Construisez un modèle de scoring

Openclassrooms - Projet 4

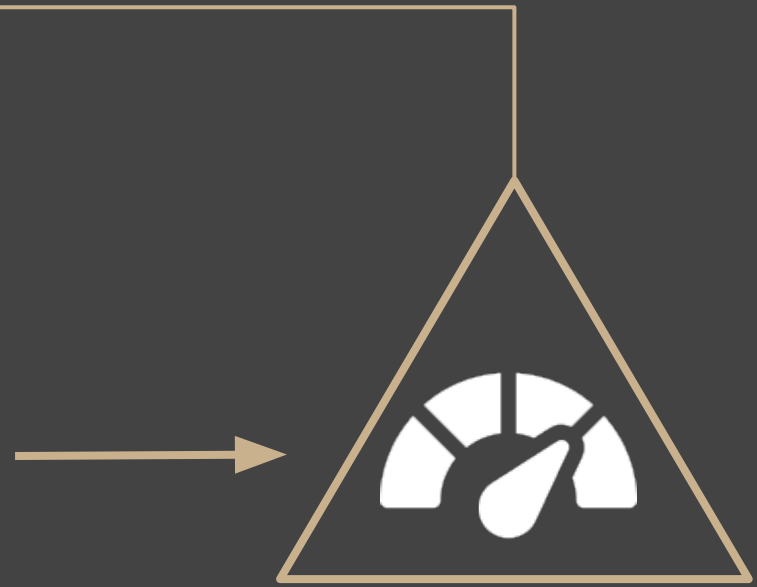
Problématique et Contexte

Contexte

Problématique métier

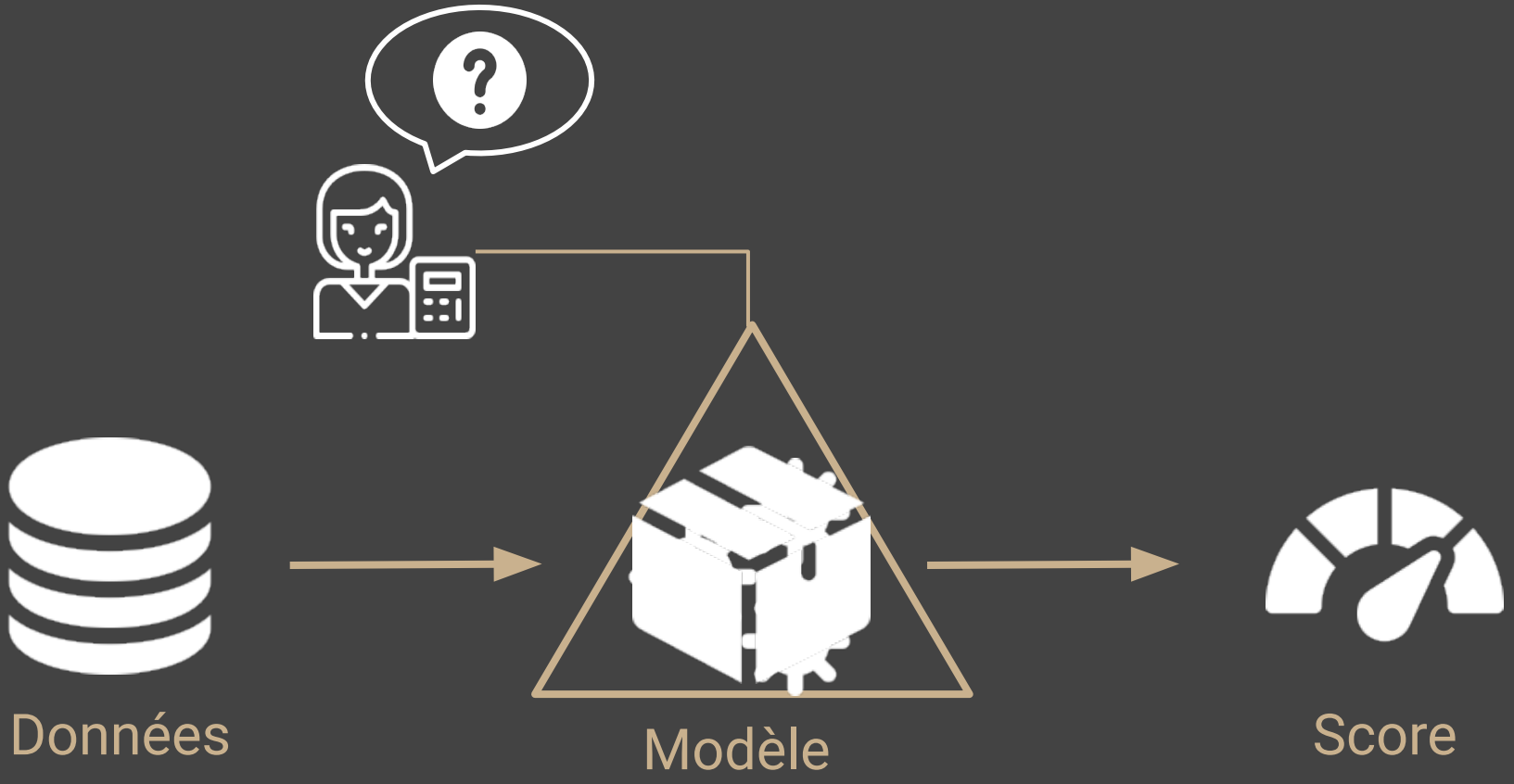
Données

Modèle



Score

Problématique métier



Enjeu métier

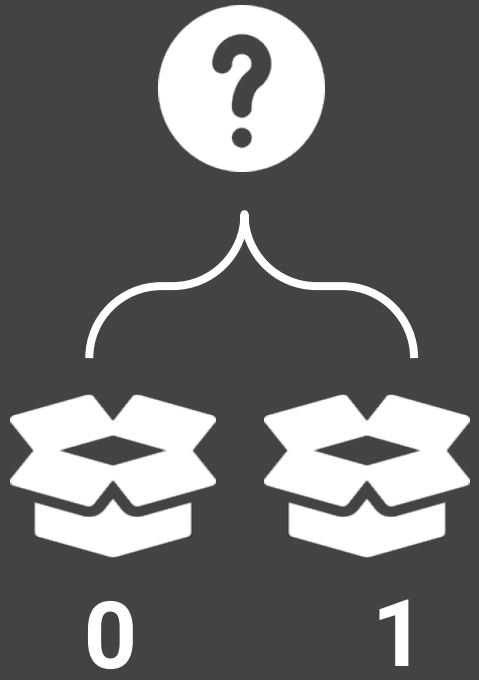




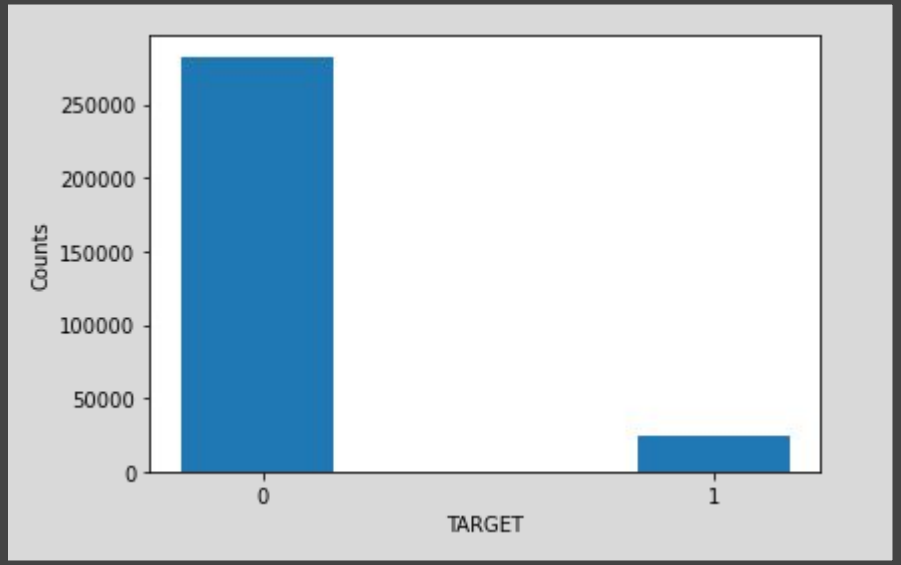



Formalisation

Classification supervisée



Classes déséquilibrées



92%

8%

Méthodologie

Méthodologie

Data

- nettoyage
- encodage
- pre-processing

Importance des variables

Variables

- comparaison
- selection
