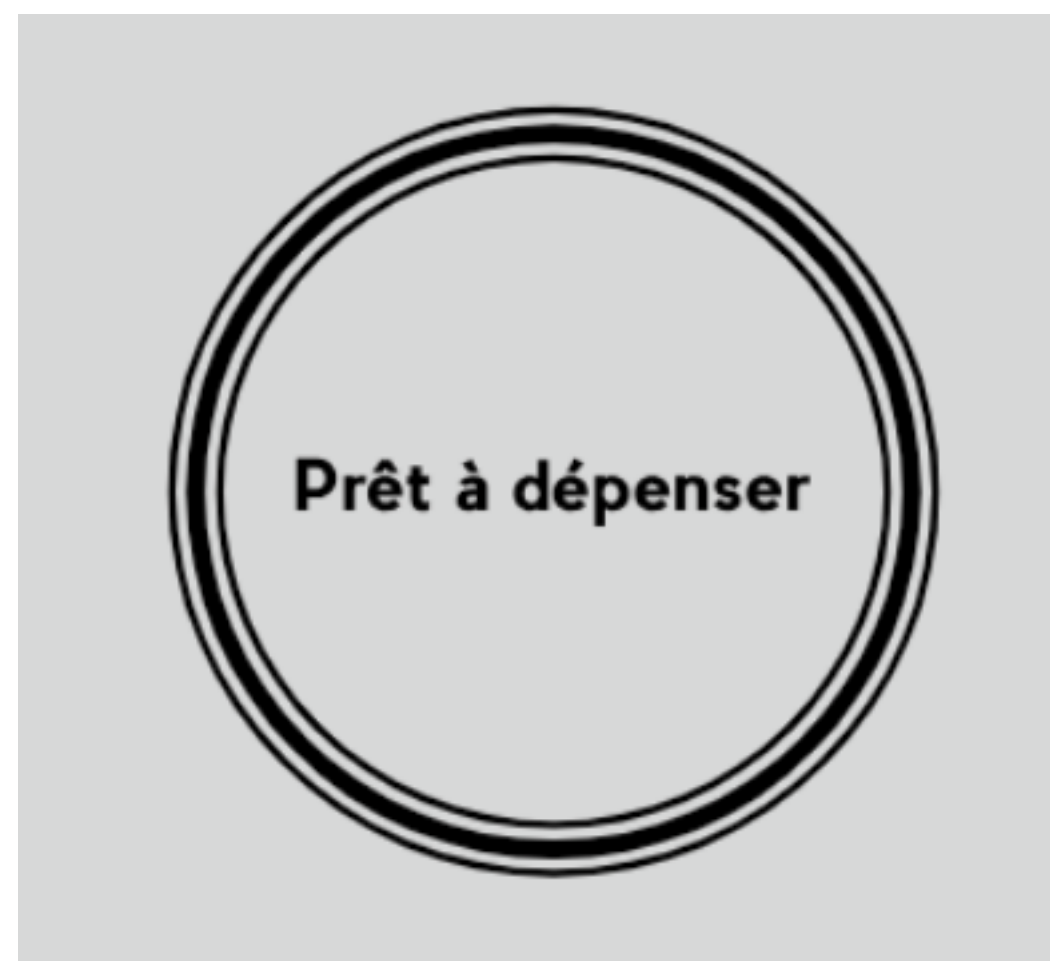


# Implémentation d'un Modèle de Scoring



**OPENCLASSROOMS**

**Eloi Le Quilleuc 23/11/2021**

# Objectif du Projet

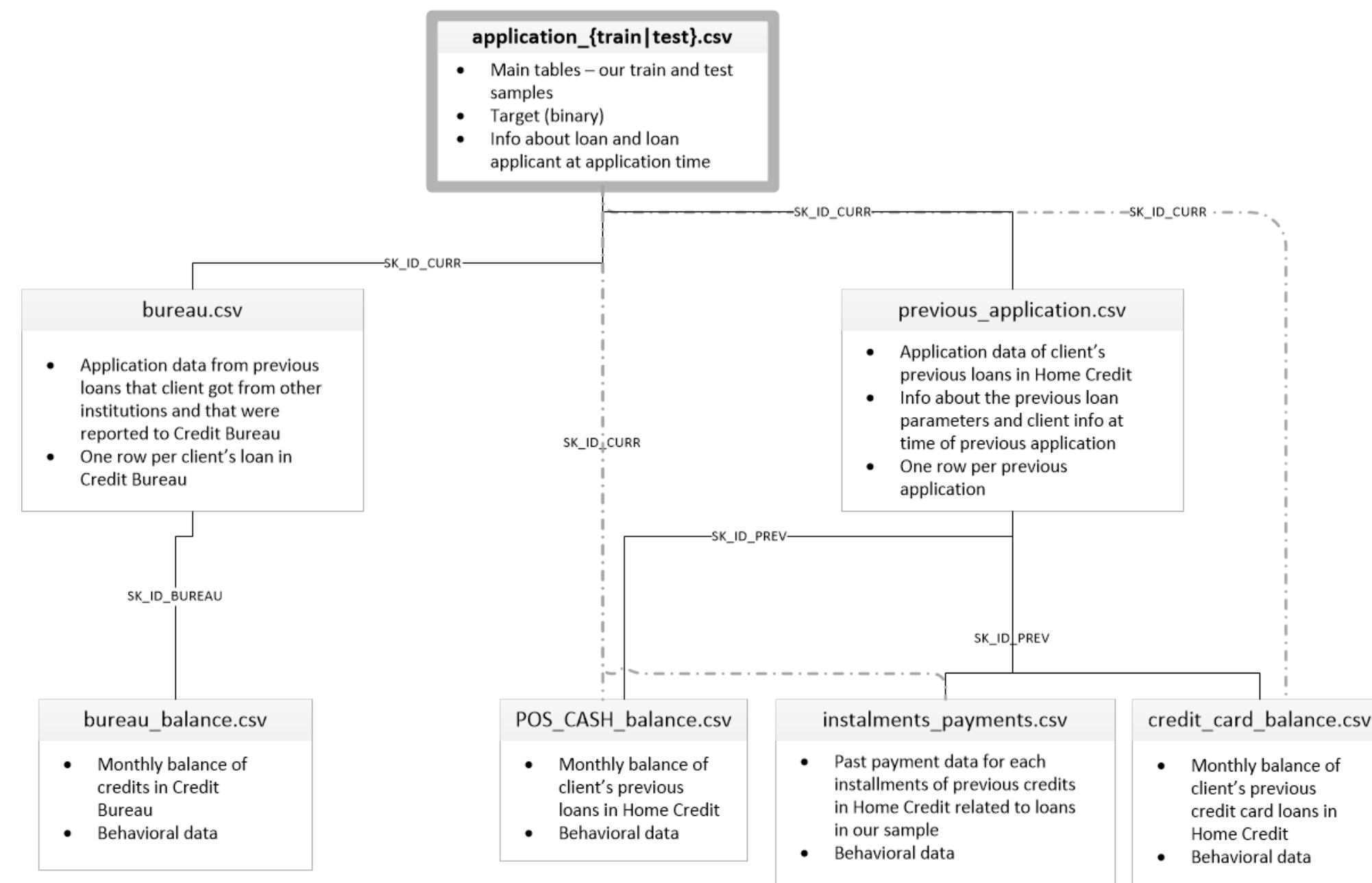
***Prêt à Dépenser*** : société financière qui fournit des crédits à ses clients

