# Netflix Movies and TV shows

Netflix is a popular online platform which provides people with variaces kinds of movies and tv shows. During the last year, many of us have planty of chances to explore the many movies and tv shows, but what bothers me from time to time is to choose a movie or tv show to watch. Through this project, my goal is to find the key factors that separate movies from tv show.

The data is three csv files. The first one is from https://www.kaggle.com/shivamb/netflix-shows, and this contains all the movies and tv shows in netflix, but it does not have the IMDb rating and Rotten Tomatoes Rating. Thus I downloaded two more dataset, https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney for movies, and https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney for tv shows.

After inner join these three datasets, the columns I selected for my analysis are (1) type, (2) title, (3) released year, (4) rating, (5) IMDb rating, (6) Rotten Tomatoes. The type is whether it is a movie or tv show, title is the name of the movie or tv show, released year is the year it is released on, the rating are the age groups it is targeted at, IMDb rating is the IMDb score, and Rotten Tomatoes is the Rotten Tomatoes score. For the rating column I combined tv shows rating with movies rating, I use G, PG, PG-13, R to replace TV-G, TV-Y, TV-Y7, TV-PG, TV-14, TV-MA so that in the third step, the rating column will not tell the program whether it is a tv show or not by only the prefix. After these steps, our data observations were reduced from 7787 to 1754.

First, we are exploring whether the time is a significant factor that separate movies from tv shows. The blue dots in **Figure 1** represents movies and the orange dots represents tv show. We can see that most of the movies and tv shows are released close to the right. In fact, 98.36% of tv shows are released after 2010, and 94.50% of movies are released after 2010. These two observations looked closely and seems like time will not be a significant factor when determining whether it is movie or tv show.

Second, we investigate the dimensionality of our data by doing PCA (principal component analysis) over 7 variables (actually 4 variables, the rating type are being seen as 4 different age groups). The orange line of **Figure 2** shows the unscaled results, which 98.01% of the variance is being captured by the first two variable, that is because the last 4 columns are 0 or 1 and that makes the first two columns have a bigger weight. Whereas in the blue line where I use the standard scaling, and 49.46% variance are captured by the first two columns, but the last column almost captures no variance with a value of $1.488 * 10^{-32}$ , and that tells us that we may be able to reduce one part of the rating.

Third, I am going to predict whether a given show is a movie or tv show based on its IMDb, Rotten Tomatoes, released year and targeted age group. A skylearn pipeline consisting of (1) a StandardScaler to normalize the data and (2) a LogisticRegression is used for this task. After applying a 75/25% train/test split, I performed a permutation testing to get a score of 0.8040. Although the data has about 75.66% movies, so 80.40% may not seems great, but with a p-value of 0.0099, the score is statistically significant. And by looking at the testing results, I found that the predictions are not all movies, there are 17.31% being predicted as tv show, which means this is not too bad and could be useful to determine whether it is a movie or tv show.

**Figure 3** shows the coefficient weight for each of the features used. The observation shows that IMDb is the most import indicator which positively shows whether it is a tv show or not. And the released year is not only vital, but also has a positive correlation with tv show.

In conclusion, though most of movies and tv shows are both in recent years, but their released time is still an crutial indictor of its type. By looking at all the indicators, I found that they are all important, and there is little possibility of dimensional reduction. At last, after model training, the most import factor in determining whether it is a movie or tv show is the IMDb rating.

Figure1: IMDb rating of different released year and relation to their targetted age groups

