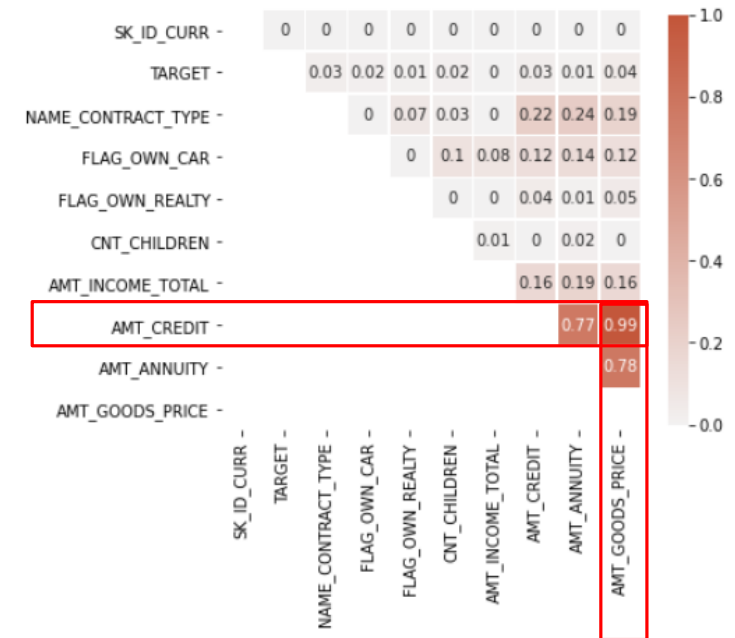
Méthode : Création d'une matrice de corrélation diagonale

Calcul mathématique : Formule de la corrélation de Pearson

Seuil d'exclusion : 0,8

Taille de la base d'entraînement après la transformation :
(307511, 199)

Corrélogramme pour les 10 première variables



A supprimer

FEATURE ENGINEERING — AJOUT DE NOUVELLES VARIABLES

Rappel

Bureau.csv : credits précédents dans d'autres établissements

Récupération de la table bureau.csv

Taille de la table bureau : (1716428, 17)

Suppression de la colonne ID de la table bureau

Encodage d'étiquettes et à chaud

Taille de la table bureau après la transformation : (305811, 71)

Regroupement des données par ID de la base Application_Train par max et count

Ajout des variables bureau à la base Application_Train

Taille de la base d'entraînement après transformation : (307511, 236)

Exclusion des variables colinéaires

ÉQUILIBRAGE DES DATASETS D'ENTRAINEMENT

Résoudre le problème du déséquilibre des classes

Sous-échantillonnage : éliminer aléatoirement de quelques exemples de la classe majoritaire pour diminuer leur effet sur le modèle.

Suréchantillonnage (SMOTE) : compléter la base des données originale par des observations synthétiques des classes minoritaires

