Rotten Tomatoes vs. Audience

Caedon Ott, Jacob Shankles, Jay Jemila

STAT 472

Dr. Aaron Nielsen

May 2023

ABSTRACT

Using Rotten Tomatoes data, we investigated the factors that contribute to differences in audience and critic ratings, predicting classifications for audience higher versus critic higher ratings, and predicting the differences between audience and critic ratings. Our data was obtained from Kaggle, where a user scraped factors like genre, movie descriptions, ratings, and any other information that is provided for a movie when you would search it on Rotten Tomatoes. We used lasso regression to find the relevant factors, linear regression and XGboost for predicting the differences between audience and critic scores, and KNN for classifying if a movie had a higher audience or critic score. The prediction for classifying audience higher versus critic higher ratings was better than random chance, and the lasso regression found that genre was the most informative indicator of whether or not a movie would be audience or critic favored.

INTRODUCTION

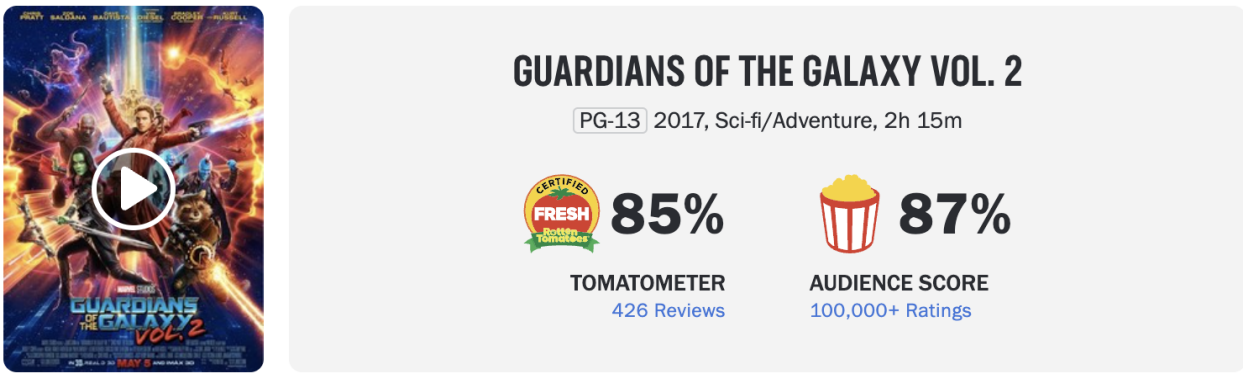
Movies are a beloved part of our culture. They come in many different flavors, whether that be something like a high budget blockbuster, or a low budget horror. With so much variety, there is also volume. Hundreds of movies are released every year, which forces people to choose which movies they will decide to watch. With all of the movies available to watch, how do they choose?

One common way of assessing a movie's quality is through online reviews, which are typically broken up into two categories: critics, and the audience. Through these reviews people can get a sense of how well a movie was received. Often, these ratings can be very similar, which conveys that the audience and critics feel the same. Conversely, there are movies that have large differences between these scores. Maybe an action movie had really cool visuals but poor storytelling, which led to the audience favoring the movie over critics. There are many reasons why there could be a difference between these scores, so we decided to look at some existing studies to illuminate which factors could be of importance.

In an exploratory study conducted by Rao Issan, through observing word clouds for audience and critics, he noticed that there were strong associations for genre for movie ratings. After performing a linear regression with ratings as the outcome and genre as the predictors, he found that comedy and kids & family were audience favored, and mystery, classics, and cult movies were critic favored.

For movie success, a study was conducted by two students from Stanford that used a subset selection to predict movie gross and movie ratings. In this study, they found that factors such as genre, movie description, and runtime were used to predict movie success and ratings with a success rate anywhere from 0.66-0.88.

Drawing inspiration from these two studies, we wanted to approach ratings from a different perspective. When looking at a generic movie description on Rotten Tomatoes, it looks something like this:



Scrolling down a bit, you can also find the general movie description.

