


PARCOURS DATA SCIENTIST 2021-2022

SOMMAIRE

- 
- Projet 1** DÉFINIR une **stratégie d'apprentissage**
 - Projet 2** ANALYSER des **données** de **systèmes éducatifs**
 - Projet 3** CONCEVOIR une **application** au service de la **santé publique**
 - Projet 4** ANTICIPER les **besoins en consommation électrique** de bâtiments
 - Projet 5** SEGMENTER des **clients** d'un **site e-commerce**
 - Projet 6** CLASSIFIER automatiquement des **biens de consommation**
 - Projet 7** IMPLÉMENTER un **modèle de scoring**
 - Projet 8** DÉPLOYER un modèle dans le **cloud**

PROJET 7

IMPLÉMENTER un modèle de scoring

- 1 • **CONTEXTE** - Rappel de la problématique et présentation du jeu de données (5 min)
- 2 • **MODÉLISATION** - Explication de l'approche de modélisation (10 min)
- 3 • **DASHBOARD & DÉPLOIEMENT** - Présentation du dashboard et des outils utilisés pour le déploiement (5 min)
- 4 • **CONCLUSION/Q&A** (5-10 min)

Préliminaire - Les liens utiles pour la bonne lecture du projet

- Instructions Open Classroom
- Données utilisées
- Notebooks issus de Kaggle
- API
- Dashboard
- Tous les documents de travail produits dans le cadre de ce projet ont été déposés sur un repository Git

1. Contexte

Rappel de la problématique

Feature engineering

Traitement du déséquilibre des données

Présentation de la problématique



« *Prêt à dépenser* » est une
société de crédits à la
consommation



Missions

1 Construire un **modèle de scoring**

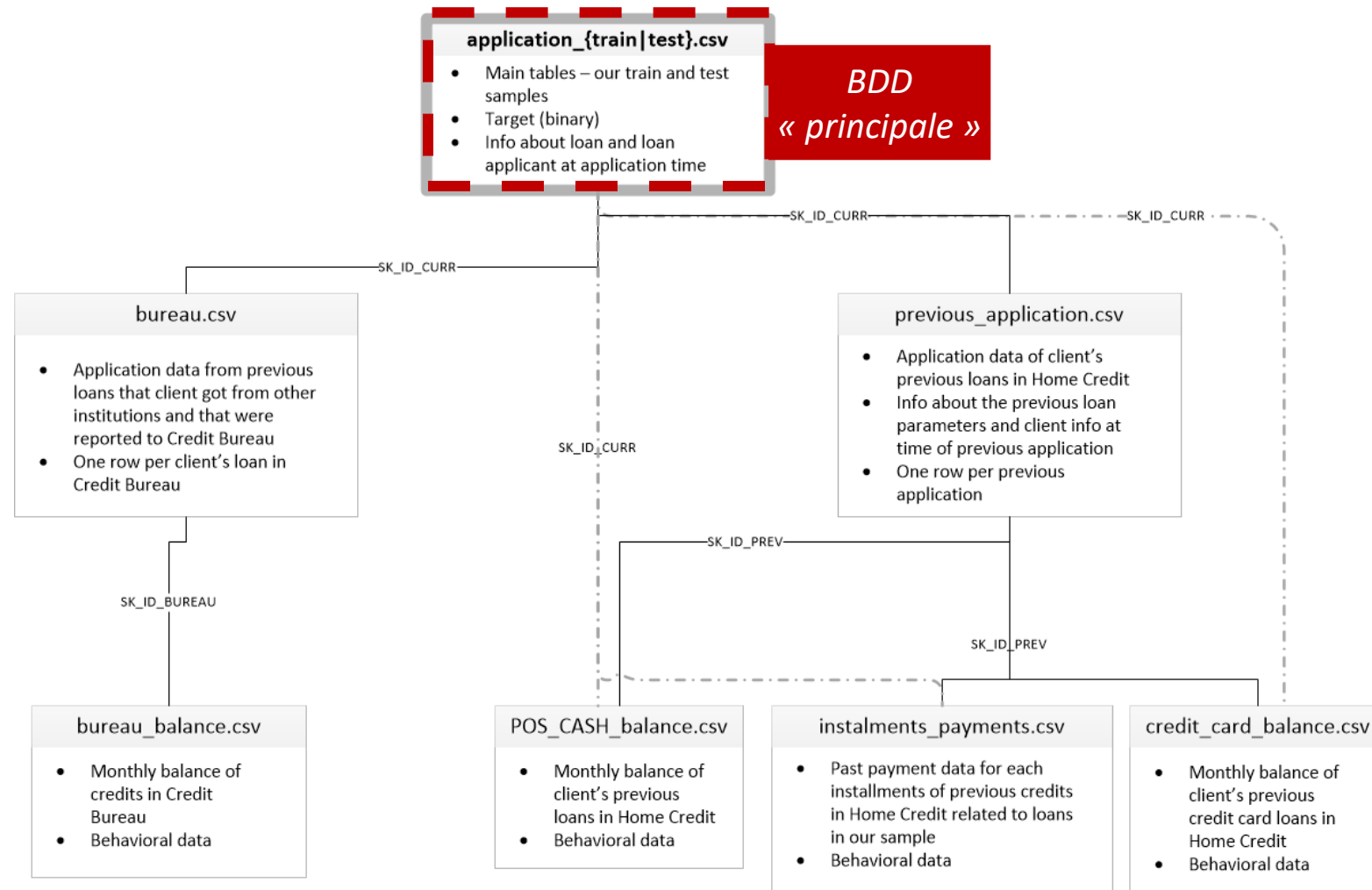
2 Construire un **dashboard interactif**

