



PRÊT A DEPENSER

Implémentation d'un score client et création d'un tableau de bord



Problématique



- Mettre en place un modèle de scoring des clients afin de décider l'octroi d'un prêt.
- Contraintes imposées
 - Tenir compte de la spécificité métier
 - Veiller à l'interprétabilité du score
 - Présenter une interface à destination des conseillers client
- Problématique spécifique
 - Tenir compte du déséquilibre des effectifs



Fichiers de données

- Application_train.csv:
 - 307 511 clients(282 686 classe [0]; 24825 classe [1]), soit un ratio de 8,78% de défaut.
- bureau.csv et bureau_balance.csv:
 - Informations sur tous les emprunts actuels ou passés (tous organismes confondus)
- previous_application.csv:
 - Informations sur tous les emprunts actuels ou passés (Home Credit)
- POS_CASH_balance.csv
 - Balance mensuelle des crédits en point de vente et en cash de Home Credit
- credit_card_balance.csv
 - Balance mensuelle des crédits revolving auprès de Home Credit
- installments_payments.csv
 - Historique des paiements des précédents emprunts auprès de Home Credit

Analyse exploratoire

L'analyse exploratoire des données a été menée de manière à trouver, ou créer, un nombre raisonnable de variables explicables permettant la modélisation de la problématique.

- ▶ feature engineering:
 - ▶ limité à des agrégations simples(min, max, moyenne, médiane, somme, ...)
 - ▶ Création de variables ratio interprétables
- ▶ Filtrage des variables:
 - ▶ Quantitatives, corrélées à plus de 3% avec la cible
 - ▶ Qualitatives, p-value < 5%, F-value > 200

Afin de limiter la perte d'information 20% de valeurs manquantes maximum
(stratégie de non imputation retenue)
- ▶ Toutes les tables sont jointes sur SK_ID_CURR
- ▶ Suppression d'une variables de variance nulle
- ▶ Reste 39 variables expliquant la cible, renseignées pour 166 167 clients

Modélisation:

Métrique spécifique

Original	Repayed	Default
	True Neg	False Pos
Repayed		
Default	False Neg	True Pos

Original	Repayed	Default
	2	-5
Repayed		
Default	-6	1

- Calcul de la fréquence de chaque vraie classe suivant les prédictions.
- Pondération des fréquences en pénalisant les cas défavorables au business model et en récompensant les cas favorables.
- On obtient un score variant de -11 à +3 que l'on normalise sur [0;1]

Original	Repayed	Default
	True Neg 0 0.00%	False Pos 2 100.00%
Repayed		
Default	False Neg 2 100.00%	True Pos 0 0.00%

Original	Repayed	Default
	True Neg 7 100.00%	False Pos 0 0.00%
Repayed		
Default	False Neg 0 0.00%	True Pos 6 100.00%

- Les modèles sont ensuite évalués et optimisés sur cette métrique



Modélisation

Modèles retenus

- Modèles retenus:
 - Régression Logistique
 - Random Forest classifier
 - Ligth Gradient Boosting classifier



Modélisation

Procédure d'évaluation

- Séparation des données:
 - Entraînement : 132 933 clients avec un taux de défaut de 8.552180%
 - Validation : 33 234 clients avec un taux de défaut de 8.551084%



Modélisation

Procédure d'évaluation

Solutions de rééquilibrage des classes:

- Sous échantillonnage de la classe majoritaire
- Sur échantillonnage de la classe minoritaire
- Pondération inversement proportionnelle aux effectifs des échantillons

Afin de limiter le biais de chaque option la stratégie retenue est la suivante:

- Sur-échantillonnage modéré afin de minimiser le biais introduit
- Sous-échantillonnage modéré afin de ne pas omettre trop d'informations
- Le déséquilibre restant entre les effectifs de classes est compensé par pondération



Modélisation

Procédure d'évaluation

Pour chaque modèle, toutes les options précédentes sont évaluées:

- Les scores moyens sont établis par cross validation (5 StratifiedKfolds).
- L'évolution des effectifs des matrices de confusion permettra de vérifier la bonne prise en charge par le modèle des impératifs métier de la classification.

Régression logistique

Base

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30600 99.95%	False Pos 16 0.05%
	Default	False Neg 2591 98.97%	True Pos 27 1.03%
		Repayed	Default
		Predicted	

Pondération

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21215 69.29%	False Pos 9401 30.71%
	Default	False Neg 887 33.88%	True Pos 1731 66.12%
		Repayed	Default
		Predicted	

Sous échantillonnage

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21235 69.36%	False Pos 9381 30.64%
	Default	False Neg 899 34.34%	True Pos 1719 65.66%
		Repayed	Default
		Predicted	

Approche retenue

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21173 69.16%	False Pos 9443 30.84%
	Default	False Neg 893 34.11%	True Pos 1725 65.89%
		Repayed	Default
		Predicted	

Sur échantillonnage

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21191 69.22%	False Pos 9425 30.78%
	Default	False Neg 888 33.92%	True Pos 1730 66.08%
		Repayed	Default
		Predicted	

