Rapport : TP Classification des Prêts Bancaires

1. Prétraitement des Données

(a) Chargement et Analyse Exploratoire des Données (EDA)

- Les données ont été chargées depuis donnees_pret_avance.csv.
- Une analyse exploratoire a été réalisée :
    - La répartition des classes de la cible Pret_Approuve a révélé un **déséquilibre** important : 67% de prêts approuvés (1) contre 33% refusés (0).
    - Visualisation de la distribution des variables via des statistiques descriptives et un **countplot** des classes.
    - Variables influentes : Montant_Pret, Revenu, et Score_Credit.

(b) Gestion des Valeurs Manquantes

- Les valeurs manquantes des colonnes numériques (Age, Revenu, etc.) ont été remplacées par la **médiane** pour éviter les biais.
- Les valeurs manquantes des colonnes catégorielles (Statut_Emploi, Niveau_Education) ont été remplacées par le **mode**.
- La colonne Statut_Marital a également été remplie avec la valeur la plus fréquente.

(c) Encodage des Variables Catégorielles

- Les variables catégorielles (Statut_Emploi, Niveau_Education, Statut_Marital) ont été encodées avec un **One-Hot Encoding**.
- Le nombre total de variables après encodage est passé de 9 à **16 variables**.

(d) Ingénierie des Caractéristiques

- Une nouvelle variable, Ratio_Endettement (**Montant_Pret / Revenu**), a été créée pour évaluer le poids de l'endettement d'un client.
- Les valeurs infinies et NaN ont été remplacées par 0.

(e) Division des Données

- Les données ont été divisées en **70% pour l'entraînement** et **30% pour le test**, en conservant la distribution des classes.

2. Traitement du Déséquilibre des Classes

(a) Analyse du Déséquilibre

- Répartition initiale :
    - 67% des prêts sont approuvés.
    - 33% des prêts sont refusés.

(b) Techniques de Rééchantillonnage

Trois techniques ont été appliquées pour équilibrer les classes :

1. **Sous-échantillonnage** : Nombre égal de classes positives et négatives (469/469).
2. **Sur-échantillonnage** (SMOTE) : Synthèse de nouvelles données pour équilibrer les classes (931/931).

- Résultat : **SMOTE** a montré de meilleures performances lors de l'évaluation des modèles.

3. Modélisation

(a) Régression Logistique

- Entraînement effectué avec **validation croisée stratifiée (k=5)**.
- Précision obtenue :
    - Sous-échantillonnage : **0.89**.
    - SMOTE : **0.90**.

(b) XGBoost avec Optimisation des Hyperparamètres

- Optimisation via **GridSearchCV** (paramètres : n_estimators, max_depth, learning_rate).
- Précision obtenue :
    - Sous-échantillonnage : **1.00**.
    - SMOTE : **1.00**.

(c) Perceptron Multicouche (MLPClassifier)

- Réseau de neurones avec hidden_layer_sizes=(50, 30) et learning_rate_init=0.01.
- Précision obtenue :
    - Sous-échantillonnage : **1.00**.
    - SMOTE : **1.00**.

4. Évaluation

(a) Métriques Utilisées

- **Précision**, **Rappel**, **F1-Score**, **AUC**.
- XGBoost et Perceptron Multicouche ont atteint une **précision parfaite** sur les ensembles équilibrés.

(b) Courbes ROC et Precision-Recall

- Les courbes **ROC** et **Precision-Recall** ont été tracées pour chaque modèle :
    - XGBoost et Perceptron Multicouche affichent une **courbe ROC parfaite** (AUC proche de 1).

(c) Comparaison des Modèles

- **XGBoost** et **Perceptron Multicouche** surpassent largement la **Régression Logistique**.
- **SMOTE** semble plus adapté pour le traitement du déséquilibre.

5. Interprétation et Explicabilité

(a) Importance des Caractéristiques

- **Gini Importance** (XGBoost) :
    - Les variables les plus influentes sont : Statut_Emploi_Etudiant, Niveau_Education_Secondaire, et Ratio_Endettement.
- **SHAP Summary Plot** :
    - Statut_Emploi et Ratio_Endettement ont le plus grand impact sur les prédictions.

(b) Applicabilité Bancaire

- Le modèle est applicable pour :
    - Identifier les clients présentant un **risque d'endettement élevé**.

    - Prédire la probabilité d'approbation du prêt en fonction des variables clés comme le **revenu**, l'**âge**, et le **statut d'emploi**.
- Il aide les institutions financières à prendre des décisions basées sur des variables explicites et interprétables.

6. Déploiement

(a) Sauvegarde du Modèle

- Le meilleur modèle XGBoost a été sauvegardé avec **joblib** (meilleur_modele_xgb.pkl).

(b) Fonction de Prédiction

- Une fonction a été créée pour :
    - Prendre les caractéristiques d'un client en entrée.
    - Retourner la **probabilité d'approbation du prêt**.
- Exemple de résultat :
    - Caractéristiques : Age=45, Revenu=60000, Montant_Pret=20000, etc.
    - **Probabilité d'approbation : 0.99 (99%)**.