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

Background

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

Data Cleaning (see cleaning_data.Rmd for details)

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

The Kaggle Movie dataset was downloaded on December 4th, 2020. This dataset contained information on movies available on various streaming services as of May 22, 2020 from US locations. In the dataset there were 16,631 movie titles (entries) and 16 variables including imdb_id (IMDb-specific id), title, year (of release), age_rating (7+, 13+, 16+, 18+, or all), imdb_score (rating from IMDb, which is aggregated from IMDb registered users who cast a vote from 1 to 10 per title), rt_score (Tomatometer score from Rotten Tomatoes), netflix (whether it was available on Netflix), hulu (whether it was available on Hulu), amazonprime (whether it was available on Amazon Prime Video), disneyplus (whether it was available on Disney+), type (whether it was a movie or not), directors, genres, country (of origin), languages (in which the movie is available, either voice dubbed or subtitles), and runtime (length of movie). During cleaning, movies that did not match to titles on IMDb were excluded. The final movie dataset consisted of 11448 titles.

Data made publicly available on the website IMDb is updated daily and available for download. The dataset "name.basics.tsv.gz" was downloaded from IMDb on 12/4/2020 and used in this project. This dataset 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, director). 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.

Using the actors dataset, we created three variables specific to each actor/actress to use to describe movie titles. Score1 takes into account the number of actors/actresses involved in the titles the specific actor/actress was known for, score2 considers the number of titles and additional roles the actor/actress took on, and score3 prioritizes producers, directors, and writers, respectively.

Data Merging

We merged the movies dataset from Kaggle and the actors dataset from IMDb using the unique imdb_id associated with each title. The final merged dataset contained 139,772 observations. See merging_data.Rmd for details.

Prediction

We attempted two different algorithms for predicting IMDb scores: linear regressions and random forests. For each of these, we considered a pool of covariates that included hulu, disneyplus, netflix, amazonprime, year, rt_score, actor_score1, actor_score2, and actor_score3. Other variables might have been predictive as well, such as runtime, genre, or country, but would have required extensive cleaning or correction to be of use. In order to meet the current time constraints, we chose to omit these covariates from our analysis.

##Linear Model Results for our linear models are included in the Shiny app (described below). The model is built to be flexible, such that the user can specify which variables they would prefer to include or omit, and app will produce estimates and p values for those particular predictors. We found that netflix, amazonprime, hulu, and disneyplus were all statistically significant on their own, indicating that IMDb ratings do differ by streaming service. Tomatometer score is even more predictive, with high values corresponding to similarly high IMDb scores. Although critical reviews often differ from audience ratings, those differences appear to be quite small by and large, since Tomatometer score and IMDb rating show a strong, positive correlation. Please view the Shiny app for specific estimates and p values.

Random Forest

Random forest is a method of fitting multiple decision trees using a technique called bagging, which introduces randomness into each tree building process and averages the results across multiple trees. Random forest can be used for regression and classification tasks but requires complete datasets. We chose to implement random forest to observe the predictive ability of variables we chose to include in the linear regression models. Random forest is ideal for our situation because our final dataset is not too large.

