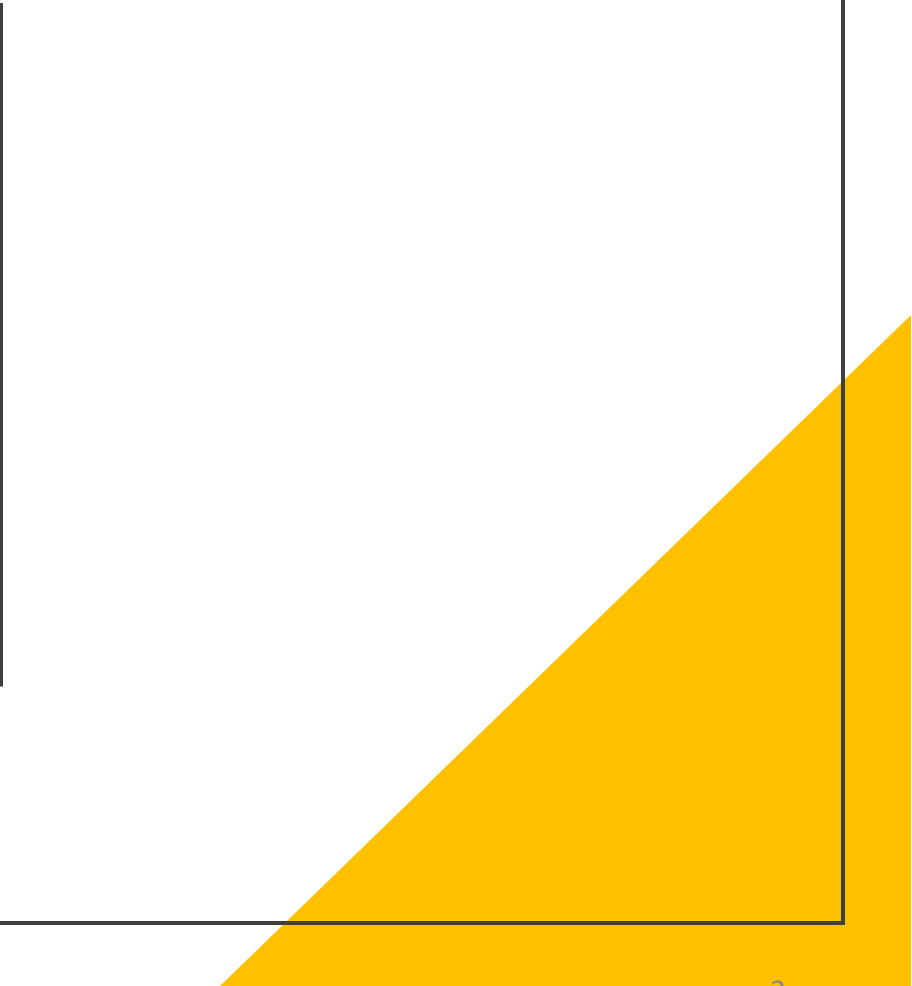


Projet 7 : Implémentez un modèle de scoring

Fanjamalala Rajaonalison
01/2022

Présentation de l'étude

A vertical line is positioned to the right of the title. In the bottom right corner, there is a yellow triangle pointing upwards and to the left, with its hypotenuse forming a diagonal line across the corner.

Contexte

Entreprise
"Prêt à dépenser"

Crédit à la consommation



avec peu ou pas
d'historique de prêt

Problématiques :



Comment classifier
les demandes?



Comment expliquer la
décision aux clients?



Objectifs :



