

29/12/2024

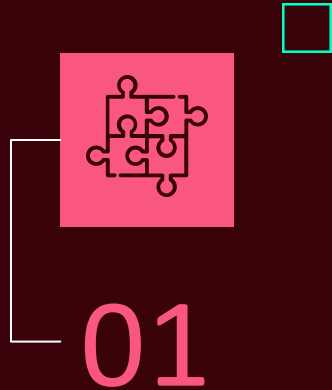
Implémentez un modèle de scoring

Mentor : Nassim LOUATI

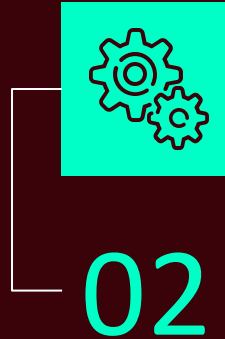
Examineur : Sitou AFANOU

Élève : Aymen TLILI

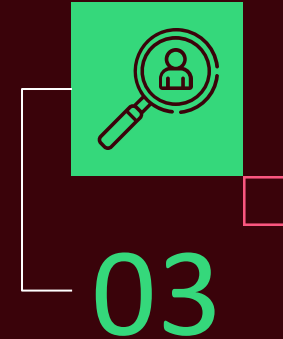
Sommaire



Problématique

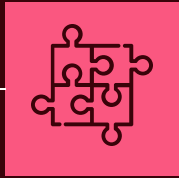


Analyse exploratoire



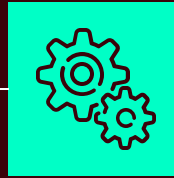
Modélisation

Sommaire



04

Interpretation des
résultats



05

Dashboard



06

Conclusion

Problématique

01

Problématique

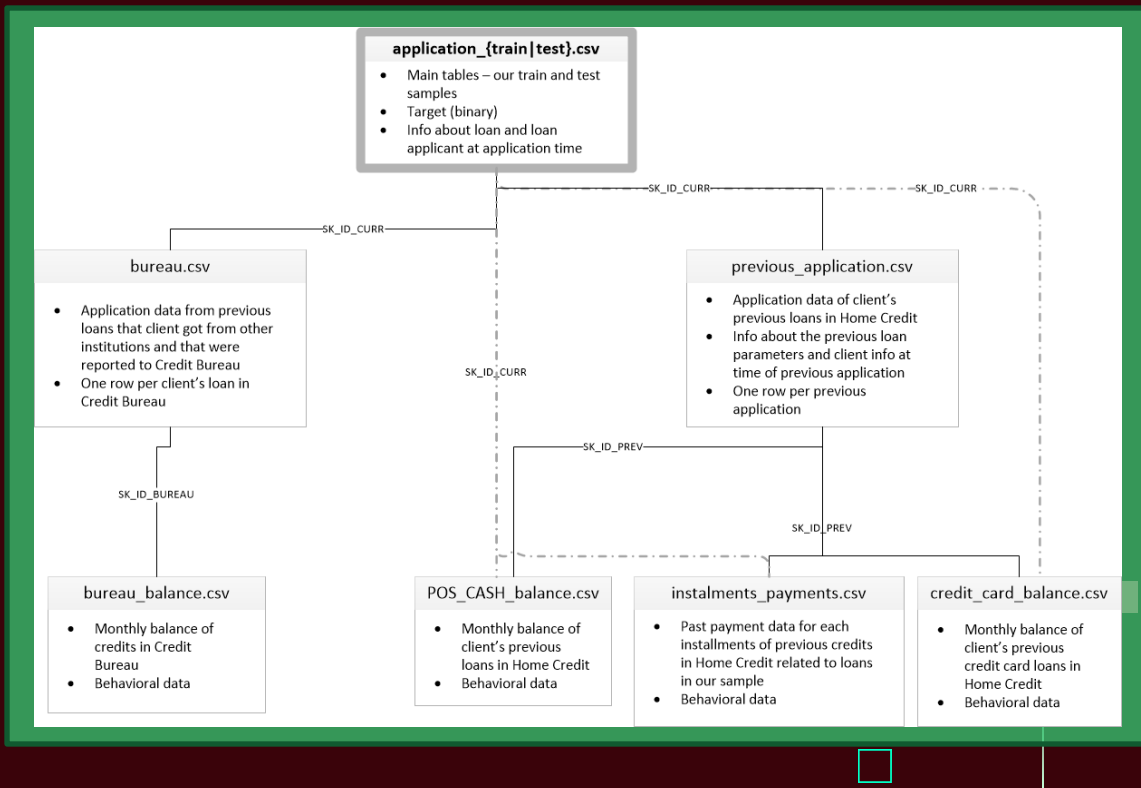
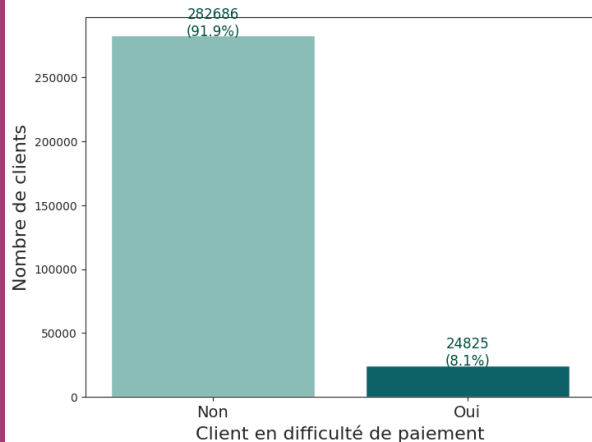
Prêt à dépenser, une société financière, propose des crédits à des clients avec peu ou pas d'historique de prêt. L'objectif est de développer un outil de scoring crédit capable de prédire la probabilité de remboursement pour accorder ou refuser un crédit. Face à une demande croissante de transparence des décisions d'octroi, un **dashboard interactif** sera conçu pour expliquer clairement les décisions aux chargés de relation client et permettre aux clients d'accéder et d'explorer facilement leurs informations personnelles, en phase avec les valeurs de l'entreprise.