- création

Métrique de performance

Modèles

- expérimentation
- selection
- optimisation

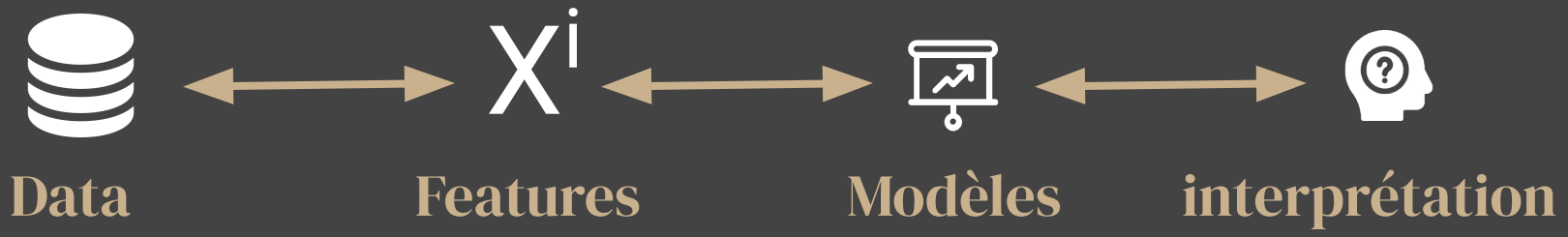
interprétation

- permutation
- partial dependancy
- SHAP

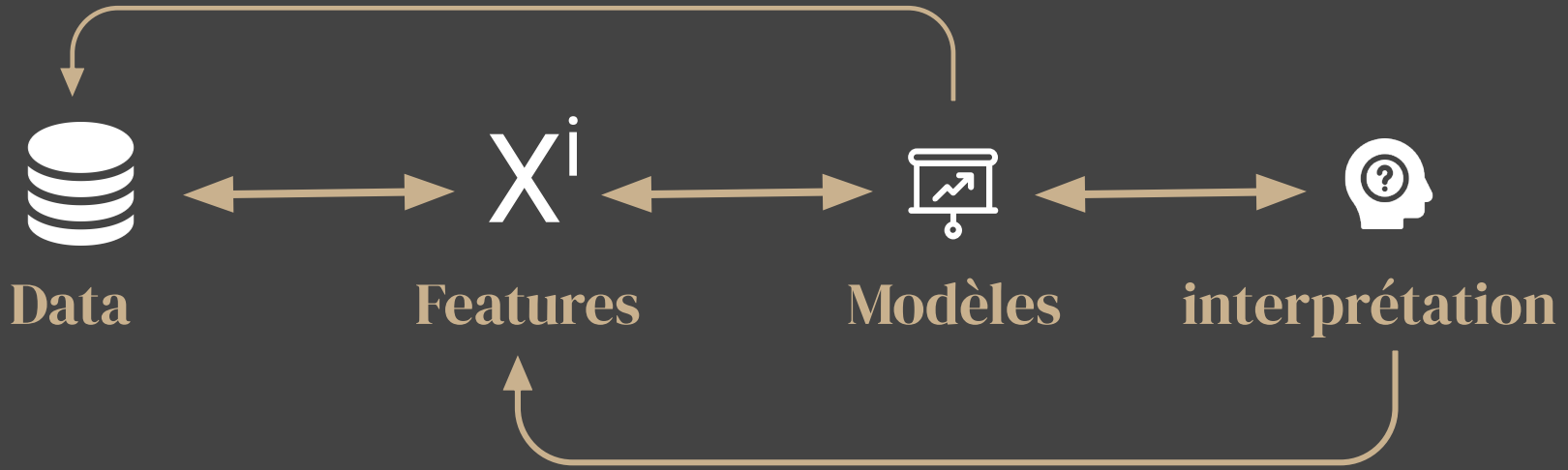
Comparaison de modèles

Méthodologie

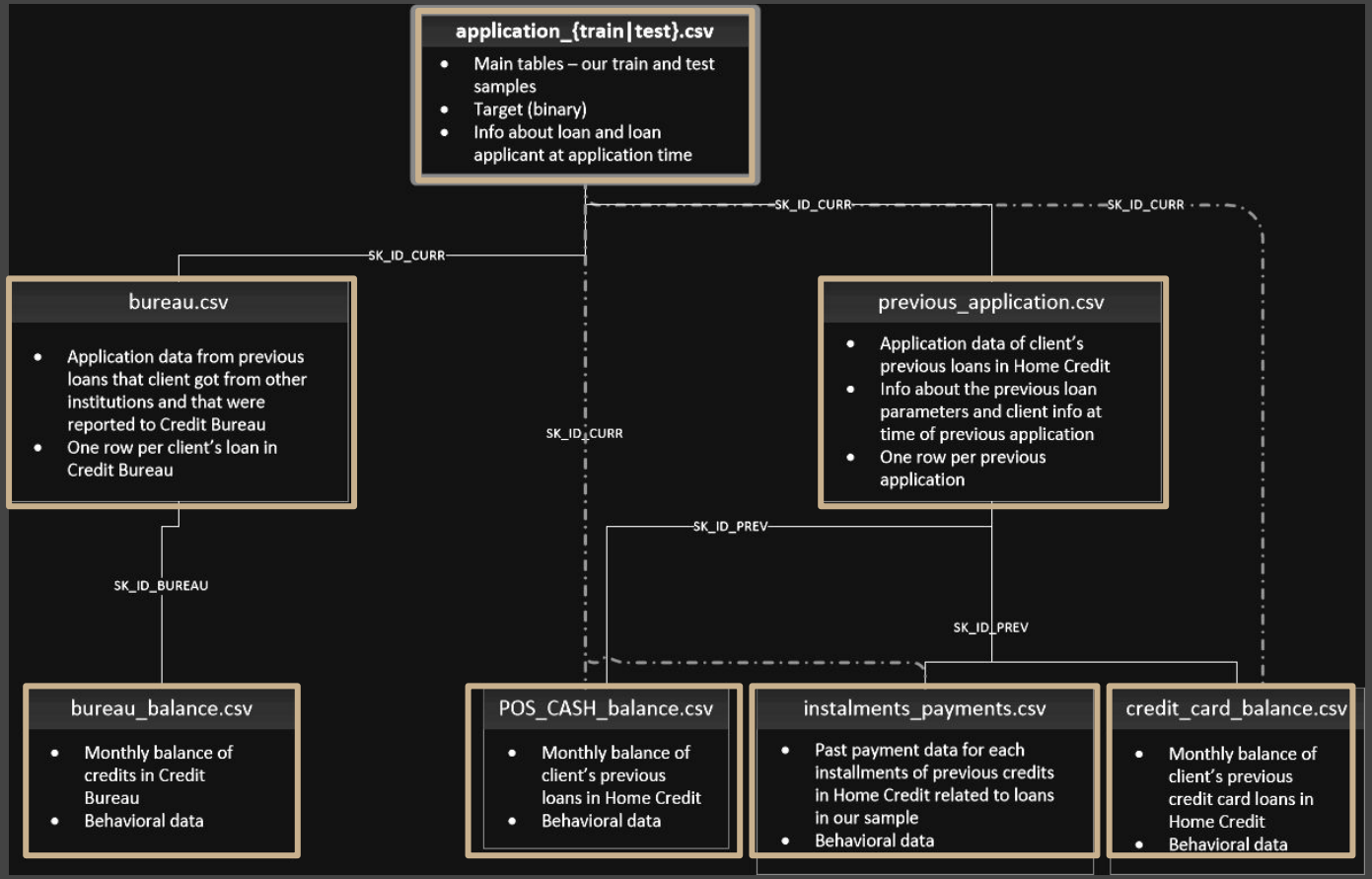
Méthodologie



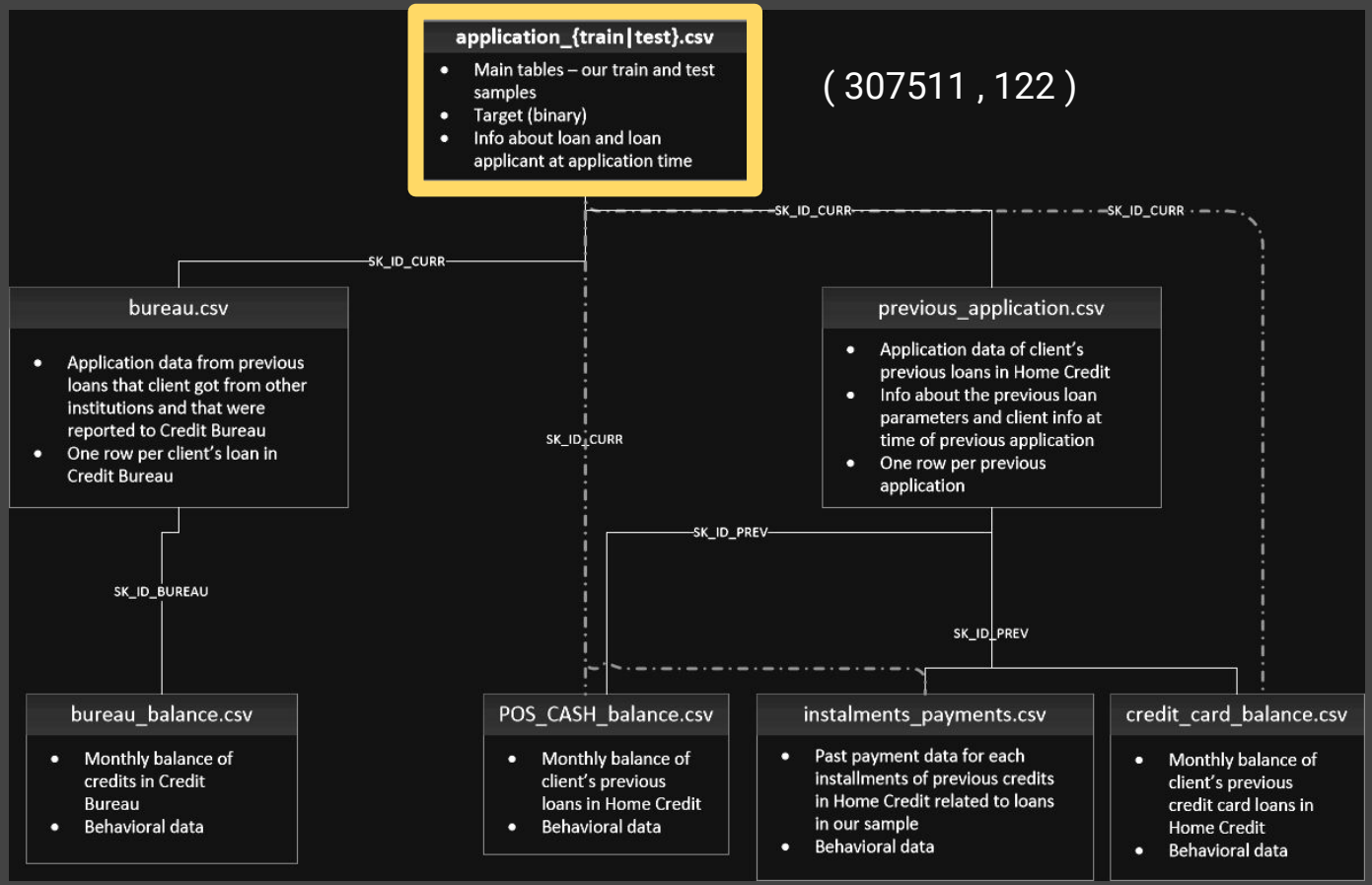
Méthodologie



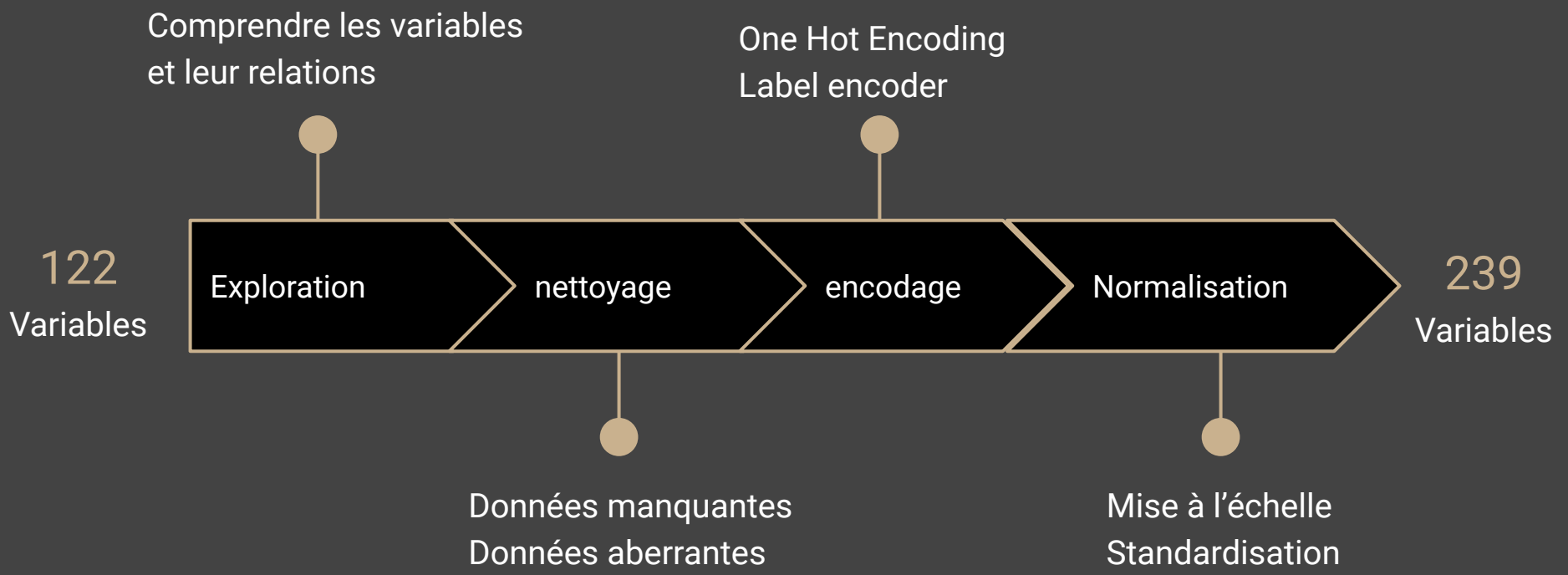
Base de données



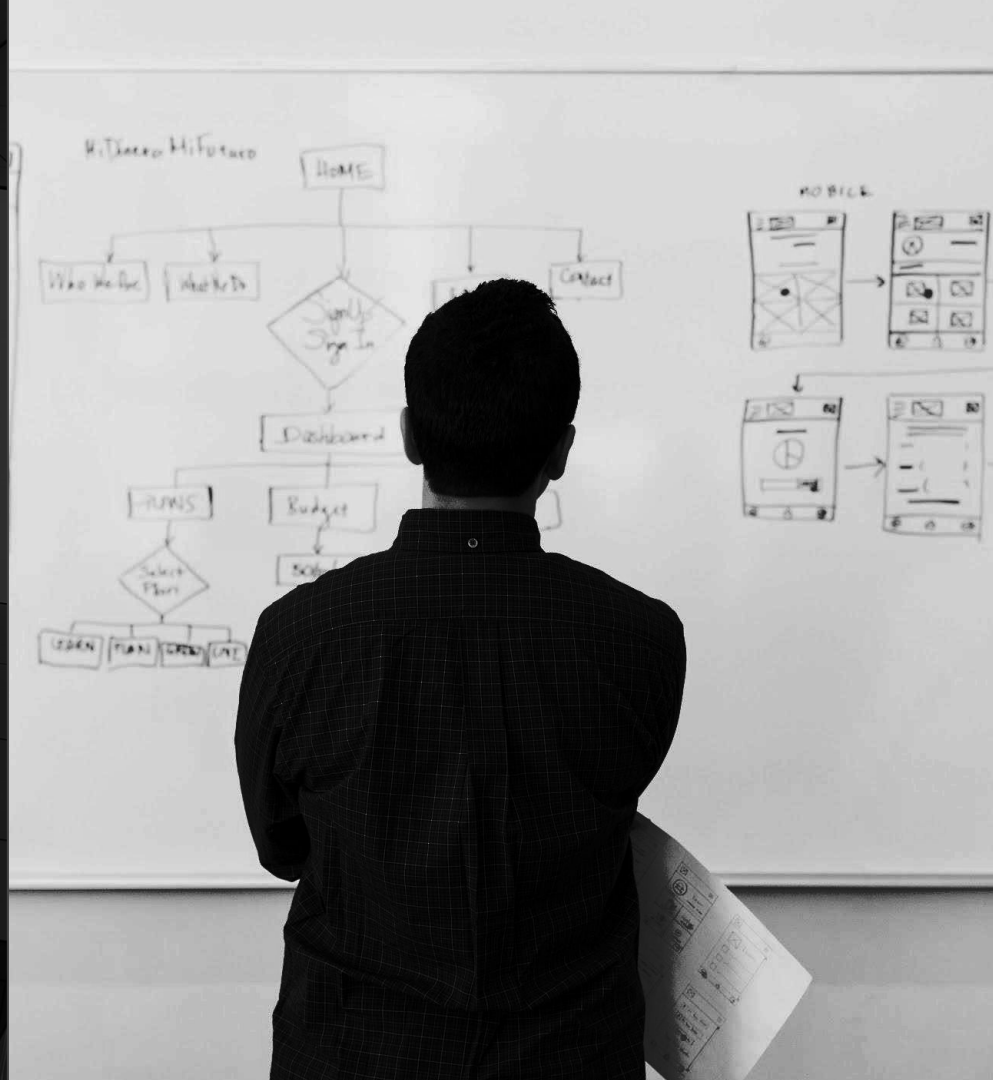
Base de données



Préparation des données



Les modèles



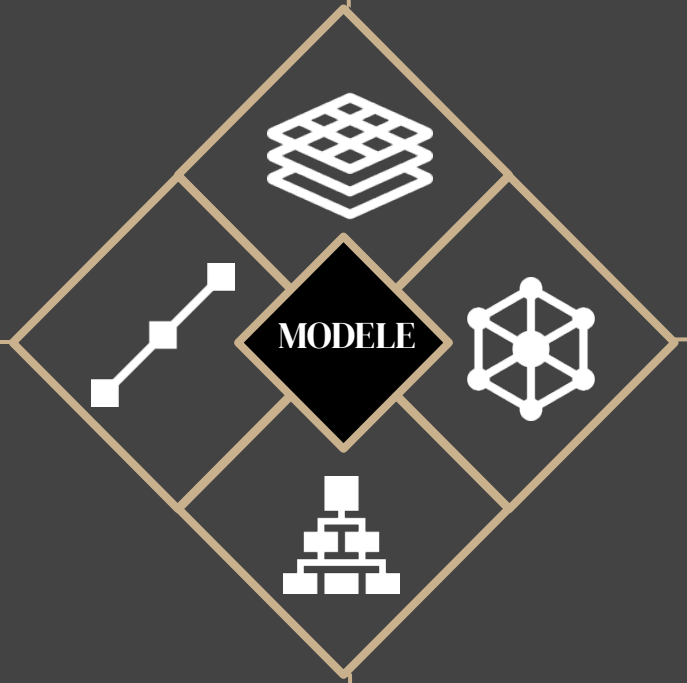
