Journal of Science, Computing and Engineering Research (JSCER)
Volume-7, Issue-10, October 2024.
DOI: *https://doi.org/10.46379/jscer.2023.07080*
ISSN: 2708-1079

BOX OFFICE PREDICTION USING RIDGE AND LASSO REGRESSION

Rajeswari G[1], Santhosh S.A[2], Yuvan Sankar S.K.G[3], Ramana C[4], Sharukeshavalingam A[5]

[1]Assistant Professor, Dept of Computer Science and Engineering,

[2-5]UG students, K L N College of Engineering, Sivagangai, Tamil Nadu

Article Information

Received : 16 Oct 2024
Revised : 20 Oct 2024
Accepted : 21 Oct 2024
Published : 31 Oct 2024

Corresponding Author:
G. Rajeswari
Email: rajeegs@gmail.com

Abstract— Accurately predicting the field workplace overall performance of films is a important venture within the leisure industry because it drives selections concerning production, marketing, and distribution. This paper proposes the usage of Ridge and Lasso regression techniques to are expecting container workplace sales primarily based on diverse capabilities along with production budget, genre, forged, recognition, director history, launch date, and greater. Ridge and Lasso are regularized regression techniques that cope with problems which include overfitting and multicollinearity with the aid of penalizing large coefficient values. This paper compares the overall performance of those fashions, offering insights into how they handle excessive-dimensional records and enhance prediction accuracy.

Keywords: ***Box Office Prediction, Ridge Regression, Lasso Regression, Regularization, Multicollinearity, Feature Selection, Machine Learning, Overfitting.***

Citation: Rajeswari G, Santhosh S.A, Yuvan Sankar S.K.G, Ramana C, Sharukeshavalingam A, "Box Office Prediction Using Ridge And Lasso Regression", Journal of Science, Computing and Engineering Research, 7(10), October 2024.

I. INTRODUCTION

Box office sales is one of the most excellent indications of how a film has economically performed and always has an immediate impact on the strategic decisions to be made by utilizing studios and investors. Prognostication of this type of sale is, however very challenging, due to the large selection of factors that can influence the outcome of any film performance. These are production finances, megastar electricity, advertising campaigns, launch timing, and social media buzz, all of which lead to the film's box office achievement, thus making the problem complicated and multidimensional.

Traditional predicting models of field workplace performance are covered in linear regression, decision trees and even some complicated ensemble methods, random forests. With an extremely high-dimensional data set, with the many features and multicollinearity, the described fashions generally overfit. In order to counteract this effect, regularized regression fashions including Ridge and Lasso regression are used. Consequences are added to those models in order to reduce large coefficients caused by multicollinearity, therefore improving generalization performance.

In the present work, we will apply Ridge and Lasso regression methods in predicting container office revenue. The ridge regression turns out quite powerful in controlling overfitting by applying an L2 penalty that is the square of the coefficients while applying L1 penalty, that is, the absolute value of the coefficients, in lasso regression, we not only prevent overfitting but also carry out characteristic choice through shrinking some coefficients to zero. This allows Lasso to select the most relevant functions for the predictive task of field office prediction.

II. RESEARCH AND FINDINGS

Research within the field of workplace prediction relied on much system learning and statistical methodology. Initial approaches based their forecasting on linear regression with the assistance from several variables such as budget, genre, and cast involved in a movie. However, these models have frequently been restricted by their susceptibility to multicollinearity and overfitting. Multicollinearity occurs when predictor variables are almost perfectly correlated with each other, which results in unreliable coefficient estimates in conventional linear regression models.

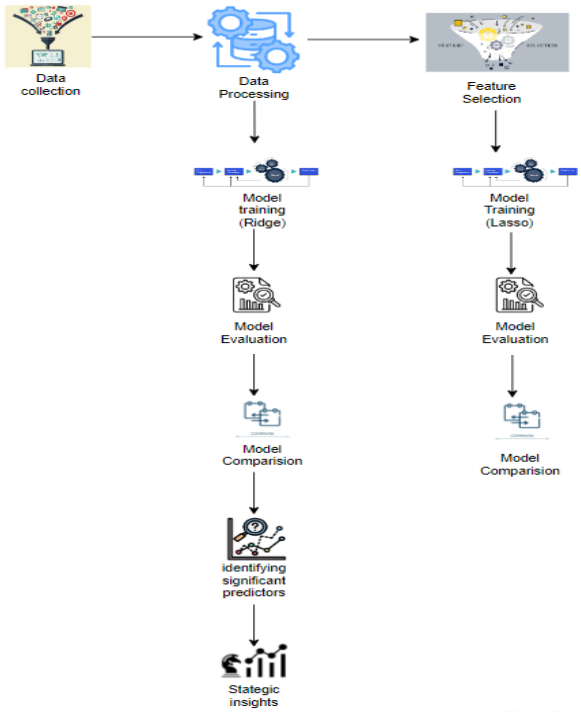
Later works focused on ensemble methods and Random Forests and Gradient Boosting, which increase prediction by using the combining multiple decision bushes to reduce variance. Though those models have a higher accuracy, they suffer from interpretability issues because the relationships of individual features to the outcome become hidden inside the complicated tree structures.

The interesting thing is that regularized regression models-Ridge and Lasso-offer an appealing alternative practically; with the evils of overfitting and multicollinearity not being delayed attacks while allowing model interpretability. With the introduction of a penalty on huge coefficients, Ridge regression has proven particularly powerful in controlling overfitting and resulting in even more robust models. Lasso regression goes further with this

function by not only providing least penalizing the significant coefficients but automatically functioning as some sort of function selection by making some coefficients equal to zero.

