Decoding IMDb Scores and Understanding the Key Factors Behind Movie Ratings with Regression

MGSC401 - Midterm Project

Kaibo Zhang
Email: kaibo.zhang@mail.mcgill.ca

Yanfei Wu
Email: yanfei.wu@mail.mcgill.ca

Tia Qiu
Email: tian.qiu3@mail.mcgill.ca

Wilson Chen
Email: zexian.chen@mail.mcgill.ca

Xuechen Hong
Email: xuechen.hong@mail.mcgill.ca

March 10, 2025

Abstract

This project evaluates regression modeling techniques to predict IMDb ratings for twelve upcoming films using a comprehensive historical movie dataset. Through careful exploratory data analysis (EDA), targeted variable selection, and appropriate data transformations and interactions, the developed model successfully identified key factors influencing audience ratings, including film attributes, cast characteristics, and production details. During the modeling process, special attention was given to addressing statistical issues such as collinearity, heteroskedasticity, and overfitting, ensuring both robustness and interpretability. Cross-validation and out-of-sample testing procedures confirmed the stability and reliability of the predictive performance across different subsets of data. Overall, this project highlights the practical value of statistical modeling techniques in accurately forecasting audience preferences, providing meaningful insights into the drivers behind IMDb ratings and offering guidance for future film releases.

1 Introduction

We explored the performance of regression modeling techniques on the IMDb Winter 2025 Dataset [1], which comprises detailed attributes for approximately 2,000 movies, including production budgets, film genres, cast characteristics, audience engagement metrics, and release year details. Preprocessing steps involved checking for NaN values and removing unrelated variables (e.x., id's) in preparing the dataset for modeling. Through careful EDA and rigorous cross-validation, we developed a robust predictive model capable of accurately forecasting IMDb scores for 12 newly released movies in 2025. The final model demonstrated strong predictive reliability and interpretability, effectively capturing the key drivers influencing audience ratings.

2 Exploratory Analysis

2.1 Data Cleaning and Preprocessing

Preprocessing involved removing non-informative identification columns and addressing categorical variables with excessive levels, which could distort statistical tests. An ANOVA test initially suggested all categorical features were significant, an implausible result given the data structure. Investigation revealed that some features had over 1,500 unique levels with highly imbalanced distributions, likely inflating Type I errors. To mitigate this, categories with fewer than ten occurrences were merged into an "Other" category, as seen with the "Language" variable, where only "English" and "Other" remained. This transformation improved interpretability, with boxplots confirming distinct IMDb score distributions between the two groups (Figure 1). Additionally, columns like "Director" and "Production Company" were dropped due to excessive variability, which could complicate model performance without providing meaningful insights.

2.2 Analysis of Dependent Variable IMDb Score

The IMDb score is a user-generated rating that reflects audience perceptions of a film's quality, measured on a scale from 1 (worst) to 10 (best). Its distribution is left-skewed, with most ratings concentrated between 5 and 8, indicating a bias toward mid-to-high scores (Figure 2). The center around 6-7 suggests that audiences generally rate films positively, likely due to selection bias. Viewers tend to watch films they expect to enjoy. The spread is limited, with few extreme ratings, reflecting industry trends where poorly rated films receive less visibility and high-budget productions maintain a baseline quality. Psychological and social factors also shape this distribution. Central tendency bias leads users to avoid extreme ratings unless a film is exceptionally good or bad. Additionally, review aggregation and pre-release marketing influence audience perception, reinforcing clustering around moderate scores. The scarcity of very low ratings may also result from poor films receiving fewer votes, further skewing the observed distribution.

2.3 Continuous Variable Distributions and IMDb Score Relationships

2.3.1 Distribution Insights

Figure 3 displays the distributions of the continuous variables in the Movie dataset. The distribution of "Movie Budget" appears relatively uniform, indicating that films are produced across a wide financial spectrum rather than clustering around specific budget levels. This variability is expected, as budgets are heavily influenced by genre, production scale, and marketing expenses. In contrast, "Duration" follows a right-skewed distribution, with most films falling between 90 and 120 minutes, aligning with industry norms for theatrical releases. While outliers with significantly longer runtimes exist, they are rare, likely due to audience retention concerns and theater scheduling constraints.

Two variables, "Number of Articles" and "Movie Rank 2023 by IMDbPro", exhibit extreme right skewness,

meaning that most films receive little media coverage while a select few dominate in terms of publicity and IMDb rankings. This follows a power-law distribution, where a small number of highly anticipated films, often blockbusters, capture a disproportionate share of public attention. The disparity in media exposure suggests that marketing strategies and franchise reputation heavily influence a film's visibility. The presence of extreme distributions like this may imply the need for variable transformations to ensure that the relationship captured by the model is well represented, as highly skewed predictors can distort regression coefficients, reduce model interpretability, and violate assumptions of linearity, leading to biased or unstable predictions.

2.3.2 Relationships with IMDb Score

The relationship between "Movie Budget" and IMDb score is nearly flat, implying that larger financial investments do not necessarily translate into higher audience ratings (Figure 4). This is likely because while bigger budgets enable improved production quality, other factors such as storytelling and audience engagement play a more significant role in determining reception.

Movie duration exhibits a nonlinear relationship with IMDb scores, where very short films (under 80 minutes) tend to receive lower ratings, but standard-length films (90-120 minutes) stabilize at higher scores. This suggests that films below a certain runtime may feel underdeveloped, while excessively long runtimes could test audience patience, leading to mixed reviews. For "Number of Articles", the relationship with IMDb scores is curved, where moderate publicity correlates with higher ratings, but excessive media attention may lead to declining scores. This could indicate that while promotional efforts boost engagement, extreme media coverage—possibly linked to controversy or franchise fatigue—does not always translate into positive reception. A similar non-monotonic trend appears in the IMDb Movie Meter rankings, where highly ranked films (low "Movie Rank 2023 by IMDbPro" values) initially correlate with higher IMDb scores, but mid-ranked films see a decline before ratings rise again at the extreme tail. This suggests that while popularity increases visibility, some widely discussed films attract divisive opinions, leading to inconsistent audience ratings.

In short, nonlinear effects in duration and media exposure suggest that audience perception follows more complex patterns than simple linear relationships. Future modeling approaches should incorporate nonlinear transformations or interaction terms to better capture these dynamics and improve predictive accuracy.