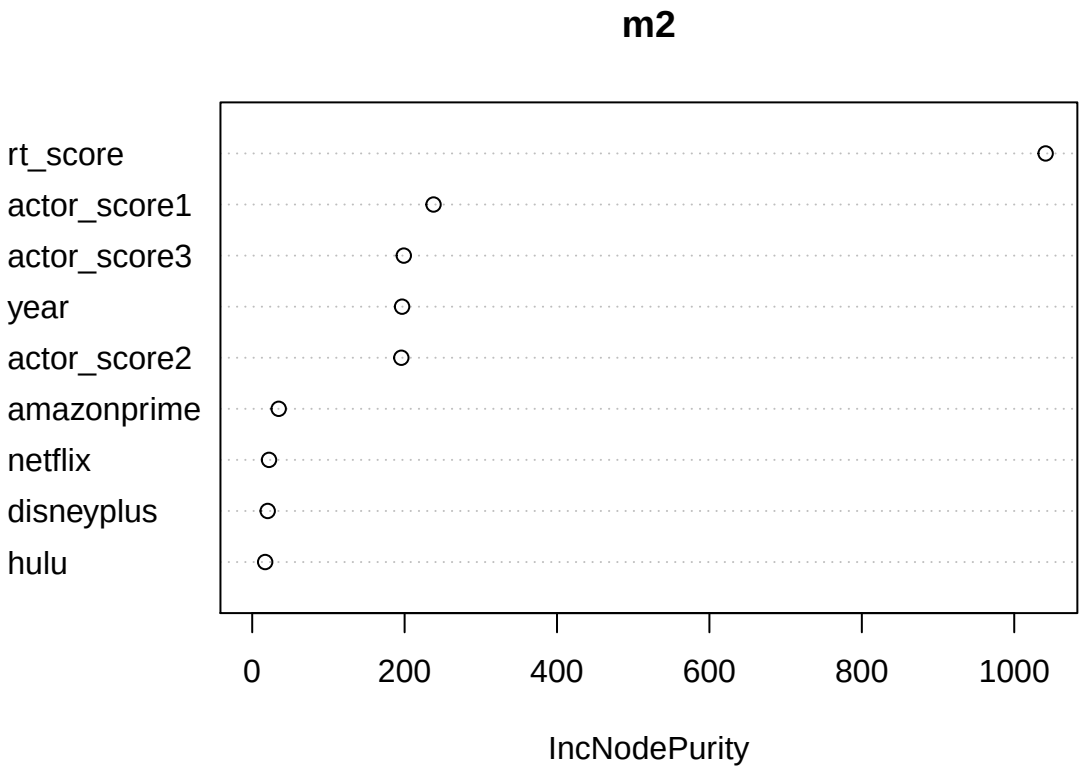
```
load("lm_data.rda")
library (randomForest)

set.seed(123)

model_form <- imdb_score ~ rt_score + year + disneyplus + hulu +
  amazonprime + netflix + actor_score1 + actor_score2 + actor_score3

m1 <- randomForest(
  formula = model_form,
  data = lm_data[complete.cases(lm_data),]
```

```
)
which.min(m1$mse)
```

```
## [1] 452
```

