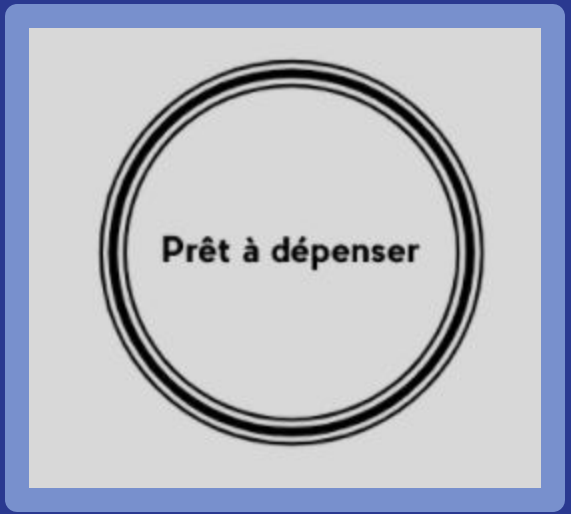
Construction d’un modèle de scoring

Prédire la solvabilité d’un client



Projet 4 - Sylwia Bankowska - Avril 2022

Sommaire

1. Compréhension de la problématique métier
2. Description du jeu de données
3. Transformation du jeu de donnée
4. Comparaison et synthèse des résultats pour les modèles utilisés
5. Interprétabilité du modèle
6. Conclusion

Problématique métier

Objectifs du projet

Développer

un algorithme de classification

pour identifier les clients solvables ou non

Fournir

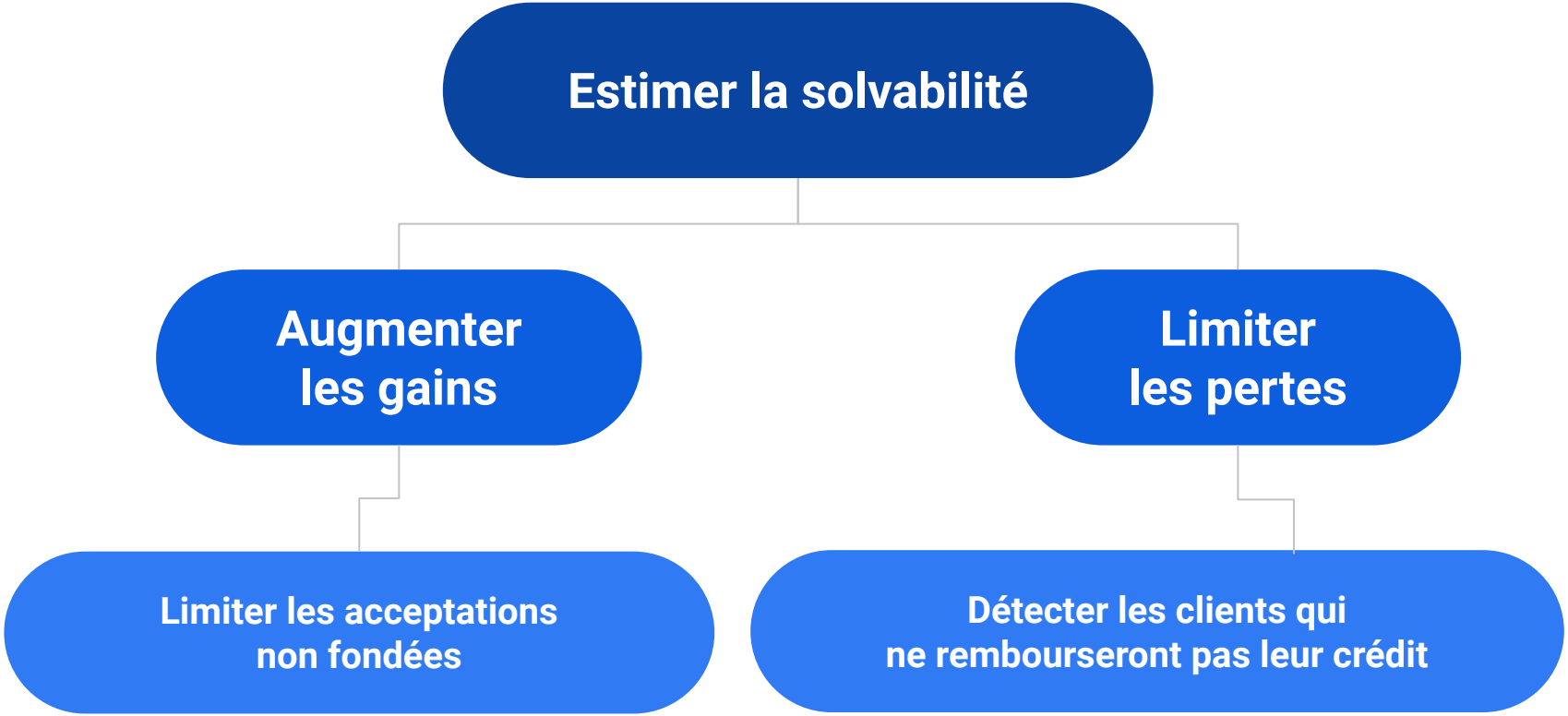
une mesure de l'importance

des variables

qui ont poussé le modèle

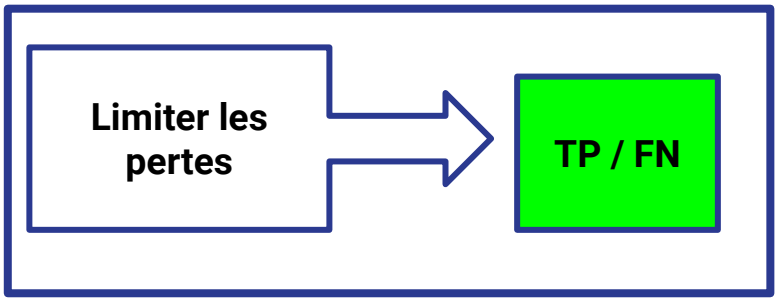
à donner une probabilité à un client

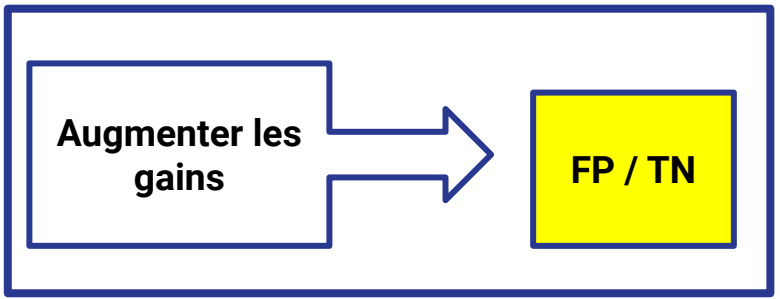
Enjeux


Estimer la solvabilité
Augmenter les gains
Limiter les pertes
Limiter les acceptations non fondées
Détecter les clients qui ne rembourseront pas leur crédit

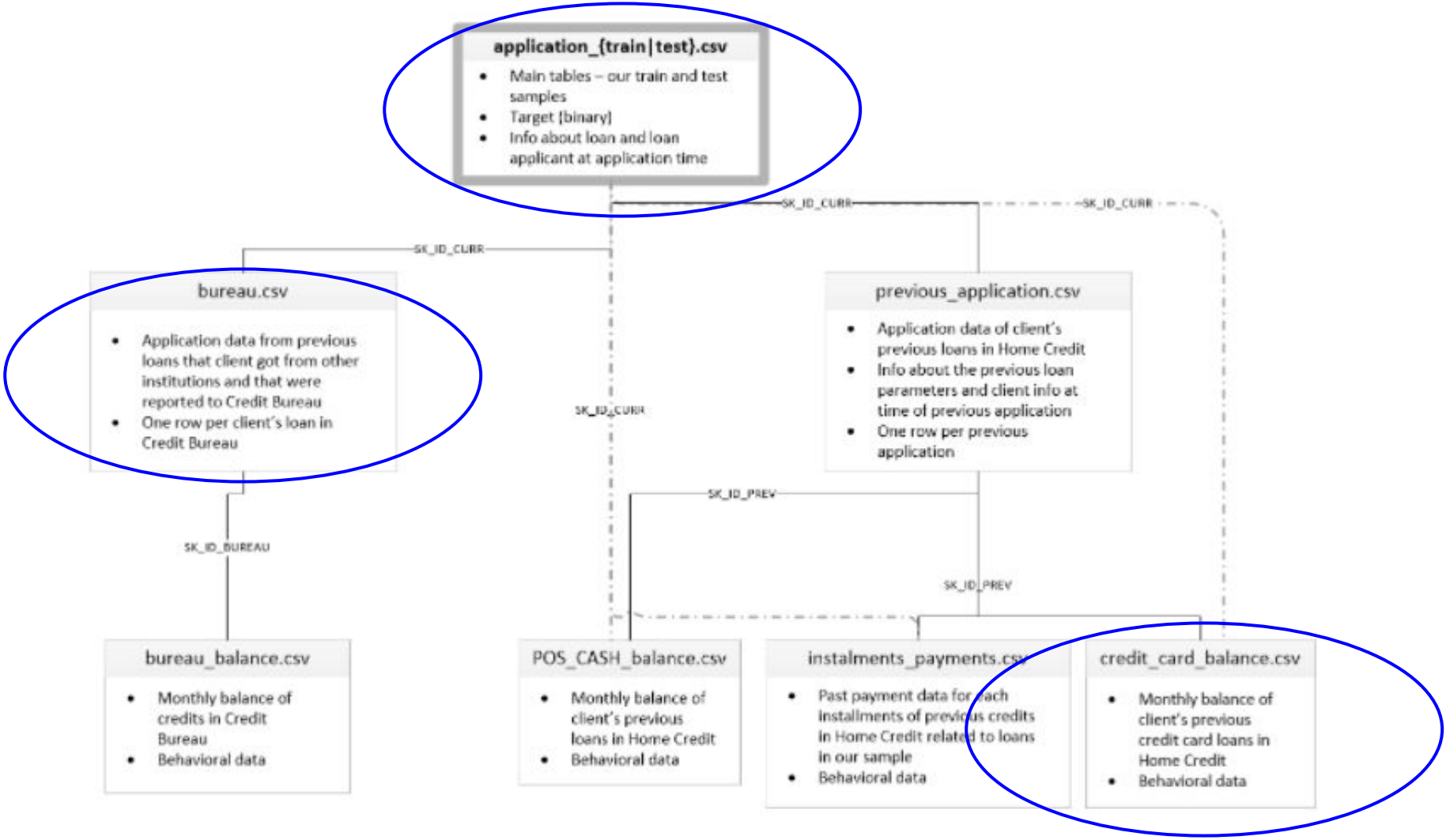
Matrice de confusion

	CLIENT A RISQUE	CLIENT SOLVABLE
REFUS DE CRÉDIT	True Positive	False Positive
ACCEPTATION DE CRÉDIT	False Negative	True Negative