Objectifs

- Donner une **prédiction sur la probabilité de faillite** d'un client de façon automatique

- **Interpréter plus facilement les prédictions** faites par le modèle
- **Améliorer la connaissance client** des chargés de relation client

Jeux de données disponibles



Feature engineering



Suppression des colonnes avec un **taux de NaN supérieur à 40%** (considérées comme moins utilisables que les autres)



Imputation de la médiane par colonne dans les NaN restants



Vérification de la « **qualité** » de nos données, avec gestion des valeurs absurdes



Obtention de nouvelles variables à partir d'agrégats numériques et catégoriels



Création de features métiers

- **+ d'autres traitements apportés** : visualisation des corrélation avec notre cible, encodage des variables catégorielles, standardisation des données...

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 356255 entries, 0 to 356254
Data columns (total 28 columns):
#   Column                                  Non-Null Count  Dtype
---  -
0   applicant_loan_id                       356255 non-null  int64
1   target                                  307511 non-null  float64
2   credit_payment_type                     356255 non-null  int64
3   applicant_gender                         356255 non-null  int64
4   flag_car_owner_applicant                356255 non-null  int64
5   applicant_total_income                  356255 non-null  float64
6   total_credit_amount                     356255 non-null  float64
7   applicant_best_education                 356255 non-null  int64
8   applicant_family_status                 356255 non-null  int64
9   applicant_housing_type                  356255 non-null  int64
10  level_pop_living_region                 356255 non-null  float64
