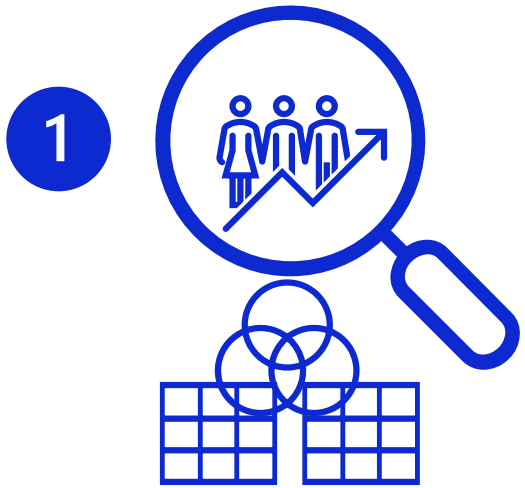
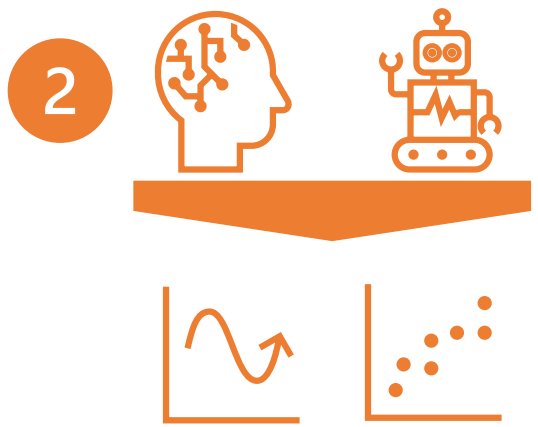
Implémentez un modèle de scoring

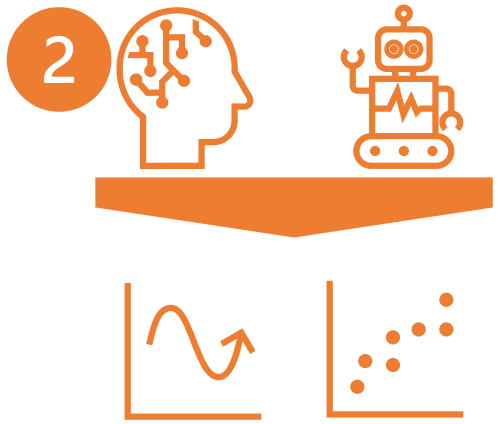
Prêt à dépenser





Analyse exploratoire

- Construction DataFrame
  - Acquisition données
  - Nettoyage &
  - Feature Engineering
- Analyse des distributions



Entrainement / Sélection modèle

- Validation croisée + SMOTE
- Fonction coût métier
- Optimisation du seuil
- Tests des modèles et sélection

Dashboard & déploiement

- Interprétation globale / locale
- Serveurs Flask / Streamlit
- Hébergement Cloud AWS
- Limites & préconisations

1 Analyse exploratoire

Chargement données :
8 fichiers .csv

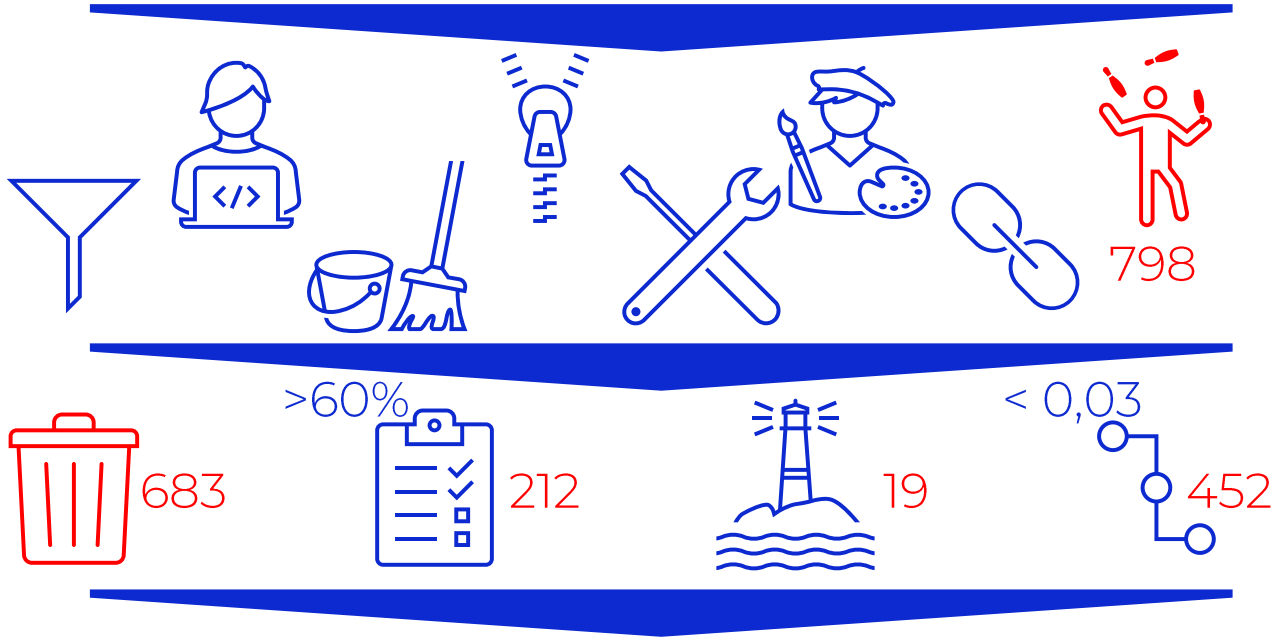
3 clés :
SK_ID_CURR (n° demande)
SK_ID_PREV
SK_ID_BUREAU