2.4 Exploratory Analysis of Categorical Predictors and IMDb Scores

The boxplots (Figure 5) illustrate how IMDb scores vary across categorical predictors, revealing differences in predictive power. While some categories show clear separation, others exhibit substantial overlap, suggesting limited explanatory value in a regression model.

For film colour, black-and-white movies have a higher median IMDb score and lower variance than colour films, likely due to their association with classic cinema and arthouse productions. Colour films display a broader range of scores, reflecting the diversity of modern filmmaking, from blockbusters to lower-rated commercial releases.

A similar trend emerges in "Language" and "Country." English-language films show greater variance, likely due to the sheer volume and variety of productions, while non-English films tend to have higher median scores, potentially due to selection bias—only the most successful international films gain mainstream visibility. At the country level, USA-produced films have the highest variance, capturing both major Hollywood productions and smaller independent films, whereas *France and Germany* exhibit more concentrated distributions, suggesting more consistent quality perceptions. The influence of directors appears less pronounced. While notable figures such as *Spielberg, Eastwood, and Soderbergh* show distinct distributions, their overlapping ranges suggest that directorial influence alone is not a strong predictor of IMDb scores. Finally, "Maturity Ratings" (*PG-13*, *PG*, *R*) exhibit considerable overlap, with similar medians and interquartile ranges, indicating that maturity ratings do not strongly differentiate IMDb scores. This could lead to statistically insignificant coefficients in a predictive model, suggesting the need for transformation or grouping to enhance interpretability.

2.5 Analysis of Relationships between Independent Variable

The correlation heatmap (Figure 6) provides insights into the relationships among predictor variables and their influence on IMDb scores. At first glance, no variables exhibit particularly strong correlations (i.e., $r > 0.5$), though subtle patterns emerge. For instance, "Action Genre" and "Movie Budget" are positively correlated, suggesting that high-budget films are more likely to belong to the action genre. This likely stems from the genre's reliance on costly visual effects and large-scale productions. Their co-movement suggests an interaction effect, where budget's influence on IMDb scores may depend on genre classification. Including an interaction term could improve model accuracy by capturing this joint effect. "Movie Budget" has weak correlation with IMDb scores, reinforcing that higher spending does not directly translate to better ratings. However, budget correlates with factors like "Duration" (longer films often have higher budgets) and "Release Year" (newer films tend to have larger budgets due to inflation and evolving industry trends). Collinearity is also observed among maturity ratings (e.g., *PG-13* vs. *R*) and genre labels (e.g., drama and horror show negative correlation), indicating potential redundancy in modeling. Multicollinearity may inflate variance in regression models, necessitating dimensionality reduction or variable selection techniques. In addition, the weak correlations across most features suggest that IMDb scores may be driven by nonlinear or interaction effects rather than simple additive relationships.

3 Model Selection

3.1 Baseline Model and Stepwise Regression

To build the regression model for predicting IMDb scores, we performed an 80-20 train-test split to evaluate performance on unseen data. Initially, all predictors were included, establishing a baseline R^2 of 35.18%. Given the dataset's complexity, stepwise selection was used instead of best subset selection to improve efficiency and interpretability. This iterative method added or removed variables based on their contribution to model performance.

Using Bayesian Information Criterion (BIC), we tested forward, backward, and bidirectional stepwise selection, all of which converged on the same final model (Table 1). The reduced model achieved predictive performance similar to the full model while reducing unnecessary complexity (Figure 7).

3.2 Addressing Multicollinearity

To diagnose multicollinearity, we computed Variance Inflation Factors (VIFs), flagging variables with VIF greater than 5. The analysis revealed strong collinearity among maturity ratings (*PG-related* and *R*). Initially, we merged *PG* categories, but the coefficient remained insignificant. Summary statistics and insights from EDA confirmed that IMDb score distributions across these levels were nearly identical, making mean differences negligible. To mitigate multicollinearity and improve interpretability, we consolidated all maturity levels (*PG*, *PG-13*, and *R*) into a single category, redefining "PG and R" as the reference due to its larger sample size. To further assess collinearity, we performed an eigenvalue decomposition of the design matrix and analyzed eigenvectors corresponding to eigenvalues below 1. Since the trace of the matrix equals the sum of its eigenvalues, close zero eigenvalues indicate high correlation among predictors, leading to a close singular matrix and unstable coefficient estimates (Table 2). For instance, Eigenvector 1 exhibited high values for "Action Genre" and "Movie Budget," consistent with the EDA correlation heatmap. Based on this, we introduced an interaction term between these two variables, which was statistically significant and retained in the model. Additional insights from the decomposition were tested sequentially, and variables exceeding the 5% significance threshold were removed to prevent unnecessary complexity and overfitting.

3.3 Nonlinearity and Transformations

3.3.1 Transformation on Target Variable

After examining residual plots, we revealed a downward curve in the higher range of fitted values and dense tails on the left in the Q-Q plot, indicating violations of normality assumptions. Given the left-skewed distribution of IMDb scores, we first applied a square root transformation as an initial trial. To further refine the transformation, we conducted a Box-Cox analysis, which identified an optimal power transformation of 4.5 by minimizing the Maximum Likelihood Estimation (MLE). This adjustment significantly improved residual normality (Figure 8). However, the transformation led to a noticeable increase in VIF values, prompting the need for feature scaling. Z-score standardization was applied, as it rescales features to have zero mean and unit variance, ensuring that variables with large magnitudes do not dominate the regression estimates, thereby stabilizing coefficient estimates and mitigating multicollinearity.

3.3.2 Transformation on Independent Variables

To assess potential nonlinearity between predictors and IMDb scores, we examined partial regression plots, which visualize the relationship between each covariate and the response variable after accounting for all other predictors.

The plot for "Movie Rank 2023 by IMDbPro" exhibited anomalous behavior (Figure 9), with extreme dispersion and an unusual spread of residuals, suggesting a poor linear fit . To address this, we tested polynomial terms up to the 9th degree, all of which remained statistically significant. However, recalling insights from the EDA, the distribution of this predictor was highly right-skewed, likely complicating the regression relationship by amplifying the influence of extreme values. To mitigate this, we applied a log transformation, which is widely recognized for stabilizing variance and normalizing right-skewed distributions. After transformation, non-significant coefficients were removed to improve model interpretability. Checking the residual plots revealed persistent nonlinear patterns, prompting us to reassess whether the initial transformation on $\boldsymbol{y}$ remained optimal. We re-ran the Box-Cox test, which suggested a revised power transformation of 3 as the new optimal choice. After applying this transformation, the curve was flattened, indicating that the nonlinearity issue was effectively mitigated (Figure 10).

