SUP MTI
المدرسة العليا لإدارة الأعمال، المعلوميات والاتصالات
École Supérieure de Management, de Télécommunications et d'Informatique
RECONNUE PAR L'ÉTAT

PROJET BIG DATA FINANCE

Détection de fraude sur les transactions bancaires

Analyse et traitement Big Data

Présenté par :

Zalla Ngodi Astrel du Chemin
Diarrassouba Nassong Aminata
Okamba Grâce Caleb

Année universitaire 2025 – 2026

Table des matières

Introduction

Dans un monde de plus en plus numérisé, les systèmes de paiement électronique sont devenus le pilier de l'économie mondiale. Cependant, cette dématérialisation des échanges s'accompagne d'une recrudescence des activités criminelles, rendant la sécurité des transactions bancaires plus critique que jamais. La détection manuelle des fraudes étant devenue impossible face au volume massif de données générées chaque seconde, l'utilisation du **Machine Learning** et des technologies **Big Data** s'impose désormais comme une nécessité stratégique.

C'est précisément dans ce contexte que s'inscrit notre étude, soulevant la problématique majeure de la distinction, en temps réel et avec une précision extrême, entre une transaction légitime et une tentative de fraude. Ce défi est d'autant plus complexe que les activités frauduleuses sont souvent noyées dans une masse de données licites, créant un déséquilibre de classe important qui peut fausser les analyses classiques.

Pour répondre à cet enjeu, l'objectif central de ce travail est de construire un modèle prédictif capable de détecter automatiquement ces anomalies à partir de données historiques. Le problème est abordé sous l'angle d'une **classification binaire** sur la variable cible `Class` extraite du dataset `creditcard.csv`

Chapitre 1

Contexte général des transactions bancaires

1.1 Historique des systèmes de paiement

L'évolution des moyens de paiement s'est faite par étapes marquantes, passant d'échanges physiques à une numérisation totale :

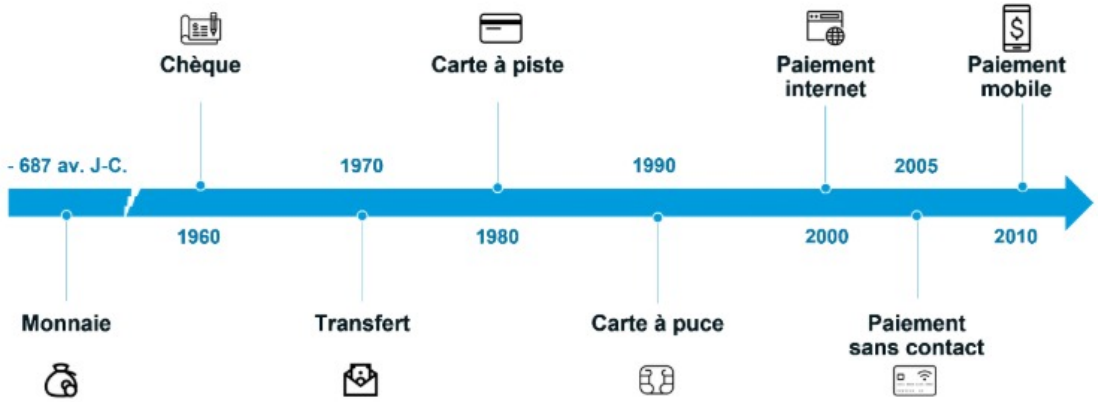


FIGURE 1.1 – Évolution historique des moyens de paiement

1.2 Le flux d'une transaction monétique

Le processus d'une transaction par carte bancaire implique plusieurs acteurs principaux : le porteur (client), le commerçant, la banque acquéreuse (celle du marchand) et la banque émettrice (celle du client). Chaque étape du flux génère des métadonnées cruciales que nous utiliserons pour notre analyse Big Data.

Chapitre 2

La fraude bancaire et ses enjeux

2.1 Typologie des fraudes

La fraude bancaire se décline sous plusieurs formes, allant de la manipulation psychologique à l'interception technique des données. Voici les principales techniques rencontrées :

Le Phishing (Hameçonnage)

Le *phishing* est une technique d'ingénierie sociale consistant à usurper l'identité d'une institution de confiance (banque, administration) via un email ou un SMS frauduleux. L'objectif est d'inciter la victime à fournir ses identifiants de connexion ou ses coordonnées bancaires sur un faux site web.

