OPENCLASSROOMS

Parcours Data Scientist

Projet 7 : Implémentez un modèle de scoring

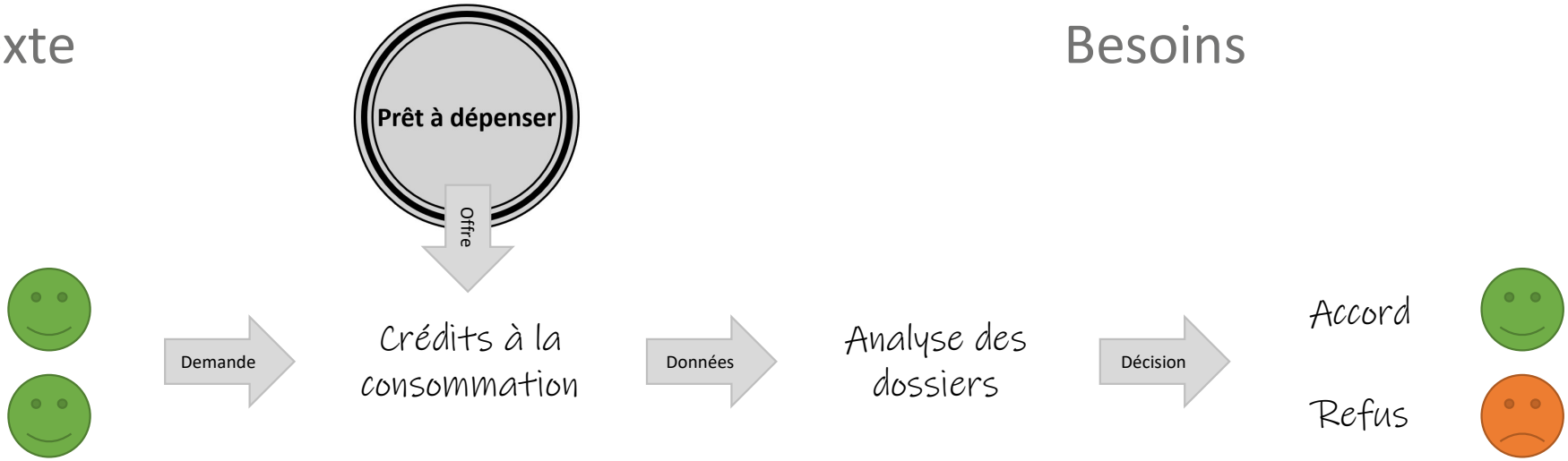
Etudiant: Eric Wendling
Mentor: Julien Heiduk
GitHub: https://github.com/leerik/OC_DS_P7
Date: 20/10/2020

Crédit Scoring

Définition du projet

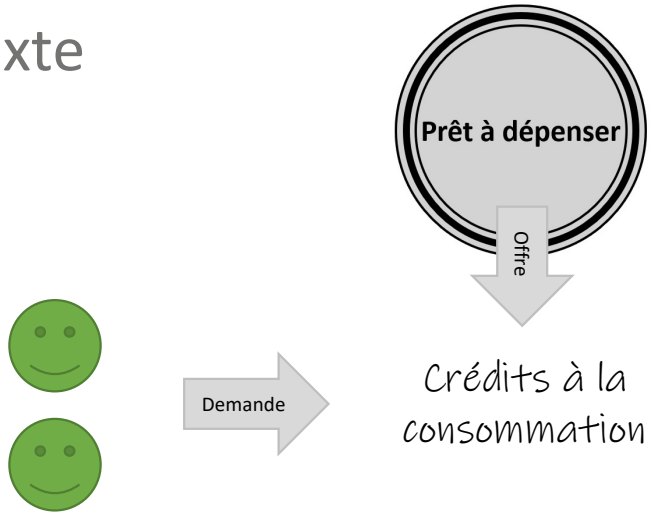
Contexte

Besoins

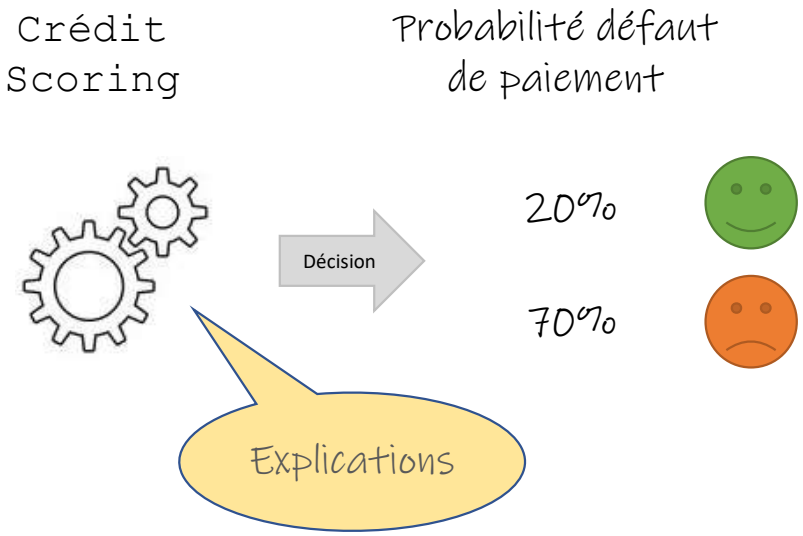


Définition du projet

Contexte



Objectifs



Projet

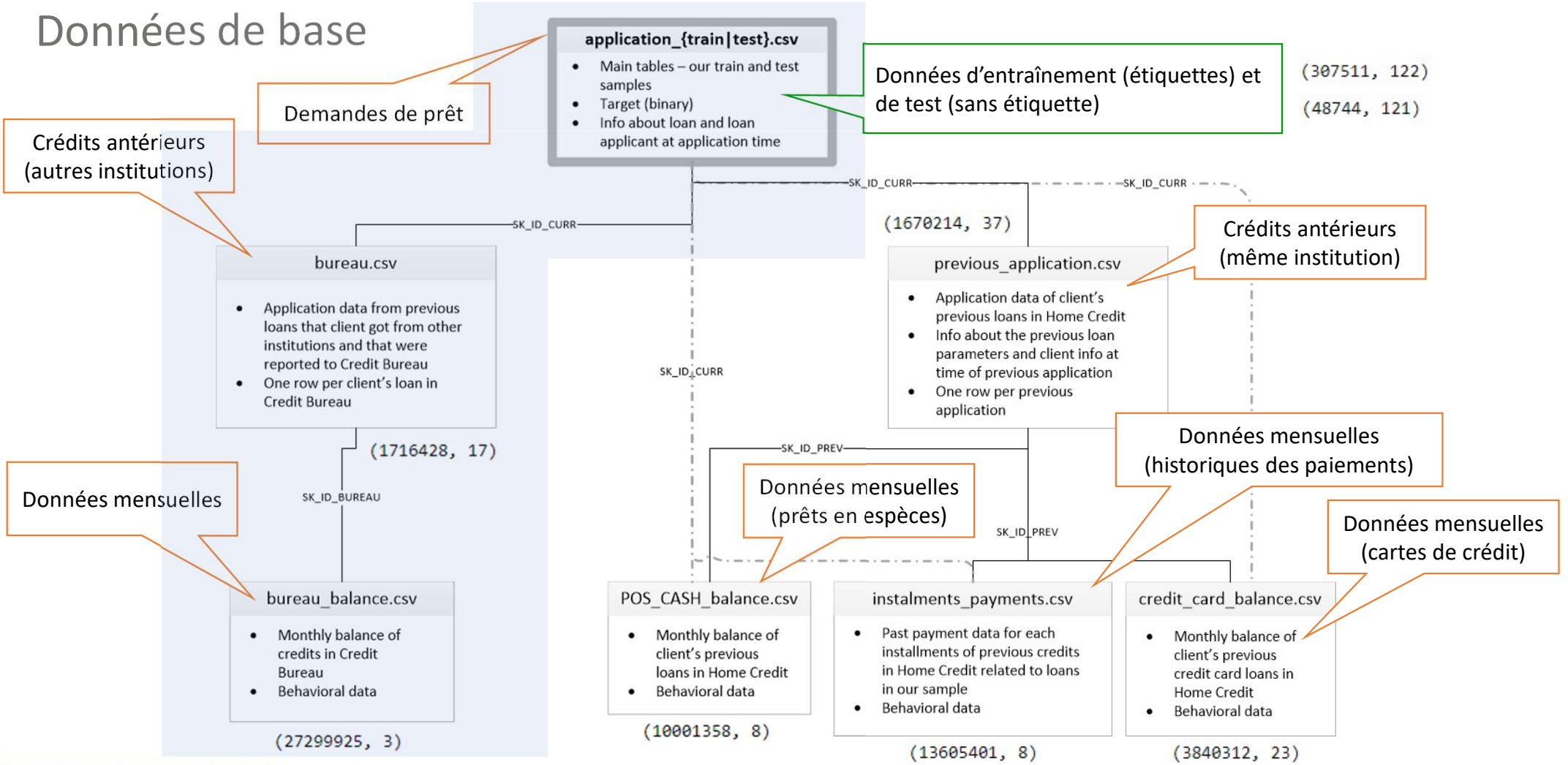
Analyse et traitement des données existantes

