# Note méthodologique : Méthodologie d'entraînement du modèle de scoring crédit

## Introduction

L’entreprise “Prêt à dépenser” souhaite mettre en place un modèle de scoring de crédit pour évaluer la probabilité de faillite d'un client et décider de l'octroi ou non d'un crédit. Dans le cadre de cette mission, la méthodologie d'entraînement du modèle est cruciale pour la qualité des prédictions et la fiabilité du modèle. Cette note méthodologique décrit les étapes de l'entraînement du modèle, en mettant l'accent sur les points importants pour la prise en compte du déséquilibre des données et la minimisation du coût métier.

## I. Préparation des données La qualité des données est essentielle pour la construction d'un modèle fiable et performant. La préparation des données comprend plusieurs étapes :

1. **Exploration des données** Il est important de comprendre les données avant de les utiliser pour entraîner le modèle. Cette étape comprend l'analyse des données manquantes, l'identification des données aberrantes et la compréhension des relations entre les variables.
2. **Nettoyage des données** Cette étape comprend le traitement des données manquantes, des données aberrantes et des valeurs extrêmes, ainsi que la normalisation des données.
3. **Traitement du déséquilibre des données** Comme indiqué dans les spécifications, les données peuvent être déséquilibrées en termes de nombre de bons et de moins bons clients. Pour éviter un biais de prédiction, différentes techniques peuvent être utilisées pour équilibrer les données, telles que le sur-échantillonnage de la classe minoritaire ou la sous-échantillonnage de la classe majoritaire. Ces techniques sont appliquées à l'étape de préparation des données.

## II. Construction du modèle La construction du modèle comprend plusieurs étapes :

1. **Choix de l'algorithme** Le choix de l'algorithme est basé sur la nature des données et les spécifications de l'entreprise. Pour la construction du modèle de scoring crédit, les algorithmes de classification tels que la régression logistique, les arbres de décision, les forêts aléatoires, les SVM ou les réseaux de neurones peuvent être utilisés.
2. **Entraînement du modèle** L'ensemble de données préparé est divisé en deux parties : un ensemble d'entraînement et un ensemble de validation. L'ensemble d'entraînement est utilisé pour entraîner le modèle, tandis que l'ensemble de validation est utilisé pour évaluer la performance du modèle.
3. **Optimisation des hyperparamètres** L'optimisation des hyperparamètres est une étape importante pour améliorer la performance du modèle. Les hyperparamètres tels que le taux d'apprentissage, le nombre de couches, la profondeur de l'arbre, etc. peuvent être optimisés à l'aide d'une recherche systématique de grille ou d'algorithmes d'optimisation tels que la descente de gradient stochastique.

# III. Evaluation du modèle La performance du modèle est évaluée à l'aide de différentes métriques telles que l'AUC, la précision, le rappel, la F1-score, etc. Dans le cadre de cette mission, l'évaluation doit prendre en compte le déséquilibre.

La construction d'un modèle de scoring crédit est une tâche essentielle pour toute entreprise de crédit à la consommation. Le but de ce modèle est de prédire la probabilité de défaut de paiement d'un client, afin d'identifier les demandes de crédit à haut risque et d'en minimiser le nombre. Dans ce contexte, la méthode de régression logistique est une des approches les plus couramment utilisées pour construire des modèles de scoring crédit.

La régression logistique est une technique de modélisation statistique qui permet de prédire une variable binaire, c'est-à-dire une variable qui ne peut prendre que deux valeurs possibles (0 ou 1). Dans le cas d'un modèle de scoring crédit, la variable binaire à prédire est la probabilité de défaut de paiement d'un client, qui peut être soit "défaut" (1) ou "pas de défaut" (0). La régression logistique calcule la probabilité de défaut en fonction de variables explicatives, telles que le montant du crédit demandé, le revenu du client, l'historique de crédit, etc.

