CONSTRUISEZ UN MODELE DE SCORING

Présentation “Projet 4” chez “OPENCLASSROOM”

Jaoid KRAIRI

(Septembre 2021)

SOMMAIRE

Rappel du contexte

Aide à la décision de crédit accordé ou non??

COMPREHENSION DE LA PROBLEMATIQUE METIER : 2/Problématique

Organisme de crédit « Prêt à dépenser »

Crédits à la consommation: Personnes ayant peu d'historique de prêt

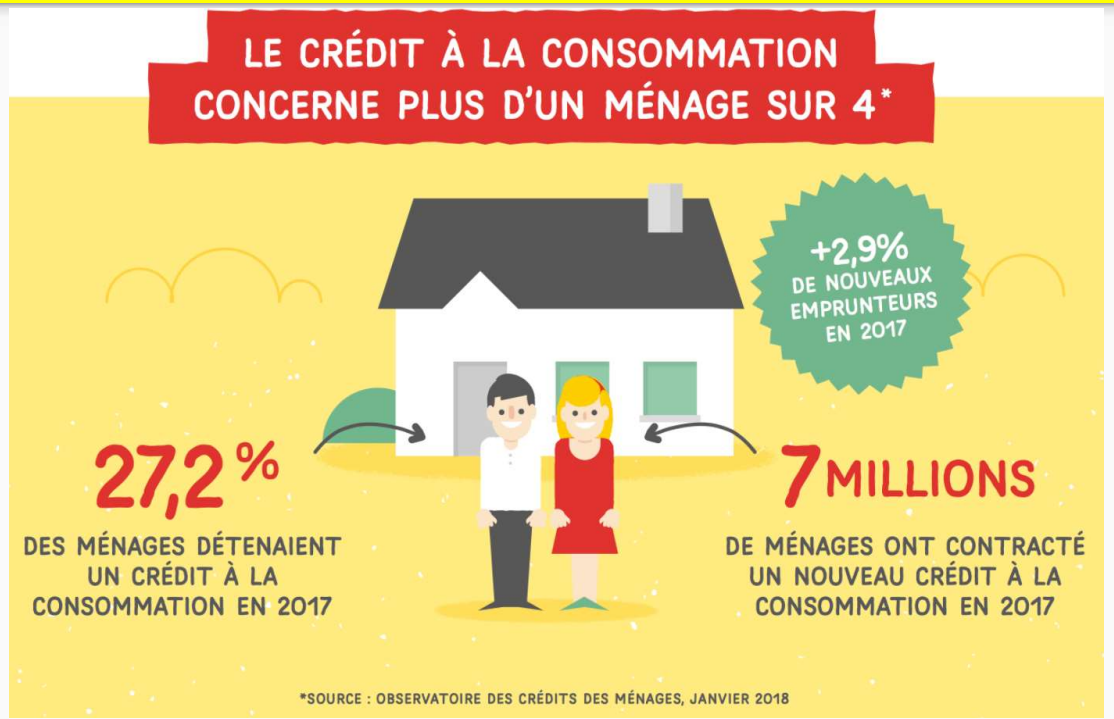
Besoin

Modèle de Scoring: Probabilité de défaut de paiement du client

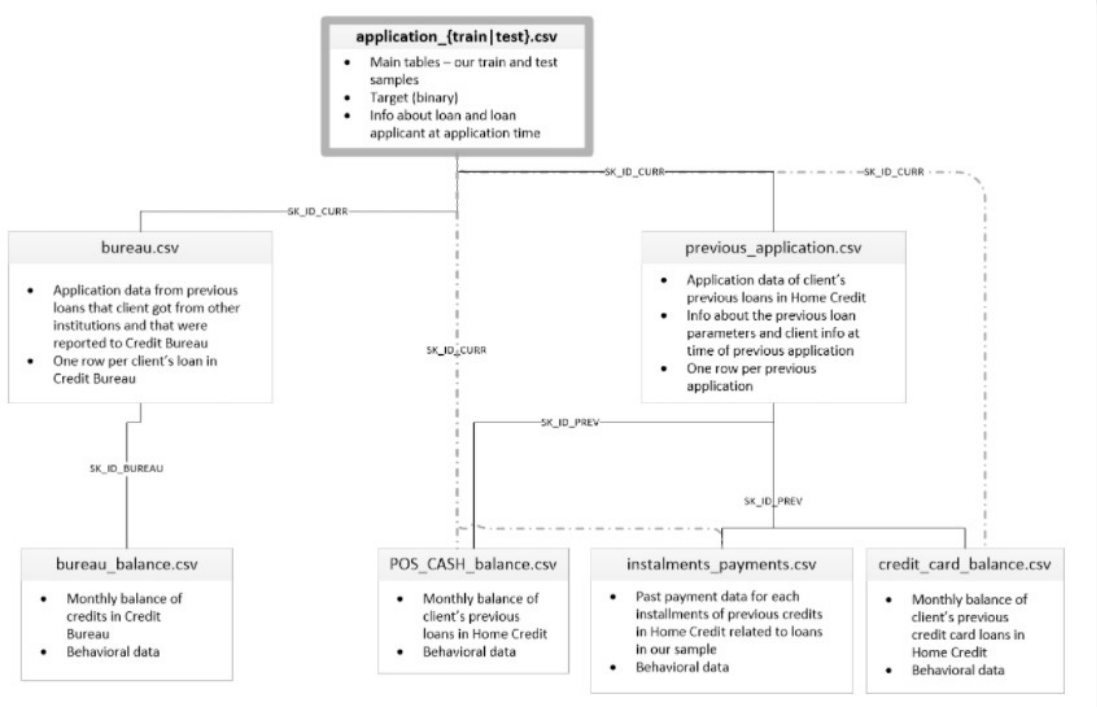
Objectif

Tableau de bord interactif: Destiné aux chargés de relation client

COMPREHENSION DE LA PROBLEMATIQUE METIER :
3/Impact sur le marché



DESCRIPTION DU JEU DE DONNEES :
1/ Le schéma du jeu de donnée

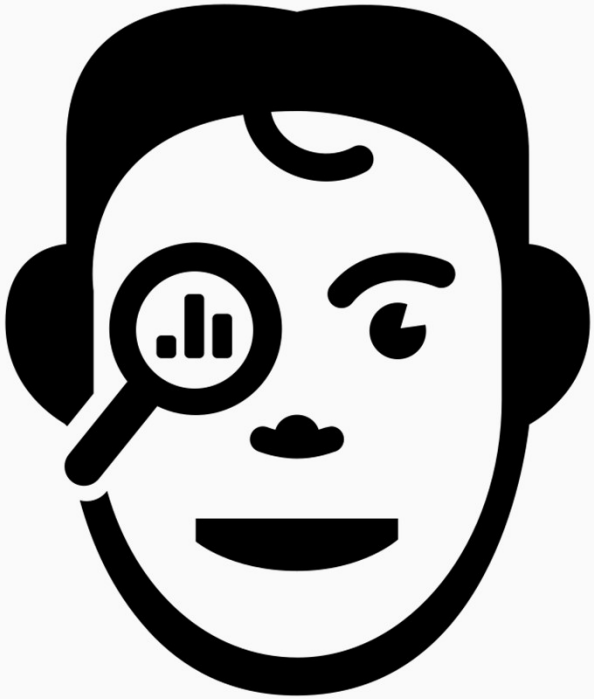


