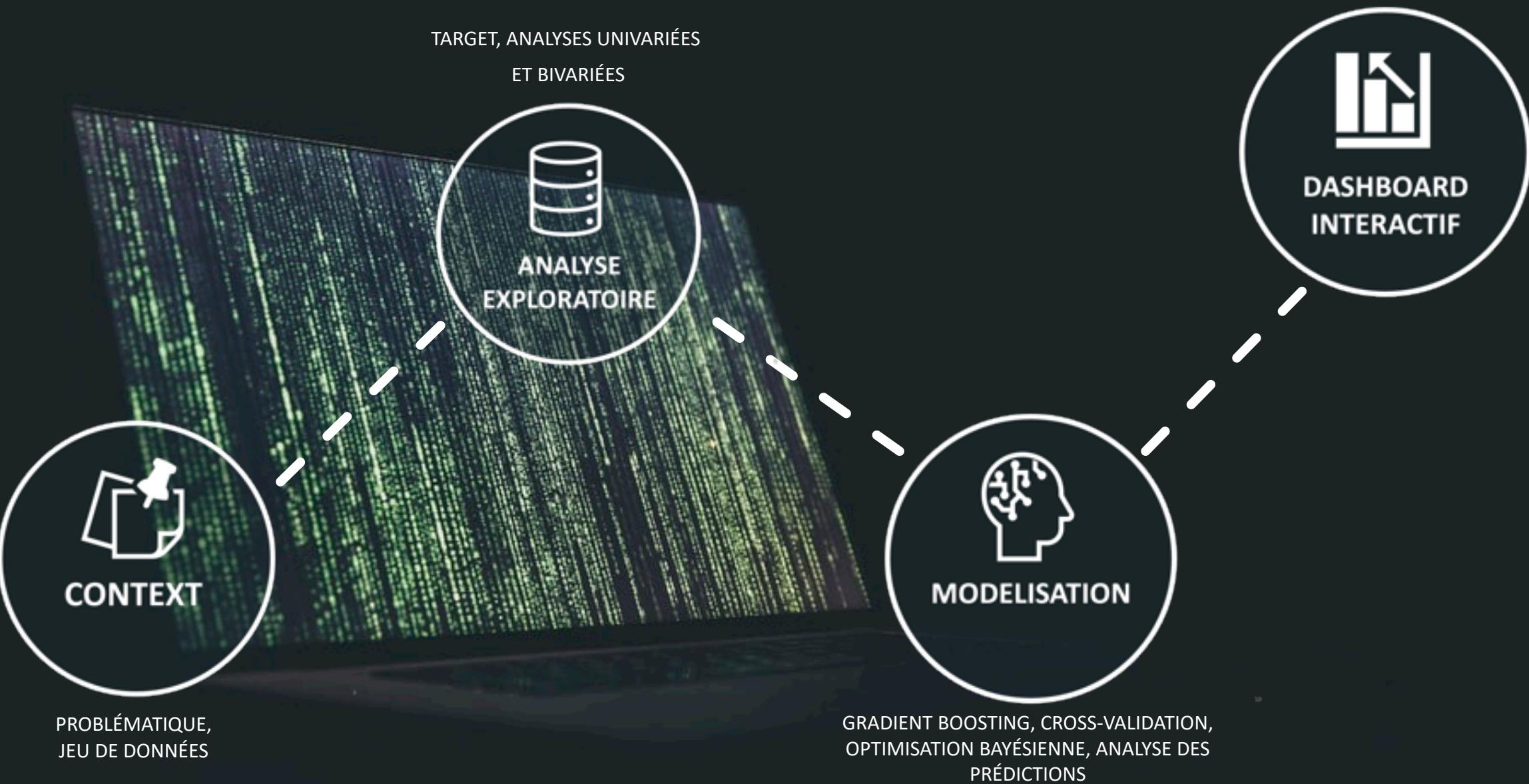




# P7 IMPLÉMENTEZ UN MODÈLE DE SCORING

Formation Data Scientist  
OpenClassrooms





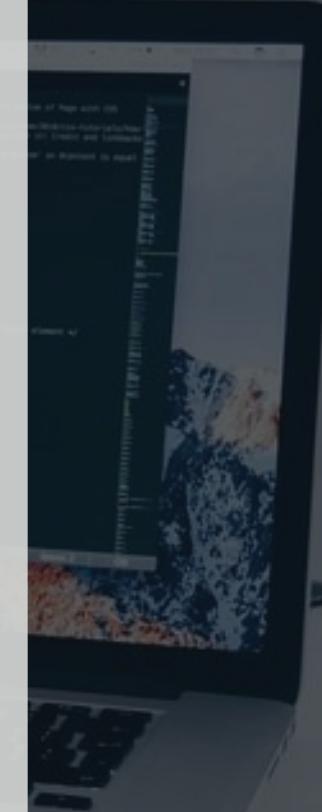
CONTEXT



# CONTEXT



Prêt à dépenser





# CONTEXT

**Problématique :** développer un modèle de scoring de la probabilité de défaut de paiement du client et un dashboard interactif



