

Implémentation d'un modèle de scoring

Probabilité de défaut de paiement du client



Prêt à dépenser

Prêt à dépenser



Problématique

Solutions

Offre
Crédits à la consommation

Clientèle
Ayant peu ou pas d'historique de prêt

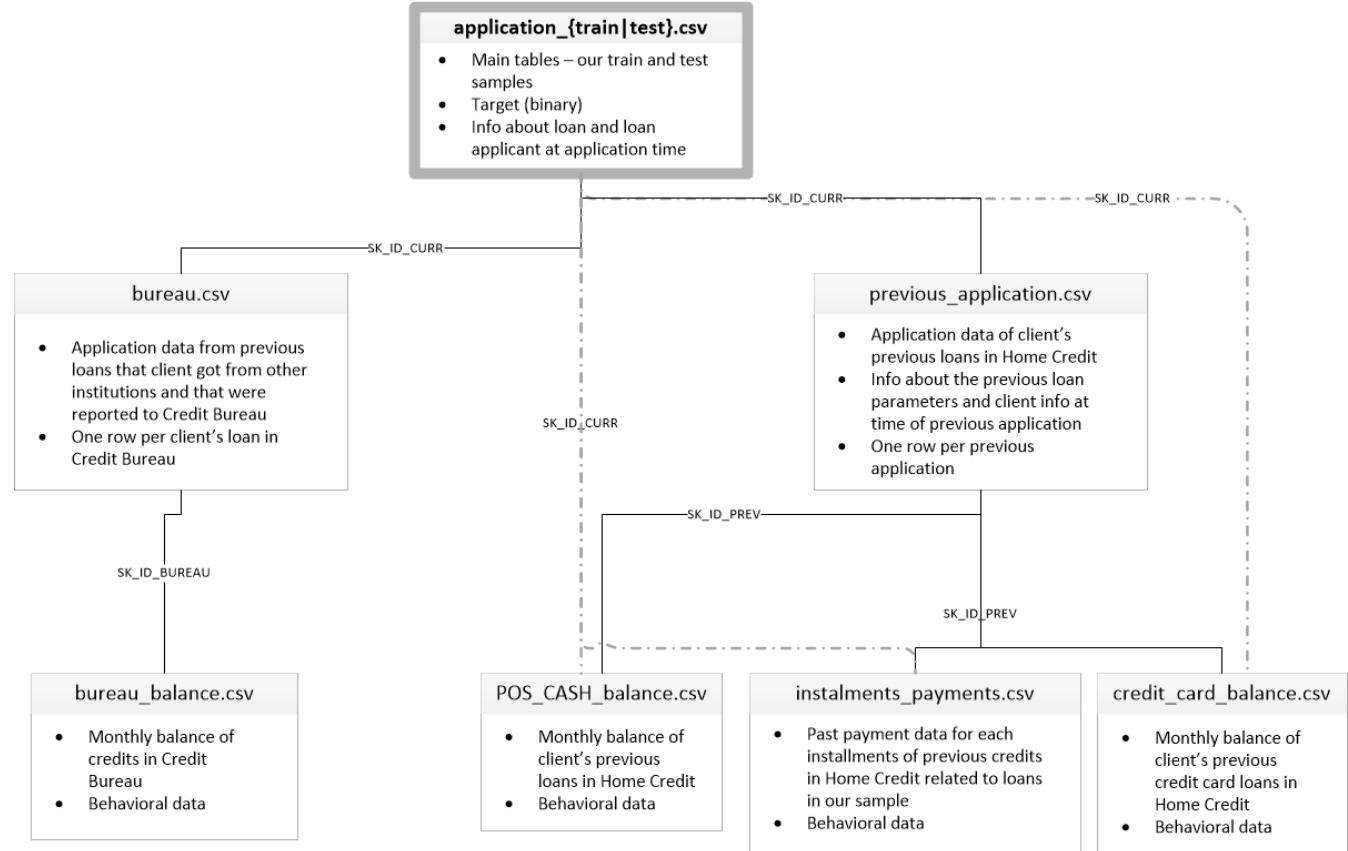
Estimer la probabilité de défaut de paiement d'un client