FIGURE 2.1 – Mécanisme d'une attaque par phishing

Le Skimming (Copie de carte)

Le *skimming* consiste à installer un dispositif physique discret sur un distributeur automatique (DAB) ou un terminal de paiement. Ce "skimmer" lit et copie les données de la piste magnétique de la carte, tandis qu'une micro-caméra filme souvent la saisie du code confidentiel.

FIGURE 2.2 – Mécanisme d'une attaque par Skimming

La fraude au paiement sans présence physique (CNP)

La fraude *Card Not Present* (CNP) concerne les transactions effectuées à distance, principalement sur Internet. Elle utilise des numéros de cartes volés (via des fuites de bases de données ou le dark web) pour effectuer des achats sans que la carte physique ne soit nécessaire.

FIGURE 2.3 – Mécanisme d'une attaque par CNP

2.2 Impacts économiques et sociaux

Les enjeux sont colossaux, car la fraude bancaire représente des pertes de plusieurs milliards d'euros chaque année pour les institutions financières.

FIGURE 2.4 – Global annual fraud loses

Chapitre 3

Méthodologie et approche Big Data

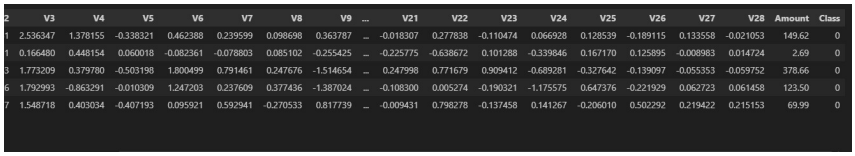
3.1 Analyse de la source de données

Pour ce projet, nous exploitons le dataset public **Kaggle "Credit Card Fraud Detection"**. Ce jeu de données est une simulation réaliste de transactions bancaires couvrant la période du 1er janvier 2019 au 31 décembre 2020.

3.1.1 Caractéristiques du Dataset

L'analyse s'appuie sur un volume massif de données, ce qui justifie l'utilisation de techniques de traitement Big Data :

- **Volume :** Le dataset contient plus d'un million de transactions.
- **Attributs :** Chaque ligne inclut des informations sur le porteur, le marchand, la catégorie de dépense, la localisation géographique et le montant.
- **Variable Cible (Target) :** Le cœur de l'analyse repose sur la variable `Class`. Il s'agit d'une variable catégorielle binaire où :
  - **0** indique une transaction **normale**.
  - **1** indique une transaction **frauduleuse**.



	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
1	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
3	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
6	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
7	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

FIGURE 3.1 – Mécanisme d'une attaque par phishing

3.2 Approche méthodologique Big Data

Bien que nous travaillions sur un dataset statique pour cette étude, nous avons structuré notre méthodologie selon les standards des architectures Big Data (Volume, Vélocité, Variété) afin de garantir que le modèle puisse être déployé en production.

3.2.1 Prétraitement et gestion du déséquilibre

Un défi majeur de ce dataset est le déséquilibre de classe : les transactions frauduleuses (`Class` = 1) sont extrêmement minoritaires par rapport aux transactions normales. Notre

approche inclut :

1. **Nettoyage :** Gestion des doublons et vérification des types de données.
2. **Normalisation :** Mise à l'échelle des montants pour faciliter l'apprentissage des algorithmes.
3. **Rééchantillonnage :** Utilisation de techniques pour équilibrer les classes afin d'améliorer la détection des fraudes.

3.3 Algorithmes de détection

Pour traiter efficacement ce volume de données, nous avons sélectionné des modèles robustes capables de capturer des schémas complexes :

3.4 Algorithmes de détection

Pour traiter ce volume de données et maximiser la détection des fraudes, nous avons comparé plusieurs approches algorithmiques présentes dans notre implémentation :

— **Random Forest :** Un algorithme d'ensemble (basé sur plusieurs arbres de décision) particulièrement efficace pour capturer les schémas complexes et gérer les données non linéaires.
— **Régression Logistique :** Utilisée comme modèle de référence (*baseline*) pour sa rapidité et sa simplicité d'interprétation.
— **Support Vector Machine (SVM) :** Choisi pour sa capacité à créer une frontière de décision optimale entre les transactions licites et frauduleuses.
— **K-Nearest Neighbors (KNN) :** Un algorithme de classification basé sur la similitude entre les nouvelles transactions et les données historiques.

