Midterm: IMDb Prediction Challenge

Presented to

Professor Juan Serpa

By

The R-3POs

MGSC 661

McGill University - Desautels Faculty of Management

Introduction

IMDb (Internet Movie Database) is among the most well-known entertainment rating platforms, incorporating various criteria including critic perception, user perception and interest, and sales into its verdicts. For rating movies, IMDb provides a score on a scale of 1 to 10, and this score is contingent on users' perceptions of the movies being rated. Several factors can affect the IMDb score of a movie, and they include the movie budget, the main actors, directors, producers, and more. However, regardless of the buildup that precedes the release of a movie, user acceptance and ratings can be unpredictable. In this context, developing a predictive model that can accurately forecast the IMDb score of a movie before its release could prove highly valuable. This is the primary focus of this project.

To actualize the primary goal of this project, historical IMDb datasets were leveraged. These datasets contain data points such as directors, producers, genre, main actors, and more. These are generally some of the readily observable factors that shape viewers' perceptions of a movie. However, there are instances where the relationship between these apparent factors and viewer acceptance is not straightforward. Approaching this through the lens of statistical modeling offers an opportunity to try to capture this complex relationship.

Along those lines, this project involved various analyses being conducted to understand the ideal approach for doing this within the context of regression modelling. This report details the processes and explorations which drove the creation of a final model, evaluates its viability, and speaks to the implications that can be drawn from interpreting its results.

Data Description, Processing, and Model Selection

The dataset comprises data from 1,930 movies, encompassing various variables capturing film characteristics, production details, cast information, and more. A comprehensive understanding of the data's distribution and variable relationships were crucial in the initial phases of model building.

To begin, three variables—movie title, movie ID, and IMDb link—were excluded from the dataset as they were identifiers which provided no meaningful insights into predicting IMDb scores. After removing these variables, the dataset contained 39 potential predictor variables. These variables required an in-depth analysis to determine their suitability for predicting the final IMDb score.

To gain a better understanding of each variable, exploratory data analysis was conducted to assess their frequency, distribution, and their relationships with the IMDb score. The distribution of IMDb scores was

analyzed using summary statistics, histograms, and boxplots. The analysis revealed a slight left-skew in the IMDb score distribution, with most scores clustered toward the higher end. This is supported by the median IMDb score of 6.6, which is slightly higher than the mean score of 6.5. The IMDb scores range from 1.9 to 9.3, indicating a wide range, encompassing both very low and very high IMDb scores.

The distribution of numerical independent variables was similarly examined. This examination highlighted skewness in the distributions (see Exhibit 1 in Appendix). Notably, variables such as movie duration, the number of news articles related to the movie, and the star meters of actors exhibited right-skewness. Additionally, the aspect ratio, which represents screen size, was explored with a hypothesis that IMAX movies (defined by an aspect ratio greater than or equal to 1.90) might influence higher ratings. A regression discontinuity test was conducted using a dummy variable for if the movie is IMAX or not and the interaction term between aspect ratio and the dummy term. However, both the dummy and the interaction term showed higher p-values, suggesting minimal impact on IMDb scores (see Exhibit 2 in Appendix). The aspect ratio remained statistically significant and was retained. The analysis of the remaining numerical variables indicated that movie budget exhibited right-skewness, while the day of release showed a normal distribution. Additionally, scatterplots were created to test the linearity of these variables which showed that almost all variables were non-linear (see Exhibit 3 in Appendix). A correlation matrix was constructed to investigate potential strong correlations between numerical variables, but no significant correlations were found (see Exhibit 4 in Appendix).

For categorical independent variables, a distinct approach was adopted. Genre dummies were refined by matching plot keywords with genre keywords (see Exhibit 5 in Appendix). Processed genres were amalgamated into more comprehensive genre dummies, including action, adventure, sci-fi, thriller, musical, romance, western, sport, horror, drama, war, animation, and crime. Release day was deemed irrelevant and excluded. Release year and release month were examined, with release year identified as a significant predictor and easily interpretable, leading to its retention (see Exhibit 6 in Appendix). Black and white movies were found to receive higher ratings than color films, which was statistically confirmed. However, variables like country and language displayed skewed distributions and limited significance, resulting in their exclusion.

