Prêt à dépenser

Scoring crédit

Presentation

Problématique : Défauts de paiement des credit => Pertes financières

Solution : Modèle de prédiction interprétable

Objectif : Réduction des risques

Données => Modélisation => Interprétation => Décision

Presentation Dataset

307511, 122

int64	41	} 102
float64	65	
object	16	

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	...
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	...
...	...	...	...	...	...	...	...	...	...	...	...
307506	456251	0	Cash loans	M	N	N	0	157500.0	254700.0	27558.0	...
307507	456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	12001.5	...
307508	456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	29979.0	...
307509	456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	20205.0	...
307510	456255	0	Cash loans	F	N	N	0	157500.0	675000.0	49117.5	...

307511 rows × 122 columns

Préparation data

kaggle

- EncodageVariables catégorielles 16

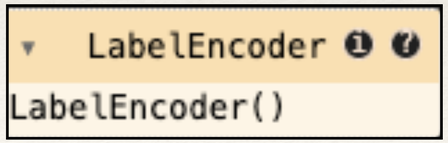
NAME_CONTRACT_TYPE 2

FLAG_OWN_CAR 2

FLAG_OWN_REALTY 2

One-Hot Encoding

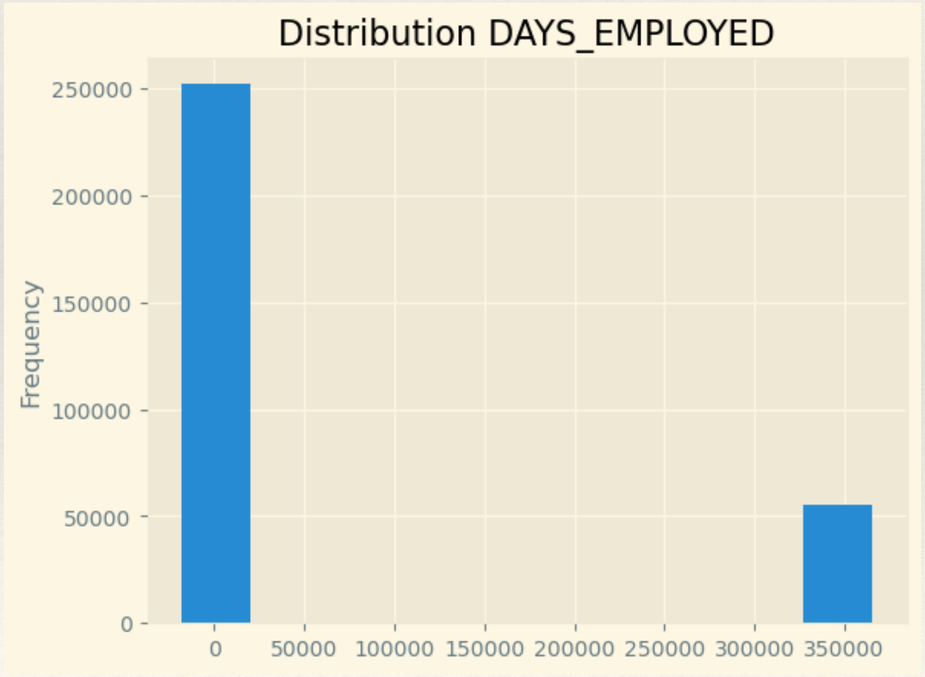
122 => 242

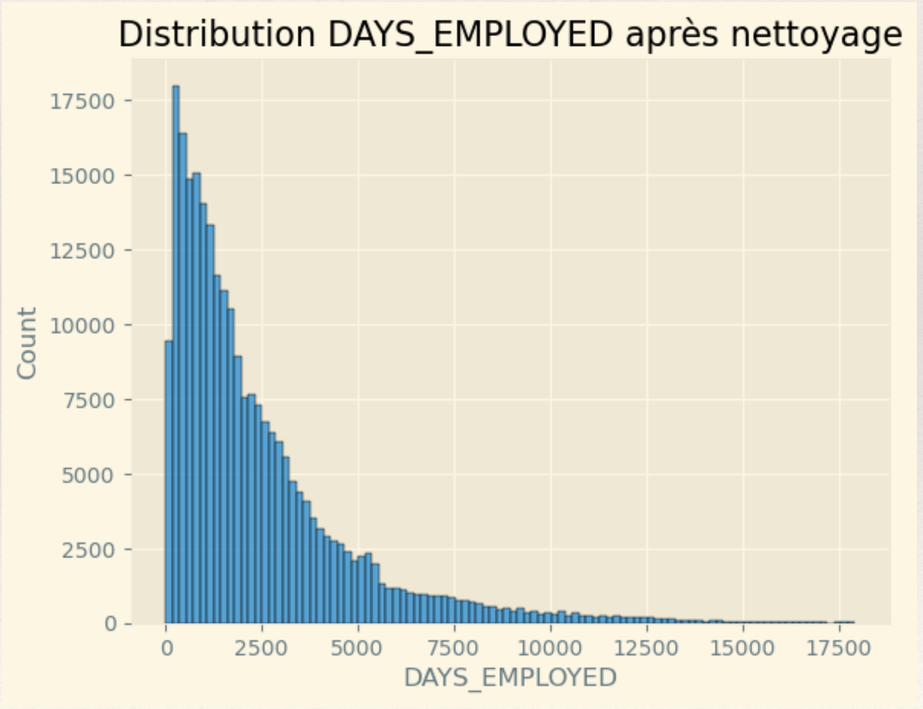


Préparation data

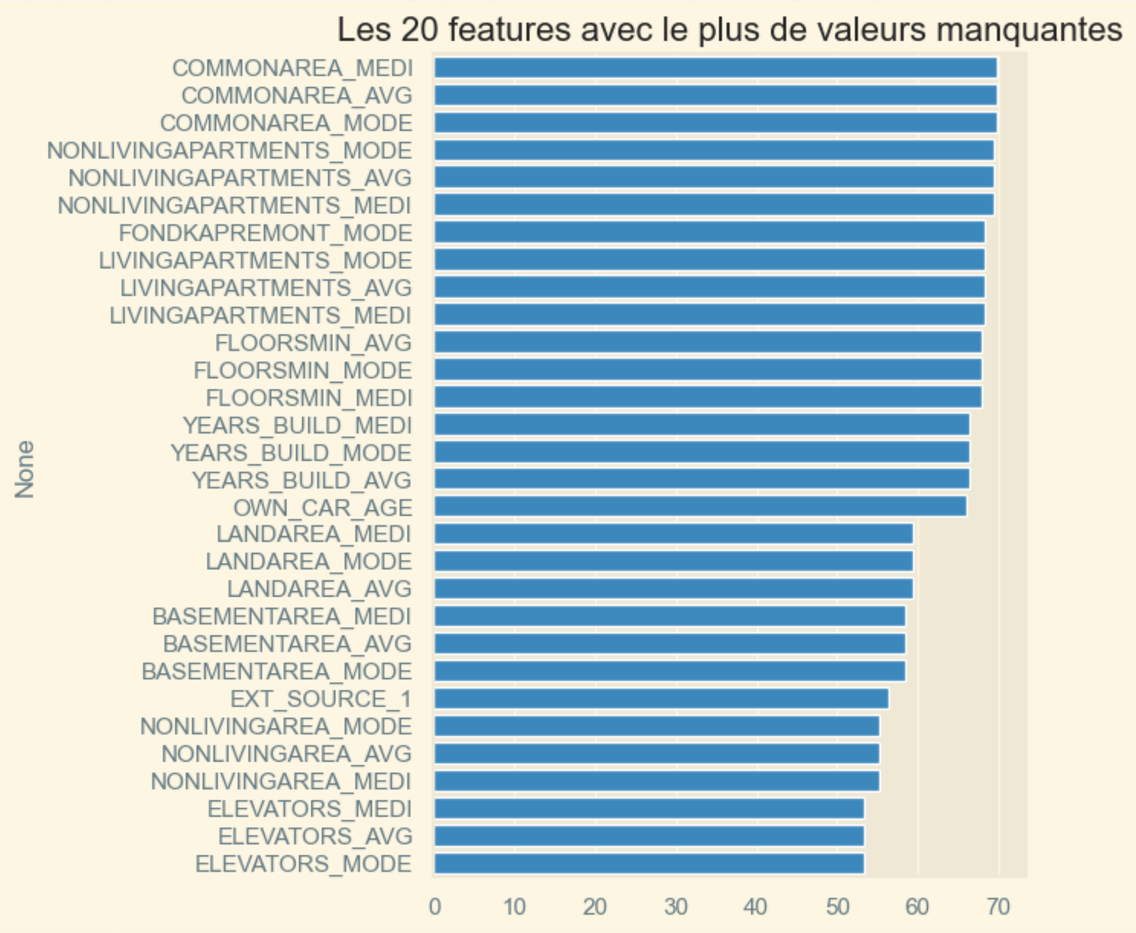
Outliers

```
count    307511.000000
mean      63815.045904
std      141275.766519
min      -17912.000000
25%       -2760.000000
50%       -1213.000000
75%        -289.000000
max      365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```





Imputation et normalisation



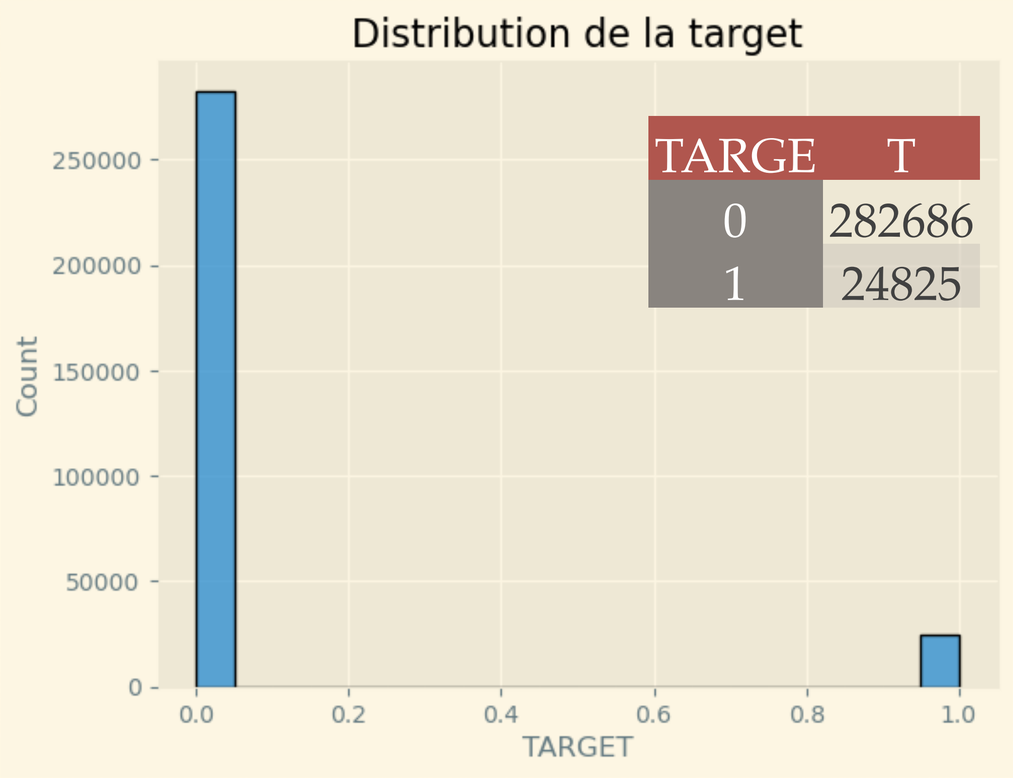