Transparence

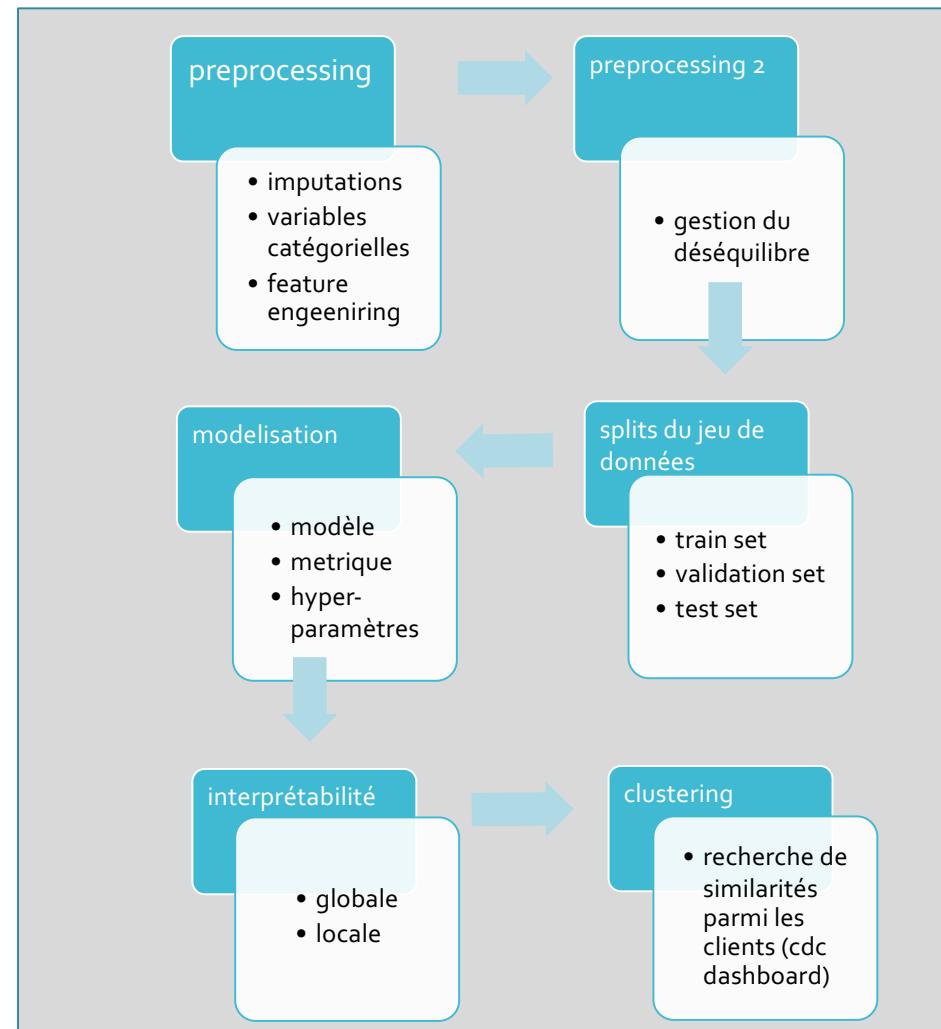
Implémenter un modèle de scoring

Dashboard interactif

Données



Stratégie



Preprocessing

Kernel Kaggle*



Valeurs manquantes

Imputation (médiane)

Variables catégorielles

Label encoding (2 catégories)
One Hot encoding (+ de 2 catégories)

Feature engineering

Création de variables composites
% du montant du crédit/revenu
% des jours employés/âge...

*Will Koehrsen - Boston, Massachussets

<https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction>

Preprocessing



Inégale répartition des classes
-> risque de prédiction systématique de la classe 0

Mise en œuvre de méthodes de gestion du déséquilibre

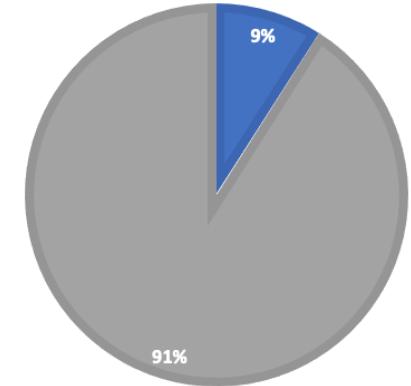
Undersampling

Oversampling

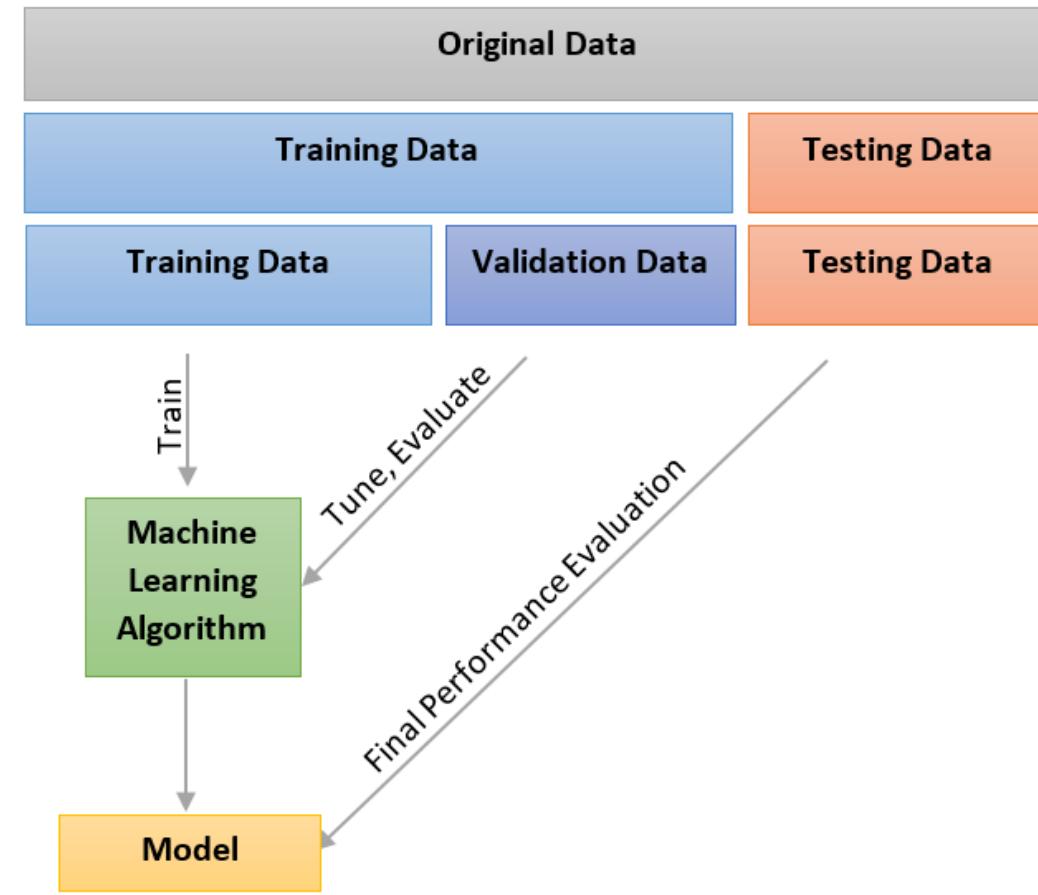
Sample Weights
spécifie un poids différent selon le montant du crédit

