IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Statistical Modelling of Movie Ratings: Lasso and Ridge Regression Perspectives

Jatan Patel[1], Kaustubh Shinde[2], Dr. Bhushan Jadhav[3]
Student[1], Student[2], Professor[3]
[1-3] Thadomal Shahani Engineering College, Bandra (West), Mumbai, India

Abstract: It is crucial to accurately predict how people would receive movies in the entertainment industry. In this study, Lasso and Ridge Regression models are used to forecast IMDb ratings, which are an important indicator of public opinion based on a comprehensive set of movie metadata. The dataset contains variables like budget, revenue, runtime, vote count, director and lead actors. We used Lasso and Ridge Regression as regression analysis techniques to create a model that predicts IMDb ratings from movie attributes. When using Lasso Regression penalization terms for the size of coefficients for individuals features rather than groups to encourage sparsity and provide greater interpretability to the model; but when using Ridge Regression all these coefficients can be shrinked so as to reduce the effect of collinearity among them. Continuous preprocessing, feature engineering, and model training enable accurate IMDb ratings predictions. Performance measures such as mean absolute error (MAE), mean squared error (MSE), and R-squared test statistic are utilized in evaluating models' performances. The results showed that both Lasso and Ridge Regression models could effectively predict IMDb ratings with their mean absolute errors (MAE) being 0.453 and 0.449 respectively, mean squared errors (MSE) were 0.573 and 0.591 while R-squared scores were 0.942 and 0.940 respectively. These figures underscore the value of regression analysis in enhancing predictive accuracy within the entertainment industry, offering valuable insights for filmmakers, production studios, and distributors.

Index Terms **- IMDb Ratings Prediction, Lasso Regression, Ridge Regression, Movie Metadata, Regression Analysis.**

I. INTRODUCTION

The film industry plays a significant role in the entertainment sector, consistently engaging audiences through a mix of creativity and variety. In this rapidly changing environment, where a movie's success often depends on accurate predictions of audience reception, IMDb ratings serve as a critical measure of audience appreciation. These ratings summarize viewers' opinions about a film and heavily influence their decisions on what to watch. To uncover preferences for high-quality movies within vast movie datasets—including budget, revenue, runtime, vote count, director, and lead actors—advanced predictive models such as Lasso and Ridge Regression can be effectively applied.

With the increasing prominence of digital platforms and the expansion of streaming services, the competition for viewers' attention has grown fiercer. Filmmakers, production companies, and distributors are increasingly relying on data-driven strategies to optimize their content and enhance its appeal. Predictive modeling serves as a powerful tool for understanding the many factors that influence a film's success, helping stakeholders make informed decisions throughout the filmmaking process. Machine learning, particularly regression analysis, has proven indispensable in this effort, as it allows for the analysis of extensive datasets and the identification of trends that traditional methods may overlook. Models like Lasso and Ridge Regression are especially effective for this purpose due to their ability to handle multicollinearity and overfitting, providing clear, actionable insights that support decision-making in the industry.

This research leverages the capabilities of Lasso and Ridge Regression models to forecast IMDb ratings based on a detailed dataset of movie attributes. By carefully examining factors such as budget, revenue, runtime, vote count, director, and lead actors, we aim to uncover the key relationships influencing audience perceptions of a movie's quality. We also plan to investigate the role of other variables, such as genre, release date, and critical reviews, in shaping IMDb ratings. Through a process of data cleaning, feature engineering, and model training, our goal is to deliver accurate predictions of IMDb ratings, offering filmmakers, production studios, and distributors valuable insights to refine their strategies and boost audience engagement. By understanding the drivers behind audience preferences, industry professionals can make informed choices that resonate with viewers, thereby fostering success in an increasingly competitive entertainment landscape.

NEED OF THE STUDY.

With the increasing prominence of digital platforms and the expansion of streaming services, the competition for viewers' attention has grown fiercer. Filmmakers, production companies, and distributors are increasingly relying on data-driven strategies to optimize their content and enhance its appeal. Predictive modeling serves as a powerful tool for understanding the many factors that influence a film's success, helping stakeholders make informed decisions throughout the filmmaking process. Machine learning, particularly regression analysis, has proven indispensable in this effort, as it allows for the analysis of extensive datasets and the identification of trends that traditional methods may overlook. Models like Lasso and Ridge Regression are especially effective for this purpose due to their ability to handle multicollinearity and overfitting, providing clear, actionable insights that support decision-making in the industry.

