Traitement des Données Déséquilibrées (Focus sur Smote): Une Étude de Cas sur la Détection de Fraude

Jules Odje

December 8, 2024

Abstract

Cette étude explore les stratégies de traitement des données déséquilibrées à travers une application de détection de fraudes bancaires. L'analyse compare différentes approches, avec un focus particulier sur la technique SMOTE, et évalue leur efficacité dans l'amélioration des performances de classification.

1 Introduction

1.1 Contexte

Le déséquilibre des classes est un défi majeur en apprentissage automatique, particulièrement présent dans des domaines comme la détection de fraude, où les cas positifs sont naturellement rares.

1.2 Choix du Dataset

Le dataset de fraudes par carte de crédit a été choisi pour plusieurs raisons :

- Déséquilibre naturel (0.17% de fraudes)
- Problématique réelle et critique
- Données standardisées (via PCA)
- Taille significative (284,807 transactions)

2 Méthodologie

2.1 Données et Prétraitement

- Dataset de transactions par carte de crédit
- Standardisation des variables
- Split train-test (80-20%)

2.2 Modèles Développés

- Modèle de base : Réseau neuronal dense
- Modèle avec SMOTE : Même architecture + rééquilibrage

3 Résultats

3.1 Performance du Modèle de Base

Matrice de confusion :

- Vrais Négatifs : 56850
- Faux Positifs : 14
- Faux Négatifs : 24
- Vrais Positifs : 74

3.2 Performance avec SMOTE

Métriques pour la classe frauduleuse :

- Précision : 0.71
- Recall : 0.85
- F1-score : 0.77

4 Analyse Comparative

4.1 Améliorations Observées

- Augmentation du recall ($0.76 \rightarrow 0.85$)
- Meilleure détection des fraudes
- Compromis précision-recall

4.2 Compromis et Limitations

- Légère baisse de précision
- Temps de traitement accru
- Complexité additionnelle

5 Recommandations d'Amélioration

5.1 Optimisations Techniques

- Ajustement de l'architecture du réseau
- Fine-tuning des hyperparamètres
- Expérimentation avec d'autres ratios de SMOTE

5.2 Approches Alternatives

- Techniques d'ensemble
- Autres méthodes de rééquilibrage
- Apprentissage sensible aux coûts

6 Conclusion

Le traitement des données déséquilibrées nécessite une approche réfléchie et adaptée au contexte. Cette étude démontre l'efficacité de SMOTE, tout en soulignant l'importance d'une évaluation complète des compromis impliqués.