MOVIE INFO

Peter Quill and his fellow Guardians are hired by a powerful alien race, the Sovereign, to protect their precious batteries from invaders. When it is discovered that Rocket has stolen the items they were sent to guard, the Sovereign dispatch their armada to search for vengeance. As the Guardians try to escape, the mystery of Peter's parentage is revealed.

From this information, can we identify what creates the differences between the critic score (tomatometer) and the audience score? If so, can we predict it? To answer this idea, we present three research questions:

- Can we predict the higher movie score between the general audience and movie critics?
- Can we predict the difference between audience and critics scores? (audience - critic)
- What are the factors that influence a decision in movie scores?

DATA

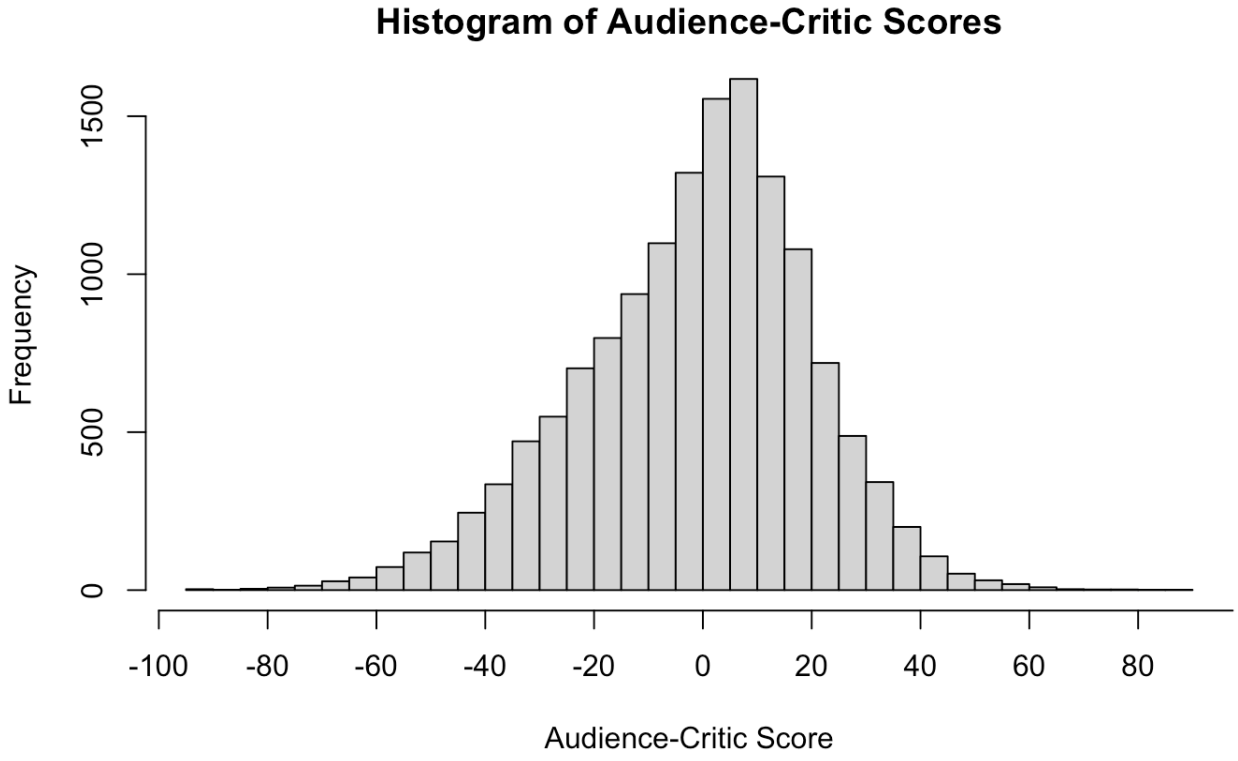
Conveniently, we were able to find a dataset on Kaggle from Stefano Leone that scraped information off of Rotten Tomatoes for 17,000+ movies, dating from the early 1900's all the way to 2022. Genres were grouped in a single category, so we looped through each row and split the genres into binary indicators of whether or not the movie was a part of that genre. Next, we created a classifier for audience higher and critic higher by performing a simple greater than less than operation and storing the results for each movie. Finally, any movie that had a missing value for content rating, audience score, critic score, runtime, production company, number of critics, number of audience, and movie description were dropped. There were not many missing values, and we are only interested in movies that have ratings for audience and critics, and enough of them where people are likely to be googling the movie. Thus, we felt alright in dropping a few movies.

