Credit Card Fraud Detection and Prevention

Predictive Analytics and Visualization for Proactive Banking Fraud Detection

Hackathon #1: PSTB Gen AI Bootcamp 2025-Team "Mirage"

0-Dataset Identification and Merging

```
Début
  ↓
1. Identifier la clé de jointure
(ex : `user`, `amount`, `merchant`, `is_fraud`,
etc.)
  ↓
2. Vérifier la qualité de la clé
- Doublons ?
- Types compatibles ?
- Clés correspondantes ?
  ↓
3. Fusionner les datasets
`pd.merge(df1, df2, how='left', on='clé')`
  ↓
4. Nettoyer le résultat
- Gérer les `NaN`
- Supprimer doublons
- Vérifier dimensions
  ↓
5. Analyse exploratoire des nouvelles colonnes
- Statistiques
- Visualisations
- Corrélations
  ↓
Fin – Dataset enrichi prêt pour la modélisation
```

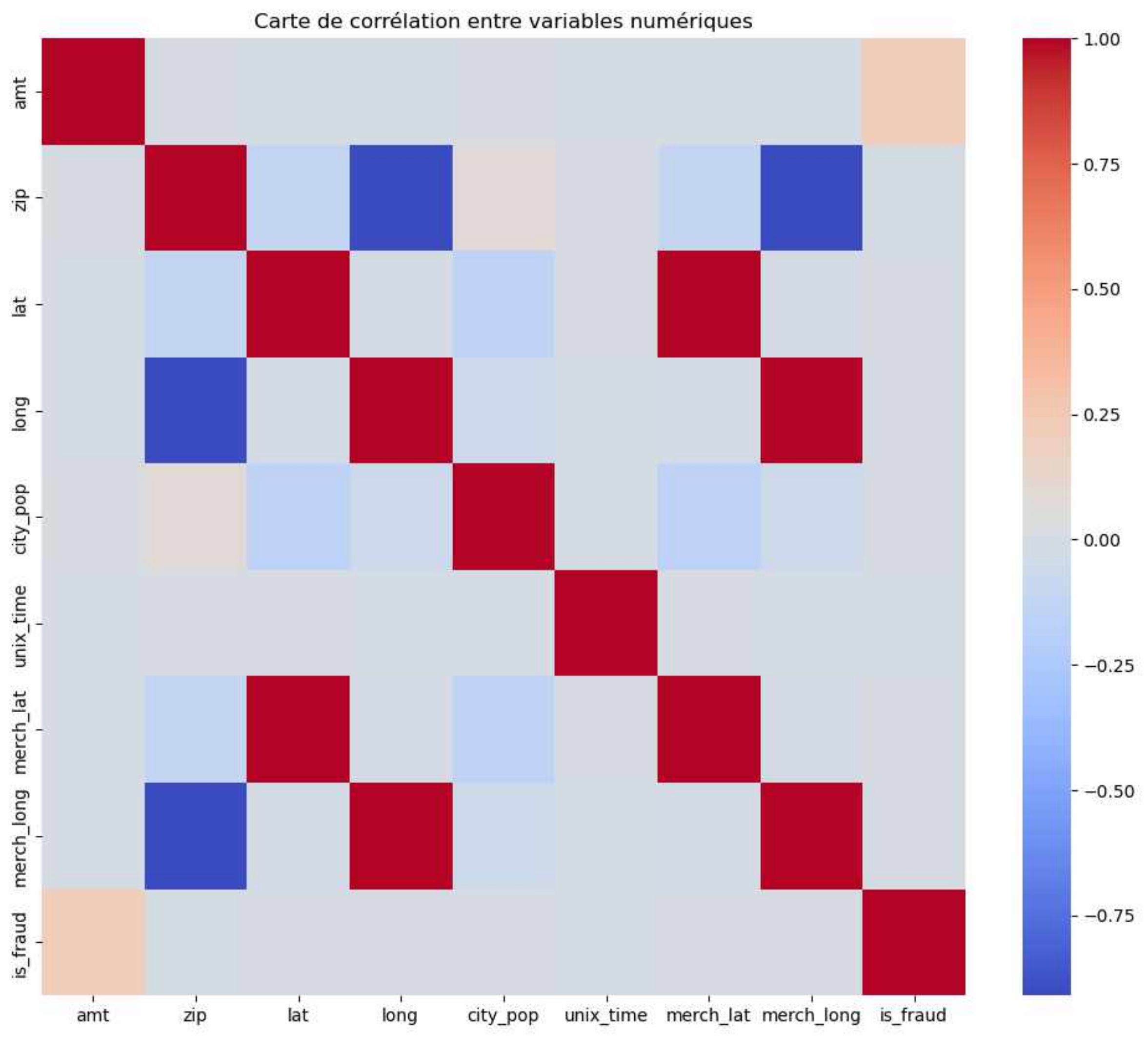
Dataset fusionné
df_merged contient
10 lignes et 58 colonnes, aucune valeur manquante, et toutes les colonnes attendues sont présentes.

