Predicting IMDb Ratings

Desautels Faculty of Management, McGill University

MGSC661: Multivariate Statistics

Instructor: Juan Camilo Serpa

October 24th, 2024

Introduction 3

Data Description 3

Numerical variables 4

Categorical variables 4

Outliers 5

Model Selection 5

Results 6

Model Performance 7

Cross-validation & Model Predictive Power 7

Significance of Predictors: 7

Conclusion 8

Introduction

IMDb, or the Internet Movie Database, is one of the most popular platforms for accessing extensive information on films, including cast details, plot summaries, and user-generated reviews and ratings. As a critical indicator of public and critical reception, IMDb ratings are often referenced when audiences choose which movies to watch. Given the platform's influence, the ability to predict these ratings holds significant value for filmmakers, studios, and marketers. IMDb scores can shape promotional strategies, influence audience expectations, and impact box office performance.

In this project, the primary objective is to develop a robust predictive model that can accurately estimate IMDb ratings for upcoming films. By analyzing historical IMDb data on variables such as budget, cast, duration, genre, and release dates, this model will help industry professionals gain insights into the factors that drive higher movie ratings. To achieve this, we will employ a range of statistical and machine learning models, including linear regression, polynomial models, and spline regression, to find the best fit for predicting IMDb scores. Understanding these factors not only informs better decision-making around film production and marketing but also helps refine creative and production choices to align with audience preferences and trends. While the model offers predictive power, we also acknowledge the limitations, such as the subjective nature of ratings in data, which may affect the model's precision. Ultimately, the predictive model serves as a valuable tool for anticipating a film's potential success and guiding data-driven strategies within the highly competitive film industry.

Data Description

The data set used for this project contains information on approximately 2000 films, spanning from 1936 to 2018. The data contains a variety of features that can influence its IMDb rating. These features fall into several categories; film characteristics, cast characteristics, and production-related details. By analyzing these characteristics, we can determine which factors have the strongest impact on a movie's final rating.

The dependent variable is the IMDb Rating (imdb_score) which is the target variable we aim to predict. Scores are user-generated ratings on a scale of 1 to 10, reflecting the overall public perception of the film. The ratings in the dataset range from 1.9 to 9.3.

Film characteristics include the budget, release date (month, day, and year), the duration (in minutes), the language, the country, the genre, maturity rating, whether the film is in color or black and white, the number of news articles, number of faces in the main poster, plot keywords, and the movie meter. Cast characteristics include the lead actors and the actor star meter which represents the 2022 ranking of the actor/actress. Production characteristics include the director, the distributor, the cinematographer, and the production company. Given the extensive number of categories that exist within the

‘Production Characteristics’ variables-including 1115 different directors, 768 production companies, 737 cinematographers, and 334 distributors- we chose to exclude this category from further analysis. The predictors are categorized as numerical and categorical groups to examine each type more effectively.

Numerical variables

Key genres in the dataset include Action, horror, drama, and animation. The budget indicates the film’s production budget. It is right skewed with 50% of the movies having budgets ranging between $8.7 million and $30 million. The range is wide, spanning from $560,000 to $55 million. The number of news articles represents the number of articles in the news of the main country about the film. The distribution is highly right skewed with values that range from zero to 60,620 with a median of 286 which indicates that most movies have a lower number of news articles written about them while only a few have a large number. Half of the movies from the dataset were released between the 9th and the 23rd day of the month and between the years 1997 and 2010. The earliest release year is 1936 while the latest release year is 2018. The duration indicates the length of the movies in minutes. Half of the movies last between 96 and 118 minutes. The shortest film is 37 minutes while the longest film lasts 330 minutes. Aspect ratio represents the ratio of the image (width to height). The minimum is 1.18 while the maximum is 2.76. The number of faces represents the number of faces in the main poster of the movie. There are some movies with zero faces in the main poster while the maximum number of faces in a movie’s poster is 31. The star meter for Actors 1, 2, and 3 represents the 2022 ranking of actors and actresses made by IMDb pro where a lower score indicates more fame. The movie meter IMDb pro represents the popularity of movies, where lower scores indicate higher popularity. 50% of the movies have a score that ranges between 2,836 and 10,198. The best ranking for a movie received a score of 71 while the worst rating is 849,550. The presence of outliers in this variable suggests that while most movies achieve moderate popularity, a small number of movies reach exceptional levels of public interest, achieving very low scores. The top 5 movies in the dataset that have the lowest IMDb Pro rankings, indicating they are the most popular movies among the entire dataset are The Purge, Mean Girls, Vanity Fair, Sicario, and The Witch. On the contrary, the 5 least popular films are Strangerland, Exiled, Richard III, The Brothers, and Jeepers Creepers II.

Categorical variables

The release month shows when a movie was first shown in theaters. October is the most common release month, while May is the least common. The primary language of the movie is also important. There are 19 distinct languages in the dataset. Some examples are Aramaic, Cantonese, Dari, Dutch, English, French, and German. Other languages include Hindi, Indonesian, Italian, Japanese, Korean, Mandarin, Mongolian, Portuguese, Spanish, and Zulu. Most films are in English, with 1,892 movies and there are 7 films in French. Other languages appear in 1 to 3 films each. Additionally, 2 films are silent. The country variable shows where the movie was produced. The dataset includes films produced in 34 different countries. The United States produced the most films, with 1,555 movies and the UK comes next with 177 films. Greece, Indonesia, and Taiwan are the least represented. The maturity rating indicates the content rating of the film. The most frequent categories are a rating R (1013), PG-13 (582), and PG (255), while the least frequent ratings include GP (2), M(2), and NC-17(3). The data shows whether a movie is in color or black and white. Most movies are in color, with 1,867 films. Only 63 films are in black and white.

Outliers

(Refer to Figure *2,3,4,5, and 6* in the Appendix)

We identified outliers in several numerical predictors. The number of news articles, the release year, the movie meter from IMDb Pro, and the duration have a moderate number of outliers. The number of news articles could vary greatly depending on the film's visibility or controversy, leading to outliers in this category. Since films span a large time range, older films might appear as outliers, especially when newer films tend to dominate the market. The IMDb Pro variable measures the visibility and popularity of movies on IMDb, which can vary widely between blockbuster hits and lesser-known films, leading to extreme values. Outliers in duration may indicate either very short films (e.g., documentaries or special features) or extremely long ones (epics).

In addition to outliers, we found perfect collinearity between some languages and countries. For example, the language Hindi is perfectly correlated with the country India.

