



# Soutenance Projet 7

...

Pierre Schifflers



# Agenda

- Problématique
- Présentation du jeu de données
- Modélisation
- Outils
- Présentation Dashboard
- Aller plus loin



# Problématique

- L'entreprise "*Prêt à dépenser*" propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.
- Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.
- Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle.
- La transparence et l'interprétabilité sont essentiels.



# Jeu de Données

Données du Home Credit Group disponibles sur Kaggle:

1. **Table principale** → Table avec une ligne par prêt + infos sur le prêt et le client.
2. **Tables historiques** → Données sur les prêts précédents des clients et leurs habitudes de remboursement. Pas utilisées dans ce projet par souci de simplicité.
3. **Description des features** → Utilisé lors de l'interprétation des prédictions

Table principale: 307,000 observations sur 122 colonnes



# Modélisation

## Feature Engineering

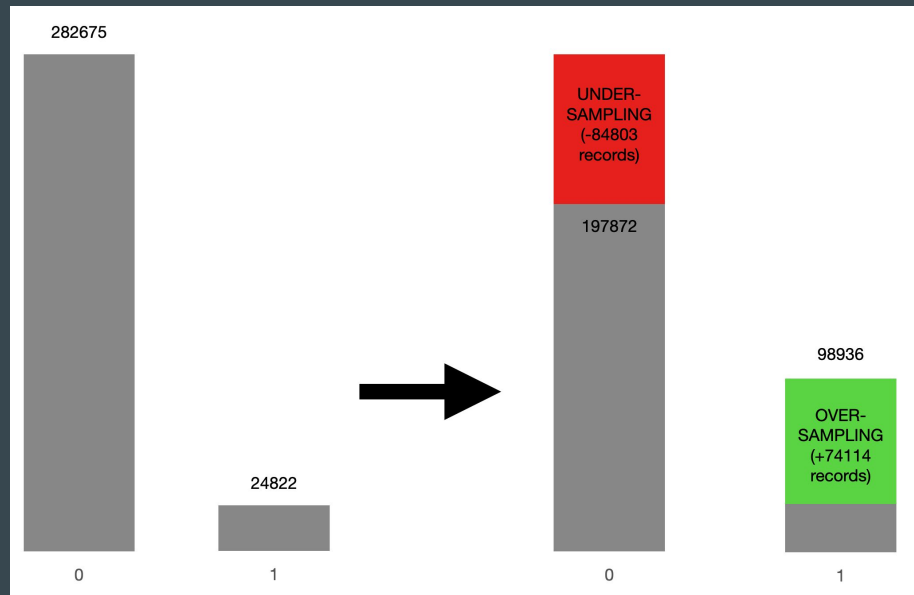
- Basé sur [kernel Kaggle](#)
- Label Encoding, One-Hot Encoding, Scaling, création de features spécifiques au domaine financier
- Nettoyage des données (valeurs aberrantes, imputation par la médiane)



# Modélisation

## SMOTE

- Données très inégales: **92%** des données dans la catégorie 0.
- Utilisation de Synthetic Minority Oversampling Technique pour résoudre ce problème.
- Over-sampling: 30% → Under-sampling 50%.
- Distribution finale: 67% des données dans la catégorie 0.



# Modélisation

## Définition de la métrique

- Faux Positif = Perte de revenus potentiels
- Faux Négatif = Coûts additionnels

**Hypothèse:** Les faux négatifs ont un impact plus important sur l'entreprise. Ils doivent donc avoir plus de poids dans l'évaluation du modèle.

**Métrique:** F-Beta où  $\text{Beta} = 2$ . Le recall a 2x plus de poids que la précision.

		Classe Réelle	
		Non-Défaut	Défaut
Classe Prédite	Non-Défaut	Vrai Négatif	Faux Négatif
	Défaut	Faux Positif	Vrai Positif



# Modélisation

## Evaluation des modèles

- Séparation en train/validation de 70/30%
- Tuning des hyperparamètres avec cross-validation 4 folds + SMOTE dans le pipeline
- Test des performances (score F-Beta) sur jeu de validation

	model	score	train_score	test_score
0	Logistic Regression - No Smote	F-Beta	0.013918	0.015908
1	Logistic Regression	F-Beta	0.352568	0.354126
2	XG Boost	F-Beta	0.190002	0.117907
3	Light Gradient Boosting	F-Beta	0.282701	0.261707
4	Random Forest Classifier	F-Beta	0.989360	0.040362

**Conclusion:** SMOTE permet de gagner énormément en performance et la **régression logistique** est le modèle le plus performant.

