Predicting Movie Ratings from Characteristics

Chloe Gorgen in collaboration with Keegan McDowell, TIngyan Wang, Jay Peeler, and Daisy Xue

August 30, 2021

Introduction

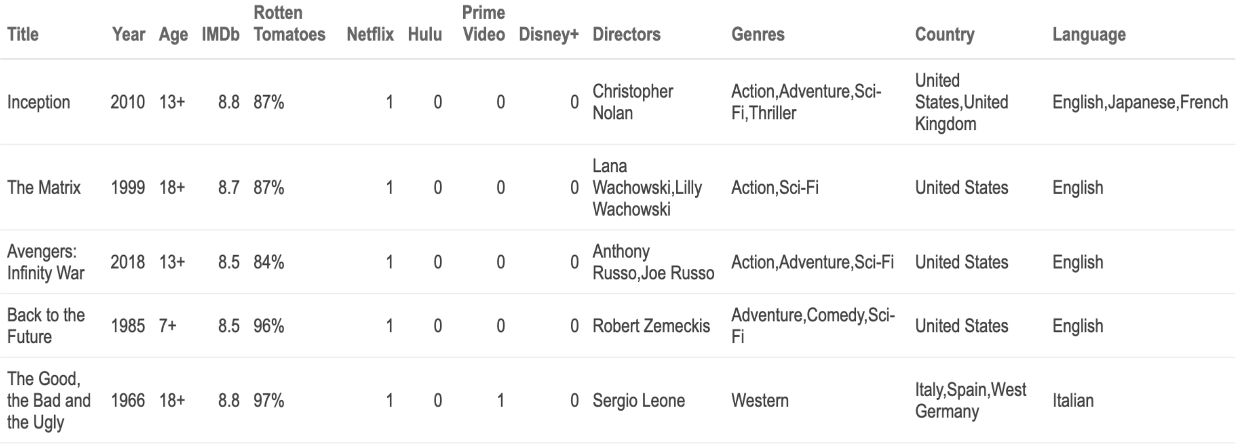
In order to maximize entertainment and minimize costs, it is important for consumers to choose the right streaming service that shows the highest quality films. This analysis aims to further understand how streaming services and critical ratings of movies are interconnected, and how these connections can allow for more strategic purchases on the consumer end. For the purpose of exploring these connections, the following tables and models attempt to answer the questions: What are the most influential factors that IMDb uses to determine their ratings of films? And how can we use these ratings to choose the best streaming service?

These questions were answered using movies currently available on the top four subscribed streaming services. For each title, we used IMDb rating as the metric for the quality of the film because of its standardized and quantified scale, as well as its wide acceptance among critics and audiences alike. After finding which of the streaming services has the highest quality selection of movies as decided by IMDb ratings, two predictive modelling techniques were used to attempt to formulate and predict IMDb rating. These followup investigations give insights to the most important factors in determining critical rating, and allow the consumer to decide whether or not they agree with IMDb, and therefore whether or not they agree with the findings of the initial investigation.

Understanding the way IMDb, one of the most widely accepted and sought out critical reviewing groups, creates their rating system allows the user to align their spending habits and interests by deciding which streaming services best suit their needs.

