Baptiste Hurel

M1 IoT

16/11/2020

Rapport de projet

Problématique : Récupération d'un dataset sur la NBA et dans un 2em temps un dataset sur les séries et films Netflix.

Mon modèle à pour but de définir la quantité et la répartition des séries et films sur la plateforme Netflix et quel est le leader en termes de production. En fonction de retour utilisateurs sur leurs préférences on pourrait adapter notre production de film ou séries.

- Travail de Cleaning de data
- Travail de visualisation sur les données

Introduction :

J'ai commencé avec un dataset sur la NBA puis j'ai pris un dataset Netflix.

Certaines parties (1 et 3) sont plus pertinentes dans le notebook NBA mais la majorité du devoir est sur le dataset Netflix.

Tous les documents sont disponibles à cette adresse :

https://github.com/BaptisteHurel/Python/tree/master/Mini%20Projet

---

1) Analyse graphique des données (Dataviz)

Dans cette partie j'ai commencé par faire un petit clean du dataset afin de récupérer que les données utiles pour les parties suivantes.

Suppression de la colonne director car trop de données manquantes

1.1) Diagrammes de répartition des données (type gaussienne sur les données)

J'ai configuré et affiché les données permettant d'obtenir un histogramme représentant le nombre de films ou séries en fonction de leurs durées (en minute).

1.2) Vérification du nombre de données, si plusieurs données sont peu représentées (<3%) alors regrouper dans une seule et même catégorie, 1 pie chart avant/après

Dans cette partie on cherche à représenter la proportion entre les films et les séries au sein du catalogue Netflix.

On obtient un diagramme camembert (pie chart)

1.3) Nettoyage des données manquantes, encodage (OneHot, dictionnaire ou Sklearn Encoder)

On veut remplacer les données catégories Movie et TV Show en des données numériques (O et 1), on utilise pour cela un encodage de type One_Hot.

Cela permettra de mieux traiter les données pour plus tard.

1.4) Boites à moustache avec données extrêmes (Box Plot)

L'utilisation d'une box plot permet d'avoir un visuel du profil statistique de ce dataset et en particulier de la durée moyenne des films

On y observe que la grande majorité des films à une durée comprise entre 80 et 120min.

1.5) Heatmap + observations sur les corrélations

Ce graphe Heatmap montre une correlation de 0.57 soit 57% entre le show_id et la release_year.

Cela n'est pas très représentatif et pertinents car ce dataset manque de données pour analyser des corrélations entre différents éléments.

2) Model Building

2.1) 2 algorithmes avec 2 paramètres différents (ex: max_depth, n_estimators,….) que vous expliquerez en commentaire

Dans cette partie on à 2 algorithmes :

- Un Algorithme permettant de gérer des problèmes de classification à plusieurs classes
- Algorithme de régresseur de vote par prédiction pour les estimateurs non ajustés

2.2) Affichage des coefficients/ accuracy

(Voir dataset NBA pour plus de pertinence)

On entraine ici un modèle de donnée et on cherche le score du modèle linéaire permettant de définir la pertinence de notre jeu de donnés.

On obtient ici un score de 0.57 entre les (points/rebonds/passes) et le nombre de match joué par un joueur, le lien entre ses données n'est pas totalement lié.

3) Feature Importance

3.1) Affichage sous forme de barplot

On affiche ici les 10 plus importants producteurs de séries puis de films sur Netflix en fonction de leur nombre de séries puis films.

On y observe que les états dominent très largement ce marché et que la grande l'inde est largement plus importante dans le classement du nombre de film que du nombre de séries par pays.

3.2) Autre forme d'affichage si vous avez le temps

Simulation d'un dataset random car mon dataset ne permet pas d'obtenir cet affichage

On affiche ici l'arbre de régression avec une coupe de profondeur 3.

4) Model Réexécution avec les features sélectionnés

4.1) Affichage des metrics standard et commentaire sur la pertinence

(Voir dataset NBA pour plus de pertinence)

On cherche ici le score de précision moyen pour conclure sur la pertinence de nos données ainsi sur le score de précision équilibré.

Concernant le dataset on obtient un score assez proche du score moyen équilibré qui est de 0.625, celui de la NBA étant de 0.578.

On peut en conclure que sur mon dataset NBA la pertinence de mes données est correcte.