Modélisation d'un système de classification binaire

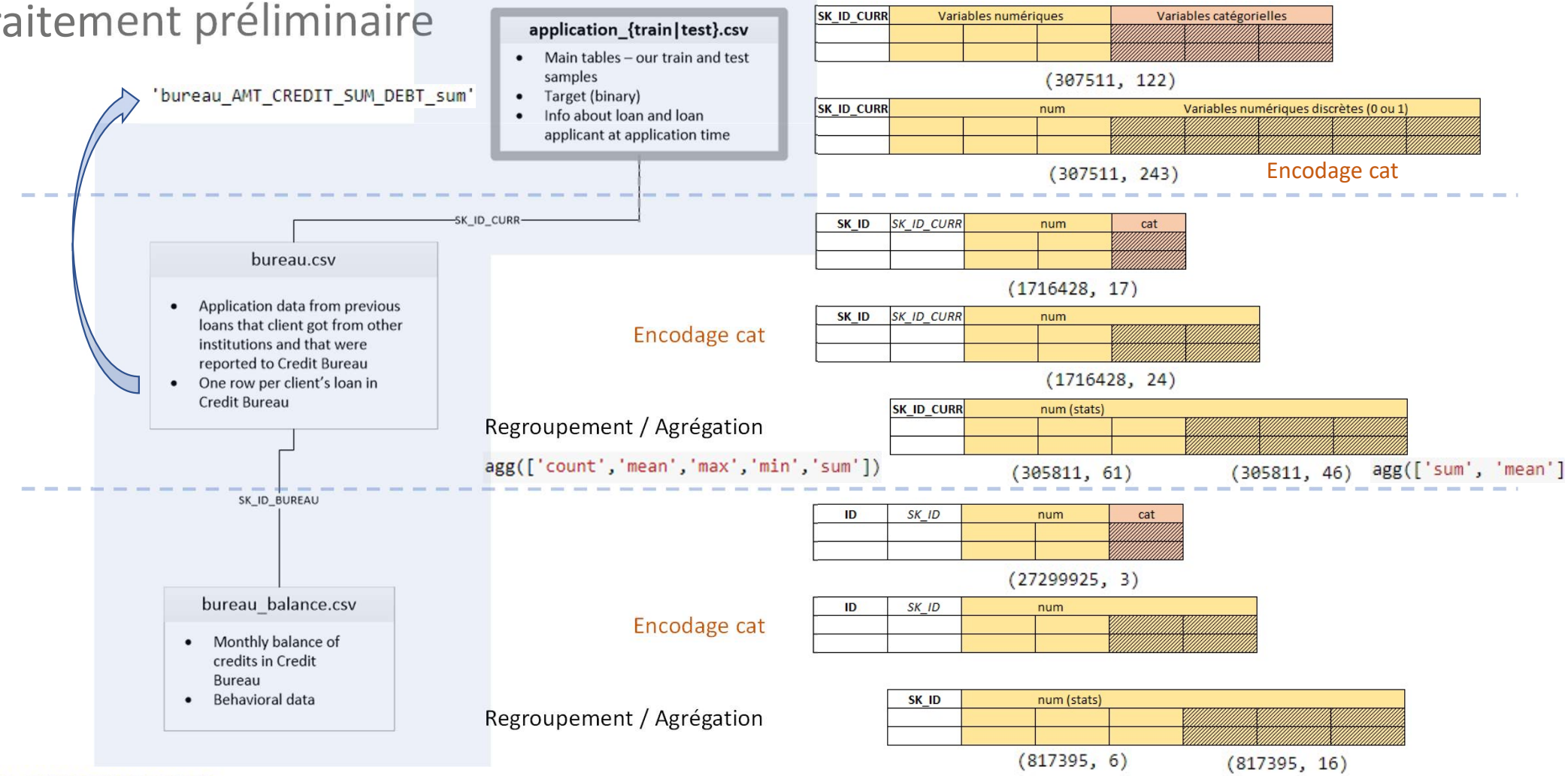
Réalisation d'un dashboard interactif

- Prédiction de scores
- Informations clients
- Explicabilité des décisions

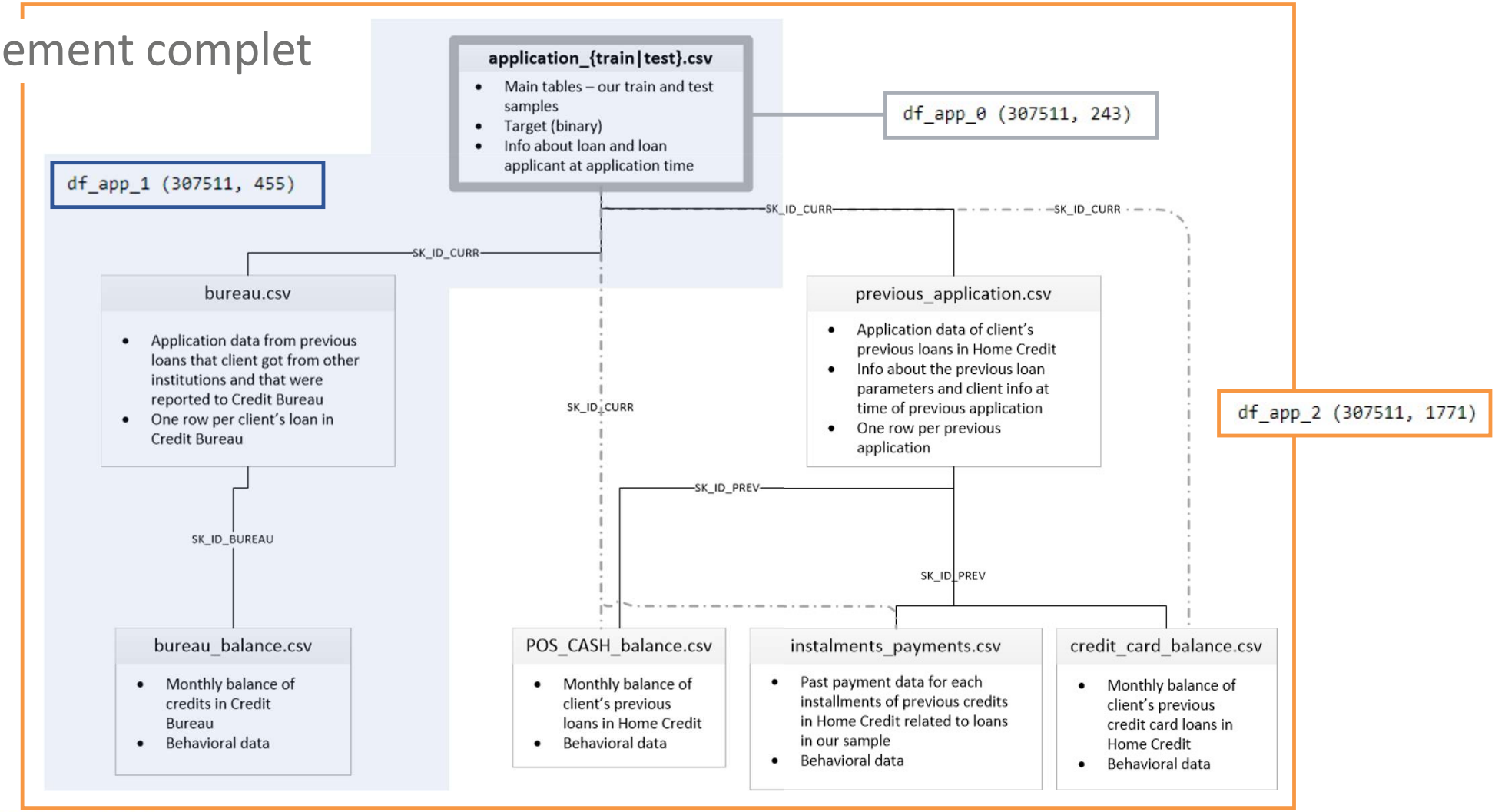
Données de base



Traitement préliminaire



Traitement complet



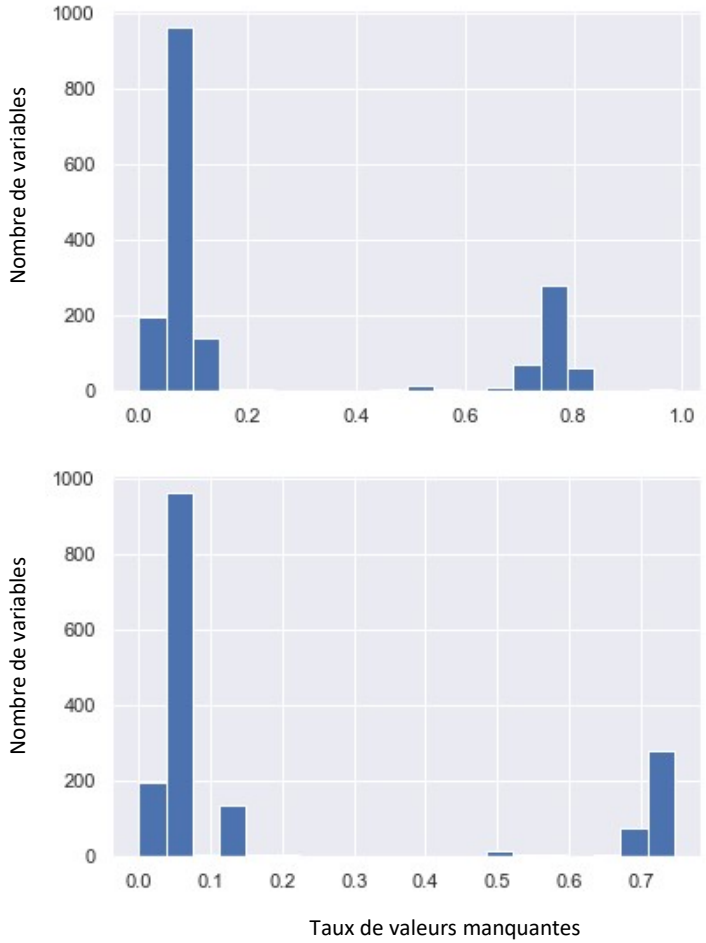
Valeurs manquantes

Dataframe

- df_app_2
  - Le dataframe comprend **1771** variables (avec ID et Cible)
  - Il y a **1588** variables avec des valeurs manquantes

Réduction de dimension

- Réduction du nombre de variables
  - On supprime les variables qui ont plus de 80% de valeurs manquantes
  - Après suppression des variables, le dataframe comprend **1702** variables
- Remplacement des valeurs manquantes
  - Valeurs de remplacement
    - 0
    - Moyenne ou médiane
      - Traitement par classe
      - Déduction par similarité (proches voisins)

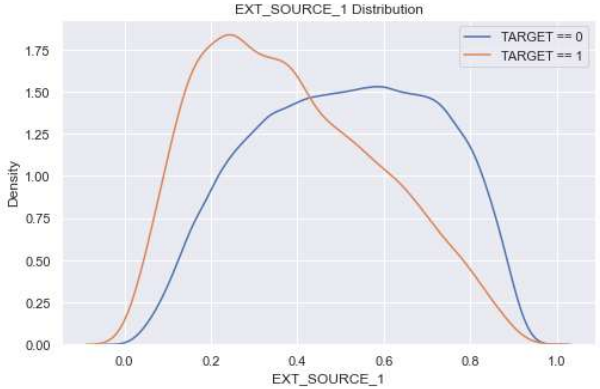


Valeurs nulles

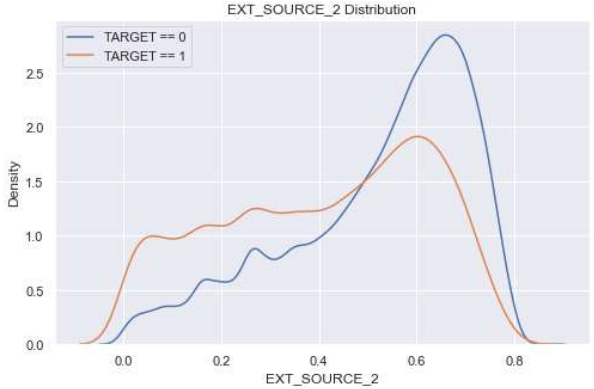
```
'EXT_SOURCE_1': 173378 (ratio = 0.56)
'EXT_SOURCE_2': 660 (ratio = 0.00)
'EXT_SOURCE_3': 60965 (ratio = 0.20)
```


Corrélations des variables avec la cible (références KDE)

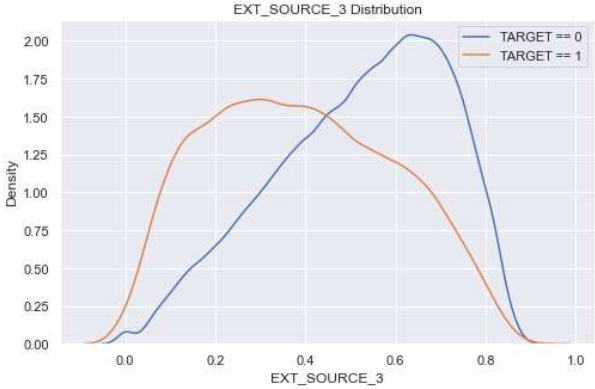
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.1553
Valeur médiane pour les crédits non remboursés = 0.3617
Valeur médiane pour les crédits remboursés = 0.5175
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1605
Valeur médiane pour les crédits non remboursés = 0.4404
Valeur médiane pour les crédits remboursés = 0.5739
```

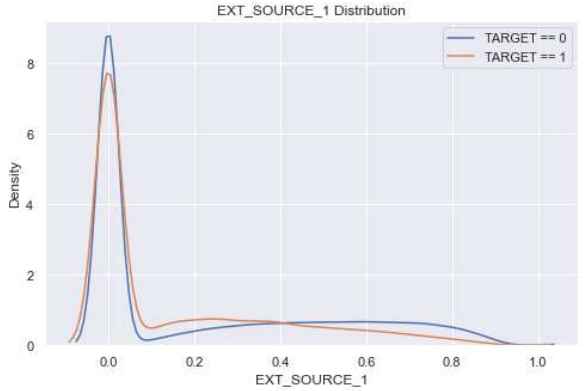



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1789
Valeur médiane pour les crédits non remboursés = 0.3791
Valeur médiane pour les crédits remboursés = 0.5460
```

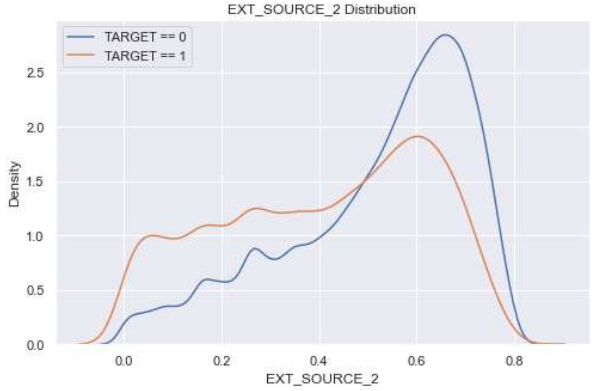


Après remplacement des valeurs nulles par 0

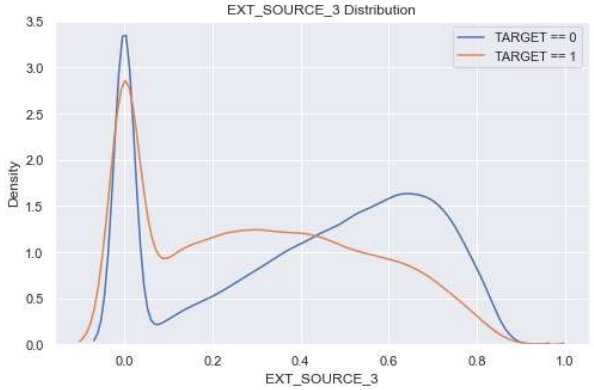
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.0647
Valeur médiane pour les crédits non remboursés = 0.0000
Valeur médiane pour les crédits remboursés = 0.0000
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1590
Valeur médiane pour les crédits non remboursés = 0.4395
Valeur médiane pour les crédits remboursés = 0.5734
```




```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1196
Valeur médiane pour les crédits non remboursés = 0.2881
Valeur médiane pour les crédits remboursés = 0.4741
```

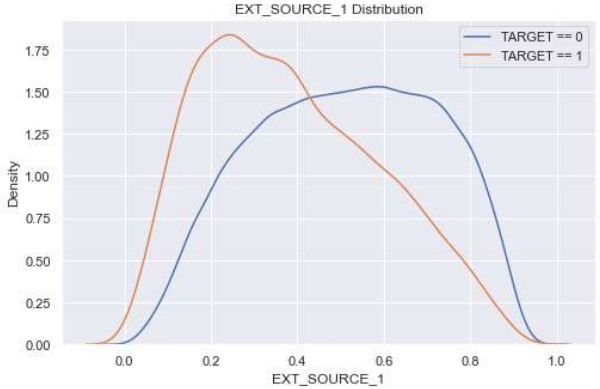


Valeurs nulles

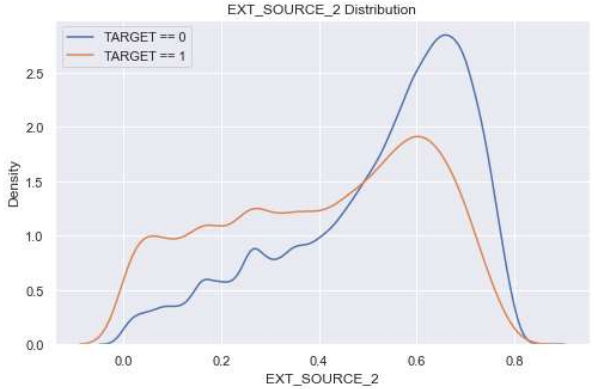
```
'EXT_SOURCE_1': 173378 (ratio = 0.56)
'EXT_SOURCE_2': 660 (ratio = 0.00)
'EXT_SOURCE_3': 60965 (ratio = 0.20)
```


Corrélations des variables avec la cible (références KDE)

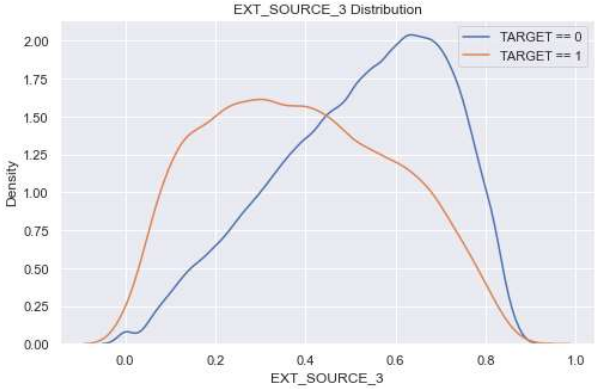
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.1553
Valeur médiane pour les crédits non remboursés = 0.3617
Valeur médiane pour les crédits remboursés = 0.5175
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1605
Valeur médiane pour les crédits non remboursés = 0.4404
Valeur médiane pour les crédits remboursés = 0.5739
```

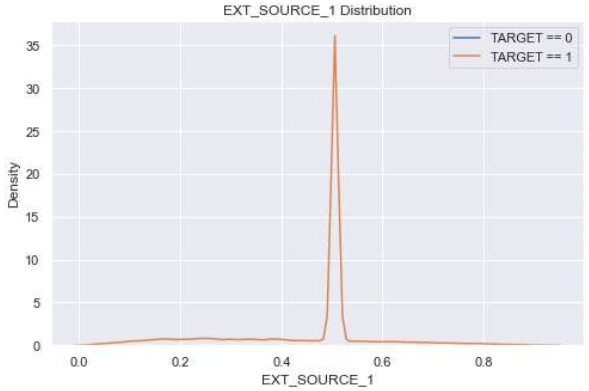



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1789
Valeur médiane pour les crédits non remboursés = 0.3791
Valeur médiane pour les crédits remboursés = 0.5460
```

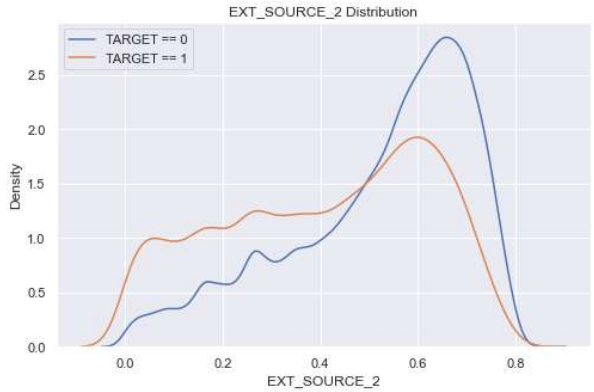


Après remplacement des valeurs nulles par la valeur médiane

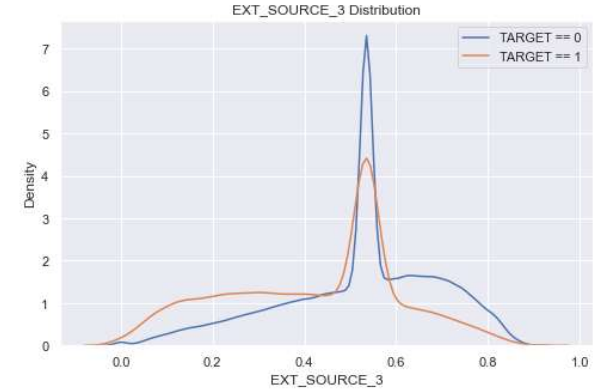
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.0989
Valeur médiane pour les crédits non remboursés = 0.5060
Valeur médiane pour les crédits remboursés = 0.5060
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1603
Valeur médiane pour les crédits non remboursés = 0.4411
Valeur médiane pour les crédits remboursés = 0.5734
```

