Implémentez un modèle de scoring

Projet 07
N'Gouda BA

Prêt à dépenser

- Offre de crédits à la consommation
- Public : client ayant peu d'historique de prêts bancaires
- Aider la décision des chargés de clients

Enjeux :

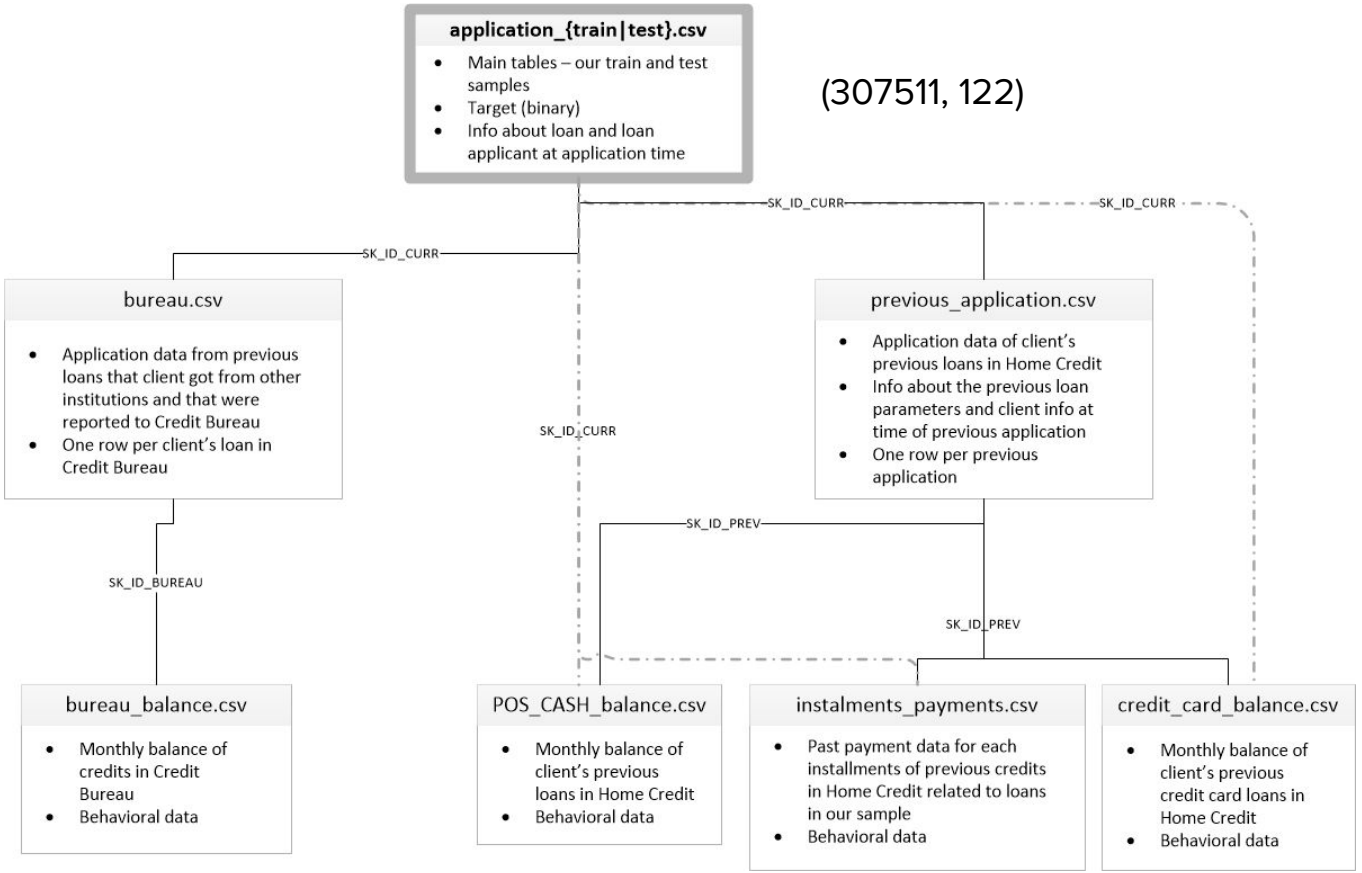
- Modèle de scoring
- API
- Dashboard

Plan

1. Les données
    1.1. Nettoyage
    1.2. Exploration
2. Problématique du crédit
3. Comparaison des modèles
4. Modélisation des bénéfices
5. Dashboard et API

Les données

Structure


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
(307511, 122)
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_PREV
SK_ID_BUREAU
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

