

Présenté par ABDOULLATUF Maoulida

Aout 2023

Implémentez un modèle de scoring

Soutenance projet 7

Parcours Data Scientist – OpenClassrooms – Centrale Supélec



Table de matières

1. Introduction

- Contexte du projet
- Importance du scoring dans l'entreprise

2. Données et Prétraitement

- Sources de données
- Nettoyage et prétraitement
- Exploration des données

4. Choix du Modèle

- Modèles candidats
- Raisons du choix du modèle spécifique

4. Optimisation du Modèle

- Description du LightGBM et ses avantages
- Métriques utilisées
- Méthode d'optimisation
- Comparaison
- Importance des caractéristiques

7. Data drift

6. Limitations et Défis

- Points faibles du modèle
- Défis rencontrés

Table de matières (suite)

7. Futurs Travaux et Améliorations

- Propositions pour les recherches futures

8. Conclusion

- Résumé des principaux résultats
- Contribution clés
- Améliorations futures
- Questions

Context et problématique

Entreprise “Prêt à dépenser”

Crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt.

Besoin

Modèle de scoring de la probabilité de défaut de paiement du client.

Objectif

Dashboard interactif à destination des chargés de relation client.

Importance du scoring

1. **Evaluation objectives des emprunteur**

Le scoring permet d'évaluer objectivement et systématiquement la capacité de remboursement d'un client

2. **Réduction du Risque de Crédit**

Le model de scoring aide à identifier les emprunteurs à faible risque et à réduire ainsi le risque de défaut de paiement.

3. **Optimisation des décision**

Le scoring permet de prendre de décision plus rapide et mieux informées sur l'octroi de crédits

4. **Automatisation du Processus**

le scoring permet d'automatiser une partie du processus de décision, ce qui réduit les coûts opérationnels.

5. **Amélioration de la Rentabilité**

En réduisant les taux de défaut et en optimisant l'allocation des ressources, le scoring peut contribuer à améliorer la rentabilité de l'entreprise.

...

Source et prétraitement des données:

Kaggle « Home Credit Default Risk »: <https://www.kaggle.com/c/home-credit-default-risk/data>

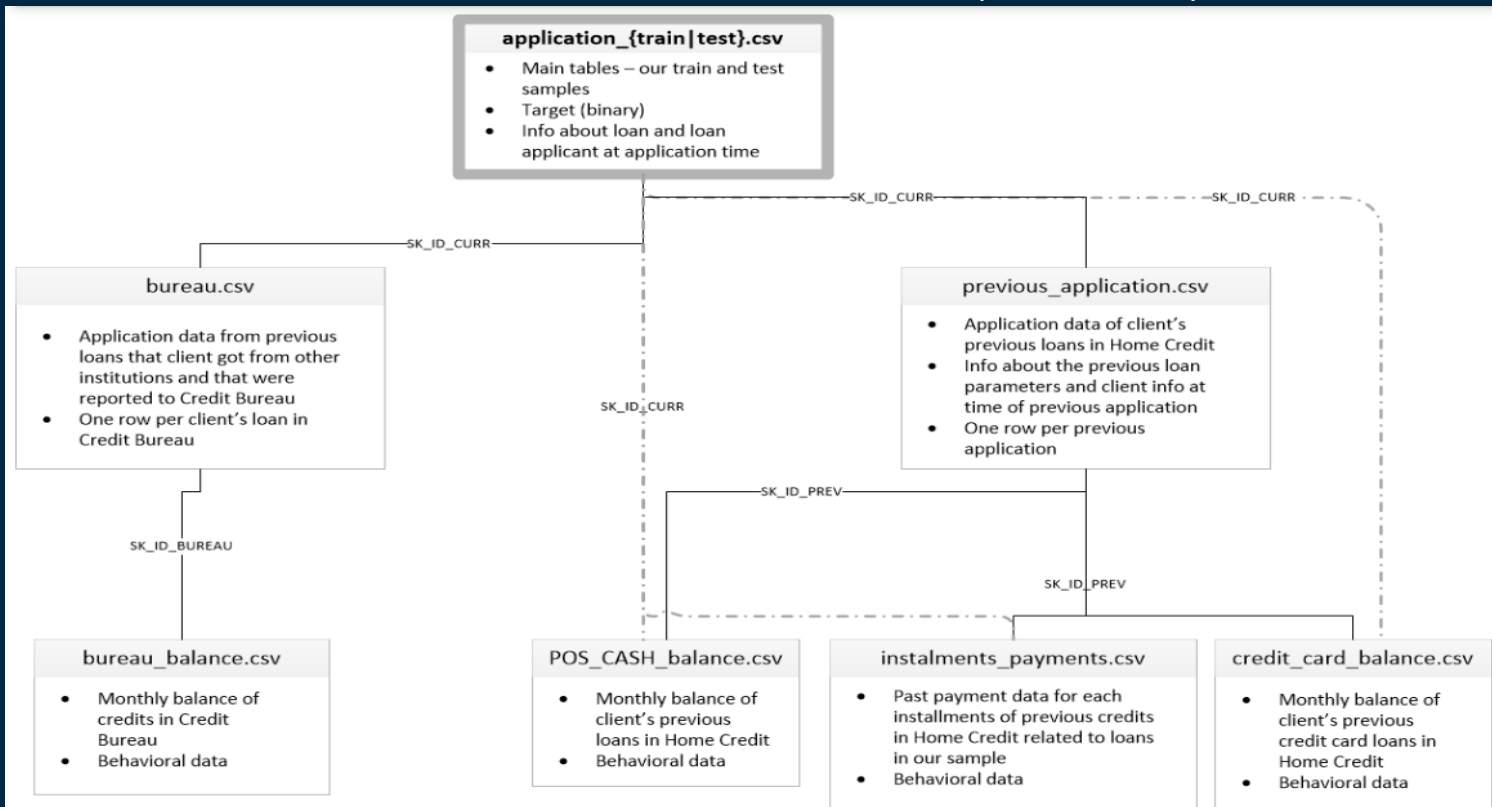
Description rapide des données

	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/POS_CASH_balance.csv	10001358	8	0.07	0.0	1	2	5	0	610.435
./data/credit_card_balance.csv	3840312	23	6.65	0.0	1	15	7	0	673.883
./data/installments_payments.csv	13605401	8	0.01	0.0	0	5	3	0	830.408
./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/previous_application.csv	1670214	37	17.98	0.0	16	15	6	0	471.481
./data/bureau_balance.csv	27299925	3	0.00	0.0	1	0	2	0	624.846
./data/sample_submission.csv	48744	2	0.00	0.0	0	1	1	0	0.744