1. réaliser un modèle pour identifier les clients
2. rendre le modèle *accessible* et *interprétable* depuis le web

# Le Modèle de Classification

# Les Données

On réalise un modèle statistique, entraîné sur les données de Prêt à Dépenser



Ensemble des tables d'information  
client de Prêt à Dépenser

Tables pré-  
traitées  
récupérées à  
partir d'un  
kernel Kaggle  
(lien)

kaggle

# Le Pré-traitement

**Sélection des clients** : ceux enregistrés dans application\_train.csv

**Valeurs manquantes** : prises en compte

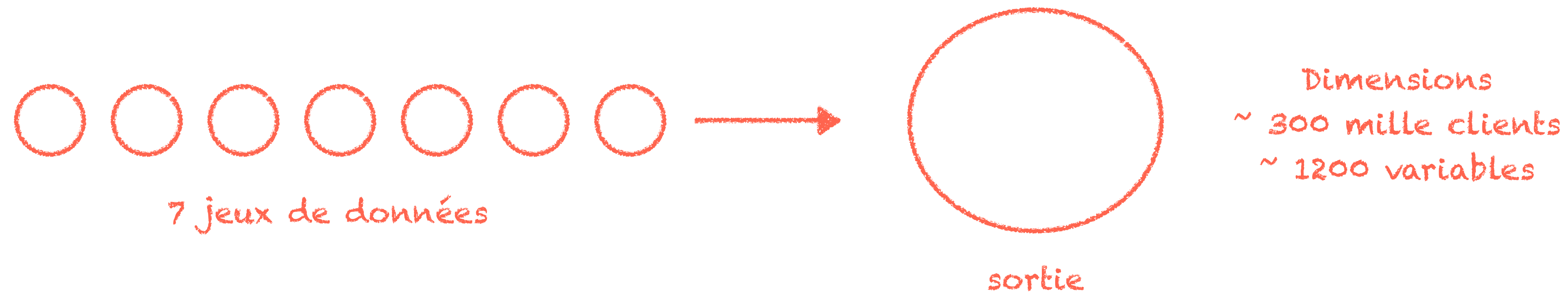
**Encodage** : One Hot pour les variables catégorielles

**Feature Aggregation** : une ligne par client pour chaque table (sum, mean)

**Feature engineering** : combinaison de variables, mean, max

# Les Données de Modélisation

Les 7 jeux de données pré-traités sont fusionnés.