Description du jeu de données


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_CURR
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Agrégation de données

application_train
- 307511 lignes
- 122 variables

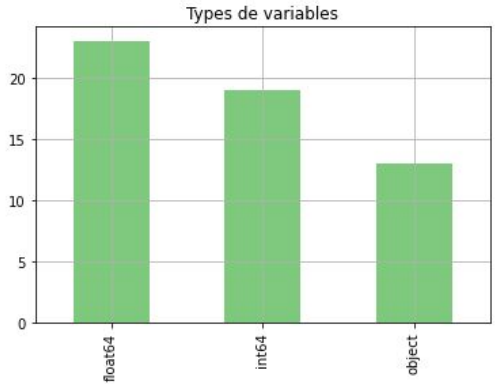
bureau
- 1 716 428 lignes
- 17 variables

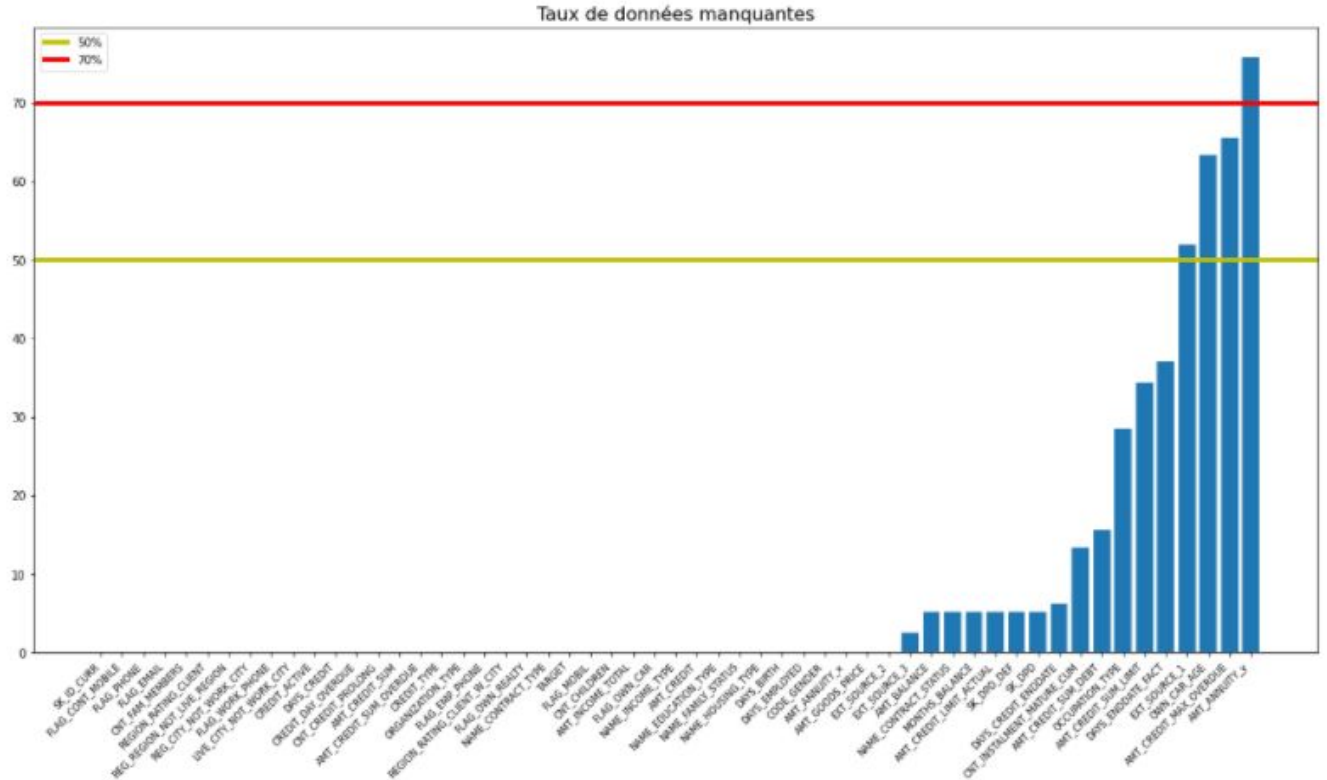
credit_card_balance
- 3 840 312 lignes
- 23 variables

DATASET FINAL
- 3 912 016 lignes
- 55 variables

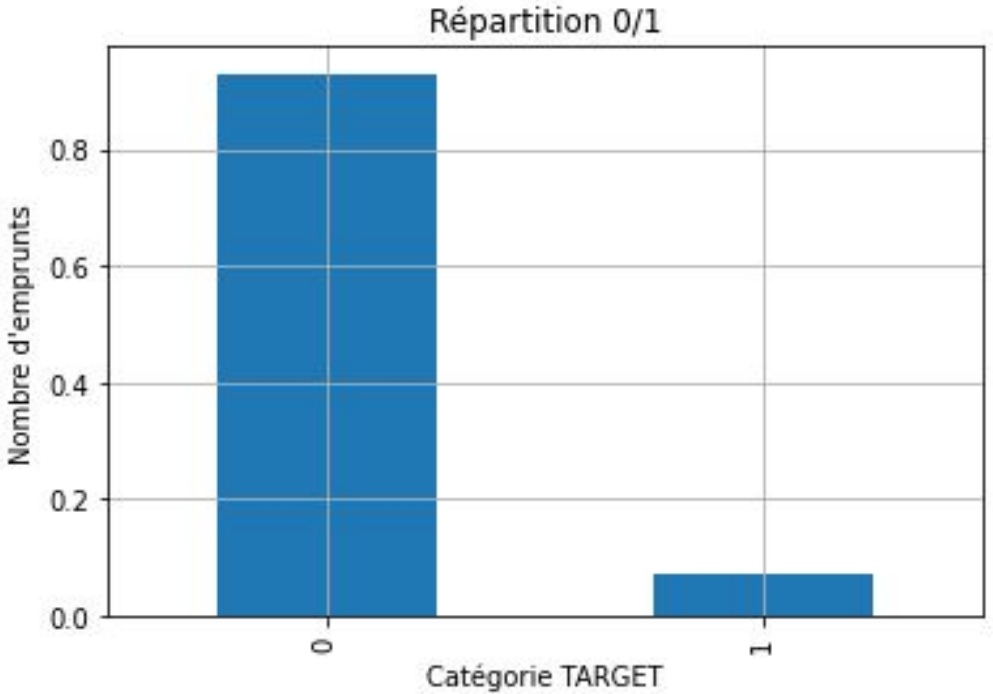
- EDA : 10% => 391 202
- MOD : 40% => 152 285

Analyse univariée





Variable cible



- Classe majoritaire
  - 93% de clients 0
- Classe minoritaire
  - 7% de clients 1