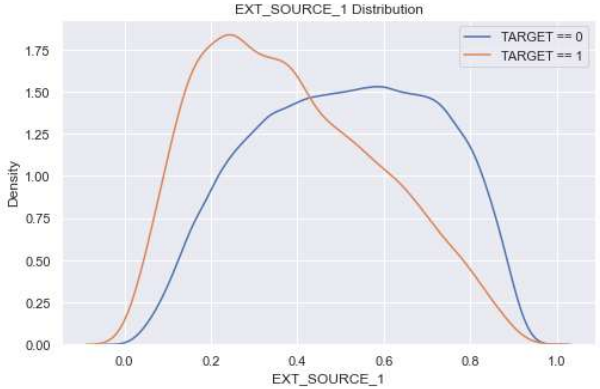



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1559
Valeur médiane pour les crédits non remboursés = 0.4758
Valeur médiane pour les crédits remboursés = 0.5353
```

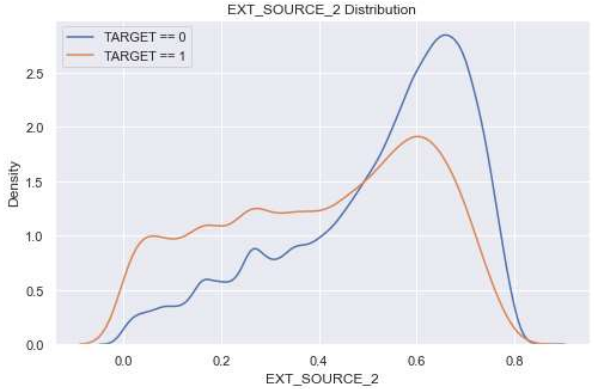


Corrélations des variables avec la cible (références KDE)

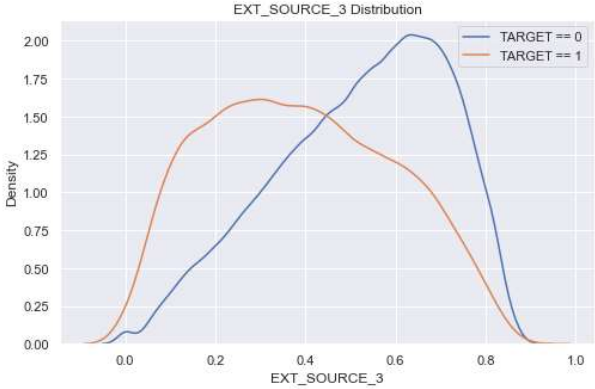
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.1553
Valeur médiane pour les crédits non remboursés = 0.3617
Valeur médiane pour les crédits remboursés = 0.5175
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1605
Valeur médiane pour les crédits non remboursés = 0.4404
Valeur médiane pour les crédits remboursés = 0.5739
```

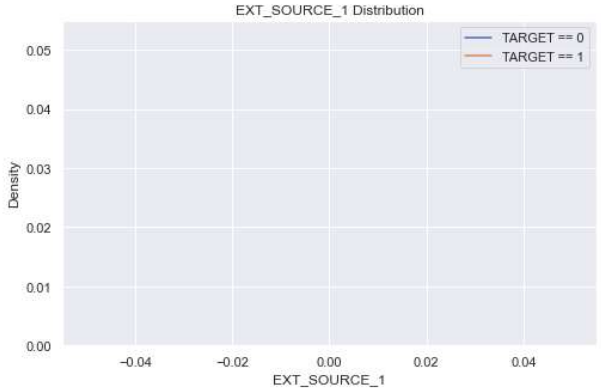



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1789
Valeur médiane pour les crédits non remboursés = 0.3791
Valeur médiane pour les crédits remboursés = 0.5460
```

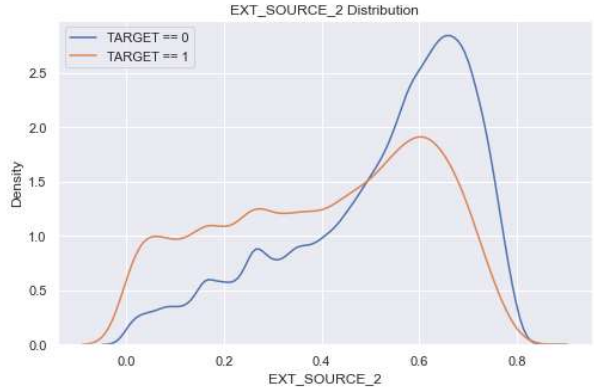


Après remplacement des valeurs nulles par la valeur médiane par classe

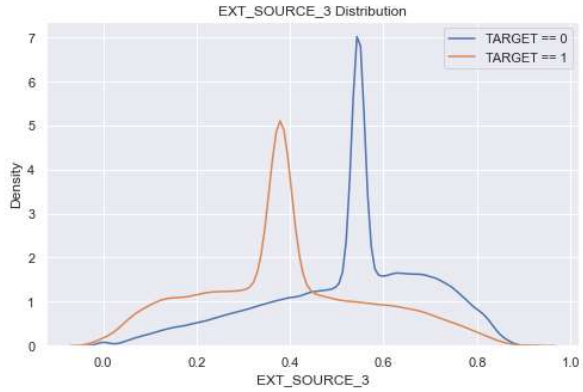
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.2719
Valeur médiane pour les crédits non remboursés = 0.3617
Valeur médiane pour les crédits remboursés = 0.5175
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1607
Valeur médiane pour les crédits non remboursés = 0.4404
Valeur médiane pour les crédits remboursés = 0.5739
```

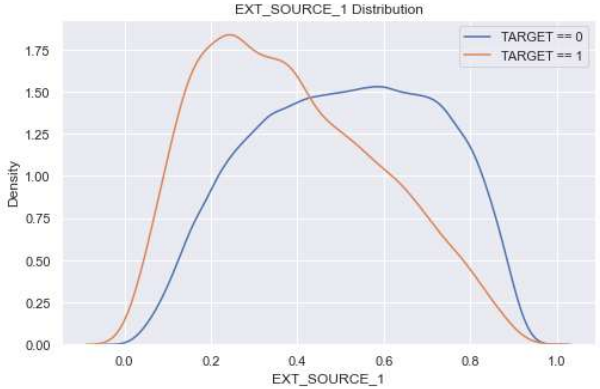



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.2133
Valeur médiane pour les crédits non remboursés = 0.3791
Valeur médiane pour les crédits remboursés = 0.5460
```

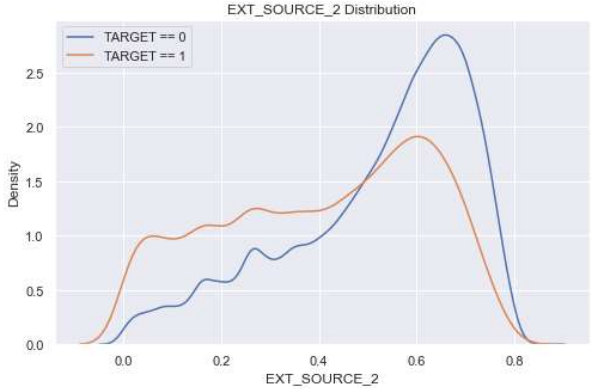


Corrélations des variables avec la cible (références KDE)

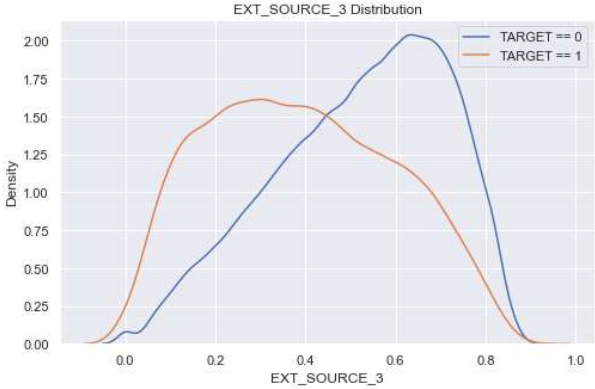
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.1553
Valeur médiane pour les crédits non remboursés = 0.3617
Valeur médiane pour les crédits remboursés = 0.5175
```




```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1605
Valeur médiane pour les crédits non remboursés = 0.4404
Valeur médiane pour les crédits remboursés = 0.5739
```

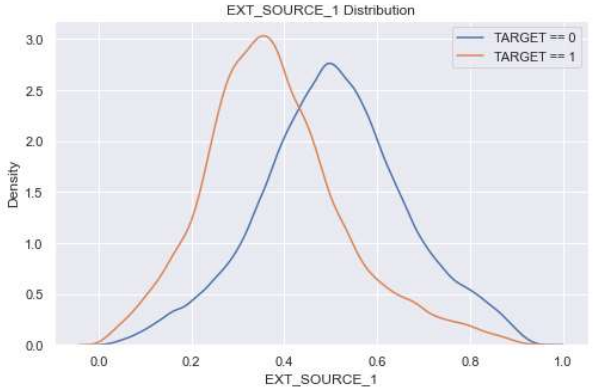



```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1789
Valeur médiane pour les crédits non remboursés = 0.3791
Valeur médiane pour les crédits remboursés = 0.5460
```

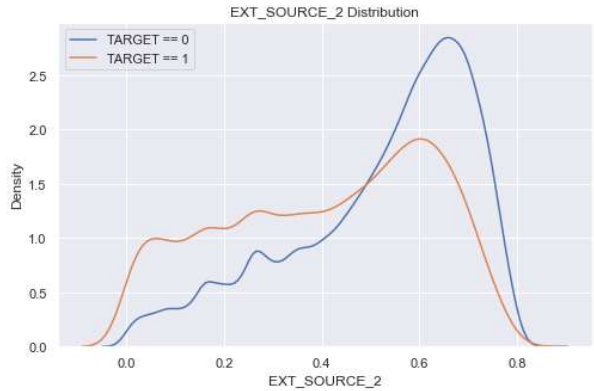


Après remplacement des valeurs nulles par la moyenne des valeurs des plus proches voisins, par classe

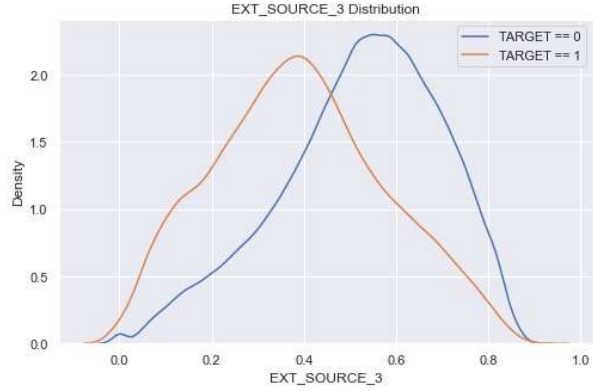
```
La corrélation entre la variable EXT_SOURCE_1 et la cible est de -0.2108
Valeur médiane pour les crédits non remboursés = 0.3640
Valeur médiane pour les crédits remboursés = 0.5028
```



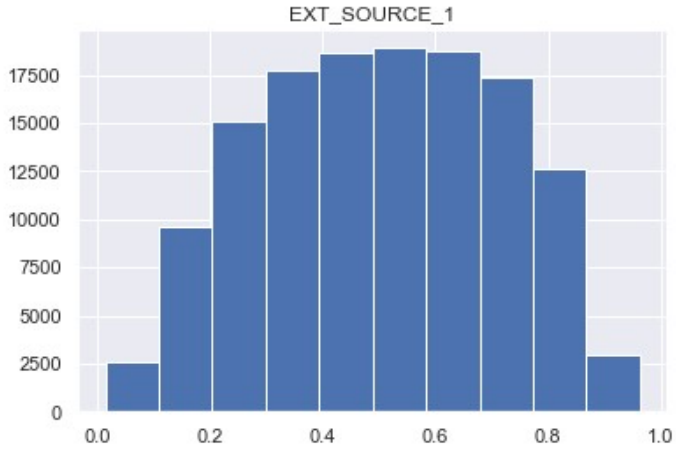

```
La corrélation entre la variable EXT_SOURCE_2 et la cible est de -0.1606
Valeur médiane pour les crédits non remboursés = 0.4400
Valeur médiane pour les crédits remboursés = 0.5737
```

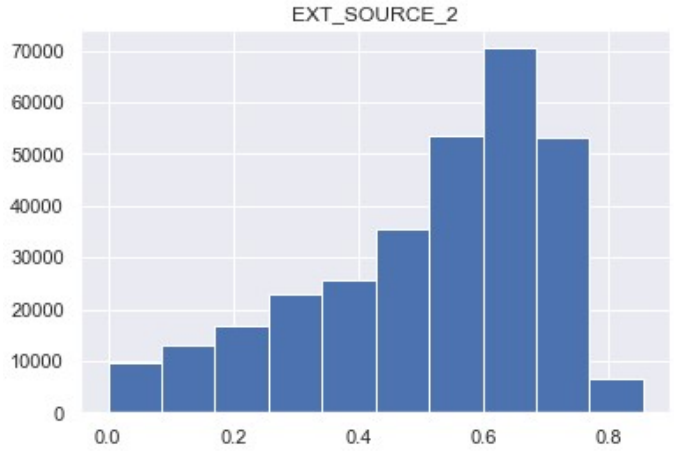


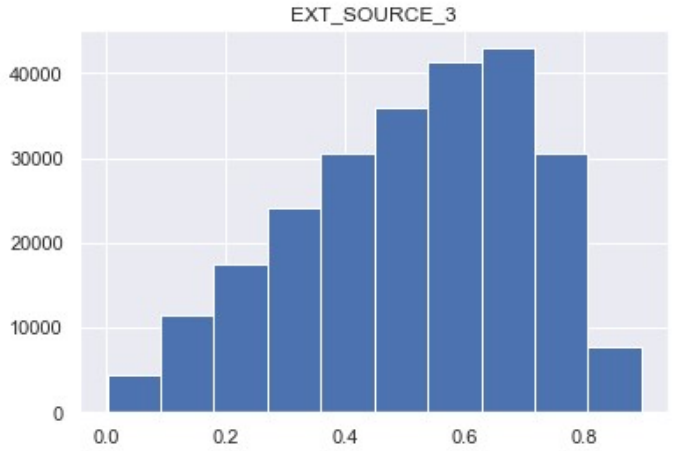

```
La corrélation entre la variable EXT_SOURCE_3 et la cible est de -0.1979
Valeur médiane pour les crédits non remboursés = 0.3858
Valeur médiane pour les crédits remboursés = 0.5394
```



