

# P7: Implémentez un modèle de scoring

OpenClassrooms  
Formation Data Scientist  
20 Juillet 2021 - 18 Mai 2022

Romain Vaillant  
Avril 2022

# P7: Implémentez un modèle de scoring

1

## Sommaire

- 1 Objectifs
- 2 Jeu de données
- 3 Analyse exploratoire
- 4 Nettoyage du jeu de données
- 5 Approche de modélisation
- 6 Présentation du dashboard
- 7 Conclusion

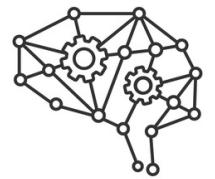
# 1. Objectifs



## Contexte

Data Scientist au sein d'une société financière, nommée "Prêt à dépenser", qui propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt.

---



## Modèle

Construire un modèle de “scoring” qui donnera une prédition sur la probabilité de remboursement du crédit d'un client de façon automatique.

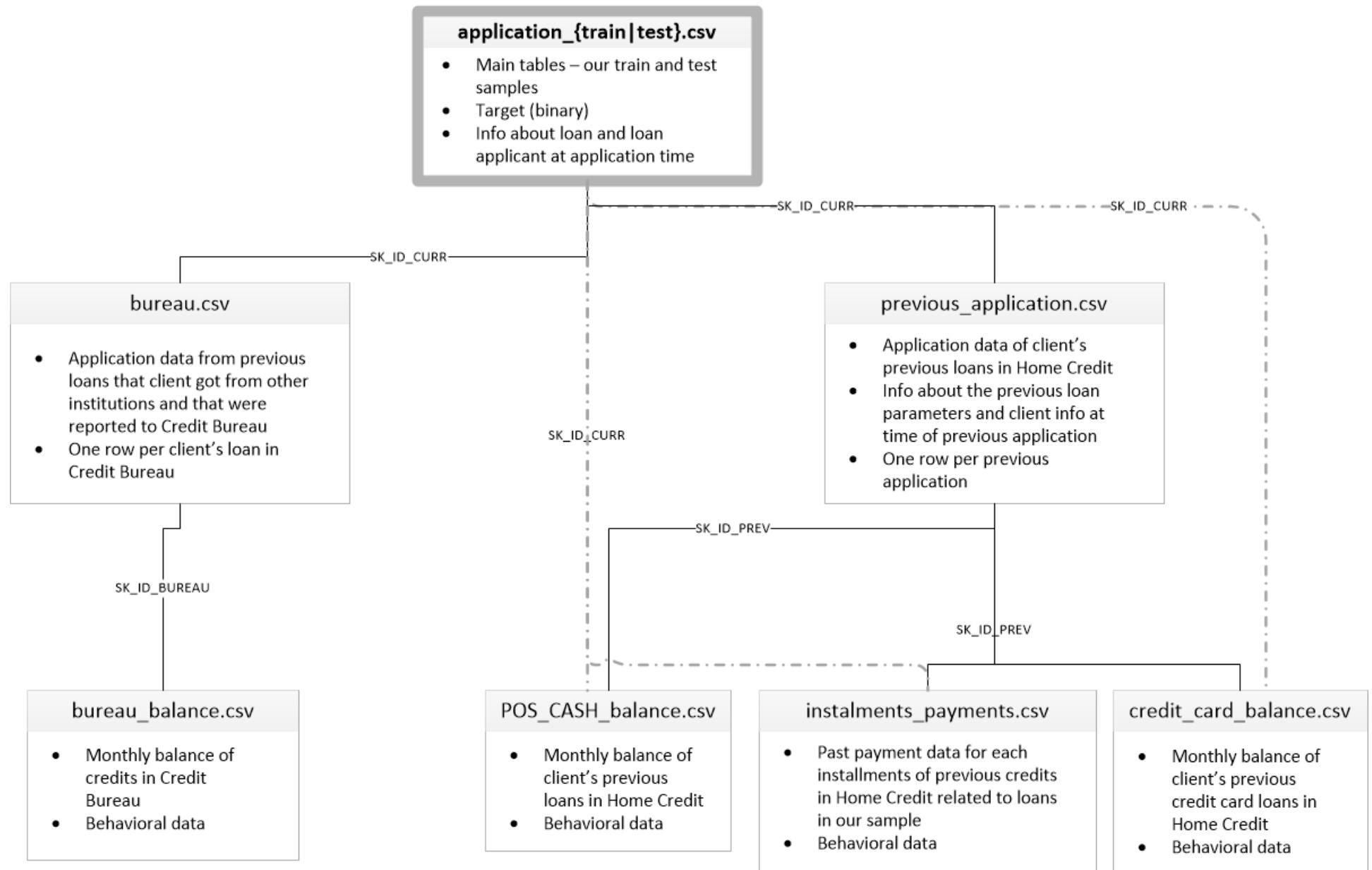
---



## Dashboard

Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle,

# 2. Présentation du jeu de données



## Calcul des agrégats

Kernel Kaggle  
Cf. note méthodologique

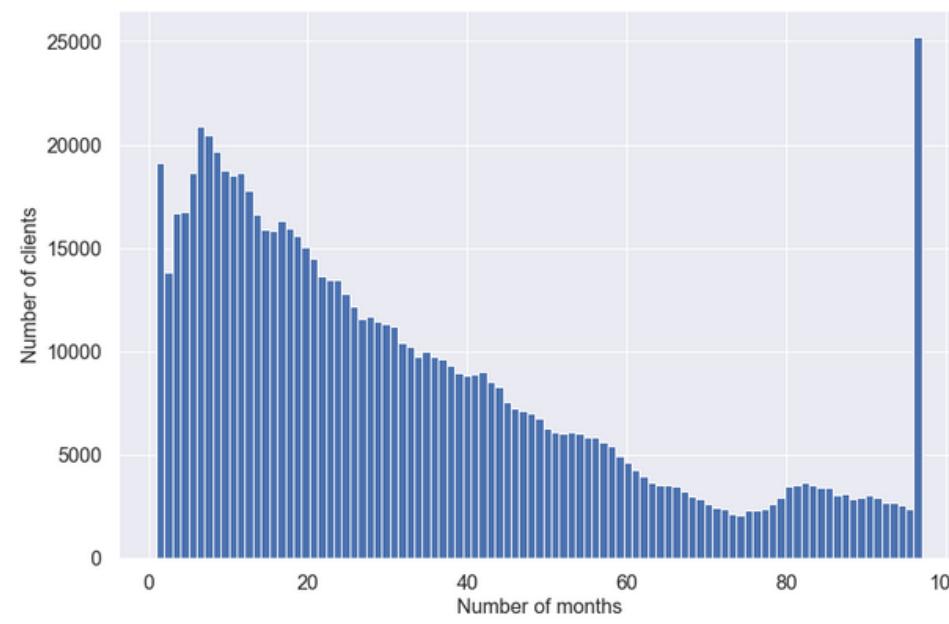
## Merge

### df\_merge

- 433 colonnes
- 307 507 lignes
- Taux de valeurs manquantes: 54.4 %