3.4 Leverage and Influence Analysis

Despite these refinements, partial regression plots of the updated model still revealed nonlinear relationships with covariates (Figure 11), particularly for "Number of Articles," where outlying points far from the main clusters suggested potential influence issues. The plot matrix further highlights strong correlations between the fitted values and the response variable ($\boldsymbol{y}$), but also dense tail distributions in residuals, suggesting that some influential observations are distorting parameter estimates (Figure 12). To systematically assess their impact, we conducted influence diagnostics using the following thresholds:

$$|\text{DFFITi}| > 3\sqrt{\frac{p}{n-p}}, \quad |1 - \text{COVRj}| > \frac{3p}{n-p}, \quad \text{hii} > \frac{3k}{n}$$

Each metric captures different aspects of point influence: high leverage indicates potential influence on model predictions, large residuals suggest poor fit, and influential points exert disproportionate effects on coefficient estimates. Notably, a point can have high leverage or be an outlier without necessarily being influential, as seen in the .hat vs. .cooksd plot, where many high-leverage points do not correspond to high Cook's distances (Figure 12). After removing 70 influential data points, model fit improved (Table 3), and the interaction term for "Movie Budget" and "Number of Articles" became spurious. However, residual diagnostics still indicated remaining nonlinearities, warranting further investigation.

3.5 Rationale for Nonlinear Modeling and Spline Selection

Figure 13 illustrates the complex relationships between key predictors and IMDb scores. Several predictors exhibited distinct nonlinear trends, necessitating appropriate transformations to improve model fit while maintaining interpretability. For "Release Year", a simple linear term failed to capture the observed trend, where IMDb scores showed a mild downward trajectory and were flattened around 0, with non-uniform variance. To accommodate this, we applied piecewise linear splines, placing knots at the mean release year (i.e., 0 after the standard scaling) to allow flexibility in modeling local changes without overfitting. For the log of "Movie Rank 2023 by IMDbPro",

a pronounced nonlinear decline was evident, with diminishing marginal effects at higher values. Polynomial regression was tested, and coefficients remained significant up to the third degree, beyond which they provided no additional explanatory power. Similarly, "Number of Articles" followed a cubic pattern, justifying a third-degree polynomial transformation to capture its increasing but asymptotic trend. Movie duration displayed a concave upward shape, which was effectively modeled using a quadratic term. These refinements improved the R^2_{adj} from 57.1% to 57.91%, demonstrating enhanced predictive accuracy.

3.6 Final Refinement on Model Structure

After implementing these transformations, we conducted another eigenvalue decomposition to reassess multicollinearity. The analysis revealed a near-zero eigenvalue associated with the last eigenvector, suggesting a strong dependency between variables. Examining the corresponding eigenvector entries, we identified high contributions from "Movie Budget" and the two splines for "Release Year". This indicated potential multicollinearity arising from the spline transformation. To address this, we introduced interaction terms between "Movie Budget" and the splines of "Release Year". Upon evaluating their statistical significance, we found that only the interaction term with the spline after the mean of "Release Year" remained significant. Consequently, we retained this interaction while discarding the other to enhance model parsimony without sacrificing predictive power.

4 Results

4.1 Model Performance

The final regression model (Equation 4) exhibited strong explanatory power, capturing approximately 58.52% of the variation in IMDb scores within the training dataset (Table 5). A similar value of R^2_{adj} further reinforced the relevance of the selected predictors, highlighting their meaningful contributions to explaining IMDb ratings. Notably, all included predictors were statistically significant at the $\alpha = 5\%$ threshold, demonstrating their individual importance in shaping the model's predictions.

Cross-validation assessments (5-Fold, 10-Fold, and LOOCV) revealed highly consistent results, indicating that the model is not excessively reliant on any specific subset of the data (Table 6). With IMDb scores ranging from 1 to 10, an RMSE of 0.75 across validation methods suggests the model's predictions typically deviate by less than one rating point, indicating strong reliability. The lower RMSE in LOOCV likely stems from reduced bias, as each observation serves as a test set once, minimizing variability in training data and leading to a smoother, more stable model fit. The test set evaluation provides a more unbiased assessment of the model's true predictive power, as the selection of covariates depends partly on patterns unique to the training data (Table 7). As expected, MSE and RMSE are higher on the test set than in cross-validation, reflecting the natural performance drop when transitioning to unseen samples. However, this gap is reasonable, suggesting that the model generalizes well rather than overfitting to training patterns. The Mean Absolute Percentage Error (MAPE) further supports the model's

predictive utility. Despite a slight increase in test-set MAPE, it is still well below the 15% threshold, suggesting that the model maintains promising predictive power on unseen data.

4.2 Validating the Inductive Biases of Linear Regression

Figure 14 provides key insights into the model's adherence to fundamental linear regression assumptions. The Residuals vs. Fitted plot suggests no clear pattern, indicating that the linearity assumption holds, as systematic structure in residuals would suggest model mis-specification. Homoscedasticity is supported by the Scale-Location plot, where residuals exhibit a fairly constant variance across fitted values, with no pronounced funneling effect. The Normal Q-Q plot shows residuals mostly aligning with the theoretical normal distribution, though slight deviations in the tails suggest the presence of outliers. While Cook's distance and Residuals vs. Leverage plots highlight a few high-leverage observations (e.g., points 702, 1480, and 1525), their impact remains limited, as no extreme Cook's distance values indicate undue influence over parameter estimates. Lastly, an examination of the model's VIF values showed no excessively high entries (i.e., > 10), indicating that multicollinearity is not a concern (Table 8). Together, these diagnostics confirm that the model maintains robust coefficient estimates and is unlikely to be significantly distorted by individual data points.