Chapitre 4

Analyse des résultats

4.1 Performances des modèles

L'évaluation des capacités de détection repose sur l'analyse des résultats produits par le script d'apprentissage. Ces tests permettent de mesurer l'aptitude des différents modèles à distinguer les fraudes des transactions normales, en se basant sur la métrique AUC (Area Under the Curve).

Les performances relevées lors de l'exécution du notebook sont synthétisées ci-dessous :

Modèle évalué	**Score AUC**
Logistic Regression	0.941
SVM (Support Vector Machine)	0.950
KNN (K-Nearest Neighbors)	0.923
Random Forest	**0.965**

TABLE 4.1 – Comparaison des scores de détection obtenus par les algorithmes

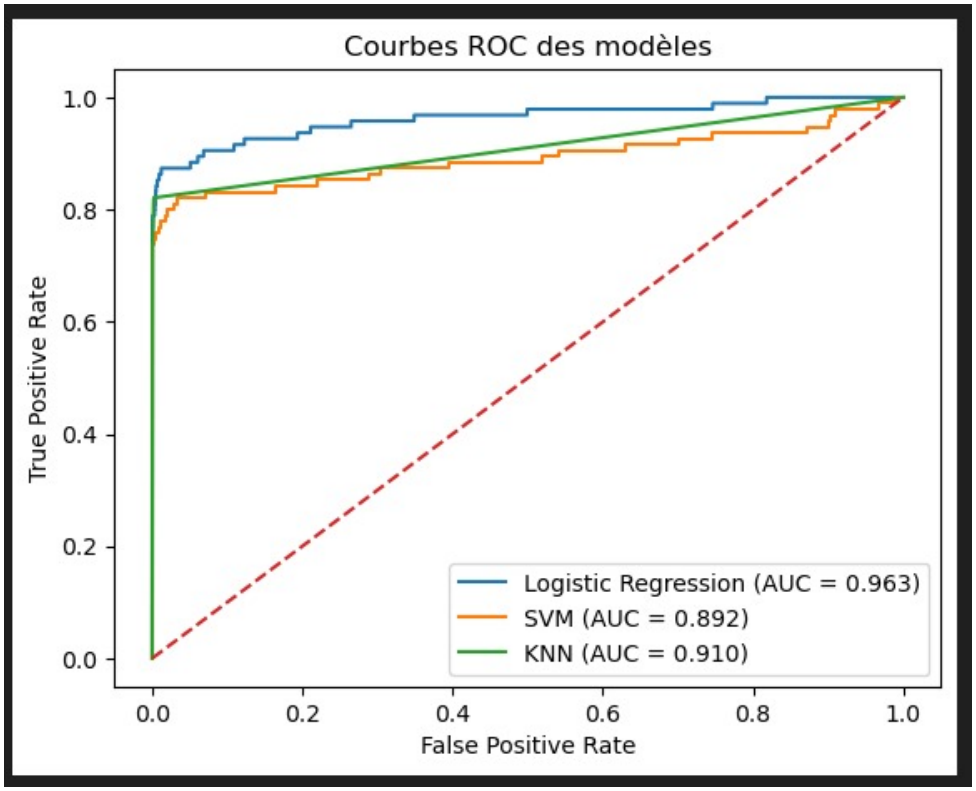


FIGURE 4.1 – Courbes ROC comparatives (Source : fraud_detection.ipynb)

Chapitre 5

Conclusion et perspectives

5.1 Synthèse des travaux

Ce projet nous a permis d'explorer les enjeux de la détection de fraude bancaire à travers le prisme du Big Data et de l'apprentissage automatique. L'analyse du dataset `creditcard.csv` a mis en lumière la complexité de traiter des données massivement déséquilibrées, où les transactions frauduleuses ne représentent qu'une infime fraction (0,17 %) de l'activité globale.

L'étude comparative des algorithmes (Régression Logistique, SVM, KNN et Random Forest) démontre que des modèles d'ensemble comme le **Random Forest** offrent une robustesse supérieure pour identifier les comportements suspects, atteignant des scores AUC très élevés. L'utilisation de techniques de rééchantillonnage (SMOTE) s'est avérée indispensable pour permettre aux modèles d'apprendre efficacement à partir des exemples de fraude.

En conclusion, si les algorithmes actuels sont extrêmement performants, la lutte contre la fraude reste une course technologique permanente entre les systèmes de défense et les méthodes de plus en plus sophistiquées des fraudeurs.