Note Méthodologique

Ismail Can OGUZ

15/11/2023

Ce document s'inscrit dans le cadre du projet "Implémentez un modèle de scoring" du parcours Data Scientist d'OpenClassrooms. Il expose la méthodologie complète adoptée pour la modélisation et l'interprétabilité.

1 Contexte

Le projet a été initié en réponse aux besoins de la société "Prêt à Dépenser", une entreprise spécialisée dans les crédits à la consommation. L'objectif principal était de développer un modèle de scoring capable de prédire la probabilité de défaut de paiement pour des clients ayant peu ou pas d'historique de prêt. Cette initiative s'inscrit dans une démarche visant à renforcer la prise de décision en matière d'octroi de crédit et à accroître la transparence dans les processus de la société.

Les données nécessaires pour mener à bien ce projet comprenaient une base de données de 307 000 clients, chacun décrit par 121 caractéristiques, telles que l'âge, le sexe, l'emploi, le type de logement, les revenus, ainsi que des informations relatives au crédit et des notations externes.

2 Analyse Exploratoire des Données (EDA)

Pour mieux comprendre les données et faciliter le processus de modélisation, une analyse exploratoire des données (EDA) a été effectuée. Cela a impliqué plusieurs étapes clés de prétraitement et d'analyse, dont voici un résumé :

- **Encodage des Variables Catégorielles :** Les variables catégorielles ont été encodées à l'aide de la méthode de label encoding pour les variables binaires et one-hot encoding pour les autres. Cela garantit une représentation numérique appropriée des données.
- **Traitement des Anomalies :** Les données présentant une anomalie dans la variable DAYS EMPLOYED ont été identifiées et séparées. Le taux de défaut a été calculé pour les anomalies et les non-anomalies.
- **Exploration des Relations avec l'Âge :** Une analyse de la distribution de l'âge par rapport à la variable cible TARGET a été réalisée. Cela a permis de visualiser les différences dans le remboursement des prêts en fonction de l'âge.
- **Création de Nouvelles Colonnes :** Deux ensembles de données supplémentaires, *app_poly* et *app_domain*, ont été créés en ajoutant des colonnes supplémentaires basées sur une corrélation élevée avec la variable cible. Cependant, pour des raisons de simplicité, la préférence a été donnée à l'utilisation du premier ensemble de données (*app_train* et *app_test*).

3 Modélisation et Entraînement

Après une analyse approfondie des données, la phase suivante du projet consiste à construire et entraîner des modèles de machine learning pour prédire la probabilité de défaut de paiement des clients. Cette étape est cruciale pour prendre des décisions éclairées en matière d'octroi de crédit.

3.1 Choix des Caractéristiques et Ensembles d'Entraînement

Pour simplifier le processus tout en garantissant une performance adéquate, nous avons opté pour l'utilisation du jeu de données d'entraînement (*app_train*) en choisissant les 50 colonnes les plus corrélées avec la variable cible. Ce choix a été motivé par la connaissance préalable des valeurs cibles pour tous les identifiants de clients dans l'ensemble d'entraînement.

La séparation des données d'entraînement a été réalisée en utilisant une division de 80% pour l'entraînement et 20% pour la validation (*test_split*). Cette approche permet de réserver une partie des données pour évaluer la performance des modèles.

3.2 Modèles Évalués

Quatre modèles de classification ont été sélectionnés pour évaluation, chacun offrant des avantages distincts :

- **Dummy Classifier (Baseline) :** Ce modèle simple permet d'établir une référence de performance minimale.
- **Régression Logistique :** Un modèle linéaire adapté à notre problème de classification binaire.
- **Random Forest Classifier :** Un modèle d'ensemble robuste basé sur des arbres de décision.
- **XGBoost et LightGBM :** Deux algorithmes de boosting efficaces qui améliorent la performance prédictive.