Les modèles

Référence

DummyClassifier

Ensembliste

Random Forest Classifier
Light Gradient Boosting



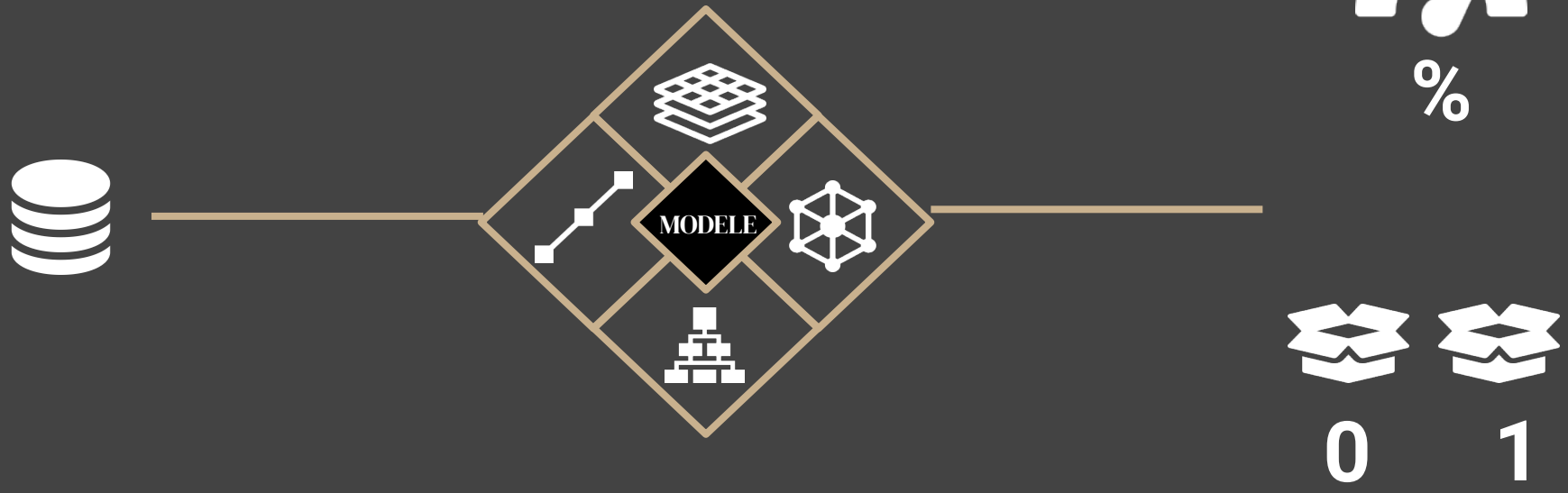
Linéaire

LDA Classifier
SGD Classifier
Logistic Regression

Arbres de décision

Tree Classifier

Les modèles

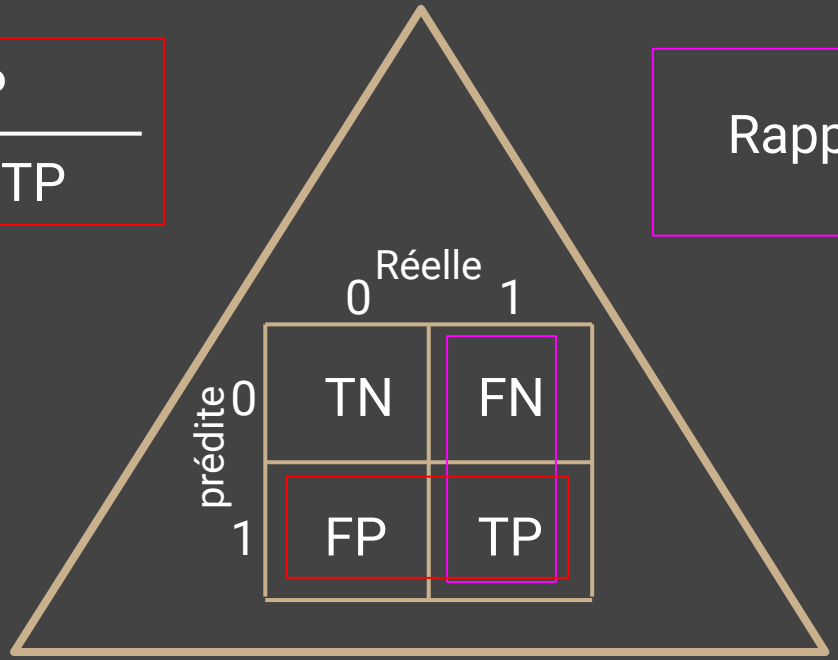


Choix de la métrique

F1-score

$$\text{Precision} = \frac{TP}{FP + TP}$$

$$\text{Rappel} = \frac{TP}{FN + TP}$$



Métier

Identifier les mauvais payeurs

Ne pas refuser de bons payeurs

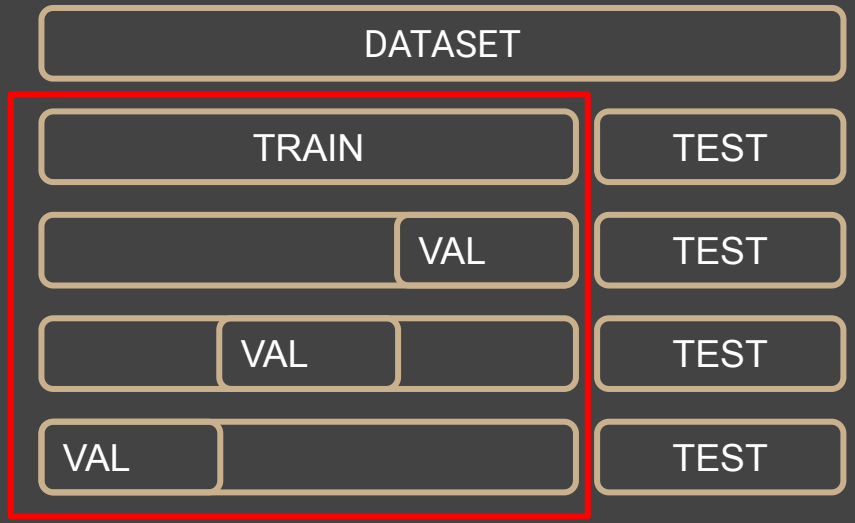
Technique

