Implémentez un modèle de scoring.

Rappel de la problématique

- « Prêt à dépenser » : société proposant des crédits aux personnes ayant peu d'historique de prêt.
- Volonté de cette société : développer un modèle de scoring de la probabilité de défaut de paiement du client pour renforcer sa décision d'accorder ou non un prêt.
- Besoin de cette société : création d'un dashboard interactif à l'intention de ses chargés de relation clients. Ceux-ci pourront expliquer à leurs clients, avec plus de transparence, les conditions d'octroi de crédit.

Mission

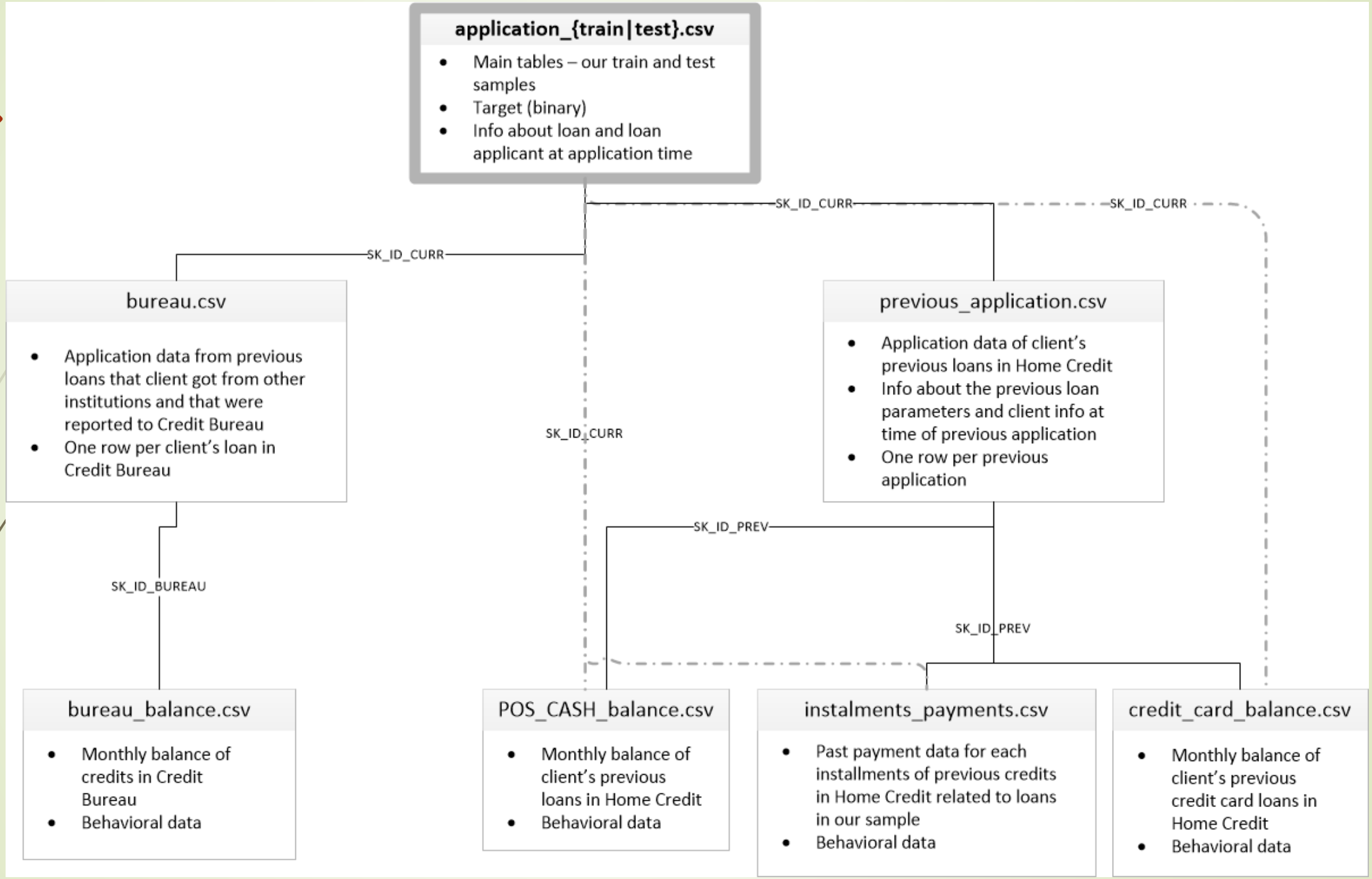
- Construction d'un problème de classification comprenant en entrée des données variées et en sortie la solvabilité ou non du client. Pour cela, notre manager nous demande de sélectionner un kernel kaggle et de l'adapter afin de répondre au mieux au besoin.
- Construction d'un dashboard interactif permettant d'interpréter les prédictions faîtes par le modèle et améliorer la connaissance de la clientèle.

Récapitulatif :

Présentation du jeu de données

- Le jeu de données se présente sous la forme de 7 fichiers au format csv.
- Les données sont des données comportementales, données personnelles sur le clients, données provenant d'autres institutions financières…

	Rows	Columns	%NaN	%Duplicate	type object	type float	type int	MB_Memory
application_train	307511	122	24.40	0.0	16	65	41	286.227
application_test	48744	121	23.81	0.0	16	65	40	44.998
bureau	1716428	17	13.50	0.0	3	8	6	222.620
bureau_balance	27299925	3	0.00	0.0	1	0	2	624.846
pos_cash_balance	10001358	8	0.07	0.0	1	2	5	610.435
credit_card_balance	3840312	23	6.65	0.0	1	15	7	673.883
installments_payements	13605401	8	0.01	0.0	0	5	3	830.408
previous_application	1670214	37	17.98	0.0	16	15	6	471.481


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Analyse exploratoire des données

- Cette analyse des données est inspirée par un notebook kaggle.
- Elle révèle que :
  - la classe cible du problème de classification est une classe déséquilibrée,
  - beaucoup de valeurs non définies dans les datasets et des anomalies, quelques valeurs atypiques à traiter.
- Elle détermine :
  - les principales valeurs corrélées avec la classe cible