Modèle de scoring de la
probabilité de défaut de
paiement



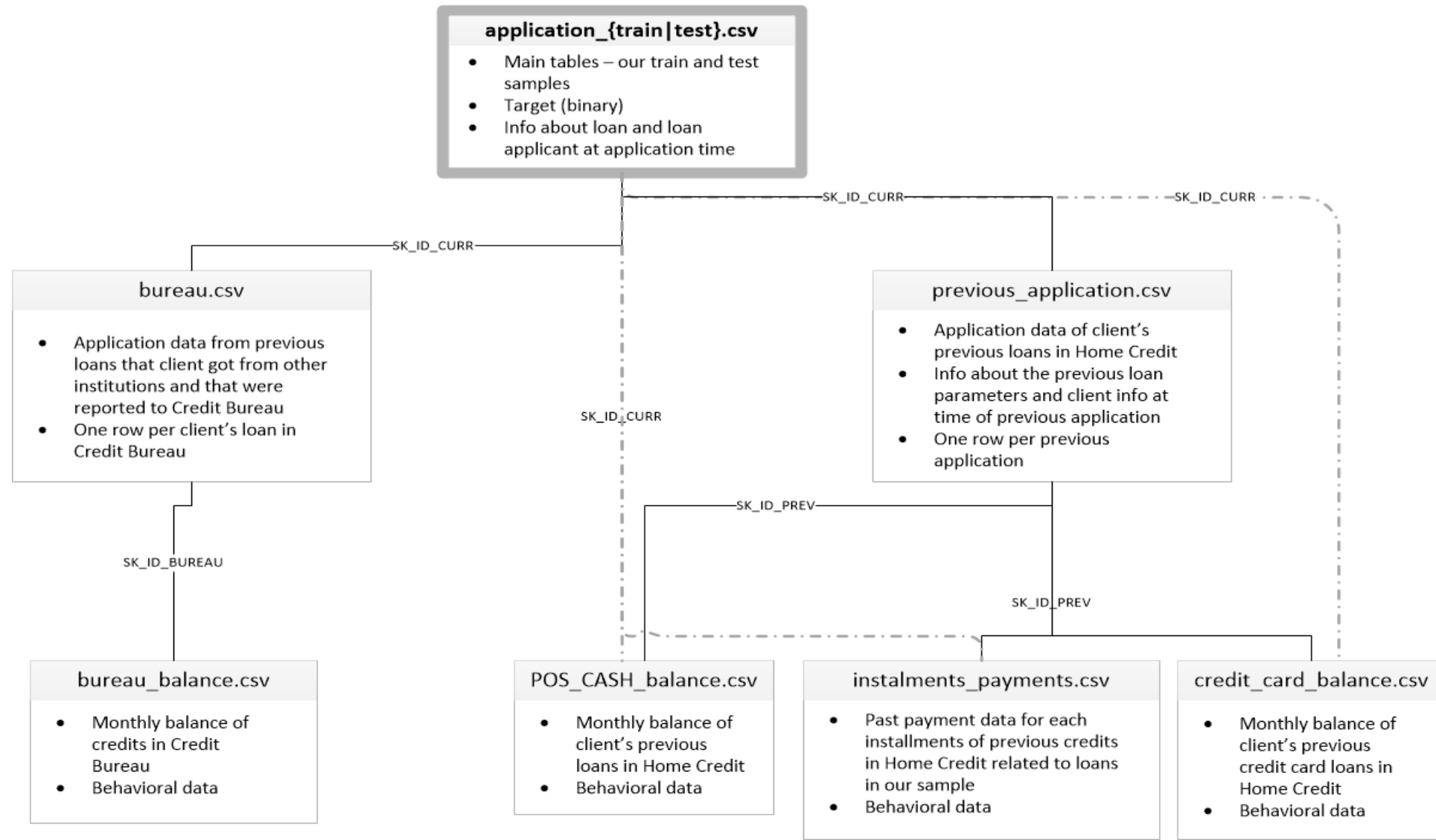
Dashboard interactif



Dataset ([source: https://www.kaggle.com/c/home-credit-default-risk](https://www.kaggle.com/c/home-credit-default-risk))