Scale-pos-weight
redimensionne les erreurs faites sur la classe minoritaire

■ classe 1 ■ classe 0



Preprocessing



Modèle & Métriques

Modèle implémenté: XGBoost Classifier

Hyper-paramètres ajustés: n_estimators, learning rate, max_depth

Métrique

	Prédits sans défaut (0)	Prédits en défaut (1)
Réellement sans défaut (0)	Vrais négatifs (TN)	Faux positifs (FP)
Réellement en défaut (1)	Faux négatifs (FN)	Vrais positifs (TP)

A MINIMISER

A MAXIMISER

$$Rappel = \frac{TP}{TP+FN} = \frac{\text{nb prédites en défaut et en défaut}}{\text{nb réellement en défaut}}$$

RISQUE

augmentation significative des faux positifs

A MAXIMISER

$$\text{Précision} = \frac{TP}{TP+FP} = \frac{\text{nb prédites en défaut et en défaut}}{\text{nb prédites en défaut}}$$

FINALEMENT
MAXIMISER

$$F1 score = \frac{2*P*R}{P+R}$$

Fonction coût métier

	Prédits sans défaut (0)	Prédits en défaut (1)
Réellement sans défaut (0)	+ 30% Intérêts	- 30% Coût d'opportunité
Réellement en défaut (1)	- 80% Défaut de paiement	0

$$Gain = 0.3 * TN - 0.3 * FP - 0.8 * FN$$

Crédit accordé et remboursé
Gain: 30%

Crédit accordé partiellement remboursé
Coût: 80%

Crédit refusé mais remboursable
Coût opp.: 30%

Scores

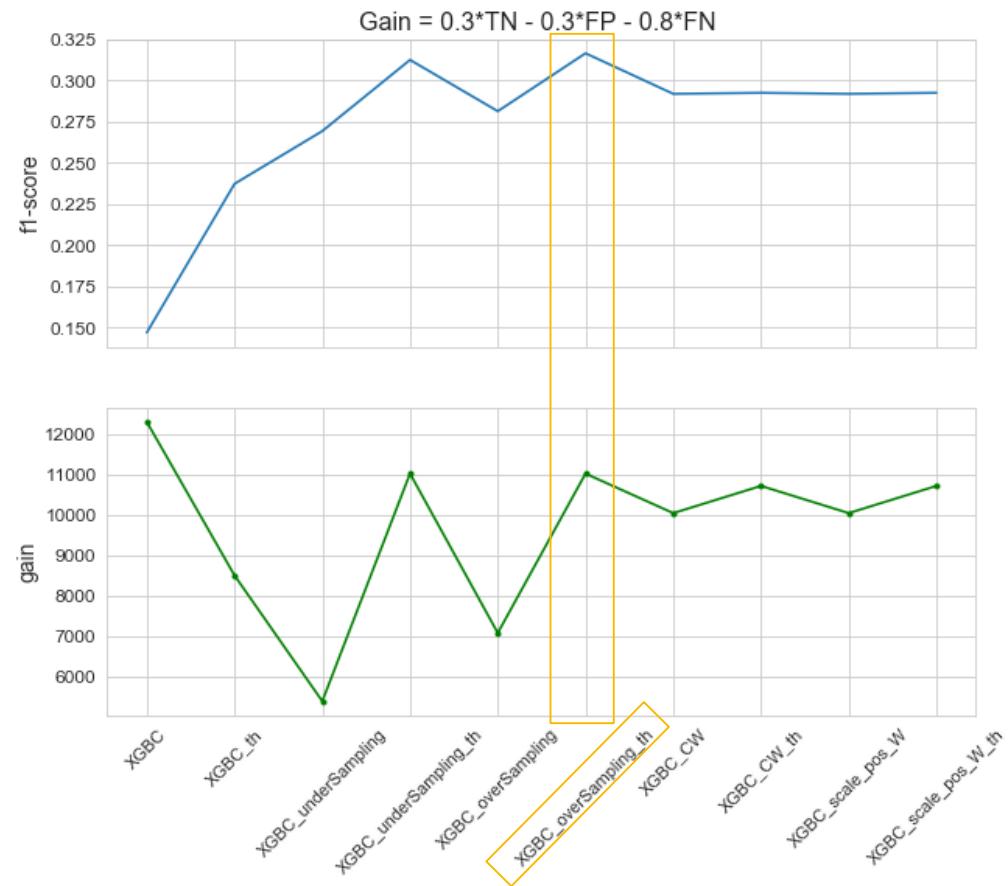
Méthode de gestion du déséquilibre par oversampling