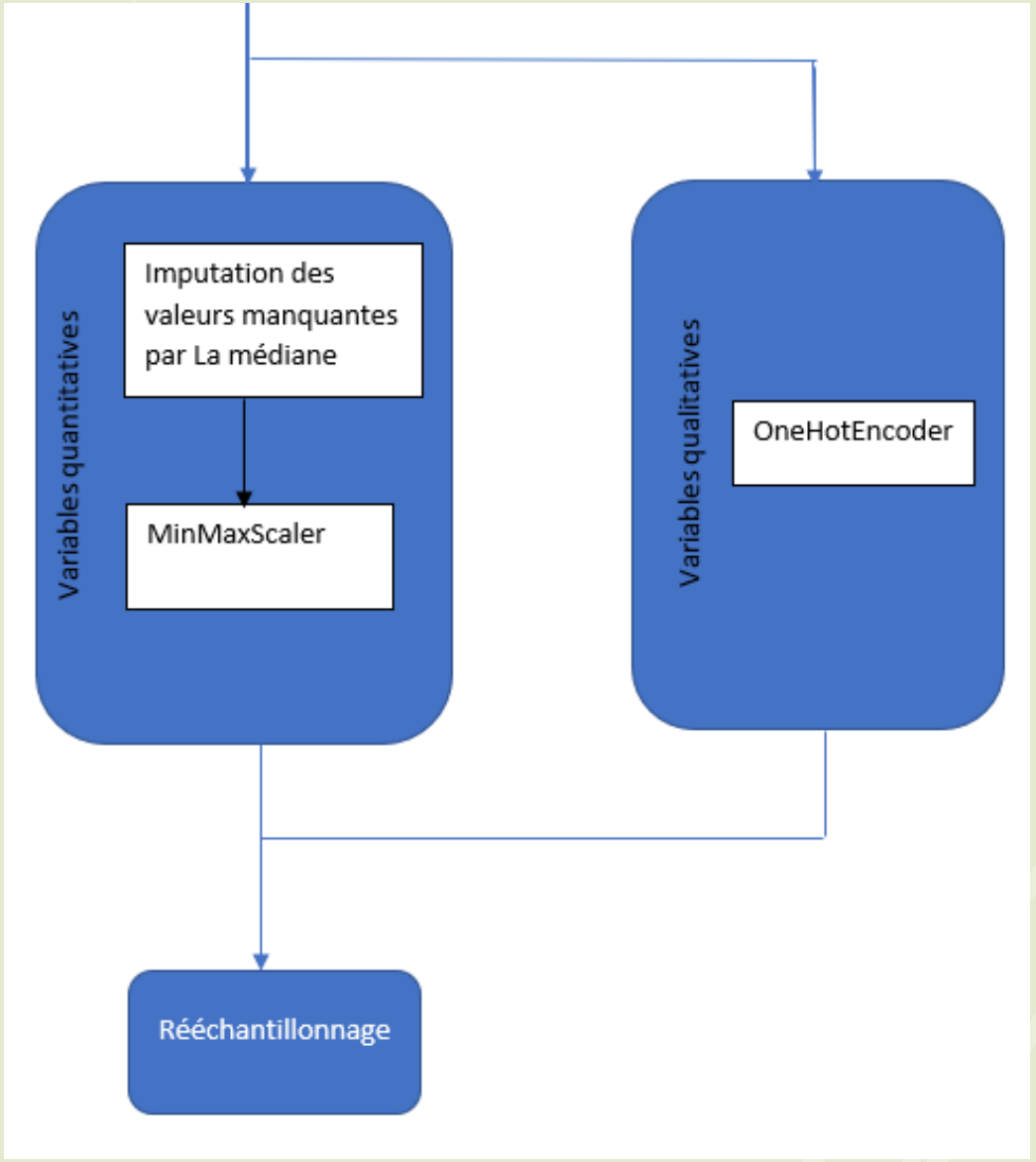
Merging et feature engineering automatique

- Traitement de tous les datasets autres que application_train et application_test :
  - Remplacement des valeurs aberrantes par la valeur Nan.
  - Création de nouvelles features issues des données quantitatives en utilisant la fonction mean.
  - Fusion de ces datasets entre eux pour obtenir le dataset train_prev_app_inst_pos_credit_bureau.
- Fusion du dataset application_train avec le dataset train_prev_app_inst_pos_credit_bureau pour obtenir le dataset train et du dataset application_test avec train_prev_app_inst_pos_credit_bureau pour obtenir le dataset test.
- Feature engineering manuel sur les dataset train et test

Feature engineering manuel

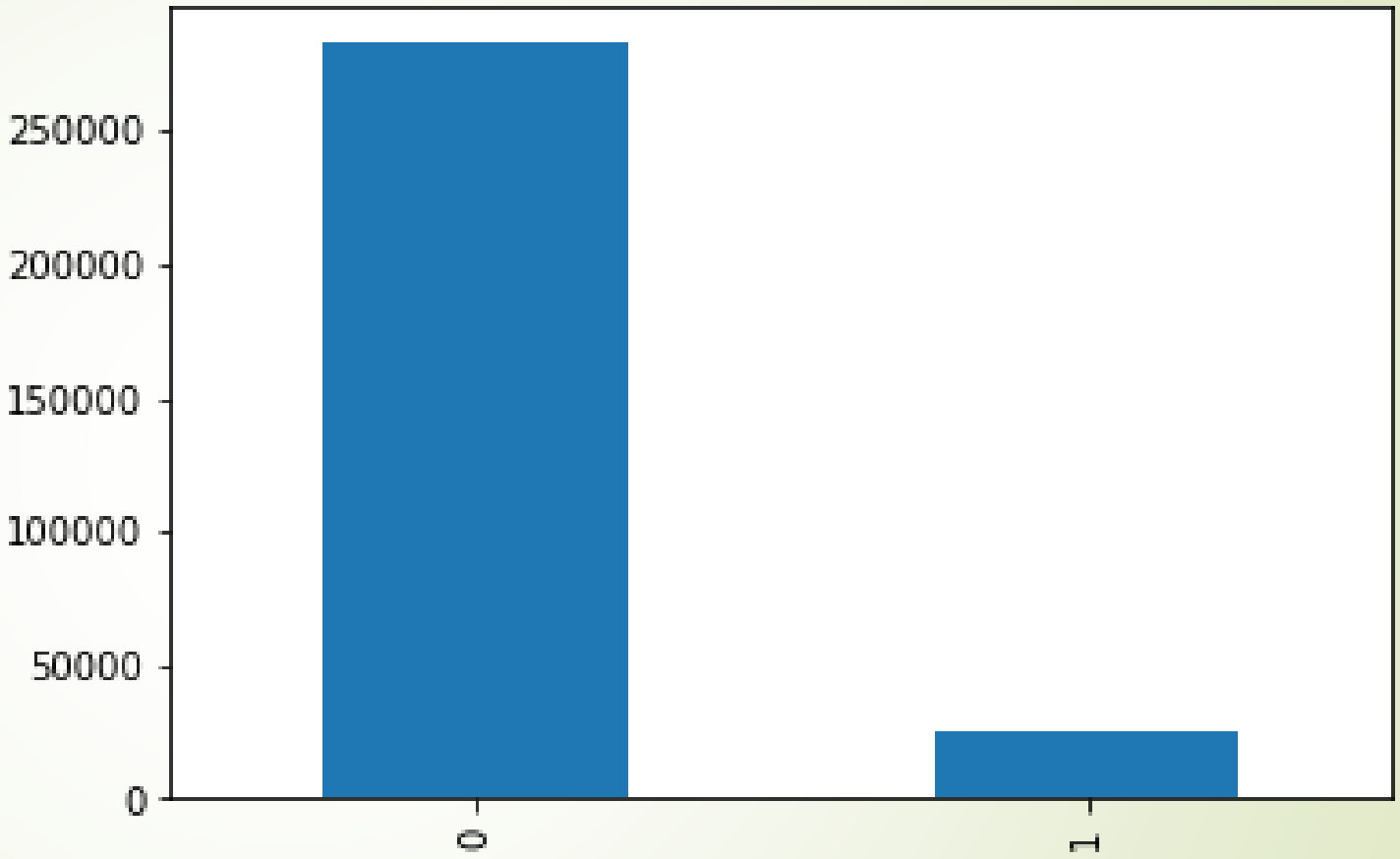
- Ajout de ratios explicatifs :
  - ratio montant du crédit du prêt /revenu du client
  - ratio montant du crédit du prêt/ prix des biens pour lesquels le crédit est accordé
  - ratio rente de prêt(annuité) /revenu du client
  - ratio rente de prêt (annuité) / montant du crédit du prêt
  - ratio revenu du client/ montant du crédit du prêt
  - ratio revenu du client / nb de personnes de la famille du client.