Nettoyage et création de features

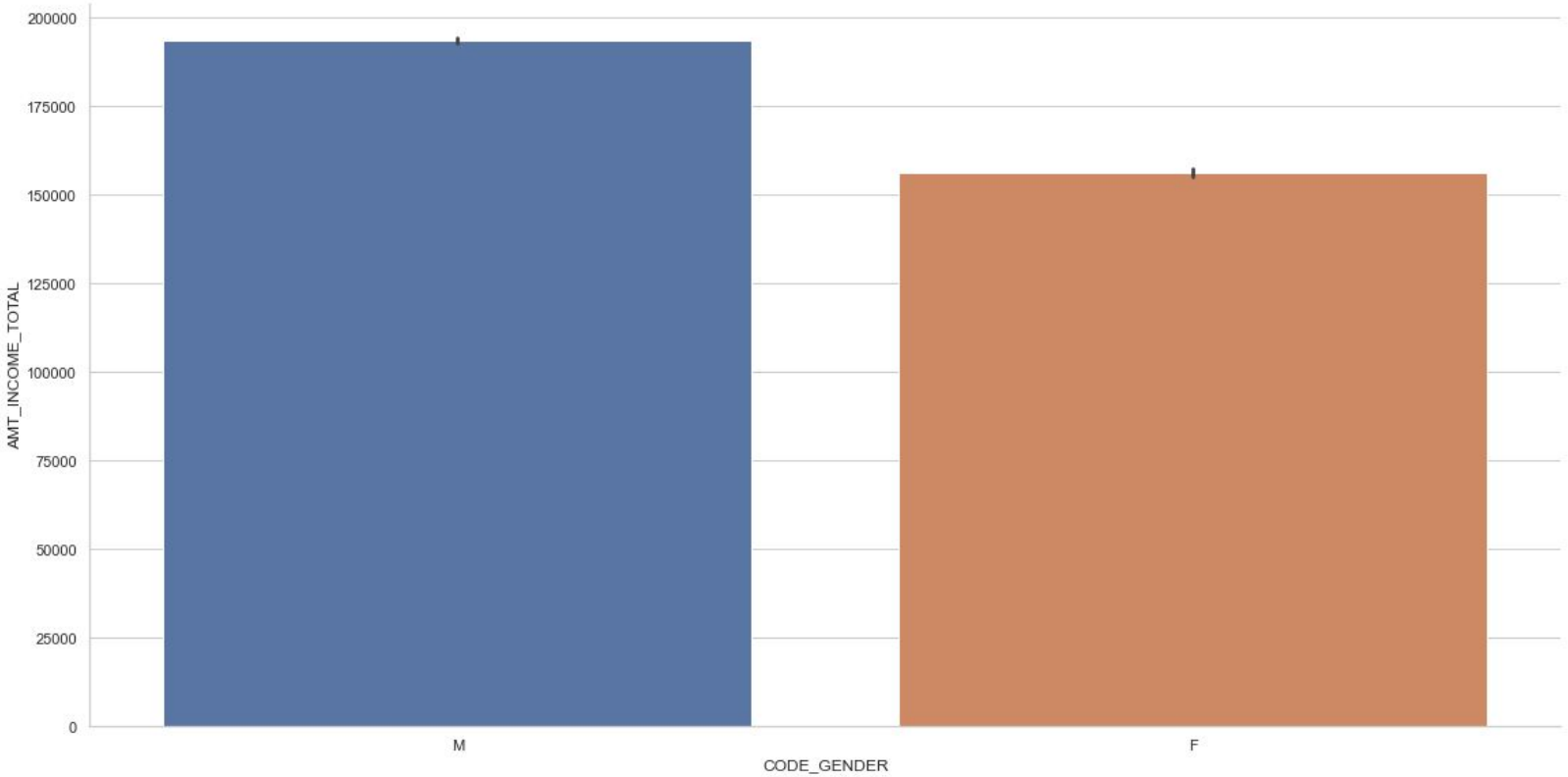
- 24% de NAN
- Suppression des features > 20% de Nan (1% de Nan)
- Deux nouvelles features
  - "credSURrevenu" = "AMT_CREDIT"/"AMT_INCOME_TOTAL"
  - "annuitySURrevenu" = "AMT_ANNUITY"/"AMT_INCOME_TOTAL"
- 307499 lignes et 66 colonnes