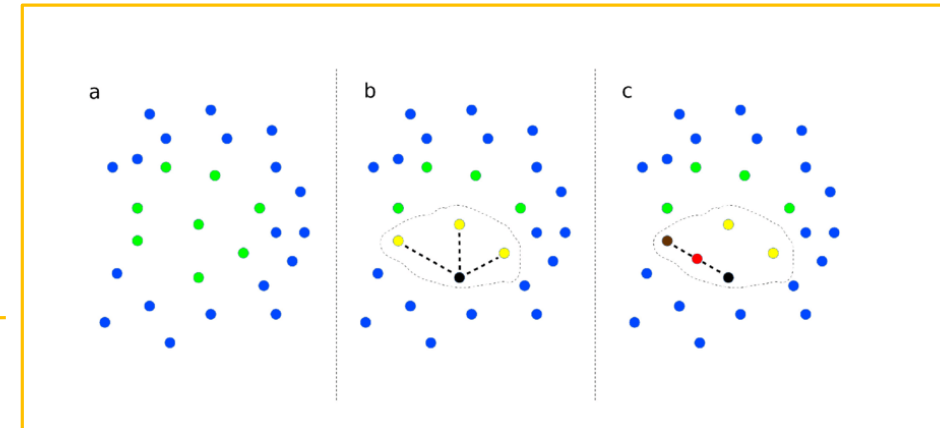
Undersampling



Oversampling



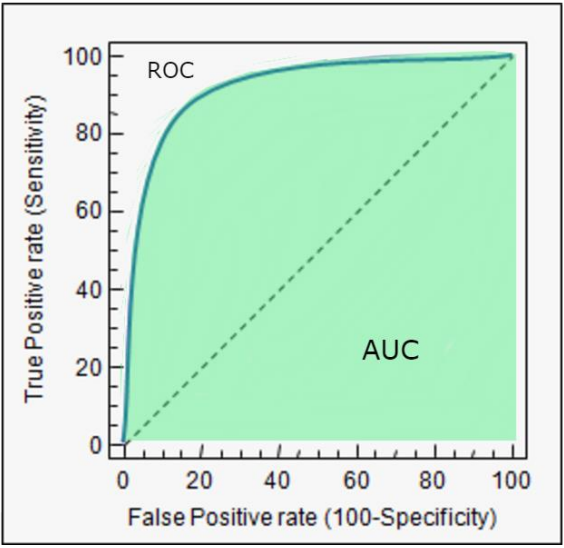
SMOTE



COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES DATASETS UTILISÉS

		With added features			Default features		
		Default sample	SMOTE	under-Resampling	Default sample	SMOTE	under-Resampling
Trainig time		28s	70s	15s	22s	49s	12s
Traing result	(tn, fp, fn, tp)	(131309, 66571, 5307, 12070)	(197255, 625, 17200, 177)	(131295, 66585, 5329, 12048)	(130482, 67398, 5354, 12023)	(196155, 1725, 16753, 624)	(129948, 67932, 5316, 12061)
	Precision	0.153	0.221	0.153	0.151	0.266	0.151
	Recall	0.695	0.010	0.693	0.692	0.036	0.694
	F1 Score	0.251	0.019	0.251	0.248	0.063	0.248
	Accuracy	0.666	0.917	0.666	0.662	0.914	0.660
	AUC	0.740	0.695	0.738	0.737	0.701	0.736
Testing result	(tn, fp, fn, tp)	(56112, 28694, 2380, 5068)	(84562, 244, 7389, 59)	(56047, 28759, 2374, 5074)	(56013, 28793, 2350, 5098)	(84047, 759, 7208, 240)	(55935, 28871, 2364, 5084)
	Precision	0.150	0.195	0.150	0.150	0.240	0.150
	Recall	0.680	0.008	0.681	0.684	0.032	0.683
	F1 Score	0.246	0.015	0.246	0.247	0.057	0.246
	Accuracy	0.663	0.917	0.663	0.662	0.914	0.661
	AUC	0.729	0.696	0.731	0.731	0.694	0.728

AUC : Area under the ROC Curve



Confusion Matrix		Predicted Class		
		No	Yes	
Observed Class	No	TN	FP	
	Yes	FN	TP	