Problématique

- Application "train" : 307 511 clients avec Target
- Application "test" : 48 744 clients sans information sur la décision d'octroi

Répartition des clients selon leur situation de paiement

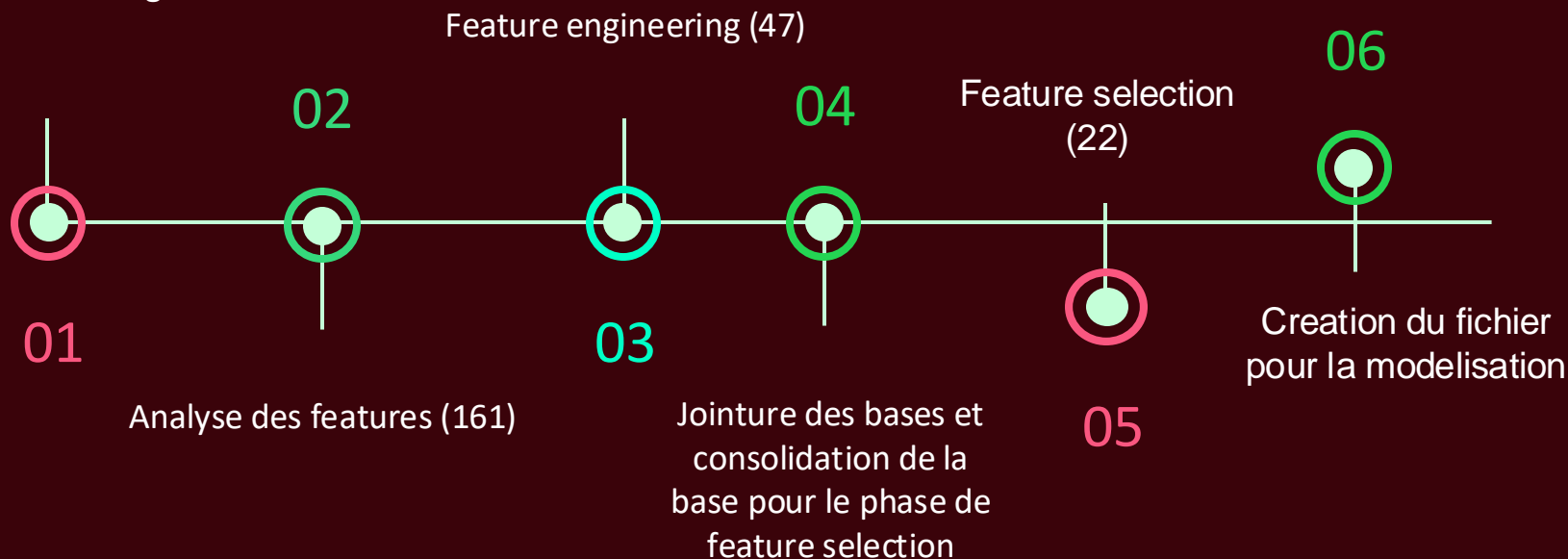




Analyse
exploratoire

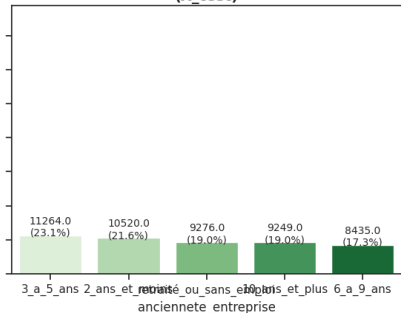
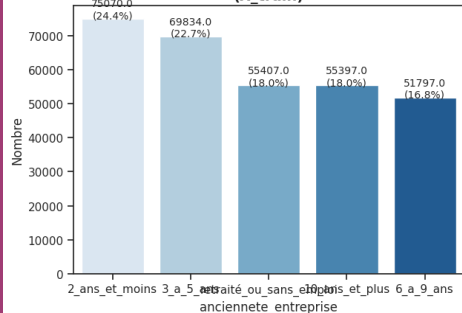
Analyse exploratoire

Analyse de la target

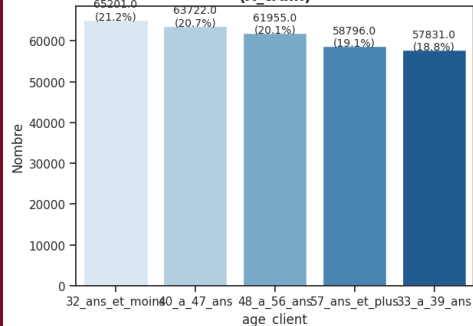


Analyse exploratoire