Random Forest

Base

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30608 99.97%	False Pos 8 0.03%
	Default	False Neg 2607 99.58%	True Pos 11 0.42%
		Repayed	Default
		Predicted	

Pondération

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30613 99.99%	False Pos 3 0.01%
	Default	False Neg 2615 99.89%	True Pos 3 0.11%
		Repayed	Default
		Predicted	

Sous échantillonnage

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21029 68.69%	False Pos 9587 31.31%
	Default	False Neg 909 34.72%	True Pos 1709 65.28%
		Repayed	Default
		Predicted	

Approche retenue

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30596 99.93%	False Pos 20 0.07%
	Default	False Neg 2597 99.20%	True Pos 21 0.80%
		Repayed	Default
		Predicted	

Sur échantillonnage

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30573 99.86%	False Pos 43 0.14%
	Default	False Neg 2578 98.47%	True Pos 40 1.53%
		Repayed	Default
		Predicted	

Light Gradient Boosting

Base

		Validation	
		Repayed	Default
Original	Repayed	True Neg 30587 99.91%	False Pos 29 0.09%
	Default	False Neg 2578 98.47%	True Pos 40 1.53%
		Repayed	Default
		Predicted	

Pondération

		Validation	
		Repayed	Default
Original	Repayed	True Neg 21833 71.31%	False Pos 8783 28.69%
	Default	False Neg 948 36.21%	True Pos 1670 63.79%
		Repayed	Default
		Predicted	

Sous échantillonnage

		Validation	
		Repayed	Default
Original	Repayed	True Neg 20932 68.37%	False Pos 9684 31.63%
	Default	False Neg 864 33.00%	True Pos 1754 67.00%
		Repayed	Default
		Predicted	

Approche retenue

		Validation	
		Repayed	Default
Original	Repayed	True Neg 22036 71.98%	False Pos 8580 28.02%
	Default	False Neg 966 36.90%	True Pos 1652 63.10%
		Repayed	Default
		Predicted	

Sur échantillonnage

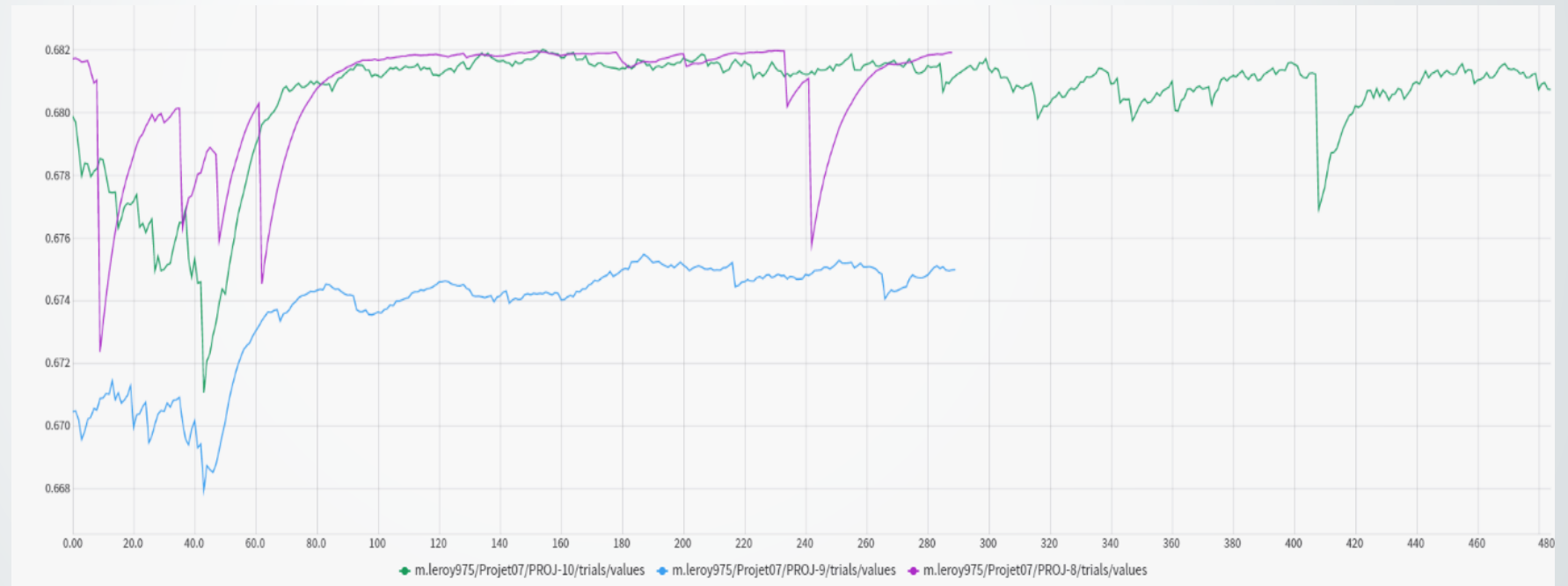
		Validation	
		Repayed	Default
Original	Repayed	True Neg 21901 71.53%	False Pos 8715 28.47%
	Default	False Neg 954 36.44%	True Pos 1664 63.56%
		Repayed	Default
		Predicted	