Exploration

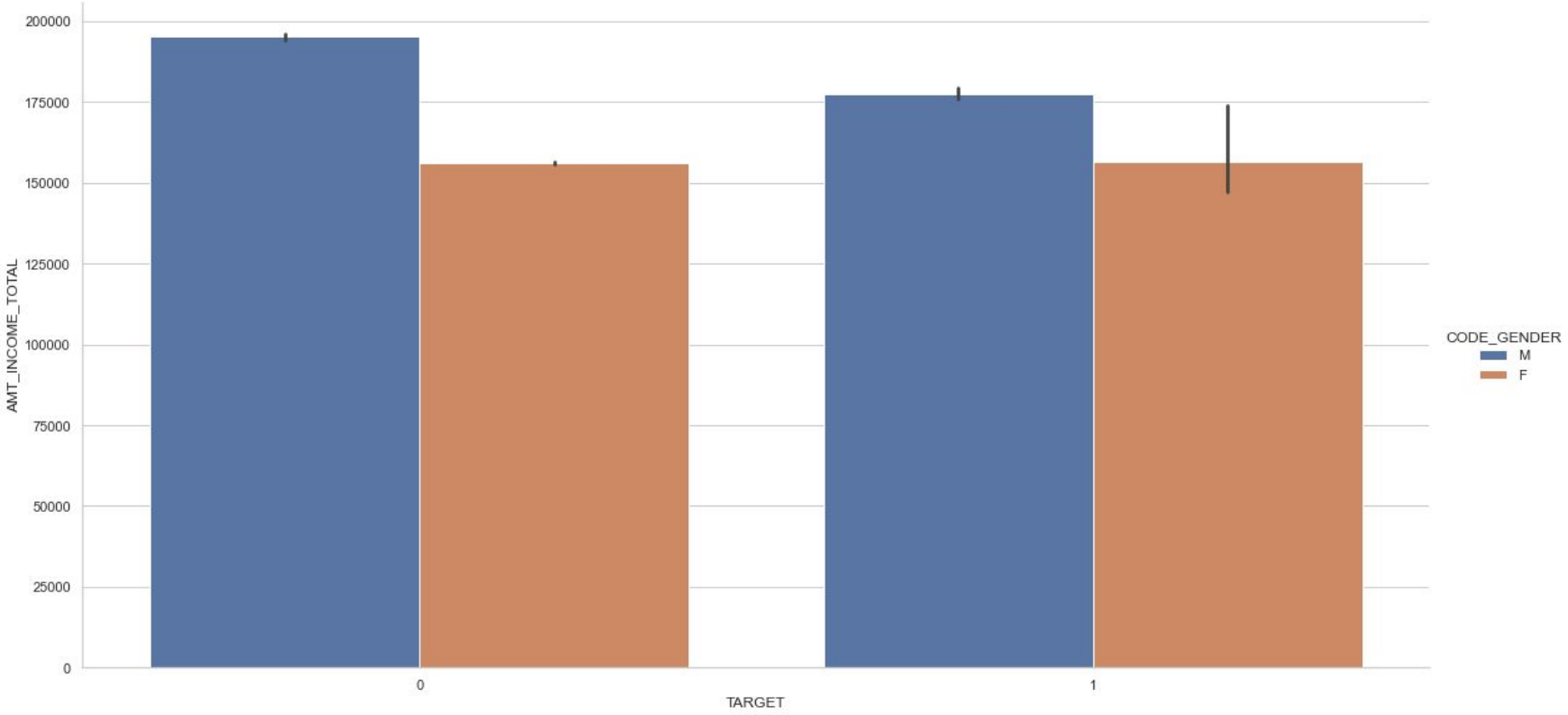
Analyse univariée

- TARGET
  - 92% Non défaut, 08% de défaut
- CODE_GENDER
  - 65% de femmes, 35% hommes

Analyse univariée



Analyse univariée



Analyse univariée

F : 07% de défaut
M : 10% de défaut

Analyse univariée

Éducation

Academic degree :

