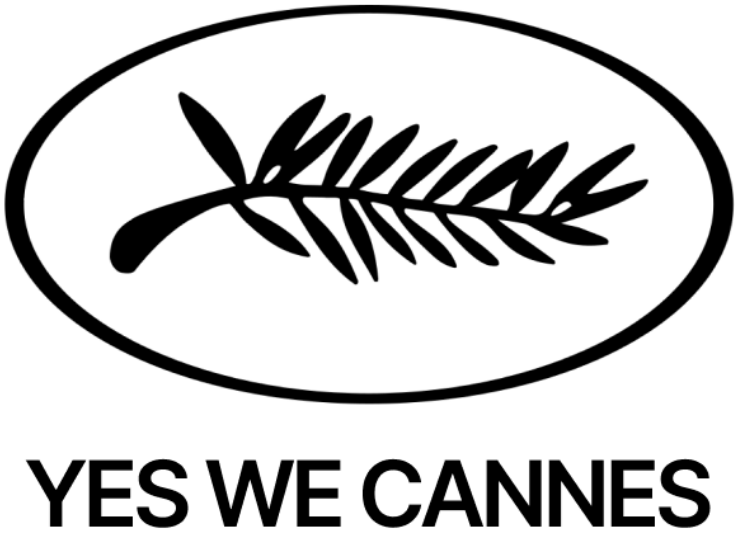
IMDB Prediction Challenge



Sakshi Kumar
Peter Heglund-Lohman
Kabir Nain

I. Introduction

Has anyone managed to find the magical "cheat code" for producing blockbuster films? Does such a code even exist? And just because a film does well at the box office, does that mean it will score well with the movie fanatics on IMDB?

"Cinema is a matter of what's in the frame and what's out."
-Martin Scorsese

In this report, we attempt to predict the IMDB ratings of ten movies planned to be released in November 2022, based on previous data of thousands of movies which were rated by IMDB. Our goal is to create a sufficiently simple and easy to understand model which provides the highest R squared value but also considers heteroskedasticity, collinearity and does not overfit/underfit, allowing us to make the optimal prediction for the IMDB scores of the upcoming movies.

II. Data Description

The initial dataset contained 42 variables, ranging from movie budget to descriptive keyword strings. Our first step was to look at the data table, read the data dictionary and eliminate any variables which seem redundant or unusable at the very outset. Based on our analysis, we eliminated nine variables.

Movie ID and imdb Link were removed because they are identifying variables which don't add any additional information for predictive analysis. *Release year* was removed because even if there was a connection between the year in which a movie was released and IMDB score, it would not give depth to predicting the test movies - all of which are to be released in 2022. *Actor1, actor2, and actor3* were removed because they are simply categorical variables which represent the names for the starmeter numbers - their effect will be difficult to capture. *ColourFilm* was removed since all movies in the test set are colored. *Genre* is also getting captured through the genre-related dummy variables in the data. Finally, *plotKeywords* may add predictive power but it will be difficult to produce them into dummy variables since they are composite texts.

After creating a duplicate dataset "imdb2", the irrelevant variables were removed from "imdb2" and the group moved forward with observing the distribution of each variable as shown

in *Figure 1*. Beginning with the dependent variable "imdbScore" and working through all the predictor variables, the following was observed: (please go to appendix for figures and tables)

Dependent Variable

As stated earlier, our dependent variable is IMDB Score. We observed that the data is fairly distributed between 3.9 to 9.3, with a trail of outliers. (See in Figure2(a), 2(b)).

Independent/Predictor Variables

As seen in *Figure 3a, 3b,* we noticed that most of the movies had a budget between $10~$30 million, and were 60 to 150 minutes long. The variables related to actors' star meter also had many outliers. The main actors, antagonist (actor1,2) have lower starmeter scores in general, but starting from actor 3, more actors are distributed across higher starmeter scores. All three variables are extremely skewed to the right. News articles had one extreme outlier - Star Wars with 60K articles. In terms of genre, we had the maximum number of movies classified under drama. Thriller romance, action and crime were also very popular categories.

For categorical variables, our goal was to observe each variable and create the dummy variables with the level that has the highest frequency to lower the number of dummy variables to be created. For release month, we notice that most movies are released in October, but since all movies in our test set are november releases, we created a dummy variable for November. (see Table 1 in appendix)

We also decided to drop Director and Cinematographers since there were no levels that had a frequency large enough to provide much depth to our analysis.