4.3 Understanding Audience Preferences and Industry Trends from Coefficients

The model's coefficients provide key insights into the factors shaping IMDb scores, highlighting both direct influences and nuanced nonlinear relationships. Movie budget, while generally correlated with higher ratings, interacts significantly with action films and recent productions made after 2001, suggesting that well-funded blockbusters tend to perform better within these contexts. Interestingly, colour films are associated with lower ratings compared to black-and-white films, which may reflect how the introduction of color and advanced visual effects shifted industry priorities. This transition may have led to a greater emphasis on spectacle over script quality, whereas black-and-white films, produced in an era with fewer technological distractions, often relied more heavily on strong storytelling and compelling plots. Genre-specific trends reveal systematic differences in audience reception. Action and horror films tend to receive lower ratings, likely due to variability in expectations and reliance on visual spectacle, while drama and animation score higher, reflecting audience preference for narrative depth and artistic execution. These insights highlight the strategic importance of aligning production with genre-driven consumer preferences.

The cubic transformation of IMDb scores captures the nonlinear nature of audience grading habits, suggesting that perceptions of movie quality intensify at both ends of the rating spectrum, while also amplifying the interactive effects of key predictors, such as budget, genre, and publicity, on audience reception. The model's inclusion of polynomial terms captures how "Movie Rank 2023 by IMDbPro" and media coverage ("Number of Articles") exhibit diminishing returns, reinforcing the idea that while popularity boosts ratings, excessive media exposure does not guarantee sustained acclaim. Film duration also follows a nonlinear trend, indicating that audiences favor films

of balanced length, avoiding extremes. The spline at 2001 indicates a structural shift in audience reception, where films released after this threshold face steeper declines in ratings, likely reflecting shifts in cinematic standards, increased competition from digital content, and changing audience review behaviors influenced by online platforms and social media discourse.

4.4 Predictions and Managerial Insights

Data for the 12 newly released movies in 2025 was collected, and "Movie Budget", "Duration", and "Release Year" were scaled using the same mean and standard deviation of the training data. The final regression model was used to forecast IMDb scores based on these attributes (Table 9).

The predictive analysis highlights key structural inefficiencies in the film industry, particularly the diminishing returns on high-budget productions. *Snow White* ($250M) receives only a moderate predicted IMDb score, reinforcing the broader trend that financial investment alone does not secure strong audience reception. This suggests that studios must prioritize narrative quality and market positioning over sheer production scale. Conversely, mid-budget films such as *The Day the Earth Blew Up* and *Novocaine* achieve comparable predicted scores, supporting the strategic viability of controlled-budget productions when paired with targeted audience engagement.

Thriller saturation is evident, with *The Alto Knights* and *A Working Man* performing modestly despite the genre's prevalence. This suggests a need for genre diversification to mitigate cannibalization effects within the market. Meanwhile, the lower scores of *High Rollers* and *Ash* emphasize the inherent risk of niche films, where mainstream appeal remains a challenge. Studios should consider multi-platform distribution models, using streaming services to extend market reach and compensate for lower theatrical draw.

Additionally, the relatively strong projection for *The Day the Earth Blew Up* despite its shorter runtime (91 minutes) suggests shifting audience preferences toward concise, high-engagement storytelling. This aligns with digital consumption trends favoring efficient narrative structures. From a managerial standpoint, these insights reinforce the need for data-driven investment strategies, emphasizing genre viability, audience segmentation, and cross-platform content strategies over budget-intensive production models.

5 Discussion and Conclusion

This study demonstrates the effectiveness of regression modeling in analyzing IMDb ratings, identifying key factors influencing audience reception. Through exploratory analysis and model selection, we examined how production attributes, media exposure, and genre shape ratings. The final model captured nonlinear patterns, improving predictive accuracy but at the cost of interpretability due to the cubic transformation. Given its moderate explanatory power, external factors such as audience demographics and marketing strategies, including actor promotion, likely contribute significantly. Future research should refine transformations for better interpretability and integrate external data sources, such as social media sentiment and critic reviews, to enhance predictive insights into movie success.

6 Appendix

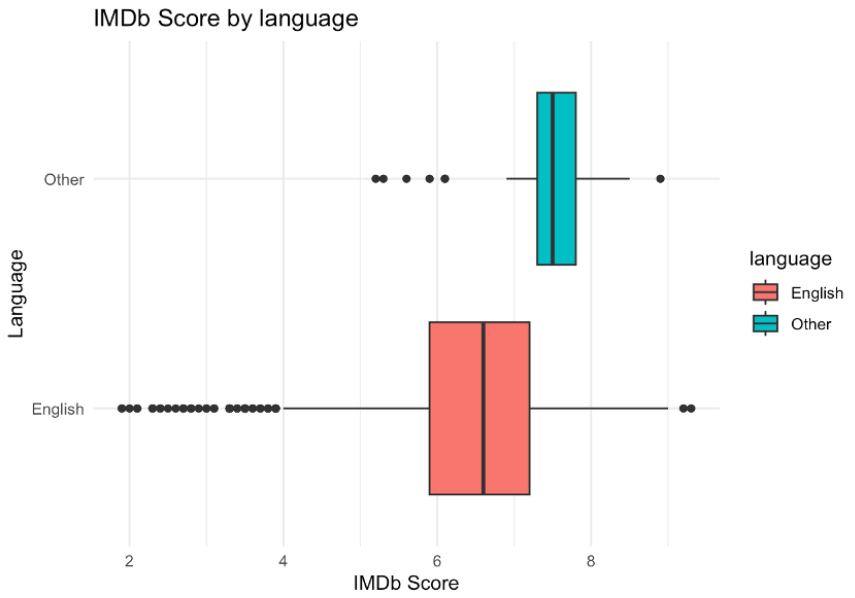

Figure 1: Box Plot of IMDb Score by Language

	Dependent variable:
	IMDb Score
Movie Budget	-0.00^{***} (0.00)
Release Year	-0.02^{***} (0.002)
Duration	0.01^{***} (0.001)
Other Language	0.70^{***} (0.16)
G and Other Maturity Rating	-0.11 (0.26)
PG Maturity Rating	0.10 (0.24)
PG-13 Maturity Rating	0.07 (0.24)
R Maturity Rating	0.40^{*} (0.24)
Number of Articles	0.0001^{***} (0.0000)
Colour Film	-0.46^{***} (0.13)
Number of Faces	-0.04^{***} (0.01)
Action Genre	-0.35^{***} (0.06)
Horror Genre	-0.39^{***} (0.08)
Drama Genre	0.37^{***} (0.05)
Animation Genre	0.80^{***} (0.21)
Movie Rank 2023 by IMDbPro	-0.0000^{***} (0.0000)
Constant	38.06^{***} (4.67)
Observations	1,544
R^2	0.35
Adjusted R^2	0.34
Residual Std. Error	0.88 (df = 1527)
F Statistic	51.55^{***} (df = 16; 1527)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Stepwise Baseline Model Summary