Ajustement du seuil de probabilité pour prédire les classes

Prise en compte du gain métier

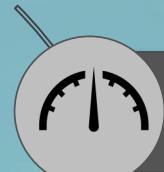


f1-scores et gains selon la gestion du déséquilibre et l'ajustement du seuil de probabilité



Dashboard

- Cahier des charges:



Visualiser le score et son interprétation de façon intelligible pour une personne non experte



Visualiser les informations descriptives relatives à un client (filtre)



Comparer ces informations descriptives à l'ensemble des clients et/ou à un groupe similaire

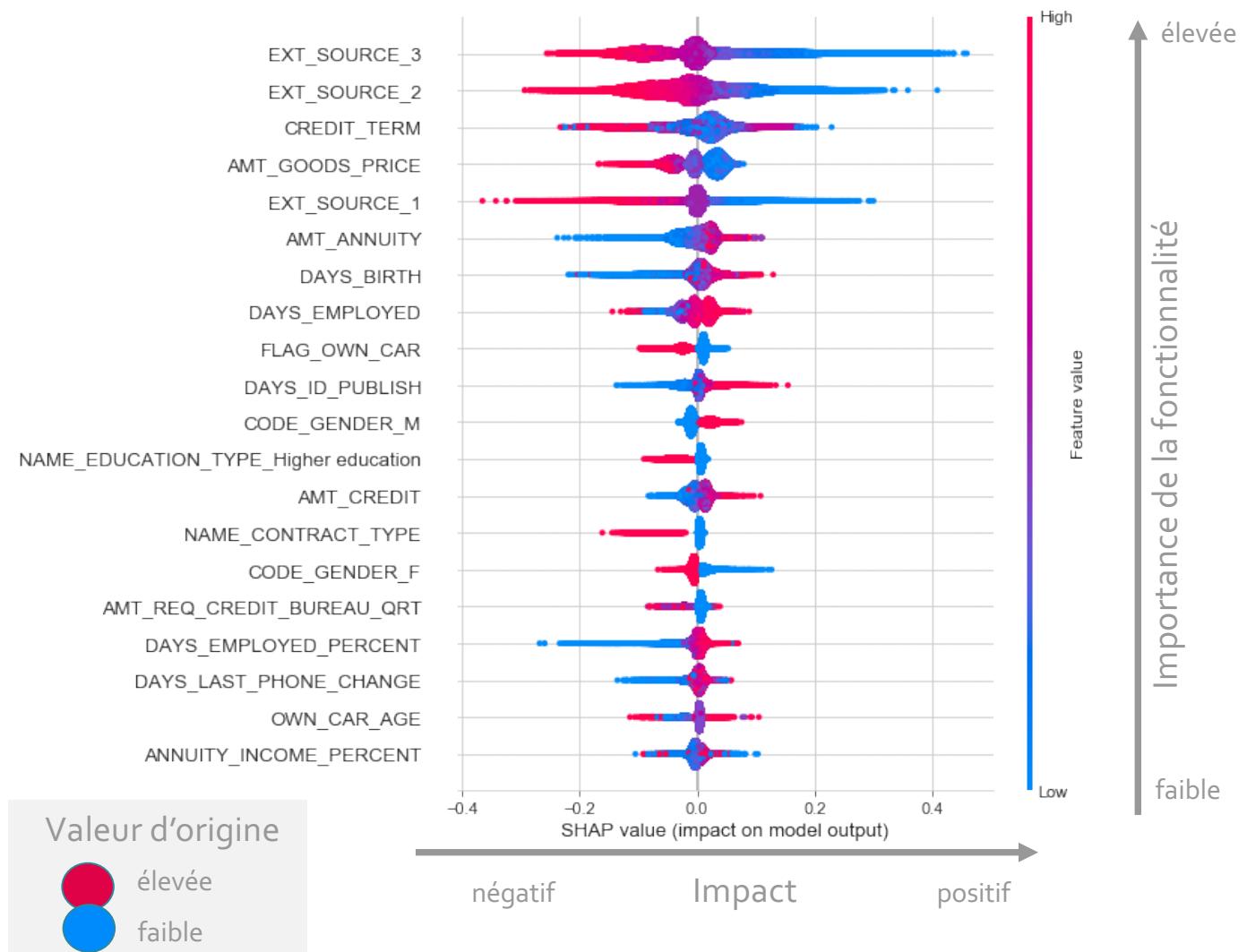
Interprétabilité globale

summary plot

Visualiser les variables les plus importantes et l'amplitude de leur impact sur le modèle



