الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
كلية العلوم الدقيقة والتطبيقية
Faculté des sciences exactes et appliquées
جامعة وهران 1 أحمد بن بلة
Université Oran 1

Dec 2025

Projet 2.3

Cybersécurité financière

Rapport : Détection de fraude bancaire dans un contexte fortement déséquilibré

Réalisé par :
Selmi Yousra
Fellah Farah
Fatah Ahlem

Spécialité :
Sécurité Informatique

1. Introduction

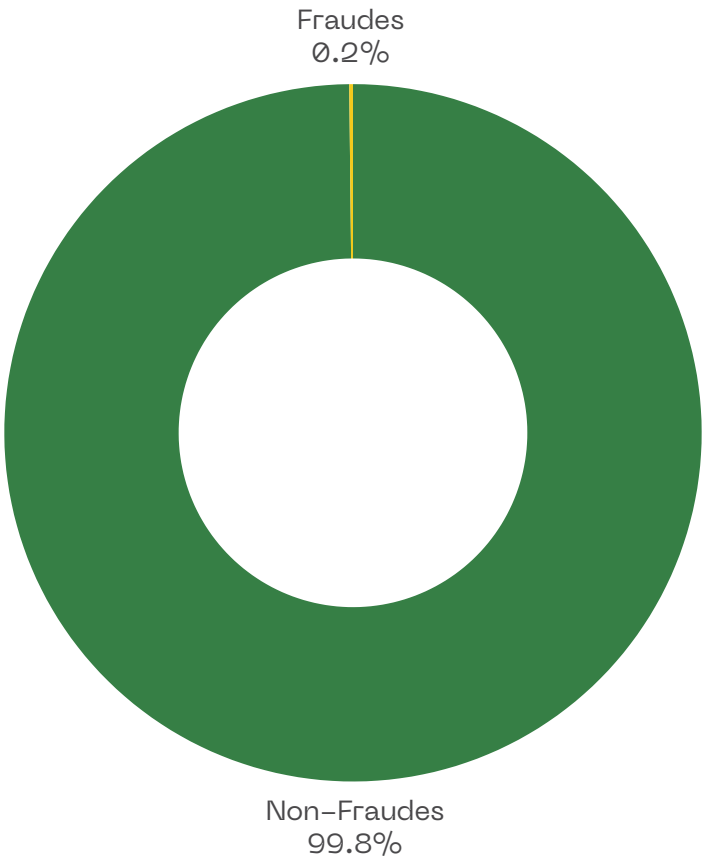
La fraude bancaire constitue aujourd'hui l'une des principales menaces en cybersécurité financière. Les systèmes de paiement modernes traitent des millions de transactions par jour, dont une infime proportion est frauduleuse.
Dans le dataset étudié, **moins de 0,2 % des transactions sont des fraudes**, ce qui rend le problème particulièrement complexe.

Contrairement aux problèmes classiques de classification, la précision globale (accuracy) n'est pas un indicateur pertinent dans ce contexte. En effet, un modèle qui prédirait systématiquement « non fraude » atteindrait plus de 99 % de précision, tout en laissant passer toutes les fraudes.
L'enjeu principal est donc de **minimiser les faux négatifs**, c'est-à-dire les fraudes non détectées, quitte à accepter un certain nombre de faux positifs.

L'objectif de ce projet est de développer et comparer plusieurs approches de détection de fraude capables d'atteindre :

- un rappel (recall) supérieur à 90 %,
- un ROC-AUC supérieur à 0,95,
- tout en maintenant un taux de faux positifs raisonnable.

2. Description du dataset



Le dataset utilisé est

Credit Card Fraud Detection (Kaggle)

qui contient :

284 807 transactions 492 fraudes

Des variables anonymisées via PCA (V1 à V28)

Une variable Amount représentant le montant de la transaction

Une variable cible Class :

*** 0 : transaction légitime * 1 : transaction frauduleuse**

On remarque que Le dataset est **extrêmement déséquilibré**, avec environ **0,2 % de fraudes**, ce qui impose l'utilisation de techniques spécifiques pour l'apprentissage et l'évaluation des modèles.