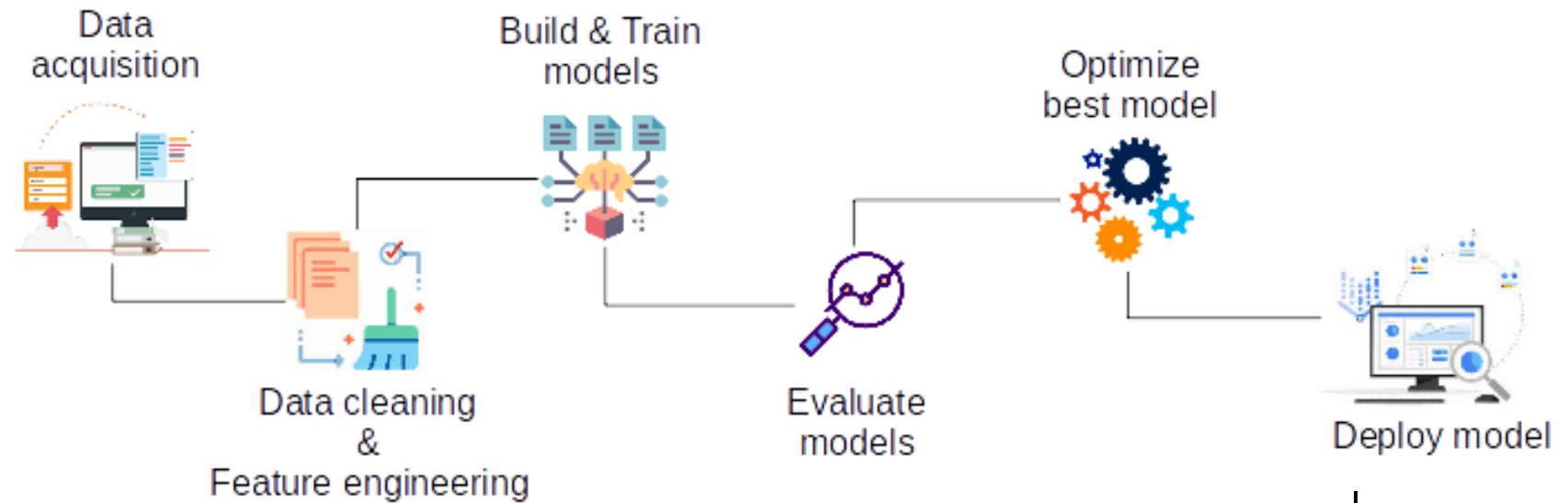
Données variées :

- données comportementales
- données provenant d'autres institutions financières
- etc.

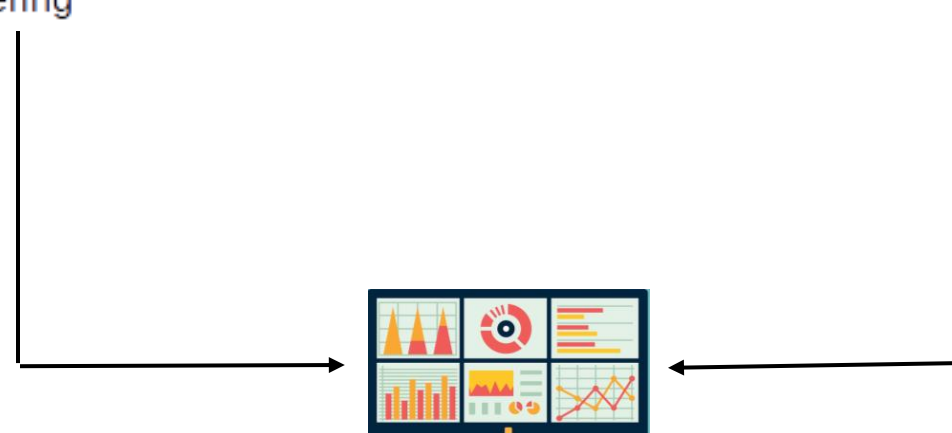


Méthodologie

1. Construction du modèle de scoring



2. Création du dashboard



Acquisition Exploration des données



Description des données

(*source* : <https://www.kaggle.com/c/home-credit-default-risk>)

	Rows	Columns	%NaN	%Duplicate	object_dtype	float_dtype	int_dtype	bool_dtype	MB_Memory
../Dataset/sample_submission.csv	48744	2	0.00	0.0	0	1	1	0	0.744
../Dataset/bureau.csv	1716428	17	13.50	0.0	3	8	6	0	222.620
../Dataset/HomeCredit_columns_description.csv	219	5	12.15	0.0	4	0	1	0	0.008
../Dataset/application_train.csv	307511	122	24.40	0.0	16	65	41	0	286.227
../Dataset/credit_card_balance.csv	3840312	23	6.65	0.0	1	15	7	0	673.883
../Dataset/POS_CASH_balance.csv	10001358	8	0.07	0.0	1	2	5	0	610.435
../Dataset/installments_payments.csv	13605401	8	0.01	0.0	0	5	3	0	830.408
../Dataset/previous_application.csv	1670214	37	17.98	0.0	16	15	6	0	471.481
../Dataset/application_test.csv	48744	121	23.81	0.0	16	65	40	0	44.998
../Dataset/bureau_balance.csv	27299925	3	0.00	0.0	1	0	2	0	624.846