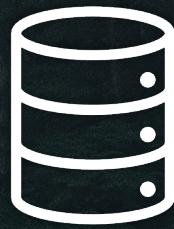
ANALYSE  
EXPLORATOIRE

# JEU DE DONNÉES

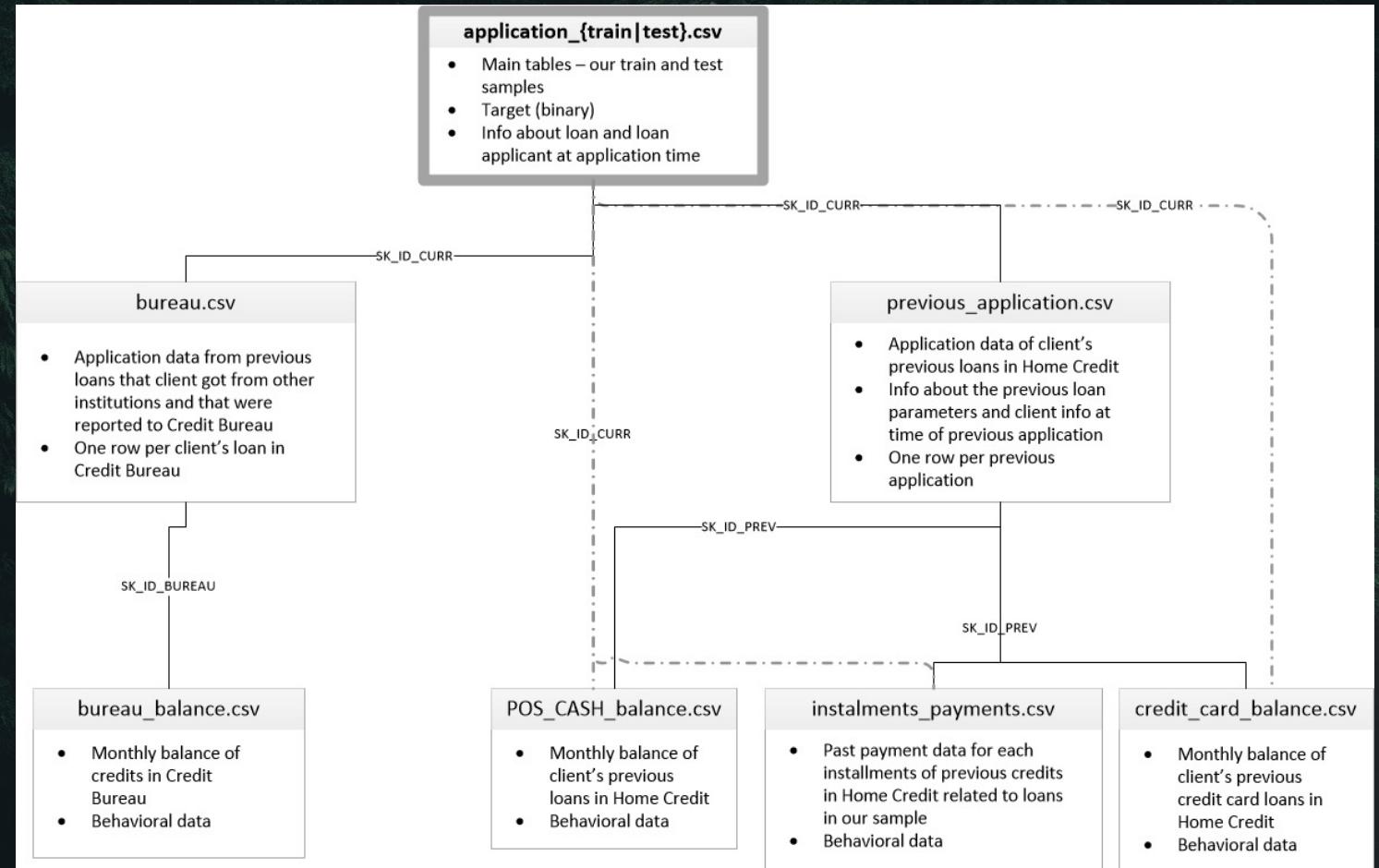
- Kaggle compétition  
<https://www.kaggle.com/c/home-credit-default-risk/data>