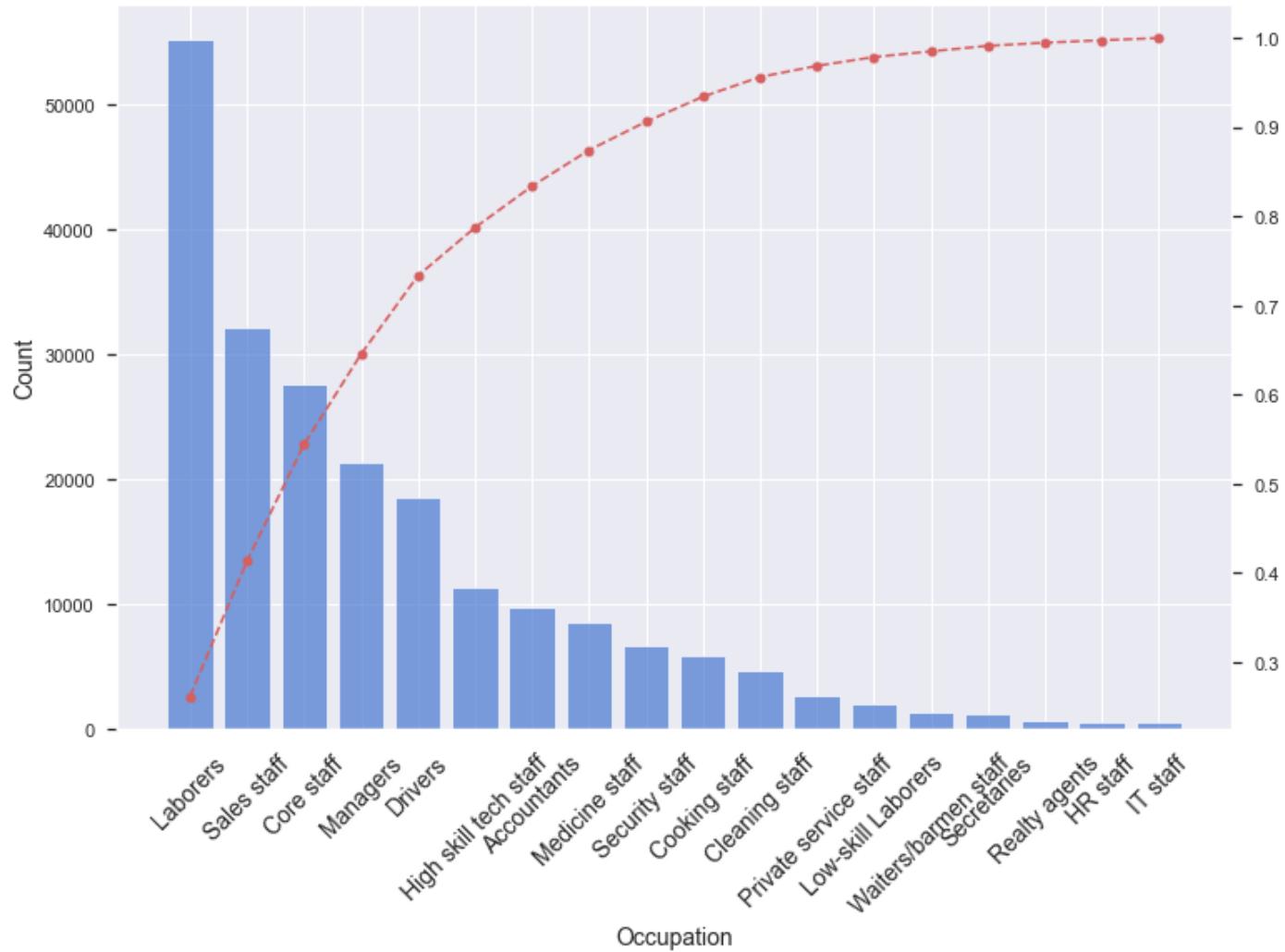
# 3. Analyse exploratoire

Distribution de la durée des prêts en mois



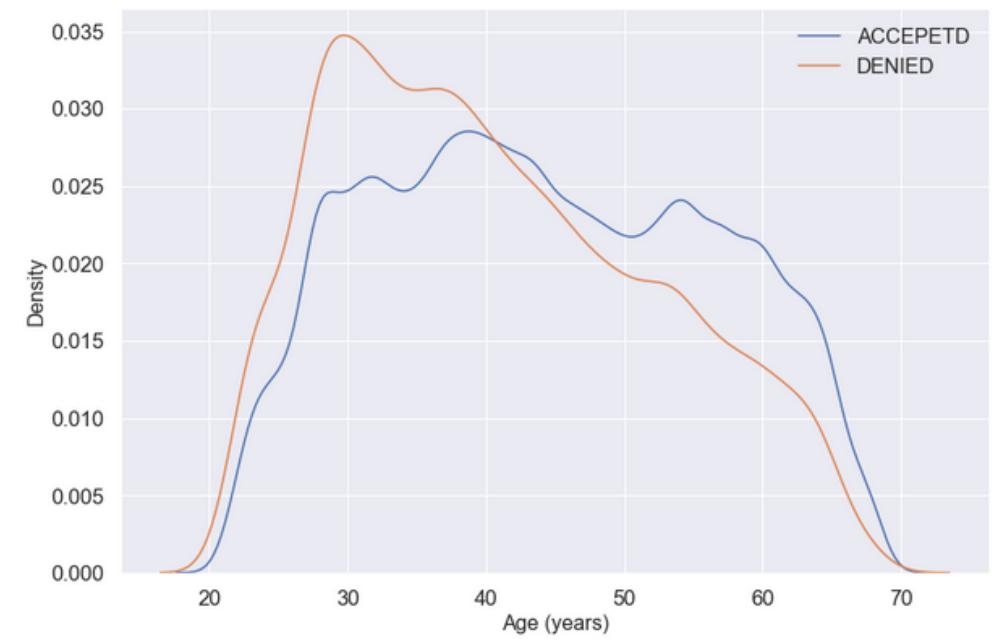
- Des prêts à court terme

Nombre de prêts par profession



- Un pool de métiers restreint

Distribution des âges par décision de prêt

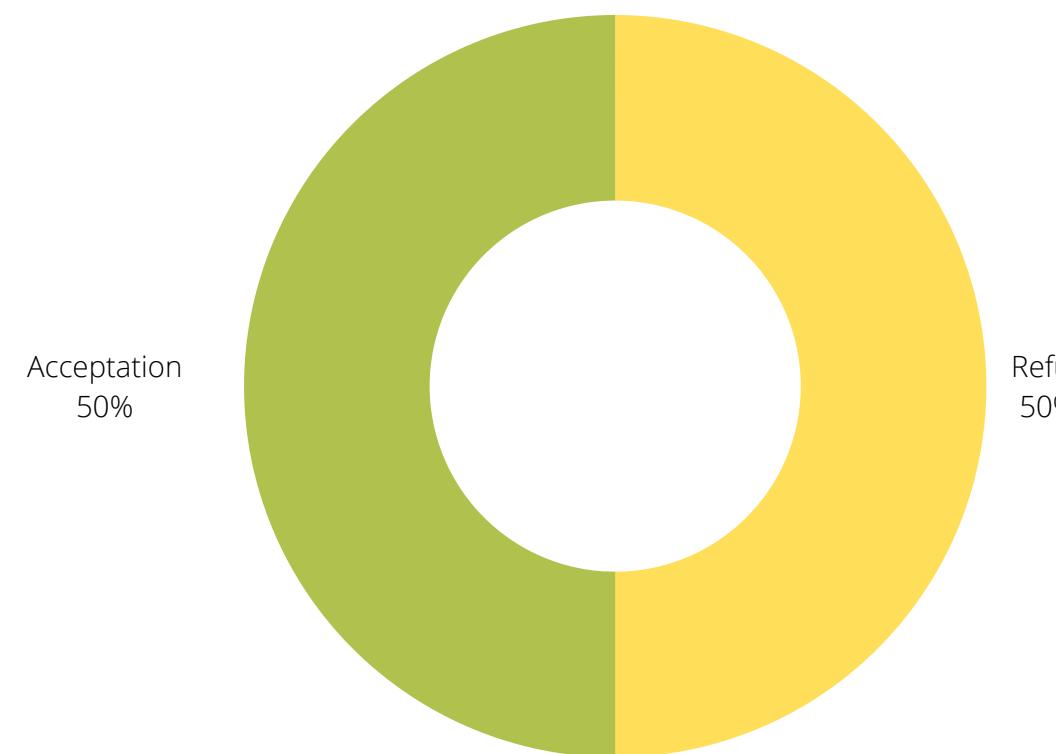
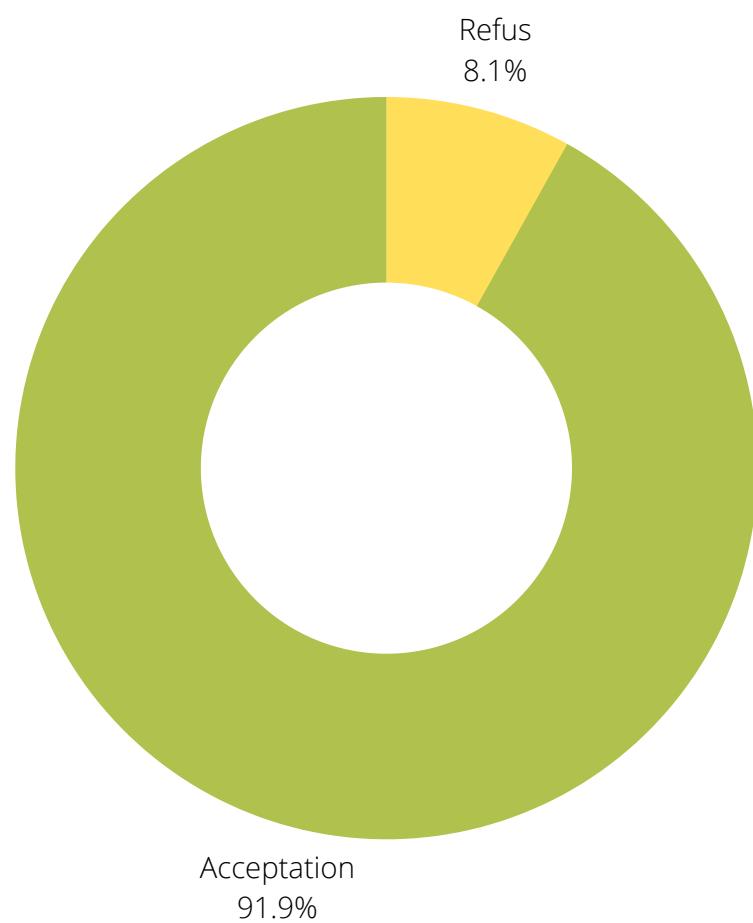


- L'âge, un facteur de décision

# Distribution de l'acceptation des prêts



Variable cible: TARGET



Imblearn

- 
- Oversampling
    - SMOTE() (0.2)
  - Undersampling
    - RandomUnderSampler() (1)