This research leverages the capabilities of Lasso and Ridge Regression models to forecast IMDb ratings based on a detailed dataset of movie attributes. By carefully examining factors such as budget, revenue, runtime, vote count, director, and lead actors, we aim to uncover the key relationships influencing audience perceptions of a movie's quality. We also plan to investigate the role of other variables, such as genre, release date, and critical reviews, in shaping IMDb ratings. Through a process of data cleaning, feature engineering, and model training, our goal is to deliver accurate predictions of IMDb ratings, offering filmmakers, production studios, and distributors valuable insights to refine their strategies and boost audience engagement. By understanding the drivers behind audience preferences, industry professionals can make informed choices that resonate with viewers, thereby fostering success in an increasingly competitive entertainment landscape

3.2 Data Analysis

The dataset offers a comprehensive set of features that are crucial for analyzing patterns in movie ratings and box office performance. It includes details such as the movie title, release date, budget, running time, director, lead actor, genre, language, and production companies. These attributes are vital in understanding and predicting a movie's rating, success, and overall popularity.

3.3 Research Methodology

The methodology outlines the approach taken to predict movie ratings using a large dataset of movie metadata. It covers key aspects such as the dataset itself, the steps involved in data preprocessing, feature engineering, and the selection of models. It also discusses the process of training and evaluating the models, while providing insights into the features within the dataset.

3.3 Technical Framework

Scaling techniques, such as Max scaling and Standard scaling, help improve model performance by adjusting the data's range and distribution. Max scaling normalizes data values to fall between 0 and 1, while Standard scaling standardizes the data by setting the mean to 0 and the standard deviation to 1. These methods help models better understand and learn from the data. Additionally, feature engineering enhances predictions by generating new features or extracting valuable information from existing ones. For instance, identifying the release season or the day of the week from the release date can provide insights into how audiences rate movies. Splitting datasets into training and testing sets (usually in a 4:1 or 8:2 ratio) is important for checking how _well models work. This way, models learn from one set of data and then get tested on another set they've never seen. After splitting, normalization makes sure all features are scaled properly. This is super important for sensitive models like regressions or neural networks.

Normalization scales features so they all fit within a similar range, often between 0 and 1, using a simple formula.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Here, scaled represents the scaled or normalized data value, is the original data value, and min and max are the minimum and maximum data values for that attribute, respectively. This preprocessing ensures that the data is appropriately formatted and scaled for input in the predictive model.

3.4Statistical tools and econometric models

This section elaborates the regression models which are being used to study from data towards predictions. The detail of methodology is given as follows.

3.4.2 Regression Models

Lasso regression, is also known as L1 regularization, it adds a penalty term to the standard linear regression cost function. This penalty term is proportional to the absolute value of the regression coefficients. Donnarumma By penalizing the magnitudes of the regression coefficients, Lasso regression encourages sparsity in the model, effectively selecting a subset of the most relevant

features while setting the coefficients of less important features to zero. This results in a simpler and more interpretable model, reducing the risk of overfitting.

Ridge regression, is also called as L2 regularization, it tries adds a penalty term to the linear regression cost function. However, unlike Lasso regression, the penalty term in Ridge regression is connecting to the square of the regression coefficients. This technique help in shrinks the coefficient towards zero, and help lesser their magnitudes without setting them exactly to zero. Ridge regression is effective in dealing with multicollinearity, where predictors are highly correlated, by spreading the impact of correlated features around multiple predictors.

By incorporating Lasso and Ridge regression techniques into the predictive model, we try to balance between model complexity and predictive performance. These regularization methods help simpler models with less parameters, thereby reducing the risk of overfitting and improving the model capacity to understand to unseen data. Through careful tuning of the regularization parameters, we aim to achieve optimal predictive accuracy while maintaining model interpretability

3.4 Proposed Solution

The proposed approach leverages machine learning techniques, specifically ridge and lasso regression, to predict movie ratings accurately. These regression methods are chosen due to their ability to handle high-dimensional data, mitigate overfitting, and select relevant features.

