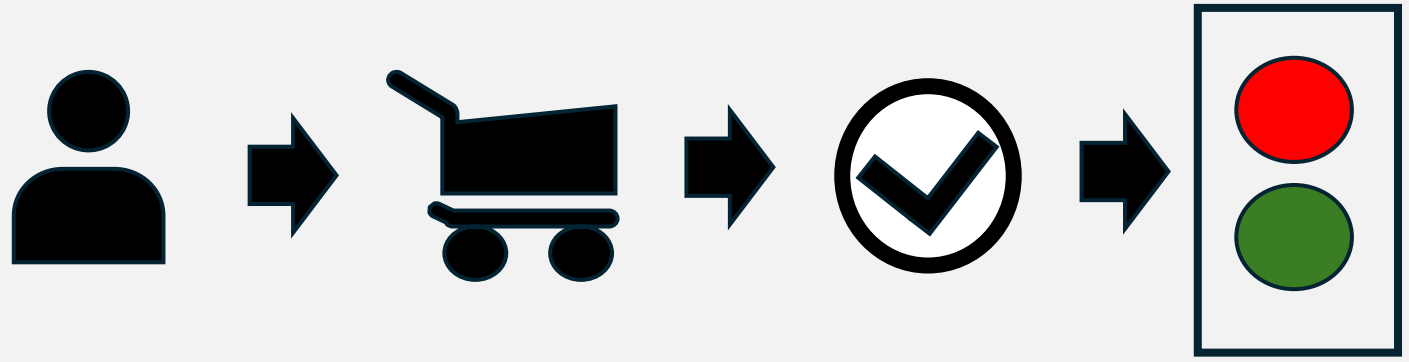
Fraud Data Analyst

Romain RAMAT
08/02/2026

Contexte

- La fraude, un enjeu majeur et stratégique :
  - Perte financière
  - Risque de réputation et d’image

Le process



Données d’entrées du challenge :

4 Fichiers sont disponibles en entrées :

- **X_train :** variables explicatives pour l’entraînement
- **y_train** : variables cibles pour l’entraînement
- **X_test** : variables explicatives pour le test ➔ **Non exploitable**
- **y_test :** variables cibles pour le test ➔ **Non exploitable**

Le jeu « y_test » étant un exemple aléatoire composé de pourcentage et non de cas de fraude avéré, n’a pas été considéré dans le déroulé de l’exercice.

Analyse du jeu d'entrainement

X_train

(92790,146)

Y_train

(92790,2)

Merge

df_train

(92790,147)

Variable	Description	Exemple
ID (num)	ID unique	1
Item1 à 24	Catégorie du bien de l'item de 1 à 24	Computer
Cash_price1 à 24	Prix de l'item de 1 à 24	850
Make 1 à 24	Fabriquant de l'item	Apple
Goods_code1 à 24	Nombre de produits dans l'item 1 à 24	2
Nb of items	Nombre total d'items	7

- 98,57 % (91471) cas non-Fraude
- 1,4% (1319) cas fraude

Déséquilibre des cas de fraudes

Gestion des items dans le fichier
Objectif ➔
Se rapprocher sur une vue Panier

En l'état, les données décrivent un panier sous forme éclatée (item1,2,3,...24)

Mise en place de features focus « Panier » :

Un bon modèle de détection de fraude :
- N'interdit pas
- Ne juge pas

Il observe des écarts

On ne dit pas : Téléphone Iphone 16 = Fraude

Mais :
Le contenu de ce panier ressemble statistiquement aux paniers frauduleux observés

Structure du panier :	Le prix:	Les quantités:	Diversité:
• Nombre réel d'items • Nombre d'items unique • Taux de répétition	• Montant total du panier • Prix max • Prix min • Prix moyen • Dispersion des prix • Ratio prix max/prix total	• Quantité totale • Quantité moyenne • Quantité max	• Nbre catégorie unique • Nbre marque unique • Nbre de modèle unique • Dominance d'une marque

Transformation des items en feature exploitables par un algorithme de Machine Learning

Néanmoins il existe des catégories à risque :

CATEGORIE	TAUX_FRAUDE	NB_TRANSACTION
KITCHEN SCALES & MEASURES	100%	2
KITCHEN UTENSILS & GADGETS	35%	31
STORAGE & ORGANISATION	21%	39
LAUNDRY & CLOTHESCARE	15%	34
LUGGAGE	13%	15
PRESERVING & BAKING EQUIPMENT	11%	9
IMAGING EQUIPMENT	8%	85
MEN S CLOTHES	7%	59
BLANK MEDIA MEDIA STORAGE	6%	17
GAMING	6%	36
AUDIO ACCESSORIES	5%	1670
BARWARE	5%	41
NURSERY ACCESSORIES	5%	111
MEN S FOOTWEAR	4%	26
NURSERY LINEN	3%	29
HI-FI	3%	135
TABLEWARE	3%	681
TELEPHONES, FAX MACHINES & TWO-WAY RADIOS	3%	3138
JEWELLERY WATCHES	3%	36
BABY CHILD TRAVEL	3%	1161
TELEPHONES FAX MACHINES TWO-WAY RADIOS	3%	1513
DECORATIVE ACCESSORIES	3%	592
FRAGRANCE	3%	79
CARPETS, RUGS & FLOORING	2%	86
COMPUTERS	2%	50221
FULFILMENT CHARGE	2%	25023