Kernel Kaggle Dataset cible :
https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script

application_{train|test}.csv
bureau.csv
bureau_balance.csv
POS_CASH_balance.csv
credit_card_balance.csv
previous_application.csv
installments_payments.csv
HomeCredit_columns_description.csv

Construction DataFrame

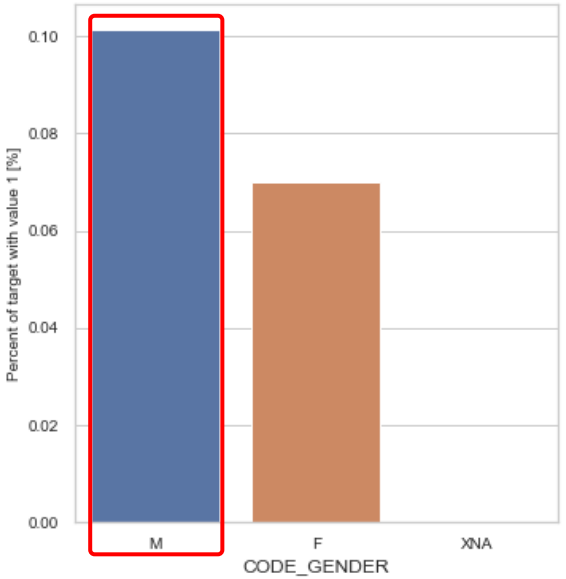
	Rows	Columns	%NaN	%Duplicate	Object _dtype	Float _dtype	Int _dtype	Bool _dtype	MB _Memory
./data\application_test.csv	48744	121	23.81	0.0	16	65	40	0	44.998
./data\application_train.csv	307511	122	24.40	0.0	16	65	41	0	286.227
./data\bureau.csv	1716428	17	13.50	0.0	3	8	6	0	222.620
./data\bureau_balance.csv	27299925	3	0.00	0.0	1	0	2	0	624.846
./data\credit_card_balance.csv	3840312	23	6.65	0.0	1	15	7	0	673.883
./data\HomeCredit_columns_description.csv	219	4	15.18	0.0	4	0	0	0	0.008
./data\installments_payments.csv	13605401	8	0.01	0.0	0	5	3	0	830.408
./data\POS_CASH_balance.csv	10001358	8	0.07	0.0	1	2	5	0	610.435
./data\previous_application.csv	1670214	37	17.98	0.0	16	15	6	0	471.481
./data\sample_submission.csv	48744	2	0.00	0.0	0	1	1	0	0.744

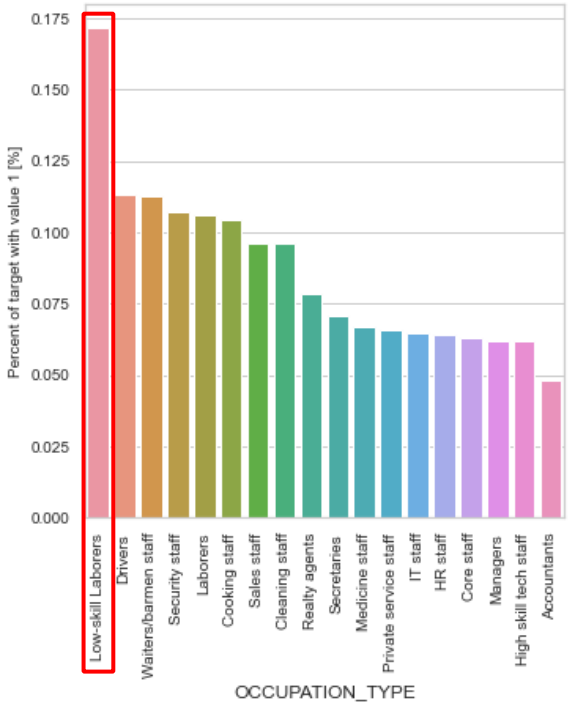


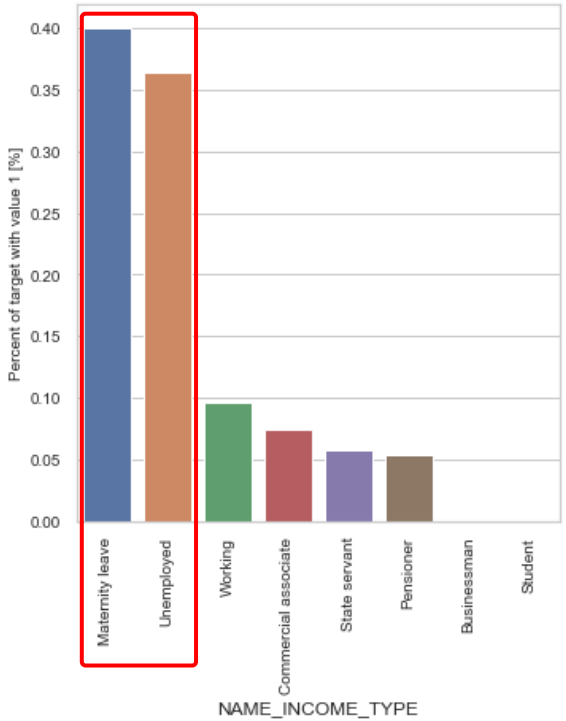
DataFrame final (train + test)
356 251 individus x 115 features