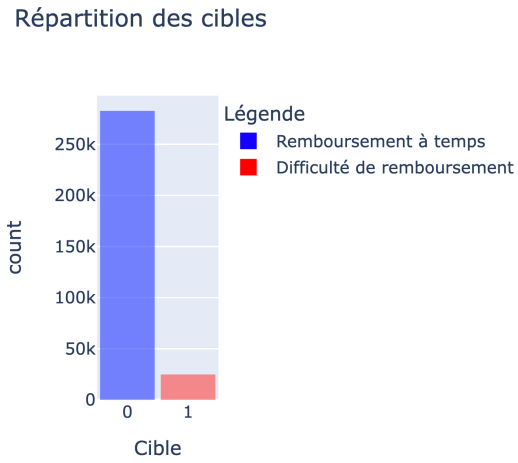

Figure 1 – Distribution of the Target Variable

3.3 Analyse de la Distribution de la Variable Cible

La Figure 1 illustre la distribution de la variable cible (*TARGET*). En examinant les valeurs normalisées, on observe une disparité significative, où environ 92% des échantillons appartiennent à la classe 0

(remboursement du prêt à temps) et seulement 8% à la classe 1 (difficulté de remboursement).

Cette disparité souligne le déséquilibre des classes dans le jeu de données, ce qui peut entraîner des problèmes de performance pour les modèles de machine learning, car ils peuvent être biaisés vers la classe majoritaire.

3.4 Traitement du Déséquilibre des Classes

Pour atténuer le déséquilibre des classes, deux approches ont été utilisées :

— **Réglage des Hyperparamètres :** Nous avons exploré l'ajout de nouveaux hyperparamètres aux modèles de machine learning, tels que *class_weight* ou *scale_pos_weight*, pour donner plus de poids à la classe minoritaire.

— **Suréchantillonnage et Sous-échantillonnage :** Nous avons appliqué la technique SMOTE (Synthetic Minority Over-sampling Technique) pour suréchantillonner la classe minoritaire et RandomUnderSampler pour sous-échantillonner la classe majoritaire. Cette combinaison vise à équilibrer les classes et à améliorer la performance du modèle sur les données déséquilibrées.

4 Évaluation des Performances des Modèles

Ce projet de science des données présente un défi majeur lié au déséquilibre de la variable cible, le défaut de paiement. Afin de résoudre ce problème, deux approches ont été explorées, et la performance des modèles a été évaluée en utilisant des métriques spécifiques prenant en compte ce déséquilibre.

4.1 Choix des Métriques d'Évaluation

Lors de l'évaluation des modèles, il est crucial de choisir des métriques adaptées au problème de déséquilibre de la variable cible. Les métriques traditionnelles telles que l'exactitude (*Accuracy*) et le score F1 ne sont pas suffisantes dans ce contexte, car elles peuvent être biaisées en faveur de la classe majoritaire. Dans ce scénario, où seulement environ 8% des clients présentent un défaut de paiement, un modèle peut atteindre une bonne précision en simplement prédisant la classe majoritaire.

Pour surmonter cette limitation, les métriques suivantes ont été privilégiées :

4.1.1 AUC (Area Under the Curve)

La métrique AUC évalue la capacité du modèle à discriminer entre les classes. Une valeur élevée d'AUC indique une bonne capacité du modèle à distinguer entre les clients qui remboursent et ceux qui ne remboursent pas.

$$\text{AUC} = \int_0^1 \text{Courbe ROC}\, dx$$

4.1.2 Score F2

Le score F2 met davantage l'accent sur la diminution des faux négatifs. Dans le contexte des prêts, un faux négatif (prédire qu'un client remboursera alors qu'il ne le fera pas) peut avoir des conséquences financières graves pour la banque. Ainsi, maximiser le score F2 contribue à minimiser les faux négatifs.

$$\text{Score F2} = \frac{(1+\beta^2) \cdot \text{Précision} \cdot \text{Rappel}}{\beta^2 \cdot \text{Précision} + \text{Rappel}}$$

(où $\beta = 2$ pour donner plus de poids au rappel)

4.1.3 Rappel (Recall)

Le rappel mesure la capacité du modèle à identifier tous les exemples positifs. Un rappel élevé est souhaitable car il réduit les risques de faux négatifs.