1-Dataset Final

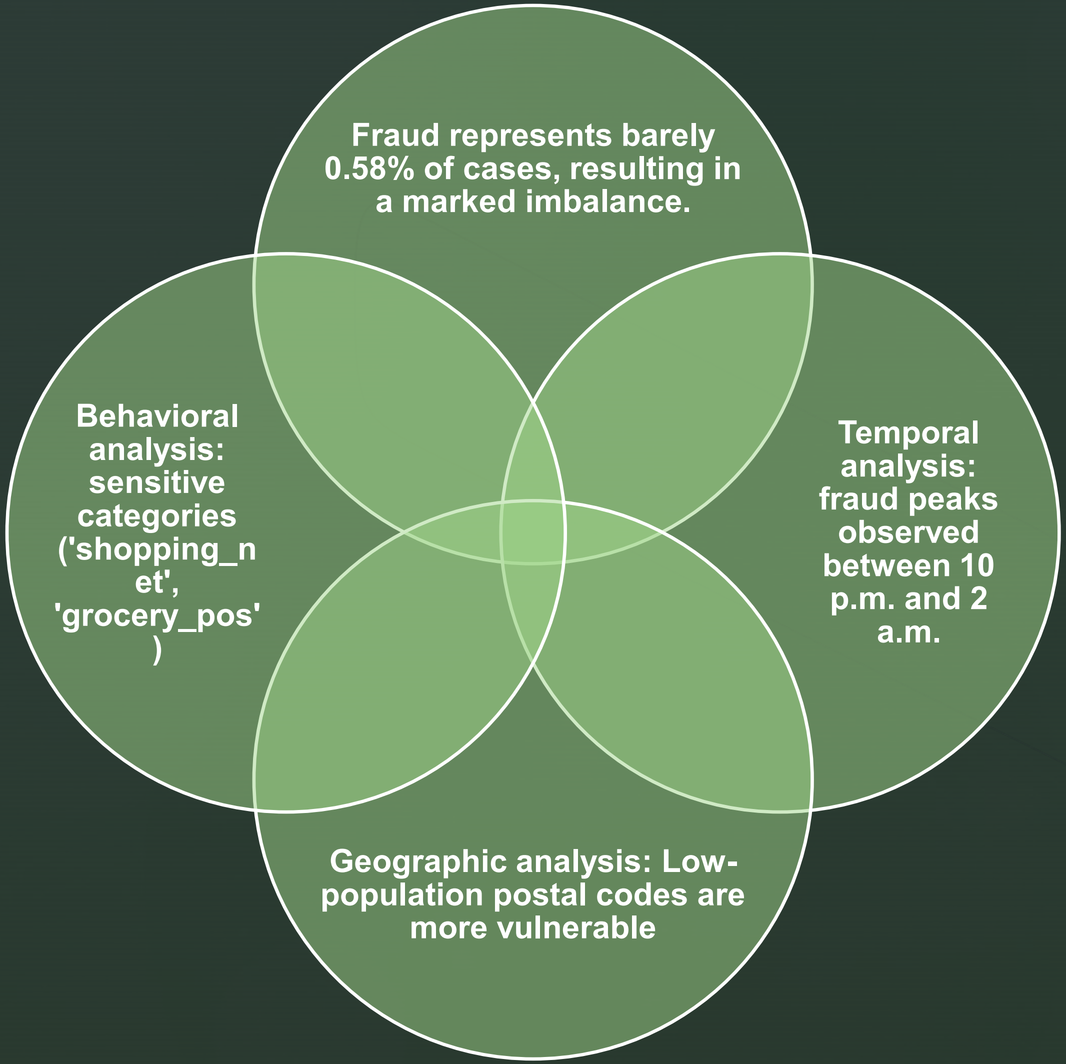
Background: Simulated dataset covering two years of banking transactions (2019-2020)