Model Selection

We approached the task of building a predictive model for IMDB scores by first running a baseline linear regression model to understand which predictors will be of interest. We evaluated the p-value of these regressions to assess predictive power of each variable, focusing on variables with p-value lower than 0.05 (Figure 16). For each selected variable, we evaluated linear, polynomial, and spline models based on the scatter plot of that variable in relation to the IMDb score to determine the best approach, balancing fit and complexity (refer to Figure 8-14 in the Appendix).

While evaluating spline regressions, we initially placed knots at quantiles. However, much of the data was concentrated in specific areas, causing knots to cluster too closely, reducing their effectiveness. To resolve this, we visually inspected the data and/or where polynomial lines curved and manually placed knots at points where trends shifted. We then visually identified where polynomial lines curved and placed knots at these points. We adjusted knot placement to maximize R^2 while minimizing the number of knots to prevent overfitting.

Model types were chosen based on the R-squared values and their statistical significance (refer to Figure 16 in the Appendix). Moving to higher-degree regressions was considered when ANOVA p-value was lower than 0.05. After deciding the best model type for each individual variable, we built the final model by iteratively adding the predictors and evaluating changes in the adjusted R^2. Finally, to validate our models, we applied leave-one-out cross-validation (LOOCV) and monitored the impact on MSE. Our goal was to minimize MSE, while maximizing R-squared to ensure that model generalized well and did not underfit or overfit.

To adjust for heteroskedasticity, we used the HC1 type from the heteroskedasticity-consistent covariance matrix estimation. This ensures that our coefficient estimates remain reliable, even if the variance of the residuals is not constant across observations.

The final model's predictors are the release month, colors (whether the movie is in color or black and white), maturity rating, duration, IMDb Pro's rating, the number of news articles, the release year, and the 12 different genres (Drama, Biography, Crime, Comedy, Horror, Action, Family, Music, Romance, Adventure, Animation, and Documentary). To view the selected model for each predictor, see Figure 16 in the Appendix.

Results

Results for 12 movies

Name of the Movie	Predicted Score
Venom: The Last Dance	5.5
Your Monster	5.2
HitPig!	5.6
A Real Pain	6.3
Elevation	4.7
The Best Christmas Pageant Ever	5.7
Kanguva	5.7
Red One	4.0
Heretic	5.8
Bonhoeffer	6.0
Gladiator II	4.6
Wicked	6.1

Model Performance

The final model, referred to as reg13, predicts IMDb movie ratings using a combination of categorical and continuous predictors.

The model yielded an adjusted R-squared of 0.505. This indicates that approximately 50.5% of the variance in IMDb ratings is explained by the model.

The residual standard error was 0.746. This indicates that most predictions deviate by less than one point from the actual ratings.

The F-statistic of 47.51 and p-value of < 2.2e-16 demonstrate that the model is statistically significant overall.

Cross-validation & Model Predictive Power

The cross-validated model using LOOCV returned a mean squared error (MSE) of 0.575. It shows that the model performs reasonably well in predicting IMDb ratings. This MSE value implies that, on average, the model's predictions deviate by approximately 0.758 (RMSE) points from the actual ratings, which is a relatively small margin considering IMDb ratings are typically on a scale from 1 to 10.

Significance of Predictors:

Several predictors were found to have a high statistical significance in the model. To better understand the individual impact of each predictor, we analyzed their effects while holding all other variables constant. This allows us to isolate the contribution of each factor to the IMDb rating, giving us a clearer picture of the drivers behind movie success (refer to Figure 8 in the Appendix).

- Language ('IsEnglish'): Since most of the movies are in English, we created a new column with binary values to indicate whether the language is English or not. For every movie that is in English, the rating decreased by approximately 0.71.
- Color: Color films have a negative impact on the rating, where holding everything else equal, the rating decreases by 0.37.
- Maturity Rating: For every movie classified as PG-13, the rating will decrease by 0.15. Other maturity ratings don't have as much statistical significance.
- IMDbPro Popularity: first two spline terms for IMDbPro were significant. The model found strong evidence that movies with higher IMDbPro popularity were rated significantly lower.
- Number of news articles: Strong positive relationship with IMDb score, indicating that with every additional news article, the rating increases by 0.67.
- Budget: Also significantly affects IMDb rating. An increase of $100 million in the budget would result in only about a 0.094 point. This suggests that although the

relationship is statistically significant, the effect is practically negligible. It indicates that budget alone does not strongly influence movie ratings.

- Genres:
  - Drama: This genre had a significant positive effect on IMDb ratings.
  - Biography: Movies classified as biographies also saw a significant boost in ratings.
  - Horror: Horror films, in contrast, were associated with significantly lower ratings ($p < 0.001$).
  - Animation and Documentary: Both genres had highly positive effects on ratings, with Animation contributing to a 0.985 increase and Documentary contributing a 1.326 increase to IMDb ratings.

Conclusion

The model identifies critical factors influencing the ratings on IMDb, allowing stakeholders to understand the role of different characteristics in predicting a successful movie. By leveraging these insights, movie producers, marketers, and distributors can optimize budgets, target release dates, and promote specific genres to maximize ratings and audience appeal.

However, there are some limitations to consider. We couldn't separate production and marketing budgets, nor account for inflation, which may impact the accuracy of budget-related predictions. Additionally, large categorical variables like directors were excluded due to their size, though incorporating external data to rank them could improve future models.

Looking ahead, future models could incorporate additional predictors like social media presence or whether a movie is a remake or a sequel for a more comprehensive view. Exploring non-linear machine learning models, such as random forests or gradient boosting, could also offer improvements in predictive power. These considerations would allow for even more nuanced predictions and a deeper understanding of the elements that influence movie ratings.

