Prédiction d'approbation de crédit

Auteur :
DIF Akli

13 décembre 2024

1 Introduction

Ce rapport présente en détail le travail réalisé dans le cadre du projet de prédiction de crédit. L'objectif principal du projet était de développer un modèle de machine learning précis pour déterminer si une demande de crédit serait approuvée ou non. Les données utilisées pour ce projet proviennent du site Kaggle, plus précisément du fichier CSV "train.csv", qui contient des informations sur les emprunteurs.

2 Méthodologie

2.1 Analyse Univariée

Nous avons effectué une analyse univariée des données pour mieux comprendre la distribution de la variable cible "Loan_Status". Voici quelques résultats clés :

- La majorité des demandes de crédit ont été approuvées (68.73%).
- La distribution des emprunteurs par genre montre une prédominance masculine (81.76%).
- La plupart des emprunteurs sont mariés (65.43%).
- La variable "Credit_History" indique que la plupart des emprunteurs ont un historique de crédit positif (84.85%).

2.2 Gestion des Données Manquantes

La première étape de notre prétraitement des données a consisté à gérer les valeurs manquantes. Nous avons adopté les approches suivantes :

- Variables Catégoriques (Gender, Married, Dependents, Credit_History, Self_Employed) : Les valeurs manquantes ont été remplacées par le mode de chaque variable respective.
- Variables Numériques (LoanAmount, Loan_Amount_Term) : Les valeurs manquantes ont été remplacées par la moyenne de chaque variable respective.

2.3 Prétraitement des Données

Pour préparer les données pour la modélisation, nous avons effectué les étapes suivantes :

- Suppression de la colonne 'Loan_ID'.
- Application de l'encodage One-Hot sur le DataFrame.
- Suppression des colonnes redondantes après l'encodage One-Hot.
- Renommage des colonnes pour plus de clarté.
- Calcul de l'écart interquartile (IQR) et suppression des valeurs aberrantes.
- Transformation des colonnes 'ApplicantIncome', 'CoapplicantIncome' et 'LoanAmount' en utilisant la racine carrée.
- Gestion du déséquilibre des classes en utilisant la technique SMOTE.
- Normalisation des données numériques en utilisant la technique StandardScaler.
- Division des données en ensembles d'entraînement (80%) et de test (20%).

3 Modélisation

Dans ce projet, plusieurs modèles de machine learning ont été testés pour déterminer lequel serait le plus approprié pour la prédiction de crédit. Les modèles suivants ont été testés :

- Régression Logistique
- Random Forest
- Support Vector Machine (SVM)
- Réseau de neurones

L'Implémentation des différenyts modéles est disponible dans le code dans le fichiers main.ipynb et main.py.

4 Résultats

En comparant les résultats des différents modèles, le model de Random-Forest a montré la meilleure performance avec un score de précision de 93%. Le modèle de régression logistique a montré le score le plus bas avec 68%. Le modèle de SVM a également montré de mauvais résultats avec un score de précision de 77%. Le réseau de neurones a montré un score de précision de 90%.

Suite à ces résultats, le modèle de Random Forest a été sauvegardé pour une utilisation dans une interface graphique pour la prédiction de crédit réalisé avec la bibliothèque PyQt5.

Ce projet démontre l'application de diverses techniques d'apprentissage automatique pour la prédiction de l'approbation de prêt, en mettant en lumière les avantages et les inconvénients de chaque méthode.