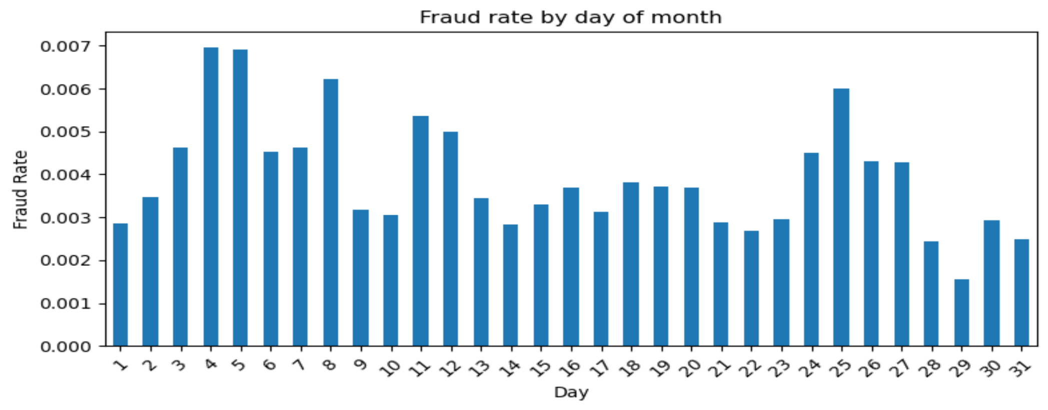
Volume: More than 1.2 million transactions, involving 1,000 customers and 800 merchants

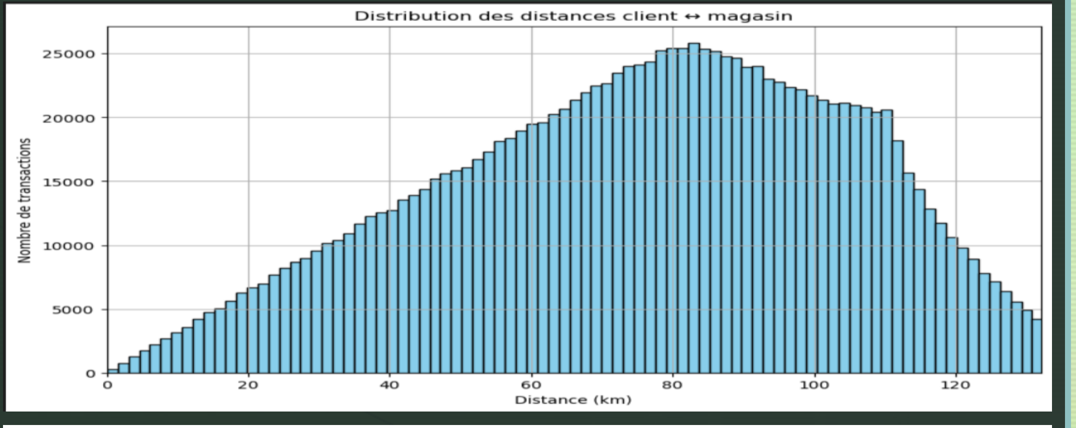
Challenge: Effectively identify rare frauds while minimizing false positives

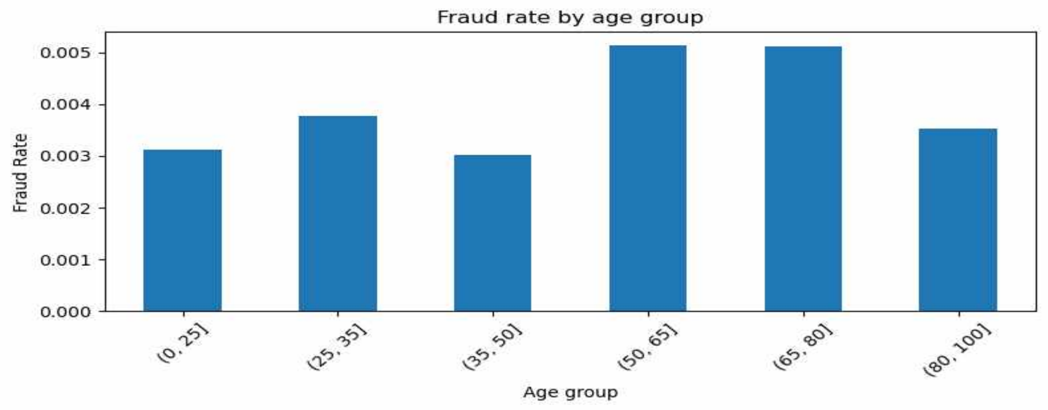


2- Exploratory Analysis (EDA)