Preprocessing des données



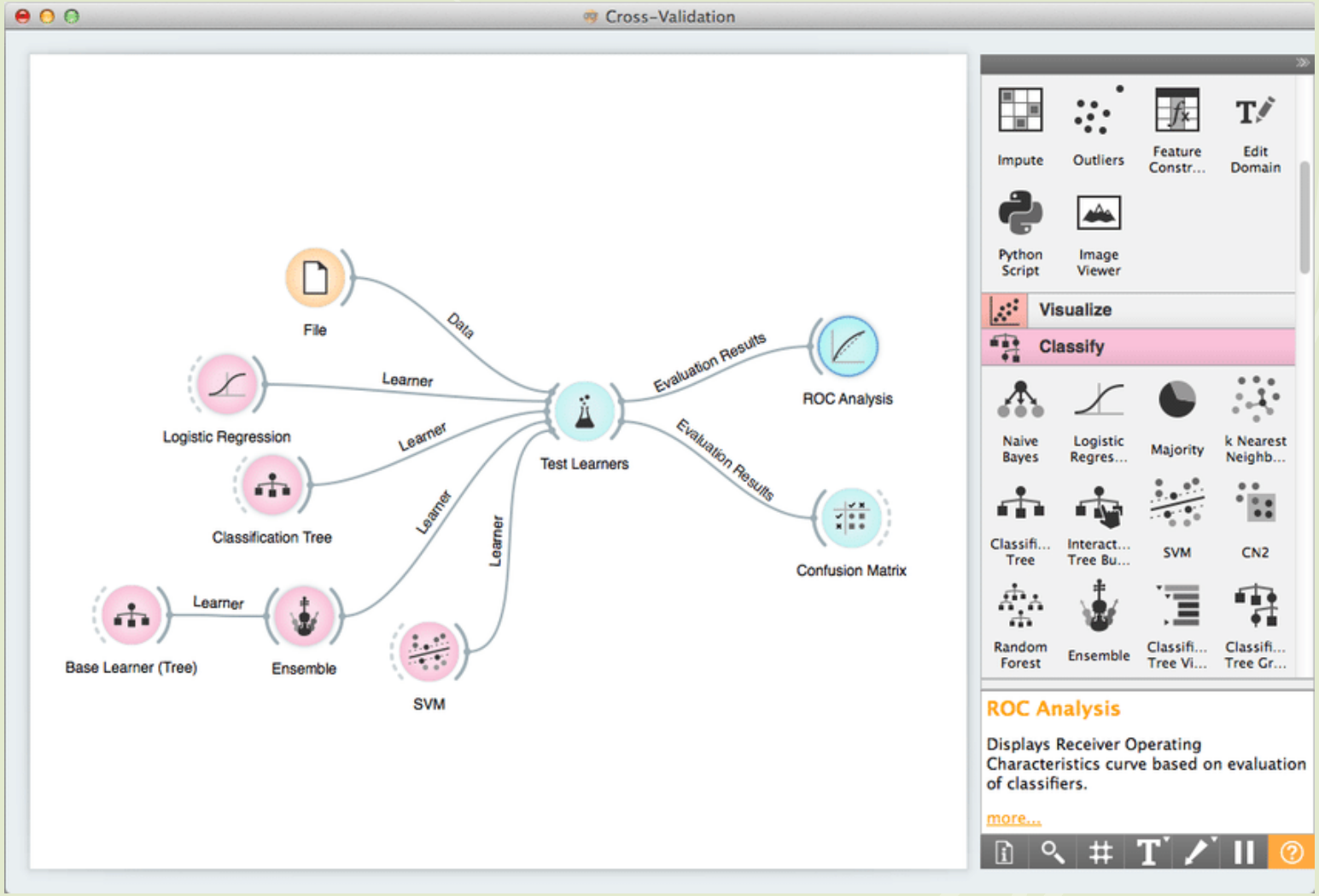
Le rééchantillonnage du jeu de données

- Pour faire face au déséquilibre des classes, on peut adapter l'étape du traitement de données du projet.
  - Le suréchantillonnage : rééquilibre du jeu de données en augmentant artificiellement la classe minoritaire.
  - Utilisation de SMOTE qui créé un nombre d'éléments d'observations synthétiques à partir de la classe minoritaire.



Distribution du Target

Modélisation

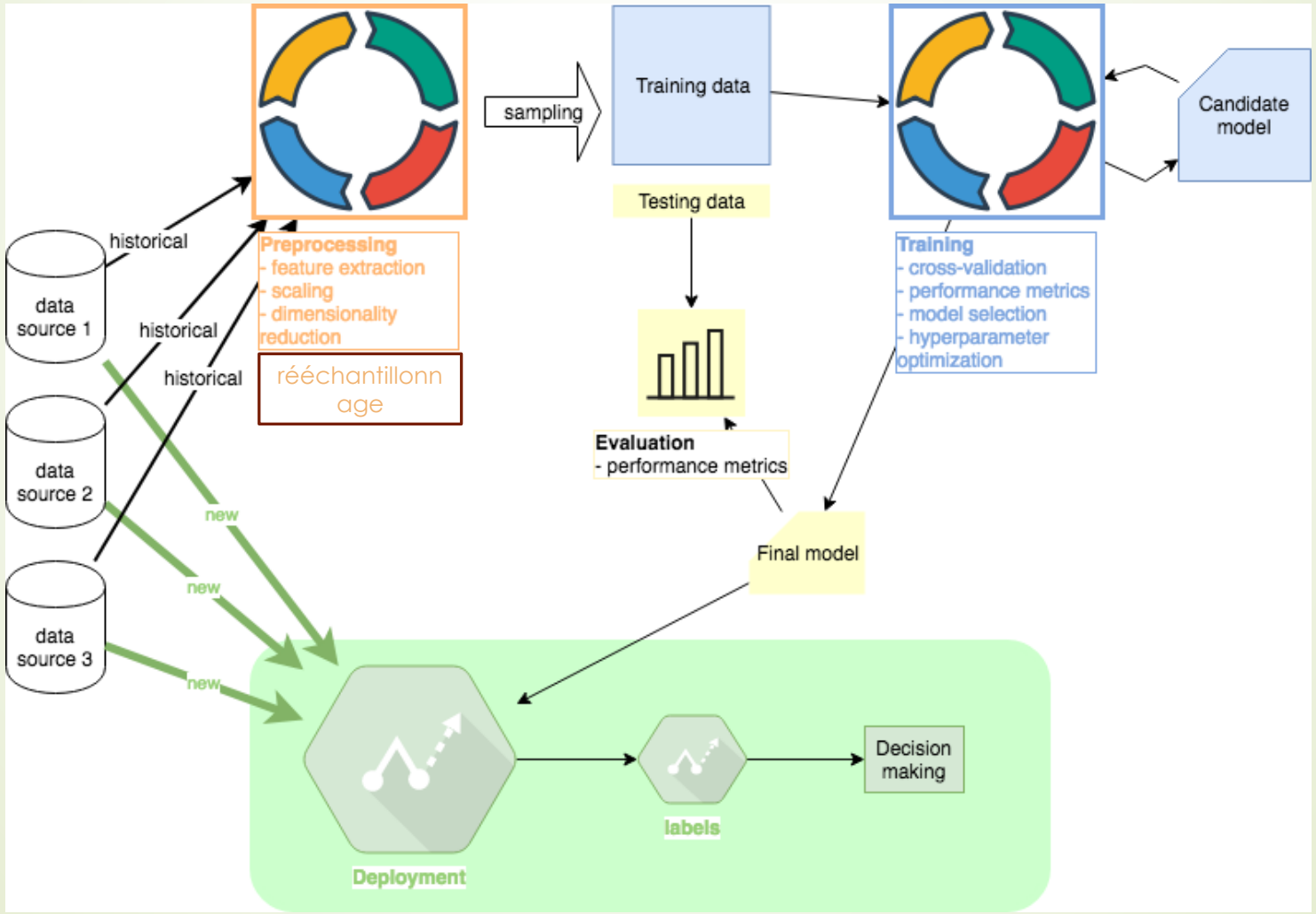


Modélisation

Nous allons utiliser comme modèle une régression logistique :

c'est un modèle simple s'adaptant aux situations où la variable se limite à deux choix,	particulièrement indiqué lorsque l'on veut exprimer une variable dichotomique sous forme de probabilité,	donne non seulement une mesure de la pertinence d'un prédicteur (taille du coefficient), mais aussi de sa direction d'association (positive ou négative).