The data was almost ready to be used. To analyze the movie description, some additional information would need to be extracted. Like the study conducted by Rao Issan, we were interested if certain words had associations for higher ratings between audience and critics. We looped through every movie, took its description, and ran the words through a library provided by the syuzhet package with sentiments attached to each word using the NRC word-emotion

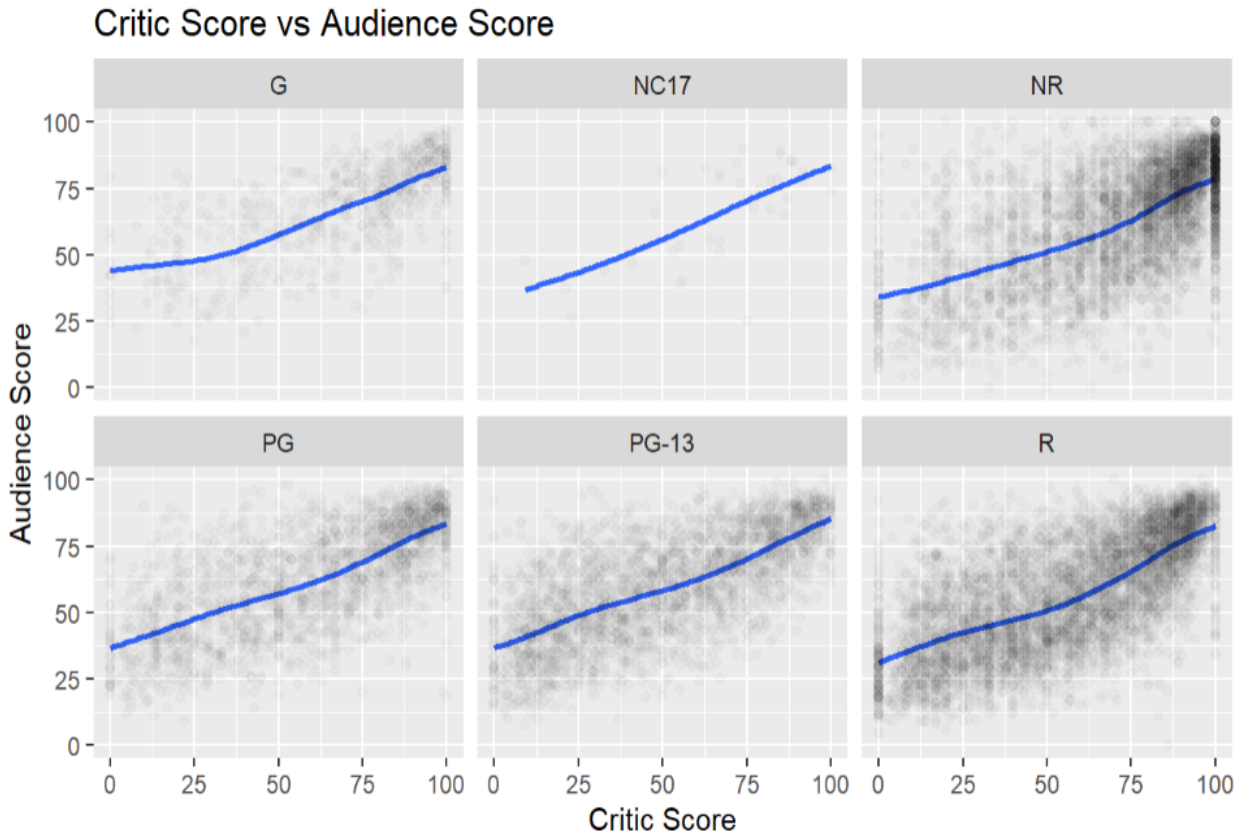
association lexicon. There were 10 sentiments; anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, and positive. Each word could take on a value from 0-3 for all 10 sentiments, with 0 being no association, and 3 being a strong association. We then took the sum of the overall scores across sentiments from every word, and stored each sentiment score in a separate column that pertained to the movie being analyzed (the row). This process took a few hours to complete, but once it finished, we were able to begin our analysis.

