PART 1: DATASET INFORMATION

Topic: NETFLIX TV SHOWS AND MOVIES

Dataset: NETFLIX TV SHOWS AND MOVIES DATASET

Source: Kaggle Datasets

Dataset information: TV Shows and Movies listed on Netflix

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. (https://www.kaggle.com/shivamb/netflix-shows)

PART 2: HYPOTHESIS TESTING

The aim of this study is to analyze whether rating is predict by looking at the description of the movies and tv shows.

PART 3: EXPLORATORY DATA ANALYSIS

Number of Columns: 12
Number of Rows: 7787

Data types: Varies
Missing values: Some values are dropped some values are filled with a not known category
Value counts of y : There are 10 ratings.

Data visualization: Visualized the data to show duration, most common county which produced TV shows and movies, top directors, top genres etc.

PART 4: MODELING

DATA PREPARATION FOR NLP: Downloaded IMDB ratings data from IMDB website and merged the data with netflix data.

NLP : Selected description column and rating column. Cleaned description column tokenized and coded each word as a column. Selected 50 common words.

A. RECURSIVE FEATURE ELIMINATION

Aim: To conduct feature engineering and eliminate unnecassary X variables (words), I performed a Recursive Feature Elimination model.

Results: The RFE eliminated only five(5) X variables which is not enough to conduct the study.

B. MODELING USING PYCARET

Aim: Compare machine learning algorithms on the netflix dataset by using the test data.

Results: The finding showed that Gradient Boosting Classifier is the best model which explains the relationships between description (Xs) and IMDB ratings (y) with the best accuracy of 0.33