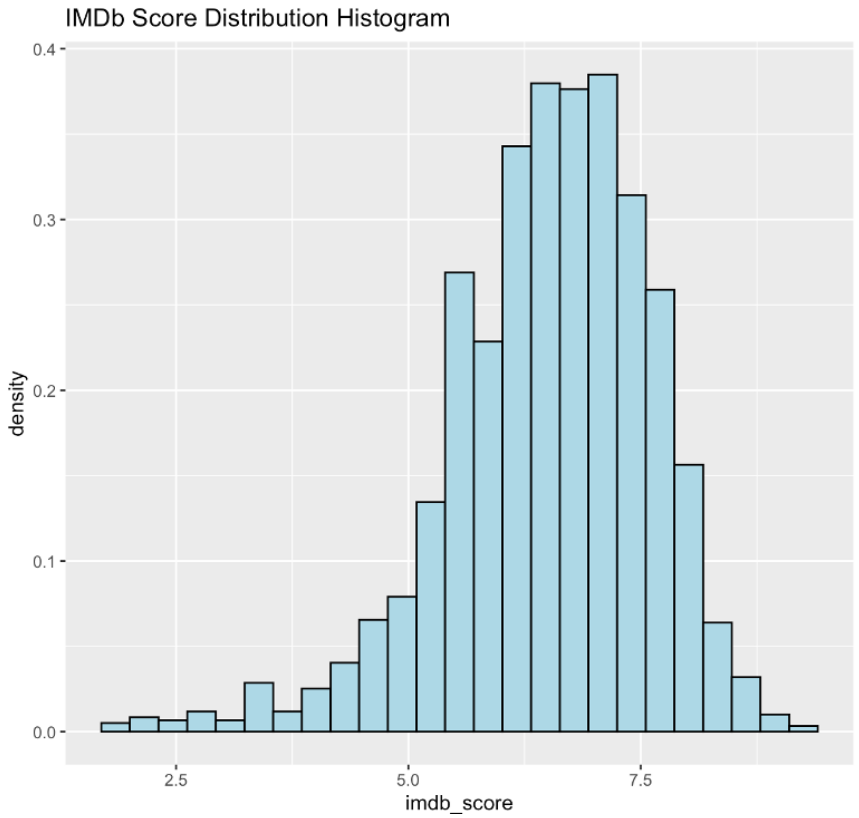

Figure 2: Distribution Histogram of IMDb Scores

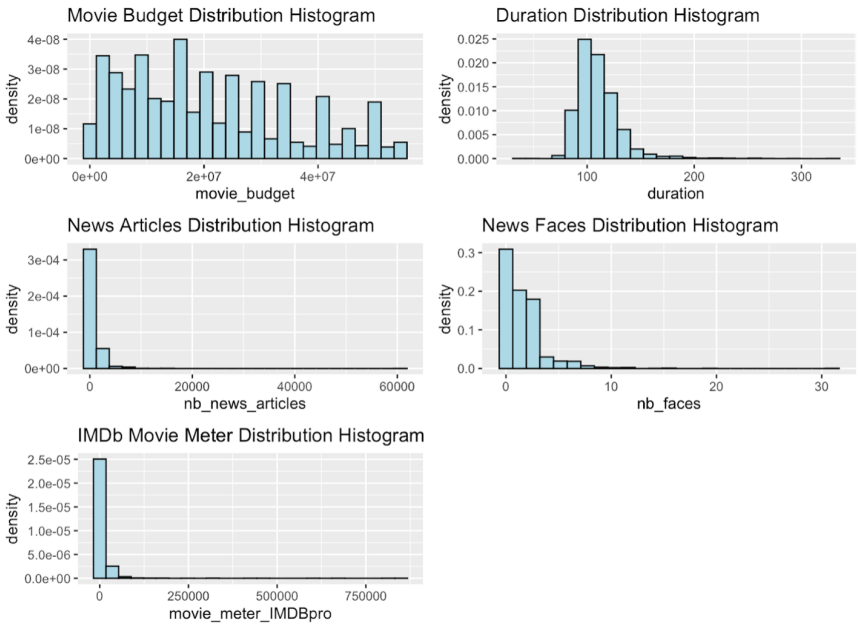

Figure 3: Matrix Plot of Distribution Histogram of Independent Variables

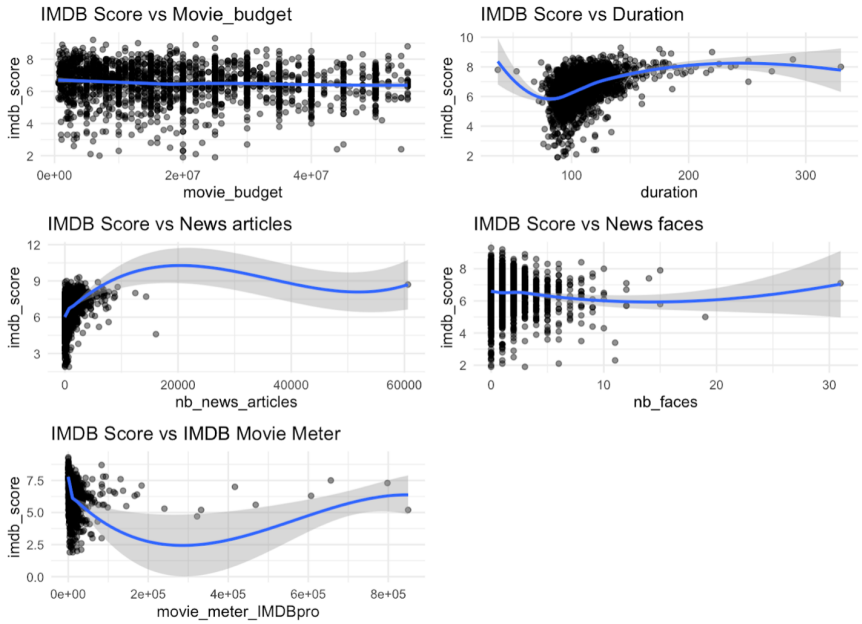

Figure 4: Matrix Plot of Scatter Plots of Continuous Variables VS. IMDb Scores

Figure 5: Matrix Plot of Box Plots of IMDb Scores by Categorical Levels

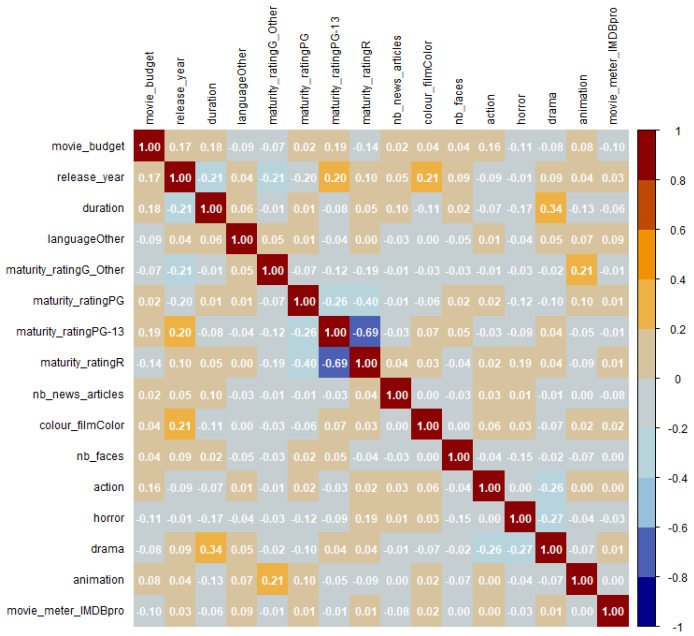

Figure 6: Correlation Heat Map of Variables

Regression Model Comparison

Metric	Full Model	Stepwise Model
R-squared	0.3518	0.3507