EXPLORATORY ANALYSIS

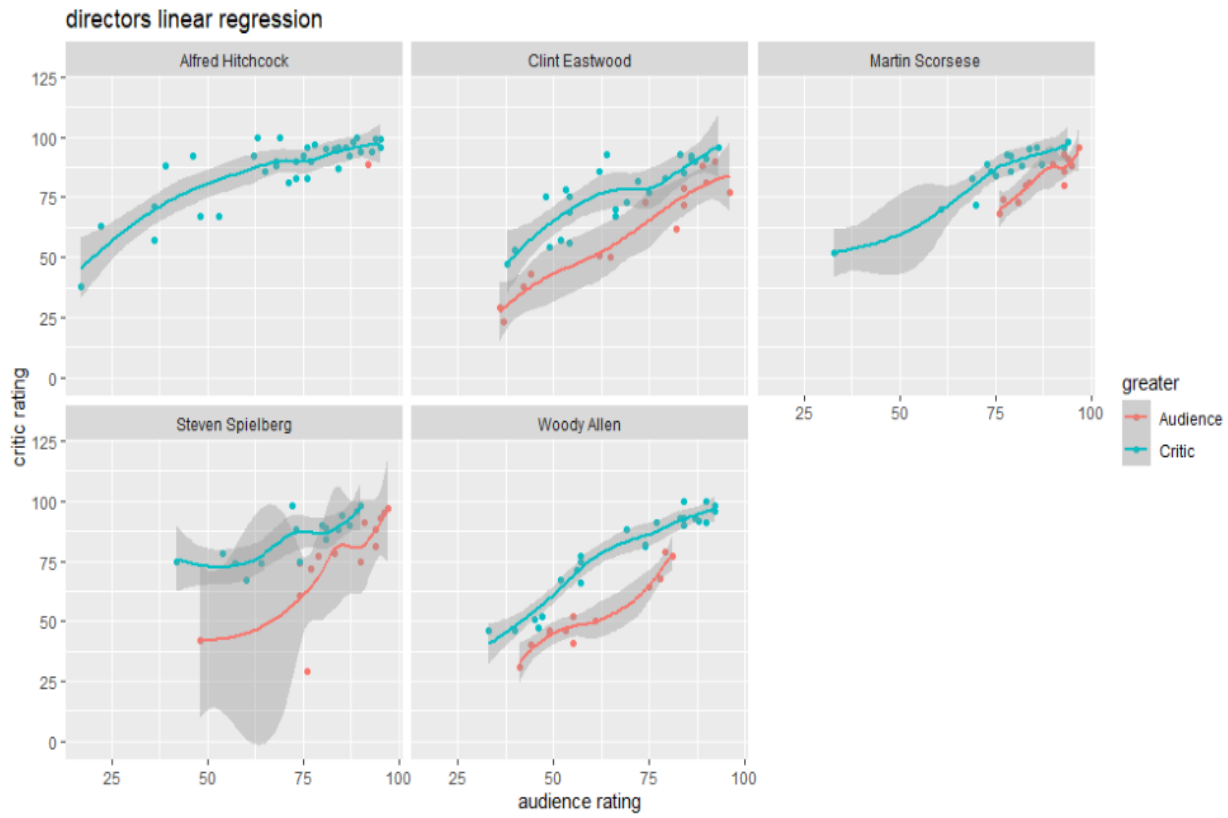
A histogram of the difference in the audience and critic scores shows us that the mean is centered around the audience having higher scores. This contrasts the fact that the frequency of the critics giving higher ratings in our data is higher compared to the audience.



