MGSC 661 Midterm Final Report

Date: 24 October 2024

1. Introduction

The film industry has rapidly grown into a billion-dollar industry as a result of technological advancements such as virtual and augmented reality and 360-degree video experiences. The motion picture industry had profits of over 49.68 billion USD in 2023 and is expected to reach 81 billion by 2029. In such a high-stakes complex industry, filmmakers increasingly use predictive analytics to enhance their creative processes, and consolidate their competitive edge by analyzing the public's preferences.

12 blockbuster movies are set to be released between October and November this year[1] but there is currently no established predictive model that accurately forecasts the reception of these movies. To address this gap, a sample of 1930 movies released from 1936 to 2018 was collected from IMDb to build a predictive model. Each film was described by 27 components, broadly grouped into four categories: film identifiers, film characteristics, cast characteristics, and production characteristics. Using statistical tools and analyses executed through R, predictors that had low predictive power or highly correlated with others were removed. A refined set of predictors was then used to build and refine the predictive model that obtains the lowest mean squared error (MSE). The following report outlines our approach to develop the predictive model and presents the predicted IMDb ratings for the 12 upcoming movies.

2. Data Description

2.1 Data inconsistencies

No missing value was found in the training data. However, there were some inconsistencies:

1. Dummified genre columns were inconsistent with the "genres" column. There were a total of 21 unique genres separated from the "genres" columns compared to the 13 unique columns that came with the dataset. Thus, to ensure data integrity and accuracy, the analysis was conducted using the "genres" column that was manually processed in favor of the original dummified columns.
2. One movie with "Official website" listed as its country of production was corrected to "USA."
3. Different names were used for the same distributor. For example, "Lionsgate" was also listed as "Lions Gate Films". Thus, distributor names were standardized to ensure consistency.

2.1 Removal of Predictors

Film identifiers held no predictive power and thus, were all removed. Plot keywords were also disregarded as natural language processing would be required to utilize the information fully. Since IMDb star meters were present in the dataset, actor names were removed as they provided a more quantifiable measure.

In the determination of a film's success, the distributor has more influence than the production company as the former is responsible for marketing, promoting, and securing the film's release in various markets.

[1] The 12 movies are as follows: Your Monster, Venom: The Last Dance, Hitpig!, A Real Pain, Elevation, The Best Christmas Pageant Ever, Kanguva, Red One, Heretic, Bonhoeffer: Pastor. Spy. Assassin, Gladiator II, Wicked.

Additionally, with 768 production companies compared to 334 distributors in the dataset, retaining the production company as a predictor increases the complexity of reclassification and the risk of overfitting. Thus, it was removed.

Less than 2% of the films were in languages other than English. Similarly, around 3% were in black and white instead of color, rendering these two predictors largely unary, and thus were dropped as predictors.

Separate simple linear regression models were also generated between IMDb score and (1) aspect ratio, (2) release year, and (3) release day. However, no meaningful relationship was observed between these predictors and IMDb score, and the r^2 values were extremely low, with these predictors explaining less than 4% of the variance in IMDb scores. Additionally, the regression model with release year suggested that IMDb score decreases over time, which does not align with logical expectations. Furthermore, Motion Picture Association film rating system only came into effect after 1968, suggesting that maturity ratings before that are likely inaccurate (1.76% of data). Thus, all movies before 1968 and the three predictors were removed.

2.2 Dummification and processing of remaining variables

The number of films released each month varied significantly, with October having the highest number of releases (Appendix Fig 3). Using it as a reference, a simple regression model between IMDb score and release month revealed significant differences in average scores across certain months. This is likely due to seasonality, as films are often released to coincide with specific seasons or festivities. Thus, the release month variable was retained.

With films prior to 1968 removed, the number of distinct maturity ratings was reduced to 10. As the maturity rating system evolved over time, the remaining maturity ratings were grouped into three broader categories to reflect the change: PG, PG-13, and R, with R serving as our reference category (see Appendix Fig 4). Using a simple regression model, it revealed notable differences among these categories, indicating that this reclassification was appropriate for a clearer understanding of the data.

There were 33 different film production countries, with the USA accounting for over 80% of the data points. To simplify the analysis, the countries were dummified and evaluated using a regression model, with the USA set as the reference category. Only countries that exhibited a significantly different IMDb score compared to the USA were retained as individual categories, narrowing it down to four main groups: USA, UK, Canada, and Others. Comparing the regression models before and after reclassification, the coefficients remained significant, suggesting that the reclassification retained key relationships, helped simplify the model, and reduced the risk of overfitting.

A histogram plot of the number of faces on promotional materials revealed a right-skewed distribution (Appendix Fig 5), with 40% of the films showing no faces. A simple regression model indicated a weak correlation with IMDb score ($r^2 = 0.009$). However, the variable was retained considering the potential influence of the number of faces on a film's marketing appeal, which affects the eventual rating. To simplify the model, it was transformed into a binary variable, with 1 indicating the presence of at least one face, and 0 indicating the absence of faces.

Similarly, histogram plots of film duration, number of news articles, budget, and movie meter revealed right-skewed distributions (Appendix Fig 5). In the respective simple regression models, outliers and heteroskedasticity were observed, and the r^2 values were mostly below 0.01, indicating that these predictors explained very little of the variance in IMDb scores when assessed through a simple linear regression model. To address these issues, logarithmic transformations were applied, reducing skewness and mitigating the influence of outliers (0.95% of films with 0 news articles were removed in the process). Thereafter, the use of non-linear models or splines may be necessary to more accurately capture complex relationships between these predictors.

After standardizing the distributor names, they were further reclassified. Based on the market share distribution from Nash Information Services, the top 15 distributors were retained as individual categories. Distributors that performed significantly differently from those producing fewer than five films were retained in their respective categories. All remaining distributors were then grouped under the category 'Others,' resulting in a final total of 20 distinct subgroups—a significant reduction from the original 334. This reclassification process was essential, as the IMDb scores were found to differ across various distributors (Appendix Fig 6) thus, valuable information would have been lost if this predictor were dropped.

For the processing of the cinematographer and director variables, separate lists of famous directors and cinematographers were extracted from IMDb. Binary variables were then created, assigning a value of 1 to films that featured a famous cinematographer or director. This approach was favored as renowned directors or cinematographers are known to significantly influence a film's quality, marketing, and audience reception. Simple regression models subsequently confirmed the validity of this classification, as the coefficients for both variables were statistically significant.

