Does Movie Runtime Impact IMDb Ratings?

Thomas Farrar

08/12/2023

1 Introduction

The dynamic between a movie's runtime and its reception, through IMDb ratings, is intriguing for both filmmakers and audiences. Understanding the factors that influence a movie's success is key as the film industry continues to evolve. This executive summary researches the question whether Movie runtime impact IMDb Ratings. This analysis aims to find patterns and correlations using various regression models.

2 Significance of the Research

Understanding the impact of movie runtime is a fascinating topic for a few reasons. Filmmakers and producers are continually seeking ways to enhance audience engagement. Recognising the influence of movie runtime on viewer satisfaction can guide the development of more captivating experiences and deliver content that resonates with modern viewing habits. Additionally, movie production involves substantial financial investments. Knowing how runtime correlates with the movie's reception can influence budget allocation, marketing strategies, digital platforms, can improve a movie's economic success.

3 Dataset and Preprocessing

The dataset used for this investigation was taken from online on kaggle.com [1]. The dataset contains 25,000 movies, with details on the movie's title, total runtime, average user rating, total number of user's that rated the movie, types of genres, a short overview, plot keywords, movie director name, names of top five casts, writer's name, and the year of release. The data was last updated 9 months ago and original source of the data was scraped from the IMDB.com website [3].

After loading the csv file, the dataset revealed that it only contained 24,402 movies. The first step was to check for any duplicate entries. While there were movies with the same title, there were no duplicate entries in the dataset. Some of the movies did not have a rating, therefore I removed these from the dataset, bringing the total movies down to 22,662. In the runtime column, a lot of movies contained data other than the runtime, such as 'non-released', for movies that never released, and '$60,000,000 (estimated)', which unfortunately appears to be a mistake, so removing these left the total at 11335 movies.

None of the columns contained null entries, except for 6 movies with missing years of release, so removing these ensured that there weren't any movies that hadn't released. Checking that the year of each movie was not 2023 or later also removed any movies that hadn't released. While it resulted in the exclusion of a few movies that were released in the early months of 2023 when the dataset was updated in March, this was preferable to including movies from 2023 that had not yet been released.

As the runtime values were all formatted as hours and minutes, it was necessary to convert them into a numerical format of total minutes. Removing movies that had a runtime of less than 40 minutes, left only what is considered a feature length film, and not short films [2]. Certain movies, like 'The Clock' [4], had extreme runtimes, reaching up to 24 hours. Recognising these as outliers,

they were excluded to ensure a more focused and representative examination of the majority of movies.

4 Machine Learning

The chosen and most suitable machine learning technique to address the research question and analyse the dataset is regression. Regression provides a robust framework for conducting a quantitive examnation of the relationship between two continuous variables, these being the movie runtime and IMDb ratings. It also facilitates predictive modeling, enabling filmakers to anticipate the potential impact of runtime decisions on viewer reception and make informed decisions about the optimal movie duration.

Every model that was used employed cross-validation to provide a more robust and reliable estimate of the model's performance compared to a single train-test split, as well as aiding in identifying which models generalise well to unseen data.

The first model used was a Linear Regression and resulted in an R-squared value of 0.10582460822809847 and showed that there was a weak positive correlation between movie runtime and IMDb rating. However, these variables do not have a linear relationship so this model can't effectively model the trend.

The next model used was Polynomial Regression, which gave an R-Squared value of 0.11872786162600257. Using the Bayesian Information Criterion (BIC) to select the best polynomial degree, the model identified a degree of 2 as the most favourable. When incorporating the projected line into visualizations of the dataset, runtime versus IMDb rating, the resulting curve reveals an inverted U-shape. This pattern illustrates an initial descent, followed by a peak, and ultimately a subsequent decline.

The other two models used were Ridge Regression and Lasso Regression which introduce regularisation. With only one predictor variable in this research, being the movie runtime, multicollinerity and the need for sparsity in the model is not crucial. Therefore these models do not provide significant benefits over the others. Nevertheless, the Ridge Regression model provided an R-Squared value of 0.13258178106971608 and the Lasso Regression model 0.12981277909420486.

