MGSC661 Midterm: The 2024 IMDb Prediction Challenge

Hannah Wang
JaeYoon Lee
Yifei Liu
Jintao Li
Shuxi Chen

Introduction

The 2024 IMDb Prediction Challenge seeks to predict the ratings of twelve upcoming blockbuster films on the IMDb platform, with the goal of determining public reception. Drawing on data from approximately 2,000 previous movies, the project will focus on building a robust predictive model by applying data analytics techniques designed to minimize bias, heteroskedasticity, overfitting, and other common modeling issues. To better understand the various characteristics of these films, statistical methods such as linear regression and non-linear modeling will be employed. Through comprehensive and thoughtful analysis, the most significant variables for predicting IMDb ratings will be identified, ensuring strong out-of-sample performance. The aim is to create a model with a high R-squared value with the lowest MSE, enabling accurate predictions of the ratings for the twelve upcoming movies.

Data Description

The dataset used for the 2024 IMDb Prediction Challenge contains detailed information on 1,930 films, comprising one dependent variable (`imdb_score`) and 41 independent variables. These variables offer insights into various aspects of each movie, such as identifiers, film characteristics, cast details, and production-related features. To refine the dataset, irrelevant label variables like `movie_title`, `movie_id`, and `imdb_link` were excluded, as they do not contribute to the prediction task based on the data descriptions. Second, release_day and release_month were dropped considering the lack of prediction power (see Figure 1), release_year was also dropped because the time range of the whole dataset is too big, in terms of predicting futrue movies which are relased in similar years, this cannot be consider as a powerful predictor. Additionally, the `plot_keywords` column was dropped due to overlap with genre information, making it redundant. Similarly, the `colour_film` column was removed, as future movies are unlikely to be black-and-white, making it irrelevant for prediction. Also, we concluded language is not a significant predictor, as overwhelming dominance of English makes it unlikely to contribute meaningfully to the prediction model (see Figure 2).

After performing high-level data preprocessing, we have dived deeper into the dataset to identify potential issues such as skewed distributions, outliers, or columns that may need to be recategorized or simplified. This step was crucial to refine the dataset and improve the overall accuracy of our predictions.

Looking at the dependent variable, `imdb_score` (see Figure 3), the scores typically range from 5.9 to 7.3, with a mean value of 6.512. The data is slightly left-skewed, with a skewness value of -0.8645. This skewness may be attributed to the tendency of viewers to rate good movies higher, while lower ratings are reserved for particularly bad experiences, which are less common. As for the correlation between all the numerical variables (see Figure 4), in general there is no collinearity between the independent variables. However, we can see some correlation between them. For example, the three 'star meter scores' exhibit correlations among themselves, indicating that selecting just one of these as a predictive variable may be more prudent to avoid redundancy. Furthermore, a correlation exists between 'movie budget' and 'aspect ratio', suggesting that the predictive power of these two variables may overlap, potentially diminishing their individual contributions to the model.

Examining the correlations between the independent variables and 'imdb_score' (see Figure 5), the strongest positive predictors (> 0.3) are `duration`, the genre 'drama', and 'number of news articles', while the strongest negative predictors (> -0.3) are 'comedy', 'horror', and 'PG' rating. Although other variables show weaker correlations, they may still contribute through interactions or in non-linear models. Looking at Figure 4, 'release day' and 'aspect ratio' have nearly no correlation with the target variable, which suggested that maybe we should consider dropping them in the further model selection.

Looking at the numerical dataset (see Figure 6), 'budget', 'nb_news_articles', 'actor1_star_meter', "actor2_star_meter", "actor3_star_meter", "nb_faces" and 'movie_meter_IMDBpro's' histogram and bar chart were heavily right skewed, indicating a concentration of values at the lower end with a long tail towards higher values. For managing outliers in our dataset, we adopted a straightforward approach by removing all values that fell above the third quartile (Q3). This method was particularly effective in handling the extreme data points because these variables initially showed the right skewness, also ensuring a more normalized data set that better represents the central tendency and variability. After dropping the outliers, we applied log transformation to address the skewness issues and reduce the impact of extreme values, resulting in more symmetrical distributions that can be a better fit for our predictive modeling techniques (see Figure 7).

Upon reviewing the categorical data, we found that eight genres, including 'comedy' and 'biography,' were missing from the dummified genres in the raw dataset. Since genre is a key

predictor of IMDb ratings, we re-dummified all genres to ensure accuracy. Genres like "animation" and "documentary," with fewer than 30 occurrences (see Figure 8), were grouped into "other" category. Additionally, p-value tests showed that 'musical', 'romance', 'animation', 'music', and 'mystery' lacked significant predictive power, so these were also consolidated into the "other" category to serve as a baseline for genre in the prediction model. For all personnel-related predictors "cinematographer," "director," "production company", "distributor", "actor1", "actor2", and "actor3", we decided to focus on the top 10 in each category from the dataset (e.g. see Figure 9). These individuals or companies are considered the most influential, as they likely have a relatively greater impact on IMDb scores. However, for the reclassified top 10 actors—variables 1, 2, and 3—the actual data included a metric called the Actor Star Meter Rating, which indicates the ranking of individual actors. We chose to use this more specific data rather than relying on the top 10 actors for our predictors. For the other reclassified categorical predictors, we created new dummy columns indicating whether a movie involves a top 10 contributor for each of these predictors. As for the country, we selected the top 5 countries as dummified variables and all other countries as the baseline (see Figure 10). Additionally, for the maturaity ratings, because of the natural of similar definitions of different rating scales, we reclassified the maturity ratings into three categories (see Figure 11). After this pre-processing, due to the minimal representation of the G rating, PG and R are left as effectively dummified variables. This results in the nearly perfect negative correlation between the maturity ratings PG and R (Correlation, -0.94), indicating these categories are mutually exclusive, which suggests that only one should be included in the model to avoid multicollinearity.

Model Selection

After performing Simple Linear Regressions between the dependent variable (IMDb Score) and each remaining numerical independent variable, we dropped non-significant variables and those with an R-squared value of less than 1% (see Table 1). Next, we built a Multiple Linear Regression model using the remaining numerical variables.