0	0.981707
1	0.018293

Lower secondary

0	0.890723
1	0.109277

Analyse univariée

Statut

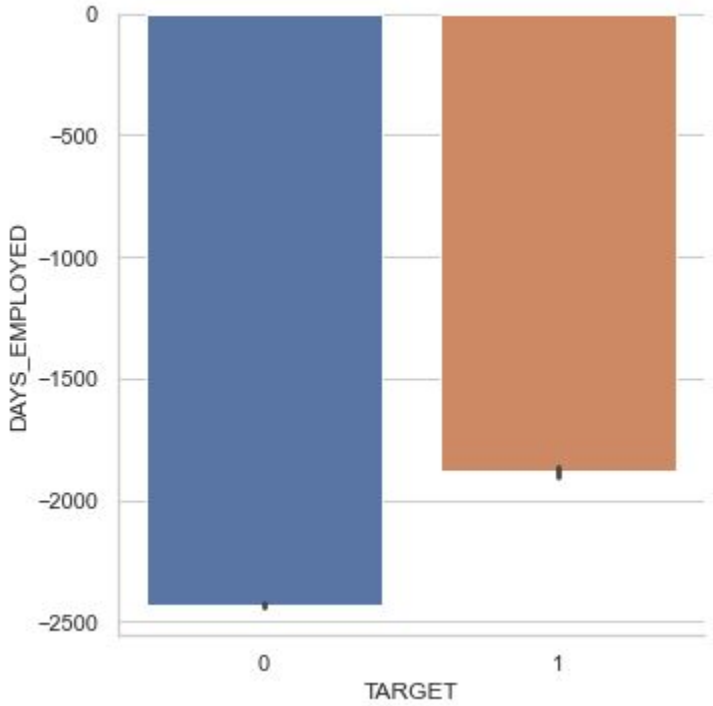
Widow

0	0.941758
1	0.058242

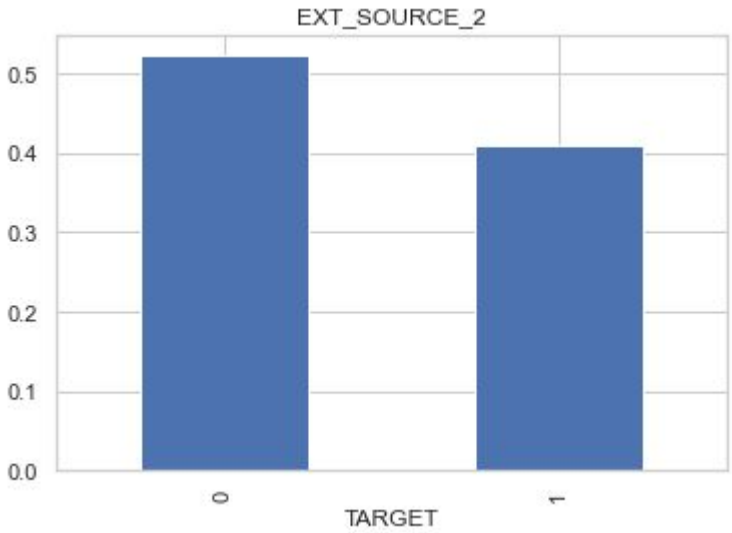
Civil marriage

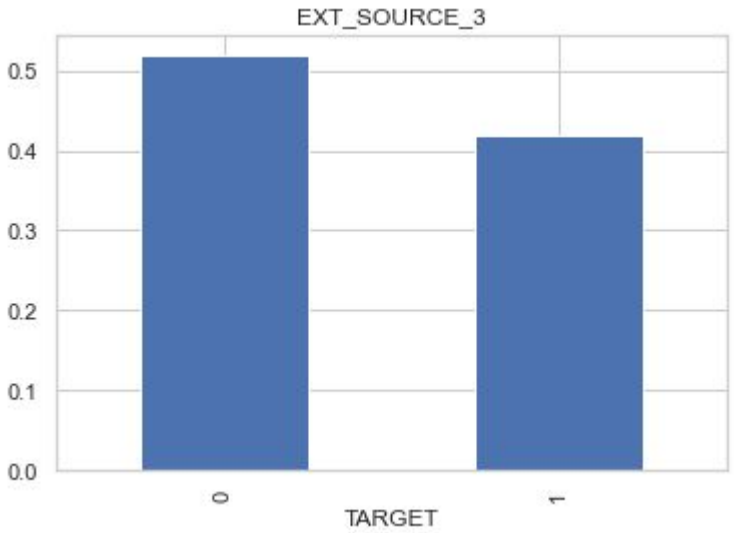
0	0.900544
1	0.099456

Analyse univariée

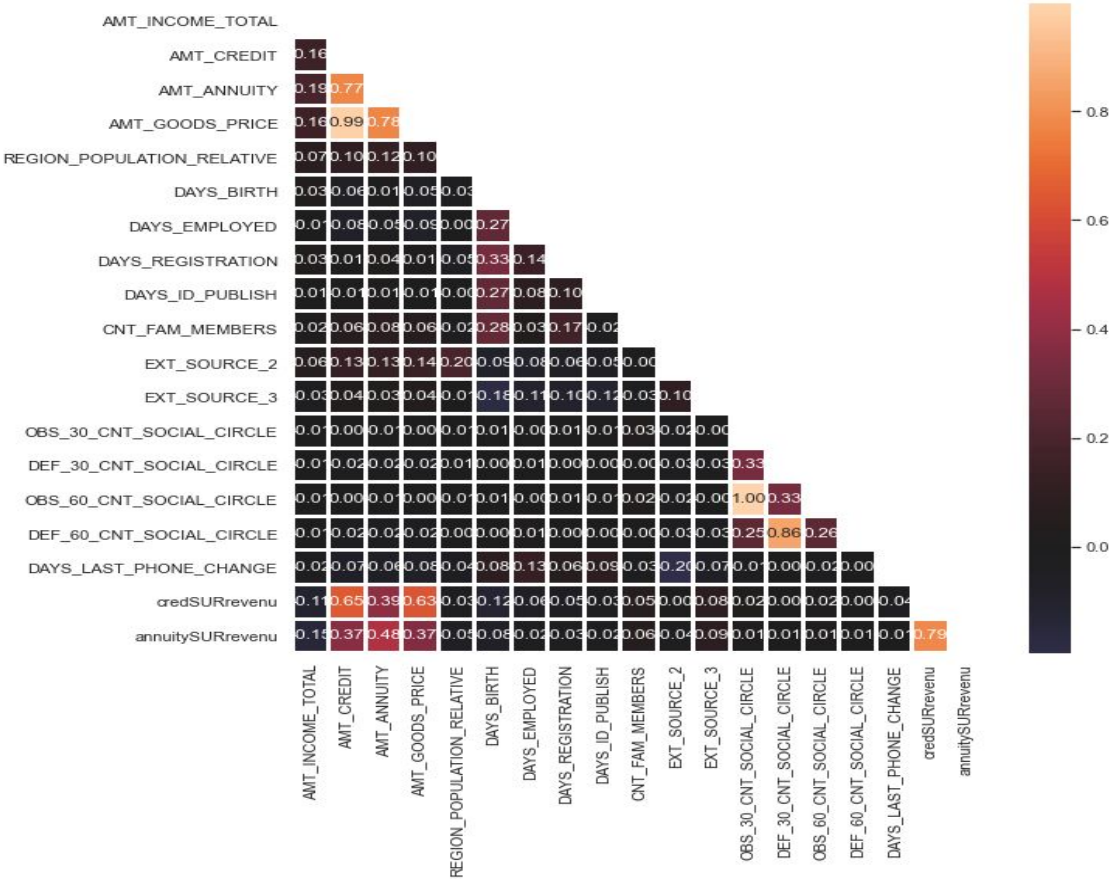


Analyse univariée





Analyse bivariée



PROBLÉMATIQUE

4 types de prédiction

PRÉDICTION

RÉEL	0	1
0	TN = GAINS	FP = OPPORTUNITÉS PERDUES
1	FN = PERTES	TP = PERTES ÉVITÉES

PRECISION = TP/(TP + FP)

RECALL = TP/(TP + FN)

Comparaison des modèles

Comparaison modèle

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.6742	0.7377	0.6731	0.6733	0.6731	0.3484	0.3484	3.1520
lightgbm	Light Gradient Boosting Machine	0.6718	0.7378	0.6690	0.6714	0.6701	0.3435	0.3436	0.5880
ridge	Ridge Classifier	0.6698	0.0000	0.6591	0.6722	0.6655	0.3395	0.3396	0.0800