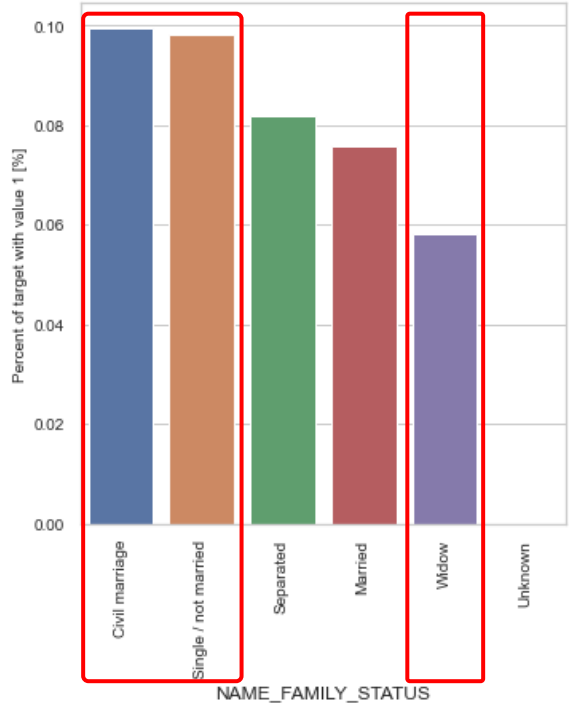
1 Analyse exploratoire

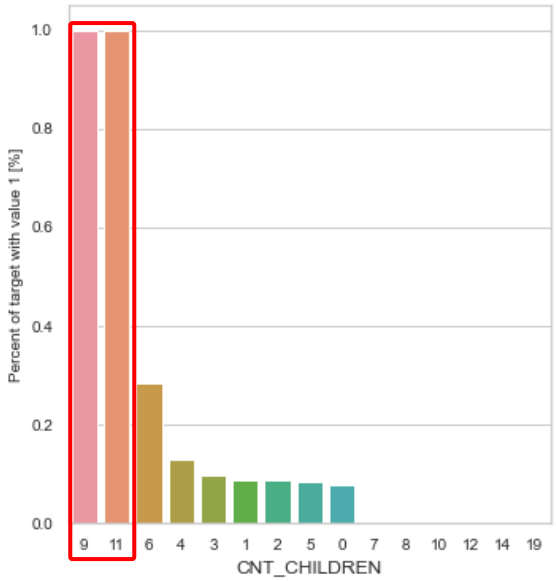
Analyse des distributions

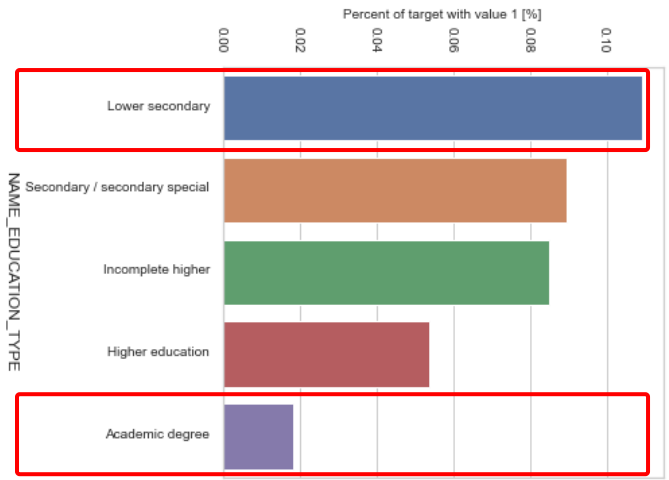


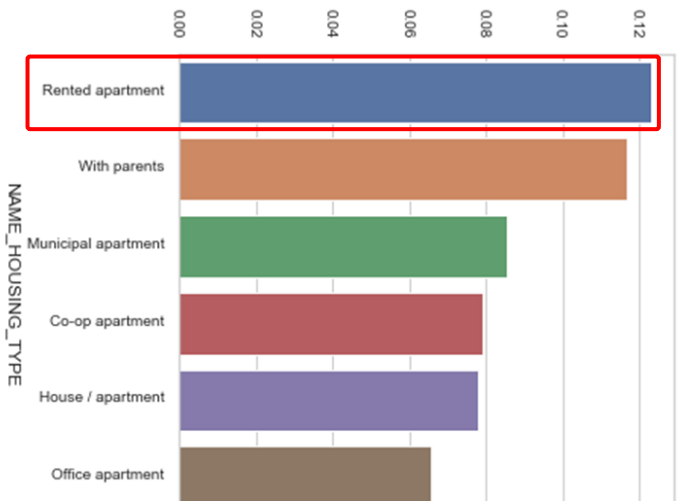




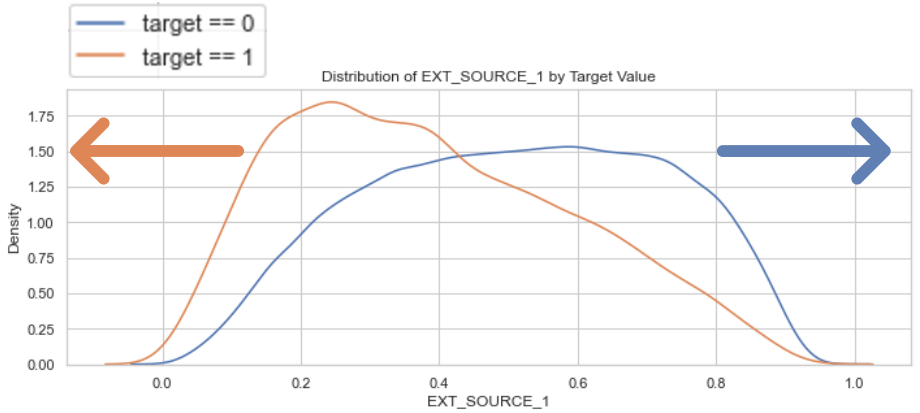


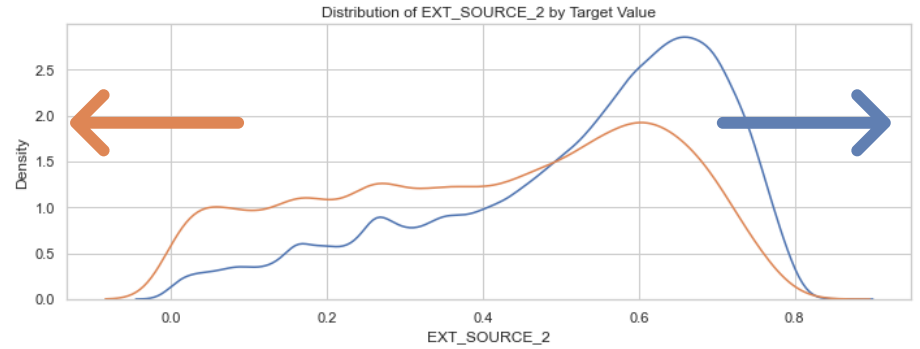


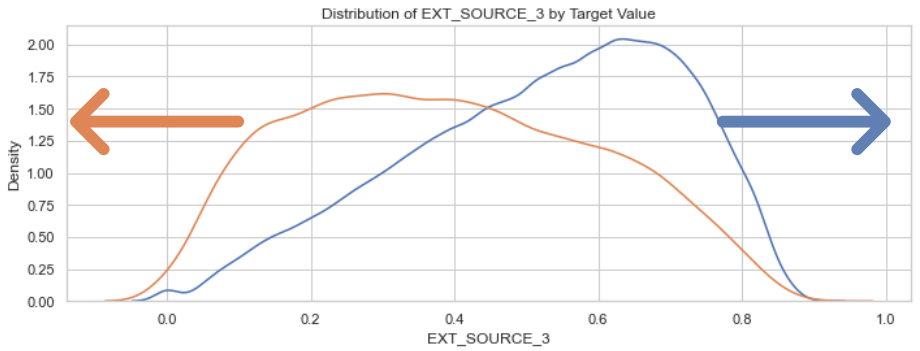




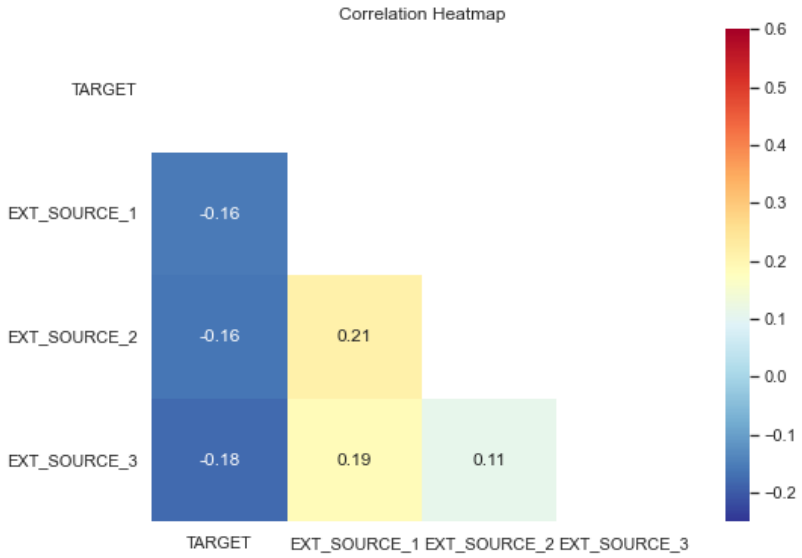