To check for possible collinearity in the model, we calculated VIF scores, none of which exceeded 5, indicating no collinearity among the variables (see Table 2). We assessed heteroskedasticity by examining the residual plots and found a noticeable funnel shape for Movie Duration (see Figure 12(a)). To address this issue, we applied the log transformation to movie duration, which improved the funnel shape, though some heteroskedasticity remained (see Figure 12(b)). A linearity test for the revised model revealed that both the log of movie duration and the log of the star meter for actor 1 were non-linear, suggesting that the model itself is non-linear (see Table 3).

Before addressing nonlinearity, we first introduced selected categorical variables, which raised the Adjusted R-squared from 0.380 to 0.474, indicating that adding categorical variables enhanced the model's predictive power. We then performed non-linear regression models with different polynomial degrees. Using ANOVA tests, we determined that the highest degree of polynomial for both non-linear variables was 2 (see Table 4). We then obtained the regression results for the model (see Table 5). An NCV test showed a p-value of less than 0.001, indicating the presence of heteroskedasticity. To correct for heteroskedastic errors, we performed a coefficient test (see Table 6), and after addressing heteroskedasticity, the significance levels for each variable remained unchanged. Therefore, our final model for predicting IMDb scores is as follows:

$$\begin{aligned} IMDb\ Score = {} & b_0 + b_1 ln(Movie\ Budget) + b_2 ln(Movie\ Budget) + b_3 ln(Movie\ Budget)^2 \\ & + b_4 ln(Star\ Meter{:}\ Actor\ 1) + b_5 ln(Star\ Meter{:}\ Actor\ 1)^2 \\ & + b_6 ln(Number\ of\ News\ Articles) + b_7 ln(IMDbPro\ Movie\ Meter) \\ & + b_8 Maturity\ Rating{:}\ R + b_{8,i} \sum_{i=0}^{15} Genre_i + b_9 Country{:}\ USA \\ & + b_{10} Top\ Cinematographer + b_{11} Top\ Director + b_{12} Top\ Distributor \end{aligned}$$

Result

Based on our predictive model, the predictive IMDb score of the 12 upcoming new blockbuster movies is presented as follows:

Table 7: Predicted IMDb Score

Movie Name	Predicted IMDb Score
Venom: The Last Dance	7.045025
Your Monster	6.123169
Hitpig!	4.747022
A Real Pain	7.12499
Elevation	5.327322
The Best Christmas Pageant Ever	6.531037
Kanguva	6.419244
Red One	6.520043
Heretic	6.885692
Bonhoeffer: Pastor. Spy. Assassin.	6.733261
Gladiator II	7.905175
Wicked	7.741898

Based on the predictive scores, the rationale behind may suggest that the most significant factors influencing higher predicted IMDb scores are a combination of budget, star power, media coverage, and genre. Movies with larger budgets typically have the resources to enhance production quality, cast well-known actors, and invest in extensive marketing campaigns, all of which contribute to higher ratings. For example, movie "Gladiator II" (7.90) has the highest budget, "Wicked" (7.74) has the highest amount of news articles, "A Real Pain" (7.12) falls into genre ‘drama’, and "Venom" (7.04) has high budget and high star meter score. Lead actors with high star meter rankings also attract more audience attention, boosting a film's visibility and appeal. Additionally, movies that generate significant media coverage, such as news articles and reviews, are more likely to gain public interest and achieve better ratings. Certain genres like action, drama, and biography tend to receive more critical acclaim and resonate with audiences, contributing further to higher scores. In contrast, movies with smaller budgets, less popular actors, fewer media mentions, or niche genres (like horror or family films) tend to have lower predicted IMDb scores. For example, the movies with lower predicted IMDb scores include "Hitpig!" (4.74) who has lowest duration and number of news articles, and "Elevation" (5.33) who also has low duration, number of news articles, and low IMDbPro movie meter. Thus, a combination of strong financial

backing, popular talent, substantial media attention, and genre alignment is key to achieving high scores in the predictive model.

Given the relatively small dataset, the model's performance was evaluated using Leave-One-Out Cross-Validation (LOOCV). The model's performance is summarized below:

Table 8: Model Performance (LOOCV)

Statistics	Values
R-squared	0.4857
Adjusted R-squared	0.476
MSE	0.5967

By using our final model, and using a new test data, the average squared residual of these new data points would be 0.5967. This value suggests that on average, the squared difference between predicted and observed values for new data points is relatively low, demonstrating that the model has reasonable predictive power in terms of out-of-sample performance. We can also see that 48.57% of the variability in the target variable is explained by the model's predictors. This generally tells that the model's predictors only capture less than half of the total variability, highlighting necessity of including other relevant variables. However, the reason behind this value is to avoid the dangers of overfitting. We tried to construct a model that maintains a balance between flexibility and simplicity by limiting the number of predictors and the polynomial degree.

Taking a deeper look at the significant level of each predictor, we divide the predictors into numerical category and categorical category.

The predictor 'log movie budget' shows that, holding other variables fixed, for each unit increase in the movie budget, the IMDb rating is expected to increase by 0.76 points. The p-value of these variables shows a statistically significant relationship, suggesting that higher budgets will make the movie have a better rating. This makes sense because the audience may value content quality if the production scale is larger. For the 'duration' predictor, our model includes first-order and second-order coefficient terms, and the p-value shows the statistical significance of the first term. The first-order term coefficient is 10.1132. This non-linear relationship indicates that the impact of movie duration on IMDb ratings is not simply linear, and there might be an optimal duration range. The 'log of number of news articles' variable shows that, holding other variables constant, for each unit increase in the news article, the average IMDb rating is expected to increase

by 1.07 points. Its p-value also shows the statistical significance of the variable. For the 'actor1 star power' predictor, our model also includes two polynomial coefficient terms. Although the statistical significance is relatively weak, the influence of actor1 star power shows a non-linear pattern. This indicates that if the actor exists in the top 10 list we combined, it will decrease the IMDb score. The coefficient for the 'log of IMDbpro movie popularity' variable tells us if we hold other variables fixed, for each unit increase in the movie popularity, the average IMDb rating is expected to increase by 0.68 points. The corresponding p-value shows the variable's significance and suggests that audiences have stricter evaluation standards for highly anticipated movies because they have higher expectations.