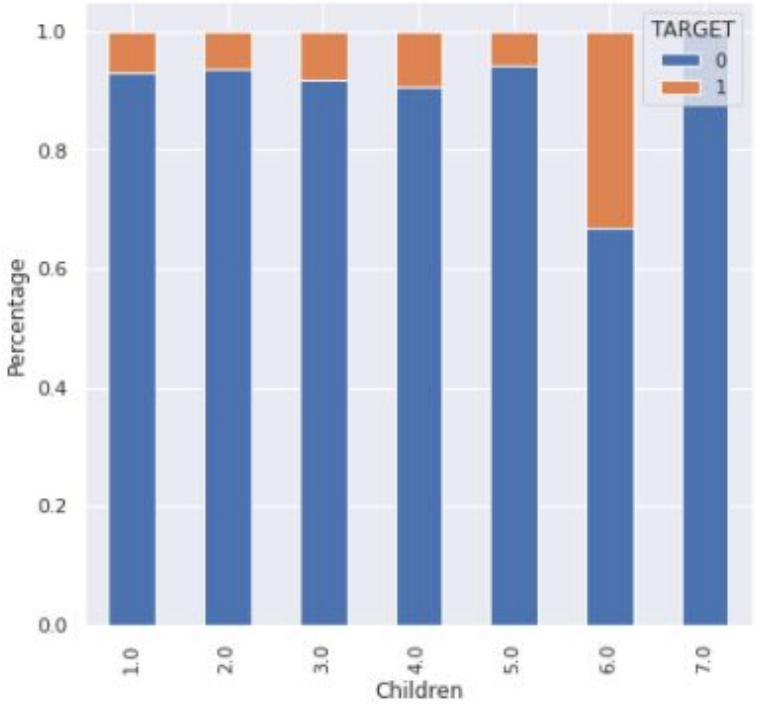
Anomalies

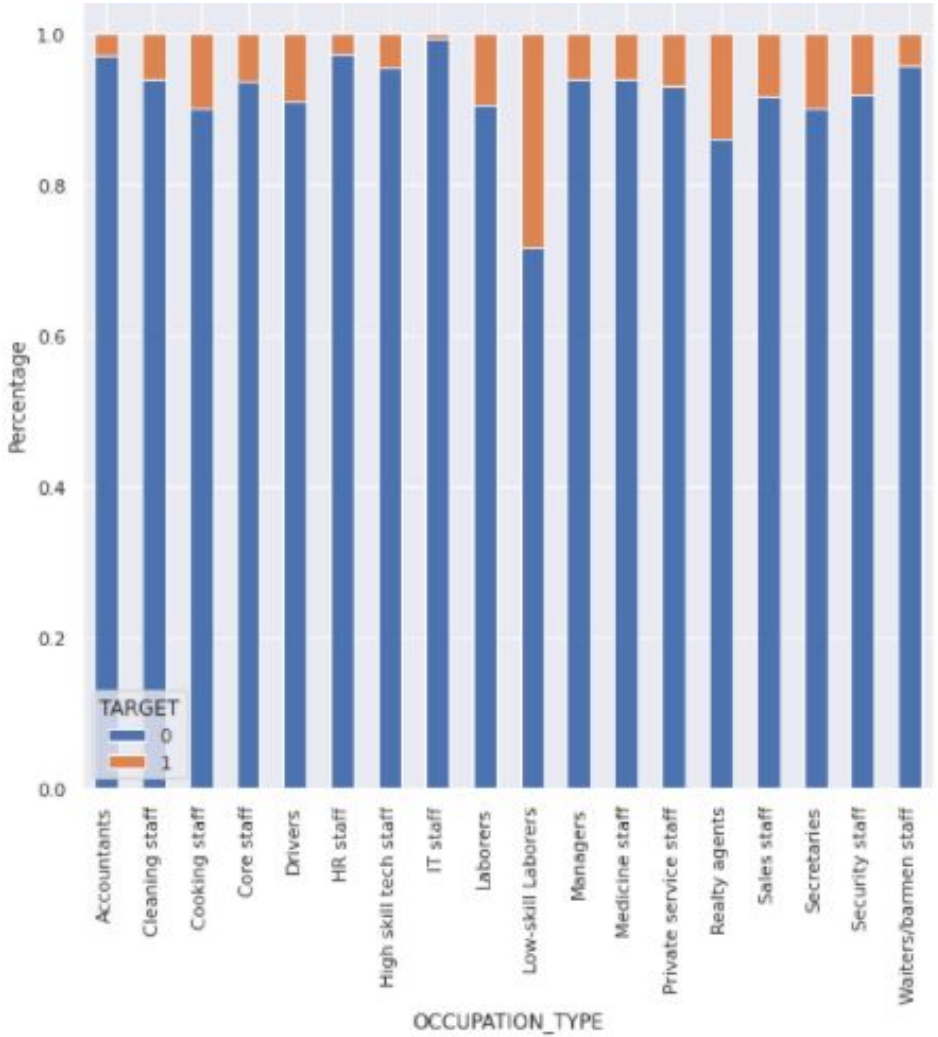
		Feature	Correction	Solution apportée

1	**Chiffres négatifs**	DAYS_BIRTH
		DAYS_EMPLOYED
			✓	• Conversion en chiffres positifs :
				  ○ diviser par 365
				  ○ multiplier par -1

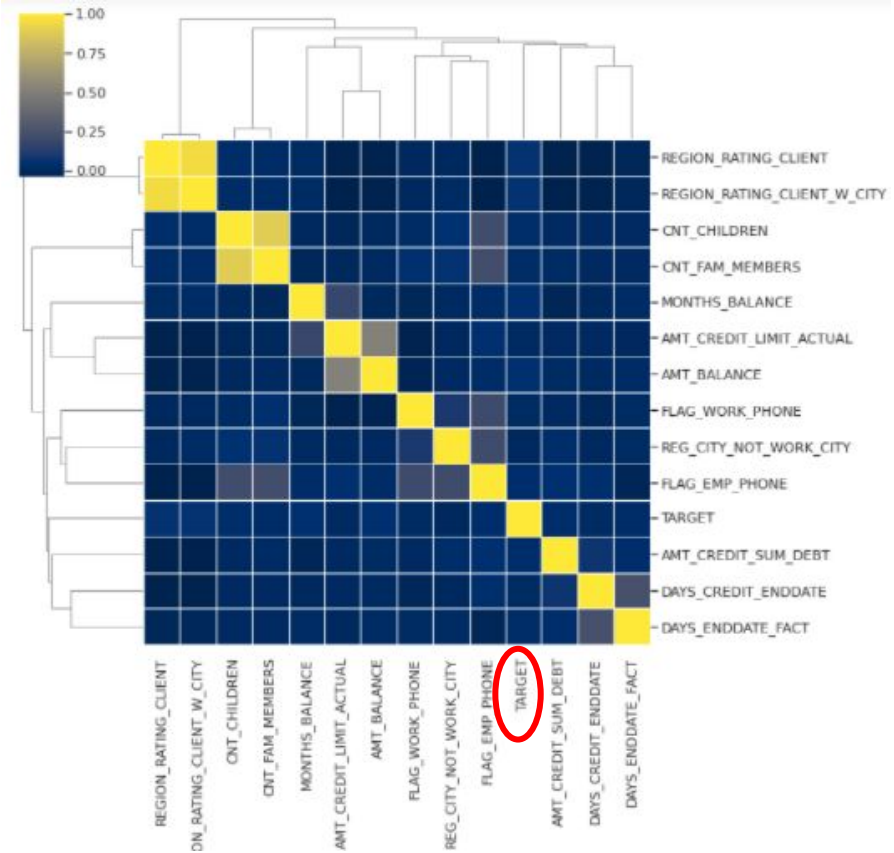
2	**Présence des nan**		✓	• Imputation par médiane / mode • Suppression de features de plus de 60% de nan
3	**Présence des inconnus**	ORGANIZATION_TYPE CODE_GENDER	✓	• “XNA” • Imputation par mode
4	**Présence des doublons**		✓	• Suppression
5	**Présence des outliers**		✓	• Remplacement par médiane