Principes de la modélisation



Modélisation

Baseline : un modèle de classification aléatoire.

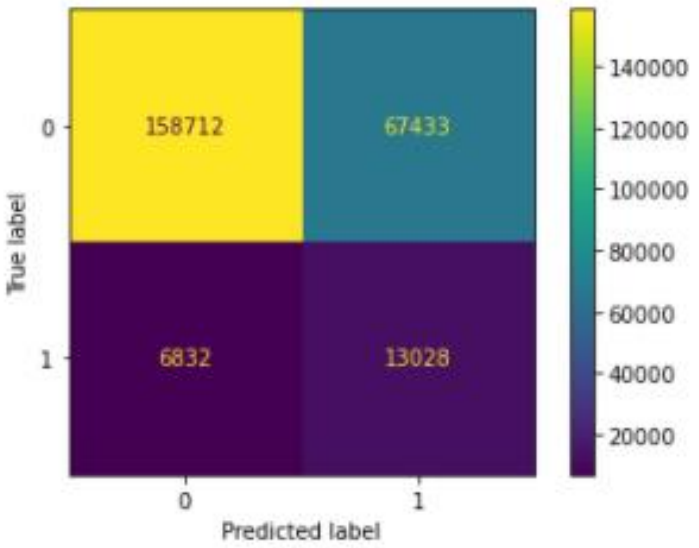
Métrique d'évaluation :

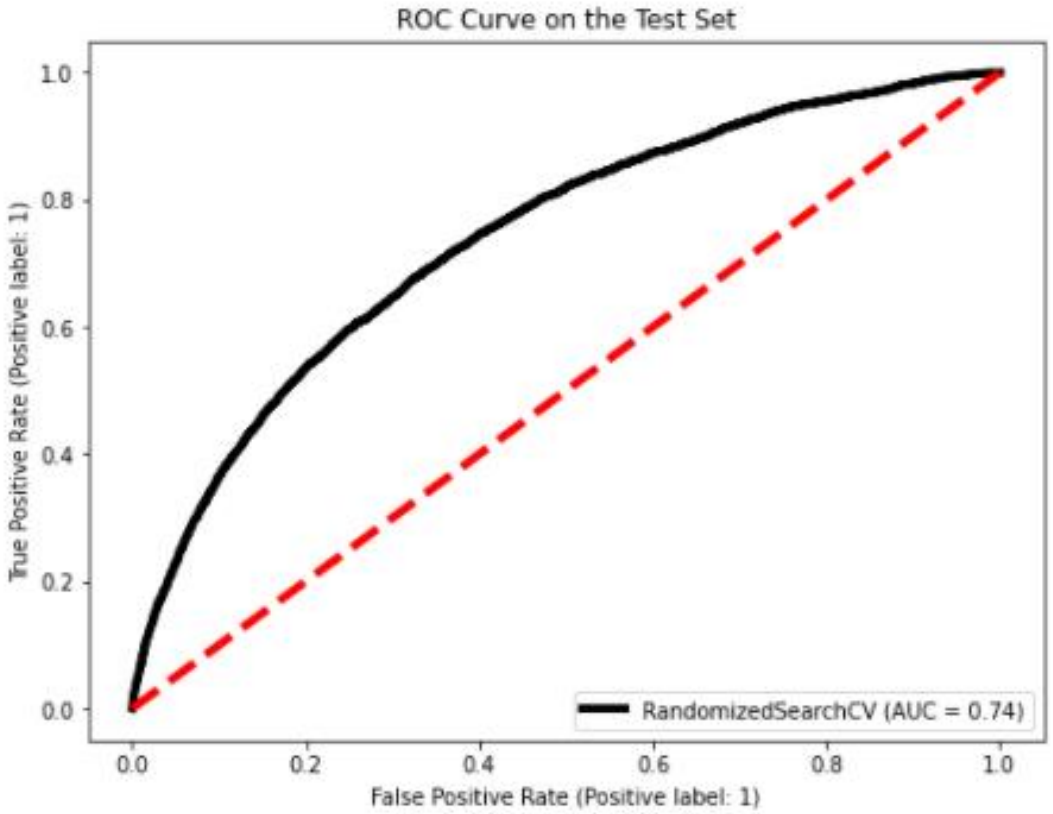
- La métrique dépend de la matrice de confusion (visualisation des prédictions du modèles par rapport aux labels).
- On cherche à connaître les clients non solvables, à maximiser les vrais positifs de notre matrice de confusion .
- Nous cherchons à maximiser le Recall = $\frac{TP}{TP+FN}$ et la précision = $\frac{TP}{TP+FP}$
- Le coût des FN étant supérieur au coût des FP, on cherche surtout à maximiser le Recall.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Métriques d'évaluation

- AUC = 0,74
- Accuracy = 0,68
- Précision = 0,16
- Recall = 0,68
- F1 = 0,26
- Ici nous avons un Recall élevé et une precision faible indique que la classe positive est bien détecté mais inclus des observations de la classe négative.





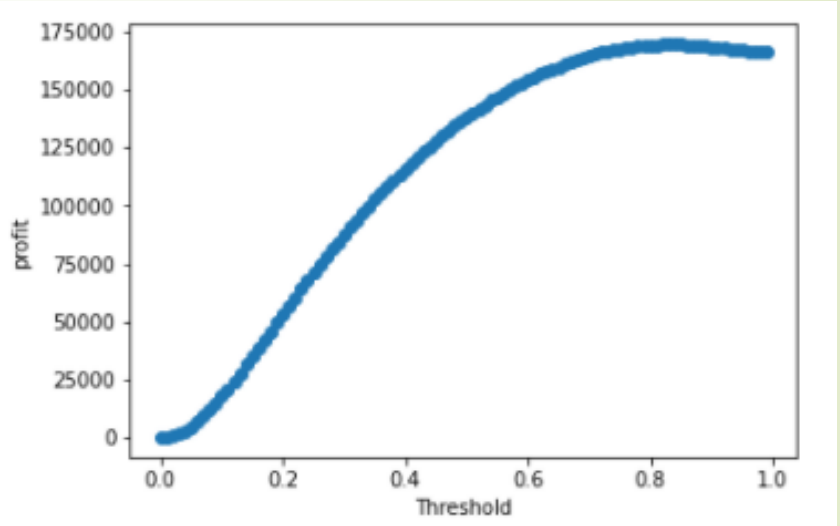
La fonction coût

- Accorder un crédit à un client non solvable et refuser un crédit à un client solvable n'a pas le même coût, il faut limiter les pertes financières.
  - Estimer le coût d'un défaut de paiement,
  - estimer le coût de refuser un crédit à un client solvable,
  - Estimer le gain d'accorder un crédit à un client solvable,
  - estimerle gain de refuser un crédit à un client non solvable.
- Création d'une fonction profit (- fonction coût) rendant compte de tout cela:

$$J = C_0 * TN + C_1 * FN + C_2 * FP + C_3 * TP$$

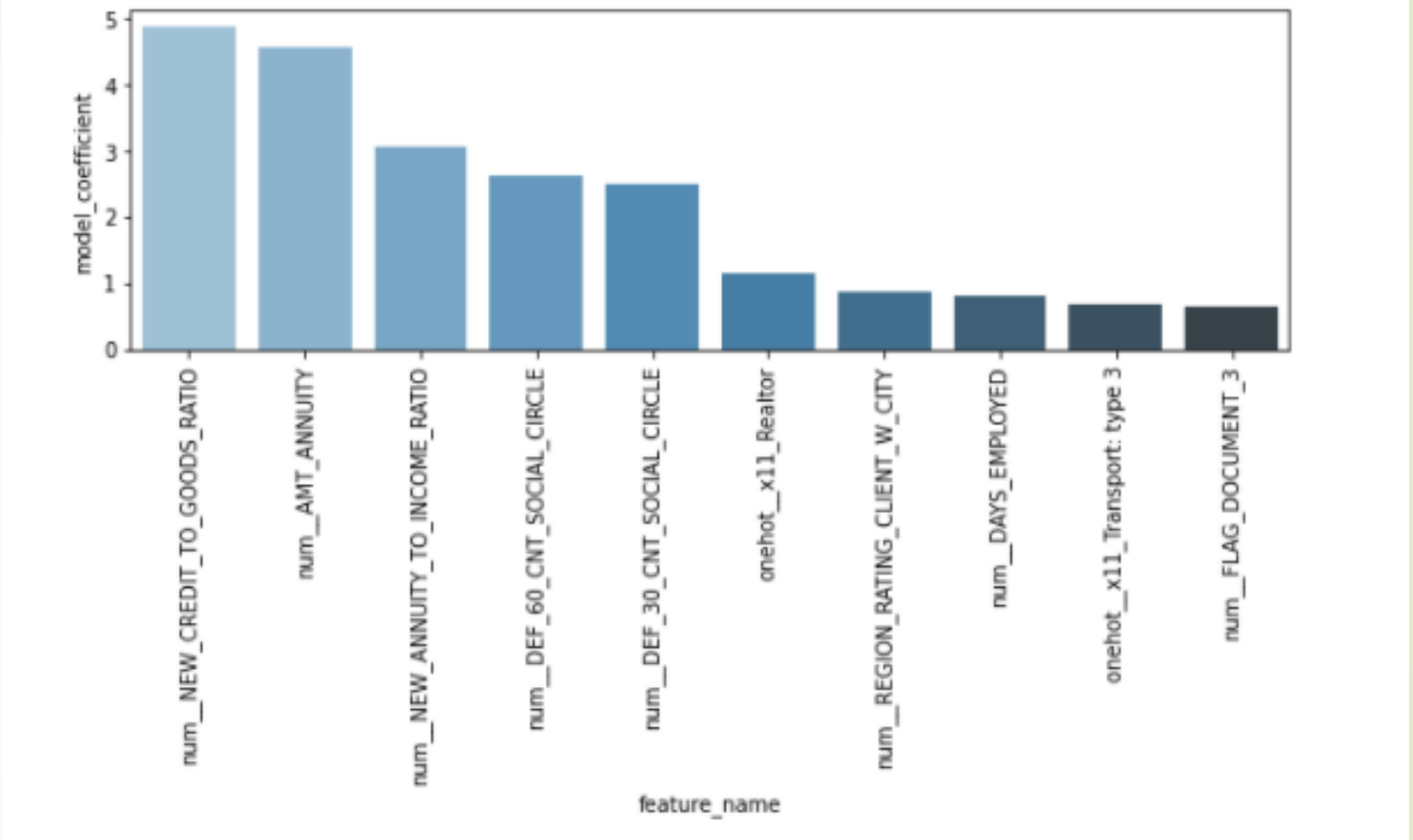
Prenons ici comme valeurs arbitraires $C_0 = 1,\ C_1 = -3,\ C_2 = C_3 = 0$.

On obtient un seuil optimal de 0,84.



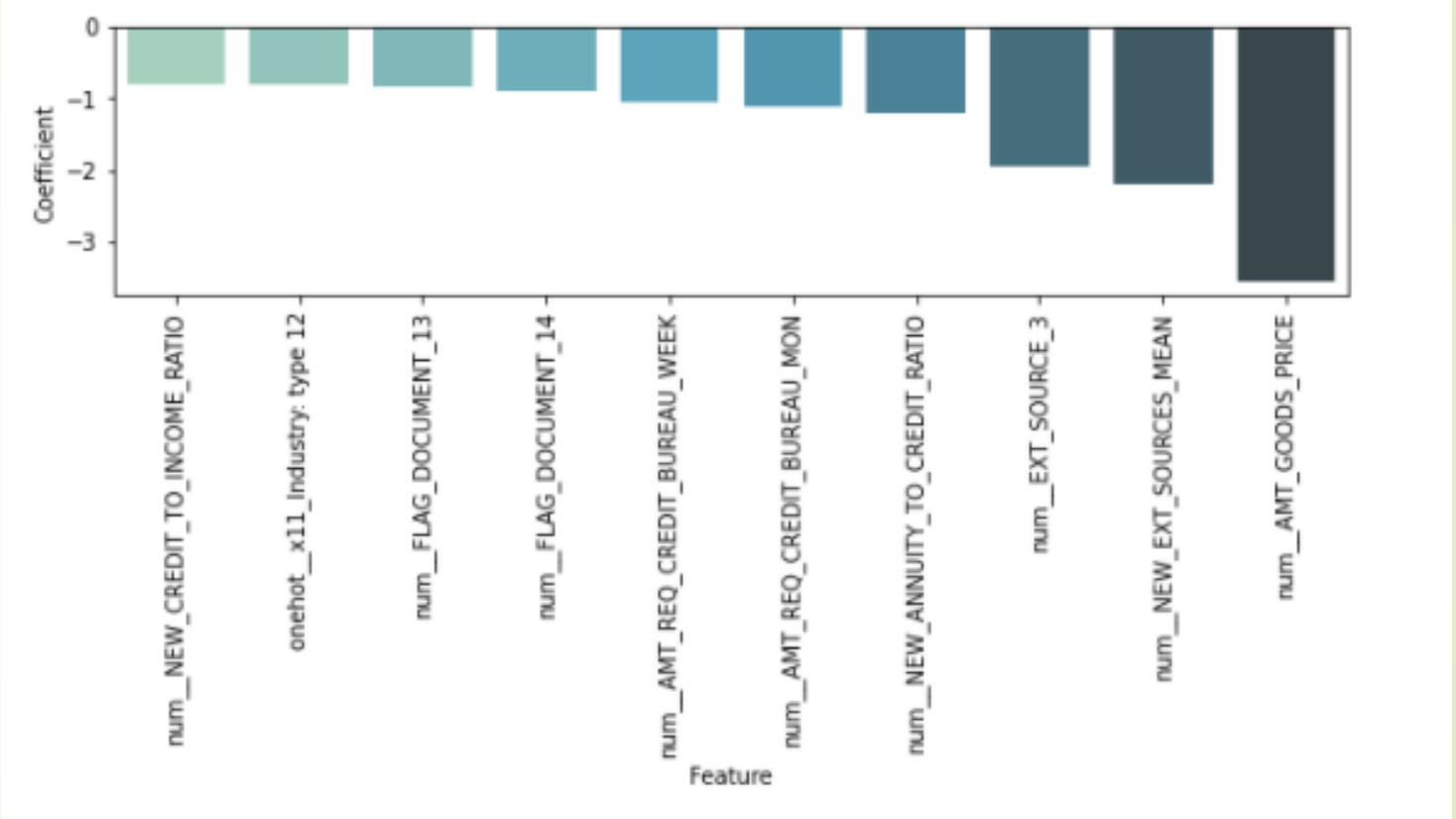
Importance des features

- Regardons les 10 coefficients positifs qui ont un large impact sur la régression logistique.
- Nous voyons que, par exemple, plus le ratio new_credit_to_goods_ratio est grand plus le client a des chances d'être non solvabilité du client.