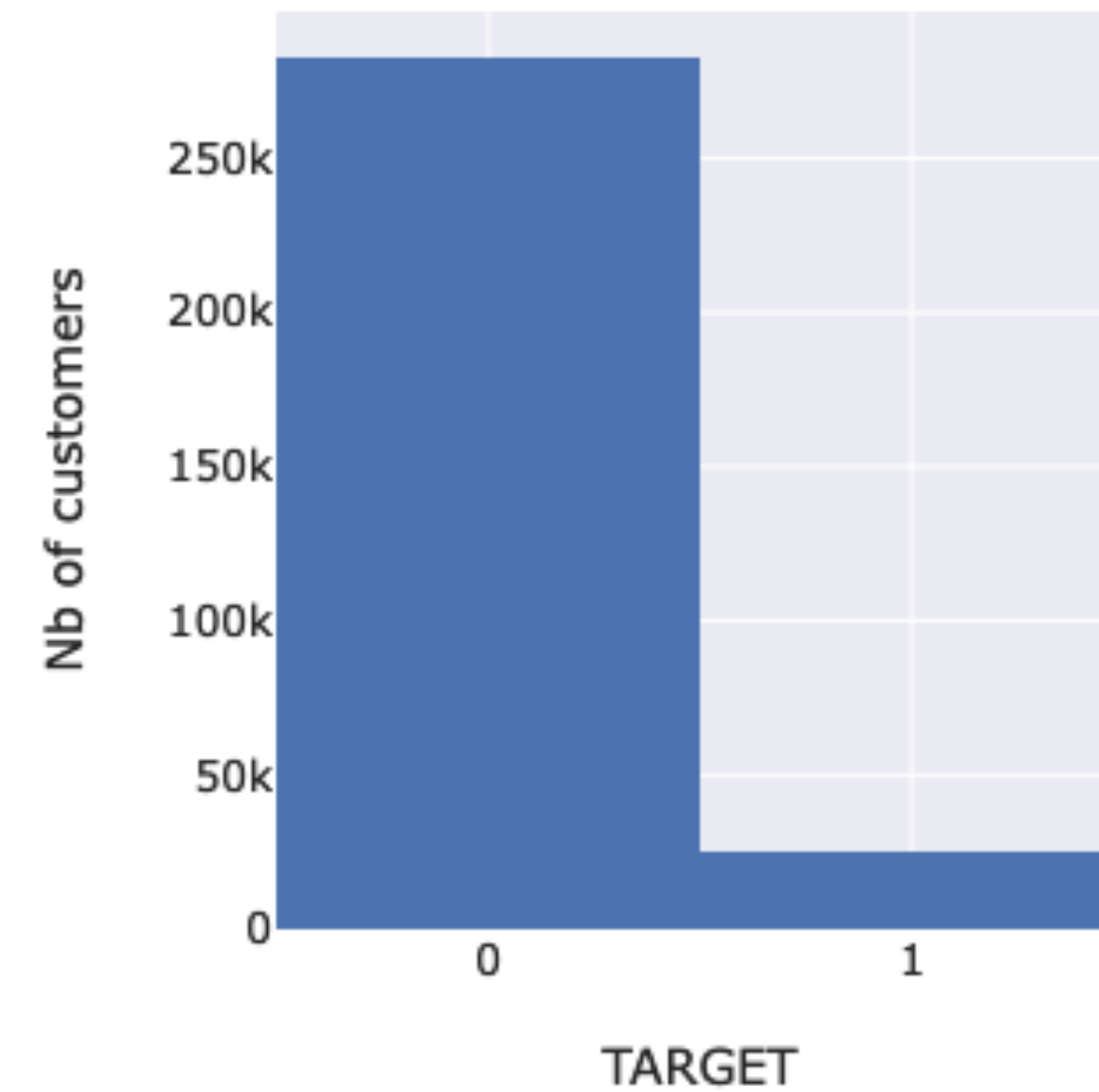
Séparation entraînement - test, avec les proportions 93 % - 7 %

