



Détection de Drift

Projet Entreprise

Bilal SAYOUD

Tuteur : Thomas KASTNER - Samir KHERCHAOU

Tuteur Ecole : Maria MALEK

Connected to the Future

www.ca-leasingfactoring.com

SOMMAIRE

- 1 Etat de l'art
- 2 Etapes d'action du projet
- 3 Mise en œuvre des méthodes
- 4 Résultats
- 5 Proposition de solution





1

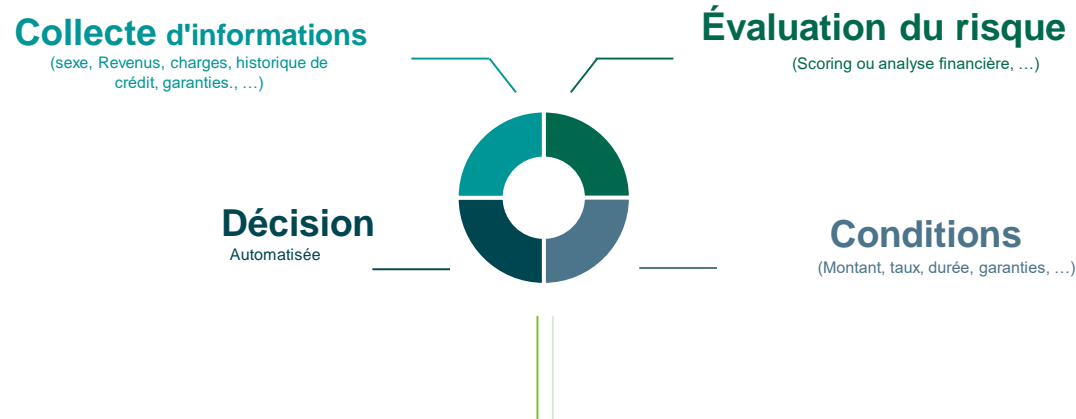
Etat de l'art

Modèle d'octroi

Définition

Acceptation des demandes de leasing en crédit bail mobile
Objectif : Évaluer la **solvabilité** et minimiser le **risque de défaut**.

Étapes principales



Importance

Réduit les pertes liées aux défauts.
Optimise la rentabilité de la banque.



Modèle d'octroi CALF



Modèle utilisé

Algorithme : XGBoost – Classification binaire
Objectif : **Prédire** le risque de défaut à 20 mois
Variable cible : Événement de défaut (0 / 1)

Structure des données

Nombre de features : 11 variables, continues et discrètes
Exemples de variables : montant demandé, revenu, forme juridique, situation financière...

Déploiement

Volume scoré : plusieurs dizaines de milliers de demandes par an
Intégration : modèle accessible via API, appelée à la demande par le logiciel de gestion utilisé par les équipes métiers



Monitoring du modèle d'octroi



Objectif

Surveiller la performance des modèles d'octroi.
Détecter les **anomalies** et **ajuster** les stratégies.

Bénéfices

Surveillance du fonctionnement des modèles pour repérer rapidement un défaut sur le processus d'octroi.
Réduction des risques financiers.



Backtesting

Définition

Évaluation d'un modèle d'octroi sur des **données historiques** pour **valider** sa performance.

Processus



Importance

Valide la robustesse du modèle après son déploiement.
Identifie les failles potentielles (ex. : sur-apprentissage).

Détection de drift

C'est quoi le drift détection :

Le drift est un changement dans **la distribution des données** ou **les performances d'un modèle** au fil du temps.

Exemples :

COVID-19 => Perte d'emplois entraînant **une baisse des revenus** des clients, modifiant la distribution des revenus => variation de variable **revenu**

Importance :

Affecte la fiabilité des modèles d'octroi de crédit

Objectif : Identifier les dérives pour maintenir la robustesse du modèle.



Backtesting VS Détection de drift

Outil de suivi du modèle



Backtesting



Pourquoi est-ce nécessaire ?

Etude ponctuelle qui vise à analyser la fiabilité du modèle sur des **périodes historiques**



Détection de drift



Monitore la performance d'un modèle **en temps réel** dans un environnement changeant (Covid, enjeux géopolitiques...) .



Type de détection de drift

1

Drift de données:

Le drift de données désigne **un changement dans la distribution des variables** d'entrée du modèle, sans forcément affecter la relation avec la cible.