11  applicant_occupation                     356255 non-null  int64
12  internal_rating_living_region            356255 non-null  int64
13  weekday_starting_process                 356255 non-null  int64
14  type_of_set                              356255 non-null  int64
15  applicant_age                            356255 non-null  float64
16  annuity_share_to_income                 356255 non-null  float64
17  children_in_household_rate              356255 non-null  float64
18  bureau_count_past_loans                  356255 non-null  float64
19  bureau_count_credit_prolongations        356255 non-null  float64
20  bureau_seniority_past_loans              356255 non-null  float64
21  bureau_share_active_loans                356255 non-null  float64
22  applicant_bank_account_seniority         356255 non-null  float64
23  avg_amount_available_bank_account        356255 non-null  float64
24  cumulative_number_of_days_late           356255 non-null  float64
25  previous_application_accepted_share      356255 non-null  float64
26  previous_application_credit_term         356255 non-null  float64
27  share_previous_refused_applications      356255 non-null  float64
dtypes: float64(17), int64(11)
memory usage: 76.1 MB
```

Obtention d'un jeu de données de 356k clients (train + test) et 28 variables

Présentation des variables obtenues

1 variable id pour **chaque demande de prêt**

(applicant_loan_id)

1 variable à prédire **répartie en deux classes**

(target)

10 variables **déjà présentes** dans nos différents data sets

(credit_payment_type, applicant_gender...)

9 variables métiers créées à partir de nos connaissances du domaine

(applicant_age, annuity_share_to_income...)

6 variables créées sur la **base d'agrégats**

(bureau_count_past_loans, bureau_count_credit_prolongations...)

