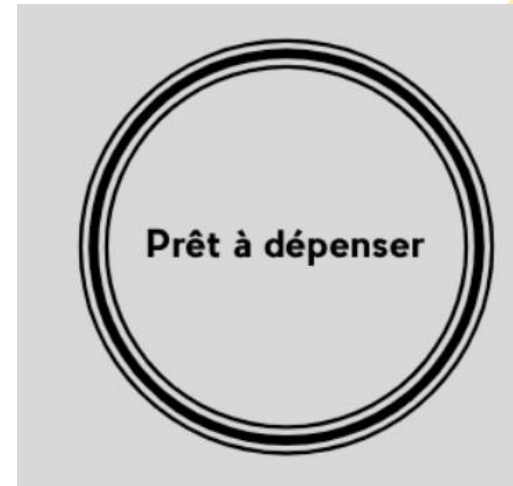


# Implémentez un modèle de scoring

Nathalie MAVEL  
Parcours Data Scientist – Projet 7

Soutenance : le vendredi 2022 à  
Evalueur :

# Rappel de la mission



- **Mettre en oeuvre un outil de “scoring credit”**

==> Calculer la probabilité qu’un client rembourse son crédit

==> Classifier les probabilités par un accord ou un refus

- **Mettre en œuvre un dashboard interactif** à destination des chargés de clientèle

Etude réalisée à partir des données fournies par la société : <https://www.kaggle.com/c/home-credit-default-risk/data>

- **Dossier git** : Code EDA + modélisation  
Code Dashboard (streamlit) et de son déploiement (Heroku)

- **Note méthodologique**



# **1. Présentation des données**

## **2. La mise en place de modélisations**

2.1 Preprocessing

2.2 Evaluation

2.3 Sélection du modèle

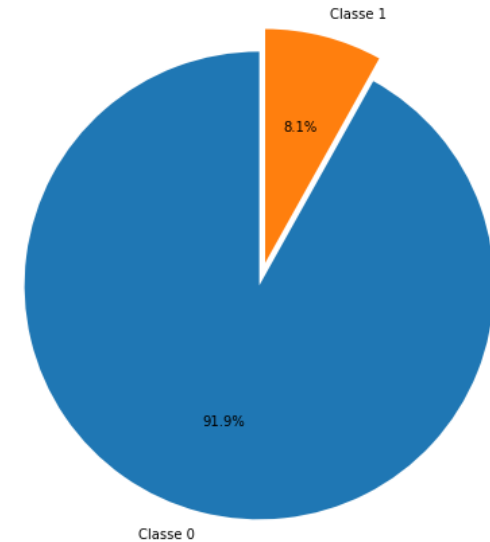
## **3. Présentation du dashbaord**

## **4. Conclusion**



# 1. Présentation des données

- 10 fichiers CSV : 2 retenus, application\_train : analyse exploratoire des données, mise en place de la modélisation  
La target est présente  
application\_test : utiliser pour la mise en place du dashboard  
La target est absente
- EDA : utilisations d'un notebook Kaggle (gold, le plus consulté)
  - Application\_train : 307511 individus et 122 colonnes
  - Id client unique
  - Cible très déséquilibrée
  - Transformations des variables catégorielles : label encoder et one hot encoder
  - Conversion de la variable Day\_birth en Age\_Client
  - Détection d'anomalies : Days employment : certains clients ont été employés pendant 1000 ans  
Ajout d'une variable anomalies days of employment



# 1. Présentation des données

➤ EDA : utilisations d'un notebook Kaggle (gold)

- **variable d'intérêt (présence de corrélation avec la cible)** : ext\_1, ext\_2, ext\_3, DAYS\_BIRTH

- **Features engennerring** :

Construction de 35 variables polynomiales en lien avec les variables d'intérêt.

