



MACHINE LEARNING

1. Contexte

Vous êtes l'équipe "Data & Sécurité" de la startup *PoketraFinday*.

La Startup

PoketraFinday est une fintech malgache innovante qui vise à démocratiser l'accès aux services financiers. En transformant chaque téléphone mobile en un portefeuille électronique intelligent, elle permet aux populations non-bancarisées d'accéder au micro-crédit instantané et aux paiements digitaux.

La Problématique de Sécurité (Mission Critique)

Le succès de *PoketraFinday* repose entièrement sur la confiance. Or, cette confiance est aujourd'hui fragilisée. Une recrudescence de fraudes plus ou moins sophistiquées (vols de comptes nocturnes, ingénierie sociale ciblant les seniors, ...) ralentit le développement du système.

Votre mission est de créer un modèle qui peut prédire la cible **is_fraud**. Cela aidera à sauver la réputation de la plateforme en bloquant les attaquants sans pénaliser les utilisateurs honnêtes.

Modalités Logistiques

- Durée :** 8 Heures (08h00 - 16h00).
- Lieu :** Distanciel complet. Pas de déplacement à l'ISPM.
- Deadline :** Le dernier commit sur votre repository doit être effectué à **16h00 précises**.

2. Description du Dataset

Vous disposez d'un jeu de données avec les colonnes suivantes :

Colonne	Description
transaction_id	Identifiant unique de la transaction (UUID).
step	Unité de temps (1 heure). De 1 à 744 (31 jours).
type	PAYMENT, TRANSFER, CASH_OUT, DEBIT.
amount	Montant en Ariary (MGA).
customer_id	Identifiant unique du client émetteur.
age	Âge du client émetteur.
is_fraud	Cible : 0 = Légitime, 1 = Fraude.

Indice crucial : Pour vos analyses de Feature Engineering, supposez que le **Step 1** correspond à la première heure d'un **LUNDI**.

3. Votre Mission

Votre objectif est de maximiser le F1-Score sur la détection de fraude et de faire un rapport sur les fraudes que vous détectez à partir de vos analyses. Le fichier du train set est disponible dans *ressources/train.csv*.

Étape 1 : EDA et préparation initiale des données

Faites une Exploratory Data Analysis orientée vers la cible **is_fraud**. Préparez les données en appliquant les bonnes pratiques (préparations des variables catégorielles, feature engineering basiques, gestion des variables manquantes, suppression des données inutiles).

Étape 2 : Baseline : Régression Logistique

Vous devez commencer par entraîner une Régression Logistique.

- Ce modèle servira de référence (Baseline).
- Vous devez analyser ses résultats avant de passer à plus complexe.

Étape 3 : Exploration & Modélisation

Une fois la baseline établie, vous êtes libres d'explorer d'autres approches selon votre savoir faire :

- Advanced Feature Engineering (Création de variables temporelles, etc.).
- Modèles avancés (Decision Tree, Random Forest, XGBoost, Réseaux de Neurones...).
- Hyperparameter tuning (si besoin est).
- Stratégies de gestion du déséquilibre (SMOTE, etc.).

Étape 4 : Génération de la Soumission

Générer le fichier **submission.csv** en utilisant le test set (sans cible) fourni dans *ressources/test.csv*.

Exemple de code (Python/Pandas) :

```
submission = pd.DataFrame({  
    "transaction_id": test_df["transaction_id"],  
    "is_fraud": model.predict(X_test)  
})  
submission.to_csv("submission.csv", index=False)
```

4. Livrables (Avant 16h00)

Tout doit être disponible sur votre Repository GitHub.

1. **submission.csv** : Les prédictions sur le test set (colonnes transaction_id, is_fraud).
2. **Vidéo de Présentation (3-5 min)** :
 - Remplace la soutenance physique.
 - Présentez votre équipe, votre analyse (EDA), les types de fraudes identifiés, votre Baseline vs Modèle Final.
 - Lien de la vidéo dans le README ou fichier vidéo dans le repo (si < 100Mo).
3. **Notebook & Code** : Le code doit être clairement commenté et structuré logiquement.
4. **README.md** : Utilisez le modèle *ressources/readme-model.md* fourni pour votre rapport. Il inclut les questions sur F1-Score, FP/FN et la décision opérationnelle.

5. Critères d'Évaluation

Critère	Poids
Performance (F1-Score)	30%
Feature Engineering	20%
Présentation & Vidéo	20%
Qualité du Code	15%
Réponses README	15%

Bon courage!