Présentation du jeu de données:

Kaggle « Home Credit Default Risk »: <https://www.kaggle.com/c/home-credit-default-risk/data>

Schéma de la base de données (source Kaggle)

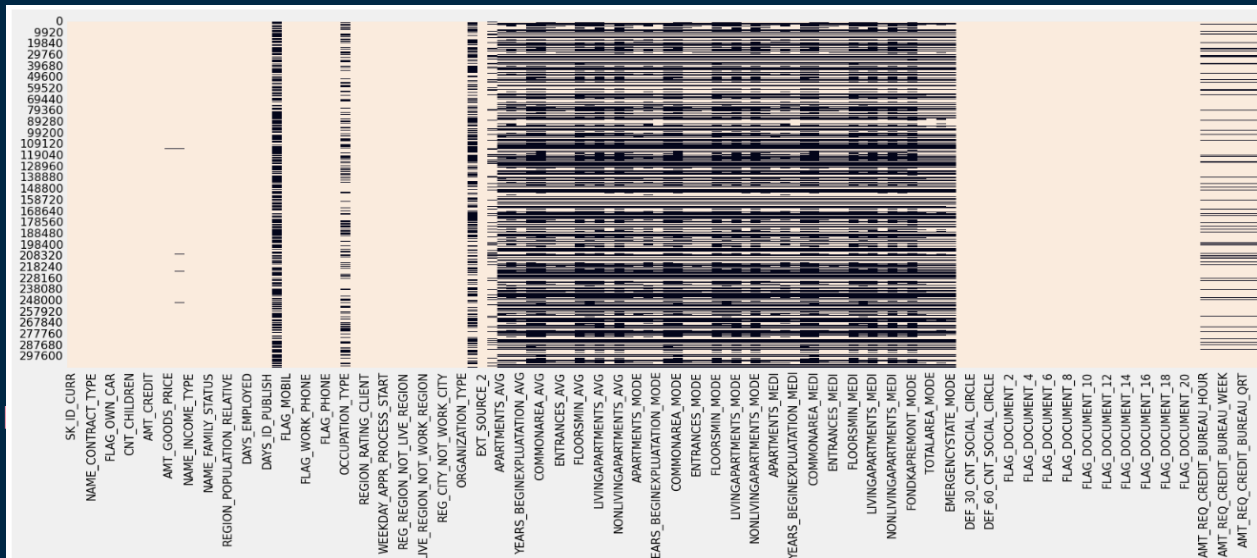


Nettoyage des données



Objectif: Garantir des données de haute qualité pour une meilleure prédiction et modélisation

- **Valeur manquante:** Imputation par moyenne et valeurs plus fréquentes
- **Valeur aberrantes:** Détection et correction
- **Normalisation:** Standardisation
- **Encodage:** One hot encoding

