

CONSTRUISEZ UN MODELE DE SCORING

Présentation “Projet 4” chez “OPENCLASSROOM”
Jaoid KRAIRI
(Septembre 2021)

SOMMAIRE

- ➡ Compréhension de la problématique métier
- ➡ Description du jeu de données
- ➡ Transformation du jeu de données (nettoyage et feature engineering)
- ➡ Comparaison et synthèse des résultats pour les modèles utilisés
- ➡ Interprétabilité du modèle
- ➡ Conclusion
- ➡ Remerciements

COMPREHENSION DE LA PROBLEMATIQUE METIER :
1/Contexte

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

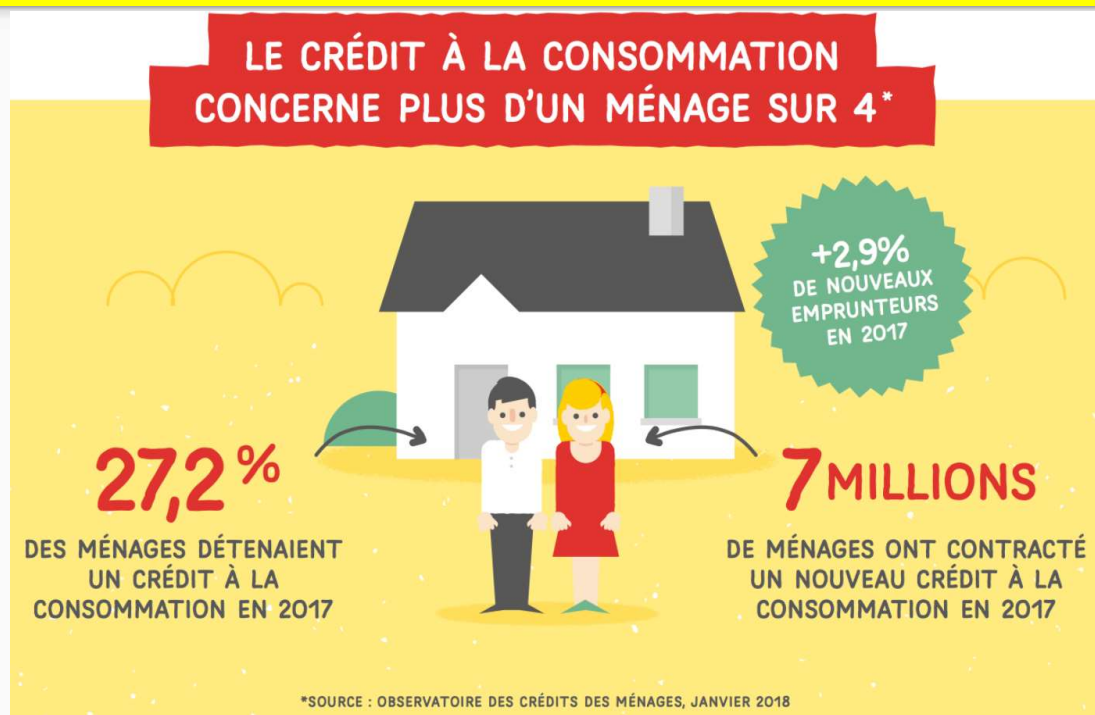
Modèle de Scoring: Probabilité de défaut de paiement du client

Objectif

Tableau de bord interactif: Destiné aux chargés de relation client

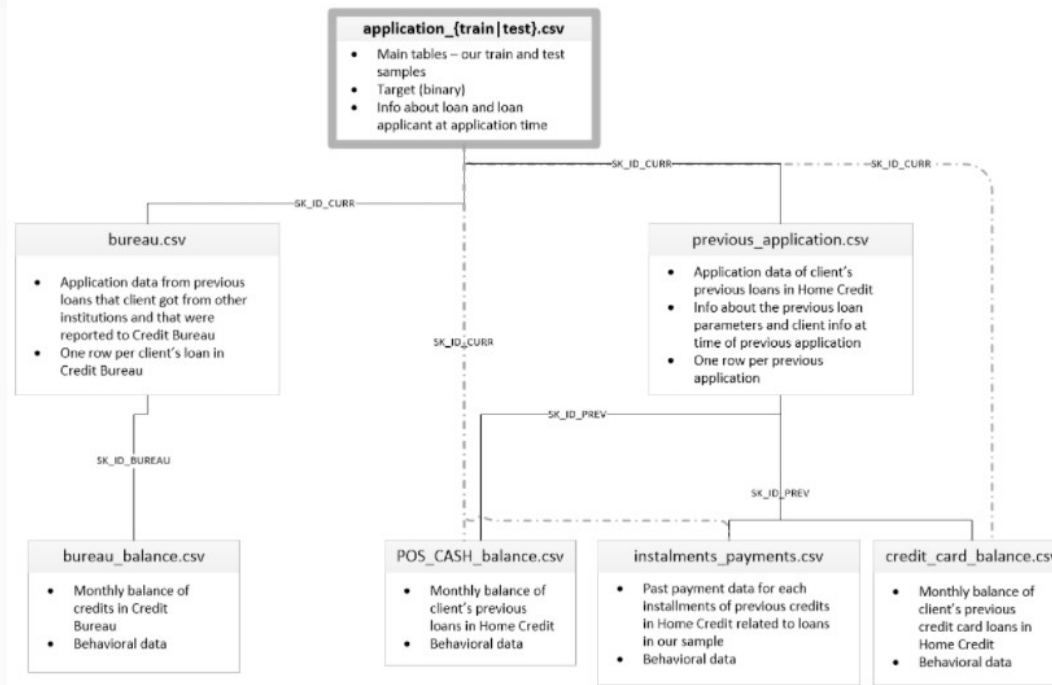
COMPREHENSION DE LA PROBLEMATIQUE METIER :

