Predicting Movies' IMDb Scores

Matt Melody

5/07/2021

Which Aspects of a Film Best Predict its IMDb Score?

Given the influx of new films each year, it has become increasingly important for consumers to evaluate the quality of a film prior to watching it to understand whether that film will be worth the value of their time. While individual critics and consumers may have radically differing opinions on the overall quality of a film, particularly with many current films that concern themselves with topical, socially relevant subject-matter, the aggregate consensus of a film usually provides a decent understanding of the film's quality. While previously limited to an overall critical consensus to understand the "value" of a film, the onset of the internet has given rise to various databases and "aggregation" reviewers that combine multitudes of different review-scores into a single digit. While Rotten Tomatoes and Metacritic place focus on the critical consensus, with a smaller component focused on consumers, the Internet Movie Database (IMDb) places primary focus on aggregating scores from individual consumers to form an overall "IMDb Score."

This IMDb Score is a number "out of ten" with a single decimal provided to give deeper insight into the varying qualities of films. Unlike the score itself, consumers can only rate a movie with whole numbers, ranging from 1 to 10 (10 being the best and 1 being the worst). Unlike other "aggregation" review-sites like Rotten Tomatoes and Metacritic, IMDb does not provide a filter that forces consumers to prove that they have seen the movie that they are rating; however, IMDb does force individual reviewers to create an IMDb Account with an associated Email to help prevent repeat-votes. Furthermore, IMDb places equal weight to each vote cast, unlike Metacritic, which increases the weight of longtime members' votes compared to newcomers.

In this analysis, I plan to predict the aggregate IMDb scores of movies using two separate methods: **conditional means**, and **linear regression**. The data that I am analyzing originates from a Kaggle Dataset (https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset/download), specifically from the "IMDb movies.csv."

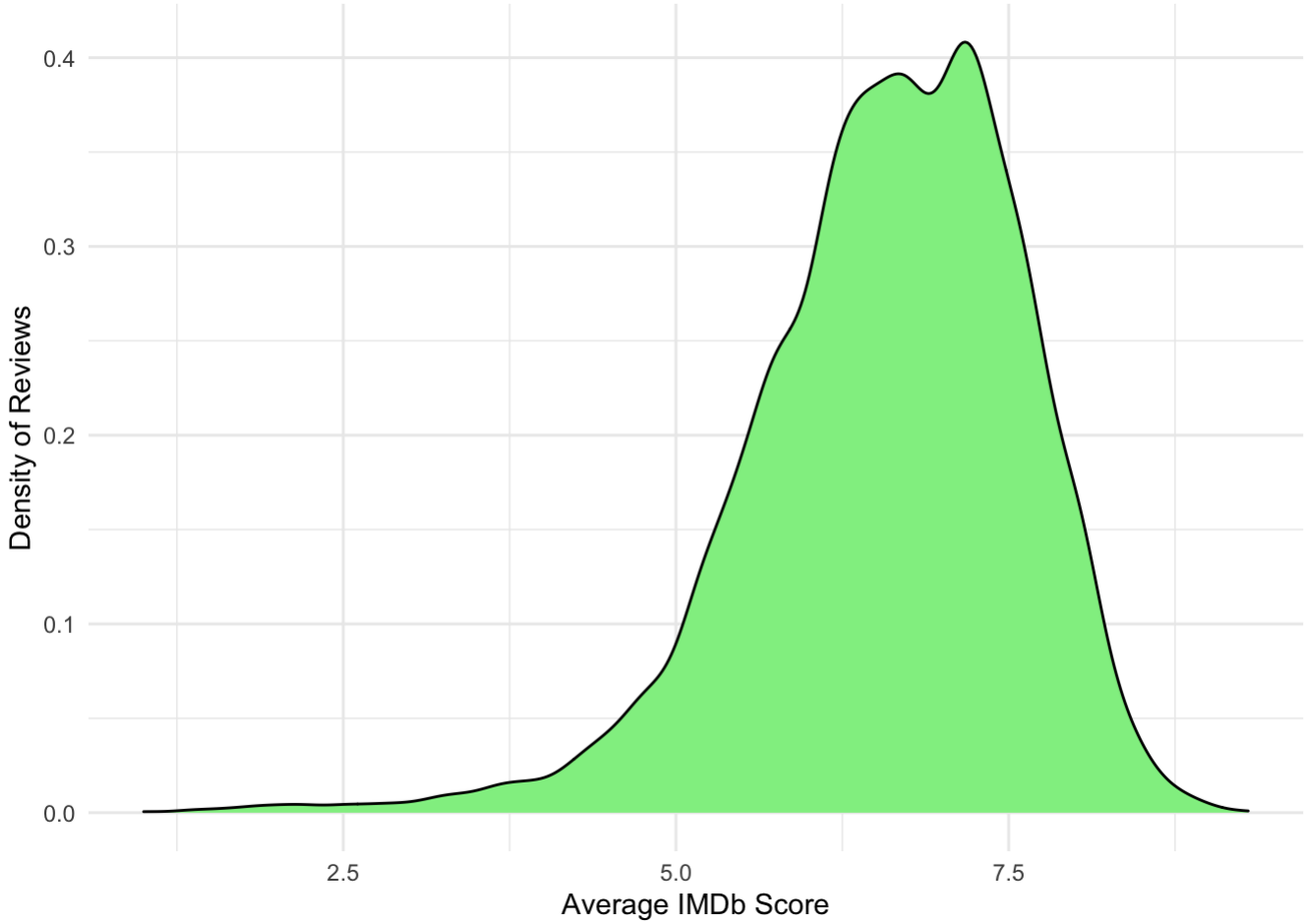
Conditional means will predict the aggregate IMDb score, using certain attributes to increase accuracy, which will be measured by calculating the root-mean-square-error (rmse). The following variables will be used in individualized, conditional mean analyses: worldwide box-office income, number of IMDb votes cast, genre, and year of release. Ideally, each of these analyses will produce an rmse value considerably less than the standard deviation of the IMDb score variable in the given dataset; the standard deviation of IMDb scores will be equal to its rmse when the unconditional mean is used as the sole predictor.

Linear Regression will combine all the variables used as conditional means into a single, linear model to predict the aggregate IMDb scores of movies, once again using the rmse to estimate the accuracy of the predictions. In addition to the rmse calculation, an r-squared value will be calculated to determine the percent of variation in the IMDb score variable that is being accounted for by the model. To ensure this linear model has not been "over-fit" to the training data—causing for misleading rmse and r-squared values—, I will incorporate cross validation through bootstrap resampling, which will test the model against differing segments of the dataset to understand its precision.

All in all, these various methods will display the variables of a film that best predict its IMDb score and help provide further information to the trends involved with the scoring of movies via IMDb itself.

Characteristics of the IMDb Score Variable

The figure below depicts the distribution of the average IMDb scores of movies that have greater than 5000 total votes from individuals on the website. Movies with less than 5000 votes were omitted to ensure that the average IMDb score accurately depicts the overall quality of the movie—with relatively equal weight given to individual votes.



