IMDB Midterm Write Up

Mahrukh Shamas, Aasna Shah, Hao Duong, Yichen Yu, Xinran Yu

November 02, 2023

MGSC 661-075

Group: Rick Rollers

Table of Contents

Part I
Introduction

The purpose of this report is to document the development of a statistical model to forecast the IMDb ratings for twelve upcoming films scheduled for release in November 2023. Using a dataset of approximately two thousand films, we aim to identify significant predictors of IMDb ratings and construct an accurate predictive model using regression techniques. The model will be trained and tuned through cross-validation to maximize out-of-sample predictive accuracy while avoiding overfitting to the training data. Feature engineering and selection methods will be systematically explored to construct an optimal set

of predictors. The final model will be used to generate predicted IMDb ratings for the twelve films of interest. These predictions will then be evaluated against the actual post-release IMDb ratings to assess the predictive accuracy of the model. Our detailed statistical analysis and modeling aims to construct an accurate, robust, and interpretable model for forecasting IMDb ratings of new films prior to release. The systematic testing of variable transformations and model comparisons yielded the optimal balance of flexibility, interpretability, and predictive performance. The methods, results, and insights garnered throughout the model development process will be documented in this report along with our predictions for IMDB score for the twelve upcoming films.

Part II

Data Description

The dataset contains information about two thousand movies, including variables related to budget, release date, genres, and cast. The dependent variable we will focus on predicting is the IMDb score (imdb_score). This score reflects the overall user ratings on IMDb.com, ranging from 1 to 10, with higher values corresponding to more popular and well-liked movies. The imdb_score is skewed left and despite this, the distribution appears unimodal (Figures 1, 2 and 13). The skew reflects the fact that truly exceptional or terrible movies are outliers and rare, while moderately good movies are common. Thorough exploration of the relevant independent variables is a crucial step in identifying the factors that have the strongest predictive power for the imdb_score. The upcoming paragraphs will dive deeper into analyzing each key variable to determine its correlation and predictive relationship with the dependent variable.

The movie budget variable represents the film's production budget in US dollars. The positive skew in the budget distribution indicates that the majority of movies have lower budgets, while a minority require significantly higher investments (Figures 6 and 13). These high-budget outliers typically represent blockbuster productions or movies with exceptional production quality. There exists a moderate correlation between higher budgets and IMDb scores, suggesting that increased production resources, including visual effects (VFX), tend to result in higher user ratings (Figure 22). However, this correlation is

not exceptionally strong, indicating that a high budget alone does not guarantee high ratings if other quality factors are lacking. Nevertheless, in general, the right-skewed budget distribution and its positive correlation with ratings provide evidence that, on average, high-budget movies tend to receive better user ratings.

The year the movie was released is represented by the release year variable. This variable is highly left-skewed because the dataset includes recent and older movies, but far fewer movies were made decades ago due to limited filmmaking technology/infrastructure (Figures 12 and 13). There is also greater audience demand today supporting more releases, however, classic and culturally impactful movies from earlier decades, also draw consistent user interest and ratings over time. There is a slight negative correlation between the release year and imdb_score, suggesting newer movies tend to have slightly lower scores, on average (Figure 22). This may be because the surge in modern production includes many formulaic sequels and remakes, rather than groundbreaking original films. However, the correlation is small in magnitude. In summary, the skew arises from past limitations, but release year also reflects the difference between enduring classics and recent releases.

The duration variable represents the runtime of the movie in minutes. The duration variable itself has a right-skewed distribution, with most movies having moderate runtimes, but a small number of movies having exceptionally long durations (Figures 4 and 13). These unusually long movies likely represent historical epics or movies with complex storylines that require greater runtimes and are outliers. The negative correlation between duration and imdb_score suggests extremely long movies tend to have slightly lower user ratings on average, perhaps because viewers lose interest, or the lengthy pace detracts from enjoyment (Figure 22). However, this relationship is fairly weak, indicating runtime alone does not dictate user ratings. Rather, the right-skewed duration variable and its correlation shows that while most good movies have standard runtimes, extremely long movies tend to score slightly worse despite having more time for character and plot development.