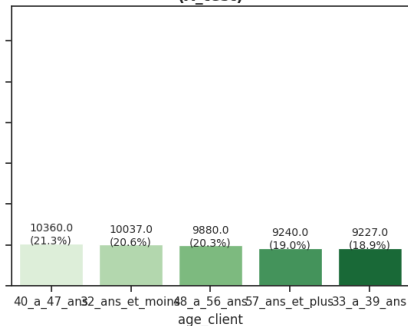
Répartition de l'ancienneté en entreprise dans le train (X_train)



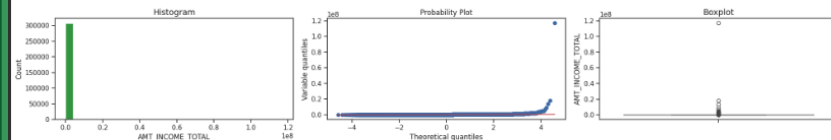
Répartition de l'age dans le train (X_train)



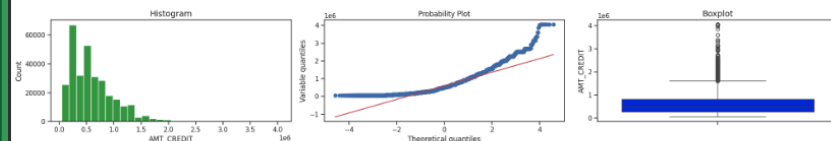
Répartition de l'age dans le test (X_test)



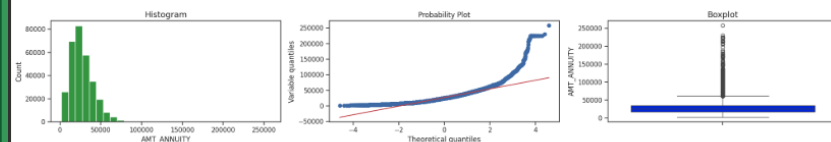
Analyse de la distribution de AMT_INCOME_TOTAL



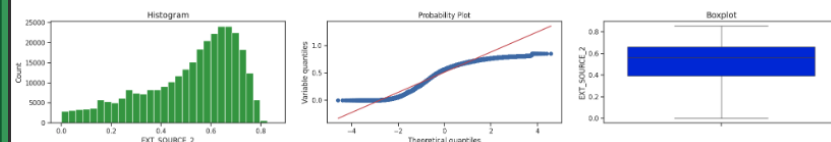
Analyse de la distribution de AMT_CREDIT