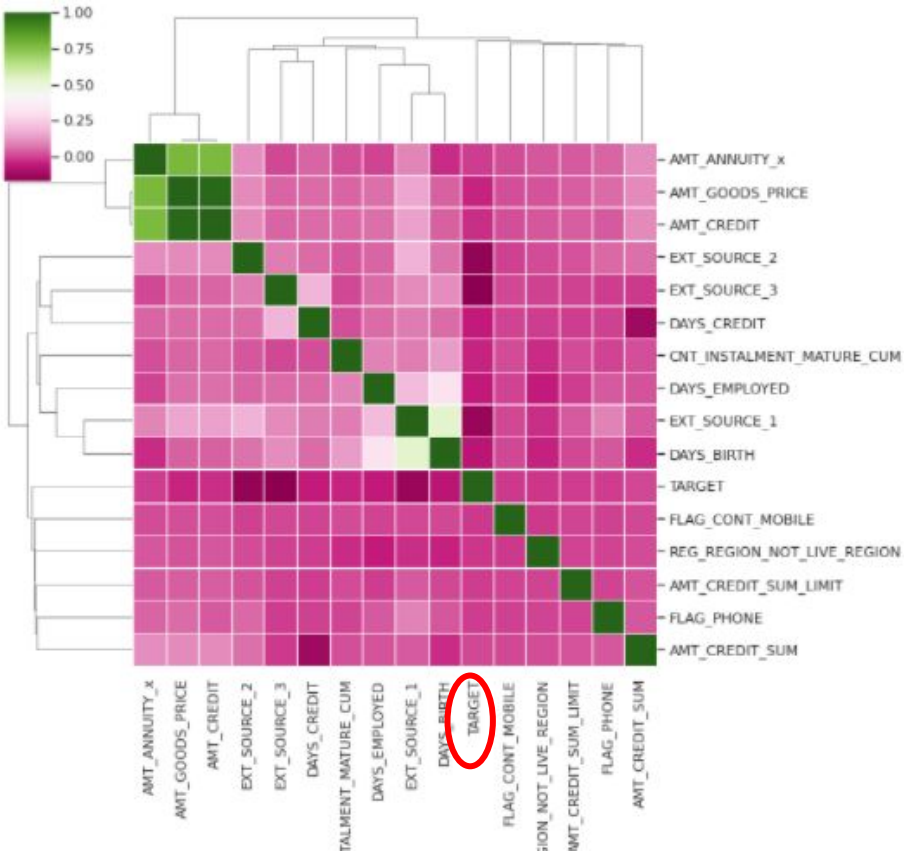
Analyse bivariée



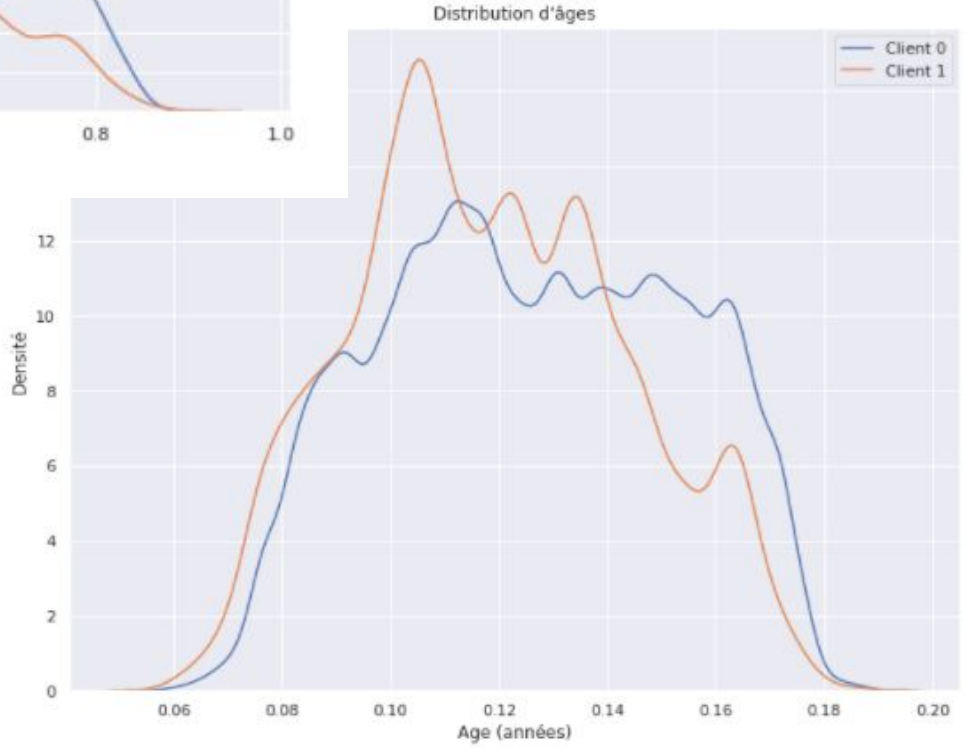


Corrélations



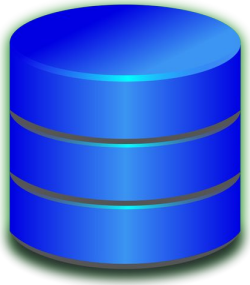


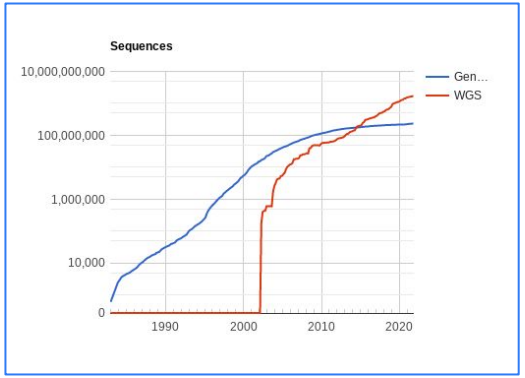

Distribution of EXT_SOURCE_3 by Target Value
Client 0
Client 1
Density
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
EXT_SOURCE_3


Distribution d'âges
Client 0
Client 1
Densité
12
10
8
6
4
2
0
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0.20
Age (années)

Transformation du jeu de données

Preprocessing





1. **Imputation des NaN** — Identification et élimination des nan
2. **Encodage** — One-Hot
3. **Feature engineering** — Création de nouvelles variables métier
4. **Feature selection** — Détermination de l'importance et choix de variables
5. **Normalisation** — Centrage d'une variable à zéro et de standardisation de la variance à 1

Imputation des NaN

1. **Variables catégorielles**
   - Remplissage par mode
   - fillna()
2. **Variables numériques**
   - Remplissage par médiane
   - SimpleImputer()

Encodage

id	color
1	red
2	blue
3	green
4	blue

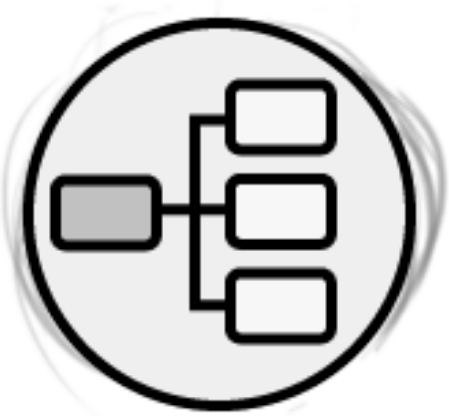
id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

```
NAME_CONTRACT_TYPE------------------------------- 2
CODE_GENDER------------------------------------- 3
FLAG_OWN_CAR------------------------------------ 2
FLAG_OWN_REALTY--------------------------------- 2
NAME_INCOME_TYPE-------------------------------- 6
NAME_EDUCATION_TYPE----------------------------- 5
NAME_FAMILY_STATUS------------------------------ 5
NAME_HOUSING_TYPE------------------------------- 6
OCCUPATION_TYPE--------------------------------- 19
ORGANIZATION_TYPE------------------------------- 58
CREDIT_ACTIVE----------------------------------- 4
CREDIT_TYPE------------------------------------- 13
NAME_CONTRACT_STATUS---------------------------- 8
```

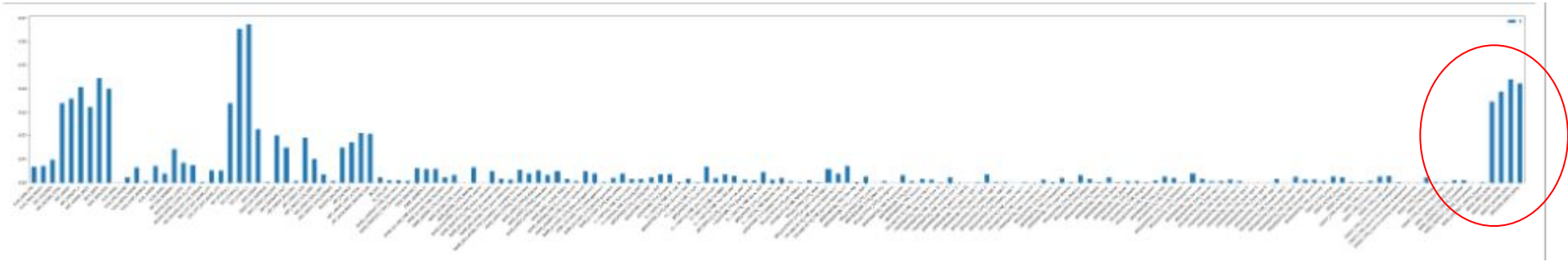
```
trainset.FLAG_OWN_CAR.unique(), trainset.FLAG_OWN_REALTY.unique()
```

```
(array(['Y', 'N'], dtype=object), array(['Y', 'N'], dtype=object))
```

Feature engineering



- Ratio **montant emprunté** vs **prix du bien acheté** : AMT_CREDIT / AMT_GOODS_PRICE
- Ratio **annuités** vs **montant emprunté** : AMT_ANNUITY / AMT_CREDIT
- Ratio **annuités** vs **revenus annuels** : AMT_ANNUITY / AMT_INCOME_TOTAL
- Ratio **ancienneté au travail** vs **âge** : DAYS_EMPLOYED / DAYS_BIRTH