```
SK_ID_CURR                      0.0
OCCUPATION_TYPE_Managers        0.0
OCCUPATION_TYPE_Private service staff    0.0
OCCUPATION_TYPE_Realty agents   0.0
OCCUPATION_TYPE_Sales staff     0.0
                                ...
FLAG_DOCUMENT_5                 0.0
FLAG_DOCUMENT_6                 0.0
FLAG_DOCUMENT_7                 0.0
FLAG_DOCUMENT_8                 0.0
TARGET                          0.0
Length: 244, dtype: float64
```

```
SimpleImputer
SimpleImputer(strategy='median')
```

9.152.465 => 0

```
MinMaxScaler
MinMaxScaler()
```

Exploration de la Target



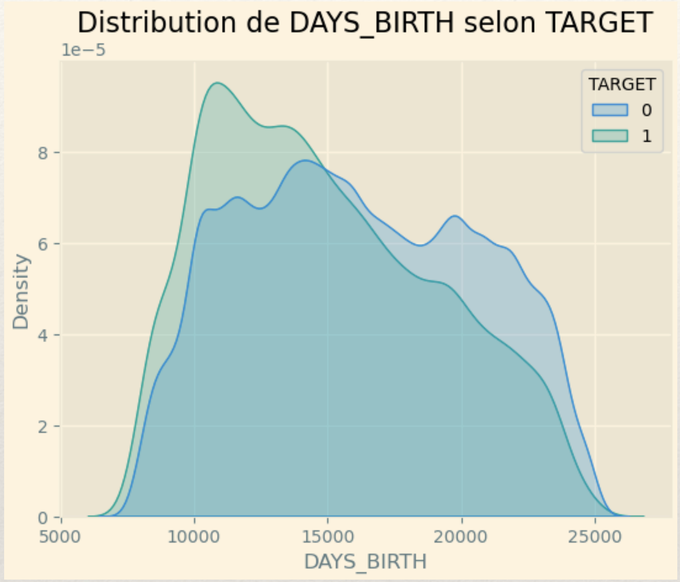
TARGET	T
0	282686
1	24825

=> Déséquilibre entre les classes

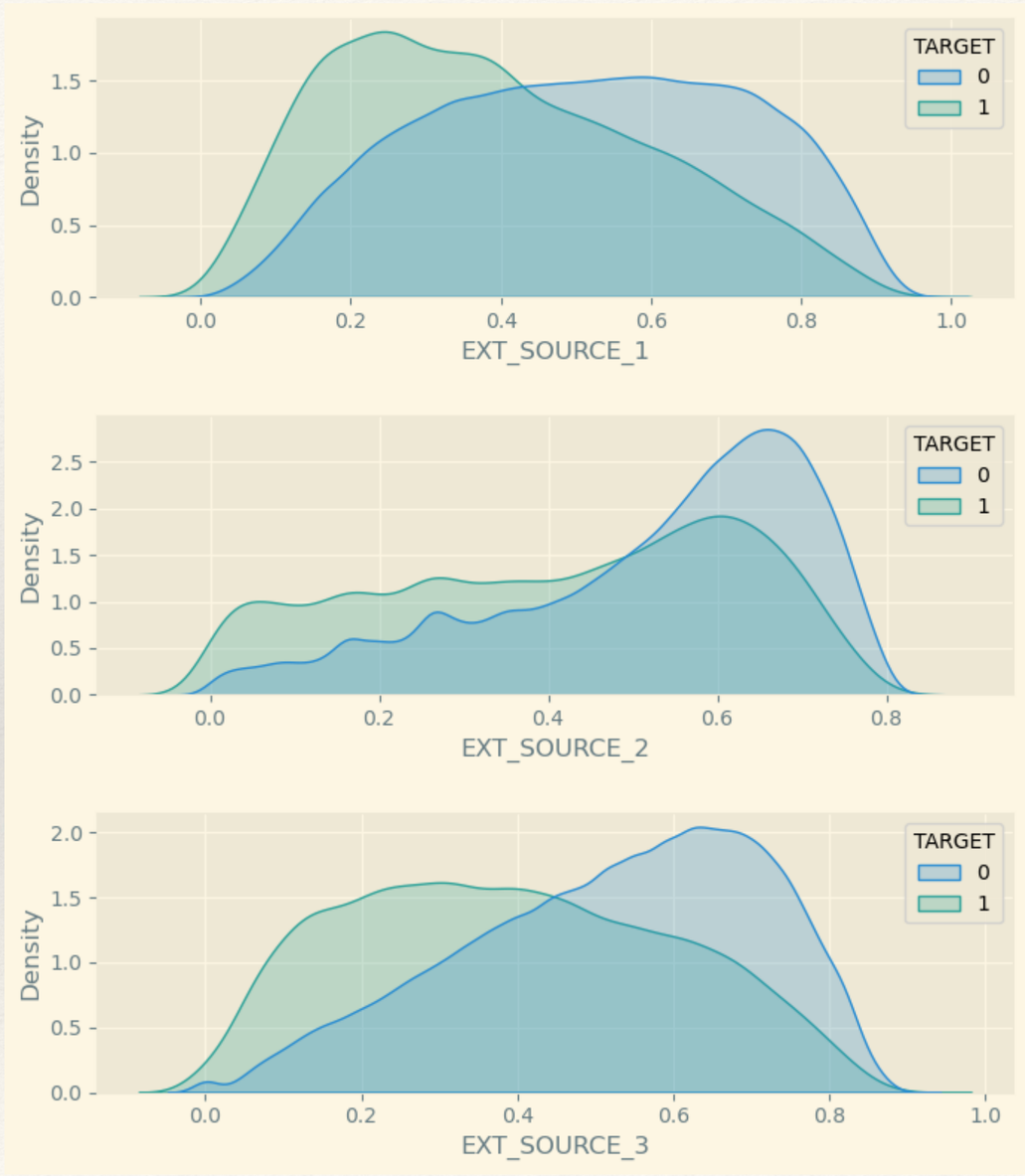
92% remboursent

8% défaut

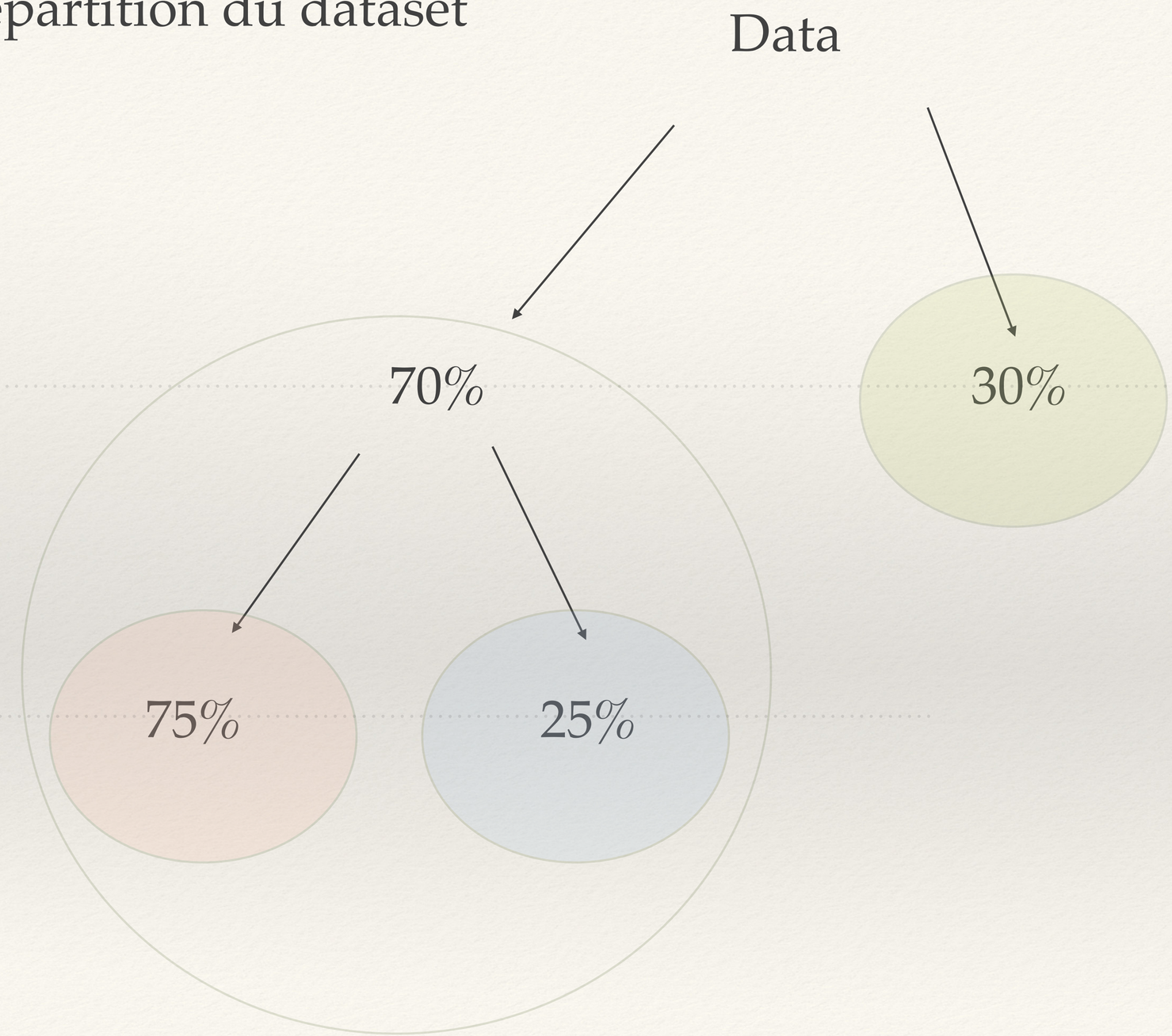
Corrélation avec TARGET



Corrélation : -0.0782393083098270 9



Répartition du dataset



Train => 52%

Validation => 18%

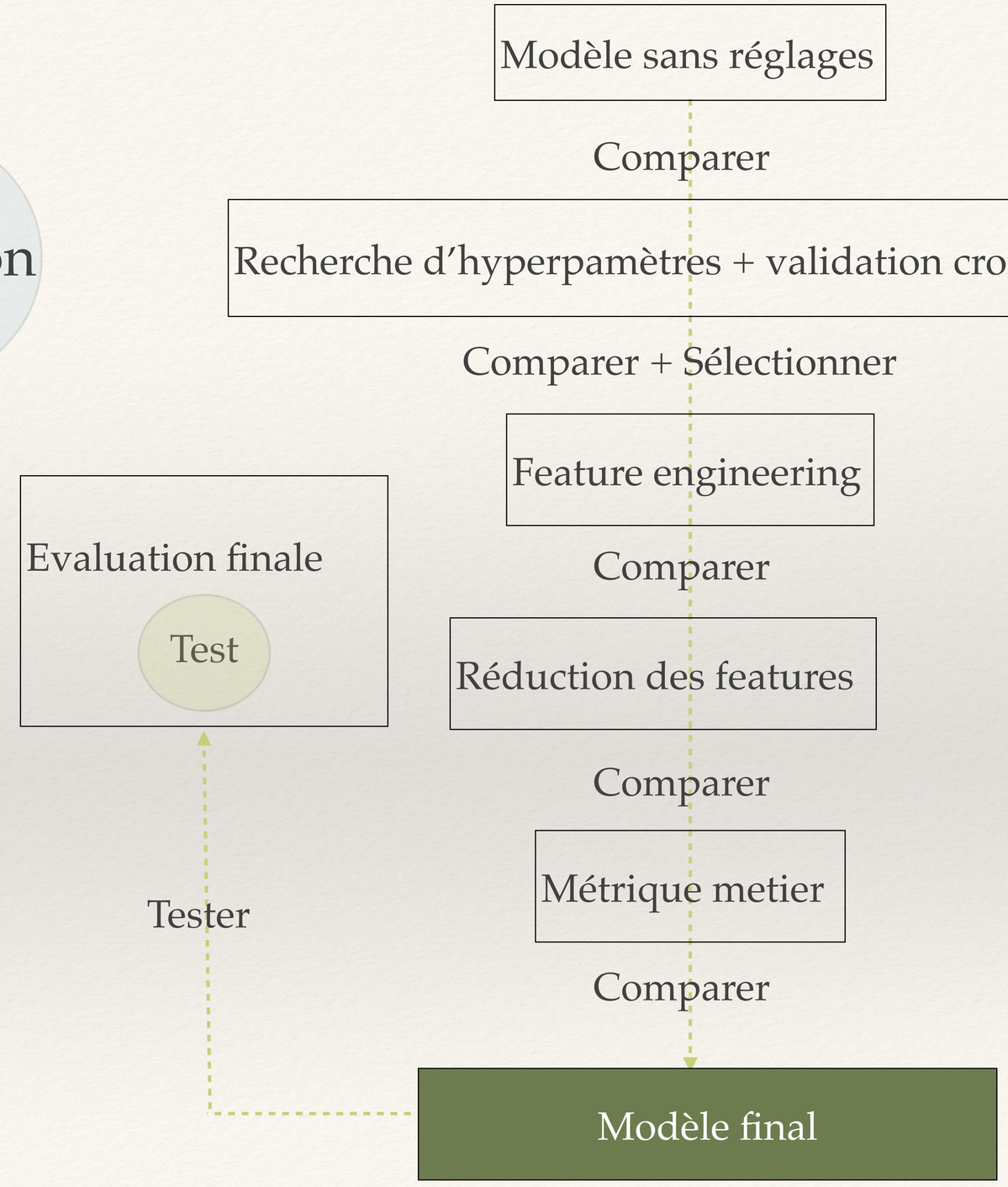
Test => 30%

Méthodologie

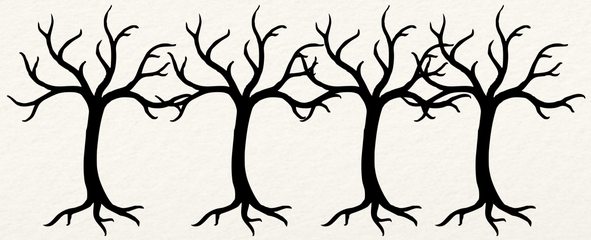
Dataset utilisé = Train + Validation

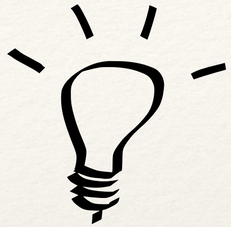
Evaluation

- accuracy
- recall
- precision
- f1_score
- roc_auc
- Confusion matrix
- Fbeta_score



Présentation des modèles





LogisticRegression	*DecisionTreeClassifier*	*RandomForestClassifier*	*LightGBM*
linéaire	*non_linéaire*	*Ensemble d'arbres*	*Gradient boosting*
Coefficient	*Règles hiérarchiques*	*Random*	*Fonction de perte*
Interprétable	*Interprétable*	*Bagging*	*Correction erreur*

Modèle de référence

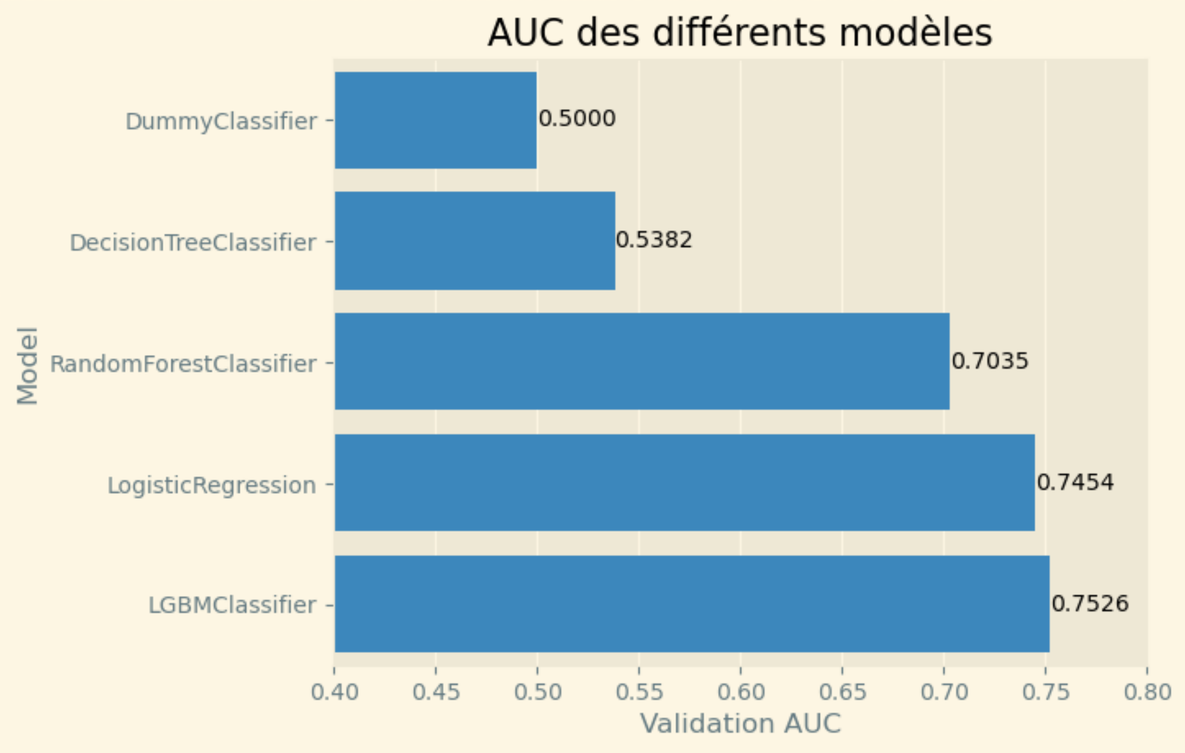
comparaison *DummyClassifier* strategy='uniform'

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
0	DummyClassifier	{'strategy': 'uniform'}	0.500155	0.491132	0.080899	0.138916	0.500000	0.5	NaN	[[24877 24540]\n [2238 2160]]	sans réglages

50/50

Performance sans réglages

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
0	DummyClassifier	{'strategy': 'uniform'}	0.500155	0.491132	0.080899	0.138916	0.500000	0.5	NaN	[[24877 24540]\n [2238 2160]]	sans réglages