2

Drift de modèle:

Le drift de modèle correspond à une **dégradation du comportement** ou **des performances d'un modèle** au cours du temps, souvent causée par **des changements dans les données**, mais aussi par une perte d'adéquation au contexte réel.

3

Drift de label

Le drift de label correspond à un **écart constaté** au niveau **de la variable cible du modèle**.

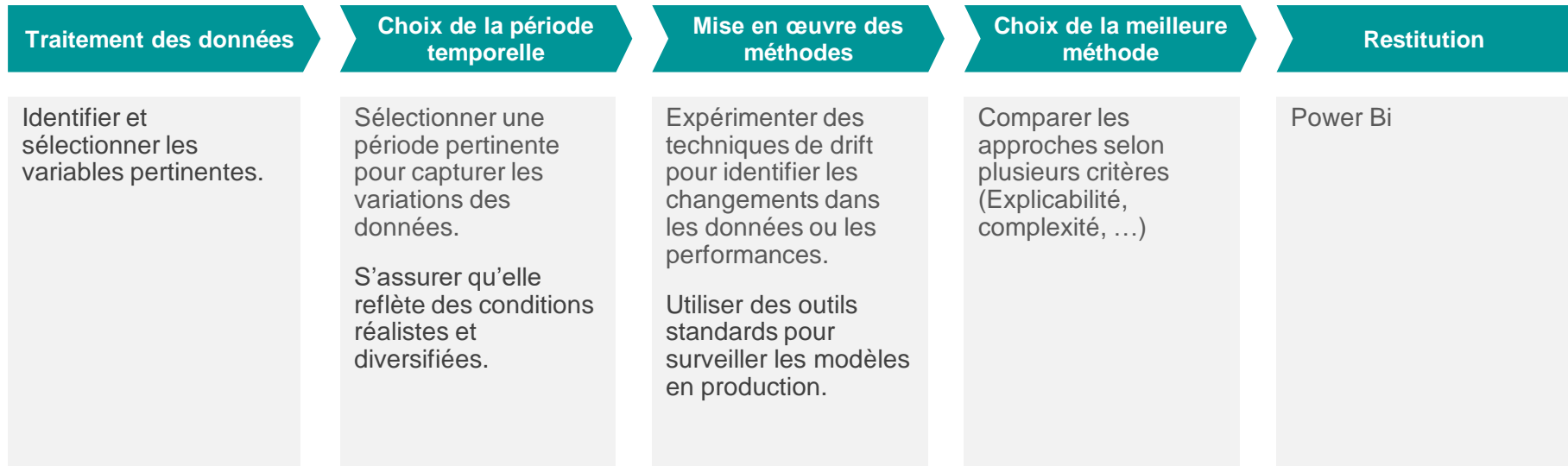
On ne peut pas l'appliquer en temps réel car la cible du modèle nécessite 20 mois de recul.



2

Etapes d'action du projet

Etapes d'action du projet



Traitement des données



Deux fichiers CSV : un de référence (2017–2020), un de test (2020–2024)

11 variables analysées : montant, revenu, endettement, etc.

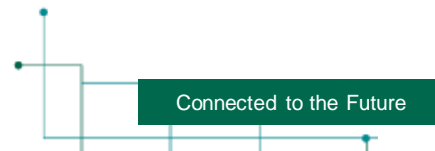
Ajouter de variables temporelles : année, trimestre, ancienneté du client