Appendix

Figure 1: Correlation Matrix

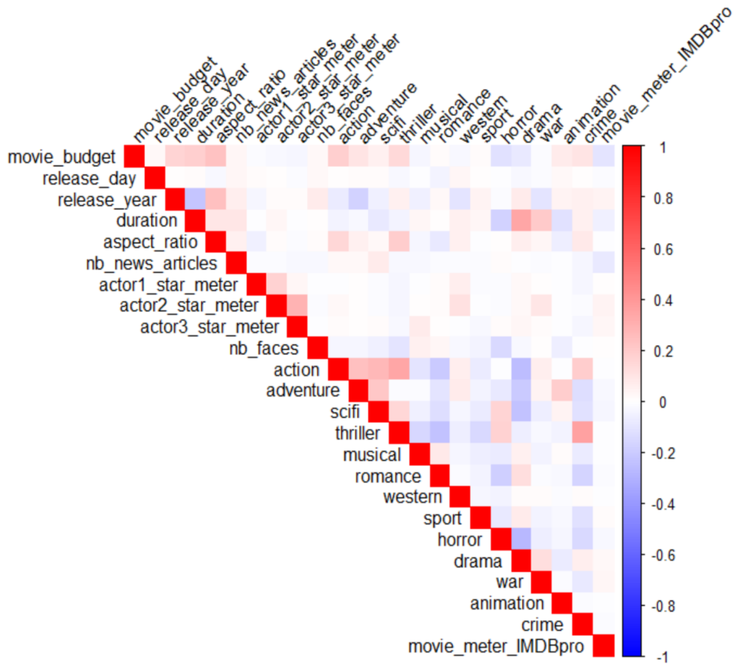


Figure 2: Actor 1, 2, and 3 Outliers

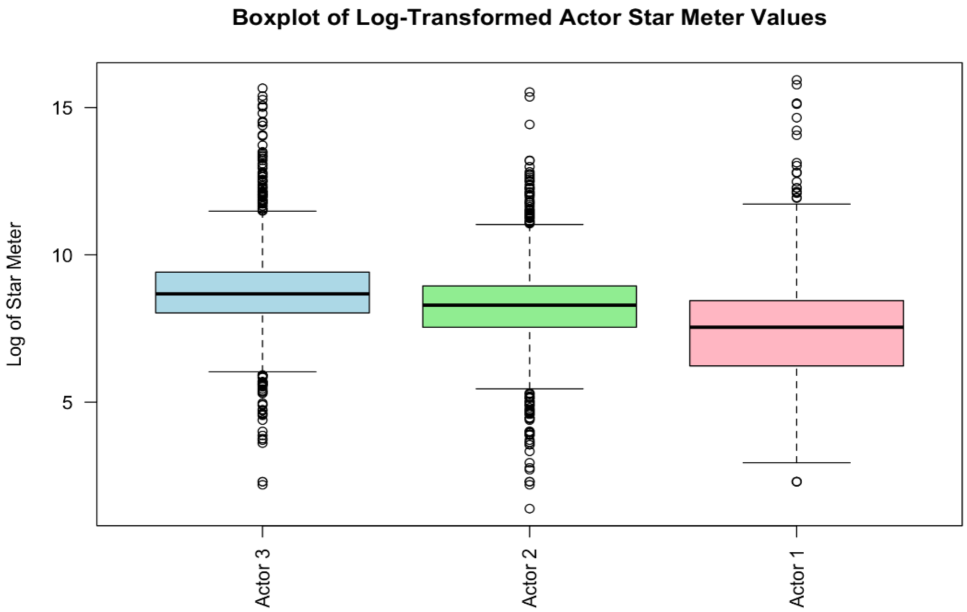


Figure 3: Number of News Articles Outliers

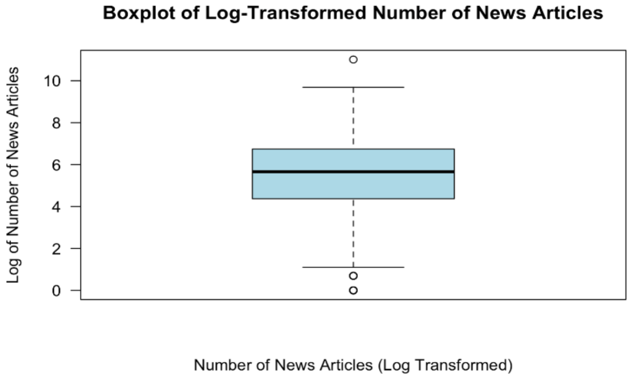


Figure 4: Release Year Outliers

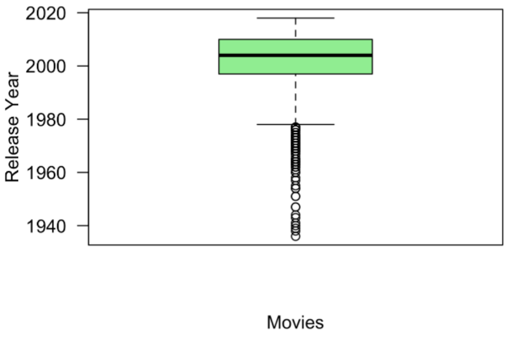


Figure 5: Duration Outliers

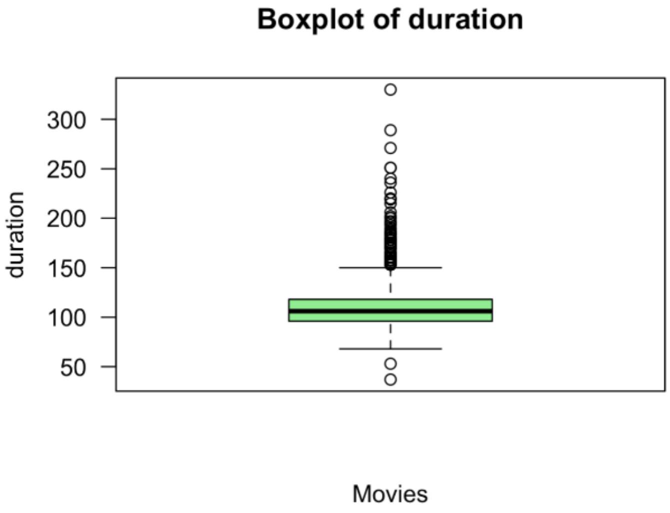


Figure 6: Movie Meter IMDb Pro Outliers

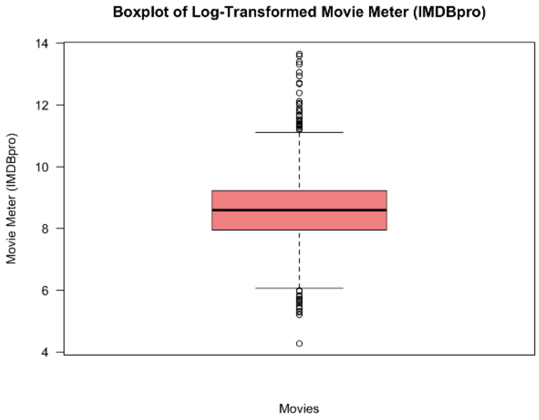
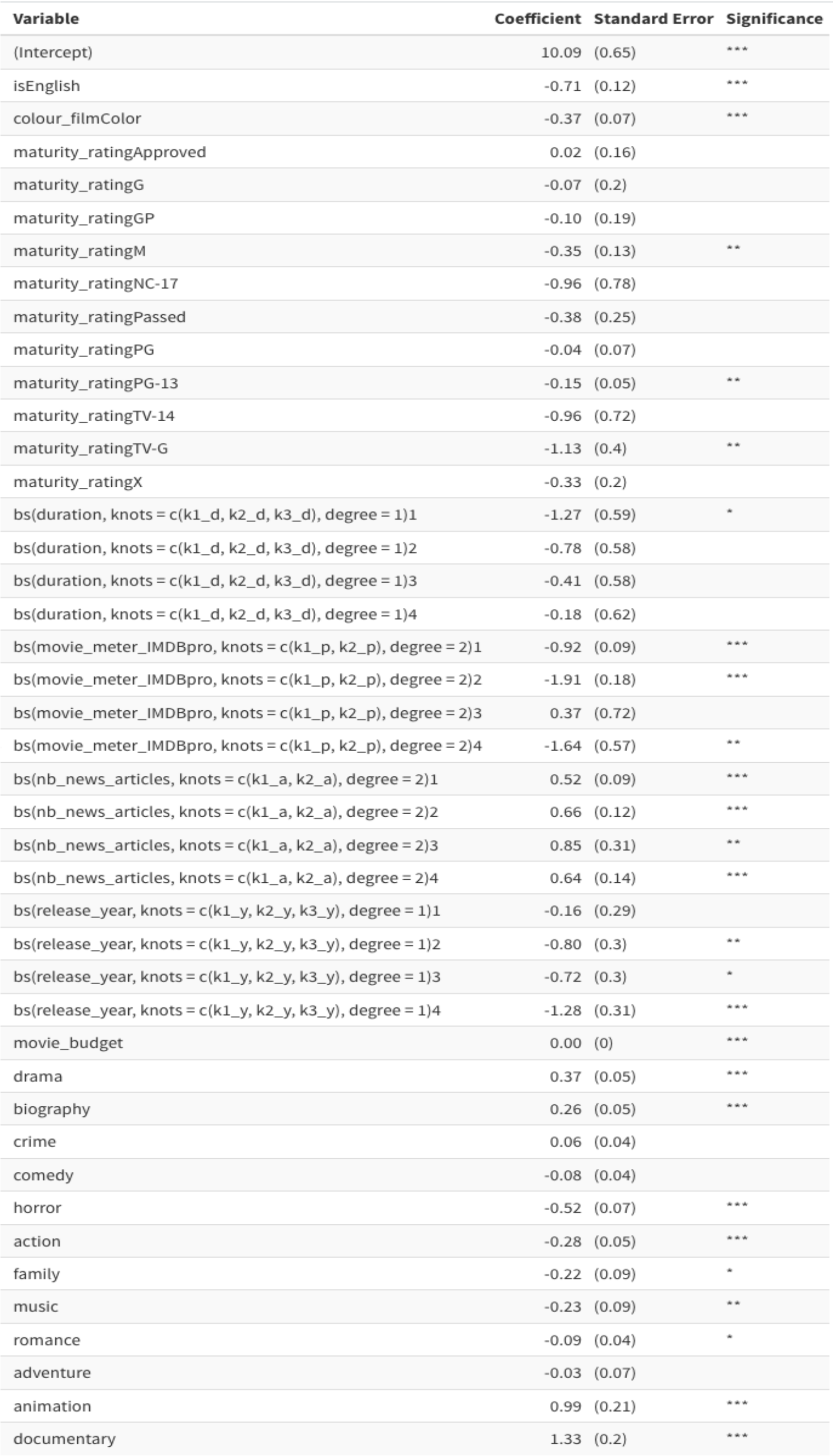


Figure 6: Final IMDB Model Predictor Summary



Variable	Coefficient	Standard Error	Significance
(Intercept)	10.09	(0.65)	***
isEnglish	-0.71	(0.12)	***
colour_filmColor	-0.37	(0.07)	***
maturity_ratingApproved	0.02	(0.16)	
maturity_ratingG	-0.07	(0.2)	
maturity_ratingGP	-0.10	(0.19)	
maturity_ratingM	-0.35	(0.13)	**
maturity_ratingNC-17	-0.96	(0.78)	
maturity_ratingPassed	-0.38	(0.25)	
maturity_ratingPG	-0.04	(0.07)	
maturity_ratingPG-13	-0.15	(0.05)	**
maturity_ratingTV-14	-0.96	(0.72)	
maturity_ratingTV-G	-1.13	(0.4)	**
maturity_ratingX	-0.33	(0.2)	
bs(duration, knots = c(k1_d, k2_d, k3_d), degree = 1)1	-1.27	(0.59)	*
bs(duration, knots = c(k1_d, k2_d, k3_d), degree = 1)2	-0.78	(0.58)	
bs(duration, knots = c(k1_d, k2_d, k3_d), degree = 1)3	-0.41	(0.58)	
bs(duration, knots = c(k1_d, k2_d, k3_d), degree = 1)4	-0.18	(0.62)	