# Variable cible

**0** : le client rembourse ses crédits

**1** : le client n'a pas remboursé un crédit

~ 10 % des clients ne  
remboursent pas leur  
crédit



# Algorithme d'Apprentissage Machine

**Classifieur LightGBM** : arbre de décision boosté

**Hyperparamètres** : par défaut (100 arbres, 32 feuilles, taux d'app. de 0.1)

**Fonction de coût** : régression logistique

**Métrique d'évaluation** : aire sous la courbe ROC

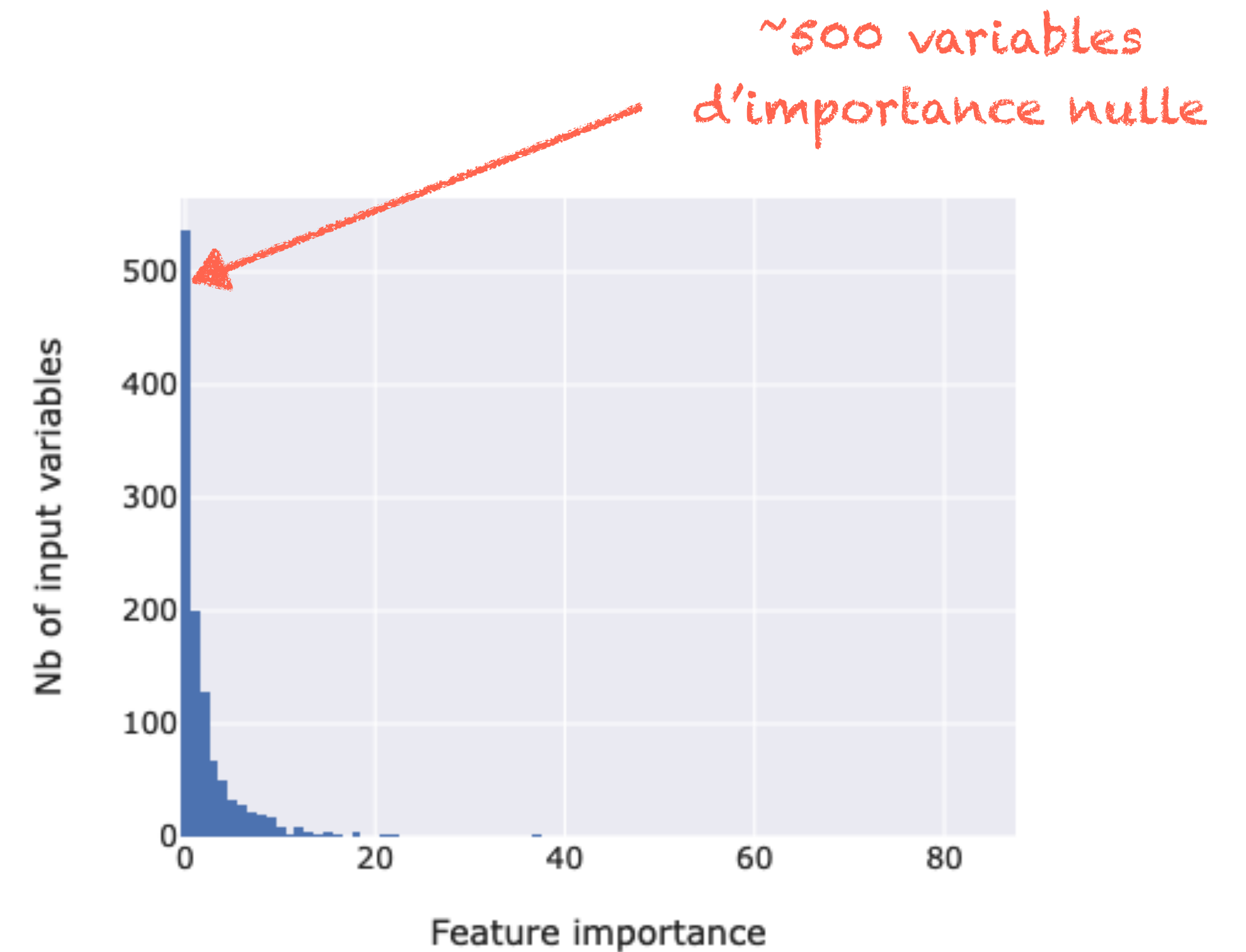




# Premiers Résultats

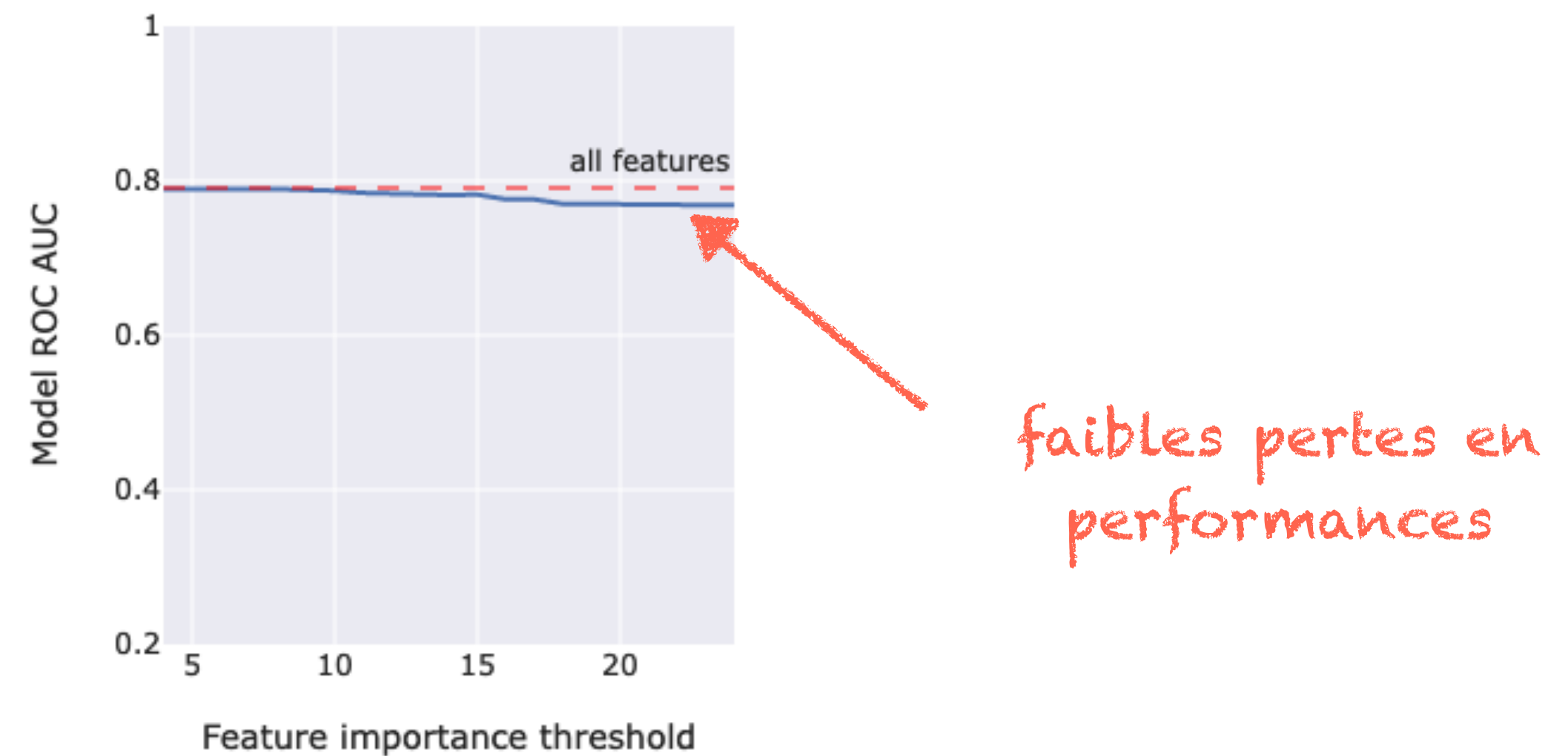
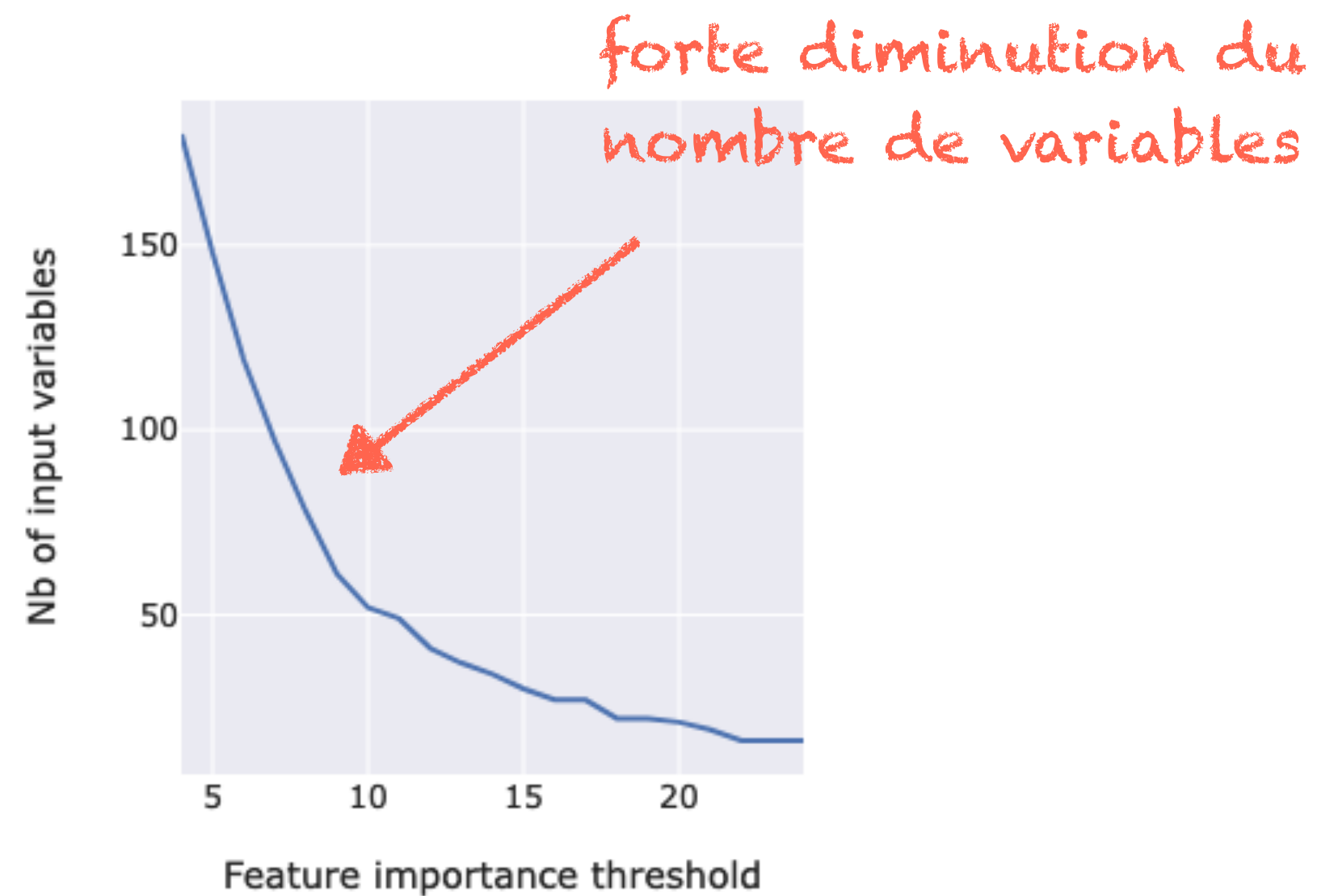
**ROC AUC : 0.79** -> bonnes performances

