IMPLEMANTATION D'UN MODELE DE SCORING, API ET DASHBOARD

Objet du projet :

Mettre en œuvre un outil de scoring crédit pour calculer la probabilité qu'un client rembourse son crédit, classifier la demande en crédit accordé ou refusé, et communiquer les résultats sur un dashboard de façon claire et transparente à destination des chargés de relation client et du client.

L'outil s'appuie sur des sources de données variées : données comportementales, données provenant d'autres institutions financières…

L'outil permet au chargé de relation client d'expliquer de façon la plus transparente possible la décision d'octroi d'un crédit à son client. Il doit pouvoir explorer avec lui ses données personnelles.

METHODOLOGIE D'ENTRAINEMENT DU MODELE 1/2

Deux modèles ont été entraînés pour le projet, avec les objectifs suivants :

	Objectif	Caractéristique	Conséquence
Modèle n°1	Rapidité d'exécution	Données prétraitées	Pas de modification possible des données par le chargé de relation client
	Maximisation de la performance du modèle	Le modèle choisi présente une courbe de performance en fonction du seuil de probabilité en « pic »	Peu de possibilité de « négociation » pour le chargé de relation client
Modèle n°2	Permettre de modifier des variables avant prédiction	Modèle pipeline intégrant le pré-traitement des données	Plus lent pour la prédiction
	Offrir une marge de manœuvre pour la « négociation »	Le modèle choisi présente une courbe de performance en fonction du seuil de probabilité plus « plane »	Les performances du modèle sont légèrement amoindries. Le chargé de relation garde sa crédibilité face au client. Le chargé de relation est responsabilisé face à « l'outil ».

Pour ces deux méthodes d'entrainement, un algorithme identique a été utilisé. Les hyperparamètres ont été obtenus par optimisation bayésienne. En voici les caractéristiques :

ALGORITHME :

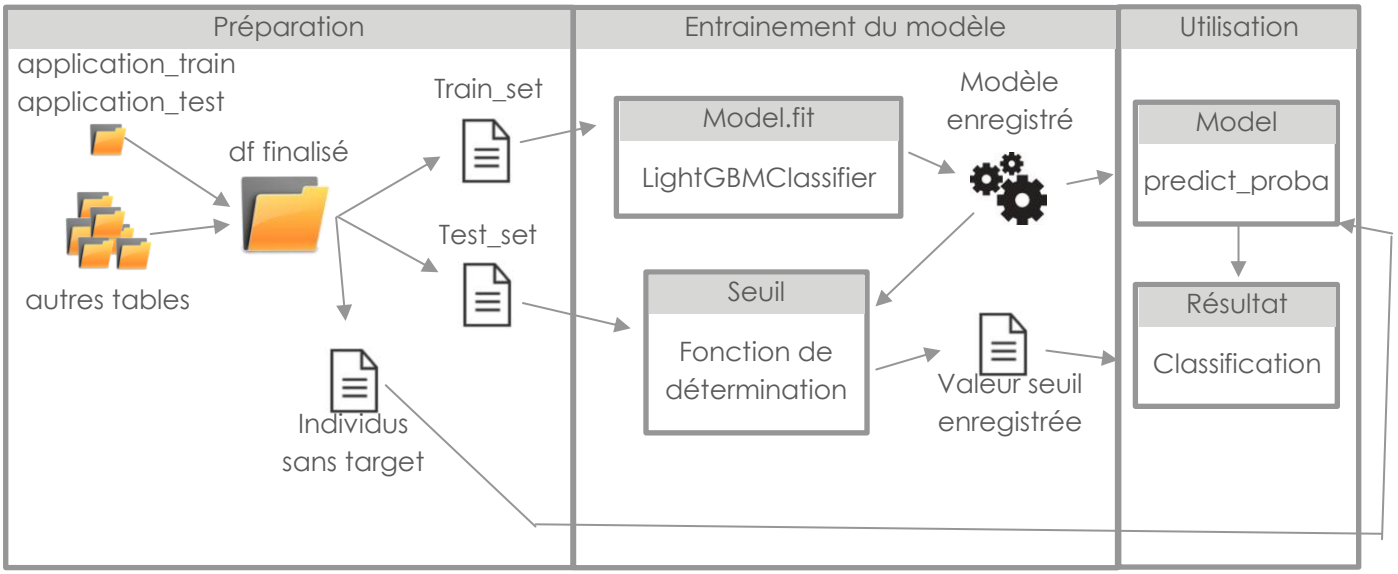
LightGBM.LGBMClassifier

HYPERPARAMETRES:

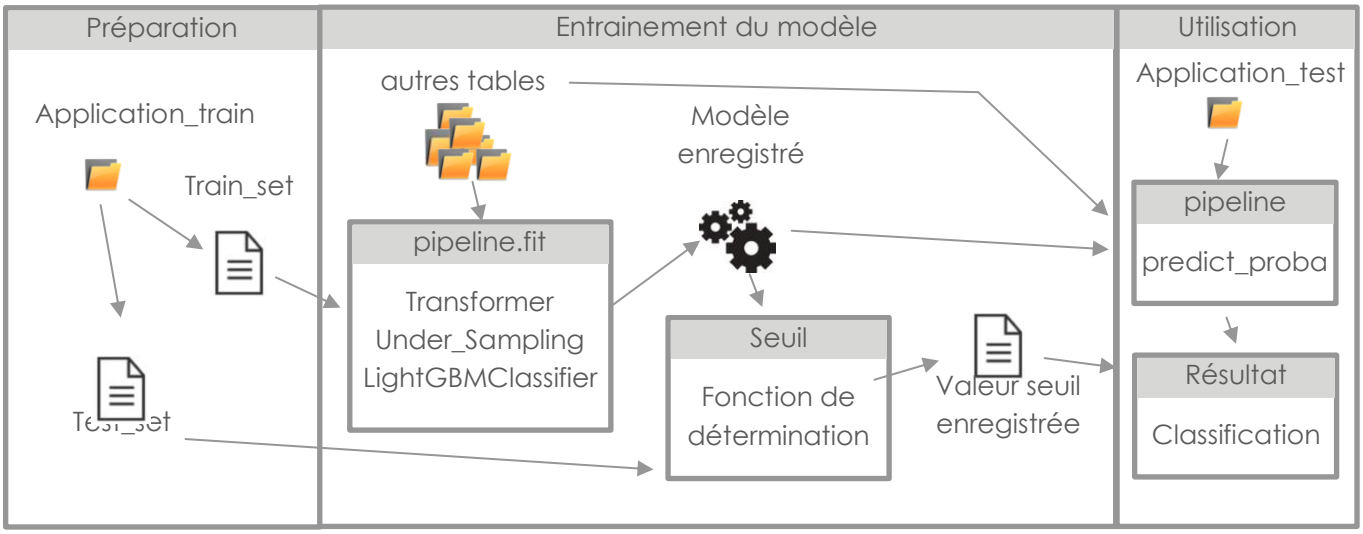
n_estimators	10000	max_depth	8
learning_rate	0.02	reg_alpha	0.041545473
num_leaves	34	reg_lambda	0.0735294
colsample_bytree	0.9497036	min_split_gain	0.0222415
subsample	0.8715623	min_child_weight	39.3259775

METHODOLOGIE D'ENTRAINEMENT DU MODELE 2/2

MODELE N° 1



MODELE N° 2



FONCTION COUT METIER

OBJECTIF : Maximiser les bénéfices de l'entreprise

POSTULAT : gain d'un dossier sans défaut de paiement : +1
perte d'un dossier avec défaut de paiement : -10