1	LogisticRegression	{'random_state': 42}	0.919439	0.010232	0.500000	0.020053	0.746980	0.745403	NaN	[[49372 45]\n [4353 45]]	sans réglages
2	DecisionTreeClassifier	{'random_state': 42}	1.000000	0.163256	0.143428	0.152701	1.000000	0.538242	NaN	[[45129 4288]\n [3680 718]]	sans réglages
3	RandomForestClassifier	{'random_state': 42}	0.999988	0.000455	0.400000	0.000908	1.000000	0.703476	NaN	[[49414 3]\n [4396 2]]	sans réglages
4	LGBMClassifier	{'random_state': 42}	0.921142	0.016144	0.550388	0.031367	0.814207	0.752646	NaN	[[49359 58]\n [4327 71]]	sans réglages

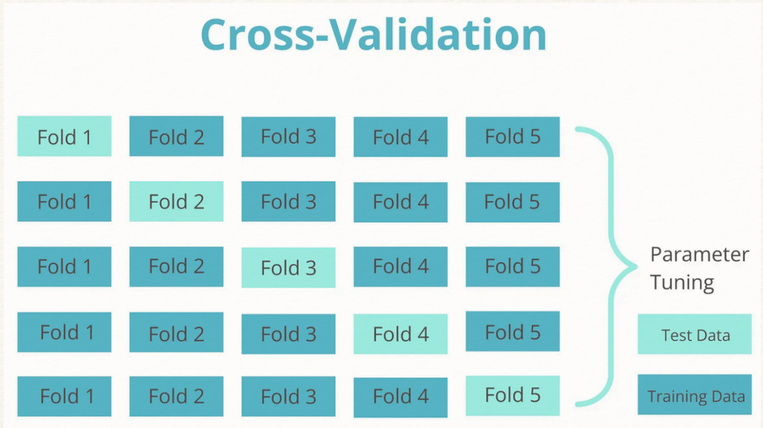


TN	FP
FN	TP

	TN	FP	FN	TP
LogisticRegression	49372	45	4353	45
DecisionTreeClassifier	45129	4288	3680	718
RandomForestClassifier	49414	3	4396	2
LGBMClassifier	49359	58	4327	71

Recherche d'hyperparamètres

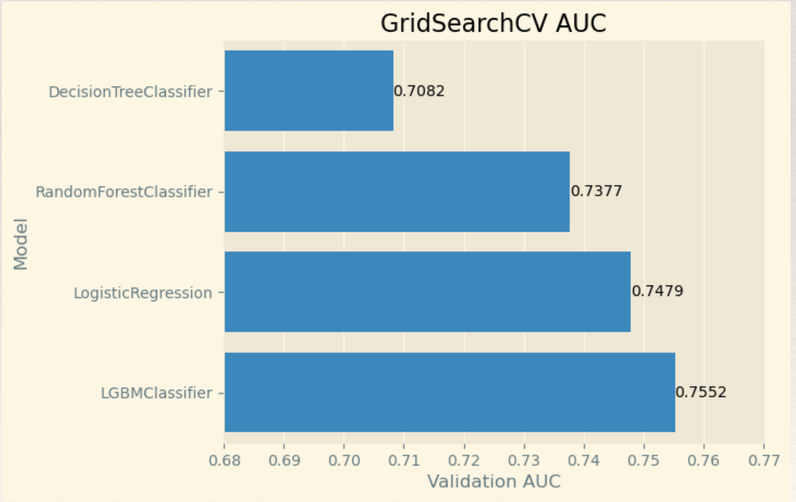
GridsearchCV



TN	FP
FN	TP

	TN	FP	FN	TP
LogisticRegression	49367	50	4346	52
DecisionTreeClassifier	49240	177	4328	70
RandomForestClassifier	49412	5	4390	8
LGBMClassifier	49369	48	4349	49

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
5	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.918313	0.011824	0.509804	0.023111	0.749472	0.747904	0.744518	[[49367, 50], [4346, 52]]	gridsearch
6	DecisionTreeClassifier	{'max_depth': 8, 'min_samples_leaf': 2, 'rando...	0.916287	0.015916	0.283401	0.030140	0.738189	0.708186	0.708735	[[49240, 177], [4328, 70]]	gridsearch
7	RandomForestClassifier	{'max_depth': 10, 'max_features': 0.3, 'n_esti...	0.918331	0.001819	0.615385	0.003627	0.814666	0.737734	0.736693	[[49412, 5], [4390, 8]]	gridsearch
8	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.918294	0.011141	0.505155	0.021802	0.794170	0.755158	0.753647	[[49369, 48], [4349, 49]]	gridsearch



LogisticRegression 'C': 10.0, 'penalty': 'l1', 'random_state': 42, 'solver': 'liblinear'

DecisionTreeClassifier max_depth': 8, 'min_samples_leaf': 2, 'random_state': 42

RandomForestClassifier 'max_depth': 10, 'max_features': 0.3, 'n_estimators': 200, 'random_state': 42