When comparing the critic and audience score models based on the movies' content ratings, we see a similar trend in all of them, with the audience rating at about the same as the critics, though leaning a bit more towards the critics.



A linear model on the directors with the most movies in our dataset from the Directors variable shows us that critics tend to greatly favor them compared to the general audience; e.g. out of 36 of his movies, only 1 of director Alfred Hitchcock's movies was favored by the audience compared to critics - Dial M for Murder, a 1954 movie that became available for streaming in 2008.



METHODS AND RESULTS

a) Linear Regression

One of the models we fit was a linear regression model that used each genre as an indicator variable as well as categorical variables for 3 different runtimes and 6 content ratings. The first runtime category was short movies with runtime under 95 minutes which was roughly 5700 movies. The medium length movies were between 95 minutes and 120 minutes with roughly 8000 movies. The last category was long movies with any movie over 120 minutes which was a surprisingly small group at 2200 movies. The dependent variable for this model is the difference between the mean audience score and the mean critic score (audience - critic). The goal of this model was to further satisfy our third research objective which was to determine the factors that have the most influence on the difference between critic and audience scores. The intercept of this model takes on the long runtime category and the g rating category.

Predicting the Difference, Audience - Critic Scores

Predictor	Estimate	Std Error	t stat	p-value
Intercept	0.83	1.06	0.79	0.431
Midlength	-0.64	0.49	-1.32	0.187
Short Runtime	-2.63	0.54	-4.91	<0.001
NC-17	2.50	3.43	0.73	0.467
NR	-4.96	0.91	-5.44	<0.001
PG	3.96	0.92	4.32	<0.001
PG-13	10.16	0.96	10.57	<0.001
R	3.20	0.93	3.42	<0.001
Action & Adventure	2.46	0.42	5.89	<0.001
Art House International	-2.92	0.46	-6.34	<0.001
Classics	-5.69	0.57	-9.96	<0.001
Comedy	1.93	0.39	4.90	<0.001
Drama	-1.50	0.38	-3.93	<0.001
Horror	-4.19	0.56	-7.46	<0.001
Science Fiction & Fantasy	-1.89	0.53	-3.55	<0.001
Kids & Family	3.18	0.75	4.23	<0.001
Mystery & Suspense	-1.36	0.42	-3.21	0.001
Romance	1.52	0.52	2.94	0.003
Documentary	-6.65	0.78	-8.50	<0.001
Cult Movies	-2.23	1.98	-1.13	0.259
Television	-0.91	1.33	-0.68	0.496
Gay & Lesbian	7.77	2.40	3.24	0.001
Musical Performing Arts	2.75	0.70	3.92	<0.001
Special Interest	1.82	0.81	2.26	0.024
Faith & Spirituality	8.66	2.34	3.70	<0.001

b) Lasso Regression

A large number of variables in a model can lead to overfitting. Using subset selection, we can choose an optimal model with a subset of the original variables that will be considered of more importance.

A Least Absolute Shrinkage and Selection Operator (LASSO) regression model eliminates variables that cause a large variance by shrinking their coefficient to zero.