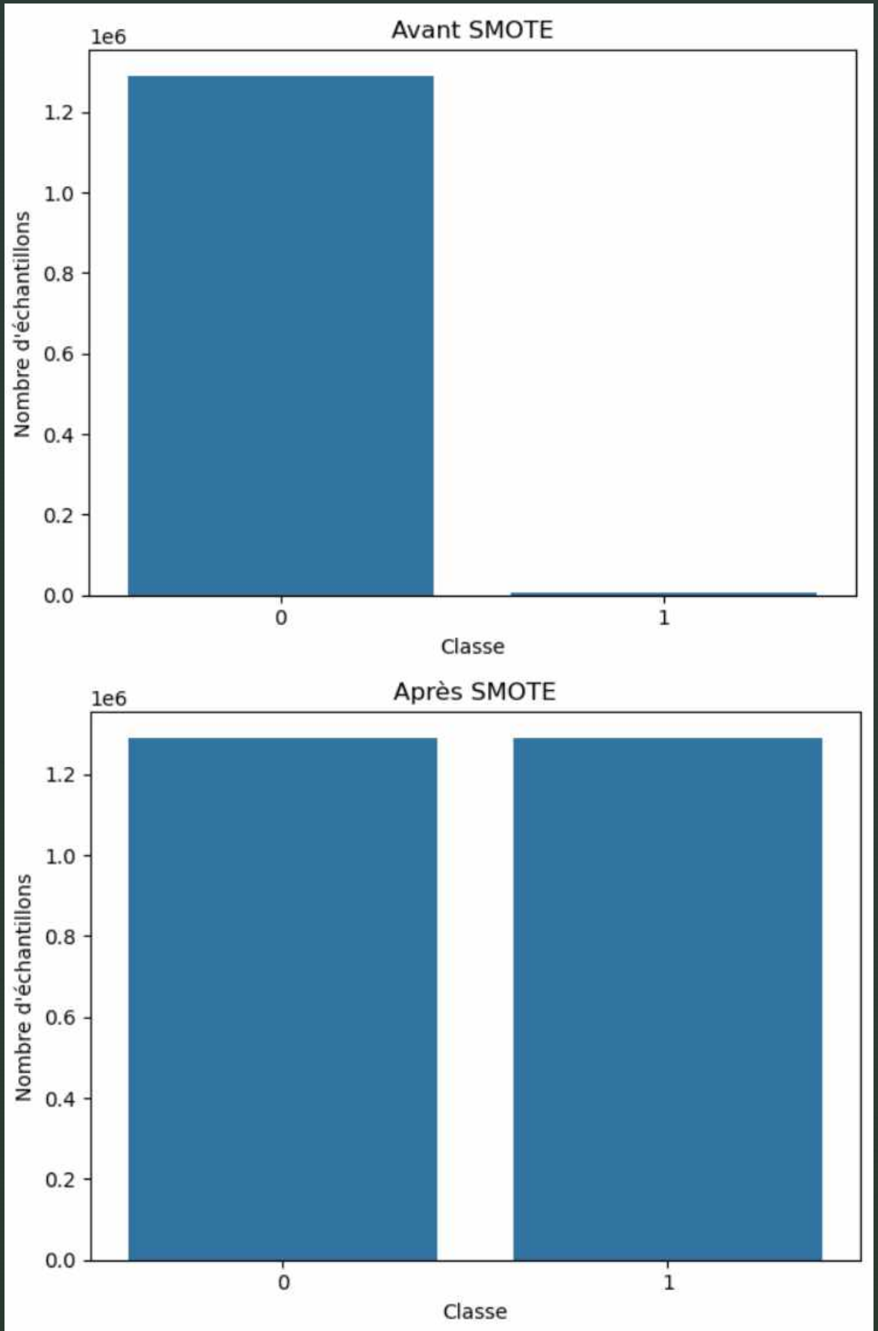


3- Pretreatment & Balancing

No missing values in the dataset

Standardization of transaction