Python for data-science
Devoir à la maison
Date de rendu : 1er mars 2019

L'objet du devoir est de mettre en pratique les différentes compétences nécessaires au métier de data-scientist, vues en cours cette année :

- Web-scrapping
- Data-visualisation
- Modélisation
- Passage en production via la création d'une API
- Usage de Git

Une base de donnée distincte est affectée à chaque étudiant du cours selon le fichier excel joint. Chacun doit :

- Créer un script python qui automatise toute la procédure de modélisation de la base de données :
    - Téléchargement via webscrapping de la base de donnée sur le pc local (via selenium ou beautiful soup)
    - Data-visualisation des données (via matplotlib, seaborn ou bokeh …)
    - Data-préparation (pandas)
    - Modélisation (scikit learn)
    - Optimisation des hyperparamètres (grid search)
    - Visualisation des performance (courbe roc …)
- Créer une API permettant d'interviewer un web service pour obtenir une prédiction sur une nouvelle instance de la base de donnée :
    - Avec django et django-rest-framework

Le résultat attendu est en plusieurs points :

- Deux repositories github :
    - Un pour la partie modélisation
    - Un pour la partie API
    - Chacun doit avoir :
        - Du code propre, commenté et fonctionnel
        - Un fichier readme expliquant le contexte, et les objectifs
- Un ppt :
    - Présenter l'analyse descriptive des données dans un powerpoint qui explique le contexte de la base de donnée, la cible à prédire, les différentes features disponibles, les étapes de récupération de données, de features engeneering, les tests des différentes hyper paramètres, et les gains de performance des différents modèles.
    - Le nom du ou des élèves
    - Ce fichier ppt doit être dans le github

Il y a en général deux étudiants par base de données. Vous avez le choix de faire le travail en groupe ou seul. Le rendu sera envoyé par mail à romain.jouin@memorandum.pro avec dans le titre du mail :

[Python for data-science - Devoir à la maison] + votre nom et prénom + nom de la base