- 7 fichiers
- + 300 000 clients

- Agrégation -> transformation de variables



## ANALYSE EXPLORATOIRE



# JEU DE DONNÉES

Target très inégalement distribuée

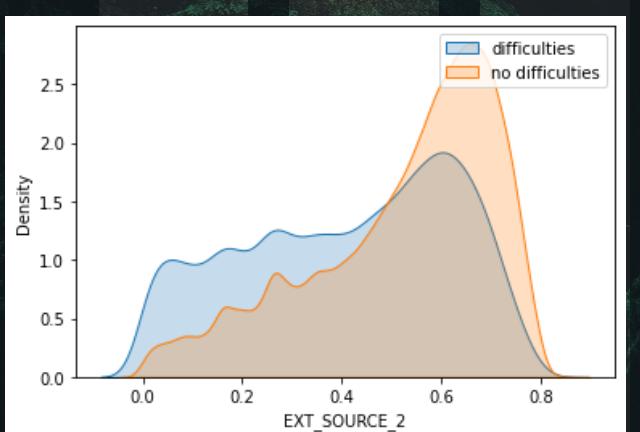
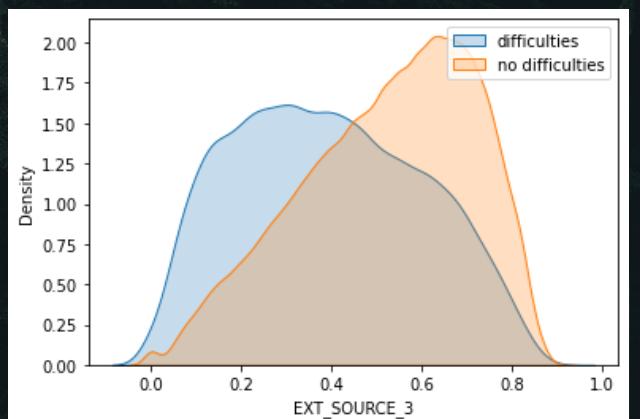
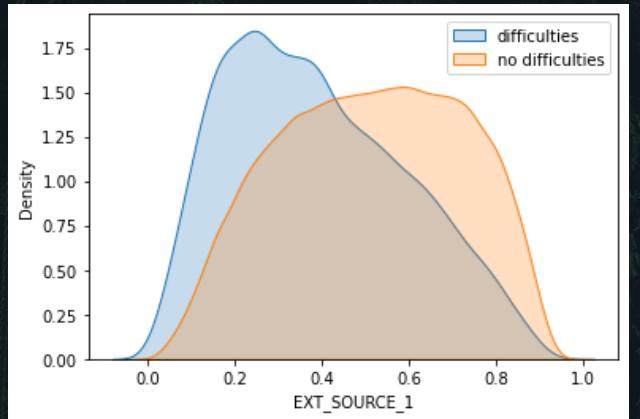
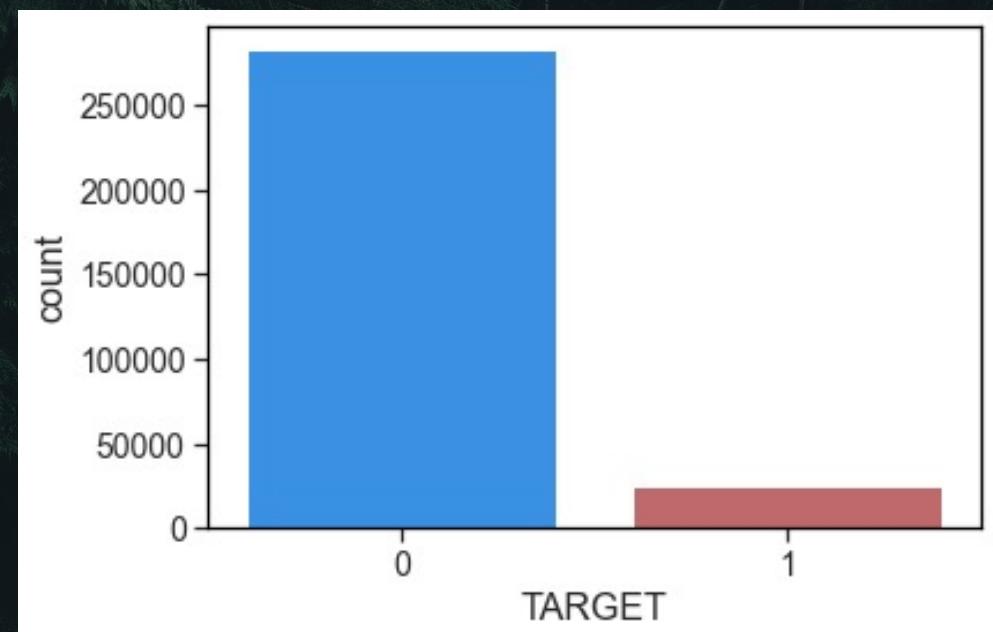
**Client solvable = 0, Client non-solvable = 1**

Analyse des variables ayant un impact sur la target (ANOVA, Cramer's V)

Analyse des corrélation entre les variables  
(matrice de Spearman)



## ANALYSE EXPLORATOIRE





# JEU DE DONNÉES POUR MODÉLISATION

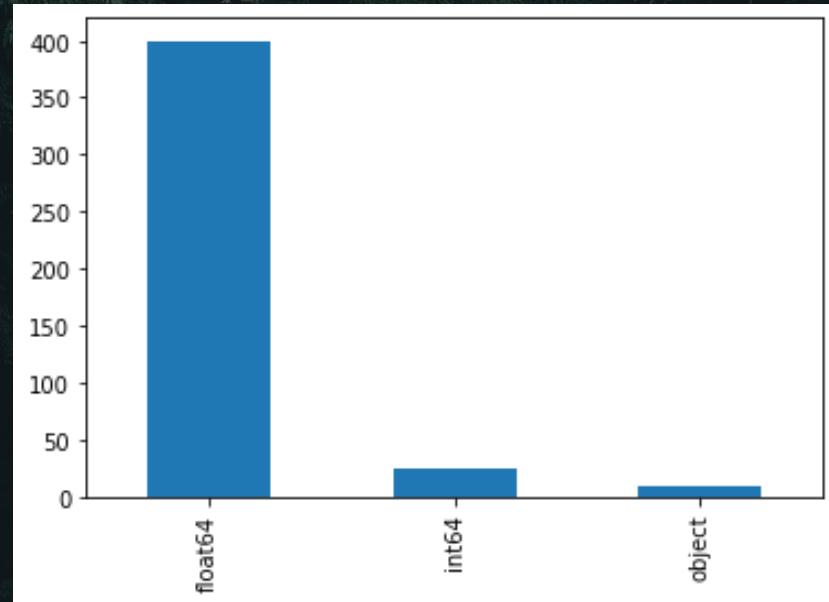
Valeurs manquantes :

- suppression si > 40 % de valeurs manquantes sauf si fortement corrélé à la Target (ANOVA).

Multi colinéarité :

- supprimer de variables si coefficient de corrélation > 0,9.

Le jeu de données final comportait 431 variables dont 9 catégorielles, 421 continues et la variable Target.