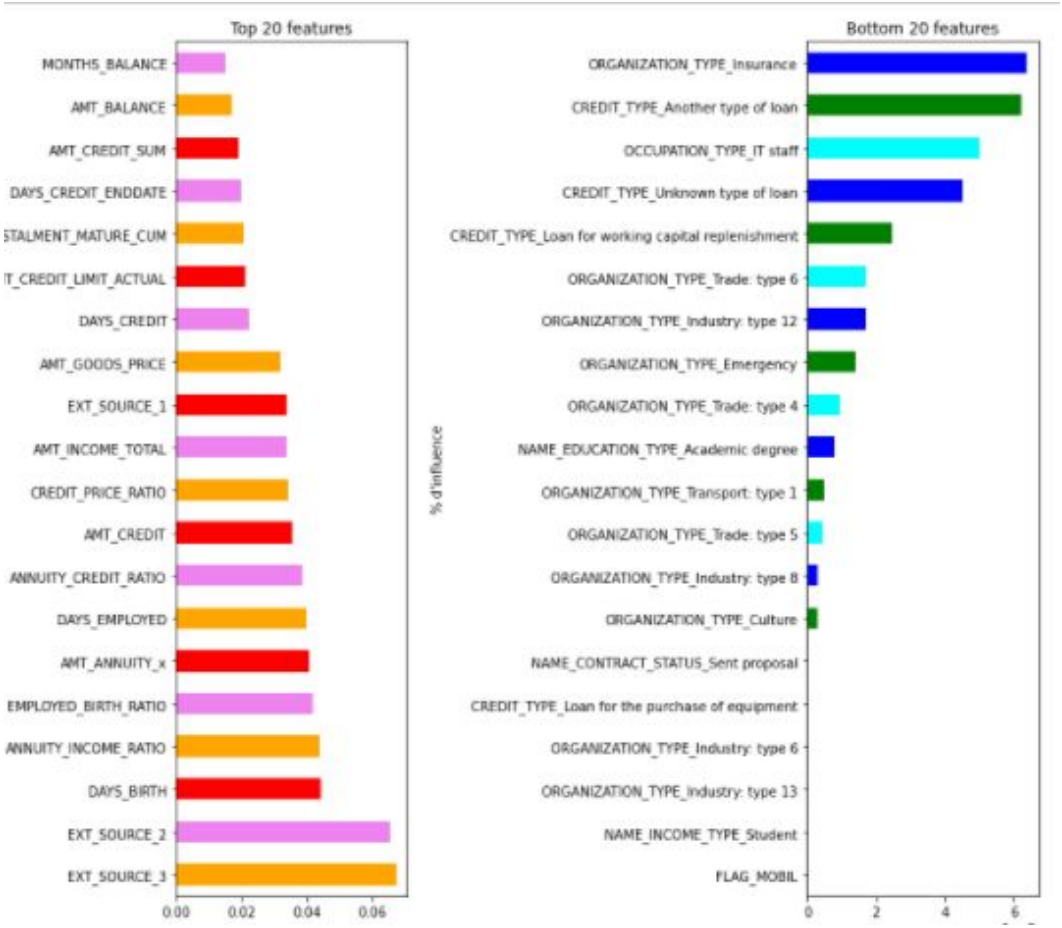
Feature selection

SelectFromModel

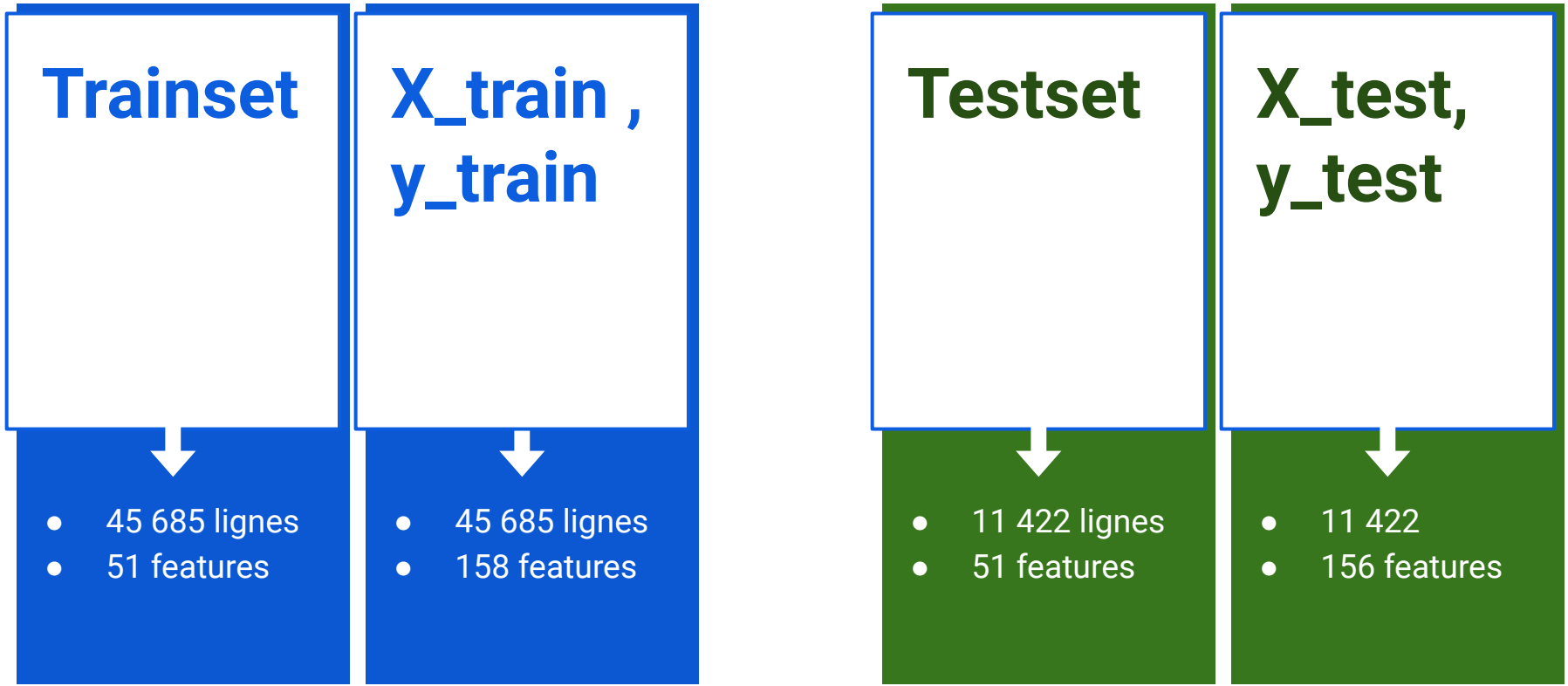
Entraîne un estimateur puis sélectionne les variables les plus importantes pour cet estimateur.

SelectKBest

Sélectionne les K variables X dont le score du test de dépendance (ici Chi2) avec y est le plus élevé



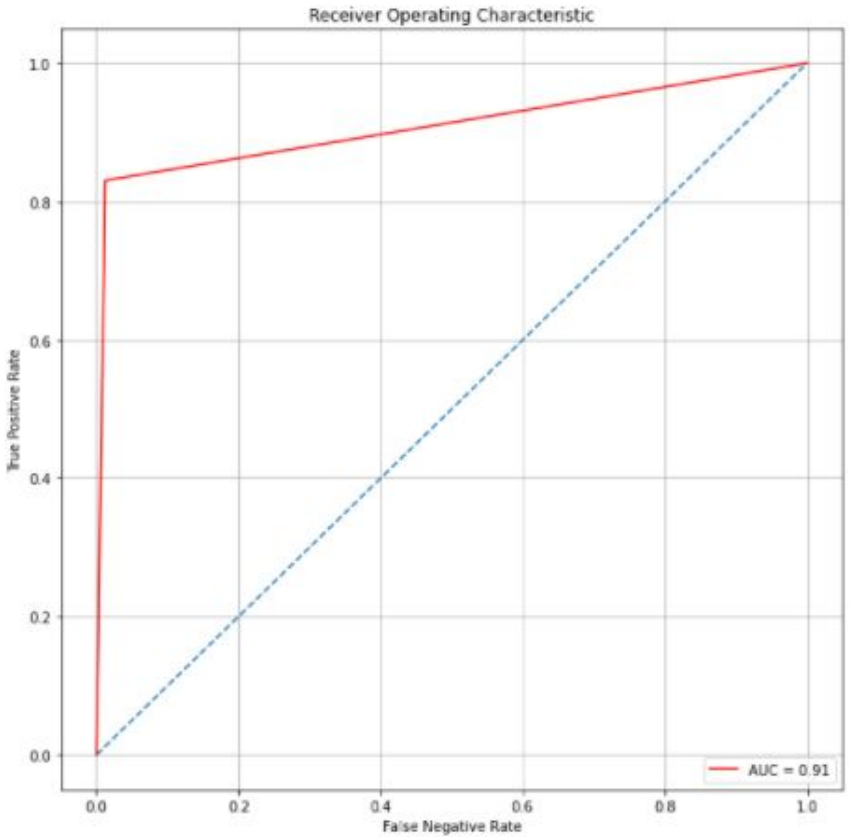
Séparation du jeu de données

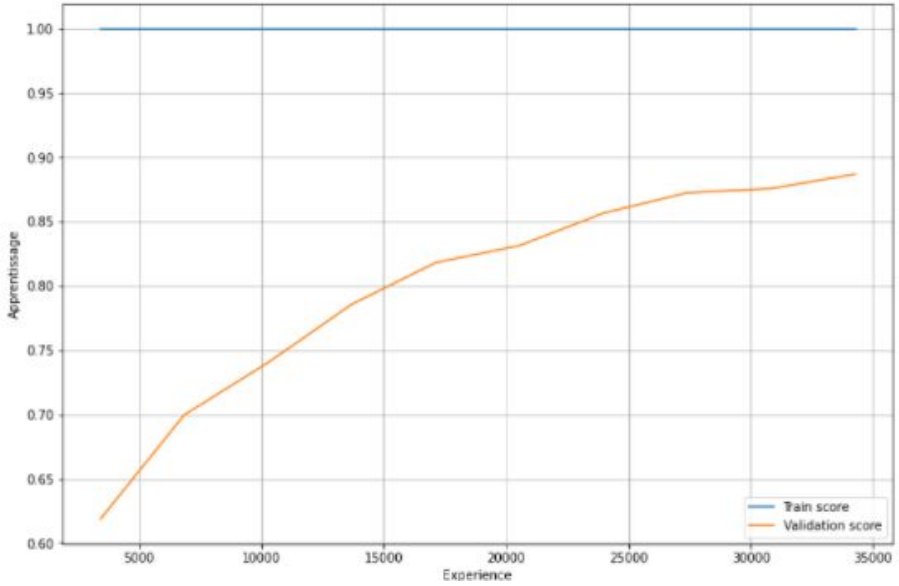


Première modélisation

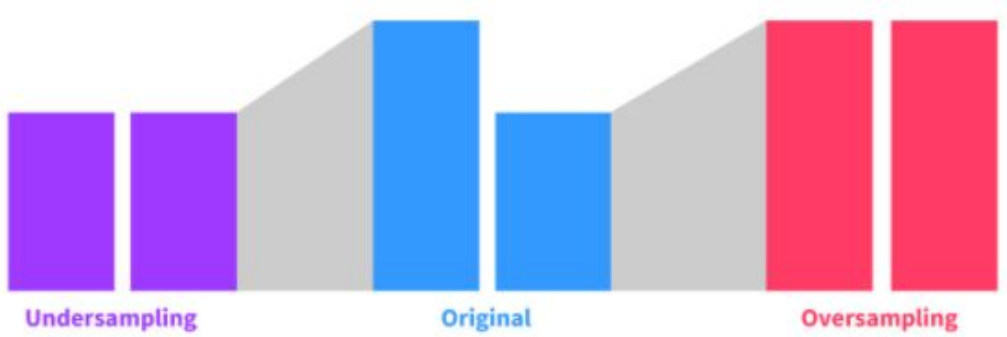
```
------------------------- 1. Matrice de confusion ------

[[10513   131]
 [  132   646]]
```

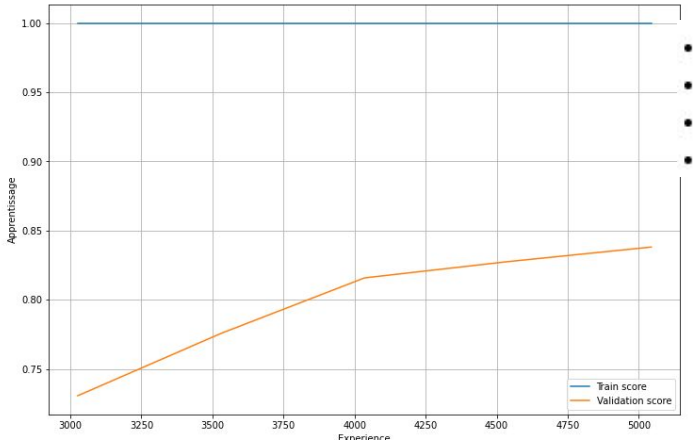
	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	10644
1.0	0.83	0.83	0.83	778
accuracy			0.98	11422
macro avg	0.91	0.91	0.91	11422
weighted avg	0.98	0.98	0.98	11422