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



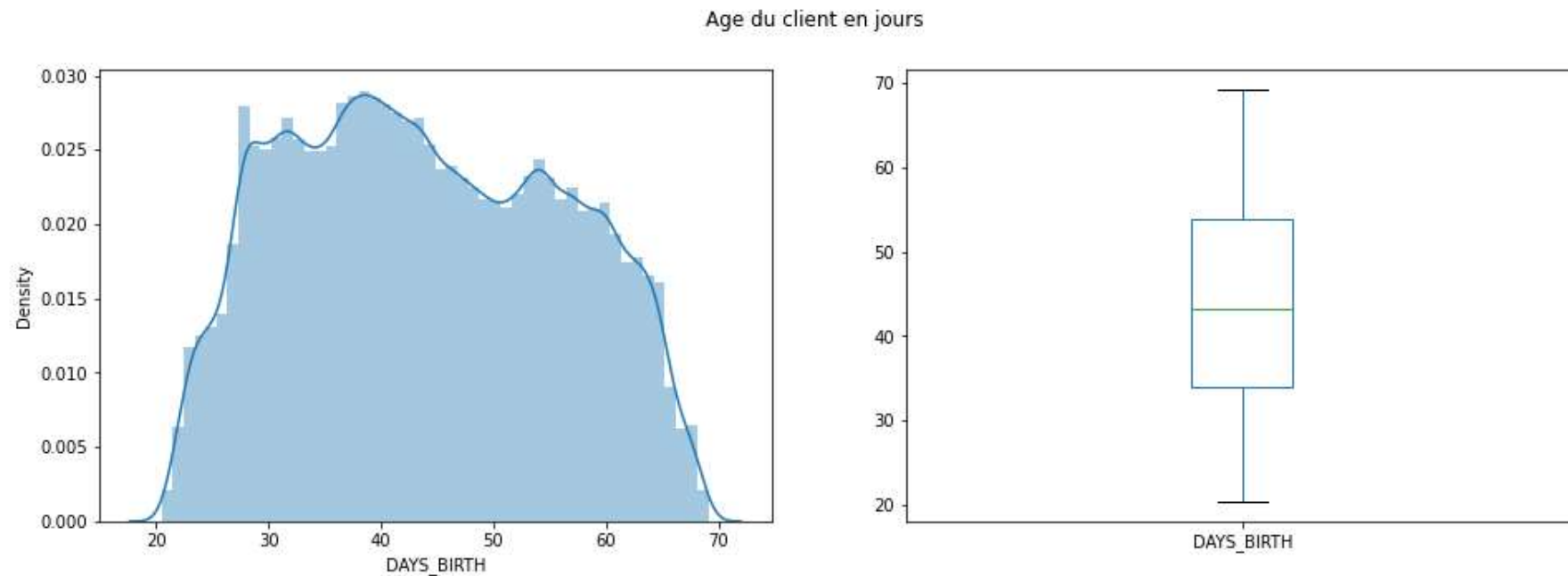
DESCRIPTION DU JEU DE DONNEES :