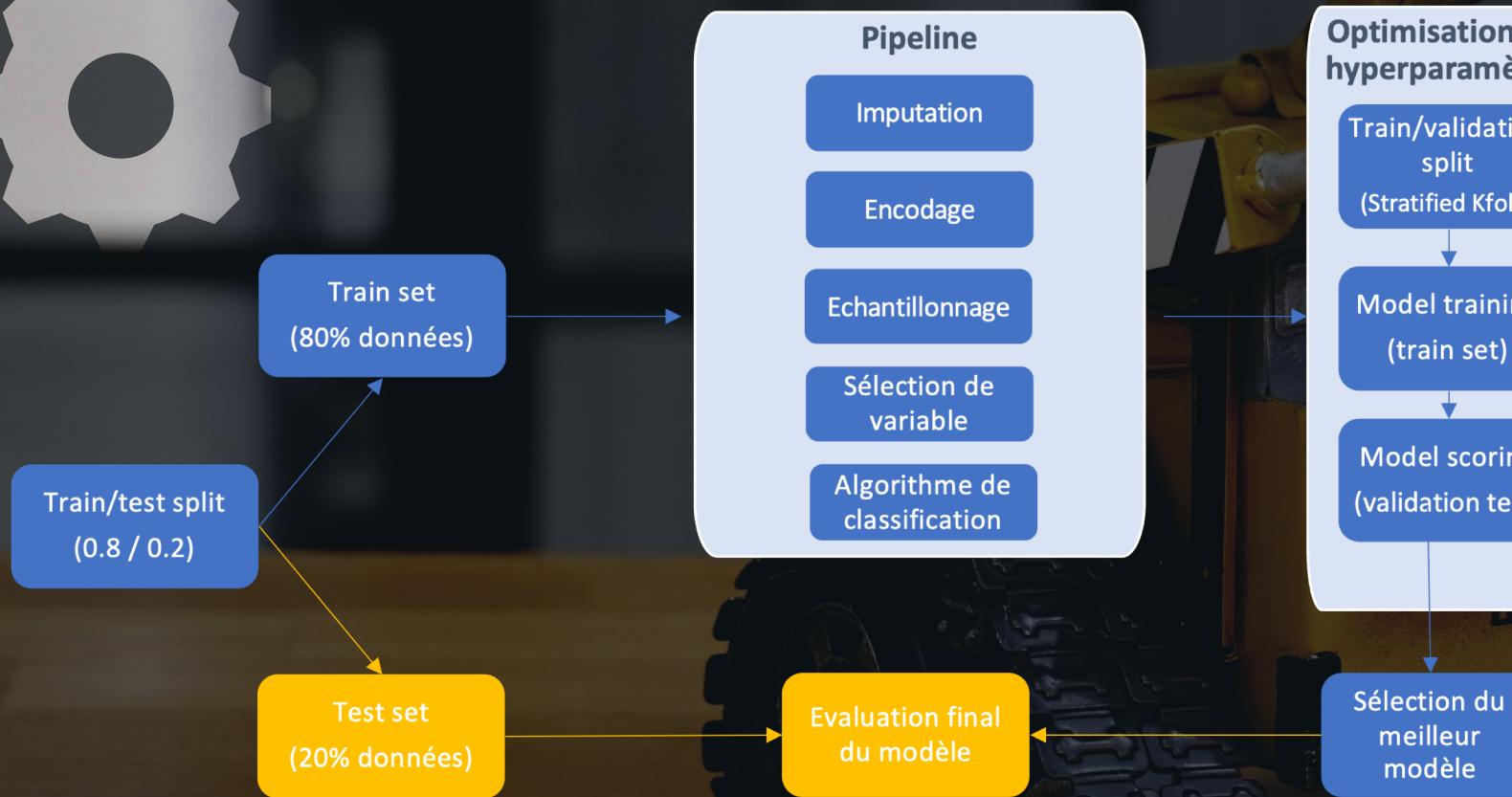




# MODELISATION

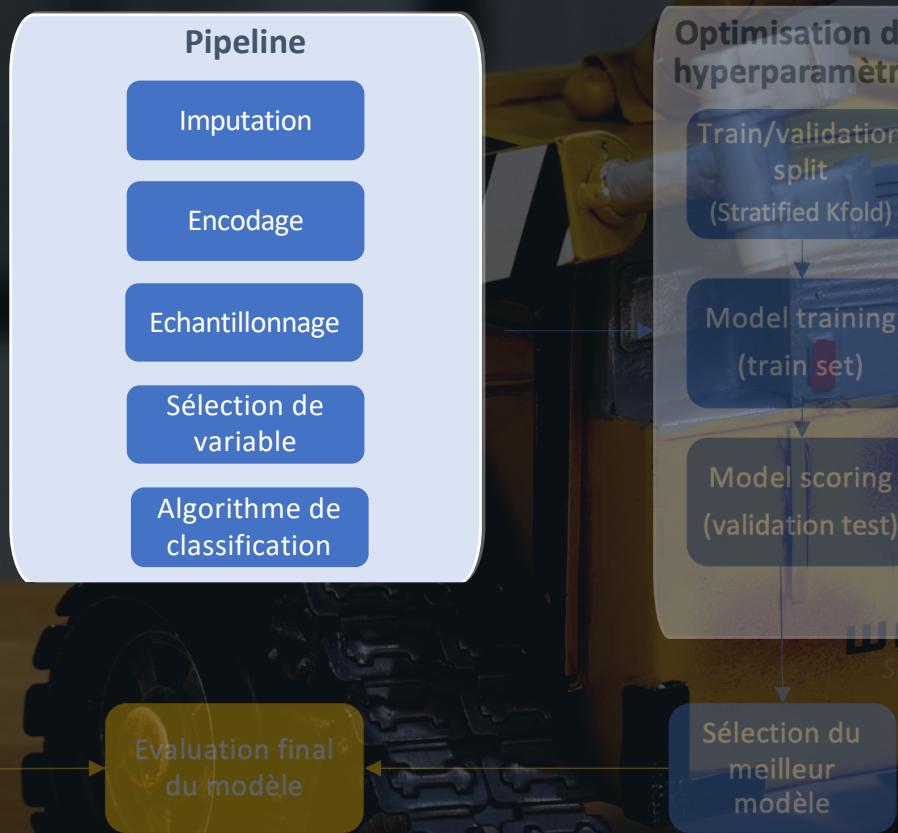


# MODELISATION





# MODELISATION



## EVALUATION DU PRÉTRAITEMENT :

- imputation ou non des valeurs manquantes restantes ou suppression
- encodage (variables catégorielles transformées : One Hot Encoder et variables numériques Standard Scaler).