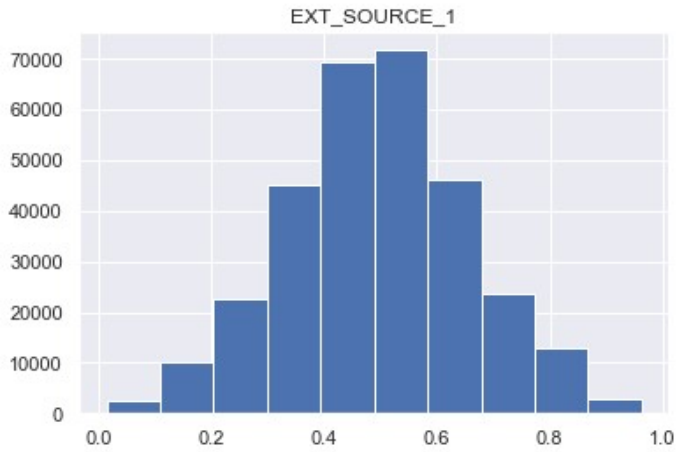
Histogrammes avant traitement

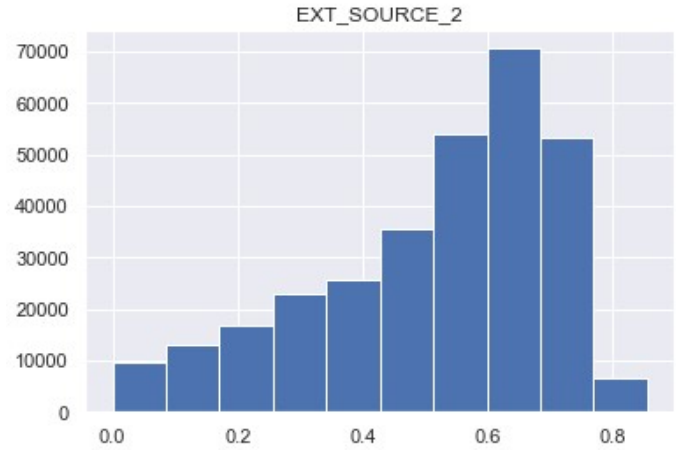


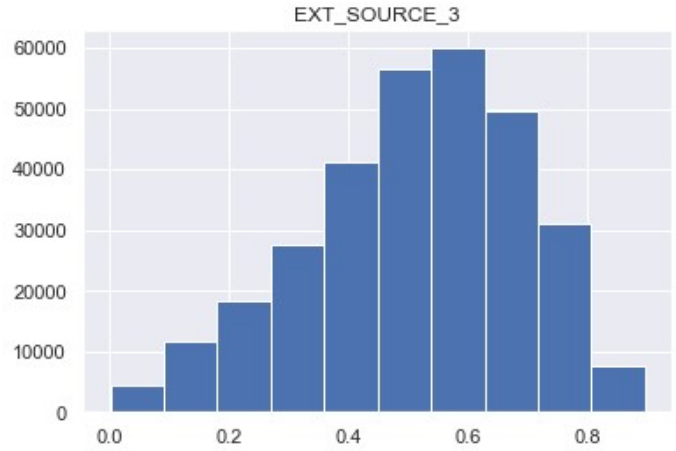




Histogrammes après remplacement des valeurs nulles par la moyenne des valeurs des plus proches voisins







Corrélations

Dataframe

- df_app_2
  - Le dataframe comprend **1702** variables
  - (307511 dossiers)

Corrélations entre variables

- Lien entre les variables
  - Nature du lien: linéaire ou plus complexe
- Identification des variables corrélées
  - Corrélation de Pearson (relation linéaire)
  - Corrélation de Spearman (relation monotone)

Réduction de dimension

- Identification des variables corrélées
  - Identification des variables corrélées à plus de 95%
  - Suppression d'une variable pour chaque paire de variables corrélées
  - Après suppression des variables, le dataframe comprend **913** variables (avec la Cible)

	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
SK_ID_CURR	1.0	-0.0011	-0.0018	-0.00034	-0.00041	-0.00039
CNT_CHILDREN	-0.0011	1.0	0.013	0.0021	0.021	-0.002
AMT_INCOME_TOTAL	-0.0018	0.013	1.0	0.16	0.19	0.16
AMT_CREDIT	-0.00034	0.0021	0.16	1.0	0.77	0.99
AMT_ANNUITY	-0.00041	0.021	0.19	0.77	1.0	0.78
AMT_GOODS_PRICE	-0.00039	-0.002	0.16	0.99	0.78	1.0

```
{'AMT_CREDIT': ['AMT_CREDIT', 'AMT_GOODS_PRICE']}
```

Valeurs des variables corrélées

	AMT_CREDIT	AMT_GOODS_PRICE
0	406597.5	351000.0
1	1293502.5	1129500.0
2	135000.0	135000.0
3	312682.5	297000.0
4	513000.0	513000.0
...	...	...
307506	254700.0	225000.0
307507	269550.0	225000.0
307508	677664.0	585000.0
307509	370107.0	319500.0
307510	675000.0	675000.0

Variance

Définition

- Mesure de la dispersion des valeurs d'une variable

Dataframe

- df_app_2
  - Le dataframe comprend **912** variables (sans la Cible)
  - Il y a **14** variables avec une variance = 0

Réduction de dimension

- Réduction du nombre de variables
  - On supprime les variables inférieures à un seuil de variance
  - On a fixé le seuil à **0** afin de supprimer les variables avec une variance = 0
  - Après suppression des variables, le dataframe comprend **898** variables

	index	Variance
911	client_credit_card_balance_SK_DPD_min_sum	0.000000e+00
898	client_credit_card_balance_SK_DPD_min_min	0.000000e+00
899	client_credit_card_balance_SK_DPD_DEF_min_mean	0.000000e+00
900	client_POS_CASH_balance_NAME_CONTRACT_STATUS_X...	0.000000e+00
901	previous_application_NAME_GOODS_CATEGORY_House...	0.000000e+00
902	client_POS_CASH_balance_SK_DPD_DEF_min_min	0.000000e+00
903	previous_application_NAME_GOODS_CATEGORY_House...	0.000000e+00
910	client_credit_card_balance_SK_DPD_DEF_min_max	0.000000e+00
905	client_credit_card_balance_SK_DPD_min_mean	0.000000e+00
909	client_credit_card_balance_SK_DPD_DEF_min_min	0.000000e+00
908	client_credit_card_balance_SK_DPD_DEF_min_sum	0.000000e+00
904	client_POS_CASH_balance_NAME_CONTRACT_STATUS_X...	0.000000e+00
906	client_POS_CASH_balance_SK_DPD_min_min	0.000000e+00
907	client_credit_card_balance_SK_DPD_min_max	0.000000e+00
897	client_credit_card_balance_NAME_CONTRACT_STATU...	1.390690e-08
896	client_POS_CASH_balance_NAME_CONTRACT_STATUS_X...	3.583340e-08
895	client_POS_CASH_balance_NAME_CONTRACT_STATUS_C...	4.582500e-07
894	client_POS_CASH_balance_SK_DPD_DEF_min_mean	8.129790e-07
893	client_bureau_balance_STATUS_4_count_norm_min	8.519733e-07
892	client_POS_CASH_balance_NAME_CONTRACT_STATUS_C...	9.412349e-07

Jeux de données

Dataframe

- df_app_2
  - X (307511, **898**)
  - y (307511,)
- df_app_2_test
  - (48744, **898**)

⇦ Sans étiquette

Jeux d'entraînement et de test

- df_app_2
  - Réservation de 20% des données pour le jeu de test

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(X, y, train_size = 0.8, random_state = rs_)
```

```
Jeu d'entraînement: x_train (246008, 898)
Etiquettes: y_train (246008,)

Jeu de test: x_test (61503, 898)
Etiquettes: y_test (61503,)
```

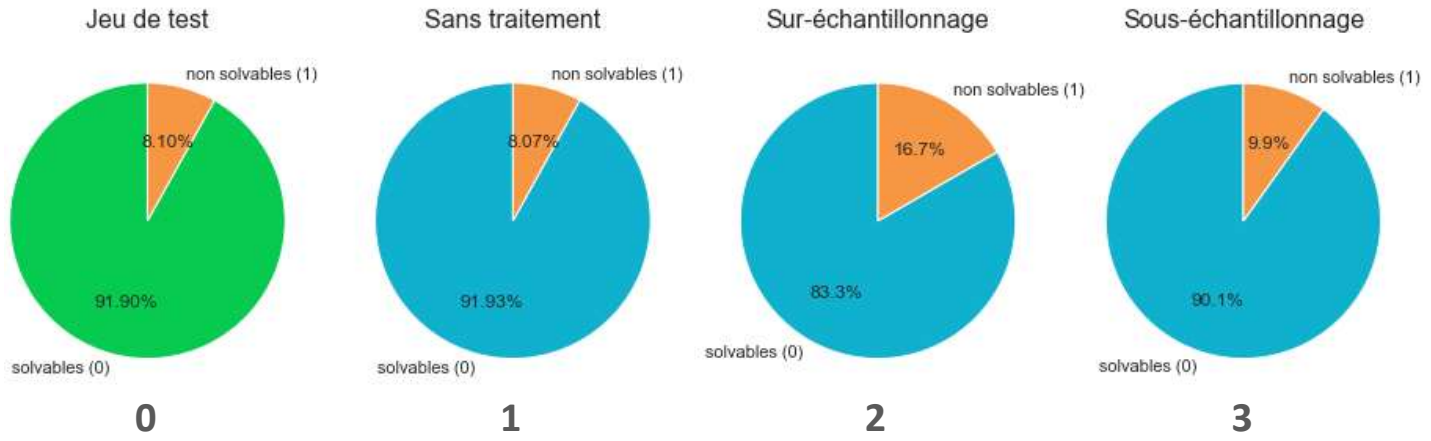
Jeux de données

Déséquilibre des données

- Sur-représentation d'une classe
  - Crédit remboursé (client solvable): 91,93%
  - Crédit non remboursé (client non solvable): 8,07%
- Méthodes
  - Sur-échantillonnage de la classe minoritaire (Smote)
  - Sous-échantillonnage de la classe majoritaire (NearMiss)

	Données	Traitement	Individus	Non solvables (1)	Solvables (0)	% Non solvables (1)
0	Test	Non	61503	4982	56521	8.10
1	Train	Non	246008	19843	226165	8.07
2	Train	Sur-échantillonnage	271398	45233	226165	16.67
3	Train	Sous-échantillonnage	200775	19843	180932	9.88

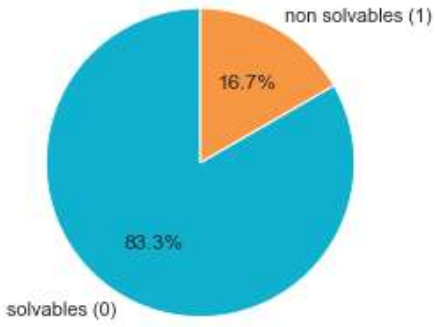
Sur-éch. Sous-éch.



Caractéristiques du modèle

Jeux de données

- Entraînement
  - (246008, **897**)
- Test avec étiquettes
  - (61503, **897**)
- Test sans étiquette
  - (48744, **897**)



Objectifs

- Classification de dossiers de crédits en 2 classes
  - ➢ Classification binaire
  - ➢ Apprentissage supervisé
- Compréhension des critères de décision
  - Interprétabilité (ou explicabilité)
  - ➢ Importance des variables
    - Global
    - Local
- Volumes de données importants
  - ➢ Temps d'exécution raisonnables
- Mesures des performances
  - Optimisation fonction de perte
  - ➢ Définir les Métriques

Mesure de performance d'un classifieur

Dossier	Négatif Remboursé	Positif Non remboursé	Seuil	Prédiction	Réel	Vrai Positif	Vrai Négatif	Faux Positif	Faux Négatif
1	0,3	0,7	0,6	1	0			1	
2	0,6	0,4	0,6	0	0		1		
3	0,8	0,2	0,6	0	0		1		
4	0,1	0,9	0,6	1	1	1			
5	0,5	0,5	0,6	0	0		1		
6	0,3	0,7	0,6	1	0			1	
7	0,4	0,6	0,6	0	1				1
8	0,8	0,2	0,6	0	0		1		
9	0,6	0,4	0,6	0	0		1		
10	0,3	0,7	0,6	1	1	1			
						2	5	2	1

Seuil = 0,6

Sensibilité = TP / (TP + FN) = 2/3 = 0,67
Spécificité = TN / (TN + FP) = 5/7 = 0,71
Précision = TP / (TP + FP) = 2/4 = 0,5

	Réel Négatif		Réel Positif		
Prédiction Négatif	TN	**5**	FN	**1**	6
Prédiction Positif	FP	**2**	TP	**2**	4
		7		3	

Type de modèle

Seuil = 0,3

Sensibilité = TP / (TP + FN) = 3/3 = 1
Spécificité = TN / (TN + FP) = 2/7 = 0,29
Précision = TP / (TP + FP) = 3/8 = 0,38

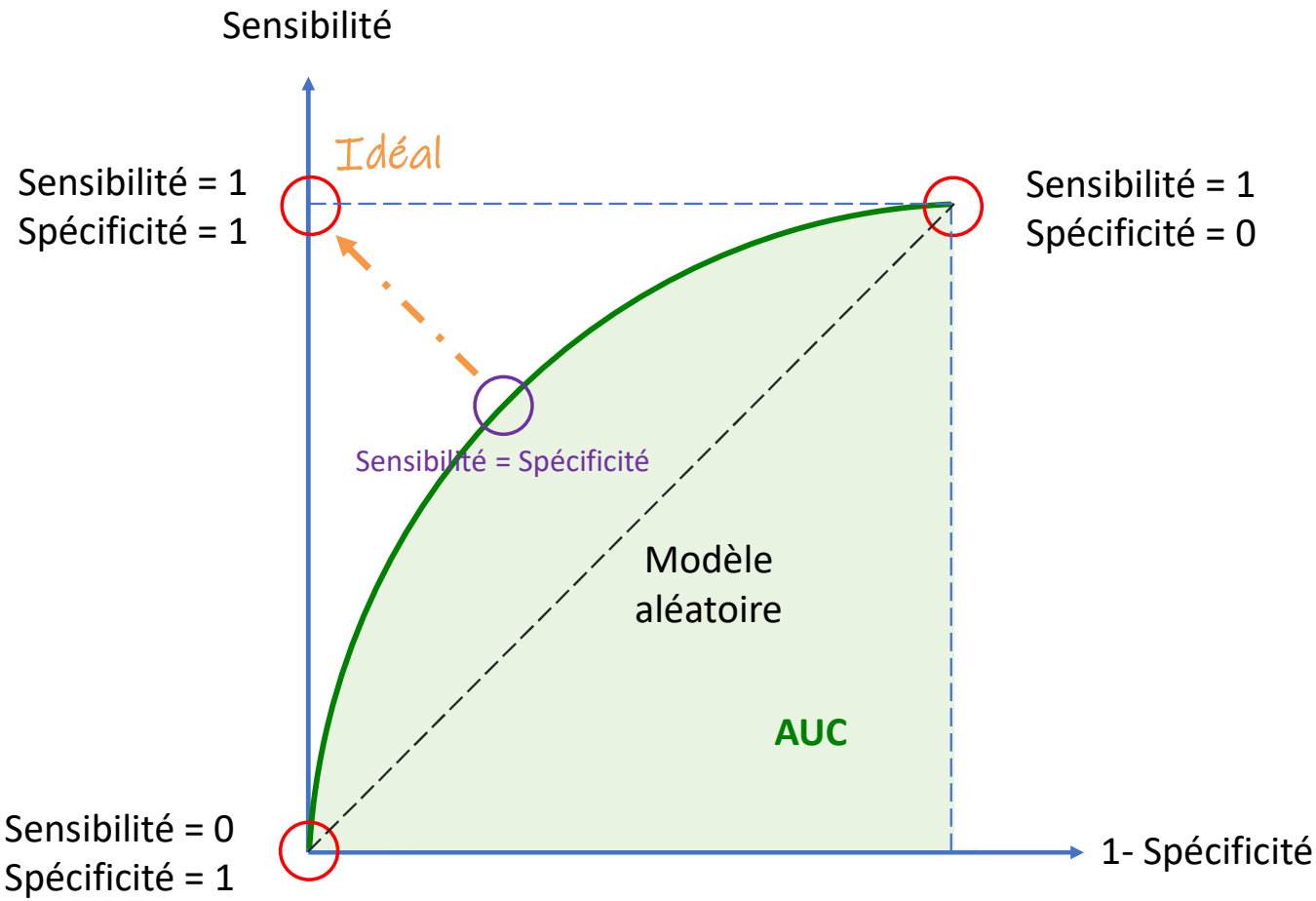
	Réel Négatif		Réel Positif		
Prédiction Négatif	TN	**2**	FN	**0**	2
Prédiction Positif	FP	**5**	TP	**3**	8
		7		3	

Seuil = 0,6

Sensibilité = TP / (TP + FN) = 2/3 = 0,67
Spécificité = TN / (TN + FP) = 5/7 = 0,71
Précision = TP / (TP + FP) = 2/4 = 0,5

	Réel Négatif		Réel Positif		
Prédiction Négatif	TN	**5**	FN	**1**	6
Prédiction Positif	FP	**2**	TP	**2**	4
		7		3	

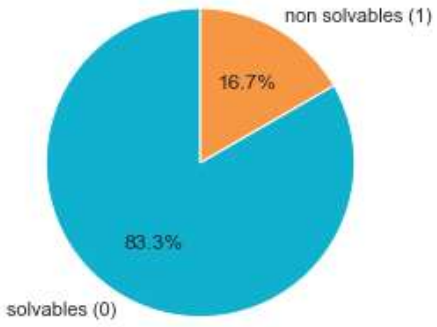
Courbe ROC


Sensibilité
Idéal
Sensibilité = 1
Spécificité = 1
Sensibilité = 1
Spécificité = 0
Sensibilité = Spécificité
Modèle
aléatoire
AUC
Sensibilité = 0
Spécificité = 1
1- Spécificité

Caractéristiques du modèle

Jeux de données

- Entraînement
  - (246008, 897)
- Test avec étiquettes
  - (61503, 897)
- Test sans étiquette
  - (48744, 897)



Objectifs

- Classification de dossiers de crédits en 2 classes
  - ➢ Classification binaire
  - ➢ Apprentissage supervisé
- Compréhension des critères de décision
  - Interprétabilité (ou explicabilité)
  - ➢ Importance des variables
    - Global
    - Local
- Volumes de données importants
  - ➢ Temps d'exécution raisonnables
- Mesures des performances
  - Optimisation fonction de perte
  - ➢ Métriques: Mesures **ROC** et **AUC**

Classifieurs

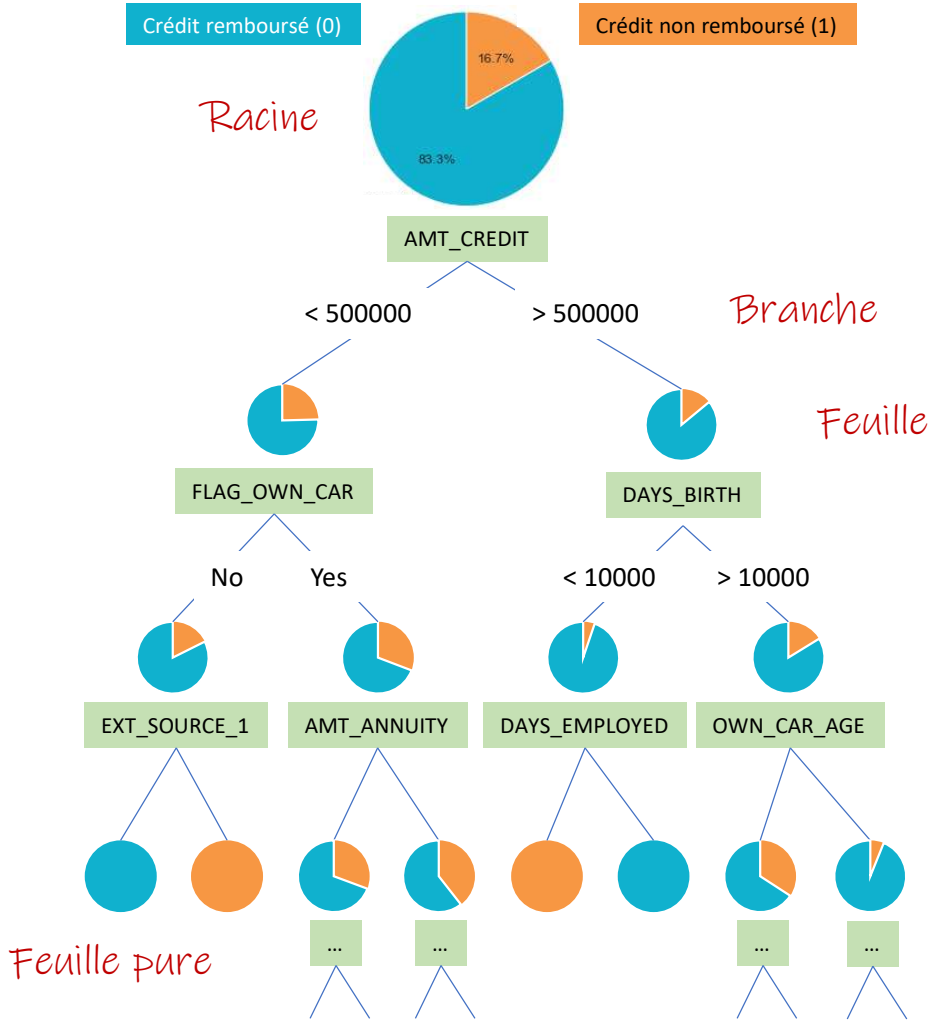
Arbres de décision

Données	Traitement	Individus	Non solvables (1)	Solvables (0)	% Non solvables (1)
Test	Non	61503	4982	56521	8.10
Train	Sur-échantillonnage	271398	45233	226165	16.67

- Forêts aléatoires
- Méthodes ensemblistes

Modèles testés

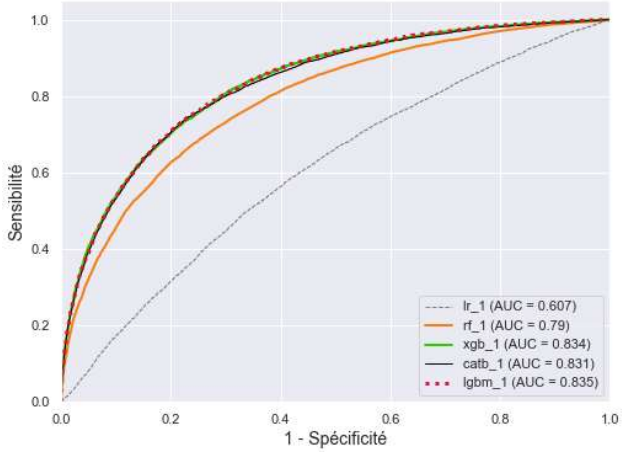
- Random Forest
- XGBoost
- CatBoost
- LightGBM



Comparaison des performances AUC

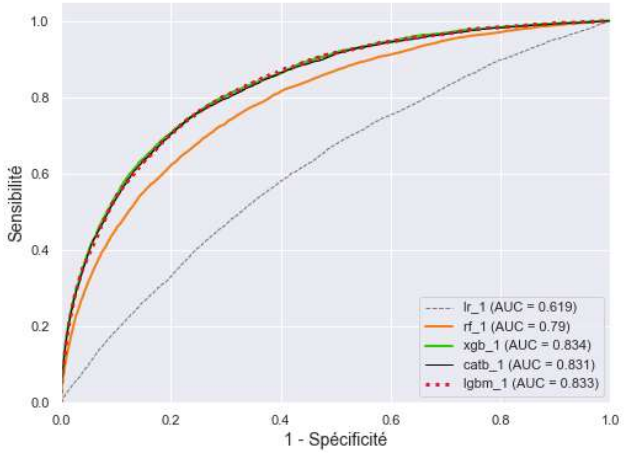
Sans traitement

	Features dim	Model	Train score	Test score	Run time
4	(246008, 897)	lgbm_1	0.868	0.835	289.0
2	(246008, 897)	xgb_1	0.895	0.834	3389.0
3	(246008, 897)	catb_1	0.874	0.831	1854.0
1	(246008, 897)	rf_1	1.000	0.790	1041.0
0	(246008, 897)	lr_1	0.613	0.607	132.0



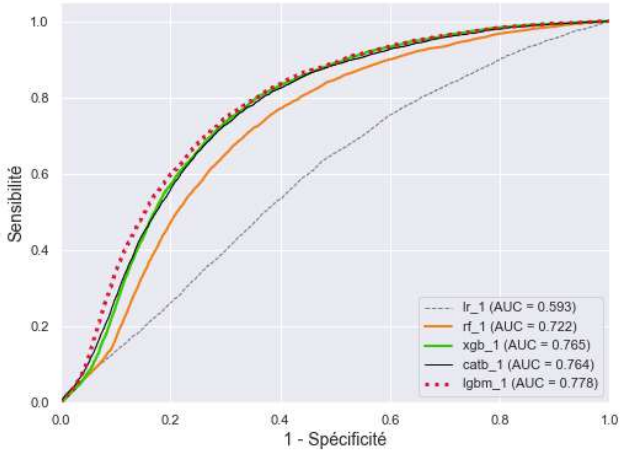
Sur-échantillonnage Smote

	Features dim	Model	Train score	Test score	Run time
2	(271398, 897)	xgb_1	0.955	0.834	3934.0
4	(271398, 897)	lgbm_1	0.936	0.833	354.0
3	(271398, 897)	catb_1	0.942	0.831	2499.0
1	(271398, 897)	rf_1	1.000	0.790	971.0
0	(271398, 897)	lr_1	0.631	0.619	143.0



Sous-échantillonnage NearMiss

	Features dim	Model	Train score	Test score	Run time
4	(200775, 897)	lgbm_1	0.885	0.778	231.0
2	(200775, 897)	xgb_1	0.912	0.765	2728.0
3	(200775, 897)	catb_1	0.883	0.764	1580.0
1	(200775, 897)	rf_1	1.000	0.722	853.0
0	(200775, 897)	lr_1	0.657	0.593	102.0



Réduction de dimensions

Importance des variables

- Combinaison des variables importantes des différents classifieurs
  - Sélection N variables les plus importantes (par classifieur)
    - N = 300

➢ Option 1: On combine l'ensemble des variables et on supprime les doublons