4/ Analyse graphique données d'apprentissage

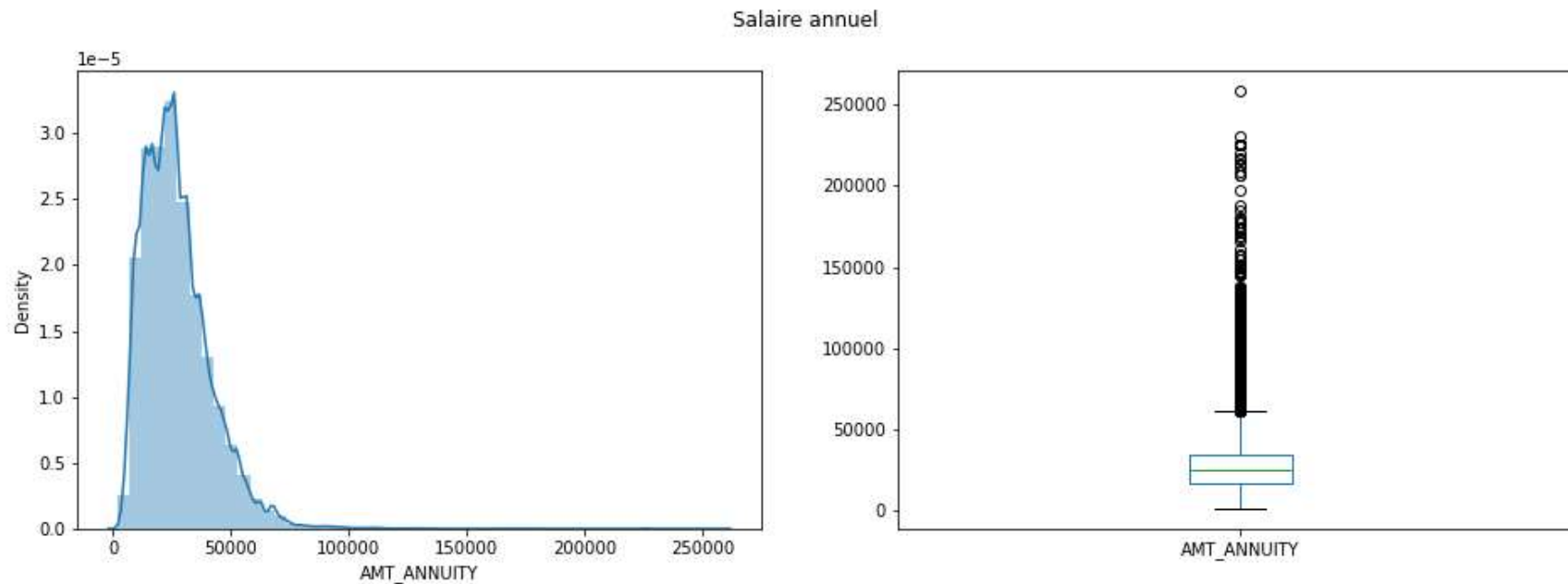
1/ Analyse des valeurs manquantes

133371	SK_ID_CURR
267422	CODE_GENDER
401113	CNT_CHILDREN
534884	AMT_ANNUITY
668555	NAME_INCOME_TYPE
802226	NAME_HOUSING_TYPE
980507	DAYS_EMPLOYED
106968	OWN_CAR_AGE
120033	FLAG_WORK_PHONE
133710	FLAG_EMAIL
147081	REGION_RATING_CLIENT
160452	HOUR_APPR_PROCESS_START
173823	LIVE_REGION_NOT_WORK_REGION
187194	LIVE_CITY_NOT_WORK_CITY
200565	EXT_SOURCE_2
213936	BASEMENTAREA_AVG
227307	COMMONAREA_AVG
440678	FLOORSMAX_AVG
554049	LIVINGAPARTMENTS_AVG
674200	NONLIVINGAREA_AVG
807901	YEARS_BEGINEXPLUATION_MODE
94162	ELEVATORS_MODE
	FLOORSMIN_MODE
	LIVINGAREA_MODE
	APARTMENTS_MEDI
	YEARS_BUILD_MEDI
	ENTRANCES_MEDI
	LANDAREA_MEDI
	NONLIVINGAPARTMENTS_MEDI
	HOUSETYPE_MODE
	EMERGENCYSTATE_MODE
	OBS_60_CNT_SOCIAL_CIRCLE
	FLAG_DOCUMENT_2
	FLAG_DOCUMENT_5
	FLAG_DOCUMENT_8
	FLAG_DOCUMENT_11
	FLAG_DOCUMENT_14
	FLAG_DOCUMENT_17
	FLAG_DOCUMENT_20
	AMT_REQ_CREDIT_BUREAU_DAY
	AMT_REQ_CREDIT_BUREAU_QRT

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



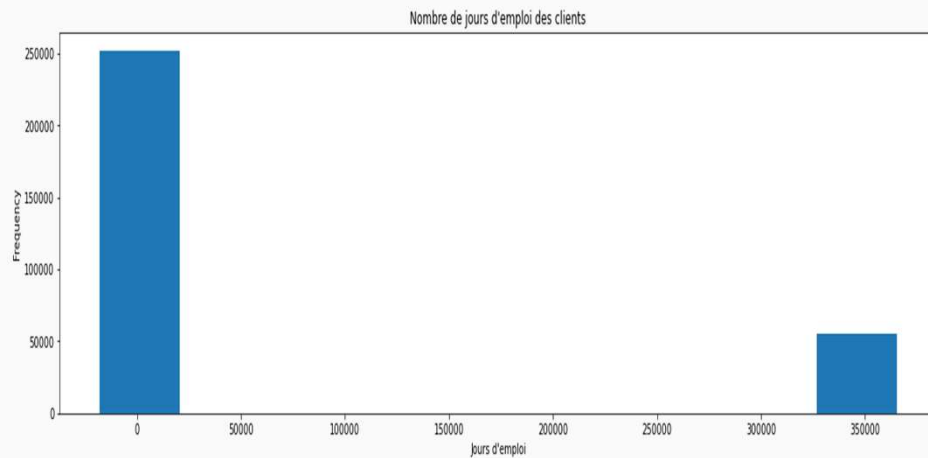
DESCRIPTION DU JEU DE DONNEES :
4/ Analyse graphique données d'apprentissage
2/ Analyse des outliers (aberrantes, atypique)
2/ Salaire annuel client



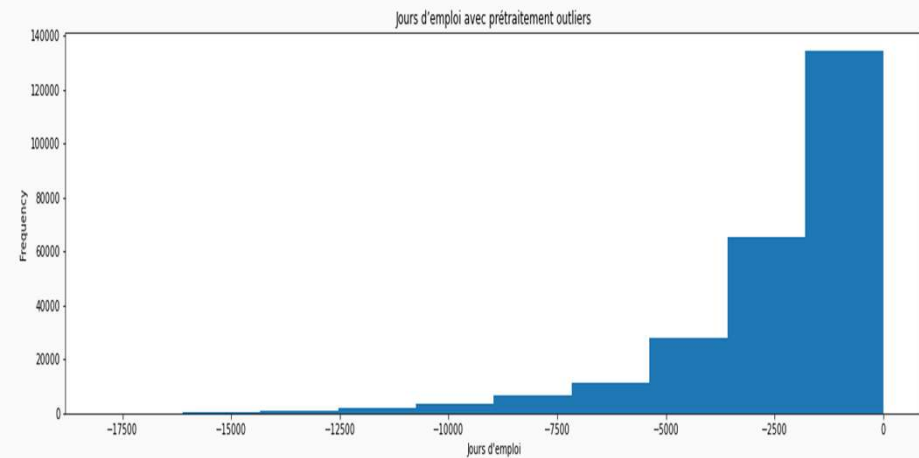
DESCRIPTION DU JEU DE DONNEES :

- 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

0	
6598	FLG_OVR_CAR
13196	FLG_OVR_CAR
19184	FLG_OVR_CAR
26192	FLG_OVR_CAR
32990	FLG_OVR_CAR
39588	FLG_OVR_CAR
46186	FLG_OVR_CAR
52784	FLG_OVR_CAR
59382	FLG_OVR_CAR
65980	FLG_OVR_CAR
72578	FLG_OVR_CAR
79176	FLG_OVR_CAR
85774	FLG_OVR_CAR
92372	FLG_OVR_CAR
98970	FLG_OVR_CAR
105568	FLG_OVR_CAR
112166	FLG_OVR_CAR
118764	FLG_OVR_CAR
125362	FLG_OVR_CAR
131960	FLG_OVR_CAR
138558	FLG_OVR_CAR
145156	FLG_OVR_CAR
151754	FLG_OVR_CAR
158352	FLG_OVR_CAR
164950	FLG_OVR_CAR
171548	FLG_OVR_CAR
178146	FLG_OVR_CAR
184744	FLG_OVR_CAR
191342	FLG_OVR_CAR
197940	FLG_OVR_CAR
204538	FLG_OVR_CAR
211136	FLG_OVR_CAR
217734	FLG_OVR_CAR
224332	FLG_OVR_CAR
230930	FLG_OVR_CAR
237528	FLG_OVR_CAR
244126	FLG_OVR_CAR
250724	FLG_OVR_CAR
257322	FLG_OVR_CAR
263920	FLG_OVR_CAR
270518	FLG_OVR_CAR
277116	FLG_OVR_CAR
283714	FLG_OVR_CAR
290312	FLG_OVR_CAR
296910	FLG_OVR_CAR
303508	FLG_OVR_CAR
310106	FLG_OVR_CAR
316704	FLG_OVR_CAR
323302	FLG_OVR_CAR
329900	FLG_OVR_CAR
336498	FLG_OVR_CAR
343096	FLG_OVR_CAR
349694	FLG_OVR_CAR
356292	FLG_OVR_CAR
362890	FLG_OVR_CAR
369488	FLG_OVR_CAR
376086	FLG_OVR_CAR
382684	FLG_OVR_CAR
389282	FLG_OVR_CAR
395880	FLG_OVR_CAR
402478	FLG_OVR_CAR
409076	FLG_OVR_CAR
415674	FLG_OVR_CAR
422272	FLG_OVR_CAR
428870	FLG_OVR_CAR
435468	FLG_OVR_CAR
442066	FLG_OVR_CAR
448664	FLG_OVR_CAR
455262	FLG_OVR_CAR
461860	FLG_OVR_CAR
468458	FLG_OVR_CAR
475056	FLG_OVR_CAR
481654	FLG_OVR_CAR
488252	FLG_OVR_CAR
494850	FLG_OVR_CAR
501448	FLG_OVR_CAR
508046	FLG_OVR_CAR
514644	FLG_OVR_CAR
521242	FLG_OVR_CAR
527840	FLG_OVR_CAR
534438	FLG_OVR_CAR
541036	FLG_OVR_CAR
547634	FLG_OVR_CAR
554232	FLG_OVR_CAR
560830	FLG_OVR_CAR
567428	FLG_OVR_CAR
574026	FLG_OVR_CAR
580624	FLG_OVR_CAR
587222	FLG_OVR_CAR
593820	FLG_OVR_CAR
600418	FLG_OVR_CAR
607016	FLG_OVR_CAR
613614	FLG_OVR_CAR
620212	FLG_OVR_CAR
626810	FLG_OVR_CAR
633408	FLG_OVR_CAR
640006	FLG_OVR_CAR
646604	FLG_OVR_CAR
653202	FLG_OVR_CAR
659800	FLG_OVR_CAR
666398	FLG_OVR_CAR
672996	FLG_OVR_CAR
679594	FLG_OVR_CAR
686192	FLG_OVR_CAR
692790	FLG_OVR_CAR
699388	FLG_OVR_CAR
705986	FLG_OVR_CAR
712584	FLG_OVR_CAR
719182	FLG_OVR_CAR
725780	FLG_OVR_CAR
732378	FLG_OVR_CAR
738976	FLG_OVR_CAR
745574	FLG_OVR_CAR
752172	FLG_OVR_CAR
758770	FLG_OVR_CAR
765368	FLG_OVR_CAR
771966	FLG_OVR_CAR
778564	FLG_OVR_CAR
785162	FLG_OVR_CAR
791760	FLG_OVR_CAR
798358	FLG_OVR_CAR
804956	FLG_OVR_CAR
811554	FLG_OVR_CAR
818152	FLG_OVR_CAR
824750	FLG_OVR_CAR
831348	FLG_OVR_CAR
837946	FLG_OVR_CAR
844544	FLG_OVR_CAR
851142	FLG_OVR_CAR
857740	FLG_OVR_CAR
864338	FLG_OVR_CAR
870936	FLG_OVR_CAR
877534	FLG_OVR_CAR
884132	FLG_OVR_CAR
890730	FLG_OVR_CAR
897328	FLG_OVR_CAR
903926	FLG_OVR_CAR
910524	FLG_OVR_CAR
917122	FLG_OVR_CAR
923720	FLG_OVR_CAR
930318	FLG_OVR_CAR
936916	FLG_OVR_CAR
943514	FLG_OVR_CAR
950112	FLG_OVR_CAR
956710	FLG_OVR_CAR
963308	FLG_OVR_CAR
969906	FLG_OVR_CAR
976504	FLG_OVR_CAR
983102	FLG_OVR_CAR
989700	FLG_OVR_CAR
996298	FLG_OVR_CAR
1002876	FLG_OVR_CAR
1009474	FLG_OVR_CAR
1016072	FLG_OVR_CAR
1022670	FLG_OVR_CAR
1029268	FLG_OVR_CAR
1035866	FLG_OVR_CAR