DESCRIPTION DU JEU DE DONNEES :
2/ Description des données de manière rapide

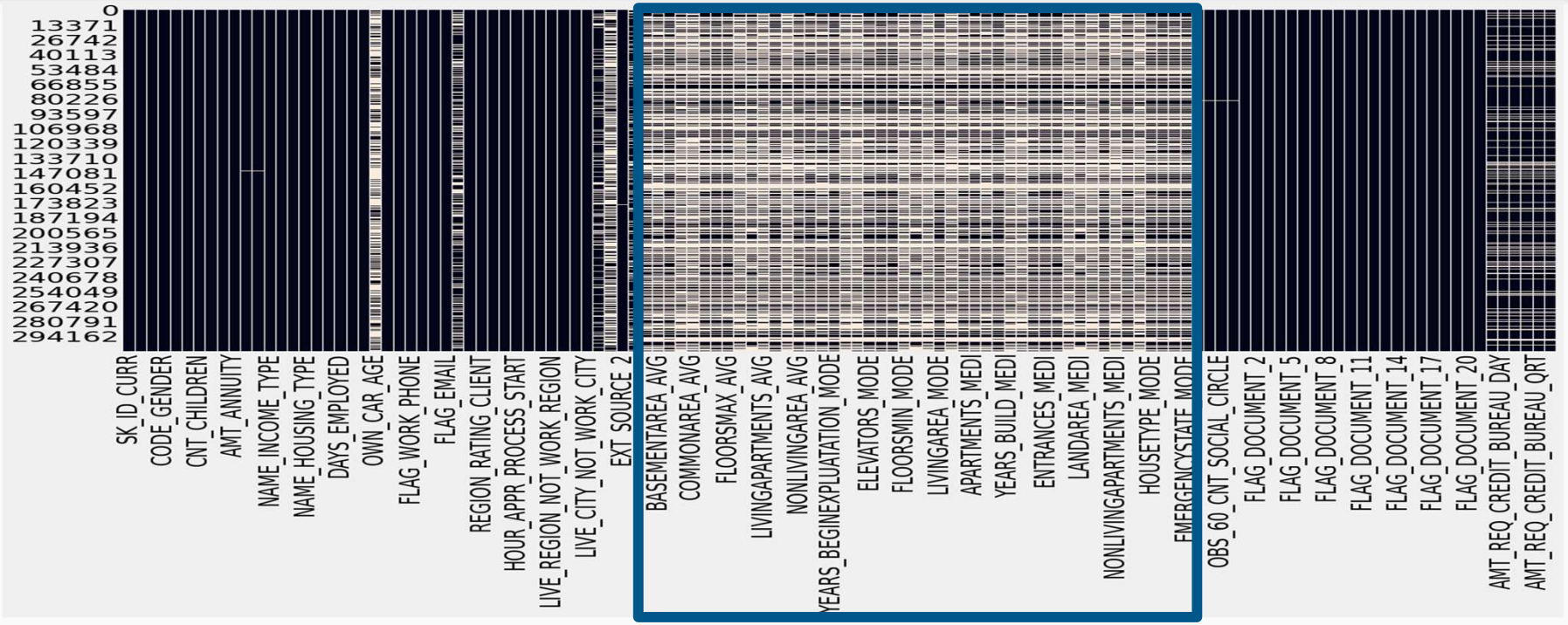
	Rows	Columns	%NaN	%Duplicate	object_dtype	float_dtype	int_dtype	bool_dtype	MB_Memory
./data/application_test.csv	48744	121	23.81	0.0	16	65	40	0	44.998
./data/POS_CASH_balance.csv	10001358	8	0.07	0.0	1	2	5	0	610.435
./data/credit_card_balance.csv	3840312	23	6.65	0.0	1	15	7	0	673.883
./data/installments_payments.csv	13605401	8	0.01	0.0	0	5	3	0	830.408
./data/application_train.csv	307511	122	24.40	0.0	16	65	41	0	286.227
./data/bureau.csv	1716428	17	13.50	0.0	3	8	6	0	222.620
./data/previous_application.csv	1670214	37	17.98	0.0	16	15	6	0	471.481
./data/bureau_balance.csv	27299925	3	0.00	0.0	1	0	2	0	624.846
./data/sample_submission.csv	48744	2	0.00	0.0	0	1	1	0	0.744