1 variable **utilisée pour le traitement** mais supprimée avant la modalisation

(type_of_set)

Résolution du déséquilibre entre les classes à prédire

Problématique

- Notre base de données est déséquilibrée du point de vue des classes qu'elles contient : 92% de nos clients remboursent leur prêt (valeur = 0) ; **seulement 8% ne le remboursent pas** (valeur = 1)
- Alors que notre objectif reste bien de classer ces clients en particulier, on sait que **la plupart des algos fonctionnent mieux avec un nombre d'échantillon à peu près égal dans chaque classe**
- Il y aura donc bien ici **potentiellement une problématique de surreprésentation de la classe majoritaire dans la prédiction**

Possibles solutions

- 1 Changer la **structure au global** de l'algorithme
- 2 **Réduire le nombre d'individus** dans la classe majoritaire
- 3 **Collecter plus de données** sur la classe minoritaire
- 4 **Dupliquer les individus sous-représentés**
- 5 **Pondérer les observations** dans le training
(dans les hyperparamètres de nos algorithmes de ML)
- 6 Choisir une **métrique de performance adaptée**
(expliqué plus loin dans la présentation)
- 7 **Créer des individus « synthétiques »**
(suréchantillonnage de minorité synthétique SMOTE)

Les solutions retenues

2. Modélisation

Métrique de performance

Méthodologie

Modèle retenu

Quel est la mesure la plus adaptée à notre problème ?

		Classe réelle	
		Rembourse (0)	Ne rembourse pas (1)
Classe prédite par le modèle	Rembourse (0)	Vrais positifs (TP)	1 Faux négatifs (FN) <i>On cherche avant tout à détecter le nombre de faux négatifs</i>
	Ne rembourse pas (1)	2 Faux positifs (FP) <i>Dans une moindre mesure, on veut limiter le nombre de faux positifs</i>	Vrais négatifs (TN)

Les métriques possibles

Sensitivity/recall