Analyse de la distribution de AMT_ANNUITY

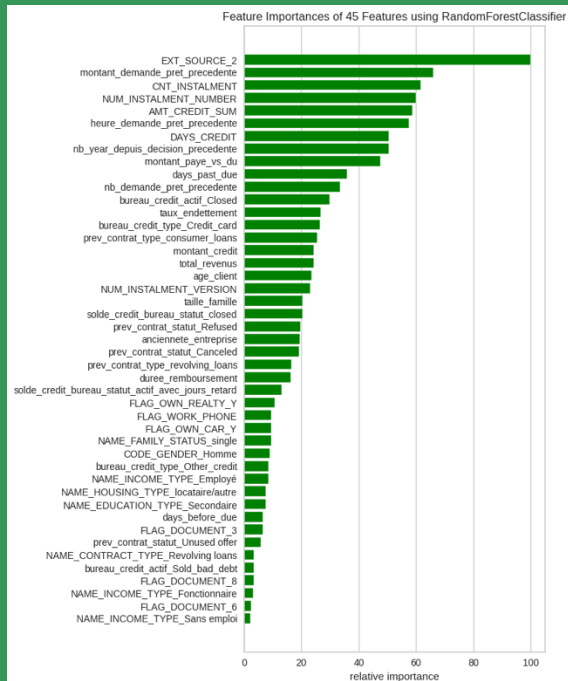


Analyse de la distribution de EXT_SOURCE_2

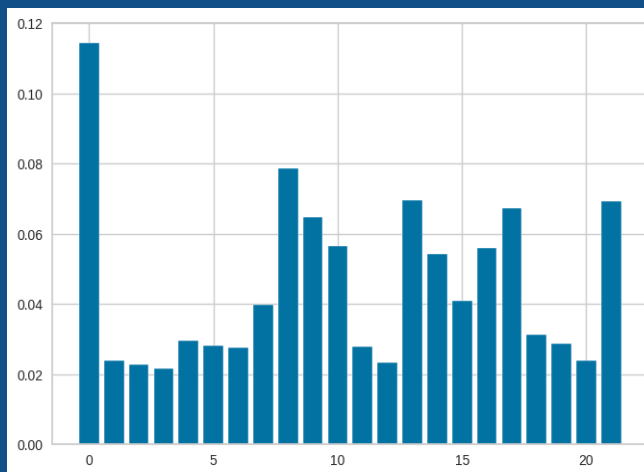
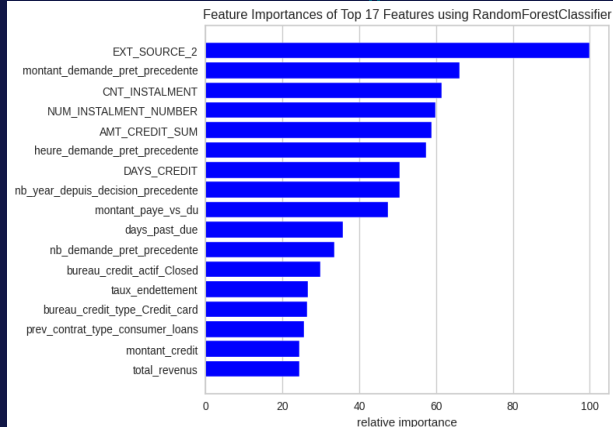


- Feature engineering

Analyse exploratoire



➤ Nous utilisons les méthodes Wrapped et Embedded pour la sélection des variables afin d'identifier les plus pertinentes pour notre modèle.



Feature
selection (22)

The background is a dark, atmospheric scene featuring several wireframe cubes floating in space. These cubes are illuminated from within, casting a soft glow. Bright, star-like lens flares are scattered across the upper right portion of the image. In the lower left, there are overlapping geometric shapes: a teal square, a smaller maroon square, and a larger maroon rectangle that serves as a backdrop for the text. The overall aesthetic is futuristic and digital.