Relationship of predictor variables with IMDB Score

We now moved on to analyze the relationship of each predictor variable with IMDB Score to see what stands out. We noticed that movie budgets, movie duration and news articles had a positive relationship with IMDB score though not a very strong one. In genre, it was interesting to observe that western, sport, drama, war and animation movies generally had higher ratings while action and horror movies had lower ratings. (Figure 6) Release day seemed to be unrelated to IMDB score. Also, all our movies in the test set either have an aspect ratio of 1.85 or 2.35. These

aspect ratios don't seem to be related to IMDB scores, so we won't be considering them for our model.

III. Model Selection

We wanted to find a model which has a high predictive power but is not complex, relying on too many variables or high degree polynomials. To achieve this, we began by drawing individual linear regression plots to understand the statistical significance of each individual predictor variable.

Afterwards, we started building simple regression models to identify independent variables with high P-Values and then excluded them one by one. During this process, we also kept a close eye on adjusted R-squared values and checked that it did not decrease in response to the elimination of high p-value variables.

The best linear regression model that we found using this method did not have a satisfactory predictive power. After excluding insignificant predictors, we were able to confirm non-linearity and heteroskedasticity within the model through an NCV test(see Figure 8). To solve this issue, we looked for variables with outliers and placed thresholds to ensure that our model was not influenced by the outliers(Figure 9). Then we looked for collinearity by using VIF but surprisingly did not find any(Figure 7).

Finally, we built a polynomial regression model which had degrees from one through four. We stopped adjusting the polynomial degrees at four because beyond this point it did not add to our predictive power and only increased overfitting. We were able to confirm this by running ANOVA tests with models which had up to 10 degree polynomials, which showed insignificant changes to the predictive power. As a result, we made the final selection by comparing the sum of squared residuals between polynomial degree three and four.

No special shape was seen between any of our data variables and IMDB score, so we did not use splines for our model.

IV. Results

Our final regression equation is as follows:

$$imdbScore = 7.28 - 0.000000012 \cdot movieBudget + 14.25 \cdot duration - 4.3 \cdot duration^2 + 9.49 \cdot nbNewsArticles - 4.11 \cdot nbNewsArticles^2 + 2.05 \cdot nbNewsArticles^3 - 0.04 \cdot nbFaces - 3.31 \cdot movieMeter + 5.47 \cdot movieMeter^2 - 6.20 \cdot movieMeter^3 + 3.78 \cdot movieMeter^4 - 0.25 \cdot action + 0.36 \cdot western + .21 \cdot sport - .46 \cdot horror + 0.31 \cdot drama + 0.93 \cdot animation - 0.53 \cdot EnglishLanguage - 0.13 \cdot USAProd + .20 \cdot RRating$$

This final model gives us the following predictions for our test set:

IMDB Challenge Predictions

Falling for Christmas	Black Panther: Wakanda Forever	Spirited	Paradise City	Poker Face	Que viva Mexico!	Slumberland	Blue's Big City Adventure	The Menu	The Fabelmans	Devotion	Strange World
6.1	4.4	5.7	5.8	5.0	7.6	4.9	6.7	5.5	6.8	6.0	5.7

The R^2 of the model is relatively low at 0.41, with an also relatively poor out-of-sample performance with a MSE of 0.71, considering the IMDB scores are in the range of 0-10. Some variables are inflating and deflating the final predictions more than others. For example, although the coefficient for movieBudget is -1.27e-08, when multiplied by Black Panther's $250 million dollar budget (highest in the test set), the IMDB score is brought down by 3.175. All tested movies with a budget above $100 million scored below the average score. On the other hand, the Mexican movie Que viva Mexico! has the lowest budget in the test set at $10 million, and its score goes down by only 0.127. The effect of the factor variables of genre, EnglishLanguage, USAProd and RRating is also direct, with just the coefficients being considered if values are 1 and not if 0.

In terms of the significance of the predictors, the p-values of each predictor are the most significant being <0.001 except the genres of western and sport, and USA_Prod. This is due to our selection of only the most significant predictors in our model selection.

In terms of limitations to the model, we could have improved the model further with more time by testing spline regression models with varying knots to see how it would affect the R^2 value.