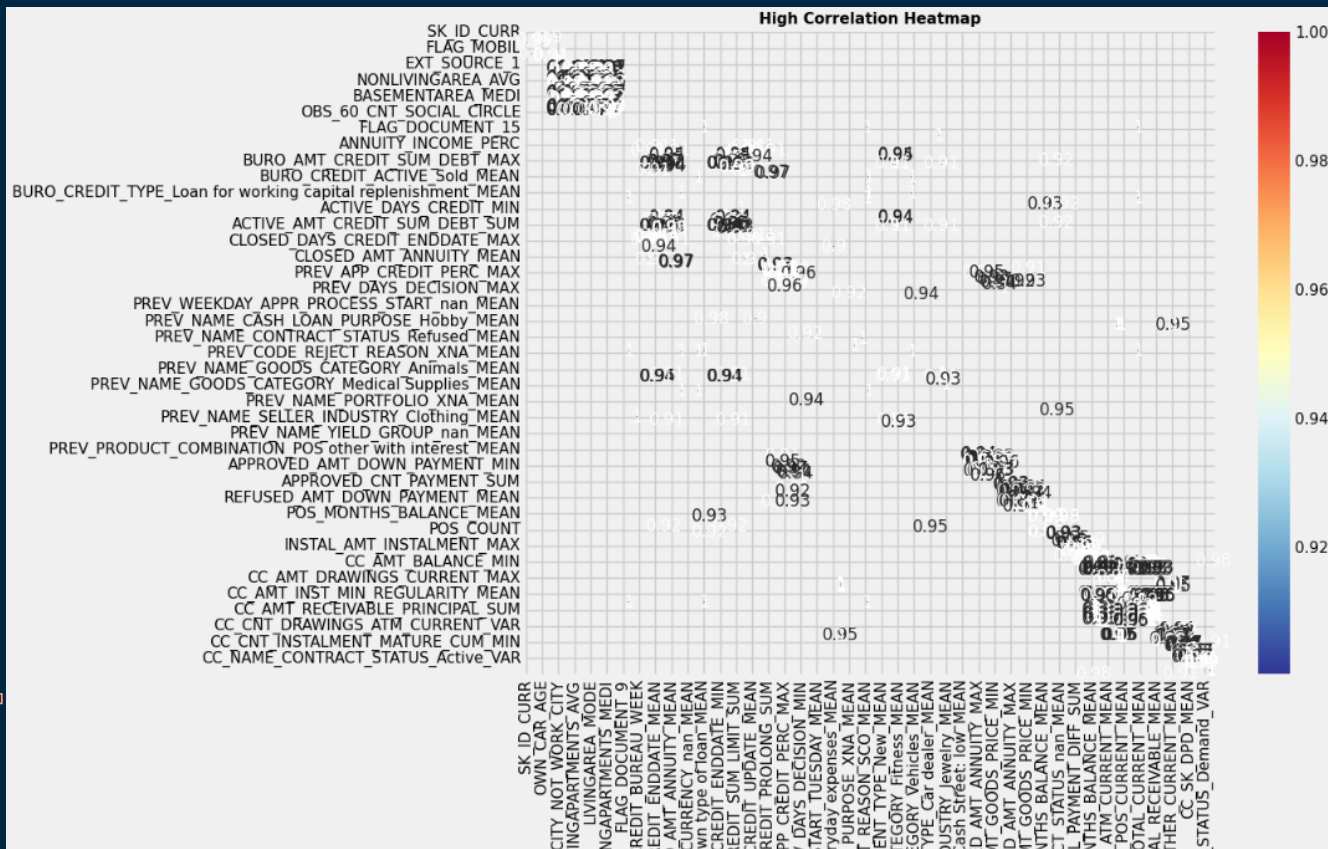


TOP 10 Missing values from Training dataset

	Total	%
COMMONAREA_MEDI	214865	69.87
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_AVG	213514	69.43
NONLIVINGAPARTMENTS_MEDI	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_MODE	210199	68.35
LIVINGAPARTMENTS_AVG	210199	68.35
LIVINGAPARTMENTS_MEDI	210199	68.35

Nettoyage des données

Corrélations des caractéristiques



Analyse exploratoire des données:

Opération de merging – feature engineering :

Enrichissement de l'échantillon de travail: Combinaison des 7 jeux de données

Avant 122 features – Après 797 features

Dont 3 features de moyenne et de comptage:

- PREVIOUS_LOANS_COUNT: nombre des précédent crédits pris par le client
- MONTHS_BALANCE_MEAN: solde mensuel moyen des précédents credits
- DAYS_EMPLOYED_PERCENT: % jours employés par rapport à l'âge du client
- PREVIOUS_APPLICATION_COUNT: nombre de demandes antérieures au crédits immobilier

Analyse exploratoire des données:

Préprocessing – opération de merging – feature engineering :

Enrichissement de l'échantillon de travail par 4 ratios explicatifs:

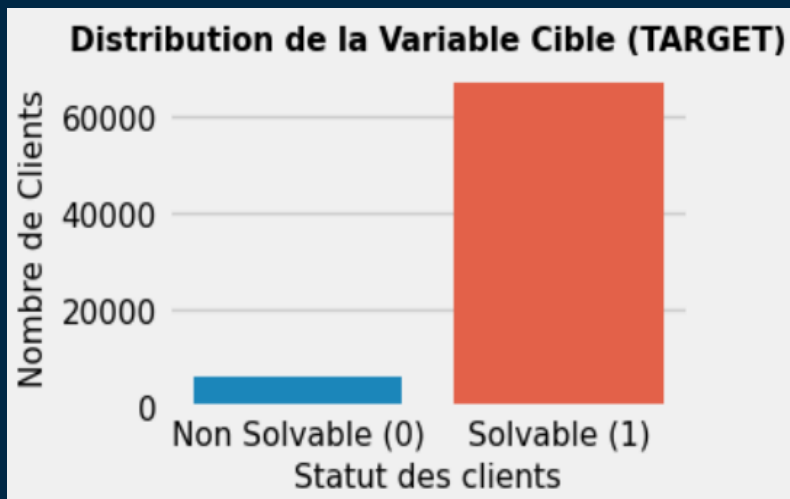
- CREDIT_INCOME_PERCENT: % montant du crédit par rapport au revenu d'un client
- ANNUITY_INCOME_PERCENT: % rente de prêt par rapport au revenu d'un client
- DAYS_EMPLOYED_PERCENT: % jours employés par rapport à l'âge du client
- CREDIT_TERM: durée du paiement en mois

Echantillon obtenu: 496 features obtenus après pré-traitement, merging des data sets et suppression des variables corrélées à plus de 90%.

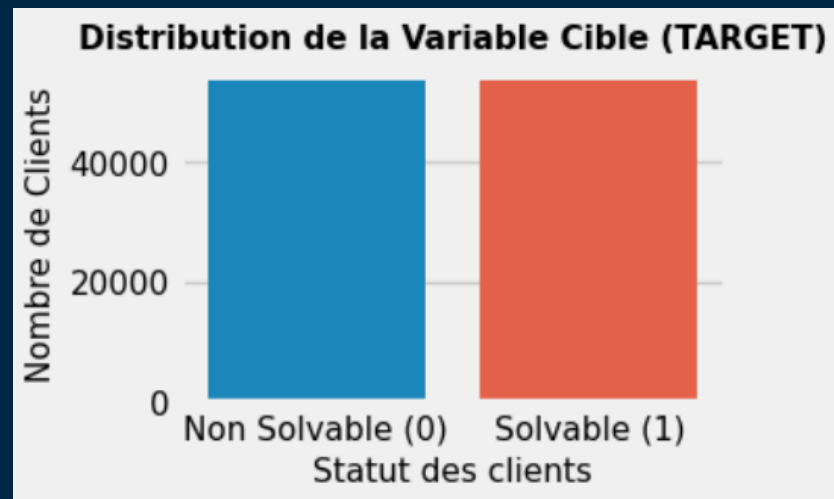
Exploration des données

Prétraitement : Distribution de la 'TARGET'

