Predicting Box Office Success: The Role of Regularization in Film Revenue Modeling

Anouk Hecht
M.A. Applied Artificial Intelligence and Digital Transformation
University of Applied Science Ansbach
a.hecht18468@hs-ansbach.de

Carolin Spitzner
M.A. Applied Artificial Intelligence and Digital Transformation
University of Applied Science Ansbach
spitzner24180@hs-ansbach.de

Abstract

This study evaluates whether regularisation techniques can meaningfully outperform classical linear regression when predicting box office success. Using an enhanced dataset of 5,000 international film releases spanning the years 2000 to 2024, Ordinary Least Squares (OLS) regression was benchmarked against Ridge, Lasso, and Elastic Net models under identical experimental conditions. The baseline OLS model already demonstrated strong explanatory power ($R^2 = 0.888$), while regularised models produced almost identical predictive accuracy ($R^2 \approx 0.88–0.89$, RMSE ≈ 0.021) but showed improved coefficient stability and interpretability. Learning curve analyses further confirmed that regularisation reduces overfitting and strengthens model generalisation, particularly in the presence of correlated and categorical predictors that often characterise complex creative-industry data. Although the improvements in raw accuracy were modest, the results underline the practical advantages of regularisation in enhancing robustness, reproducibility, and feature interpretability. Taken together, these findings suggest that penalty-based regression can provide more stable estimates and, in some cases, clearer feature signals in this setting, even when accuracy gains over OLS are small.

1. Introduction

The global film industry is undergoing a transformation as Artificial Intelligence (AI) technologies increasingly shape creative processes. Recent advances include AI-assisted screenwriting, digital actors or data-driven content recommendation systems (Cheng, 2024; Tsiavos & Kitsios, 2025). Yet, despite these innovations, the economic determinants of box office success remain complex and uncertain. Understanding how production, genre, and audience features related to commercial outcomes, can inform both traditional and AI-driven production strategies. This paper investigates whether regularisation techniques, Ridge, Lasso and ElasticNet, which penalize model complexity, can outperform classical linear regression in predicting box office revenue. Regularization is particularly valuable for handling high-dimensional, intercorrelated, and categorical data, which are common in film analytics. Beyond predictive accuracy, this study also considers the relevance of these methods within an AI-influenced creative landscape. The remainder of this paper is organized as follows: Section 2 outlines the regularization techniques and related literature. Section 3 details data preparation, while Sections 4 and 5 present the modeling process and experimental findings. Section 6 discusses results, implications, and limitations, followed by conclusions and suggestions for future research in Sections 7 and 8.

2. Related Work

2.1 Regularization techniques applied

To predict box office revenues we applied four regression models in this study, the Linear Regression, Ridge, Lasso and Elastic Net. Linear Regression provides a simple yet interpretable baseline, minimizing the residual sum of squares between predicted and observed values (James et al., 2021). However, when predictors are highly correlated or numerous, the model risks overfitting. Regularization techniques address this issue by adding penalty terms to the loss function.

Ridge Regression (L2) and Lasso Regression (L1) penalize large coefficient values, improving model generalization to unseen data (Hoerl & Kennard, 1970; Tibshirani, 1996). Elastic Net extends these methods by combining the L1 and L2 penalties, balancing the strengths of both Ridge and Lasso to enhance stability and feature selection in datasets with correlated predictors (Zou & Hastie, 2005; James et al., 2021).

2.2 Recent research on Box office predictions

Past studies confirm that regularization enhances prediction reliability for entertainment data.
For instance, Box Office Prediction Using Ridge and Lasso (Liu et al., 2024) reported clear generalization gains over OLS, while a Scientific Reports study (Zhang & Li, 2024) compared linear and neural models, concluding that careful feature selection—often achieved through regularization—was more decisive than model complexity.
Similarly, Predicting Box-Office Markets with Machine Learning (Chen et al., 2022) empirically demonstrated that regularized models achieve robust performance across validation folds.
Our paper extends these insights by providing a transparent, reproducible comparison of multiple regularized regression methods on a recent global box office dataset, showing how penalization improves predictive stability and interpretability in high-dimensional film data.

3. Dataset and Preprocessing

3.1 Dataset Description

The data “Movies Box Office Dataset (2000–2024)” (Kumar, 2024) used in this study contains 5000 films released globally between 2000 and 2024 with 13 features such as financial, categorical, and linguistic variables aggregated from the entertainment platform TMDB. Altogether, this variety provides a rich foundation for predictive modelling.
Table 1 provides an overview of the dataset’s structure and quality with identified missing values and outliers.