LightGBM learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100, 'random_state': 42, 'reg_alpha': 0.1, 'reg_lambda': 1

Déséquilibre des classes

class_weight

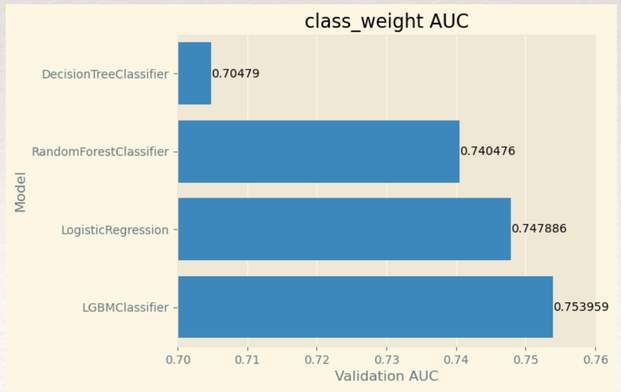
TN	FP
FN	TP

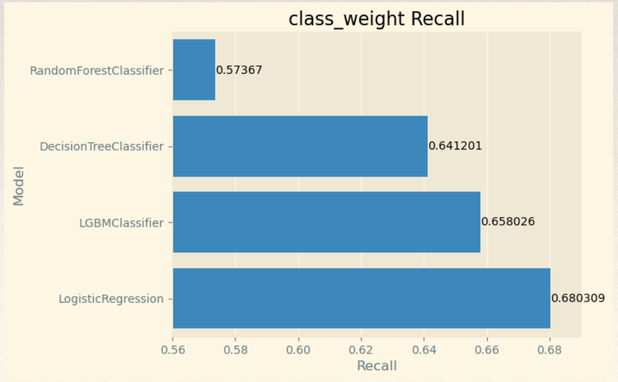
	TN	FP	FN	TP
LogisticRegression	34121	15296	1406	2992
DecisionTreeClassifier	33326	16091	1578	2820
RandomForestClassifier	37942	11475	1875	2523
LGBMClassifier	35360	14057	1504	2894

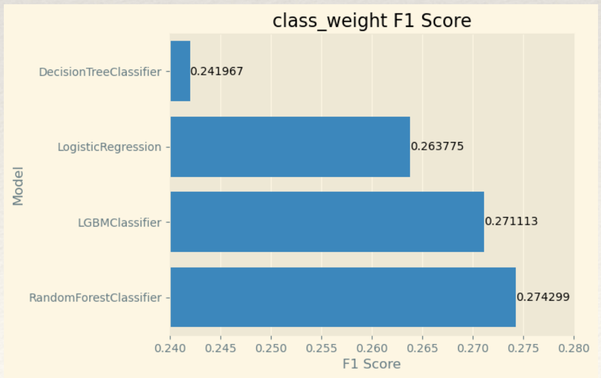
FN	TP
4346	52
4328	70
4390	8
4349	49

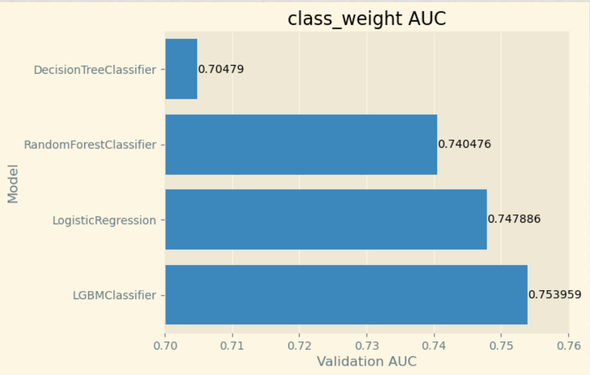
	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
5	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.918313	0.011824	0.509804	0.023111	0.749472	0.747904	0.744518	[[49367, 50], [4346, 52]]	gridsearch
6	DecisionTreeClassifier	{'max_depth': 8, 'min_samples_leaf': 2, 'rando...	0.916287	0.015916	0.283401	0.030140	0.738189	0.708186	0.708735	[[49240, 177], [4328, 70]]	gridsearch
7	RandomForestClassifier	{'max_depth': 10, 'max_features': 0.3, 'n_esti...	0.918331	0.001819	0.615385	0.003627	0.814666	0.737734	0.736693	[[49412, 5], [4390, 8]]	gridsearch
8	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.918294	0.011141	0.505155	0.021802	0.794170	0.755158	0.753647	[[49369, 48], [4349, 49]]	gridsearch

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
9	★ LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.688706	0.680309	0.163605	0.263775	0.750107	0.747886	NaN	[[34121 15296]\n [1406 2992]]	class_weight
10	DecisionTreeClassifier	{'max_depth': 8, 'min_samples_leaf': 2, 'rando...	0.678875	0.641201	0.149120	0.241967	0.751153	0.704790	NaN	[[33326 16091]\n [1578 2820]]	class_weight
11	RandomForestClassifier	{'max_depth': 10, 'max_features': 0.3, 'n_esti...	0.768493	0.573670	0.180240	0.274299	0.828263	0.740476	NaN	[[37942 11475]\n [1875 2523]]	class_weight
12	★ LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.717731	0.658026	0.170727	0.271113	0.796659	0.753959	NaN	[[35360 14057]\n [1504 2894]]	class_weight

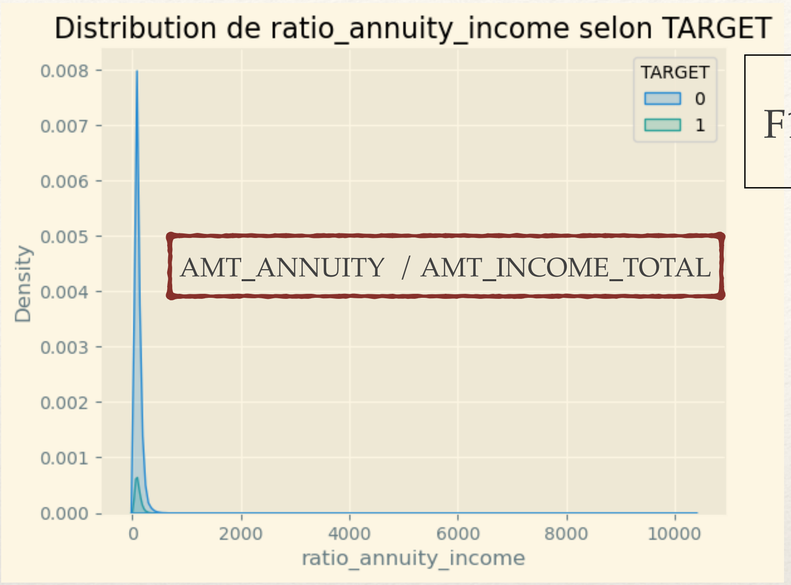




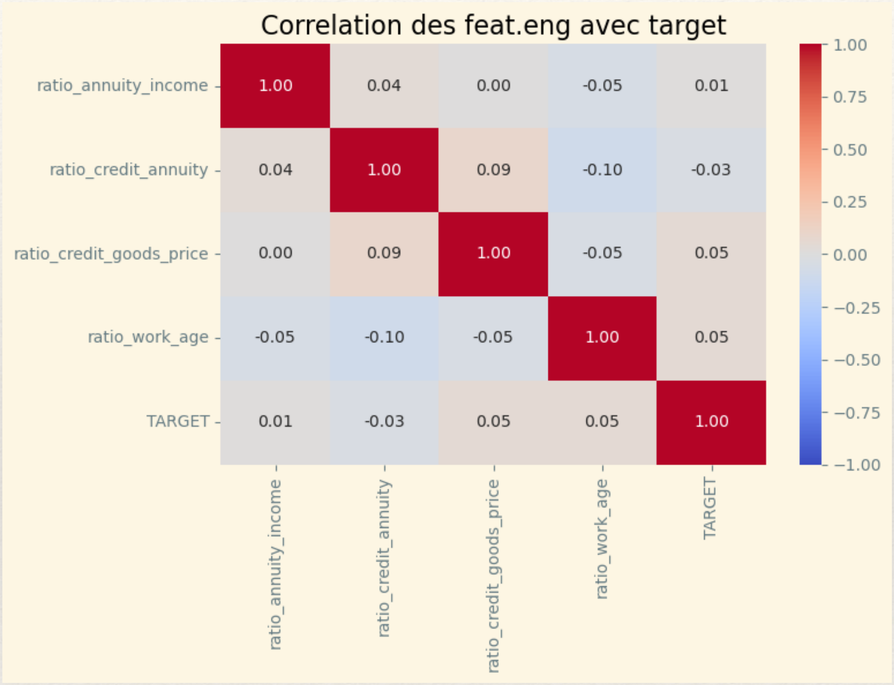




