TP Introduction à la data Science

Master MSI – M. Page

On s'intéresse aux objets trouvés dans les trains et gares de la SNCF. On dispose d'un jeu de données disponible sur Moodle : "objets-trouves-restitution.csv". La description des champs contenus dans ce jeu de données est la suivante :

- Date – date et heure de quand l'objet a été trouvé.
- Date et heure de restitution – date et heure de la restitution de l'objet, si il a été rendu.
- Gare – gare où l'objet a été trouvé.
- Code UIC – identifiant de la gare.
- Nature d'objets – nature de l'objet trouvé.
- Type d'objets – type de l'objet trouvé

Travail demandé :

1. Charger les données dans pandas et produire quelques informations de base sur les différentes colonnes du jeude données (type des colonnes, statistiques descriptives, volumétrie).
2. Analyser la présence de valeurs manquantes dans chaque champ du jeu de données. Indiquez la stratégie de traitement des valeurs manquantes que vous pensez judicieuse pour cette étude de cas, et réalisez le traitement correspondant.
3. Analyser la présence de valeurs incorrectes dans chaque champ du jeu de données. Indiquez la stratégie de traitement des valeurs incorrectes que vous pensez judicieuse pour cette étude de cas, et réalisez le traitement correspondant.
4. Extraire un échantillon aléatoire de 100 000 enregistrements du jeu de données. La suite du TP se fera sur cet échantillon. **(Attention, si vous voulez être sûr de toujours travailler sur les mêmes 100k enregistrements, utilisez une seed pour votre fonction aléatoire.)**
5. Ecrire des requêtes pandas répondant aux questions suivantes:
    a. top 10 des types d'objets les plus retrouvé.
    b. top 10 des types d'objets les plus rendus
    c. découper Dates en Day (n° du jour dans le mois), Week (n° de semaine dans l'année), Month (n° du mois dans l'année), Hour (entre 0 et 24).
    d. transformer Hour en Plage Horaire.
    e. calculer l'attribut Returned comme suit : 0 si l'objet n'a pas été rendu ; 1 sinon.
    f. tableau donnant pour chaque valeur de Type d'objets et de Returned le nombre de cas et le pourcentage de cas **par rapport au nombre de cas de la catégorie**.
6. Construire 6 graphiques matplotlib permettent d'analyser la relation entre *a* et Returned pour chacun des attributs *a* suivant : Day, Week, Month, Hour, Type d'objets, Gare.
7. A partir de l'échantillon, construire un jeu de données d'entraînement (2/3 des enregistrements) et un jeu de données de test (1/3 des enregistrements).
8. Construire un modèle de type régression logistique contenant les attributs que vous jugerez pertinents pour prédire la variable Returned puis appliquer ce modèle sur le jeu de données de test et donner la matrice de confusion.
9. Même question en utilisant un modèle de type random forest, avec les mêmes attributs.
10. Conclure sur la comparaison des deux modèles pour l'étude de cas.

Question bonus : refaire les questions 8 et 9 en créant vos propres types d'objets en partant de nature d'objets. Cela change-t-il quel que chose ?

Le TP est à rendre par binôme.