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

Precision

$$PPV = \frac{TP}{TP + FP} = 1 - FDR$$

Specificity

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

Negative predictive value

$$NPV = \frac{TN}{TN + FN} = 1 - FOR$$

Accuracy

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

Equilibre à trouver entre ces 2 métriques

Adaptation du F-Beta Score au métier

Formule F-Beta-Score

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Beta : poids du recall dans le score combiné ($\beta \geq 1 \rightarrow$ on accorde plus d'importance au recall)



Coût moyen du défaut de paiement d'un client



Coût d'opport. d'un client accidentellement écarté

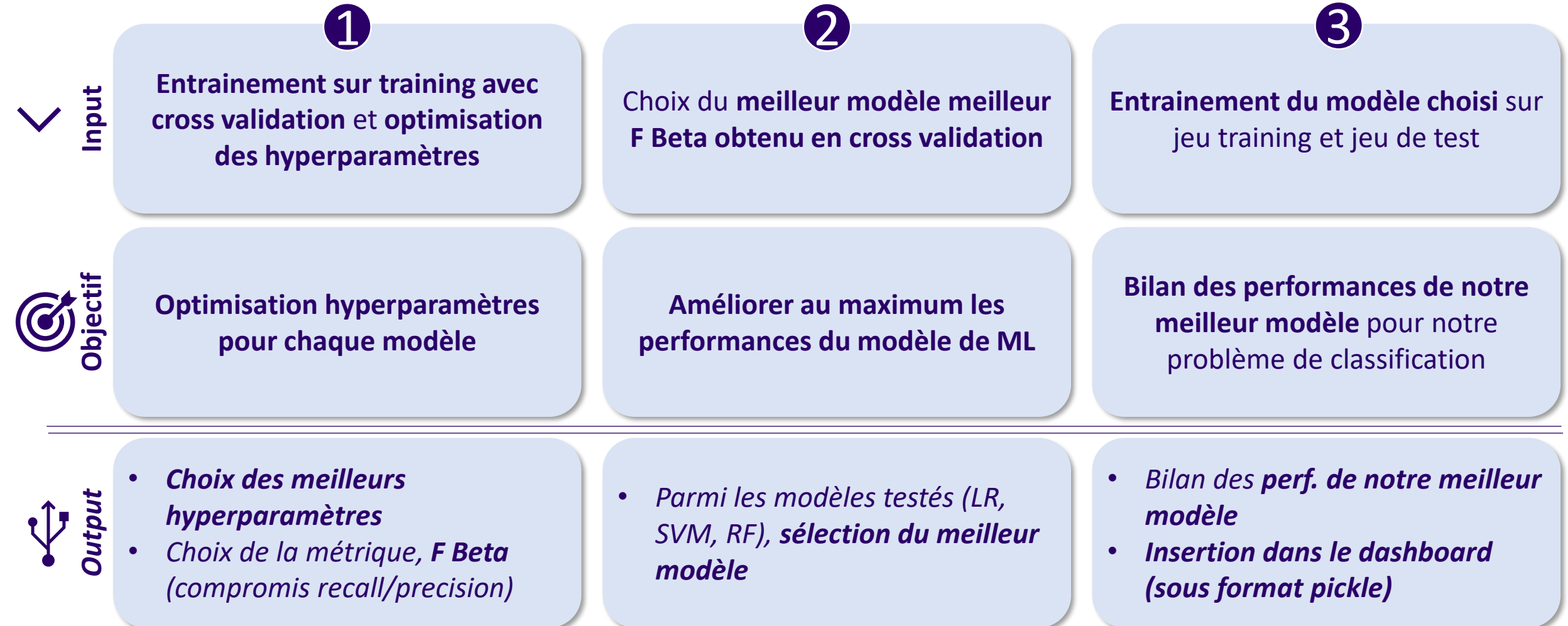
Hypothèses métiers

On obtient donc : Beta = coefficient recall / coefficient precision = 2,7



Hypothèses à faire vérifier par les équipes métiers

Méthodologie de modélisation



Hyperparamètres sélectionnés et résultat de modélisation



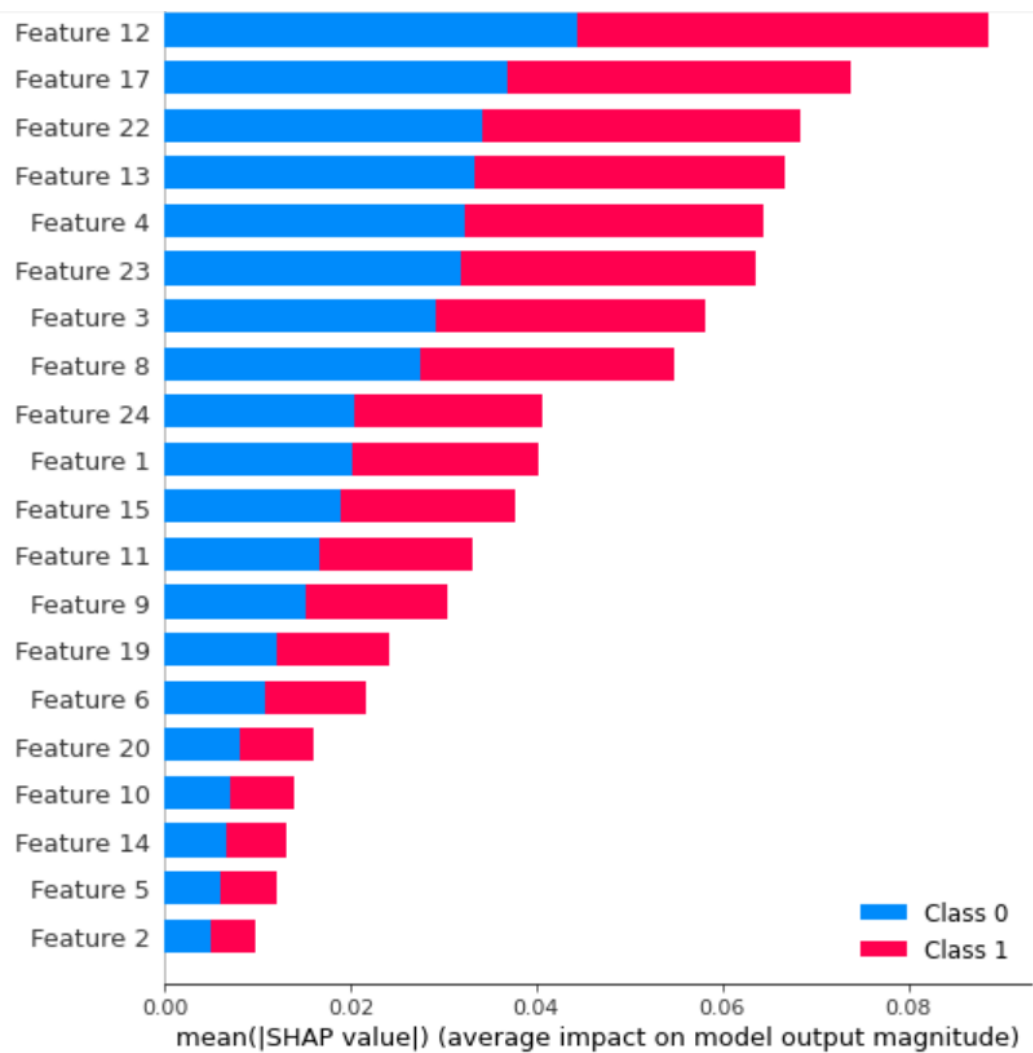
Régression logistique	SVM	Forêt Aléatoire
<ul style="list-style-type: none">• C• Solver• Class Weight• Penalty	<ul style="list-style-type: none">• C• Gamma• Class_weight• Kernel	<ul style="list-style-type: none">• n_estimators = 1000• Class_weight = 'balanced'• Max_depth = 15• Bootstrap = 'True'

Interprétabilité du modèle (1/3) – Feature importance

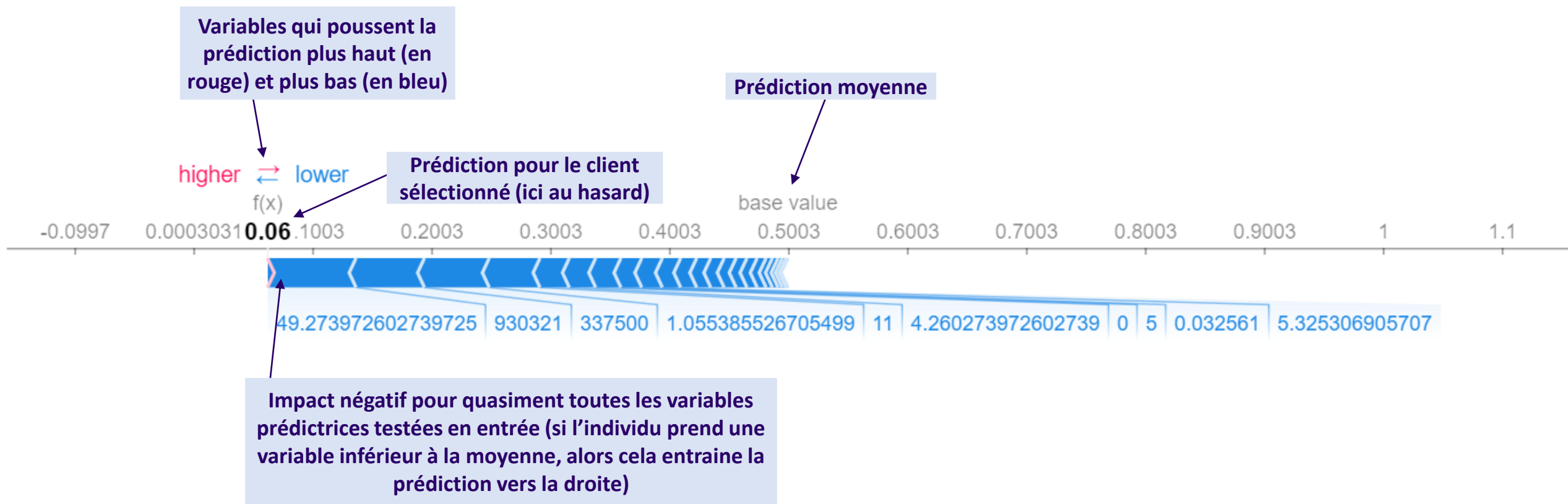
	Feature	Poids
12	applicant_age	0.111030
13	annuity_share_to_income	0.087087
22	previous_application_accepted_share	0.084529
4	total_credit_amount	0.083232
17	bureau_seniority_past_loans	0.079224
23	previous_application_credit_term	0.078378
3	applicant_total_income	0.069703
8	level_pop_living_region	0.069406
15	bureau_count_past_loans	0.046111
24	share_previous_refused_applications	0.041114