# 4. Nettoyage des données



## FILTRAGE 1

Suppression des colonnes  
peuplées à moins de 80%

**433**



**160 Colonnes**

## FILTRAGE 2

Suppression des lignes  
incomplètes

**307 507**



**157 129 Lignes**

## FILTRAGE OPT.

Sélection restreinte de variables  
avec un sens métier fort

**160**



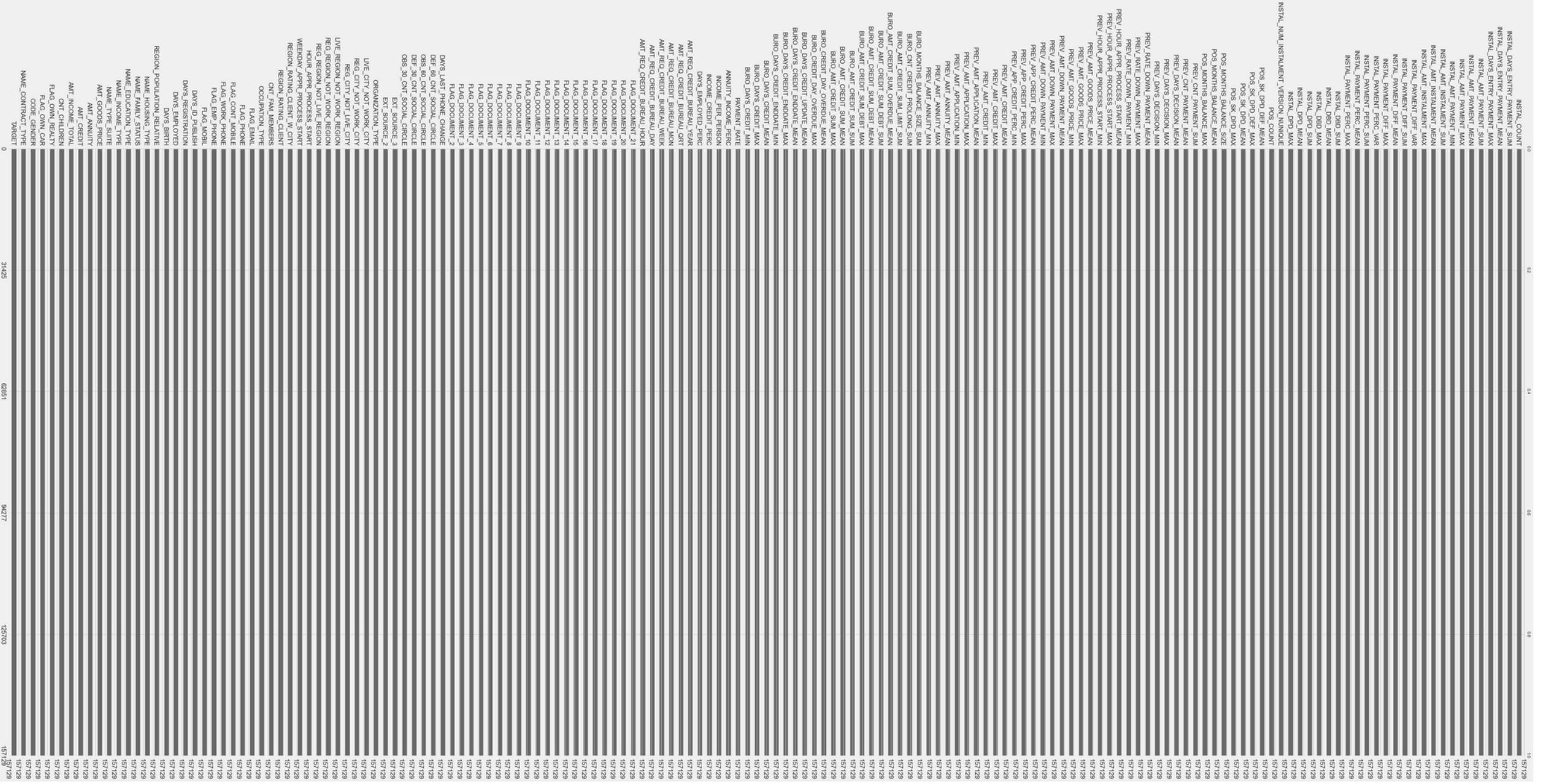
**25 Colonnes**

## CASTING

Conversion des variables  
numériques non calculable en (str)  
(ex: variables binaires, dates,...)

-

# Dataset final : df\_merge



NETTOYAGE

433  
307 507  
54.4 %



159 colonnes  
157129 lignes  
NaN : 0.0 %

## 5. Approche de modélisation



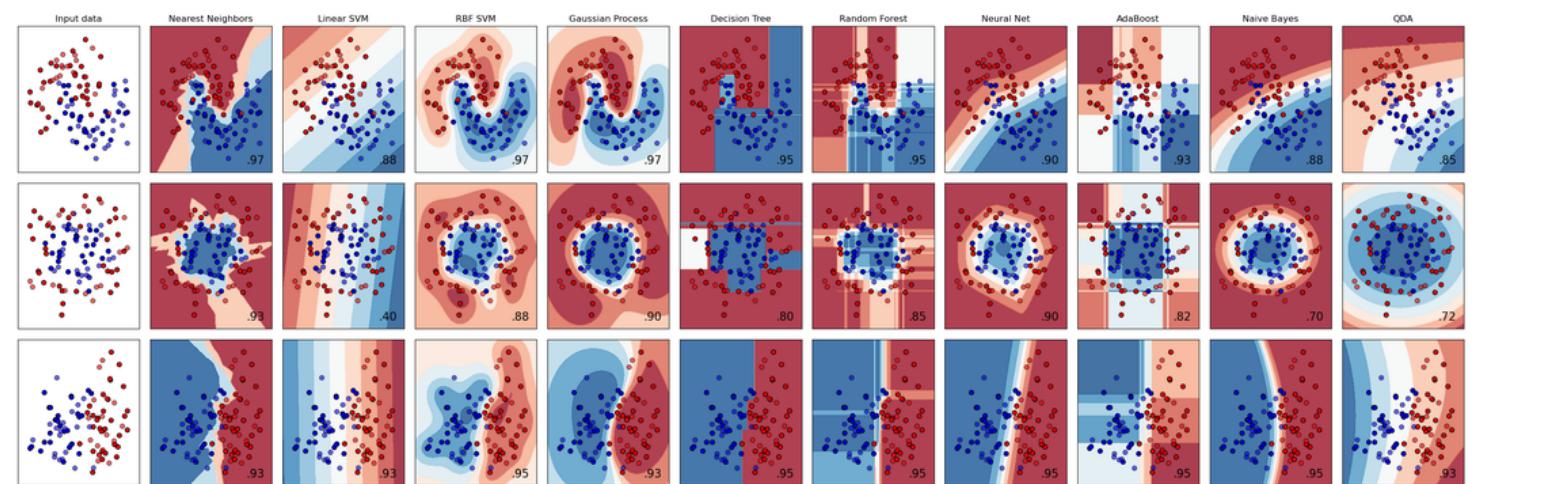
**Selection des  
algorithmes de  
classification**