Next, we jump into the detail about the impact of categorical variables on the IMDb score. For the 'maturity rating' variable, based on our model, we evaluate whether the maturity rating is R or not. The coefficient for this variable indicates that R-rated movies score on average 0.16 points higher than non-R-rated movies, suggesting that R-rated movies might contain more complex and diverse themes, which makes the movie have a higher score. Movie genres also have a significant impact on IMDb scores. The genres with a significant positive impact include Drama, War, Adventure, Crime, and Biography. If the movie theme includes any of the variables with a significant positive impact, its IMDb score will increase by the corresponding coefficient. Similarly, Horror, Action, and Comedy are the genres that have a negative impact on the IMDb score, which makes the IMDb score to have a decrease if the movie includes any of the negative impacting genres. Predictors 'top director', 'cinematographer', and 'distributor' represent the impact of the production team, and they all have a positive impact on the IMDb score. If the movie is filmed by a director who is in the top 10 director, or the movie has a cinematographer in the top 10 cinematographer list, it will increase the IMDb score. Similarly, if the movie is distributed by a distributor in the top 10 list, we will also see an increase in the average IMDb score. This demonstrates that an excellent production team can significantly improve movie quality. Lastly, we evaluate whether the movie is produced in the USA or another region. For USA production movies, the coefficient indicates that USA movies score on average 0.16 points lower than non-American movies. This tells that audiences may have higher expectations for USA movies.

Overall, most of the variables in the model have statistical significance, such as budget, duration, number of articles, etc. The polynomial terms help us handle more complex relationship patterns between predictors and the target variable. The coefficients of categorical variables show

the rating score differences across different movie types, providing us with more important insights about audience preferences.

Conclusion

The 2024 IMDb Prediction Challenge allowed us to develop a predictive model for IMDb ratings based on various movie characteristics, resulting in a relatively robust and interpretable model. Our analysis identified key factors influencing IMDb scores, including duration, budget, media coverage, IMDb Pro meter and genre.

Overall, the predictive model provides valuable insights into the factors driving movie success on IMDb with an adjusted R-squared of 0.476. By understanding the importance of budget, casting, and media presence, as well as how different genres resonate with audiences, stakeholders in the film industry can make more informed decisions to maximize their movie's potential for higher ratings.

Appendix

Figure:

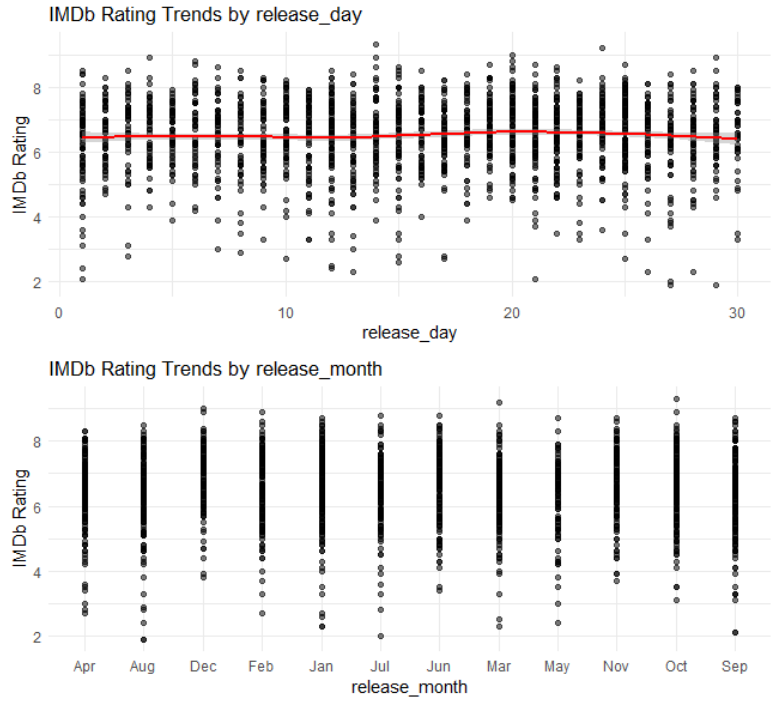


Figure 1: Time-Series vs. IMDb Score

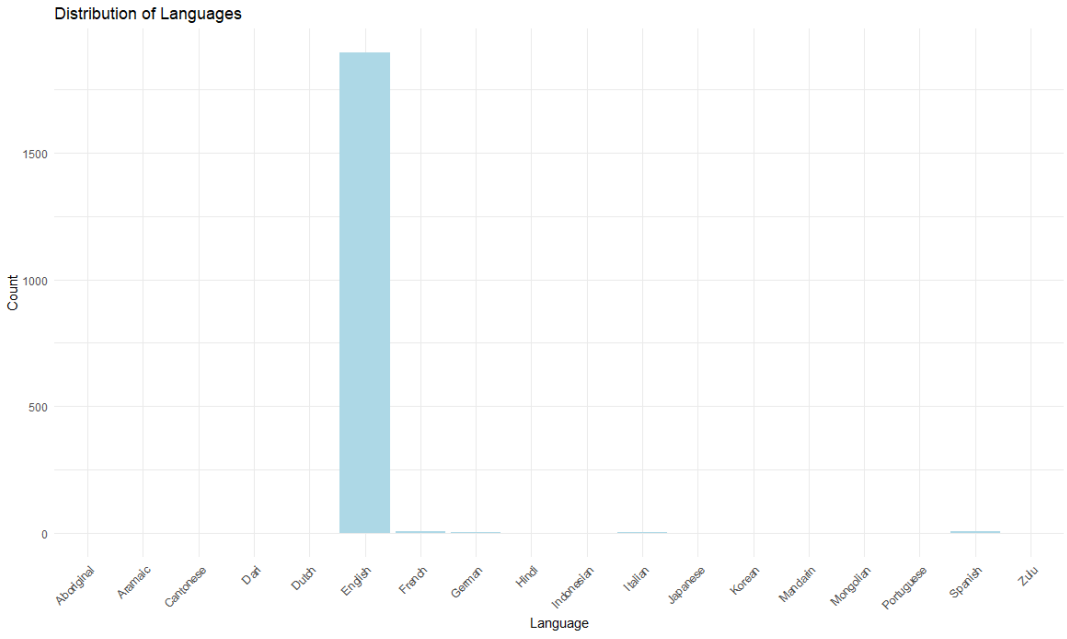


Figure 2: Distribution of Languages

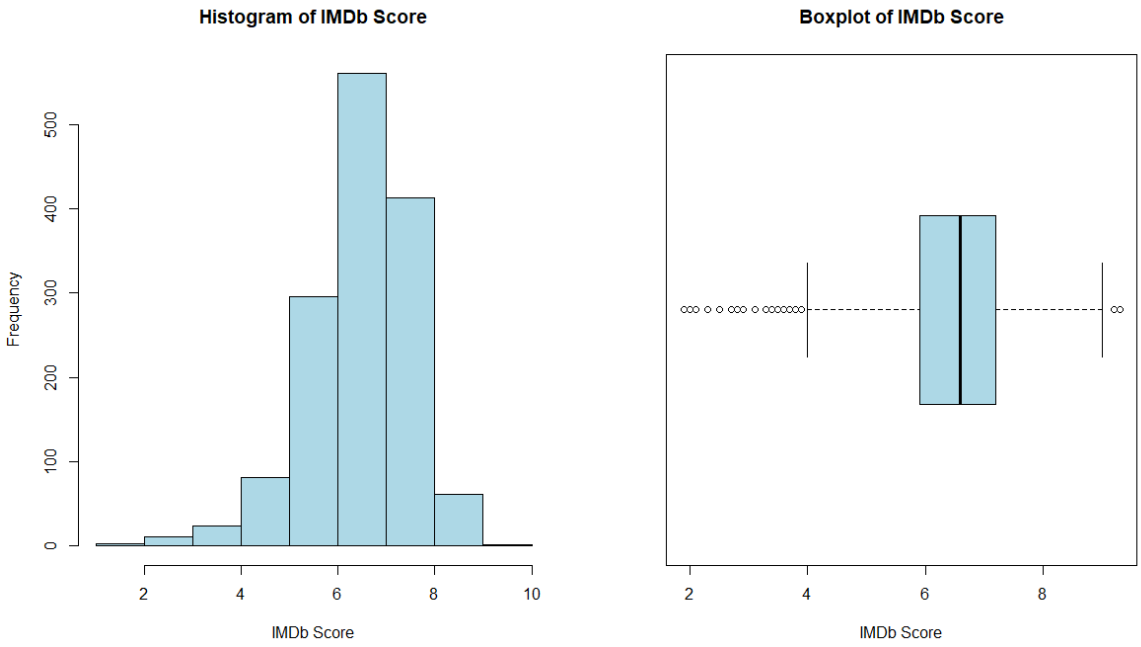


