UQAC

Université du Québec à Chicoutimi

8PRO408-Outil de programmation pour la science des données

Rapport du Projet

Travail réalisé par,

Nguiffo Gilles Roy NGUG22060400

Meijomo Hillary Alexandra MEJM28530300

Rapport Complet – Analyse Exploratoire et Détection de Fraude sur Transactions Bancaires

1. Introduction

Ce rapport présente une analyse exploratoire complète (EDA) du dataset Credit Card Fraud Detection, contenant 284 807 transactions bancaires anonymisées, dont une proportion très faible de fraudes. L'objectif principal est de comprendre la structure du dataset, les relations entre ses variables, et d'identifier les patterns potentiellement associés à la fraude. Une modélisation optionnelle via un Random Forest est également incluse pour illustrer comment l'EDA peut guider la construction d'un modèle prédictif.

2. Présentation du Dataset

Le dataset contient 31 variables : Time, Amount, Class et 28 composantes principales (V1 à V28) issues d'une transformation PCA. L'anonymisation empêche l'interprétation directe des variables V1–V28, mais elles capturent des dimensions statistiques utiles à la détection de fraude.

3. Analyse Univariée

• La variable Amount présente une distribution fortement asymétrique, concentrée vers de petits montants.
• La variable Time représente le nombre de secondes écoulées depuis la première transaction.
• Les variables V1 à V28 montrent des distributions centrées autour de 0 (effet du PCA), mais certaines composantes comme V14, V10, V12 affichent des queues lourdes révélatrices de comportements anormaux.
• La variable Class est extrêmement déséquilibrée : environ 0.17% de transactions frauduleuses.

4. Analyse Bivariée et Corrélations

L'analyse bivariée a révélé des différences significatives entre les distributions des composantes PCA pour les transactions frauduleuses et normales, en particulier pour V14, V10, V12, V17, V11. Ces variables sont fortement corrélées à la classe (fraude) d'après la matrice de corrélation. La variable Amount influence modérément la probabilité de fraude, tandis que la variable Time n'apporte que peu d'information discriminante.

5. Modélisation (Optionnelle) – Random Forest

Bien que la modélisation ne soit pas exigée dans le cadre du projet, un modèle Random Forest a été construit afin d'illustrer comment l'EDA guide les choix de variables. Le dataset étant très déséquilibré, la technique SMOTE a été appliquée sur l'ensemble d'entraînement.

5.1. Résultats du Modèle

• Accuracy : 99.95%
• ROC-AUC : 0.975
• Rappel (fraude) : 82.65%
• Précision (fraude) : 87.10%
• F1-score (fraude) : 84.82%
Ces scores indiquent une excellente capacité du modèle à détecter les fraudes tout en minimisant les faux positifs.

5.2. Importance des Variables

Le modèle utilise principalement : V14, V4, V10, V12, V17, V3, V11, V16, V2 et V7. Ces variables regroupent plus de 60% de l'importance totale, confirmant les observations issues de l'EDA. Les variables Amount et Time apparaissent très marginales dans le processus de décision.

6. Application Streamlit

Une application interactive a été développée via Streamlit. Elle inclut :
• Un aperçu général du dataset
• Une visualisation interactive du montant des transactions
• Une analyse temporelle basée sur la transformation de Time en heures
• Une représentation des classes (normales vs frauduleuses)
Cette interface permet une exploration dynamique du dataset.

7. Conclusion Générale

L'analyse exploratoire a permis d'identifier clairement les variables les plus discriminantes pour la fraude, notamment V14, V10, V12 et V17. La distribution extrêmement déséquilibrée des classes souligne l'importance de techniques adaptées pour la modélisation. Le Random Forest, bien qu'optionnel, démontre que les patterns extrais de l'EDA permettent de construire un modèle performant. L'application Streamlit offre un moyen moderne et interactif d'explorer les données. Ce projet constitue une base solide pour la mise en production d'un système de détection de fraude ou l'ajout futur de modèles avancés (XGBoost, seuils adaptatifs, etc.).