The dataset contains variables for the top one to three billed actors in each movie. Actor star meters 1, 2, and 3 are variables which essentially quantify the fame level of the actor. They are extremely right-skewed since most actors remain obscure, especially due to the smaller reach of indie productions, but very few achieve mega-stardom (Figures 9,10,11 and 13). Famous actors also lend a visible stamp of quality reassuring audiences. Long-term career reputation

also matters more than short-term fame. On average, acclaimed actors boost imdb_scores up to a point, as their talent and fame attract audiences initially. However, star power has diminishing returns - once known, more fame does not directly lead to higher ratings if other quality factors like script lag or direction. Movies rely on more than just names to achieve high critical and audience acclaim which is prevalent in its weak correlation with the IMDb score(Figure 22). The same can also be stated for the director variable; it's skewed right and is weakly correlated with the IMDb score due to the reasons stated above.

The number of faces in the poster variable represents the number of cast member faces pictured on the movie's poster, serving as a quantifiable measure of ensemble size. This variable exhibits a highly right-skewed distribution because most movies have just one to five core actors with a few faces on the poster (Figures 8 and 13). However, large ensemble films in major franchises require expansive casts of more than ten lead and supporting roles to carry the story, leading to an outlier high number of faces on the posters. These major ensemble movies are rare compared to more compact indie films. The large cast is a key part of the appeal for heavily marketed ensemble films, so highlighting the numerous famous faces on posters helps attract viewers. The variable of number of faces on poster correlates negatively with IMDb score which suggests movies with more individual faces on the poster may prioritize marketing/recognition over substance of the movie (Figure 22).

The 2023 ranking of movie made by IMDbPro variable quantifies page views and clicks for each movie on the IMDbPro site. It exhibits an extremely right-skewed distribution because most obscure indie films attract negligible traffic, while a few unprecedented blockbuster hits dominate entertainment news and online discussion, driving outlier high traffic to their IMDbPro pages (Figures 5 and 13). These are rare cultural phenomena that generate intense viewer demand, translating to enormous, disproportionate traffic spikes on IMDbPro. This variable has a slight negative correlation, though weak, implies counterintuitively that higher popularity associates with marginally lower IMDb scores (Figure 22). The outsized hype for a few mega blockbusters may inflate their IMDbPro traffic beyond what their more moderate scores warrant. For most films, the traffic aligns more directly with audience enjoyment and ratings.

The variable of the number of articles in the news of the main country about the film represents the number of online news articles about each movie around its release. It has an extremely right-skewed distribution because most indie films get negligible coverage, moderate wide-releases get hundreds, but

unprecedented blockbuster hits receive an outlier surge of tens of thousands of articles (Figures 7 and 13). This extreme skew occurs because the media heavily capitalizes on the intense audience demand for content about the few breakout hits. It is weakly positive correlated which implies that more articles are associated with higher IMDB scores (Figure 22). The media buzz and hype surrounding a major release may influence initial audience reception and ratings. But negligible press coverage does not necessarily doom a film if it connects with audiences through other means.

The maturity rating variable indicates the MPAA rating assigned to a movie based on its content. In order to assess the interaction between the various ratings and the target variable, as well as the other predictors, the maturity ratings variable was dummified. Through the creation of a correlation matrix, we discovered that the ratings correlated with genres like drama and thriller and also was highly correlated with imdb_score. Next, we assessed the collinearity, and discovered that the maturity rating variables PG, PG-13, and R have extremely high VIFs over 10, signaling severe multicollinearity issues with other predictors (Figure 22). Thus, we did not find it helpful to include in further analyses as the severe multicollinearity issue would distort any model outputs.

The genre variable represents the main genre categorization for each movie. The prevalence of dramas and comedies shows these are the most common types of movies produced. The wide range but average scores for dramas indicates this genre includes both exceptional and poorly rated films. The higher-than-average ratings for animated, biographical, and historical movies suggest these genres tend to be crowd-pleasers. However, certain genres suffer biases – horror films consistently receive lower imdb_scores on average, portraying the contrast between critics' views as either sensational and unsophisticated or refined and complex dramas. Meanwhile, dramas earn elevated ratings and acclaim due to being seen as intellectual and culturally important by reviewers. These correlations demonstrate how biased value judgements shape the rating differences across genres. Although certain genres like Animation tend to be higher-rated, overall the data shows a weak correlation between genre and IMDb score (Figure 22).

