Netflix Movies and TV Shows Clustering

Unsupervised ML

Project Description

Business Context

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.

In this project, you are required to do

- Exploratory Data Analysis

- Understanding what type content is available in different countries
- If Netflix has been increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

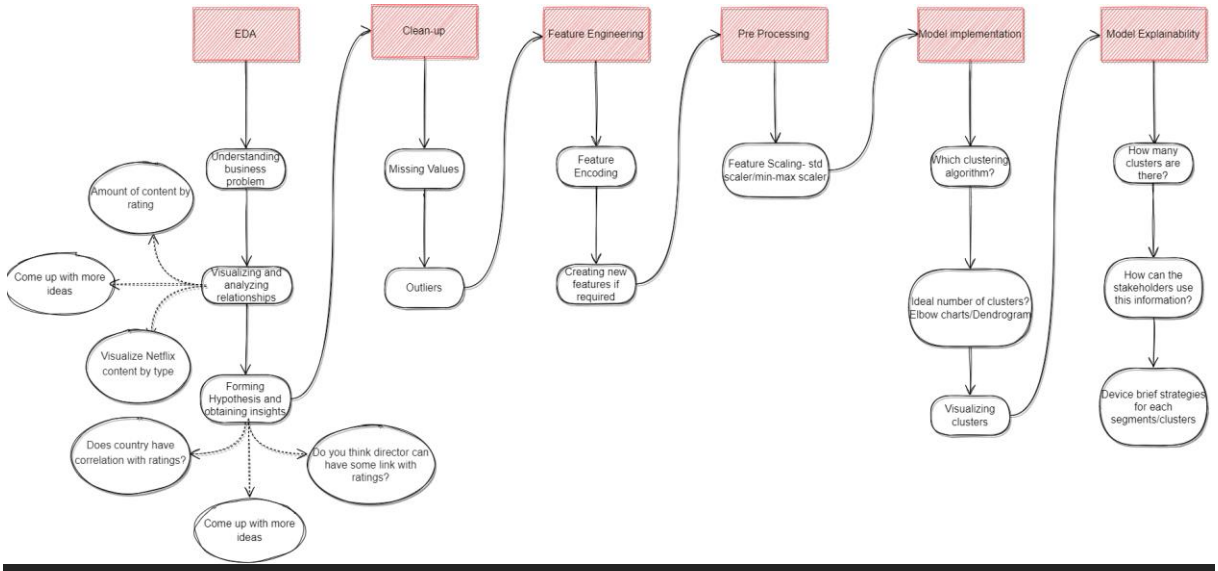
Dataset Description

Fields	Description
show_id	Unique ID for every Movie / Tv Show
type	Identifier - A Movie or TV Show
title	Title of the movie/show
director	Director of the show
cast	Actors involved
Country	Country of production
date_added	Date it was added on Netflix
release_year	Actual release year of the show
rating	TV rating of the show
duration	Total duration in minutes or number of seasons
listed_in	Genre
Description	The summary description

Main Libraries used:

- Pandas for data manipulation, aggregation
- Matplotlib and Seaborn for visualization and behavior with respect to the target variable
- NumPy for computationally efficient operations

Project Architecture:


EDA
Understanding business problem
Amount of content by rating
Come up with more ideas
Visualizing and analyzing relationships
Visualize Netflix content by type
Forming Hypothesis and obtaining insights
Does country have correlation with ratings?
Do you think director can have some link with ratings?
Come up with more ideas
Clean-up
Missing Values
Outliers
Feature Engineering
Feature Encoding
Creating new features if required
Pre Processing
Feature Scaling- std scaler/min-max scaler
Model implementation
Which clustering algorithm?
Ideal number of clusters? Elbow charts/Dendrogram
Visualizing clusters
Model Explainability
How many clusters are there?
How can the stakeholders use this information?
Device brief strategies for each segments/clusters

Help