1. **25th Percentile = 6**
2. **Median = 6.7**
3. **75th Percentile = 7.3**
4. **95th Percentile = 8**
5. **Unconditional Mean = 6.58**
6. **Standard Deviation = 1.03**

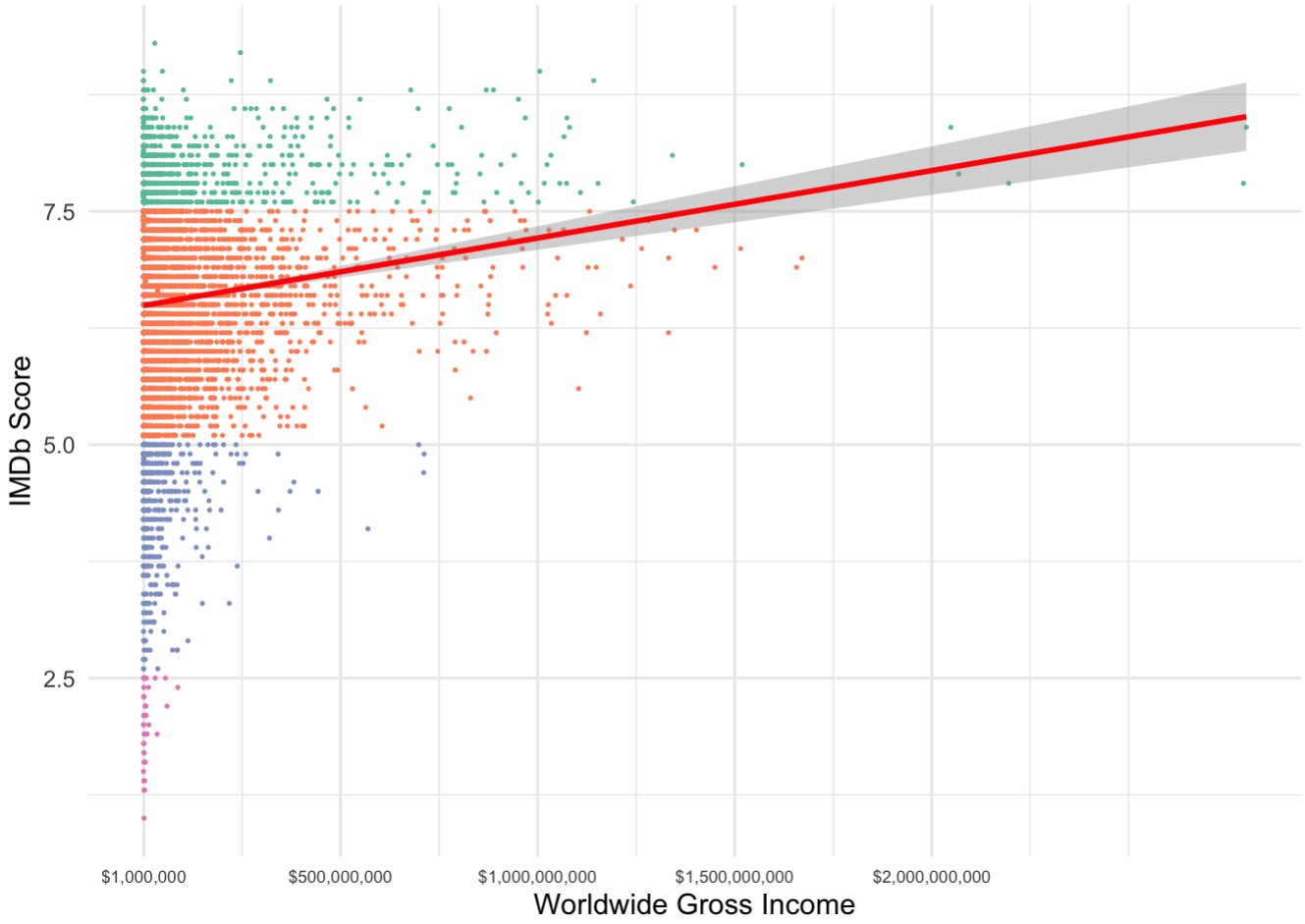
As showcased above, half of the movies recorded on IMDb have an average IMDb score above a 6.7 out of 10, while only the **top 5% of movies** have an average score above an 8 out of 10. Furthermore, the top 25% of movies will have an average IMDb score above 7.3, while the bottom 25% will have an IMDb score below 6 out of 10.

Given that the mean and median values for average IMDb score (6.58 and 6.7 respectively) are within 0.12 in score, we can also observe that the data is relatively normal, slightly skewed left. Assuming that the distribution is approximately normal based on the graphic and related metrics, it is clear that approximately 68% of films have IMDb scores within one standard deviation (1.03) of the unconditional mean (6.58).

Does Worldwide Box Office Income Accurately Predict a Movie's IMDb Score?

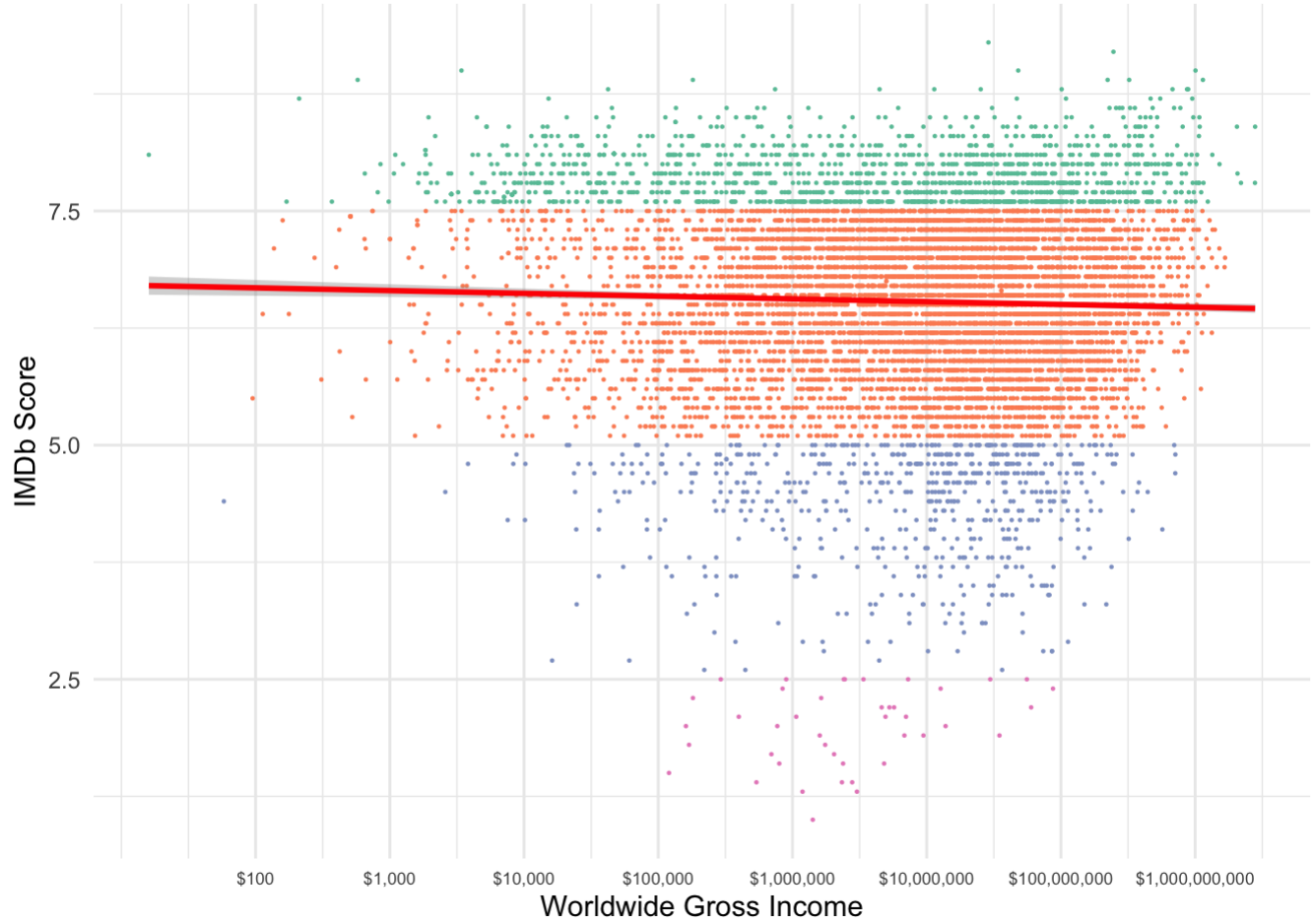
It may seem intuitive to think that the "best," most well-received movies would make the largest box-office gross income, yet in my conditional mean analysis, this was found to be far from the truth. As can be seen in the graphic below, using a film's global box office income to predict its average IMDb score produces a figure that appears to have a positive correlation (showcased via the upward trend-line); however, this figure is *also* highly skewed to the left due to the vast differences in income amounts, creating a potentially misleading correlation.