1 Analyse exploratoire

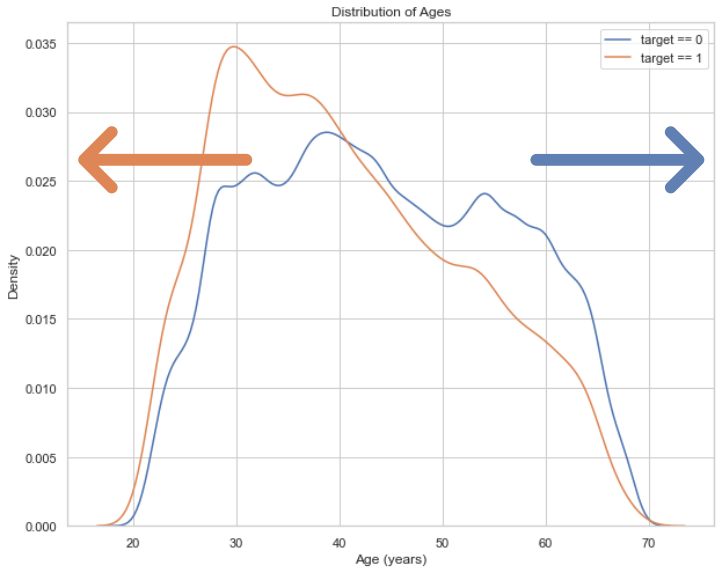






Analyse des distributions





2 Entrainement / Sélection modèle

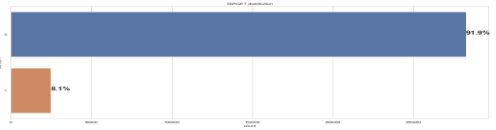
Validation croisée + SMOTE

Outliers

- remplacer outliers par des Nan

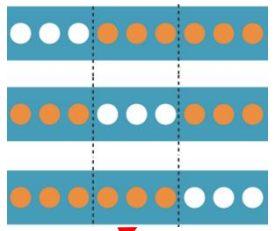
Echantillonnage Stratifié

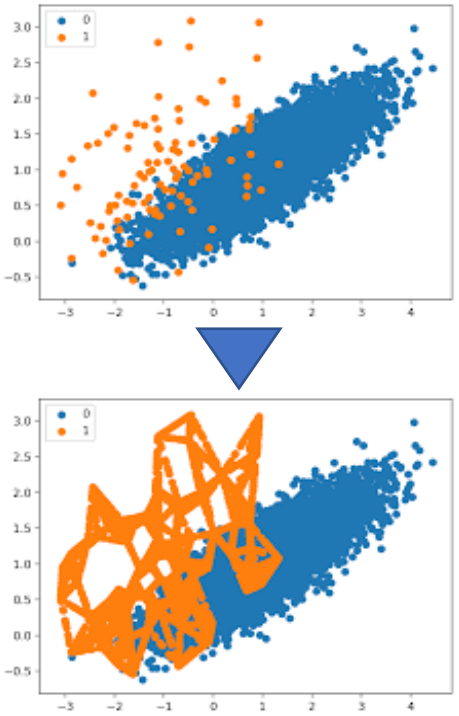
- respecter proportions « TARGET »



