NOTE METHODOLOGIQUE

I- Méthodologie d'entrainement du modèle

A- Exploration des données

Objectif : Comprendre la nature des données à disposition

1) Dimension des tableaux

Une analyse du nombre de tableaux et de leurs dimensions permet de comprendre la charge liée à la mémoire machine nécessaire pour effectuer le travail. Ainsi, il y a 8 fichiers csv de tailles relativement grandes.

2) Information sur les variables

Le type de variable et leurs nombres par tableau est à noter afin éventuellement de regrouper en variables catégorielles et variables numériques. Cela permet des approches de normalisations adaptées. Le tableau HomeCredit_columns_description fournit les explications des variables contenues dans les colonnes des autres fichiers. C'est un élément utile pour la compréhension métier.

3) Analyse des valeurs manquantes

Il s'agit de déterminer la taille des valeurs manquantes afin de choisir des méthodes de traitement adaptées. En effet, avec un nombre de valeurs manquantes, on pourrait être amené à juste les supprimer, sinon pour un nombre relativement grand, on pourra adopter d'autres approches. L'objectif étant d'avoir une qualité de données autant meilleure que possible. Un aller-retour entre la modélisation et les méthode de traitement des valeurs manquantes/aberrantes pourra être fait. La méthode retenue est de remplacer les valeurs manquantes de la variable cible TARGET par la valeur 0. Pendant que pour les variables catagorielles, nous avons le remplacement par le mode et la moyenne pour les variables numériques. Par ailleurs, la feature engineering conduit à des valeurs infinies. Aussi, inf est remplacé par la grande valeur 10 000.

4) Analyse de la distribution de la variable cible

L'analyse de la distribution de la variable cible conduit à sélectionner ou pas une méthode de rééquilibrage. Dans le cas ici, la répartition 95 % - 5 % à suggérer un rééquilibrage par la méthode SMOTE avec une majoration de données. En effet, la méthode SMOTE est l'une des plus utilisées dans ce cas. Elle donne de bons résultats. Elle peut être utilisée par minoration ou par majoration. Toutefois, dans note cas, le déséquilibre étant très grand, la minoration conduirait à réduire très signification notre échantillon d'entrainement, conduire à un sous-ajustement et à une mauvaise généralisation de l'ensemble de test. L'approche de majoration est donc retenue.
Une autre approche serait d'optimiser la métrique de la précision (Faux Négatifs vs Faux Positifs). Mais son handicap est sa non-généralisation à de nouveaux échantillons. En effet, étant une métrique empirique, sa validité sur un échantillon donnée ne peut se généraliser théoriquement à d'autres échantillons.

5) Distribution de certaines variables pertinentes

Une analyse univariée de certaines variables, notamment leurs graphes permet de comprendre leurs caratériques (moyenne, écarttype, max, min, proportion des classes,…) et ainsi éventuellement en déduire d'autres variables ou approches de normalisation.

6) Feature engineering et traitement des valeurs aberrantes

Un kernel a été fourni avec des features ajoutées. Il s'agit donc de l'adapter à notre étude-ci. Un traitement des valeurs aberrantes (infinis,…) a été fait pour rendre ces variables exploitables.

B- Chargement des données

Les fichiers csv ont ensuite été chargés traités puis concatenés. Toutefois, la taille du fichier final a induit de choisir un sous échantillon de 100 000 données pour l'entrainement des modèles.

C- Entrainement des modèles

Une baseline modèle a été établie afin de comparer les modèles à tester. Ainsi, la baseline modèle est la probabilité empirique d'obtenir une classe (0 ou 1). Le traitement par la SMOTE du déséquilibre des échantillons conduit à une baseline indiquant une probabilité de 50 % pour chaque classe.

Les modèles de classification entraînés sont le KNN et la régression logistique. Une première approche est de tester avec des paramètres arbitraires afin d'en percevoir la pertinence par rapport à la baseline modèle.

II- La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation

Les modèles testés affichent des performances prometteuses relativement à la baseline modèle. Pour les évaluer de manière plus pertinente, nous avons utilisé des mesures notamment la matrice de confusion et la classification_report qui regroupe plusieurs métriques.

Toutefois dans le cas du scoring de crédit, le risque serait d'accorder un crédit à quelqu'un qui aurait en réalité un mauvais rating, mais que le modèle classerait avec un bon rating. Aussi, il est pertinent le minimiser les faux négatifs dans la matrice de confusion.

L'utilisation du gridSearchCV avec des intervalles de données pour les paramètres conduit à choisir les paramètres optimaux. Par ailleurs, les métriques précédemment citées permettent de sélectionner le modèle définitif. Ainsi, une attention particulière est accordée aux f1, aux précisions et aux temps de calcul selon les modèles. De fait, les métriques sur les deux modèles sont bons, le temps de calcul a été le facteur déterminent du choix de la Logistic Regression.

III- L'interprétabilité globale et locale du modèle

Lorsqu'on travaille avec des techniques de classification et/ou de régression, il est toujours bon de pouvoir « expliquer » ce que fait le modèle. Grâce au Local Interpretable Model-agnostic Explanations (LIME), nous avons la possibilité de fournir rapidement des explications visuelles du modèle.

Nous pouvons tester l'exactitude et être sûrs que le classificateur et/ou le modèle sont "bons"... mais pouvons-nous décrire ce que le modèle fait réellement aux autres utilisateurs ?

L'utilisation de la librairie LIME permet cette interpretation globale et locale. Elle permet de comprendre la pertinence globale du modèle choisi appliqué. Mais, également, d'expliquer les variables qui influencent davantage la prédiction du modèle pour un élément de l'échantillon.

On pourrait également choisir le SHAP pour cette partie. Mais il fallait faire un choix. L'affichage et la simplicité d'affichage du LIME m'a paru plus pertinent pour l'usage qu'en ferait le conseiller de clientèle. Ainsi, j'ai opté pour cette méthode en me plaçant du point de vue utilisateur.

IV- Les limites et les améliorations

- L'utilisation d'autres métriques pour la sélection du modèle pourrait aider à répondre au mieux à la demande du client.
- D'autres modèles/algorithmes
- D'autres variables a créé par features engineering.

- Une aide métier pour la définition de la fonction coût.