Construction de 4 nouvelles variables métier :

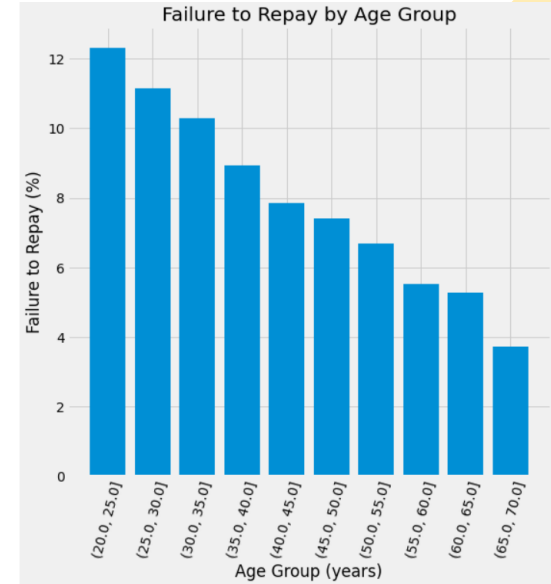
CREDIT\_INCOME\_PERCENT : le pourcentage du montant du crédit par rapport au revenu du client.

ANNUITY\_INCOME\_PERCENT : le pourcentage de l'annuité du prêt par rapport au revenu du client.

CREDIT\_TERM : la durée du paiement en mois (l'annuité étant le montant mensuel dû).

DAYS\_EMPLOYED\_PERCENT : le pourcentage des jours d'emploi par rapport à l'âge du client.

Les modifications ont été réalisées en parallèle sur le fichier application\_train.csv et application\_test.csv



## 2. Les modélisations : algorithmes de classification supervisées

- calculer la probabilité qu'un client rembourse son crédit
- classer les probabilité par un accord ou un refus

### 2.1 Pre processing

- Imputer les valeurs manquantes
- Transformer les boolens en intégrales,
- Centrer-réduire le jeu de données
- Diviser le jeu de données en 2 : train et test
- Équilibrer les données

Équilibrer les données ==> indispensable pour limiter le sur ou sous apprentissage du modèle

Stratégie d' **undersampling** retenue : forte importance du jeu de données qui permet de le diminuer,  
temps de calcul diminué

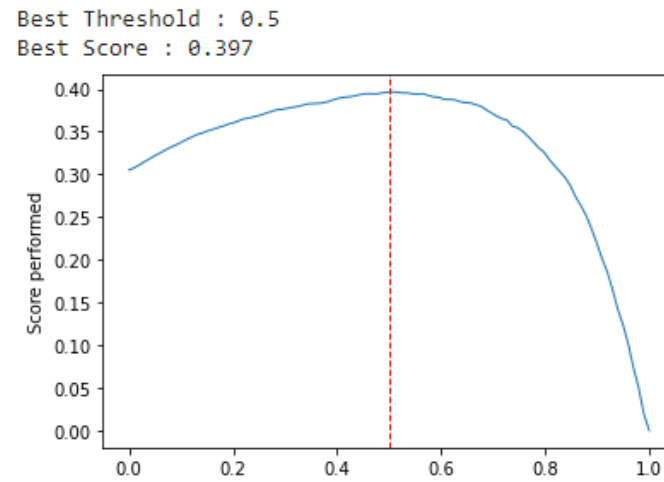
**==> 19860 individus dans chacune des 2 catégories.**

## 2. Les modélisations : algorithmes de classification supervisées

- calculer la probabilité qu'un client rembourse son crédit
- classer les probabilité par un accord ou un refus

### 2.1 modélisations

- 3 modèles testés : Regression logistique, le Random forest Classifier et le LitghGBM
- Grid search CV : sélection des meilleurs hyperparamètres en testant des paramètres aléatoirement (RandomizedSearchCV), Validé par une validation croisée basée sur le score Fbeta score=2
- Entrainement du modèle via `.predict_proba ==>` valeur de probabilité
- **définir un seuil** qui permet de classer les probabilités en 2 catégories (l'individu va rembourser, ou non)  
==> Sélection du score beta F2 maximum. En effet, plus le score F2 est fort, plus les faux négatifs (un faux bon client) sont faibles.