Ridge regression**,** a form of linear regression, is like adding a little extra something to regular math to improve our predictions. It's handy when there's a bunch of different factors influencing movie ratings. This helps_the penalty term which is proportional to the square of the magnitude of the coefficients, which helps to shrink them towards zero while still maintaining all of them in the model. This regularization technique is particularly useful when dealing with _multicollinearity, where predictors are highly correlated.

Lasso regression, imposes an L1 penalty on the coefficients, which leads to some coefficients being exactly zero. It zeroes out some of the less important factors, so we only pay attention to the significant ones. This comes in handy when we have a ton of data, but not all of it matters for predicting movie ratings. Ridge and lasso regression helps us to handle these // features effectively and identify the most influential ones.

The training process involves halve the dataset into training and testing sets to evaluate model performance. The models are trained on the training set and then evaluated on the testing set to assess their predictive accuracy >/ Metrics such as RMSE, MSE, and MAE are used to calculate the difference between actual and predicted ratings. Regularization parameters such as alpha are tuned to optimize model performance. For ridge regression, higher values of alpha lead to more regularization, while for lasso regression, the sparsity of the solution increases with higher alpha values.

Overall, the proposed models present a strong framework for predicting movie ratings by utilizing both ridge and lasso regression techniques. These models manage high-dimensional data efficiently and focus on selecting the most relevant features, offering meaningful insights into the factors that contribute to a movie's success.

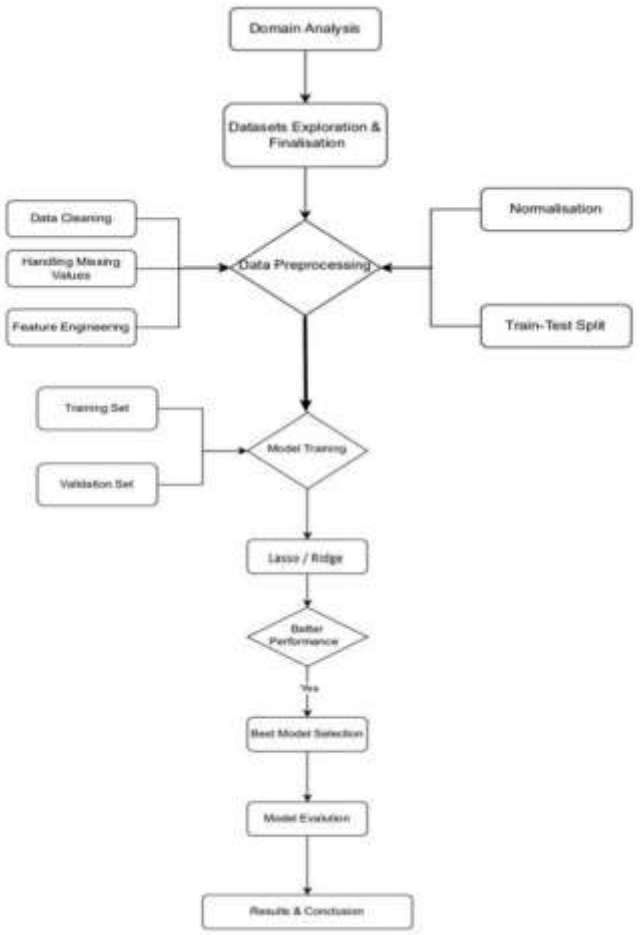


Fig. 1. Flow diagram

IV. RESULTS AND DISCUSSION

In this section, we evaluate the performance and efficiency of the Lasso and Ridge regression models. We assess how different levels of regularization impact the models' predictive accuracy. Additionally, we examine the effectiveness of various evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared score, to measure the overall performance of the models.

