Voici une note méthodologique décrivant les différentes parties de l'analyse :

# Partie 1 : Entraînement d'un modèle de classification binaire

Dans cette partie, nous allons entraîner un modèle de classification binaire pour établir si un client est à risque de défaut de paiement pour un crédit (valeur 1) ou non (valeur 0). Nous allons comparer plusieurs modèles : LGBM, régression logistique, random forest et un modèle de base naif.

Pour sélectionner le meilleur modèle, nous utiliserons une méthode de cross-validation avec RandomizedSearch. Cela nous permettra d'explorer différentes combinaisons d'hyperparamètres pour chaque modèle et d'évaluer leur performance.

Cependant, nous ne choisirons pas le modèle en fonction de sa précision globale, mais plutôt en utilisant un score métier qui minimise le coût induit par les taux de faux positif et de faux négatif. Un faux négatif aura un coût dix fois supérieur à celui d'un faux positif.

Après avoir sélectionné le meilleur modèle, nous identifierons les 10 features les plus importantes. Ensuite, nous réentraînerons le modèle en utilisant uniquement ces 10 features et calculerons le seuil de probabilité de prédiction pour lequel la fonction de coût est minimisée.

# Partie 2 : Traitement du déséquilibre des classes

Dans cette partie, nous abordons le déséquilibre des classes dans nos données, avec seulement 3% de clients à risque de défaut de paiement dans la base. Pour rétablir l'équilibre, nous appliquerons la méthode SMOTE (Synthetic Minority Over-sampling Technique) avec un SVM (Support Vector Machine).

Le SMOTE permettra de générer des échantillons synthétiques pour les clients à risque de défaut de paiement afin d'augmenter leur nombre et d'équilibrer les classes. En utilisant un SVM, nous apprendrons à séparer efficacement les deux classes équilibrées pour la classification.

# Partie 3 : Interprétabilité globale et locale du modèle

Dans cette partie, nous évaluerons l'interprétabilité globale et locale du modèle. Tout d'abord, nous comparerons les clients sur les dix features les plus importantes par rapport à la population globale. Cela nous permettra de déterminer les caractéristiques clés qui distinguent les clients à risque de défaut de paiement des autres.

Ensuite, nous analyserons les clients par rapport à des sous-populations spécifiques, en les comparant à des clients du même sexe et de la même tranche d'âge, à plus ou moins un an près. Cela nous donnera une perspective locale et nous aidera à comprendre comment les caractéristiques importantes peuvent varier selon ces sous-groupes.

# Partie 4 : Limites et améliorations possibles

Dans cette partie, nous discuterons des limites de notre approche actuelle et des possibles améliorations. Parmi les limites potentielles, nous pourrions mentionner la sensibilité des modèles aux choix des hyperparamètres et le fait que la performance du modèle pourrait varier dans le temps en raison de l'évolution des données.

Quant aux améliorations possibles, nous pourrions explorer d'autres techniques de traitement

# Partie 5 : Analyse du Data Drift

Dans cette partie, nous analyserons le phénomène de data drift, c'est-à-dire les variations dans les données entre l'analyse initiale et les données actuelles. Nous avons observé un data drift sur deux features spécifiques, ce qui nécessite une attention particulière.

Tout d'abord, il est essentiel d'identifier les deux features concernés, afin de comprendre comment ils ont évolué au fil du temps. En comparant les caractéristiques initiales utilisées dans notre analyse et les données actuelles, nous pouvons détecter les changements significatifs.

Une fois que nous avons identifié les features affectés, nous devons évaluer l'impact du data drift sur notre modèle. Les variations dans ces features peuvent potentiellement affecter la performance et la stabilité du modèle de prédiction du risque de défaut de paiement.

Pour gérer le data drift, plusieurs approches peuvent être envisagées. Tout d'abord, nous devons déterminer si les variations observées sont dues à des erreurs de mesure ou à de véritables changements dans le comportement des clients. Une analyse plus approfondie des raisons sous-jacentes du data drift peut aider à prendre des décisions éclairées.

Ensuite, nous pouvons envisager de mettre à jour périodiquement notre modèle en utilisant les nouvelles données disponibles. Cela implique de réentraîner le modèle en intégrant les données les plus récentes et de réévaluer ses performances. Une mise à jour régulière du modèle peut garantir qu'il reste adapté aux données actuelles et aux changements dans les comportements des clients.

En outre, il peut être nécessaire d'effectuer une réévaluation des features les plus importantes sélectionnées précédemment. Les variations dans les données peuvent influencer la pertinence de ces features, et il peut être nécessaire de les mettre à jour en fonction des nouvelles tendances observées.

Enfin, un suivi continu du data drift est crucial pour détecter rapidement les variations et prendre des mesures appropriées. L'utilisation de techniques de surveillance en temps réel peut nous alerter sur les changements significatifs dans les données et nous permettre de prendre des mesures correctives plus rapidement.

En conclusion, l'analyse du data drift est une étape cruciale pour maintenir la performance et la pertinence du modèle de prédiction du risque de défaut de paiement. En identifiant les variations dans les features et en adaptant notre modèle en conséquence, nous pouvons garantir sa robustesse et sa précision dans un environnement en évolution constante.