TN True Negative

FP False Positive

FN False Negative

TP True Positive

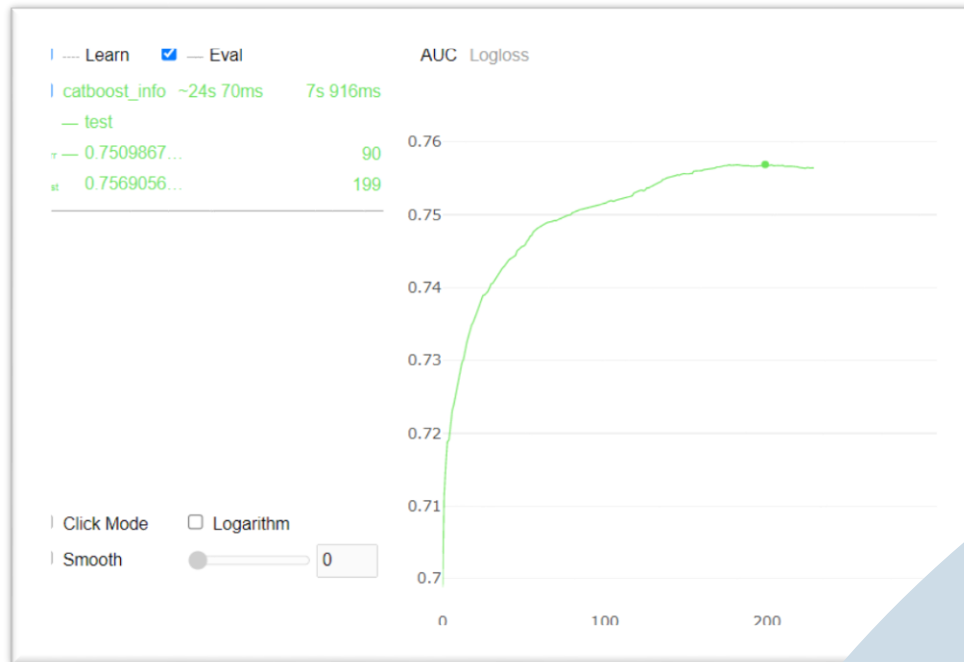
Model Performance

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$



CatBoost



**Catégorique
Boosting**

Entraînement

- s'appuie sur des **arbres de décision (Decision tree) boostés par gradient (Gradient boosting)**

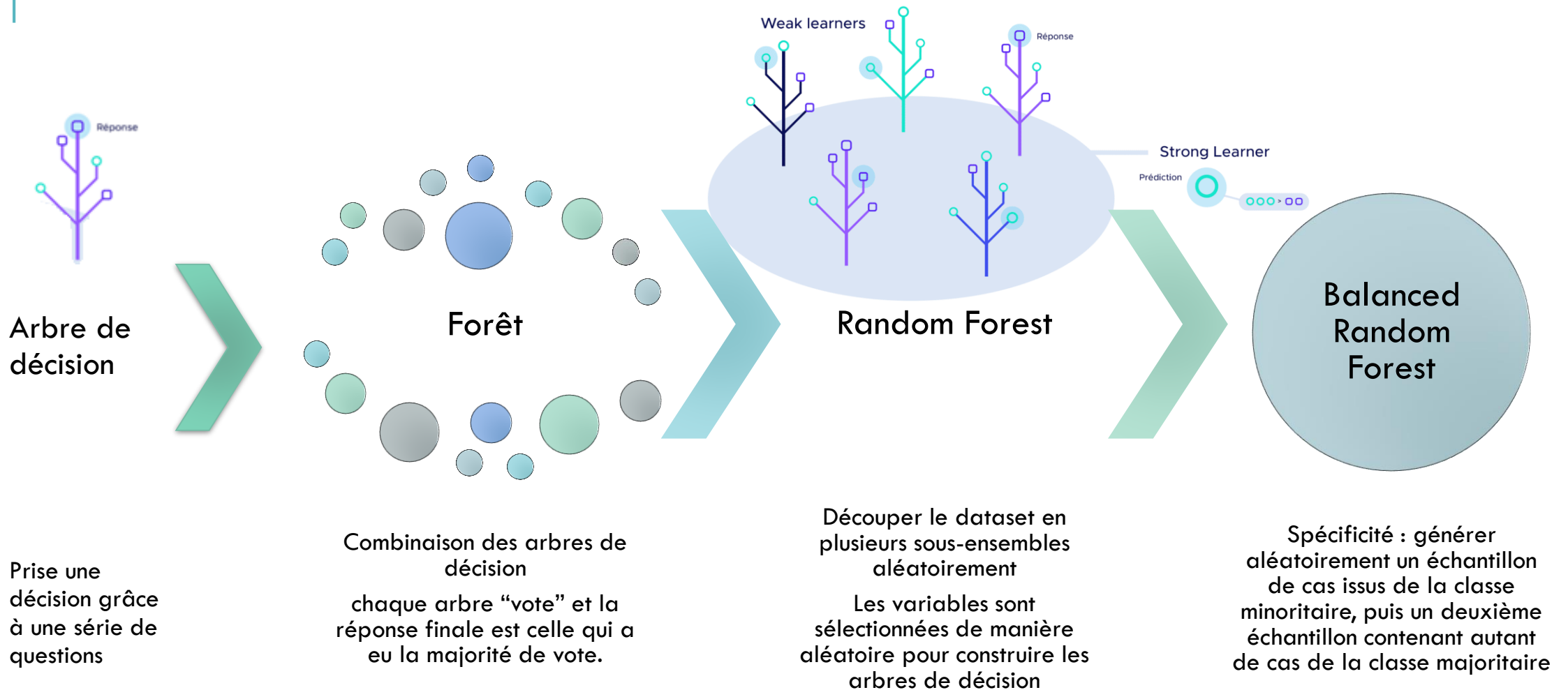
Spécificité technique

- Pour les données catégoriques, une permutation est effectuée de manière aléatoire, puis un calcul est effectué à partir d'une formule unique à CatBoost