Before Logarithmic Transformation

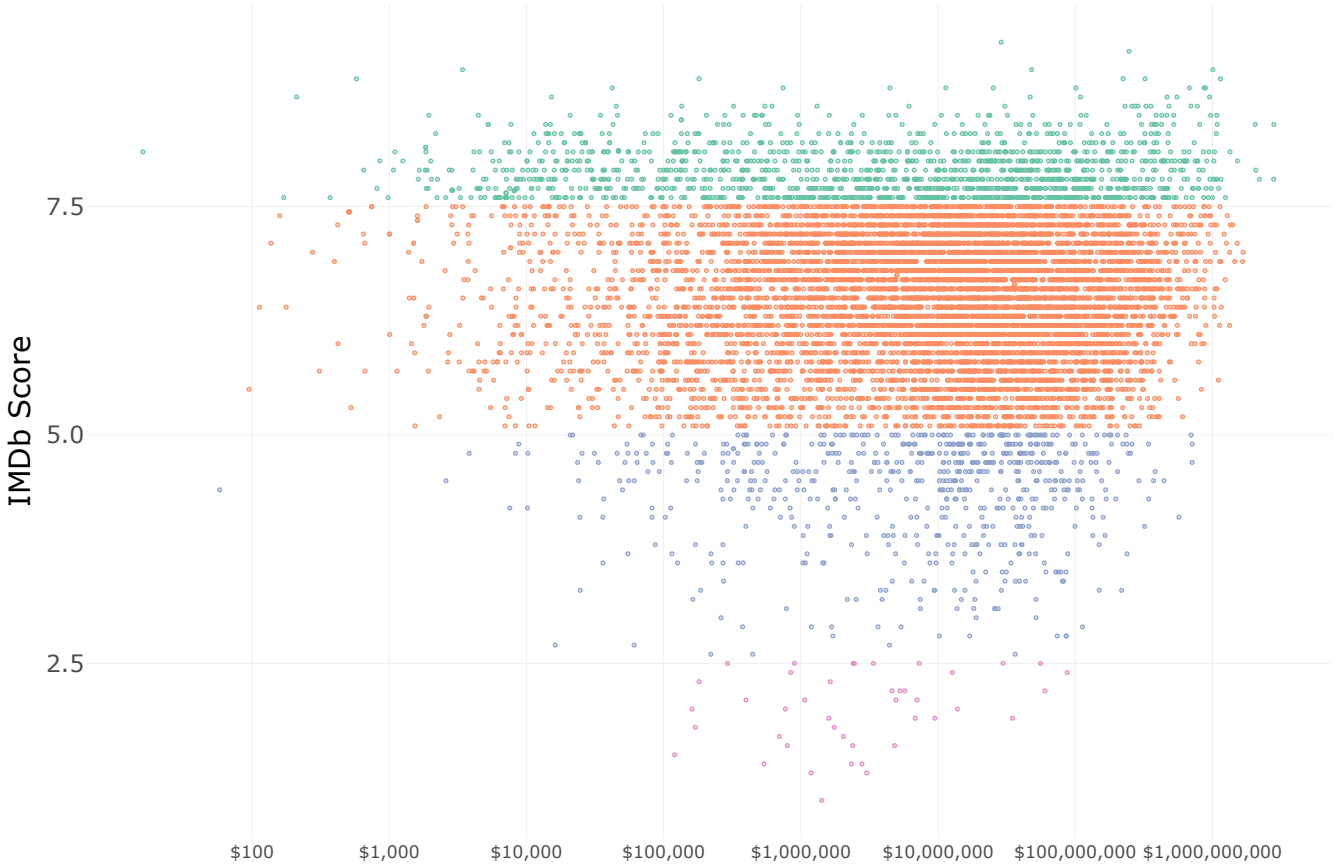


After introducing a logarithmic transformation to Worldwide Gross Income to lessen the highly skewed data-points, the graphics reveal that little correlation appears to exist between Worldwide Box Office Income and IMDb Score. Once gross income is logged to account for the skewing effect of outlying box-office earnings, the resulting trend-line becomes horizontal, showcasing no major, identifiable correlation. Ultimately, after the logarithmic transformation, the graphics showcase that no consistent trend can be found within the given data.