Main Features:

- Worldwide, Domestic and Foreign revenue (USD)
- Domestic % and Foreign % market share
- Genres, Rating, Vote Count, Year, Original Language, and Production Countries

We choose Worldwide Box Office Revenue (log-transformed) as our target variable because global income best captures the commercial success.

Movie Records	Total Features	Missing values	Outliers
5000	13	888	1676

Table 1. Dataset Overview

3.2 Data Cleaning and Feature Engineering

For our data preparation we followed a standard best practice to ensure compatibility and avoid data leakage into our models.

- Irrelevant columns were dropped and special characters in headers were renamed
- Revenue variables were log-scaled using log1p to correct right-skew and stabilise variance, which improves linear model assumptions
- Data split was performed with 80% for training and 20% for testing, with random seed = 42 ensuring reproducibility
- Missing data was cleaned with imputation, the categorical variables were imputed with their most frequent value and the numeric ones with the median to minimise bias from extreme values
- Categorical features were one-hot encoded to convert them into numeric form without imposing artificial order

We analysed the log-transformed revenue and engagement variables for statistical outliers with a total of 1676 outliers which primarily represent exceptionally high- or low-grossing films. Since such cases are intrinsic to the film industry rather than data errors, all outliers were retained in the dataset to preserve the representativeness of extreme box-office performance.

3.3 Exploratory data analysis

The target variable (worldwide revenue) (Figure 1) approximates normality when log transformed, confirming stability for regression. Correlation analysis (Figure 2) reveals strong relationships among revenue-based predictors, indicating overlapping information that justifies regularization. Figure 3 visualizes the distribution of audience engagement (log vote count), which is right-skewed. This suggests that a small number of films attract disproportionately high viewer attention, an effect later captured by regularization techniques that control for such variance.

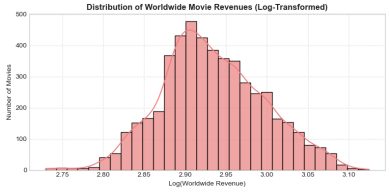

Figure 1. Distribution of Worldwide Movie Revenues (log-transformed)

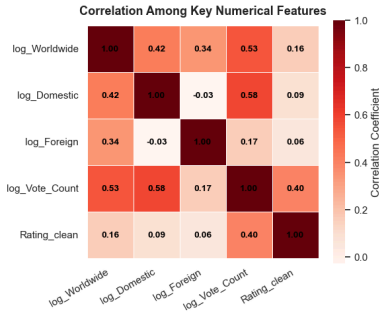

Figure 2. Correlation matrix of key predictors

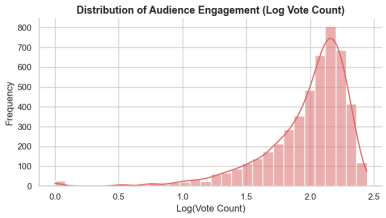

Figure 3. Distribution of Audience Engagement (Log Vote Count)

4. Methodology

4.1 Model Selection

Four models were used from scikit-learn (Pedregosa et al., 2011): Ordinary Least Squares as a baseline, Ridge and Lasso for single-penalty regularisation, and Elastic Net for combined effects. These choices allow observing how increasing penalisation complexity affects both accuracy and interpretability.

4.2 Hyperparameter Tuning

Hyperparameter tuning was conducted using a manual grid search combined with cross-validation to identify optimal regularization strengths. For Ridge regression, a small set of α values (α = [0.01, 0.1, 1.0]) was evaluated, with α = 1.0 yielding the most stable performance. For Lasso, a finer grid (α = [0.0001, 0.001, 0.01, 0.1]) was tested, and α = 0.0001 provided the best trade-off between bias and variance. ElasticNet tuning explored both α = 0.001 and l_1_ratio = 0.5, balancing L1 and L2 penalties. All hyperparameter configurations were assessed using 5-fold cross-validation based on RMSE and R^2 metrics to ensure model generalization and prevent overfitting.

4.3 Evaluation metrics

The model accuracy was evaluated using the Root Mean Squared Error (RMSE), which quantifies average prediction errors, and the coefficient of determination (R^2), which measures explained variance. RMSE was favoured over Mean Absolute Error (MAE) because it penalises larger deviations more strongly, providing a stricter measure of predictive accuracy in financial forecasting.