Optimisation des paramètres

- ▶ Pour chacun des modèles on va tester différents paramètres sur la préparation des données :
 - Sur-échantillonnage de 50% à 100% de la classe minoritaire.
 - Sous-échantillonnage de 0 à 50% de la classe majoritaire.
- ▶ Différentes transformations des distributions des données :
 - MinMaxScaler, StandardScaler, QuantileTransformer, RobustScaler, PowerTransformer
- ▶ Intégration des différentes étapes dans un pipeline:
 - [Sur échantillonnage, Sous échantillonnage, scaler, Modèle(class_weight='balanced')]
- ▶ Utilisation de la bibliothèque Optuna pour la recherche la meilleure valeur pour les paramètres testés (algorithme d'optimisation TreeParser).
- ▶ Tous les résultats intermédiaires et finaux sont suivis et enregistrés via Neptune ai

Suivi des scores durant l'optimisation des paramètres



Résultats des modèles optimisés

	Model name	Cost on train	Auc on train	Cost on test
0	Logistic Regression [Out of bag]	0.503890	0.742715	0.504895
1	Logistic Regression [Under]	0.680324	0.742279	0.675100
2	Logistic Regression [Over]	0.681392	0.742983	0.676482
3	Logistic Regression [Balance]	0.681711	0.742946	0.677065
4	Logistic Regression [Samplers & Weight]	0.681730	0.742893	0.675233
5	Logistic Regression [Optimisation]	0.682181	0.743257	0.676767

	Model name	Cost on train	Auc on train	Cost on test
0	Random Forest [Out of bag]	0.501989	0.712528	0.501970
1	Random Forest [Under]	0.673418	0.732286	0.669826
2	Random Forest [Over]	0.505893	0.720975	0.506937
3	Random Forest [Balance]	0.500439	0.717160	0.500524
4	Random Forest [Samplers & Weight]	0.503469	0.723905	0.503684
5	Random Forest [Optimisation]	0.674362	0.737256	0.673935

	Model name	Cost on train	Auc on train	Cost on test
0	LightGradientBoosting [Out of bag]	0.505965	0.741017	0.507166
1	LightGradientBoosting [Under]	0.675307	0.737229	0.676836
2	LightGradientBoosting [Over]	0.675787	0.739987	0.675472
3	LightGradientBoosting [Balance]	0.677587	0.740348	0.675508
4	LightGradientBoosting [Samplers & Weight]	0.677038	0.740666	0.675385
5	LightGradientBoosting [Optimisation]	0.677858	0.741548	0.679327

		Entrainement		Validation	
Original	Repayed	True Neg 84347 68.88%	False Pos 38113 31.12%	True Neg 21180 69.18%	False Pos 9436 30.82%
	Default	False Neg 3406 32.52%	True Pos 7067 67.48%	False Neg 888 33.92%	True Pos 1730 66.08%
		Repayed	Default	Repayed	Default
		Predicted		Predicted	

		Entrainement		Validation	
Original	Repayed	True Neg 87026 71.06%	False Pos 35434 28.94%	True Neg 21725 70.96%	False Pos 8891 29.04%
	Default	False Neg 3268 31.20%	True Pos 7205 68.80%	False Neg 947 36.17%	True Pos 1671 63.83%
		Repayed	Default	Repayed	Default
		Predicted		Predicted	

		Entrainement		Validation	
Original	Repayed	True Neg 87645 71.57%	False Pos 34815 28.43%	True Neg 21833 71.31%	False Pos 8783 28.69%
	Default	False Neg 2796 26.70%	True Pos 7677 73.30%	False Neg 928 35.45%	True Pos 1690 64.55%
		Repayed	Default	Repayed	Default
		Predicted		Predicted	

Création d'un méta modèle

► Voting Classifier

- La classification est faite sur la somme de la probabilité d'appartenance à la classe 1 donnée par chacun des modèles.

		Entraînement		Validation	
Original	Repayed	True Neg 86892 70.96%	False Pos 35568 29.04%	True Neg 21762 71.08%	False Pos 8854 28.92%
	Default	False Neg 3097 29.57%	True Pos 7376 70.43%	False Neg 936 35.75%	True Pos 1682 64.25%
		Repayed	Default	Repayed	Default
		Predicted		Predicted	

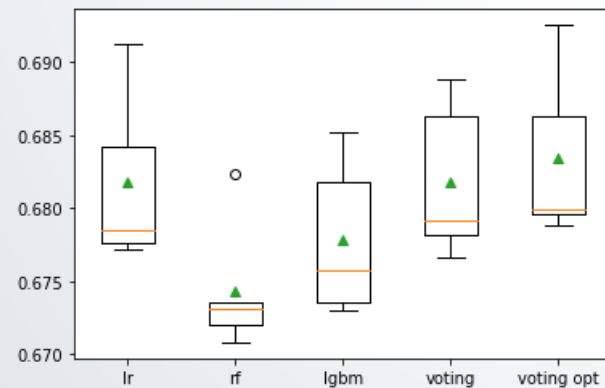
- Avec optimisation des poids des votes

		Entraînement		Validation	
Original	Repayed	True Neg 86255 70.44%	False Pos 36205 29.56%	True Neg 21663 70.76%	False Pos 8953 29.24%
	Default	False Neg 3136 29.94%	True Pos 7337 70.06%	False Neg 919 35.10%	True Pos 1699 64.90%
		Repayed	Default	Repayed	Default
		Predicted		Predicted	