Mean Squared Error (MSE),

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - y_j)2$$

Mean Absolute Error (MAE),

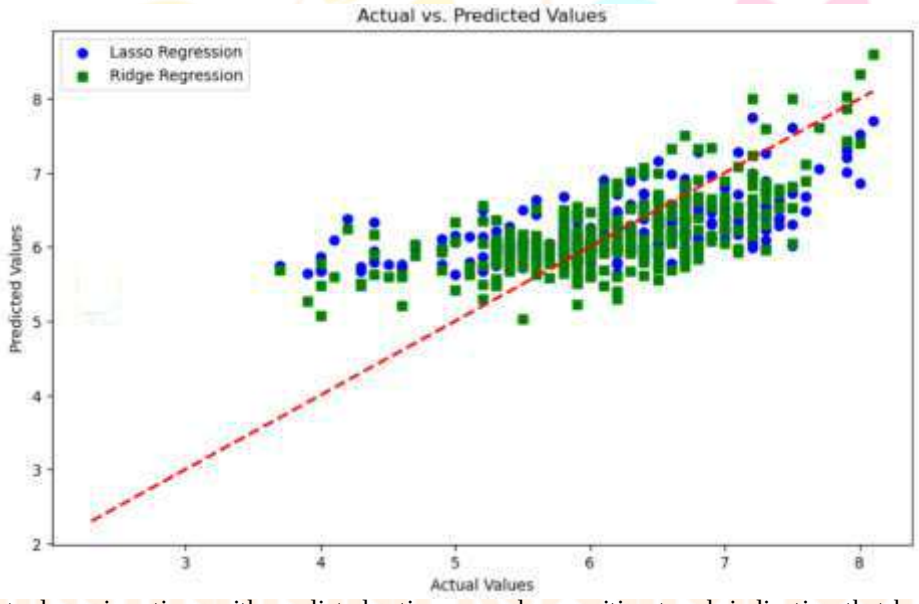
$$\frac{1}{n}\sum_{i=1}^{n}|(y_i - y_j)|$$

We saw two regression models: Lasso and Ridge.

Regression	MAE	MSE	R^2 Score
Lasso Regression	0.264	0.228	0.977
Ridge Regression	0.242	0.195	0.980

Table 4.1: Performance Metrics of Regression Models

The results suggest that both models demonstrate strong predictive capabilities. The Ridge regression model performs slightly better than the Lasso model, showing lower values for mean absolute error and mean squared error. Despite this, both models achieve high R-squared scores, indicating their effectiveness in capturing the variability in movie ratings data. These findings offer valuable insights into the models' predictive abilities, which can help filmmakers, producers, and distributors make informed decisions about a movie's potential success, ultimately guiding resource allocation and strategic planning within the film industry.



The plot comparing actual movie ratings with predicted ratings reveals a positive trend, indicating that both the Lasso and Ridge regression models performed well in predicting movie ratings in most cases. Overall, the models appear to be quite accurate for the majority of movies, although there is still room for improvement.

REFERENCES

[1] Rijul Dhir and Anand Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," May 2019. DOI: 10.1109/ICSCCC.2018.8703320

[2] Carl Jernbäcker et al., "Predicting movie success using machine learning techniques," Stockholm, Sweden, 2017.

[3] Steven Yoo, R. K. Kanter, D. C. Cummings, and A. Maas, "Predicting Movie Revenue from IMDb Data," 2011.

[4] Mohanbir S. Sawhney and Jehoshua Eliashberg, "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," vol. 15, no. 2, pp. 113–131, 1996.

[5] Márton Mestyán, T. Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PloS one, vol. 8, p. e71226, 08 2013.

[6] Jeffrey S. Simonoff and Ilana R. Sparrow, "Predicting Movie Grosses: Winners and Losers, Blockbusters and Sleepers," Chance, vol. 13, no. 3, pp. 15-24, 2000.

[7] M. H. Latif and H. Afzal, "Prediction of movies popularity using machine learning techniques," International Journal of Computer Science and Network Security (IJCSNS), vol. 16, no. 8, p. 127, 2016.

[8] S. Pramod, A. Joshi, and A. Mary, "Prediction of movie success for real world movie dataset," Int. J. of Advance Res., Ideas and Innovations in Technol, vol. 3, no. 3, 2017.

[9] R. Parimi and D. Caragea, "Pre-release box-office success prediction for motion pictures," in International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, 2013, pp. 571–585.

[10] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, "Movie success prediction using data mining," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017, pp. 1–4.