--------Jeu d’apprentissage

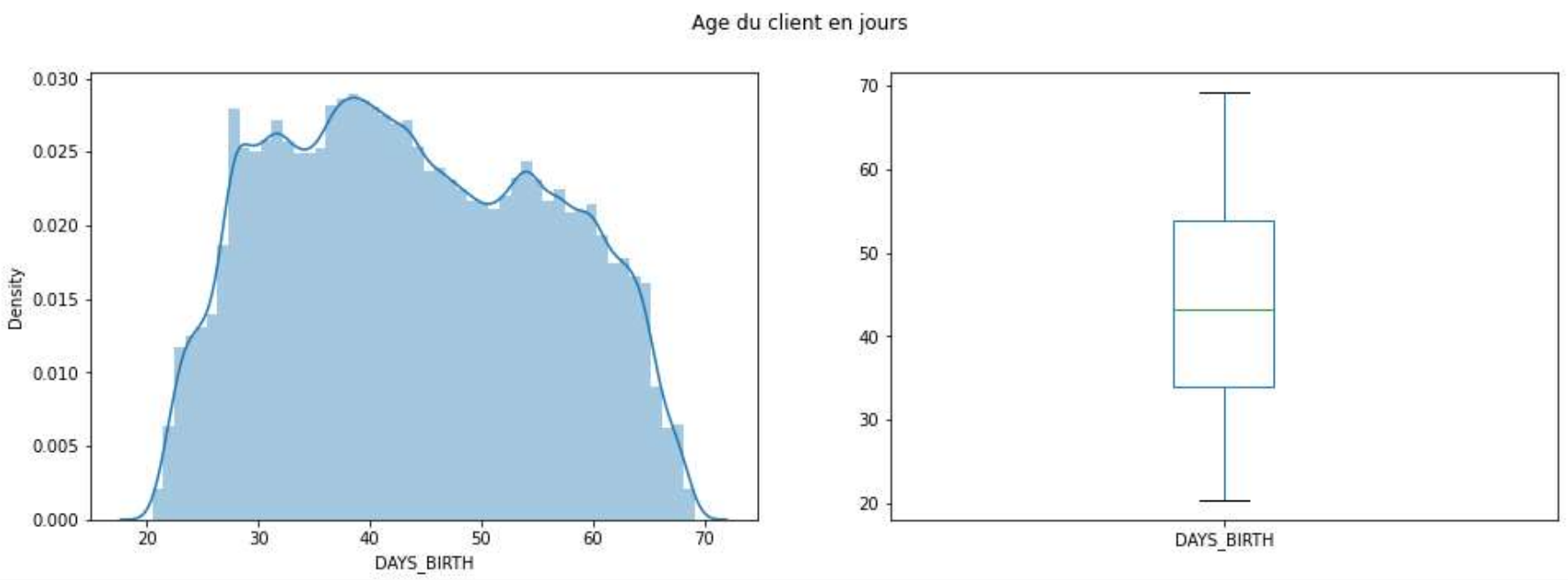
DESCRIPTION DU JEU DE DONNEES :
3/ Analyse exploratoire des données



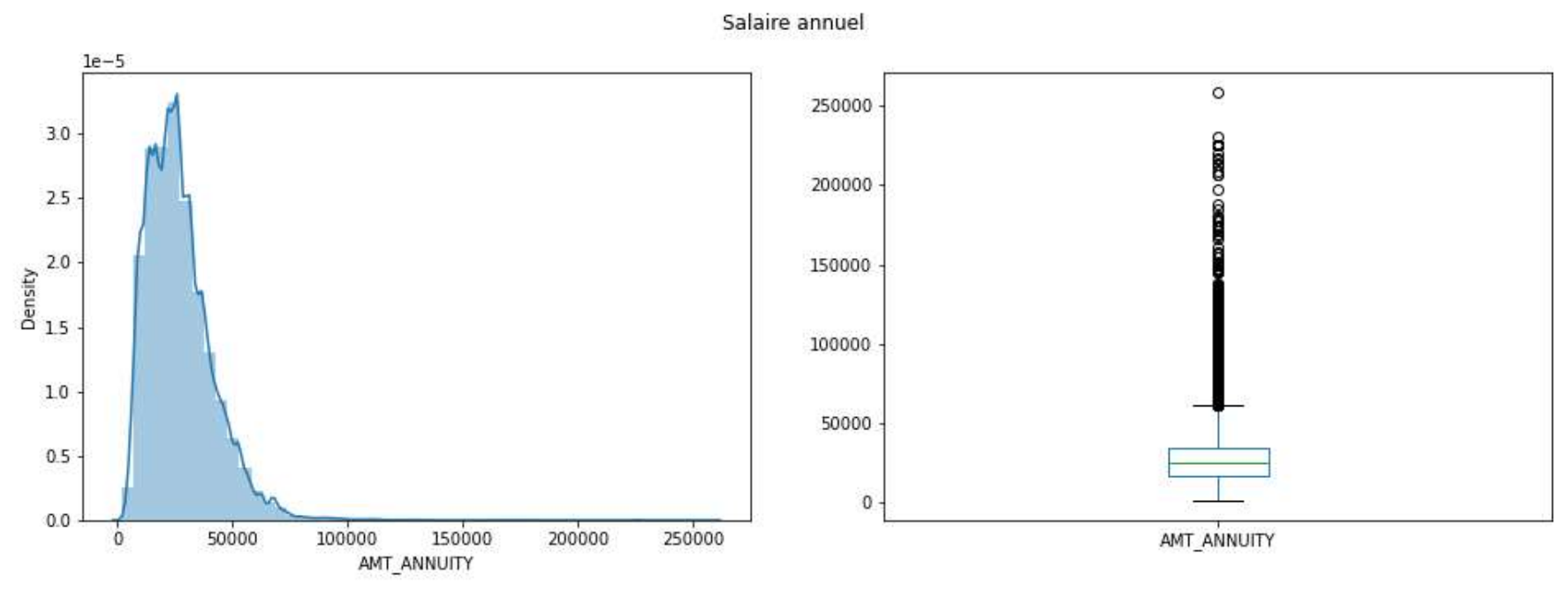
DESCRIPTION DU JEU DE DONNEES :
4/ Analyse graphique données d'apprentissage
1/ Analyse des valeurs manquantes



DESCRIPTION DU JEU DE DONNEES :
4/ Analyse graphique données d'apprentissage
2/ Analyse des outliers (aberrantes, atypique)
1/ Âge client

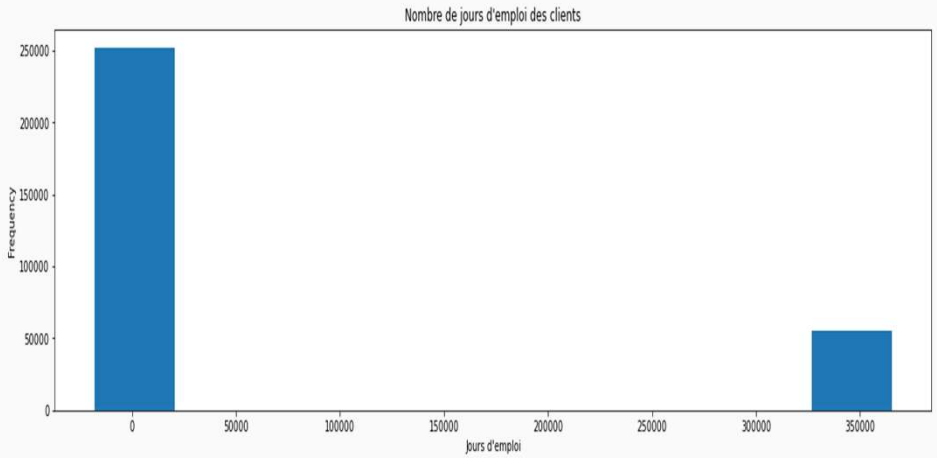


DESCRIPTION DU JEU DE DONNEES :
4/ Analyse graphique données d'apprentissage
2/ Analyse des outliers (aberrantes, atypique)
2/ Salaire annuel client