bs(movie_meter_IMDBpro, knots = c(k1_p, k2_p), degree = 2)1	-0.92	(0.09)	***
bs(movie_meter_IMDBpro, knots = c(k1_p, k2_p), degree = 2)2	-1.91	(0.18)	***
bs(movie_meter_IMDBpro, knots = c(k1_p, k2_p), degree = 2)3	0.37	(0.72)	
bs(movie_meter_IMDBpro, knots = c(k1_p, k2_p), degree = 2)4	-1.64	(0.57)	**
bs(nb_news_articles, knots = c(k1_a, k2_a), degree = 2)1	0.52	(0.09)	***
bs(nb_news_articles, knots = c(k1_a, k2_a), degree = 2)2	0.66	(0.12)	***
bs(nb_news_articles, knots = c(k1_a, k2_a), degree = 2)3	0.85	(0.31)	**
bs(nb_news_articles, knots = c(k1_a, k2_a), degree = 2)4	0.64	(0.14)	***
bs(release_year, knots = c(k1_y, k2_y, k3_y), degree = 1)1	-0.16	(0.29)	
bs(release_year, knots = c(k1_y, k2_y, k3_y), degree = 1)2	-0.80	(0.3)	**
bs(release_year, knots = c(k1_y, k2_y, k3_y), degree = 1)3	-0.72	(0.3)	*
bs(release_year, knots = c(k1_y, k2_y, k3_y), degree = 1)4	-1.28	(0.31)	***
movie_budget	0.00	(0)	***
drama	0.37	(0.05)	***
biography	0.26	(0.05)	***
crime	0.06	(0.04)	
comedy	-0.08	(0.04)	
horror	-0.52	(0.07)	***
action	-0.28	(0.05)	***
family	-0.22	(0.09)	*
music	-0.23	(0.09)	**
romance	-0.09	(0.04)	*
adventure	-0.03	(0.07)	
animation	0.99	(0.21)	***
documentary	1.33	(0.2)	***

Figure 8 indicates the HC1-adjusted final predictor coefficients for the model with their respective significance in the current model.