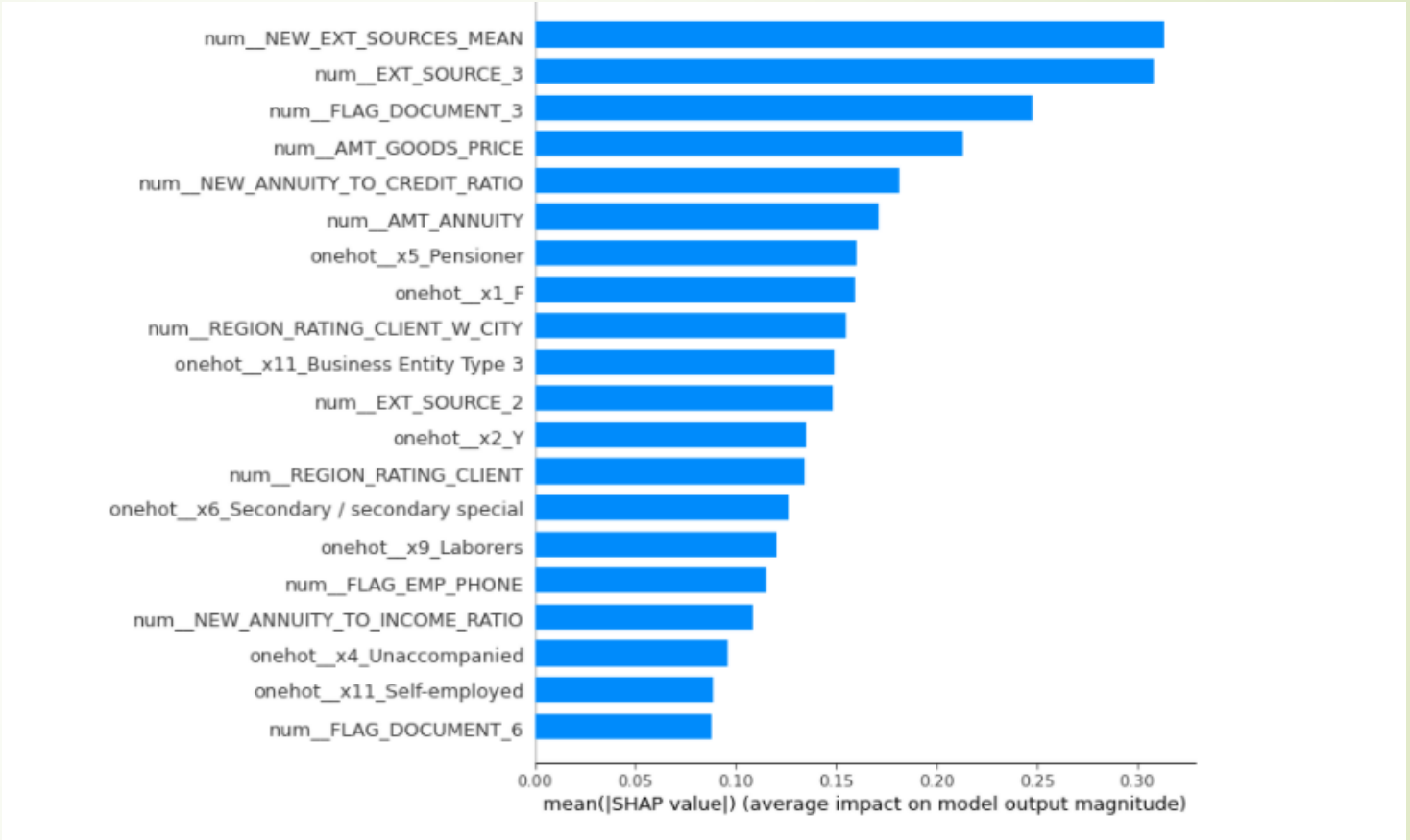
Features importance

- Regardons les 10 coefficients négatifs de la regression logistique.