Classes déséquilibrées

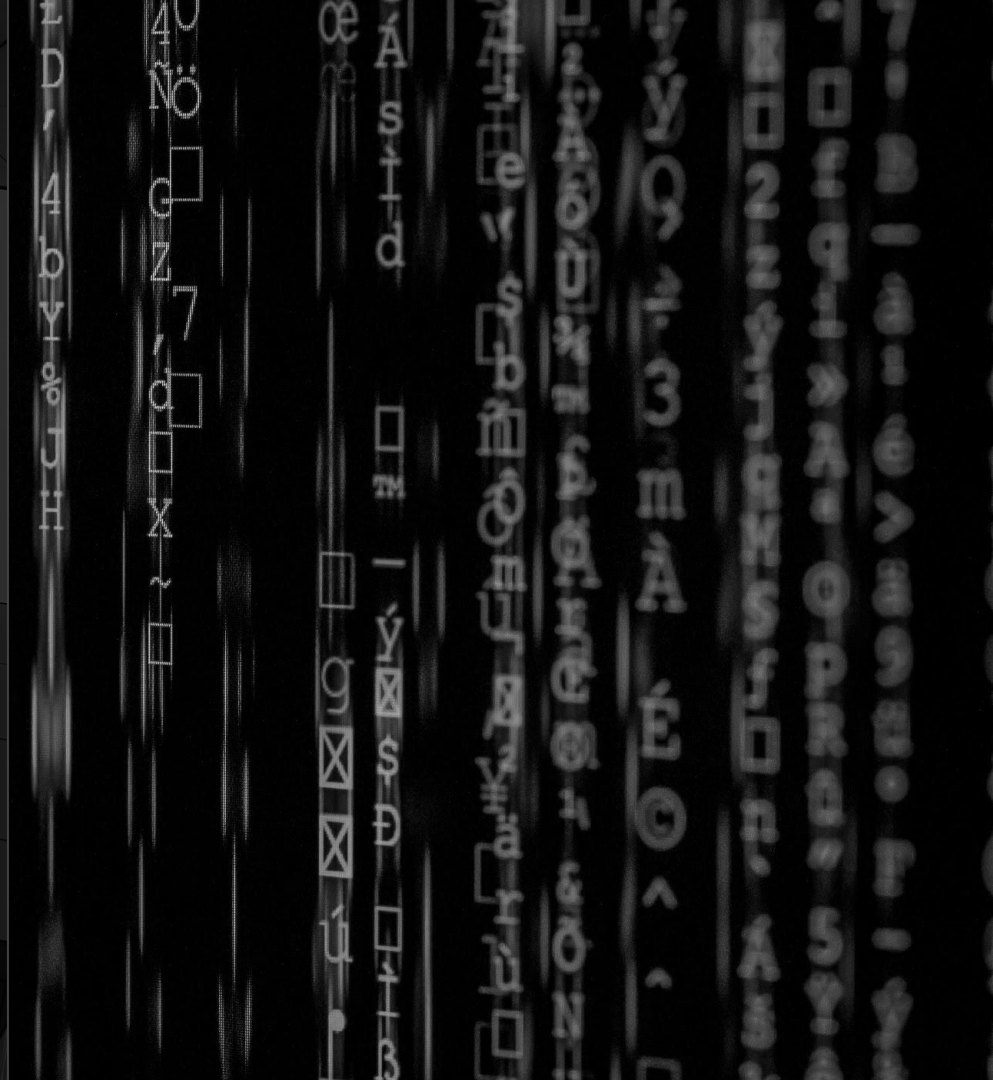
Feature engineering

Comparaison de modèle

Validation croisée

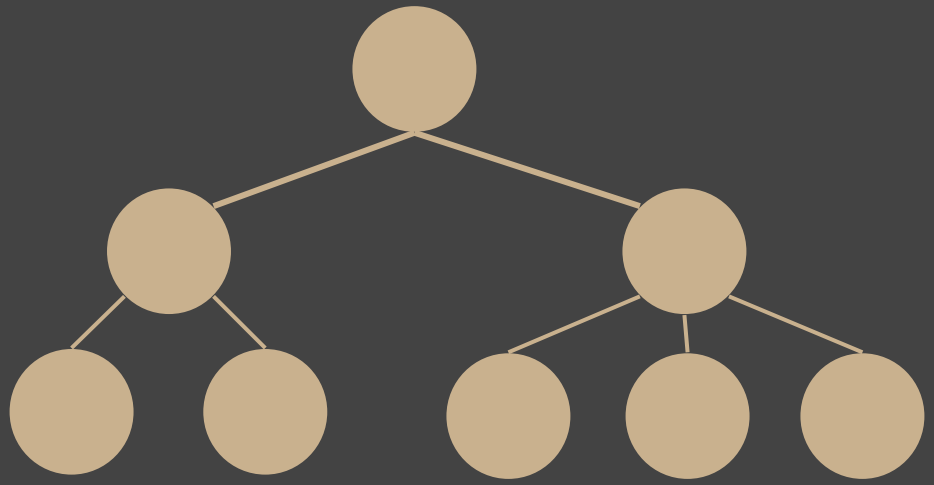


Les variables



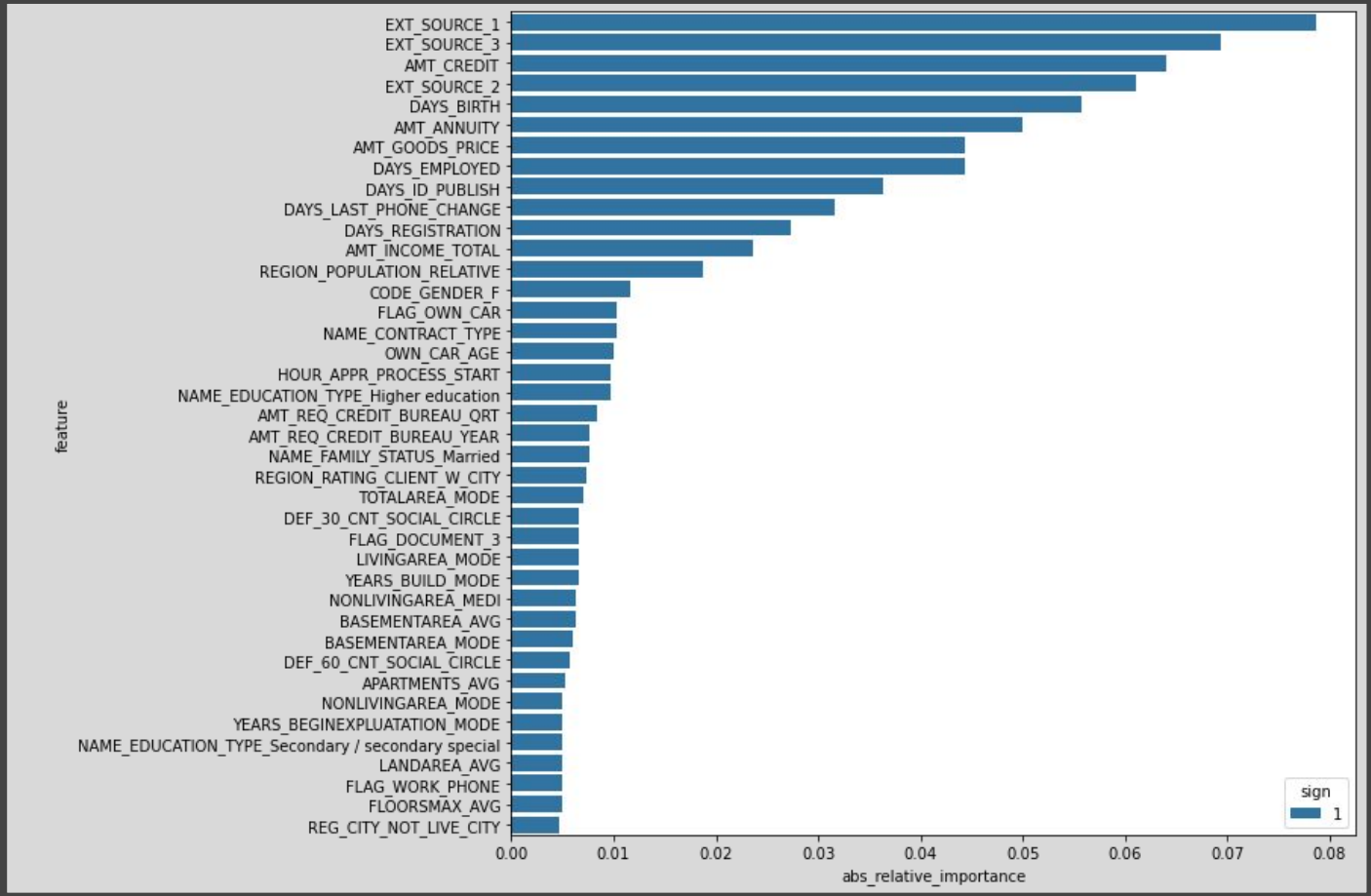
Importance des variables

Arbre de decision



Qualité de la séparation

Importance des variables



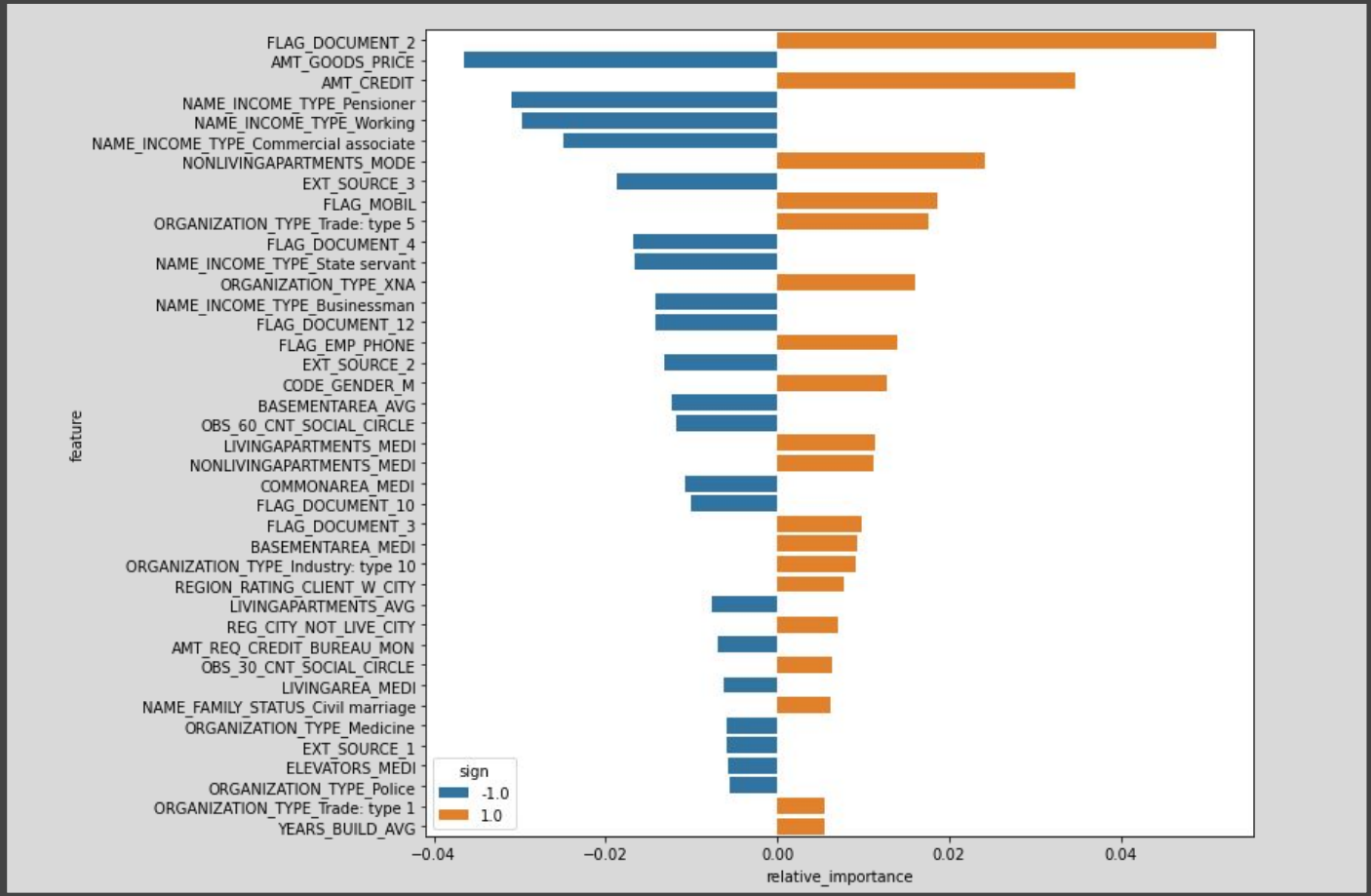
Importance des variables

Regression

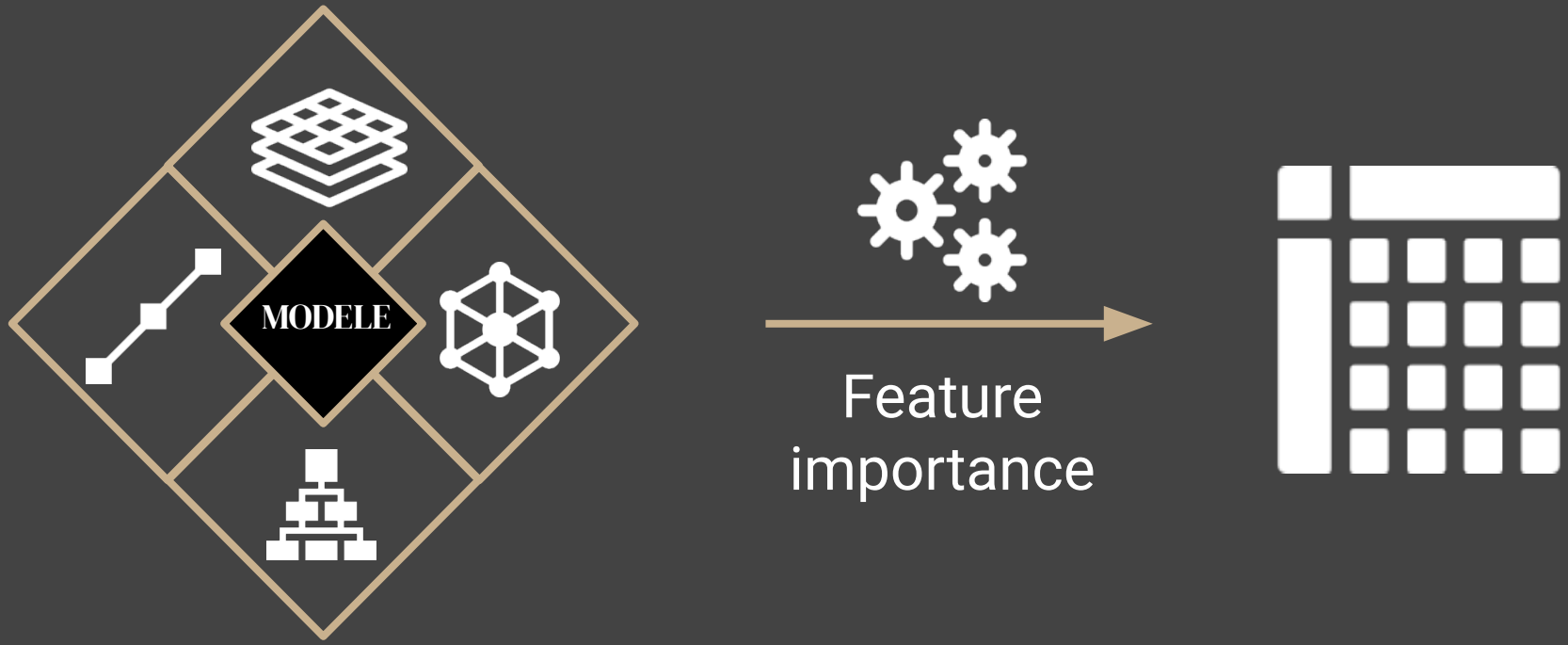
$$Y = a\,X1 + b\,X2 + c\,X3$$

Coefficient de la variable dans l’équation

Importance des variables



Feature engineering