Dans notre modèle, **calcul de l'importance générale des variables**, qui nous donne les **variables les importantes dans l'explication de notre variable cible** (*facile d'utilisation et calcul rapide, mais a tendance à gonfler l'importance des variables continues*)

Interprétabilité du modèle (2/3) – Valeurs de SHAP (interprétabilité globale)



Interprétabilité du modèle (3/3) – Valeurs de SHAP (interprétabilité globale)



3. Dashboard & Déploiement

Outils utilisés

Schéma explicatif

Présentation du dashboard

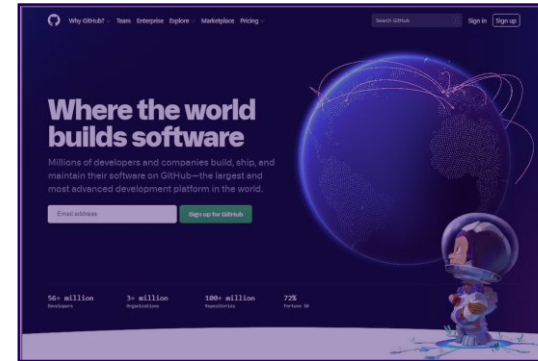
Les outils utilisés



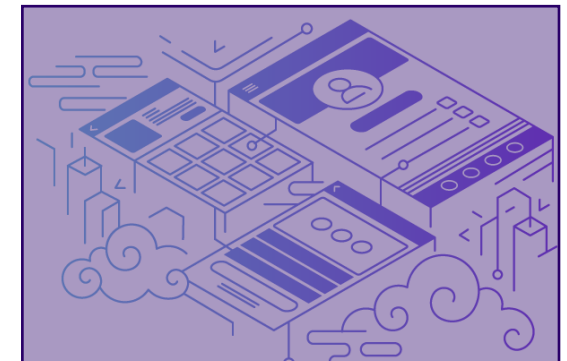
API pour appeler la
prédiction à partir de l'ID
client



Dashboard présenté à
l'utilisateur



Versioning



Déploiement de l'API et
du dashboard dans le
Cloud



Présentation du dashboard réalisé

Disponible en
cliquant sur ce lien



Choix du client

100005



ID sélectionné: 100005

Choix de la page

Sélectionner la page correspondante

Introduction



Si vous constatez un bug ou avez un besoin spécifique, contactez-nous!

Dashboard interactif à destination des gestionnaires de la relation client

Prêt à dépenser a développé pour vous un dashboard interactif qui permettra une explication transparente des décisions d'octroi de crédit et une présentation claire des informations personnelles de chacun de vos clients

Ce dashboard vous permettra de :

- Visualiser le score et l'interprétation du score pour chaque client
- Visualiser des informations relatives à un client
- Comparer les informations relatives à un client à un groupe de client similaire

Produit par





Choix du client

100005

ID sélectionné: 100005

Choix de la page

Sélectionner la page correspondante

Prédiction score

Si vous constatez un bug ou avez un besoin spécifique, contactez-nous!