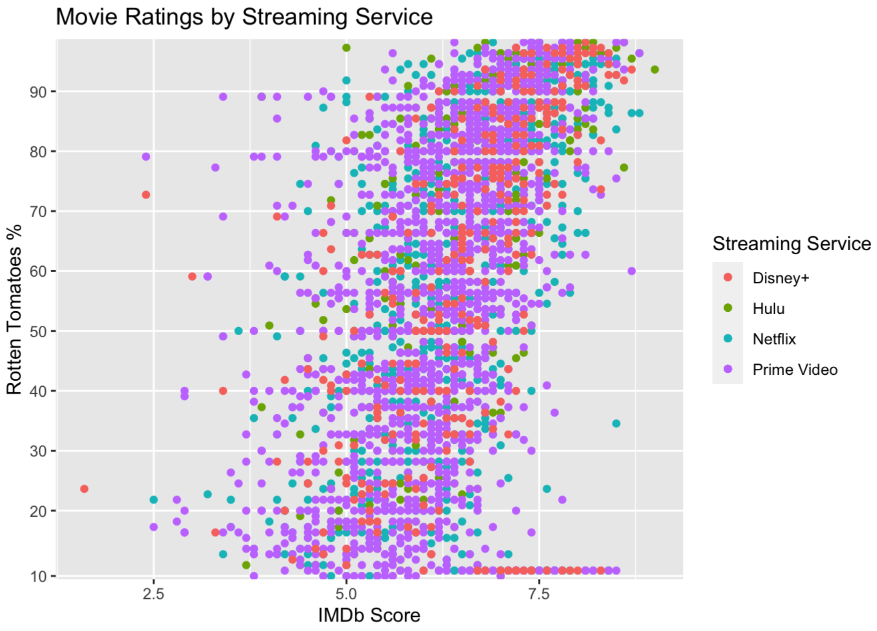
Data

The data set used for this analysis was web scraped in 2020 by Ruchi Bhatia, a data scientist from Mumbai, India. She created the data set with the intention of determining which platforms movies can be found on, as well as the average ratings for movies based on the country of production. It contains 16,774 unique movies that are included with a subscription to Netflix, Hulu, Prime Video, or Disney+. These specific platforms are currently the four most subscribed to streaming services in the US. The table below gives a look into the format of the data set we worked with.



Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Directors	Genres	Country	Language
Inception	2010	13+	8.8	87%	1	0	0	0	Christopher Nolan	Action,Adventure,Sci-Fi,Thriller	United States,United Kingdom	English,Japanese,French
The Matrix	1999	18+	8.7	87%	1	0	0	0	Lana Wachowski,Lilly Wachowski	Action,Sci-Fi	United States	English
Avengers: Infinity War	2018	13+	8.5	84%	1	0	0	0	Anthony Russo,Joe Russo	Action,Adventure,Sci-Fi	United States	English
Back to the Future	1985	7+	8.5	96%	1	0	0	0	Robert Zemeckis	Adventure,Comedy,Sci-Fi	United States	English
The Good, the Bad and the Ugly	1966	18+	8.8	97%	1	0	1	0	Sergio Leone	Western	Italy,Spain,West Germany	Italian

For each movie title, the data set indicates the platform(s) it is on, as well as the average audience ratings from both IMDb and Rotten Tomatoes. The streaming service that offers the movie is designated by a 1 in the respective column, otherwise a 0 will be present. The IMDb ratings are on a scale from 0 to 10 and rounded to the nearest tenth, while the Rotten Tomatoes score is a percentage rounded to the nearest integer. The data set also contains several other characteristics of each movie. Year is a numeric variable that indicates the year each movie was produced, not when it was added to the streaming service. Age is a categorical variable showing recommended minimum age to watch the movie based off the Movie Picture Association film rating system. This is the same system that categorizes films into the widely accepted categories of G, PG, PG-13, and R. The options for age are all, 7+, 13+, and 18+ in alignment with these ratings. The Directors variable lists the director(s) credited with the film, each separated by a comma in the case of multiple directors. Similarly, the Genres variable indicates the genre(s) of the movie, once again separated by a comma. The same goes for Country and Language, which list the countries and languages that the movies are available in.



By visualizing the critical ratings against each other in this manner, we can see that these groups rate films in a similar way, but these rating systems are not the same. In order to find out which rating we as the consumer trust more, we need to understand the factors that lend to these rankings. Further, after grouping these points by streaming service, it is difficult to determine if any streaming service has an objectively better movie selection. It is therefore up to the consumer to place themselves within these rating systems and decide for themselves what they care about and which streaming service is best catered to their needs.