Avant

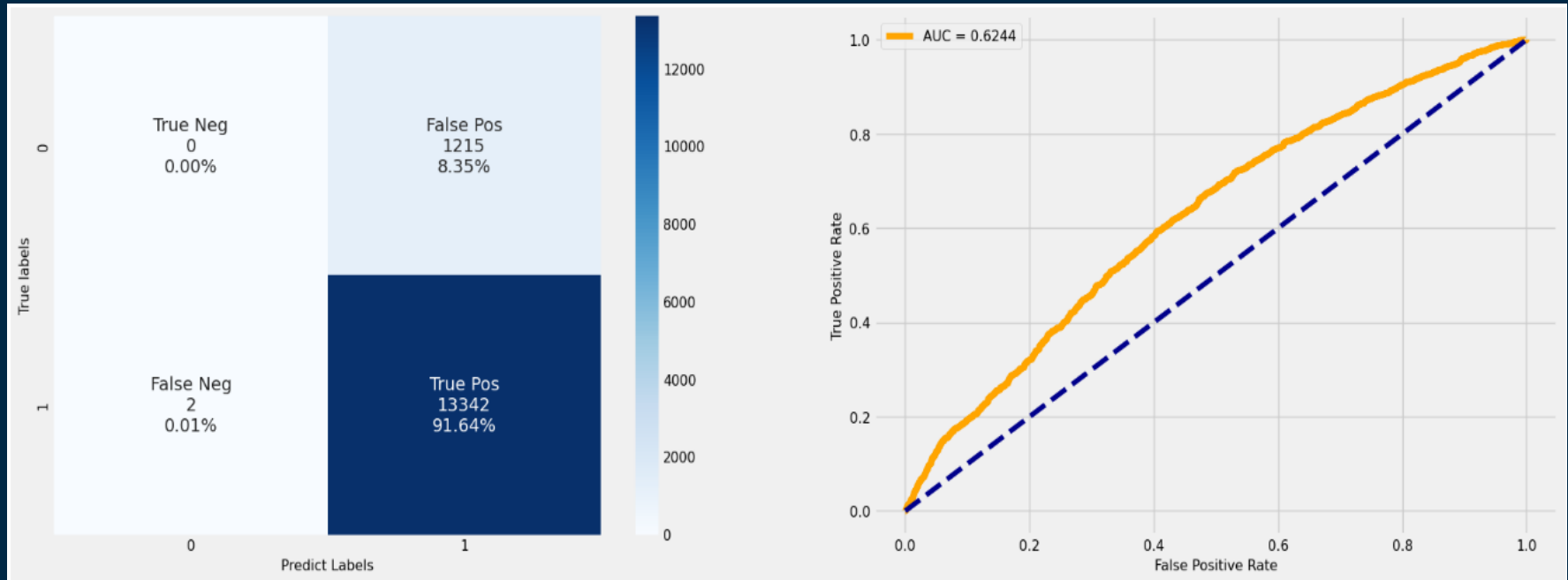


Après rééquilibrage



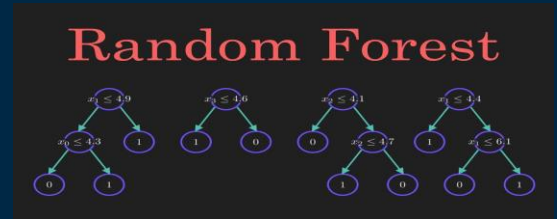
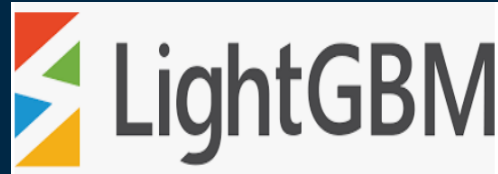
Oversampling par smote

Modélisation: Baseline fixée par Logistic Regression



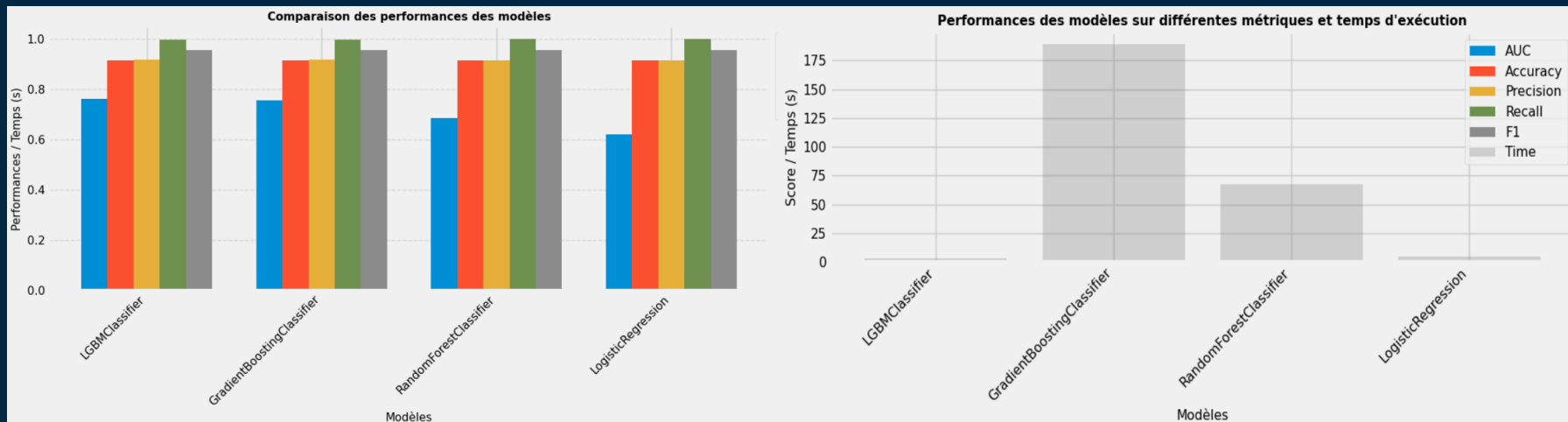
Performance de la « baseline » avec toutes les features

Modèles candidats



Sélection du modèle (avec les paramètres par défaut)