**Feature  
Engineering**

**Choix d'une  
métrique  
personnalisée**

**Tuning (Hyperopt)**

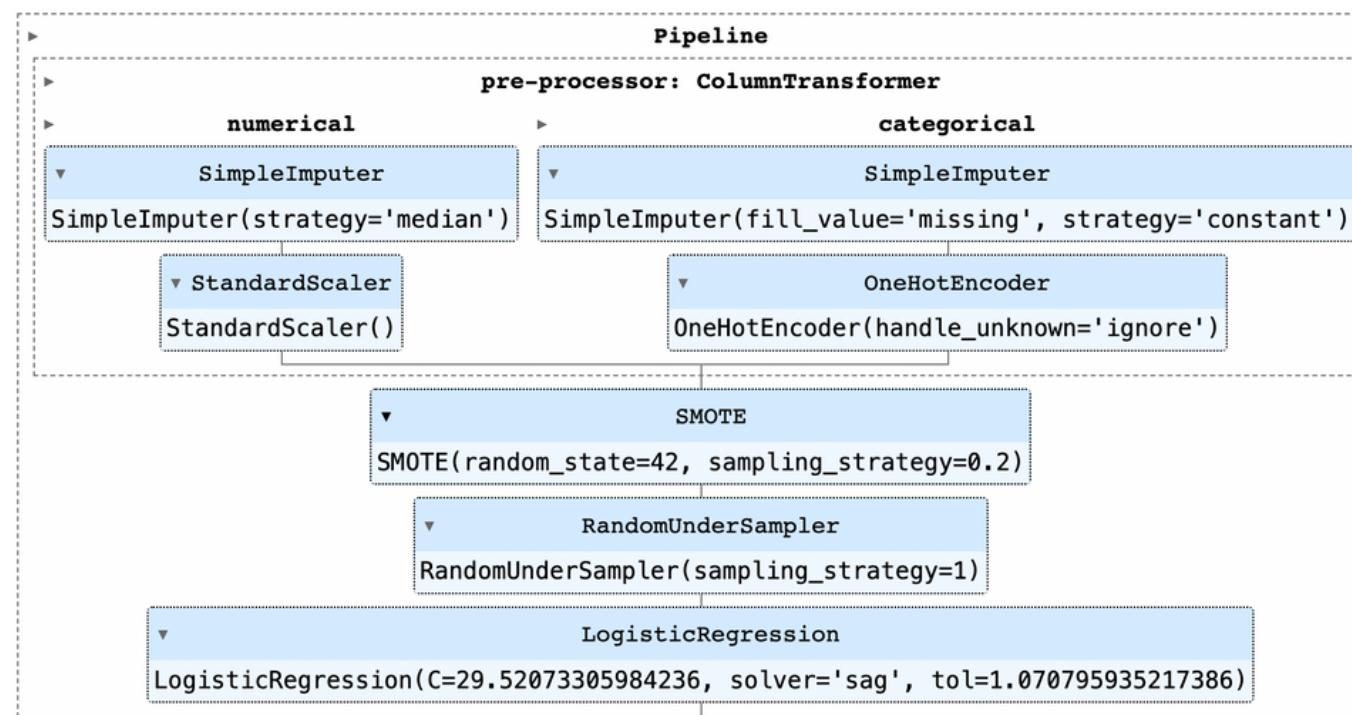
# Selection des algorithmes de classification



Exemples d'algorithmes de classification

- Dummy
- Logistic Regression
- RandomForest
- XGBoost

# Feature Engineering



Exemple de Pipeline LogisticRegression)

- StandardScaler
- RobustScaler
- MinMaxScaler

# Choix d'une métrique personnalisée

## FBETA\_SCORE

### Classification binaire

		Classe réelle	
		-	+
Classe prédictive	-	True Negatives <i>(vrais négatifs)</i>	False Negatives <i>(faux négatifs)</i>
	+	False Positives <i>(faux positifs)</i>	True Positives <i>(vrais positifs)</i>

Matrice de confusion

- PRECISION:
  - $TP/(TP+FN)$
  - Minimiser les faux positifs
- RECALL:
  - $TP/(TP+FN)$
  - Minimiser les faux négatifs
- FBETA SCORE:
  - $((1 + \beta^2) * \text{PRECISION} * \text{RECALL}) / (\beta^2 * \text{PRECISION} + \text{RECALL})$
  - Moyenne harmonique pondérée