Outliers and heteroskedasticity were prevalent for the remaining continuous variables on actor star meters. None of these variables exhibited a significant relationship with the IMDb score in the individual regression models—only the model for Star Meter 2 produced a positive r^2 value. Despite this, as actors are known to substantially impact the reception of a film through their acting skills and fame, an average of the three star meters was calculated to provide a more balanced and representative measure rather than dismissing these variables entirely.

2.3 Correlation between predictors, and between predictors and the target variable

After the predictors have been pre-processed, a correlation matrix was executed to assess potential collinearity between predictors. Using a threshold of 0.3, it was found that certain variables such as thriller and crime (coefficient: 0.37), drama, and the 'duration_more_median' variable (coefficient: 0.33) exhibited potential collinearity. Despite these findings, none of the affected variables (primarily dummified genre variables) were removed, as they carried significant information for separate films which would otherwise be lost. When examining the correlation between each predictor and IMDb score, the following variables stood out with higher correlations: film duration (0.36), director's fame (0.23), and the number of news articles (0.23), alongside genres such as drama (0.34). These variables were deemed more likely to enhance the predictive power of the model (assuming linearity holds). In contrast, predictors such as the star meters had much lower correlation values (below 0.05). With these insights, the most relevant predictors were carefully selected for inclusion in the final predictive model.

3. Model Selection

To assess each continuous variable's relationship with the IMDb score, the following approach was adopted: first, determine if there was a linear relationship between the target and each independent variable. If the relationship was not linear, the variable was transformed with a logarithmic to evaluate whether it could linearize. This transformation also helped address potential issues of heteroskedasticity in the variables. Each variable was modeled with different polynomial degrees to capture more complex non-linear relationships if necessary. The Tukey Test for linearity served as the basis to determine the best model for each predictor while maintaining the linearity assumption (See Appendix Fig 7).

The variables representing the release month number and the categorical variables—country, distributor, and maturity categories—along with dummy variables related to director fame, cinematographer fame, and the presence of faces in the movie poster were modeled without any transformation. These variables significantly influenced the IMDb score, highlighting their importance in predicting film performance (Appendix Fig 5).

In addition to applying a logarithmic transformation to the four predictors mentioned in Section 2 – movie meter IMDBPro, movie budget, duration, and number of articles – the average fame of actors also underwent a logarithmic transformation to meet the linearity assumption (Appendix Fig 8). Furthermore, two of these variables were modeled using polynomials: movie duration was modeled with a polynomial of degree four and the number of articles in news with a polynomial of degree two, enhancing the model's adherence to the linearity assumption (See Appendix Fig 13).

The model also incorporated four interaction terms, with three of them involving the movie meter IMDbPro. A film’s popularity may vary based on cultural differences related to the country of production, with higher budgets correlating to increased audience expectations. Specifically, the interaction between movie meter and country highlights how cultural context influences a film's reception, with different countries exhibiting varying audience responses. Additionally, the interaction between movie meter and month indicates that certain times of the year, such as summer or holiday seasons, enhance a film’s popularity, reflecting strategic release timing. The interaction between movie meters with budget further underscores that films with higher production budgets are often anticipated to perform better, aligning audience expectations with financial investment. Lastly, the interaction between country and duration reveals that a film's length can influence its popularity based on the filming location, suggesting that different locales may prefer varied pacing and storytelling styles. Each interaction enriches the model, offering a more nuanced understanding of the factors that drive a film’s success.

For the choice of model, the generalized linear model (GLM) was preferred over ordinary least squares (OLS) regression due to its superior handling of heteroskedasticity and more accurate standard error estimations. Additionally, issues of collinearity and the high dimensionality of the model arose from the large number of dummy variables. The initial model was established using all dummy variables from the genre to address this, and a subsequent one incorporated only the significant dummy variables. Though ANOVA estimation showed no major improvement between the two models (MSE of the second model (0.63) was just slightly lower than that of the first model (0.64)), cutting down the number of genres from 21 to 8 reduced dimensionality while ensuring that the model still captured the key genres that influence IMDb score. Thus, the second model was chosen and only significant genres were retained.

The final issue addressed was the presence of outliers. A scatter plot of studentized residuals (see Appendix Fig 18) revealed data points three standard deviations away from the mean residuals, identified as outliers negatively impacting model performance. After removing these outliers beyond three standard deviations, the MSE improved significantly to 0.48, and the adjusted r^2 increased to 0.539, notably higher than that of each individual linear regression model. No additional outliers were removed beyond this threshold to prevent overfitting and ensure the model's ability to predict new movies not included in the dataset.

4. Results & Discussion

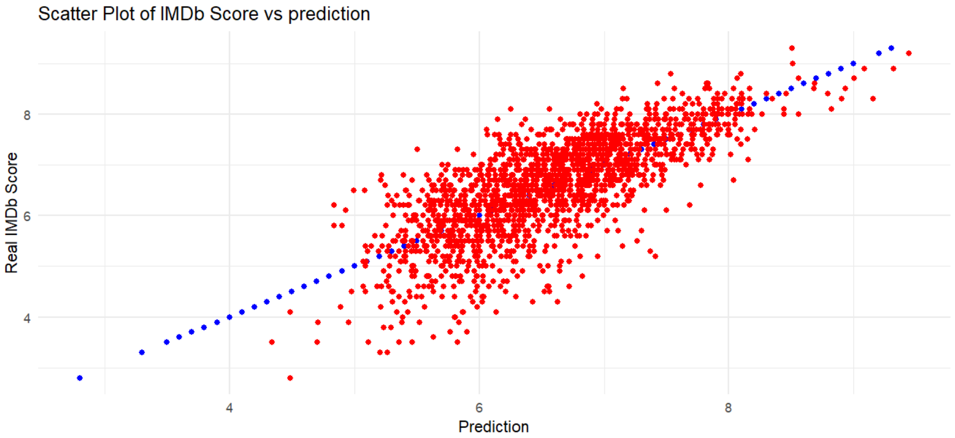
4.1 Model components

Following the above analysis, the predictive model consisted of the following components: (i) significant dummified genres, (ii) distributor, (iii) director fame, (iv) cinematographer fame, (v) maturity category, (vi) number of news articles, (vii) faces in the promotional materials (dummy), (viii) average fame of actors,

(ix) interaction terms between movie meter IMBDPro and country, release month and budget, (x) interaction term between duration and country.