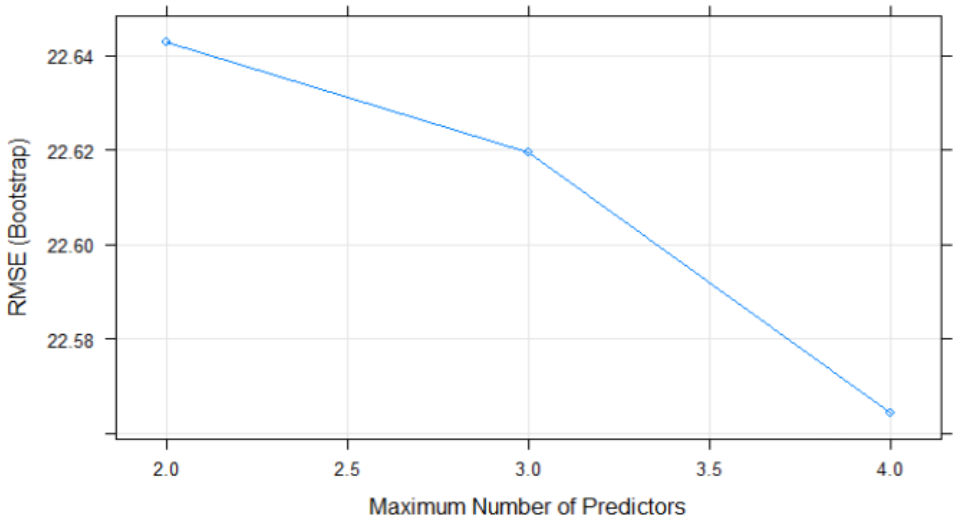
In our resulting optimal model, the variables with positive coefficients are variables that influence a large difference in the audience and critic ratings, while those with negative co – efficients lead to a small difference.

The predictor with the highest coefficient was the Higher variable, which was a dummy variable indicating which group of people had given the higher rating, 1 for audience and 0 for

critics. This could be interpreted to mean that if the audience gives a higher rating than critics, then the difference in their ratings would be much higher compared to the opposite scenario.

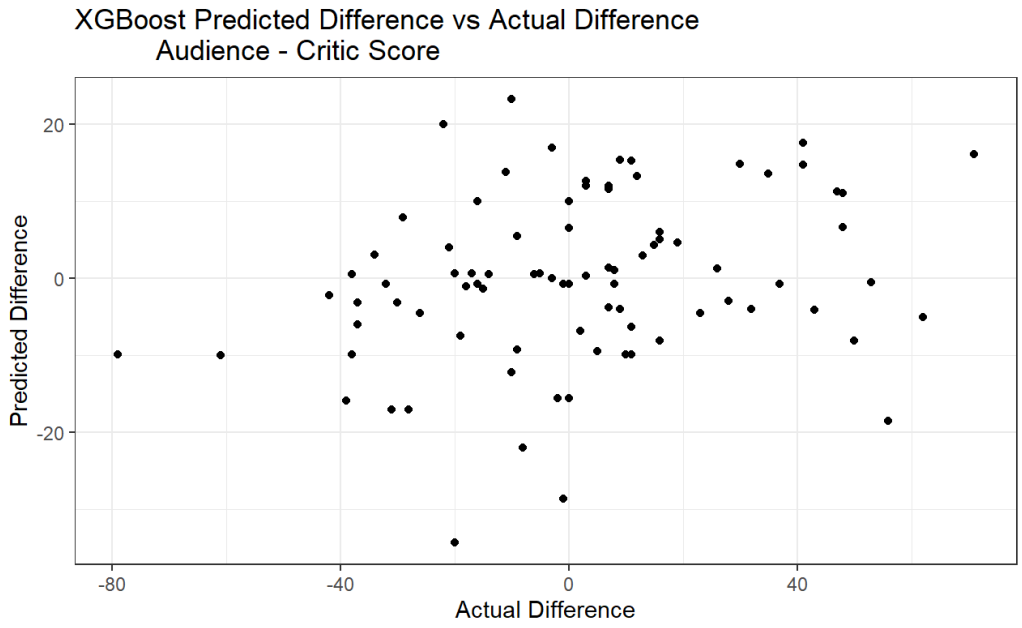
	co-eff estimates
Intercept	11.3417688
content rating	-2.0657880
run time	0.1507788
production company	-0.0001748
days to stream	0.0009951
higher rating	31.3276449
no of critics	0.0752854
action/adventure	-3.5103799
art house	8.5404331
classics	-1.1965355
comedy	-2.3743430
drama	4.8136829
horror	-5.1303886
fiction/fantasy	-2.3372690
kids/family	-1.7935184
romance	-1.6155235
mystery/suspense	-3.5653238
documentary	20.0968657
cult movies	0.3197626
television	1.2226166
LGBTQ+	3.3076392
musical	0.9591782
special interest	0.6299622
spirituality	-5.3674112