A similar approach was applied to directors, cinematographers, and other categorical variables such as language and country, leading to the discarding of these variables for predicting IMDb scores. Text engineering techniques were employed on plot keywords to identify influential keywords. Out of curiosity, it was hypothesized that a movie featuring an excessive number of genres might lack focus and consequently receive lower ratings. To explore this, the variable representing the number of genres was converted into a numeric measure. Subsequent analysis on all the other text variables' character length

revealed that the variables 'genre length,' 'distributor length,' and 'actor1 length' exhibited statistical significance at the 5% level. Consequently, these variables were retained, while the remaining ones were excluded from the analysis. The production company was retained for the final model due to its significant contribution to IMDb score variance. While initial considerations involved dropping this variable due to its sparse distribution, a simple regression test revealed that it alone accounted for nearly 20% of the variance in IMDb score, with most dummies proving significant. However, it's worth noting that only one production company, Miramax, from the test set appeared in the training set. To address this, we opted to transform it into a single dummy variable, the production company either is Miramax, or it is not. Consistency was maintained by performing the same processing of categorical variables with the test set, ensuring that the model remains consistent and allowing for IMDb score predictions using the predict function.

Further analysis examined relationships between independent variables and the dependent IMDb score, identifying statistically significant factors. Among categorical variables, release month, language, maturity rating, and actor 3 were found to lack statistical significance, while country, aspect ratio, distributor, director, actor 1, actor 2, color film, cinematographer, and production company were deemed statistically significant based on their p-values. A correlation matrix revealed weak negative correlations between IMDb score and the budget of the movie, release year, number of faces in the movie poster, and IMDb Pro's movie meter, indicating higher values in these variables were associated with slightly lower IMDb scores. Actor 3's star meter showed a very weak negative correlation, suggesting no significant relationship. Release day, number of news articles, actor 1's star meter, and actor 2's star meter had slight positive correlations with IMDb score, indicating higher values in these variables corresponded to slightly higher IMDb scores. The duration of the movie exhibited a strong positive correlation, indicating that longer movies were associated with higher IMDb scores (see Exhibit 4 in Appendix).

Numerical variables displaying skewness and non-linearity underwent log transformations, resulting in linearity according to the Tukey test and significant improvements in their p-values, rendering them statistically significant predictors (see Exhibit 7 in Appendix). These transformed variables were considered for inclusion in further modeling and testing, addressing skewness while enhancing predictive power. Outliers were also examined, revealing potential outliers for release year, duration, number of news articles, actor 1's star meter, actor 2's star meter, actor 3's star meter, number of faces in the movie poster, and IMDb Pro's movie meter. This analysis provided valuable insights into the distribution of these numerical predictors and their potential impact on the predictive model.

The final model selection was based on three criteria: R-squared, MSE, and the Akaike Information Criterion (AIC). R-squared provides insight into the model's ability to capture variability within the dependent variable, whereas MSE evaluates the model's predictive accuracy on new data points. On the

other hand, AIC assesses the model's parsimony, aiming to achieve maximum predictive capacity with the minimal number of predictors.

Initially, a model incorporating all available features was configured to gauge the maximum attainable R-squared, which stood at approximately 48%. Subsequently, insignificant variables were methodically eliminated to achieve a balance between an R-squared nearing 48%, a desirable cross-validation MSE, and a minimized AIC (see Exhibit 8 in Appendix). This elimination was done by performing linear regression modeling to assess the relationship between various predictors and IMDb scores for a dataset of movies. P-values and R-squared values were calculated for each predictor to determine their statistical significance and ability to explain the variability in IMDb scores.

Prior to this analysis, categorical variables related to movie genres were processed by encoding them into binary form and incorporating them into the training set. The findings revealed that certain predictors exhibited higher p-values and lower R-squared values, indicating their limited significance. These less significant predictors included release day, release month, language, country, actor 1's star meter, actor 2's star meter, actor 3's star meter, and various movie genres (e.g., action, adventure, scifi, thriller, western, sport, drama, war, musical, romance, animation, and crime). Other predictors were identified as significant due to their low p-values and moderate to high R-squared values. These influential predictors included movie budget, movie meter IMDb pro, release year, duration of the movie, maturity rating, number of news articles, color films, aspect ratio, and horror films. Based on these findings, the variables deemed insignificant were removed from consideration as predictors in the final model and the variables identified as significant were subject to further exploration and incorporation into the modeling process.

