Rapport Exercice de recrutement TICTACTRIP

Réalisé par Khalil Allah Abbes

Introduction

Je suis très intéressé par le poste de Data Scientist chez Tictactrip. Pour démontrer ma motivation, j'ai travaillé sur votre jeu de données et je souhaite partager mes démarches avec vous. Mon travail est divisé en trois parties : l'étude du dataset (dataset.ipynb), la visualisation des données (dataviz.ipynb) et la modélisation (modelisation.ipynb).

I) Etude du dataset

J'ai eu accès à quatre fichiers Excel: ticket_data.csv, cities.csv, stations.csv et providers.csv. Le fichier principal contenant des informations sur les différents voyages est ticket_data. Toutefois, il est relié à d'autres données par des clés (id). Mon premier objectif était donc de créer un unique dataset comportant toutes les informations nécessaires. J'ai utilisé de nombreuses boucles for pour trouver, pour chaque voyage, la ligne correspondante aux villes et aux compagnies. Cette approche rend le programme long. Cependant, je pense que cela peut être optimisé en utilisant des fonctions déjà existantes. Néanmoins, ce programme n'est utile que pour écrire un nouveau fichier Excel complet et il ne sera utilisé qu'une seule fois.

Une fois les nouvelles informations chargées dans le DataFrame pandas, j'ai choisi de rajouter des données que je considère importantes pour une analyse complète. Ces données sont :

la durée du trajet,

la catégorie de durée du trajet (court, long, ...),

la distance approximative du trajet (grâce aux longitudes et latitudes des villes de départ et d'arrivée),

la catégorie de ces distances (court, long, ...),

le nombre de stations intermédiaires pour chaque trajet,

le nombre de compagnies intermédiaires.

Une fois les nouvelles données traitées et vérifiées, je les ai sauvegardées sous le nom de tickets.csv dans le registre des données. Il faut lancer ce notebook pour créer le nouveau fichier Excel avant de lancer les prochains programmes.

II) Visualisation des données

Nous pouvons maintenant jeter un coup d'œil approfondi à nos données. À l'aide des fonctions pandas, il est facile de déterminer la moyenne, le minimum et le maximum des features numériques. Par exemple, voici la moyenne des prix en centimes :

La librairie seaborn propose de nombreuses solutions de dataviz (avec des prix en € pour une meilleure lisibilité). J'ai principalement utilisé deux types de visualisation pour approfondir l'étude. D'une part, la boîte à moustaches (boxplot) :

Cela permet d'obtenir de manière plus visuelle la moyenne et la variance de nos features avec jusqu'à trois paramètres simultanément. D'autre part, j'ai utilisé des histogrammes pour étudier la répartition de certaines features :

Il existe de nombreuses façons de visualiser les données et je me suis arrêté là. En effet, sans cahier des charges ou objectifs clairs, l'exercice peut être infini. Toutefois, je tiens à souligner qu'il existe des outils de power BI permettant des visualisations très efficaces. Les combiner à Python permettrait d'être plus efficace dans les présentations.

III) Modélisation

Pour choisir le meilleur modèle, j'ai utilisé une méthode de validation croisée (cross-validation) qui consiste à séparer les données en plusieurs ensembles d'apprentissage et de validation, pour évaluer la performance du modèle de manière plus robuste.

Après avoir entraîné différents modèles et évalué leur performance à l'aide de la validation croisée, j'ai finalement choisi d'utiliser le modèle XGBRegressor pour prédire le prix des trajets. Ce modèle est basé sur un algorithme d'arbres de décisions boostés qui permet de prendre en compte des relations non linéaires entre les différentes caractéristiques des trajets.

Enfin, j'ai utilisé le modèle choisi pour prédire le prix des trajets pour un jeu de données de test. Les résultats ont montré que le modèle était capable de prédire le prix avec une bonne précision, avec une erreur moyenne de l'ordre de 5 à 10 euros sur des prix moyens de 30 à 40 euros.

En conclusion, cette étude montre qu'il est possible d'utiliser des techniques de data science pour analyser et modéliser des données de voyage. Le travail réalisé ici peut être amélioré en utilisant des techniques plus avancées de prétraitement de données et de modélisation, ainsi qu'en intégrant d'autres caractéristiques des voyages.