Choix de la période temporelle

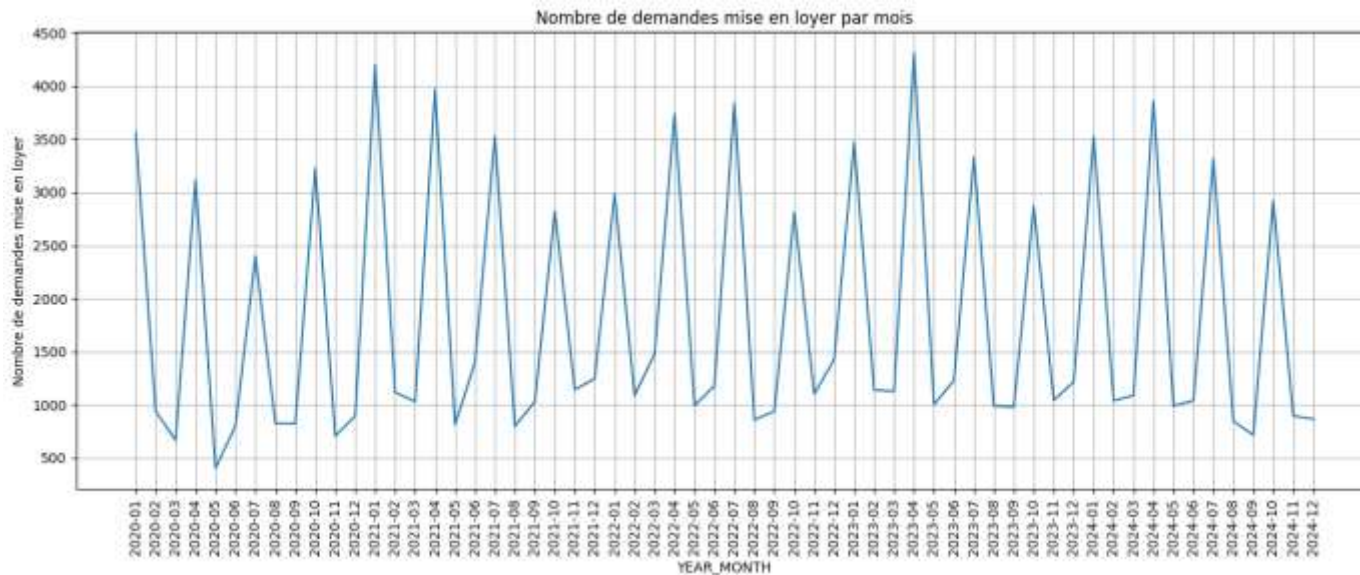


Définir **la bonne taille** de la fenêtre temporelle à étudier.

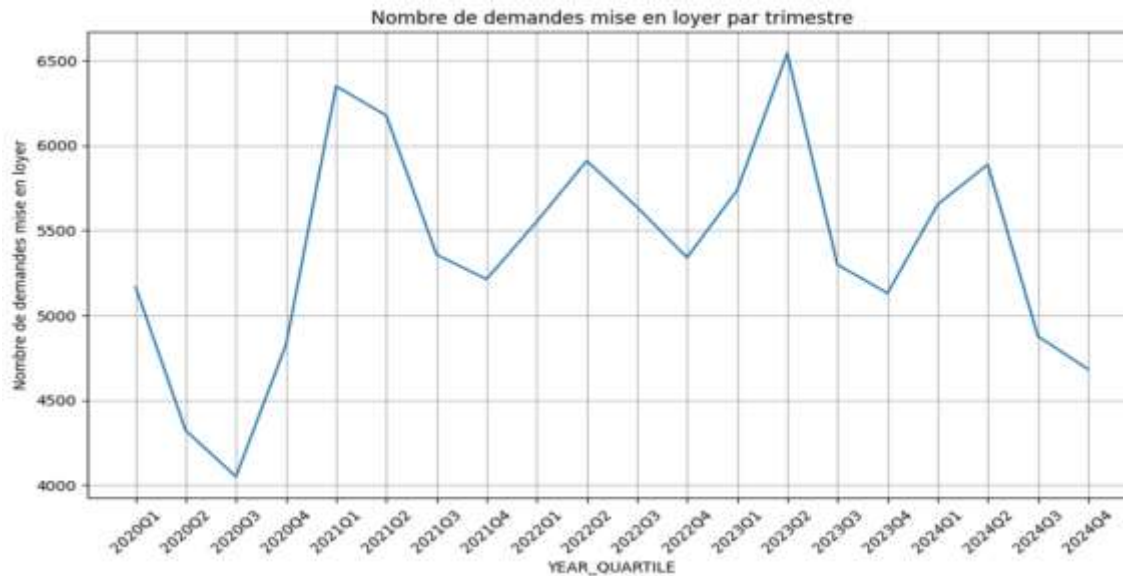
Choisir **la bonne granularité temporelle** pour que les analyses soient statistiquement valides et représentatives.



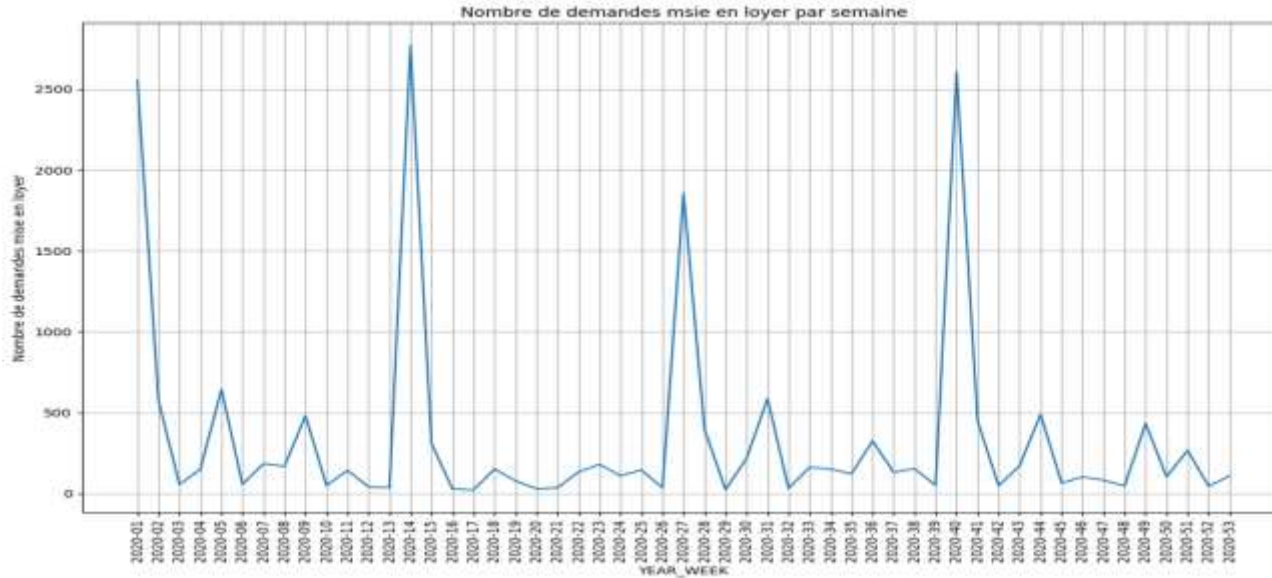
Choix de la période temporelle



Choix de la période temporelle



Choix de la période temporelle

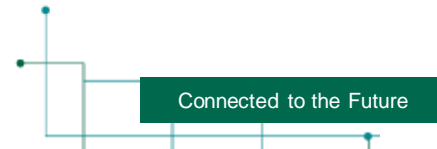


Choix de la période temporelle



Le période en **trimestre** a une volumétrie satisfaisante pour le **calcul du drift**.

Chaque période dans le **trimestre** comporte au minimum **4000** observations.





3

Mise en œuvre des méthodes

Méthodes de détection du drift

Méthode	Principe	Seuils (alerte / critique)	Complexité	Remarques
Population Stability Index	<u>Compare</u> les distributions	0.15 / 0.3	$O(n)$	Très rapide, facile à interpréter
Kullback-Leibler	<u>Mesure</u> l'écart entre deux distributions	0.15 / 0.3	$O(n)$	Sensible aux zéros
Jensen-Shannon	Version symétrique de KL	0.15 / 0.3	$O(n)$	Plus stable que KL
Kolmogorov-Smirnov	Compare les fonctions de répartition cumulées	p-value < 0.05	$O(n \log n)$	Robuste
XGBoost AUC	Mesure <u>la séparations</u> entre les données	0.6 / 0.75	$O(n \log n)$	Difficile d'interpréter quelle variable a dérivé

Kullback-Leibler

Objectif :

Mesurer à quel point une distribution de probabilité **s'écarte d'une référence.**

Équation :

$$D_{KL}(P||Q) = \sum_x P(x) \ln \left(\frac{P(x)}{Q(x)} \right)$$

où :

P(x) : Distribution de référence.

Q(x) : Distribution actuelle.

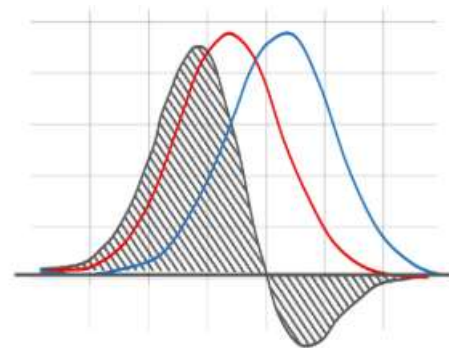
Fonctionnement :

Calcule la divergence pour variables continues ou discrètes.

Sensible aux zéros (Q(x) = 0 pose problème).

Seuils : Alerte : 0.15, Drift : 0.3.

Limites : Moins interprétable, sensible aux zéros.



Jensen-Shannon Divergence

Objectif :

Version **symétrique** et lissée de la divergence KL.

Équation :

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

où :

$$M = \frac{1}{2}(P + Q)$$

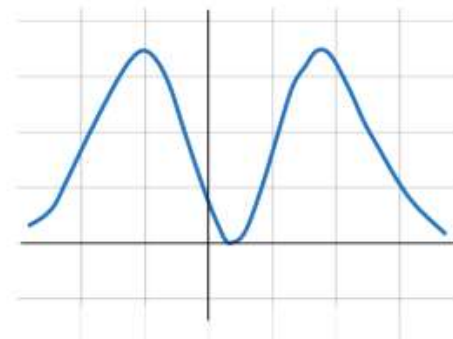
Moyenne des deux distributions.

Fonctionnement :

Symétrise la divergence KL pour **plus de stabilité**.

Seuils : Alerte à 0.15, Drift à 0.3.

Limites : Abstraite, moins intuitive pour les métier



Population Stability Index

Objectif :

Détecter le drift des données en **comparant** les distributions d'une variable entre deux périodes.

Équation :

$$PSI = \sum_{i=1}^k \left((p_i - q_i) \cdot \ln \left(\frac{p_i}{q_i} \right) \right)$$

où :

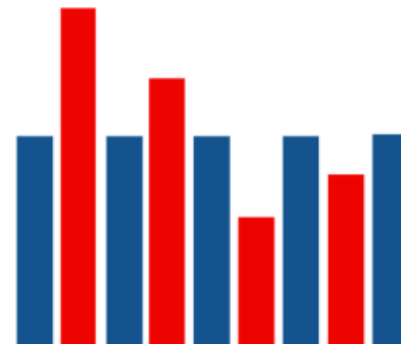
P(x) : Distribution de référence.

Q(x) : Distribution actuelle .

Fonctionnement :

Discrétisation des variables en bins.

Seuils : Alerte à 0.15, Drift à 0.3.



Kolmogorov-Smirnov

Objectif :

Comparer les fonctions de **répartition cumulées** de deux ensembles de données.

Équation :

$$D_{KS} = \sup_x |F_1(x) - F_2(x)|$$

où :

$F_1(x)$, $F_2(x)$: FRC des données de référence et actuelles.

Hypothèses :

H0 : Les deux distributions sont **identiques**.

H1 : Les deux distributions diffèrent donc **drift**.

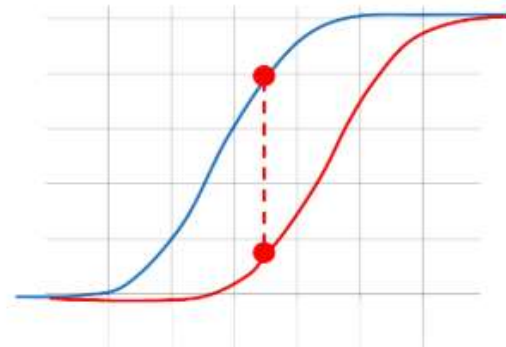
Acceptation : Rejeter H0 si la **p-valeur** < 0,05.

Fonctionnement :

Compare les distributions entières.

Seuil : Critique à 0,05.

Limites : Résultats statistiques difficiles à exploiter pour les métiers.



Xgboost AUC

Objectif :

Surveiller la dérive des performances du modèle via l'Aire Sous la Courbe .

Équation :

$$AUC = \int_0^1 TPR(t) dFPR(t)$$

où :

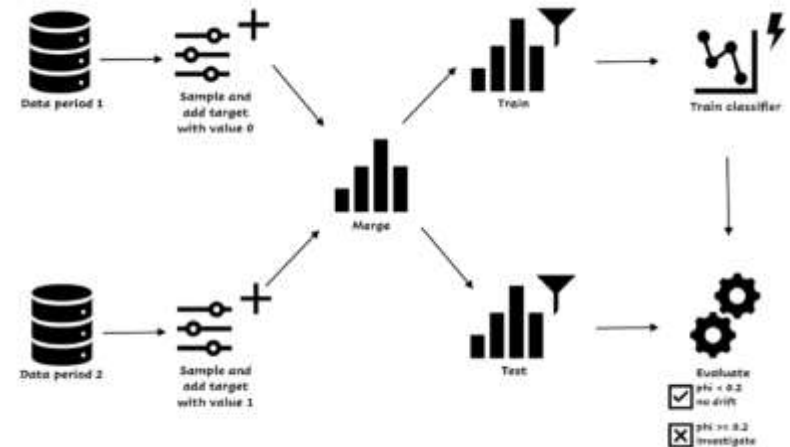
TPR : Taux de vrais positifs.

