Travail dirigé : Tests sur une Application déjà existante

© Pierre-Antoine Guillaume

2024-10-31

TD : le test appliqué au data

Objectif

— Aucun rendu ne vous est demandé à cette séance
— Un travail prendra pour sujet votre projet *fil rouge* à la prochaine séance, pour y adapter une stratégie de test automatisés.
— Vous êtes libres de vos outils : Pendant le cours, nous avons utilisé python, pytest et un aperçu de great-expectations, ici la seule contrainte c'est d'avoir des tests automatisés, des logs, et du monitoring adapté.
— Certains des exercices sont plutôt vus à destination de test de validation de données et de monitoring, et d'autres plutôt d'intelligence artificielle. Vous êtes encouragés à tout traiter.

L'objectif de ce TD est de mettre en pratique plusieurs concepts de test :
— les test sur les pipelines de données
— les tests sur le bien fondé des données
— les tests sur les modèles de données
— la possibilité de monitoring des solutions

Plusieurs détails :
— pendant ce TP, vous êtes libre de choisir vos outils et vos formats.
— si vous êtes indécis vous pouvez utiliser par exemple de faire des pipelines depuis les CSV originaux vers de nouveaux CSV
— d'utiliser pandas
— d'utiliser pytest
— d'utiliser scikit learn
— il est fait mention de logs et de monitoring dans le TP, vous pouvez utiliser l'objet logger de python, et partir du principe qu'une écriture de log de type error déclenche une notification. (c'est le rôle de bibliothèques comme sentry par exemple.)

Contexte

Il vous est fourni dans le repertoire `https://github.com/PierreAntoineGuillaume/methodologie-du-test/tree/main/session_3`
des CSV avec des données concernant des transactions.

Vous travaillez avec un dataset transactionnel fictif généré aléatoirement contenant les colonnes suivantes :
— **transaction_id** : Un identifiant unique pour chaque transaction
— **client_id** : Un identifiant unique pour chaque client
— **product_id** : Un identifiant unique pour chaque produit
— **amount** : Montant de la transaction
— **date** : Date de la transaction
— **label** : Indicateur de légitimité ou de fraude (catégorique, déterminé en fonction d'autres caractéristiques)

Première Partie : Exploration des données

Objectif

Analysez les fichiers `transactions-fair.csv` et `transactions-fishy.csv`.

Vous devez identifier :
— Les anomalies (ex : valeurs nulles, incohérences dans les colonnes)
— La distribution des labels dans chaque dataset
— Les caractéristiques critiques pour l'entraînement d'un modèle de machine learning

Deuxième Partie : Validation des données

— Écrivez des tests pour valider la qualité des données dans les datasets.
— Le set transaction-fishy est une tentative de *model poisonning*, écrivez un test qui permet d'éviter une corruption
— rajoutez des éléments de logs et de monitoring pour :
  — être prévenus en cas de problèmes (absence de données)
  — écrire des logs pour permettre la tracabilité des processus

Troisième Partie : Machine Learning

Objectif

Entraînez un modèle simple de classification pour prédire les labels (`label`). Comparez les performances entre les datasets `transactions-fair.csv` et `transactions-fishy.csv`.