5 Conclusion

Out of all the models used in this research, with Ridge Regression having the the highest R-Squared value it indicates a better fit to the data. The line in Figure 1 suggests that the relationship between runtime and IMDb rating is complex. Like the Linear Regression, the curve still generally demonstrates a weak positive trend. Notably, there are fluctuations in the ratings as runtime increases, suggesting that movies with a higher runtime do recieve better IMDb ratings, however when the movie is too long the curve begins to decline.

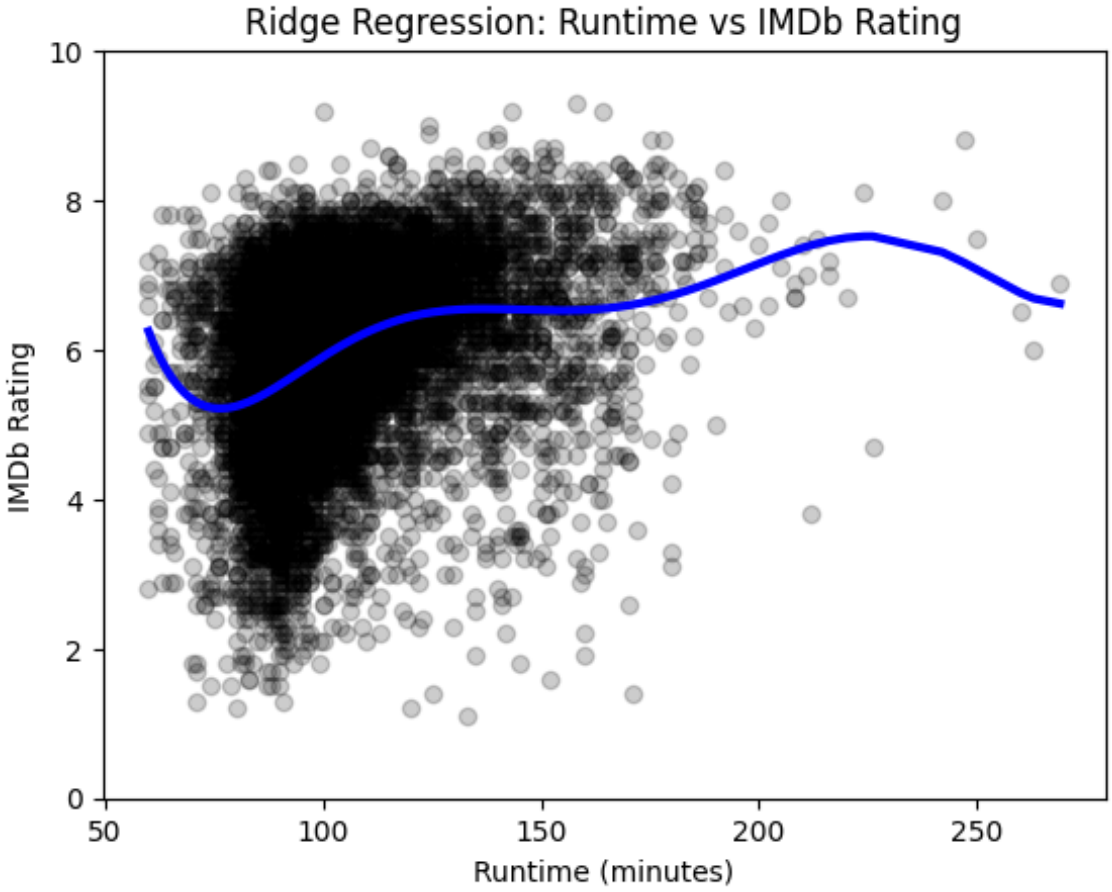

Figure 1: The figure shows a scatter plot of the IMDb rating against movie runtime with the fitted Ridge Regression curve overlayed.

References

[1] 25k imdb movie dataset. https://www.kaggle.com/datasets/utsh0dey/25k-movie-dataset/data. Accessed: 2023-12-08.

[2] British film institute filmography faq. https://www.bfi.org.uk/bfi-national-archive/search-bfi-archive/bfi-filmography/bfi-filmography-faq. Accessed: 2023-12-08.

[3] International movie database. https://www.imdb.com/. Accessed: 2023-12-08.

[4] The clock, 2010.