Since this LASSO model retained all of the variables from the original model, the next step is to run a forward subset selection model, which tells us the maximum number of predictors to take from the model. In this case, we are to work with 3 predictors.



c) XGboost

Another model was fit using a machine learning algorithm called XGBoost. This algorithm uses gradient boosting decision trees to predict a dependent variable. In this case the object classifier was a regular linear dependent variable which was the difference in mean audience and mean critic scores (audience - critic). The predictor variables were an indicator for each genre as movies most oftenly are a part of multiple genres, and categorical variables for runtime (3 categories) and content rating (6 categories). The model was trained using all movies in the dataset without NA values, prior to the year 2020. The training set had over 15,000 movies while the test set had only 85 movies. The parameters of the XGBoost model were set to ETA = 3, Nrounds = 48, Max.depth = 21 and these parameters were chosen because they provided the lowest RMSE possible, while also using the least number of rounds. The RMSE for this model came out to be 27.5. I believe part of the reason the RMSE is so high is that movies in the year 2020 were more difficult to predict, and with further testing using a randomly split data set the model would have been more accurate.



d) KNN with Sentiment Analysis

For the final model, we were interested if we could extract information from the movie descriptions and use that information for predicting the classification of audience higher vs. lower. KNN, or k-nearest neighbors, is a non-parametric supervised learning algorithm that is typically used for classification. It makes no assumptions about the data, which works perfectly for the sentiment scores, as we only want the information to come from the descriptions. KNN takes the euclidean distance, which finds the distance to all other values, finds the k closest neighbors, and uses that information to form a classifier. Below is the Euclidean formula.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

From what we saw in the lasso regression, genre played a big factor. So we first began by predicting movies past 2019, and training on all movies before then. Then, we would split the data based on a specific genre, and retrain a new classifier and predict all movies of that genre past 2019. Each neighbor was selected on highest prediction accuracy, as it only affected the training data classifier. Here are the results:

KNN Sentiment Analysis, Predicting Audience Higher or Lower for 2019 or later movies

	Accuracy	Sensitivity	Specificity	Train_Ratio	Test_Ratio
No Genre	52.40	0.30	0.68	0.46	0.42
Action/Adv	58.11	0.60	0.56	0.54	0.47
Art House/Intl	63.16	0.17	0.85	0.35	0.32
Documentary	75.00	0.40	0.91	0.30	0.31
Fiction	56.86	0.35	0.68	0.49	0.33
Horror	57.69	0.23	0.75	0.41	0.33
Mystery	60.58	0.39	0.77	0.44	0.43

This table is showing overall prediction accuracy, sensitivity, specificity, the ratio of audience higher in the training set, and the ratio of audience higher in the testing set. Sensitivity is the proportion of true positives identified (in this case, audience higher) and specificity is the proportion of true negatives identified (critic higher). From what we can see, the specificity is larger for all categories except for action/adventure, which is an audience favored category. If we compare back to the audience-critic histogram, this makes sense, as the distribution of differences is left skewed. While this model has fairly high accuracy scores, we were still curious about how the distribution of the scores were affecting the results.