lda	Linear Discriminant Analysis	0.6698	0.7342	0.6593	0.6722	0.6656	0.3395	0.3397	0.4500
ada	Ada Boost Classifier	0.6694	0.7313	0.6662	0.6692	0.6676	0.3388	0.3389	0.6960
xgboost	Extreme Gradient Boosting	0.6671	0.7289	0.6682	0.6655	0.6668	0.3343	0.3344	13.7880
rf	Random Forest Classifier	0.6662	0.7259	0.6523	0.6696	0.6608	0.3324	0.3325	1.9330
et	Extra Trees Classifier	0.6540	0.7094	0.6482	0.6544	0.6512	0.3080	0.3080	2.8430
dt	Decision Tree Classifier	0.5801	0.5801	0.5818	0.5784	0.5800	0.1602	0.1603	0.2370
nb	Naive Bayes	0.5675	0.6094	0.7412	0.5491	0.6308	0.1360	0.1451	0.0470
lr	Logistic Regression	0.5624	0.5843	0.4902	0.5713	0.5275	0.1245	0.1258	0.4860
knn	K Neighbors Classifier	0.5527	0.5713	0.5746	0.5491	0.5615	0.1056	0.1057	0.3200
qda	Quadratic Discriminant Analysis	0.5091	0.5671	0.9377	0.5042	0.6551	0.0207	0.0418	0.4550
svm	SVM - Linear Kernel	0.5039	0.0000	0.5954	0.4636	0.4126	0.0083	0.0242	0.8060