Figure 7: Movie Budget and IMDB Score Relationship

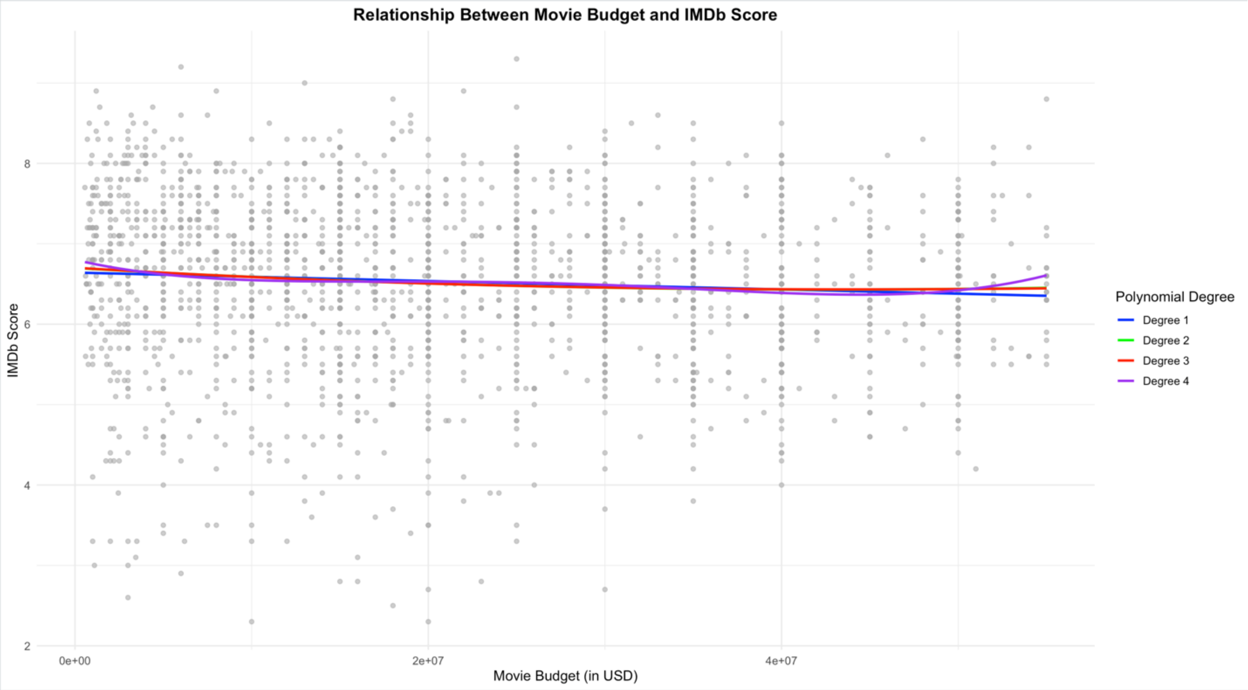


Figure 8: Movie Duration and IMDB Score Relationship

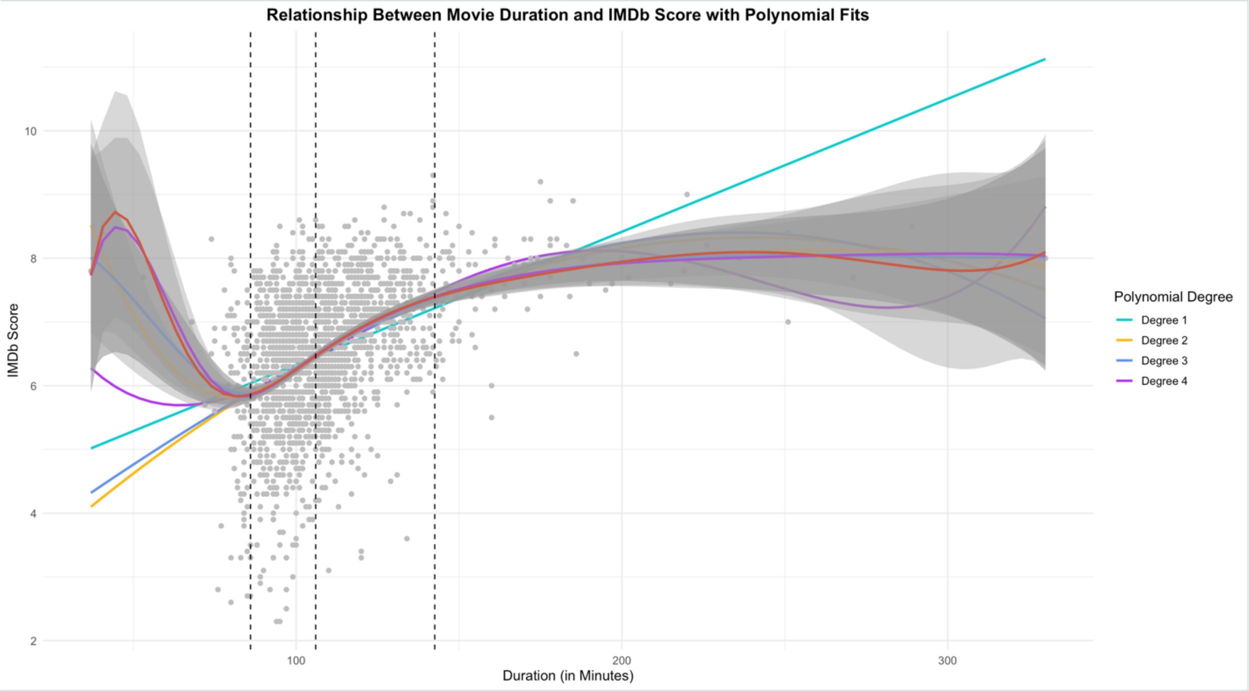


Figure 9: Relationship Between Release Year and IMDB Score

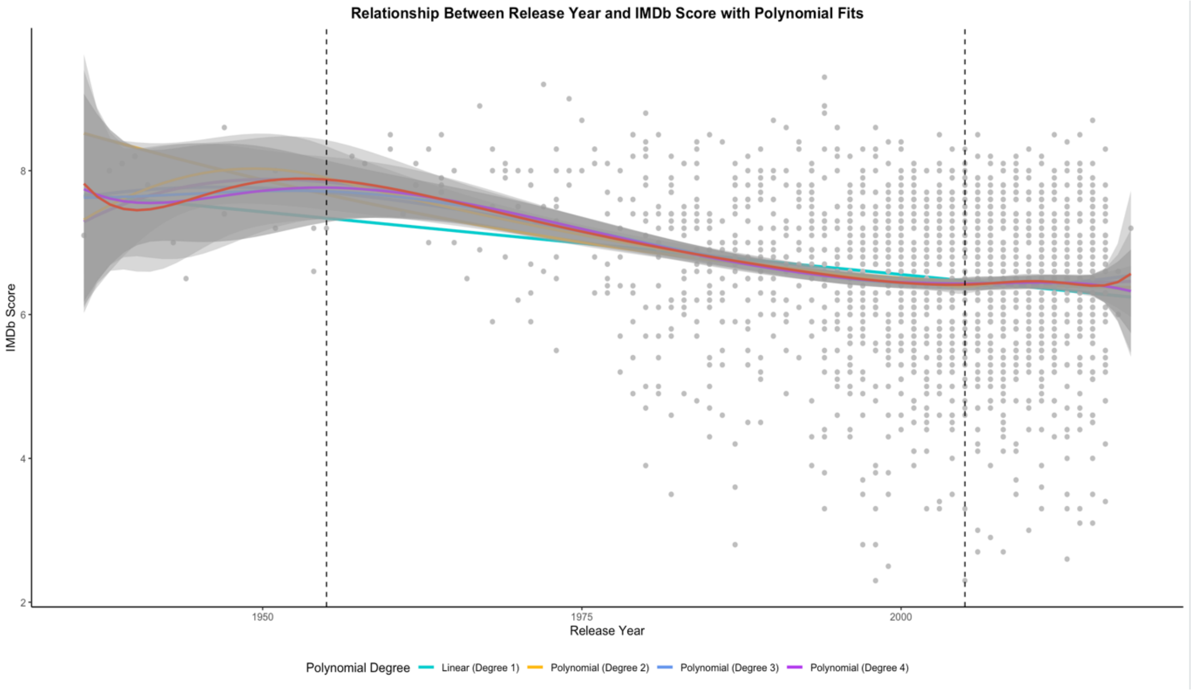