After Logarithmic Transformation



Interactive Graph



Worldwide Gross Income

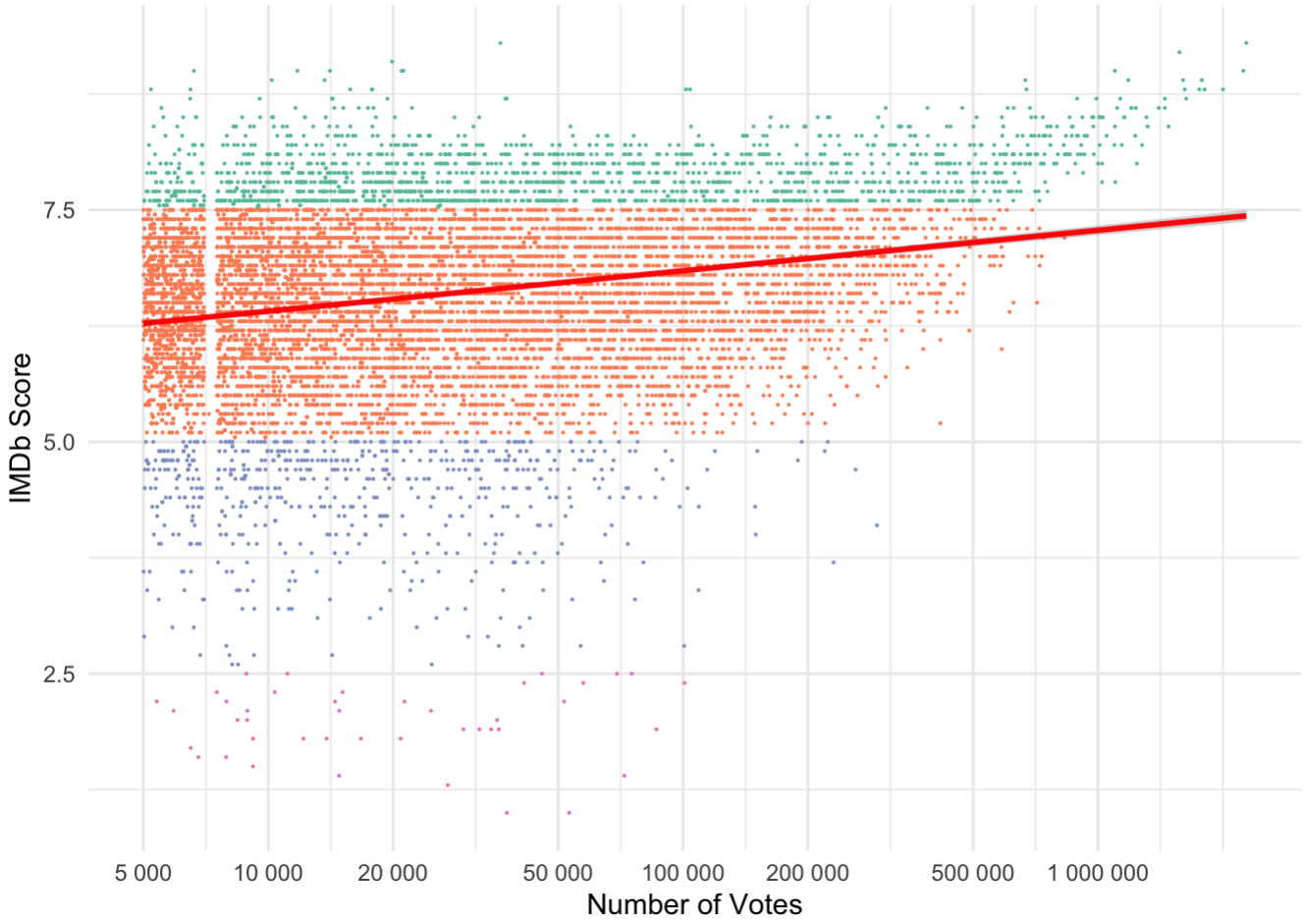
RMSE of IMDb Scores = 0.443

Despite the lack of correlation between the two variables, the use of Worldwide Box Office Income to estimate via conditional means yields an rmse of 0.443, significantly less than the standard deviation of 1.03. This significantly lower rmse indicates that using Global Box Office Income manages to predict the IMDb Score much more accurately on average than solely using the unconditional mean to estimate.

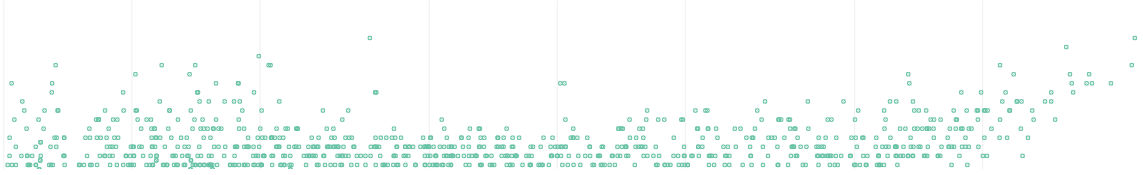
Does the Number of IMDb Votes Accurately Predict a Movie's IMDb Score?

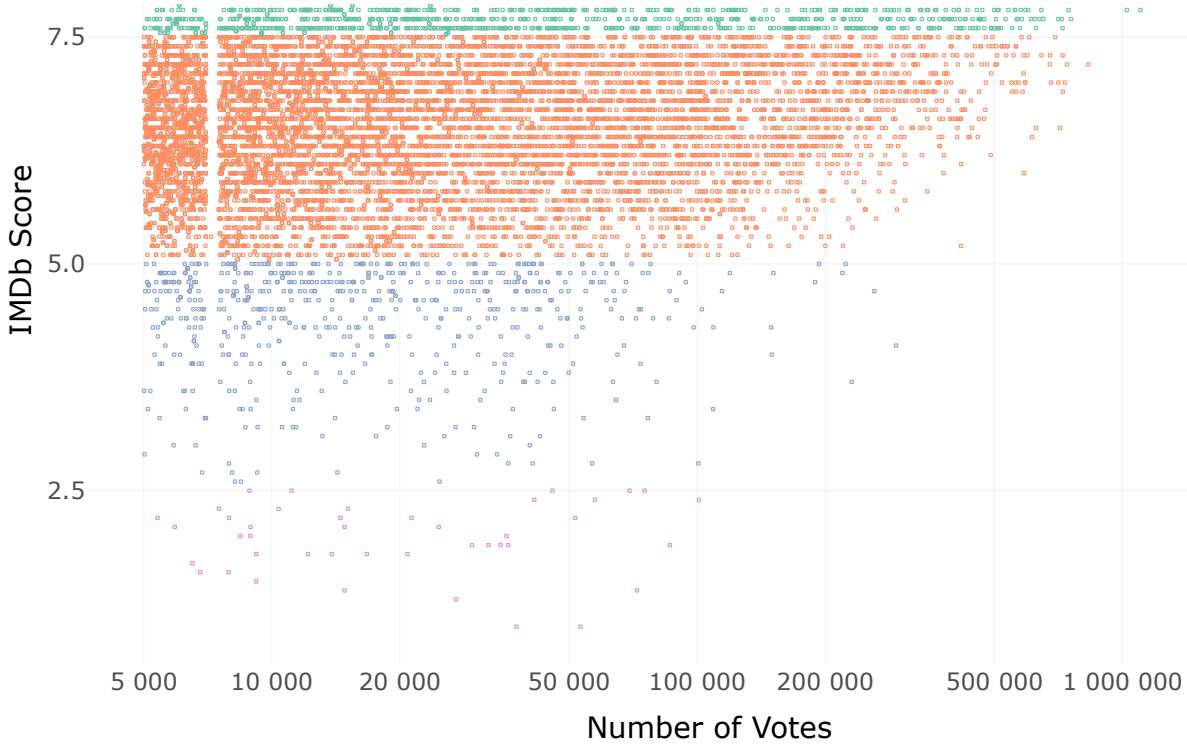
Logically, one can infer that as more individuals contribute to a movie's IMDb score, the score will trend towards a higher or lower value. Voting on IMDb takes effort, and, typically, more individuals will vote on a project that they *really enjoyed* or *really hated*.

Line of Best Fit (With Logarithmic Transformation)



Interactive Graph (With Logarithmic Transformation)