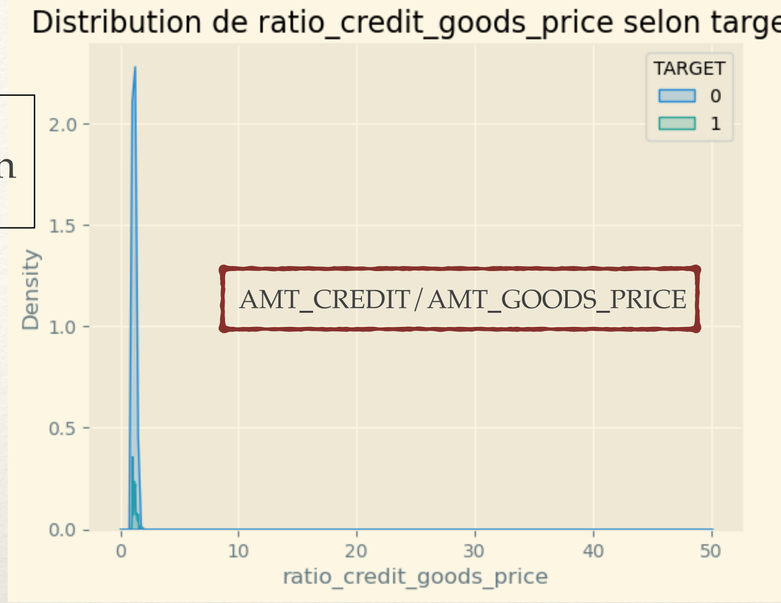
Features engineering

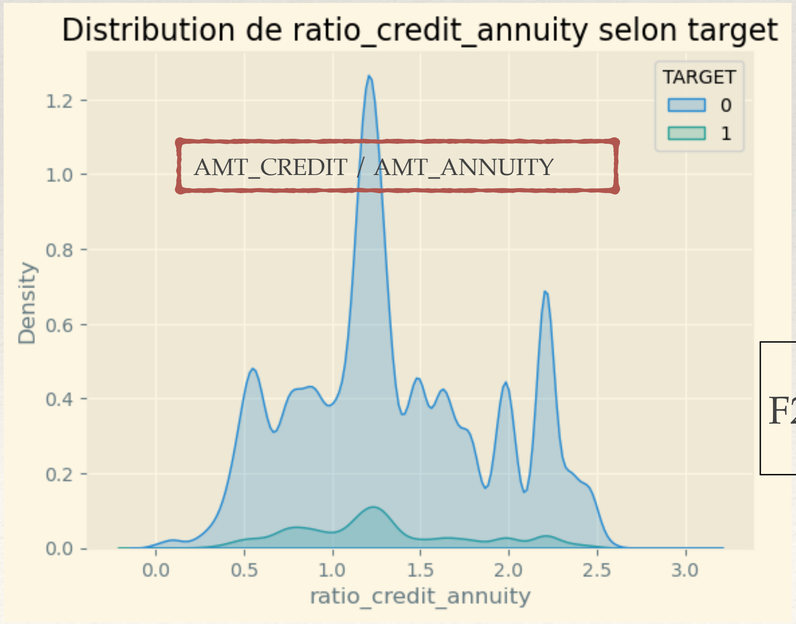


F1 mensualité / revenu



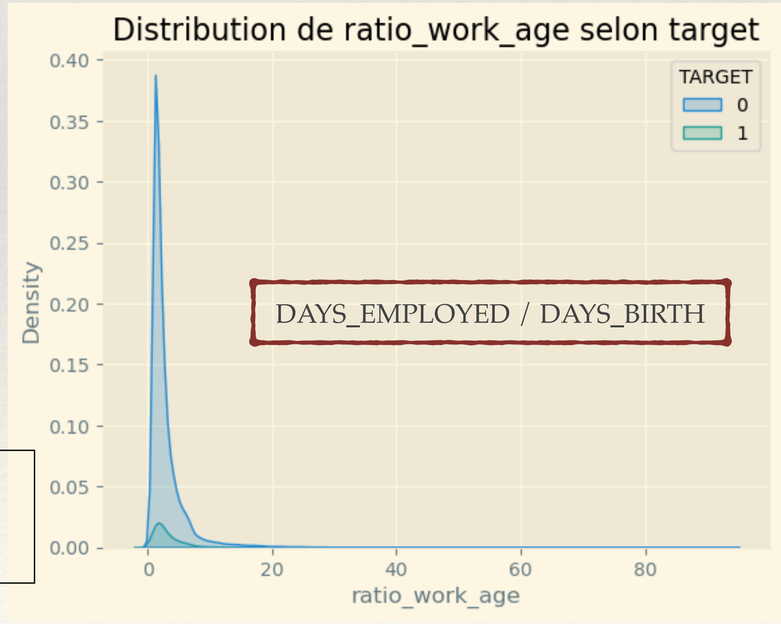
F3 credit / valeur du bien





F2 credit total / mensualité

F3 durée travaillée / age



Performance feature engineering

LogisticRegression

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
9	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.688706	0.680309	0.163605	0.263775	0.750107	0.747886	NaN	[[34121 15296]\n [1406 2992]]	class_weight
14	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.689418	0.676671	0.163463	0.263316	0.750677	0.748337	NaN	[[34187 15230]\n [1422 2976]]	avec feature engineering

Sans feature ratio_work_age

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
16	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.689381	0.675762	0.163019	0.262672	0.750593	0.747972	NaN	[[34158 15259]\n [1426 2972]]	feat.engineering - ratio_work_age

=> Diminution de toutes les métriques

LightGBM

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
12	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.717731	0.658026	0.170727	0.271113	0.796659	0.753959	NaN	[[35360 14057]\n [1504 2894]]	class_weight
15	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.721107	0.660073	0.172695	0.273765	0.801247	0.757241	NaN	[[35510 13907]\n [1495 2903]]	avec feature engineering

LogisticRegression => Augmentation de l'AUC avec une légère diminution de autres métriques (sauf accuracy)

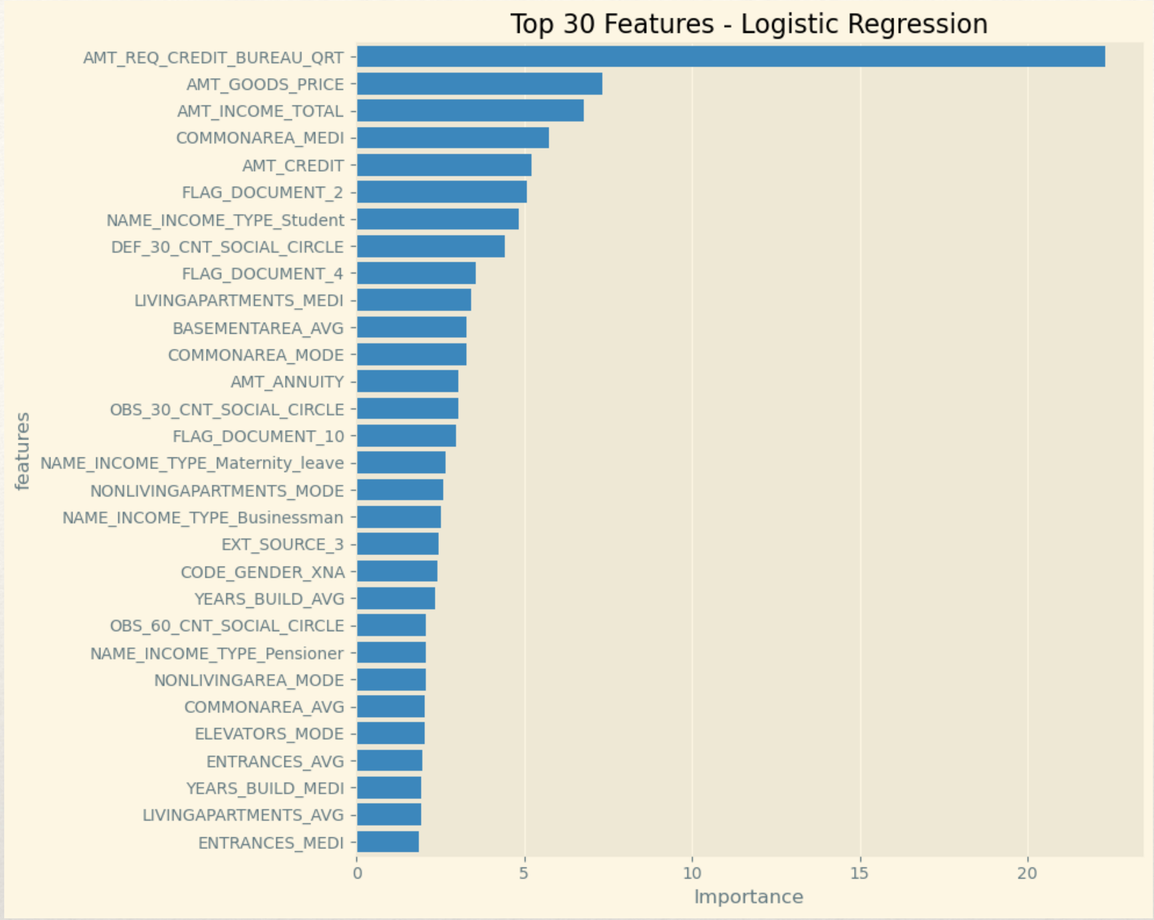
LightGBM => Augmentation de toutes les métriques ciblées

Réduction des features

LogisticRegression

247 => 124

ratio_credit_goods_price



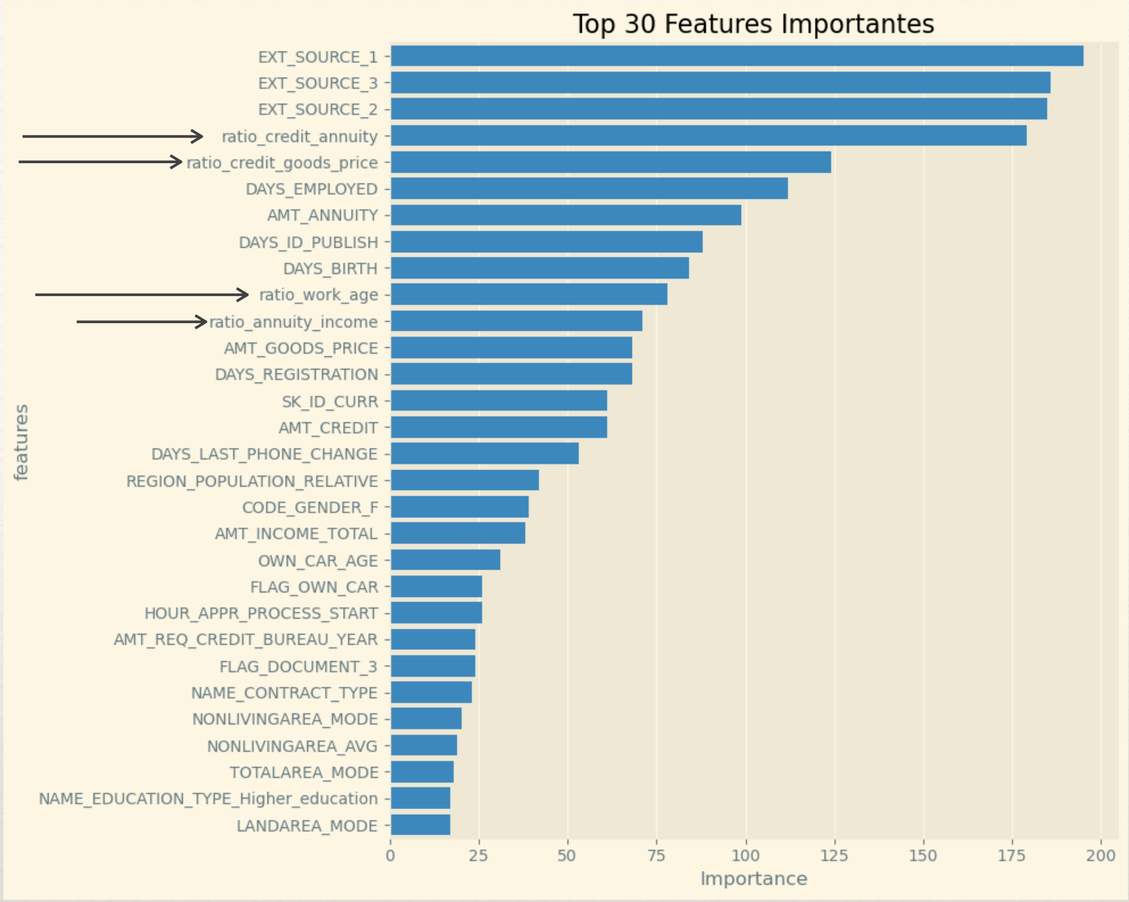
	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
9	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.688706	0.680309	0.163605	0.263775	0.750107	0.747886	NaN	[[34121 15296]\n [1406 2992]]	class_weight
14	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.689418	0.676671	0.163463	0.263316	0.750677	0.748337	NaN	[[34187 15230]\n [1422 2976]]	avec feature engineering