La méthodologie d'entraînement d'un modèle de scoring crédit avec une régression logistique implique plusieurs étapes importantes :

1. **Collecte et préparation des données :** Le premier pas consiste à collecter les données relatives aux clients et aux crédits précédents, ainsi que les données descriptives (variables explicatives) qui peuvent affecter la probabilité de défaut de paiement. Les données peuvent provenir de différentes sources, telles que les demandes de crédit précédentes, les agences de crédit, les informations démographiques, etc. Il est essentiel de préparer les données en les nettoyant, en les filtrant et en les transformant en un format standard.
2. **Exploration des données** : La prochaine étape consiste à explorer les données afin de comprendre les relations entre les variables explicatives et la variable cible. Il est important de visualiser les données à l'aide de graphiques, de tableaux de fréquence et de corrélations pour identifier les variables les plus importantes à inclure dans le modèle.
3. **Préparation des données pour l'entraînement du modèle** : Dans cette étape, les données sont divisées en ensembles d'entraînement et de test. L'ensemble d'entraînement est utilisé pour ajuster le modèle aux données, tandis que l'ensemble de test est utilisé pour évaluer la performance du modèle. Il est également important de normaliser les données pour garantir que les variables ont la même échelle et d'effectuer une sélection de variables pour identifier les variables les plus importantes à inclure dans le modèle.
4. **Entraînement et validation du modèle :** La régression logistique est ajustée aux données d'entraînement pour estimer les coefficients de régression. La performance du modèle est évaluée en utilisant des mesures de précision, telles que l'AUC et l'accuracy, sur l'ensemble de test. Il est également possible d'utiliser des méthodes plus avancées, telles que la validation croisée, pour évaluer la robustesse du modèle et réduire le risque de surapprentissage.
5. **Optimisation des hyperparamètres** : À ce stade, le modèle a été entraîné sur les données d'entraînement et évalué sur les données de validation. Plusieurs métriques peuvent être utilisées pour évaluer les performances du modèle, telles que l'AUC, l'accuracy, la précision, le rappel, le F1-score, etc. Il est important de choisir la métrique qui est la plus pertinente pour le problème métier. Ensuite, les différents modèles peuvent être comparés en fonction de leur performance sur cette métrique.  
   Une fois le meilleur modèle sélectionné, les hyperparamètres du modèle peuvent être optimisés pour améliorer encore davantage les performances du modèle. Les hyperparamètres sont des paramètres qui ne sont pas appris par le modèle lui-même, mais qui sont définis avant l'entraînement du modèle. Ils peuvent avoir un impact important sur les performances du modèle. Des techniques telles que la recherche de grille ou l'optimisation bayésienne peuvent être utilisées pour trouver les meilleurs hyperparamètres.
6. **Validation croisée :** Pour s'assurer que le modèle est généralisable, il est important de le valider sur des données qui n'ont pas été utilisées pendant l'entraînement. Pour cela, on peut utiliser une technique de validation croisée, où les données sont divisées en plusieurs ensembles et le modèle est entraîné et évalué sur chaque ensemble tour à tour.

# Méthodologie d'entraînement pour un modèle LGBMClassifier :

## Le LGBMClassifier est un modèle de classification qui utilise un algorithme de boosting pour améliorer les performances du modèle. Voici la méthodologie d'entraînement pour un modèle LGBMClassifier :

1. **Prétraitement des données :** Avant d'entraîner un modèle LGBMClassifier, il est important de prétraiter les données. Cela peut inclure la normalisation des caractéristiques, la gestion des valeurs manquantes, la suppression des caractéristiques redondantes, etc. La normalisation des caractéristiques est particulièrement importante pour les modèles de type arbre, car les caractéristiques qui ont des échelles très différentes peuvent avoir un impact disproportionné sur le modèle.
2. **Choix des caractéristiques :** Le choix des caractéristiques est crucial pour la performance du modèle. Il est important de sélectionner les caractéristiques qui sont les plus pertinentes pour la tâche de prédiction et qui ont une forte corrélation avec la variable cible. Il est également possible de créer de nouvelles caractéristiques à partir des caractéristiques existantes, par exemple en utilisant des techniques de transformation telles que la transformation logarithmique ou la transformation en puissance.
3. **Paramètres du modèle :** Le LGBMClassifier a plusieurs hyperparamètres qui peuvent être ajustés pour améliorer la performance du modèle. Certains de ces paramètres comprennent la profondeur de l'arbre, le taux d'apprentissage, le nombre d'estimateurs, etc. Il est important d'optimiser ces paramètres pour obtenir la meilleure performance possible du modèle.
4. **Entraînement du modèle :** Une fois que les données ont été prétraitées et les paramètres du modèle ont été sélectionnés, il est temps d'entraîner le modèle. L'entraînement du LGBMClassifier est assez rapide par rapport à d'autres modèles de type arbre. Il est important de surveiller la performance du modèle pendant l'entraînement pour détecter tout surajustement ou sous-ajustement.
5. **Évaluation du modèle :** Une fois que le modèle est entraîné, il est important de l'évaluer sur des données de test pour évaluer sa performance. Les métriques d'évaluation courantes pour les problèmes de classification binaires incluent l'exactitude, la précision, le rappel, le score F1, le score Fbeta et la courbe ROC-AUC.
6. **Déploiement du modèle** : Une fois que le modèle est entraîné et évalué, il peut être déployé sur une plateforme pour une utilisation en production. Cela peut inclure la création d'une API pour permettre l'accès au modèle via des requêtes HTTP ou la création d'une application web ou mobile utilisant le modèle.
7. **Surveillance en production :** Il est important de surveiller la performance du modèle en production pour détecter tout changement dans les données qui pourraient affecter sa performance. Cela peut inclure la détection de dérive de données, la surveillance des erreurs de prédiction et la mise en place d'un système d'alerte pour signaler les problèmes potentiels.