Part III
Model Selection

The modeling approach of our team was to use linear regression to predict the continuous IMDb score variable based on key movie attributes. Initial exploratory analysis informed the choice of linear, polynomial, and spline functional forms for modelling relationships while avoiding overfitting. In order to use categorical predictors effecively, we introduced a new predictor, distributor rate, into our model. This is the average IMDB score of the movies each distributor produced and highly correlated to our dependent variable. Although we tried to apply the same preprocessing to director, cinematographers, and production companies, these new predictors can not be applied to the test dataset due to the lack of data.In assessing different polynomial functions, to understand which one would provide a higher r-squared value, it was determined that for several variables a higher order model would be favourable (Figure 33). For example, distributor_rate, release_year, duration, release_month, nb_news_articles and movie_meter_IMDBpro were all found to peak in their r-squared values at a quadratic model. However, movie_budget and nb_faces were optimal at linear models. The process to test the performance of different polynomial functions on the predictors was repeated, however this time out of sample performance was tested by separating the data into training and test sets. We found that duration had the highest R-squared on both training (0.22) and test (0.23) data, indicating it is by far the most predictive individual variable. Most other variables like budget, actor ratings, nb_faces had very low R-squared on both train and test data, indicating possible minimal predictive value.

Furthermore, the previous variables identified as performing better at higher order polynomials continued to perform in a similar fashion for both training and test sets. However, polynomial fits beyond the quadratic degree tended to overfit training data while degrading test performance, thus it was found that the quadratic model was sufficient. Lastly, we carried out spline regression on a few notable predictors to assess the non-linearity of these predictors and if a spline model could better capture the relationship between these variables and imdb_score (Figure 34). It was found that for duration the df=3 spline model with 4 knots is optimal, with a r-squared increasing as df increased. This shows duration has a strong nonlinear relationship with IMDb score that a spline

model is better able to capture versus a linear model. This was also the case for the nb_newsarticle variable and release_year. Overall, through the spline modelling it was determined that the variables showing significant R-squared improvements with splines (duration, news articles, release details, actor ratings, movie meter) clearly have predictive relationships worth including in the model. Variables with minimal to no spline improvements (budget, number of faces) provide weaker predictive signal but may still be useful to retain.

Heteroskedasticity, or non-constant variance, was also assessed and detected using an NCV test. Several predictors including movie budget, duration, number of articles in the news about the film, the 2022 ranking of actors/actresses made by IMDb, the 2023 ranking of movie made by IMDbPro, scifi, thriller, romance, drama, war, crime, maturity Approved, maturity PG and distributor rate were identified as potentially exhibiting heteroskedasticity. To address heteroskedasticity and ensure robust standard errors, the vcovHC function was employed to compute a heteroskedasticity-robust covariance matrix. This resulted in larger robust standard errors for predictors that showed non-constant variance.

Predictors were selected based on stronger individual predictive power assessed through R-squared metrics. Highly collinear variables like director were removed using VIF test analysis to prevent multicollinearity. Categorical predictors were incorporated efficiently as dummy variables. Outliers were also identified by utilizing the outlier test and model residual analysis followed by subsequential removal from the training dataset to improve model performance and better interpretability. The model complexity was carefully tuned to the constraints of the dataset size using 10-fold cross-validation.

Part IV
Final Results

The final model developed to predict IMDb movie scores performed reasonably well, achieving an R-squared of 0.4091 and a mean squared error (MSE) of 0.7210 with the K-fold test on the original dataset IMDB_data. The model made predictions ranging from 5.55 to 7.66 for the twelve films we need to predict in the real world (Figure 31). The plot of the predictions versus the actual IMDb scores (Figure 30) showed that the model was able to capture the

general trend quite well, with some over and underestimation at the very low and very high ends of the ratings distribution. The final predictive model was developed through an iterative process involving feature selection, polynomial modeling, interaction effects, dummy variables, model comparison, and spline modeling.

Our model predicted the following IMDb scores for the upcoming twelve movies (Figure 31): The documentary Pencils vs Pixels received a predicted score of 5.5516. The crime movie The Dirty South earned a predicted rating of 6.6119. The drama The Holdovers achieved the highest predicted score of 7.6652. The dystopian prequel The Hunger Games: The Ballad of Songbirds and Snakes followed closely behind with a score of 7.3940. The upcoming superhero film The Marvels forecasted rating is 6.2391. The comedy Thanksgiving scored 5.9610 while the adventure movie Wish garnered 5.9952. The sports comedy Next Goal Wins earned 6.5625 and the animated Leo received 6.1052. For the sequel Trolls Band Together the prediction was 6.0249. The historical drama Napoleon earned 6.9669. Finally, the comedy Dream Scenario achieved an IMDb score prediction of 7.2373.

Key continuous predictors like distributor rating and duration were identified through R-squared scores, p-values, and domain expertise. The most significant predictors in the final model were distributor rating, quadratic duration, drama genre, and an interaction between quadratic release year and distributor rating. Distributor rating had a significant positive correlation with IMDb score, indicating that films from better rated distributors tend to achieve higher audience scores on IMDb. This aligns with the intuition that large, well-respected studios with a strong track record tend to produce higher quality films. The quadratic duration term captured an inverted-U shaped relationship between length and score, suggesting that very short and very long films are penalized while the optimal length is in the middle.