This paper extends this already existing literature where the authors apply Ridge and Lasso regressions on the task of prediction of box office sales, putting their relative performances against the accuracy of the models, feature selection, and model interpretability.

III. SYSTEM IMPLEMENTATION



A. Data Collection

To analyze this, a dataset containing the various functions involved with film making, launching, and marketing was collated from public databases, including IMDb and Box Office Mojo. The dataset encompasses the following capabilities:

- Production Budget: The entire cost incurred in the making of the movie.
- Genre: Categories such as Action, Drama, Comedy, and so on playing fields..
- Release Date: Information regarding whether the movie hit theaters during the holiday season or otherwise.
- Release Date: Information regarding whether the film changed into released throughout a holiday season or now not.
- Past Box Office Performance: General performance statistics of movies that belong in similar genres, budgets, or casts.

B. Data Preprocessing

For version training, very massive information preprocessing was conducted to obtain the high quality of the data and the performance of the model. The following steps included:

• Handling Missing Values: The movies with missing entries were either imputed or excluded based on the extent of missingness involved

• Feature Encoding: The categorical variables plus genre and fabricated were converted into a numeric version using techniques like one-hot encoding

• Outliers Detection: In the case of continuous data, outliers in container office sales were picked up and either excluded or capped to limit the distortion of model estimates.

• Feature Normalization: The features of manufacturing budget and box office sales were normalized to ensure that they appear within a comparable range and no feature dominates the model.

C. Ridge and Lasso Regression

Ridge Regression: Ridge regression adds an L2 penalty on the magnitude of the coefficients, so that no single person predictor can have too much influence on the model.

Lasso Regression: The Lasso regression applies an L1 penalty that is maximally shrinking large coefficients but also forces some to be precisely zero, while doing correct function selection. The L1 penalty helps identify the most important predictors; therefore, the version is easier to interpret and lowers the number of inappropriate or redundant functions.

Model Training and Hyperparameter Tuning Ridge and Lasso regression models were trained on the given dataset. A move-validation strategy was used to avoid overfitting. Grid search over lambda values was performed to find best regularization parameter values for each of the models. The idea of grid search is that multiple lambda values are tried out, and the smallest validation error is selected from among them.

IV. RESULTS AND DISCUSSION

A. Model Performance Metrics

The overall performance of each fashions were evaluated based on the following metrics:

• MSE: To calculate the average squared difference between the identified and the anticipated container office revenues.

• R-squared (R^2): To determine the percentage of the variance in the field office revenue that is explained with the help of the model.

• Cross-Validation Score: To test the version's ability to generalize to unseen records by utilizing acting k-fold pass-validation.

B. Ridge vs. Lasso Comparison

• Both Ridge and Lasso fashions proved to be improved generalization, away from the standard linear regression. Due to their regularization properties, they proved to be effective in relation to existing linear regression models. Ridge regression was strong at reducing overfitting, especially when multicollinearity became extreme, because it retained all predictors but shrunk their coefficients to minimize their impact.

• Lasso regression, however, worked well in feature selection. It reduced many coefficients to zero by removing redundant or inappropriate functions and provided a more parsimonious version. This ability at the feature selection aspect makes Lasso very effective in dealing with high-dimensional statistics where inappropriate features make it tough to understand the true relationships among the predictors and the outcome..

C. Feature Importance

• The Lasso model's two most critical features--the largest determining container workplace sales--applied to protect the production budget range, lead actor identification, and time of release (or lack thereof, in alignment with a major vacation or summer). Other functions, including directors' music file as well as social media presence, also considerably factored into the version's quality.

• Ridge regression, on the other hand, had retained all capabilities but whose coefficients had been more uniformly shriveled, which made it harder to discern the relative significance of each predictor.

V. CONCLUSION

This paper has shown that both Ridge and Lasso regression are effective tools for the prediction of box workplace sales. Ridge regression is very useful when multicollinearity becomes a problem, whereas Lasso regression excels in feature selection, providing a more interpretable version. These models could help movie studios and buyers better to make more informed choices that would lead to the appropriate sources of advertisement and marketing, optimal time launches, and more strategic investments. Future research may focus on adding more records resources, for example social media sentiment analysis or audience demographic statistics, to further amplify the predictive energy of those models. Further utility of hybrid models, such as Elastic Net that combines both L1 and L2 consequences, may potentially offer similar upgrades in model overall performance.

REFERENCES

[1]. .The Taiwanese Film Market Case Study , Shih-Hao Lu , Hung-Jen Wang , Anh Tu Nguyen - Machine learning Application on Box Office Revenue Forecasting , Springer Access , Volume : 483, pp : 384 – 402, (2023)

[2]. Dawei Li , Zhi-Ping-Liu, MDPI Reference on predicting Box-Office Markets with Machine Learning Methods, Volume : 24, Number : 711,(2022)

[3]. P. Nirmala, T. Manimegalai, J. R. Arunkumar, S. Vimala, G. Vinoth Rajkumar, Raja Raju, "A Mechanism for Detecting the Intruder in the Network through a Stacking Dilated CNN Model", Wireless Communications and Mobile Computing, vol. 2022, Article ID 1955009, 13 pages, 2022. https://doi.org/10.1155/2022/1955009

[4]. Boning JIANG1, Research and Prediction of Influential Factors of Film Box Office: Based on Machine Learning

[5]. Algorithms Such as XGB, LGB and CAT, PP: 749 – 758 ,(2024)

[6]. Runzhi Xie , Mengke Wang Empirical Analysis of Factors Influencing Box Office Revenue Of Imported Animated Films, Resource Data Journal, PP : a hundred and forty – 156 , Volume : three, (2024)