amounts to harmonize data

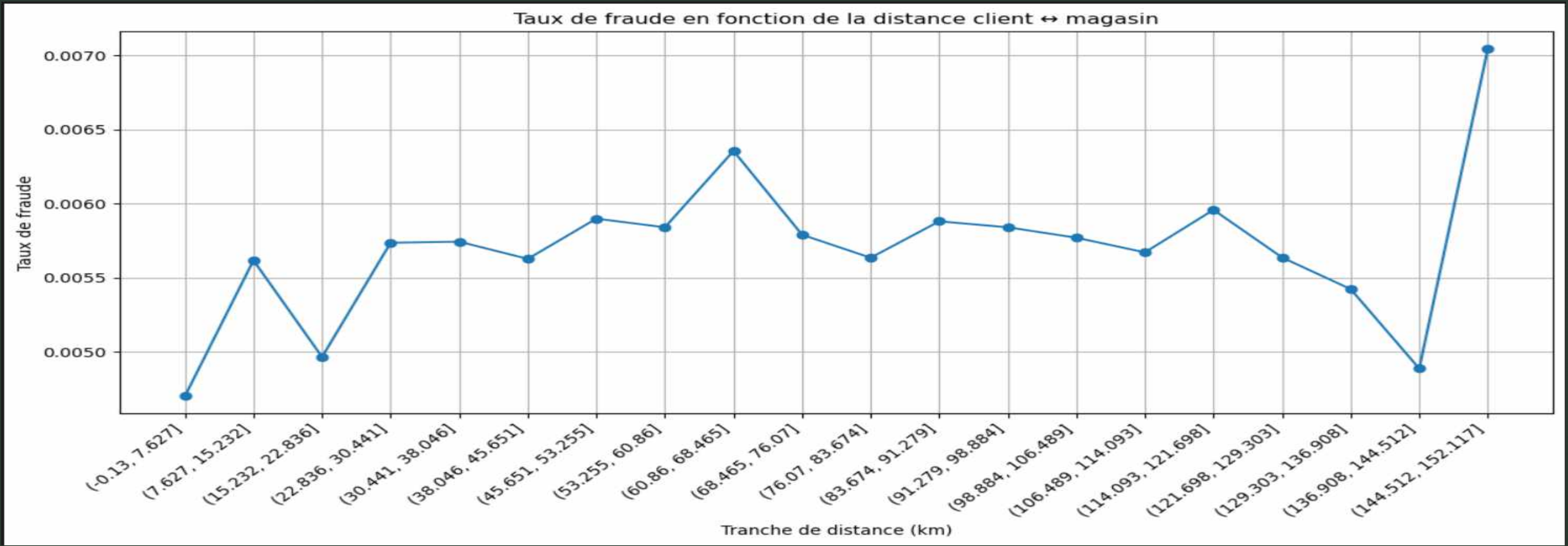
SMOTE applied to balance the training game (cheating/non-cheating)