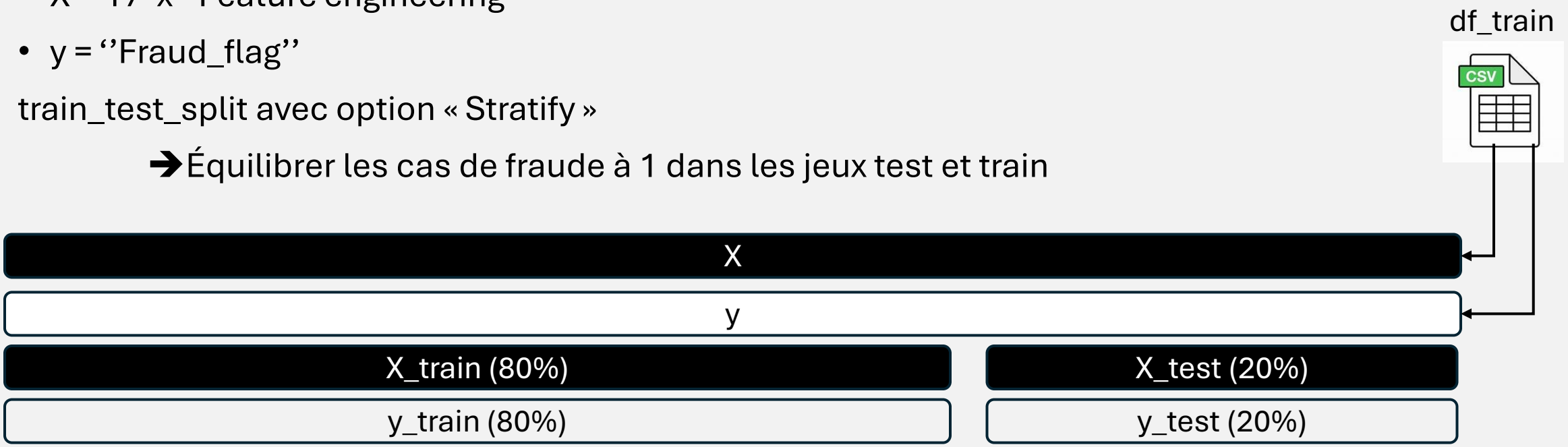
Sont considérés à risque si :
Taux de fraude supérieur à la moyenne globale (1,4%) et + de 100 transactions

Nettoyage des données et mise en place des jeux de test

- X = 17 x ‘’Feature engineering’’
- y = ‘’Fraud_flag’’

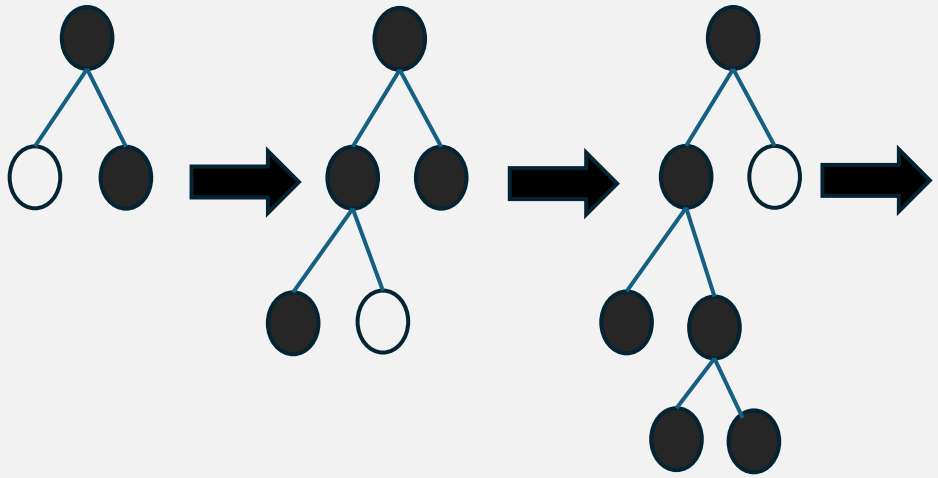
train_test_split avec option « Stratify »

