Outils de programmation avancée pour l'IA

Master 2 Ingénierie en Intelligence Artificielle

Université Paris 8

Thème

Modèles de machine et deep learning pour la classification de fraude en Credit Cards

- CHABANE Khaled

13 décembre 2024

1 Introduction

Dans le cadre de ce projet, j'ai abordé un problème de classification visant à détecter des fraudes sur des cartes de crédit. Le jeu de données utilisé a été récupéré sur la plateforme Kaggle, et a fait l'objet d'un traitement minutieux afin de le préparer pour l'analyse. Trois modèles de machine learning ont été construits et évalués : un arbre de décision, un modèle de forêt aléatoire (Random Forest) et un réseau de neurones. L'objectif de ce projet est de comparer les performances de ces modèles pour déterminer celui qui offre la meilleure précision dans la détection des fraudes.

2 Dataset

Le dataset utilisé contient des transactions réalisées par des titulaires de cartes de crédit européennes en septembre 2013. Il présente un total de 284 807 transactions sur une période de deux jours, dont 492 sont des fraudes.

Toutes les variables d'entrée sont numériques et ont été obtenues à l'aide d'une transformation PCA (Analyse en Composantes Principales). Les caractéristiques V1, V2, ... V28 représentent ces composantes principales, tandis que les deux variables non transformées par PCA sont 'Time' et 'Amount'. La variable 'Time' indique le nombre de secondes écoulées entre chaque transaction et la première transaction du dataset. La variable 'Amount' correspond au montant de la transaction. Enfin, la variable cible 'Class' indique si la transaction est frauduleuse (1) ou non (0).

2.1 Choix de modèles

J'ai choisi trois modèles pour ce projet en raison de leurs forces complémentaires. L'arbre de décision est simple et interprétable, ce qui permet de comprendre facilement les règles de classification. La forêt aléatoire, en combinant plusieurs arbres, améliore la précision et réduit le surapprentissage, ce qui est utile pour des données déséquilibrées. Enfin, le réseau de neurones, bien que plus complexe, est capable de capturer des relations non linéaires complexes, ce qui peut améliorer la détection des fraudes dans des ensembles de données volumineux et variés.

2.2 Discussion

Le dataset utilisé présente un déséquilibre de classes, avec une classe majoritaire (transactions non frauduleuses) et une classe minoritaire (transactions frauduleuses). Ce déséquilibre peut entraîner des biais dans les modèles, favorisant la classe majoritaire. Avant de construire les modèles, j'ai pris des mesures pour traiter ce déséquilibre en calculant les poids nécessaires pour chaque classe à l'aide de la fonction computeclassweights. Ces poids ont ensuite été appliqués pour ajuster l'importance de chaque classe dans le modèle, en multipliant les classes par ces poids. Cette approche a permis d'améliorer la détection des fraudes en équilibrant mieux les classes, assurant ainsi une meilleure performance du modèle.

2.3 Résultats

RN :
Accuracy : 0.992205 F1 Score : 0.271335 Recall : 0.837838
Arbre de décision :
Accuracy : 0.996922 F1 Score : 0.468687 Recall : 0.783784
Random Forest :
Accuracy : 0.992205 F1 Score : 0.802817 Recall : 0.770270

Ici dans notre cas :Le modèle Random Forest semble être le meilleur choix parmi les trois. Il offre un bon compromis entre précision et rappel, avec le plus haut F1 Score (0.8028), ce qui est particulièrement important pour un problème de détection de fraude où il est crucial d'éviter à la fois les faux positifs et les faux négatifs.

3 Conclusion