Results

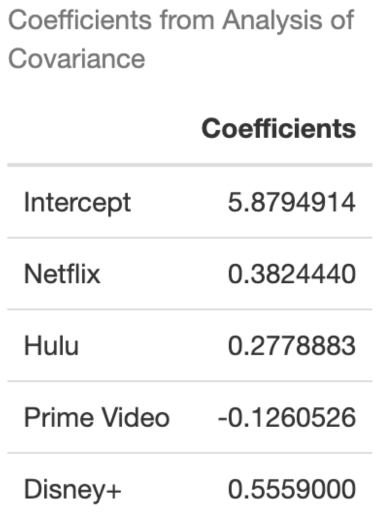
The results of this analysis first attempt to find the streaming service with the best selection of movies, with the quality of movies measured by the IMDb rating system. This portion of the investigation is carried out by completing an Analysis of Covariance, which is able to create regression coefficients associated with the impact of availability on each streaming service on a movie's IMDb rating. Then, Lasso Regression and Random Forest Regression are used to attempt to predict IMDb ratings of movies from the other available characteristics. These predictive modelling techniques are useful to determine which factors are most impactful on a movie's critical rating and how standardized the rating system is. These two insights can then determine how much a consumer trusts IMDb, and therefore how reliable the results of the Analysis of Covariance is.

Analysis of Covariance

The first modeling technique is an Analysis of Covariance, designed to show the relationship between IMDb and the availability of movies on Netflix, Hulu, Amazon Prime Video, and Disney+.

	Netflix	Hulu	Prime Video	Disney+
Intercept	5.8080669	5.8890125	6.2708558	5.8833248
Streaming Service Coefficient	0.4448965	0.2491041	-0.4999455	0.5580606

The table above shows all of the linear coefficients calculated from an Analysis of Variance between each of the streaming services and IMDb rating. These calculations were produced after initial linear models were generated and the significance of each of the potential variables was tested. All four of the streaming services proved to be significant. Therefore, the above table can give an indication of the influence that availability of a movie on each of the four streaming services has individually on IMDb rating. However, this is an incomplete picture of the potential relationships between streaming services and movie rating because most of the movies in the set were included on multiple streaming platforms. In order to account for interactions between these streaming services, an Analysis of Covariance was conducted, the results of which are shown in the table below.



Coefficients from Analysis of Covariance

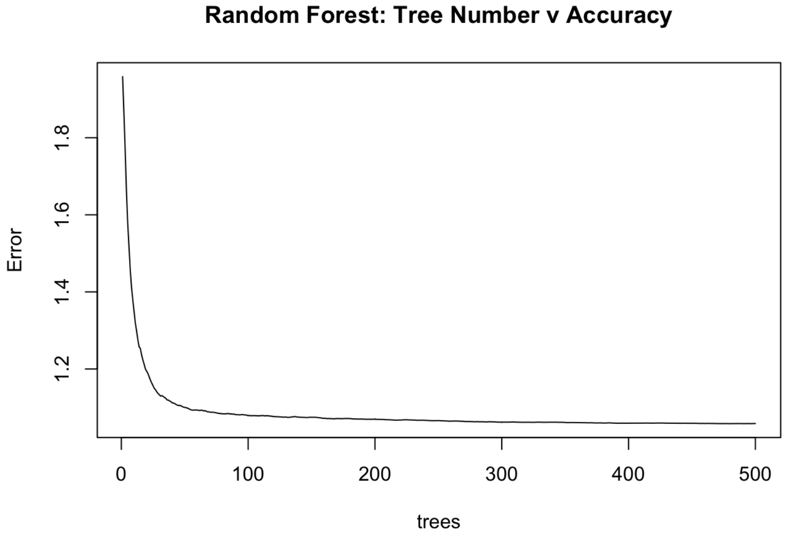
	Coefficients
Intercept	5.8794914
Netflix	0.3824440
Hulu	0.2778883
Prime Video	-0.1260526
Disney+	0.5559000

Our Analysis of Covariance resulted in the following coefficients and intercept. From these results, we can see that Disney+ has the largest positive impact on IMDb, followed by Netflix and Hulu. Amazon Prime Video had the only negative impact. Pragmatically, this means that Disney+ has the highest average critical rating, making it the best investment for a consumer wanting the highest rated selection of movies. It is worth noting that Amazon Prime Video has a significantly higher number of movies available through their subscription, which could explain the negative coefficient.

