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 détection de fraude sur les cartes de crédit

- CHABANE Khaled

13 décembre 2024

Table des matières

1 Introduction

Pour ce projet, j'ai travaillé sur un problème de classification portant sur la détection de fraudes liées aux cartes de crédit. J'ai récupérer jeu de données depuis la plateforme Kaggle, et je l'ai préparé pour l'analyse grâce à un traitement approfondi.

J'ai développé et évalué trois modèles de machine learning : un arbre de décision, une forêt aléatoire (Random Forest) et un réseau de neurones. L'objectif principal était de comparer leurs performances afin d'identifier celui qui se montre le plus précis pour détecter les 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).

3 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.

4 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 compute-class-weights. 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.

5 Résultats

RN :
Accuracy : 0.992334 F1 Score : 0.271413 Recall : 0.824324

Arbre de décision :
Accuracy : 0.996922 F1 Score : 0.468687 Recall : 0.783784

Random Forest :

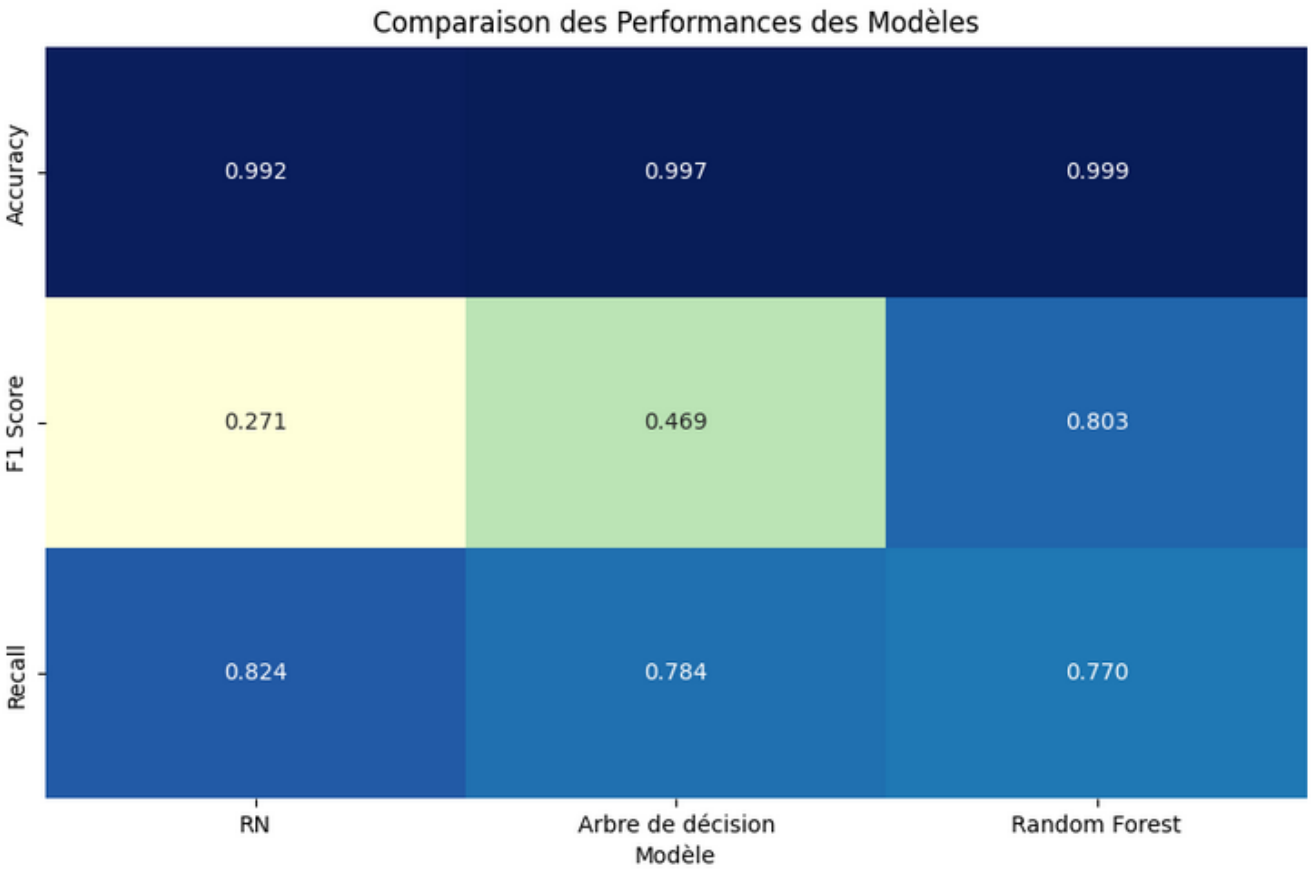

FIGURE 1 – Plot de comparaison.

Accuracy : 0.999345 F1 Score : 0.802817 Recall : 0.770270

Accuracy : Dans le cadre de la détection de fraude, l'accuracy n'est pas le critère principal. Étant donné que les ensembles de données sont souvent déséquilibrés (les fraudes étant rares par rapport aux transactions normales), un modèle peut avoir une haute accuracy en prédisant principalement la classe majoritaire sans jamais identifier les fraudes réelles.

F1 Score : Le F1 Score est plus pertinent dans des contextes de classes déséquilibrées, car il équilibre la précision et le rappel. Ici, Random Forest a le meilleur F1 Score (0.8028), ce qui montre qu'il est bien équilibré entre la capacité à détecter les fraudes et à minimiser les faux positifs.

Recall : Le recall est crucial pour la détection de fraude. Un recall élevé signifie que le modèle détecte une grande proportion des fraudes réelles. Le modèle RN présente le meilleur recall (0.8243), ce qui le rend très performant pour identifier les fraudes. Cependant, son faible F1 Score suggère qu'il y a un nombre élevé de faux positifs.

6 Conclusion

Dans le contexte de détection de fraude sur les cartes de crédit, où il est essentiel de minimiser les faux négatifs (fraudes non détectées), le modèle RN, avec son recall élevé (0.8243), est le plus adapté, car il capture une plus grande proportion des fraudes, bien qu'il présente un nombre élevé de faux positifs. Cependant, le modèle Random Forest offre un compromis intéressant, avec un excellent F1 Score (0.8028), assurant un bon équilibre entre la

détection des fraudes et la minimisation des faux positifs.

Donc, vu que l'objectif ici est de détecter un maximum de fraudes, même au prix d'un nombre plus élevé de faux positifs alors, le Réseau de Neurones est le meilleur choix.