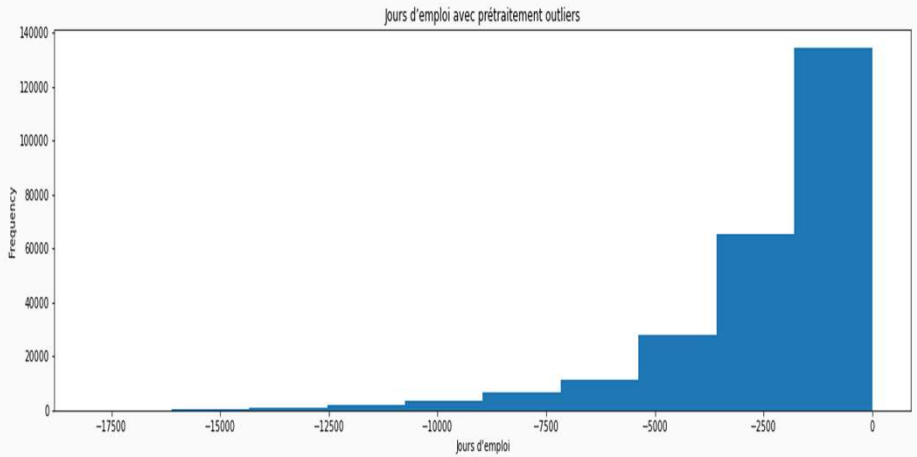


DESCRIPTION DU JEU DE DONNEES :
4/ Analyse graphique données d'apprentissage
2/ Analyse des outliers (aberrantes, atypique)
3/ Nombre de jours d'emploi

Avant



Après



TRANSFORMATION DU JEU DE DONNEES :
1/ Fusion et agrégations (assemblage) des données plus de nouvelles variables

Enrichissement de l'échantillon de travail :

Combinaison des 7 jeux de données.

Avant 122 variables - Après 193 variables.

Dont 3 variables fonctionnelles extraites :

PREVIOUS_LOANS_COUNT de bureau.csv: Nombre total des précédents crédits pris par chaque client

MONTHS_BALANCE_MEAN de bureau_balance.csv: Solde moyen mensuel des précédents crédits

PREVIOUS_APPLICATION_COUNT de previous_application.csv: Nombre de demandes antérieures des clients au crédit immobilier

Dont 4 nouvelles variables métiers:

PREVIOUS_LOANS_COUNT de bureau.csv: Nombre total des précédents crédits pris par chaque client

MONTHS_BALANCE_MEAN de bureau_balance.csv: Solde moyen mensuel des précédents crédits

PREVIOUS_APPLICATION_COUNT de previous_application.csv: Nombre de demandes antérieures des clients au crédit immobilier

TRANSFORMATION DU JEU DE DONNEES : 2/ Encodage des variables catégoriques par étiquette

Avant	
float64	137
int64	39
object	16
bool	1

Après	
float64	137
int64	40
object	13
int32	3

TRANSFORMATION DU JEU DE DONNEES :
3/ Encodage des variables catégoriques à chaud

Avant

float64	137
int64	40
object	13
int32	3

Après

float64	137
uint8	132
int64	40
int32	3

TRANSFORMATION DU JEU DE DONNEES :
4/ Vérification de mon jeu de données « Data »