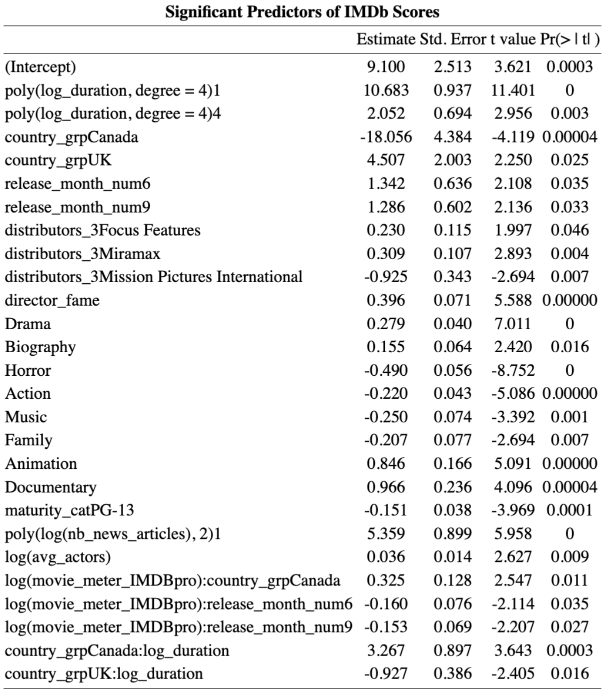
4.2 Cross Validation

After running the model with 10-fold cross-validation, it retrieved an MSE of 0.48. With a LOOCV approach, the MSE was 0.46. While LOOCV showed a slight improvement in model performance, the computational cost associated with this method outweighs its benefits, given the insignificant difference in results. The graph below illustrates the model's performance in predicting IMDb scores compared to actual scores. The predictions closely align with the line representing the real IMDb scores, indicating a low variance between the predicted and actual values. This suggests that the model effectively captures the underlying trends in the data.



Though our model performed moderately well in predicting average IMDb scores, there remains room for improvement, especially at the lower and higher ends of the score distribution. Additionally, certain influential variables, such as plot words, were excluded from the current model. Future iterations could benefit from incorporating dummy variables representing the most common plot themes to assess whether this inclusion improves performance and reveals any interactions between plot themes and movie genres. However, it would be necessary to exercise caution in incorporating these variables to prevent overfitting arising from the introduction of excessive dummy variables, which can impair the model's performance on unseen data.

4.3 Predictive power of predictors



Significant Predictors of IMDb Scores

	Estimate	Std. Error	t value	Pr(> ǀ tǀ)
(Intercept)	9.100	2.513	3.621	0.0003
poly(log_duration, degree = 4)1	10.683	0.937	11.401	0
poly(log_duration, degree = 4)4	2.052	0.694	2.956	0.003
country_grpCanada	-18.056	4.384	-4.119	0.00004
country_grpUK	4.507	2.003	2.250	0.025
release_month_num6	1.342	0.636	2.108	0.035
release_month_num9	1.286	0.602	2.136	0.033
distributors_3Focus Features	0.230	0.115	1.997	0.046
distributors_3Miramax	0.309	0.107	2.893	0.004
distributors_3Mission Pictures International	-0.925	0.343	-2.694	0.007
director_fame	0.396	0.071	5.588	0.00000
Drama	0.279	0.040	7.011	0
Biography	0.155	0.064	2.420	0.016
Horror	-0.490	0.056	-8.752	0
Action	-0.220	0.043	-5.086	0.00000
Music	-0.250	0.074	-3.392	0.001
Family	-0.207	0.077	-2.694	0.007
Animation	0.846	0.166	5.091	0.00000
Documentary	0.966	0.236	4.096	0.00004
maturity_catPG-13	-0.151	0.038	-3.969	0.0001
poly(log(nb_news_articles), 2)1	5.359	0.899	5.958	0
log(avg_actors)	0.036	0.014	2.627	0.009
log(movie_meter_IMDBpro):country_grpCanada	0.325	0.128	2.547	0.011
log(movie_meter_IMDBpro):release_month_num6	-0.160	0.076	-2.114	0.035
log(movie_meter_IMDBpro):release_month_num9	-0.153	0.069	-2.207	0.027
country_grpCanada:log_duration	3.267	0.897	3.643	0.0003
country_grpUK:log_duration	-0.927	0.386	-2.405	0.016

Table 1: Coefficients, standard error, test statistic, and p-values of predictors with p-value < 0.05

Upon analyzing the coefficients in the final model, several predictors emerged significant in terms of their predictive power.[2] Notably, the polynomial terms for log(duration) highlighted that the first-degree term had a strong positive coefficient of 10.683, suggesting that audiences may favor extended narratives, perhaps perceiving them as more comprehensive or engaging. The fourth-degree polynomial term also demonstrated significance with a coefficient of 2.052. This suggests that while longer films generally receive higher ratings, the effect of duration may not be linear and could vary at different lengths. Furthermore, the interaction terms between movie metrics and country of production revealed significant relationships, particularly with the Canadian production group showing a notable increase when combined with log(duration) (coefficient of 3.267). This finding suggests that Canadian films, when longer in duration, may particularly resonate with audiences, potentially due to cultural storytelling preferences or unique narrative structures.

Additionally, the number of news articles covering a film substantially positively affects ratings. The first polynomial term has a significant coefficient of 5.36 ($p = 3.09 \times 10^{-9}$), indicating that increased media coverage is strongly associated with higher ratings. In contrast, the second term, with a p-value > 0.05, lacks statistical significance. This suggests that while more media coverage positively impacts IMDb

[2] Refer to Appendix 5.4 for further elaboration on the remaining pertinent predictors in the model

scores, the effect diminishes after a certain point, highlighting the crucial role of initial coverage in shaping audience perceptions of film quality.