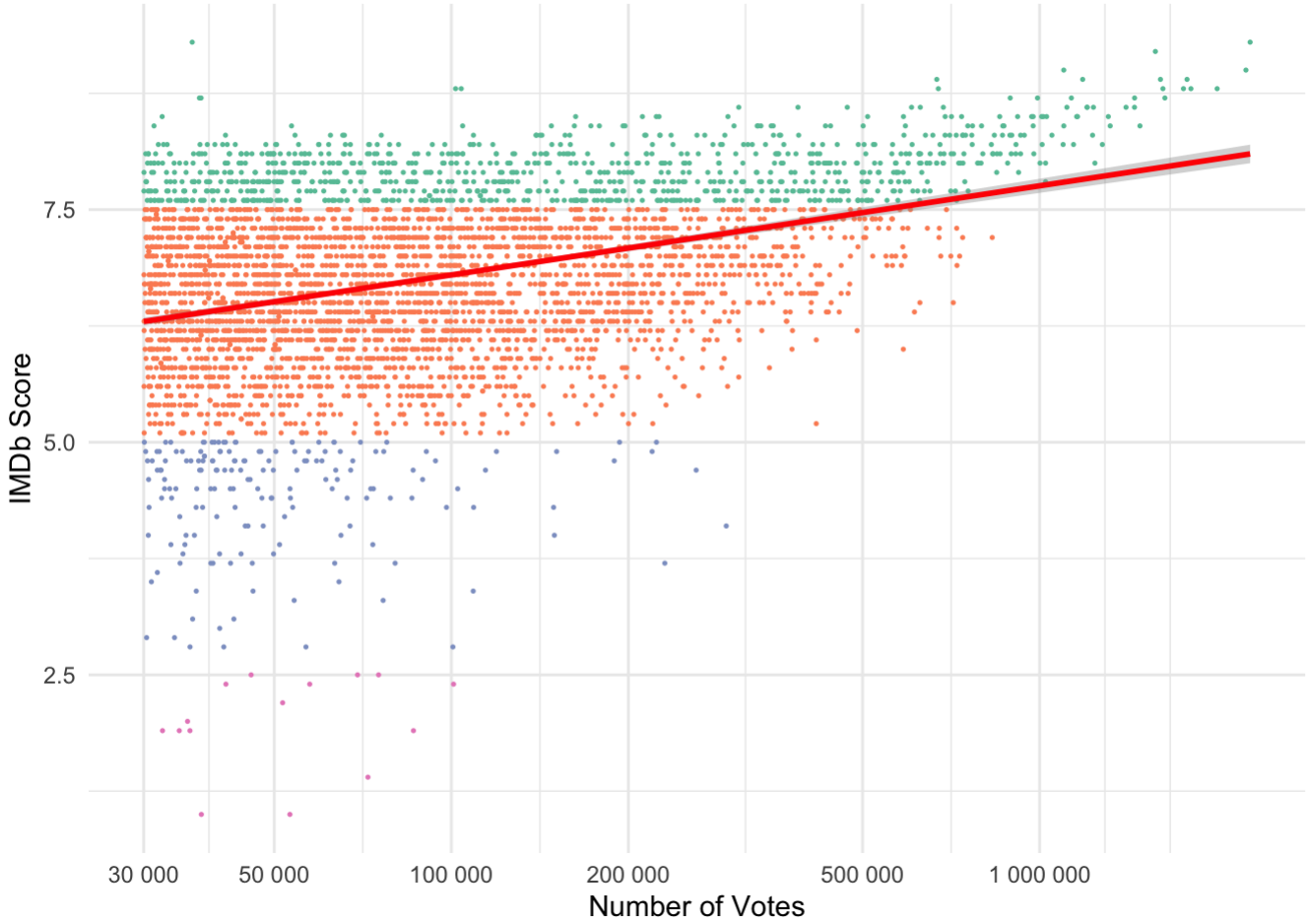




RMSE of IMDb Scores = 0.391

As can be seen above, the number of votes provides an excellent rmse of 0.391, considerably below the standard deviation of 1.03, illustrating its considerable impact in predicting the IMDb score. The figure above has two distinct sections that contain conflicting information, which is examined in detail in the graphics below.

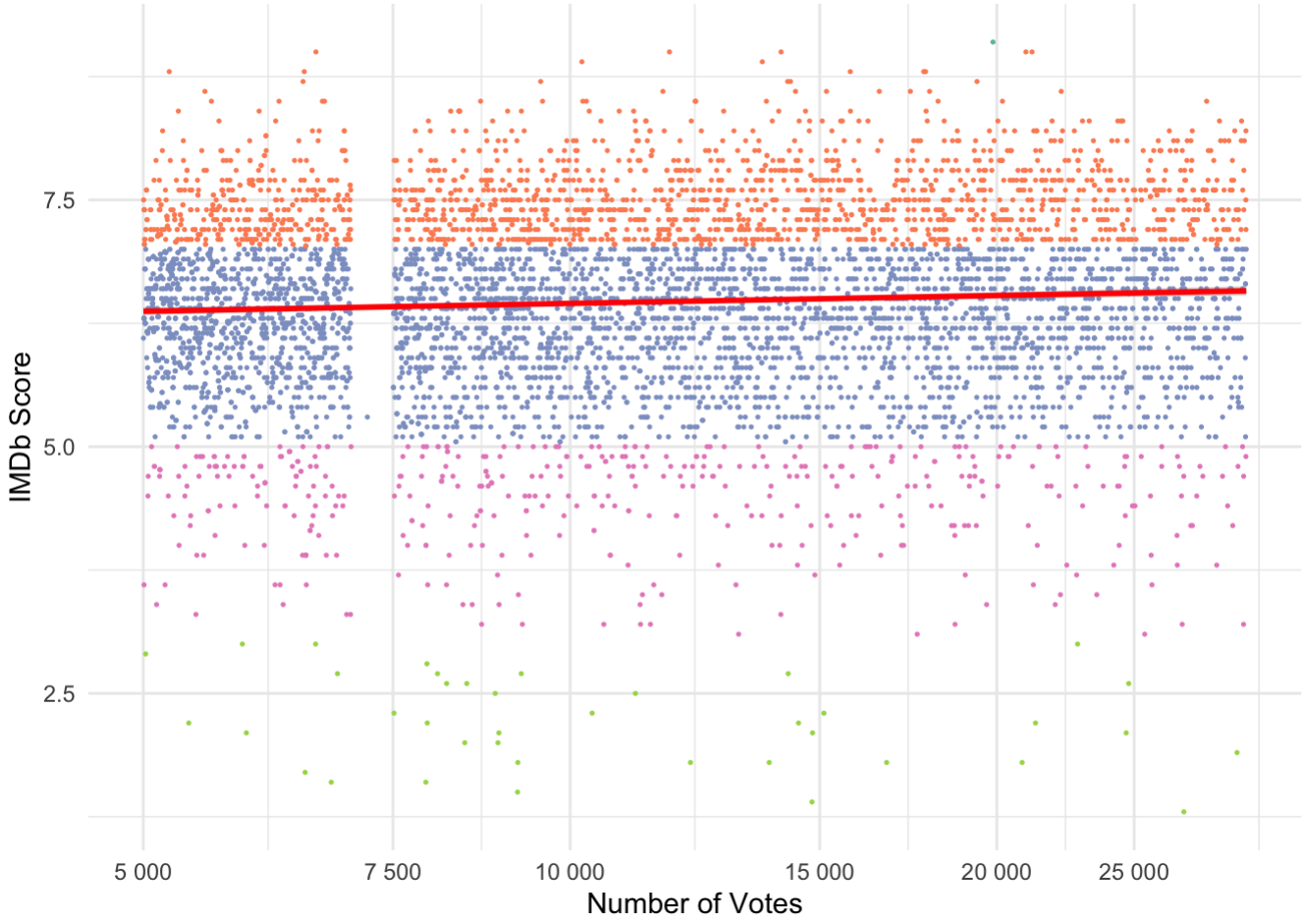
Above 30,000 Vote Threshold (With Logarithmic Transformation)



RMSE of IMDb Scores = 0.114

Based on the figure above, while not a directly linear relationship, the number of votes, past the 30,000 vote threshold, can help approximate the Average IMDb Score of the movie very accurately. Above 30,000 votes, the number of votes can produce an rmse of 0.114. However, below the 30,000 vote threshold, almost no correlation can be observed—as can be seen in the graphic below—despite its rmse of 0.488.

Below 30,000 Vote Threshold (With Logarithmic Transformation)



RMSE of IMDb Scores = 0.488

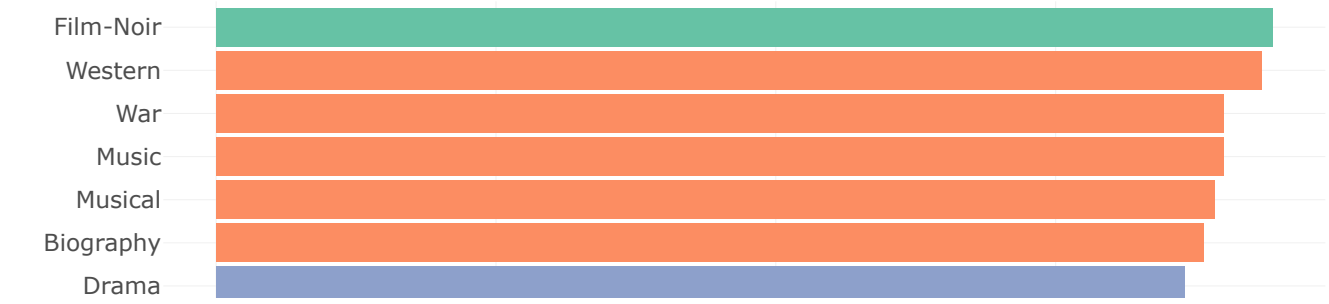
Note: The data used only includes movies on IMDb that have above 5000 votes. This was done to ensure that obscure films with very few votes did not skew the average rating of the movie to be inconsistent with its actual quality.

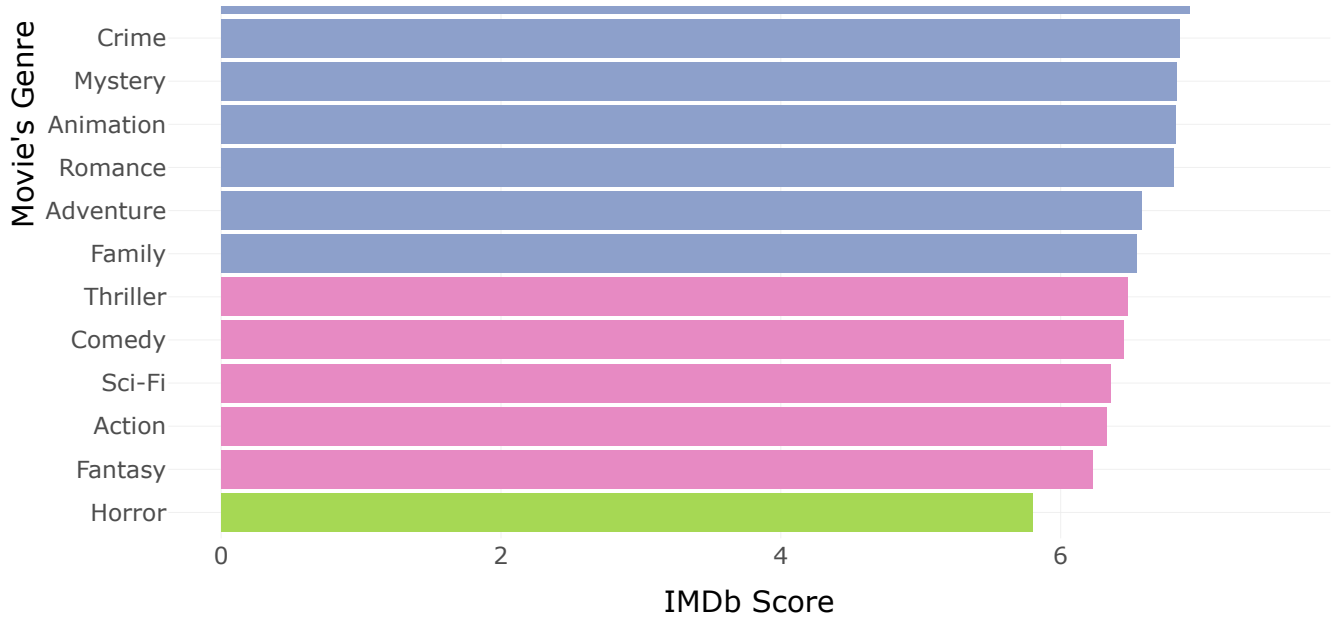
How Does a Movie's Genre Impact its IMDb Score?

Given that certain genres, specifically Horror and Comedy, are commonly known to have a high abundance of poorly-received films, it is logical to think that a film's genre may be an effective predictor for a film's IMDb score.

The following information provides somewhat conflicting information. The graphic below indicates a significant divide between certain genres based on their average IMDb scores, with Horror having the lowest average IMDb Score (5.8/10) and Film Noir having the highest average IMDb Score (7.55/10).

Interactive Graph





RMSE of IMDb Scores = 0.981

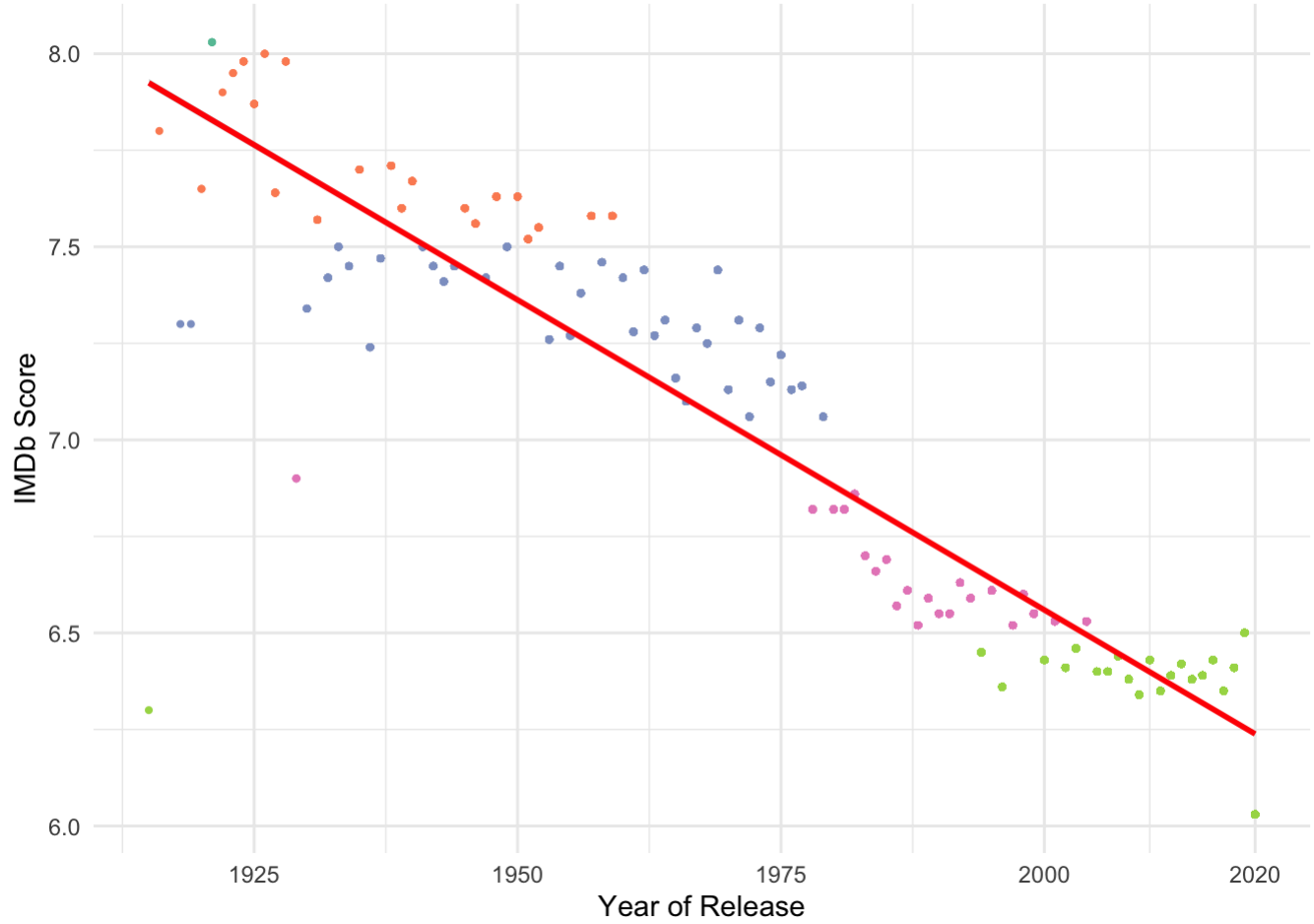
Despite the loose association indicated by the bar-chart above, using a film's Genre to predict its IMDb Score yields an insignificant rmse of 0.981, slightly improving upon the standard deviation of IMDb Scores: 1.03. Despite the high rmse, relative to the IMDb data, I will utilize Genre in my model due to the association indicated by the bar-chart.