Choix du modèle

Light Gradient Boost Machine

Librairie : lightgbm

taille du jeu de données : **307495, 140**

	precision	recall	f1-score	support
Non-Default	0.96	0.69	0.80	56534
Default	0.16	0.68	0.26	4965
accuracy			0.69	61499
macro avg	0.56	0.69	0.53	61499
weighted avg	0.90	0.69	0.76	61499

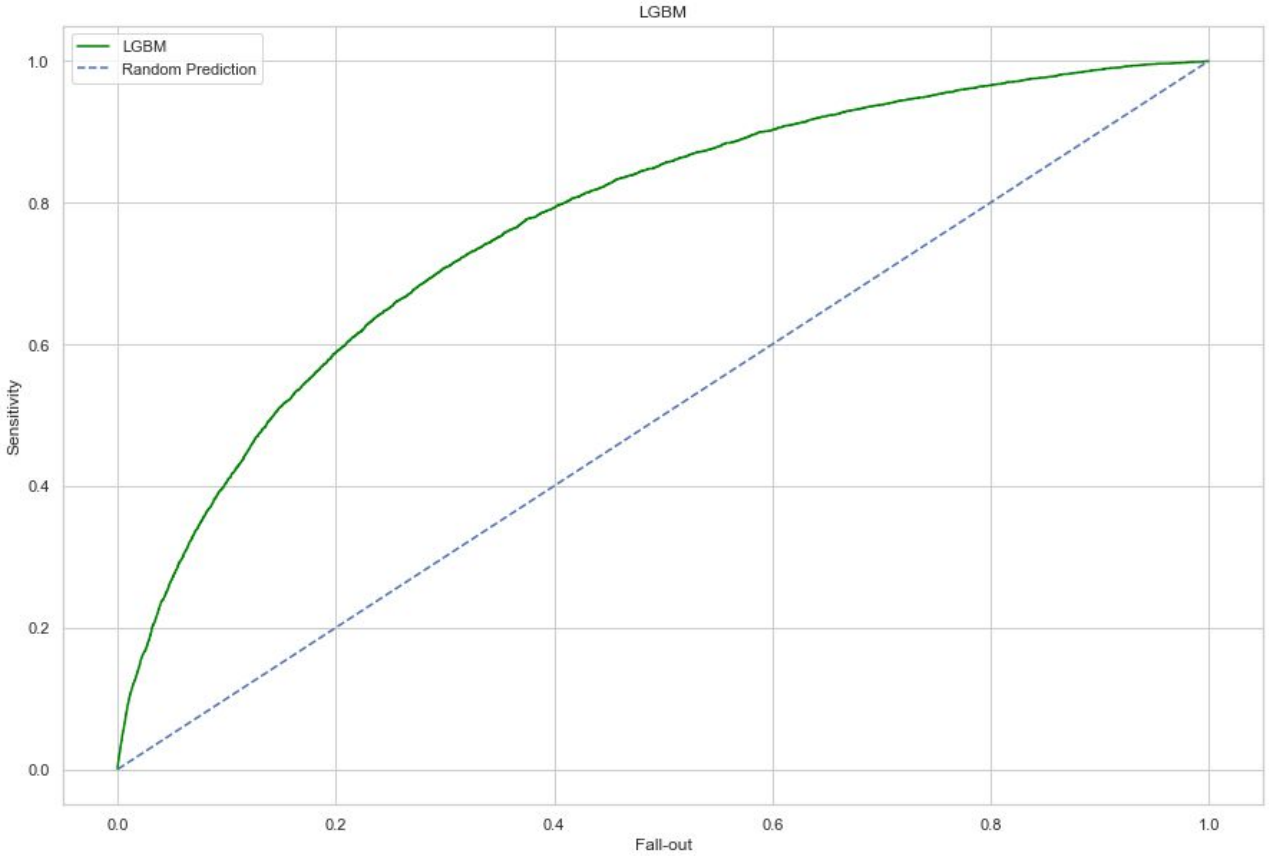
Test size : 61499

taille du jeu de données : **307495, 1224**

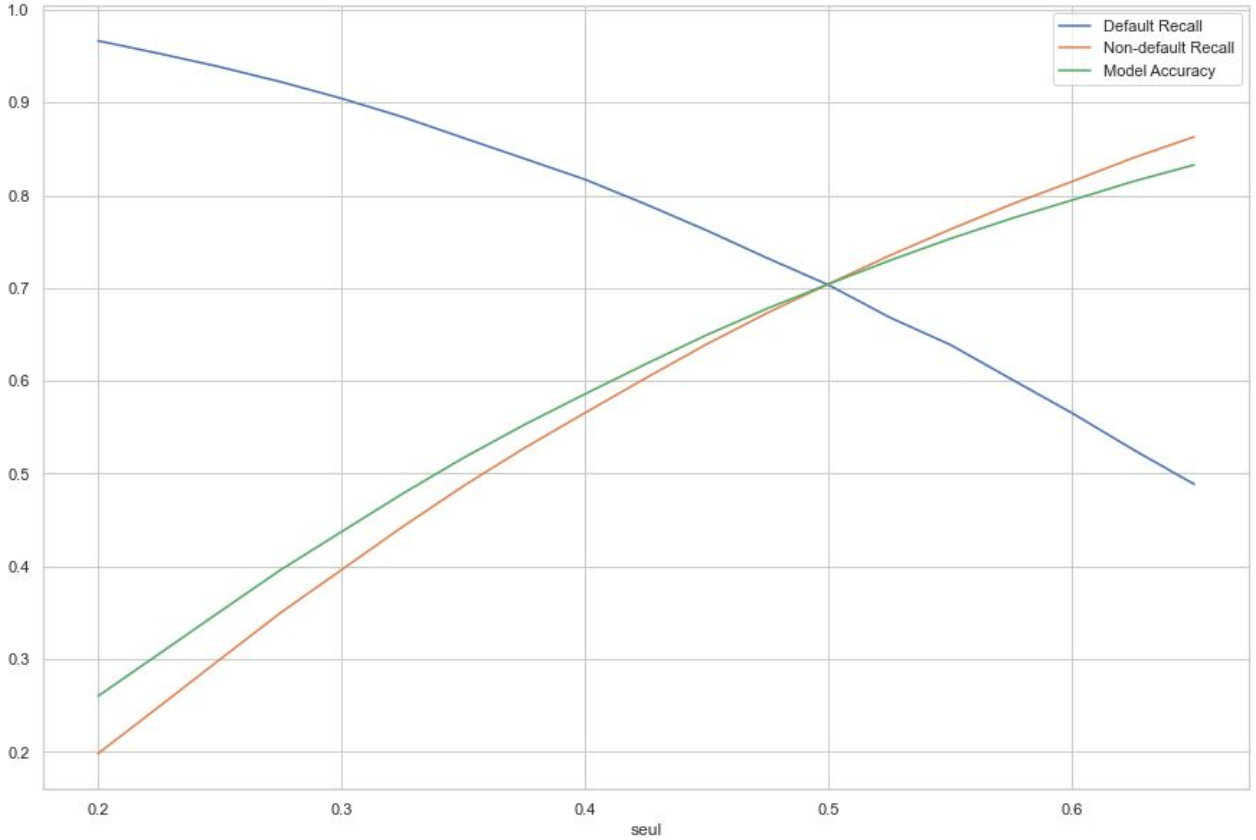
	precision	recall	f1-score	support
Non-Default	0.96	0.71	0.82	56534
Default	0.17	0.70	0.28	4965
accuracy			0.71	61499
macro avg	0.57	0.70	0.55	61499
weighted avg	0.90	0.71	0.77	61499

Courbe ROC

score ROC AUC :
0.70



Théorie VS Métier



Theorie VS Métier

Seuil : 0.5

pred_TARGET	0	1
TARGET		
$0.00	$23,868,480,866.11	$10,013,422,847.16
$1.00	$883,396,293.79	$2,092,222,836.93

Seuil : 0.8

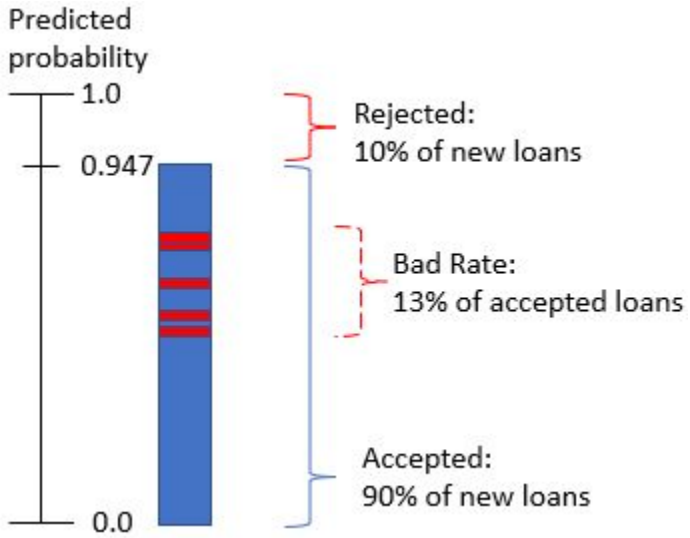
pred_TARGET	0	1
TARGET		
$0.00	$29,717,834,887.44	$4,164,068,825.84
$1.00	$1,610,969,632.10	$1,364,649,498.62

Modélisation des bénéfices

Bad rate

Bad rate = (Faux négatif)/(Total des crédits acceptés)

Acceptance Rate	Threshold	Bad Rate
1.00	0.999	0.219
0.95	0.988	0.177
0.90	0.947	0.133
0.85	0.503	0.097
0.80	0.330	0.078
0.75	0.227	0.066
0.70	0.163	0.055