# Note méthodologique : Traitement du déséquilibre de classes pour la modélisation de crédit

Lorsqu'on travaille sur des problématiques de crédit, un des challenges majeurs est la présence de déséquilibres entre les classes. En effet, la plupart des ensembles de données dans ce domaine sont déséquilibrés, c'est-à-dire que les classes positives (clients qui remboursent leur crédit) sont souvent moins nombreuses que les classes négatives (clients qui ne remboursent pas). Ce déséquilibre peut rendre la modélisation difficile, car les algorithmes ont tendance à privilégier la classe majoritaire, ce qui peut entraîner des prédictions biaisées.

Pour résoudre ce problème, différentes méthodes peuvent être utilisées. Dans ce qui suit, nous allons décrire les principales techniques pour traiter le déséquilibre des classes dans le cadre de la modélisation de crédit.

1. **Sous-échantillonnage** La méthode de sous-échantillonnage consiste à réduire la taille de la classe majoritaire pour équilibrer la distribution des classes. Il existe différentes stratégies pour sous-échantillonner les données, notamment la sélection aléatoire d'un sous-ensemble de la classe majoritaire ou l'utilisation d'algorithmes spécifiques tels que Tomek Links, Condensed Nearest Neighbor ou One-Sided Selection.
2. **Sur-échantillonnage** La méthode de sur-échantillonnage consiste à augmenter la taille de la classe minoritaire pour équilibrer la distribution des classes. Il existe différentes techniques de sur-échantillonnage, notamment la duplication aléatoire d'observations de la classe minoritaire ou l'utilisation d'algorithmes spécifiques tels que SMOTE, ADASYN ou Borderline-SMOTE.
3. **Ré-échantillonnage** **mixte** Une autre approche consiste à combiner les deux techniques précédentes pour équilibrer la distribution des classes. Cette méthode consiste à sous-échantillonner la classe majoritaire et à sur-échantillonner la classe minoritaire de manière équilibrée.
4. Une autre méthode pour traiter le déséquilibre des classes consiste à utiliser des techniques de pénalisation pour les algorithmes d'apprentissage automatique. Ces techniques consistent à **attribuer des poids différents aux classes en fonction de leur distribution**. Ainsi, la classe minoritaire sera pénalisée plus lourdement lors du calcul de la fonction de coût.

Dans le cadre de ce projet, nous avons choisi d'utiliser une combinaison de sur-échantillonnage pour traiter le déséquilibre des classes. Nous avons également opté pour un algorithme LightGBM comme algorithme de classification pour ce problème de crédit. La prochaine section de cette note méthodologique décrira en détail la méthodologie d'entraînement pour ce modèle.

# Note méthodologique : Fonction coût métier, algorithme d'optimisation et métrique d'évaluation

Introduction   
  
Dans le cadre de l'élaboration d'un modèle de prédiction de crédit, il est important de prendre en compte les coûts métier associés aux erreurs de prédiction. En effet, un mauvais client prédit comme bon peut entraîner des pertes en capital pour l'organisme de crédit, tandis qu'un bon client prédit comme mauvais peut entraîner un manque à gagner en termes de marge. Pour intégrer ces coûts dans le processus de modélisation, nous allons définir une fonction coût métier qui tiendra compte du déséquilibre des classes et des coûts associés aux erreurs de prédiction.