The model's linearity, along with that of its predictors, was evaluated using the Tukey test. As mentioned above, the numerical variables that exhibited skewness in their distributions, approximating a chi-square distribution with 2 degrees of freedom, indicating a high occurrence of a few values, and suggesting non-linearity and low prediction power, had log transformations applied to them as an exploration of potential transformations. This log transformation serves as an index for these skewed numerical variables and tends to increase as they increase, addressing the issue of skewness and non-linearity. It became evident that with these transformations, they became relevant to be considered as potential variables in the model. However, for these persistently non-linear variables, various polynomial degrees were experimented with until an optimal fit was achieved. Given the nature of the dataset, the inclusion of splines was avoided due to potential overfitting concerns. After visually inspecting the relationships between the numerical predictors and the target variable, it appeared that polynomial functions might be more suitable for capturing the nonlinear patterns without overfitting to the noise. In contrast, splines might overfit by closely following

the data, potentially incorporating the noise. Given these observations, splines were excluded from the model.

Following model finalization, diagnostic tests were performed to identify outliers, potential collinearity, and signs of heteroskedasticity (see Appendix for Exhibit 9). Mild heteroskedasticity was observed, but no traces of collinearity were found. Outliers were detected and subsequently addressed. Following this, the final model was achieved.

Interpretation of Results

The final model, which includes five numerical and nine categorical variables achieves an R-squared of 0.4649, a 5-fold Cross-Validation MSE of ~0.63, and an AIC of 4543. As multiple R-squared only increases or remains level with the addition of predictors, the R-squared of the final model is compared to that of a model including all predictors available. A model with all predictors explained roughly 48% of the variability in IMDb scores, and the final selected model also captures nearly half of the score variability while achieving impressively low prediction error and high parsimony. By applying iteratively improved modelling techniques on just a few predictors, a high level of predictive power was achieved.

Each included variable achieves a minimum 5% significance level, indeed the majority of them are highly significant with a p-value of less than 0.001. Among the most statistically significant predictors is the IMDBpro Movie Meter score. This metric captures behavioral indicators about public interest in a movie through the number of page views a movie gets on the IMDb site. It is worth noting that this metric reflects interest but does not necessarily indicate sentiment. It is also important to note that this variable may experience periodic fluctuations from one week to another based on spikes of interest not only in the movie being observed but in other movies, given that it is a relative ranking. Interestingly, this means that the Movie Meter score can help capture the level of competition a movie release is facing from other releases on the market. These may be from new releases, or from viral spikes in interest for older films.

The duration and number of news articles each have a positive effect on the predicted IMDb score, but curiously this model indicates an inverse relationship between movie budget and IMDb score. This may feel counterintuitive. It is possible that this variable is capturing some information about movies with a higher budget being held to a higher standard by movie audiences. It is also possible that this predictor is capturing information held by categorical “covariates” in the original dataset which were eliminated from the model predictors but together may have a more intuitive relationship to IMDb score despite not translating well in a mathematical model. A movie being in color as opposed to in black & white has a

slight downward effect on scores, however it is worth noting that approximately 97% of the training data is movies in color. The 3% which were black & white films have a higher mean IMDb score and less variation in the range of scores achieved. The Miramax predictor has similar challenges of improving model performance but not having a clear real-world extrapolation concerning how it affects IMDb scores.

The genre and maturity rating predictors have greater interpretability as to their effect on IMDb scores and, in contrast to those previously discussed, hold implications which can be leveraged by decision makers across the movie industry. Horror is the only genre with a negative impact on scores as compared to genres not in the dataset, while animation has a relatively strong upward impact in the same context. The constructed genre length predictor shows a strong negative impact on IMDb scores, which could imply that when a film lacks clear categorization and can instead be categorized across multiple genres its overall public perception could suffer. Findings for maturity ratings, with TV-14 exerting strong downward pressure on IMDb score and being rated R having a positive effect, seem to tell a similar story of penalization in IMDb score for being technically suitable for most types of audiences but not homing in a particular audience. Executives in the movie industry can use such information when making decisions about which projects are more likely to be well received, be profitable, and are worth greenlighting. Similarly for creatives, this information can provide guidance about which of their projects have more commercial appeal and should be prioritized for pitching.