The drama genre indicator predicted higher scores compared to non-dramas, controlling for other variables. This may reflect viewer preference for serious dramas versus action or comedy films. The interaction term showed that more release year positively amplified the effect of distributor rating on score. More media buzz appears to heighten the reputation bump that big studios get. Finally, PG-13 and R-rated films achieved higher scores than G and PG, controlling for other factors. This likely indicates that audiences find films tailored for adults with more violence, language, etc. to be more engaging than all-ages films.

Polynomial terms were tested to capture nonlinear relationships, with a quadratic model for duration found to be significant. Interaction effects between predictors were evaluated, with the interaction between release year and distributor rating found to be significant. Dummy variables like genre improved predictive ability and were included. Multiple models using different combinations of these predictors were constructed and compared based on R-squared and MSE to select the best final model. Splines were tested as an alternative to quadratic relationships but did not improve the polynomial model for variables such as duration. This systematic and iterative modeling process resulted in a final model that balanced predictive power with interpretability and aligned with theory about influences on movie ratings. The variables selected and relationships captured provide insights into the key factors that may influence IMDb scores based on this dataset.

The final model achieved an R-squared value of 40.91%, indicating that it explained approximately 41% of the variance in IMDb scores.However, several factors contributed to prediction errors. To address this issue, collecting more comprehensive training data is crucial. In contrast, the model that includes the three additional predictors performed exceptionally well on the original dataset, incorporating these three additional predictors, boasting an R-squared value of 0.8323 and a minimal Mean Squared Error (MSE) of 0.2047 (Figure 32). However, these predictors couldn't be applied to the test movies. With an expanded dataset, the model could better capture and leverage these effects for more accurate predictions.

In our pursuit to ensure the model's robustness against overfitting, we subjected it to use K-fold test for the out-of-sample performance evaluation. Given the model's design—encompassing polynomial transformations, interaction terms, and dummy variables. We used the validation set test for performance evaluation, while recognizing the overfitting risk. The model produced an in-sample mean squared error (MSE) of 0.7244, closely aligned with an out-of-sample MSE of 0.7108. This tight convergence between in-sample and out-of-sample metrics is encouraging, suggesting our model generalizes well to new data. However, the proximity of these values also indicates that while our model has achieved a commendable level of stability, there's potential to enhance its predictive power. As we move forward, refining the model by honing predictor selection and possibly integrating richer training data sets will be instrumental in elevating its predictive accuracy in real-world applications.

