# ACMS 40790 - Spring 2025

Aoife Mannion and Lissette Hernandez

April 28, 2025

## 1 Introduction

IMDb ratings are user-submitted ratings, offering a diverse range of perspectives on public opinion. However, Rotten Tomatoes focus on reviews from professional film critics. Essentially, Rotten Tomatoes’ critic scores can be more informative while IMDb ratings are more useful overall.

In this project, we analyze four different streaming services to determine which one is the best and determine if you can predict the IMDb ratings for a TV show.

## 2 Dataset

For this project, we are using data from TV shows found on Netflix, Prime Video, Hulu and Disney+. This data was obtained from Kaggle using public information from each streaming service. We chose this dataset because it contains IMDb ratings, these ratings are important because they give a better insight on the opinions of the general public. First, we took our original dataset and sorted it in descending order based on IMDb ratings. The issue with this was that since there were shows that aren’t well-known, they could have few reviews and a biased rating leading the IMDb ratings to be inaccurate. Later, we saw that the dataset could be grouped by streaming service and by the popularity of shows within each streaming service this way we could see well-known shows at the top and unheard shows at the end. These steps are important to note to perform our data clean up.

This dataset consists of two main data frames. Cleaned up versions are displayed below:

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Top Shows in Cleaned Dataset (Highest IMDb) df

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: TV Shows with no Rotten Tomatoes Rating df

## 3 Data Cleanup Performed

To keep only relevant shows, we deleted the last 40% of TV shows from each streaming service when the dataset was sorted by index. We dropped rows that had null values for both columns of IMDb and Rotten Tomatoes. There was also a column named “Unnamed: 0”, which was identical to the index, so we dropped it as well and dropped a column named “type” that just contained the number 1 for every show. Additionally, we renamed two of the columns to make them more specific to what we were trying to find; “Year” to “Release Year” and “Age” to “Age Rating.” Finally, for the “Age Rating” and “Rotten Tomatoes” columns since they were initially strings, we changed them to floats to better perform calculations. This was done by removing the plus and percentage signs and for “Age Rating” we changed the output “all” to “0.”

## 4 Numerical Analysis

Once our data clean-up was finalized, we went from 5611 TV show titles to 3316 TV show titles. Later, we filtered out shows with an IMDb rating of less than 3 to avoid outliers dragging down the mean. From this we saw Netflix had the 2nd greatest number of shows with an IMDb rating higher than 3 and had the highest average IMDb rating.

## 5 Data Visualization

A screenshot of a computer

AI-generated content may be incorrect.The first plot is a boxplot that displays each streaming service and its designated IMDb rating. We decided to overlay them with swarm plots to represent the number of shows each to visually see what streaming service had most of the highest rated TV shows. From this visualization, we can see that all the services have similar medians except for Disney+ and it also has significantly less shows than the other platforms. On the other hand, Prime Video and Netflix have the highest rated TV shows.

A screenshot of a computer

AI-generated content may be incorrect.Our second visualization was a bar graph that shows the Top 50 TV shows available per streaming service. From this graph, we can see that Netflix is the best streaming service by having the highest rated TV show.

## 6 Data Model and Prediction

A chart with red blue and white text

AI-generated content may be incorrect.Using a Correlation Heatmap, we determined if we could predict IMDb ratings. We started by selecting our numeric columns and dropping rows with missing values. After, we computed the correlation matrix and extracted only IMDb correlations. Once this was done, we plotted the IMDb correlation matrix into the heatmap as seen below. Since correlation matrices show linear relationships, we can see from our results that there isn’t a strong linear relationship between IMDb ratings and any of the other features.

A graph with a line

AI-generated content may be incorrect.To determine if we could predict an IMDb score based on other columns, we created a function using RandomForestRegressor that takes in different tree depths along with training and validation data to return the mean absolute error between the predictions and the actual values. This showed us that errors were minimized at a depth of 4, so we set that as our max depth.

Furthermore, we created a tree from the random forest that shows how Rotten Tomatoes, the squared errors, sample sizes, and values change as we progress through the different features. We see that 498 shows from the training data had an average IMDb rating of 7.573. The tree starts with Rotten Tomatoes, since this is our strongest predictor and the first way it splits the data is between shows that are above or below the Rotten Tomato value of 77.5. The other features used to split the data within the tree are age, rating, and release year.

A diagram of a structure

AI-generated content may be incorrect.

## 7 Conclusion

From our data analysis, we found Netflix to be the best streaming service and that the IMDb rating for a tv show can only be moderately predicted. After filtering out our data, creating a boxplot, and bar graph we visually could depict that Netflix had the highest rated TV show (Breaking Bad) with a rating of 9.5/10, almost half of the Top 50 TV shows (24/50), and had the highest average IMDb with a rating of 7.52/10. While the IMDb rating can only be moderately predicted for a tv show because after performing our regression we had an r-squared output of 0.35, meaning that only 35% of the variation in IMDb ratings is explained by our model. Additionally, we also obtained a mean absolute error of 0.47 indicating that a 0.47 difference in rating would not greatly affect a viewer’s decision of watching a show. While we successfully found the best streaming service and if an IMDb rating for a tv show could be predicted, we did face some limitations. The first one being that we only had a Rotten Tomatoes rating for 18% of the dataset and the second one being that our dataset did not contain the number of review or genre of each show. Having these components could have been useful for filtering out the unpopular shows and to determining the best streaming service based on genre.