Figure 3: IMDb Score Distribution

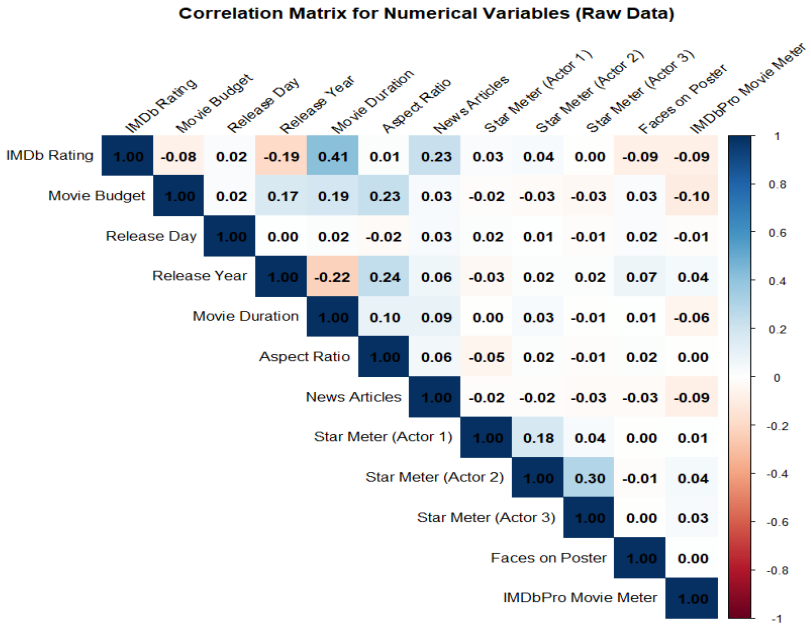


Figure 4: Correlation for numerical variables

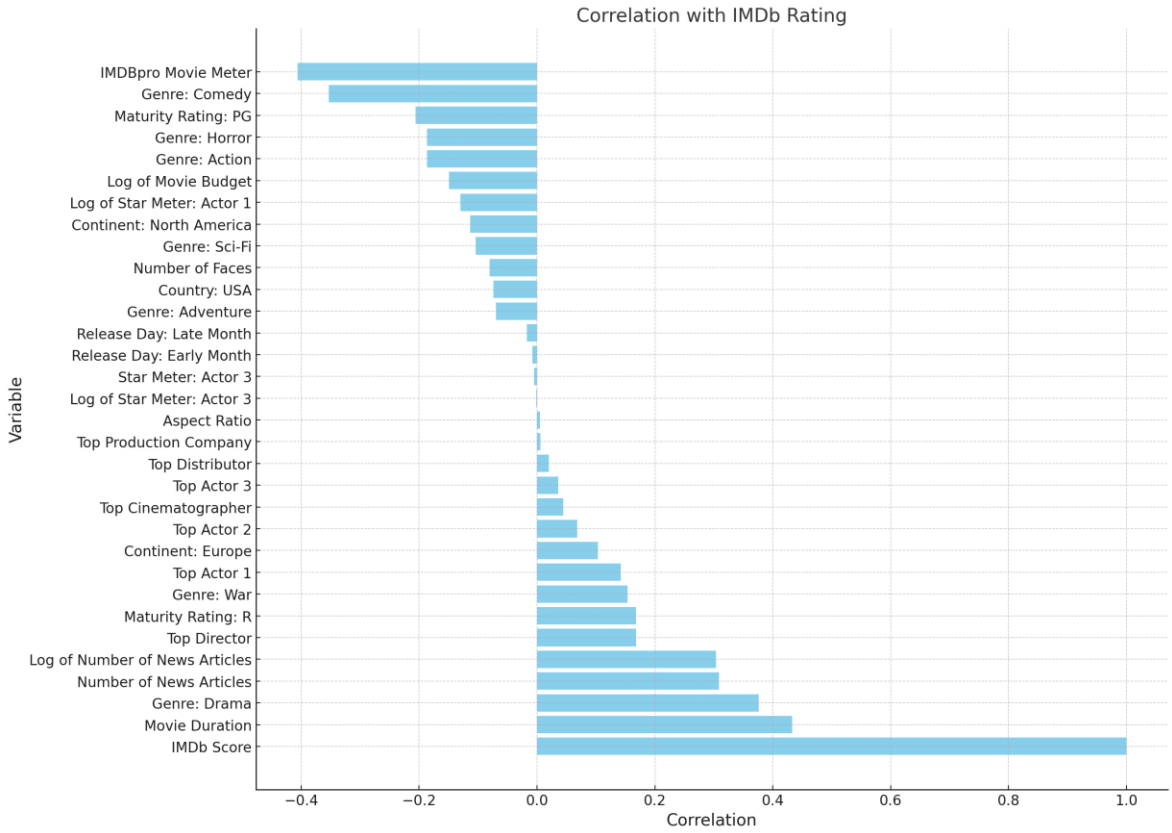


Figure 5: Correlation between independent