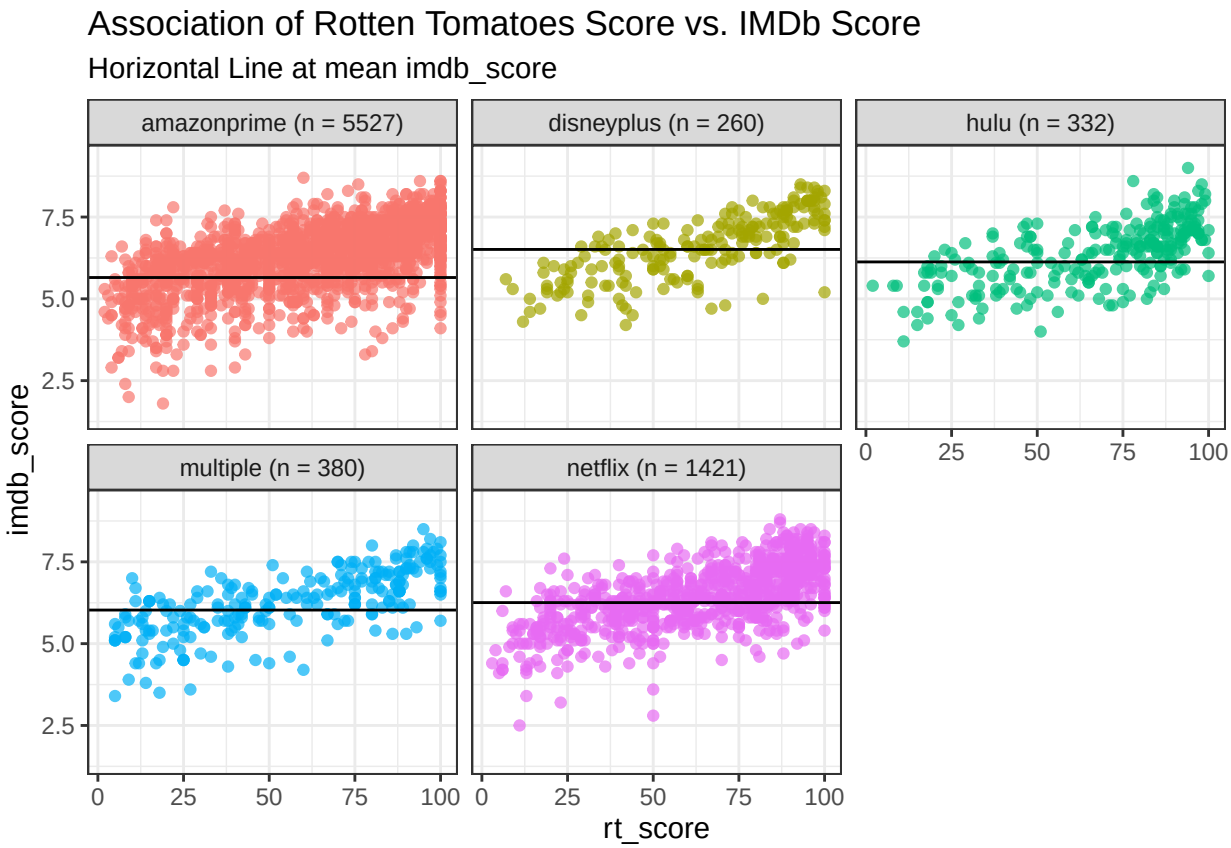
The default forest used 500 trees and chose 3 predictor variables at each split. When we plot the model, we see that we obtain the lowest error rate with 452 trees.

Because our goal for the random forest was to determine important variables, we did not continue with tuning the random forest and chose to look at the variable importance plot using a slightly modified random forest.



From this plot we see that rt_score (Tomatometer score from Rotten Tomatoes) is highly influential in determining imdb_score (rating from IMDb), followed by actor_score1, year (of release), actor_score2, and actor_score3. Availability on streaming service was not found to be highly influential in determining IMDb rating.

Vizualization



When we plot rt_score vs. imdb_score, we see that there is a positive association between the two across all platforms.

RShiny Application

An RShiny Application was created using our dataset 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 number of bins. The second tab, Ratings By Category, presents three sections for visualizing IMDb scores across different variables. The user can see ratings distributions for each streaming platform and each age group, as well as a scatterplot of IMDb ratings by Tomatometer scores. The third tab, Score Prediction, shows the coefficient estimates of a linear model for predicting IMDb scores based on streaming platform. The user can additionally select other variables to be included as coviariates, such as release year, Tomatometer score, age rating, and cast score. The final tab, About, briefly summarizes the background, application, model, and contributors.

Discussion

As mentioned in the prediction section, the TOMATOMETER score is highly influential in determining the rating from IMDb as shown by the RandomForest model. Actor score 1, year of release, actor score 2 and actor score 3 followed in prediction capabilities. In the linear regression model, a fairly similar conclusion was reached. TOMATOMETER score remains to be influential and significant as well as age groups and release year. If each actor score is added to the linear model individually, we see that actor score 1 and actor score 2 appear to be significant but have low magnitude estimates. Going forward, it may be interesting to consider the impact of variables like genre or country of release could have on ratings. Similarly, it may be interesting to explore the relationship of the cast outside of actors, for example, roles like directors, writers, producers could have on the ratings of a particular movie.

Contributions

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