Avantages

- Plus rapide
- Efficace et fonctionne particulièrement bien avec des variables catégorielles
- Précision déterminée sur un ensemble de données de validation.
- superbes visualisations

BALANCED RANDOM FOREST

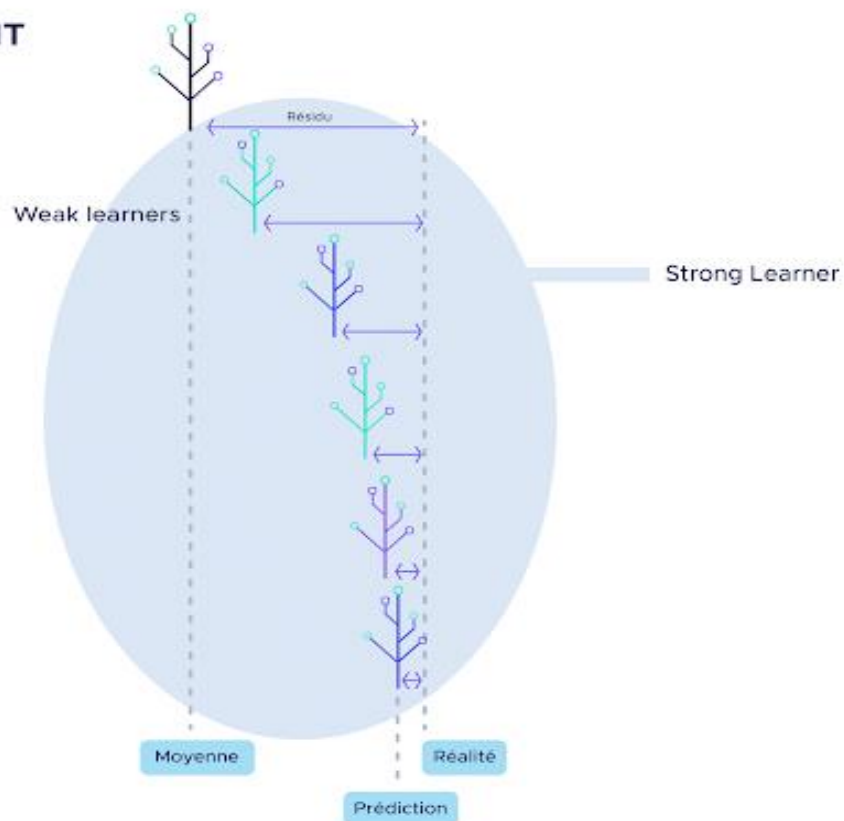


Spécificité : générer aléatoirement un échantillon de cas issus de la classe minoritaire, puis un deuxième échantillon contenant autant de cas de la classe majoritaire

XGBOOST

GRADIENT BOOST

Calcule l'écart entre la prédiction et la réalité

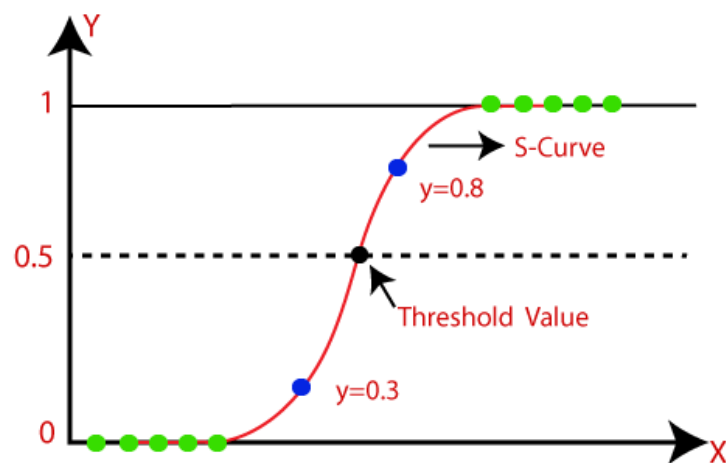


XGBoost

Version particulière de Gradient boost

- Les branches ou même arbres qui ne sont pas assez performants sont "élagués"
- Permet une souplesse de manœuvre grâce à un panel d'hyperparamètres très important