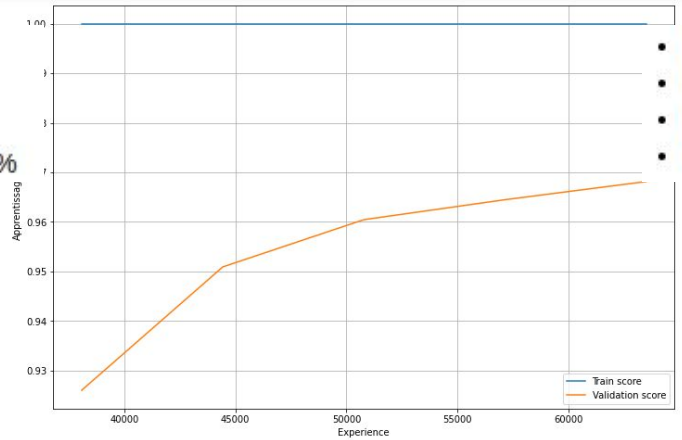


Équilibrage des classes





- Recall : 0.90
- Precision : 0.84
- ROC : 0,862
- Learning curve : 84%



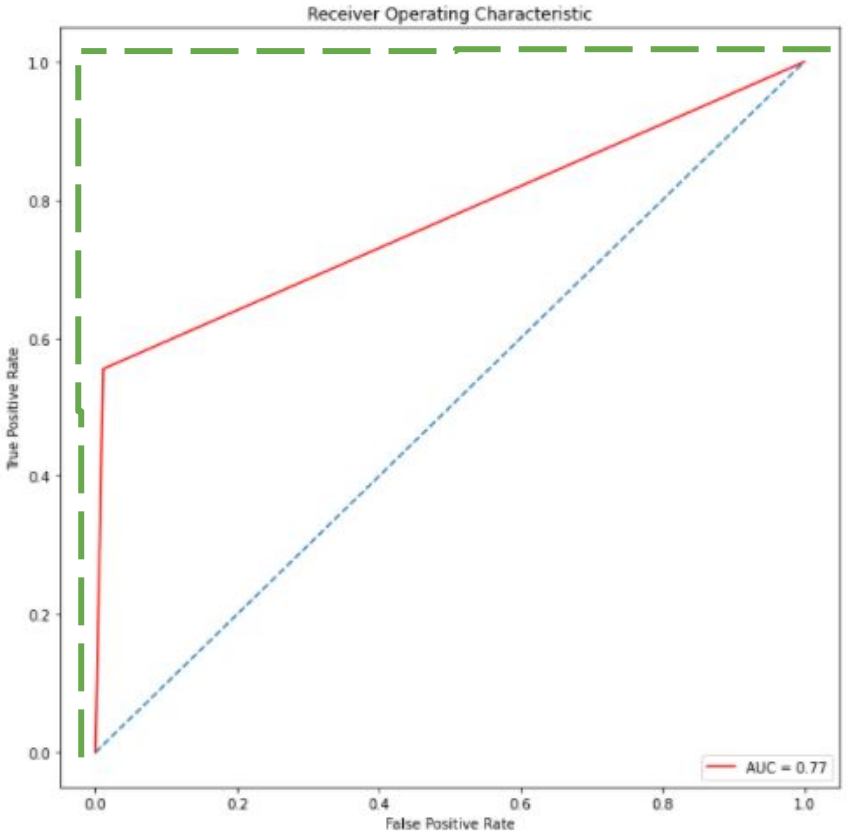
- Recall : 0.92
- Precision : 0.96
- ROC : 0,942
- Learning curve : 96,8%

source

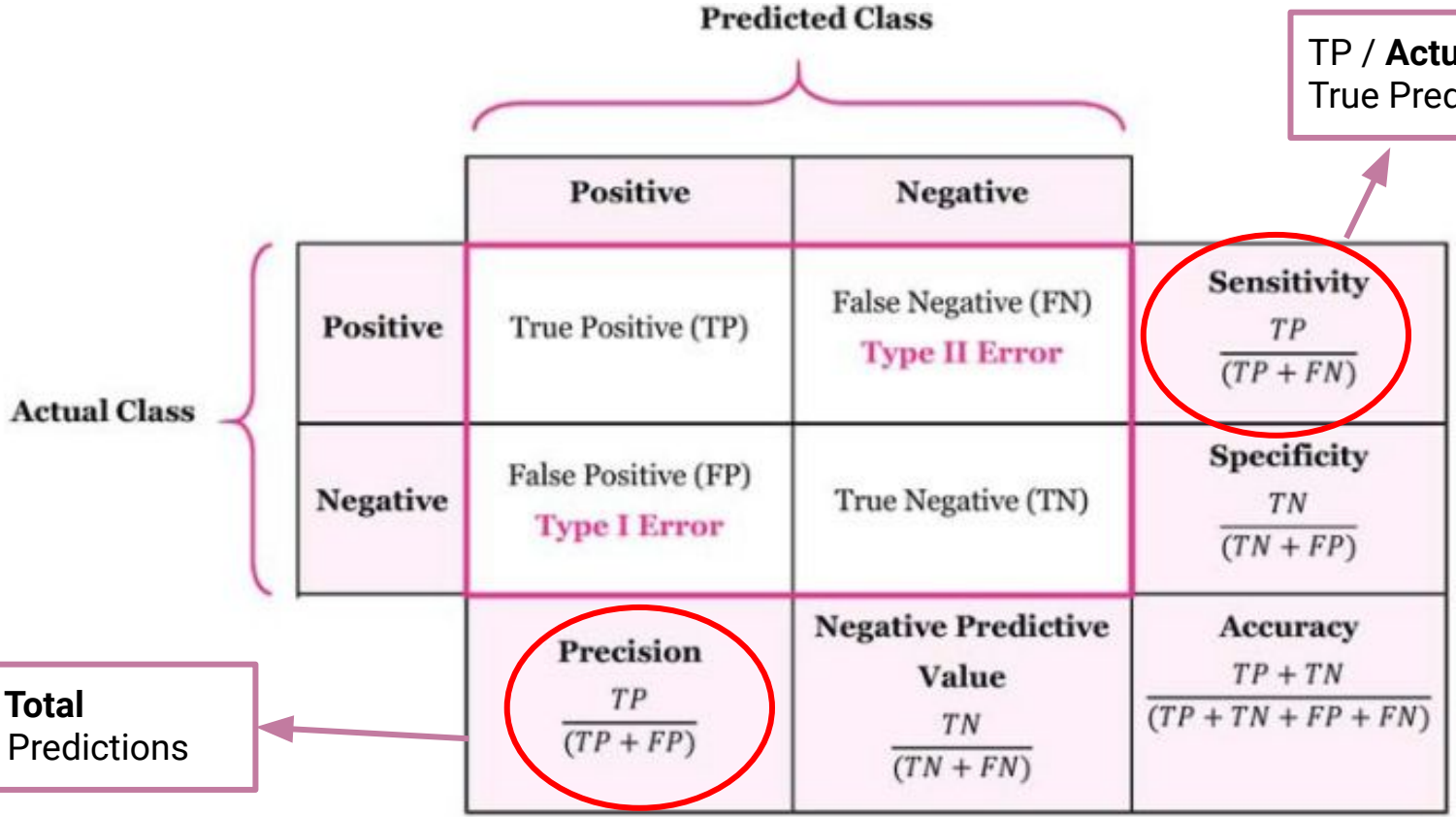
Synthèse des résultats pour les modèles utilisés

Métrique : ROC

- **modèle parfait AUC =1**
  - 100% d'observations correctements classées
- **modèle naïf AUC = 0.5**
  - TP Rate = FP Rate
  - même proportion d'observations correctement & incorrectement classées
- **modèle KNeighbors Clf AUC**
  - supérieur à 0.5
  - proportion d'observations correctement classées en TP est supérieure aux observations incorrectement classées (FP)