variables with dependent variable

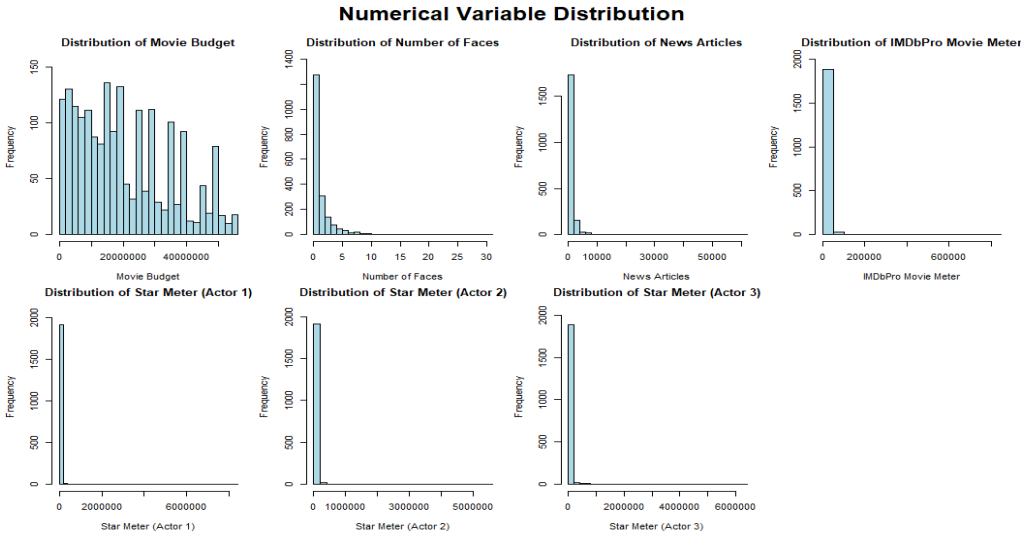


Figure 6: Histogram of numerical variables

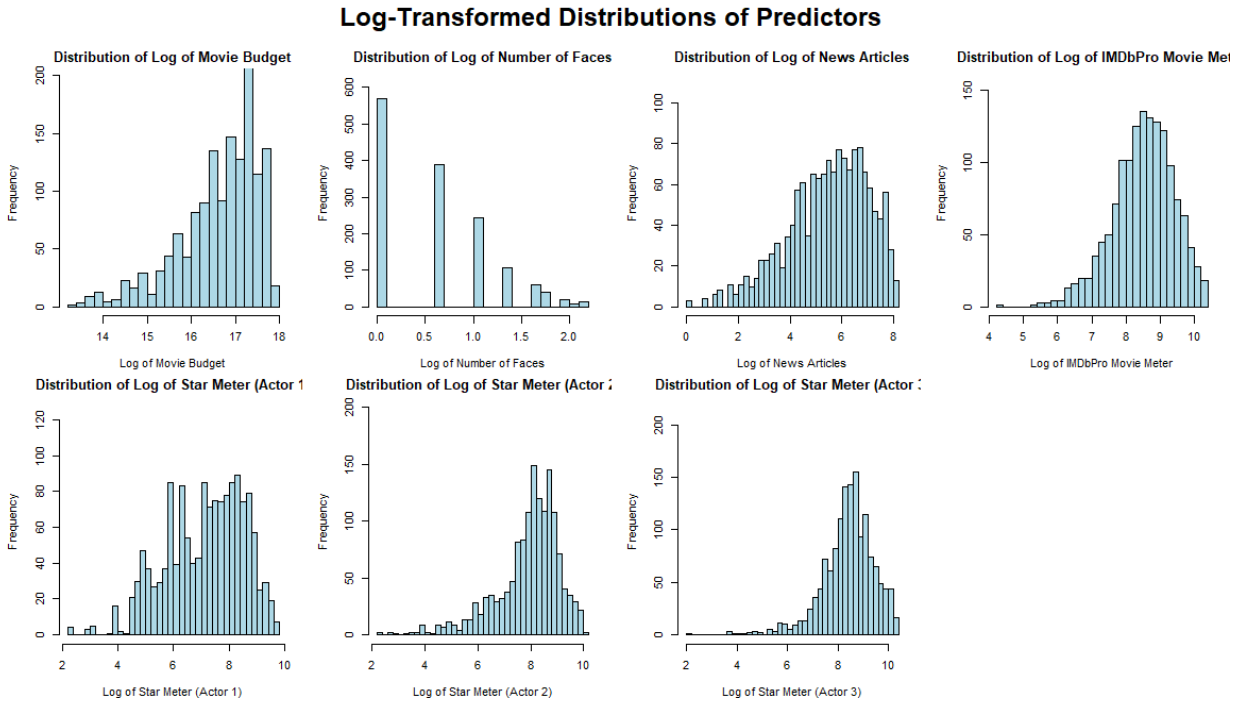
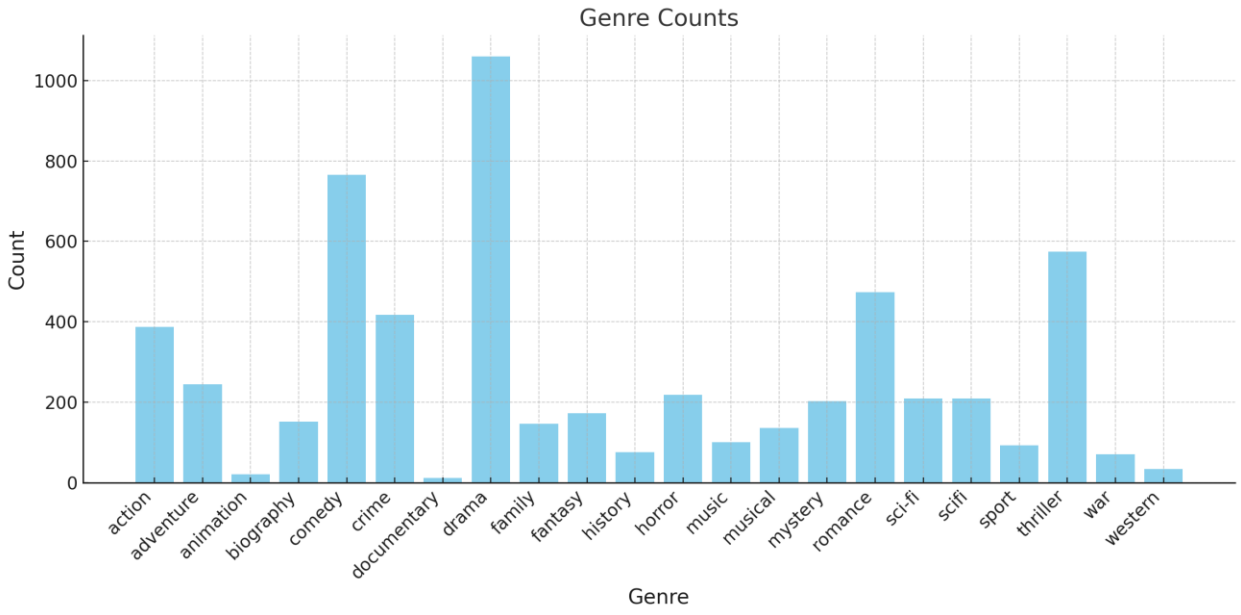