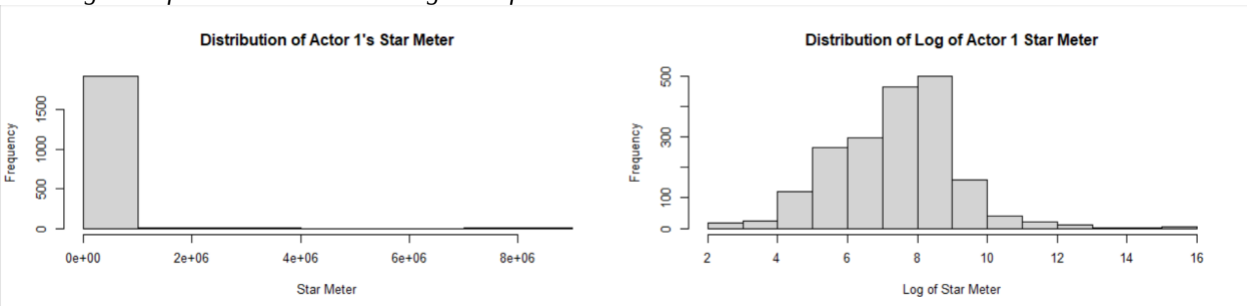
The varying influences of these variables can be used to interpret the predicted performance of the twelve forthcoming releases. The model predicts that all the movies are expected to perform fairly well, with IMDB scores all firmly above the midpoint mark. The Holdovers, a lower budget, two-genre film produced by Miramax is predicted to perform the best. Conversely, the film with the lowest predicted IMDB score is Thanksgiving, a triple genre film including horror (see Exhibit 10 in Appendix for all final movie predictions).

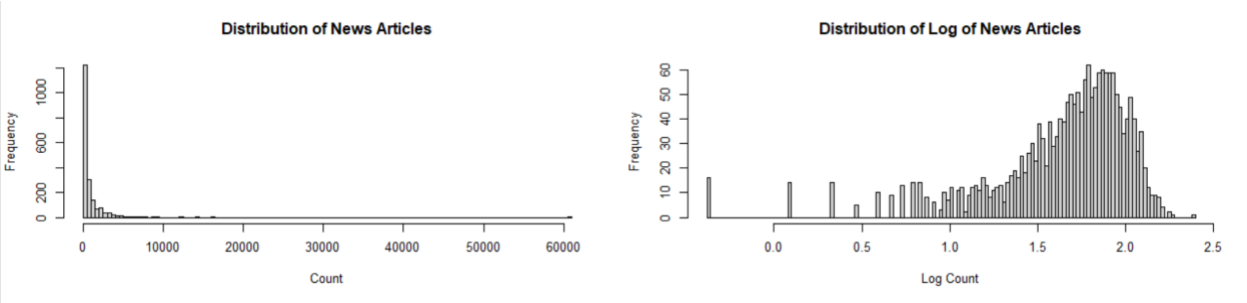
Through this model, predicting public perception to a film– a process which is commonly speculative in nature – is bolstered with analytical rigor.