Figure 10: Relationships Between Number of News Articles and IMDB Score with Polynomial Fits

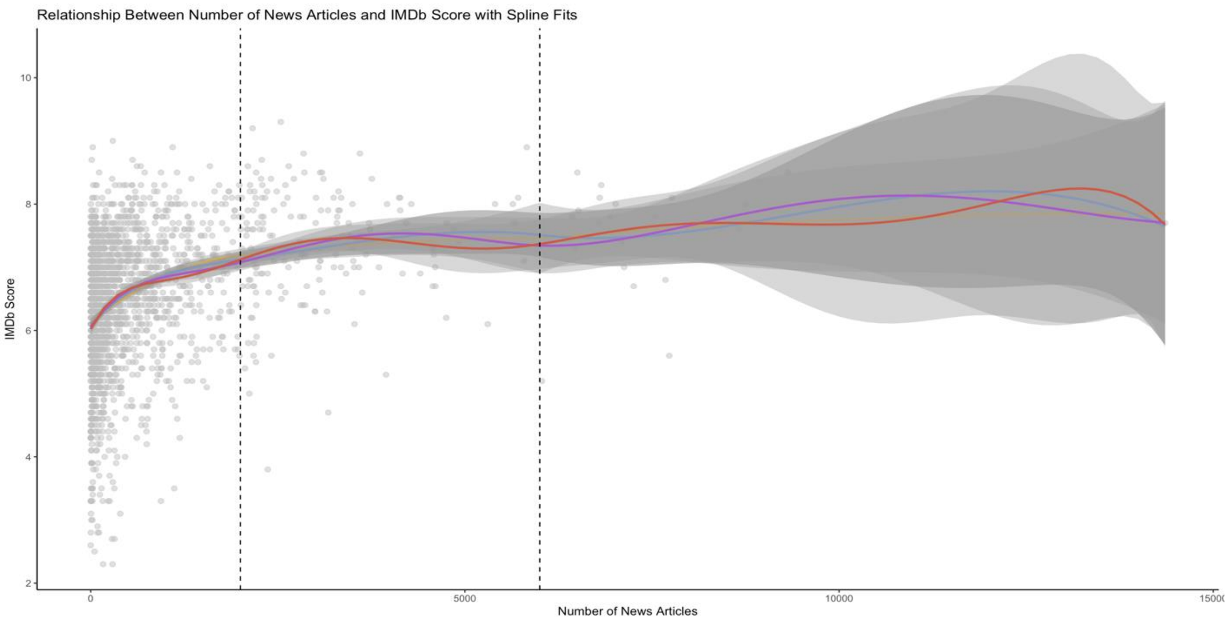


Figure 11: Relationship Between Number of Faces in the Main Movie Poster and IMDB Score

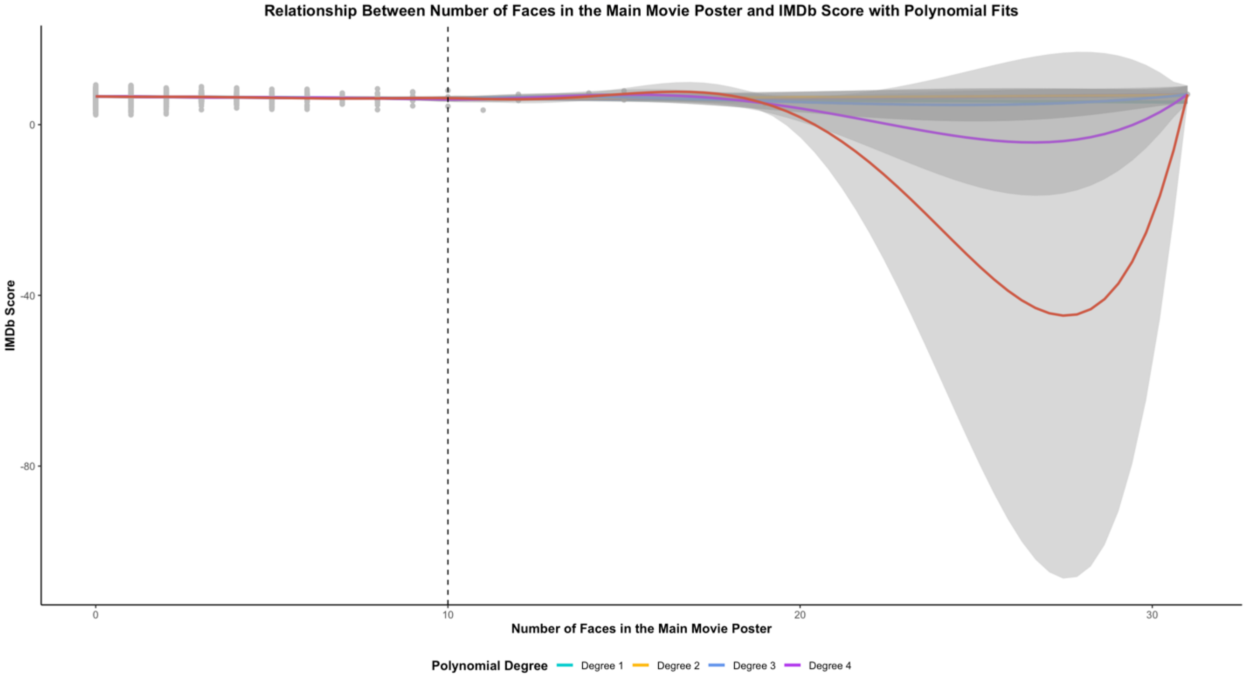


Figure 12: Relationship Between Actor 1 Star Meter vs IMDB Score with Polynomial Fits

