Streaming Service Movie Rater

Chris Shin, Dylan Clark-Boucher, Grace Bosma

Due: 12/22/2020

Objective

The purpose of this project was to both model and visualize IMDb ratings of movies and films currently available for streaming on various online platforms

Introduction

According to Statista, approximently 74% of homes have access to at least one streaming service in 2019 compared to about 52% in 2015. Additionally, as a result of the COVID-19 pandemic, many consumers have either purchased an additional streaming service subscription or increased time spent streaming. It is clear that movie and TV streaming services are increasing in popularity and are integrated into many of our day to day routines. Just last year, the movie and television streaming industry was valued $42.60 billion USD. Information on the popularity of movies available for streaming is critical to streaming platforms like Netflix, Disney+, Hulu to maintain customer interest in their platform rather than their competitors.

Methods

The data for this analysis was obtained from either Kaggle or International Movie Database (IMDb) websites. Data regarding streaming service platform was obtained from kaggle whereas information on ratings and actors involved were obtained from IMDb.

The Kaggle Movie data set contained information on 16,631 movie titles, had 16744 observations and 17 variables including XXX. During cleaning, movies with XXX were excluded. The final kaggle data set consisted of XXX instances

Actors and rating information was obtained from IMDb in two separate data sets. The first, containing information on ratings, contained information on XXX films.

The second IMBd data set contained information on actors and held information on 6,414,148 movie production team members, four "Known For" film names associated with each actor or actress, and additional roles the actor or actress held (i.e. writer, producer, ...). During cleaning, those that did not hold an actor/actress role or had no "known for" titles were excluded from analyses. The final, cleaned IMBd actors dataset held information on 2,426,996 actors or actresses.

- revisit include info on final merge. include info on final size and final variables
- include something about any new variables created?

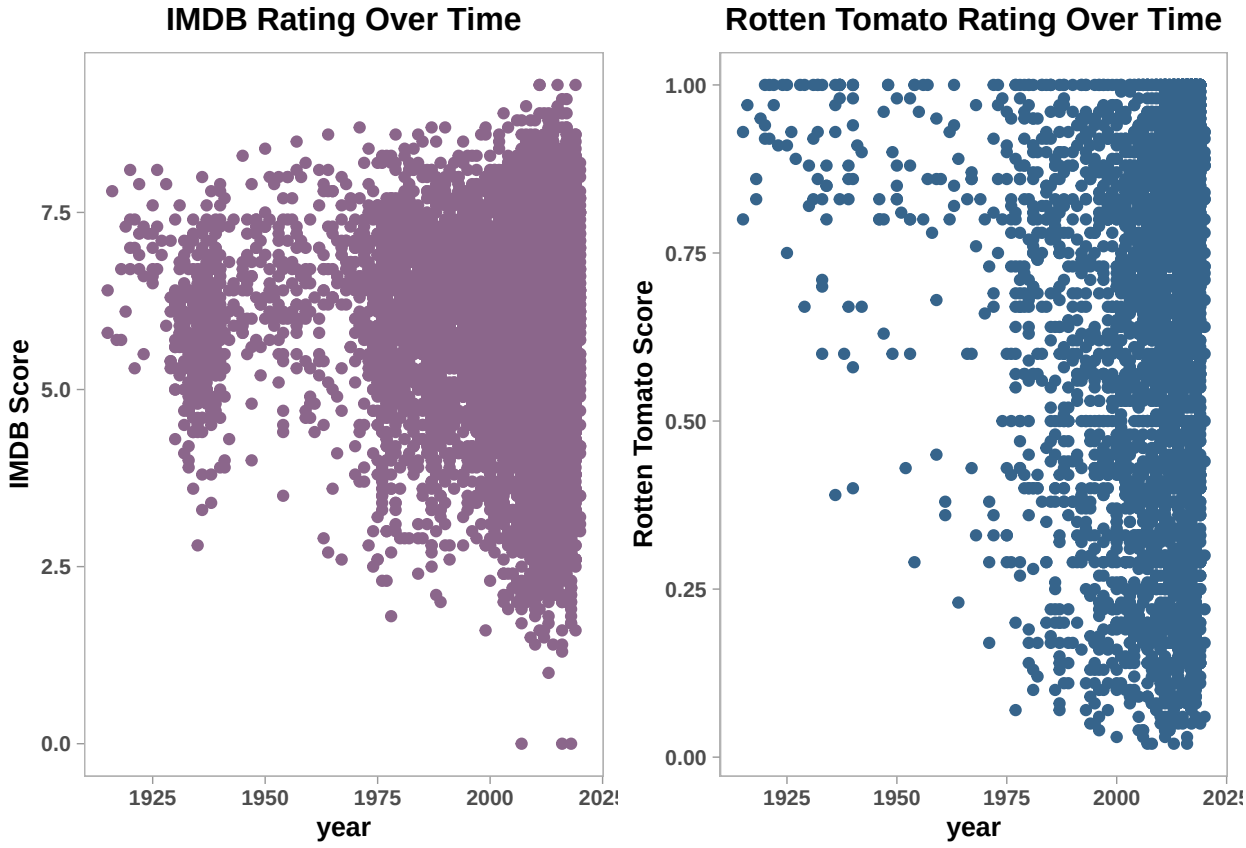
Prediction

Based on ratings data mentioned above, a flexible model was constructed for predicting overall ratings on a scale of 0 to 5 stars. These predictions depended on a user-defined set of covariates, including such factors as

a production's streaming service(s), target age group, release year, Rotten Tomato's Average Tomatometer score, and others.

- lm model

Vizualization



- not married to these plots just figured we should show something

RShiny Application

An RShiny Application was created using our data set to display our results and can be accessed at: `https://dylanclark-boucher.shinyapps.io/movie_rater_shiny/`

There are four tabs in the upper left that allows you to navigate through the application; Ratings Filter Tool, Ratings by Category, Score Prediction, and About. The Ratings filter tool presents a histogram and box plot of movie ratings and allows the user to filter based on recommended age of target audience, streaming service, Rotten Tomato Score, and release year and select a preferred choice of graph color and bin width. The Ratings By Category page presents three tabs on the left side bar. The first, Streaming Service, presents the distributions of movie ratings for each streaming service in a box plot format. The second, Age Rating, presents the distributions of movie ratings for by age group in a box plot format. The third, presents a scatter plot of IMBd movie ratings plotted against Tomatometer Score. Below the scatter plot the user is again able to filter but age group and streaming service. THe third tab, Score Prediction, [FILL IN LATER XXX]. The final tab, About, briefly summarized the background, app, model and contributors.

- potentially integrate an example here?

Discussion

- conclusionfrom lm model
  - what variables are important
  - did we predict well
- future direction

Contributions

Our group collaborated well and distributed work evenly among all members. While all members participated in searching for potential data sets, organizing, drafting and proofreading, Dylan created the RShiny app and organized a shared folder on the biostat computing cluster, Chris cleaned the Kaggle and movie rating data sets and Grace cleaned the actors data set and created a "cast score" to be used in model.

- be sure to update and include model