**Fonction coût métier** La fonction coût métier est une fonction qui attribue un coût différent aux différentes erreurs de prédiction en fonction de leur impact sur le coût global. Dans notre cas, nous supposons que le coût d'un faux négatif (mauvais client prédit comme bon) est dix fois supérieur au coût d'un faux positif (bon client prédit comme mauvais). La fonction coût métier prendra donc en compte ce déséquilibre des coûts en attribuant un poids plus élevé aux faux négatifs.

Pour définir la fonction coût métier, nous allons utiliser la matrice de confusion, qui permet de visualiser les différentes erreurs de prédiction. La matrice de confusion est une matrice 2x2 qui regroupe les prédictions correctes et incorrectes. Les deux axes de la matrice représentent les prédictions réelles (positives ou négatives) et les prédictions du modèle (positives ou négatives). La matrice de confusion permet de calculer différentes métriques, dont le taux de faux positifs (FP) et le taux de faux négatifs (FN), qui seront utilisés pour définir la fonction coût métier.

La fonction coût métier peut être définie de différentes manières, mais une approche simple consiste à attribuer un coût spécifique à chaque type d'erreur de prédiction, en multipliant le nombre de prédictions de chaque type par le coût associé. Par exemple, si le coût d'un faux négatif est dix fois supérieur au coût d'un faux positif, la fonction coût métier peut être définie comme suit :

Coût métier = 10 x FN + FP

Cette fonction attribue un poids dix fois supérieur aux faux négatifs par rapport aux faux positifs.

**Algorithme d'optimisation** Pour optimiser la fonction coût métier, nous allons devoir ajuster le seuil de classification utilisé par le modèle. Par défaut, le seuil de classification est fixé à 0,5, ce qui signifie que toute observation ayant une probabilité supérieure à 0,5 est classée comme positive, et toute observation ayant une probabilité inférieure à 0,5 est classée comme négative. Cependant, le seuil de classification optimal peut varier en fonction de la fonction coût métier, car il détermine le trade-off entre le taux de faux positifs et le taux de faux négatifs.

## Pour trouver le seuil de classification

1. **Algorithme d'optimisation**

Une fois la fonction de coût métier définie, l'étape suivante consiste à trouver le meilleur seuil pour optimiser la fonction de coût. Nous pouvons utiliser un algorithme de recherche pour trouver le seuil optimal.

L'un des algorithmes de recherche les plus couramment utilisés est la recherche par grille. La recherche par grille consiste à définir une grille de seuils possibles et à évaluer chaque seuil sur la grille pour trouver le seuil optimal qui minimise la fonction de coût métier.

Une autre approche courante est l'utilisation de la recherche aléatoire. La recherche aléatoire consiste à définir une distribution de probabilité pour les seuils possibles et à échantillonner à partir de cette distribution pour trouver le seuil optimal. La recherche aléatoire peut souvent être plus efficace que la recherche par grille, car elle permet d'explorer un plus grand espace de recherche.

1. **Métrique d'évaluation**

Pour évaluer la performance du modèle de crédit, nous utiliserons plusieurs métriques d'évaluation, notamment l'aire sous la courbe ROC (AUC-ROC), la précision, le rappel, la F-mesure et la courbe de gain.

L'aire sous la courbe ROC (AUC-ROC) est une mesure de la capacité du modèle à distinguer les bons et les mauvais clients. Une AUC-ROC de 1 indique que le modèle est capable de faire une distinction parfaite entre les bons et les mauvais clients, tandis qu'une AUC-ROC de 0,5 indique que le modèle est incapable de faire une distinction entre les deux.

La précision est le nombre de vrais positifs divisé par le nombre total de prédictions positives, tandis que le rappel est le nombre de vrais positifs divisé par le nombre total de positifs réels. La F-mesure est une mesure qui combine la précision et le rappel pour donner une mesure globale de la performance du modèle.

La courbe de gain est une mesure qui montre le pourcentage de clients que le modèle peut détecter par rapport au pourcentage de clients qui auraient été détectés au hasard. La courbe de gain est particulièrement utile pour évaluer les performances du modèle dans des situations où le coût de la décision de prêt est important.

En résumé, la mise en place d'un modèle de crédit implique plusieurs étapes, notamment la collecte et la préparation des données, la sélection de caractéristiques, la modélisation et l'évaluation de la performance. Le traitement du déséquilibre des classes et la définition d'une fonction de coût métier appropriée sont également des éléments clés pour améliorer la performance du modèle. L'optimisation du seuil de classification et l'utilisation de métriques d'évaluation appropriées sont des étapes importantes pour garantir que le modèle est capable de prendre des décisions de crédit précises et fiables.

# Note méthodologique : L’interprétabilité globale et locale du modèle

Introduction :   
  
L’interprétabilité des modèles de machine learning est une question clé dans de nombreux domaines d’application, notamment dans le cas de l’octroi de crédit où les décisions peuvent avoir des conséquences importantes pour les clients et les prêteurs. Dans cette note, nous aborderons deux aspects de l'interprétabilité du modèle : l'interprétabilité globale et l'interprétabilité locale. Nous présenterons également des techniques pour améliorer l'interprétabilité du modèle.

**Interprétabilité globale** : L'interprétabilité globale se réfère à la capacité de comprendre le modèle dans son ensemble et d'expliquer comment il prend des décisions. Dans le cas de la classification de crédit, il est important de comprendre les caractéristiques des clients qui sont utilisées par le modèle pour prendre des décisions, afin de s'assurer qu'elles sont pertinentes et non discriminatoires.

**Pour améliorer l'interprétabilité globale du modèle, plusieurs techniques peuvent être utilisées, notamment :**

* La réduction de dimensionnalité : cela permet de visualiser les données et d'identifier les caractéristiques les plus importantes qui influencent les décisions du modèle.
* La sélection de modèle : certains types de modèles, comme les arbres de décision, sont plus faciles à interpréter que d'autres.
* L'analyse des coefficients : dans le cas d'un modèle LGBM, l'analyse des coefficients permet de comprendre l'impact de chaque caractéristique sur la probabilité de défaut de paiement.

**Interprétabilité locale :** L'interprétabilité locale se réfère à la capacité de comprendre comment le modèle prend des décisions pour un cas particulier. Dans le cas de l'octroi de crédit, il est important de comprendre pourquoi un client spécifique a été accepté ou refusé.

Pour améliorer l'interprétabilité locale du modèle, plusieurs techniques peuvent être utilisées, notamment :

* Les graphes de décision : cela permet de visualiser comment le modèle prend des décisions pour un cas particulier.
* L'importance des caractéristiques : cela permet de comprendre quelles sont les caractéristiques qui ont été les plus importantes pour la décision.
* La comparaison avec des cas similaires : cela permet de voir comment le modèle a pris des décisions pour des cas similaires.

**Conclusion :** Dans cette note, nous avons discuté de l'importance de l'interprétabilité du modèle de classification de crédit, et avons présenté des techniques pour améliorer l'interprétabilité globale et locale du modèle. Il est important de souligner que l'interprétabilité ne doit pas être considérée comme une tâche annexe mais comme une étape clé du processus de développement du modèle. En effet, l'interprétabilité permet de s'assurer de la pertinence des caractéristiques utilisées, d'éviter les discriminations et de gagner la confiance des utilisateurs du modèle.

# Note méthodologique : Les limites et les améliorations possibles

Malgré l'attention portée aux différentes étapes de l'élaboration d'un modèle de prédiction de crédit, il existe toujours des limites et des améliorations potentielles à considérer. Dans cette note, nous allons passer en revue quelques-unes de ces limites et proposer des pistes d'amélioration pour les futures itérations du modèle.