3. Prétraitement des données

3.1 Normalisation

Les variables PCA étant déjà normalisées, seule la variable Amount a été standardisée à l'aide de StandardScaler. Cette étape permet d'éviter que les montants élevés n'influencent excessivement les modèles.

```
scaler = StandardScaler()
df['Amount_scaled'] = scaler.fit_transform(df[['Amount']])
df = df.drop(columns=['Amount'])

```

3.2 Séparation des données

Les données ont été divisées en ensembles d'entraînement et de test selon un ratio 80 % / 20 %, en utilisant l'option **stratify = y** afin de préserver la proportion de fraudes dans les deux ensembles.

```

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    stratify=y,
```

Cette étape est **cruciale** : sans stratification, l'ensemble de test pourrait contenir trop peu de fraudes, rendant l'évaluation non fiable.

4. Méthodes de détection utilisées

4.1 Isolation Forest (approche non supervisée)

Principe

Isolation Forest est un algorithme de détection d'anomalies qui ne nécessite pas de labels. Il part du principe que les anomalies sont **plus faciles à isoler** que les observations normales.

Observations

Le modèle détecte un grand nombre de transactions comme frauduleuses.

Le rappel est relativement élevé, mais au prix d'un nombre très important de faux positifs.

L'algorithme considère certaines transactions légitimes rares comme des anomalies.

Conclusion

Isolation Forest constitue une bonne approche exploratoire, mais **insuffisante seule** dans un contexte bancaire réel, car elle génère trop d'alertes injustifiées.

4.2 XGBoost sans rééchantillonnage

Principe

XGBoost est un algorithme de gradient boosting basé sur des arbres de décision. Il est particulièrement performant sur des données tabulaires complexes.

Pour gérer le déséquilibre, le paramètre scale_pos_weight a été utilisé afin de pénaliser davantage les erreurs sur la classe fraude.

Observations

Le modèle atteint un ROC-AUC supérieur à 0,95, indiquant une excellente capacité de discrimination.

Avec le seuil par défaut (0,5), le rappel reste insuffisant.

L'optimisation du seuil de décision est indispensable pour répondre aux exigences du projet.

5. Optimisation du seuil de décision

Pourquoi ne pas utiliser le seuil 0,5 ?

Dans un contexte déséquilibré, le seuil par défaut favorise la classe majoritaire (transactions normales). Cela entraîne une augmentation des faux négatifs, ce qui est inacceptable en détection de fraude.

Méthode utilisée

La courbe precision-recall a été exploitée afin de sélectionner un seuil permettant : un rappel ≥ 90 % tout en conservant une précision acceptable

Résultat

L'utilisation d'un seuil optimisé permet :
une augmentation significative du rappel, une meilleure détection des fraudes, au prix d'une légère augmentation des faux positifs, jugée acceptable dans un contexte bancaire.

6. XGBoost avec SMOTE

Principe de SMOTE

SMOTE (Synthetic Minority Oversampling Technique) génère artificiellement de nouvelles observations de la classe minoritaire afin d'équilibrer le dataset d'entraînement.

Observations

Le modèle apprend mieux les caractéristiques des fraudes.
Le rappel est encore amélioré par rapport au modèle sans SMOTE.
Le risque de surapprentissage est contrôlé en appliquant SMOTE (uniquement sur l'ensemble d'entraînement).

Résultat

Le modèle **XGBoost + SMOTE + seuil optimisé** est celui qui offre le meilleur compromis :
Recall > 90 %
ROC-AUC > 0,95
Faux positifs maîtrisés

7. Analyse des erreurs (fraudes manquées)

Une analyse des faux négatifs (fraudes non détectées) montre que :