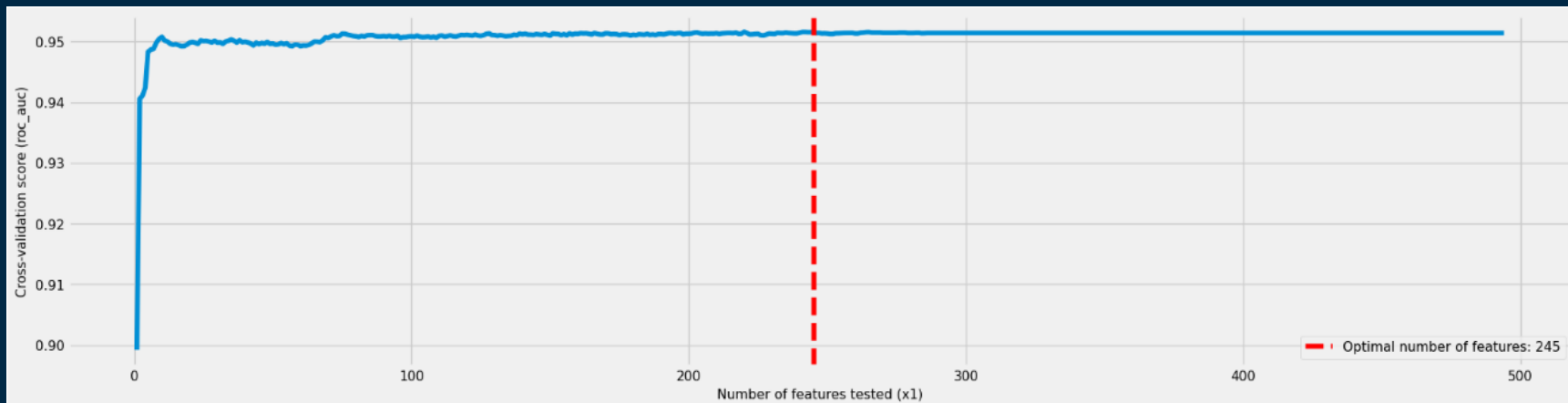
Après comparaison, LGBMClassifier se distingue comme le modèle optimal, avec un AUC de 0,76%



	Model	AUC	Accuracy	Precision	Recall	F1	Time
1	LGBMClassifier	0.76256	0.916684	0.918397	0.997752	0.956431	2.845178
0	GradientBoostingClassifier	0.756312	0.916546	0.917349	0.998951	0.956413	189.621204
2	RandomForestClassifier	0.68525	0.916615	0.916609	1.0	0.956491	67.863091
3	LogisticRegression	0.621593	0.91634	0.916587	0.9997	0.956341	4.393925

Feature sélection: Réursive Feature Elimination: RFECV

Identification des meilleurs features par validation croisée en optimisant la métrique AUC



RFECV (496 features) → 245 features

Avec RFECV, nous avons retenu 245 caractéristiques, améliorant ainsi la performance et la rapidité du modèle

Notre métrique Métier

```
def custom_metric(y_true, y_pred):  
    tn, fp, fn, tp = confusion_matrix(y_true, y_pred).ravel()  
    return 0.9 * fp / (tn + fp) + 0.1 * fn / (tp + fn)  
  
credit_score_3 = make_scorer(custom_metric, greater_is_better=False)
```

Cette métrique assure une meilleure évaluation de la performance en se concentrant sur:

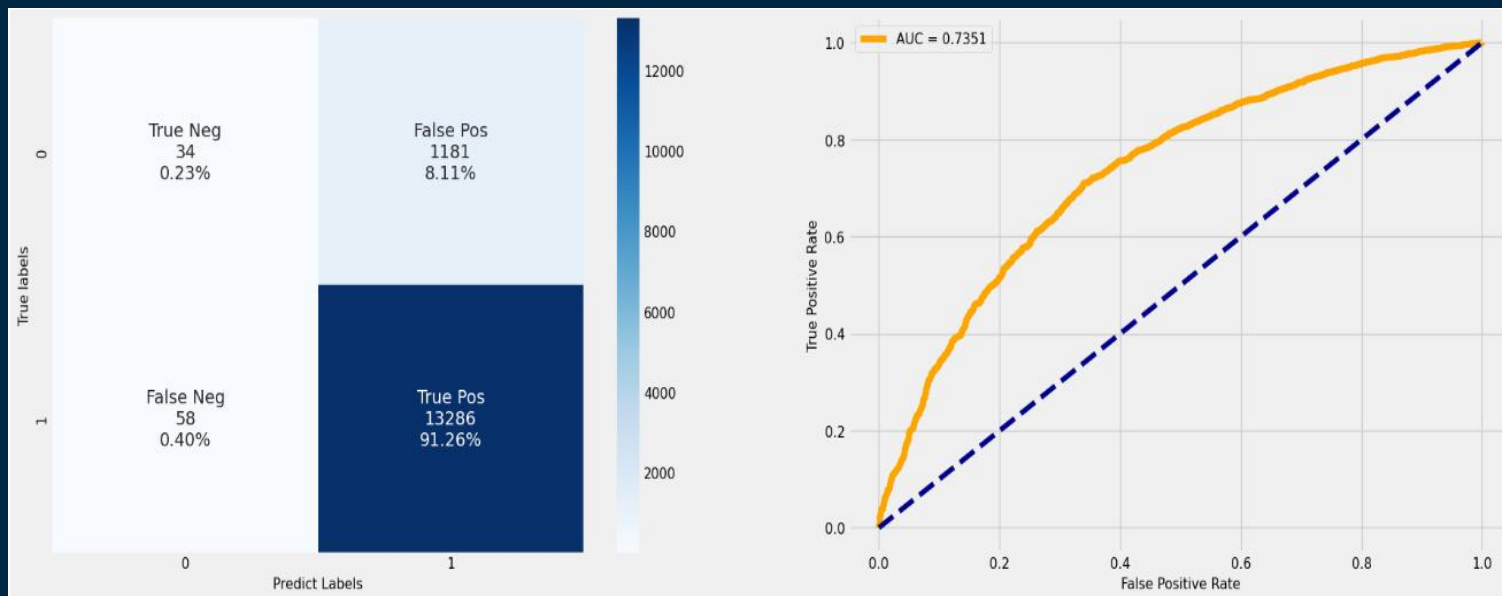
- ❑ **Limitation des risques de perte financière: Pénaliser les faux positifs et les faux négatifs**
- ❑ **Importance relatif entre Recall et Precision:**
 - Estimation du coût moyen d'un défaut de paiement
 - Estimation du coût d'opportunité d'un client refusé par erreur

Important: Connaissance métier nécessaire ou hypothèse à fixer!