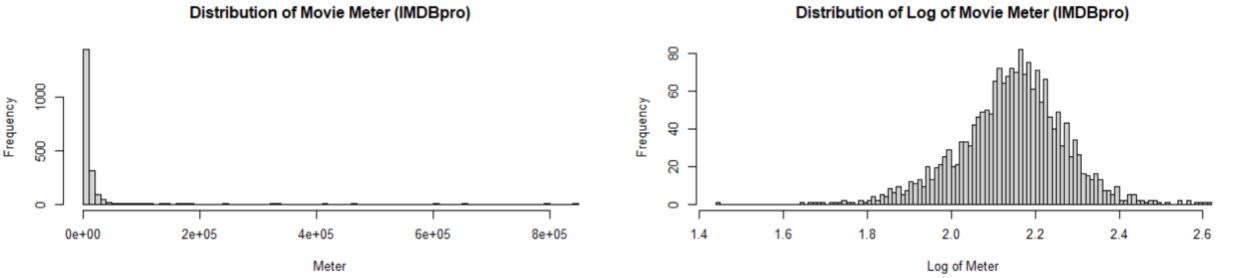
Appendix

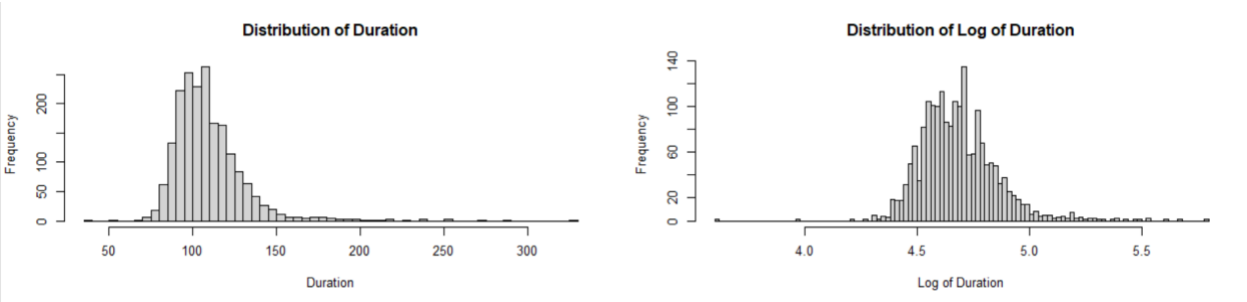
Exhibit 1: Log-Transformation of Numerical Variables for Skewness

With Log Transformation vs Without Log Transformation









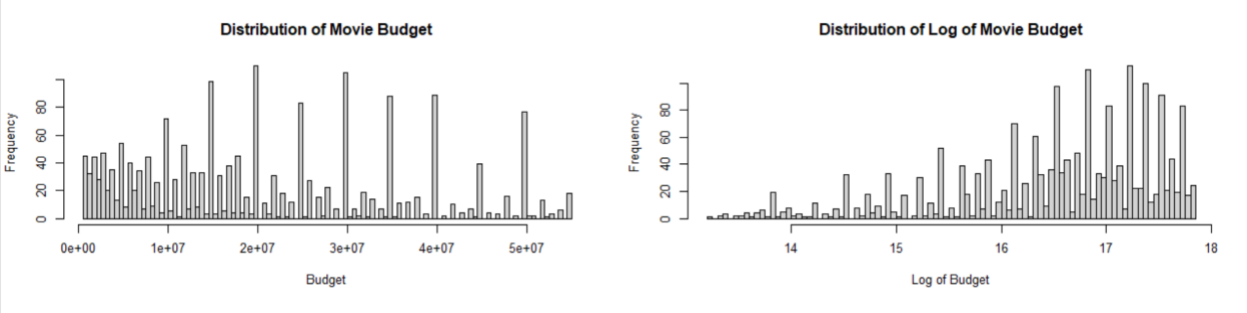
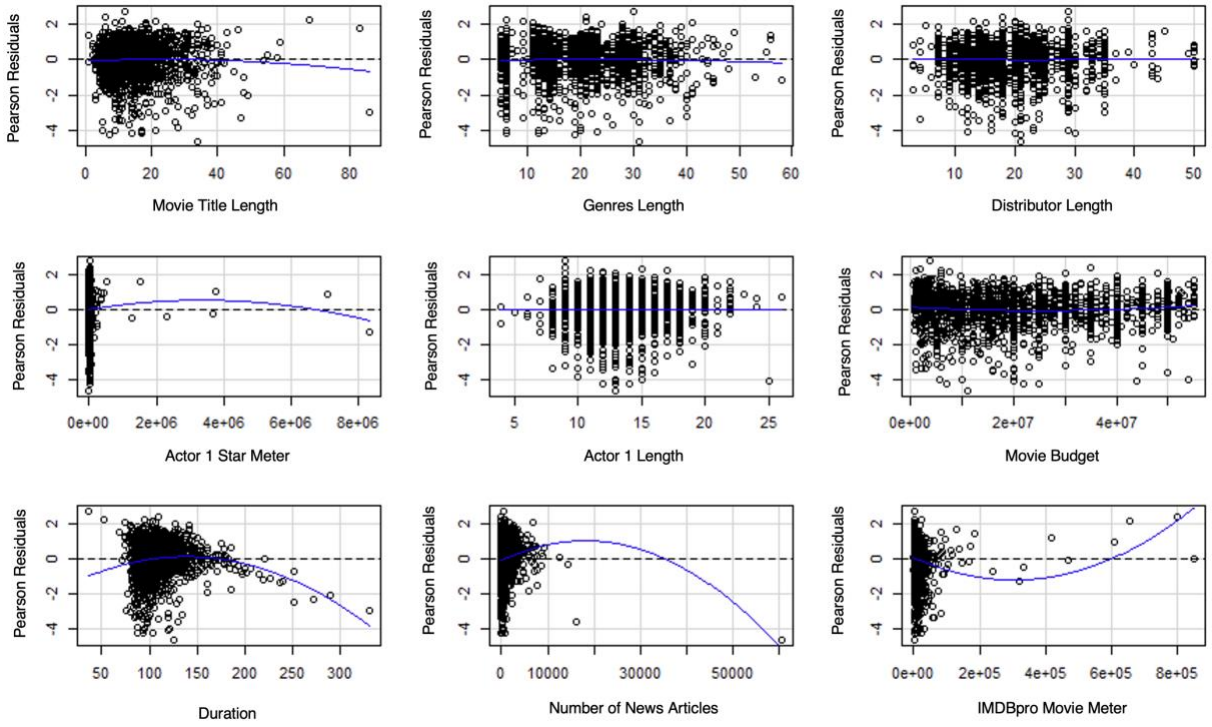


