Test technique Stage Data Scientist

Contexte business

Une start-up souhaite souhaite bâtir une offre vendue sous forme de licence à une catégorie d'entreprise faisant face à des incohérences dans leur données. Les entreprises cibles sont principalement des sites E-commerce et des retailers. On constate très souvent des incohérences sur la catégorisation d'un produit (FEDAS Code). La catégorisation incorrecte dans ce jeu de données correspond au champ **incorrect_fedas_code**. En vue de fiabiliser les données, la start-up souhaite développer un modèle ML permettant de corriger les données. Dans ce use case la donnée corrigée est la catégorie produit (**correct_fedas_code).**

Il faudra au préalable :

- Comprendre la problématique business
- Jeter un coup d'œil sur les données pour vous faire une idée de la problématique
- Faire une note succincte qui justifie vos choix techniques à défaut d'une présentation orale (README)

Il est fortement recommandé de passer exactement 4h sur le test car il s'agit d'un mini use case. Vos choix techniques doivent être cohérents avec le temps de développement alloué et vous serez jugé en conséquence.

Critères d'évaluation : fournir un repository git correspondant aux features suivantes (à la fin du développement, partager le repo avec l'adresse romuald.krappa@teamzen-services.com)

- Job de lancement de l'entrainement
- Job de sauvegarde et de chargement du modèle entrainé
- Job de predict à partir du modèle entrainé
- Qualité de code (docstrings, typing, linting, commit messages, etc.)
- L'architecture de votre repository

Il est fortement recommandé de créer une classe qui embarquera les principales steps du process ML. Le langage de programmation devra être Python.

Critères d'évaluation bonus :

- Création de la step d'evaluation du modèle
- Création du job permettant la promotion d'un nouveau modèle VS ancien modèle (on aimerait pouvoir déployer un nouveau modèle en production sur la base de critère pertinent)
- Packaging de l'application avec Docker
- Un accuracy à minima égale à 70%. Si vous obtenez une performance supérieure à 85.6% alors vous avez gagné le Kaggle puisque vous m'avez battu.

Si vous avez des questions, n'hésitez pas à envoyer un email à romuald.krappa@teamzen-services.com.