**Limites :**

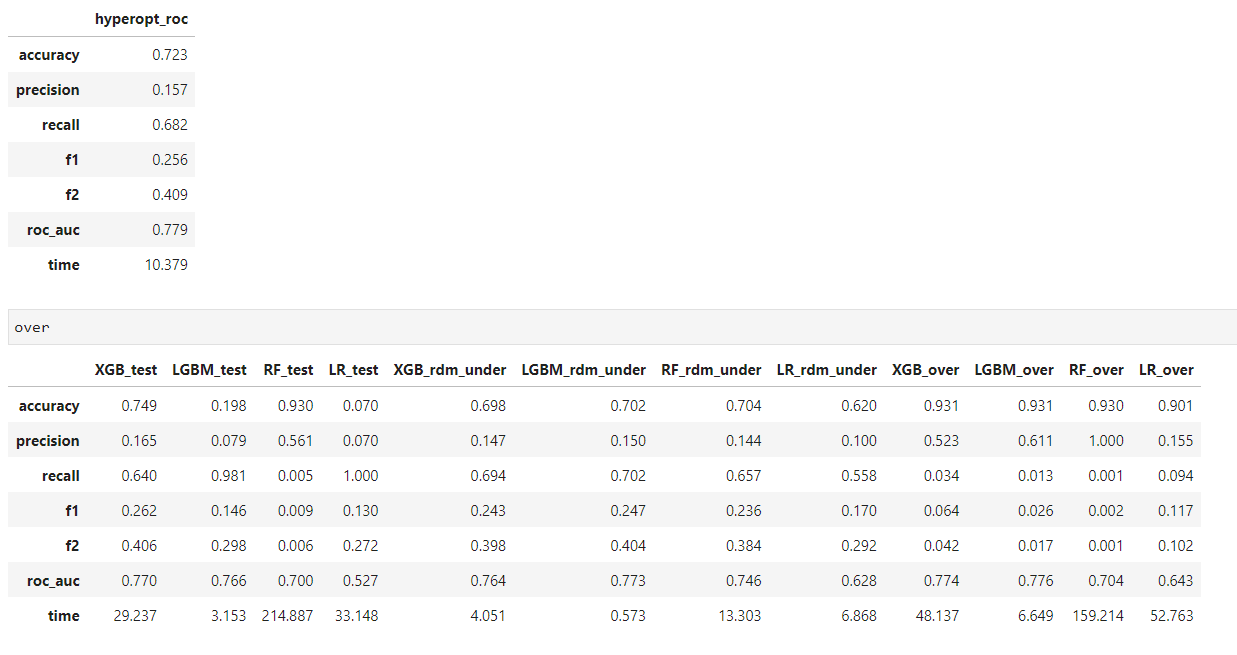
* **Le modèle est limité par la qualité des données d'entrée** : bien que nous ayons effectué des étapes de nettoyage et de traitement des données, il est possible que des erreurs ou des biais subsistent dans les données sous-jacentes. Cela peut affecter la qualité de notre modèle et entraîner des prédictions erronées.
* **Le modèle peut ne pas généraliser de manière optimale** : bien que nous ayons utilisé des techniques pour prévenir le surapprentissage, il est possible que notre modèle ne généralise pas de manière optimale sur des données de test. Cela peut être dû à des différences dans la distribution des données de test et de formation, ou à d'autres facteurs.
* **Le modèle peut être difficile à interpréter** : bien que nous ayons utilisé des techniques pour rendre le modèle plus interprétable, il est possible que certaines prédictions soient difficiles à expliquer. Cela peut limiter la confiance que les utilisateurs ont dans le modèle.

**Améliorations possibles :**

* **Collecte de données plus complètes :** pour améliorer la qualité de notre modèle, nous pourrions envisager de collecter des données supplémentaires qui peuvent améliorer la précision de nos prédictions. Cela peut inclure des données provenant de sources externes telles que les réseaux sociaux, les données financières ou les données de crédit supplémentaires.
* **Utilisation de techniques d'apprentissage en continu :** pour améliorer la capacité de notre modèle à généraliser sur de nouvelles données, nous pourrions envisager d'utiliser des techniques d'apprentissage en continu. Cela permettrait à notre modèle d'apprendre en temps réel à partir de nouvelles données, améliorant ainsi sa précision au fil du temps.
* **Utilisation de techniques d'interprétabilité plus avancées :** pour améliorer la capacité de notre modèle à être interprété, nous pourrions envisager d'utiliser des techniques d'interprétabilité plus avancées telles que l'importance des caractéristiques, les cartes de chaleur, les réseaux neuronaux interprétables ou les arbres de décision.
* **Considération de l'équité et de la responsabilité :** pour assurer que notre modèle n'a pas de biais et est juste pour toutes les parties prenantes, nous pourrions envisager d'utiliser des techniques d'équité et de responsabilité telles que l'apprentissage équitable, la transparence de l'algorithme, ou la conformité réglementaire.

En somme, malgré les efforts consacrés à l'élaboration d'un modèle de prédiction de crédit de qualité, il reste toujours des marges de progrès. En continuant d'examiner les limites et les améliorations possibles de notre modèle, nous pouvons travailler à créer un outil de prédiction plus précis,

# Un tableau de synthèse des résultats



# L’analyse du Data Drift

**Pas de résultat**