**Inconvéniant : 1200 variables** -> trop complexe



# Performances vs Nb de Variables

**Amélioration** : retirer des variables d'entrée par ordre d'importance



# Modèle Final

Nombre de variables : 37

**ROC AUC** : 0.783 ( < 1 % de perte par /  
baseline 1200 variables)

**Intervalle de confiance 95 %** :  
[0.771 - 0.794] (± 1.5 %)

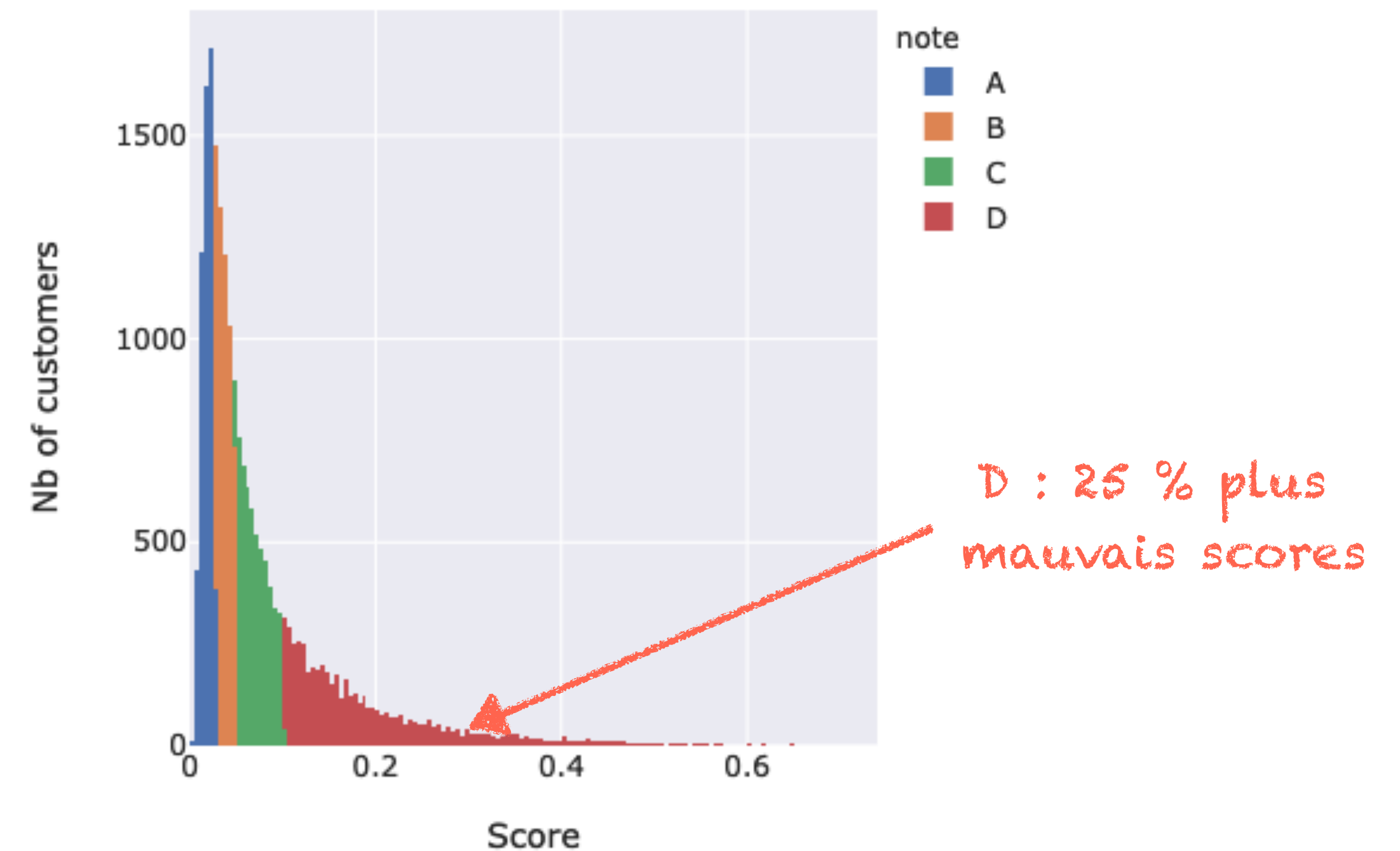
	feature_importance
NEW_EXT_MEAN	87
PAYMENT_RATE	83
PREV_DAYS_LAST_DUE_1ST_VERSION_MAX	57
AMT_ANNUITY	56
DAYS_BIRTH	54
EXT_SOURCE_3	46
EXT_SOURCE_1	44

7 variables les plus importantes pour le modèle

# Score Crédit

**Score** : probabilité de défaut de paiement

**Note** : répartition des scores en quartiles




Score médian : 5 %

Répartition des clients en fonction  
du score crédit

# L'Application Web

# Page Web : Deux Parties

1. **Dashboard** : accéder au score crédit client et son interprétation
2. **Predicteur** : utiliser le modèle statistique sur de nouveaux clients



The screenshot shows a web application interface with a sidebar on the left and a main content area on the right. The sidebar is titled 'Application ML' and contains two radio buttons: 'Dashboard' (selected) and 'Predictor'. The main content area is titled 'Dashboard Prêt à Dépenser'. It features a form with a label 'Quel est l'identifiant client ?' and a text input field containing '295201'. Below the input field, it says 'Vous avez selectionné le client : 295201'. At the bottom, there is a section titled 'Probabilité de défaut de paiement' with two columns: 'Score (%)' and 'Note'. The 'Score (%)' column shows '3.6' and the 'Note' column shows 'B'.