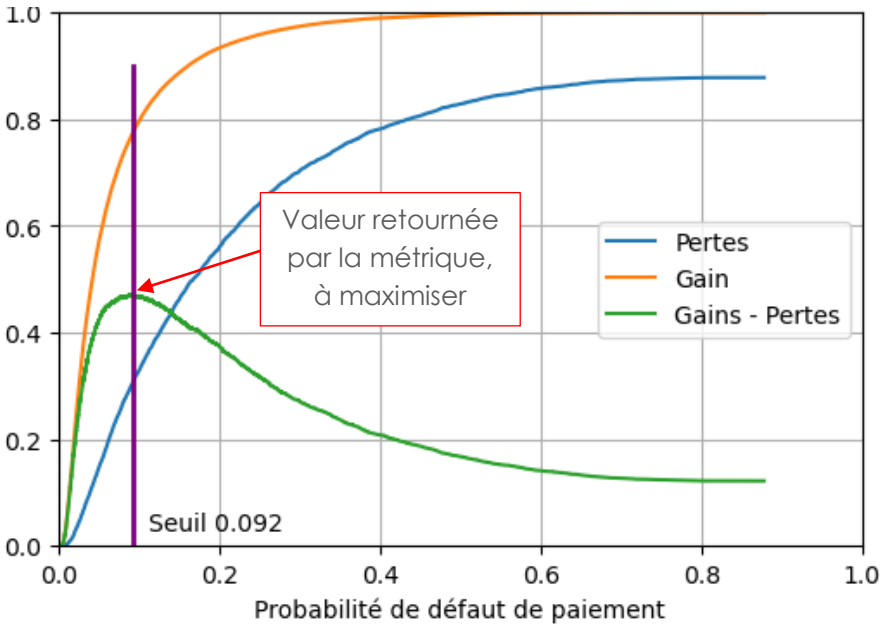
METRIQUE A MAXIMISER :

La métrique reflète le gain financier potentiel en fonction du gain possible avec la totalité des dossiers sans défaut de paiement.

Elle est inférieure ou égale à 1 et peut-être négative.
1 : résultat d'une prédiction parfaite avec tous les dossiers sans défauts réalisés, et aucun dossier avec défaut réalisé.

Sur le graphique suivant :
pertes : somme cumulée du nombre de dossiers avec défauts réels x10
gains : somme cumulée du nombre de dossiers sans défauts x1

DETERMINATION DU SEUIL IMPLIQUANT LE REFUS DU PRET



Plus la courbe verte a un sommet plat, plus le chargé de relation client a une marge de négociation confortable.

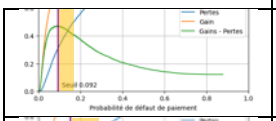
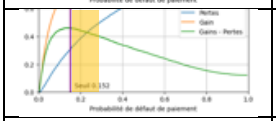
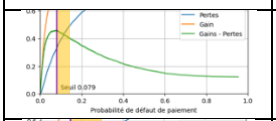
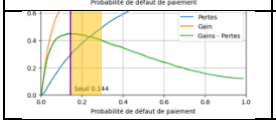
TRAITEMENT DU DESEQUILIBRE DES CLASSES

REPARTITION DES PRETS AVEC ET SANS DEFAUT DE PAIEMENT

Les classes sont fortement déséquilibrées. Néanmoins, l'algorithme LightGBM ne semble pas en être perturbé. Des essais ont été réalisés avec de l'under sampling et de l'over sampling. Les résultats sont légèrement dégradés avec ces traitements. Les temps de traitement pour l'entrainement du modèle sont fortement impactés.

Cependant, avec de l'under sampling, nous obtenons une marge de manœuvre plus importante pour le chargé de relation client. C'est pour cela que l'algorithme RandomUnderSampler de la librairie imbalanced_learn a été ajouté au pipeline utilisé lors de prédictions avec modification des données client.