For the second trial, we first began by sampling an equal amount of audience higher and critic higher movies from the overall dataset. Instead of predicting movies after 2019, we chose a random 80/20 split to ensure more audience higher and critic higher movies would make it into the training and testing sets. A KNN classification was then performed on that sample. Afterwards, when genres were split, we reclassified audience higher and critic higher by taking the audience score subtracted from the mean of the audience score, and seeing if it was greater than the critic score subtracted from the mean of the critic score. By doing this, we were able to compare significant differences within genres, which allowed for more audience higher scores overall. Here are the results:

KNN Sentiment Analysis, Predicting Audience Higher or Lower

	Accuracy	Sensitivity	Specificity	Train Ratio	Test Ratio
No Genre	50.13	0.49	0.51	0.50	0.49
Action/Adv	50.09	0.49	0.51	0.49	0.50
Art House/Intl	54.10	0.38	0.67	0.43	0.45
Documentary	59.88	0.45	0.69	0.45	0.40
Fiction	53.24	0.43	0.65	0.47	0.53
Horror	49.67	0.42	0.59	0.47	0.52
Mystery	52.27	0.46	0.58	0.47	0.48

From the table, we can see that there is a slight increase across all sensitivity scores, with the exception of action adventure. We were also able to get the train ratio and test ratios to be much closer to 0.5. This did not improve prediction accuracy, unfortunately, but from attempting

to stratify the data, we learned having more audience higher scores did not positively contribute to prediction accuracy.

CONCLUSION

Research question #1 : Can we predict the higher movie score between the general audience and movie critics?

Answer :

Using data solely obtained from the movie information, we were able to create a model that could meaningfully predict better than random chance for classifying audience higher vs. critic higher for movies released past 2019, with the highest prediction accuracies coming from documentary, art house/ international, and mystery. The model tended to have a higher accuracy in identifying critic higher classifications, which follows the distribution of audience – critic scores shown in the first histogram. When attempting to balance the data, results ended up being much worse in return for having equal test and train ratios and higher sensitivity. We believe with this data, having a left skew, a higher specificity will always lead to better predictions.

Research question #2 : Can we predict the difference between audience and critics scores?

Answer :

We were able to fit a XGBoost model with an RMSE of 27.5. This is not as accurate as we would have hoped but given how much information about any individual movie we can't include into a model it makes sense that the RMSE is high for this model. I believe that using a split at the start of 2020 to make up our training and testing data for this model also had an influence on the high RMSE as it appears that 2020 was a more difficult year to predict the difference in mean audience score and mean critic score (audience - critic). Some of the predictions were extremely accurate (within 1 percent of the actual difference) while others were much further off (almost 70 percent off).

Research question #3 : What are the factors that influence a decision in movie scores?

Answer :

At an RMSE of 22.62, our subset selection models tell us that the predictor variables that influence the difference in audience and critic score ratings are;

- Documentary, Art House, Drama (large difference in ratings)
- Spirituality, Horror, Mystery, Suspense (small difference in ratings)

Note that all these variables are genre dummy variables, telling us that genre is interpreted as the most influential variable when deciding on a movie's rating.

FUTURE RESEARCH

After doing this research and analysis there is still plenty of room to expand on our understanding on audience and critic scores. It would be valuable to dive deeper into using directors and production companies to learn more about how they can be used to predict the difference in scores we commonly see. Another potential idea would be to include casting as a variable, possibly using variables for number of A-list actors/actresses in a movie to determine difference in ratings. We could also consider a categorical variable for popularity of the lead role using either social media followers or google searches.

For the sentiment analysis, it would be useful to try other classification algorithms other than KNN. There is clear indication that information in the movie descriptions, when separated by genre, can be used to classify audience higher vs. critic higher scores. Along with trying new algorithms, we would also use a different library for extracting sentiment scores. We only used one of the basic library packages provided by r, so using a more advanced library could provide more information. Finally, we could apply the sentiment analysis to predicting differences in scores, and see how accurate it is in giving predictions of scores rather than classifications.

REFERENCES

Rao, Ishaan. “Predicting Rotten Tomatoes Audience Score.” Medium, 3 May 2021, https://ishdr08.medium.com/predicting-rotten-tomatoes-audience-score-15da882c95b

Ericson, Jeffrey. Grodman, Jesse. “A Predictor for Movie Success.” Stanford University, 14 December 2013, http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf

Data acquired from: https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_movies.csv

Rotten Tomatoes: https://www.rottentomatoes.com