The country of production also significantly influenced IMDb scores, with Canada exhibiting a notable negative impact, with films found to be 18.056 points lower than US productions, which served as the reference category. This stark difference suggests potential biases or perceptions in how Canadian films are received compared to their US counterparts. In contrast, films produced in the UK are expected to score higher by 4.507. This positive coefficient indicates that UK films tend to be rated higher, which may be attributed to the global recognition of British cinema. These findings highlight the country of production's significant role in shaping audience perceptions and ratings.

As anticipated, seasonality, which was captured with information on the month of release, played a crucial role. Movies released in June and September exhibited positive coefficients of 1.34 (p = 0.035) and 1.29 (p = 0.033), respectively, suggesting that these months were more favorable for movie releases. This may be attributed to various factors, such as the start of the summer movie season in June, which typically sees higher audience turnout, and September when audiences return from summer vacations and seek new entertainment options. These trends indicate that strategic timing of releases can significantly influence a film's performance and audience reception.

Among the 21 genres from the dataset, Documentary demonstrated the highest predictive power with a coefficient of 0.966, closely followed by Animation at 0.846. This strong association with higher IMDb ratings suggests that these genres resonate well with consumer preferences, likely due to their engaging storytelling and ability to connect with audiences on a deeper level. Documentaries often provide insightful perspectives on real-world issues, appealing to consumers' interests in learning and understanding complex subjects. Similarly, Animation captivates audiences with its creativity and visual appeal, making it a favorite across all age groups.

These findings underscore the complexity of factors influencing IMDb scores, highlighting genre and production-related variables as critical components in predicting film success.

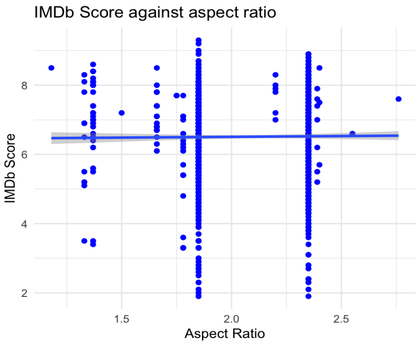
4.4 Prediction out of the dataset

With the refined model, we successfully predicted the IMDb scores for the 12 upcoming movies, as detailed below. Gladiator II is projected to receive the highest score and Elevation the lowest.

	Movie	Predicted score	Standard Error	95% LCL	95% UCL
1	Venom: The Last Dance	7.09	0.331	6.45	7.74
2	Your Monster	5.99	0.140	5.72	6.27
3	Hitpig!	6.55	0.203	6.15	6.94
4	A Real Pain	6.96	0.171	6.63	7.30
5	Elevation	5.70	0.124	5.46	5.94
6	The Best Christmas Pageant Ever	6.59	0.151	6.30	6.89
7	Kanguva	6.77	0.154	6.47	7.07
8	Red One	6.62	0.265	6.10	7.14
9	Heretic	7.35	0.327	6.71	7.99
10	Bonhoeffer: Pastor. Spy. Assassin.	6.68	0.120	6.45	6.91
11	Gladiator II	8.32	0.406	7.52	9.11
12	Wicked	7.98	0.339	7.31	8.64

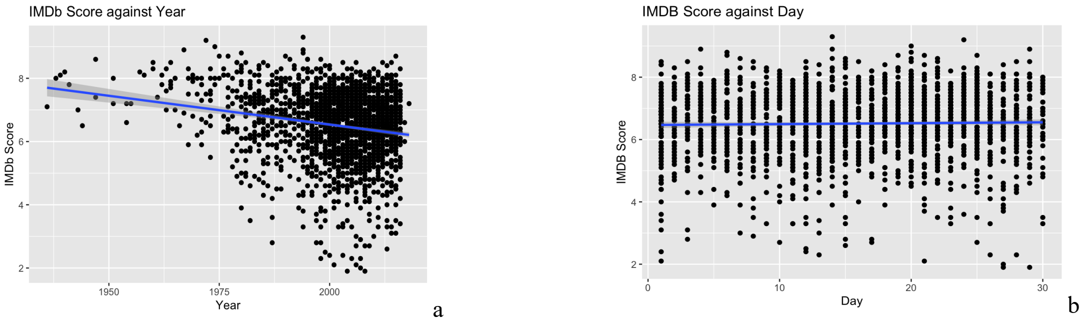
5. Appendix

5.1 Data Description



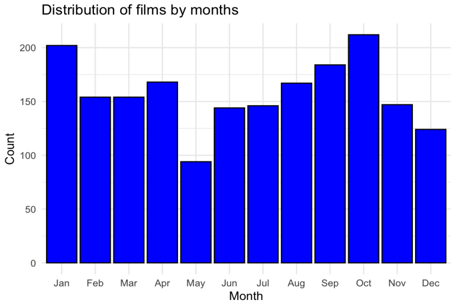
Appendix Fig 1: Scatterplot and regression line of IMDb score against aspect ratio of the films

No visible pattern was observed in Appendix Fig 1, aside from the fact that most of the movies within the dataset had an aspect ratio of 1.85 and 2.35. A simple regression model suggests that as the aspect ratio increases by 1 unit, the IMDb score increases by 0.12. Intuitively, aspect ratio should not significantly influence a movie rating because it primarily affects the film's visual presentation rather than the narrative or performance elements that typically drive audience reception and ratings. Thus, aspect ratio was removed from the model consideration.



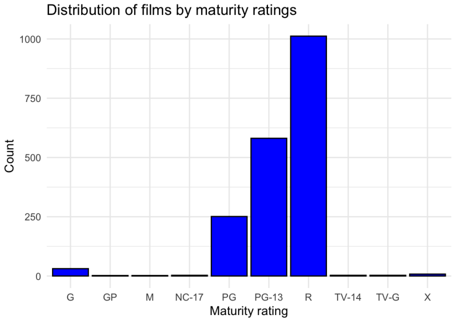
Appendix Fig 2a: Scatterplot and regression line of IMDb score against release year. Appendix Fig 2b: Scatterplot and regression line of IMDb score against release day

As anticipated, the number of films released over the decades has increased alongside the growth of the entertainment industry and advancements in technology. This rise in the volume of films contributed to a broader distribution of IMDb scores, reflecting the diversity in content, quality, and audience reception. Although newer films may seem to perform worse than those released decades ago, this deduction is misleading, particularly given that the sample size for films from earlier decades was significantly smaller. Thus, the release year was omitted as a predictor. Regarding the movie's release day, the regression model coefficient was not statistically significant. Furthermore, including this variable introduces a risk of overfitting due to the presence of 30 different categories (days). Thus, also omitted.