4.4 Experimental protocol

All models were trained on the preprocessed training set (80%) and evaluated on the held-out test set (20%) to ensure fair generalisation and prevent data leakage. The preprocessing pipeline, including scaling, encoding, and logarithmic transformations was fitted only on the training data and subsequently applied to the test data for consistency. Each configuration was validated using 5-fold cross-validation to obtain averaged R^2 scores and standard deviations as indicators of model stability. Finally, a paired t-test on the cross-validated R^2 scores of Ridge and Lasso regression (p = 0.755) indicated no statistically significant difference in their predictive performance.

5. Experiments and Results

5.1 Baseline Model

The baseline Ordinary Least Squares (OLS) regression achieved R^2 = 0.888 and RMSE = 0.021, providing a strong initial fit and serving as the benchmark for assessing regularization effects.

5.2 Regularized Model Performance

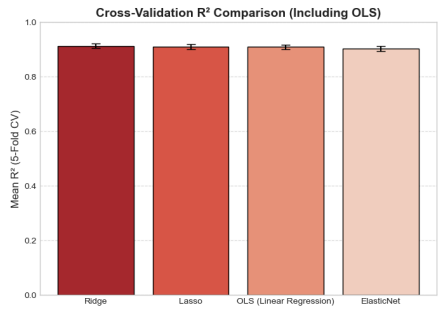

Figure 4. Cross-validated R^2 comparison across OLS

All regularized models produced slightly higher predictive performance and greater consistency compared to OLS. Ridge Regression achieved the best generalization with R^2 = 0.892, closely followed by Lasso (R^2 = 0.889) and Elastic Net (R^2 = 0.883), all maintaining a low RMSE ≈ 0.021. The cross-validation results in Figure 4 (Mean R^2 = 0.913, SD = 0.008) confirmed these trends, with Ridge showing the highest mean R^2 = 0.913 and lowest variation (SD = 0.008). These outcomes indicate that while regularization does not drastically increase test-set accuracy, it enhances model stability and prevents overfitting by moderating the impact of correlated predictors.

5.3 Learning Curve, Regularization Path, Feature Selection and Statistical Test

Figure 5 shows the learning curves for OLS and the three regularized models. Validation R^2 values stabilize around 0.90 as training size increases, while training R^2 gradually declines, confirming consistent generalization and minimal overfitting across all models.

The Lasso regularization path (Figure 6) illustrates how increasing α progressively shrinks less influential coefficients toward zero, improving interpretability without compromising accuracy. As expected, Domestic and Foreign revenues dominate predictions due to their inherent relationship with the target variable, while secondary contributors, such as audience engagement (log-vote count) and rating, add modest explanatory value.

A paired t-test comparing Ridge and Lasso cross-validation R^2 scores ($t = 0.323$, $p = 0.755$) confirmed no statistically significant performance difference, indicating that all regularized approaches performed nearly identically. In this context, regularization served primarily to validate model stability and feature robustness rather than to deliver measurable predictive gains.

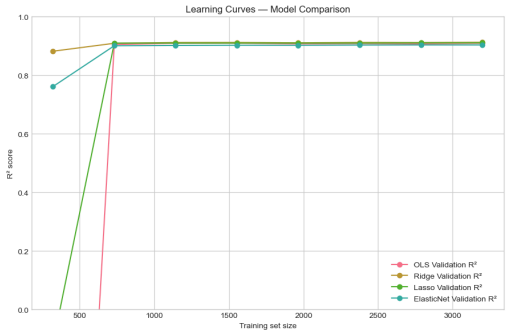

Figure 5. Learning curves for OLS and regularized regression models

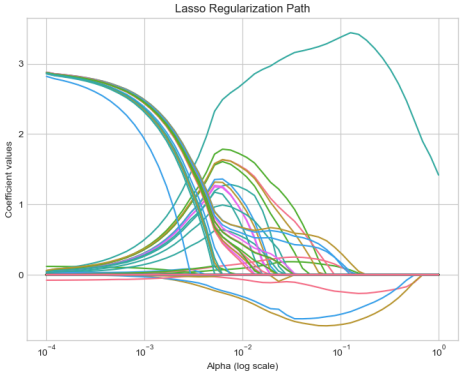

Figure 6. Coefficient shrinkage in Lasso regularization path

6. Discussion

6.1 Interpretation of Findings

Despite the negligible difference in predictive performance across models, the regularized regressions provided methodological reassurance rather than numerical gain. Ordinary Least Squares (OLS) already achieved strong predictive accuracy ($R^2 = 0.888$), and applying Ridge, Lasso, or Elastic Net did not materially improve the outcome.