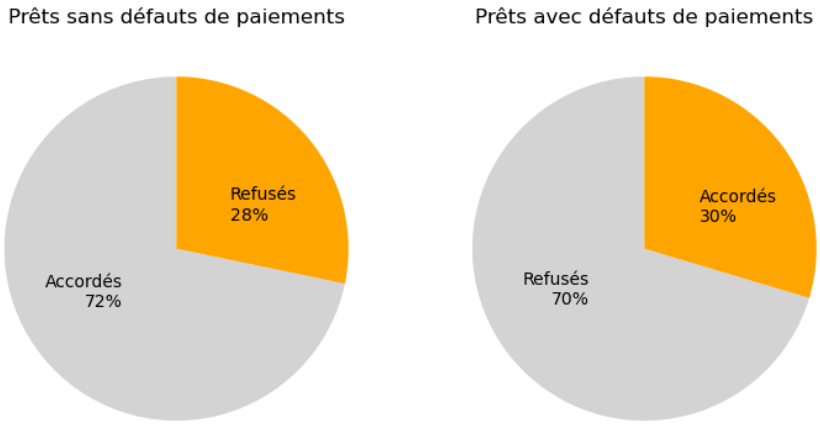
MEILLEURS RESULTATS OBTENUS PAR TYPE DE STRATEGIE D'EQUILIBRAGE DES CLASSES

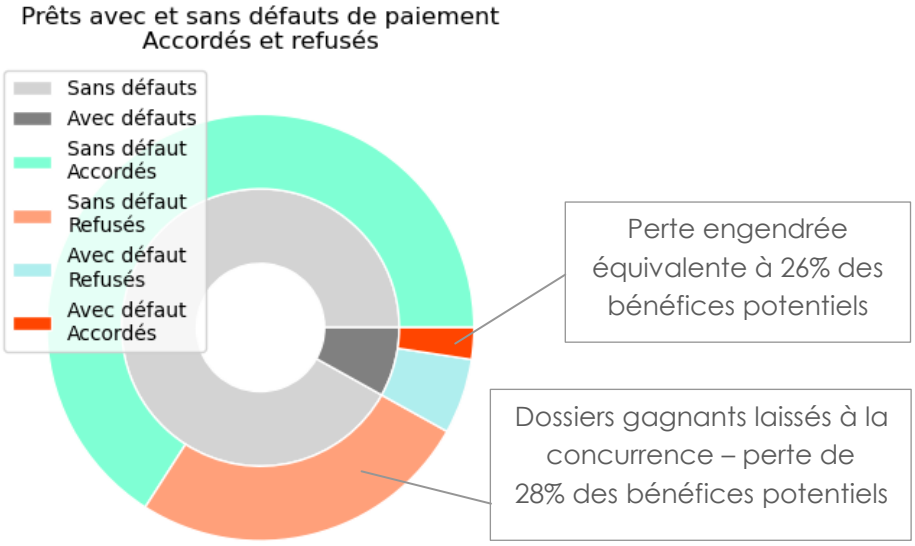
	Under sampling	Over sampling	Classe weight	Métrique métier	Temps d'exécution
	-	-	-	0.470	6 mn
	0.2	-	-	0.461	18 mn
	-	-	balanced	0.434	30 mn
	-	0.2	-	0.458	40 mn
	0.35	0.15	-	0.451	30 mn

Marge de manœuvre pour métrique > 0.4

SYNTHESE DES RESULTATS

La modélisation effectuée permet, si elle est strictement respectée, de gagner **46% des bénéfices potentiels** du marché actuel, en prenant en compte les dossiers refusés à tort et qui auraient permis des gains supplémentaires, et les dossiers acceptés à tort, qui viennent amputer les bénéfices.