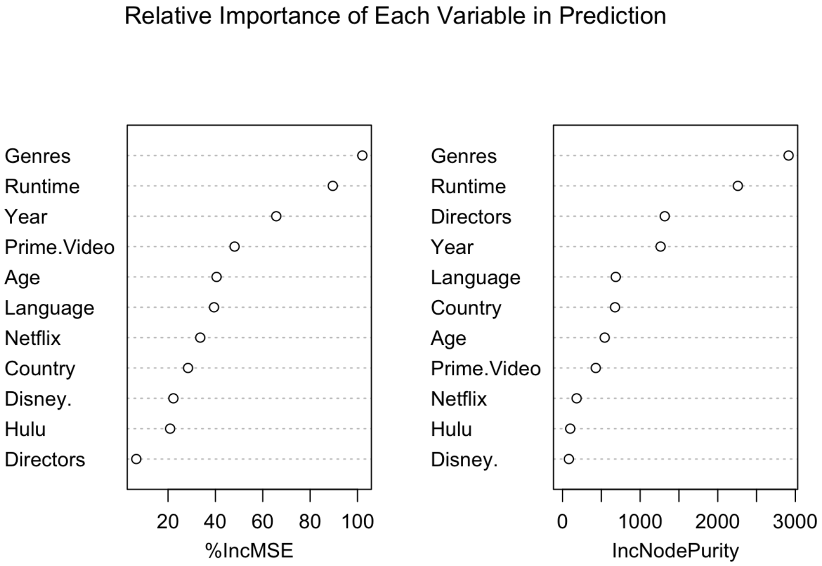
In order to be able to use the results of this model, we have to understand the factors that lead to an IMDb score. The rest of the analysis will attempt to use predictive modelling to predict IMDb score from the characteristics of movies. From this information, consumers will be able to decide if they align with the IMDb rating system.

Prediction of IMDb Rating

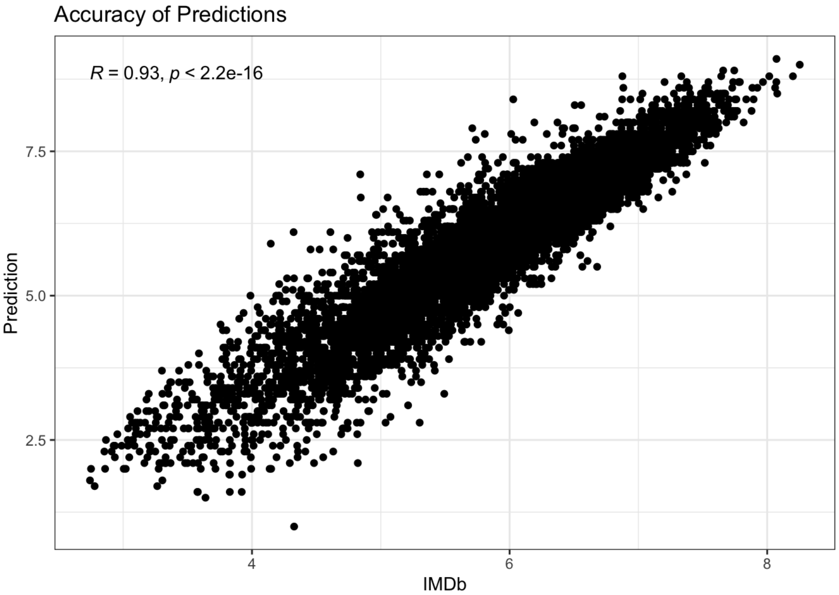
Random Forest Regression



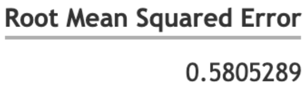
The graph above illustrates the error of predictions graphed against the number of trees used in generating that prediction. Random Forest Regression aims to find the most accurate model of prediction by generating and comparing a multitude of decision trees. The more trees added to the model, the more accurate it will be but the more overly complicated and over-fit it can become. The number of trees used in our prediction was generated to be close to the horizontal asymptote of this graph, such that the prediction was as accurate as possible but not overly complicated.