- déséquilibre de la TARGET (Class weights, Over-sampling, SMOTE, Under-sampling)

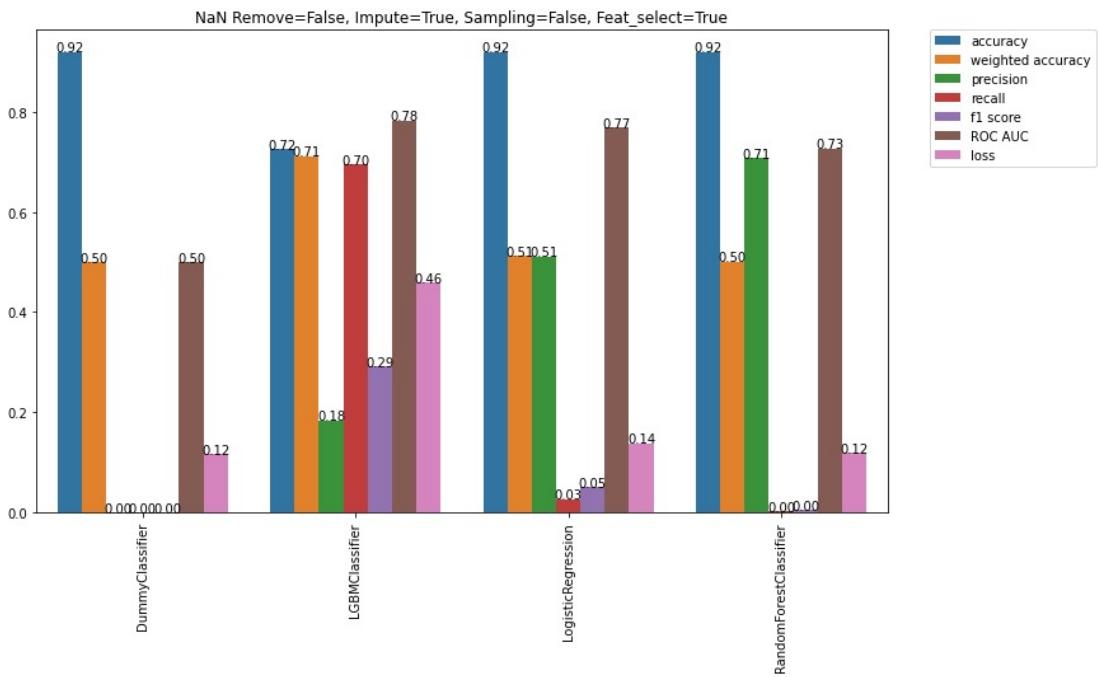
## ALGORITHME DE CLASSIFICATION :

- Un algorithme dit naïf, comme baseline
- Logistic Regression
- Random Forest Classifier
- Light-GBM