4- Feature Engineering

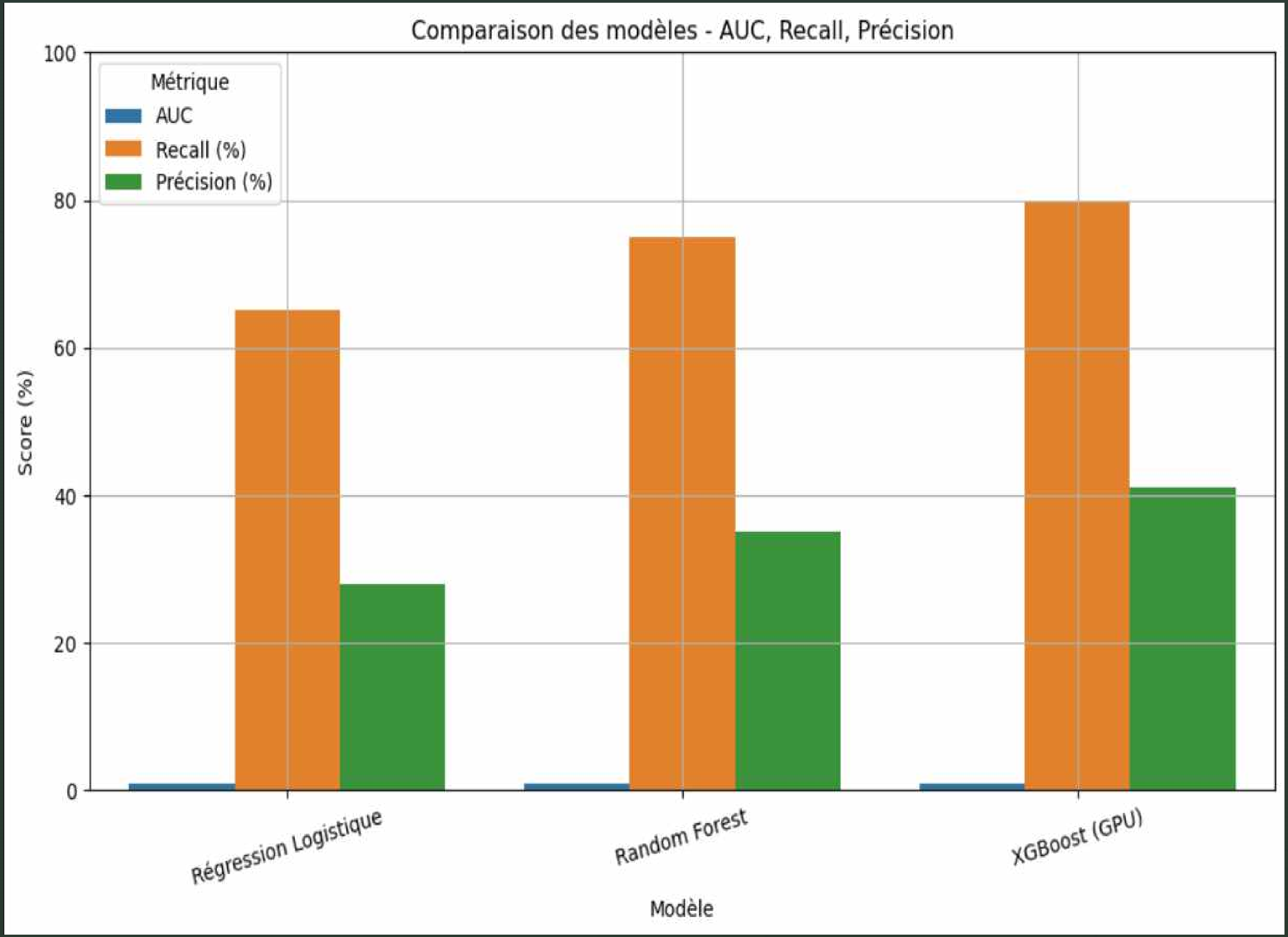
Creating key variables:

- Customer ↔ merchant distance (GPS coordinates)
- Temporal spikes (<60s)
- Hours/days, age, occupation, city population
- Binning (age, population) and categorization of amounts (round, atypical)



5- Modeling & Evaluation

- Models tested:
  - Logistic Regression: Baseline
  - Random Forest: good recall but slower
  - XGBoost (GPU): Best overall results
- Actual (test) results for XGBoost :
  - AUC = 0.98, Recall = 80%, Precision = 41%