The plot above illustrates the relative importance of each variable, which was used in the determination of coefficients to provide the most accurate model. Focusing on the percentage increased in mean squared error, it's clear to see that Genres has the highest variable importance, followed by Runtime and Year, resulting in higher predictive power in the analysis. Removing these variables of high importance generates significant loss of accuracy on the prediction results. On the other hand, looking at the table of increased node purity, we see some differences compared to percentage increase in mean square error, which is one of the results of measuring the node impurity. Splits of the data set based on each variable is biased towards variables with many classes, which also biases the importance measure. In our case, the bias results in different orders regarding the importance of variables. However, Genres and Runtime have the highest importance in both ways.



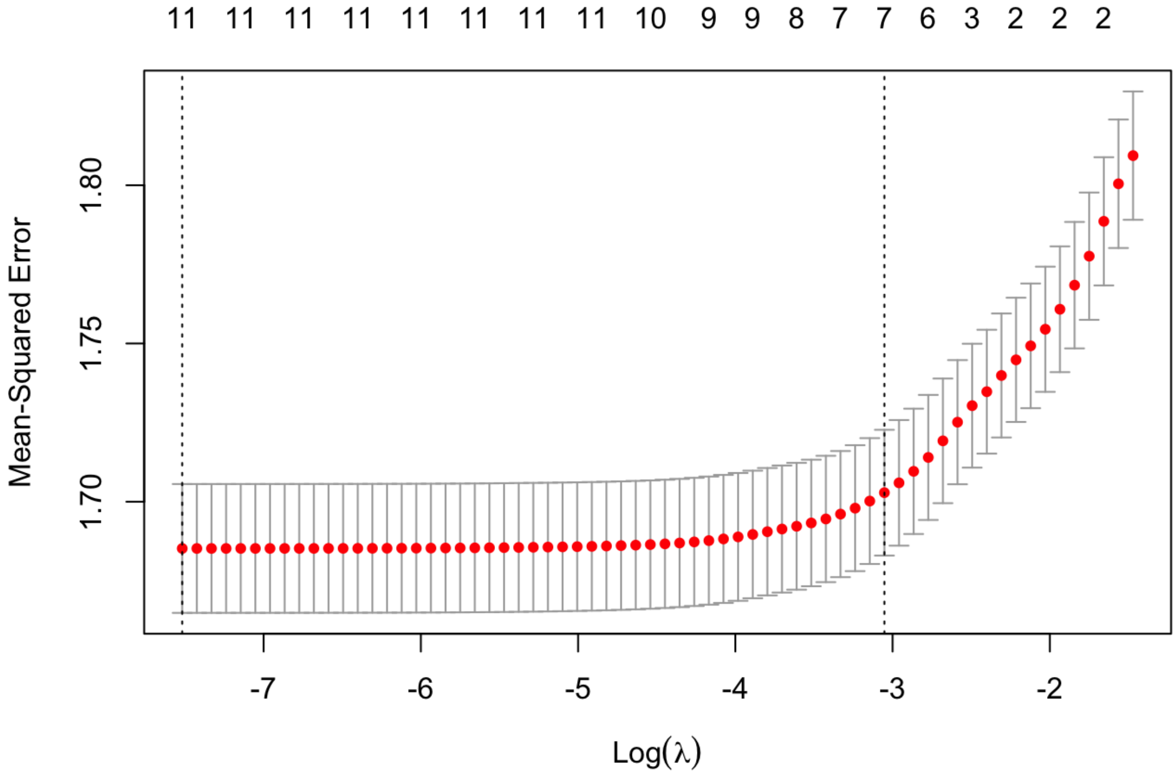
The figure above illustrates the accuracy of the Random Forest Predictions against the actual IMDb ratings for each title, as well as the R and p-values of the model. The shape of the points shows a positive correlation between the prediction and observed IMDb rating, roughly resembling a straight line which indicates a highly accurate model. This is reflected in the R value, the p-value, and the Root Mean Squared Error. These results indicate that the relative importance of each coefficient measured by Random Forest Regression is an accurate prediction of how IMDb is measured.