Optimisation des Hyperparamètres

Choix d'une méthode avancée: HyperOpt

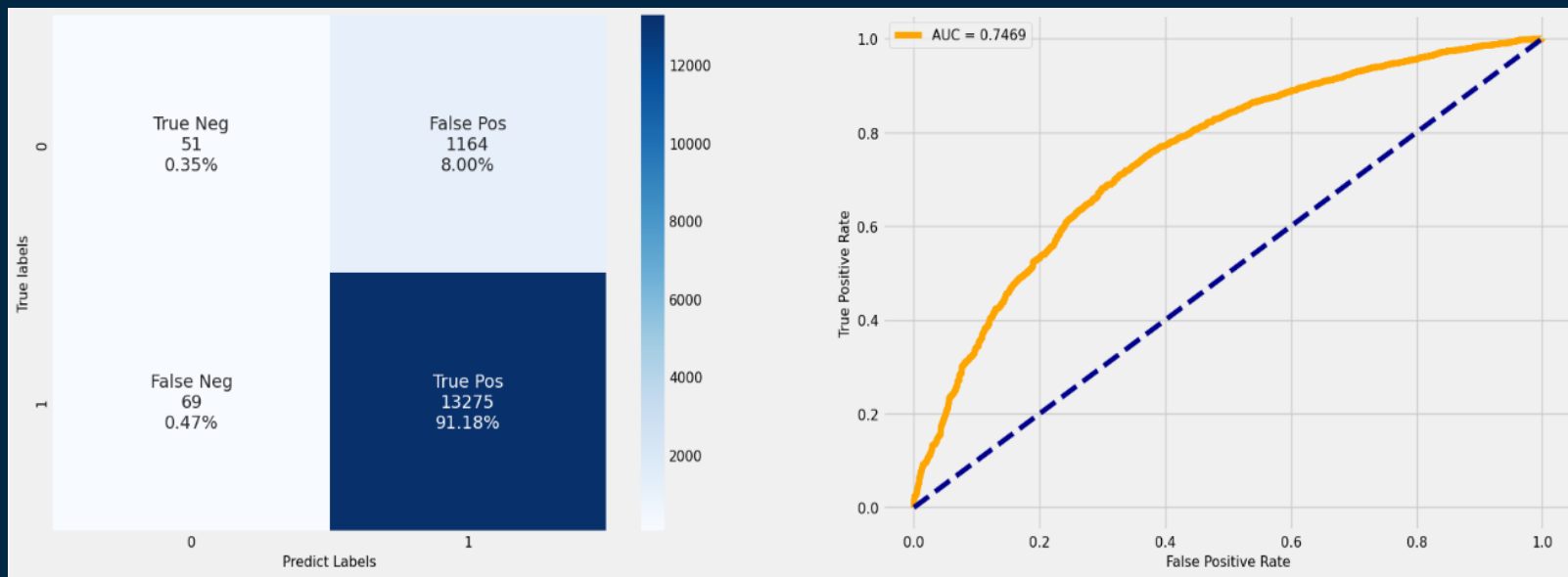
Une méthode d'optimisation bayésienne, a été utilisé pour affiner les hyperparamètres du LGBMClassifier en utilisant notre métrique métier



Avant l'optimisation

Optimisation des Hyperparamètres (suite)

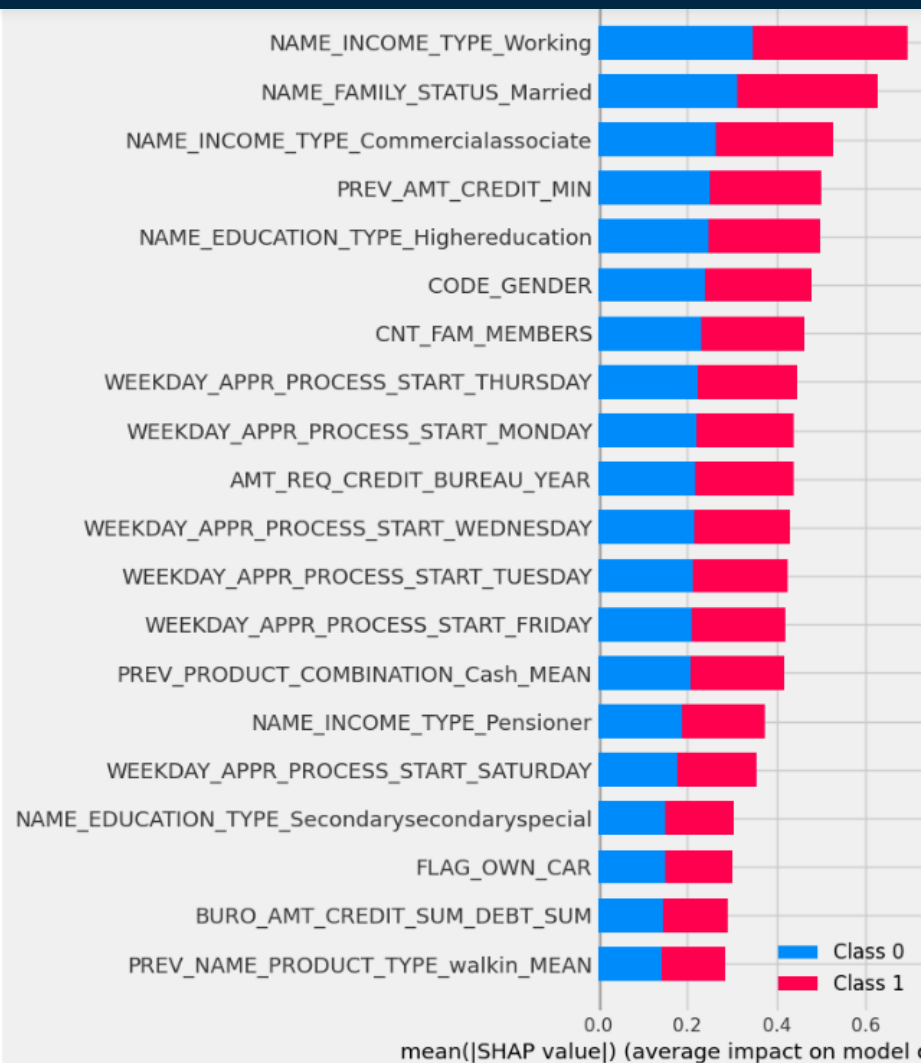
Grâce à l'optimisation, nous avons pu augmenter l'efficacité du LGBMClassifier, le rendant plus précis et adapté à nos besoins métier.