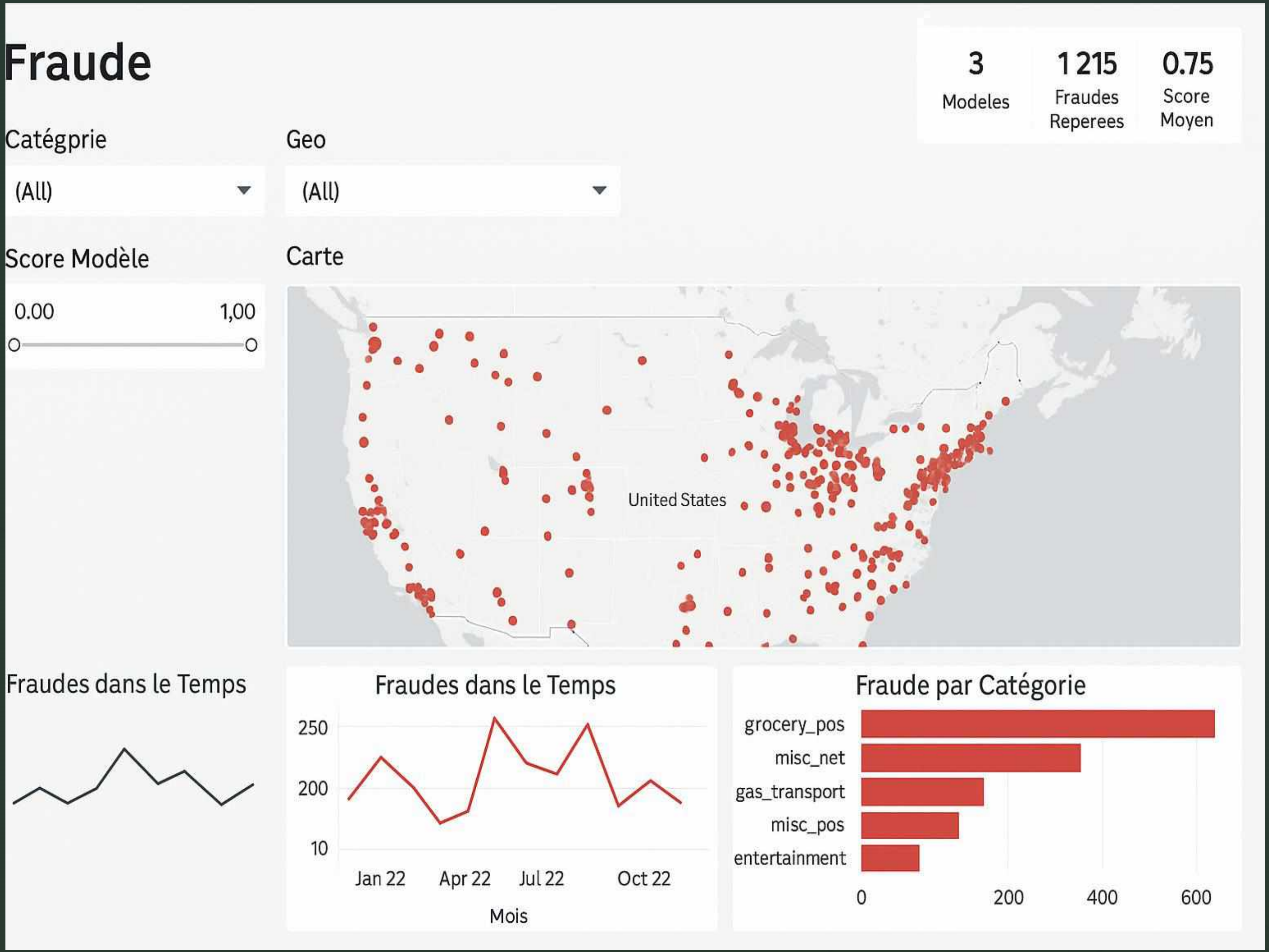


6- Visualization & Dashboard

Visualizations: Matplotlib, Seaborn ,Folium for mapping

PowerBI: interactive dashboard with filters (category, geo, model score)