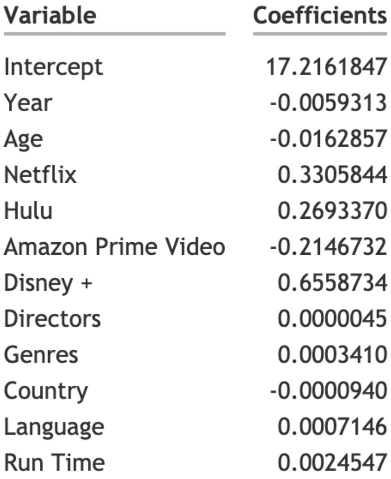
Root Mean Squared Error
0.5805289

The Root Mean Squared Error indicates that Random Forest is a very accurate tool to test the relative importance of each coefficient in predicting IMDb score. The coefficients listed in order of importance by this model are Genre, Runtime, Year, Prime.Video, Age, Language, Netflix, Country, Disney, Hulu, and Directors.

Lasso Regression



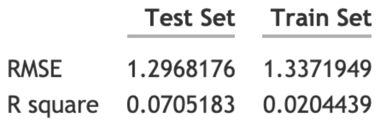
The plot above shows, with error margins, the Mean Squared Error of each Regression model against the Lambda used in creating that prediction. The goal of Lasso Regression is to find and use the Lambda value that minimizes error without making the model over-fit to the data. The best lambda calculated by the model and used in our analysis is the one which minimizes error, which appears to be along the horizontal asymptote.



Variable	Coefficients
Intercept	17.2161847
Year	-0.0059313
Age	-0.0162857
Netflix	0.3305844
Hulu	0.2693370
Amazon Prime Video	-0.2146732
Disney +	0.6558734
Directors	0.0000045
Genres	0.0003410
Country	-0.0000940
Language	0.0007146
Run Time	0.0024547

This table displays the coefficients of the regression model generated to predict IMDb. None of the coefficients are zero or close to zero, indicating that the Lasso method found each of the variables to be important in predicting a movie's IMDb rating. The highest coefficients are Disney+, Language, Hulu, and Netflix. These coefficients indicate that movies that are available on these platforms and that are available in many languages are the most popular in terms of IMDb. The lowest coefficients are Year and Age. These coefficients indicate that more popular movies are generally older, and movies that do not have age restrictions based on content.

Certain values in this are seemingly illogical in the scope of the analysis. For example, the intercept of the line is higher than the maximum IMDb score of 10. However, this value is explained by the Year variable, for which there is a minimum non-zero value (minimum Year is 1939). The unscaled year variable, and resulting intercept, are valuable to this analysis because it shows how much IMDb rating changes on average by year.



	Test Set	Train Set
RMSE	1.2968176	1.3371949
R square	0.0705183	0.0204439

In the calculation of Root Mean Squared Error and R-squared values, it is worth noting that the R-squared value for both the training set and the testing set is close to 0. This means that Lasso Regression was able to quantify the relationship between each variable and the IMDb rating, but these coefficients were not able to accurately predict IMDb ratings, and that the relative importance calculated by the Random Forest Regression may be more appropriate for comparison. In future analysis, in order for Lasso Regression to be more accurate as a predictive tool, the year variable should either be scaled or adjusted (for example by subtracting the minimum value from each of the values).