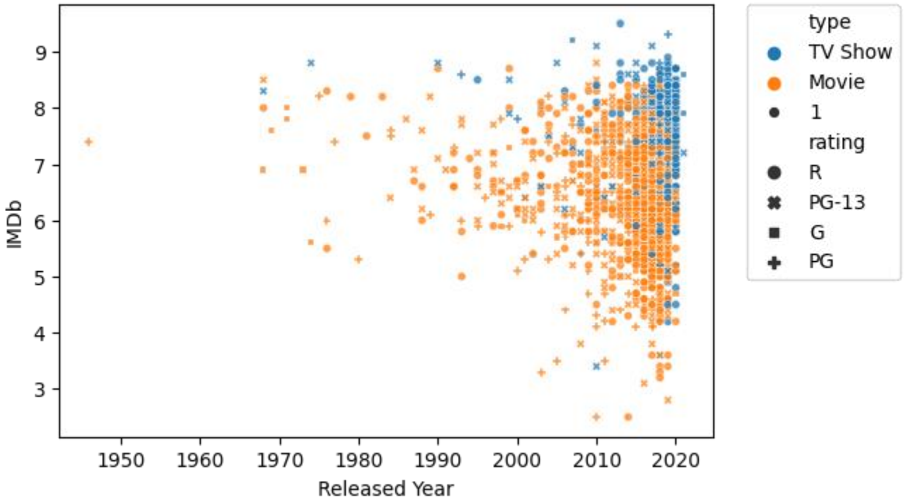


Figure2: principle components analysis

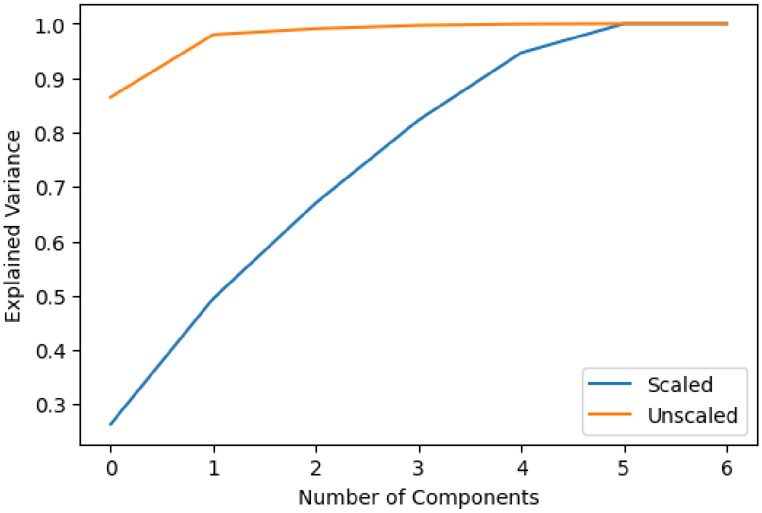
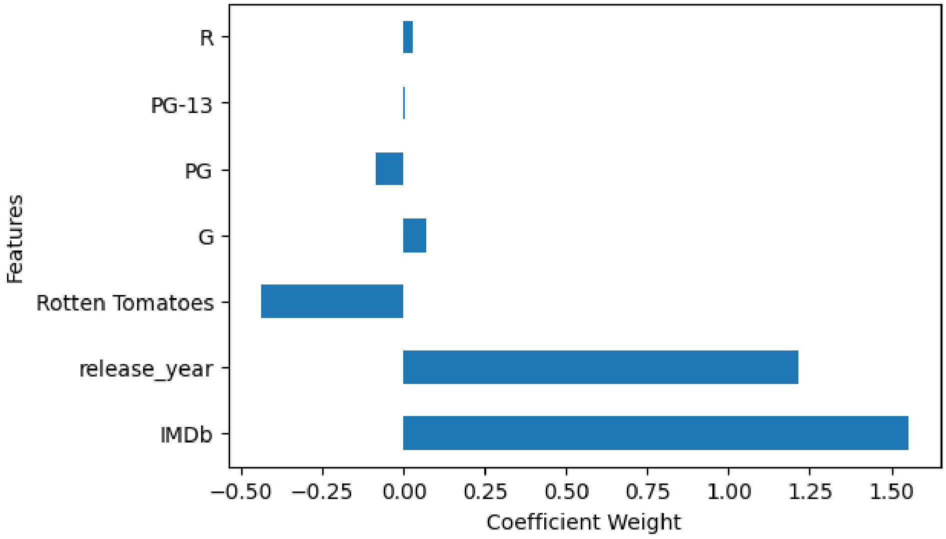


Figure3: Logisitc Regression Coefficients


Features
R
PG-13
PG
G
Rotten Tomatoes
release_year
IMDb
Coefficient Weight