TRANSFORMATION DU JEU DE DONNEES :

6/ Standardisation des données

Avant

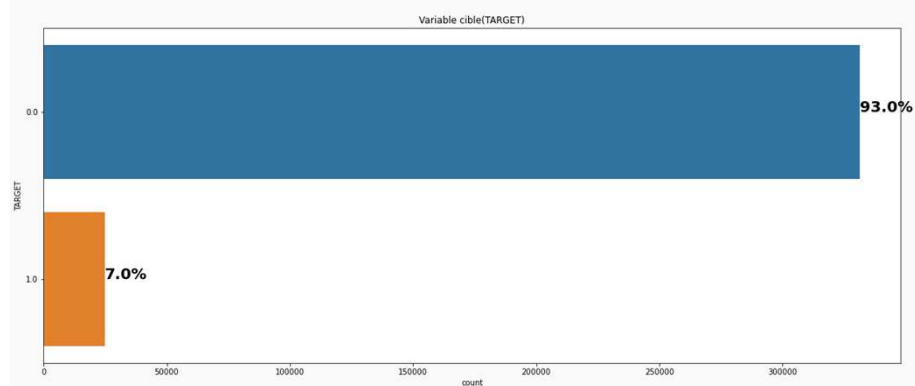
CNT_CHILDREN	AMT_INCOME_TOTAL
0.0	202500.0
0.0	270000.0
0.0	67500.0
0.0	135000.0
0.0	121500.0
0.0	99000.0
1.0	171000.0
0.0	360000.0
0.0	112500.0
0.0	135000.0

Après

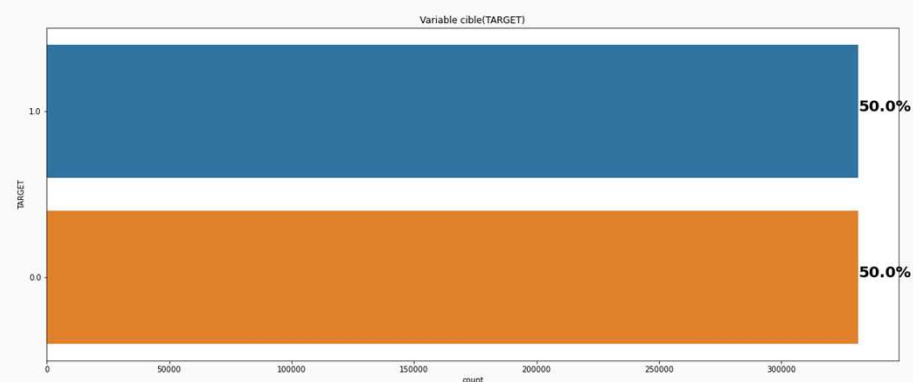
CNT_CHILDREN	AMT_INCOME_TOTAL
0.00	0.001512
0.00	0.002089
0.00	0.000358
0.00	0.000935
0.00	0.000819
0.00	0.000627
0.05	0.001243
0.00	0.002858
0.00	0.000742
0.00	0.000935

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:

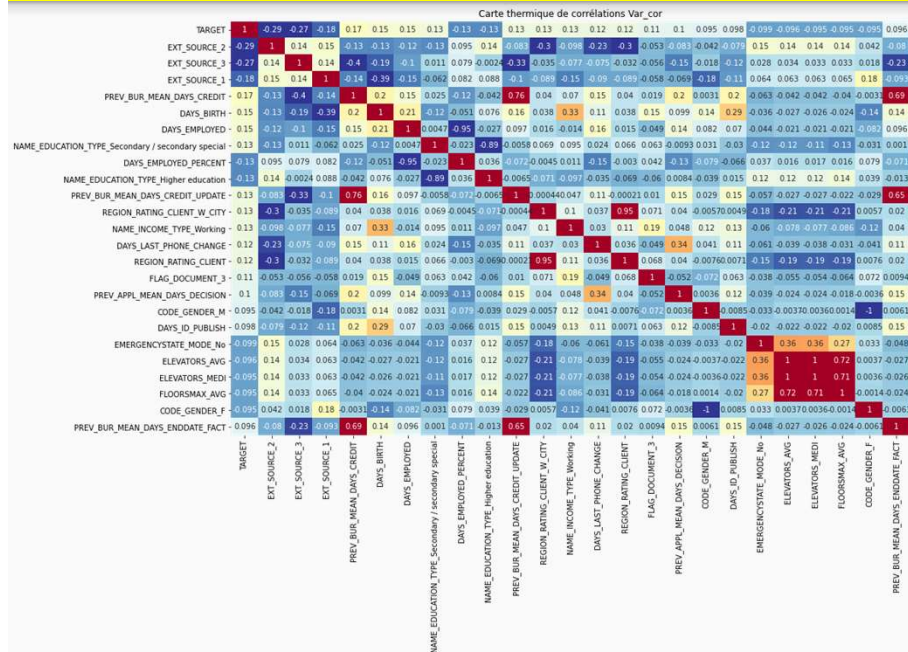
CODE_GENDER_M	0.094901
PREV_BUR_MEAN_DAYS_ENDDATE_FACT	0.095990
DAYS_ID_PUBLISH	0.097661
PREV_APPL_MEAN_DAYS_DECISION	0.102629
FLAG_DOCUMENT_3	0.107680
REGION_RATING_CLIENT	0.123151
DAYS_LAST_PHONE_CHANGE	0.124920
NAME_INCOME_TYPE_Working	0.125965
REGION_RATING_CLIENT_W_CITY	0.128619
PREV_BUR_MEAN_DAYS_CREDIT_UPDATE	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= **-0,89**
NAME_EDUCATION_TYPE_Higher education **0,95**
- 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