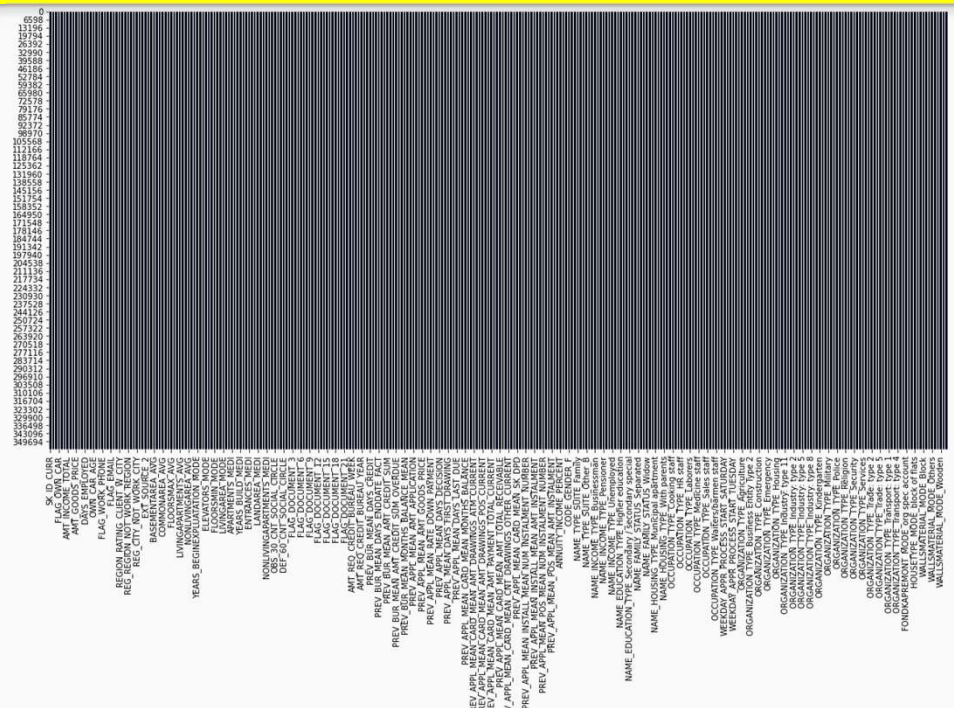
Avant

356255 Echantillons
193 Variables

Après

356255 Echantillons,
312 Variables

TRANSFORMATION DU JEU DE DONNEES :
5/ Imputation des valeurs manquantes



TRANSFORMATION DU JEU DE DONNEES :
6/ Standardisation des données

Avant

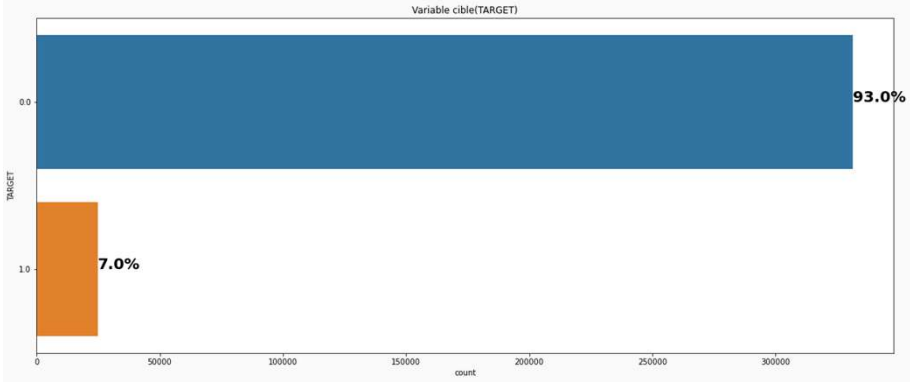
CNT_CHILDREN	AMT_INCOME_TOTAL	A
0.0	202500.0	
0.0	270000.0	
0.0	67500.0	
0.0	135000.0	
0.0	121500.0	
0.0	99000.0	
1.0	171000.0	
0.0	360000.0	
0.0	112500.0	
0.0	135000.0	

Après

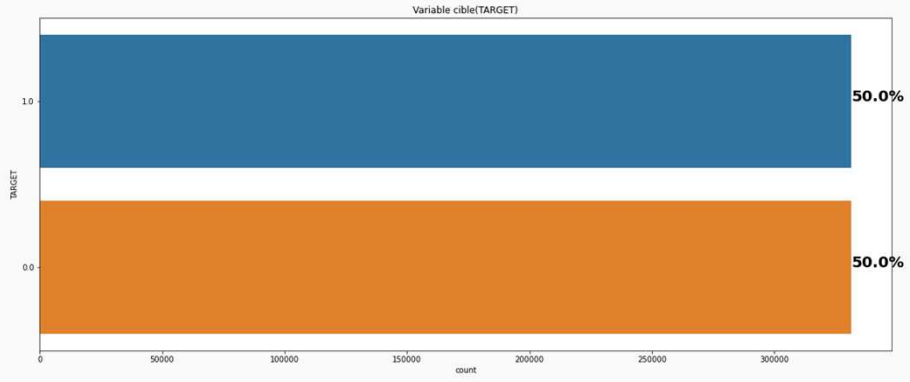
CNT_CHILDREN	AMT_INCOME_TOTAL	A
0.00	0.001512	
0.00	0.002089	
0.00	0.000358	
0.00	0.000935	
0.00	0.000819	
0.00	0.000627	
0.05	0.001243	
0.00	0.002858	
0.00	0.000742	
0.00	0.000935	

TRANSFORMATION DU JEU DE DONNEES :
7/ Equilibrage de la variable TARGET SMOTE

Avant



Après



TRANSFORMATION DU JEU DE DONNEES :
8/ Vérifier les corrélations du jeu de données en fonction de la TARGET

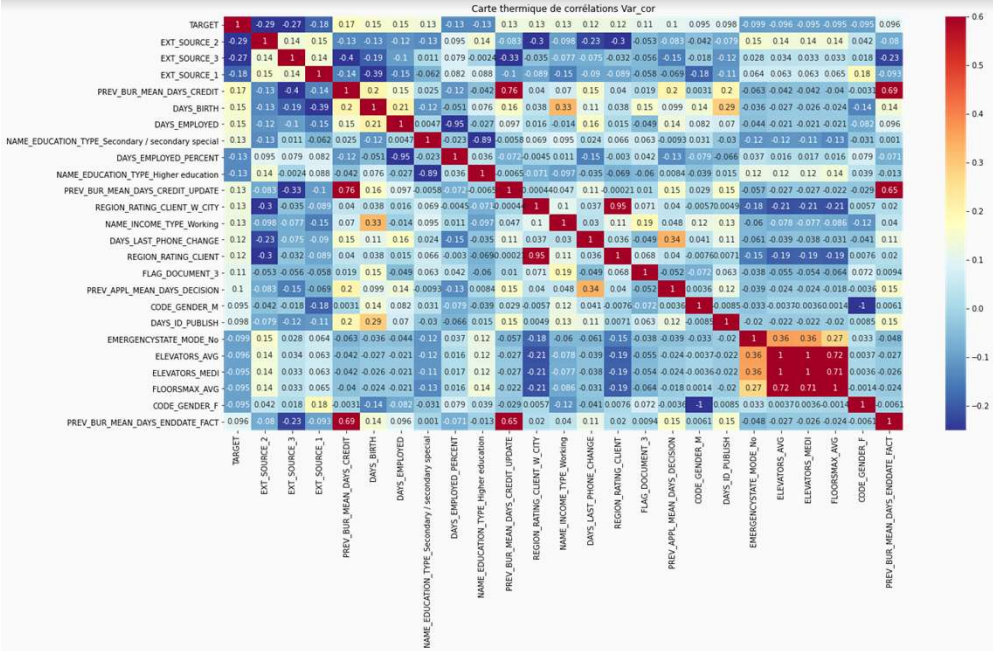
Corrélations les plus positives:

CODE_GENDER_M	0.094901
PREV_BUR_MEAN_DAYS_ENDDATE_FACT	0.095990
DAYS_ID_PUBLISH	0.097661
PREV_APPL_MEAN_DAYS_DECISION	0.102629
FLAG_DOCUMENT_3	0.107680
REGION_RATING_CLIENT	0.123151
DAYS_LAST_PHONE_CHANGE	0.124920
NAME_INCOME_TYPE_Working	0.125965
REGION_RATING_CLIENT_W_CITY	0.128619
PREV_BUR_MEAN_DAYS_CREDIT_UPDATE	0.130468
NAME_EDUCATION_TYPE_Secondary / secondary special	0.132558
DAYS_EMPLOYED	0.146549
DAYS_BIRTH	0.149723
PREV_BUR_MEAN_DAYS_CREDIT	0.168499
TARGET	1.000000

Corrélations les plus négatives:

EXT_SOURCE_2	-0.290872
EXT_SOURCE_3	-0.273895
EXT_SOURCE_1	-0.184212
DAYS_EMPLOYED_PERCENT	-0.131517
NAME_EDUCATION_TYPE_Higher education	-0.131309
EMERGENCYSTATE_MODE_No	-0.099327
ELEVATORS_AVG	-0.095889
ELEVATORS_MEDI	-0.095111
FLOORSMAX_AVG	-0.095092
CODE_GENDER_F	-0.094888
FLOORSMAX_MEDI	-0.094508
FLOORSMAX_MODE	-0.091909
ELEVATORS_MODE	-0.091589
NAME_INCOME_TYPE_Pensioner	-0.090801
HOUSETYPE_MODE_block of flats	-0.090660