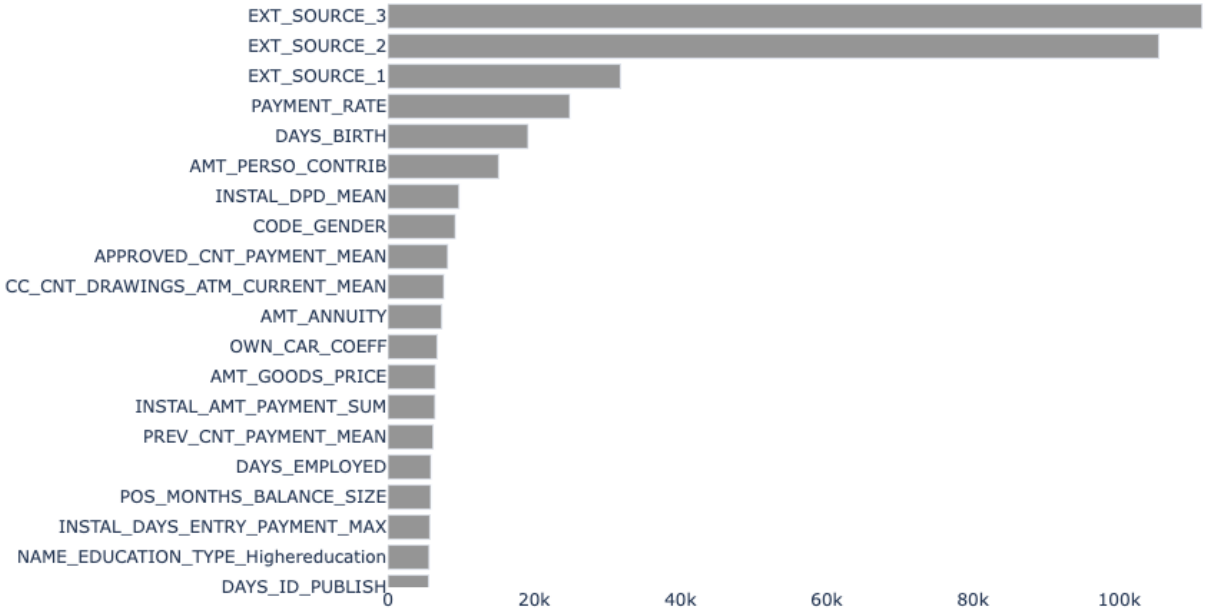




INTERPRETABILITE GLOBALE ET LOCALE DU MODELE

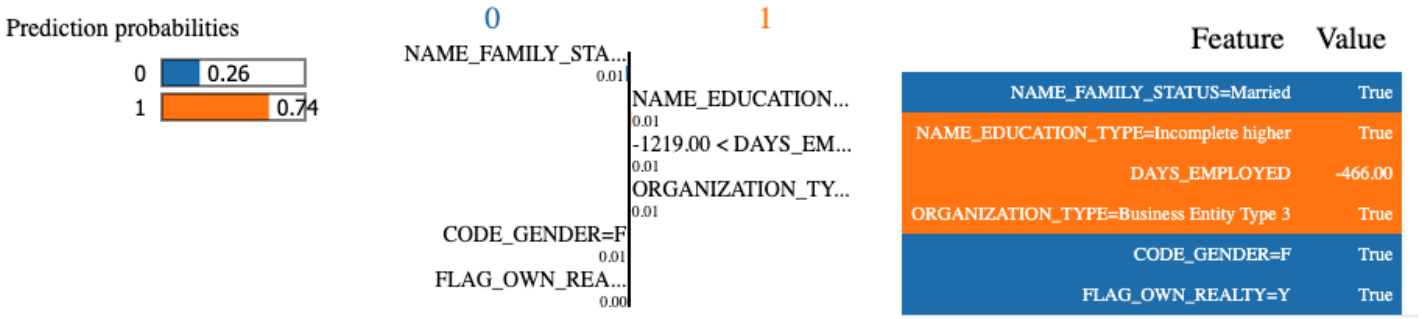
IMPORTANCE GLOBALE : importance d'une variable selon la modélisation, pour LightGBM, l'importance a été paramétrée à 'gain'. Elle correspond donc à la somme des gains apportés par la variable au fil des splits pendant l'entraînement du modèle.

20 PREMIERES VARIABLES SELON LEUR GAIN APPORTE DANS LE MODELE



IMPORTANCE LOCALE : elle correspond à l'importance d'une variable dans la prédiction pour un individu particulier. Elle est donc fonction des données propres à chaque individu.

Nous avons pu la visualiser de façon claire grâce à un pipeline et la librairie **LIME** :



LIMITES ET AMELIORATIONS POSSIBLES

DEUX MODELISATIONS : Le fait d'avoir réalisé deux modélisations peut être un danger. Les dossiers mal classifiés ne sont pas forcément les mêmes sur les deux modélisations. Le chargé de relation client ne doit pas prendre l'habitude de faire tourner les deux modélisations et accorder le prêt si une des deux modélisations classe le dossier en « Accepté ». Autrement nous risquons de sélectionner toutes les réponses fausses négatives* de chacun des modèles et d'accumuler rapidement les pertes.