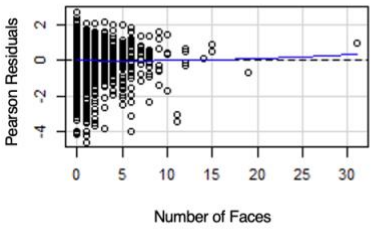
Exhibit 2: Check if IMAX has a significant impact on IMDb score (using the interaction term)

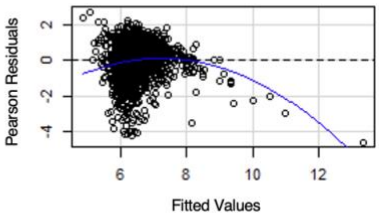
Variables	**Coefficient**	**P-value**
Aspect Ratio	-0.93	0.00726
is_MAX	1.31	0.76277
Aspect Ratio*is_MAX	-0.32	0.86279

Exhibit 3: Log-Transformation of Numerical Variables for Linearity & Significance

Residual Plots without Log Transformation







Tukey Tests

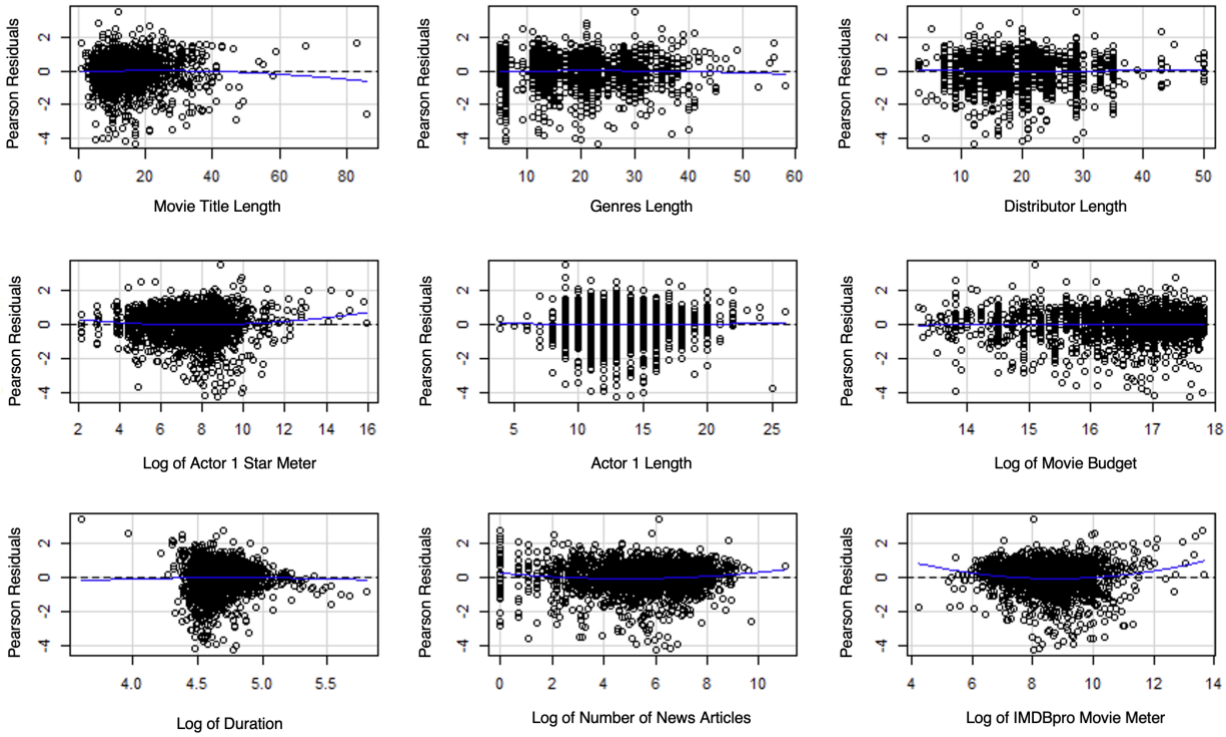
	Test stat	Pr(>Test stat)
Movie Title Length	-1.3453	0.1786968
Genres Length	-0.7670	0.443203
Distributor Length	0.3930	0.6943909
Actor 1 Star Meter	-1.3064	0.191562
Actor 1 Length	0.0118	0.9905920
Movie Budget	3.0697	0.002173 **
Duration	-7.3450	3.026e-13 ***
Number of News Articles	-8.2386	3.187e-16 ***
Movie Meter IMDbpro	6.7554	1.838e-11 ***
Number of Faces on Movie Poster	0.4667	0.640801
Tukey test	-10.4907	< 2.2e-16 ***