There were quite a few categorical variables with high variety and it would be good if we could play more with interaction variables and work with different dummy variables to improve the model. We also believe that adding external data sources to our data points would enhance the model, which would require gathering more data.

Furthermore, we basically had to disregard the effect of key words. If we had the time to properly analyze the effects of specific keywords in movies we could've built a model with higher prediction power. The variety of the keywords is pretty overwhelming, so that would require some serious time.

	Dependent variable:
	IMDB Score
Movie Budget	-0.00***
	(0.00)
Duration	14.22***
	(0.95)
Duration^2	-4.28***
	(0.87)
# of News Articles	9.51***
	(0.90)
# of News Articles^2	-4.12***
	(0.86)
# of News Articles^3	2.06**
	(0.85)
# of Faces	-0.04***
	(0.01)
Movie Meter Score	-3.30***
	(0.86)
Movie Meter Score^2	5.46***
	(0.89)
Movie Meter Score^3	-6.18***
	(0.88)
Movie Meter Score^4	3.77***
	(0.86)
Action	-0.25***
	(0.05)
Western	0.37**
	(0.14)
Sport	0.21**
	(0.09)
Horror	-0.46***
	(0.07)
Drama	0.31***
	(0.05)
Animation	0.93***
	(0.19)
English Language	-0.53***
	(0.14)
USA Production	-0.13***
	(0.05)
Rated R	0.20***
	(0.04)
Constant	7.28***
	(0.14)
Observations	1,920
R^2	0.42
Adjusted R^2	0.41
Residual Std. Error	0.83 (df = 1899)
F Statistic	68.13*** (df = 20; 1899)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 11: Stargazer output of final regression model