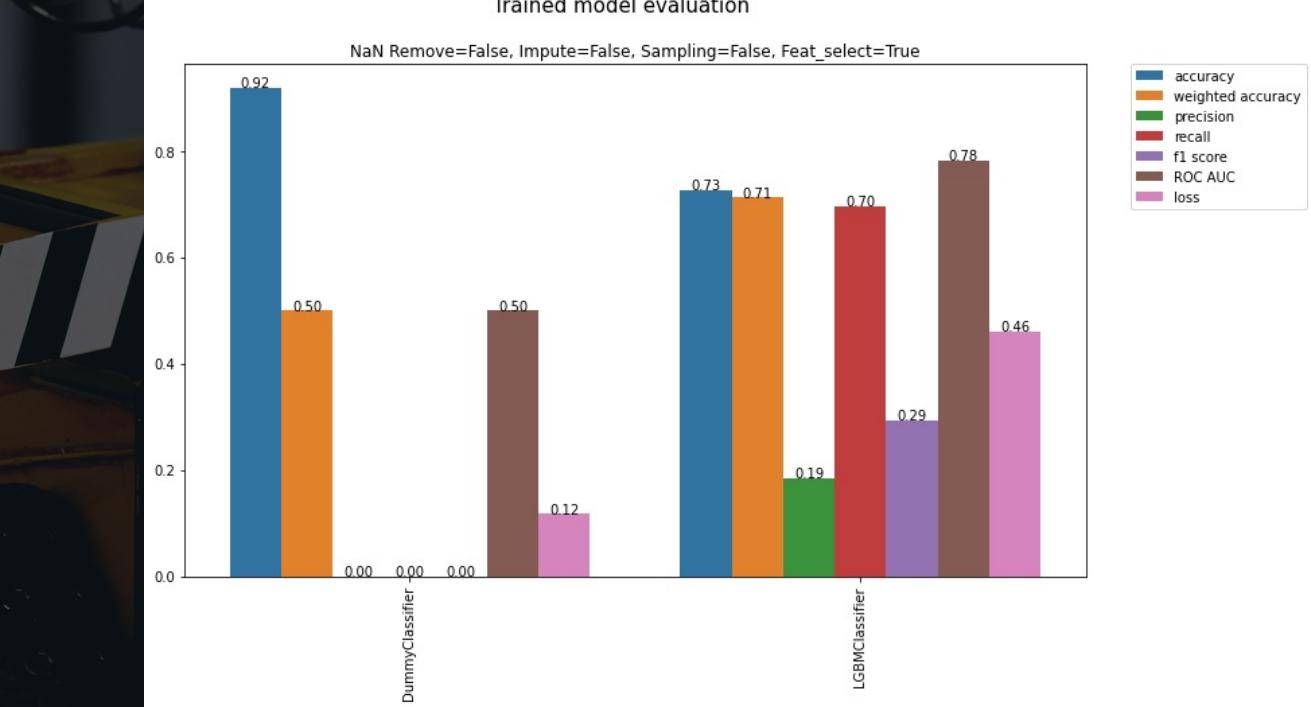


# MODELISATION

Trained model evaluation

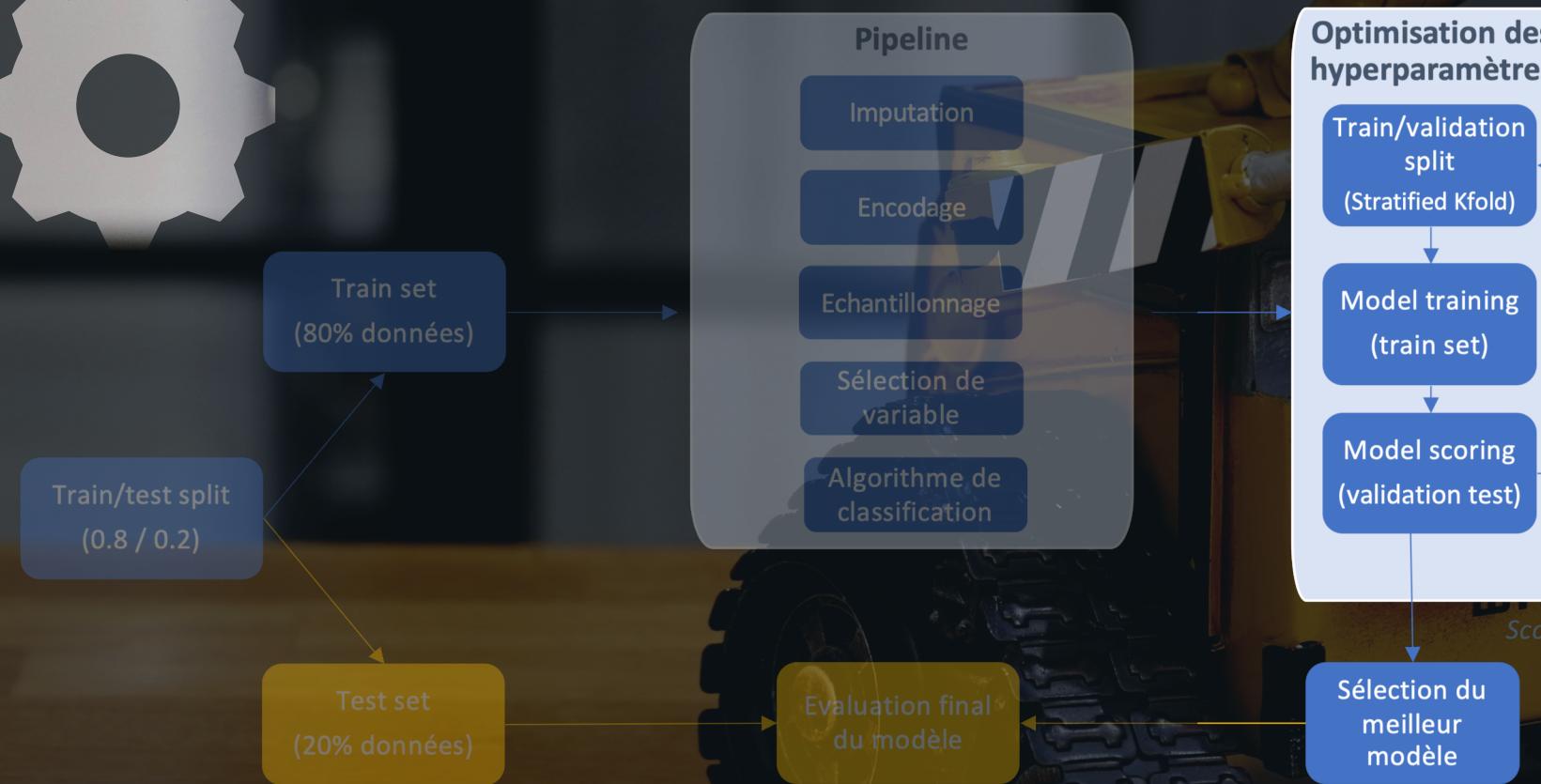


Trained model evaluation





# MODELISATION



OPTIMISATION BAYESIENNE –  
HYPERPARAMÈTRES

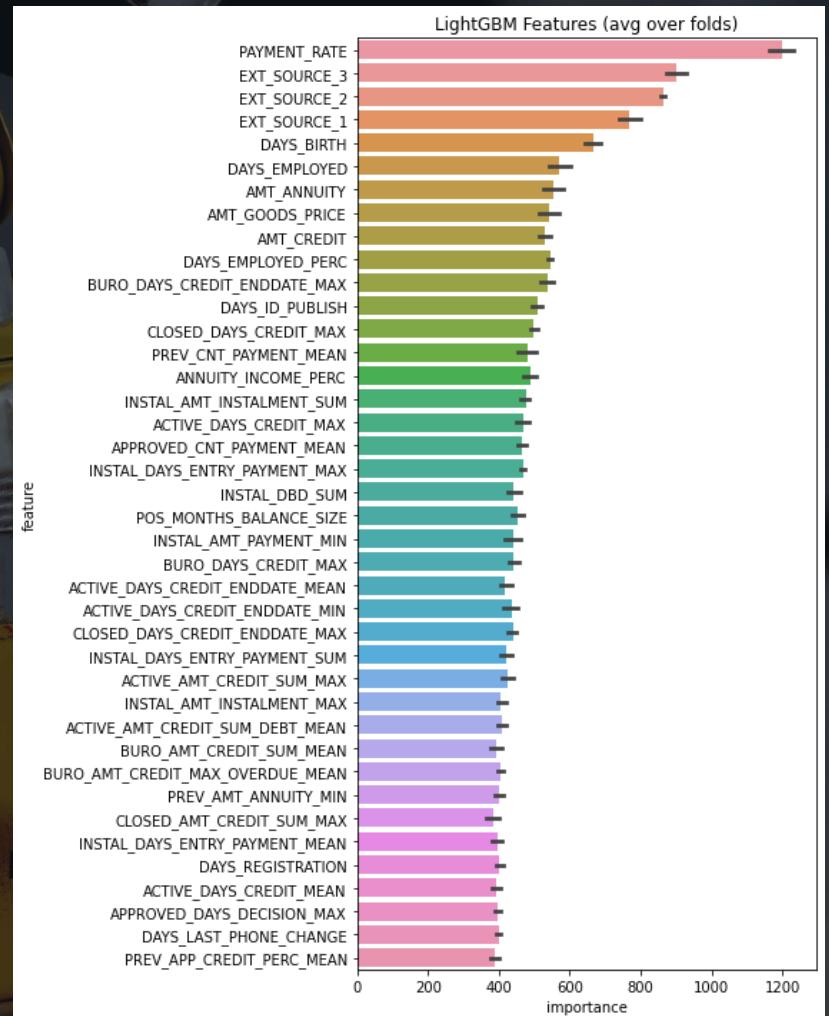
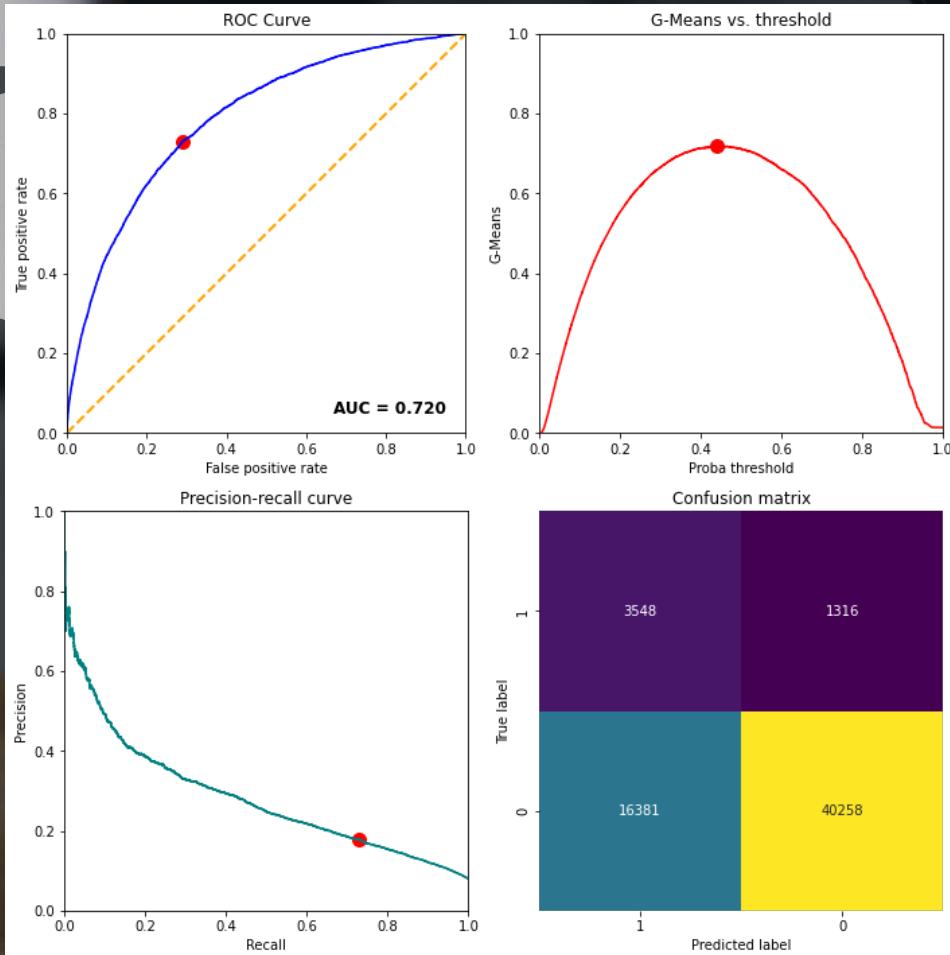
**Light GBM :**

application = binary  
n\_estimators = 100  
learning\_rate = 0.02  
**scale\_pos\_weight = 11**  
colsample\_bytree = 0.283  
max\_depth = 13  
min\_child\_weight = 41.017  
min\_split\_gain = 0.471  
num\_leaves = 30  
reg\_alpha = 0.84  
reg\_lambda = 0.0  
subsample = 0.47