Part V
Appendix

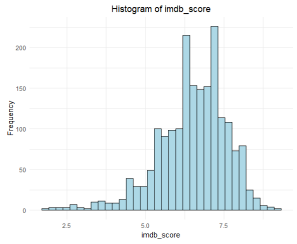

Figure 1: Distribution for IMDB Score

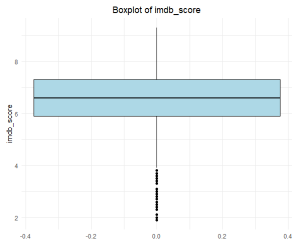

Figure 2: Boxplot for IMDB Score

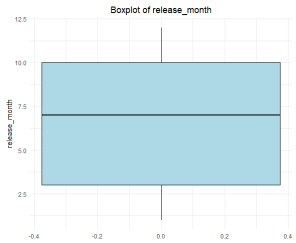

Figure 3: Boxplot for Release Month

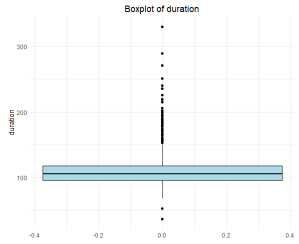

Figure 4: Boxplot for Duration

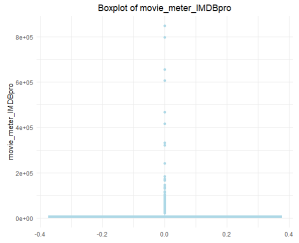

Figure 5: Boxplot of Movie Meter IMDB Pro

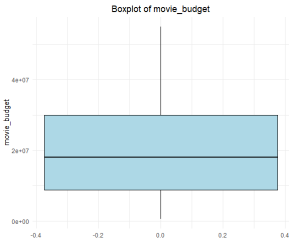

Figure 6: Boxplot of Movie Budget

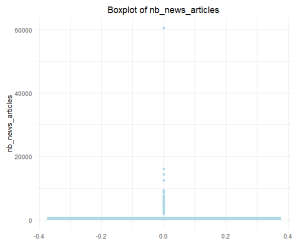

Figure 7: Boxplot of Number of News Articles

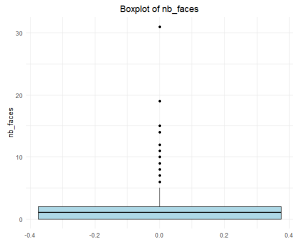

Figure 8: Boxplot of Number of Faces

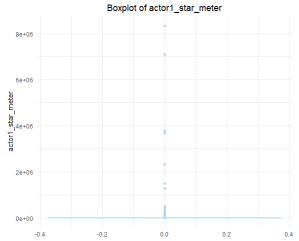

Figure 9: Boxplot of Actor 1 Star Meter

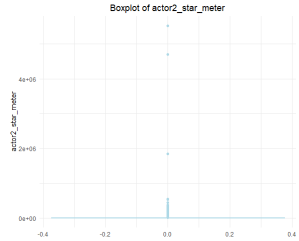

Figure 10: Boxplot of Actor 2 Star Meter

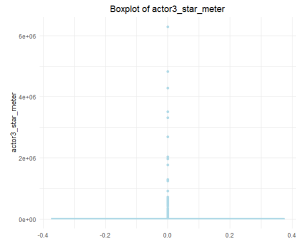

Figure 11: Boxplot of Actor 3 Star Meter

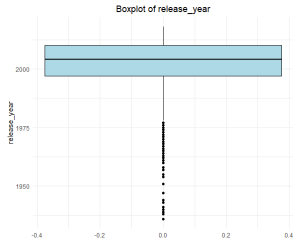

Figure 12: Boxplot of Release Year

```
[1] "imdb_score Skewness:  -0.864495722581472"
[1] "movie_budget Skewness:  0.537946396687438"
[1] "release_year Skewness:  -1.62679354265645"
[1] "release_month Skewness:  -0.0877910139426639"
[1] "duration Skewness:  2.67708199575918"
[1] "nb_news_articles Skewness:  18.6088703577527"
[1] "actor1_star_meter Skewness:  23.2678536480579"
[1] "actor2_star_meter Skewness:  27.5774777293866"
[1] "actor3_star_meter Skewness:  16.2897373811665"
[1] "nb_faces Skewness:  3.73447579184557"
[1] "movie_meter_IMDBpro Skewness:  14.532832547376"
```

Figure 13: Skewness Report

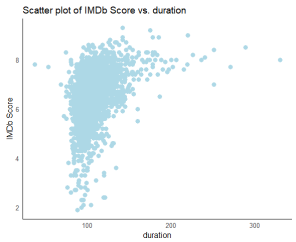

Figure 14: Scatterplot of Y against Duration

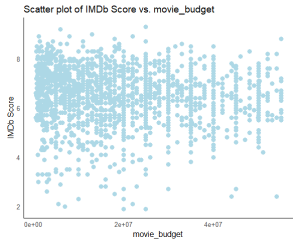

Figure 15: Scatterplot of Y against Movie Budget

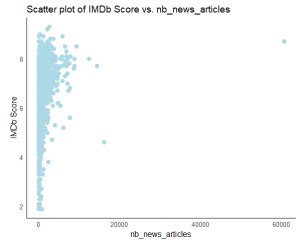

Figure 16: Scatterplot of Y against Number of News Articles

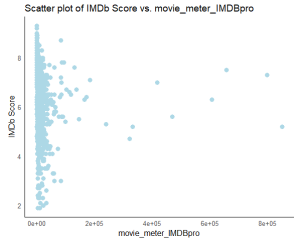

Figure 17: Scatterplot of Y against Movie Meter IMDBPro

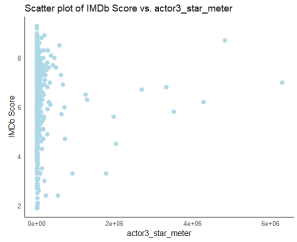

Figure 18: Scatter plot of Y against Actor 3 Star Meter

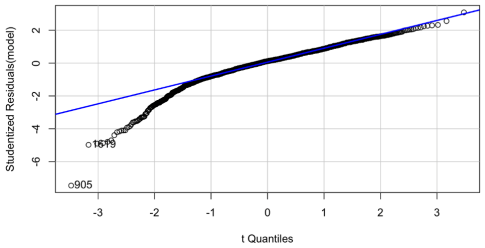

Figure 19: Studentized Residuals Plot

```
> print(outlier_test_result)
      rstudent unadjusted p-value Bonferroni p
905  -7.457440         1.3289e-13   2.5648e-10
1619 -4.977395         7.0208e-07   1.3550e-03
1854 -4.905251         1.0118e-06   1.9527e-03
890  -4.874460         1.1808e-06   2.2790e-03
1354 -4.793108         1.7685e-06   3.4132e-03
39   -4.705362         2.7153e-06   5.2406e-03
1149 -4.390458         1.1927e-05   2.3019e-02
```

