Eli Oceanak
5/14/21
COSC 4555
Final Project Report

Final Project Report
By Eli Oceanak

Problem Statement

I want to predict IMDB ratings via various other data available on any particular film's IMDB page. This is a nifty idea because it will show if there is a formula that can help a film get a higher rating, just by following the formula I come up with. The IMDB rating is on a score of 1-10.

Significance

The significance of this work is that it will show whether film ratings are truly random, or can be predicted (without even watching the movie!). This will add to the currently existing work done in this sector by using newer data.

Literature survey

There were several projects I found that already worked on this problem. They were mostly done several years ago though, so by using newer data I thought I could contribute to the current literature. I also initially meant to do this work on TV shows, which was much different than the norm of doing movies but I eventually realized TV shows were not going to work as I'll discuss later in the challenges section.

Dataset Description

I used an IMDB movies dataset available on Kaggle.

It can be found here:

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

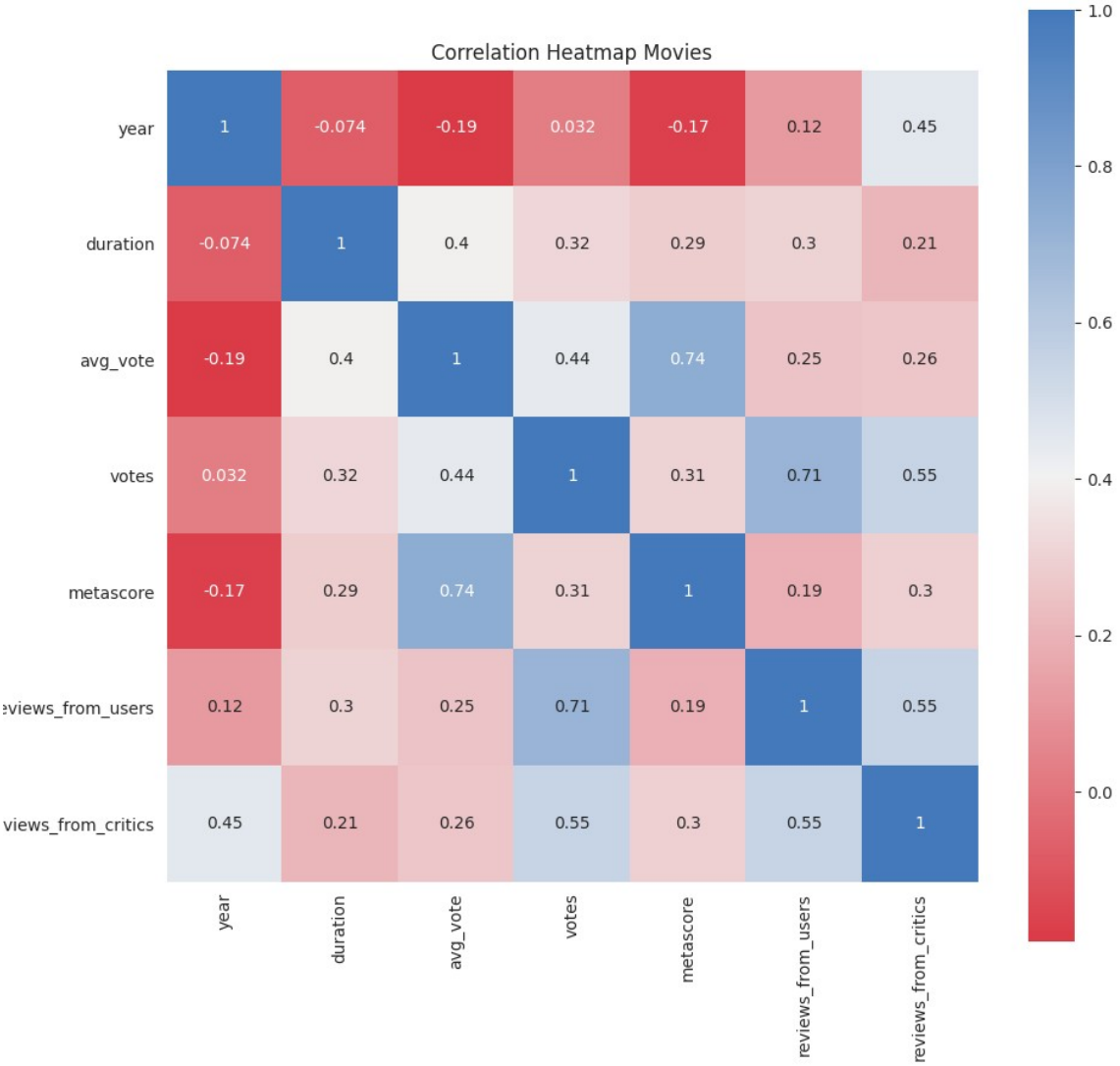
The dataset contains multiple csv files that all have different data on different movies. Some of the categories were duration (movie length), release year, language, total number of votes, and the average vote (IMDB rating). There were many other categories available as well.

Methodology

The first thing I did was clean the data. I removed all the Null/NaN valued movies, limited the minimum number of votes on any given movie to 10000, and then removed some extreme outliers that must have had bunk values input to the dataset incorrectly somehow.

I also split the data into training and testing datasets at a 70/30 ratio.

I then created a heatmap of the values and their correlations with the IMDB rating to get a good idea of which values to focus on. This heatmap can be seen below. This was done using the Seaborn package.



After creating this heatmap, I was able to proceed with linear regression while ignoring all the categories that did not have any chance of being useful to me.

I used the sklearn package to assist with linear regression.

In the end I came up with two separate linear regression models for project.

Results

I came up with two linear regression models, shown below.

Model 1:

$$\text{Rating} = 3.789 + 1.069 * 10^{-6} * \text{numVotes} + .007 * \text{duration} + .033 * \text{metascore}$$

Model 2:

$$\text{Rating} = 4.865 + 1.9 * 10^{-6} * \text{numVotes} + 0.013 * \text{duration}$$

Model 1 was a better model, with a lower MSE on the testing data of .355. However, I decided to use Model 2 as my final model. The reason I did this is that one of the predictors for Model 1 was “metascore”. Now “metascore” is a rating from another movie website, and I believe that using a movie rating score to predict another movie rating score is not very useful, so I decided against using model 1 as my final model.

Model 2 still performs well. Here are the testing and training data errors:

Training data mean squared error: 0.6506180403790429
Training data mean absolute error: 0.6206031536750252

Testing data mean squared error: 0.726093638519896
Testing data mean absolute error: 0.6474434307452884

Challenges

There was one really big challenge in this project, which was that I originally attempted to guess the rating of TV shows.

The datasets I could find that included the ratings and data for TV shows were often incomplete, and did not have the same level of pruning and polishing done on them as the movie datasets had. In hindsight this makes sense as IMDB is a movie website, so TV shows would almost certainly be a secondary concern of anyone scraping data from the website.

I went through many hours of work cleaning/working on the TV dataset to get some usable data points, but they just weren't helping. Here is an example of one of the heatmaps I got for the TV dataset and some predictors:

As you can see, there isn’t a single variable that correlates significantly with the average rating of any given TV show.

This was the biggest challenge as I had to entirely abandon a lot of work due to it being unsalvageable.

Once I switched to movies though, it was all smooth sailing.

Conclusions

IMDB ratings can be predicted with a MSE of ~ .7 using only the two predictors of total number of votes as well as duration of the movie.

This is very interesting, as it shows that IMDB ratings are somewhat predictable. It’s a scale of 1-10 so being able to predict within 1 point of the rating was not something I was expecting.

References

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv