Test Technique Quantmetry

Objectif

Cet exercice a pour objectif d'évaluer vos compétences en data science. Vous serez évalués selon 4 axes :

- *Statistiques Descriptives*
- *Machine Learning*
- *Interprétation et restitution des résultats*
- *Qualité et clarté du code*

Votre tâche consiste à produire 2 fichiers :

- Les réponses aux questions ci-dessous dans un fichier texte (Word, LibreOffice ou PDF) nommé *NOM_Prénom_réponses.ext (6 pages max)*
- Le code produit durant l'exercice dans un fichier nommé *NOM_Prénom_Code.R* ou *NOM_Prénom_code.py (merci de préciser dans le header la version de python utilisée. eg: python 2.x ou python 3.x).* Le code doit être écrit de préférence en R ou en Python (ipython notebook accepté). Si vous souhaitez utiliser un autre langage, merci de bien vouloir nous faire confirmer votre choix au préalable. Vous êtes libres d'utiliser toutes les librairies de Machine Learning que vous estimerez nécessaires. Votre script devra pouvoir être rejoué facilement. N'hésitez pas à fournir un fichier README ou à spécifier dans le corps du script les actions nécessaires à la relance de votre code.

Enoncé

Le jeu de données contenu dans `data.csv` est un ensemble de relevés de vélos partagés (type Vélib). On cherche à prédire le nombre de vélos loués par heure dans la ville (variable `count`). Voici une brève description des données :

`datetime` - date et heure du relevé
`season` - 1 = printemps , 2 = été, 3 = automne, 4 = hiver
`holiday` – indique si le jour est un jour de vacances scolaires
`workingday` - indique si le jour est travaillé (ni week-end ni vacances)
`weather` - 1: Dégagé à nuageux, 2 : Brouillard, 3 : Légère pluie ou neige, 4 : Fortes averses ou neiges
`temp` – température en degrés Celsius
`atemp` – température ressentie en degrés Celsius
`humidity` – taux d'humidité
`windspeed` – vitesse du vent
`casual` - nombre de locations d'usagers non abonnés
`registered` – nombre de locations d'usagers abonnés
`count` – nombre total de locations de vélos

Partie I – Statistiques descriptives

1. Sans commencer la partie modélisation, quels sont les facteurs qui semblent influencer la demande en vélos ? Justifiez vos choix. Présentez quelques graphiques pertinents pour illustrer votre réponse et interprétez-les.
2. En supposant que vous ayez accès au sexe et à l'âge des utilisateurs abonnés, quelle procédure statistique vous permettrait de dire si les distributions en âge des deux populations (femme et homme) sont identiques ou non ? Décrivez les étapes de votre raisonnement.

Partie II – Machine Learning

1. Concevez un modèle permettant de prédire la variable **count** et expliquez votre choix d'algorithme. Si votre modèle comporte des spécificités de paramétrage, justifiez également vos choix de paramètres.
2. Décrivez et justifiez le critère de performance utilisé.
3. Proposez deux à trois pistes d'amélioration de votre modèle.