Figure 20: Outlier Test Results Output

Table: Variance Inflation Factors

Variable	VIF
movie_budget	1.323122
release_year	1.665408
release_month	1.026941
duration	1.451941
nb_news_articles	1.046875
actor1_star_meter	1.042691
actor2_star_meter	1.165499
actor3_star_meter	1.117252
nb_faces	1.067911
movie_meter_IMDBpro	1.030408
action	1.431642
adventure	1.350441
scifi	1.233844
thriller	1.505182
musical	1.102514
romance	1.200413
western	1.073701
sport	1.119211
horror	1.356730
drama	1.454773
war	1.127895
animation	1.165370
crime	1.410046
maturityApproved	2.407479
maturityG	3.111923
maturityGP	1.136280
maturityM	1.143179
maturityPassed	1.299983
maturityPG	14.321889
`maturityPG-13`	25.397043
maturityR	29.460919

Figure 21: VIF Table for each variable

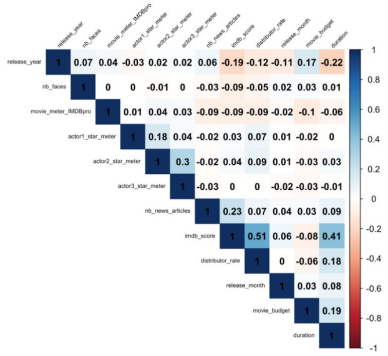

Figure 22: Correlation Matrix between variables

```
> print(importance_dataset)
                      var          rsq        pvalue degree
24    distributor_rate 2.567400e-01 2.155841e-126      1
4            duration 1.686284e-01  2.124863e-79      2
20              drama 1.143819e-01  7.482679e-53      1
5     nb_news_articles 5.082841e-02  1.154764e-23      5
2        release_year 3.795539e-02  5.812340e-18      3
19              horror 2.757971e-02  2.102335e-13      1
11              action 2.529932e-02  2.096054e-12      1
21                 war 1.178186e-02  1.756616e-06      1
13               scifi 8.798989e-03  3.669494e-05      1
10 movie_meter_IMDBpro 8.051840e-03  7.897205e-05      5
9             nb_faces 7.992422e-03  8.394547e-05      1
14            thriller 6.405711e-03  4.323968e-04      1
1        movie_budget 6.188867e-03  5.417691e-04      1
12           adventure 4.474328e-03  3.282087e-03      1
17             western 4.294571e-03  3.974098e-03      1
3       release_month 3.790037e-03  6.822270e-03      2
23              crime 3.775400e-03  6.930645e-03      1
18              sport 3.025159e-03  1.566731e-02      1
7    actor2_star_meter 1.464975e-03  9.276008e-02      1
6    actor1_star_meter 8.368206e-04  2.039799e-01      1
15            musical 5.132520e-04  3.198531e-01      1
22          animation 2.748906e-04  4.666386e-01      1
16            romance 2.215080e-04  5.134628e-01      1
8    actor3_star_meter 1.657242e-05  8.581515e-01      2
```