Jeu de données principal

Data cleaning

Feature engineering



Preprocessing

Identification / imputation des **valeurs manquantes**

- Numérique → median
- Catégorielle → most_frequent

Suppression des **outliers** / valeurs atypiques

- Ex : 'XNA' dans 'CODE_GENDER'

Encodage des variables catégorielles

Standardisation des données

Feature engineering

Application : Création de features "**Taux**" en divisant certaines features

- Ex : 'PAYMENT_RATE' = 'AMT_ANNUITY' / 'AMT_CREDIT' Division de features → taux

Bureau : Création de features spécifiques pour les **Crédits Actifs** et les **Crédits Fermés**

Autres tables : Création de features avec la **moyenne** des valeurs selon 'SK_ID_CURR'

Train samples: (150000, 122)

test samples: (48744, 121)

Bureau df shape: (34117, 74)

Previous applications df shape: (110071, 187)

Pos-cash balance df shape: (104063, 11)

Installments payments df shape: (58741, 9)

Credit card balance df shape: (64370, 27)

Join
on 'SK_ID_CURR'

(198741, 556)

Feature selection

Tri des features selon leur importance :

- Correlation avec 'Target'
- Statistique Chi2
- RFE (recursive feature elimination)
- Poids (SelectFromModel) : LogisticRegression, RandomForest, LightGBM

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	index	True	True	True	True	True	True	6
2	REGION_RATING_CLIENT_W_CITY	True	True	True	True	True	True	6
3	NAME_INCOME_TYPE_Pensioner	True	True	True	True	True	True	6
	⋮							
98	ANNUITY_INCOME_PERC	False	False	True	True	True	True	4
99	AMT_REQ_CREDIT_BUREAU_WEEK	False	False	True	True	True	True	4
100	AMT_REQ_CREDIT_BUREAU_MON	False	False	True	True	True	True	4

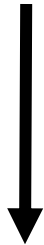
→ **100 features** importantes
selon au moins 4 méthodes

Modélisation

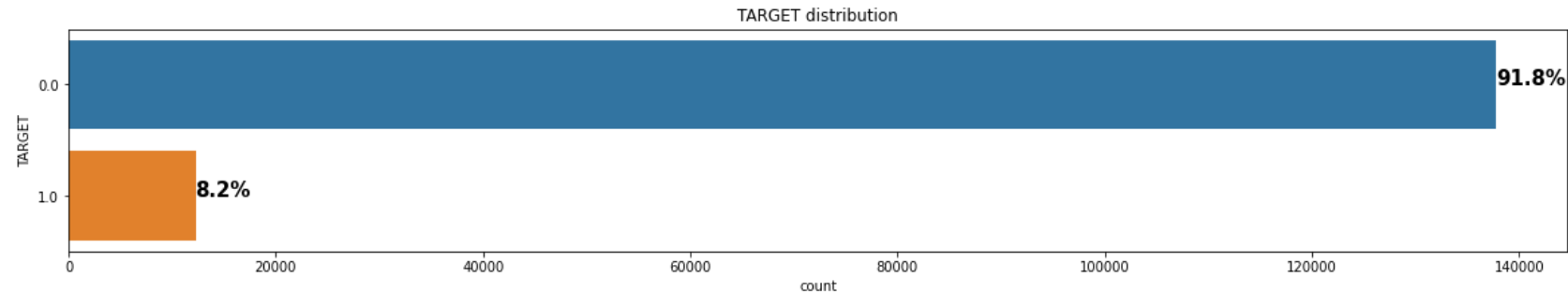
A vertical line is positioned to the right of the word 'Modélisation'. In the bottom right corner of the slide, there is a yellow right-angled triangle pointing towards the top-left.