TRANSFORMATION DU JEU DE DONNEES :
10/ Réduire mon nombre de variables importantes



Très forte corrélation:

- DAYS_EMPLOYED=DAYS_EMPLOYED_PERCENT **-0,95**
- NAME_EDUCATION_TYPE_Secondary / secondary special= NAME_EDUCATION_TYPE_Higher education **-0,89**
- REGION_RATING_CLIENT_W_CITY=REGION_RATING_CLIENT **0,95**
- CODE_GENDER_M=CODE_GENDER_F **-1**
- ELEVATORS_MEDI=ELEVATORS_AVG **1**

Légende en valeur absolu:

- ✓ 00 à 0.19 “très faible”,
- ✓ 0.20 à 0.39 “faible”,
- ✓ 0.40 à 0.59 “modéré”,
- ✓ 0.60 à 0.79 “fort”,
- ✓ 0.80 à 1.0 “très fort”

Nouveau jeu de données:

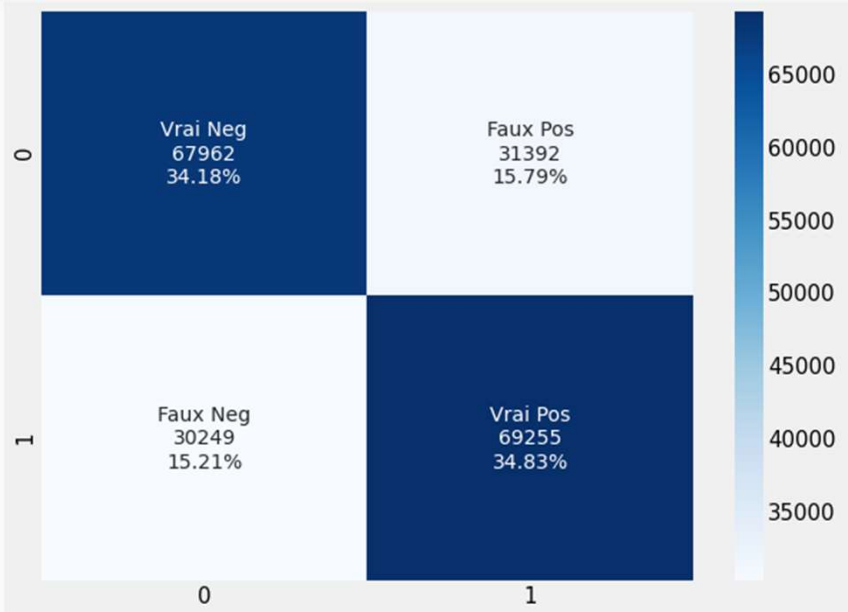
Var_imp comportant 25 variables

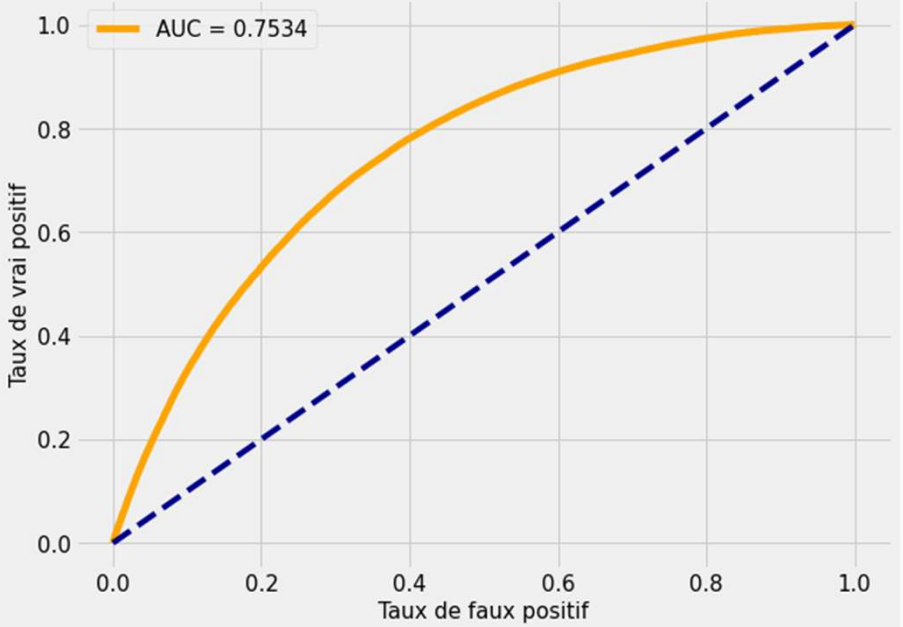
COMPARAISON ET SYNTHESE DES RESULTATS POUR LES MODELES UTILISES: 1/ Entrée en matière modélisation

CHOIX ENTRE DEUX BASELINES FIXEES PAR REGRESSION LOGISTIQUE ET ARBRE DE DECISION

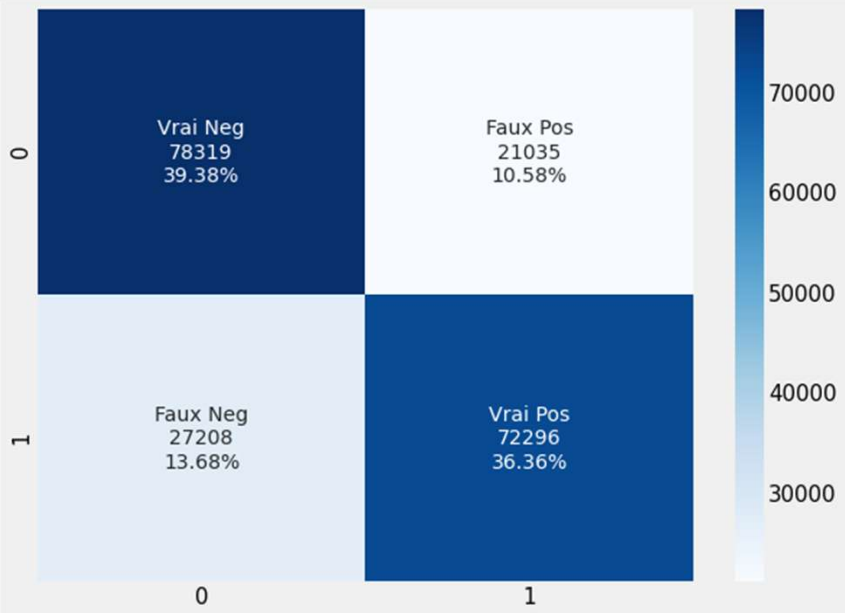
Elaboration d'un modèle, Optimisation et Compréhension