METRIQUE METIER : La métrique dépend du déséquilibre des classes de par sa formule. Il est donc important de la comparer sur des jeux d'entrainement de même équilibrage. Elle se dégrade de manière naturelle, mais non symptomatique d'une moins bonne modélisation, si le jeu de données présente davantage de dossiers avec défauts de paiement. Pour s'affranchir de ce défaut si l'on rééquilibre les données, il faut tester le modèle et calculer la métrique sur un jeu test de données sans rééquilibrage et n'appliquer le rééquilibrage que sur le jeu d'entrainement, ce qui a été fait ici. Seul point problèmatique ici : l'optimisation des paramètres avec GridSearchCV. Il faudrait donc implémenter une GridSearchCV qui prenne ce point en compte, ou alors réimplémenter une métrique qui n'ait pas ce défaut.

INTERPRETABILITE LOCALE : L'ensemble des variables a été conservé à partir du moment où celles-ci présentaient au moins deux valeurs différentes, puisque l'objectif est d'utiliser de nombreuses sources de données… En conserver uniquement 200 dégradait fortement le score (environ 4% de bénéfices potentiels en moins) et l'idée d'en supprimer a été jusqu'ici écartée. Mais l'interprétabilité locale s'en trouve néanmoins pénalisée, chaque variable ayant une faible part dans la décision.

VARIABLES EXT_SOURCE_X : ces trois variables sont les plus impactantes et nous ne maitrisons pas les calculs effectués. Il faut veiller à ce que ces calculs ne prennent pas eux-mêmes en compte notre propre métrique.

** faux négatif : dossier prédit à tort comme sans défaut de paiement*

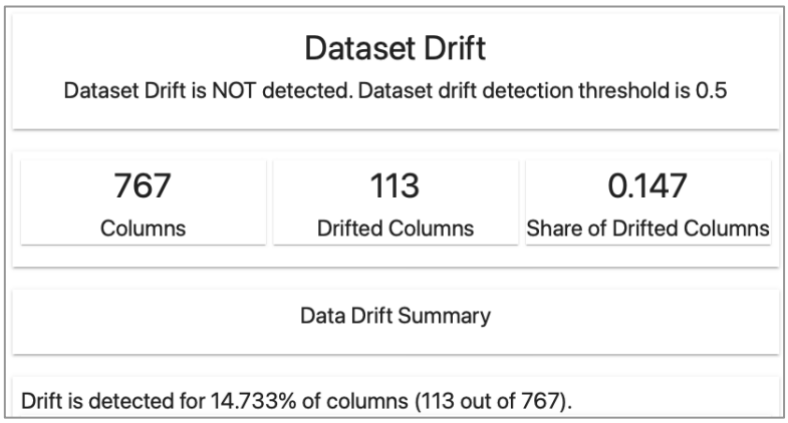
ANALYSE DU DATA DRIFT

Le data drift correspond à une dérive de la distribution des données dans le temps. Cette dérive viendra réduire les performances du modèle et doit donc être détectée lorsqu'elle a lieu.

L'analyse du Data Drift a été effectuée sur le jeu de données pré-traité, afin d'avoir toutes les relations entre les tables établies.

Voici le résultat de l'analyse :

LIBRAIRIE UTILISEE : Evidently



Plus spécifiquement, nous retrouvons du DataDrift sur les données issues de Bureau et Bureau_ balance, qui retracent un historique des anciens prêts contractés. La dérive provient peut-être du fait que nous avons davantage de recul pour les nouveaux prêts et donc plus d'historique. Ces variables ont une importance faible dans la modélisation. Cela ne perturbe pas la modélisation pour l'instant mais reste à surveiller.
Plus inquiétant, nous avons du Data Drift sur la variable Payment_Rate, qui correspond à l'inverse de la durée de paiement en année, et qui est une variable avec une importance globale forte.

Les données utilisées pour l'analyse étaient complètes. Aller plus loin dans l'analyse permettrait de confirmer ou infirmer l'absence de Data Drift impactant. Pour cela il serait préconisé de relancer une analyse avec un DataFrame intermédiaire, où les tables seraient regroupées, mais sans création de nouvelle variable et sans encodage des variables catégorielles.