Compte rendu : Détection de Fraude aux Cartes de Crédit

JAMAL YASSINE
Numéro Étudiant : 22007655

Décembre 2025

1 Analyse Complète et Contexte du Projet

Contexte

Dans un contexte de digitalisation accélérée des paiements, la **fraude aux cartes de crédit** représente un enjeu économique majeur pour les institutions financières. Ce projet de Data Science, réalisé dans le cadre d'un module de Machine Learning (CAC2), vise à développer un système prédictif capable d'identifier les transactions frauduleuses en temps réel.

- **Jeu de données** : Kaggle "Credit Card Fraud Detection".
- **Volume** : 284 807 transactions européennes sur 2 jours.
- **Taux de fraude** : 0,172 % (492 fraudes pour 284 315 transactions normales), révélant un déséquilibre extrême.

Problématique et Objectifs Prioritaires

La classification binaire supervisée a pour but de prédire la variable cible '**Class**' (0=légitime, 1=fraude). L'enjeu principal est de minimiser les **faux négatifs** (fraudes non détectées).

- **Métriques prioritaires** : En raison du déséquilibre massif, l'accuracy est insuffisante. Les indicateurs clés sont le **Recall** (pour éviter les Faux Négatifs), la **Précision**, le **F1-score**, et le **ROC-AUC** (le plus adapté aux déséquilibres extrêmes).

2 À propos du Jeu de Données

Le dataset comporte **31 variables** :

- **28 variables anonymisées** ($V1$ à $V28$), résultant d'une transformation par **Analyse en Composantes Principales (PCA)**.
- **2 variables originales** : `Time` (temps écoulé depuis la première transaction) et `Amount` (montant de la transaction).
- **1 variable cible** : `Class`.

3 Analyse Exploratoire des Données (EDA)

L'EDA a confirmé :

- **Déséquilibre massif** : La classe fraude est extrêmement minoritaire.
- **Séparabilité** : Les variables PCA sont déjà centrées-réduites. Certaines d'entre elles (`V14`, `V17`) montrent des distributions distinctes entre les deux classes.
- **Montant** : La distribution de `Amount` est très asymétrique, ce qui justifie une transformation logarithmique.
- **Corrélations** : Très peu de relations linéaires fortes sont observées du fait de la PCA.

4 Préparation et Ingénierie des Données

La phase de preprocessing inclut :

- **Nettoyage** : Suppression des doublons.
- **Feature Engineering** : Création de `Amount_Scaled` (via `RobustScaler`) et `Log_Amount` (transformation logarithmique).
- **Standardisation** : Retrait des colonnes `Time` et `Amount` brutes.
- **Découpage** : Split **80 % / 20 %** avec **stratification sur Class** pour garantir la même proportion de fraudes dans les échantillons d'entraînement et de test.

5 Méthodologie de Modélisation

Trois algorithmes supervisés ont été étudiés : la **Régression Logistique**, le **Random Forest** et le **XGBoost**.

Gestion du Déséquilibre (SMOTE)

La technique **SMOTE** (Synthetic Minority Over-sampling Technique) est utilisée pour sur-échantillonner la classe minoritaire, mais **uniquement sur les données d'entraînement** afin d'éviter la fuite d'information (Data Leakage).

Validation Croisée et Optimisation

Chaque modèle est intégré dans un pipeline standardisation-oversampling-classification. La recherche d'hyperparamètres est réalisée via **GridSearchCV**, en utilisant le **ROC-AUC** comme scoring principal, car il est le plus fiable pour ce type de déséquilibre.

6 Résultats, Limites et Recommandations

Résultats Observés

- L'accuracy n'est pas une métrique fiable.
- Le **ROC-AUC** est plus représentatif des performances globales.
- Les modèles avancés (**Random Forest** et **XGBoost**) montrent un fort potentiel pour obtenir de bons **Recall**, **Précision** et **F1-Score** sur la classe de fraude.

Limites Rencontrées

- Nécessité d'utiliser explicitement `scoring="roc_auc"` dans `GridSearchCV`.
- Le faible nombre de fraudes induit une variabilité potentiellement élevée des performances.
- Le caractère anonymisé des variables PCA limite l'interprétation métier directe.

Recommandations

- Finaliser l'optimisation des hyperparamètres (Tuning).
- Explorer d'autres algorithmes spécialisés (e.g., Isolation Forest, modèles neuronaux).
- Générer un tableau comparatif complet des métriques clés (AUC, Recall, F1).
- Mettre en place une stratégie de déploiement en production avec un seuil de décision ajustable.

7 Conclusion

Ce projet a permis de mettre en place une chaîne d'analyse robuste (nettoyage, ingénierie de variables, rééquilibrage par SMOTE, validation croisée) pour la détection de fraude. La prochaine étape sera de finaliser le tuning et de choisir la solution offrant le meilleur compromis entre une détection efficace (maximisation du Recall) et la réduction des fausses alertes (maintien d'une bonne Précision). Ce travail constitue une avancée significative vers un système fiable de détection de fraude bancaire.