Conception des modèles

Classes
Déséquilibrées



Split Train / Test avec **SMOTE**



	Avant SMOTE	Après SMOTE
Train size	X : (104997, 99) Y : (104997,)	X : (192652, 99) Y : (192652,)
Pourcentage de 1 dans Y train	8.26	50.0

Test de cinq modèles de classification :

- LogisticRegression
- RandomForest
- Boosting : LightGBM, XGBoost, CatBoost

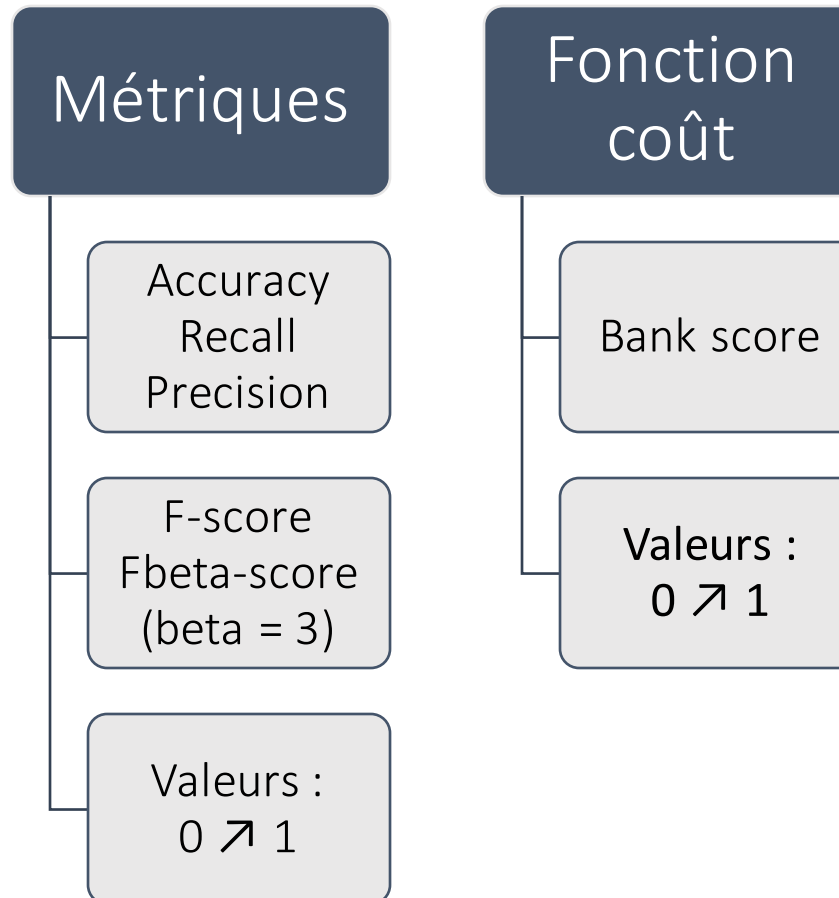
Evaluation des modèles



Limiter les risques de perte financière :

→ Pénaliser les Faux Positifs et surtout les **Faux Négatifs**

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FN + FP}$$
$$\text{Recall} = \frac{VP}{VP + FN}$$
$$\text{Precision} = \frac{VP}{VP + FP}$$
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$



$$P_{VP} = P_{VN} = 1$$

$$P_{FP} = -1 \text{ et } P_{FN} = -10$$

$$\text{model} = (VN * P_{VN}) + (FP * P_{FP}) + (FN * P_{FN}) + (VP * P_{VP})$$

$$\text{positif model} = (VN + FP) * P_{VN} + (FN + VP) * P_{VP}$$

$$\text{negatif model} = (VN + FP) * P_{FP} + (FN + VP) * P_{FN}$$

$$\text{Bank score} = \frac{\text{model} - \text{negatif model}}{\text{positif model} - \text{negatif model}}$$

Optimisation du meilleur modèle

RandomForest

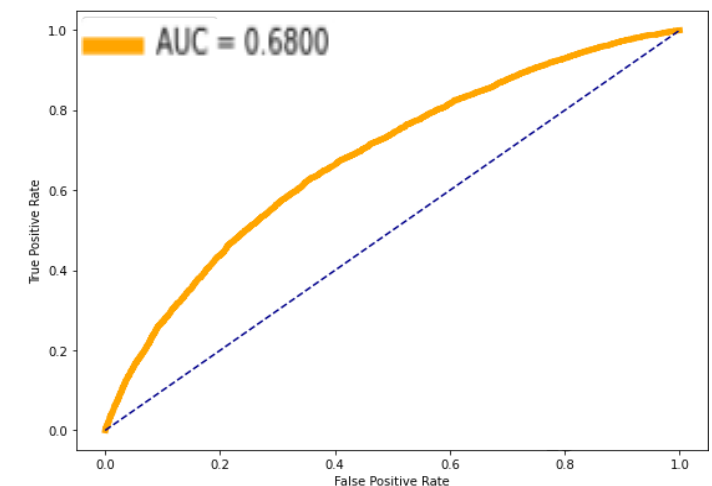
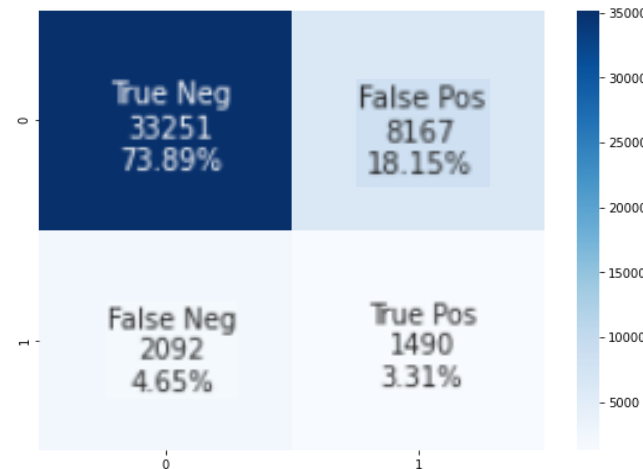
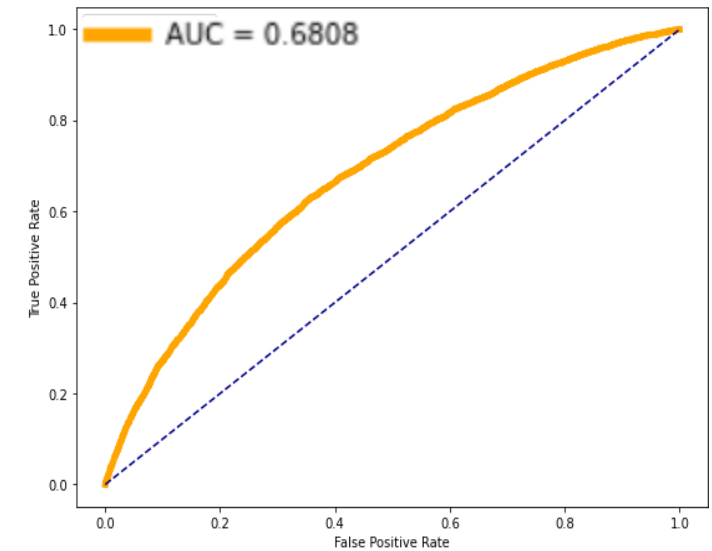
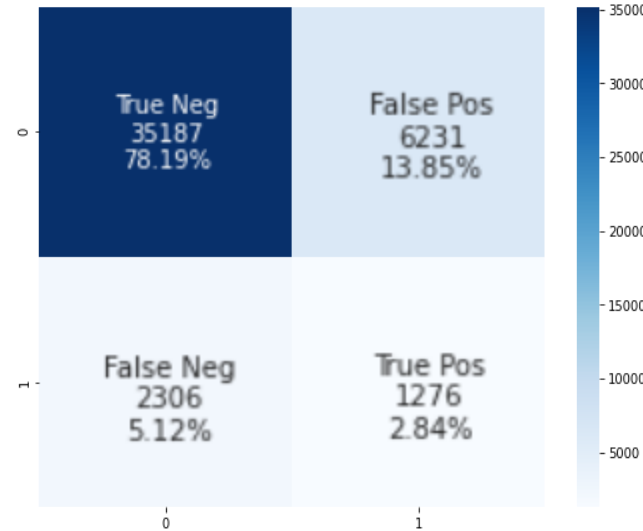
Accuracy	Precision	Recall	F-score	Fbeta	Bank score
0.81	0.17	0.36	0.23	0.31	0.69