$$\text{Rappel} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}}$$

4.1.4 Précision (Precision)

La précision mesure la précision des prédictions positives. Bien que la précision soit importante, elle doit être interprétée conjointement avec le rappel, car un modèle peut avoir une précision élevée en prédisant principalement la classe majoritaire.

$$\text{Précision} = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}}$$

4.1.5 Exactitude (Accuracy)

Bien que moins informatif dans le cas de classes déséquilibrées, l'exactitude est toujours présentée pour une vue d'ensemble de la performance globale.

$$\text{Exactitude} = \frac{\text{Vrais Positifs} + \text{Vrais Négatifs}}{\text{Total des Prédictions}}$$

4.2 Analyse des Résultats des Modèles

Le tableau ci-dessous présente les résultats obtenus par différents modèles dans le cadre de la prédiction du défaut de paiement. Les métriques évaluées incluent l'AUC, le score F2 et le rappel.

Table 1 – Performance Metrics for Different Classifiers

Name	Test AUC	Test F2-score	Test Recall
LightGBM Classifier (With Class Weight)	0.757	0.718	0.676
XGBoost Classifier (With scale_pos_weight)	0.756	0.719	0.675
Random Forest (Resampled)	0.694	0.797	0.405
Random Forest Classifier (With Class Weight)	0.740	0.738	0.614
Random Forest Classifier (No Class Weight)	0.741	0.905	0.000
Logistic Regression (Resampled)	0.739	0.703	0.663
Logistic Regression (Weighted)	0.741	0.703	0.664
Logistic Regression (No Weight)	0.740	0.905	0.007
Dummy Classifier	0.499	0.853	0.079

4.2.1 Observations et Comparaison des Modèles

Meilleur Modèle : Le modèle LightGBM avec poids de classe offre la meilleure performance globale en termes d'AUC, de rappel, et de score F2 sur l'ensemble de test.

XGBoost et LightGBM : Ces deux modèles de boosting (XGBoost et LightGBM) surpassent les autres modèles dans plusieurs métriques, montrant leur efficacité dans la gestion du déséquilibre de la variable cible.

Random Forest (Resampled) : Bien que le Random Forest avec suréchantillonnage offre une performance honorable, il est dépassé par XGBoost et LightGBM en termes de rappel et de score F2.

Impact des Poids de Classe : L'utilisation de poids de classe semble bénéfique pour Random Forest et Logistic Regression, mais LightGBM reste le plus performant même sans poids de classe.

Durée d'Entraînement : LightGBM et XGBoost atteignent des performances remarquables avec un temps d'entraînement relativement court par rapport au Random Forest (Resampled).

En conclusion, LightGBM avec poids de classe se distingue comme le choix optimal pour résoudre le problème de déséquilibre de la variable cible, offrant un équilibre optimal entre AUC, rappel et score F2.

Après l'entraînement initial sur un sous-ensemble de l'ensemble de données d'entraînement, les modèles ont été évalués en utilisant des métriques telles que l'AUC, l'exactitude, la précision moyenne, le score F1, le score F2, la précision, le rappel, etc. LightGBM a émergé comme le modèle optimal, sélectionné grâce à une recherche en grille pour les meilleurs hyperparamètres basés sur le score *roc_auc*. Par la suite, ce modèle a été appliqué à l'ensemble complet des données. L'entraînement s'est déroulé sur la totalité de l'ensemble d'entraînement, tandis que l'ensemble de test a été utilisé pour évaluer les performances du modèle. Avant l'application des modèles, une étape de prétraitement a été effectuée, impliquant l'utilisation de MinMaxScaler pour la normalisation des données et SimpleImputer pour traiter les valeurs manquantes. La figure ci-dessous présente la matrice de confusion illustrant la performance du modèle LightGBM appliqué à l'ensemble des données de test.