Après l'optimisation

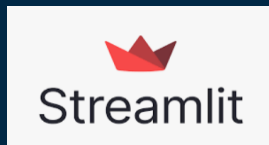
Importance des caractéristiques avec SHAP

Après optimisation du modèle, la caractéristique la plus influente est 'NAME_INCOME_TYPE_Working' suivi de 'NAME_FAMILY_STATUS_Married' et 'NAME_INCOME_TYPE_Commercialassociate',



Présentation du Dashboard

Dashboard interactif
avec Streamlit:



Analyse locale et globale de vos clients

Sélectionner l'identifiant du client

396535

	INUITY_INCOME_PERCENT	CREDIT_TERM	DAYS_EMPLOYED_PERCENT	TARGET	proba_1	prediction
673	0.1556	0.0486	0.1333	0	0.3838	0

Le client 396535 n'est pas éligible à un prêt. Sa probabilité de paiement est faible, elle est de 38.38%. Le seuil optimal est de 52%.

Interpretabilité du modèle : Quelles variables sont les plus importantes ?

☐ Interpretabilité du modèle

Analyse comparative entre le clients courant et les classes :

Présentation du Dashboard

Intégration avec FastApi:



Name	Description
id_client * required integer (path)	<input type="text" value="268858"/>
<div>Execute</div> <div>Clear</div>	
Responses	
Curl	
<pre>curl -X 'GET' \ 'https://fast-api-dashboard-final.onrender.com/predict/268858' \ -H 'accept: application/json'</pre>	
Request URL	
<pre>https://fast-api-dashboard-final.onrender.com/predict/268858</pre>	
Server response	
Code	Details
200	<div>Response body</div> <pre>{ "probaClasse1": 0.9757413948744945, "probaClasse0": 0.024258605125505484, "prediction": 1 }</pre> <div>Download</div>

Surveillance du Data drift (avec Evidently)

La dérive des données est un phénomène où le modèle perd de la précision en raison de changements dans les données d'entrées au fil de temps

Dataset Drift

Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5

245
Columns









16
Drifted Columns

0.0653
Share of Drifted Columns

Data Drift Summary

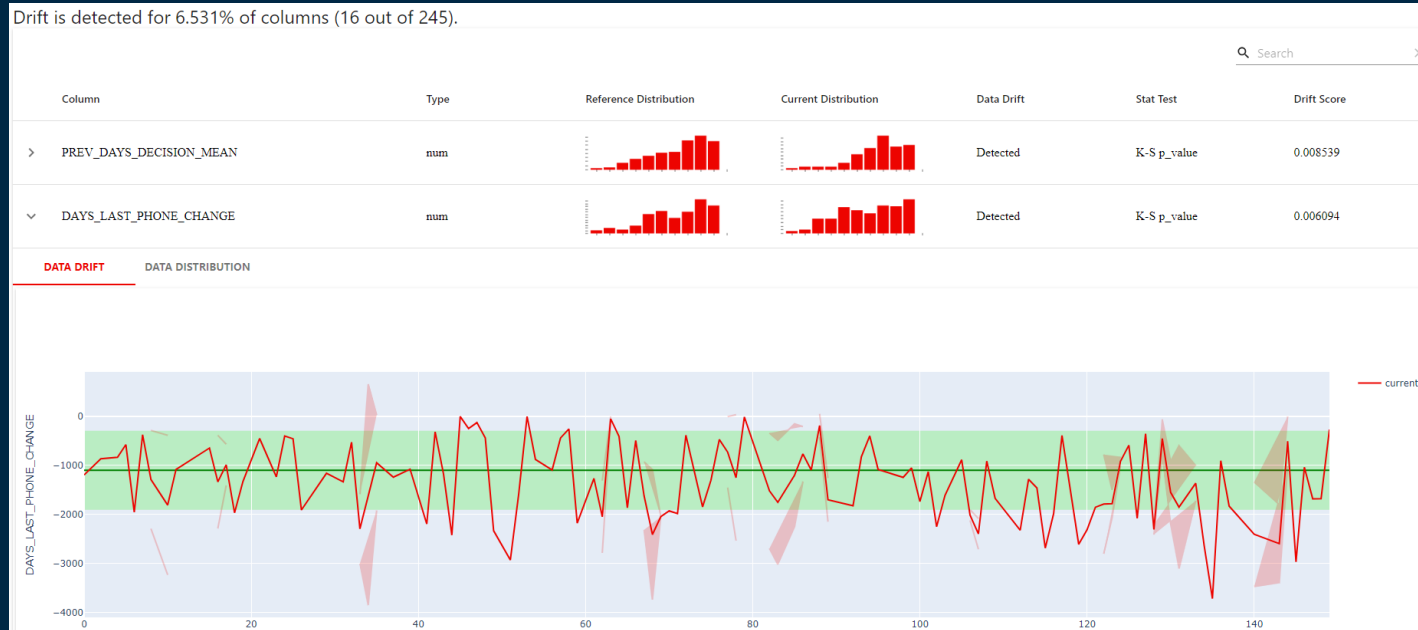
Drift is detected for 6.531% of columns (16 out of 245).