AUC = 0,786 (+0,01)  
Fonction coût métier = 0,471 (+ 0,01)



# MODELISATION





DASHBOARD  
INTERACTIF



# DASHBORD

114450

## General informations:

Gender: Male

Age: 59

Education level: Secondary / secondary special

Marital status: Married

Family members : 2 (including 0 children)

Work: Working

Work experiences: 8 years

Income: 49,500 \$

## Credit informations:

Credit amount: 553,806 \$

Annuity amount: 22,090 \$

Payment rate: 3.99%

Check credit score

Client more informations

## Evaluate your client's credit capacity.



Client #114450 has 4.13 % of risk to make default.

We recommend to accept client's application to loan.



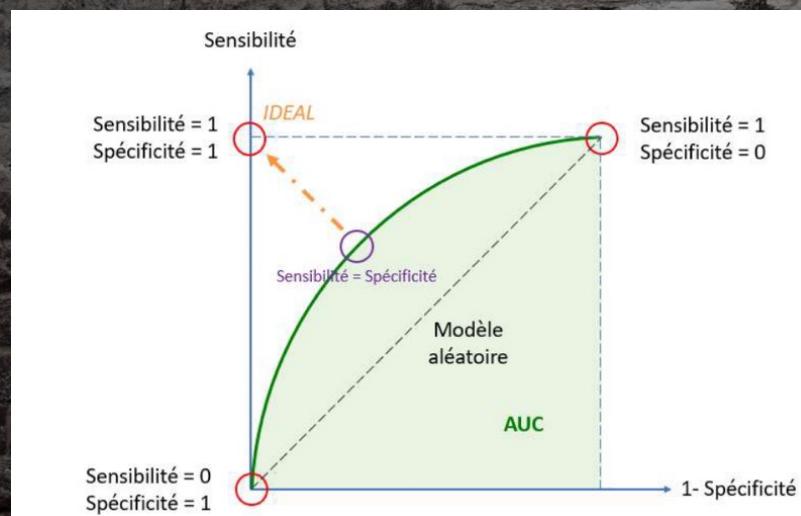
Below 30% of default risk, we recommend to accept client application. Above 50% of default risk, we recommend to reject client application. Between 30 and 50%, your expertise will be your best advice in your decision making. You can use the "client more informations" page to help in the evaluation.

## Prediction explanation



# QUESTIONS – Métriques classifier

		Classe réelle	
		-	+
Classe prédictive	-	True Negatives (vrais négatifs)	False Negatives (faux négatifs)
	+	False Positives (faux positifs)	True Positives (vrais positifs)



Pour notre problématique, le résultat attendu le plus important pour un client est la valeur de la probabilité de défaut de paiement. En appliquant un seuil à cette valeur, nous pouvons lui affecter une valeur binaire (0 ou 1) suivant que la probabilité est inférieure ou supérieure au seuil.

Si la probabilité est inférieure au seuil, on considère que le crédit sera remboursé, la prédiction est négative (0). Inversement, si la probabilité est supérieure au seuil, on considère que le crédit ne sera pas remboursé, la prédiction est positive (1).

Ainsi :

- Accorder un crédit à un client ne pouvant pas le rembourser par la suite (FN) est synonyme de perte.
- Accorder un crédit à un client qui le remboursera par la suite (TN) est un gain.
- Ne pas accorder le prêt et que le client ne peut pas rembourser (TP) n'est ni une perte, ni un gain.
- Ne pas accorder le prêt alors que le client pouvait rembourser (FP) est une perte de client donc d'argent.

Ces valeurs peuvent être traduites en des indicateurs caractérisant le modèle, dont voici les plus importants :

- Sensibilité =  $TP / (TP + FN)$  - Capacité du modèle à détecter les dossiers de crédit non remboursés (1)
- Spécificité =  $TN / (TN + FP)$  - Capacité du modèle à détecter les dossiers de crédit remboursé (0)
- Précision =  $TP / (TP + FP)$  - Capacité du modèle à détecter les vrais dossiers non remboursés (1)

# QUESTIONS – définition de la fonction coût métier

En utilisant les valeurs de la matrice de confusion nous pouvons calculer le nombre total de client qui sont capable de rembourser le prêt soit  $TN + FP$ . Soit  $\alpha$  le bénéfice moyen si un client a remboursé son prêt. Les gains maximums sont donc de  $\alpha \times (TN + FP)$ .

Cependant, les réels bénéfices fait par la société sont les gains obtenus par le remboursement d'un prête plus les pertes subit par le défaut de remboursement. En prenant cette fois  $\beta$  comme la perte moyenne si un client fait défaut les gains sont  $\alpha \times TN + \beta \times FN$ . Donc la fonction coût est :

$$\frac{\alpha \times TN + \beta \times FN}{\alpha \times (TN+FP)}$$

Sans information métier concrète nous avons utilisé des valeurs estimées pour **le coefficient alpha et beta, 20 000 et 200 000 respectivement**. Cette fonction coût a été implémentée afin de pénaliser l'impact des erreurs sur la décision d'octroi de crédit.