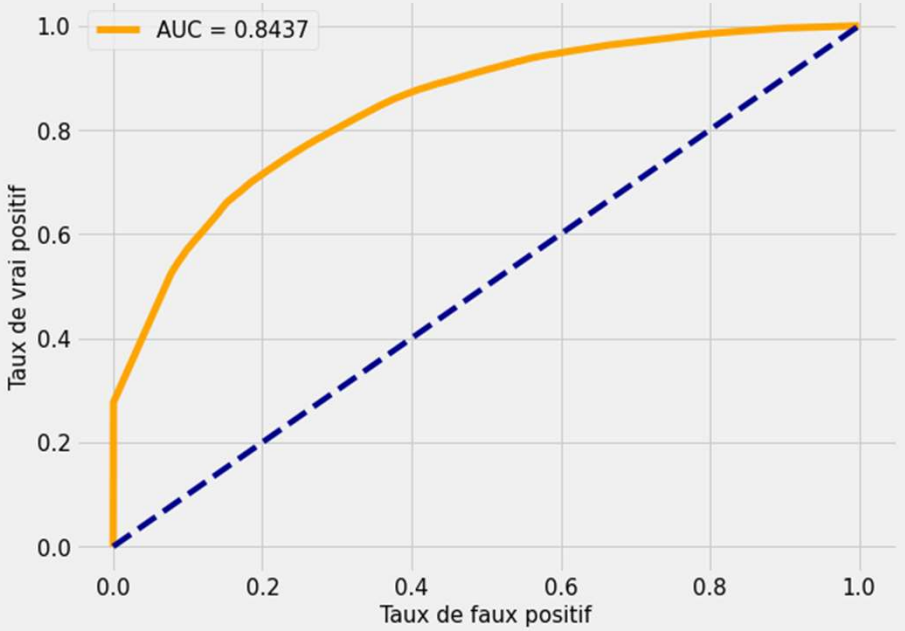
COMPARAISON ET SYNTHESE DES RESULTATS POUR LES MODELES UTILISES:
2/ Baseline Régression Logistique





COMPARAISON ET SYNTHESE DES RESULTATS POUR LES MODELES UTILISES:
3/ Arbre de décision





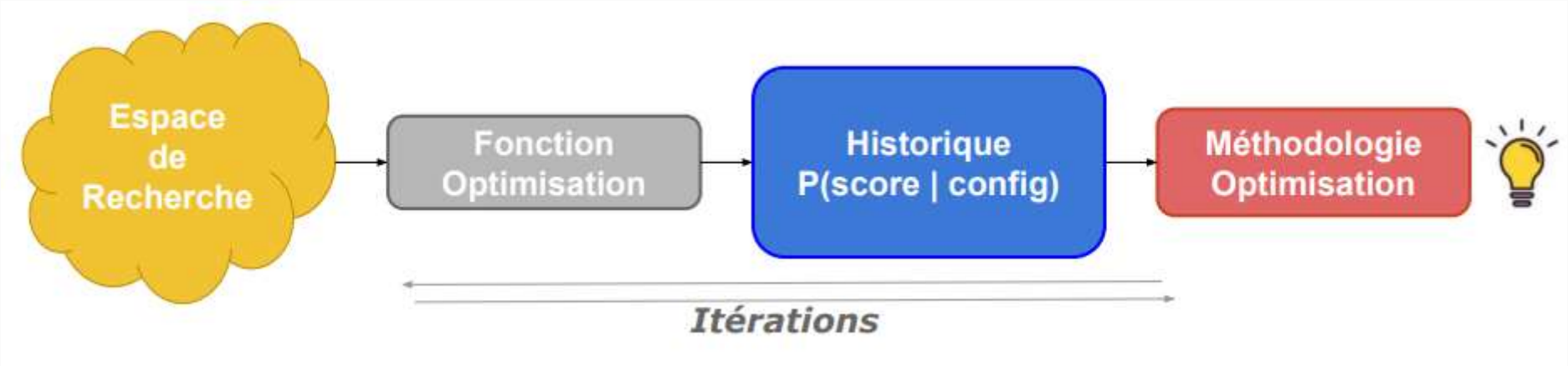
COMPARAISON ET SYNTHESE DES RESULTATS POUR LES MODELES UTILISES:
4/ Gradient Boosting

Le Gradient Boosting :

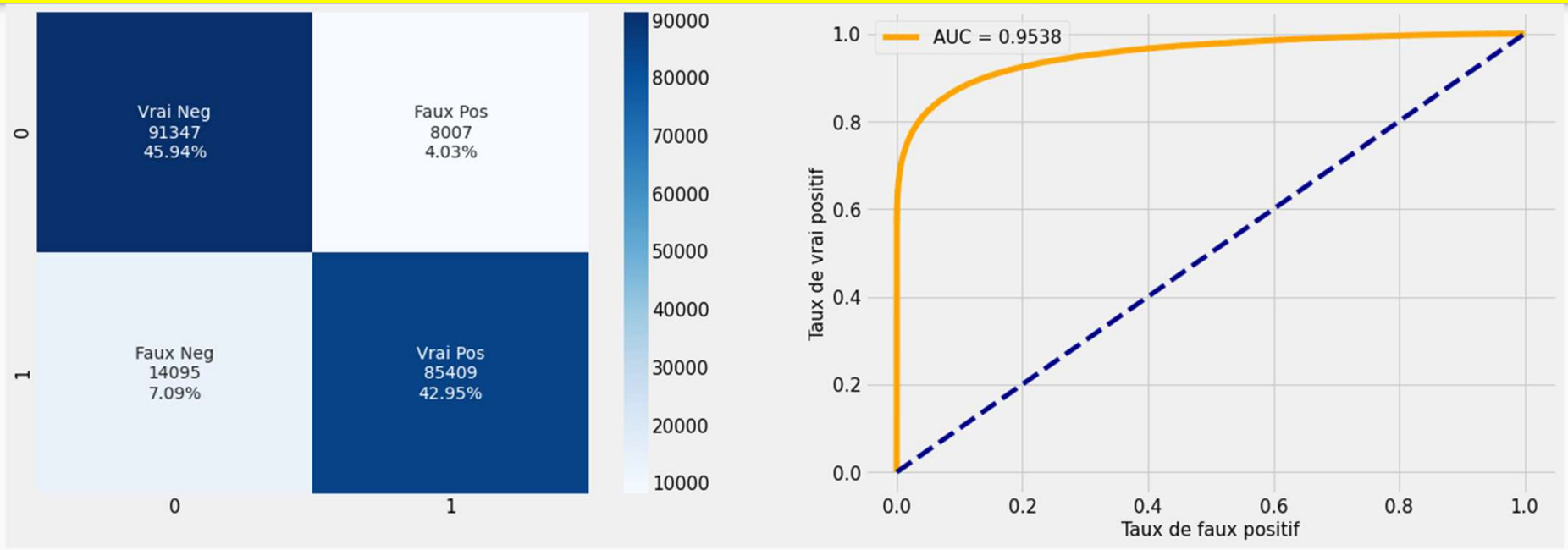
- ✓ Algorithme d'apprentissage supervisé,
- ✓ Combine les résultats d'un ensemble de modèle simple,
- ✓ Principe d'auto amélioration séquentielle.