17	LogisticRegression	{'C': 10.0, 'penalty': 'l1', 'random_state': 4...	0.688297	0.675307	0.162021	0.261340	0.746381	0.743878	NaN	[[34056 15361]\n [1428 2970]]	feat.engin+reduc

=> Légère diminution des performances avec

Réduction des features

LightGBM

247=>134

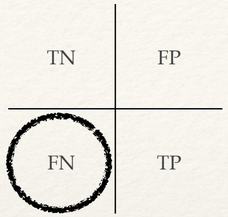


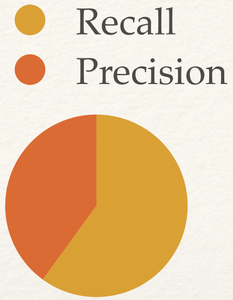
	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode
12	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.717731	0.658026	0.170727	0.271113	0.796659	0.753959	NaN	[[35360 14057]\n [1504 2894]]	class_weight
15	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.721107	0.660073	0.172695	0.273765	0.801247	0.757241	NaN	[[35510 13907]\n [1495 2903]]	avec feature engineering
18	LGBMClassifier	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...	0.721107	0.660073	0.172695	0.273765	0.801247	0.757241	NaN	[[35510 13907]\n [1495 2903]]	feat.engineering + feat.reduits

=> Performance inchangée

Métrique metier

F-Beta





$$F_\beta = \frac{(1+\beta^2)\text{tp}}{(1+\beta^2)\text{tp} + \text{fp} + \beta^2\text{fn}}$$

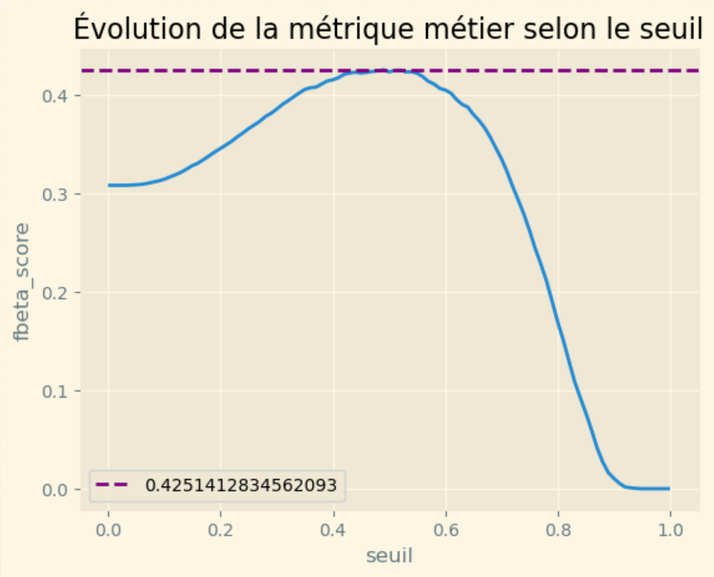
19	LogisticRegression	{'C': 2.1544346900318834, 'class_weight': 'bal...	0.687857	0.672806	0.161535	0.260521	0.746292	0.743720	0.410427	[[34058 15359]\n [1439 2959]]	Fbeta	0.4120022277917015
20	★ LGBMClassifier	{'class_weight': 'balanced', 'learning_rate': ...	0.717830	0.657344	0.174482	0.275767	0.812099	0.760289	0.421511	[[35739 13678]\n [1507 2891]]	Fbeta	0.4231433506044905

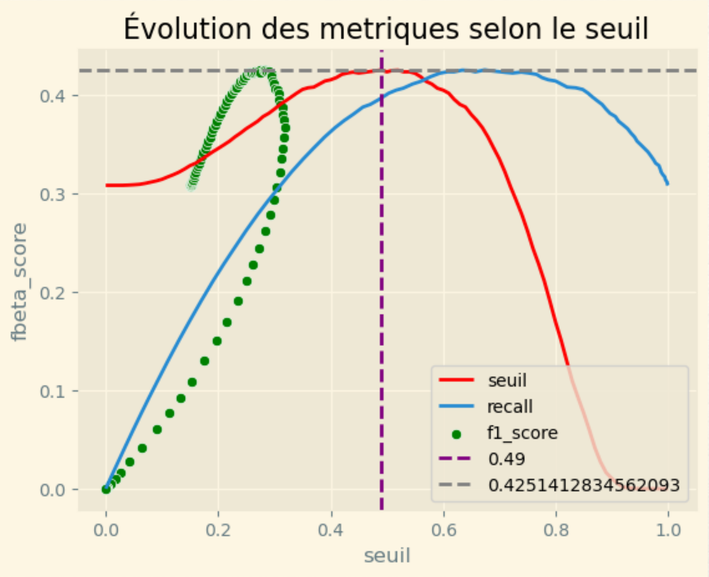
LogisticRegression 'C': 2.1544346900318834, 'class_weight': 'balanced', 'max_iter': 1000, 'penalty': 'l1', 'random_state': 42, 'solver': 'liblinear'

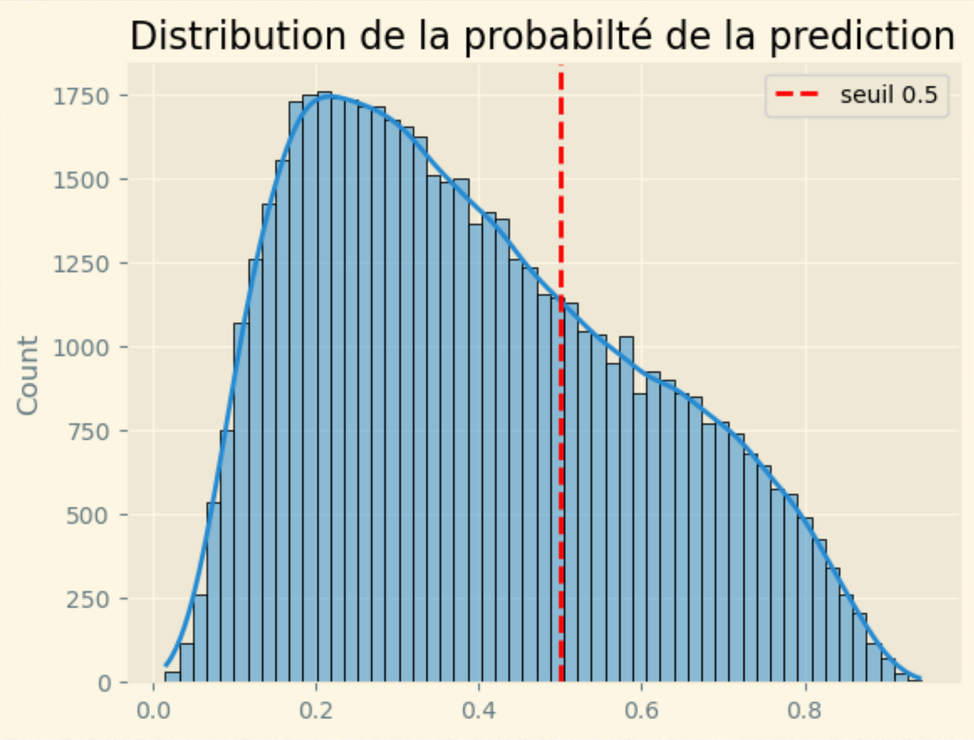
LightGBM 'class_weight': 'balanced', 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'random_state': 42, 'reg_alpha': 1, 'reg_lambda': 0.1

Métrique metier

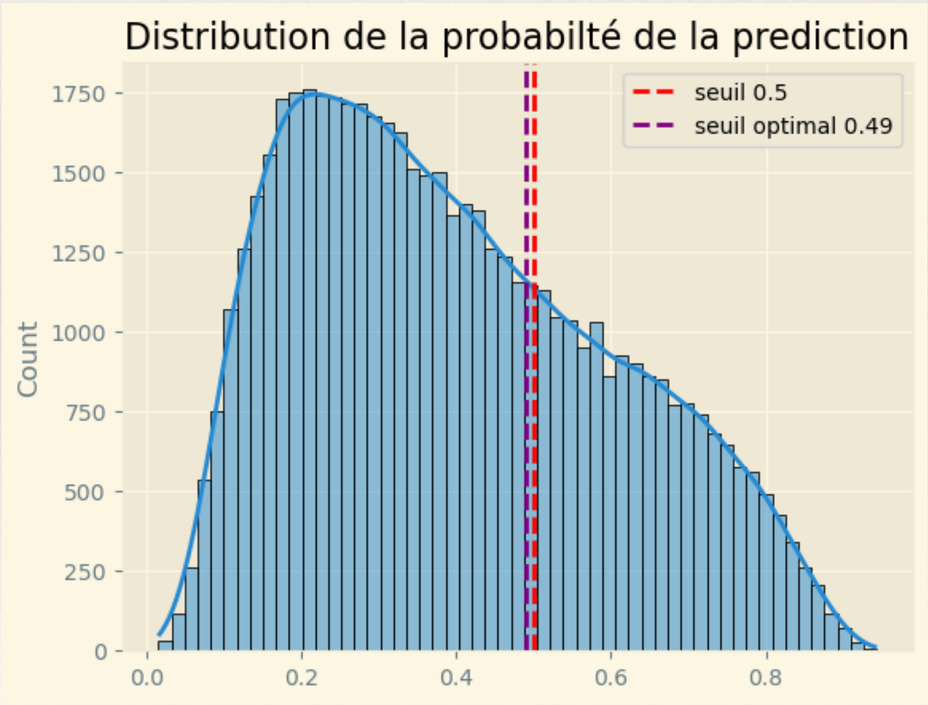
Seuil de décision





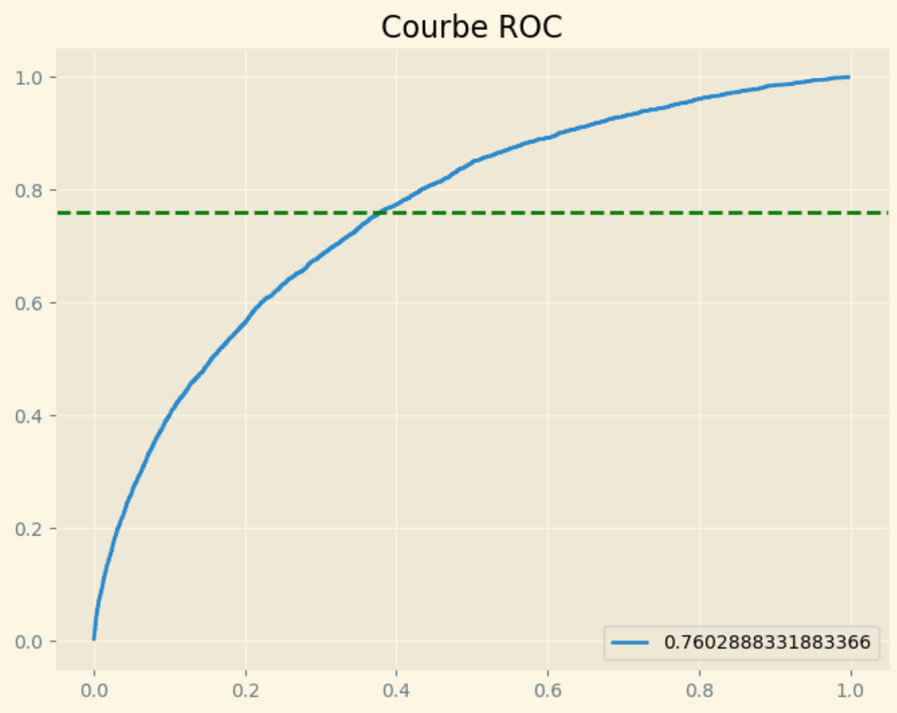