How Does a Movie's Release-Year Impact its IMDb Score?

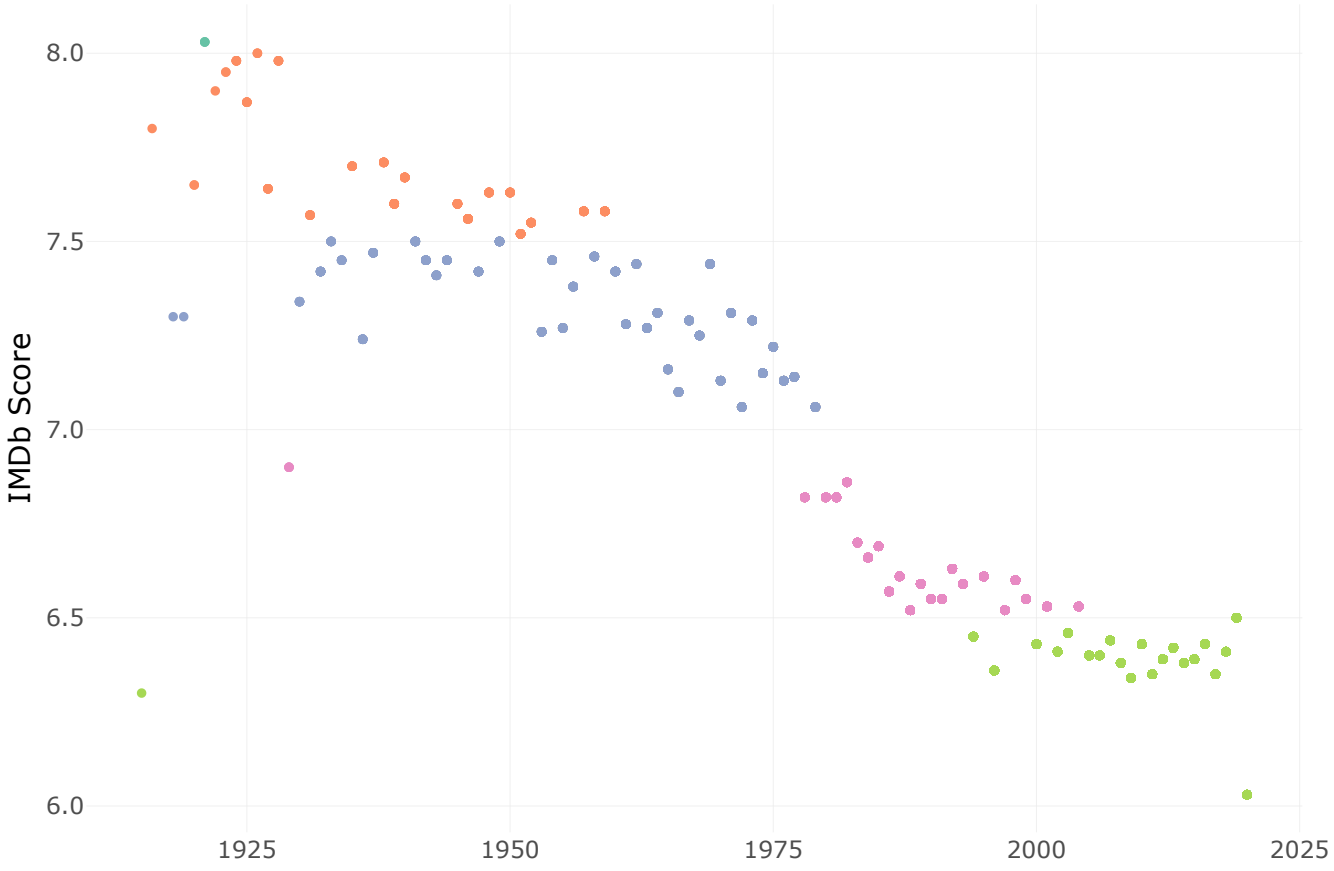
Given that film production has grown more accessible over time, allowing for more indie films to be produced, It seems logical to assume that the overall average IMDb score of films to will decrease over time. With an increased number of unprofessional film-makers, there will likely be an increase in poorly-received films after each consecutive year.

Based on the graphics below, a clear trend is showcased as the average IMDb score for films has decreased over time. The rmse given solely by year is 0.976, which is slightly below the standard deviation amount of 1.03.

With Line of Best Fit



Interactive Graph



Year of Release

RMSE of IMDb Scores = 0.976

Despite the the relatively high rmse produced (0.976) in comparison to the standard deviation of 1.03, the graphic above clearly indicates a downward trend, with the majority of movies having an estimated average IMDb score of approximately an 8.0/10 around 1925 and those today (after the year 2000) having an average score of approximately 6.3/10.

Due to the considerably linear relationship indicated in the graphic above, I have chosen to keep the movie's release-year as an independent variable in the predictive model.

The Predictive Model

Using these various predictors (worldwide box-office income, number of votes, genre, and year of release) and the insights provided from them, I formatted a linear regression model to predict the aggregate IMDb score of any released movie on the IMDb Database.

The Linear Regression Model:

RMSE of IMDb Scores = 0.8695977

R-Squared of IMDb Scores = 0.2586

Ultimately, the model produces predictions more accurate than using the unconditional mean of 6.58 (which yields an rmse/standard deviation of 1.03); the model's resulting rmse is 0.8695977, which indicates that, on average, the model's predicted IMDb Score will be "off" by 0.8695977. This is a 15.57% improvement in accuracy to the rmse of the unconditional mean.

Furthermore, the model yields an r-squared of 0.2586, which indicates that the model accounts for approximately 25.86% of the total variation in the IMDb score variable; alternatively, this value indicates that 74.14% of the IMDb score variable is not taken into account by the model. Given the complexity and variability in IMDb scores–and a movie's overall quality–this 0.2586 metric is relatively high and indicates that a considerable percentage of predictable attributes that constitute the overall quality of a movie are being considered by the model.

The results from the model provides a t-value of +34.046 to the independent variable "the number of votes," indicating that this predictor is significant in predicting the aggregate IMDb score of movies. Furthermore, the model indicates that, with every vote a movie receives, the IMDb score will, on average, increase by 0.000002713. In other words, with every 10,000 increase in the number of votes a movie receives, the IMDb score will, on average, increase by 0.02713.

Additionally, the model also indicates a high significance to the independent variable "year of release" with a t-value of -29.459. The model also suggests that, with every successive year since 1915, a movie's aggregate IMDb score will decrease by 0.01562, aligning with the previous insight provided from the conditional mean analysis that suggested a negative correlation between "year of release" and IMDb score.