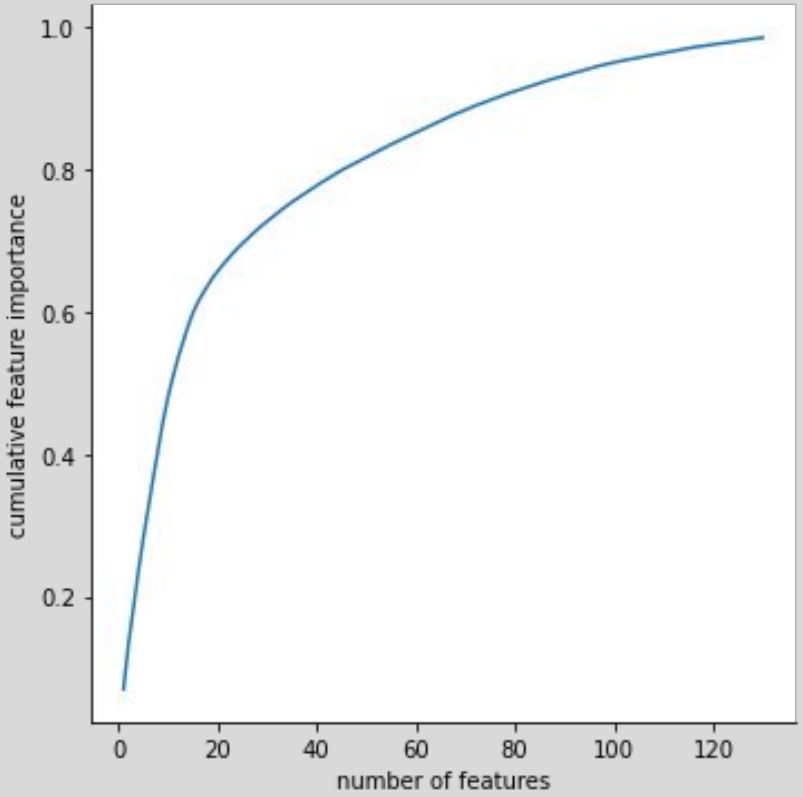


Feature engineering

	counts	rank	total importance
EXT_SOURCE_2	5	51	0.051833
EXT_SOURCE_3	5	59	0.044400
AMT_CREDIT	6	28	0.032359
DAYS_REGISTRATION	4	170	0.027430
DAYS_EMPLOYED	4	108	0.026430
AMT_GOODS_PRICE	6	49	0.024952
DAYS_ID_PUBLISH	3	311	0.020717
AMT_ANNUITY	3	297	0.020553
EXT_SOURCE_1	4	164	0.019704
YEARS_BUILD_MEDI	3	157	0.015724
YEARS_BEGINEXPLUATATION_AVG	4	280	0.012105
HOUR_APPR_PROCESS_START	3	343	0.011235
YEARS_BUILD_AVG	4	261	0.010220
OBS_60_CNT_SOCIAL_CIRCLE	4	179	0.009717
OBS_30_CNT_SOCIAL_CIRCLE	4	325	0.009698