GridSearchCV

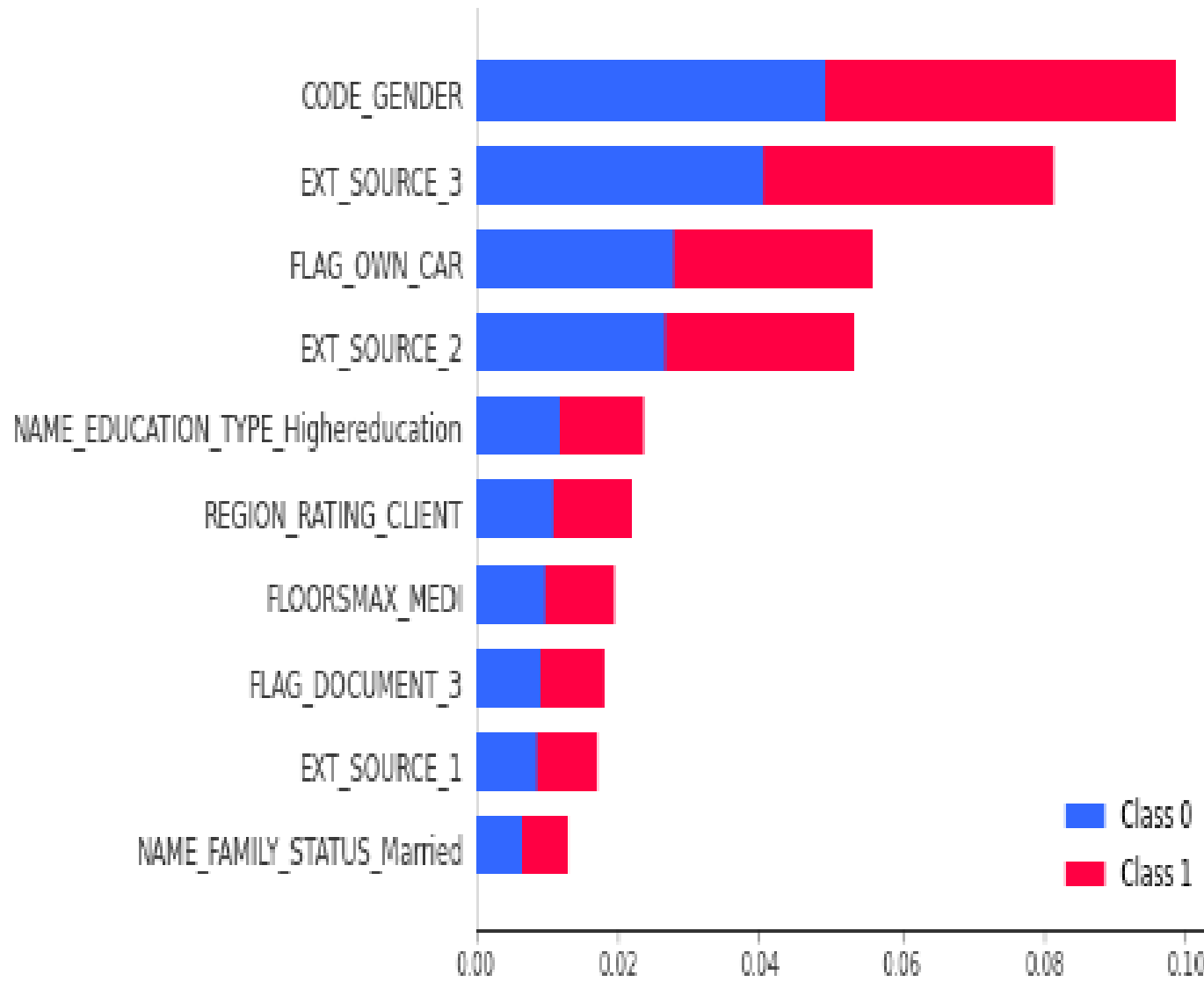
- CV = 5
- Paramètres : n_estimators, max_depth, random_stat, max_samples
- Score : Bank score