Adj. R-squared	0.3381	0.3439
AIC	4029.0245	3999.7522
BIC	4210.6570	4095.9106
RMSE	0.8726	0.8734
Num Predictors	33.0000	17.0000

Figure 7: Table for Stepwise Regression Model Comparison with Full Model

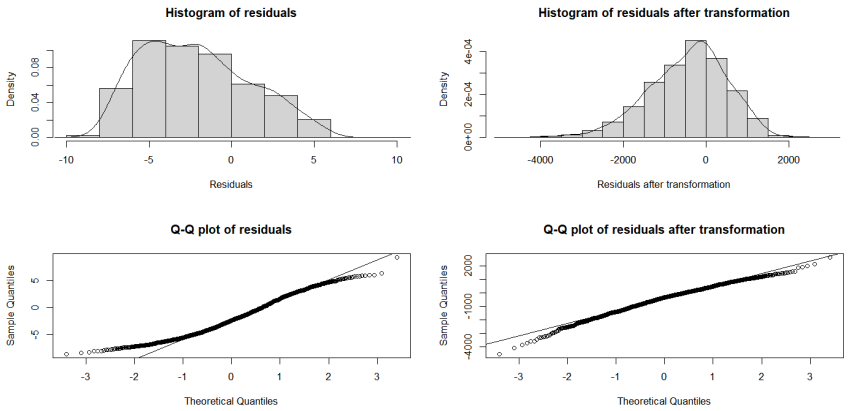

Figure 8: Box-Cox Analysis Summary Output

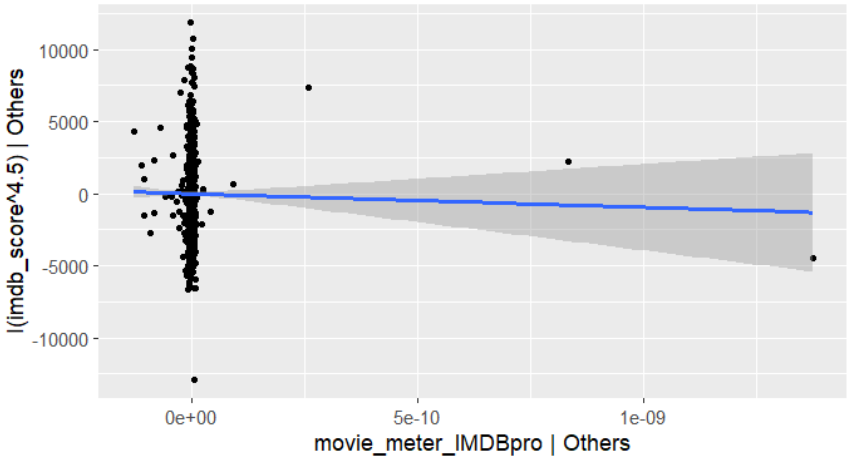

Figure 9: Partial Regression Plot for Movie Rank 2023 by IMDbPro

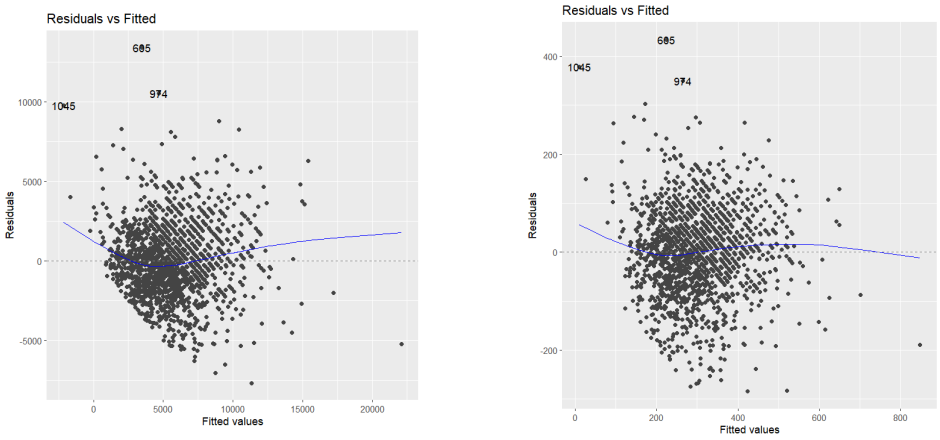

Figure 10: Left: Residual VS. Fitted Plot Before Transformation. Right: Residual VS. Fitted Plot After Transformation.