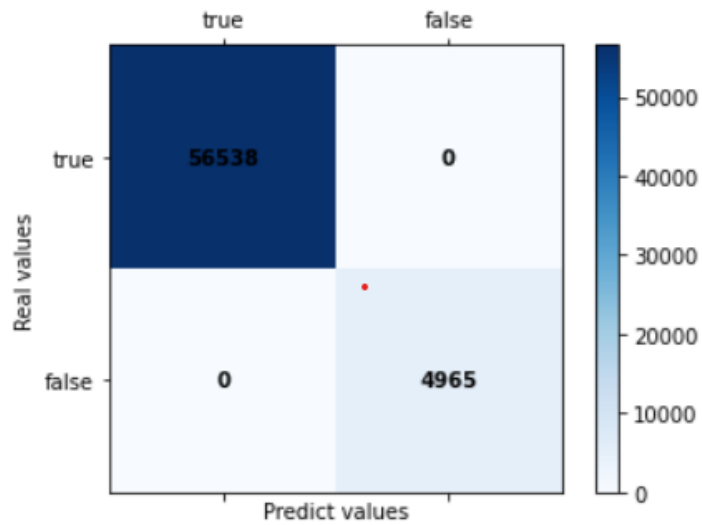
## 2. Les modélisations : algorithmes de classification supervisées

- calculer la probabilité qu'un client rembourse son crédit
- classer les probabilité par un accord ou un refus

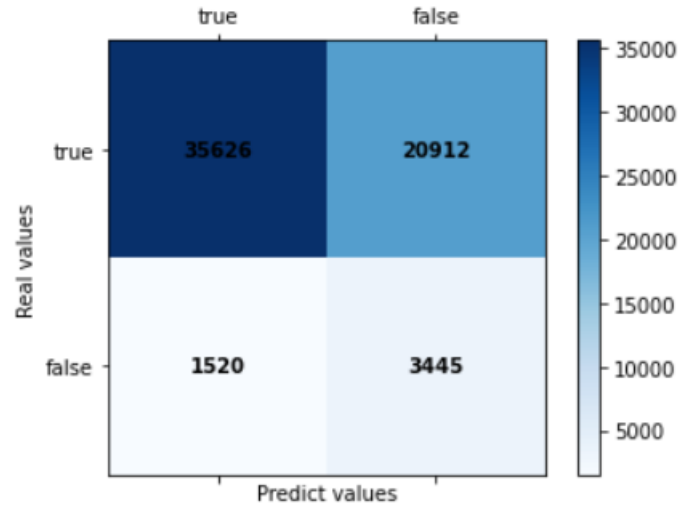
### 2.2 Evaluation des modèles

*Sur la partie non undersampler*

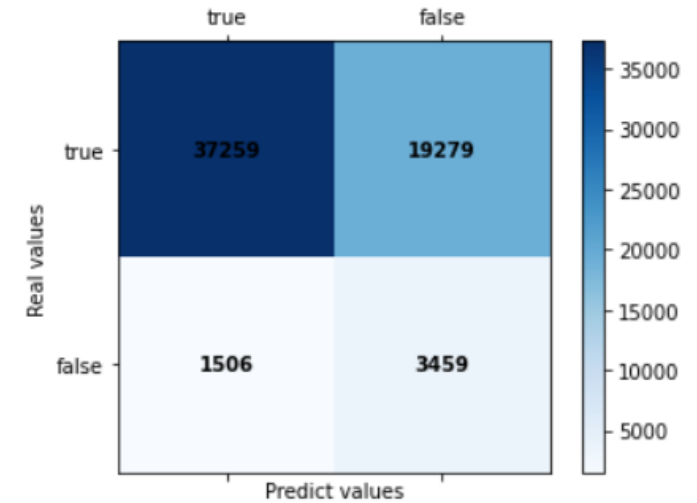
➤ Matrice de confusion :



