Tests Techniques Data-Scientist – Multi-classification

AXA Direct France

Objectif

Le principe général de ce cas d'étude est de construire un modèle prédictif à partir d'un ensemble de données.

Dans ce défi, l'objectif est de prédire la variable *target*. Vous pouvez entraîner votre modèle sur l'ensemble de données train.csv, les prédictions doivent être faites pour l'identifiant indiqué dans l'ensemble de données test.csv. Les prédictions sont les probabilités pour toutes les valeurs possibles dans la variable cible (0, 1, 2 ou 3). Nous vous donnons un exemple de la sortie finale attendue dans test_y_example.csv qui est une solution de base prédisant une probabilité de 25% pour toutes les classes.

L'évaluation se fera via une « weighted log-loss » qui pénalise de manière plus important les erreurs commises sur les labels en fonction des labels avec les poids 1, 10, 100 ou 1000 si la cible prend la valeur 0, 1, 2 ou 3 respectivement.

Livrable attendu

1. Un .csv contenant vos prédictions : Ce fichier doit être composé de 25 000 lignes, chacune contenant l'identifiant de test.csv et 3 colonnes contenant vos probabilités prédites pour cet identifiant.
2. Le code qui a permis de générer le modèle et les instructions pour l'exécution du code si nécessaire. Le format doit être un **notebook (**python ou R**)** ou un **Rmarkdown.** Joindre également une version html du notebook pour faciliter la relecture par l'évaluateur.
3. L'ensemble devra être envoyer par mail sous 72h.

Critères d'évaluations

1. L'exactitude de la prévision n'est PAS déterminante en raison des contraintes de temps. Il n'est donc pas nécessaire de consacrer beaucoup de temps et d'efforts pour affiner le modèle afin d'en accroître la précision.
2. La méthodologie mise en œuvre est plus importante que le score lui-même pour cette évaluation. Il est donc important d'expliquer votre approche. En particulier, n'hésitez pas à mentionner d'autres méthodes que vous avez testées, y compris celles qui ne se sont pas avérées pertinentes
3. Le langage de programmation est Python ou R.
4. Qualité du code : la lisibilité du code et des commentaires est très appréciée. N'hésitez pas à produire des résultats sous forme visuelle / graphique.
5. Point de bonus :
    - Créer un repository git sur le provider de votre choix (github, gitlab, bitbucket...) et nous fournir le lien
    - Un code de qualité : PEP8

Bonne Chance !