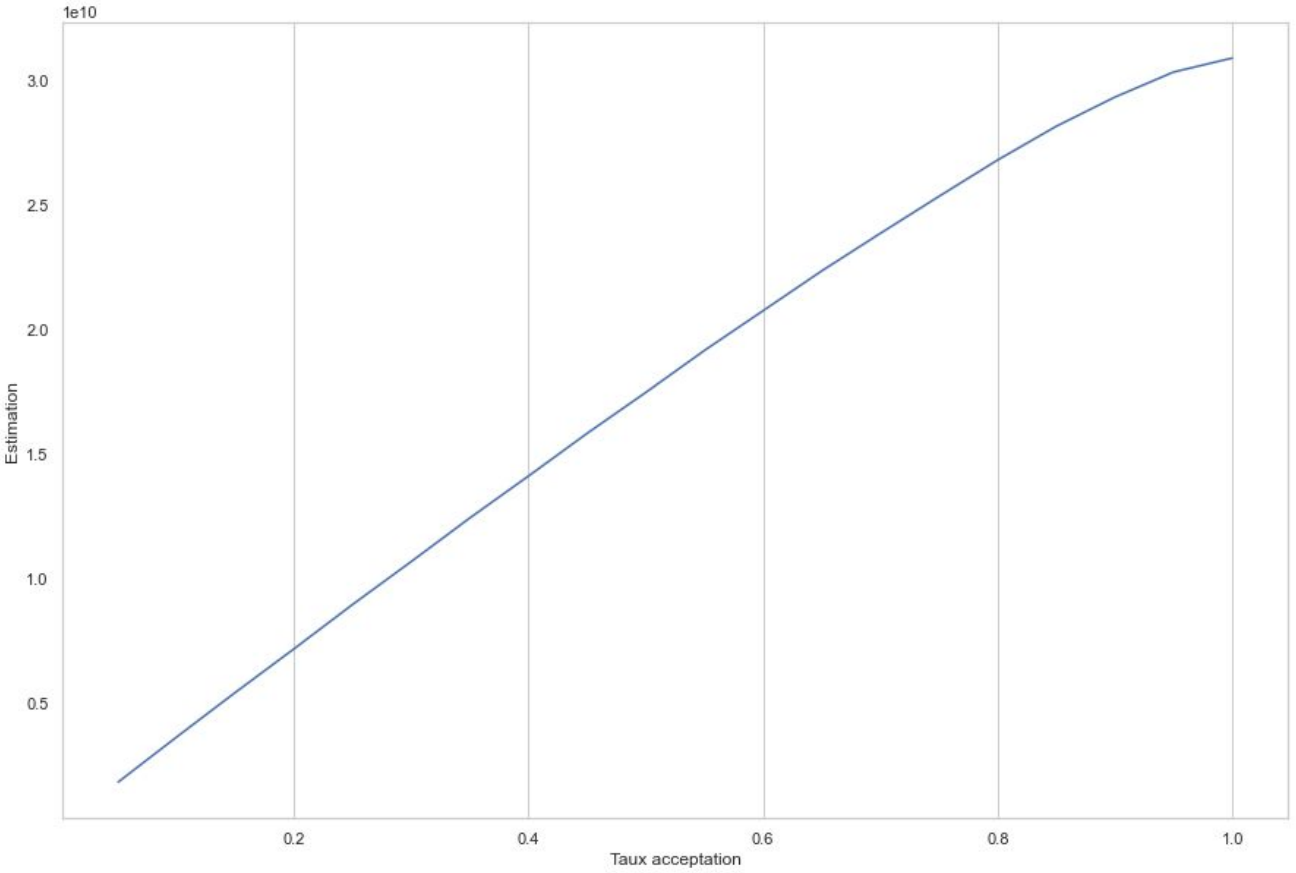
Estimation des bénéfices “nets”

	Taux acceptation	Threshold	Bad Rate	Avg	nb_accept	estimation
0	1.00	0.969	0.081	599319.059562	61499	3.088660e+10
1	0.95	0.795	0.067	599319.059562	58427	3.032422e+10
2	0.90	0.723	0.058	599319.059562	55337	2.931743e+10
3	0.85	0.666	0.051	599319.059562	52297	2.814564e+10
4	0.80	0.615	0.046	599319.059562	49231	2.679061e+10
5	0.75	0.567	0.042	599319.059562	46141	2.533031e+10
6	0.70	0.523	0.038	599319.059562	43058	2.384426e+10
7	0.65	0.483	0.034	599319.059562	39986	2.233479e+10
8	0.60	0.445	0.031	599319.059562	36887	2.073644e+10
9	0.55	0.410	0.028	599319.059562	33830	1.913957e+10
10	0.50	0.376	0.026	599319.059562	30721	1.745427e+10
11	0.45	0.345	0.024	599319.059562	27711	1.581056e+10
12	0.40	0.316	0.022	599319.059562	24603	1.409626e+10
13	0.35	0.288	0.020	599319.059562	21573	1.241195e+10
14	0.30	0.260	0.018	599319.059562	18449	1.065879e+10
15	0.25	0.235	0.017	599319.059562	15434	8.935394e+09
16	0.20	0.208	0.015	599319.059562	12302	7.151638e+09
17	0.15	0.182	0.013	599319.059562	9252	5.400733e+09
18	0.10	0.154	0.011	599319.059562	6164	3.612930e+09
19	0.05	0.121	0.007	599319.059562	3041	1.797014e+09

'estimation' = 'nb_accept' *(1-'Bad Rate')* 'Avg' - 'nb_accept'*'Bad Rate'*'Avg'