# Choix d'une métrique personnalisée

## FBETA\_SCORE

### Discussion

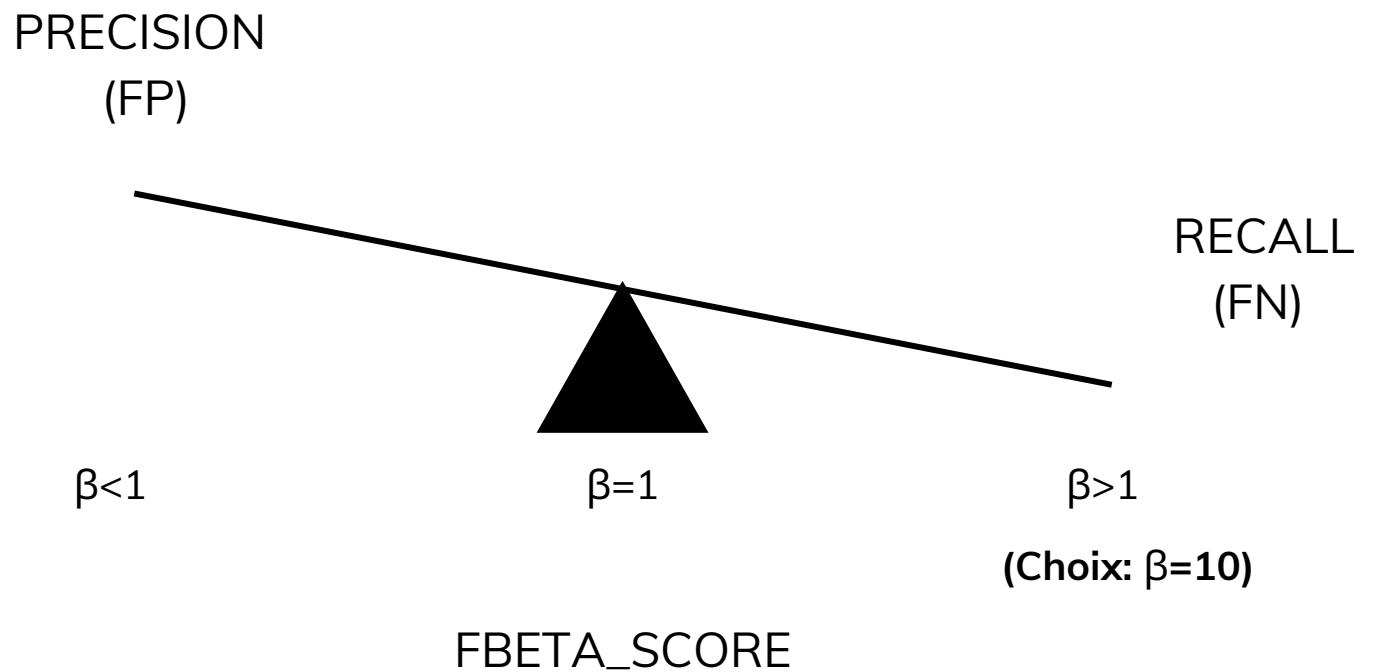
La banque peut commettre 2 erreurs:

- Refuser un prêt dû
  - Perdre le client (FP)
  - Estimation des Intérêts cumulés/composés d'un prêt:  
10% de la valeur du prêt
- Accorder un prêt indû
  - Perdre la somme prêtée (FN)

En résumé:

Perte de 10% (perte d'un client) VS Perte maximum de 100%

### Pondération



# Performance des algorithmes de classification

## FBETA\_SCORE

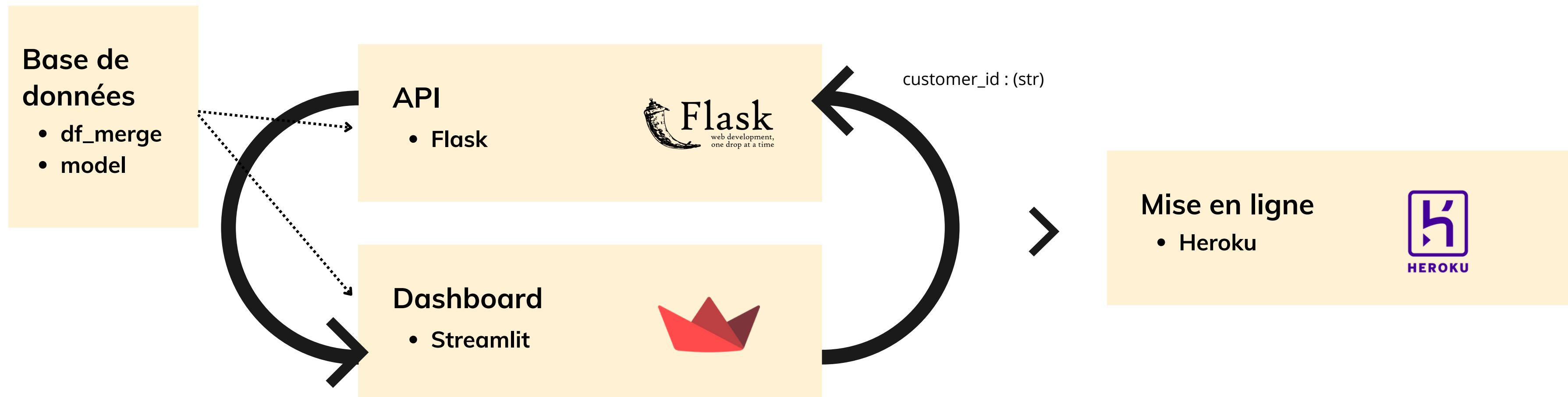
Tuning - Hyperopt



# Présentation du Dashboard

# 6. Présentation du dashboard

## Architecture de l'application



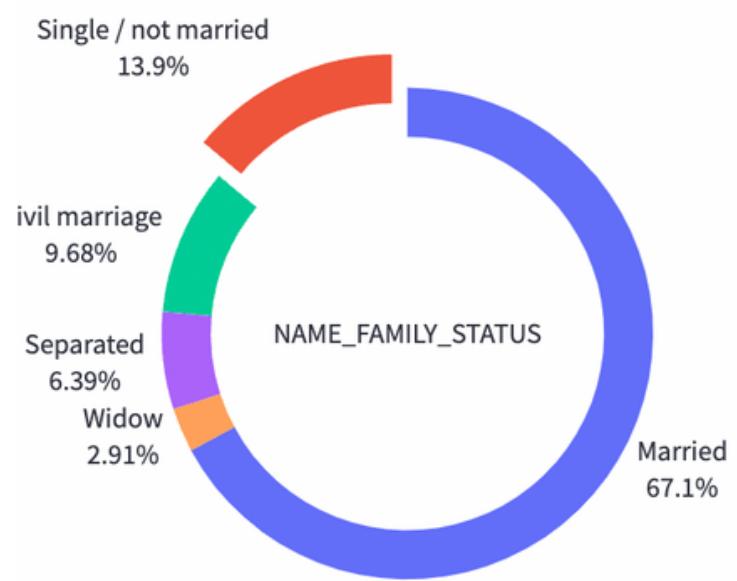
```
{  
    customer_id : (str),  
    prediction: (int),  
    prediction_probability: (float)  
}
```

# Présentation du dashboard

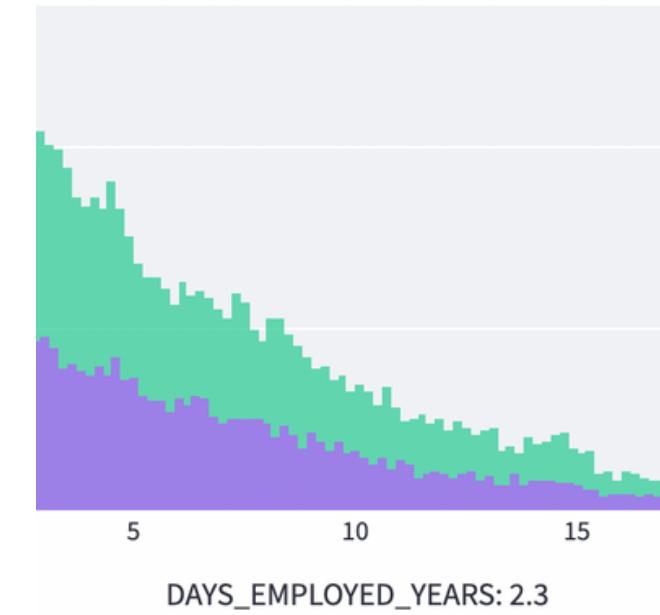
## Aperçu



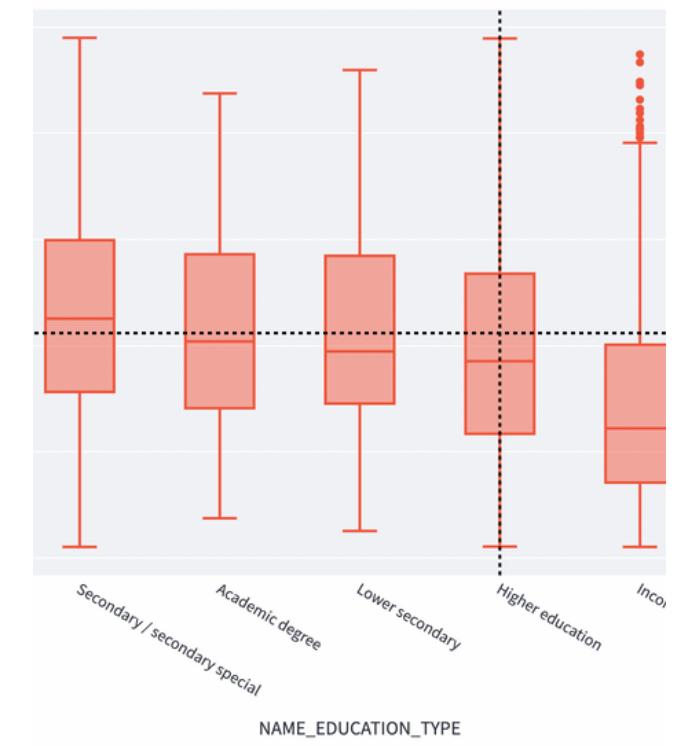
Decision



Pie Plot



Hist Plot



Box Plot

> Dashboard

# 7. Conclusion

- Etudier d'autres algorithmes de classification.
- Données déséquilibrées. Faible % de crédits refusés.
- Custom metric à revoir avec une équipe métier pour affiner le coefficient de pondération.
- LogisticRegression meilleur algorithme.
- Avoir plus de données et mieux réparties.
- Choix plus fin des variables. Calcul de nouveaux agrégats.



# **Merci de votre écoute**



**Séance de questions**

# Performance des algorithmes de classification (tuning)

	ROC_AUC_SCORE	F1_SCORE	ACCURACY_SCORE	RECALL_SCORE	Precision_Score	FBETA_SCORE
Dummy	0.5	0.5	<u>0.92</u>	0	0	<u>0</u>
LogisticRegression	0.60	0.19	<u>0.55</u>	<u>0.65</u>	0.11	<u>0.62</u>
RandomForest	0.65	0.26	0.75	0.54	0.17	0.53
XGBoost	0.58	0.19	0.71	0.42	0.12	0.41

ANNEXE