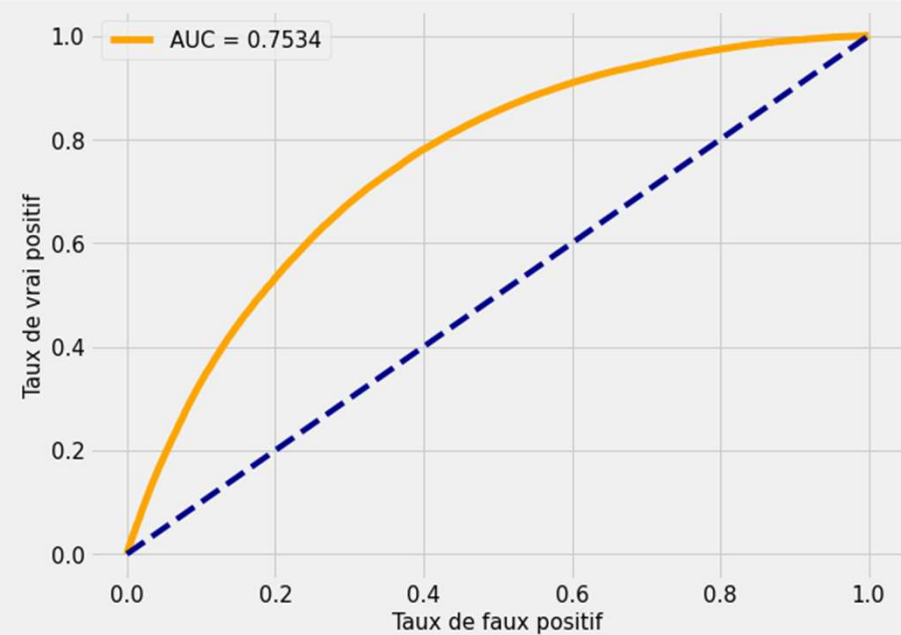
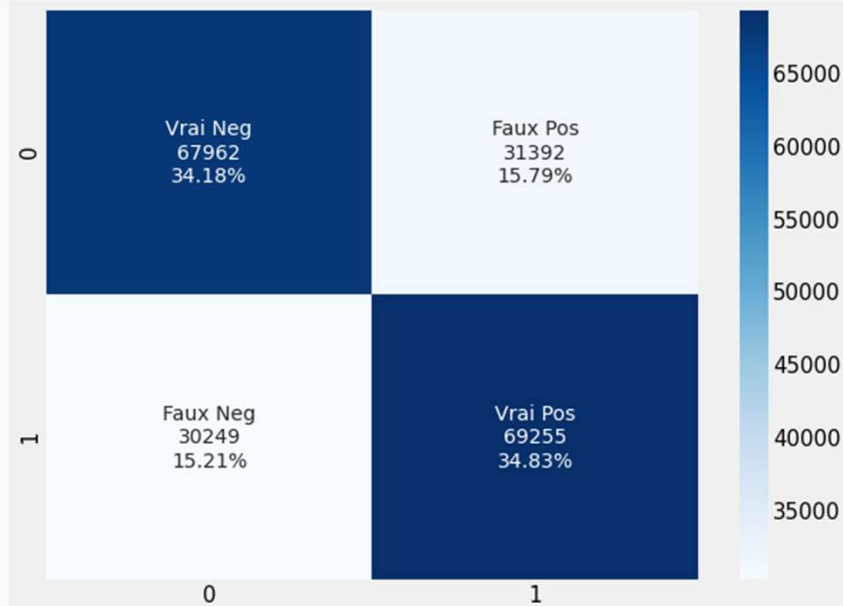
**COMPARAISON ET SYNTHÈSE DES RESULTATS
POUR LES MODELES UTILISES:
1/ Entrée en matière modélisation**