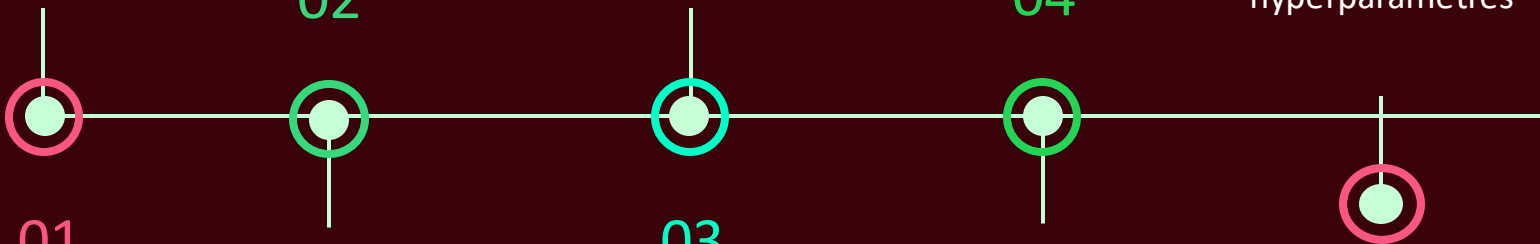
Modélisation

Modélisation

Train split test Mise en
place de Mlflow pour le
suivi

Choix des métriques

Optimisation des
modèles au travers des
hyperparamètres



Sélection des modèles
randomForest, XGBOOST,
LighjtGBM, regression

Entrainement des
modèles et gestion des
déséquilibres

Modélisation

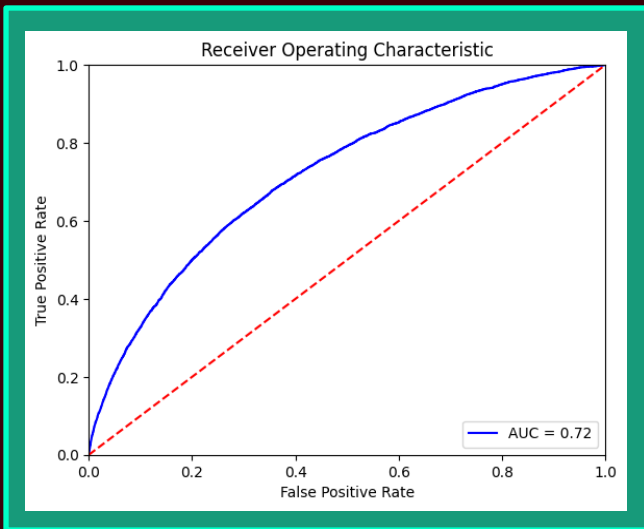
<input type="checkbox"/>	Run Name
<input type="checkbox"/>	<input checked="" type="radio"/> XGBoost MinMaxScaler
<input type="checkbox"/>	<input checked="" type="radio"/> XGBoost StandardScaler
<input type="checkbox"/>	<input checked="" type="radio"/> Dummy MinMaxScaler
<input type="checkbox"/>	<input checked="" type="radio"/> Dummy StandardScaler
<input type="checkbox"/>	<input checked="" type="radio"/> Reglog StandardScaler

		Predicted class		
		Classified positive	Classified negative	
Actual class	Actual positive	TP	FN	TPR: $\frac{TP}{TP + FN}$
	Actual negative	FP	TN	FPR: $\frac{FN}{TN + FP}$
		Precision: $\frac{TP}{TP + FP}$	Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$	

Sélection des metriques:

Accuracy ROC et AUC, Precision, Recall, F1 Score, F2 Score, Custom metric

- RandomUnderSample
- RandomOverSample
- SMOTE



A conceptual image featuring a human brain at the top, with a bundle of network cables extending downwards from its base, resembling a neural network. The cables end in RJ45 connectors on a surface. The background is a solid light pink color. On the left side, there are two overlapping squares: a teal one in front and a pink one behind it. A semi-transparent pink rectangular box is positioned in the lower-left area, containing the text 'Interpretation des résultats' in white.