Dashboard interactif à destination des gestionnaires de la relation client

Le client sélectionné appartient donc au groupe suivant :

A surveiller

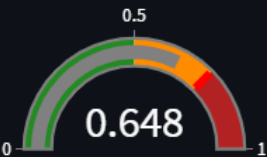


Dossiers à ré-étudier



Le client sélectionné a peu de chances de ne pas rembourser son prêt ; cela vaut le coup de retravailler et ré-étudier son dossier

Dossiers à surveiller



Le client sélectionné est dans la moyenne, mais nécessite une surveillance de la part de nos services pour cette demande de crédit et/ou des aménagements spécifiques

Dossiers à refuser



Le client sélectionné a une forte probabilité de ne pas rembourser son prêt, il est possible de rendre une décision favorable à sa demande, uniquement dans des cas exceptionnels



Choix du client

100005

ID sélectionné: 100005

Choix de la page

Sélectionner la page correspondante

Analyse client

Si vous constatez un bug ou avez un besoin spécifique, contactez-nous!

Sélectionnez une première variable

applicant_total_income

Sélectionnez une deuxième variable

total_credit_amount

Dashboard interactif à destination des gestionnaires de la relation client

Découvrez quelques indicateurs clés à propos du client sélectionné

Sexe

Age

Statut marital

Revenus

Homme

49.0

Marié(e)

99000.0

Comparez les indicateurs de votre client aux autres groupes

Chaque graphique vous permet de comparer deux variables, ainsi que les performances du groupe en moyenne sur la première variable sélectionnée

Dossiers à ré-étudier

99k

▼ -47107





Choix du client

100005

ID sélectionné: 100005

Choix de la page

Sélectionner la page correspondante

Rapport client

Si vous constatez un bug ou avez un besoin spécifique, contactez-nous!

Rédigez un rapport pour le client dont vous venez de regarder le dossier