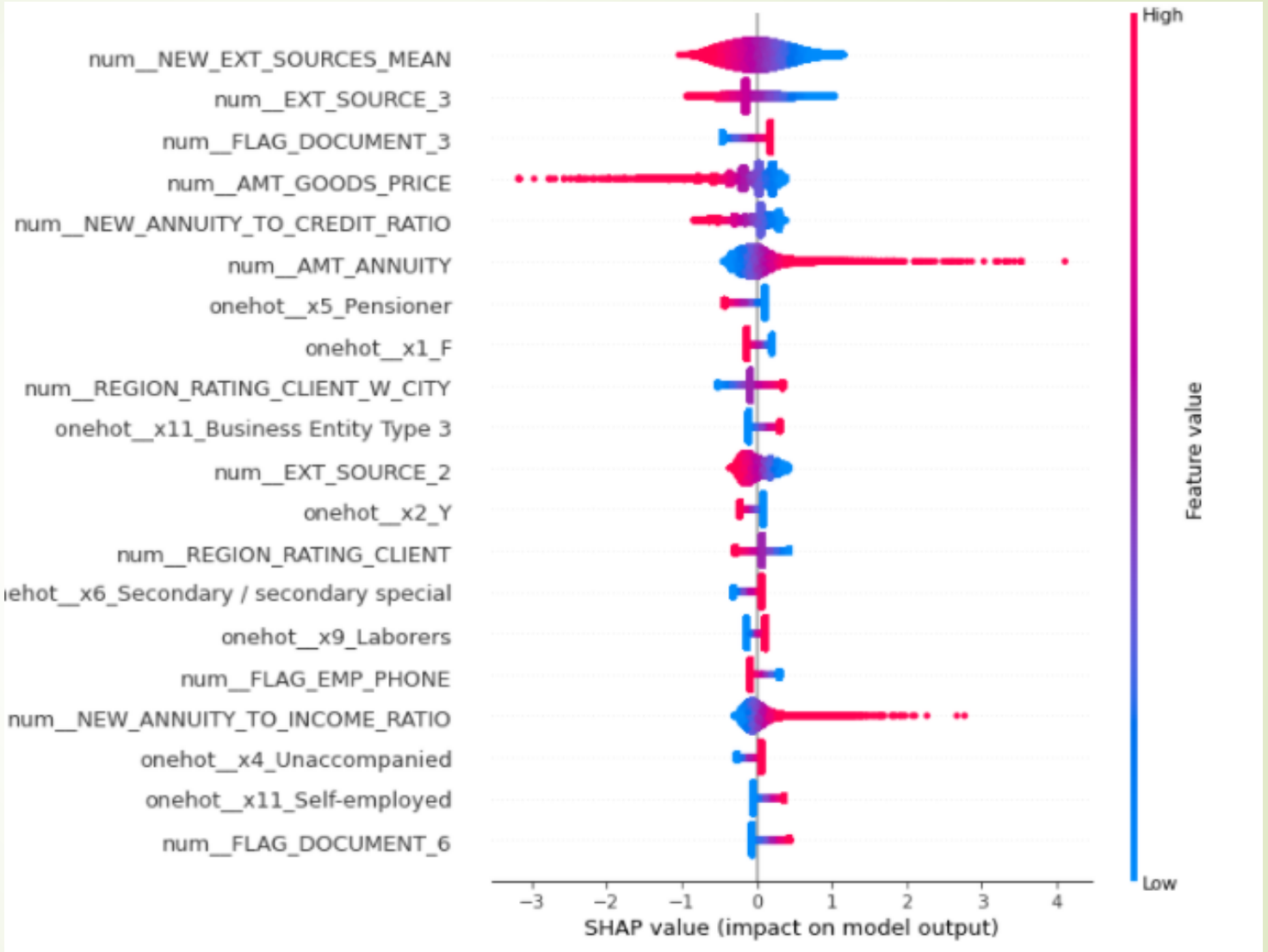


Features importance du point de vue de shap

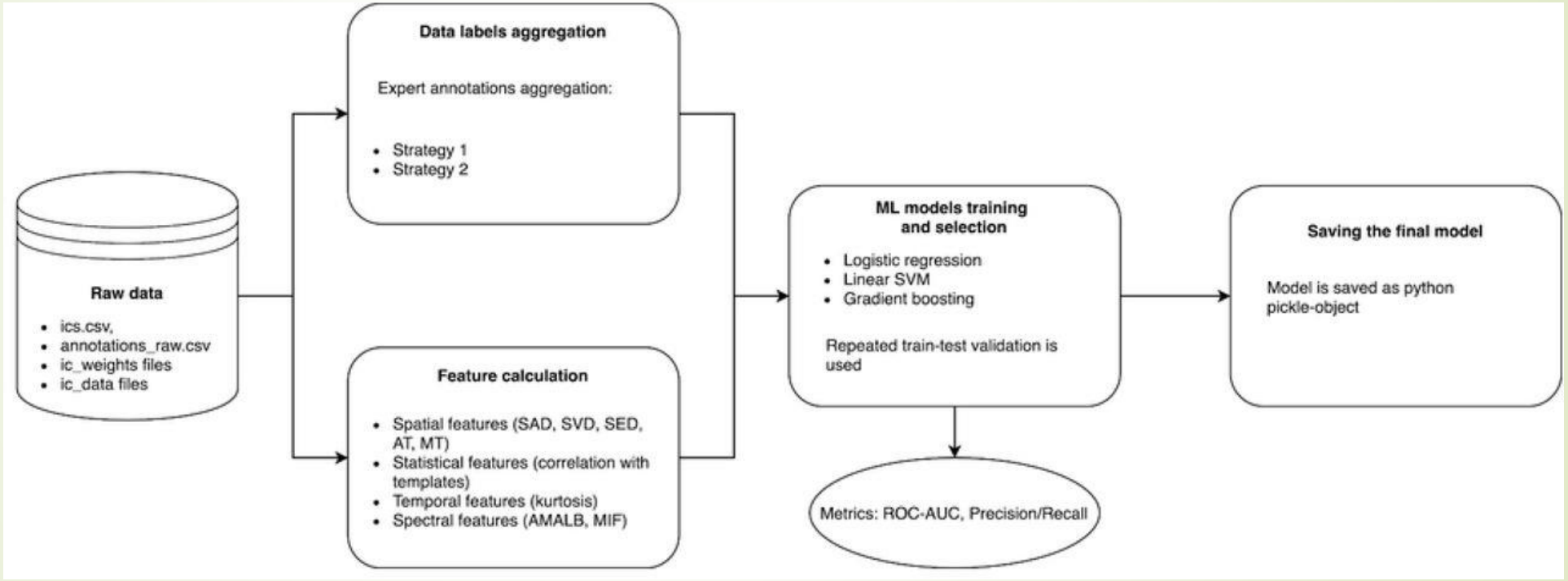
- Nous remarquons que les features avec une valeur large de shapley sont importantes.
- Ici ce sont les ressources extérieures qui ont le plus d'importance.



Features importance du point de vue de shap



Du modèle au dashboard



Dashboard

- Streamlit est une librairie open-source utilisée pour créer des applications web de machine learning. Streamlit permet de créer de belles applications web sans écrire du code HTML. Ce framework permet aussi d'avoir des applications performantes grâce à la mise en cache via une annotation.
- Heroku est une plateforme en tant que service basé sur le cloud