Classification réelle



Logistic regression



Random forest

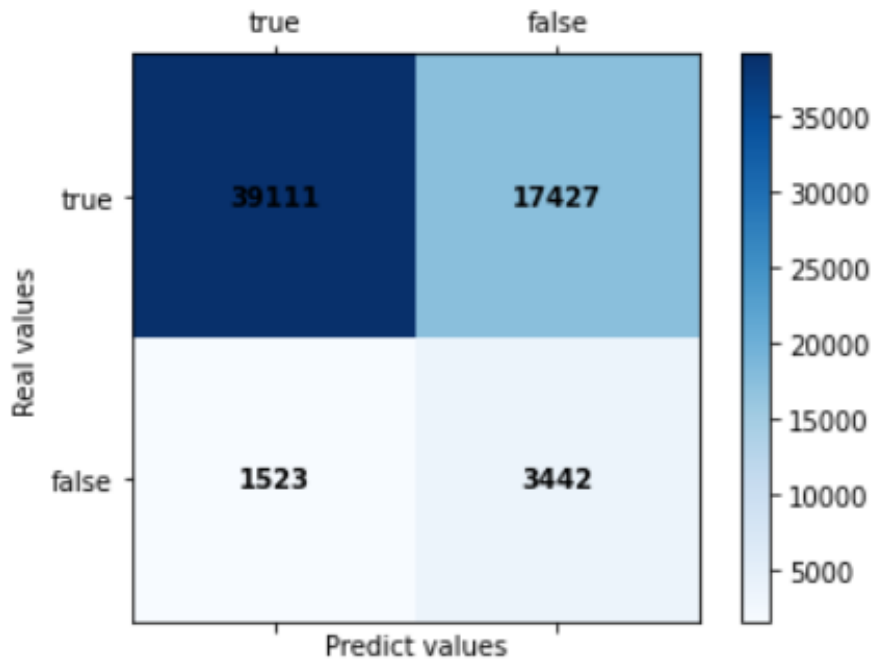


## 2. Les modélisations : algorithmes de classification supervisées

- calculer la probabilité qu'un client rembourse son crédit
- classer les probabilité par un accord ou un refus

### 2.2 Evaluation des modèles

➤ Matrice de confusion :



LGBM

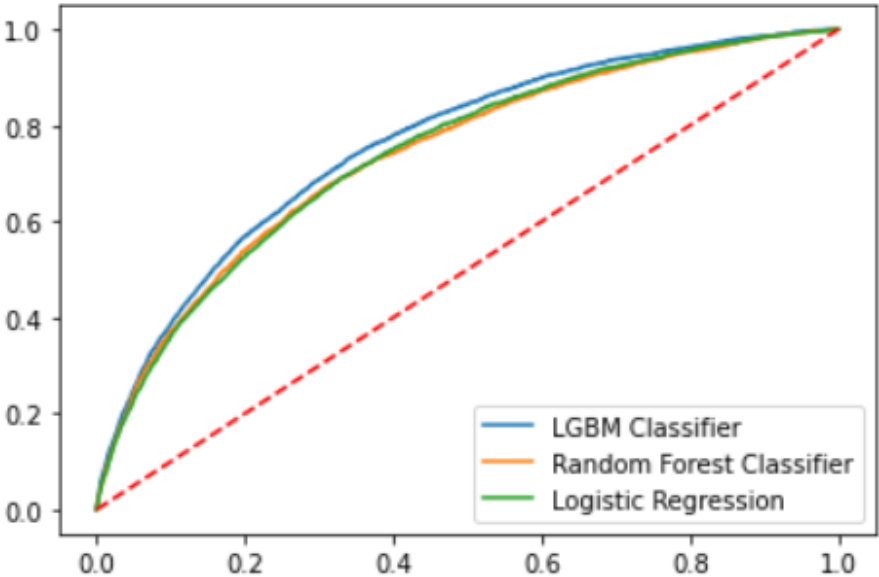
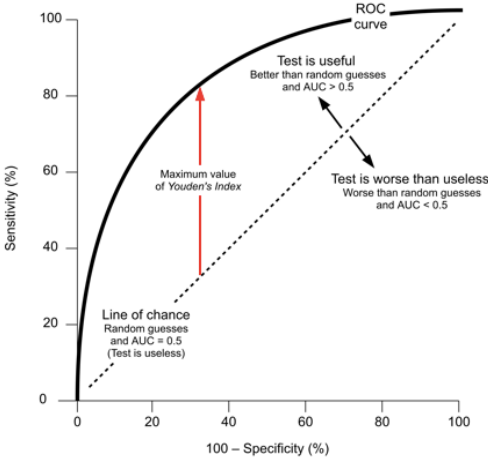
- le moins de faux négatifs et faux positifs

# 2. Les modélisations : algorithmes de classification supervisées

- Calculer la probabilité qu'un client rembourse son crédit
- Classifier les probabilités par un accord ou un refus

## 2.2 Evaluation des modèles

➤ Courbe ROC :