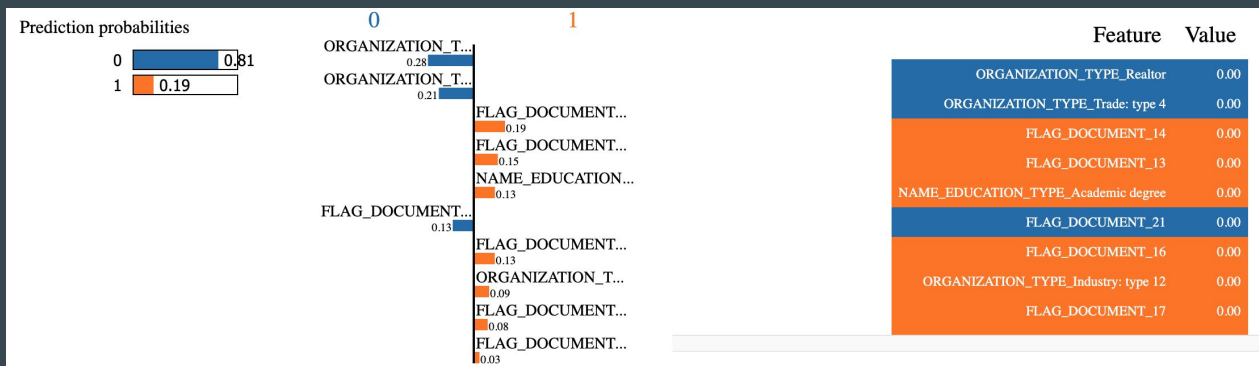


# Modélisation

## Interprétabilité des modèles



Utilisation de LIME (Local interpretable model-agnostic explanations):

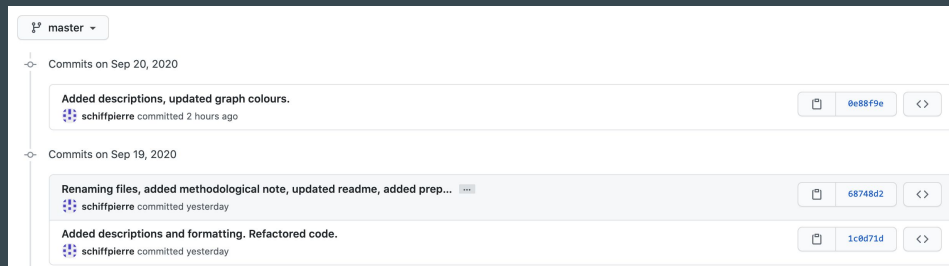
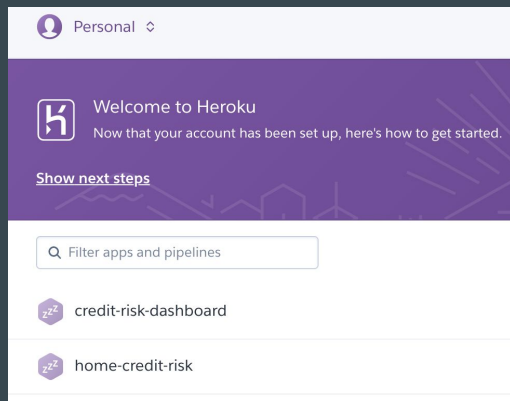
- Explication de prédiction pour une observation unique
- Bien approprié pour notre cas
- Utile pour un modèle de régression logistique où nous n'avons pas de 'feature importance'



# Outils



Fonction	Outil
Plate-forme de déploiement API + Dashboard	 <b>HEROKU</b>
Dashboard	 <b>Streamlit</b>
Versionage	<b>GitHub</b>



# Dashboard

<https://credit-risk-dashboard.herokuapp.com/>

## Credit Risk Prediction Dashboard

A dashboard to understand the factors influencing credit risk predictions

- The **Global** dashboard provides general information about loans at the Home Credit Group.
- The **Client-specific** dashboard lets you get a credit risk prediction for a specific client and provides insights into the prediction.

Please select a dashboard below:

Client-Specific Dashboard ▼

Examples of client IDs:

122913, 420515, 136802, 394032, 155090

Client ID:

Please enter a Client ID.

*Dashboard made by Pierre Schiffers as part of the Data Science track on OpenClassrooms.*

# Aller plus loin

- Tuning hyperparamètres + feature engineering
- Interprétation des features à l'échelle
- Description des features en one-hot encoding
- Hosting du dashboard sur une plate-forme permettant plus de rapidité
- Filtres pour sélectionner un groupe de clients particulier