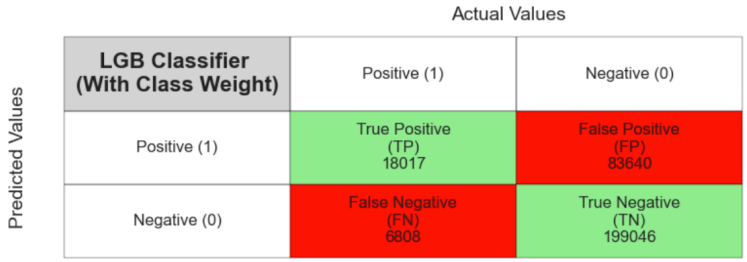

Figure 2 – Matrice de Confusion - Modèle LightGBM

5 L'interprétabilité globale et locale du modèle

Pour éclairer les facteurs qui influent le plus sur les prédictions du modèle LightGBM, nous avons utilisé la méthode SHAP (*SHapley Additive exPlanations*). Les valeurs d'importance des caractéristiques ont été obtenues en utilisant l'explorateur SHAP avec le modèle LightGBM, comme illustré par le code suivant :

```
explainer_shap = shap.TreeExplainer(lgbm_classifier)
shap_values = explainer_shap.shap_values(X_train_encoded)
```

- `EXT_SOURCE_3` : Score normalisé provenant d'une source de données externe.
- `EXT_SOURCE_2` : Score normalisé provenant d'une source de données externe.
- `EXT_SOURCE_1` : Score normalisé provenant d'une source de données externe.
- `AMT_CREDIT` : Montant du crédit demandé par le client.
- `DAYS_EMPLOYED` : Nombre de jours avant la demande où la personne a commencé son emploi actuel.
- `DAYS_BIRTH` : Âge du client en jours au moment de la demande.

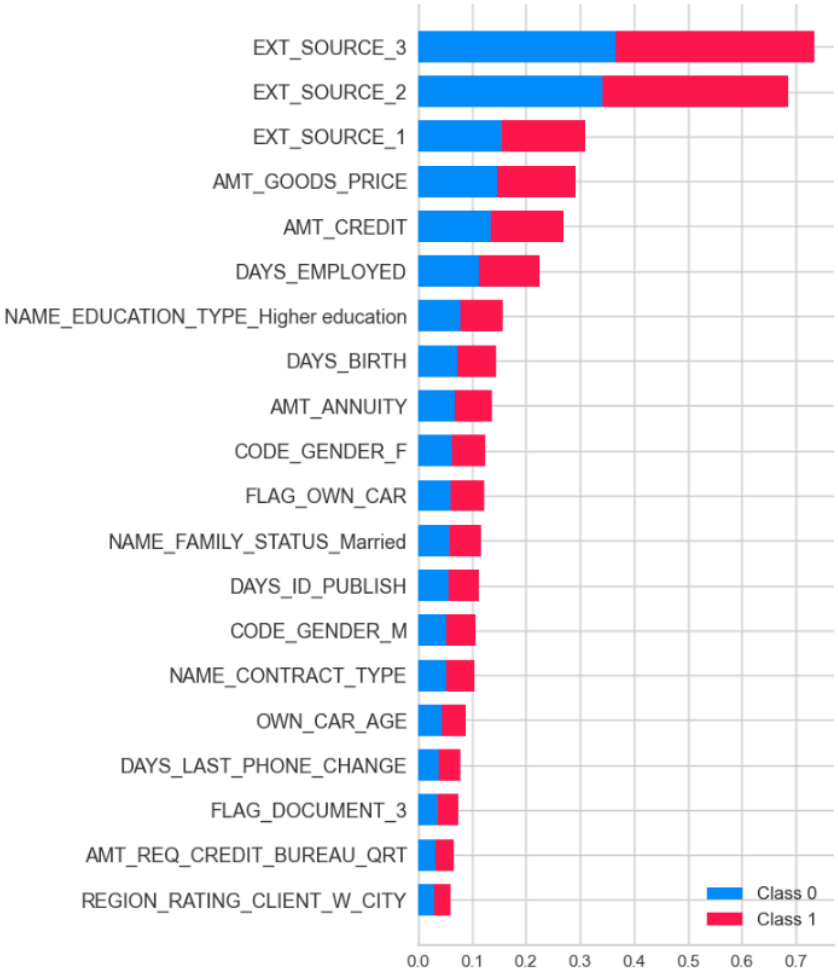

Figure 3 – Importance globale des caractéristiques par LightGBM (Top 30)

— `AMT_GOODS_PRICE` : Pour les prêts à la consommation, c'est le prix des biens pour lesquels le prêt est accordé.
— `AMT_ANNUITY` : Annuité du prêt.
— `NAME_EDUCATION_TYPE_Higher education` : Niveau d'éducation le plus élevé atteint par le client (éducation supérieure).
— `DAYS_ID_PUBLISH` : Nombre de jours avant la demande où le client a changé le document d'identité avec lequel il a demandé le prêt.