Without Log Transformation

	Test stat	Pr(>Test stat)
Log of Movie Title Length	-1.3312	0.1832925
Log of Genres Length	-0.8481	0.3965028
Log of Distributor Length	0.3730	0.7092055
Log of Actor 1 Star Meter	2.7512	0.0059940 *
Log of Actor 1 Length	0.4201	0.6744603
Log of Movie Budget	-0.4921	0.6227289
Log of Duration	-0.4090	0.6826114
Log of Number of News Articles	3.3160	0.0009301 ***
Log of Movie Meter IMDbpro	4.7191	2.54e-06 ***
Log of Number of Faces on Movie Poster	-0.0491	0.9608383
Tukey test	0.6916	

With Log Transformation

Residual Plots with Log Transformation



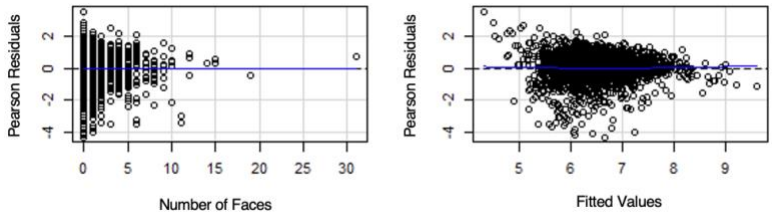


Exhibit 4: Correlation Matrix

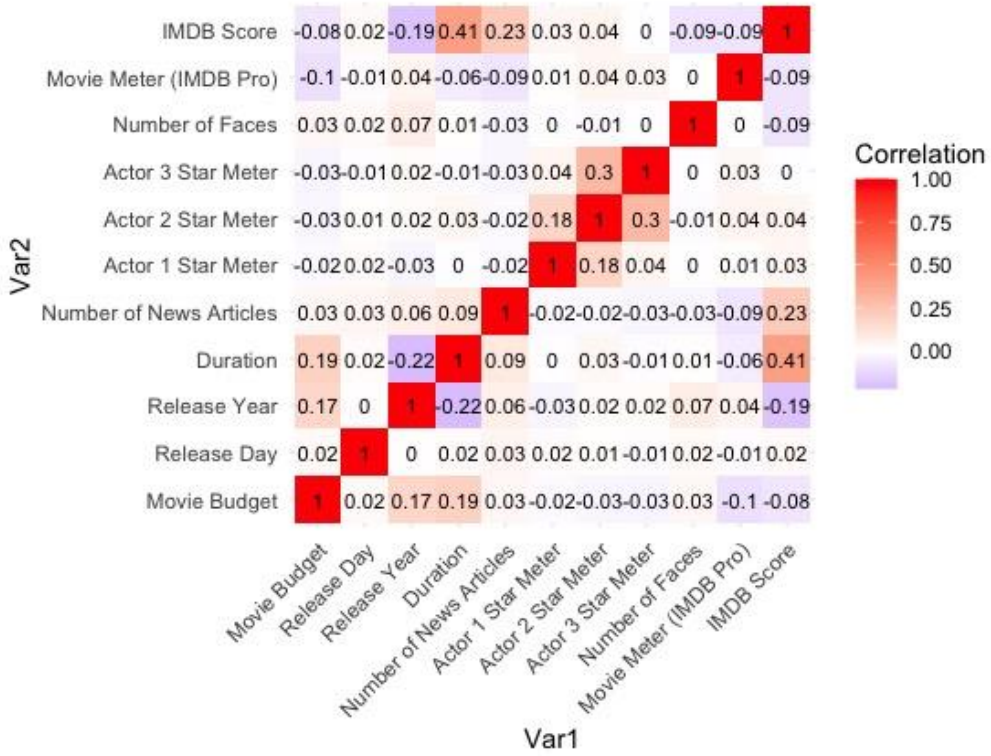


Exhibit 5: Number of Genre Inconsistencies

Genres	**Number of Inconsistencies**
action	9
scifi	0
musical	2
romance	4
western	5
sport	2
war	43
animation	1
crime	4

"Inconsistency" refers to the increase in count of films in each genre once plot keywords are also considered.