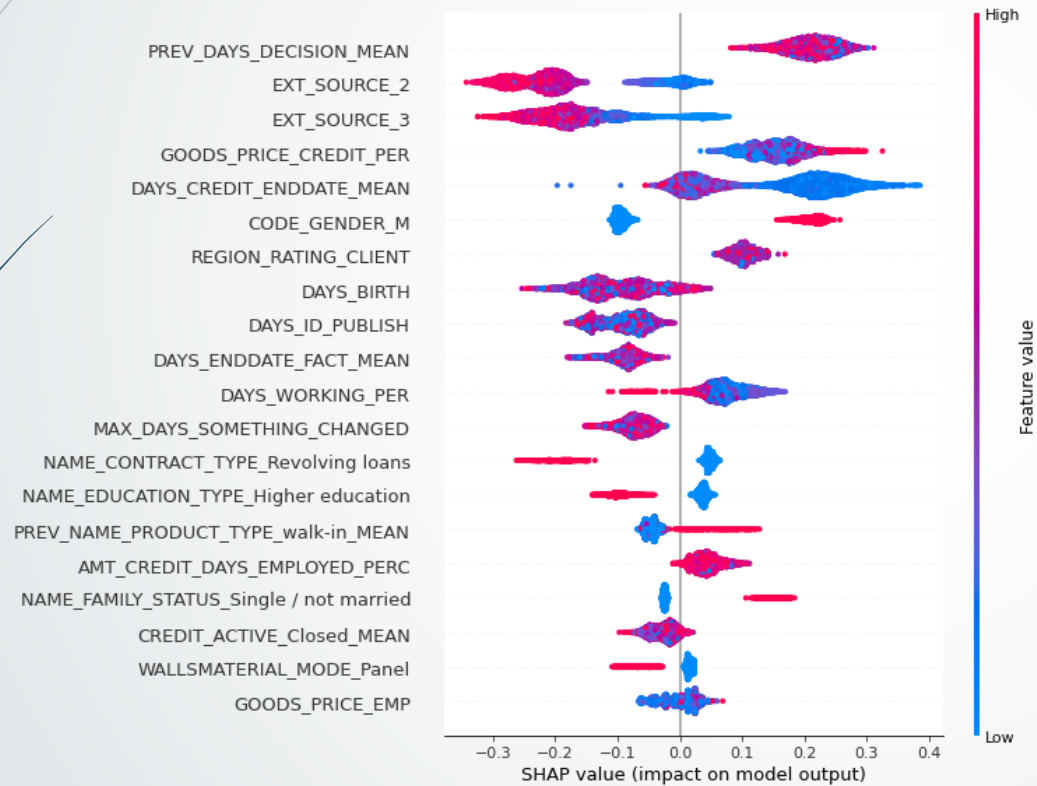
Résultats finaux

➡ Score final

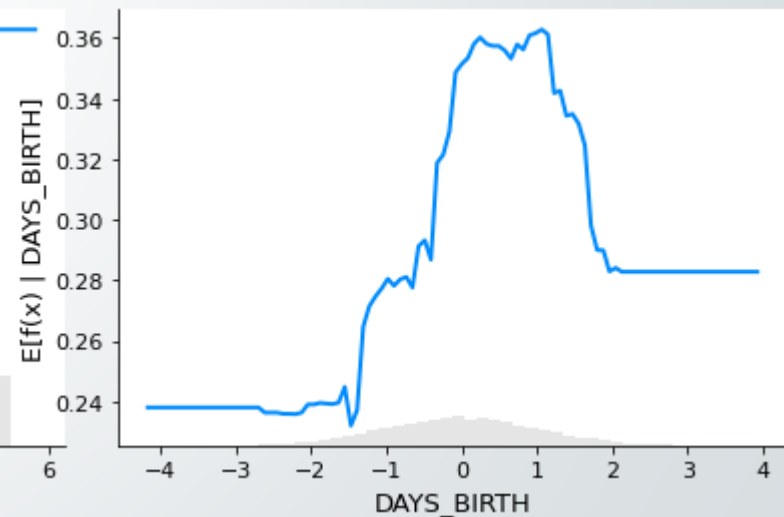
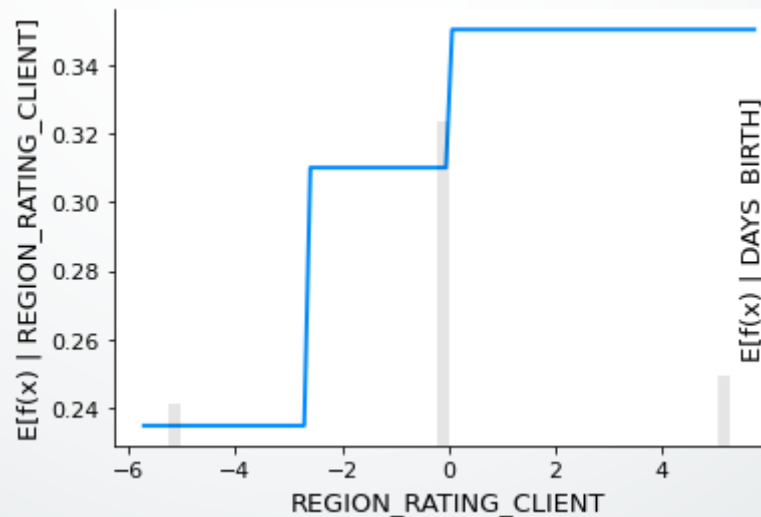
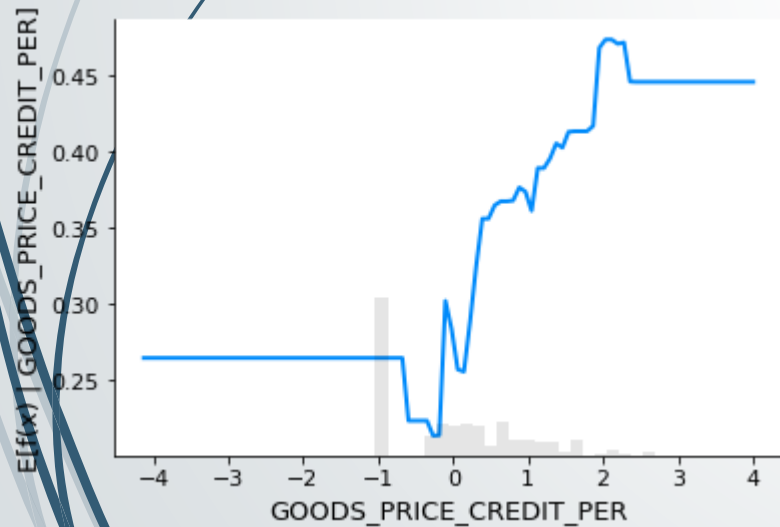
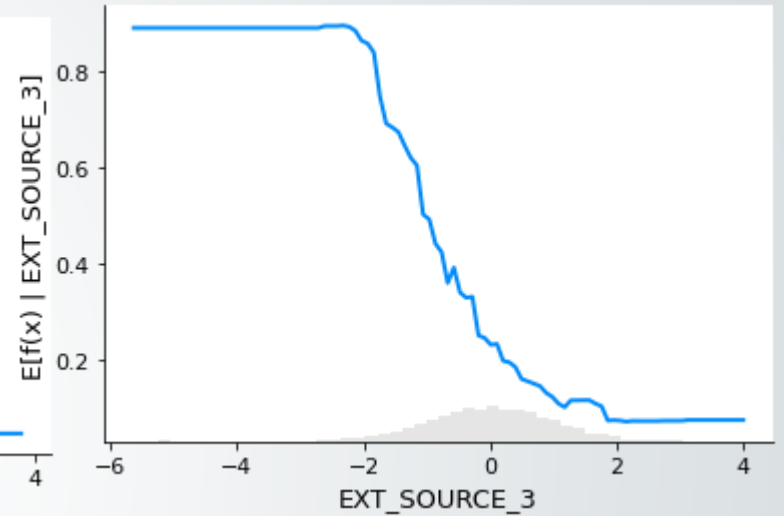
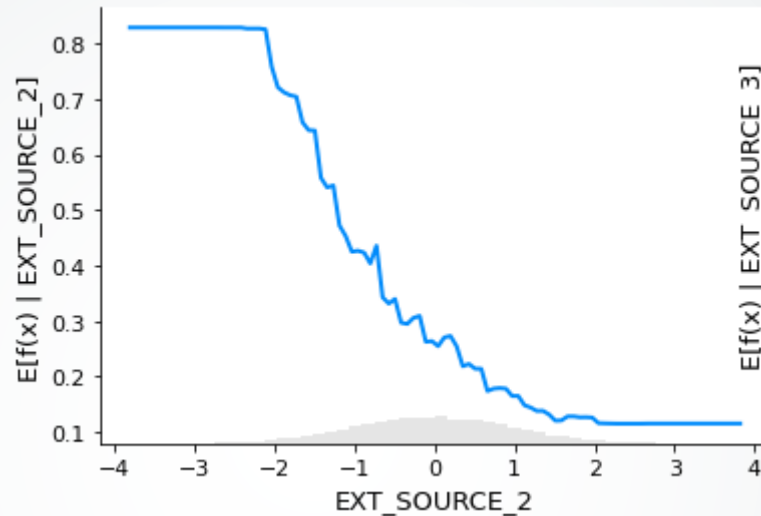
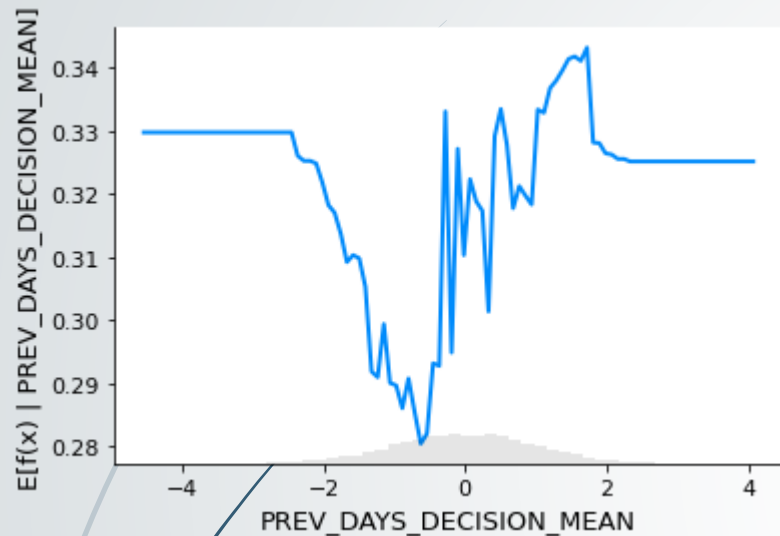
	Model name	Cost max	Cost avr	Cost min	Fit time max	Fit time avr	Fit time min	Cost on test
0	lr	0.691204	0.681747	0.677224	12.333586	10.981507	9.408714	0.676302
1	rf	0.682291	0.674362	0.670829	24.107776	23.756953	23.225049	0.673935
2	lgbm	0.685214	0.677858	0.672996	21.914827	21.320651	20.082474	0.679327
3	voting	0.688778	0.681816	0.676641	116.180967	101.697165	74.831378	0.676640
4	voting opt	0.692513	0.683433	0.678822	58.099349	56.372617	52.532451	0.678270