Figure 7: Log transformation of numerical variables




```
[1] "musical - p-value: 0.337662947281095"
[1] "romance - p-value: 0.475827407419031"
[1] "animation - p-value: 0.420847318364638"
[1] "music - p-value: 0.0939739158062037"
[1] "mystery - p-value: 0.863901049829263"
```

Figure 8: Distribution and p-value test of Genres

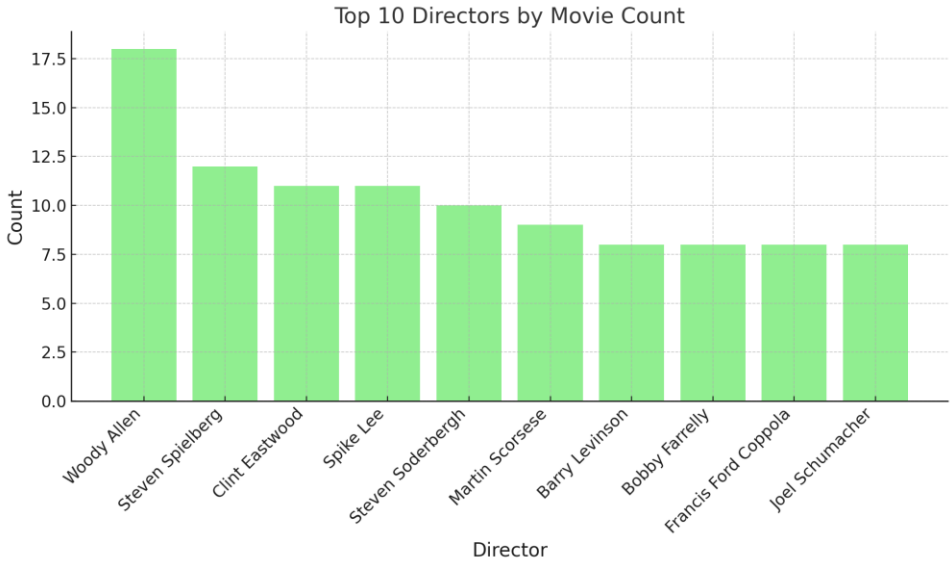


Figure 9: Top 10 Directors

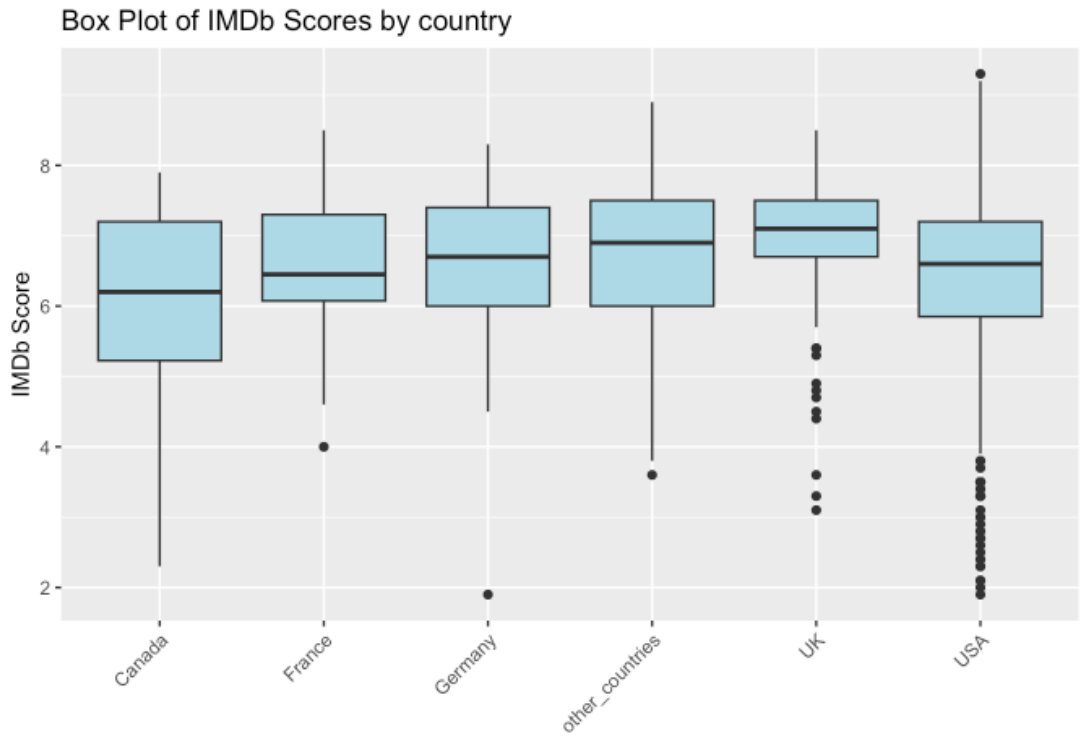


Figure 10: Box Plot of top 5 countries and other countries

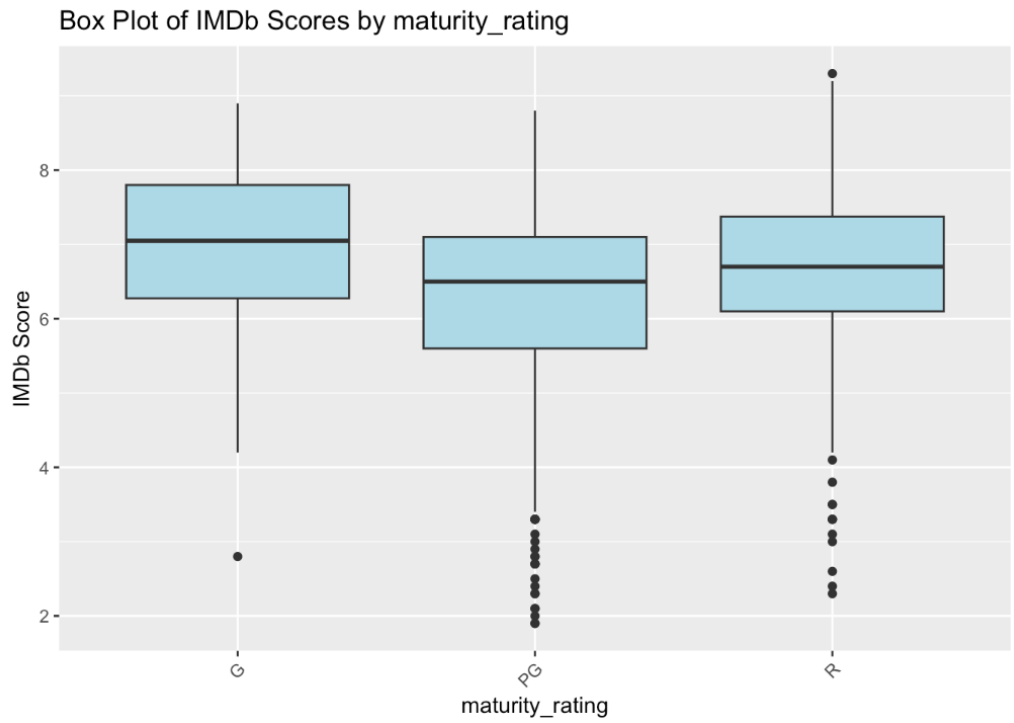


Figure 11: Reclassification of maturity rating categorical variables

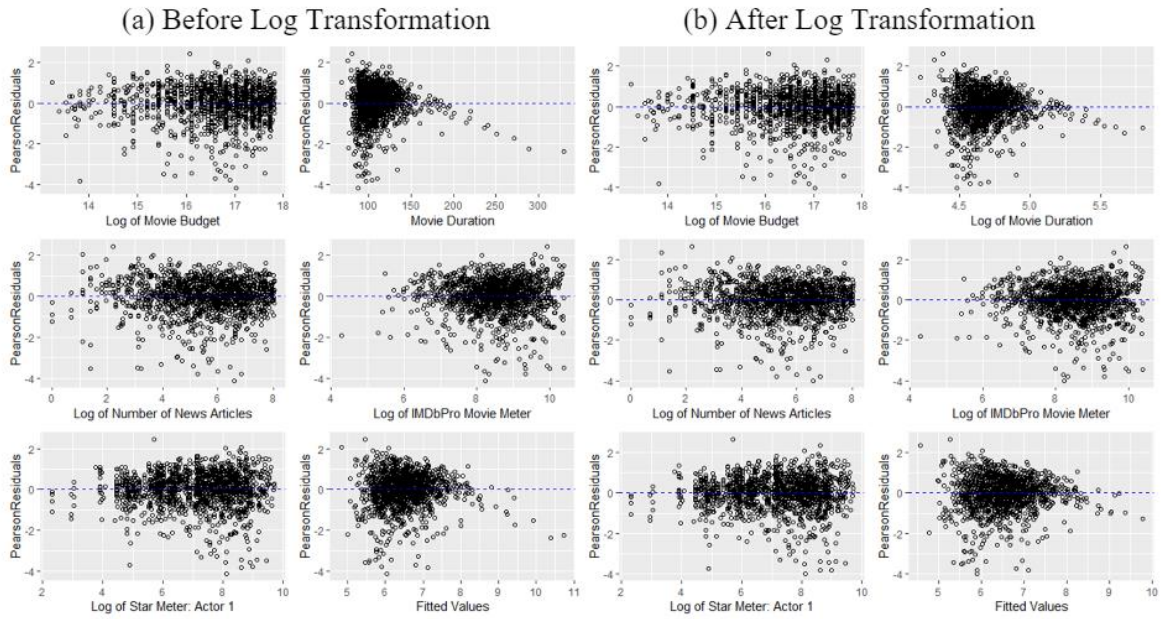


Figure 12: Residual Plots for Multi Linear Regression Models

Table:

Table 1: Simple Linear Regression Results

Simple Linear Regression Results

	Dependent variable:							
	IMDb Score							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log of Movie Budget	-0.152***							
	(0.030)							
Movie Duration		0.021***						
		(0.001)						
Log of Number of News Articles			0.207***					
			(0.017)					
Log of Number of Faces				-0.129***				
				(0.046)				
Log of IMDbPro Movie Meter					-0.480***			
					(0.028)			
Log of Star Meter: Actor 1						-0.113***		
						(0.020)		
Log of Star Meter: Actor 2							-0.063***	
							(0.023)	
Log of Star Meter: Actor 3								-0.001
								(0.027)
Constant	9.024***	4.143***	5.371***	6.588***	10.589***	7.308***	7.004***	6.517***
	(0.504)	(0.132)	(0.097)	(0.040)	(0.242)	(0.142)	(0.183)	(0.231)
Observations	1,451	1,451	1,451	1,451	1,451	1,451	1,451	1,451
R^2	0.017	0.188	0.092	0.006	0.165	0.022	0.005	0.00000
Adjusted R^2	0.016	0.187	0.092	0.005	0.165	0.022	0.005	-0.001
Residual Std. Error (df = 1449)	1.048	0.953	1.007	1.054	0.966	1.045	1.055	1.057
F Statistic (df = 1; 1449)	25.028***	334.390***	147.322***	8.047***	286.888***	33.223***	7.585***	0.003
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 2 VIF Scores