RÉGRESSION LOGISTIQUE



Définition :

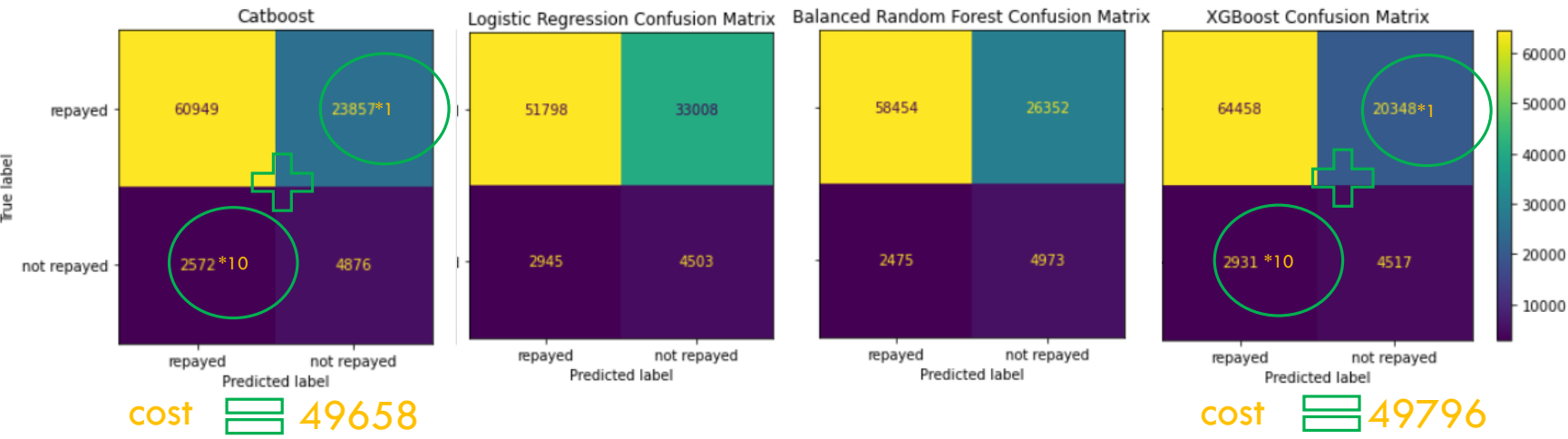
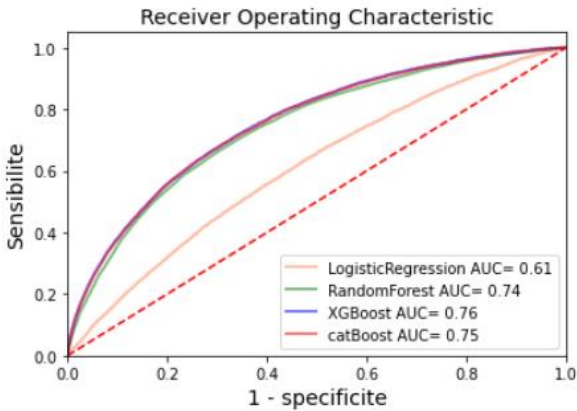
- Modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Entrainement :