V. Appendix

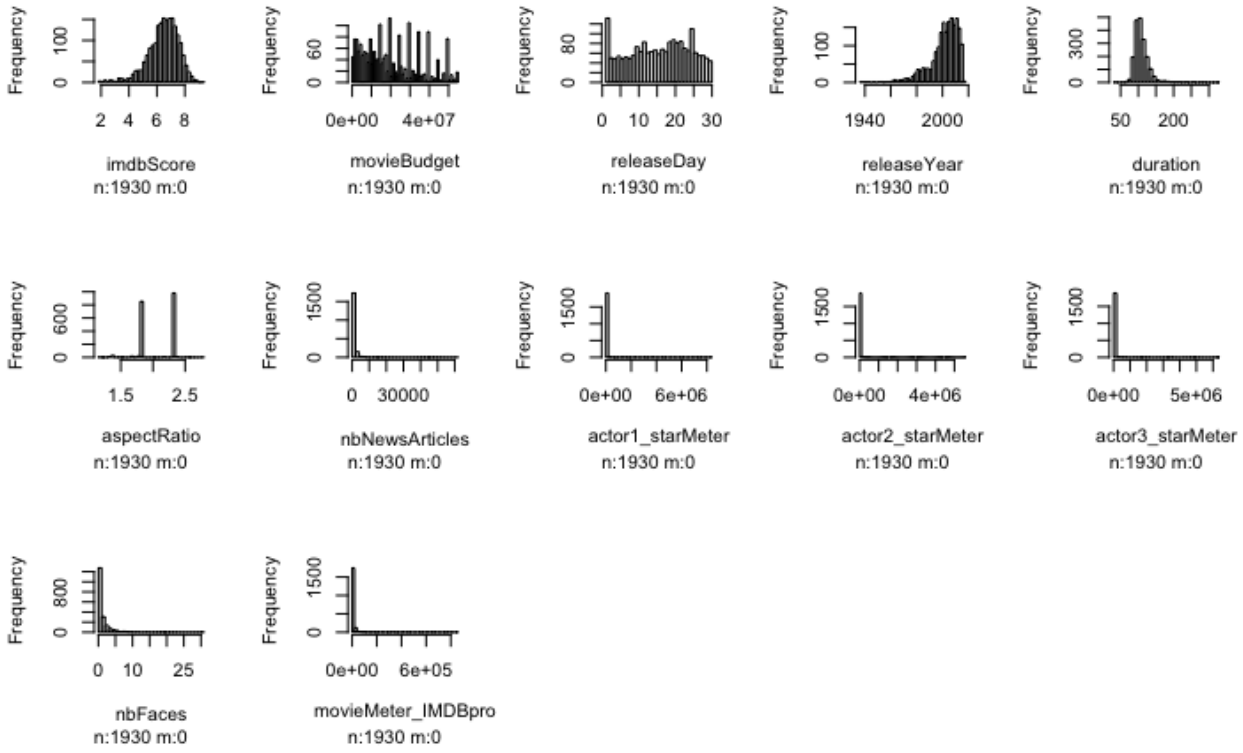


Figure 1: Predictor Variable Distributions

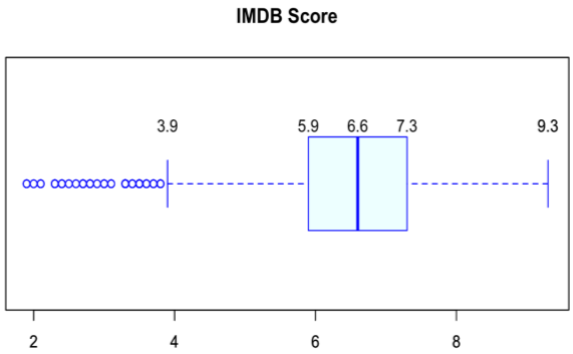


Figure 2(a) – IMDB score boxplot

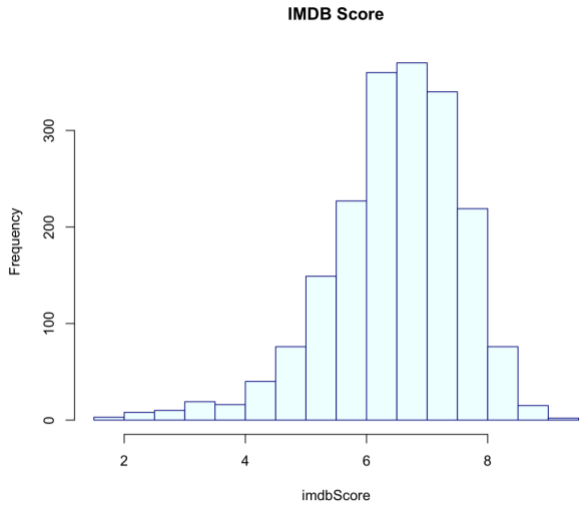


Figure 2(b) – IMDB score histogram

Table 1: New variables created

Variable Name	New Variable Created
releaseMonth	November_Release
language	Engligh_Language
country	USA_prod
maturityRating	R_rating
distributor	distribution_company

