Développement d'un modèle de scoring crédit pour Prêt à dépenser

Antoine Fusilier

Prêt à dépenser

August 12, 2024

Introduction

- **Importance du scoring crédit** : Le scoring crédit est essentiel pour les institutions financières, permettant de minimiser les risques liés aux défauts de paiement tout en optimisant les opportunités de prêt.
- **Défis associés** : Interprétabilité, gestion du déséquilibre des classes, et minimisation des coûts d'erreurs.
- **Impact sur l'entreprise** : Un modèle efficace réduit les pertes financières, améliore la satisfaction client, et renforce la compétitivité de "Prêt à dépenser".

Description des données

- **Source des données** : Historique des prêts, informations financières des clients, et leur comportement de paiement.
- **Exploration des données** :
  - **Analyse des distributions** : Identifier les variables avec des distributions asymétriques ou des outliers.
  - **Matrice de corrélation** : Détecter les relations linéaires pour éviter la multicolinéarité.
  - **Formule de corrélation** :

$$\text{Corrélation} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- **Insights clés** : Les variables liées au revenu et à l'endettement montrent une forte corrélation avec les défauts de paiement.

Nettoyage des données

- **Gestion des valeurs manquantes** :
  - **Méthode d'imputation** : Imputation par la médiane pour les variables continues, par la mode pour les catégorielles.
  - **Méthodes avancées** : Imputation avec KNN pour conserver les relations entre variables.
  - **Impact sur le modèle** : L'imputation correcte des données manquantes améliore la robustesse du modèle.
- **Traitement des outliers** :
  - **Identification** : Utilisation de l'Isolation Forest pour détecter les anomalies dans les données.
  - **Méthode** : Transformation logarithmique pour les variables à distribution asymétrique.
  - **Formule d'Isolation Forest** :

$$\text{Score} = 2^{-\frac{\text{distance}}{\text{average path length}}}$$

  - **Conséquences** : Les outliers peuvent biaiser le modèle; leur traitement est donc essentiel pour des prédictions précises.
- **Normalisation des données** :
  - Utilisation du Z-score pour standardiser les variables :

Feature Engineering

- **Création de nouvelles variables** :
  - **Exemples** : Ratio dette/revenu, score d'endettement basé sur l'historique.
  - **Justification** : Les nouvelles variables capturent des informations critiques qui ne sont pas directement présentes dans les données brutes.
  - **Impact attendu** : Amélioration significative de la capacité prédictive du modèle.
- **Encodage des variables catégorielles** :
  - **Techniques utilisées** : One-Hot Encoding pour les variables sans ordre, Target Encoding pour les variables avec influence sur la cible.
  - **Risques** : Sur-ajustement possible avec Target Encoding, nécessitant une validation croisée rigoureuse.
- **Sélection des variables** :
  - **Méthodes** : PCA pour la réduction de dimensionnalité, analyse de variance pour identifier les variables pertinentes.
  - **Pourquoi ?** : Réduire la complexité du modèle tout en conservant l'information pertinente.

Modélisation

- **Algorithmes testés** :
  - **Régression Logistique** :

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

  - **Random Forest** :

$$\text{Prediction} = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

  - **XGBoost** :

$$\text{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{k} \Omega(f_j)$$

  - **LightGBM** :

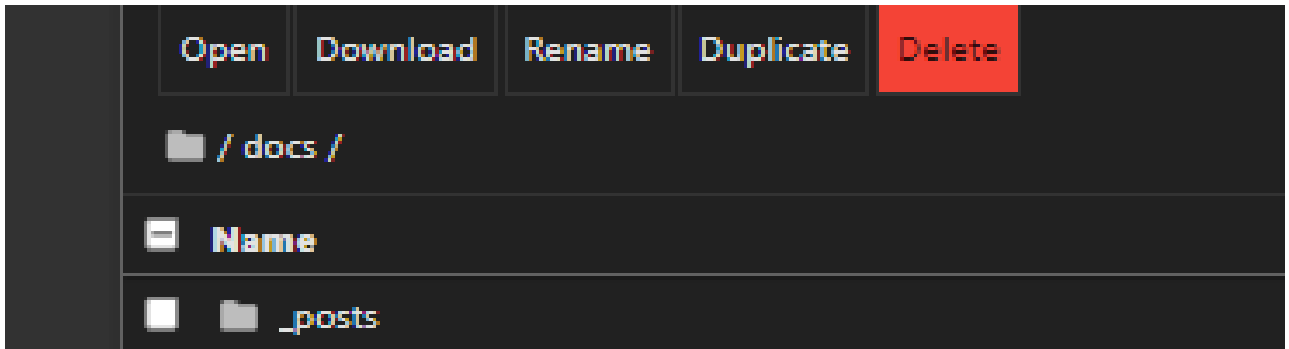
$$\text{Split Gain} = \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

- **Validation croisée** :
  - Cross-Validation avec GridSearchCV pour l'optimisation des hyperparamètres.

Comparaison des modèles

Table: Comparaison des modèles testés

Modèle	**AUC**	**Accuracy**	**F1-Score**	**Score métier**
Régression Logistique	0.75	0.70	0.65	0.68
Random Forest	0.80	0.75	0.72	0.74
XGBoost	0.82	0.77	0.74	0.78
LightGBM	0.83	0.78	0.75	0.80
SVM	0.77	0.72	0.68	0.70



Résultats des Modèles

- **Meilleur modèle retenu** : LightGBM
- **Performance** :
  - AUC: 0.83
  - Accuracy: 0.78
  - F1-Score: 0.75
  - Score métier: 0.80
- **Optimisation du seuil** : Ajustement du seuil de décision pour minimiser les coûts liés aux erreurs FN et FP.

$$\text{Score Métier} = \text{minimiser le coût} = FN \times 10 + FP$$

- **Implication** : Ce modèle offre un bon compromis entre précision et minimisation des coûts.

Interprétabilité du Modèle

- **SHAP values** : Permet d'expliquer l'importance des variables et les décisions individuelles du modèle.

$$\text{SHAP}(x_i) = \phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot [f(S \cup \{i\}) - f(S)]$$

- **Exemple d'interprétation** : Pour un client spécifique, la variable "revenu annuel" a eu un impact significatif sur la décision de refus du prêt.
- **Importance des variables globales** : Les variables financières (revenu, dette) ont le plus grand impact sur les prédictions.
- **Utilisation pratique** : Les chargés de relation client peuvent utiliser ces insights pour expliquer les décisions aux clients de manière claire et compréhensible.

Conclusion et Plan d'action

- **Modèle final retenu** : LightGBM, avec une AUC de 0.83 et un Score métier de 0.80.
- **Plan d'action pour le déploiement** :
  - **Étape 1** : Déploiement du modèle en production avec un suivi rigoureux des performances.
  - **Étape 2** : Mise en place d'un système de mise à jour continue des données pour éviter le concept drift.
  - **Étape 3** : Formation des chargés de relation client pour interpréter et expliquer les résultats du modèle.
- **Étapes futures** :
  - Amélioration continue du modèle en intégrant des données supplémentaires.
  - Test de nouvelles techniques de modélisation pour potentiellement améliorer la performance.