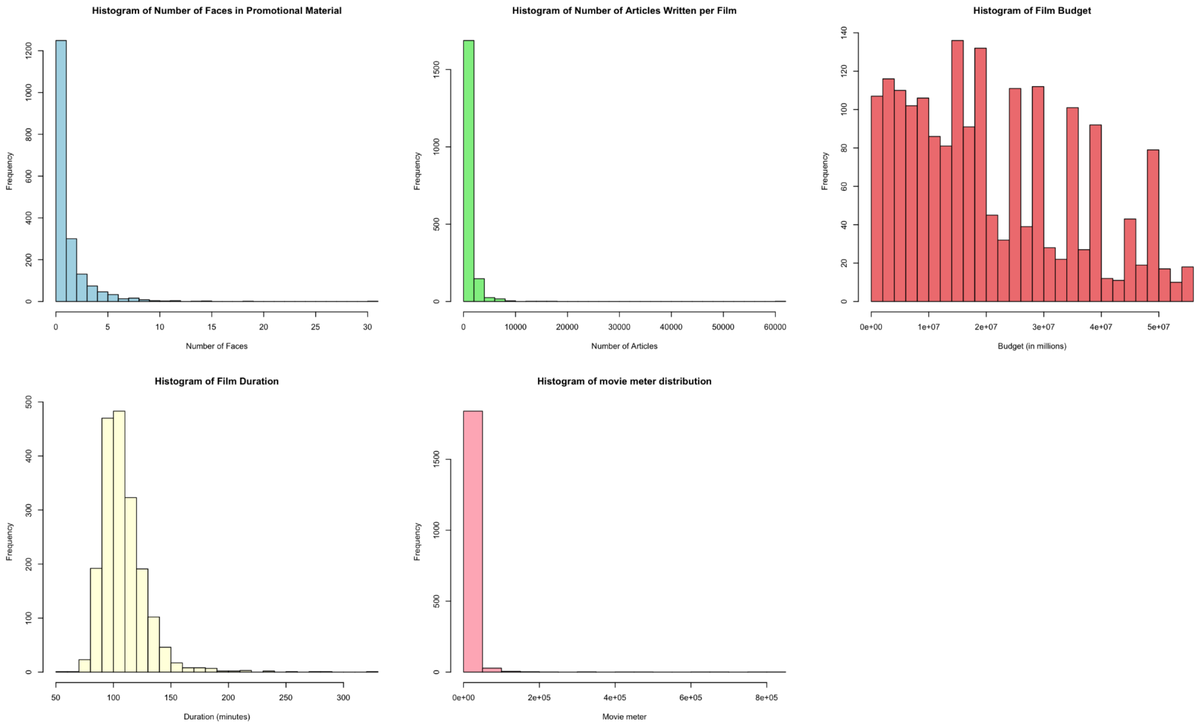
Appendix Fig 3: Distribution of movies released across the months

Based on Appendix Fig 3, it was evident that there were differences in the number of films released across the months. A simple regression model using IMDb scores with October as the reference category revealed significant differences in scores across several months (March, May, August, September, and December). These differences may be attributed to holiday seasons, special festivals, or other social factors. While it would have been possible to group the months into quarters, doing so could dilute the significance of certain months. For instance, May had the fewest films released over the decades; grouping it with April and June to form the second quarter would diminish its significance, given that both April and June had more releases. Thus, to preserve the distinct impact of each month on IMDb scores, the months were retained in their original form.



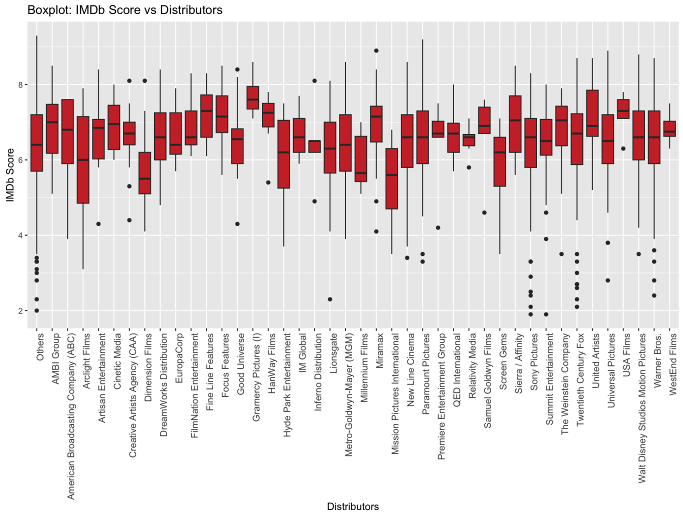
Appendix Fig 4: Distribution of movies by maturity ratings

According to the Motion Pictures Association film rating system implemented in 1968, the ratings have evolved over the last few decades. Thus, the maturity ratings were regrouped into four main categories: PG (G, GP, M, PG, TV-G), PG-13 (PG-13, TV-14), R, and Adult (NC-17, X). However, the "Adult" category only comprised 0.58% of the data. Thus, it had little bearing on the predictive model, so they were filtered out, and only the three broad categories, PG, P,-13, and R, were retained. Since the maturity rating "R" had the highest count among the maturity ratings, it was used as the reference category in our predictive model.



Appendix Fig 5: Histograms illustrating key attributes of films. (a) Number of faces in promotional material, (b) Number of articles written for each film, (c) Budget distribution, (d) Film duration, (e) Movie meter

Based on the above figures, a right skew is observed in the distribution of all five predictors. While the number of faces variable was transformed into a binary variable to account for the large proportion of films without any faces on promotional materials, a log transformation was applied to the remaining predictors after adjusting for films with zero values in their respective fields. This approach was primarily adopted to address heteroskedasticity and the impact of outliers, which are evident in the above figures.