Figure 23: Feature Importance

var1	var2	best_degree1	best_degree2	best_rsq	rsq_1_1	rsq_1_2	rsq_2_1	rsq_2_2
release_year	distributor_rate	1	1	0.279061660	0.279061660	0.279951581	0.286484798	0.286937599
release_year	duration	1	2	0.205057763	0.192511457	0.205057763	0.197942556	0.209102331
release_year	nb_news_articles	1	2	0.180457320	0.118837235	0.180457320	0.121064318	0.192703194
release_year	movie_meter_IMDBpro	1	2	0.087696799	0.045464101	0.087696799	0.056889912	0.095508162
release_year	release_month	1	2	0.054688859	0.039508998	0.054688859	0.044858250	0.059735646
release_year	movie_budget	1	1	0.048444346	0.048444346	0.048596912	0.052477955	0.054421271
release_year	nb_faces	1	1	0.045111724	0.045111724	0.047592369	0.052020237	0.055535943
release_year	actor2_star_meter	1	1	0.041567764	0.041567764	0.041690728	0.047463534	0.048013096
release_year	actor3_star_meter	1	1	0.040258621	0.040258621	0.041339867	0.046107207	0.047722327
release_year	actor3_star_meter	1	1	0.038201176	0.038201176	0.039086987	0.042781773	0.045173529
release_month	distributor_rate	1	1	0.263929589	0.263929589	0.264467595	0.270259997	0.270722355
release_month	duration	1	2	0.191050768	0.171749705	0.191050768	0.176542289	0.195383357
release_month	nb_news_articles	1	2	0.110592770	0.056406512	0.110592770	0.067822017	0.122899867
release_month	movie_meter_IMDBpro	1	2	0.048587453	0.011967324	0.048587453	0.023909435	0.059693563
release_month	release_year	1	2	0.044858250	0.039508998	0.044858250	0.054688859	0.059735646
release_month	nb_faces	1	1	0.019588791	0.019588791	0.020953450	0.031839354	0.033763891
release_month	movie_budget	1	1	0.009323004	0.009323004	0.010382309	0.020707751	0.022306732
release_month	actor3_star_meter	1	2	0.007705891	0.004245509	0.007705891	0.020659861	0.020949093

Figure 24: Interaction Importance

formula_list	list [18]	List of length 18
[[1]]	formula	imdb_score ~ distributor_rate
[[2]]	formula	imdb_score ~ distributor_rate + poly(duration, 2)
[[3]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama
[[4]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[5]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[6]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[7]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[8]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[9]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[10]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[11]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[12]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[13]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[14]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[15]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[16]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[17]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...
[[18]]	formula	imdb_score ~ distributor_rate + poly(duration, 2) + drama + poly(nb_news_article ...

Figure 25: Initial Formula List and their performance

formula	rsq	avg_MSE
1	0.2567400	9.004856e-01
2	0.3769946	7.566802e-01
3	0.3951357	7.353116e-01
4	0.4549190	1.336225e+05
5	0.4814343	3.772717e+04
6	0.4825129	1.962672e+05
7	0.4886527	4.957827e+04
8	0.4890577	4.059439e+04
9	0.4890910	4.330669e+04
10	0.5030855	3.206605e+05
11	0.5059574	2.714924e+05
12	0.5059769	2.436256e+05
13	0.5163701	2.141739e+05
14	0.5165417	1.859966e+05
15	0.5173217	2.922271e+05
16	0.5183741	1.740548e+05
17	0.5213675	3.828598e+05
18	0.5241494	2.518431e+05

Figure 26: R-Squared and Average MSE for Initial Modeling

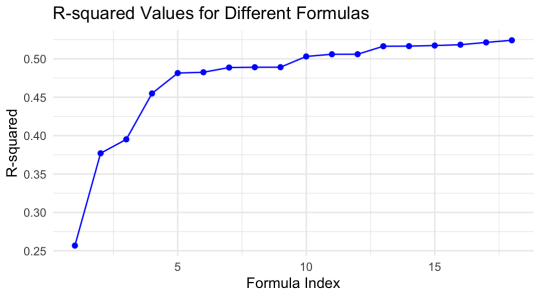

Figure 27: R-Squared for Initial Modeling Plot

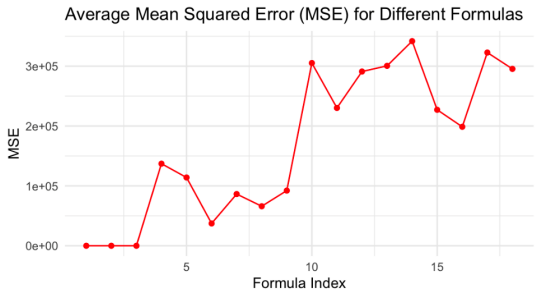

Figure 28: Average MSE for Initial Modeling Plot

Filter

	formula	rsq	MSE
1	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4024707	0.7284628
2	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4040358	0.7252863
3	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4418633	4.1984950
4	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4388129	4.7842448
5	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4091396	0.7208126
6	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4063689	0.7222993
7	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4050823	0.7883680
8	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3975715	0.7346711
9	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4002310	0.7305733
10	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4002310	0.7307605
11	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4144132	0.7124796
12	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.4144132	0.7148327
13	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3961489	0.7372377
14	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3966778	0.7384241
15	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3954020	0.7473441
16	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3954020	0.7369864
17	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3951985	0.7379462
18	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3951985	0.7380414
19	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3951730	0.7356858
20	imdb_score ~ distributor_rate + poly(duration, 2) + d...	0.3951730	0.7366635