Ces caractéristiques jouent un rôle crucial dans la prédiction du modèle, comme le montrent leurs valeurs SHAP élevées. Par exemple, les scores normalisés provenant de sources externes (`EXT_SOURCE_1`, `EXT_SOURCE_2`, `EXT_SOURCE_3`) indiquent l'importance des informations externes dans l'évaluation du risque de crédit. De même, des caractéristiques telles que l'âge du client (`DAYS_BIRTH`) et le montant du crédit demandé (`AMT_CREDIT`) jouent un rôle significatif dans la décision de crédit. Ces résultats offrent une compréhension essentielle des éléments influents pour une prise de décision plus transparente et justifiable.

5.1 Importance Locale des Caractéristiques avec Shap (LightGBM)

Pour ce client particulier, dont le profil semble être de bonne qualité, les dix principales caractéristiques les plus importantes pour la prédiction du défaut de paiement, basées sur les valeurs Shap, sont les suivantes :

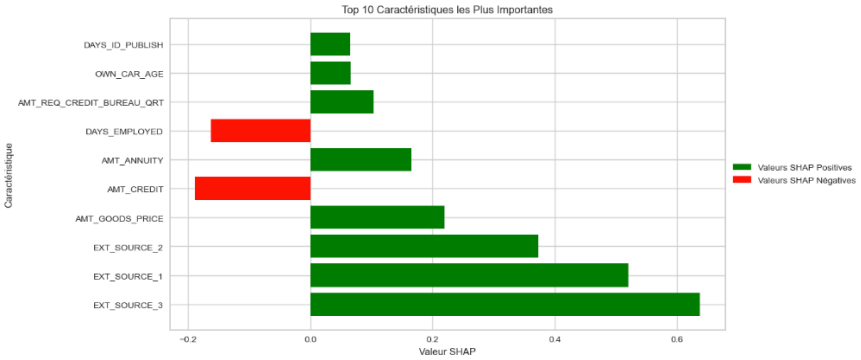

Figure 4 – Importance local des caractéristiques par LightGBM (Top 10)

Il est important de noter que, à l'exception des colonnes **AMT_CREDIT** et **DAYS_EMPLOYED**, toutes les autres caractéristiques présentent des valeurs Shap positives. Cela indique que ces caractéristiques contribuent positivement à la prédiction, suggérant que le crédit sera probablement accepté pour ce client.

Les valeurs Shap positives indiquent l'impact positif de chaque caractéristique sur la prédiction du modèle. Dans le contexte de ce projet, cela signifie que des caractéristiques telles que **EXT_SOURCE_3**, **EXT_SOURCE_1**, etc., ont des valeurs qui influencent favorablement la probabilité de remboursement du prêt. Cela renforce l'idée que le profil de ce client est favorable, et le modèle est enclin à prédire un remboursement positif du prêt.

6 Les limites et les améliorations possibles

Malgré les performances prometteuses du modèle LightGBM dans la prédiction du défaut de paiement, plusieurs limites et opportunités d'amélioration méritent d'être soulignées.

6.1 Limites du Modèle

Bien que le modèle ait montré une bonne capacité à gérer le déséquilibre de la variable cible, il peut encore présenter des défis dans des situations où les schémas de défaut de paiement évoluent avec le temps. Les modèles de machine learning, y compris celui-ci, sont sensibles aux changements dans les distributions des données, et des mises à jour fréquentes peuvent être nécessaires pour maintenir la précision.

De plus, la capacité d'interprétation des modèles complexes tels que LightGBM peut être limitée. Bien que des outils tels que SHAP permettent d'expliquer les prédictions au niveau des caractéristiques, une compréhension complète des mécanismes internes du modèle peut rester difficile.