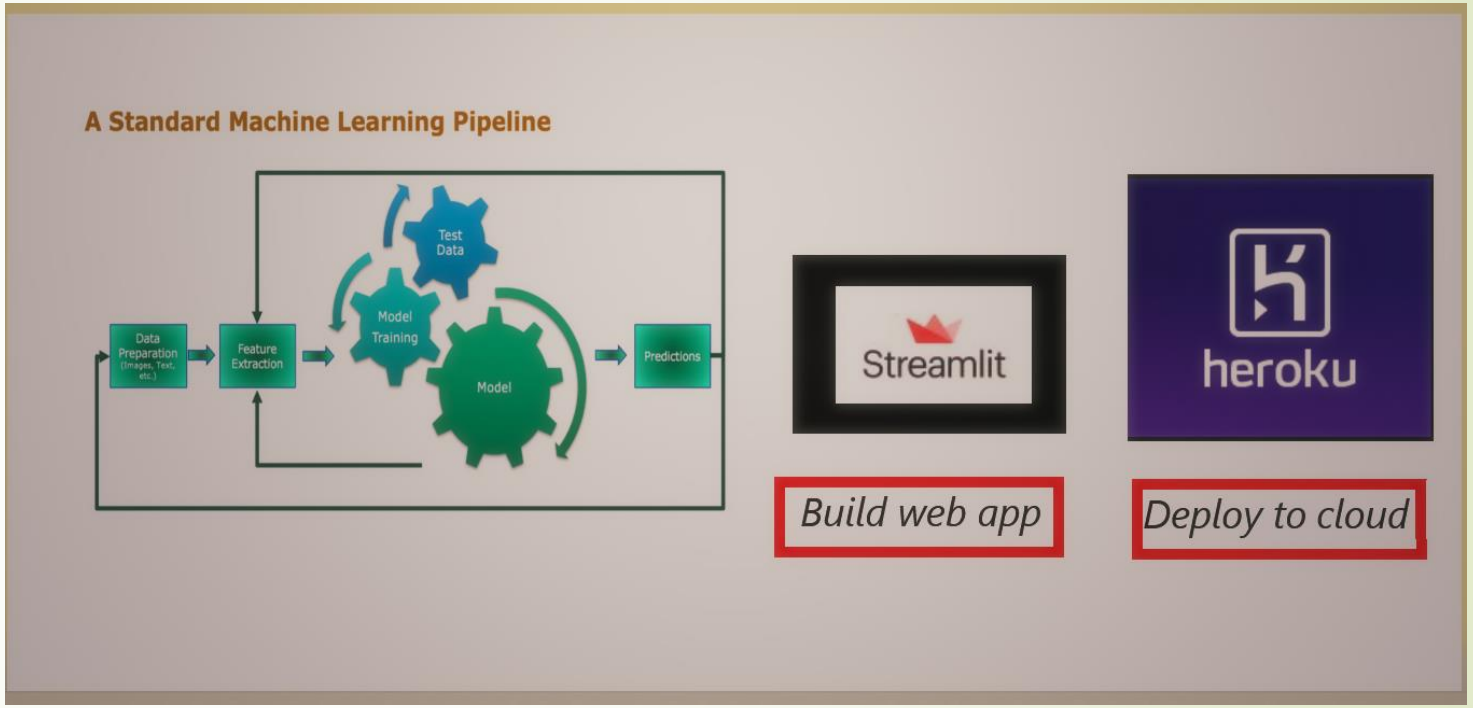
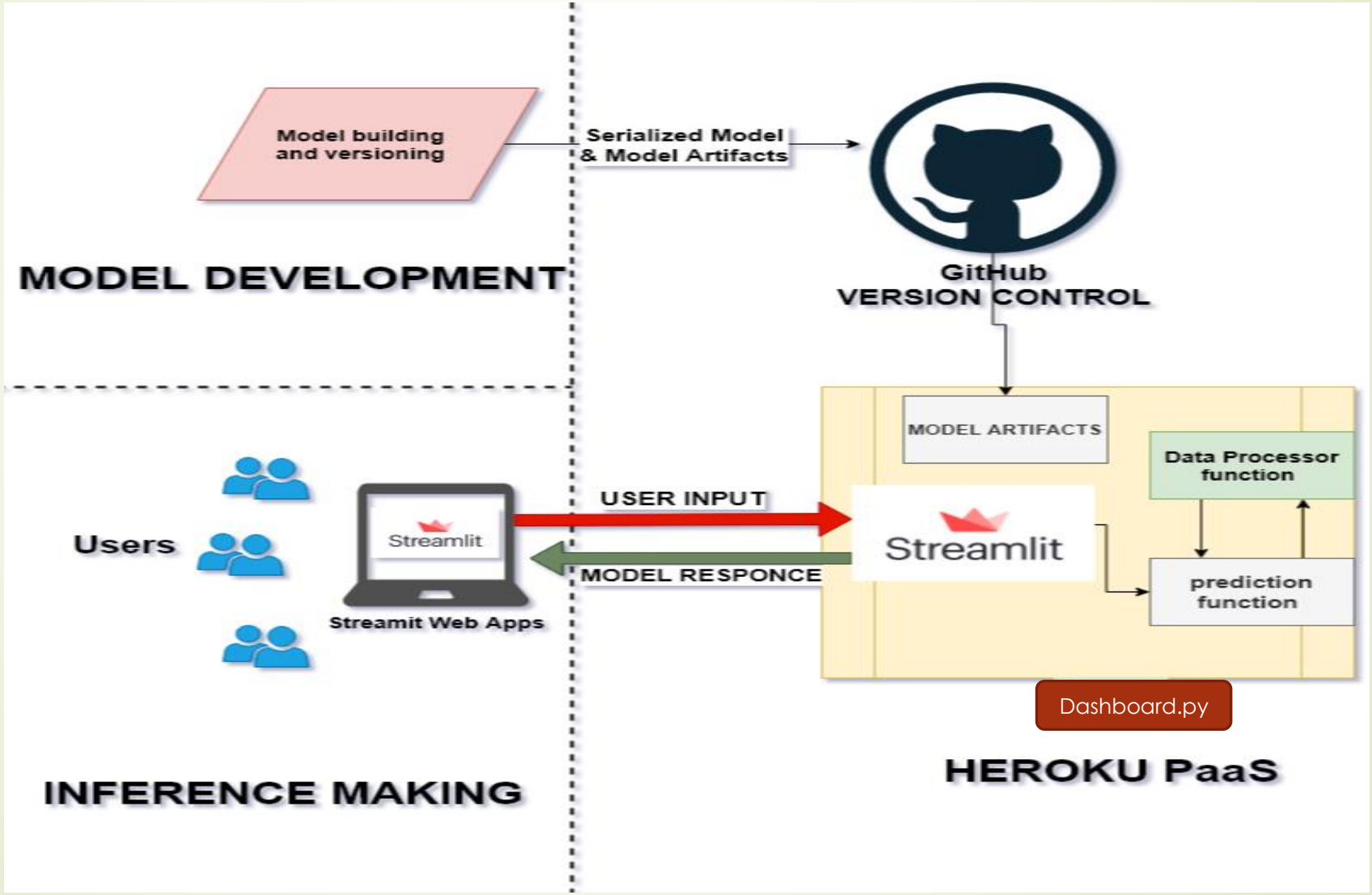
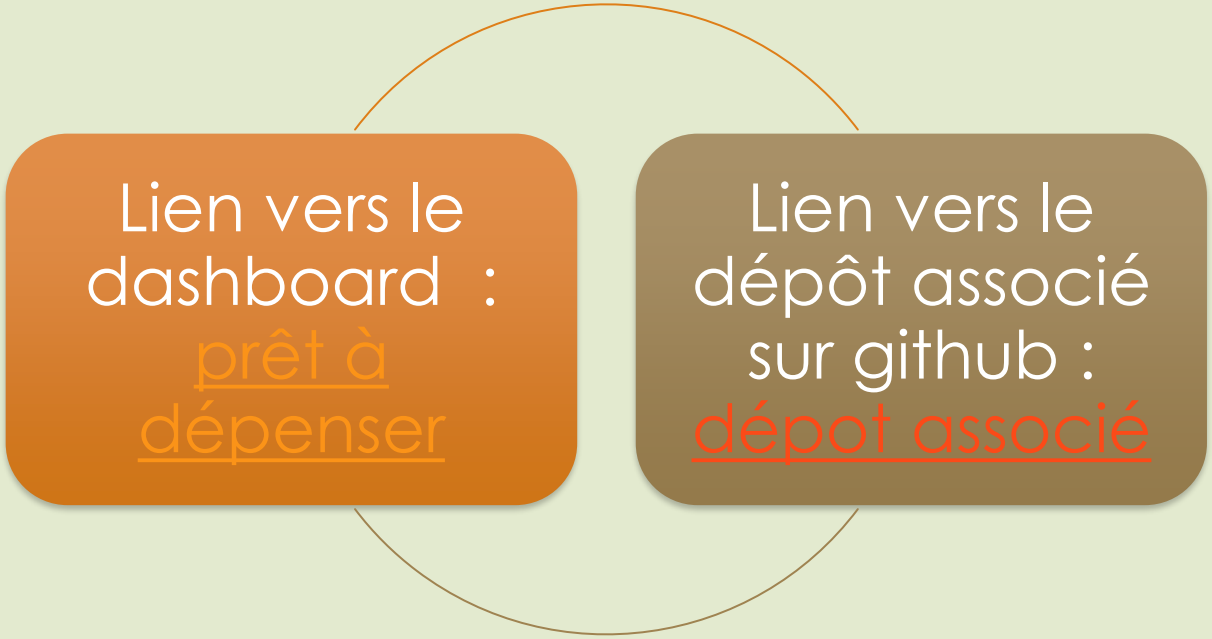


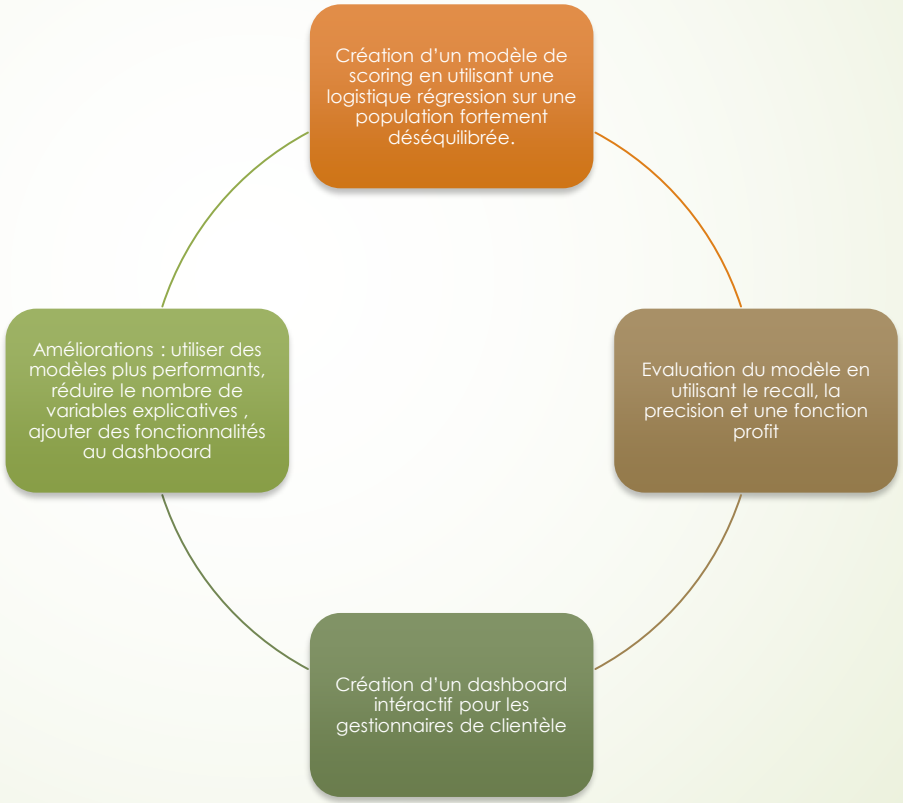
Schéma de l'application



Le dashboard



Conclusion


Création d'un modèle de scoring en utilisant une logistique régression sur une population fortement déséquilibrée.
Evaluation du modèle en utilisant le recall, la precision et une fonction profit
Création d'un dashboard intéractif pour les gestionnaires de clientèle
Améliorations : utiliser des modèles plus performants, réduire le nombre de variables explicatives , ajouter des fonctionnalités au dashboard