Diagramme de l'importance des variables



Interprétabilité locale

force-plot

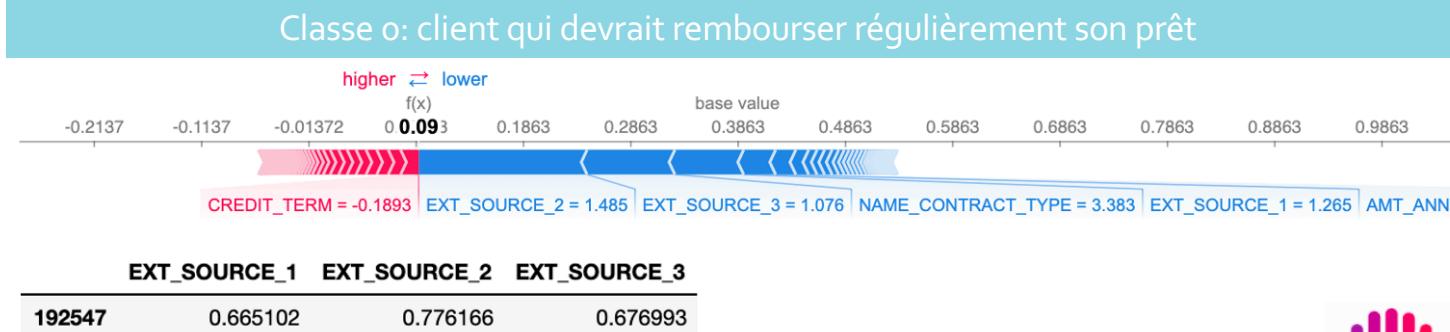
influence des paramètres, intensité et effets, sur la prédiction de la probabilité de faillite d'un client.

cas des variables EXT_SOURCE:

Globalement, un client de la classe o présente des valeurs plus élevées pour chacune de ces 3 variables qu'un client de la classe 1.



	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
132241	0.348161	0.260721	0.427657



	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
192547	0.665102	0.776166	0.676993



SHAP

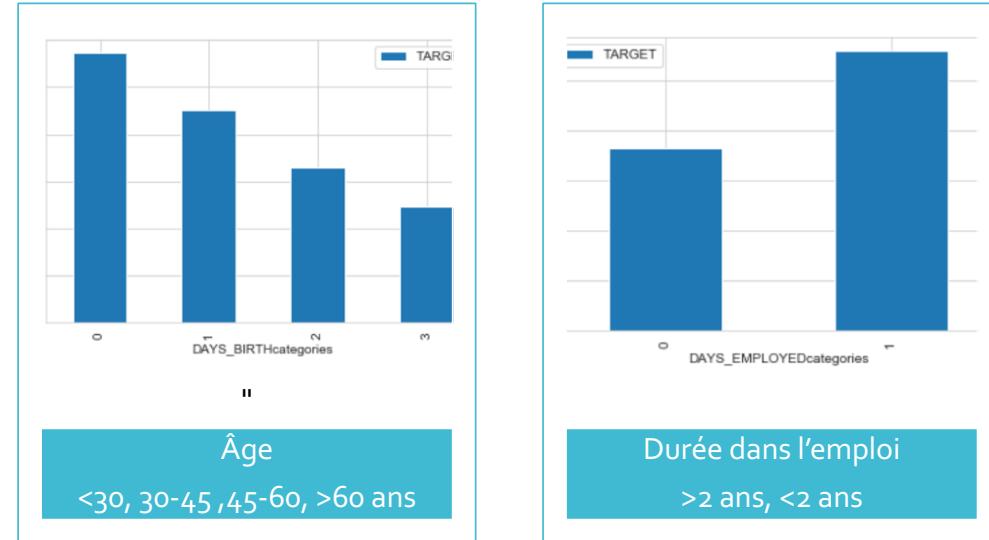
12

Clustering K-Means

Création de catégories

-> facilitation du clustering et de l'interprétation

-> proposer une comparaison de caractéristiques d'un client à celles de clients « similaires » via le dashboard