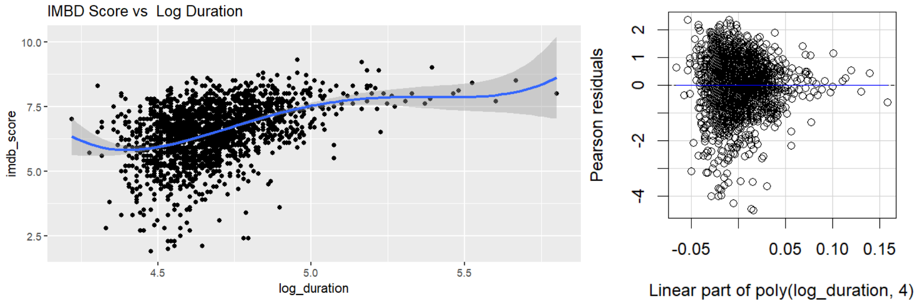
Appendix Fig 6: Boxplot of IMDb Score for preliminary reclassification of distributors

The training dataset initially comprised 334 distributors, posing the risk of overfitting without adequate pre-processing. In the preliminary investigation of the distributor variable's significance on IMDb scores, the naming conventions for distributors were standardized, reducing 56 distributors. The number of films distributed by each distributor was tallied, and a cumulative proportion was calculated, sorting the distributors from highest to lowest based on their film distribution counts. A threshold of a minimum of 5 films per distributor was established, classifying those distributing fewer than five films (accounting for less than 20% of the dataset) as "Others." This process reduced the number of distributors to 42 distinct categories. A box plot was then generated to evaluate the distributor's potential as a predictor in the predictive model. As illustrated in Appendix Fig 6, a simple regression model revealed significant variations in IMDb scores across the reclassified distributors. Given distributors' considerable influence on a movie's marketing and outreach before and after its release, retaining this information in our predictive model was crucial.

Further reclassification was conducted, using "Others" as the reference category in a simple regression model. The top 15 significant distributors, as extracted from Nash Information Services, retained their original categories. In contrast, those who did not exhibit substantial differences in IMDB scores from "Others" were further classified as "Others."

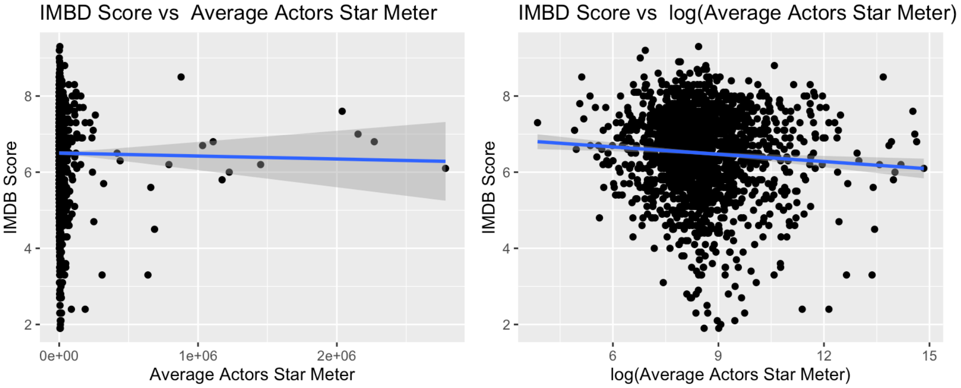
5.2 Model Selection

Duration: This variable does not follow a linear relationship with the target variable. However, if the variable is transformed with a logarithm and follows a polynomial regression of degree four, the Tukey Test for linearity outputs a p-value=0.0505. Under a 95% confidence interval, the null hypothesis cannot be rejected, supporting a linear relationship between the independent and target variables.

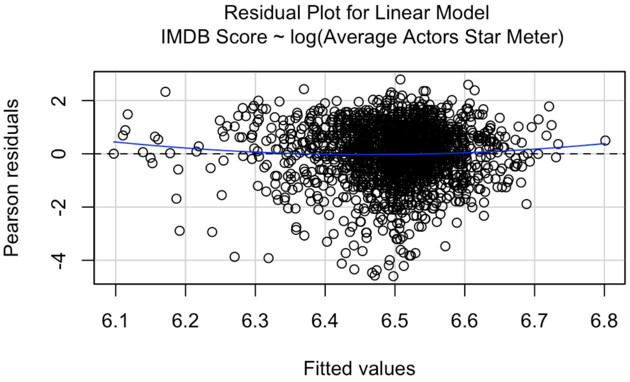


Appendix Fig 7: Left: Scatterplot of Log(Duration) vs IMDB Score. Right: Residual plots of degree 4 polynomial regression for log(duration).

Average Actors: As mentioned above, an average was calculated to use the information provided by the actor's star meters. A log transformation was applied to the independent variable to facilitate the modeling of this independent variable with the dependent variable (See Appendix Fig 8 to check the relationship between this created variable and the IMDB Score with and without transformation). Once the variable was transformed, we tested a simple linear regression against the dependent variable to assess whether a linear relationship existed. In this case, the Tukey Test for Linearity outputs a p-value = 0.07. Thus, at a 95% confidence level, the null hypothesis cannot be rejected, supporting a linear relationship between the transformed dependent variable and the dependent variable (Appendix Fig 9).

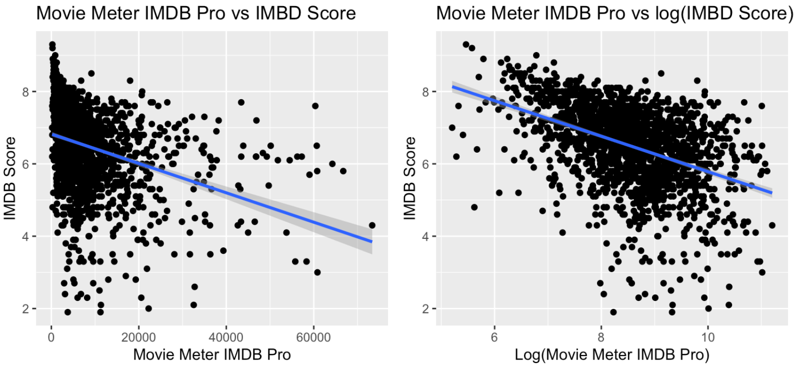


Appendix Fig 8: Scatterplots of the variable "Average Actors Star Meter" against IMDB Score with and without transformation.