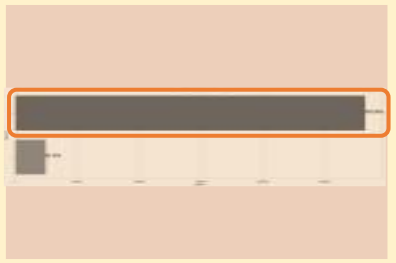
GridSearchCV

- Modèle : Pipeline [Imputer (moyenne) + Smote + Scaler + Classifier]
- Scoring : Custom Score, ROC_AUC, PR_AUC, Time
- Splits : 3 KFolds

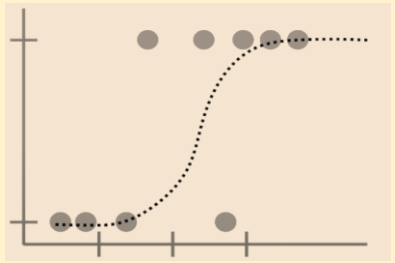




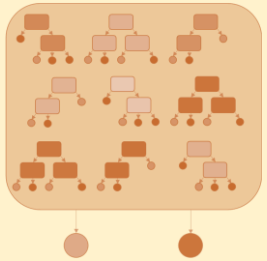
6 modèles testés :



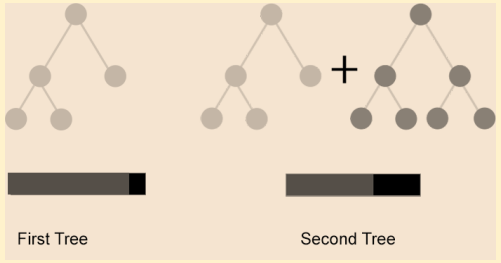
Dummy (moy)



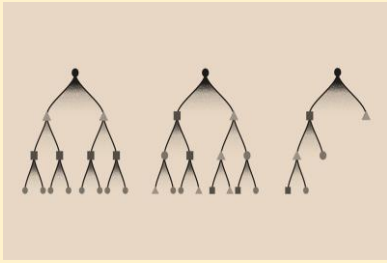
Logistic Regression



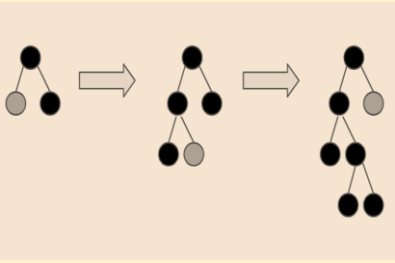
Random Forest



Cat Boost



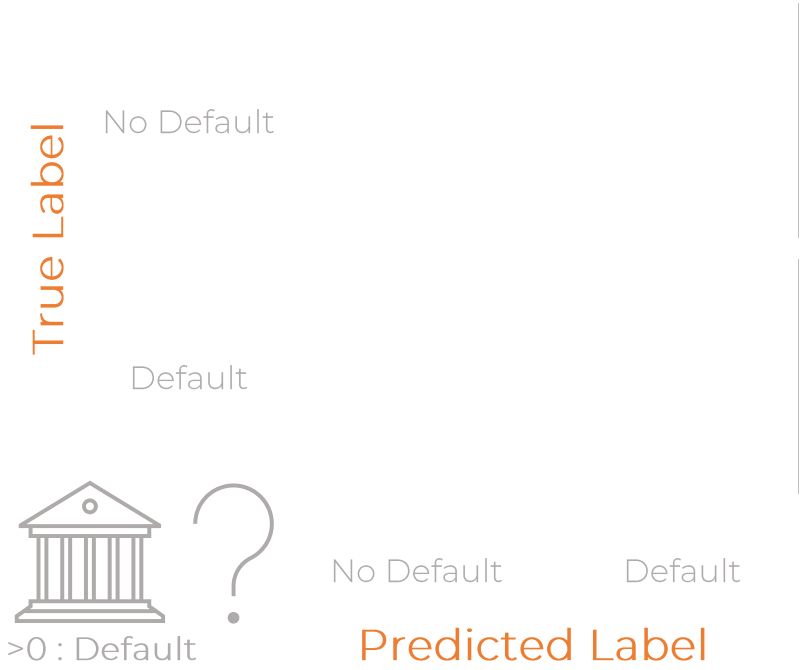
XG Boost

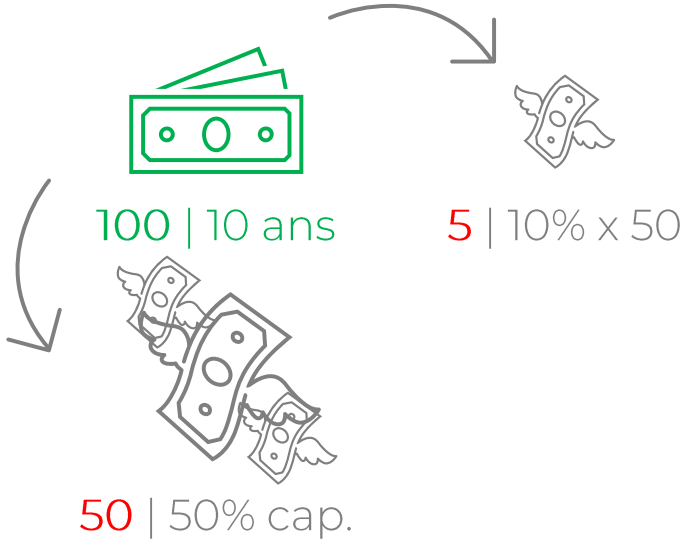


Light GBM

2 Entrainement / Sélection modèle

Fonction coût métier





10 FN pour 1 FP !

Minimiser fonction : $\frac{10\ FN + FP}{\text{Taille échantillon}}$ (entre 0 et 1)