```
Nombre de variables: 1200
Nombre de doublons: 748
Nombre de variables uniques: 452
Nombre de variables après suppression des doublons: 452
```

➢ Option 2: On retient uniquement les variables communes

```
Nombre de variables total: 897
Nombre de variables communes: 300
Nombre de variables communes: 191
Nombre de variables communes: 160
Nombre de variables communes: 150
```

Recursive Feature Elimination

- Basé sur l'apprentissage d'un classifieur
- Réduction du dataset aux variables importantes

```
Nombre de variables initial: 897
step (1): 100
n_features_to_select (1): 600
Nombre de variables restantes après l'étape 1: 600

step (2): 50
n_features_to_select (2): 300

Nombre de variables restantes après l'étape 2: 350
```

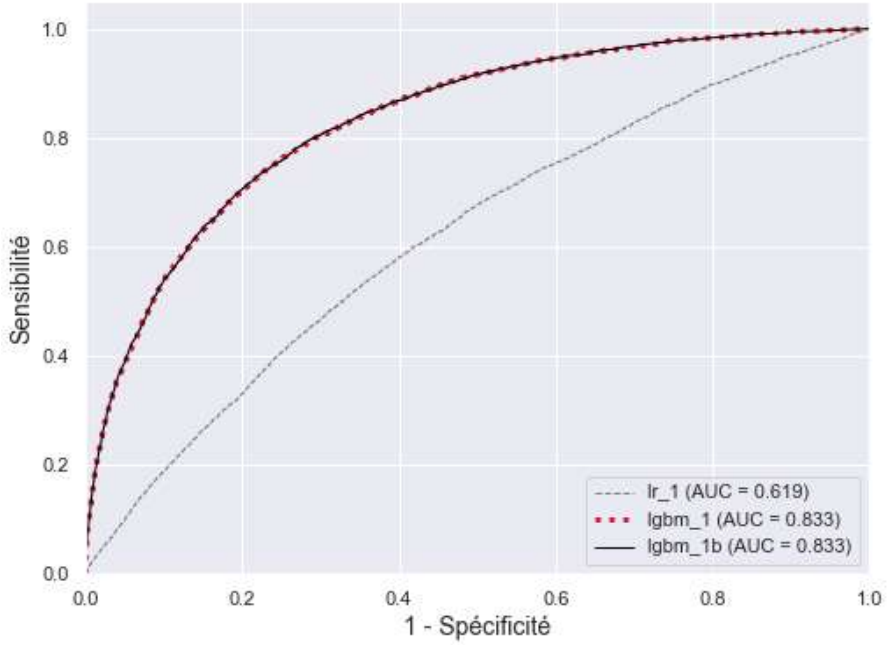
Réduction de dimensions

Jeux de données

- Entraînement
  - (246008, 350)
- Test avec étiquettes
  - (61503, 350)
- Test sans étiquette
  - (48744, 350)

Entraînement LightGBM

	Features dim	Model	Train score	Test score	Run time
2	(271398, 350)	lgbm_1b	0.936	0.833	194.0
1	(271398, 897)	lgbm_1	0.936	0.833	354.0
0	(271398, 897)	lr_1	0.631	0.619	143.0



Optimisation des paramètres

Méthodes

- Aléatoire / Grilles
  - ➢ Temps de traitement long
- Inférence bayésienne
  - Prise en compte des résultats des itérations précédentes
  - ➢ Plus performante et plus rapide
  - ➢ Implémentation **Hyperopt**

```
best = fmin(fn = objective,
            space = space,
            algo = tpe.suggest,
            max_evals = MAX_EVALS,
            trials = bayes_trials,
```

➢ Choix de la mesure d'évaluation
  - ➢ **Standard (AUC)**
  - ➢ **Custom (métier)**

Domain space

```
space = {'n_estimators': hp.quniform('n_estimators', 200, 800, 200),
         'class_weight': hp.choice('class_weight', [None, 'balanced']),
         'max_depth' : hp.quniform('max_depth', 2, 30, 2),
```

Fonction objective

```
def objective(params,
              model = LGBMClassifier(),
              x_train = features,
              y_train = labels,
              cv=skf,
              eval_metric = eval_metric_):
```

Minimisation perte (loss)

Fonction de substitution

Optimisation d'un modèle de probabilité

Sélection des prochaines valeurs à tester (principe "bayésien")

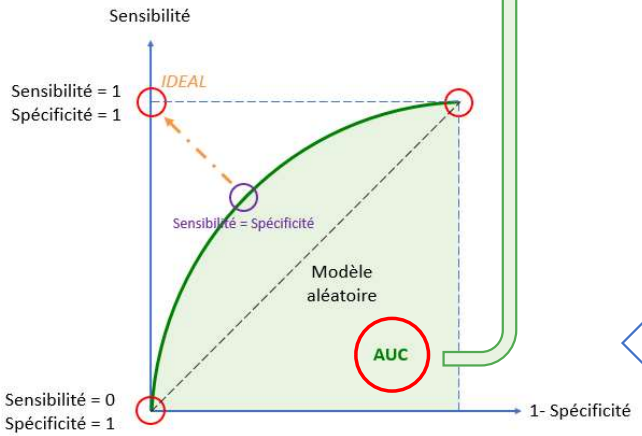
Résultats

	loss	threshold	n_estimator	class_weight	iteration	train_time
0	0.1763	0.200	400	None	13	625.0
1	0.1788	0.225	800	None	3	336.0
2	0.1874	0.100	800	None	5	1020.0
3	0.1874	0.300	400	None	8	340.0
4	0.1921	0.250	800	balanced	19	565.0

Optimisation des paramètres

```
def objective(params,
              model = LGBMClassifier(),
              x_train = features,
              y_train = labels,
              cv=skf,
              eval_metric = eval_metric_):
```

Dossier	Négatif Remboursé	Positif Non remboursé	Seuil	Prédiction	Réel	Vrai Positif	Vrai Négatif	Faux Positif	Faux Négatif
1	0,3	0,7	0,6	1	0			1	
2	0,6	0,4	0,6	0	0		1		
3	0,8	0,2	0,6	0	0		1		
4	0,1	0,9	0,6	1	1	1			
5	0,5	0,5	0,6	0	0		1		
6	0,3	0,7	0,6	1	0			1	
7	0,4	0,6	0,6	0	1				1
8	0,8	0,2	0,6	0	0		1		
9	0,6	0,4	0,6	0	0		1		
10	0,3	0,7	0,6	1	1	1			
						2	5	2	1



Cette approche est pertinente si on considère les éléments de la matrice de confusion de même importance mais…

Sensibilité = 0,67
Spécificité = 0,71
Précision = 0,5

	Réel Négatif		Réel Positif		
Prédiction Négatif	TN	5	FN	1	6
Prédiction Positif	FP	2	TP	2	4
		7		3	

Optimisation des paramètres