Interpretation des résultats

Interpretation des résultats

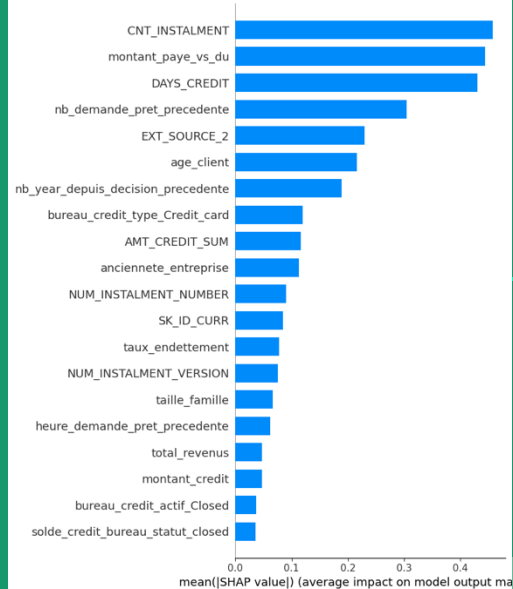
Modèles/ Métriques	Accuracy	Precision	Recall	F1 Score	F2 Score	ROCAUC	FMI
Régression Logistique							
StandardScaler	0.918	0.5	0.000638	0.00127	0.0008	0.5	0.0179
ROS	0.657	0.141	0.628	0.23	0.37	0.644	0.297
RUS	0.659	0.141	0.625	0.23	0.37	0.644	0.297
Smote	0.658	0.141	0.628	0.23	0.37	0.644	0.298
DummyClassifier							
MinMaxScaler	0.851	0.0802	0.0788	0.0795	0.079	0.499	0.0795
StandardScaler	0.851	0.083	0.0823	0.0826	0.082	0.501	0.0826
ROS	0.498	0.0803	0.493	0.138	0.24	0.496	0.199
RUS	0.501	0.0808	0.492	0.139	0.24	0.497	0.199
Smote	0.498	0.0801	0.492	0.138	0.24	0.495	0.198
RandomForest							
MinMaxScaler	0.918	0.5	0.00159	0.00318	0.002	0.501	0.0282
StandardScaler	0.918	0.471	0.00128	0.00254	0.0016	0.501	0.0245
ROS	0.918	0.35	0.00685	0.0134	0.0085	0.503	0.049
RUS	0.667	0.146	0.634	0.237	0.38	0.652	0.304
Smote	0.916	0.34	0.0249	0.0463	0.031	0.51	0.0919
XGBoost							
MinMaxScaler	0.918	0.429	0.0179	0.0343	0.022	0.508	0.0875
StandardScaler	0.918	0.429	0.0179	0.0343	0.022	0.508	0.0875
ROS	0.717	0.156	0.562	0.245	0.37	0.646	0.297
RUS	0.65	0.141	0.645	0.231	0.38	0.648	0.301
Smote	0.918	0.432	0.0188	0.0361	0.023	0.508	0.0902
LightGBM							
MinMaxScaler	0.919	0.561	0.0059	0.0117	0.0074	0.503	0.0575
StandardScaler	0.919	0.576	0.00606	0.012	0.0076	0.503	0.0591
ROS	0.685	0.154	0.637	0.249	0.39	0.664	0.314
RUS	0.664	0.148	0.655	0.242	0.39	0.66	0.311
Smote	0.918	0.455	0.00813	0.016	0.01	0.504	0.0608

- Impact du déséquilibre : Le déséquilibre des classes influence significativement les performances des modèles.
- Pertinence de l'accuracy : L'accuracy n'est pas un indicateur pertinent pour ce projet.
- Meilleurs modèles : La Régression Logistique, XGBoost, et LightGBM offrent les résultats les plus performants.

Interpretation des résultats



- Valeurs de Shapley : Elles évaluent l'importance des variables en comparant les prédictions du modèle avec et sans une variable donnée.
- Interprétation Globale : Permet de comprendre quelles variables influencent le plus les prédictions du modèle.
- Interprétation Locale : Fournit une explication spécifique pour chaque prédiction individuelle.



Analyse data drift

-> Librairie evidently

Drift is detected for 7.5% of columns (9 out of 120).

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.359052
> AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.281765
> AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.210785
> AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
> AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.161102
> AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.15426

Drift is detected for 9.091% of columns (11 out of 121).



- Drift augmente avec les bases intermédiaires
- Drift ne dépasse pas le seuil que nous avons défini 50% (7.5%)



Dashboard

Dashboard

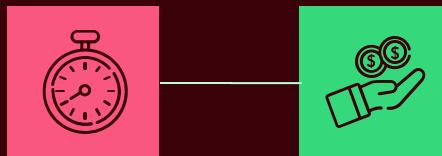




Conclusion

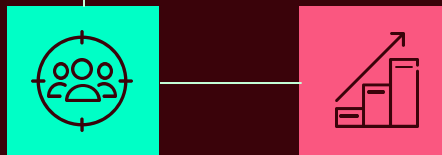
Conclusion

- Maj régulière des données pour éviter le data DRIFT



- L'optimisation des hyperparamètres à travers OPTUNA

- Avoir un feedback des équipes métiers pour confirmer le choix des métriques importantes



- Travailler sur la sécurité de l'application



Merci pour votre attention !