Top 15 des variables sur 6 modèles

Feature engineering



Optimisation du nombre de variables

	best nb of features
tree	80
lda	60
lr	40
rdf	30
sgd	120
lgb	30

Feature engineering

	f1_valid	f1_test
lgb	0.286486	0.265187
lgb_red	0.280891	0.265639
lr	0.253691	0.248996
lr_red	0.251286	0.249654
tree	0.199210	0.149072
tree_red	0.192929	0.138490

	best nb of features
tree	80
lda	60
lr	40
rdf	30
sgd	120
lgb	30

Feature engineering

Création de variables métiers

Feature engineering

annuité / revenu : taux d'endettement

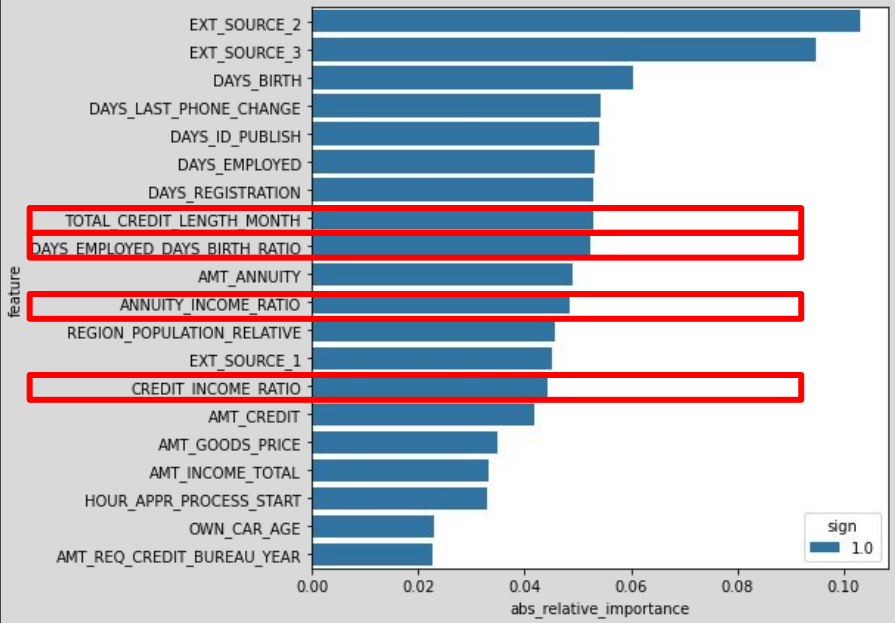
montant crédit / revenu : taille investissement

jours travaillés / jour depuis naissance : employabilité

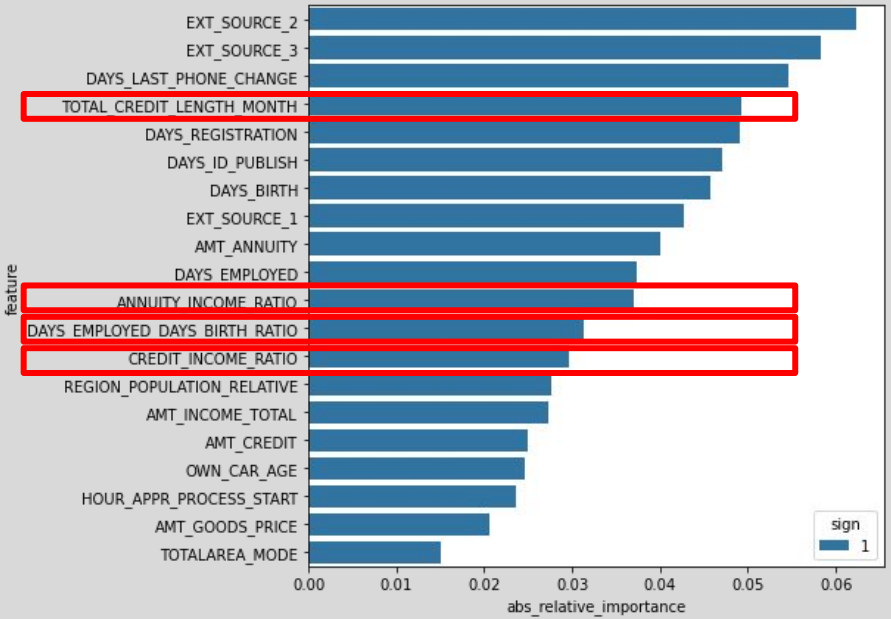
montant du crédit/ annuité : longueur du crédit

Feature engineering

Résultats



Random Forest



LGB

Feature engineering

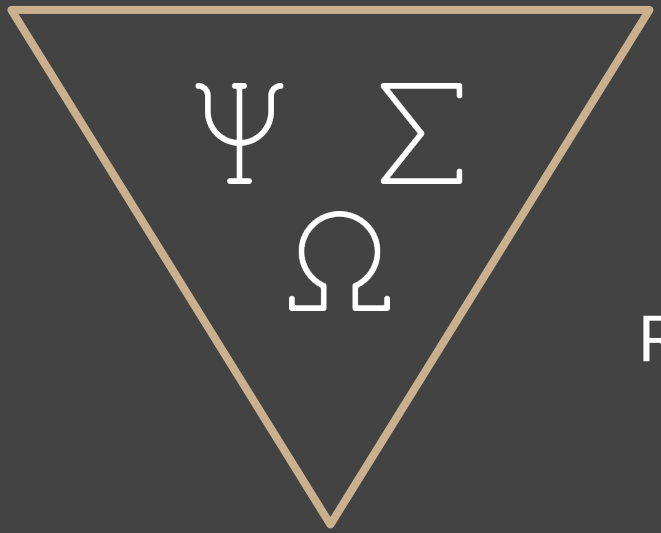
Résultats

	f1_valid	f1_test
lgb_red_busi	0.283219	0.265965
lgb_red	0.280891	0.265639
lgb	0.286486	0.265187
lr_red	0.251286	0.249654
lr	0.253691	0.248996
lr_red_busi	0.252074	0.248505
sgd	0.192715	0.191240

Optimisation du Modèle

Optimisation

Hyperparamètres



Gridsearch

Random Search

Meilleurs Hyperparamètres

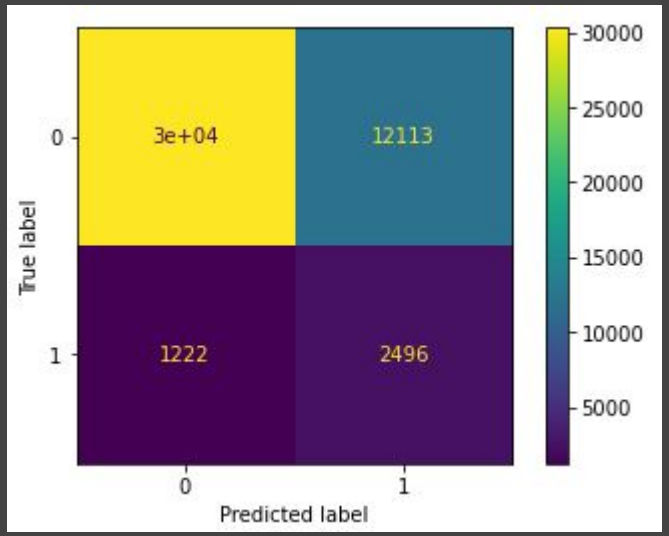
Optimisation

Résultats

	f1_valid	f1_test
lgb_red_busi_opt_grid	0.300125	0.276366
lgb_red_busi_opt_rdm	0.300569	0.273606
lgb_red_busi	0.283219	0.265965
lgb_red	0.280891	0.265639
lgb	0.286486	0.265187

Optimisation

Résultats



Classe Réelle \ Classe Prédite	0	1
0	TN	FP
1	FN	TP

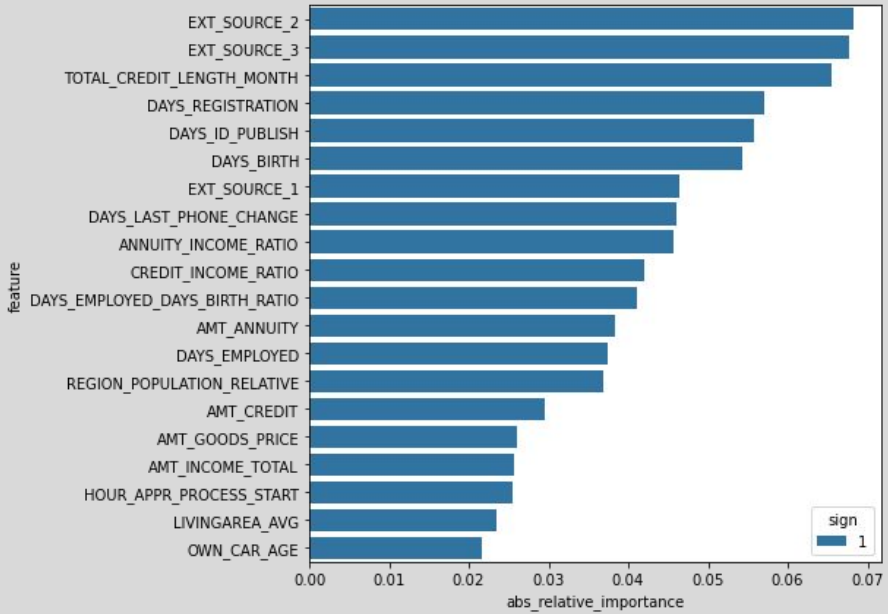
$$\text{Precision} = \frac{TP}{FP + TP} = 0.17$$

$$\text{Rappel} = \frac{TP}{FN + TP} = 0.67$$

Interprétation

Interprétation

Niveau Global - Importance des variables



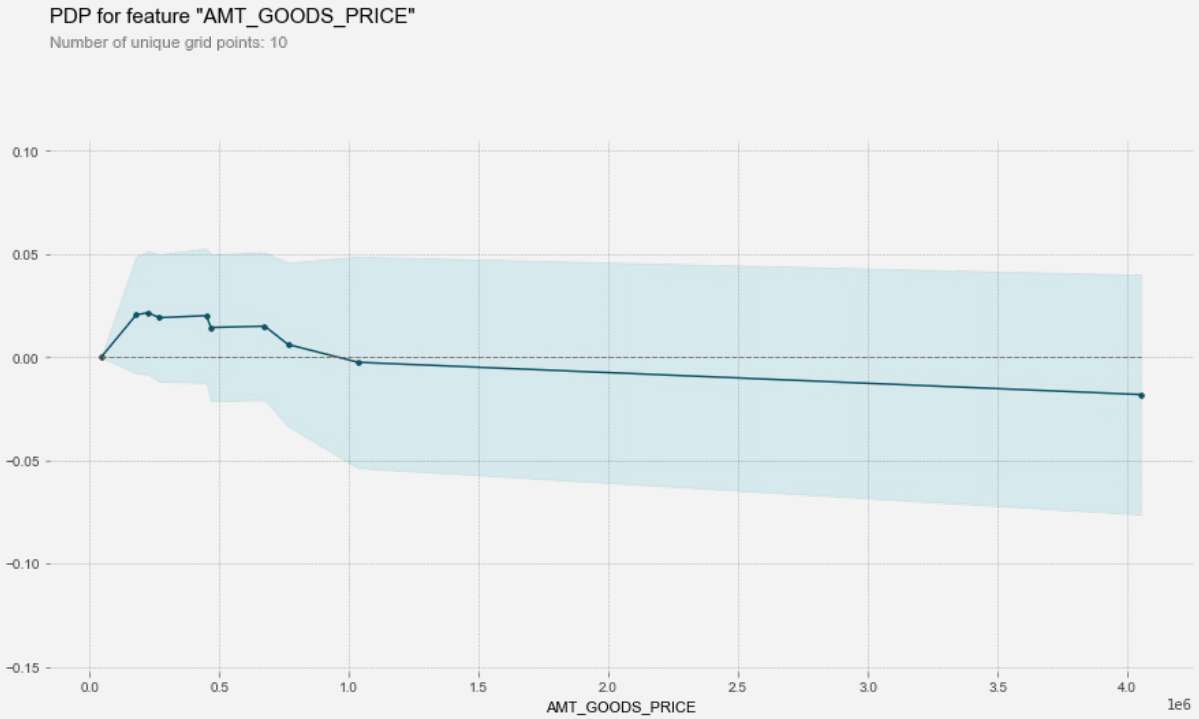
Entrainement

Weight	Feature
0.0288 ± 0.0213	EXT_SOURCE_3
0.0201 ± 0.0447	EXT_SOURCE_2
0.0134 ± 0.0145	AMT_GOODS_PRICE
0.0103 ± 0.0101	OWN_CAR_AGE
0.0080 ± 0.0112	DAYS_LAST_PHONE_CHANGE
0.0074 ± 0.0099	DAYS_EMPLOYED_DAYS_BIRTH_RATIO
0.0072 ± 0.0045	REGION_POPULATION_RELATIVE
0.0065 ± 0.0097	AMT_INCOME_TOTAL
0.0063 ± 0.0254	TOTAL_CREDIT_LENGTH_MONTH
0.0062 ± 0.0140	CODE_GENDER_F
0.0051 ± 0.0073	YEARS_BUILD_AVG
0.0043 ± 0.0065	AMT_REQ_CREDIT_BUREAU_YEAR
0.0042 ± 0.0123	DAYS_EMPLOYED
0.0036 ± 0.0067	OBS_30_CNT_SOCIAL_CIRCLE