Search

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> CC_NAME_CONTRACT_STATUS_Completed_MAX	num			Not Detected	Z-test p_value	0.117047
> POS_NAME_CONTRACT_STATUS_Active_MEAN	num			Not Detected	K-S p_value	0.112284
> PREV_CNT_PAYMENT_SUM	num			Not Detected	K-S p_value	0.112284
> PREV_NAME_PORTFOLIO_Cash_MEAN	num			Not Detected	K-S p_value	0.112284

Surveillance du Data drift

Sur certaines colonnes une dérive a été détectée et sur d'autre (une grande partie) aucune dérive n'a été détectée. Par exemple, la colonne « PREV_DAYS_DECISION_MEAN" a été détectée comme ayant subi une dérive avec une valeur de K-S p_value de 0,008, tandis que la colonne "CC_NAME_CONTRACT_STATUS_Completed_MAX" n'a pas été détectée comme ayant subi une dérive avec une valeur de Z-test p_value de 0,117047.

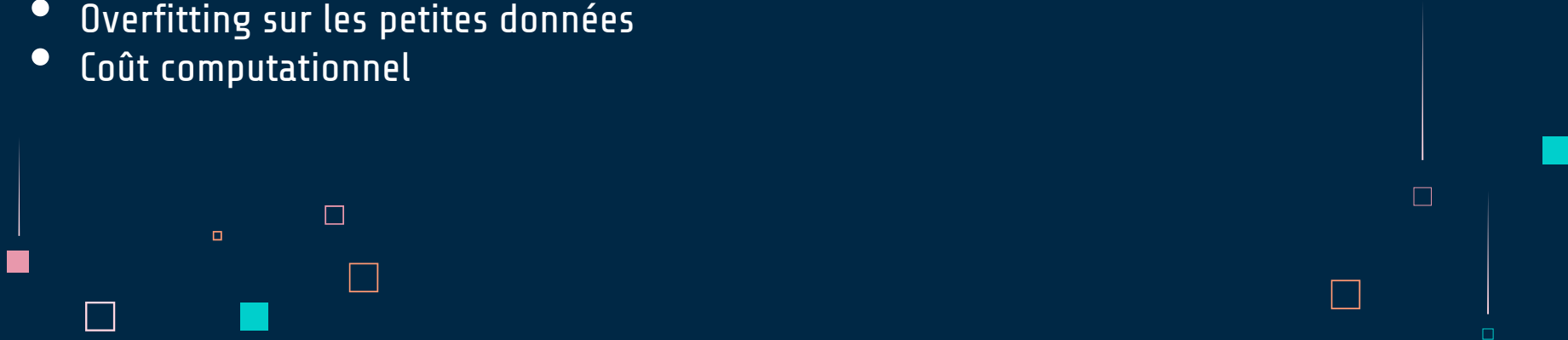


Limitation du LGBMClassifier



Le Light Gradient Boosting Machine (LGBM) est un algorithme d'apprentissage ensembliste efficace et populaire, mais comme tout modèle, il a ses limites. Voici quelques unes que nous avons rencontrées:

- Sensibilité aux données déséquilibrées
- Hyperparamétrage
- Overfitting sur les petites données
- Coût computationnel



Améliorations futures et étapes suivantes

○ **Modélisation avancée:**

- Exploration d'autre algorithme,
- technique d'ensemble pour une prédiction plus robuste,
- réévaluation périodique du modèle, ...

○ **Optimisation des features:**

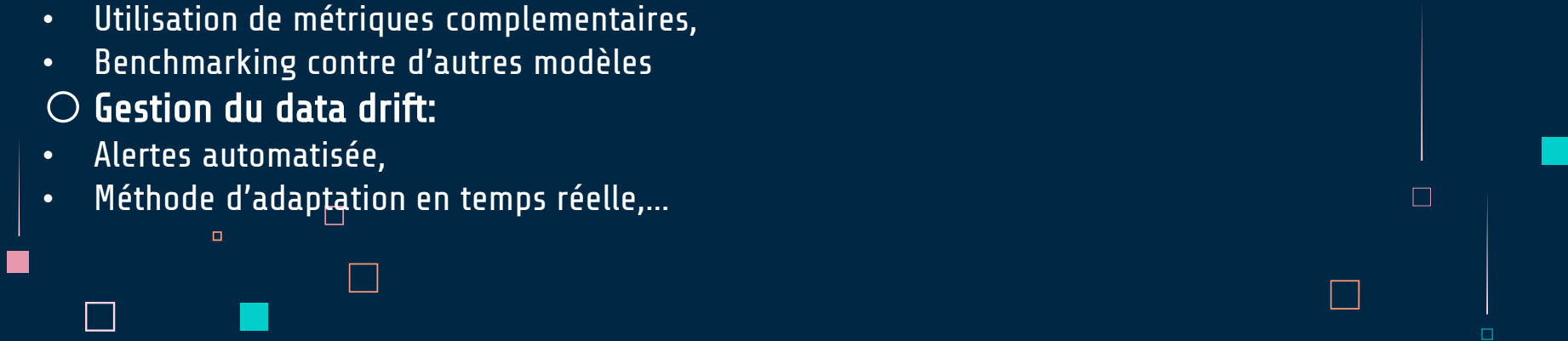
- Ingénierie des caractéristiques avancée,
- nettoyage des données plus sophistiqué, ...

○ **Métrique et évaluation:**

- Utilisation de métriques complémentaires,
- Benchmarking contre d'autres modèles

○ **Gestion du data drift:**

- Alertes automatisée,
- Méthode d'adaptation en temps réelle,...



Conclusion:

Résumé:

- Nous avons développé un modèle prédictif en utilisant le LGBMClassifier, optimisé via Hyperopt, et validé avec des métriques métier spécifiques

Contributions clés:

- Le modèle s'appuie sur une sélection rigoureuse des fonctionnalités et une optimisation d'hyperparamètres.
- Un dashboard a été créé pour une utilisation facile, et le modèle est déployé en ligne
 - Nous avons également mis en place un système pour détecter le data drift.

Améliorations Futures:

- Des améliorations sont déjà envisagées pour rendre le modèle plus robuste et efficace

Remerciements:

- Je tiens à remercier mon mentor pour son aide précieuse dans la réalisation de ce projet.

Questions:

- Je suis maintenant prêt à répondre vos questions.

Merci de votre attention, vous pouvez
me contacter à

a.ellatuf@gmail.com

(+33) 7 82 86 97 64

Merci!