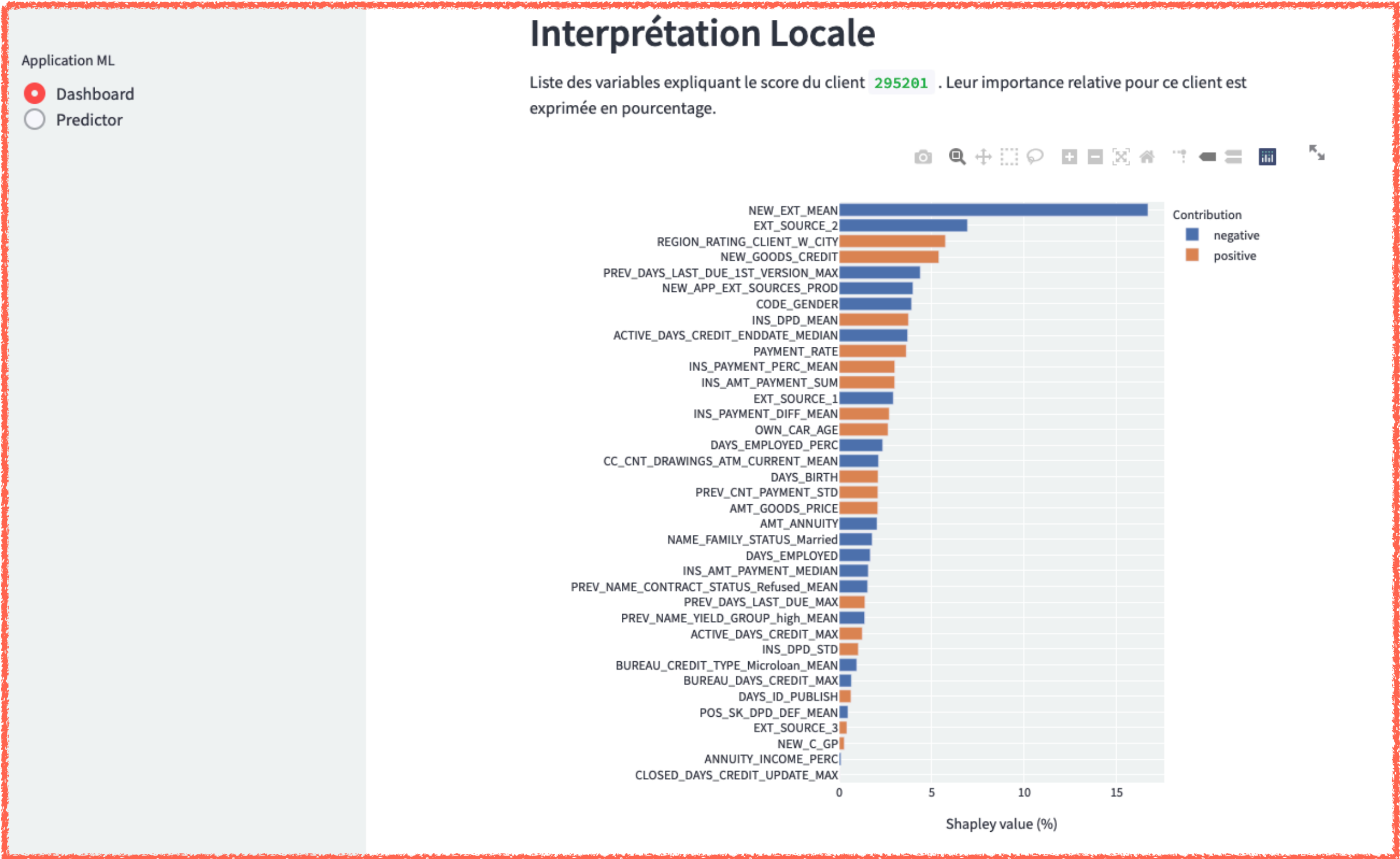
Score (%)	Note
3.6	B



# Dashboard

## Interprétation Locale : pour 1 client donné

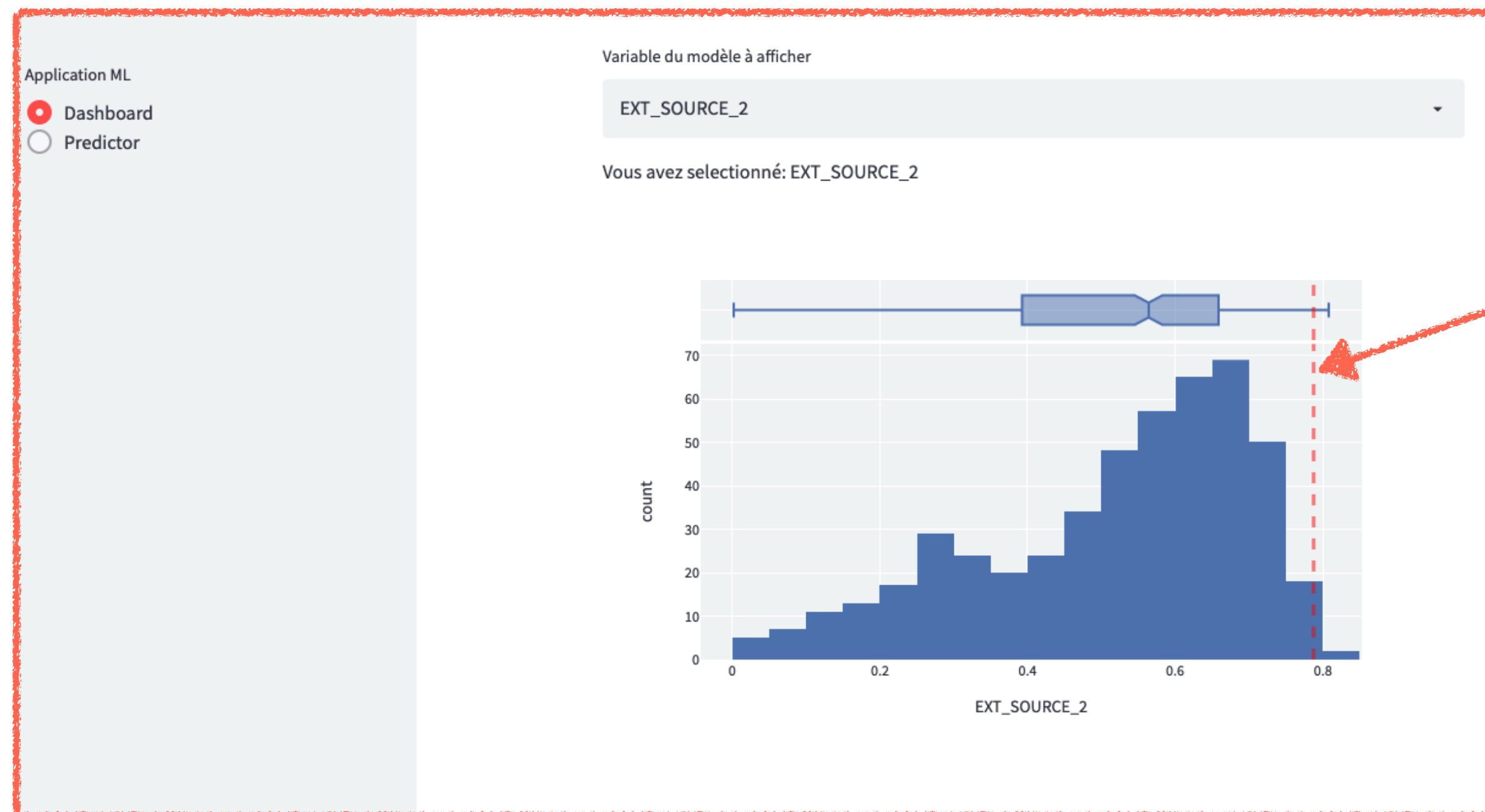
**Coefficient de Shapley** : quantifie l'influence de chaque variable sur la prédiction du score