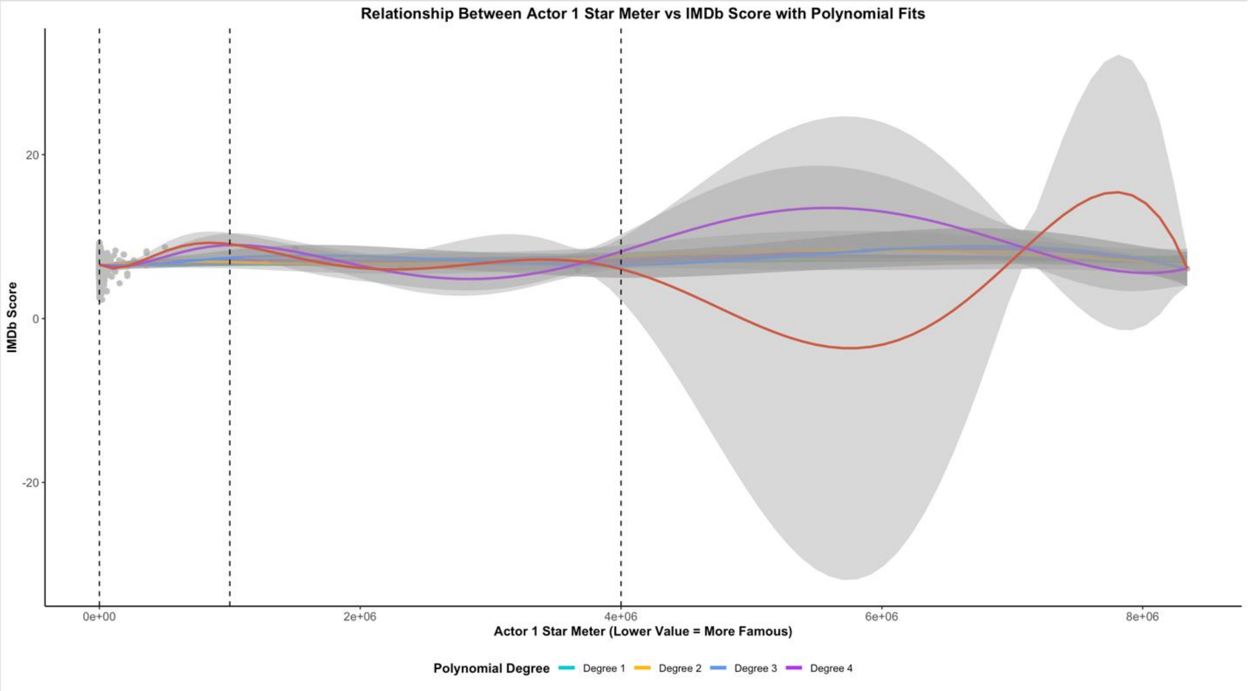


Figure 13: Relationship Between IMDBPro Movie Meter and IMDB Score

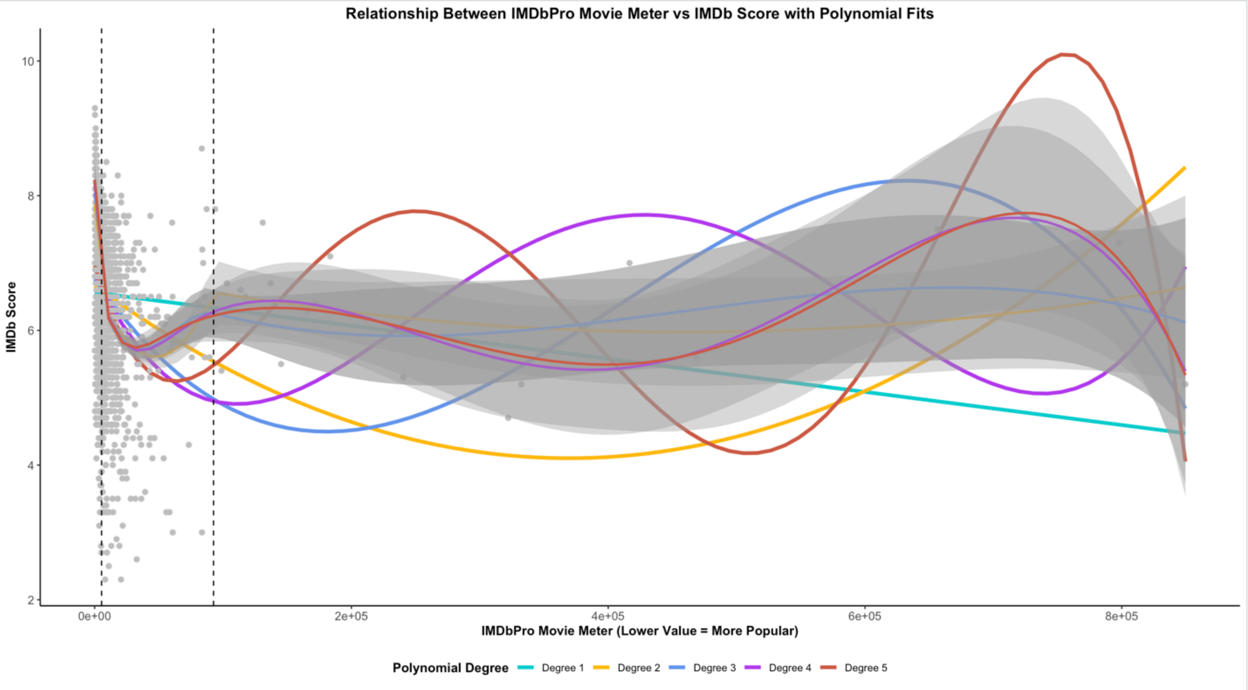


Figure 14: Variable Predictiveness Significance Test to Select for Regression Modeling

