ENSA
المدرسة الوطنية للعلوم التطبيقية بالجديدة

Examen : Visualisation et Fouille des Big Data
Professeur : Fahd KALLOUBI
Année : 2022/2023
Niveau : 3ème année (2ITE)

Vous serez amenés à prévoir le montant du tarif pour un trajet en taxi à New York. Bien que vous puissiez obtenir une estimation de base basée uniquement sur la distance entre les deux points, cela se traduira par un RMSE de 5 $ à 8 $. Toutefois, Votre défi est de faire mieux que cela en utilisant des techniques d'apprentissage automatique.

Les champs de ce dataset sont les suivants :

ID	Chaîne unique identifiant chaque ligne dans les ensembles d'apprentissage et de test. Composé de pickup_datetime plus un entier unique
pickup_datetime	Une valeur de type « timestamp » indiquant le début du trajet en taxi.
pickup_longitude	Une valeur de type « float » pour la coordonnée de longitude de l'endroit où le trajet en taxi a commencé.
pickup_latitude	Une valeur de type « float » pour la coordonnée de latitude de l'endroit où le trajet en taxi a commencé
dropoff_longitude	Une valeur de type « float » pour la coordonnée de longitude de l'endroit où le trajet en taxi a terminé.
dropoff_latitude	Une valeur de type « float » pour la coordonnée de latitude de l'endroit où le trajet en taxi a terminé.
passenger_count	Entier, indiquant le nombre de passagers dans le trajet en taxi.
fare_amount	Montant du coût du trajet en taxi. **C'est la valeur à prédire**

Vous trouverez le dataset en question dans le dossier de l'examen.

1.1 Analyse exploratoire, prétraitement et visualisation

En suivant le processus d'un projet en science de données, vous devez réaliser les prétraitements requis afin de mieux répondre à la question métier :

1. Affichez et puis Supprimez les valeurs manquantes de ce dataset
2. Supprimez les trajets ayant un coût (i.e. fare_amount) négatif
3. Visualisez ensuite le dataset (i.e. scatter) en utilisant la colonne indiquant le nombre de passagers. Est-ce qu'il y a un outlier ? si oui, supprimez ce trajet
4. Etant donné que la ville de New York est comprise entre [-90,90] de latitude et entre [-180,180] de longitude, supprimez les trajets qui correspondent à un bruit.
5. Changez le type de la colonne « pickup_datetime » vers le type « datetime » et puis affichez le résultat
6. Créez maintenant les colonnes suivantes : Year, Month, Date, day of week, Hour
7. En utilisant des visualisations, réponder aux questions suivantes :
    a. Le nombre de passagers affecte-t-il le coût du trajet ?
    b. L'heure du début du trajet affecte-elle le coût du trajet ?
    c. Le jour de la semaine affecte-t-il le coût du trajet ?

1.2 Features engineering

1. Visualiser la corrélation des caractéristiques avec la cible
2. En utilisant : Recursive features elimition, Random Forest
    a. Afficher la moyenne d'importance de chaque caractéristique en utilisant les facteurs obtenus par chaque technique
    b. Visualiser le résultat obtenu
    c. Quelles sont les 4 caractéristiques les plus importantes ?

week of the day , year , dropoffpongitude pickup Pungitude

1.3 Apprentissage du modèle et réglage des hyper-paramètres

Une fois votre analyse est terminée et vos données sont préparées, vous êtes amenés à apprendre les modèles d'apprentissage ci-dessous en utilisant les caractéristiques considérées tout en réglant les hyper-paramètres pour chaque estimateur :

1. La régression linéaire
2. La régression logistique
3. Arbres de décision

Par la suite vous devez utiliser des techniques d'agrégation (ou ensemblistes), tout en réglant les hyper-paramètres, essayez d'implémenter les algorithmes ensemblistes ci-dessous en utilisant la méthode d'évaluation holdout :

1. Voting
2. Random forest
3. XGBoost

1. En utilisant l'algorithme avec la meilleure performance, refaites le même processus en utilisant les 4 caractéristiques les plus importantes (i.e. section 1.2). Les caractéristiques considérées améliorent-elles la performance ?

1.4 Créer et consommer l'API du modèle

1. Sérialiser le modèle ayant donné la meilleure performance au format Pickle
2. En utilisant FastAPI, créez une API REST de ce modèle
3. Créer une application web dédiée (i.e., formulaire web) afin de consommer le service crée.
4. Déployer l'api récemment créée en tant que micro service en utilisant Docker

1.5 Utilisation de Amazon SageMaker

En utilisant Amazon SageMaker :

1. Entrainer le modèle ensembliste XGboost
2. Régler les hyper paramètres de ce modèle
3. Déployer le modèle ainsi obtenu
4. En utilisant directement le endpoint obtenu après déploiement, essayer de prédire sur un ensemble de lignes.