Figure 29: Performance for Modeling with Interaction

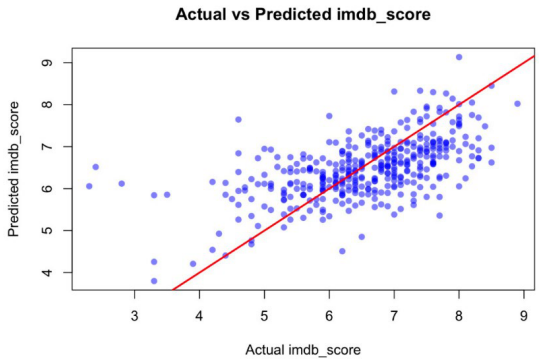

Figure 30: Final Model Prediction Plot

```
> print(result)
                                                movie_title predicted_imdb_score
1                                            Dream Scenario            7.237292
2                                           The Dirty South            6.611951
3                                                  Napoleon            6.966925
4                                             The Holdovers            7.665198
5  The Hunger Games: The Ballad of Songbirds and Snakes            7.393974
6                                                       Leo            6.105171
7                                           Pencils vs Pixels            5.551608
8                                               Thanksgiving            5.961018
9                                         Trolls Band Together            6.024917
10                                             The Marvels            6.239138
11                                            Next Goal Wins            6.562478
12                                                      Wish            5.995152
```

Figure 31: Prediction Results

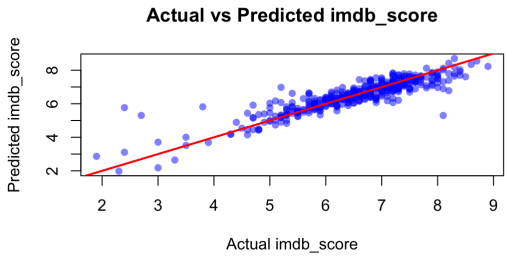

Figure 32: Model with Three Additional Predictors (director_rate, cinematographer_rate, production_company_rate)

Test of Non-Linearity and Fit

Variable	Best Degree	R-squared (Linear)	R-squared (Quadratic)	R-squared (Cubic)	R-squared (Degree 4)	R-squared (Degree 5)
movie_budget	FALSE	NA	6.2e-03	0.0068	0.0085	0.0093
release_year	TRUE	3	3.8e-02	0.0483	0.0489	0.0490
duration	TRUE	3	1.7e-01	0.1880	0.2023	0.2038
nb_news_articles	TRUE	3	5.1e-02	0.1222	0.1343	0.1359
actor1_star_meter	TRUE	4	8.4e-04	0.0014	0.0262	0.0276
actor2_star_meter	TRUE	4	1.5e-03	0.0029	0.0097	0.0110
actor3_star_meter	FALSE	NA	1.7e-05	0.0024	0.0025	0.0031
nb_faces	FALSE	NA	8.0e-03	0.0100	0.0101	0.0111
movie_meter_IMDBpro	TRUE	4	8.1e-03	0.0704	0.1789	0.1847
release_month	TRUE	3	3.8e-03	0.0163	0.0177	0.0203

Figure 33: Test of Non-Linearity and Fit Stargazer *False indicates that linear is the best model*

Spline Regression Results

Predictor	Spline Improved	Best df	R-squared (Linear)	R-squared (df=3)	R-squared (df=5)	R-squared (df=7)
movie_budget	FALSE	NA	6.2e-03	0.0068	0.0085	0.0093
release_year	TRUE	3	3.8e-02	0.0483	0.0489	0.0490
duration	TRUE	3	1.7e-01	0.1880	0.2023	0.2038
nb_news_articles	TRUE	3	5.1e-02	0.1222	0.1343	0.1359
actor1_star_meter	TRUE	4	8.4e-04	0.0014	0.0262	0.0276
actor2_star_meter	TRUE	4	1.5e-03	0.0029	0.0097	0.0110
actor3_star_meter	FALSE	NA	1.7e-05	0.0024	0.0025	0.0031
nb_faces	FALSE	NA	8.0e-03	0.0100	0.0101	0.0111
movie_meter_IMDBpro	TRUE	4	8.1e-03	0.0704	0.1789	0.1847
release_month	TRUE	3	3.8e-03	0.0163	0.0177	0.0203

Figure 34: Spline Regression Results Stargazer *False indicates that linear is the best model*