RandomForest

Accuracy	Precision	Recall	F-score	Fbeta	Bank score
0.77	0.15	0.42	0.23	0.36	0.68



Interpretabilité

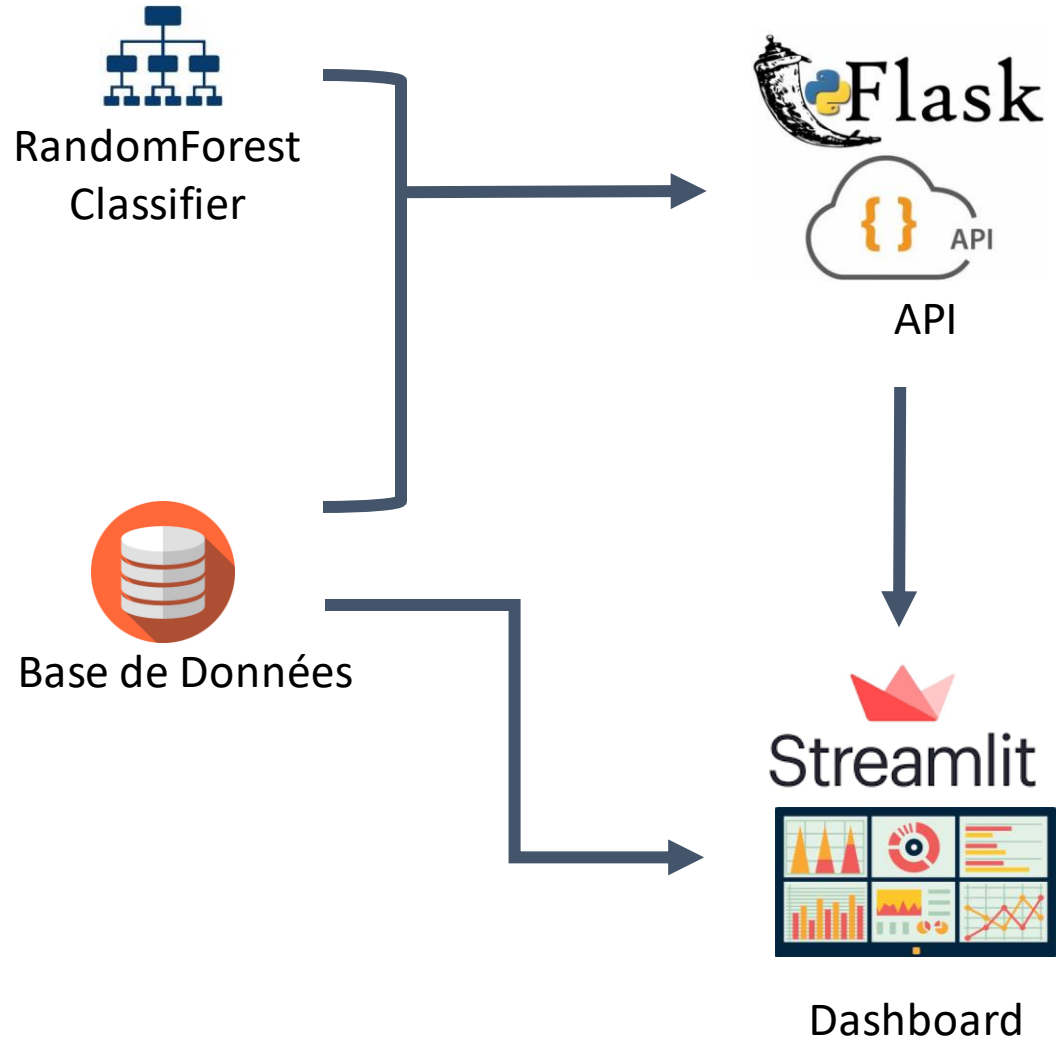


- 10 plus importants features
- Contribution dans la définition de chaque classe

Présentation Dashboard


A vertical line is positioned to the right of the title. A yellow triangle is located in the bottom right corner of the slide, pointing upwards and to the left.


Schéma fonctionnel

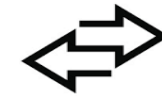


API (partie "Back end")

Prédiction à partir de l'ID du client

 [http://127.0.0.1:5000/
predict?ID=100004](http://127.0.0.1:5000/predict?ID=100004)

 [https://p7flaskapi.herokuapp.c
om/predict?ID=100004](https://p7flaskapi.herokuapp.com/predict?ID=100004)




"target": 0
"risk": 0.39

Dashboard (partie "Front end")

Partie **graphique**

 <http://localhost:8501/>

 [https://share.streamlit.io/fanjarj/oc_ds_p7
/main/Dashboard/dashboard.py](https://share.streamlit.io/fanjarj/oc_ds_p7/main/Dashboard/dashboard.py)

Limites Améliorations



Limites & Améliorations

- Feature engineering : inspiré d'un notebook Kaggle basé principalement sur une table du jeu de données



Définitions de features plus pertinentes en travaillant conjointement **avec les équipes métier**

- Bank score : basé sur des pondérations hypothétiques des prédictions (Hypothèse forte mais non confirmée)



Définition en collaboration **avec les équipes métier**

- Interprétabilité :



Prise en compte des variables issues du one hot encoding comme **un seul et même feature** (en d'autres termes revenir au feature initial)

MERCI

Questions



Incompréhensions