Explicabilité



Explicabilité



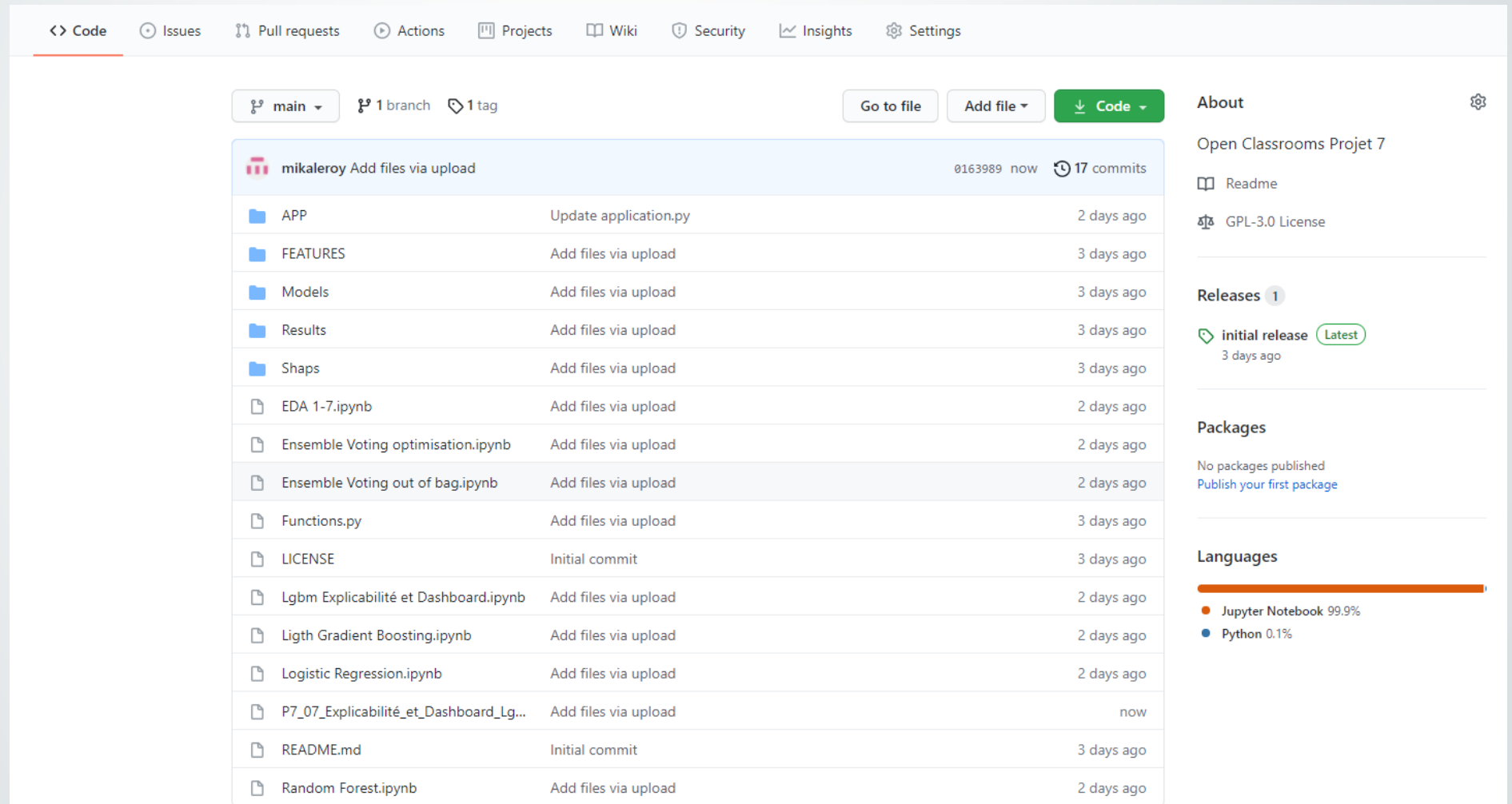


Conclusion partie modélisation

- L'optimisation des paramètres du modèle, sur la métrique créée, permet de prédire si un client fera défaut dans 64,55% des cas tout en ne rejetant que 28,696% de bons clients.
- La démarche précédente permet d'obtenir un modèle de scoring maximisant les cas favorables au modèle économique de l'entreprise, étant entendu que la pondération choisie ici pourra évoluer en fonction des objectifs.
- Pour aller plus loin :
 - Il faudrait surement changer de paradigme quant à la sélection des variables (pour des raisons de temps de calcul, le choix à peut-être été trop restrictif)
 - L'impact du choix arbitraire du random seed tout au long de cette étude serait à évaluer quant aux performances et qualité de généralisation de la modélisation.
 - De la même manière l'élargissement du corps électoral du méta modèle est lui aussi à évaluer.

Code et versions

➡ [Lien vers le dossier de versioning GitHub](#)



The screenshot displays a GitHub repository interface for the user 'mikaleroy'. The repository is named 'Add files via upload' and has 17 commits. The file list includes folders like APP, FEATURES, Models, Results, and Shaps, as well as various .ipynb files and a README.md. The right sidebar shows repository details such as the README, license (GPL-3.0), and a single release labeled 'initial release' as the latest version. A language usage bar at the bottom indicates that the repository is primarily composed of Jupyter Notebook files (99.9%) and Python code (0.1%).

Navigation: <> Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

Repository: mikaleroy Add files via upload | 0163989 now | 17 commits

Buttons: Go to file | Add file | Code

File/Folder	Action	Time
APP	Update application.py	2 days ago
FEATURES	Add files via upload	3 days ago
Models	Add files via upload	3 days ago