# Dashboard

## Explication Semi-globale

**Distributions des variables**, ou ‘visualiser où se situe le client X par rapport à la population de clients’



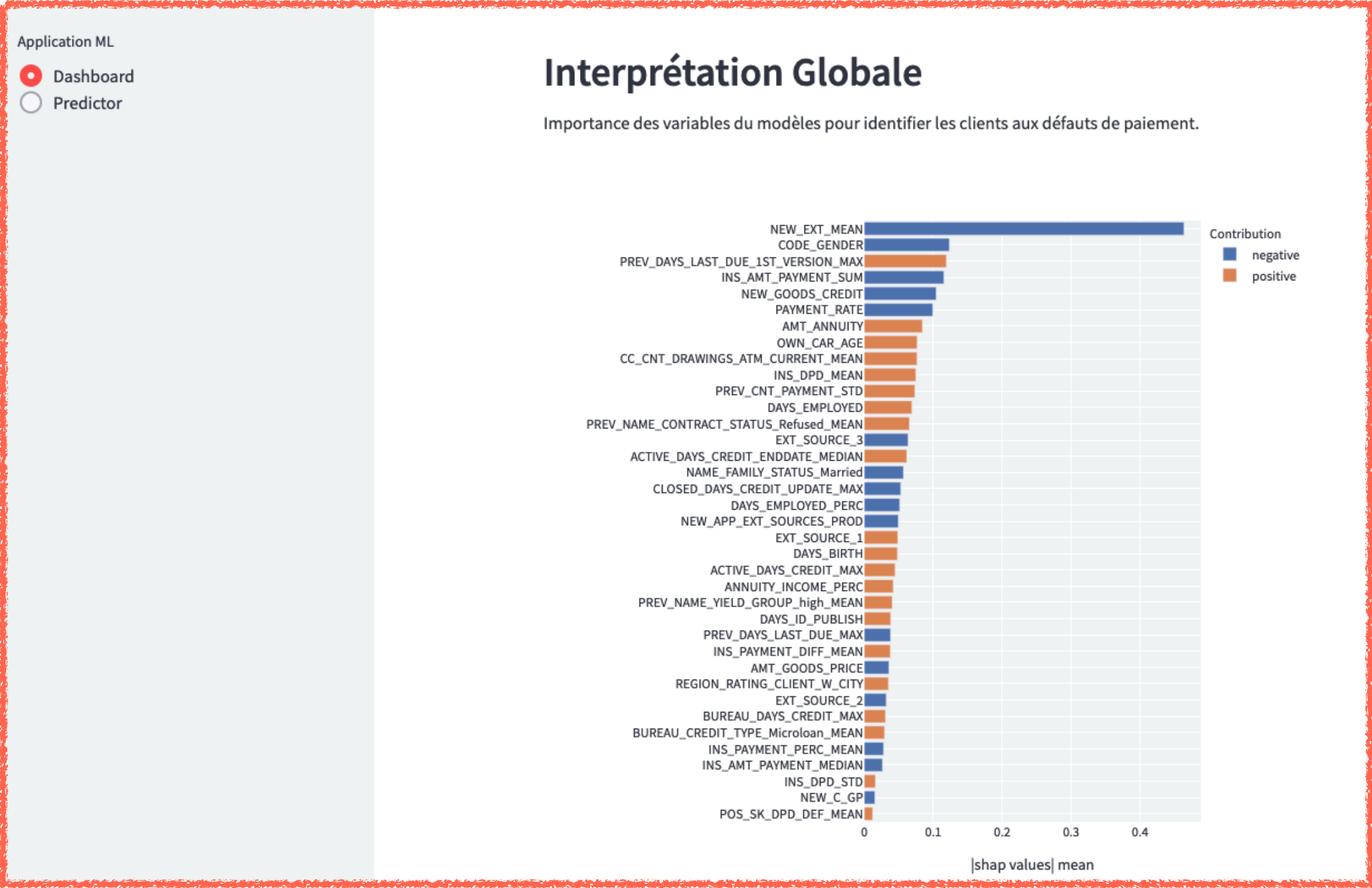
valeur pour le client X



# Dashboard

## Interprétation Globale

Coefficients de Shapley : moyenne des valeurs absolues pour chaque variable



# Prédicteur

**Entrée** : fichier .csv avec les variables d'entrée client

**Sortie** : score crédit du modèle

Application ML

Dashboard

Predictor

Prédicteur Prêt à Dépenser

Choose a file

Drag and drop file here  
Limit 200MB per file

Browse files

X\_eval\_sub.csv

259.1KB

X

Quel est l'identifiant client ?

139189

Vous avez selectionné: 139189

Score (%)

6.7

Données client

	139189
CODE_GENDER	0.0000
AMT_ANNUITY	34,308.0000
AMT_GOODS_PRICE	675,000.0000
DAYS_BIRTH	19,402.0000

# Conclusion

# Conclusion

À travers ce projet, on a réalisé :

- un modèle de classification des clients non payeurs, aux bonnes performances, relativement simple
- une application web pour utiliser le modèle et interpréter les résultats

# Perspectives

## Améliorations

- Optimisation des hyperparamètres LightGBM
- Balance 50-50 des classes à prédire
- Utilisation du bagging

**Limites** : des scores crédit faibles (grande majorité  $< 50\%$ ) -> un modèle difficile pour prendre une décision. Utile pour classer les clients