```
def objective(params,
              model = LGBMClassifier(),
              x_train = features,
              y_train = labels,
              cv=skf,
              eval_metric = eval_metric_):
```

Dossier	Négatif Remboursé	Positif Non remboursé	Seuil	Prédiction	Réel	Vrai Positif	Vrai Négatif	Faux Positif	Faux Négatif
1	0,3	0,7	0,6	1	0			1	
2	0,6	0,4	0,6	0	0		1		
3	0,8	0,2	0,6	0	0		1		
4	0,1	0,9	0,6	1	1	1			
5	0,5	0,5	0,6	0	0		1		
6	0,3	0,7	0,6	1	0			1	
7	0,4	0,6	0,6	0	1				1
8	0,8	0,2	0,6	0	0		1		
9	0,6	0,4	0,6	0	0		1		
10	0,3	0,7	0,6	1	1	1			
						2	5	2	1

Mais d'un point de vue métier, un crédit non remboursé coûte plus cher qu'un crédit non signé

	Réel Négatif		Réel Positif		
Prédiction Négatif	TN	5	FN	1	6
Prédiction Positif	FP	2	TP	2	4
		7		3	

Gain = TP * (0) + TN * (1) + FP * (0) + FN * (-10)

Optimisation des paramètres

Hyperopt

- Optimisation
  - AUC
  - F1-Score
  - Gain normalisé

- Résultats
  - Meilleurs paramètres

```
{'class_weight': 'balanced',
 'colsample_bytree': 0.8,
 'learning_rate': 0.0207747487935626,
 'max_depth': 18,
 'n_estimators': 600,
 'num_leaves': 76,
 'reg_alpha': 1.0,
 'reg_lambda': 0.30000000000000004,
 'solvability_threshold': 0.42500000000000004,
 'subsample': 0.8}
```

Profondeur

Nombre d'estimateurs

Nombre de feuilles

Seuil de classification

```
results_auc_score = optim(r_, roc_auc_score, 1)
results_hyperopt(results_auc_score).head()
```

	loss	threshold	n_estimator	class_weight	iteration	train_time
0	0.1763	0.200	400	None	13	625.0
1	0.1788	0.225	800	None	3	336.0
2	0.1874	0.100	800	None	5	1020.0
3	0.1874	0.300	400	None	8	340.0
4	0.1921	0.250	800	balanced	19	565.0

```
results_f1_score = optim(r_, f1_score, 1)
results_hyperopt(results_f1_score).head()
```

	loss	threshold	n_estimator	class_weight	iteration	train_time
0	0.2699	0.450	400	None	18	541.0
1	0.2843	0.825	400	None	3	391.0
2	0.2920	0.850	200	None	7	328.0
3	0.3015	0.550	800	balanced	4	1089.0
4	0.3166	0.525	600	balanced	1	594.0

```
results_g_norm = optim(r_, g_norm_, 1)
results_hyperopt(results_g_norm).head()
```

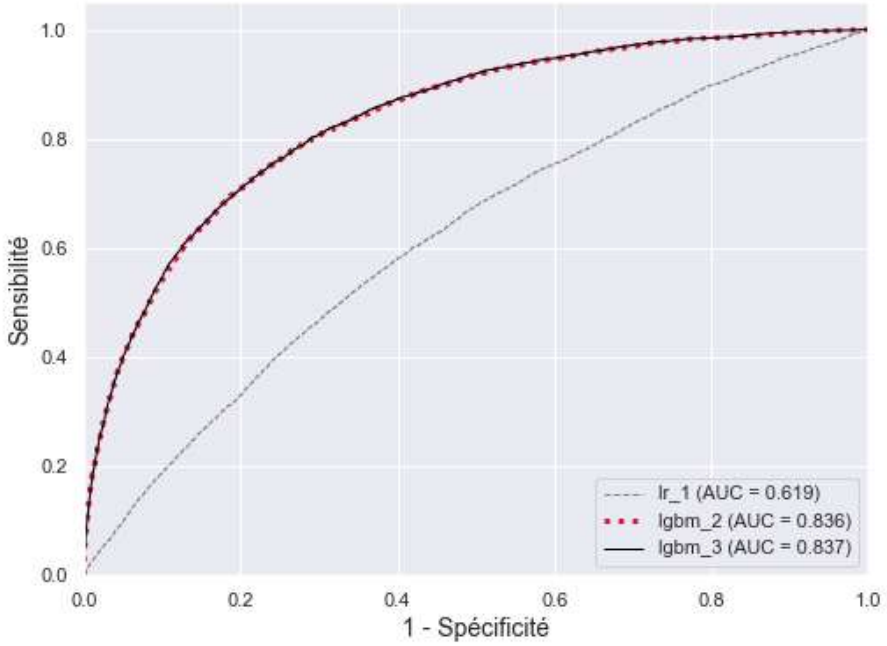
	loss	threshold	n_estimator	class_weight	iteration	train_time
0	0.1919	0.425	600	balanced	20	990.0
1	0.1939	0.275	600	balanced	13	336.0
2	0.2021	0.175	600	None	14	360.0
3	0.2080	0.075	400	balanced	10	490.0
4	0.2163	0.500	400	balanced	5	518.0

Optimisation des paramètres

Hyperopt

- Entraînement LightGBM
  - AUC (lgbm_2)
  - F1-Score (lgbm_2b)
  - Gain normalisé (lgbm_3)
- Résultats
  - Meilleurs scores

	Features dim	Model	Train score	Test score	Run time	
8	(271398, 350)	lgbm_3	0.961	0.837	1173.0	Gain
7	(271398, 350)	lgbm_2b	0.947	0.836	574.0	F1-Score
6	(271398, 350)	lgbm_2	0.959	0.836	658.0	AUC
2	(271398, 897)	xgb_1	0.955	0.834	3934.0	Défaut
5	(271398, 350)	lgbm_1b	0.936	0.833	194.0	Défaut (350 var)
4	(271398, 897)	lgbm_1	0.936	0.833	354.0	Défaut (897 var)
3	(271398, 897)	catb_1	0.942	0.831	2499.0	Défaut
1	(271398, 897)	rf_1	1.000	0.790	971.0	Défaut
0	(271398, 897)	lr_1	0.631	0.619	143.0	Défaut

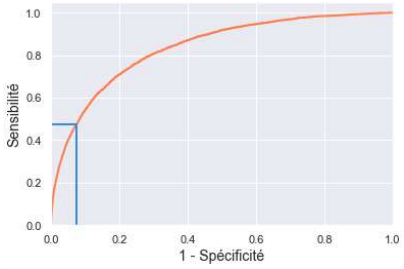


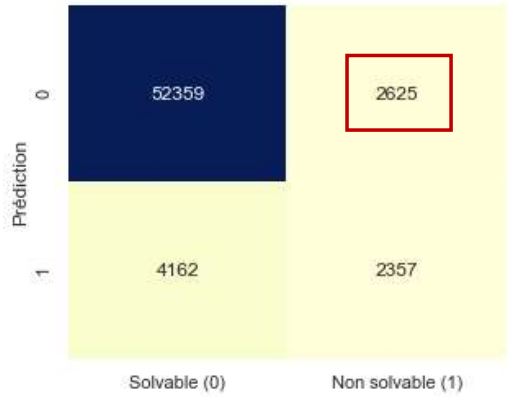
Optimisation des paramètres

Hyperopt

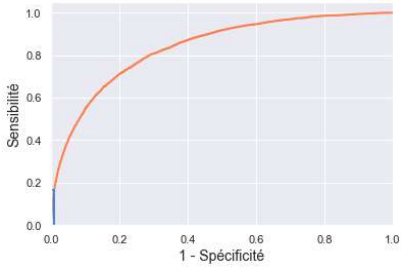
- Entraînement LightGBM
- Résultats
  - Seuil optimal
  - Matrice de confusion

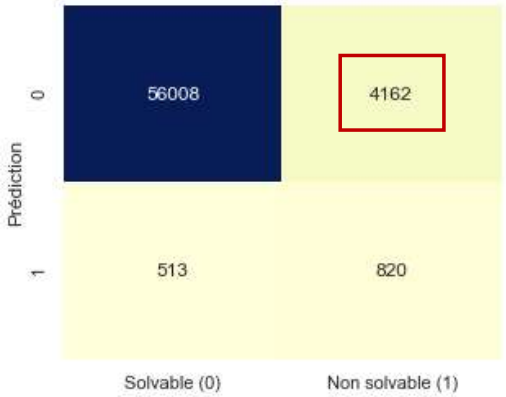
AUC



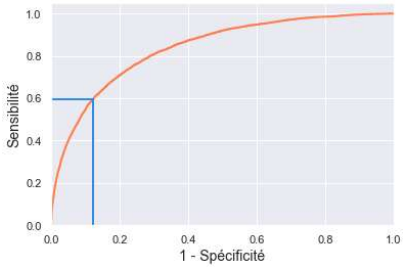


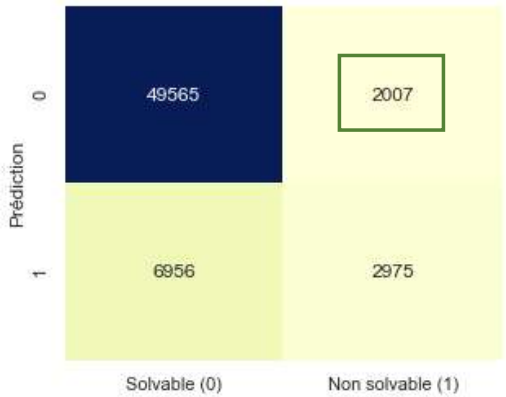
F1-Score





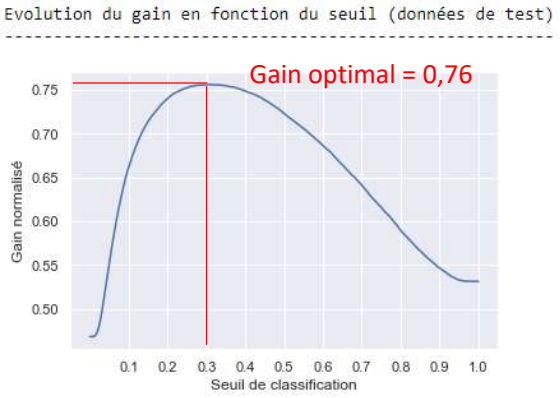
Gain



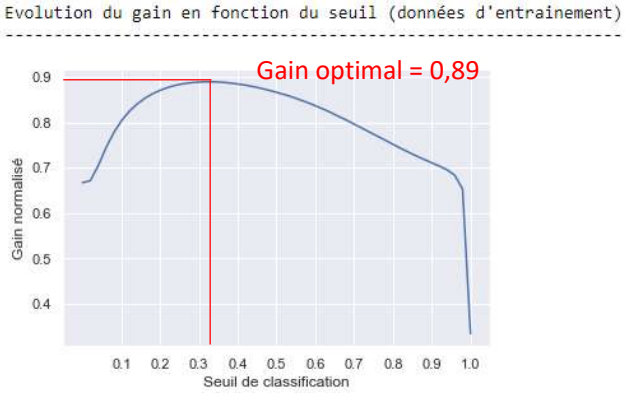


Evolution des indicateurs

Gain normalisé



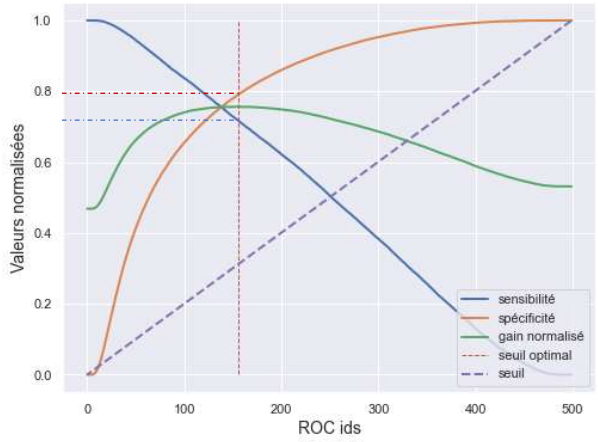
Seuil optimal = 0,31

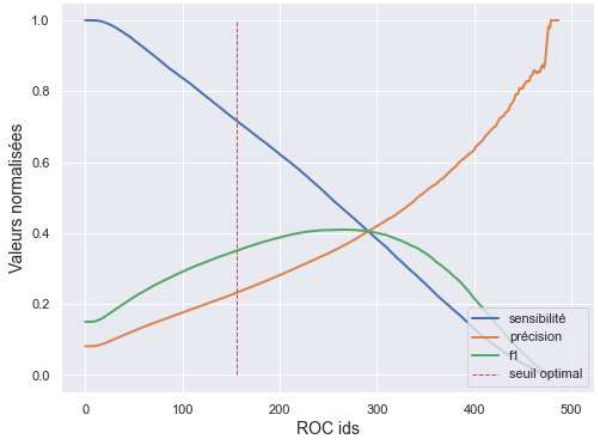


Seuil optimal = 0,33

Mesures ROC

Spécificité = 0,79
Sensibilité = 0,72





Importance des variables

Méthode utilisée pour la réduction de dimension avec RFE

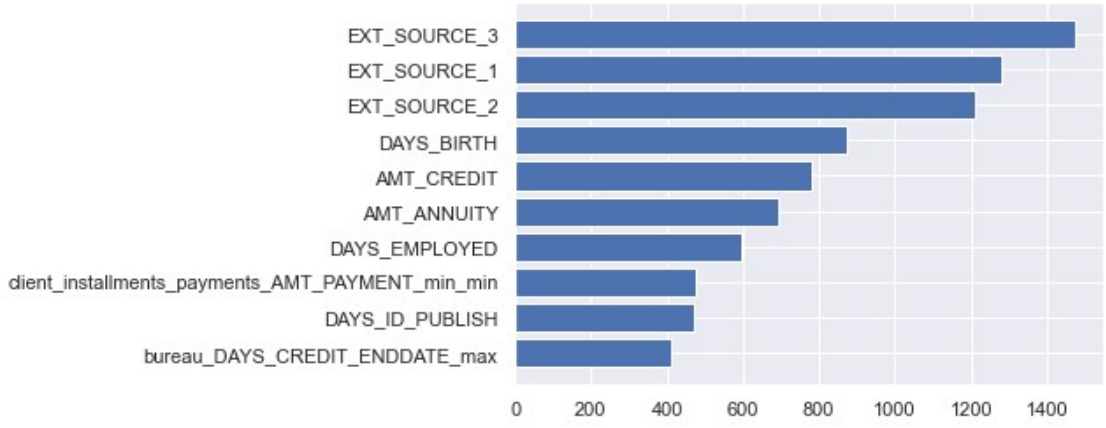
- Sélection récursive des variables
  - Score par lot

	Variable	Importance
0	CNT_CHILDREN	1
1	AMT_INCOME_TOTAL	1
2	AMT_CREDIT	1
3	AMT_ANNUITY	1
4	REGION_POPULATION_RELATIVE	1

	Variable	Importance
276	bureau_AMT_CREDIT_SUM_LIMIT_mean	2
166	ORGANIZATION_TYPE_Police	2
191	FONDKAPREMONT_MODE_regoperaccount	2
...	...	...
300	client_bureau_balance_MONTHS_BALANCE_max_mean	6
140	ORGANIZATION_TYPE_Culture	6

Variables importantes d'un modèle

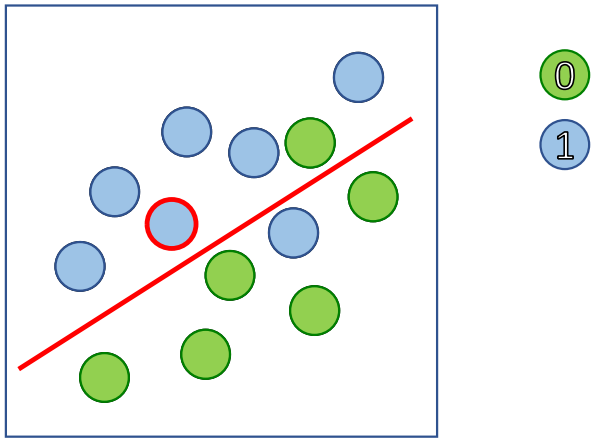
- Identification des variables importantes
  - Score par variable



Interprétabilité locale

LIME

- **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
  - Génère des individus proches
  - Prédiction
  - Modèle linéaire local
  - ➢ Facile à interpréter



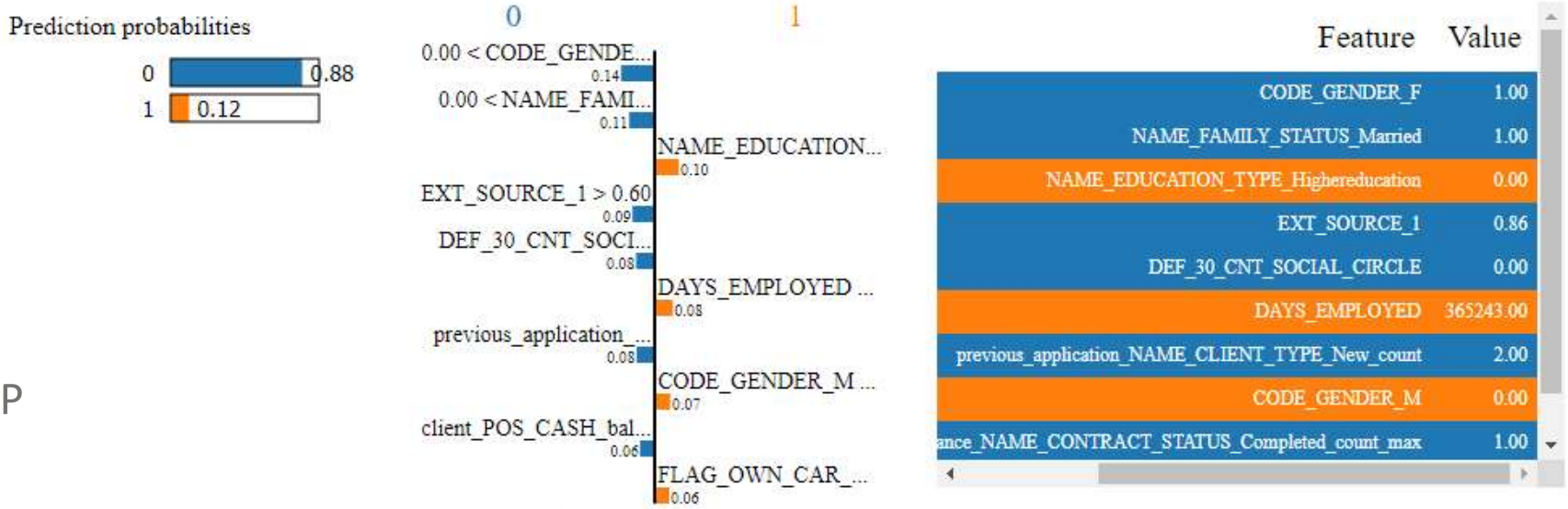
SHAP

- **Sh**apley **A**dditive ex**P**lanations
  - Calcul de la valeur de Shapley pour toutes les variables
    - Moyenne de l'impact d'une variable sur toutes les combinaisons de variables possibles
  - ➢ La somme des effets de chaque variable explique la prédiction

Interprétabilité locale

LIME