➔Équilibrer les cas de fraude à 1 dans les jeux test et train



Modèle d’apprentissage supervisé : LightGBM

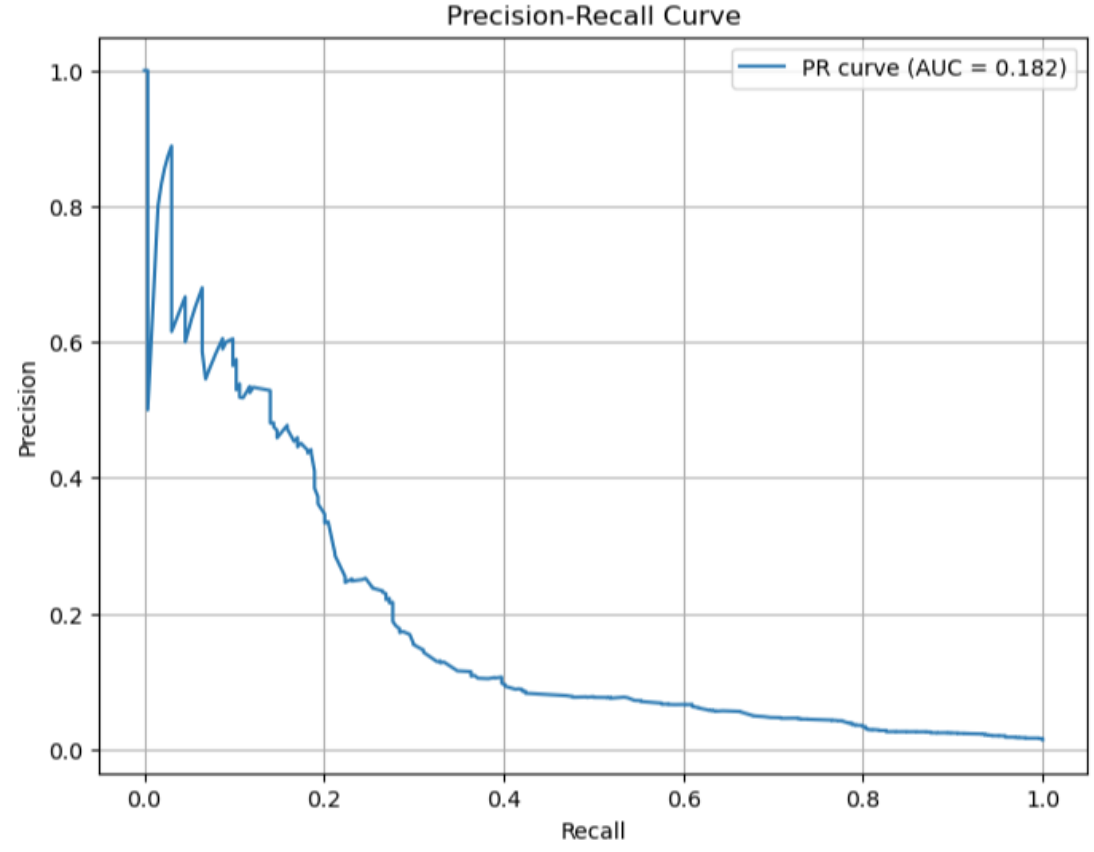
— LightGBM



Gestion du déséquilibre avec l’option scale_pos_weight = (nb cas positif / nb cas négatif) soit 69,36%

Paramètre	Catégorie	Description
objective = "binary"	Fonction de perte	Indique que la tâche est une classification binaire (fraude / non-fraude).
n_estimators = 2500	Structure du modèle	Nombre maximal d’arbres que LightGBM peut construire.
num_leaves = 127	Complexité des arbres	Nombre maximum de feuilles par arbre. Contrôle la capacité du modèle à capturer des patterns complexes.
learning_rate = 0.01	Vitesse d’apprentissage	Taille des pas de gradient. Plus c’est petit, plus l’apprentissage est stable.
min_data_in_leaf=150	Régularisation	Nombre minimum d’échantillons dans une feuille. Empêche les splits sur très peu de données.
feature_fraction=0,7	Sous-échantillonnage des features	À chaque arbre, LightGBM n’utilise que 70 % des colonnes.
bagging_fraction=0,7	Sous-échantillonnage des lignes	À chaque arbre, LightGBM n’utilise que 70 % des lignes.
bagging_freq=1	Fréquence du bagging	Active le bagging à chaque itération.
scale_pos_weight=70	Gestion du déséquilibre	Pondère la classe positive (fraude). Valeur ≈ ratio négatifs / positifs.
random_state=42	Reproductibilité	Fixe la graine aléatoire pour rendre les résultats identiques d’un run à l’autre.
Verbose=-1	Log	Désactive les logs LightGBM inutiles.

Calcul du PR-AUC et contrôle de stabilité



Contrôle de la stabilité du modèle sur 5 itérations

Itération Fold	Résultat PR-AUC
Fold 1	PR-AUC fold 1: 0.1700
Fold 2	PR-AUC fold 2: 0.1473
Fold 3	PR-AUC fold 3: 0.1630
Fold 4	PR-AUC fold 4: 0.1753
Fold 5	PR-AUC fold 5: 0.1342

Moyenne : 0.1579
Écart-type : 0.0151

Les variables qui pèsent le plus :

- **Meilleur seuil :** 1 (Fraude) >=0.930 >0 (Non-fraude)
- **Precision :** 58,1% de vraies fraudes soit : $\frac{TP}{TP+FP}$
- **Recall :** 29,9% de fraudes réelles détectées soit : $\frac{TP}{TP+FN}$

Graphique SHAP montre l'importance et le sens d'influence de chaque variable sur la prédiction de fraude.