Results	Add files via upload	3 days ago
Shaps	Add files via upload	3 days ago
EDA 1-7.ipynb	Add files via upload	2 days ago
Ensemble Voting optimisation.ipynb	Add files via upload	2 days ago
Ensemble Voting out of bag.ipynb	Add files via upload	2 days ago
Functions.py	Add files via upload	3 days ago
LICENSE	Initial commit	3 days ago
Lgbm Explicabilité et Dashboard.ipynb	Add files via upload	2 days ago
Ligth Gradient Boosting.ipynb	Add files via upload	2 days ago
Logistic Regression.ipynb	Add files via upload	2 days ago
P7_07_Explicabilité_et_Dashboard_Lg...	Add files via upload	now
README.md	Initial commit	3 days ago
Random Forest.ipynb	Add files via upload	2 days ago

About

Open Classrooms Projet 7

Readme

GPL-3.0 License

Releases 1

initial release **Latest** (3 days ago)

Packages

No packages published

[Publish your first package](#)

Languages

- Jupyter Notebook 99.9%
- Python 0.1%

Dashboard

https://share.streamlit.io/mikaleroy/home_credit_app/main/application.py

Summary

Select loan application

121700

Amount goods price : 238500.0 ₹

Goods price / Loan : 110.56%

Percent of working days : 12.84%

Total loans / working days : 0.59%

Working since : 48 months

Work : Business Entity Type 3

Income type : Working

Occupation : n.a.

Age : 33 years

Family status : Separated

Education type : Secondary / secondary special

Housetype mode : block of flats

Housing type : House / apartment

EMERGENCYSTATE_MODE : No

Force plot

-0.08538

0.01462

0.1146

0.2146

0.3146

base value
0.4146

0.5146

higher ⇌ lower
f(x)

0.6146 0.7146 **0.74**

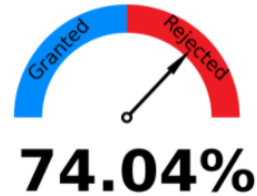
0.8146

0.9146

secondary special = 1 | NAME_EDUCATION_TYPE_Higher education = 0 | PREV_DAYS_DECISION_MEAN = -247 | CODE_GENDER_M = 1 | DAYS_BIRTH = -1.221e+4 | EXT_SOURCE_2 = 0.2651 | EXT_SOURCE_3 = 0.2609 | DAYS_ID_PUBLISH = -4.653 | GOODS_PRICE_CREDIT_PER :