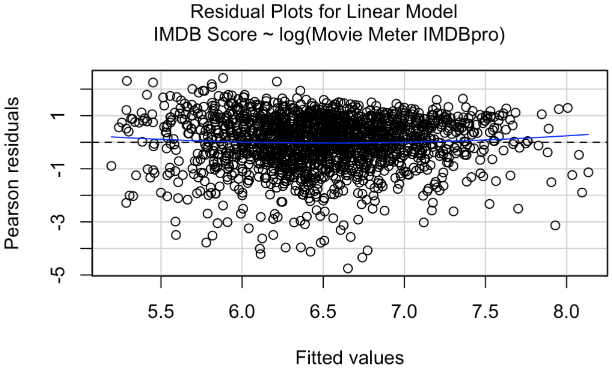


Appendix Fig 9: Residual Plot of the variable "log(Average Actors Star Meter)"

movie_meter_IMDBpro: As previously discussed, this variable was modeled using a natural logarithmic transformation (Appendix Fig 10). Once the variable was transformed, we tested a simple linear regression against the dependent variable to assess whether a linear relationship existed. In this case, the Tukey Test for Linearity outputs a p-value = 0.063. Under a 95% confidence level, the null hypothesis cannot be rejected, supporting a linear relationship between the dependent and independent variables. Therefore, in the aggregated model, this transformed variable will be modeled as a linear regression (Appendix Fig 10).

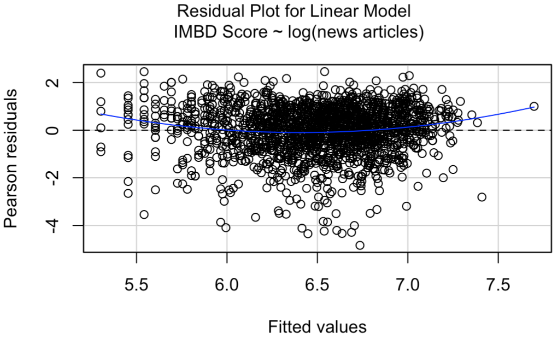


Appendix Fig 10: Scatterplots of variable "Movie Meter IMBDpro" with and without transformation.

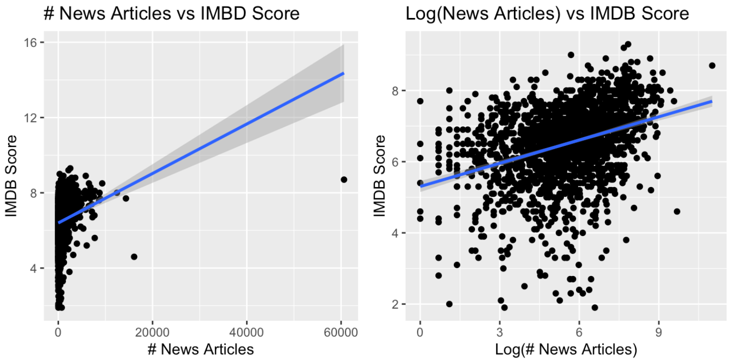


Appendix Fig 11: Residual Plot of the variable "log(Movie Meter IMBDpro)"

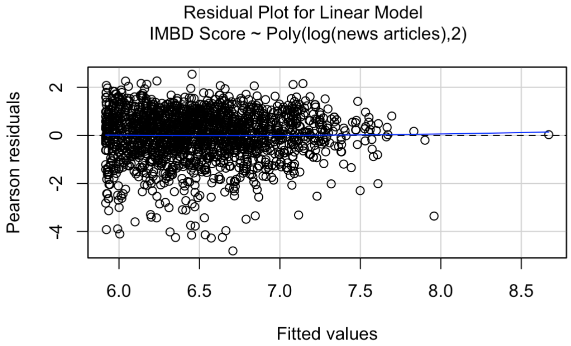
Nb_news_articles (number of articles in the news from the film's main country): As mentioned previously, natural logarithm was utilized to address non-linearity and the impact of outliers (Appendix Fig 13 shows the scatterplots of this transformed variable with and without transformation). We performed a single linear regression to confirm its significance and test for linearity. As seen in Appendix Fig 12, a quadratic relationship must be considered in the linear regression for the linearity assumption to hold (Tukey Test had a p-value = 2.441×10^{-7}). Therefore, we tried a degree 2 polynomial regression. In this case, the linearity assumption was held with a p-value of 0.6355 for the Tukey Test for Linearity (see Appendix Fig 14). Thus, we will model this relationship using a polynomial regression of degree 2 for the aggregated model.



Appendix Fig 12: Residual Plot of log(news articles)



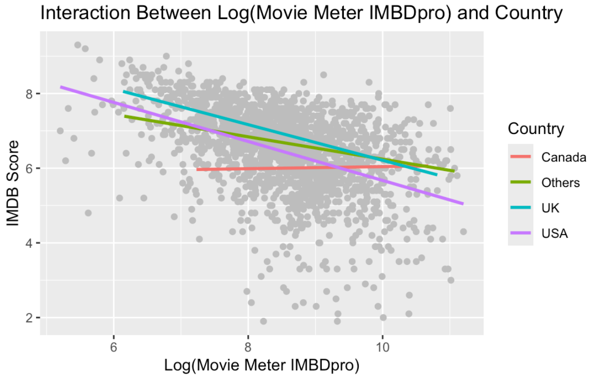
Appendix Fig 13*:* Scatterplots of variable “News Articles” with and without transformation



Appendix Fig 14*:* Residual Plot of degree 2 polynomial regression of log(news articles)

Interactions between log(movie_meter_IMDBpro) and other variables:

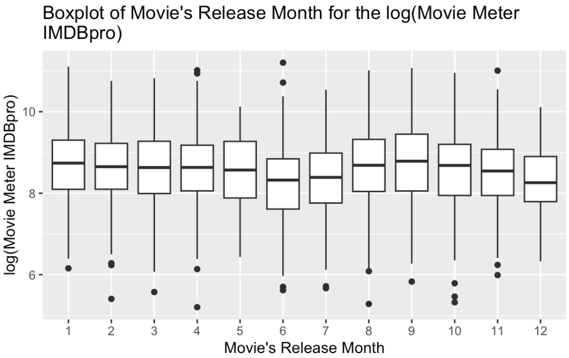
- With Country: Since the variable movie meter indicates a movie’s popularity, we hypothesized that a movie's popularity varies from country to country due to cultural differences in viewing preferences, marketing strategies, or even audience behavior. Appendix Fig 15 suggests that the linear effect of log(movie meter IMDBOpro) for IMDB Score varies per group of country(ies).