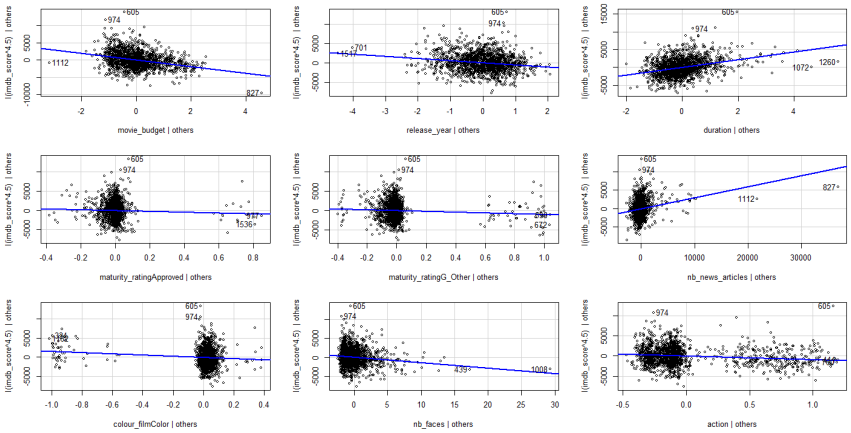

Figure 11: Excerpt of Partial Regression Plots

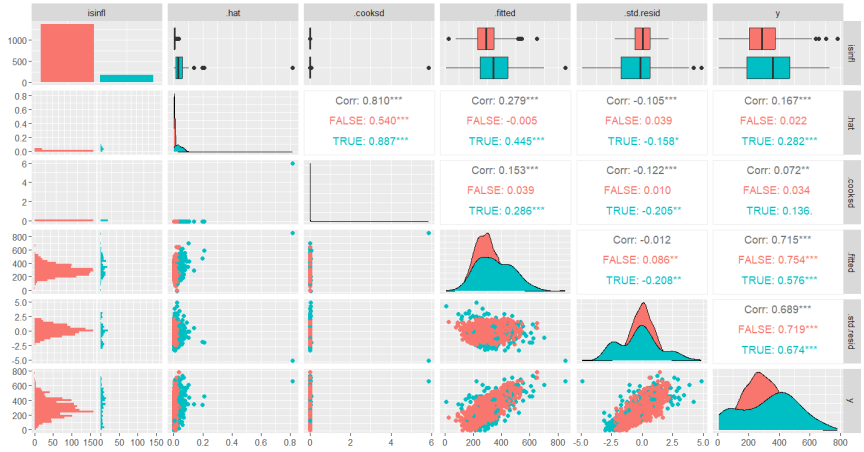

Figure 12: Matrix Plot of Influence Metrics

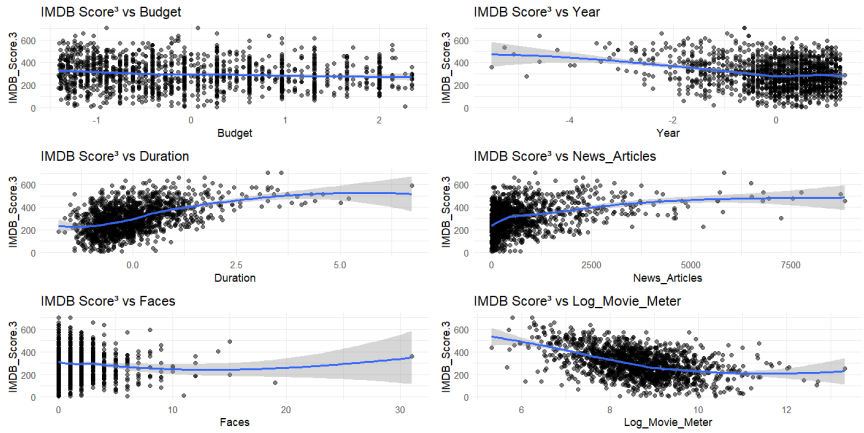

Figure 13: Matrix of Scatter Plots of Continuous Covariates with IMDb Score³

Eigenvector	High-Contributing Variables
e1	Number of Articles, Movie Budget
e2	Language, Color Film
e3	Language, Movie Rank 2023 by IMDbPro
e4	Movie Budget, Duration, Action, Horror
e5	Animation, Rating: G-rated Other
e6	Action, Horror, Drama
e7	Movie Budget, Release Year, Duration, Drama
e8	Rating: PG, PG-13, R

Table 2: Insights from Eigenvalue Decomposition

	Dependent variable:	
	IMDb Score[3]	
	Coefs Before Removal	Coefs After Removal
Movie Budget	-34.41^{***} (3.08)	-30.10^{***} (3.04)
Release Year	-19.49^{***} (3.14)	-21.85^{***} (2.93)
Duration	40.29^{***} (3.18)	42.02^{***} (3.07)
Approved Maturity Rating	-31.26 (26.12)	-47.14^{*} (27.16)
G and Other Maturity Rating	-37.29^{***} (14.00)	-27.28^{*} (15.41)
Number of Articles	0.01^{***} (0.001)	0.02^{***} (0.002)
Colour Film	-54.58^{***} (13.88)	-50.43^{***} (14.07)
Number of Faces	-5.40^{***} (1.12)	-4.39^{***} (1.03)
Action Genre	-34.42^{***} (6.28)	-37.00^{***} (5.79)
Horror Genre	-62.05^{***} (7.93)	-62.50^{***} (7.25)
Drama Genre	44.71^{***} (5.48)	45.33^{***} (5.03)
Animation Genre	117.20^{***} (22.05)	122.65^{***} (25.66)
log(Movie Rank 2023 by IMDbPro)	-43.65^{***} (2.46)	-43.51^{***} (2.64)
Movie Budget × Number of Articles	0.005^{***} (0.002)	-0.001 (0.002)
Release Year × Duration	5.73^{***} (2.05)	4.03^{*} (2.24)
Movie Budget × Release Year	5.87^{**} (2.95)	6.30^{**} (2.84)
Movie Budget × Action Genre	14.55^{**} (5.87)	15.49^{***} (5.40)
Constant	716.88^{***} (25.77)	703.31^{***} (27.70)
Observations	1,544	1,474
R^2	0.51	0.58
Adjusted R^2	0.51	0.57
Residual Std. Error	90.81 (df = 1526)	81.26 (df = 1456)
F Statistic	93.96^{***} (df = 17; 1526)	116.32^{***} (df = 17; 1456)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Regression Model Comparison Before and After Removing Influential Points