```
seuil          0.490000
fbeta_score    0.425141
recall         0.673943
f1_score       0.273621
Name: 49, dtype: float64
```



Performance

	Model	Best Params	Accuracy	Recall	Precision	F1 Score	Train AUC	Validation AUC	CV ROC AUC	Matrice de confusion	Méthode	
21	LGBMClassifier	{'boosting_type': 'gbdt', 'class_weight': 'bal...	0.707572	0.673943	0.171657	0.273621	0.812099	0.760289	NaN	[[35114, 14303], [1434, 2964]]	Application du seuil 0.49	meilleur fbeta 0.4251412834562093



35114	14303
★ 1434	2964

Performance sur le dataset test

LightGBM 'class_weight': 'balanced', 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 100, 'random_state': 42, 'reg_alpha':1, 'reg_lambda': 0.1

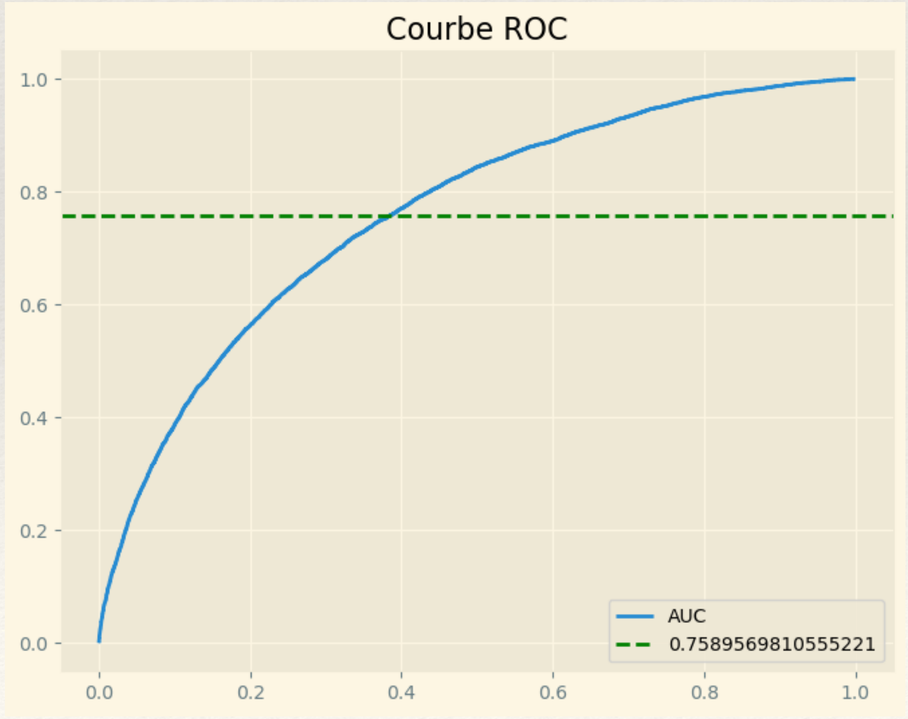
Features engineering

134 features

seuil_optimal = 0.49

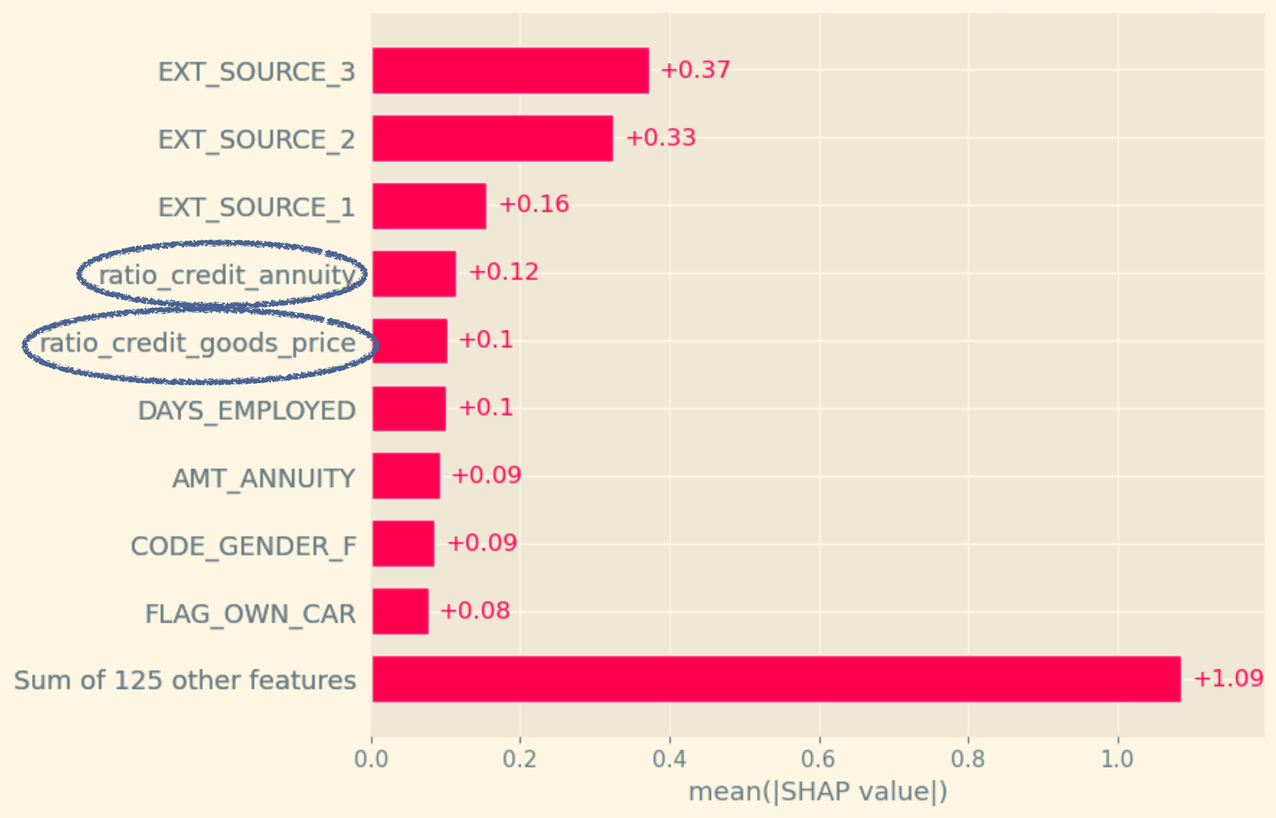
Data ⟶ Test 30%

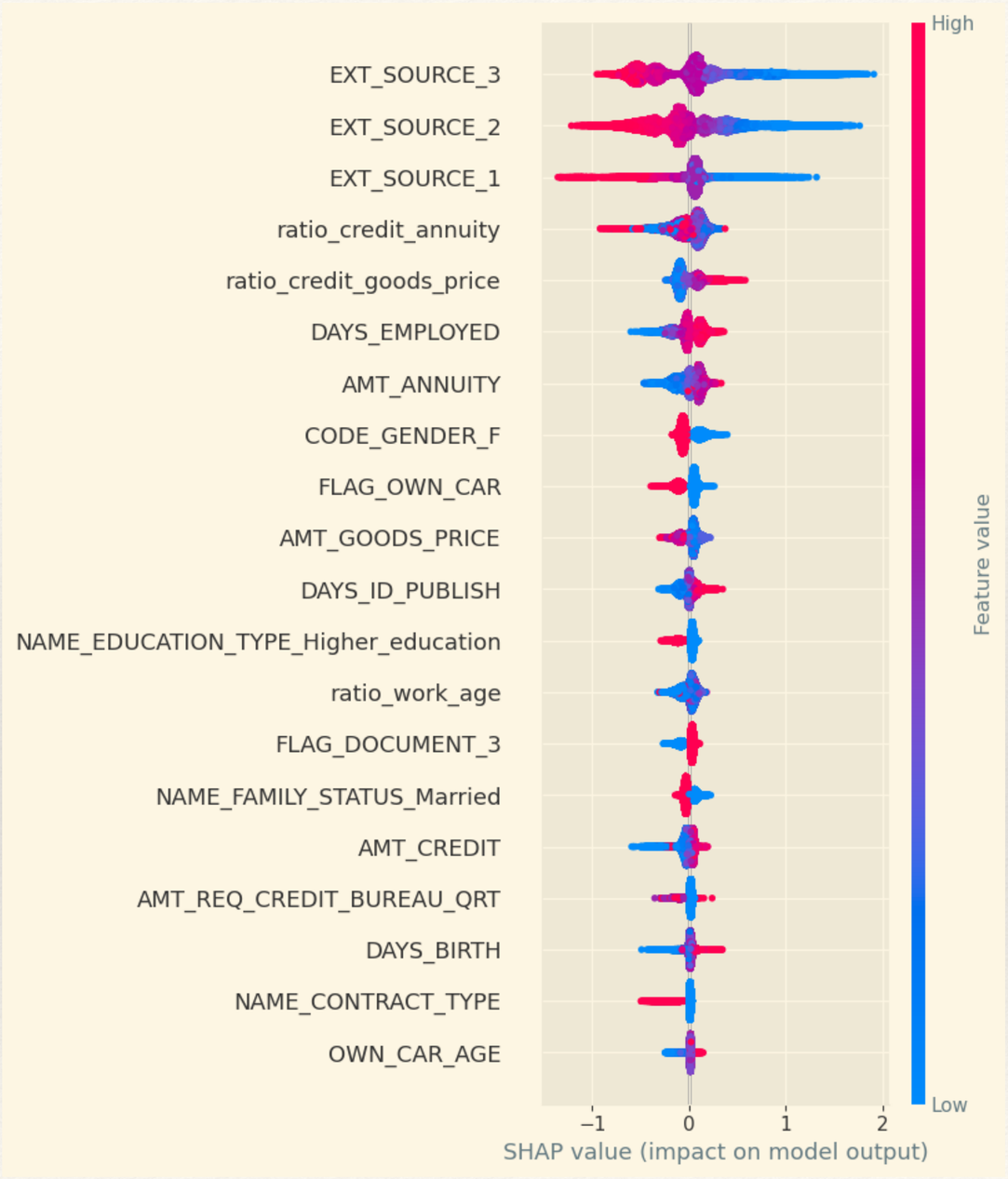