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

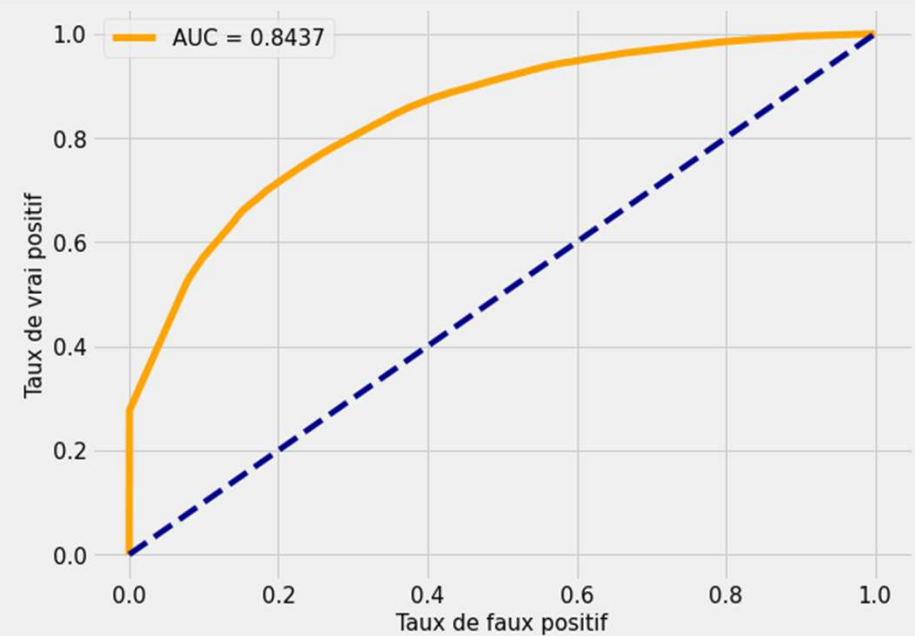
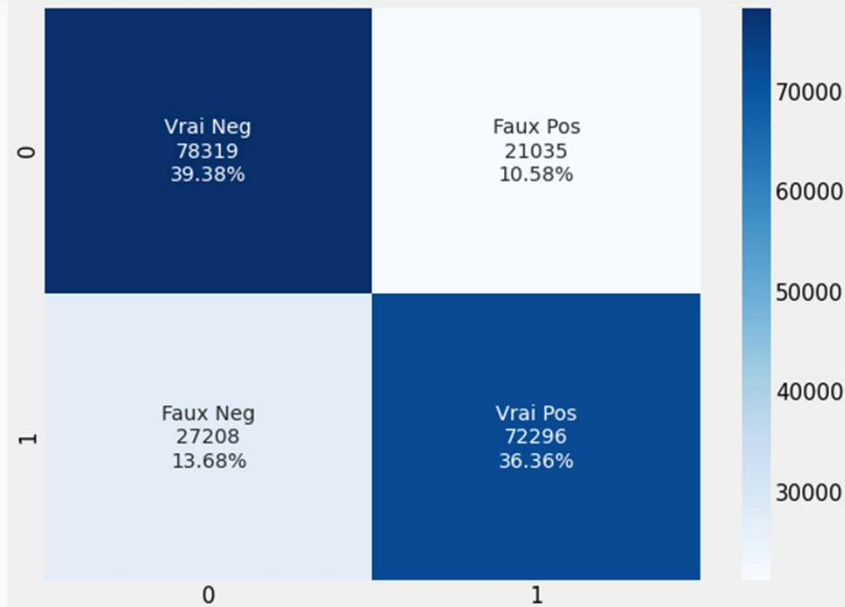


Elaboration d'un modèle, Optimisation et Compréhension

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



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



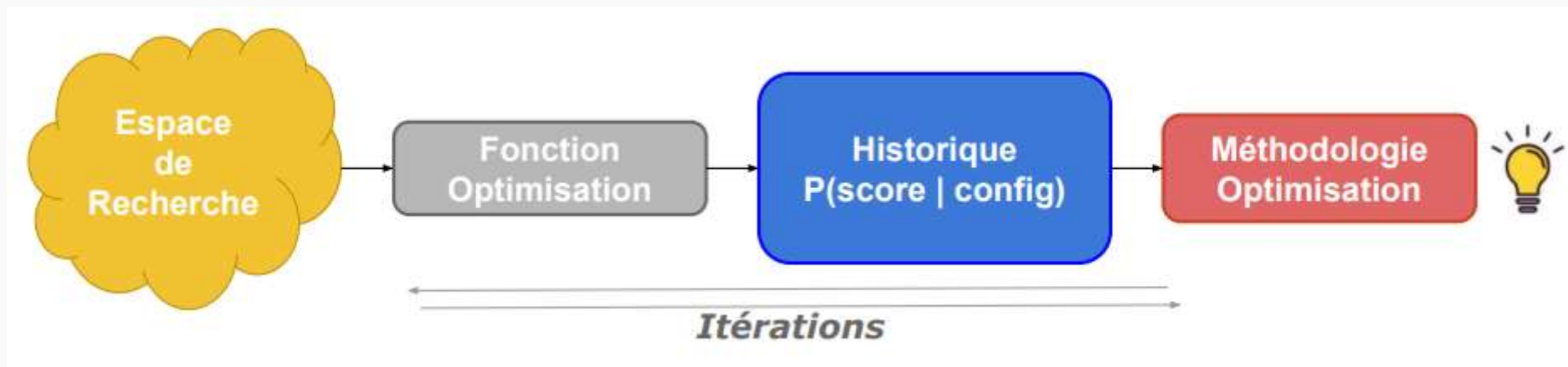
COMPARAISON ET SYNTHÈSE DES RESULTATS POUR LES MODELES UTILISES: 4/ Gradient Boosting

Le Gradient Boosting :