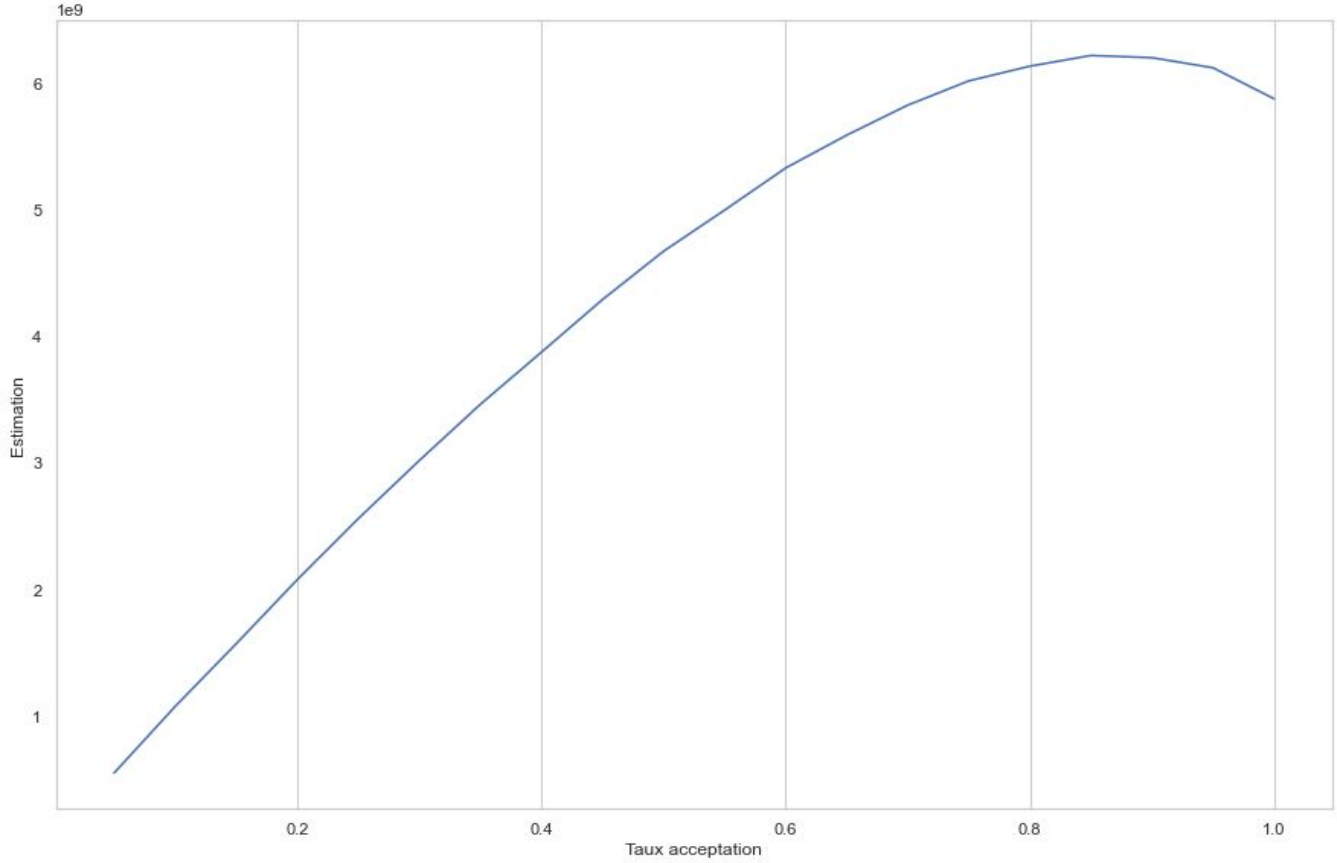
Estimation des bénéfices “nets”

8% de défaut
=
Accepter tout le monde !



Estimation des bénéfices “nets”

25% de défaut

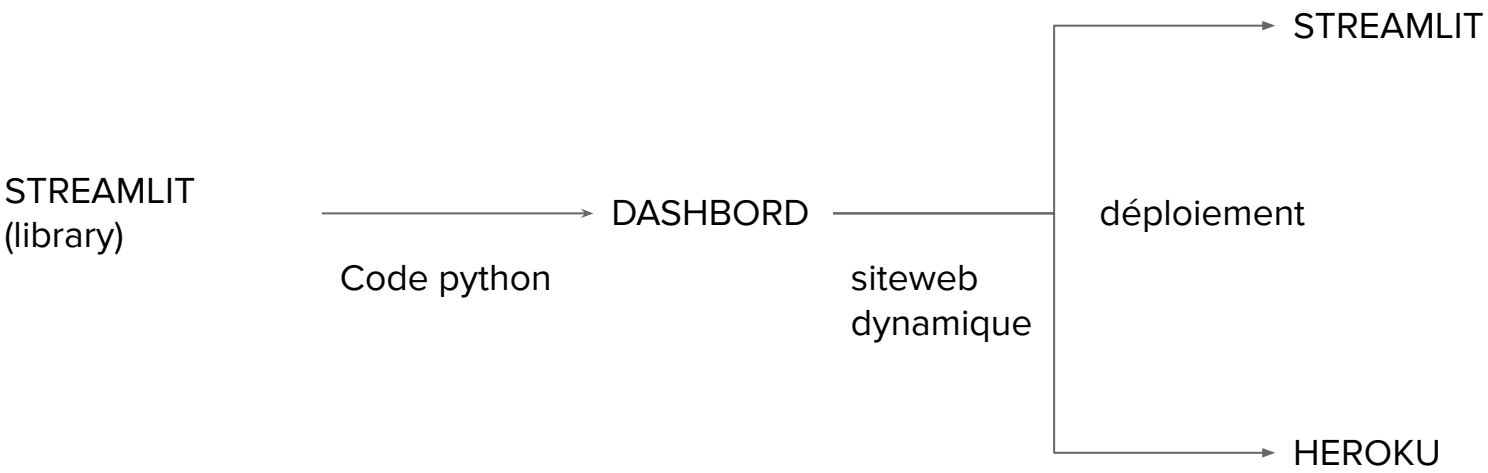


Estimation des bénéfices “nets”

	Taux acceptation	Threshold	Bad Rate	Avg	nb_accept	estimation
25% de défaut = Accepter tout le monde !	0.85	0.721	0.189	591335.4375	16893	6.213425e+09

Dashboard et API

Technologie



DASHBOARD ET API

- Motiver une décision
- Contre la déshumanisation de la prise de décision
- Rechercher des profils

Conclusion

Conclusion

- Modèle qui maximise les bénéfices
- Amélioration (FP)
- DASHBOARD orienté sur le client