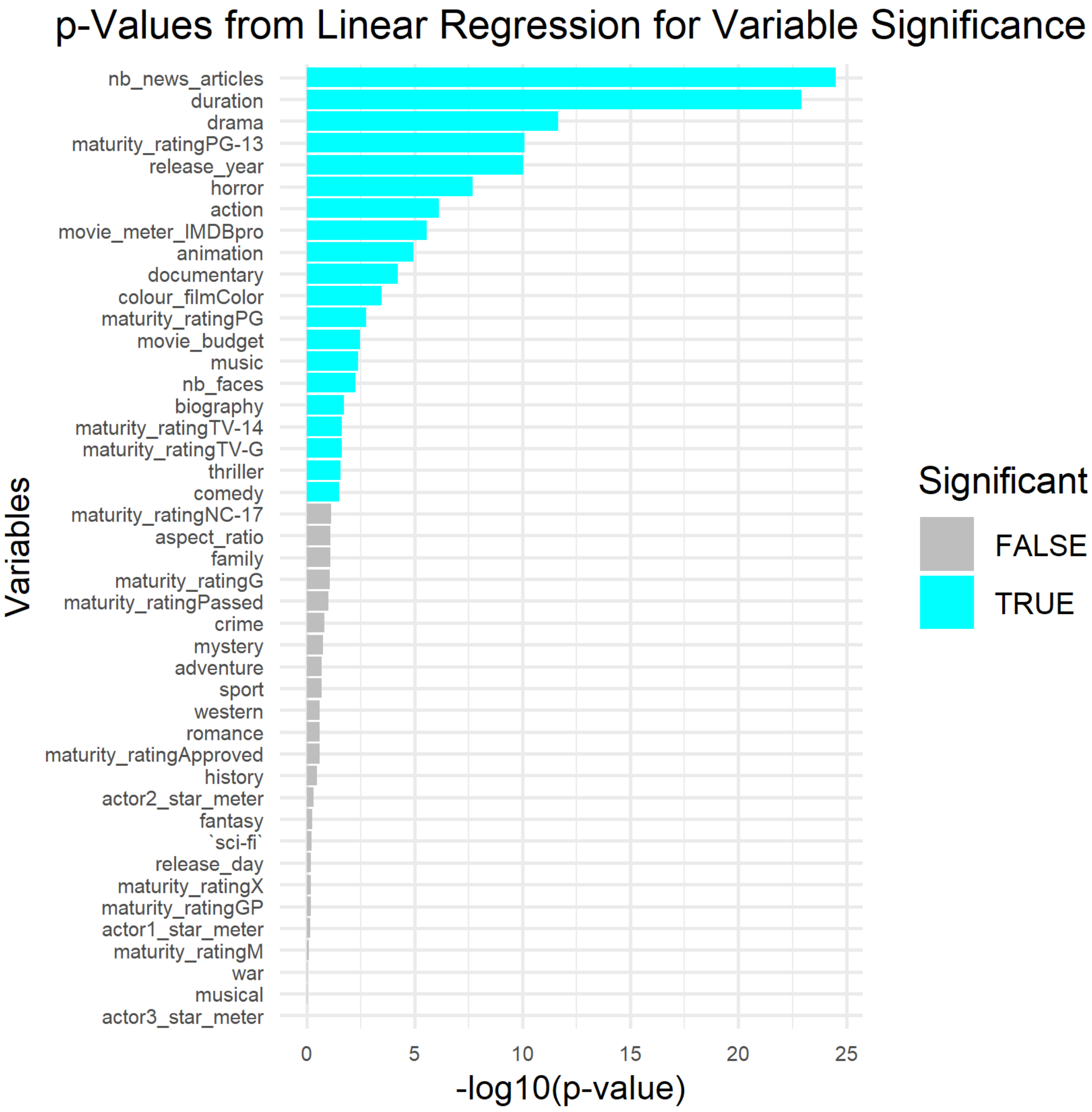
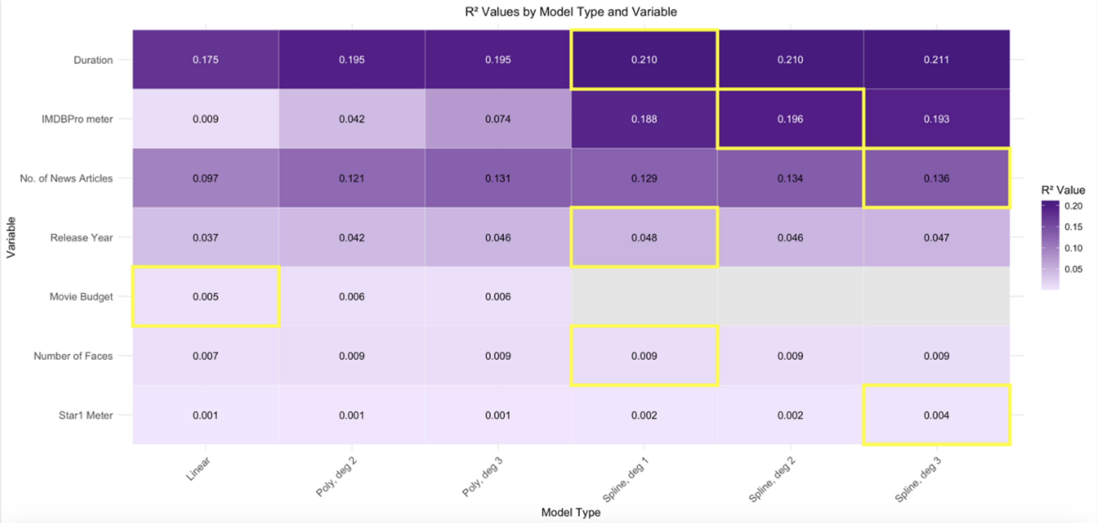

p-Values from Linear Regression for Variable Significance
Variables
nb_news_articles
duration
drama
maturity_ratingPG-13
release_year
horror
action
movie_meter_IMDBpro
animation
documentary
colour_filmColor
maturity_ratingPG
movie_budget
music
nb_faces
biography
maturity_ratingTV-14
maturity_ratingTV-G
thriller
comedy
maturity_ratingNC-17
aspect_ratio
family
maturity_ratingG
maturity_ratingPassed
crime
mystery
adventure
sport
western
romance
maturity_ratingApproved
history
actor2_star_meter
fantasy
`sci-fi`
release_day
maturity_ratingX
maturity_ratingGP
actor1_star_meter
maturity_ratingM
war
musical
actor3_star_meter
0
5
10
15
20
25
-log10(p-value)
Significant
FALSE
TRUE

Figure 16: Comparison of R^2 Values Across Different Models for Key Variables



Next Degree ANOVA p-Value*
0.2443
4.204e-05
0.0248
0.0270
0.2005
0.9954
0.0021

Figure 16 shows the comparison of different kinds of relationships with the IMDB score and how much each of the models explain variance in the dataset.