2 Entrainement / Sélection modèle

Optimisation du seuil

y_pred

0,719	0,757	0,431	0,711
0,366	0,666	0,879	0,861
0,241	0,293	0,455	0,286
0,853	0,531	0,079	0,284
0,846	0,357	0,433	0,493
0,114	0,321	0,015	0,990
0,567	0,261	0,088	0,084

y_pred_proba

- best thresh = 0
- best score = 1

Step 1

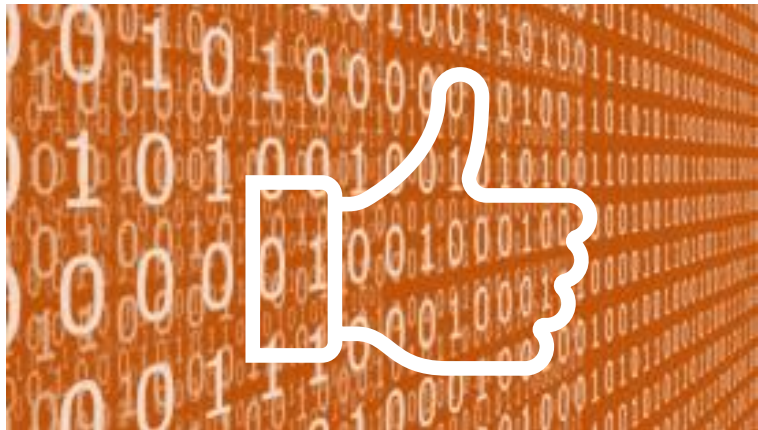
Loop : thresh Є (0, 1, 100 pas)

- score = metric(y_train, y_pred_proba > thresh)
- si score < best_score
  - best_thresh = thresh
  - best_score = score

- best thresh = 0,101010
- best score = 0,534
- y_pred = y_pred_proba > best_thresh

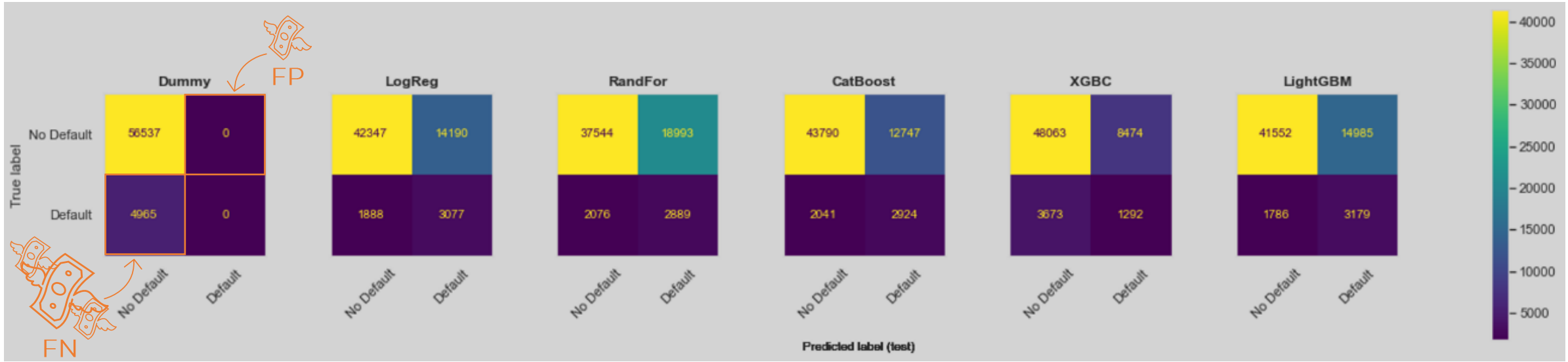
Final Step

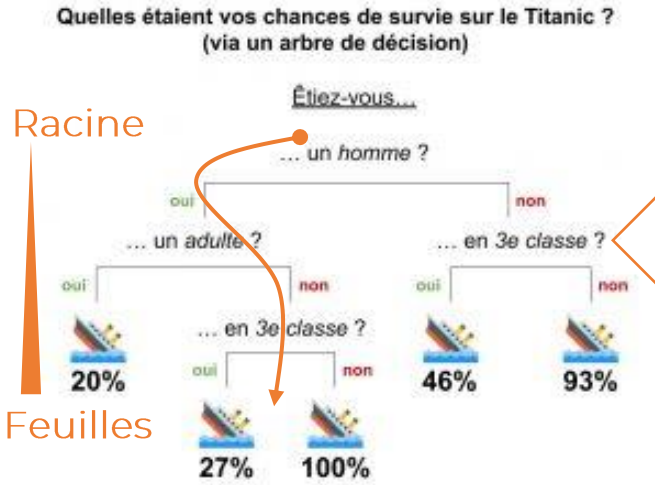
y_pred



2 Entrainement / Sélection modèle

Sélection des modèles



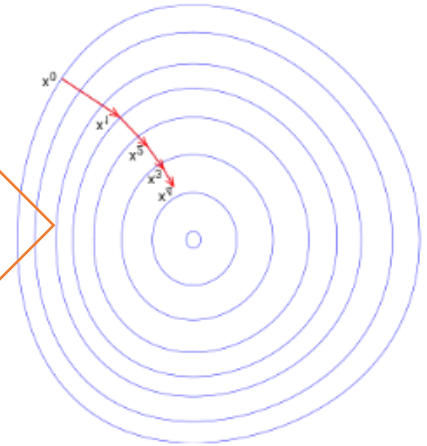


- Arbres de décision
- Départ = Racine
- Probabilités = Feuilles

Model	Custom _score	Time
LightGBM	0.534	1752.54
LogReg	0.5377	382.99
CatBoost	0.5391	415.61
RandFor	0.6464	408.85
XGBC	0.735	6966.31
Dummy	0.8073	0.009