	Log of Movie Budget	Movie Duration	Log of Number of News Articles	Log of IMDbPro Movie Meter	Log of Star Meter: Actor 1
VIF Score	1.062089	1.076030	1.392903	1.478687	1.086854

Table 3: Non-Linearity Test

	Test stat	Pr(>|Test stat|)
Log of Movie Budget	-1.2120	0.2257016
Log of Movie Duration	-3.3721	0.0007656 ***
Log of Number of News Articles	1.3087	0.1908322
Log of IMDbPro Movie Meter	-0.9837	0.3254228
Log of Star Meter: Actor 1	-2.1700	0.0301718 *
Turkey test	-3.0656	0.0021723 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4: ANOVA Tests Table

(a) Polynomial Degrees for Non-Linear Variables

	Log of Movie Duration	Log of Star Meter: Actor 1
Model 1	1	1
Model 2	2	2
Model 3	3	2
Model 4	2	3
Model 5	3	3

(b) ANOVA Test

	Res .Df	RSS	DF	Sum of Sq	F	Pr(>F)
Model 1	1425	838.27				
Model 2	1423	833.13	2	5.1381	4.3849	0.01263*
Model 3	1422	832.55	1	0.5817	0.9928	0.31923
Model 5	1421	832.55	1	0.0001	0.0002	0.98844

(c) ANOVA Test

	Res .Df	RSS	DF	Sum of Sq	F	Pr(>F)
Model 1	1425	838.27				
Model 2	1423	833.13	2	5.1381	4.3849	0.01263*
Model 4	1422	832.13	1	0.0001	0.0003	0.98726
Model 5	1421	832.55	1	0.5816	0.9927	0.31924

Table 5: IMDb Prediction Regression Models

IMDB Score Prediction Model	
	Dependent variable:
	IMDb Score
Log of Movie Budget	-0.270***
	(0.026)
Log of Movie Duration	10.113***
	(0.973)
Log of Movie Duration^2	-1.794**
	(0.799)
Log of Number of News Articles	0.064***
	(0.016)
Log of IMDbPro Movie Meter	-0.381***
	(0.029)
Log of Star Meter: Actor 1	-1.048
	(0.827)
Log of Star Meter: Actor 1^2	-1.469*
	(0.772)
Maturity Rating: R	0.162***
	(0.047)
Genre: Action	-0.308***
	(0.062)
Genre: Adventure	0.142**
	(0.069)
Genre: Sci-Fi	0.026
	(0.073)
Genre: Thriller	-0.068
	(0.059)
Genre: Western	0.214
	(0.161)
Genre: Sport	0.093
	(0.096)
Genre: Horror	-0.553***
	(0.077)
Genre: Drama	0.342***
	(0.053)
Genre: War	0.241*
	(0.123)
Genre: Crime	0.138**
	(0.058)
Genre: Biography	0.164*
	(0.088)
Genre: Comedy	-0.123**
	(0.055)
Genre: Fantasy	0.104
	(0.073)
Genre: History	0.140
	(0.123)
Genre: Family	-0.060
	(0.080)
Country: USA	-0.163***
	(0.055)
Top Cinematographer	0.147**
	(0.067)
Top Director	0.243***
	(0.085)
Top Distributor	0.108**
	(0.043)
Constant	13.768***
	(0.521)
Observations	1,451
R^2	0.486
Adjusted R^2	0.476
Residual Std. Error	0.765 (df = 1423)
F Statistic	49.783*** (df = 27; 1423)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Coefficient Test

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Log of Movie Budget	-0.27047	0.02579	-10.488	< 2e-16 ***
Log of Movie Duration	10.1132	0.9735	10.389	< 2e-16 ***
(Log of Movie Duration)^2	-1.79434	0.79903	-2.246	0.024879 *
Log of Number of New Article	0.06367	0.01575	4.044	5.54e-05 ***
Log of Star Meter: Actor 1	-1.04849	0.82747	-1.267	0.205326
(Log of Star Meter: Actor 1)^2	-1.46876	0.772	-1.903	0.057303 .
Log of IMDBpro Movie Meter	-0.38104	0.02878	-13.241	< 2e-16 ***
Maturity Rating: R	0.16154	0.0471	3.43	0.000621 ***
Genre: Action	-0.30779	0.06211	-4.956	8.06e-07 ***
Genre: Adventure	0.1422	0.06923	2.054	0.040154 *
Genre: Sci-Fi	0.02646	0.0733	0.361	0.718113
Genre: Thriller	-0.06773	0.05859	-1.156	0.247853
Genre: Western	0.21355	0.16084	1.328	0.184505
Genre: Sport	0.09277	0.09565	0.97	0.332263
Genre: Horror	-0.55338	0.0772	-7.168	1.22e-12 ***
Genre: drama	0.34153	0.05294	6.452	1.51e-10 ***
Genre: War	0.24134	0.12324	1.958	0.050381 .
Genre: Crime	0.13808	0.05759	2.398	0.016630 *
Genre: Biography	0.16393	0.08836	1.855	0.063770 .
Genre: Comedy	-0.12302	0.05529	-2.225	0.026243 *
Genre: Fantasy	0.10429	0.07271	1.434	0.151690
Genre: History	0.14023	0.12281	1.142	0.253687
Genre: Family	-0.05983	0.07991	-0.749	0.454123
Country: USA	-0.16334	0.05524	-2.957	0.003157 **
Top Cinematographer	0.14738	0.06651	2.216	0.026858 *
Top Director	0.24322	0.08533	2.85	0.004430 **
Top Distributor	0.10814	0.04339	2.492	0.012804 *
Constant	13.76825	0.5211	26.421	< 2e-16 ***