```
Dossier de crédit = 124856
Prédiction proba pour l'individu sélectionné: 0.1232
Prédiction binaire pour l'individu sélectionné: 0
Valeur réelle pour l'individu sélectionné: 0
```

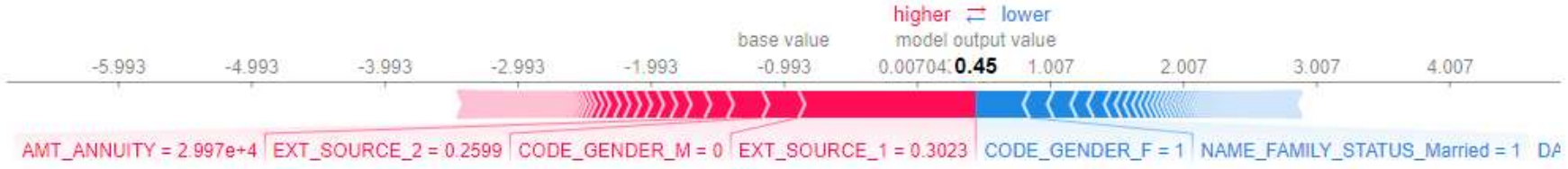


SHAP

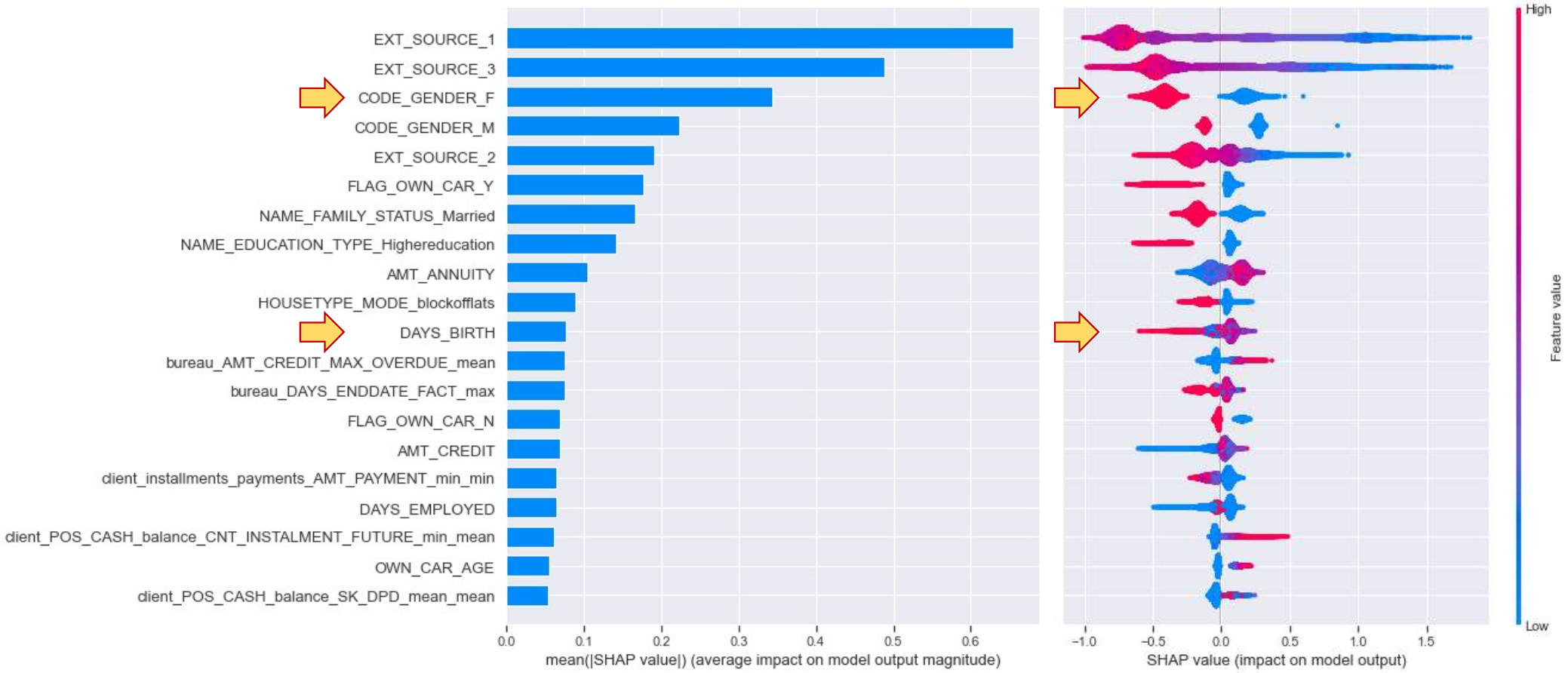
Interprétabilité SHAP (locale)

```
Dossier de crédit = 307474
Prédiction proba pour l'individu sélectionné: 0.052
Prédiction binaire pour l'individu sélectionné: 0
Valeur réelle pour l'individu sélectionné: 0
```


```
Dossier de crédit = 402778
Prédiction proba pour l'individu sélectionné: 0.6221
Prédiction binaire pour l'individu sélectionné: 1
Valeur réelle pour l'individu sélectionné: 1
```



Interprétabilité SHAP (globale)



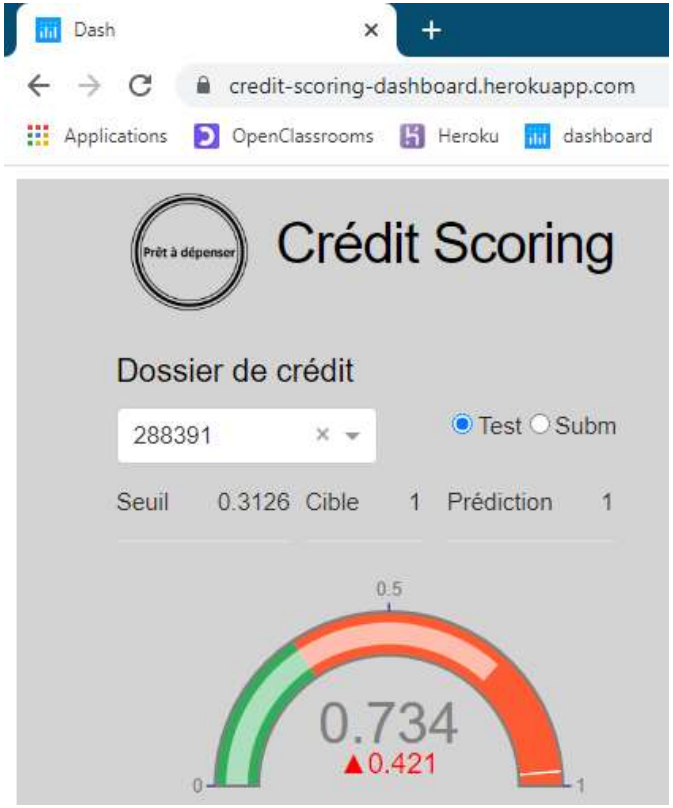
Développement d'une application web

Technologies

- Dash (Flask)
  - Langage Python
  - Interface web
- Heroku
  - Déploiement

Fonctionnalités

- Scoring
  - Prédiction automatique de la classe d 'un dossier
- Interprétabilité
  - Globale: variables importantes
  - Locale: SHAP
- Analyse
  - Simulations
  - Comparaison de dossiers



Architecture et déploiement

Modélisation

- Analyse
  - P7_01_analyse.ipynb
- Machine Learning
  - P7_02_scoring.ipynb

Application

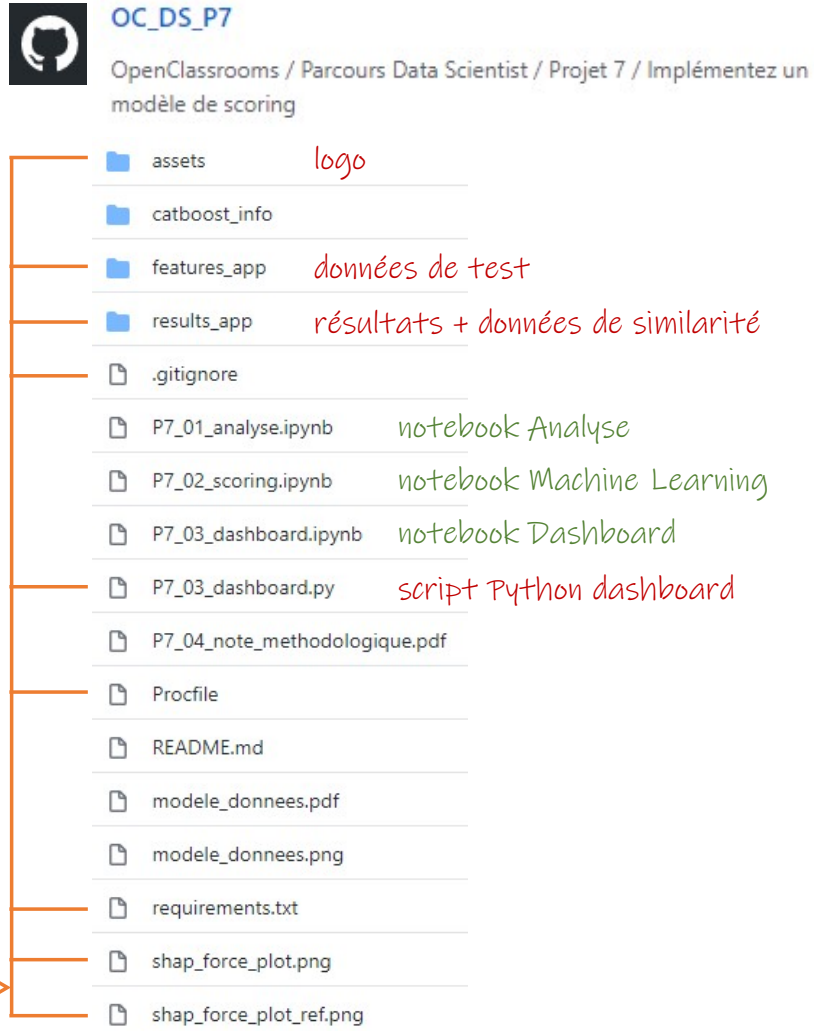
- Traitement spécifique
  - P7_03_dashboard.ipynb
    - Traitement des données
    - Fonction globale
- Application Dash
  - P7_03_dashboard.ipynb
    - Application Dash (instanciation)
    - Dash layout (présentation)
    - Dash callbacks (interactivité)

Déploiement

- Environnement virtuel