Exhibit 6: Release Year Significance Check

Predictor	**R squared**	**P-value**
Release year	0.0380	5.8E-18

Exhibit 7: Significance Check for Numerical Variables vs Log-Transformation of Numerical Variables

Predictor	**R squared**	**P-value**
Movie Title Length	0.002354	0.033077
Genres Length	0.00237	0.032487
Distributor Length	0.00305	0.015239
Actor 1 Length	0.002336	0.033731
Movie Budget	0.006189	0.000542

Exhibit 8: Statistical Summary and MSE Distribution of Final Model

	Dependent variable:
	IMDB Score
Log of Movie Budget	-0.23***
	(0.02)
Log of Duration	1.65***
	(0.13)
Log of Number of News Articles	0.06***
	(0.01)
Log of Movie Meter	-16.08***
	(1.11)
Log of Movie Meter^2	4.24***
	(0.81)
Log of Movie Meter^3	4.30***
	(0.80)
Log of Movie Meter^4	-1.41*
	(0.79)
Is Color	-0.49***
	(0.10)
Length of Genres	-1.55*
	(0.86)
Length of Genres^2	-2.37***
	(0.83)
Biography	0.30***
	(0.07)
Animation	1.24***
	(0.19)
Documentary	0.85***
	(0.21)
Produced by Miramax	0.38***
	(0.15)
Horror	-0.46***
	(0.06)
Drama	0.40***
	(0.04)
Rated R	0.13***
	(0.04)
Rated TV-14	-1.23***
	(0.46)
Constant	2.46***
	(0.62)
Observations	1,923
Log Likelihood	-2,252.51
Akaike Inf. Crit.	4,543.01
Note:	*p<0.1; **p<0.05; ***p<0.01

>

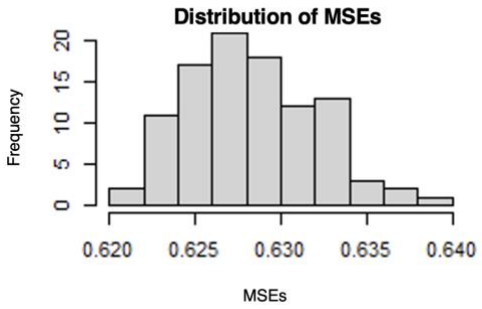
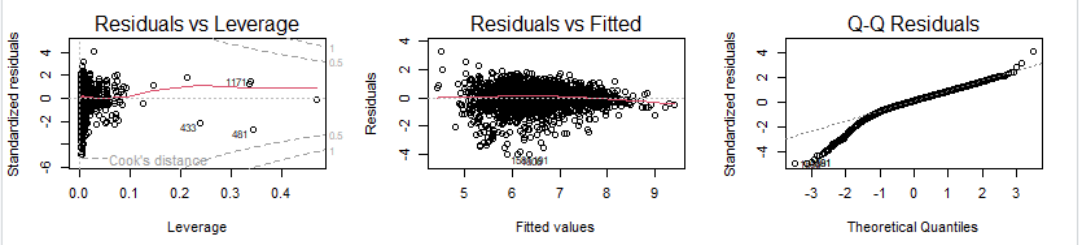


Exhibit 9: Heteroskedasticity and Linearity Check



After correcting for heteroskedasticity, Genres length and TV 14 are less significant, but not a fatal issue.

Exhibit 10: IMDB Rating Predictions with Final Model

Movie Title	**Predicted Score**
Pencils vs Pixels	7.15
The Dirty South	6.81
The Marvels	6.55
The Holdovers	7.91
Thanksgiving	6.39
The Hunger Games: The Ballad of Songbirds and Snakes	7.47
Trolls Bank Together	7.60
Leo	7.22
Dream Scenario	6.79
Wish	7.28
Napoleon	7.71