Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Team Member's Role: - ❖ Kirtesh Verma(kirteshverma12345@gmail.com) Contribution: ○ Data understanding ○ Handling null & missing values ○ Performing EDA ○ Data preprocessing ○ Silhouette score ❖ Pravin Bejjo(praveen.bejo.pb@gmail.com) Contribution: ○ Data understanding ○ Data visualization ○ Removing punctuation and stop words ○ Dendogram ○ K-means clustering ○ Visualize silhouette score and clusters ❖ Sahil Pardeshi(8623879021.sp@gmail.com) Contribution: ○ Data understanding ○ Data visualization ○ Feature engineering ○ TF-IDF vectorizer ○ Elbow method
Please paste the GitHub Repo link.
GitHub Link:- https://github.com/praveenbejo95/Netflix-Movies-and-TV-shows-Clustering

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions. It's is a very popular streaming platform especially in countries like United States, India and United Kingdom. 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. It has 7787 entries and 12 attributes.

Our objective is to do text-based clustering on the given dataset so that we can cluster similar content on Netflix. For this we began by performing Exploratory Data Analysis on the given data thus drawing meaningful insights which helped us to understand the type of content available on Netflix, the targeted audience, top genres and many more.

We featured engineered column for better intuition and followed data preprocessing steps to make to data ready for model building. We converted the data types of certain columns. We first began by handling the null values. Maximum null values were present in the 'director' and 'cast' column, as this are not important for model building, we dropped them.

For the text-based column we removed stop words and punctuations, performed stemming and calculated TF-IDF. We performed this steps on 'description' and 'listed_in' column. As this is an unsupervised machine learning problem, we used K-means to perform clustering.

We used elbow method to derive the most suitable number of clusters giving us the acceptable error term with less computational cost. The number of clusters for k-means clustering turned out to be 5. To measure the goodness of clusters we used Silhouette score method. The score for 3 clusters are 0.34 which is good.
We used 'length' and 'length_listed' column for clustering purpose.

We did a thorough analysis of the data and extensive data preprocessing to make the data ready for model deployment. After working on this dataset we concluded that Netflix has been producing more movies than TV-shows as almost 70% of the content on Netflix is of type movies and rest is TV-shows. It's most famous in United States and India. India is the only country were content released is targeted towards teens. Movies and TV-shows belonging to 'TV-MA' rating is highest on Netflix. It is for Mature audience only.

Netflix is leading in the streaming service and conclusions derived from Netflix dataset will help the competitors to plan out their strategies and future business plans.