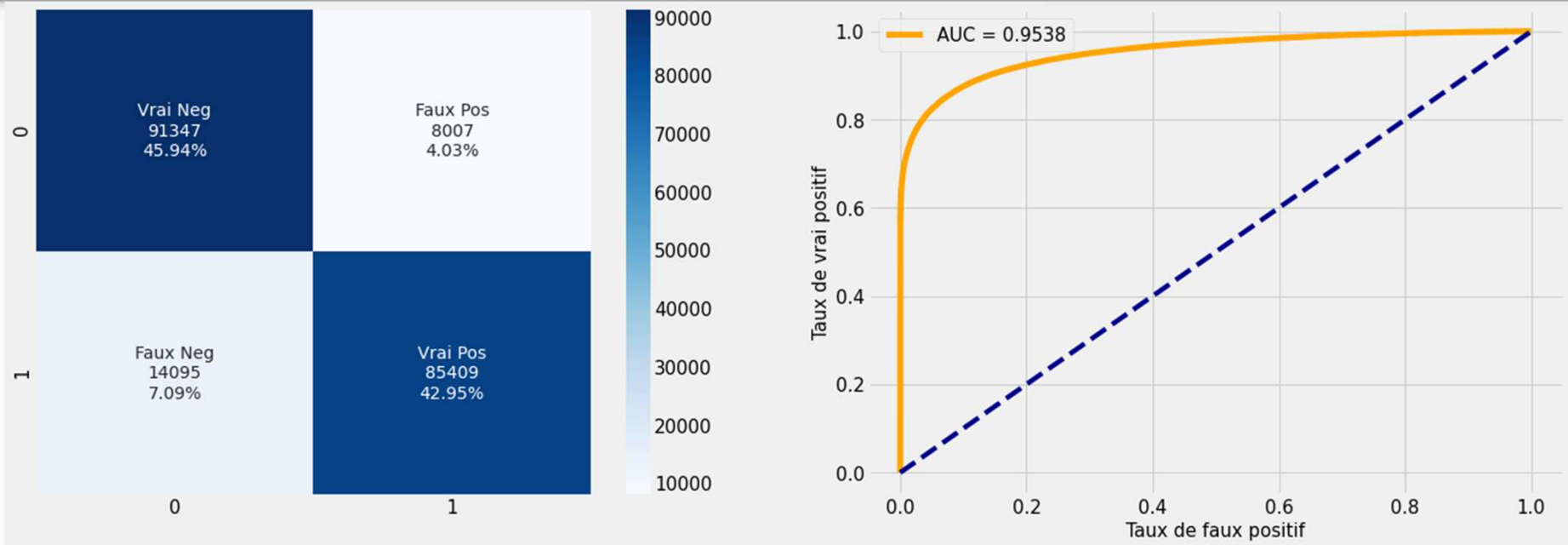
- ✓ Algorithme d'apprentissage supervisé,
- ✓ Combine les résultats d'un ensemble de modèle simple,
- ✓ Principe d'auto amélioration séquentielle.

	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

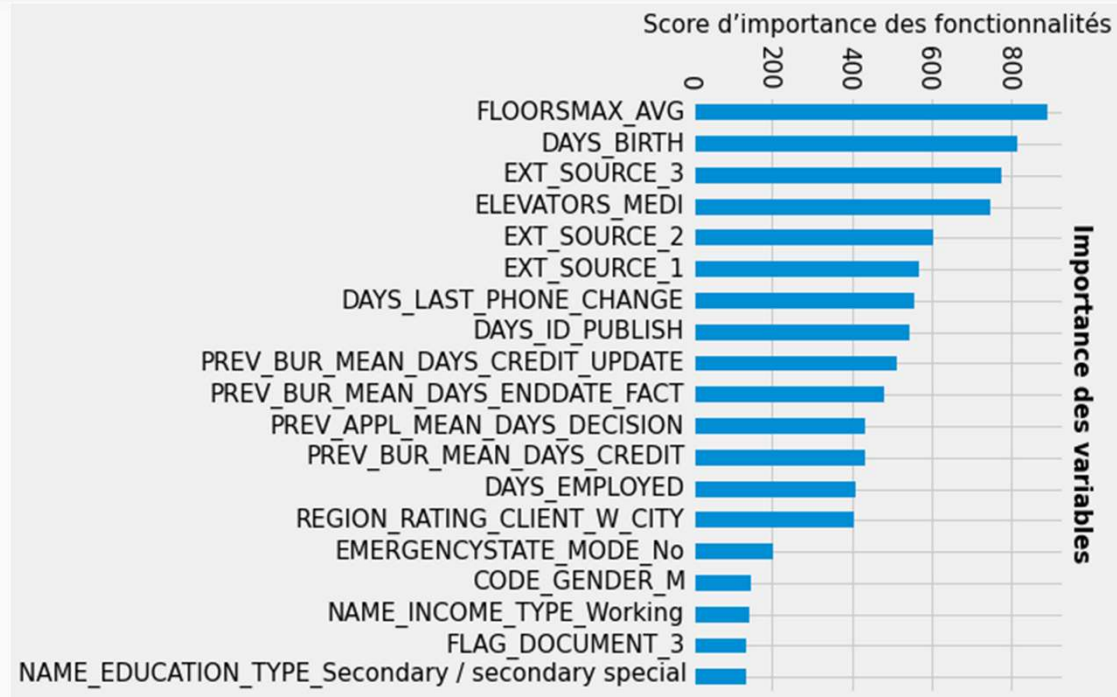


INTERPRETABILITE DU MODELE: 2/ Modèle LGBMClassifier Optimisé



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: $\text{fscore} = (1 + \text{beta}) * (tp / ((1 + 3) * tp + \text{beta} * fn + fp))$

	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,
- ✓ Entraîner mon jeu de données sur ces 25 variables,
- ✓ Réaliser une baseline optimisée faite sur deux algorithmes simple,
- ✓ Réaliser modélé 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