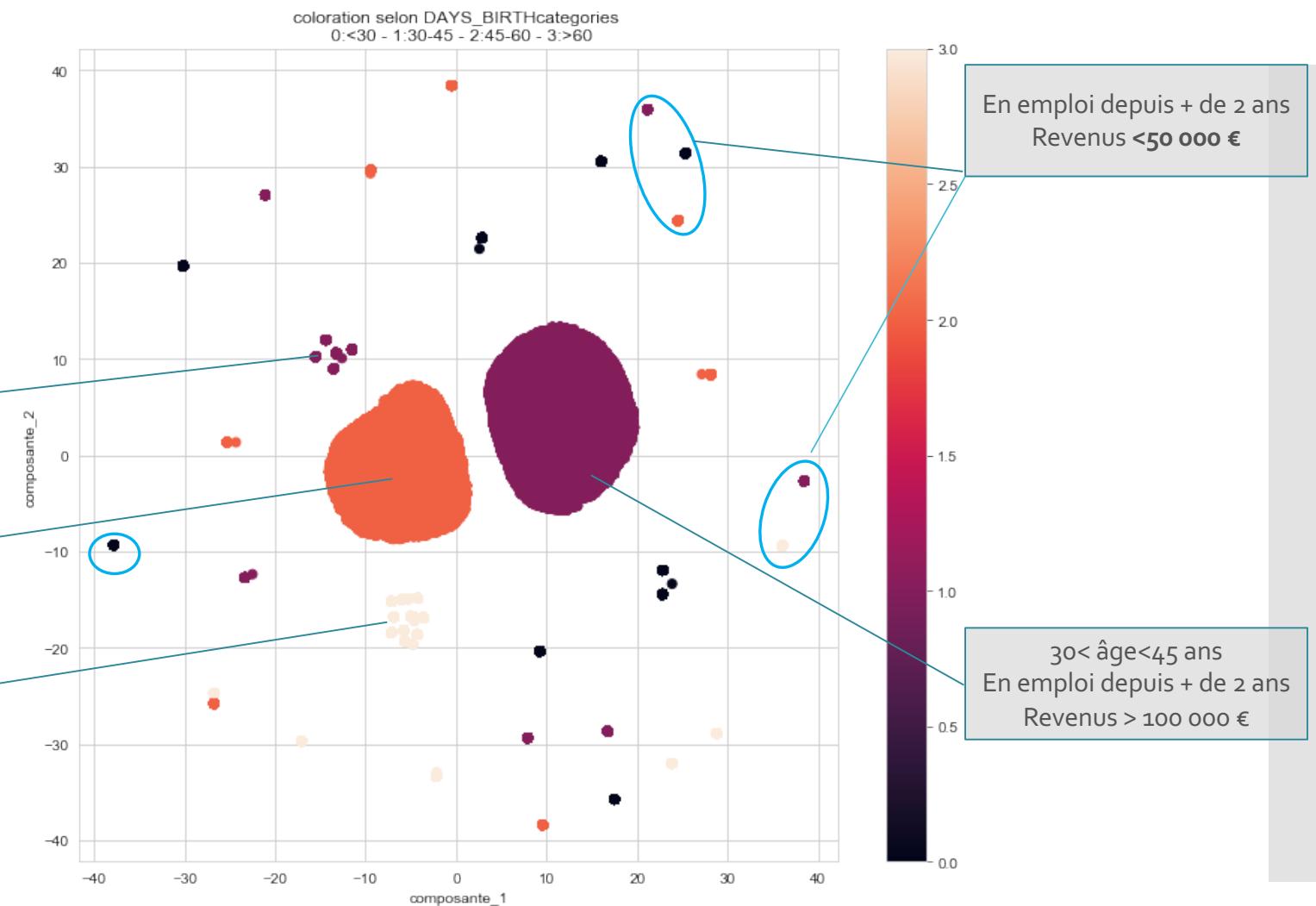
Clustering K-Means

32 clusters
silhouette : 1

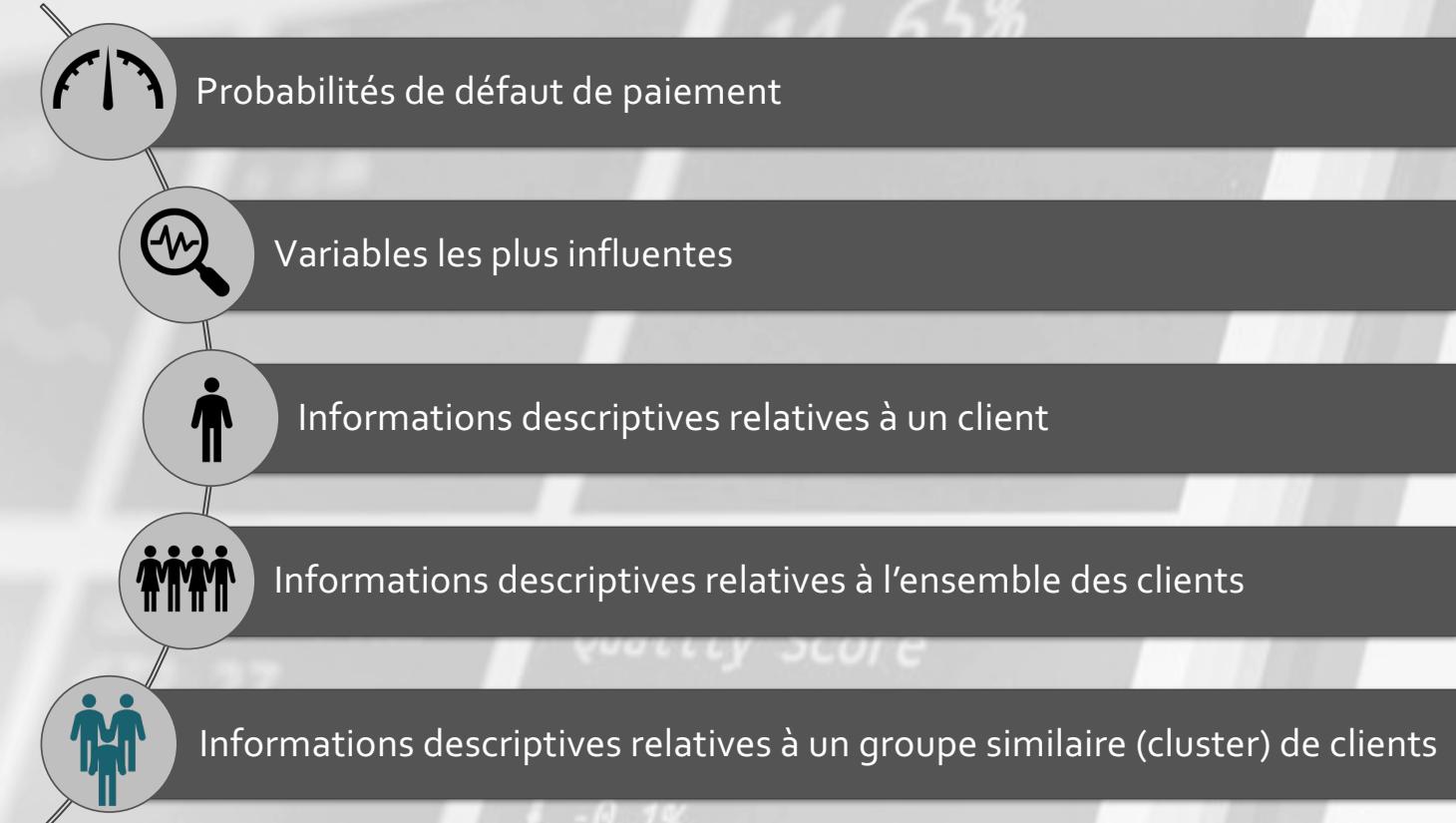
30 < âge < 45 ans
En emploi depuis - de 2 ans
Revenus > 100 000 €