```
Eric Wendling@DESKTOP-F14RHK5 MINGW64 ~/OC_DS_P7 (master)
$ source sco/Scripts/activate
(sco)
```

- Dépendances
  - Librairies Python
  - Dash
  - Plotly
  - gunicorn
- Initialisation
  - .gitignore
  - requirements.txt
  - Procfile
  - P7_03_dashboard.py
- Repository GitHub
  - ➢ OC_DS_P7
- Heroku
  - ➢ Initialisation
- Déploiement
  1) GitHub
  2) Heroku



Crédit Scoring

Dossier de crédit

441310 × ▾ ◉ Test ○ Subm

Seuil 0.3126 Cible 0 Prédiction 0

Score

Le score du dossier est la probabilité que le crédit ne soit pas remboursé.

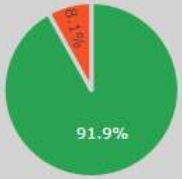
Le score du dossier en cours est situé entre le meilleur score (plus petit) et le moins bon score (plus grand).

Indicateurs

AUC	0.837	Sensibilité	0.72
Gain	0.756	Spécificité	0.79
F-Mesure	0.35	Précision	0.23

Ratios prédictions

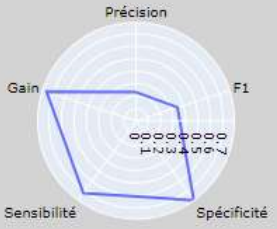
Ratios réels



Analyse statistique

Le seuil de classification permet de classer un dossier selon son score. Si le score est supérieur au seuil, le dossier présente des risques.

On peut modifier le statut du dossier en faisant varier le seuil et observer les conséquences sur d'autres indicateurs avec les simulations de seuil et de gain.



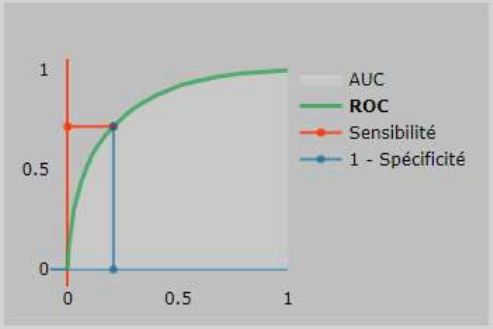
SEUIL OPTIMAL

	Réel N	Réel P
N	44763	1405
P	11758	3577

	Réel N	Réel P
N	21895	244
P	34626	4738

	Réel N	Réel P
N	55137	3604
P	1384	1378

Simulation seuil



Seuil sélectionné: 0.3126

Simulation gain

Gain sélectionné: 0.65


Crédit Scoring
Dossier de crédit
441310
Test
Subm
Seuil 0.3126 Cible 0 Prédiction 0
0.313
▼0
Score
Le score du dossier est la probabilité que le crédit ne soit pas remboursé.
400843_ 0.99297
441310_ 0.31258 0.68742
232507_ 0.97278
_N
_P
Le score du dossier en cours est situé entre le meilleur score (plus petit) et le moins bon score (plus grand).
Indicateurs
AUC 0.837
Sensibilité 0.72
Gain 0.756
Spécificité 0.79
F-Mesure 0.35
Précision 0.23
Ratios prédictions
24.9%
75.1%
Ratios réels
8.1%
91.9%
Analyse statistique
Le seuil de classification permet de classer un dossier selon son score. Si le score est supérieur au seuil, le dossier présente des risques.
On peut modifier le statut du dossier en faisant varier le seuil et observer les conséquences sur d'autres indicateurs avec les simulations de seuil et de gain.
Précision
Gain
F1
Sensibilité
Spécificité
SEUIL OPTIMAL
Réel N Réel P
N 44763 1405
P 11758 3577
Simulation seuil
AUC
ROC
Sensibilité
1 - Spécificité
Seuil sélectionné: 0.3126
Réel N Réel P
N 21895 244
P 34626 4738
Réel N Réel P
N 55137 3604
P 1384 1378
Simulation gain
Gain
Marge de gain
Seuil optimal
Gain optimal
Seuil min
Seuil max
Gain sélectionné: 0.65

Interprétabilité

Variables

Les informations des dossiers "courant" et de "référence" sont affichées avec leur taux de variation et leur importance normalisée (0 à 100).

Dossier courant: 441310 Dossier de référence: 434446 ×

Variable	Courant	Référence	Variation	Importance
filter data...				
EXT_SOURCE_3	0.11	0.492	78	100
EXT_SOURCE_1	0.809	0.733	9	85
EXT_SOURCE_2	0.096	0.815	88	79
DAYS_BIRTH	-17987	-19950	10	53
AMT_CREDIT	790830	888840	11	46
AMT_ANNUITY	52978.5	29016	45	39
DAYS_EMPLOYED	-4407	-5639	22	31
client_installments_payments_AMT_PAYMENT_min_min	2896.83	6299.865	54	22
DAYS_ID_PUBLISH	-1477	-3383	56	22
bureau_DAYS_CREDIT_ENDDATE_max	30962	1391	96	17

« ‹ 1 › »

Analyse métier

Le dossier "courant" est comparé avec des dossiers similaires, sur la base du score ou des variables.

○ Score ◉ Variables Degré de similarité: 10

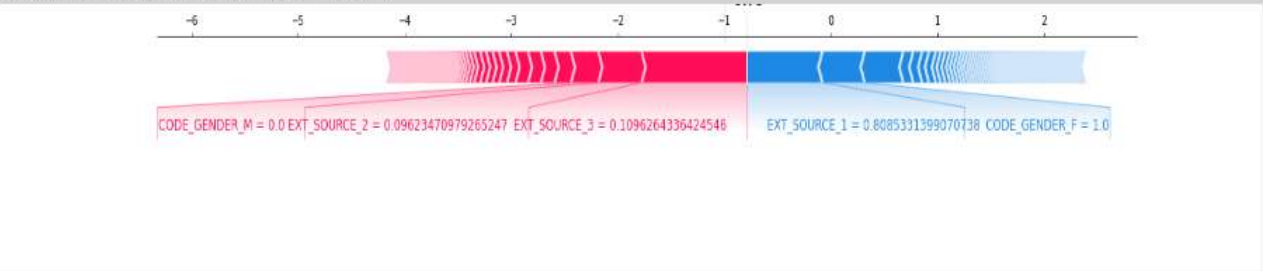
Variable	Courant	Réf. 1	Réf. 2	Réf. 3	Réf. 4	Importance
filter data...						
SK_ID_CURR	441310_	211109_	309395_	325159_	434446_	
_P	0.31258	0.38292	0.45768	0.67975	0.05858	
_Pred	0	1	1	1	0	
_True	0	0	0	1	0	
EXT_SOURCE_3	0.11	0.361	0.243	0.303	0.492	100
EXT_SOURCE_1	0.809	0.436	0.393	0.373	0.733	85
EXT_SOURCE_2	0.096	0.554	0.202	0.419	0.815	79
DAYS_BIRTH	-17987	-11583	-16050	-18285	-19950	53
AMT_CREDIT	790830	746280	852088.5	832977	888840	46
AMT_ANNUITY	52978.5	58963.5	33921	42660	29016	39

« ‹ 1 › »

Interprétabilité

Interprétation par dossier

Dossier courant: 441310 (indice = 18786)



Dossier de référence: 434446 (indice = 31873)

Mise à jour: Sun Sep 20 22:23:18 2020

Connexion: 2020-09-20 22:06:01.809209

Crédit Scoring

Dossier de crédit

441310 × ▾ ◉ Test ○ Subm

Seuil 0.3126 Cible 0 Prédiction 0

Score

Le score du dossier est la probabilité que le crédit ne soit pas remboursé.

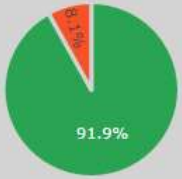
Le score du dossier en cours est situé entre le meilleur score (plus petit) et le moins bon score (plus grand).

Indicateurs

AUC	0.837	Sensibilité	0.72
Gain	0.756	Spécificité	0.79
F-Mesure	0.35	Précision	0.23

Ratios prédictions

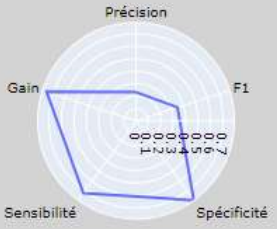
Ratios réels



Analyse statistique

Le seuil de classification permet de classer un dossier selon son score. Si le score est supérieur au seuil, le dossier présente des risques.

On peut modifier le statut du dossier en faisant varier le seuil et observer les conséquences sur d'autres indicateurs avec les simulations de seuil et de gain.



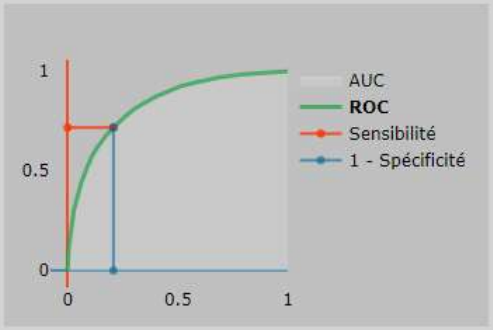
SEUIL OPTIMAL

	Réel N	Réel P
N	44763	1405
P	11758	3577

	Réel N	Réel P
N	21895	244
P	34626	4738

	Réel N	Réel P
N	55137	3604
P	1384	1378

Simulation seuil



Seuil sélectionné: 0.3126

Simulation gain

Gain sélectionné: 0.65

Crédit Scoring

Dossier de crédit

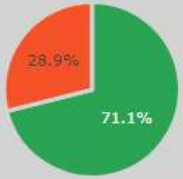
244476 × ▾ ◉ Test ○ Subm

Seuil 0.0902 Cible 0 Prédiction 0

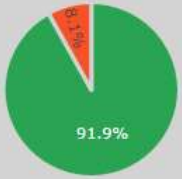
Indicateurs

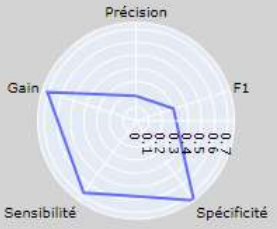
AUC	0.786	Sensibilité	0.68
Gain	0.711	Spécificité	0.75
F-Mesure	0.3	Précision	0.19

Ratios prédictions



Ratios réels

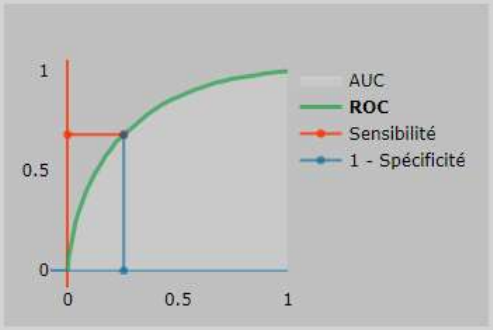




SEUIL OPTIMAL

	Réel N	Réel P
N	42116	1586
P	14405	3396

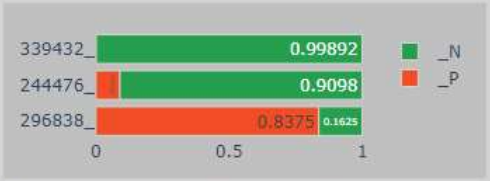
Simulation seuil



Seuil sélectionné: 0.0902

Score

Le score du dossier est la probabilité que le crédit ne soit pas remboursé.



Le score du dossier en cours est situé entre le meilleur score (plus petit) et le moins bon score (plus grand).

Analyse statistique

Le seuil de classification permet de classer un dossier selon son score. Si le score est supérieur au seuil, le dossier présente des risques.

On peut modifier le statut du dossier en faisant varier le seuil et observer les conséquences sur d'autres indicateurs avec les simulations de seuil et de gain.

	Réel N	Réel P
N	24599	497
P	31922	4485

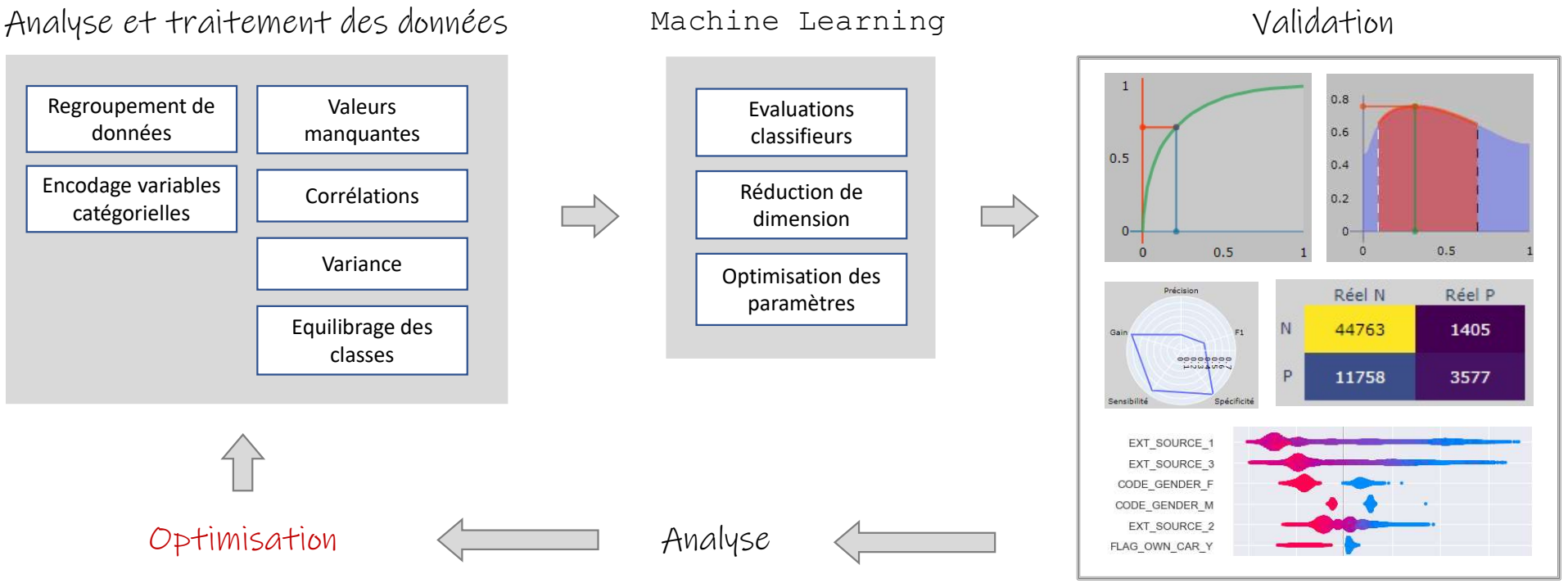
	Réel N	Réel P
N	53041	3335
P	3480	1647

Simulation gain

Gain sélectionné: 0.65


Crédit Scoring
Dossier de crédit
441310
Test
Subm
Seuil 0.3126 Cible 0 Prédiction 0
0.313
▼0
Indicateurs
(0.786) AUC 0.837
(0.711) Gain 0.756
(0.3) F-Mesure 0.35
Sensibilité 0.72 (0.68)
Spécificité 0.79 (0.75)
Précision 0.23 (0.19)
Ratios prédictions
(28,9%)
24.9%
75.1%
Ratios réels
8.1%
91.9%
Précision
Gain
F1
Sensibilité
Spécificité
SEUIL OPTIMAL
Réel N Réel P
N 44763 1405
P 11758 3577
Réel N Réel P
N 42116 1586
P 14405 3396
Réel N Réel P
N 21895 244
P 34626 4738
Réel N Réel P
N 55137 3604
P 1384 1378
Simulation seuil
AUC
ROC
Sensibilité
1 - Spécificité
Seuil sélectionné: 0.3126
Score
Le score du dossier est la probabilité que le crédit ne soit pas remboursé.
400843_ 0.99297
441310_ 0.31258 0.68742
232507_ 0.97278
_N
_P
Le score du dossier en cours est situé entre le meilleur score (plus petit) et le moins bon score (plus grand).
Analyse statistique
Le seuil de classification permet de classer un dossier selon son score. Si le score est supérieur au seuil, le dossier présente des risques.
On peut modifier le statut du dossier en faisant varier le seuil et observer les conséquences sur d'autres indicateurs avec les simulations de seuil et de gain.
Simulation gain
Gain
Marge de gain
Seuil optimal
Gain optimal
Seuil min
Seuil max
Gain sélectionné: 0.65

8 Conclusion



- Optimisation
  - Technique
    - Paramètres des modèles
    - Similarité des dossiers
  - Métier (traitement des données)
- Application
  - Analyse des besoins métier
    - Fonction de coût
  - Analyse fonctionnelle
- Interprétabilité
  - ➢ SHAP (Global + Local)
  - Déploiement web ?
  - Explorer d'autres solutions