Light GBM

- Gradient Boosting Machine
- plusieurs apprenants « faibles »
- 1 arbre = descente de gradient



3 Dashboard & déploiement

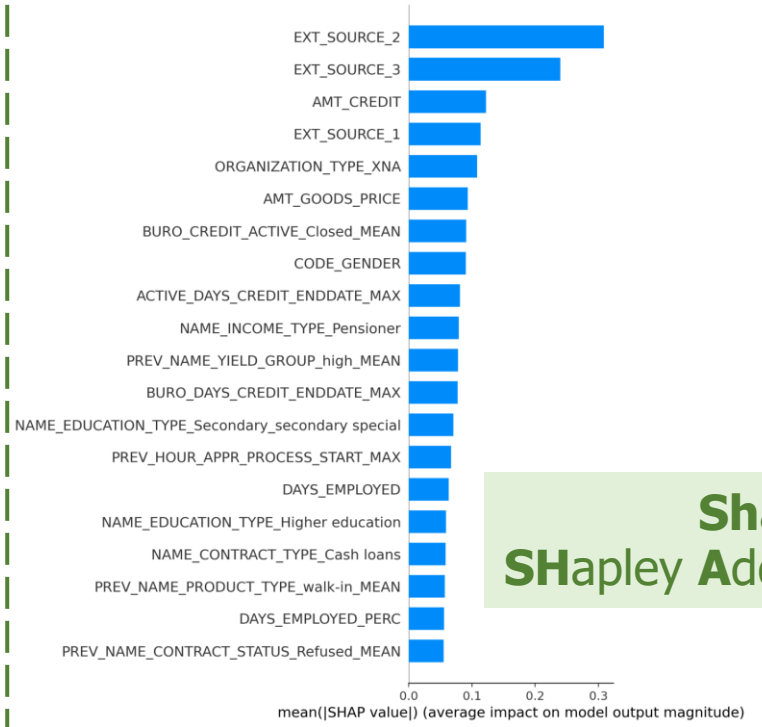
http://15.188.141.153:8501/

Bank

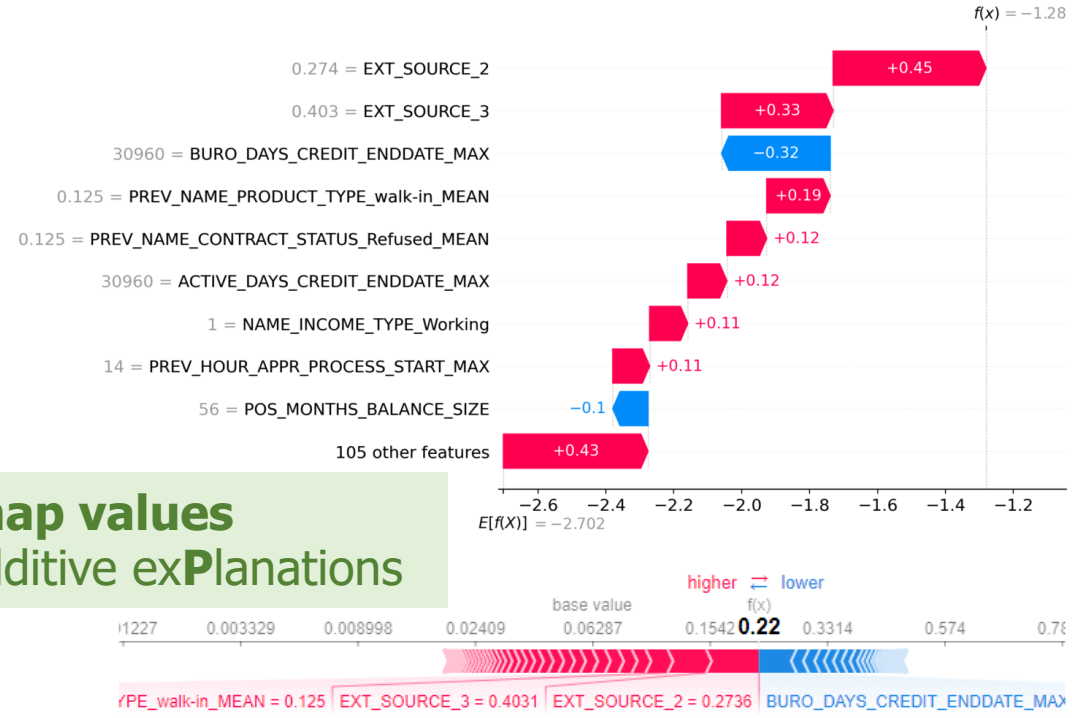
Data Customer 345565

Val / Var Mean : EXT_SOURCE_2 0.2736 ▼0	Val / Var Mean : EXT_SOURCE_3 0.403 ▼0
Val / Var Mean : AMT_CREDIT 619k ▲14210	Val / Var Mean : EXT_SOURCE_1 0.506 ▲0
Val / Var Mean : ORGANIZATION_TYPE_XNA 0 ▼0	Val / Var Mean : AMT_GOODS_PRICE 554k ▲10029
Val / Var Mean : BURO_CREDIT_ACTIVE_Closed_MEAN 0.556 ▼0	Val / Var Mean : CODE_GENDER 1 ▲0
Val / Var Mean : ACTIVE_DAYS_CREDIT_ENDDATE_MAX 31k ▲27457	Val / Var Mean : NAME_INCOME_TYPE_Pensioner 0 ▼0