45 < âge < 60 ans
En emploi depuis + de 2 ans
Revenus > 100 000 €

> 60 ans
En emploi depuis + de 2 ans
Revenus > 100 000 €



Eléments constitutifs du Dashboard



Dashboard

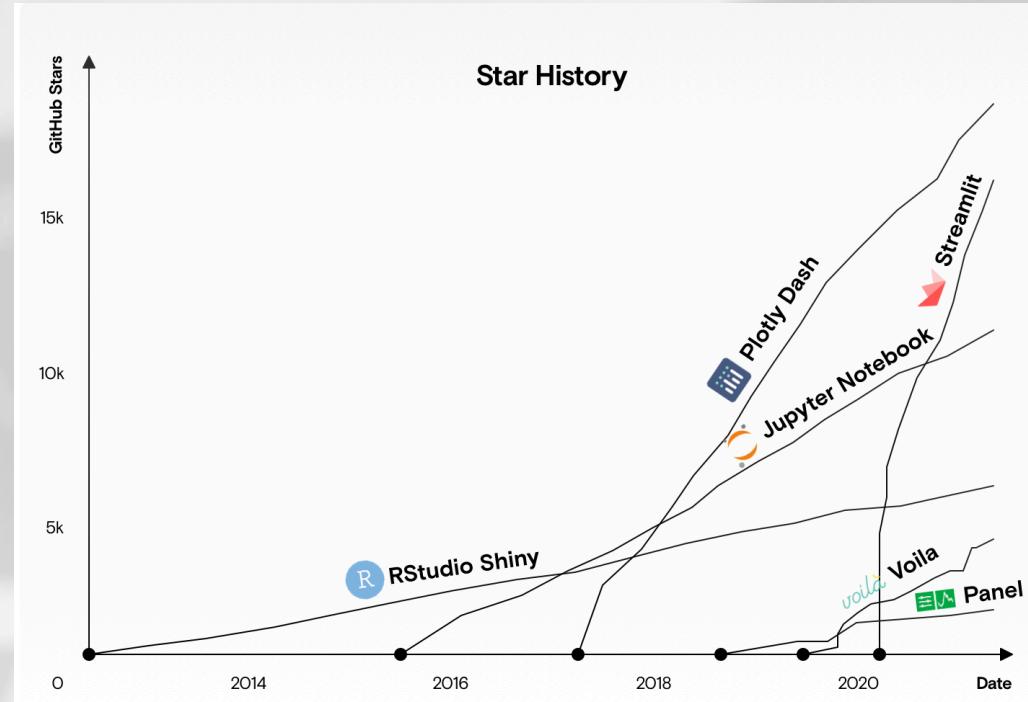
Solution :

Solutions récentes et populaires de tableau de bord complètes conçues avec Python:

Dashboard et Streamlit

Dashboard :

+flexible,
+mature,
+adapté à la production en entreprise



	Maturity	Popularity	Simplicity	Adaptability	Focus	Language support
Streamlit	C	A	A	C	Dashboard	Python
Dashboard	B	A	B	B	Dashboard	Python, R, Julia