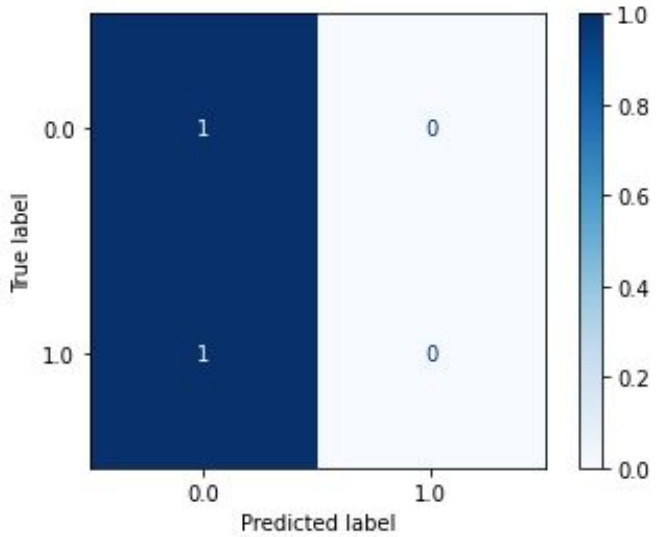


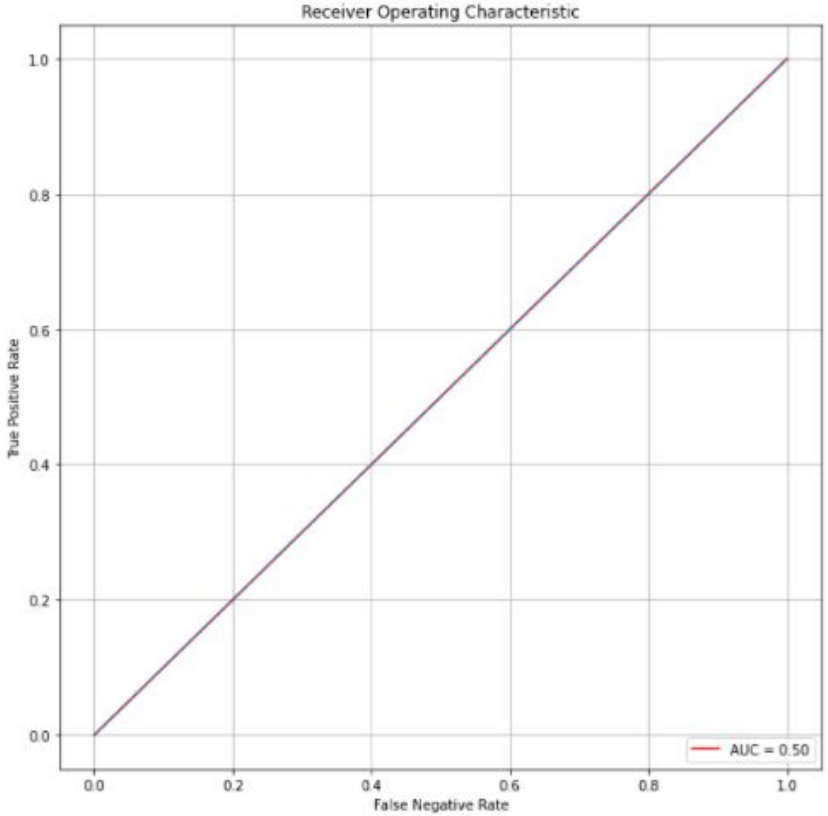
KPI : Precision - Recall (Sensitivity)