Interprétation globale / locale



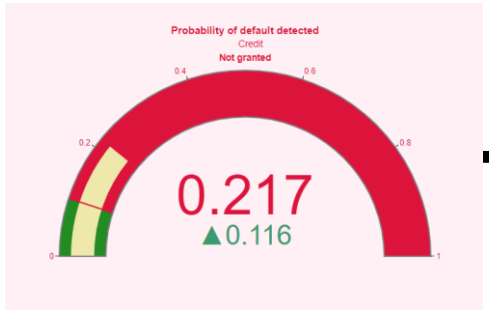
Shap values
SHapley **A**dditive ex**P**lanations



Decision Model

BLACK BOX

Credit : Not Granted

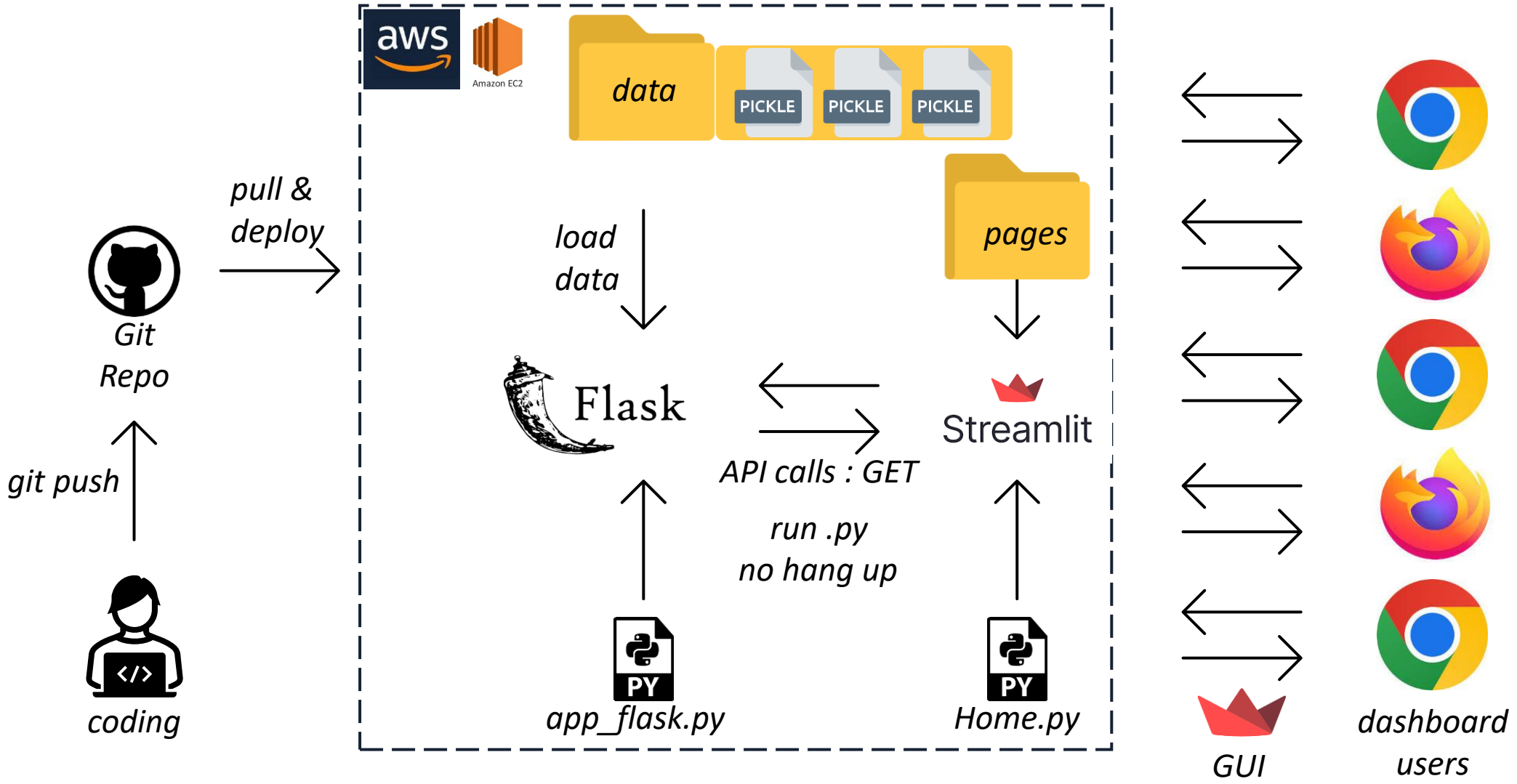


Customer 345565

3 Dashboard & déploiement

Serveurs Flask / Streamlit

http://15.188.141.153:8501/



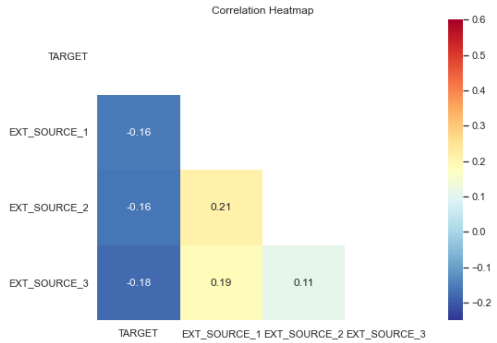
3 Dashboard & déploiement

Limites & préconisations

$$\frac{10\ FN + FP}{\text{Taille échantillon}}$$

1 | fonction coût métier

2 | détail 3 sources externes



3 | pré-traitement des données

kaggle

https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script

http://15.188.141.153:8501/

Questions ?

Merci !