Test
(Permutation importance)

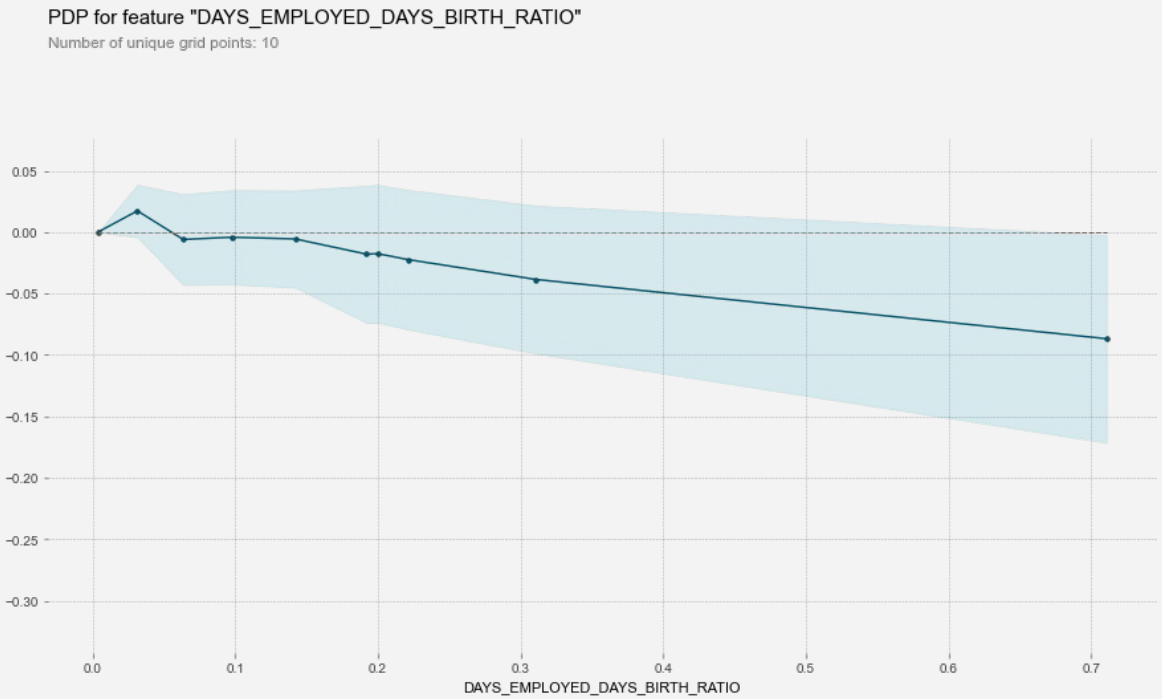
Interprétation

Niveau Global - Dépendances partielles



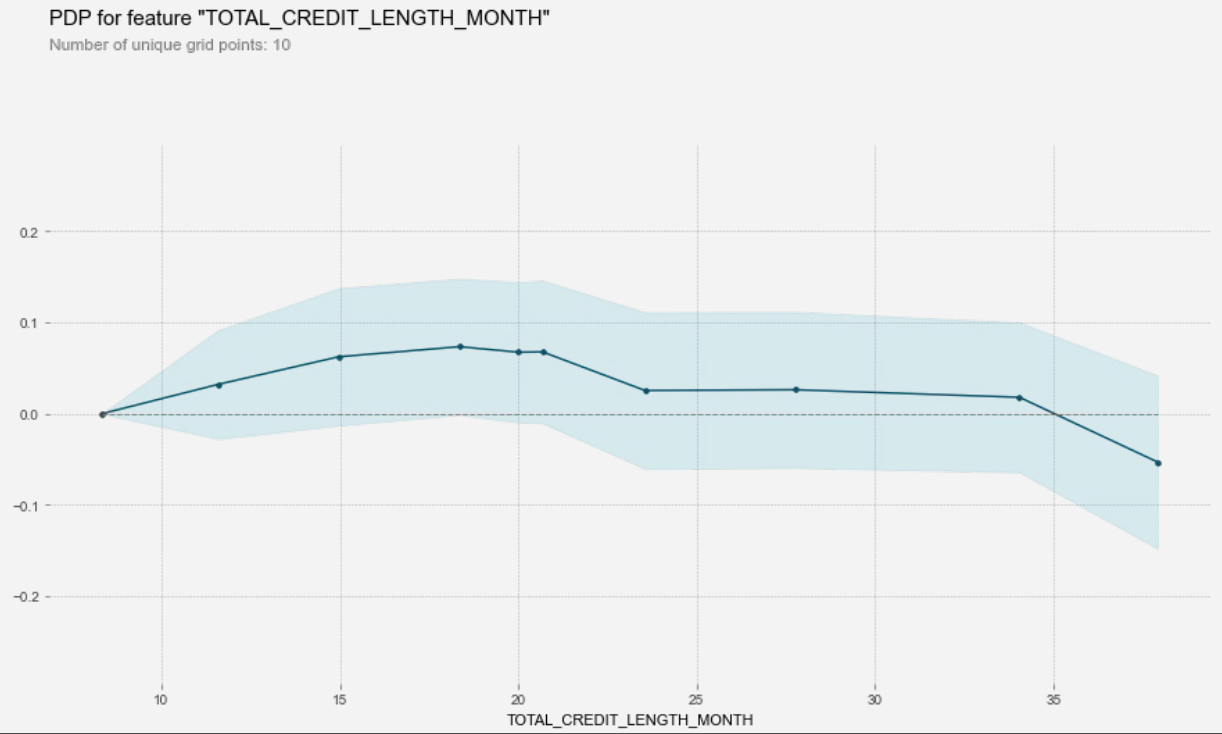
Interprétation

Niveau Global - Dépendances partielles



Interprétation

Niveau Global - Dépendances partielles

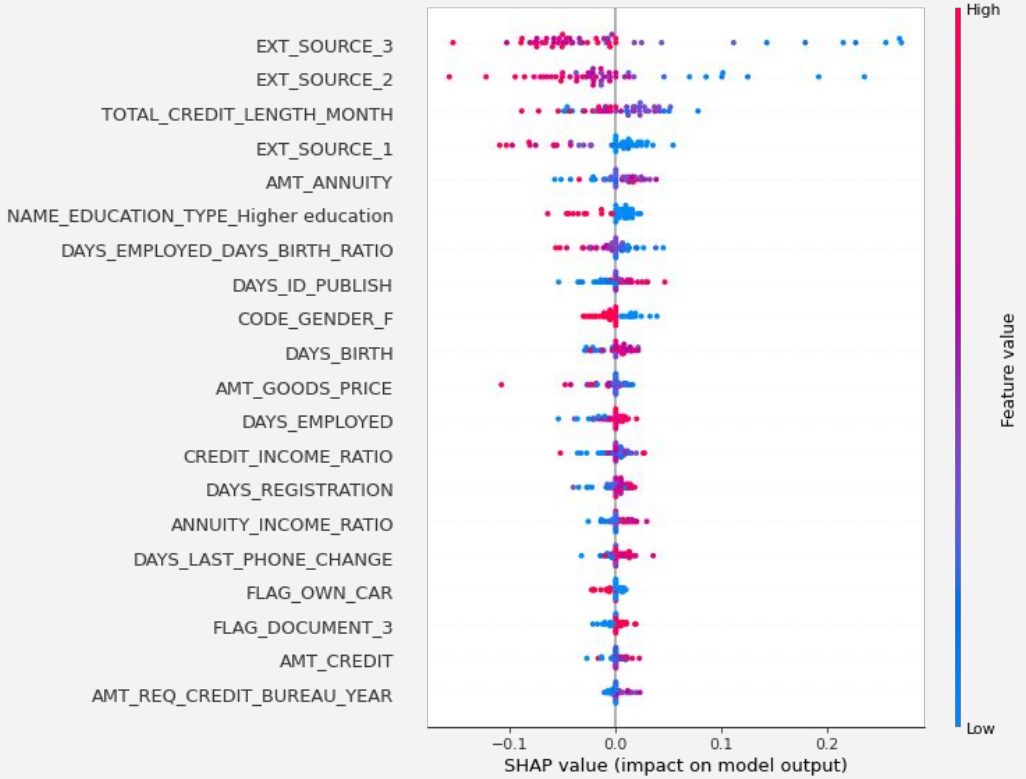


Interprétation

Niveau Local- SHAP Values

Interprétation

Niveau Global - SHAP Values

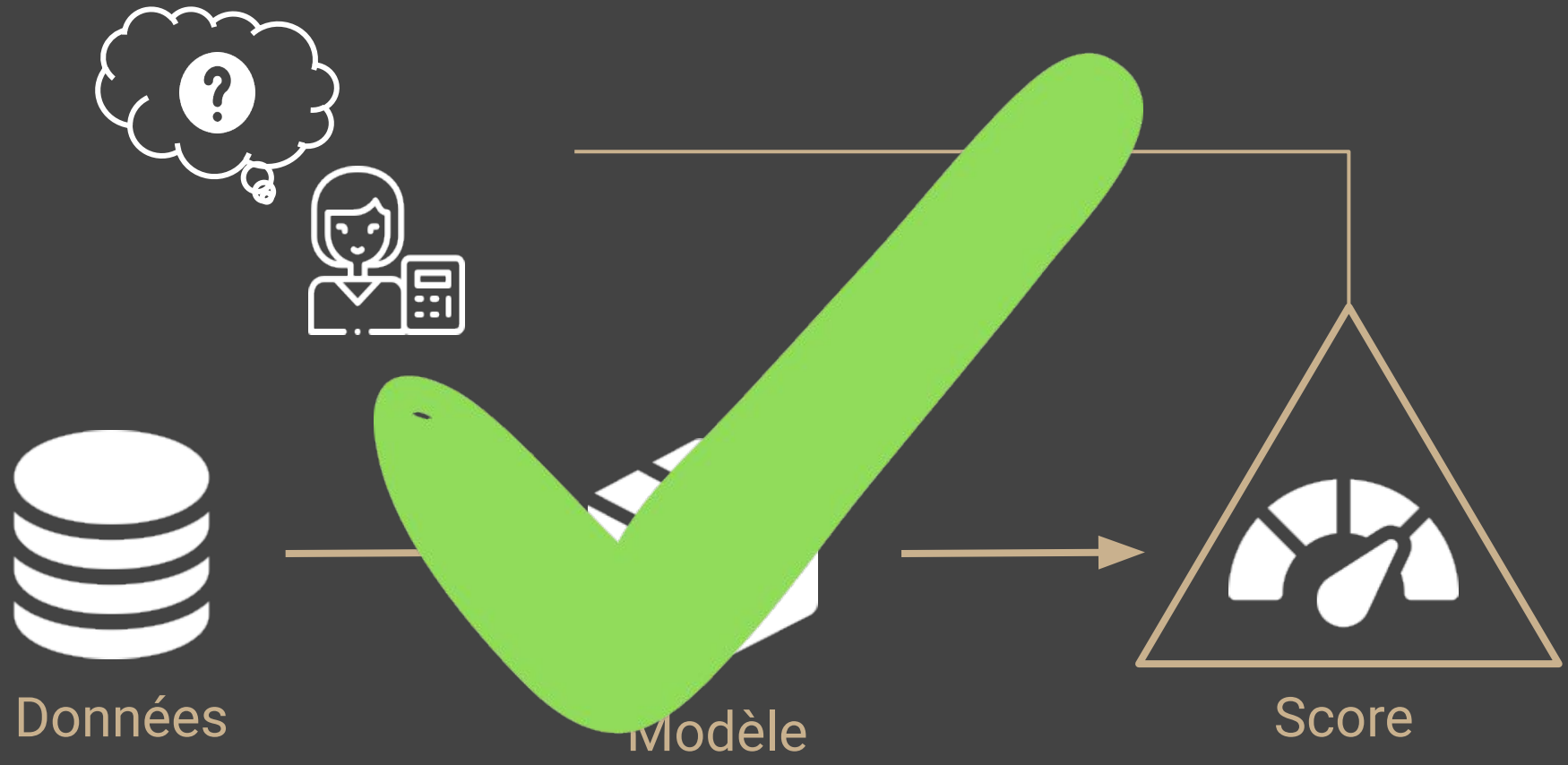


Interprétation

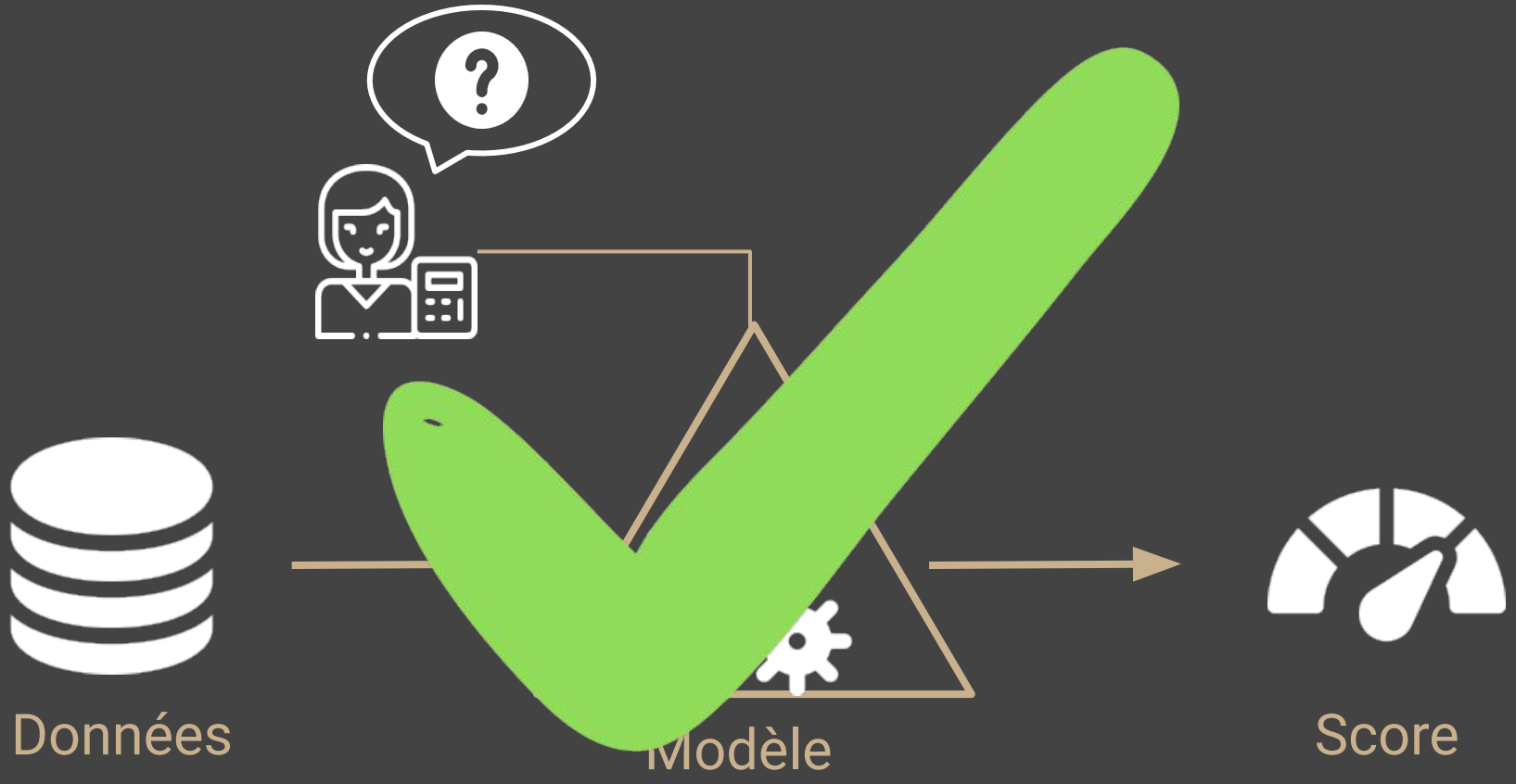
Ethique et biais

Conclusion

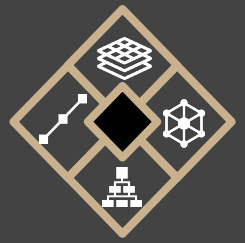
Problématique métier



Problématique métier



Pistes d’amélioration



Autres modèles

Croiser les données