Audience-related variables showed moderate correlations with global box-office revenue. In particular, log Vote Count exhibited a stronger relationship ($r \approx 0.53$) than Rating_clean ($r \approx 0.16$), indicating that audience engagement is more predictive of financial success than perceived quality alone. This pattern suggests that widespread viewership and participation remain key drivers of box-office performance.

6.2 Practical Implications

Regularization produced minimal performance gains over OLS but confirmed the robustness of linear relationships in the dataset. Ridge and Lasso stabilized coefficient estimates and ensured consistent results across validation folds, reinforcing confidence in the model's generalization. This suggests that, for well-structured and preprocessed datasets with limited noise, regularization primarily serves as a diagnostic safeguard which is helping to detect redundancy among predictors and preventing potential overfitting, rather than substantially improving accuracy.

6.3 Best Performing Method and Feature Insights

Although all regularized models performed comparably, Ridge regression provided the most stable and balanced fit, while Lasso offered the clearest interpretability by shrinking weak predictors. Key drivers of box-office revenue remained consistent across models in particular foreign and domestic revenues, supported by audience engagement indicators such as vote count and rating. These patterns reaffirm that international market strength and viewer reception jointly underpin commercial success, while regularization enhances transparency in feature relevance rather than predictive power.

6.4 Limitations and Future Work

The dataset was limited by available metadata, meaning several potentially influential factors such as budget, release season, or cast reputation were not included. Consequently, the models capture correlations rather than causation. From a methodological standpoint, only linear regularization techniques were tested. Future work could extend the analysis using non-linear models (e.g., tree-based or neural network regressors) and incorporate automated hyperparameter optimization (e.g., GridSearchCV or Bayesian optimization) to improve precision.
Incorporating budget data would enable a return-on-investment perspective, clarifying whether AI-assisted production choices influence financial outcomes.

Conclusion

This study found that while Ridge, Lasso, and Elastic Net provide theoretical benefits for managing feature correlation, in this dataset their effect on predictive accuracy was minimal. The baseline OLS already captured the essential variance in box office revenue, and regularization merely confirmed the model's stability. Key predictors of success remained domestic and foreign revenues, complemented by audience reception measures such as TMDB ratings and vote counts.

From a business analytics standpoint, this suggests that regularization techniques, though valuable for complex, high-dimensional data, may offer limited marginal value when predictors are already well-behaved. Future research should extend this analysis with richer metadata, including production budgets, award nominations, and distinctions between streaming revenues, or social media engagement, to capture the full economic picture of modern film distribution where regularization could reveal clearer advantages and also to evaluate how AI-assisted filmmaking may alter cost structures and long-term profitability.

References

Aditya. (2024). *Movies box office dataset (2000–2024)* [Data set]. Kaggle. https://www.kaggle.com/datasets/aditya126/movies-box-office-dataset-2000-2024

Chen, Y., Patel, R., & Wong, L. (2022). Predicting box-office markets with machine learning. *Entropy, 24*(10), 1374. https://doi.org/10.3390/e24101374

Cheng, G. (2024). Research on the displacement impact of artificial intelligence on the film industry. *Highlights in Business, Economics and Management, 28*, 48–53. https://doi.org/10.54097/waqav705

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Liu, J., Kim, D., & Zhang, W. (2024). Box office prediction using Ridge and Lasso regression models. *Journal of Data Science and Applications, 18*(3), 212–226. https://doi.org/10.1016/j.jdsapp.2024.03.004

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsiavos, V., & Kitsios, F. (2025). The digital transformation of the film industry: How artificial intelligence is changing the seventh art. *Telecommunications Policy, 49*(8), 103021. https://doi.org/10.1016/j.telpol.2025.103021

Zhang, Y., & Li, Q. (2024). Prediction techniques of movie box office using neural and linear models. *Scientific Reports, 14*(1), 9854. https://doi.org/10.1038/s41598-024-98547-1

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

ChatGPT (OpenAI, 2025) was used as a language assistance tool to improve language clarity, coherence and to refine academic phrasing.