Decision

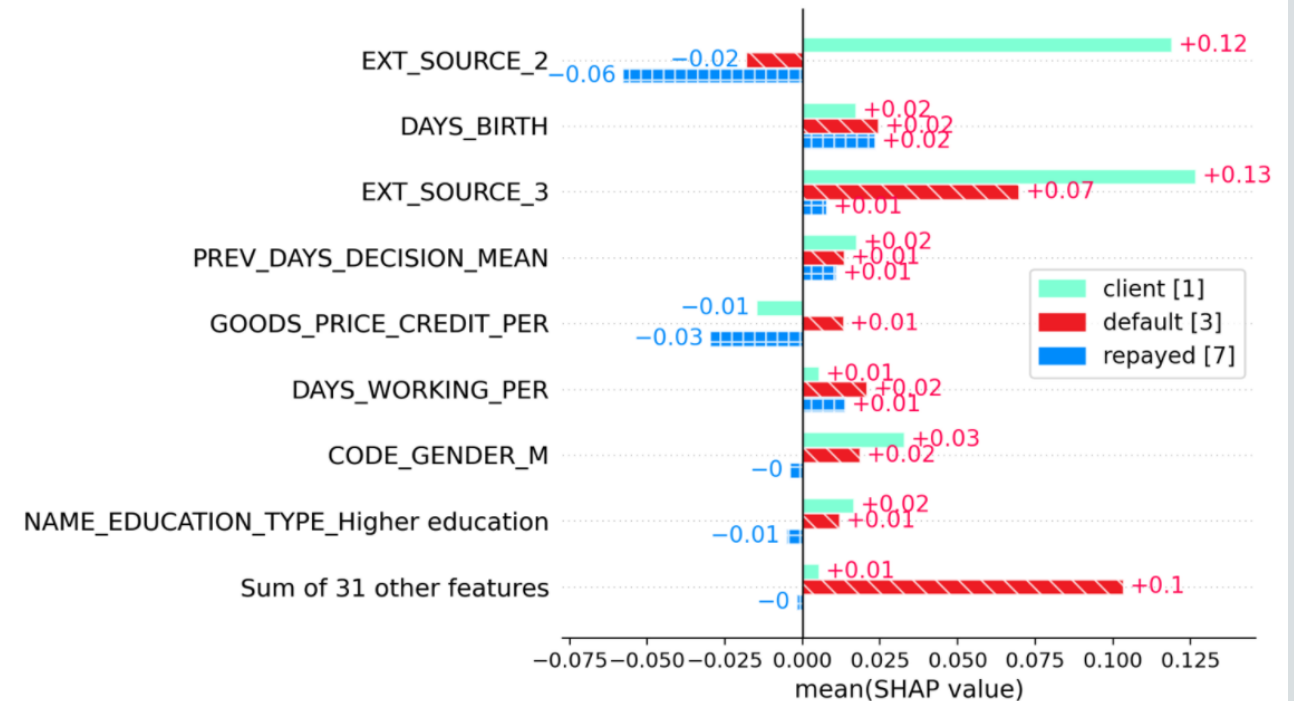
[. . .]

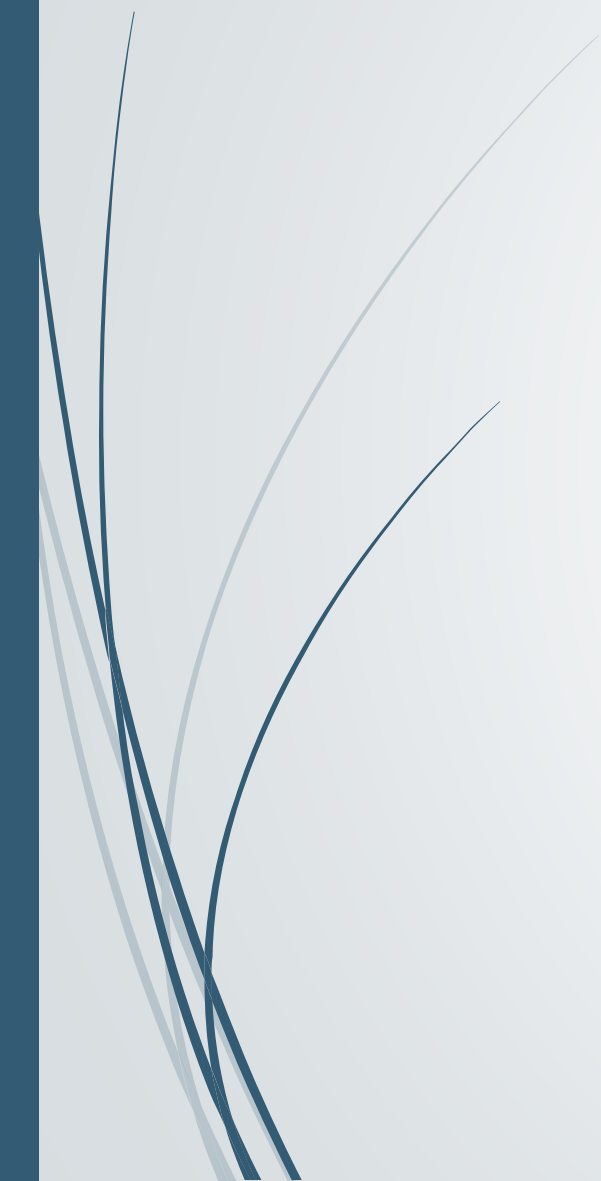
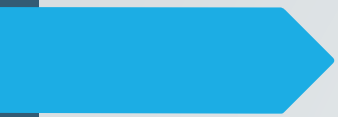


Explanations for application

Neighbors number

10





➡ Fin de la présentation