Appendix Fig 15*:* Relationship between log(Movie Meter IMDBpro) per country with IMDB Score

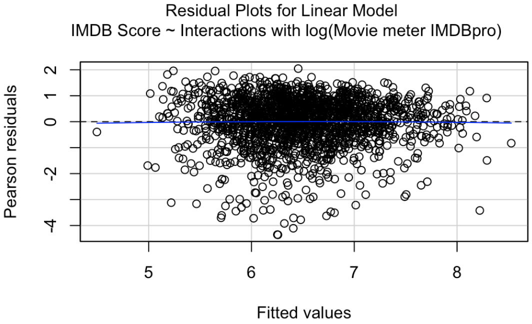
- With log(movie_budget): Although the budget did not clearly relate to the dependent variable, we hypothesized that a movie's popularity is related to its budget; the higher the budget, the higher the expectations about the film, and therefore, the higher the popularity.
- With Movie Release Month: We hypothesized that certain months are more popular for movie releases, such as summer and the holiday season when audiences have more free time. These

months often coincide with exciting releases that attract the audience's attention (see Appendix Fig 16).



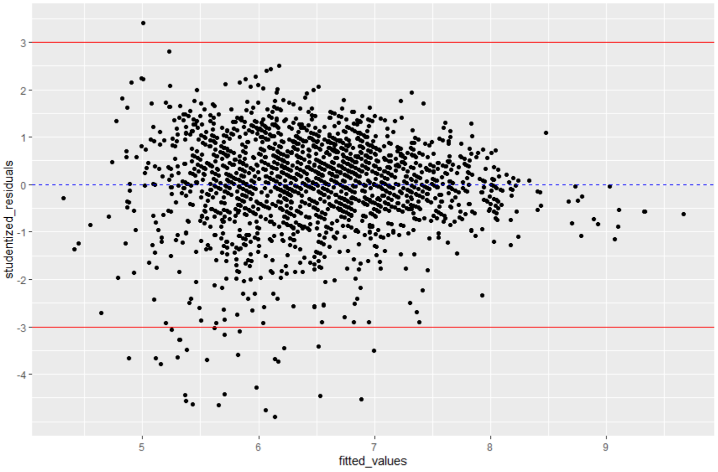
Appendix Fig 16: Relationship between log(Movie Meter IMDBpro) and release month

A linear regression was performed considering the individual variables and the interactions. The result showed that all the interactions were significant except for the interaction between log(movie's budget) and log(Movie Meter IMDBpro) (however, when considering this interaction in the aggregated model, the test MSE decreased by 0.03 units). The Tukey Test for Linearity outputs a p-value of 0.9736, showing that the relationship between the independent variables (individual variables and interactions) with the dependent variable is well explained through linear regression. Appendix Fig 17 displays the residuals from the regression, showing that the relationship is well explained by the linear regression model with no discernible pattern in its residuals:



Appendix Fig 17: Residual Plot of Interactions with Log (Movie Meter IMDBpro)

5.3 Model Selection



Appendix Fig 18: Studentized residual plot with red lines showing our threshold for removing outlier points

Data points below or above the threshold lines were removed to ensure the predictive model was not skewed by outliers. This step is crucial for enhancing the accuracy and reliability of the model by focusing on the more representative data points that fall within the expected range. By eliminating these extreme values, the model can better capture the underlying relationships among the variables, leading to more robust predictions and insights.

5.4 Predictive model

In addition to the strong positive influence of Animation and Documentary genres on IMDb scores discussed in section 4.3, the Drama genre also showed a significant positive coefficient of 0.28 ($p = 3.38 \times 10^{-12}$). This trend can be attributed to the emotional depth and complex narratives typical of drama films, which often resonate deeply with audiences and critics alike. Conversely, Horror films exhibited a notable negative coefficient of -0.49 ($p = 2 \times 10^{-16}$), suggesting that they generally receive lower IMDb ratings. Since horror films aim to evoke fear and suspense, they may fall behind in terms of storytelling or character development, leading to lower ratings. Other genres, such as Action (-0.22, $p = 4.05 \times 10^{-7}$) and Music (-0.25, p = 0.000710), are also linked to lower ratings. Action films often focus on high-octane sequences and special effects rather than character-driven narratives, which may result in mixed reviews. Similarly, while musical films can be visually and sonically captivating, they sometimes struggle to achieve critical acclaim if the plot or performances do not meet audience expectations.

Director's fame also emerged as a critical predictor; films directed by renowned directors had an average score of 0.396 points higher, underscoring the importance of a well-regarded director in elevating a film's perceived quality. A director's reputation can instill confidence in audiences and critics alike, leading to higher expectations for the film's artistic and narrative quality. Established directors often have a track record of successful films, which can create anticipation and excitement around their new projects. This positive association can translate into higher ratings, as audiences may be more inclined to view the film favorably due to the director's prior successes.

The distribution also played a significant role in influencing IMDb ratings. Notably, films released by Miramax exhibited a positive coefficient of 0.31 (p = 0.0039), indicating that these films tend to receive higher ratings. This positive association can be attributed to Miramax's reputation for producing and distributing critically acclaimed films. Their established brand recognition may lead audiences and critics to have heightened expectations, resulting in better ratings. Conversely, films distributed by Mission Pictures International demonstrated a negative coefficient of -0.93 (p = 0.0071), suggesting that they generally score lower on IMDb. This could indicate that the films from this distributor might not meet audience expectations or be perceived as lower quality, which can significantly impact their ratings.

These findings, particularly the variations across genres, highlight the importance of different factors in shaping audience perceptions, as reflected in IMDB scores. The model suggests that strategic release planning can enhance ratings, particularly for specific countries and genres. Furthermore, the significant impact of director fame and media coverage points to potential areas of focus for marketing efforts, providing valuable insights for filmmakers, distributors, and marketers aiming to improve the success of their projects. Understanding the complex relationship between film elements and audience preferences will also be crucial for filmmakers and marketers aiming to create content that resonates with viewers and achieves higher ratings.