Source: datarevenue.com



Développement

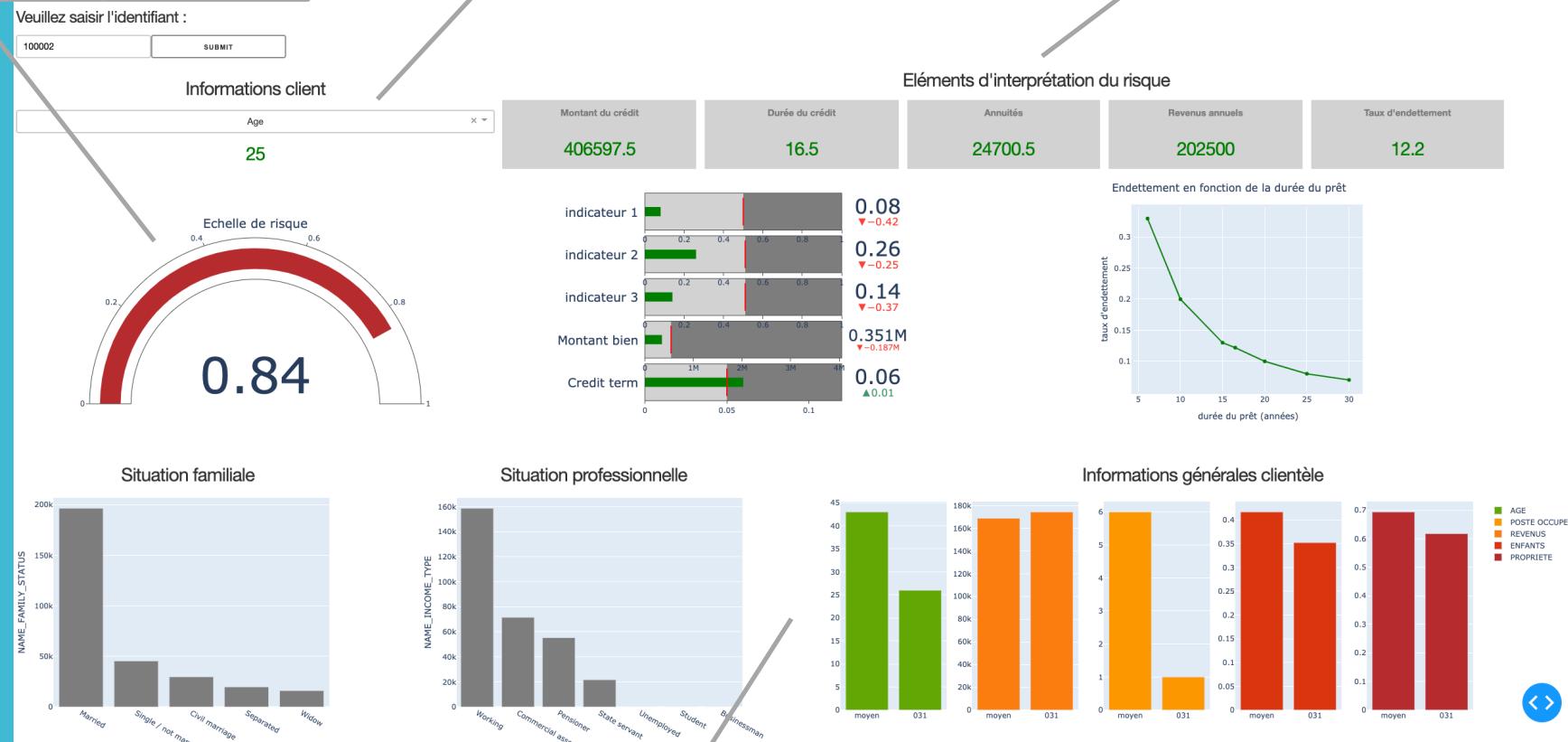


Déploiement

visualiser le score et son interprétation de façon intelligible pour une personne non expert

visualiser des informations descriptives relative à un client (filtre)

Éléments d'interprétation du risque



Comparer ces informations descriptives à l'ensemble des clients et/ou à un groupe similaire

<https://scoring-payment.herokuapp.com/>

Limites et améliorations



Déséquilibre des classes

- obstacle pour l'optimisation de la prédiction
- alimenter, dans la mesure du possible, la classe minoritaire

Amélioration du préprocessing

- réflexion sur les imputations
- révision des catégories des variables catégorielles
- autres données externes en complément
- création éventuelles de variables composites // expertise métier
->améliorer la qualité des données et l'information fournie à l'algorithme

Fonction de coût métier

- à discuter et ajuster avec l'expertise métier

Variables EXT_SOURCE

- a priori les plus influentes, à décrire

Dashboard

- Prévoir les cas où l'ID n'existe pas
- Mettre à jour une variable et réévaluer la probabilité

Limites et améliorations



Ethique

- les personnes présentant un défaut de crédit sont davantage des personnes jeunes
-> il conviendra de ne pas stigmatiser cette catégorie de clients.
- autre point d'attention : le sexe de l'emprunteur.
Selon les jeux de données utilisées pour entraîner les algorithmes, des biais peuvent naître notamment des biais liés au sexe de la personne.
Ex: programme de l'Apple card dont l'algorithme a attribué une limite de crédit 20 fois plus élevée à l'homme d'un couple marié qu'à sa femme déclarant pourtant leurs impôts conjointement.
- transparence et interprétabilité:
 - Il convient de ne pas fonder une décision d'octroi de crédit uniquement sur la probabilité fournie par l'algorithme.
 - En Europe, article 22 du RGPD:
« La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire ».

Finalement

- **si des améliorations peuvent être apportées, il va s'agir de trouver le meilleur compromis entre performance de l'algorithme et interprétabilité.**