source

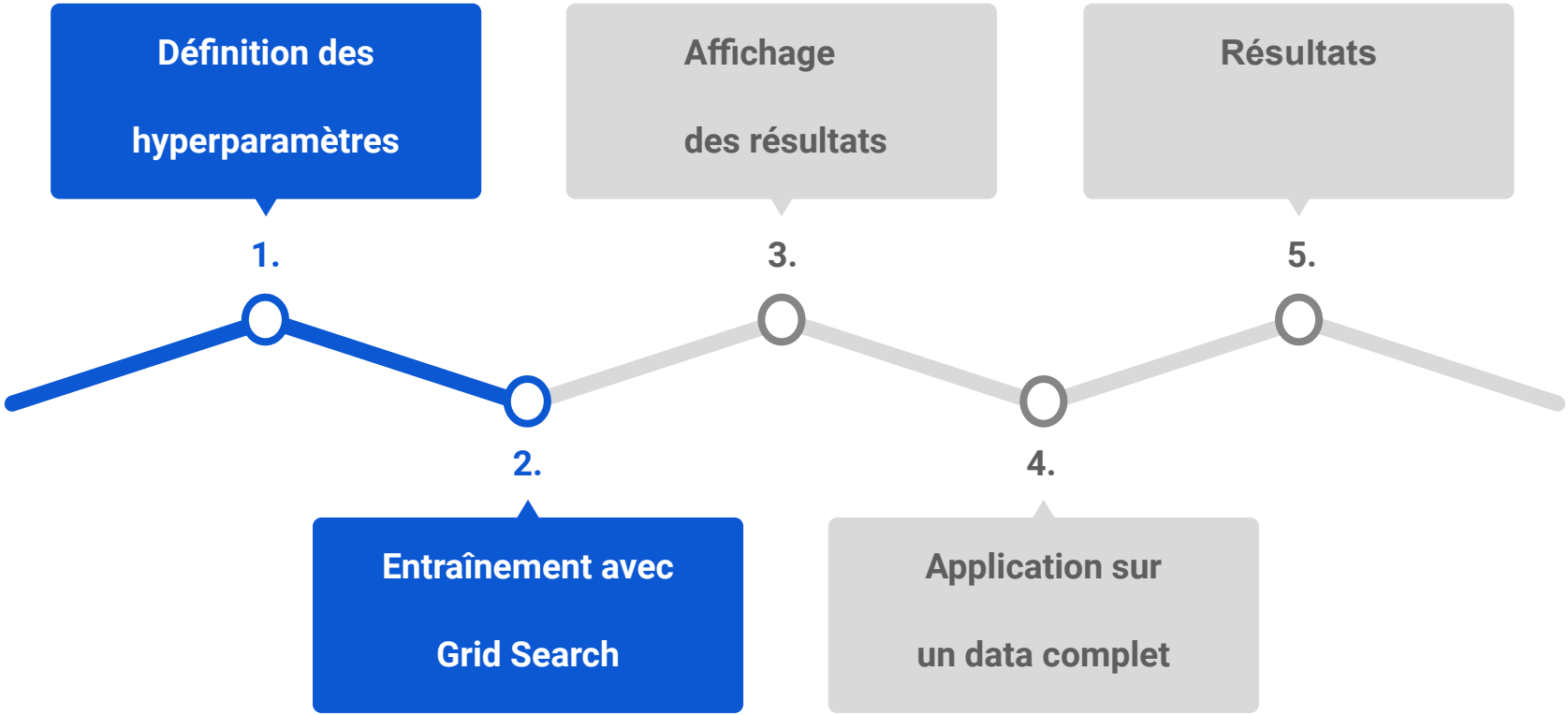
DummyClassifier



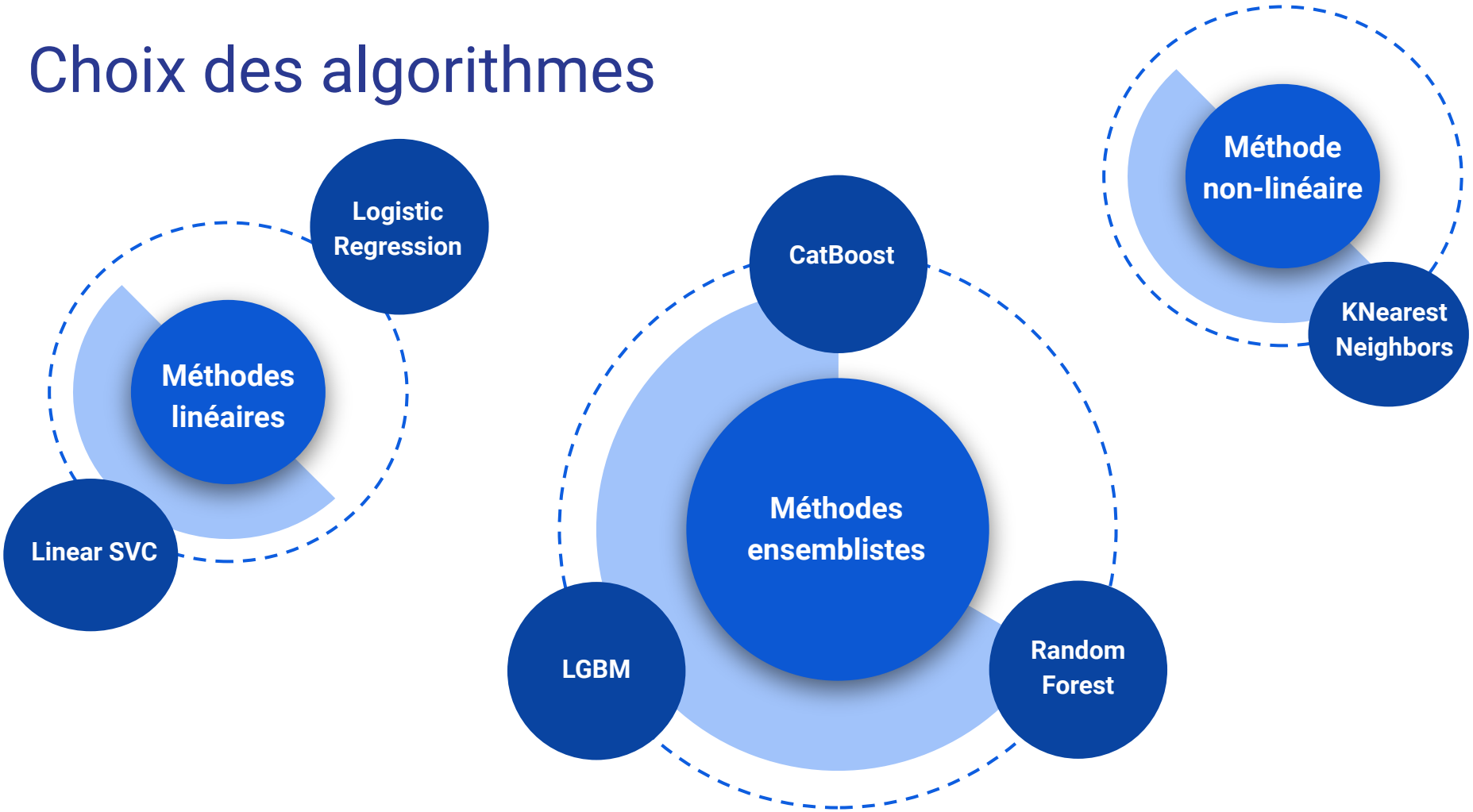


- TN = 1
- FP = 1

Méthodologie



Choix des algorithmes



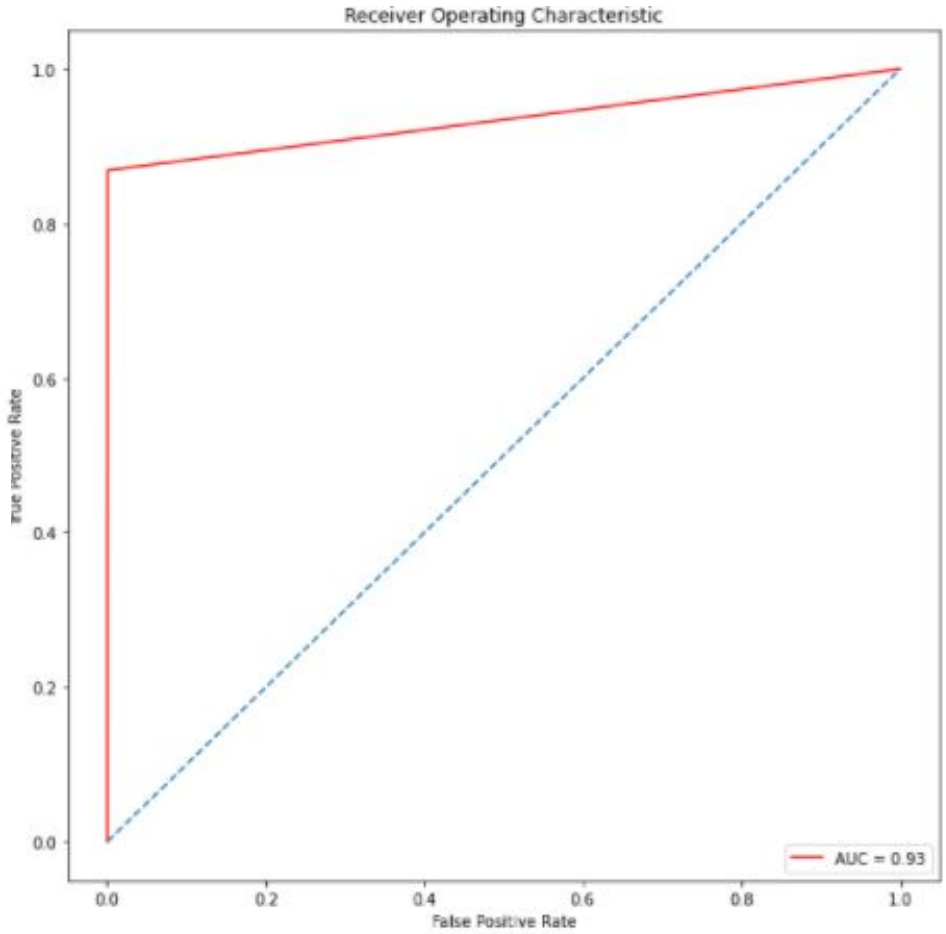
Résultats

	Recall TP	Precision FP	Accuracy	AUROC
Random Forest	0.77	0.86	0.82	0.821
Logistic Regression	0.68	0.70	0.70	0.695
SVM	0.69	0.70	0.70	0.698
KNN	0.91	0.78	0.82	0.823
CatBoost	0.96	0.96	0.96	0.960
LGBM	0.83	0.91	0.88	0.876
RandomForest Tuned	0.08	1.00	0.93	0.538
Logistic Regression Tuned	0.00	0.00	0.93	0.500
SVM Tuned	0.00	0.00	0.93	0.500
KNN Tuned	0.66	0.81	0.96	0.826
CatBoost Tuned	0.87	1.00	0.99	0.934
LGBM Tuned	0.33	0.97	0.95	0.663

CatBoost Classifier

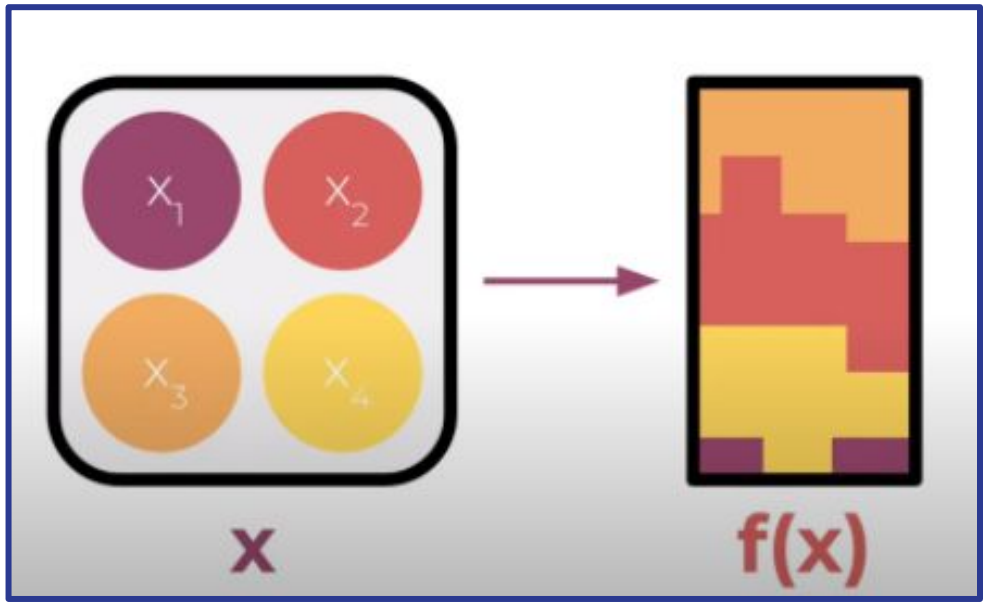
	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	28271
1.0	1.00	0.87	0.93	2186
accuracy			0.99	30457
macro avg	0.99	0.93	0.96	30457
weighted avg	0.99	0.99	0.99	30457

```
[[28263     8]
 [  287  1899]]
```



Interprétabilité du modèle

SHAP

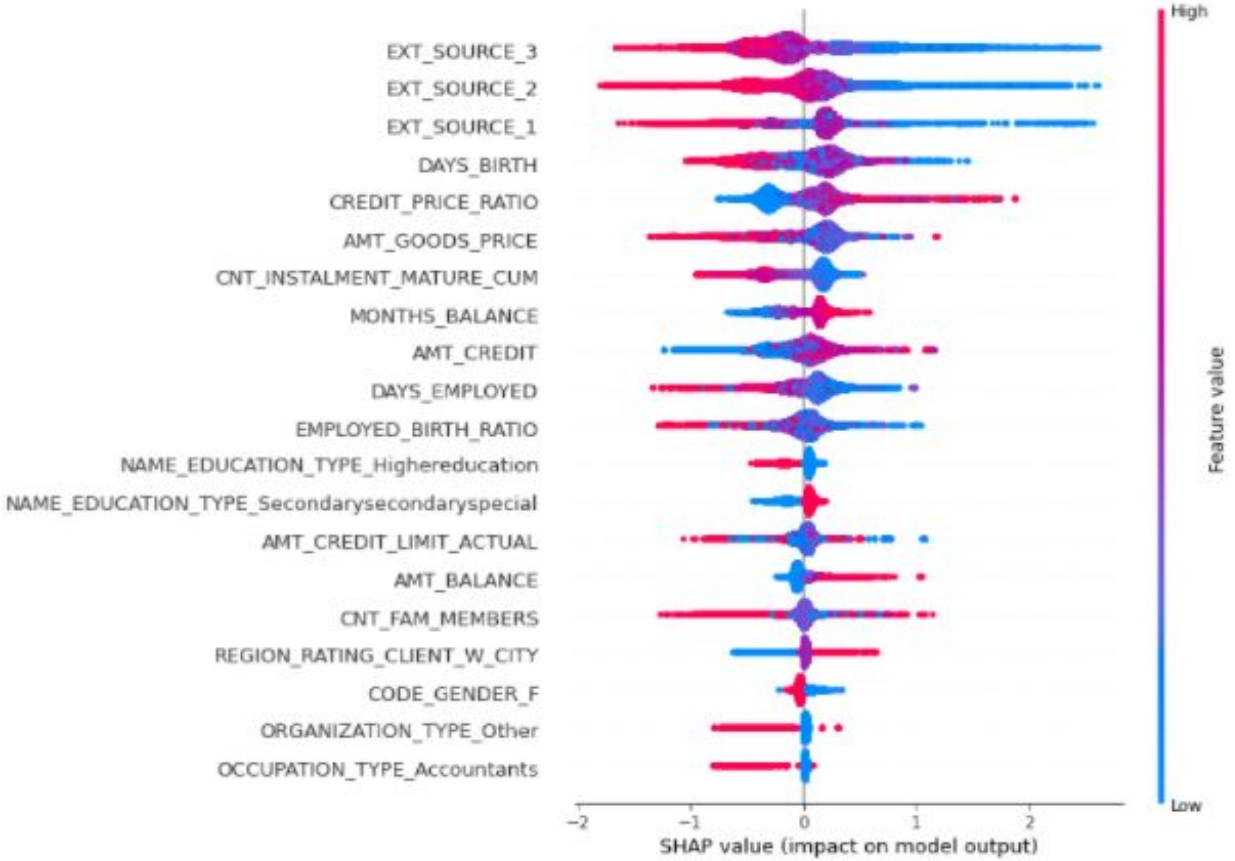


Ajouter les valeurs des contributions des variables sur toutes les prédictions, puis les **classer** par ordre d'importance.

Expliquer chaque prédiction individuellement.

Estimer la contribution de chaque variable dans la prédiction.

Importance des variables



Identification d’un TN

- Client solvable
- Probabilité de TARGET==1 : 0.01
- Prédiction brute : 1,82
- Variables en bleu

Identification d’un TP

- Client à risque
- Probabilité de TARGET==1 : 0.86
- Prédiction brute : -4.55
- Variables en rouge

Conclusion

Informations-clés
pour les conseillers de la banque

01	EXT_SOURCE_XXX	• Notes externes du demandeur d'un crédit
		  ○ favorable
		  ○ défavorable

02	**DAYS_BIRTH**	• L'âge du demandeur d'un crédit
		  ○ favorable
		  ○ défavorable

03	**CREDIT_PRICE_RATIO**	• Rapport entre le montant emprunté et le prix du bien acheté
		  ○ favorable
		  ○ défavorable