$$
\begin{aligned}
(\text{IMDb Score})^3 = {} & \beta_0 + \beta_1(\text{Movie Budget}) + \beta_2(\text{Colour Film}) + \beta_3(\text{Number of Faces}) \\
& + \beta_4(\text{Action Genre}) + \beta_5(\text{Horror Genre}) + \beta_6(\text{Drama Genre}) + \beta_7(\text{Animation Genre}) \\
& + \beta_8(\log(\text{Movie Rank 2023 by IMDbPro})) + \beta_9(\log(\text{Movie Rank 2023 by IMDbPro}))^2 \\
& + \beta_{10}(\text{Number of Articles}) + \beta_{11}(\text{Number of Articles})^2 + \beta_{12}(\text{Number of Articles})^3 \\
& + \beta_{13}(\text{Duration}) + \beta_{14}(\text{Duration})^2 \\
& + \beta_{15}(\text{Release Year before Mean}) + \beta_{16}(\text{Release Year after Mean}) \\
& + \beta_{17}(\text{Movie Budget} \times \text{Action Genre}) + \beta_{18}(\text{Movie Budget} \times \text{Release Year after Mean}) + \varepsilon
\end{aligned} \tag{1}
$$

Table 4: Final Regression Model Equation.

	Dependent variable:
	IMDb Score^3
Movie Budget	-32.83^{***} (3.29)
Colour Film	-52.04^{***} (13.68)
Number of Faces	-4.19^{***} (1.02)
Action Genre	-36.48^{***} (5.76)
Horror Genre	-62.80^{***} (7.21)
Drama Genre	44.19^{***} (5.04)
Animation Genre	118.43^{***} (24.99)
log(Movie Rank 2023 by IMDbPro)	$-1,618.87^{***}$ (108.73)
log(Movie Rank 2023 by IMDbPro)^2	362.23^{***} (84.14)
Number of Articles	912.25^{***} (103.45)
Number of Articles^2	-282.04^{***} (88.79)
Number of Articles^3	197.73^{**} (84.29)
Duration	$1,345.78^{***}$ (98.49)
Duration^2	-202.91^{**} (83.78)
Release Year before Mean	-82.44^{***} (20.90)
Release Year after Mean	-135.71^{***} (19.07)
Movie Budget × Action Genre	13.46^{**} (5.31)
Movie Budget × Release Year after Mean	15.45^{**} (7.17)
Constant	434.60^{***} (20.90)
Observations	1,474
R^2	0.59
Adjusted R^2	0.58
Residual Std. Error	80.39 (df = 1455)
F Statistic	114.06^{***} (df = 18; 1455)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Final Regression Model Summary

Cross-Validation Method	MSE	RMSE	MAPE
5-Fold	0.5755	0.7575	9.9574
10-Fold	0.5734	0.7546	9.9387
LOOCV (1474-Fold)	0.5743	0.5368	9.9402

Table 6: Cross-Validation Performance Metrics

Metric	Value
MSE	0.9021
RMSE	0.9498
MAPE	12.0736

Table 7: Test Set Performance Metrics

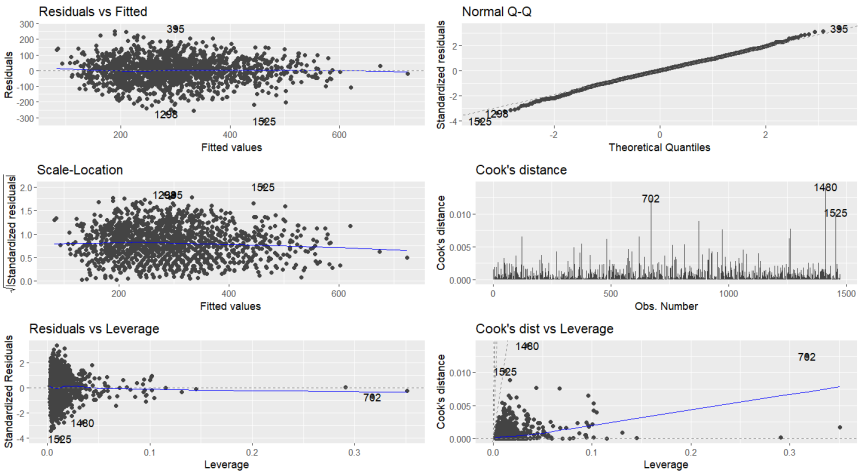

Figure 14: The Diagnostic Plots of the Final Model

Predictor	VIF
Movie Budget	2.4678
Colour Film	1.0719
Number of Faces	1.0625
Action Genre	1.1994
Horror Genre	1.1897
Drama Genre	1.4326
Animation Genre	1.0548
log(Movie Rank 2023 by IMDbPro)^1	1.8291
log(Movie Rank 2023 by IMDbPro)^2	1.0953
Number of Articles^1	1.6559
Number of Articles^2	1.2197
Number of Articles^3	1.0994
Duration^1	1.5009
Duration^2	1.0860
Release Year before Mean	7.4136
Release Year after Mean	7.7573
Movie Budget $\times$ Action Genre	1.3958
Movie Budget $\times$ Release Year after Mean	1.8527

Table 8: Variance Inflation Factors (VIFs) for Model Predictors

Movie Name	**Predicted IMDb Score**
O'Dessa	5.40
Black Bag	6.14
High Rollers	4.88
Novocaine	6.64
The Day the Earth Blew Up	6.81
Ash	4.89
Locked	5.26
Snow White	6.54
The Alto Knights	6.40
A Working Man	6.38
My Love Will Make You Disappear	5.55
The Woman in the Yard	5.34

Table 9: Predictions for the 12 Movies Released in 2025

References

[1] Internet Movie Database. Imdb datasets, 2018. Accessed: 2025-03-05.