Business view to facilitate monitoring



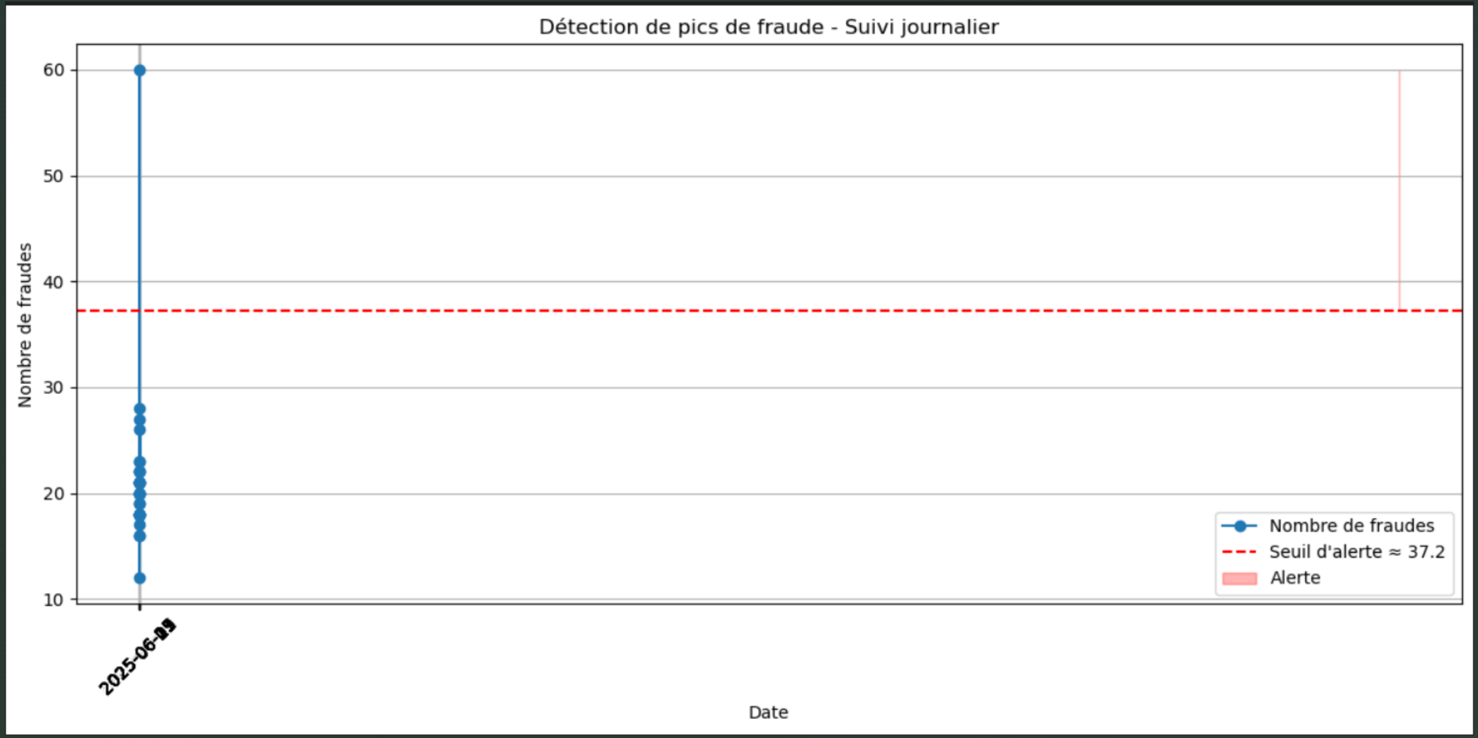
7- Prevention

Proposed strategies:

- Dynamic thresholds on time, distance, merchants
- Alert in case of spikes or sudden changes in behavior

Additional data suggested:

- Fingerprinting, merchant blacklist, network analysis



Data type	Examples	Main utility
Fingerprinting Device	Device ID, OS type, resolution, browser, session ID	Identify devices shared between accounts
User behavior	Session time, clicks, typing speed, mouse movements	Detect unusual or suspicious behavior
Blacklist	Risky merchants, banned IPs, fraudulent emails	Block or alert upon detection of a known element
Graph analysis	Shared cards/IDs, common IPs, similar addresses	Identify organized groups or related frauds
Geolocation	Customer distance ↔ transaction, country/time inconsistencies	Alert in case of impossible movements
User history	Purchase frequency, usual times, usual amounts	Detect behavioral disruptions

8-Top 5 protections through knowledge

1. **Know how to recognize a fraudulent site**
   - Check the URL, the presence of HTTPS, and avoid poorly translated or overly aggressive sites.

2. **Know phishing techniques**
   - Never click on a suspicious email link (banks, fake fines, etc.).

3. **Understand how two-factor authentication (2FA) works**
   - Know how to activate it and why it blocks most fraud.

4. **Be aware of weak signs of fraud**
   - Repeated very low amounts, unusual location, activity outside of hours.

5. **Know your rights and reflexes in the event of fraud**
   - Know that you can dispute a transaction and that the bank has obligations.

9-Behavioral Guide

❌ Guide du Mauvais Payeur (à éviter)

Comportement à éviter	Pourquoi c'est suspect	Risque détecté par l'algorithme
🔄 Transactions répétées en quelques secondes	Simule des tentatives automatisées	Spike comportemental, fraude "burst"
🌙 Paiements entre 2h et 5h du matin	Inhabituel pour la majorité des clients	Horaire à forte suspicion
🌍 Paiement soudain dans un autre pays ou région	Rupture brutale du schéma géographique	Anomalie de localisation
💰 Montants très ronds ou trop faibles	Techniques typiques de tests de carte	Détection de "test" ou de fragmentation
👥 Changement brutal de canal (ex: POS → NET)	Le modèle détecte un shift comportemental	Suspect : piratage ou usage non autorisé
👫 Paiements fréquents pour d'autres personnes	Usage tiers non déclaré	Soupçon de carte prêtée ou volée
🏪 Usage d'un marchand inconnu ou douteux	Nouveau marchand mal noté ou hors de vos habitudes	Score de risque marchand élevé
📍 Ville / ZIP code très rare pour vous	Peut être une usurpation ou une tentative	ZIP inhabituel, suspicion géolocalisée
👶 Profil jeune avec gros achat électronique	Décalage entre profil attendu et type d'achat	Détection de fraude générationnelle
📊 Usage d'un job/genre/âge surreprésenté dans la fraude	Le modèle peut associer une probabilité plus forte	Corrélation indirecte via biais appris

THANKS !

Questions

Useful links (GitHub, Notebook, Template)