Analyse des transactions bancaires et détection de fraude

1. Introduction

Ce projet a pour objectif d'explorer un dataset réel de transactions bancaires afin d'identifier les caractéristiques permettant de différencier une transaction normale d'une transaction frauduleuse.
Le dataset comprend 284 807 transactions dont seulement 492 fraudes (0,17 %), ce qui en fait un dataset extrêmement déséquilibré et représentatif des difficultés réelles liées à la détection de fraude.

Les variables sont principalement issues d'une transformation PCA, accompagnées de deux variables originales (*Time* et *Amount*) et d'une variable cible (*Class*). L'analyse se concentre sur l'étude des distributions, des montants, de l'aspect temporel et des patterns associés aux fraudes.

2. Exploration des données

2.1 Distribution des montants

L'histogramme du montant des transactions révèle une asymétrie très forte :

- la majorité des transactions concernent de petits montants,
- quelques transactions atteignent des valeurs très élevées.

Cette distribution est typique des données financières et rend utile l'utilisation d'échelles logarithmiques dans les visualisations.

2.2 Statistiques descriptives

Les statistiques montrent une forte variabilité des montants (jusqu'à environ 25 691 €). La variable *Time* correspond au temps écoulé depuis la première transaction et ne présente pas de structure temporelle marquée.

3. Analyse de la variable cible

La variable *Class* est extrêmement déséquilibrée :

- 284 315 transactions normales,
- 492 transactions frauduleuses.

Ce déséquilibre implique que les métriques usuelles (accuracy) ne sont pas adaptées, et que des techniques de rééquilibrage seraient nécessaires pour un modèle prédictif.

Le boxplot comparant *Amount* selon *Class* montre que les transactions frauduleuses ont des montants souvent plus extrêmes et une dispersion plus forte. Même si le montant ne suffit pas à lui seul pour caractériser la fraude, il reste une variable informative.

4. Analyse temporelle

L'étude du nuage de points (*Time vs Amount*) montre que :

- les fraudes apparaissent à des moments variés,
- il n'existe pas de motif temporel clair,
- la variable *Time* n'est pas un indicateur fiable de fraude.

Les fraudes semblent donc réparties de manière relativement aléatoire dans le temps.

5. Visualisations interactives (Streamlit)

Une application Streamlit a été créée afin d'explorer les données de façon interactive. Elle permet notamment de visualiser la distribution des montants, la comparaison normal/fraude, et un échantillon sur un scatter plot.

Application disponible ici :

https://fraudanalysishajarchakir-gxffzjtjviske9ay8cyr8y.streamlit.app/

6. Principales observations

- Le dataset est très déséquilibré, ce qui complique la détection automatique.
- Les fraudes présentent des montants plus extrêmes et plus dispersés.
- La variable *Time* n'explique pas la fraude.
- Certaines composantes PCA pourraient être pertinentes, mais sont difficiles à interpréter.
- Un futur modèle prédictif devra absolument prendre en compte le déséquilibre des classes.

7. Conclusion

L'analyse met en évidence plusieurs comportements associés à la fraude, notamment des montants plus élevés et une variabilité plus importante. Cependant, aucune variable ne permet à elle seule de prédire la fraude, ce qui montre la complexité du phénomène.

Cette analyse exploratoire constitue une première étape essentielle avant la construction d'un modèle de machine learning destiné à détecter les transactions frauduleuses.