Création d’autres variables

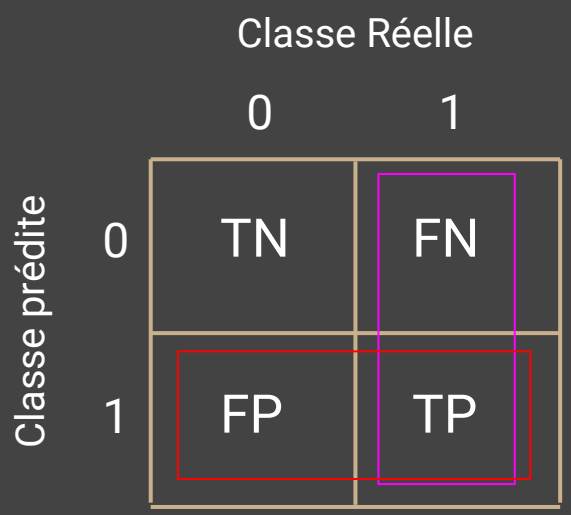
X^i

Affiner le preprocessing

Ethique

Questions

Choix de la métrique



Problématique métier:
- Identifier les mauvais payeurs
- Ne pas refuser de bons payeurs

$$\text{Precision} = \frac{TP}{FP + TP}$$

$$\text{Rappel} = \frac{TP}{FN + TP}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Rappel}}{\text{Precision} + \text{Rappel}}$$

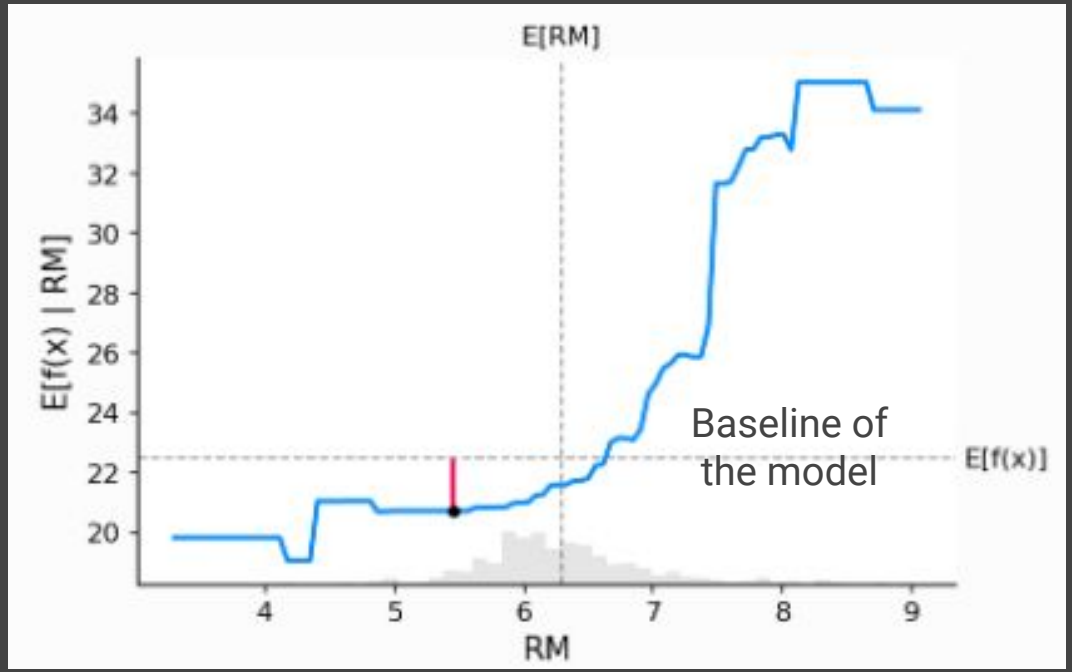
Permutation importance

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Permutation importance
=
Changement de la prédiction en permutant les valeurs de la variable

Interprétation

SHAP value



Actual prediction

SHAP value