FPR : Taux de faux positifs.

Fonctionnement :

Seuils : Alerte à 0,5, Critique à 0,75.

Limites : Interprétation global





4

Résultats

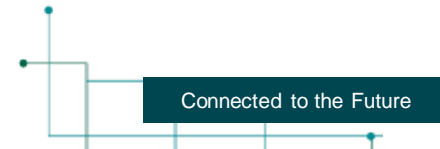
Résultats

metric	auc xgb	feat_imp	js	kl	ks	psi
feature_1	NaN	0.053898	0.811694	inf	NaN	0.008527
feature_2	NaN	0.101556	0.130462	0.071607	6.978628e-91	0.137326
feature_3	NaN	0.148058	0.068327	0.018610	5.594790e-28	0.037408
feature_4	NaN	0.086293	0.187371	0.151702	NaN	0.107511
feature_5	NaN	0.148510	0.271366	inf	NaN	0.485313
feature_6	NaN	0.069327	0.484265	inf	NaN	0.051935
feature_7	NaN	0.100038	0.391185	inf	NaN	0.131040
feature_8	NaN	0.058972	0.024353	0.002410	6.920165e-03	0.004747
feature_9	NaN	0.032501	0.148348	inf	2.524586e-01	0.007888
feature_10	NaN	0.048322	0.741639	inf	1.223548e-19	0.077602
feature_11	NaN	0.152524	0.058533	0.013372	6.882024e-13	0.027485
feature_12	0.876307	NaN	NaN	NaN	NaN	NaN

Choix de la meilleure méthode

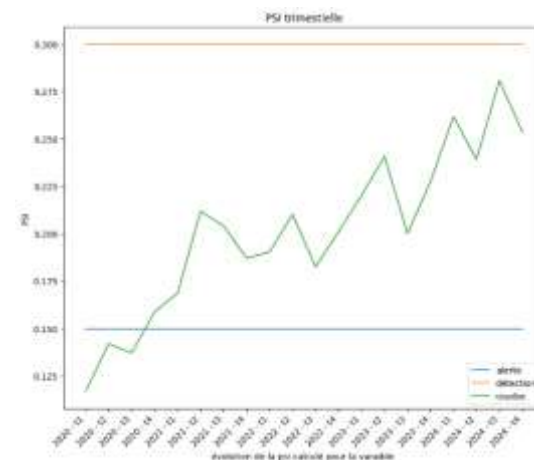
Le **PSI** donne des résultats **fiables et interprétables**, particulièrement utiles pour prioriser les analyses métier. Il est moins sensible aux problèmes de données manquantes que d'autres métriques.

Le **score AUC** mesure la **performance globale du modèle**. Dans le cadre du monitoring de drift, cette métrique permet de mesurer la capacité d'un modèle à discriminer les données d'une période récente par rapport à une période passée. En revanche, elle **ne permet pas d'identifier directement quelles variables** sont responsables de cette dérive.



Application des méthodes sélectionnées

metric	auc xgb	feat_imp	Valeur_PSI	Statut_drift_PSI
feature_1	NaN	0.053898	0.008527	Correct
feature_2	NaN	0.101556	0.137326	Correct
feature_3	NaN	0.148058	0.037408	Correct
feature_4	NaN	0.086293	0.107511	Correct
feature_5	NaN	0.148510	0.485313	Alert
feature_6	NaN	0.069327	0.051935	Correct
feature_7	NaN	0.100038	0.131040	Correct
feature_8	NaN	0.058972	0.004747	Correct
feature_9	NaN	0.032501	0.007888	Correct
feature_10	NaN	0.048322	0.077602	Correct
feature_11	NaN	0.152524	0.027485	Correct
feature_12	0.876307	NaN	NaN	Correct





5

Proposition de solution

Proposition de solution



MERCI !

Connected to the Future



**CRÉDIT AGRICOLE
LEASING & FACTORING**

MERCI

Source

Connected to the Future