➤ Tableau des résultats

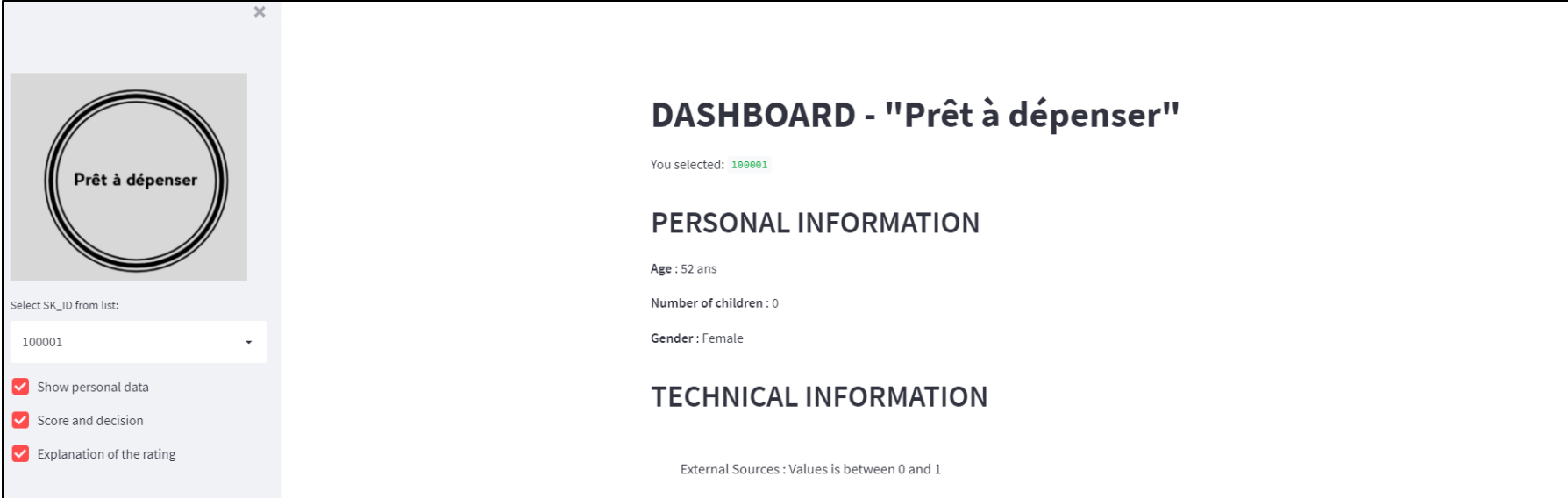
Avec ajustement : LGBM présente le meilleur F2\_score

Comportement similaire entre les modèles , accuracy forte, précision plus faible que le recall.

	accuracy	precision	recall	f1_score	f2_score	f3_score
LR	0.663	0.153	0.698	0.250	0.407	0.514
RF	0.657	0.148	0.685	0.244	0.397	0.503
LGBM	0.660	0.158	0.738	0.260	0.425	0.539

### 3. Présentation du dashboard

- Dashboard réalisé avec **Streamlit**, déployé sur **Heroku**
- MVP
- <https://pretadepenser-may2022.herokuapp.com/>



The screenshot shows a web application interface titled "DASHBOARD - 'Prêt à dépenser'". On the left, there is a sidebar with a circular logo containing the text "Prêt à dépenser". Below the logo, there is a dropdown menu labeled "Select SK\_ID from list:" with the value "100001" selected. Underneath the dropdown are three checkboxes, all of which are checked: "Show personal data", "Score and decision", and "Explanation of the rating". The main content area on the right displays the title "DASHBOARD - 'Prêt à dépenser'" followed by "You selected: 100001". Below this, there are two sections: "PERSONAL INFORMATION" which shows "Age : 52 ans", "Number of children : 0", and "Gender : Female"; and "TECHNICAL INFORMATION" which shows "External Sources : Values is between 0 and 1".

## 4. Conclusion et pistes

- ✓ **Mettre en oeuvre un outil de “scoring credit”**

==> calculer la probabilité qu'un client rembourse son crédit

==> classifiez les probabilités par un accord ou un refus

- ✓ **Mettre en oeuvre un dashboard interactif** à destination des chargés de clientèle

- Accorder plus de temps à l'analyse exploratoire
- Sélection des features à confirmer par des équipes métiers
- Détailler encore plus les hyperparamètres
- Classification multiple
- Limites éthique et juridique
- D'autres technologies : Utilisation de deep Learning