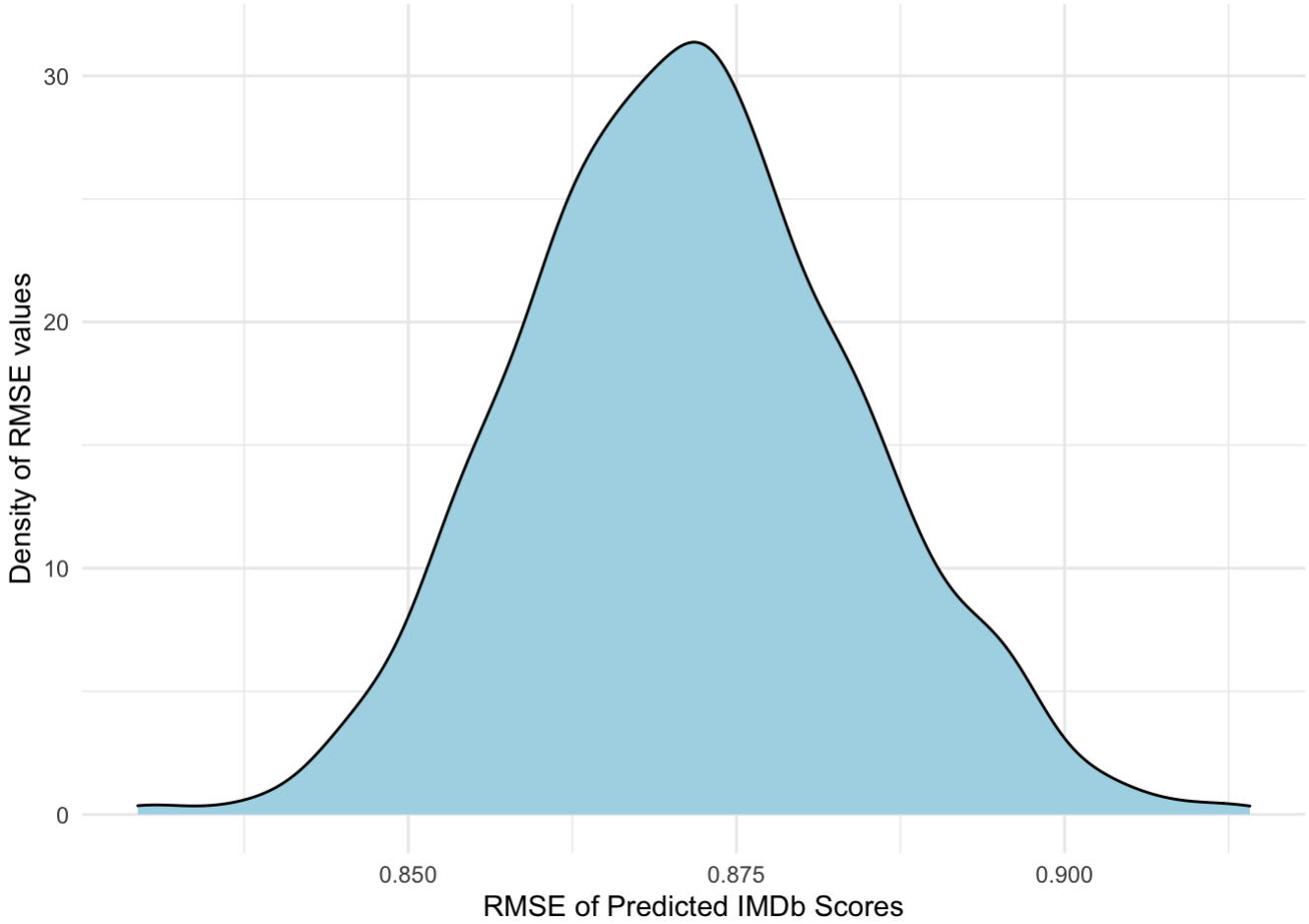
Of the various genres observed by the model, Drama and Biography had the most significant impact in predicting IMDb Scores–with t-values above +20.000 each. The model suggests that Dramas receive an additional 0.6342 in their IMDb Scores, while Biographies get an increase of 0.8245 in their scores. The model also suggests that Crime and Animation films have increased significance above other genre-types, with t-values above +10.000 each. Crime and Animation respectively received increased IMDb scores of 0.4884 and 0.5969 as indicted by the model. Alternatively, Thriller, Family, and Sci-Fi films had the least significance in predicting IMDb Scores with provided t-values less than +/-1.000.

Lastly, as indicated by the model, worldwide box office income has a reasonably significant effect on predicting an aggregate IMDb score–with a t-value of -7.016. Interestingly, unlike the conditional mean analysis performed beforehand, the model suggests a negative correlation between worldwide box office and IMDb score, indicating that, with every dollar a movie earns, its IMDb score will decrease by 0.000000000556. In other words, the model suggests that, for every 10 million dollars earned by a movie, the IMDb score will decrease by 0.00556.

Note: Due to the model having four independent variables, worldwide gross income could not be incorporated into the visual graphic above. Despite its exclusion from the visual graphic, all data and observations from the model incorporated worldwide gross income .

Bootstrap Resampling (Cross Validation)

To ensure that the previous model has not been "over-fit" to the given data and remains as precise for other datasets, I have created a bootstrap resampling model that will provide insight to the performance of the model in making out-of-sample predictions. Ultimately, the generated metrics will provide the distribution and error of a multitude of out-of-sample rmse values to understand how consistently the model performs.



RMSE of Predicted IMDb Scores = 0.871

R-Squared of IMDb Scores = 0.256

The above graphic outlines the distribution of 1457 out-of-sample rmse values for the IMDb Score variable and reveals the consistency of the model in predicting the aggregate IMDb Scores of movies through its approximately normal distribution. A unimodal distribution, the graphic reveals the average rmse value of the 1457 trials (showcased by the graphic's high-point) to be 0.871, which is very similar to the initial model's rmse value of 0.8695977.

Furthermore, from the 1457 different datasets, the bootstrap resampling produced an average r-squared value of 0.256, which, once again, remains very consistent with the r-squared value produced by the initial model: 0.2586. This precision indicates that the model consistently addresses the same approximate percentage of variation in IMDb scores.

All in all, the bootstrap resampling approach revealed that the model remains precise in its predictions over 1457 different trials, with the average rmse value and r-squared values remaining consistent. Furthermore, the unimodal and (approximately) normal distribution of the rmse values indicate that predictions nearer the model's mean are much more likely to be produced than those significantly larger or smaller.

Conclusion

Based on the conditional mean analyses as well as the linear regression model (and bootstrap resampling of the dataset), it becomes clear that, despite the improvements from the model, aggregate IMDb scores remain incredibly difficult to predict based on "hard-factors." Movies are known to be incredibly dynamic products–with multitudes of varying factors that change and transform (sometimes seemingly without any underlying cause).

Furthermore, the "public's" ability to vote on IMDb further extrapolates the dynamism of IMDb scores, as social factors and current events can manipulate the aggregate score of a film (regardless of its inherent quality). Specifically, public manipulation of IMDb scores can be seen in the "review-bombing" of movies, notably those such as "Captain Marvel" and "Star Wars: The Last Jedi."

Despite the overall average rmse value of the model via bootstrap resampling being 0.871 (and just approximately 15.44% more accurate than using the unconditional mean IMDb score to make predictions), the average r-squared value from bootstrap resampling (0.256) indicates that a large proportion of predictable variation is being considered by the model. Ultimately, while far from being a perfect model, this model *will* allow for more informed predictions of a movie's aggregate IMDb score by using the variables observed.