Explanation of Results

The results from the Random Forest Regression to predict IMDb ratings showed that the factors of each movie in order of importance are Genre, Runtime, Year, Amazon Prime Video, Age, Language, Netflix, Country, Disney +, Hulu, and Directors. On the other hand, the results from the Lasso Regression Model calculated the most important characteristics as Disney+, Netflix, Hulu, Amazon Prime Video, Year, Runtime, Language, Genre, Country, and Directors. The Lasso Regression Model added to our understanding of the influence of each characteristic in rating by showing the polarity of each coefficient. In this model, Year, and Amazon Prime Video were negative, showing that newer movies and the availability of movies on Amazon Prime Video are negatively related to IMDb.

The Root Mean Squared Error for each of the models we developed were low, but the R-square values for the Lasso Regression models we developed were close to zero. This indicates that while we are able to claim the relative importance of movie characteristics for each rating system within our set, the values of the coefficients we found for each model were poor predictive tools for predicting critical ratings of other films. However, we can determine that Disney+ has the highest positive correlation with highly rated movies, meaning that this streaming service may be a better investment for consumers who trust IMDb.

Conclusion

The purpose of this project was to gain insights of film rating systems in order to allow for more strategic content sales and purchases in the consumer end. OUr initial investigation into which streaming service has the best selection of movies found that availability of a movie on Disney+ has the highest positive relationship on IMDb calculation, meaning that among the four streaming services tested, Disney+ has the best rated selection of movies. Next, the predictive models provided insight into how IMDb is calculated, and which factors have the greatest influence on these values. These findings suggest that almost all film characteristics involved in our analysis played some role in predicting the IMDb ratings, with the exception of Directors. Using the Lasso Regression method, we found that film availability on Disney+, languages the film is available in, and availability of a movie on Hulu and Netflix were the most correlated with higher IMDb ratings. From these results, it can be concluded that Disney+, Netflix, and Hulu are better at selecting movies based on the same criteria as IMDb. Conversely, we found that year of production and age restrictions were least effective predictors of IMDb rating. With this in mind, the private sector can use this information to develop better strategies for maximizing content rating. On the consumer end, this information lets them know that Disney+, Hulu, and Netflix, have the highest rated content based on the other criteria weighted in IMDb scores.

These findings let us know that the highest rated content is typically most widely available to the public. Although the focus of this investigation was focused on informing consumer decisions, the scope of this analysis could be adjusted to benefit both the consumer and the content creation and marketing end. Further analysis should pay special attention to newer films and using this information to not only predict ratings but also predict revenue generated by newer movies. This could potentially be used by writers, marketers, and directors to improve profit margins in future films and television shows. Content creators can also use the results of this and further investigation to determine if being featured on Netflix will increase their ratings or rather being featured is simply a byproduct of high-quality content. In predicting Rotten Tomatoes scores, we found that the most influential factors were the film characteristics of genre, year produced, and run time, and that the streaming service that was most well equipped to meet these criteria was Amazon Prime Video. Therefore, if a consumer feels more aligned to the ratings of IMDb, it is a better investment to subscribe to Disney +, but if they feel more aligned with Rotten Tomatoes ratings, they should instead pay for Amazon Prime.

In the future, other statisticians can build upon this work by identifying and testing more specific variables such as how actors, franchise, production company, etc affect average critical ratings. There is also room to further investigate which demographics use which streaming services the most for the purpose of better preparing companies to reach their target audience more efficiently. If they can reach more specific demographics, companies can tailor their content in a fashion that allows for more personalized content on the consumer end. This kind of work could potentially be incorporated in future algorithms for predicting “for you” tabs on the front pages of streaming platforms. In order to complete any of this research, however, we would need access to more specific data sets regarding revenue and actor presence in content. Although difficult, such research would be to create a more harmonious relationship between consumers and producers alike.