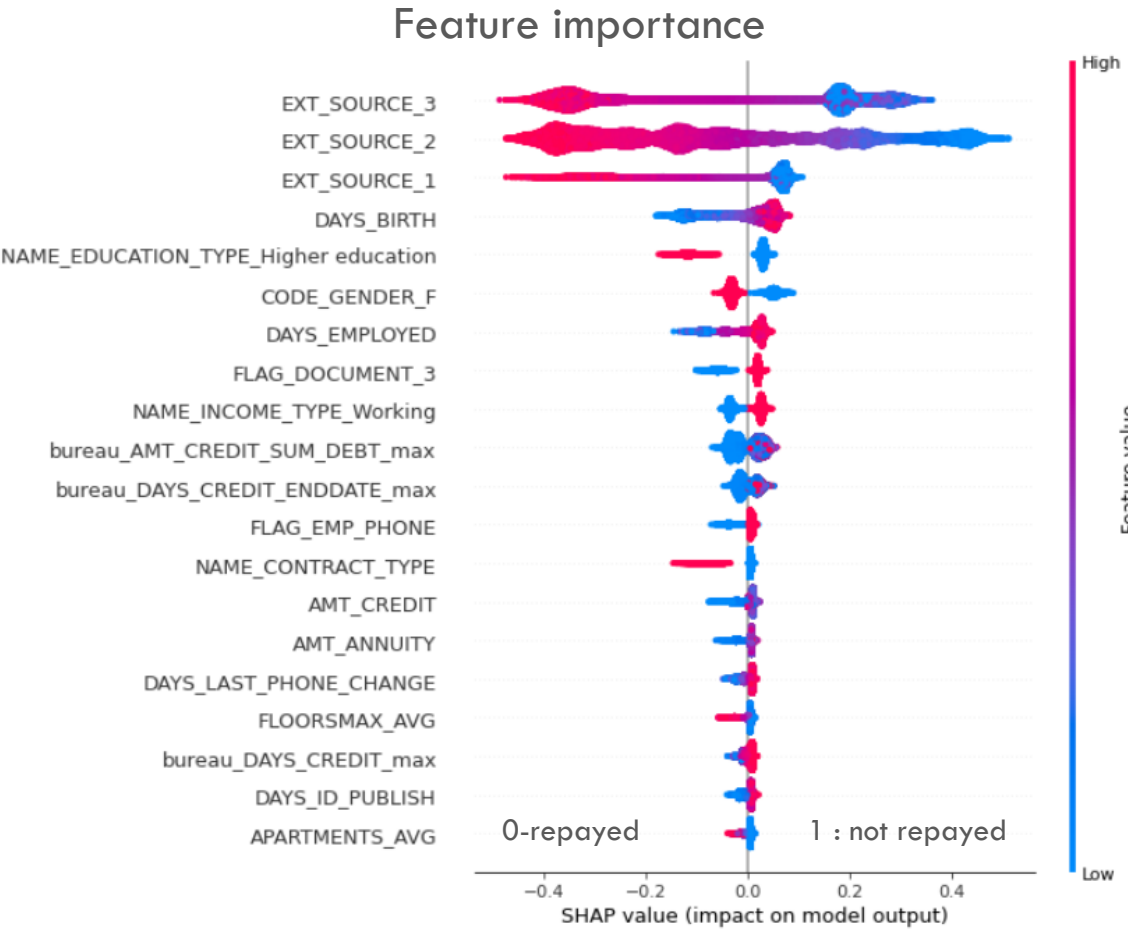
- permettre au courbe sigmoïde de coller au mieux aux données

COMPARAISON ET SYNTHÈSE DES RÉSULTATS POUR LES MODÈLES UTILISÉS

	CatBoostClassifier	LogisticRegression	BalancedRandomForestClassifier	XGBClassifier
training time	120s	78s	115s	115s
(tn, fp, fn, tp)	(60949, 23857, 2572, 4876)	(52669, 32137, 3472, 3976)	(58690, 26116, 2467, 4981)	(64458, 20348, 2931, 4517)
Precision	0.170	0.110	0.160	0.182
Recall	0.655	0.534	0.669	0.606
F2 Score	0.417	0.302	0.409	0.413
Accuracy	0.714	0.614	0.690	0.748



INTERPRÉTATION DU MODÈLE CATBOOST

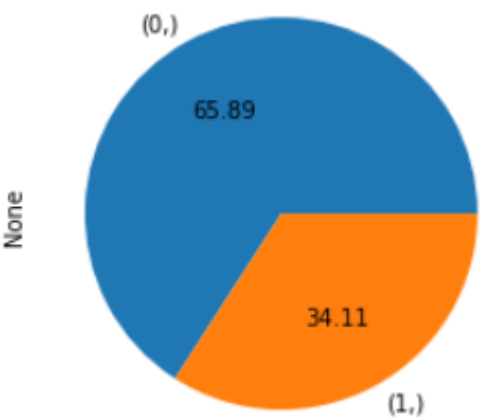


	Feature	VarImp
28	EXT_SOURCE_3	30.188797
27	EXT_SOURCE_2	23.611769
26	EXT_SOURCE_1	11.345989
9	DAYS_BIRTH	4.902138
87	NAME_EDUCATION_TYPE_Higher education	4.478136
207	bureau_AMT_CREDIT_SUM_DEBT_max	4.185965
70	CODE_GENDER_F	3.121443
45	FLAG_DOCUMENT_3	2.367703
206	bureau_AMT_CREDIT_SUM_max	1.664330
68	AMT_REQ_CREDIT_BUREAU_QRT	1.613959

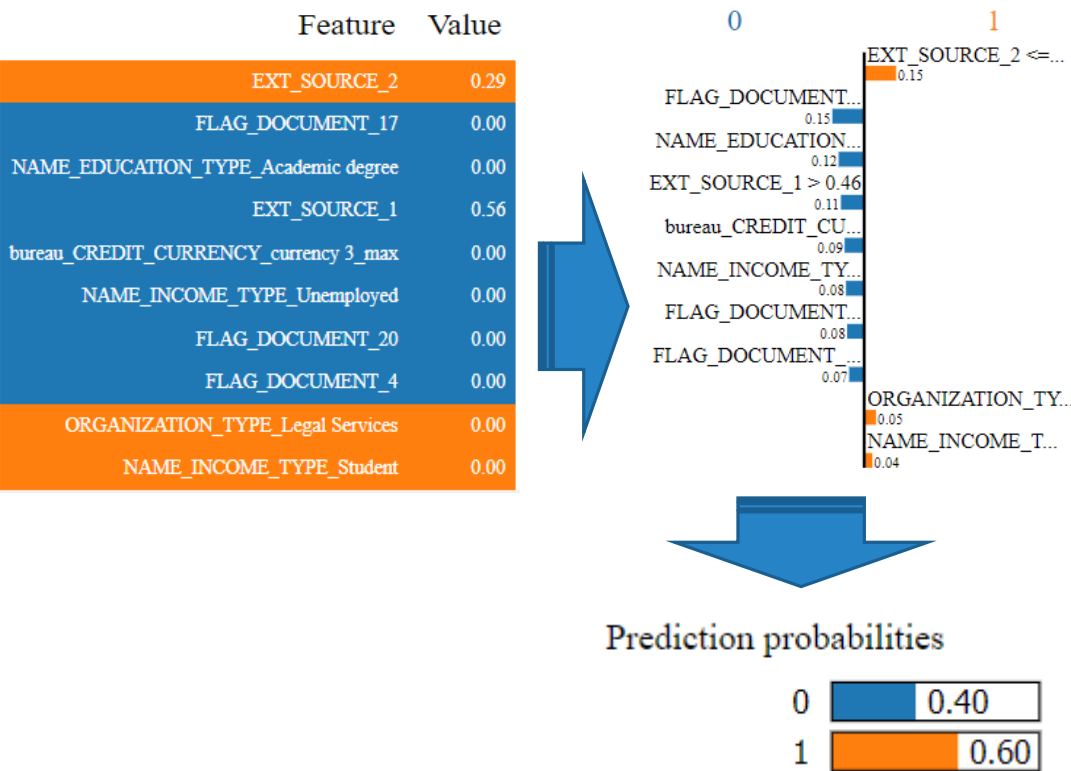
PRÉDICTION AVEC CATBOOST

Base de test : application_test.csv

Prédiction globale:



Analyse du premier prêt dans le fichier test :



CONCLUSION

L'algorithme Catboost a obtenu des résultats intéressants et du potentiel, ses performances pourraient largement s'améliorer avec suppression de l'étape encodage, ajout d'autres sources(ex, previous_application.csv) ou variables(ex, FeaturesPolynomial) , ou encore l'augmentation de la puissance des hyperparamètres comme le nombre de itinéraires.

Il est à noter qu'à part l'aspect purement technique, il est primordial d'impliquer les experts de domaine dans l'analyse de l'importance des variables afin de garantir la pertinence des choix. Leur regard en amont dans l'étape de Feature engineering pourrait aussi contribuer à la réussite du modèle.