	Model	Best Params	Accuracy	Recall	Precision	F1 Score	AUC	Matrice de confusion	F-beta Score
0	LGBMClassifier	{'boosting_type': 'gbdt', 'class_weight': 'bal...	0.70501	0.673546	0.167623	0.268441	0.758957	[[60047, 24794], [2420, 4993]]	0.42001



Interprétation globale

Feature importance



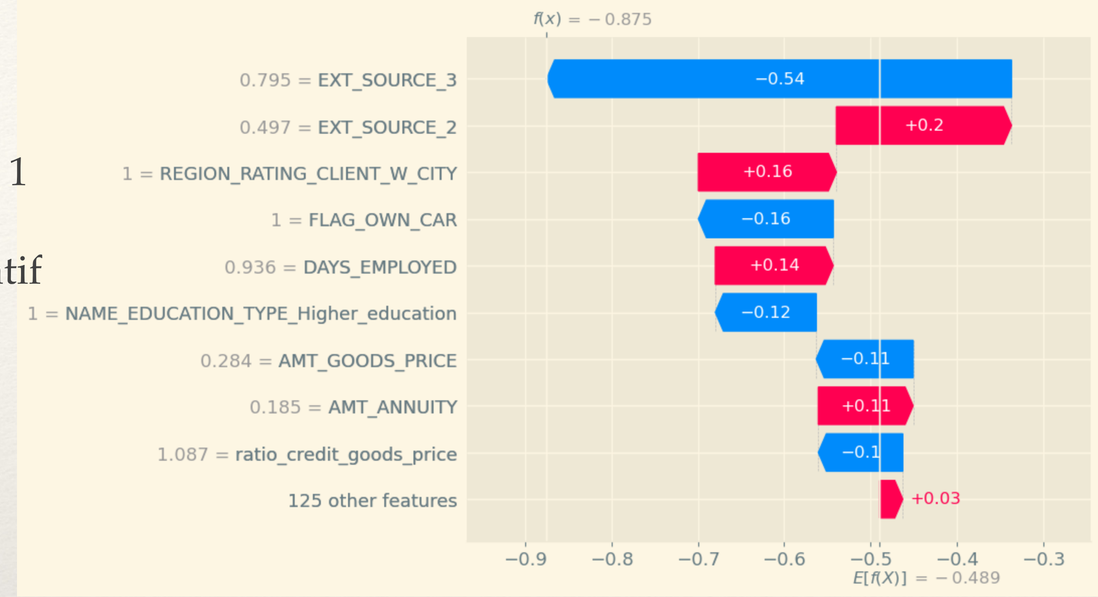


Interprétation locale

Individu 1

True negatif

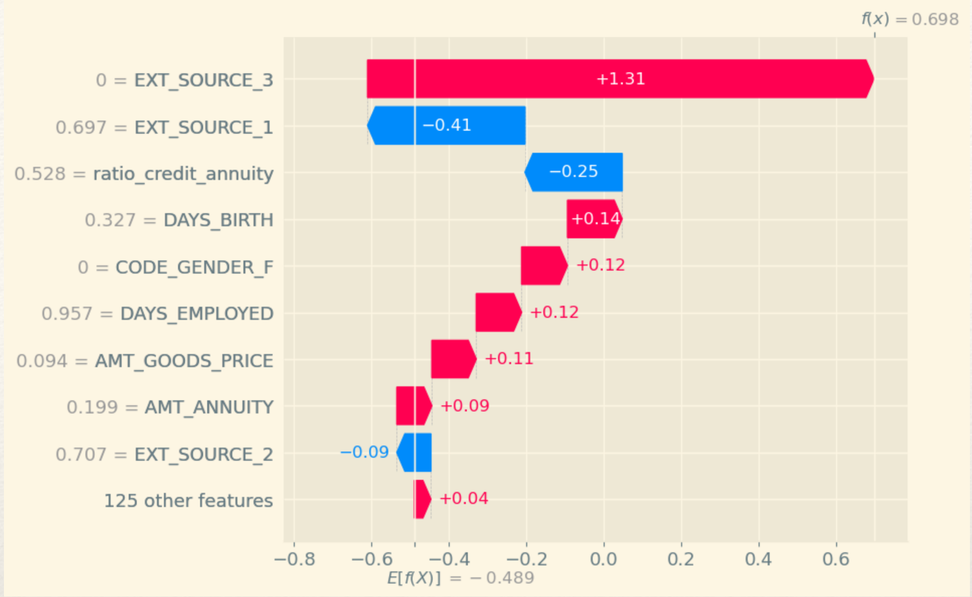
[0,0]



Individu 0

False positif

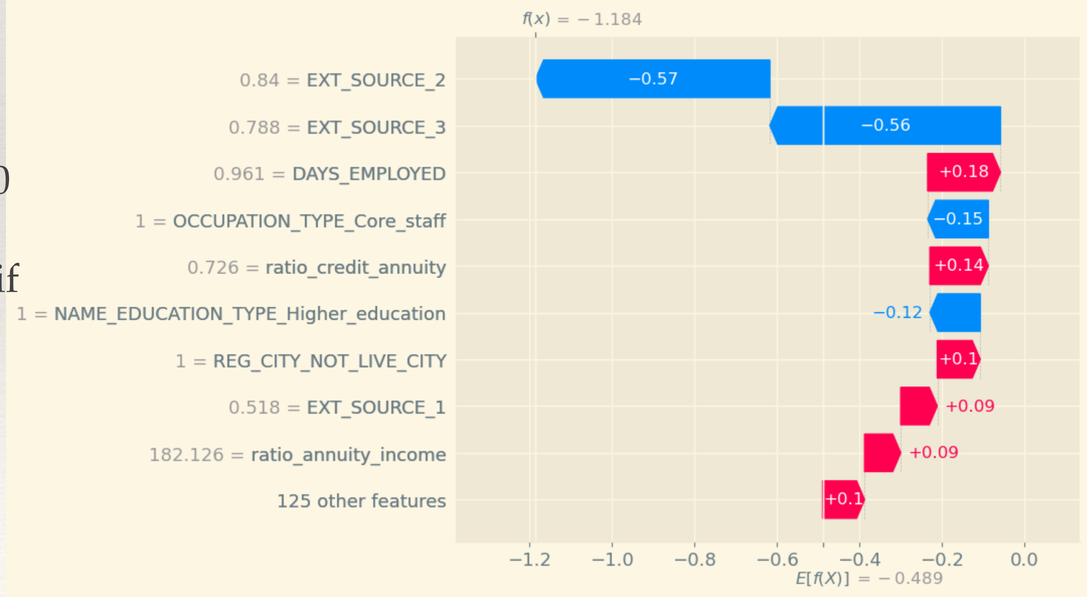
[0,1]



Individu 40

False negatif

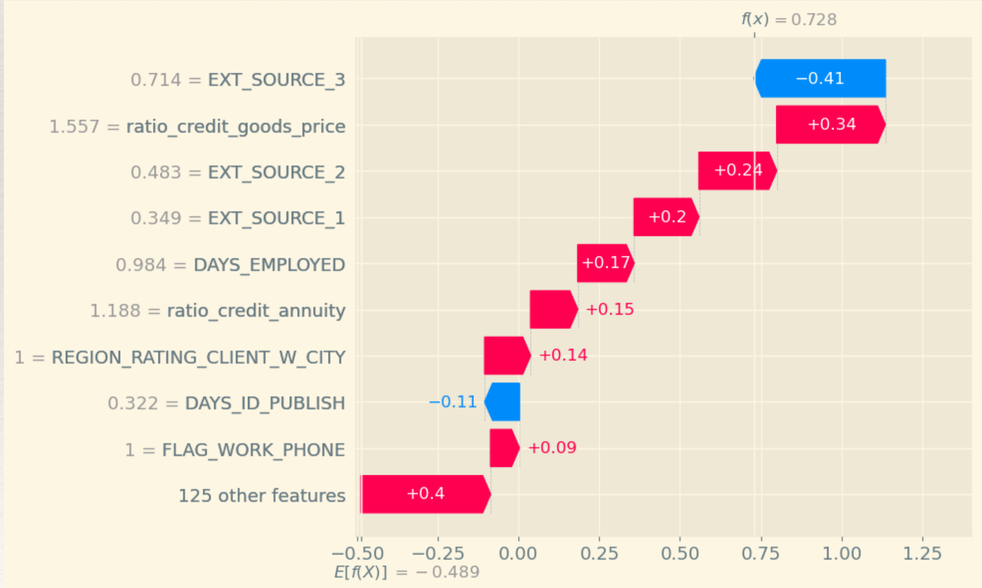
[1,0]



Individu 10

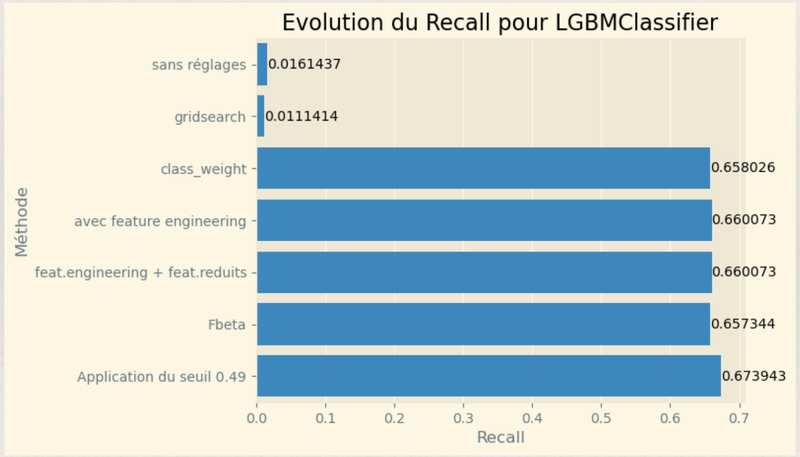
True positif

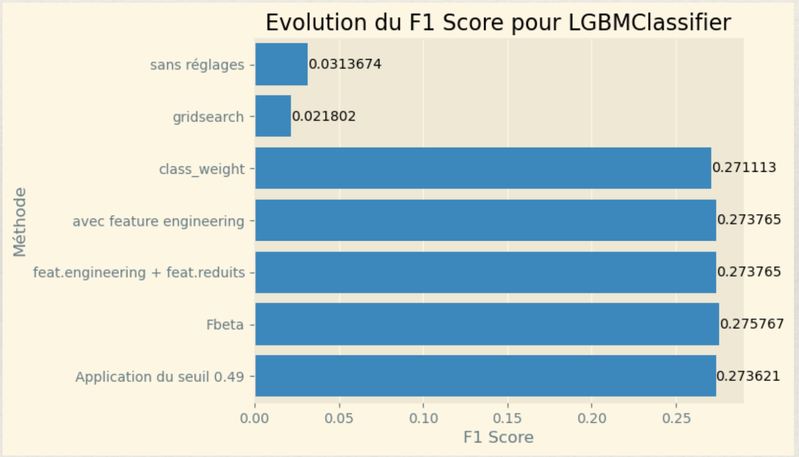
[1,1]

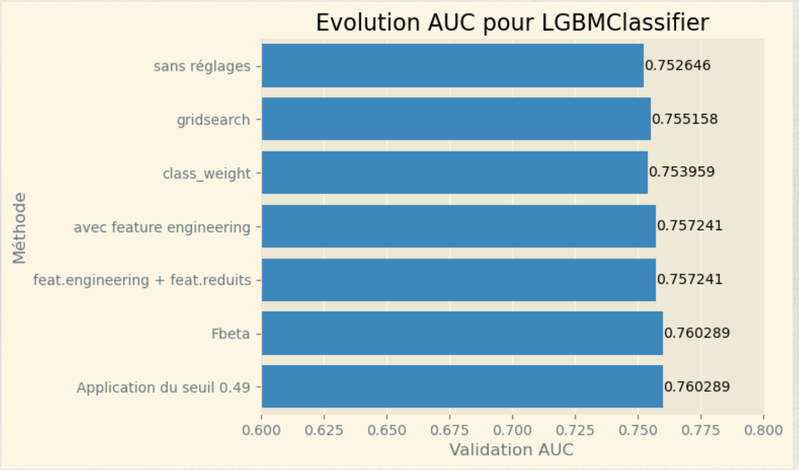


Conclusion

Data Séparation Modèles Métriques Interprétation





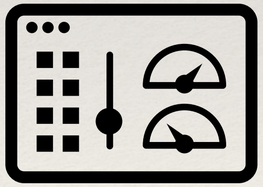


Axe d’amélioration

Amélioration du dataset

Amélioration du modèle

Dashboard



API