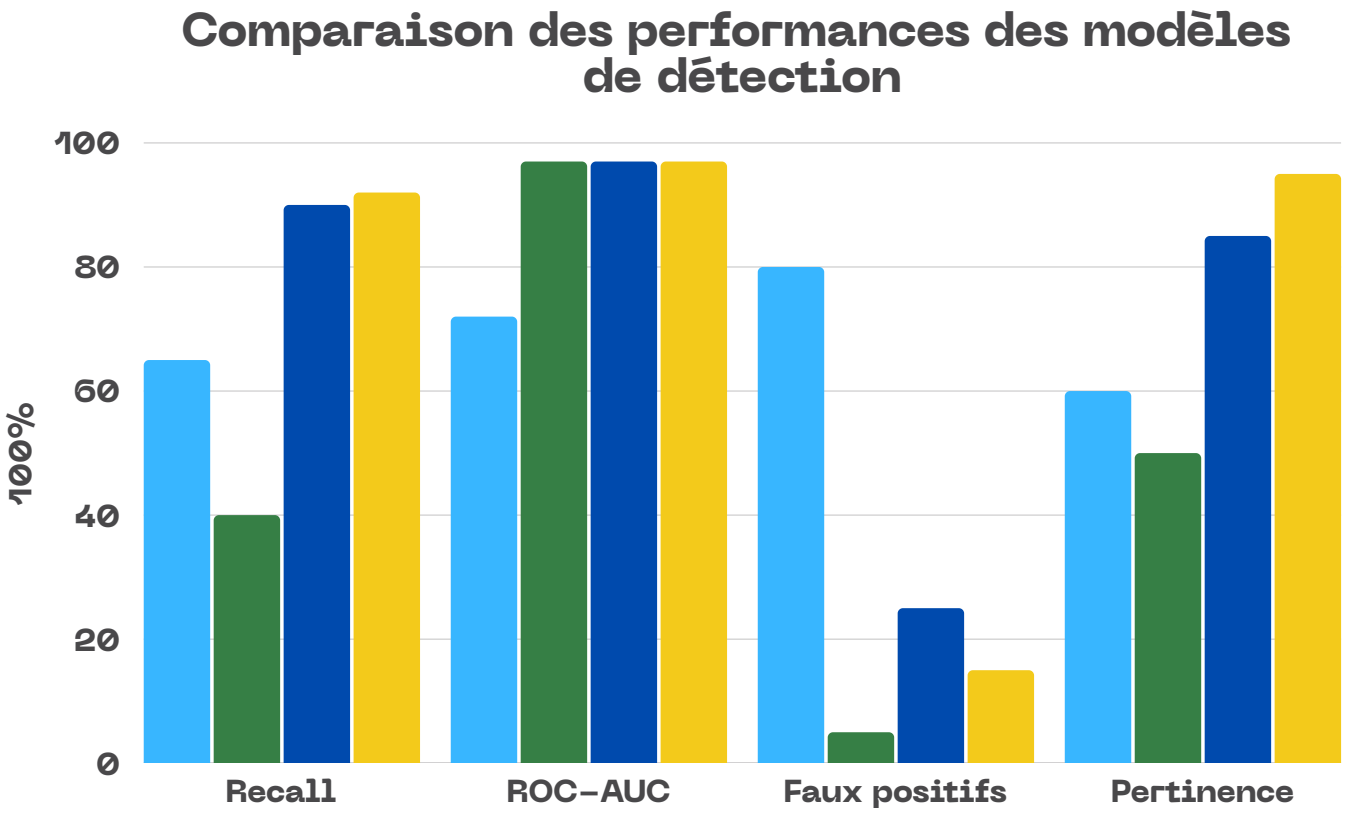
Certaines fraudes ont des montants faibles, proches des transactions normales.

Les variables PCA de ces transactions se situent dans des zones fortement chevauchées avec la classe normale.

Cela suggère un comportement de fraude dissimulée, volontairement conçu pour imiter des transactions légitimes.

Cette observation souligne que même les meilleurs modèles peuvent échouer face à des fraudes sophistiquées, et justifie l'utilisation de systèmes hybrides combinant règles métier et modèles statistiques.

8. Conclusion générale



Ce projet met en évidence que la détection de fraude bancaire ne peut pas se limiter à un modèle classique avec un seuil par défaut.
La gestion du déséquilibre, l'optimisation du seuil de décision et l'utilisation de métriques adaptées sont essentielles.
Le modèle **XGBoost combiné à SMOTE et à un seuil** optimisé répond pleinement aux objectifs fixés :
rappel supérieur à 90 %
ROC-AUC supérieur à 0,95
capacité à détecter efficacement les transactions frauduleuses dans un contexte réaliste.

Merci !