Il convient également de noter que l'analyse s'est concentrée sur le jeu de données `app_train`, qui est la version par défaut sans colonnes supplémentaires générées par des méthodes d'ingénierie des caractéristiques. D'autres fichiers, tels que `bureau.csv`, `bureau_balance.csv`, `POS_CASH_balance.csv`, `credit_card_balance.csv`, `previous_application.csv`, et `installments_payments.csv`,

contiennent des informations supplémentaires qui pourraient améliorer la performance du modèle avec une exploration plus approfondie.

6.2 Améliorations Possibles

Une amélioration potentielle réside dans l'exploration de nouvelles caractéristiques ou l'ingénierie de caractéristiques supplémentaires provenant de fichiers tels que `bureau.csv`, `previous_application.csv`, et `installments_payments.csv`. Ces fichiers fournissent des informations sur les crédits précédents des clients et leurs remboursements.

L'intégration de données externes, si disponibles, pourrait enrichir la qualité des prédictions. Des informations supplémentaires sur le comportement financier des clients provenant de sources externes pourraient renforcer la robustesse du modèle.

En ce qui concerne l'entraînement du modèle, il pourrait être bénéfique d'explorer différentes combinaisons de modèles et d'hyperparamètres, ainsi que d'essayer des valeurs de F-score alternatives en plus du F2-score.

Il convient de noter que ces suggestions d'améliorations sont prospectives et nécessitent une évaluation approfondie avant leur implémentation.

7 Surveillance de la Performance du Modèle et Détection du Data Drift

La surveillance continue de la performance d'un modèle est essentielle pour garantir des prédictions fiables au fil du temps. Dans le cadre de ce projet, nous avons mis en place une stratégie de surveillance, notamment en analysant le data drift entre le jeu de données d'entraînement (`train`) et le jeu de données de test (`test`).

7.1 Analyse du Data Drift

Nous avons sélectionné aléatoirement 48 744 lignes de données à partir des deux ensembles de données, contenant chacun 240 colonnes. À l'aide de la bibliothèque Evidently, nous avons détecté des déviations significatives dans 11 colonnes. Les principales colonnes présentant un data drift incluent `AMT_REQ_CREDIT_BUREAU_QRT`, `AMT_REQ_CREDIT_BUREAU_MON`, `TARGET`, `AMT_GOODS_PRICE`, `AMT_CREDIT`, `AMT_ANNUITY`, `AMT_REQ_CREDIT_BUREAU_WEEK`, `NAME_CONTRACT_TYPE`, `DAYS_LAST_PHONE_CHANGE`, `FLAG_EMAIL`, et `AMT_INCOME_TOTAL`.

7.2 Interprétation des Résultats

La distance de Wasserstein normalisée et la distance de Jensen-Shannon sont des mesures de divergence statistique utilisées pour évaluer la similarité entre les distributions de deux ensembles de données. Une distance plus élevée indique une plus grande divergence.

La détection de data drift dans la colonne `TARGET`, la variable cible, est particulièrement préoccupante. Cela suggère que la distribution des étiquettes de classe a changé entre les ensembles d'entraînement et de test. Un tel changement peut compromettre la performance du modèle, car il a été formé sur des données qui ne reflètent plus fidèlement la réalité des données de test.

7.3 Implications pour la Maintenance du Modèle

La détection de data drift souligne la nécessité d'une maintenance proactive du modèle. Les changements dans les distributions des données peuvent affecter la généralisation du modèle, le rendant moins efficace dans la prédiction de nouvelles données.

Pour atténuer ces effets, des mises à jour régulières du modèle doivent être envisagées, en prenant en compte les nouvelles tendances dans les données. De plus, l'utilisation d'outils de surveillance continue et d'alertes, tels que celui implémenté dans ce projet, permet de détecter rapidement les dégradations de performance et d'initier des actions correctives.

En conclusion, la surveillance du data drift est cruciale pour assurer la fiabilité continue des prédictions du modèle, et les résultats de cette analyse doivent guider les actions futures de maintenance.