	Model	AUC	Accuracy	Precision	Recall	F1	Time	fp
0	CatBoostClassifier	0.947878	0.88006	0.907442	0.846659	0.875998	9.69813	0.0432117
1	LGBMClassifier	0.947481	0.878999	0.909931	0.841474	0.874364	11.0713	0.041678
2	XGBClassifier	0.917335	0.8362	0.844973	0.823786	0.834245	2.25495	0.0756268

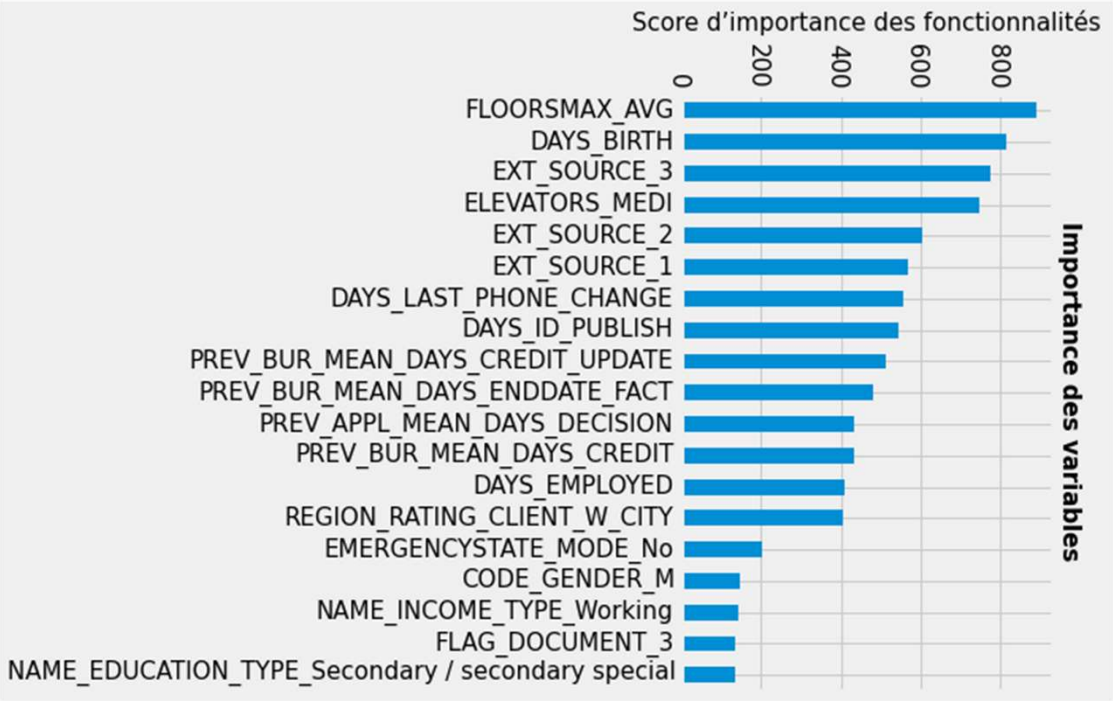
INTERPRETABILITE DU MODELE:
1/ Réglage d'hyperparamètres



INTERPRETABILITE DU MODELE:
2/ Modèle LGBMClassifier Optimisé



INTERPRETABILITE DU MODELE:
3/ Importance des variables



INTERPRETABILITE DU MODELE:
4/ La fonction coût

Faux Positif: Perte d'opportunité si le crédit client est refusé à tort, alors qu'il aurait été en mesure d'être remboursé.
Faux Négatif: Perte réelle si le crédit client accepté se transforme en défaut de paiement.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Défaut de paiement 30% du montant du crédit en pertes et autres frais de recouvrement.
10% de chance d'obtenir un crédit pour un client lambda qui souhaite emprunter.

Hypothèse d'un Beta = 3

Formule : fscore = (1+beta)*(tp / ((1+3)*tp + beta*fn + fp))

```
            tp  , tn  , fp  , fn
test_0 = [ 500 , 300 , 10  , 30  ]
test_1 = [ 500 , 300 , 30  , 10  ]
test_2 = [ 400 , 300 , 70  , 50  ]
test_3 = [ 400 , 300 , 50  , 70  ]
test_4 = [ 350 , 250 , 80  , 120 ]
test_5 = [ 350 , 250 , 180 , 90  ]
```

Résultats

```
##############################
Test 0 : [500, 300, 10, 30]
Score : 0.04761904761904767
##############################
Test 1 : [500, 300, 30, 10]
Score : 0.029126213592232997
##############################
Test 2 : [400, 300, 70, 50]
Score : 0.1208791208791209
##############################
Test 3 : [400, 300, 50, 70]
Score : 0.13978494623655913
##############################
Test 4 : [350, 250, 80, 120]
Score : 0.23913043478260865
##############################
Test 5 : [350, 250, 180, 90]
Score : 0.2432432432432432
```

CONCLUSION

- ✓ Ré-équilibrer mon jeu de données,
- ✓ Réduire mon jeu de données à 25 variables,
- ✓ Entrainer mon jeu de données sur ces 25 variables,
- ✓ Réaliser une baseline optimisée faite sur deux algorithmes simple,
- ✓ Réaliser modèlé de 3 algorithmes plus complexes de gradient boosting implémentés par LightGbm vs CatBoost vs XGBoost,
- ✓ Mon choix de modèle c'est porté sur le LGBMClassifier Optimisé,
- ✓ Détection des erreurs de prédiction.

REMERCIEMENT

Merci de m’avoir écouter

REPONDRE AUX QUESTIONS