Nom du conseiller :

Bernard

Objet du rapport :

Rendez-vous à fixer

Commentaire :

RDV à fixer car le client part en vacances à partir du 06/06

Date de rapport :

2022/04/21

Criticité :

1

2

5

Soumettre

Merci, votre rapport a bien été enregistré

Voir tous les rapports produits




	Client	Conseiller	Objet	Urgence	Commentaire	date
0	100005	Bernard	Rendez-vous à fixer	2	RDV à fixer car le client part en vacances à partir du 06/06	2022-04-21

4. Conclusion

Points d'amélioration possibles

Q&A

Points d'améliorations

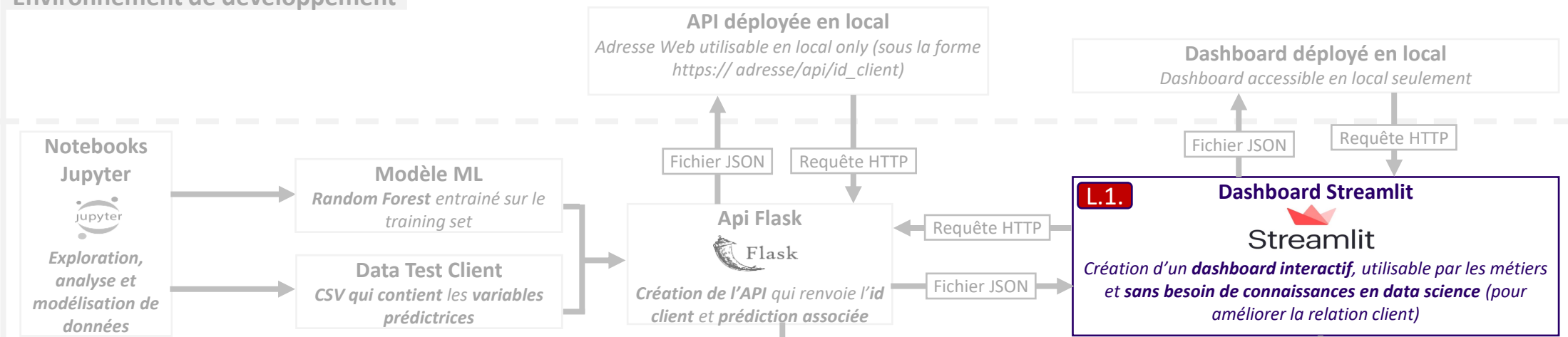
	Modélisation	<ul style="list-style-type: none">• Retravail sur l'appel aux variables d'environnement dans les notebooks• Rajout de l'ensemble des données utilisées dans la prédiction• Rajout d'un algorithme de clustering pour séparer les clients en différents groupes• Confirmation des hypothèses métiers pour le calcul du Beta dans la métrique d'évaluation
	Dashboard	<ul style="list-style-type: none">• Rajout du coût métier dans le dashboard (présenté en note de méthodologie)• Insertion des valeurs de Shapley dans le dashboard (présenté dans le notebook Modélisation)• En parallèle, allègement du dashboard (ou changement de logiciel pour passer à une version plus poussée et plus utilisable par de multiples utilisateurs)
	Déploiement	<ul style="list-style-type: none">• Gagner en agilité sur l'utilisation des différents outils mis à disposition pour le déploiement (première prise en main de Heroku seulement)

Q&A

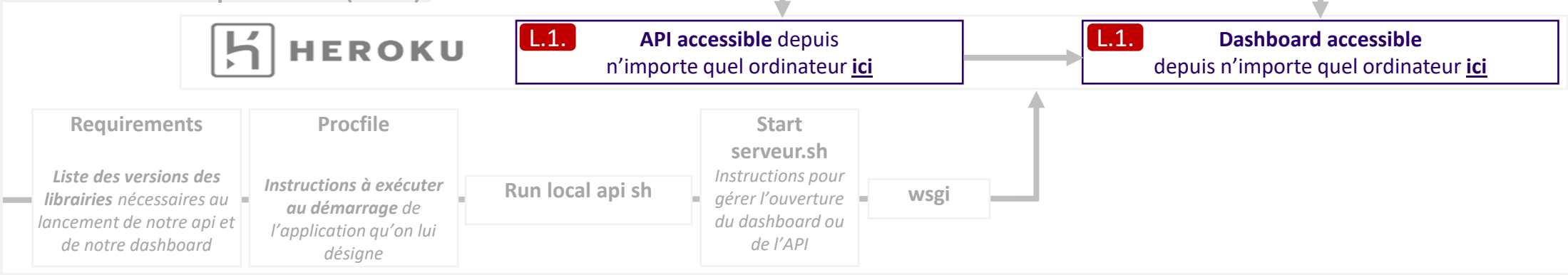
5. Annexes

L.x. Livrable du projet

Environnement de développement



Environnement de production (cloud)



Documents

L.3. Note méthodologique

L.4. Support de présentation

Enregistré dans [GitHub](#)

L.2.

Versioning