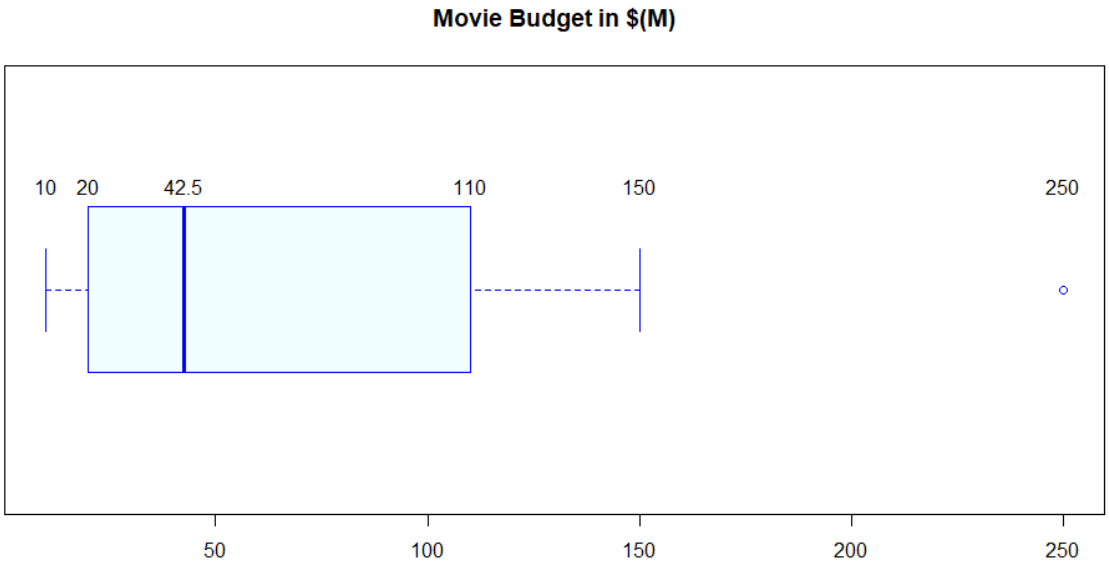


Figure 3a: Boxplot for Movie Budget

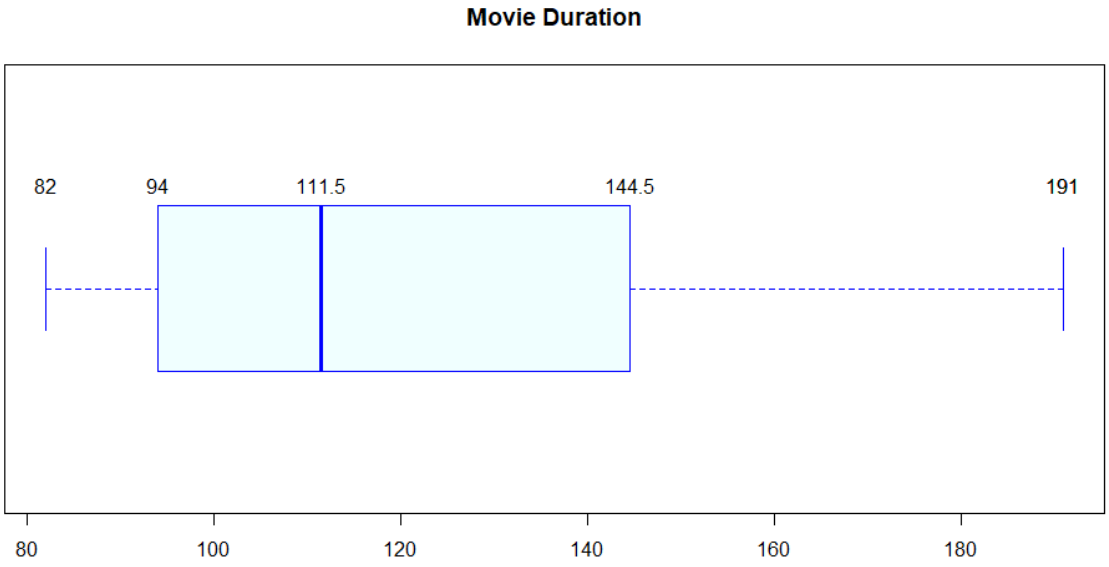


Figure 3b: Boxplot for Movie Duration

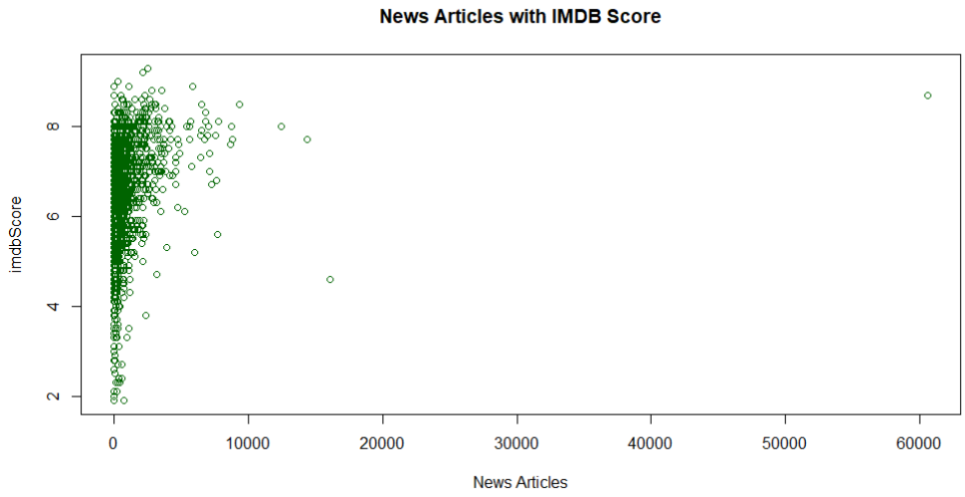


Figure 4: Scatter plot of # of news articles and IMDB score

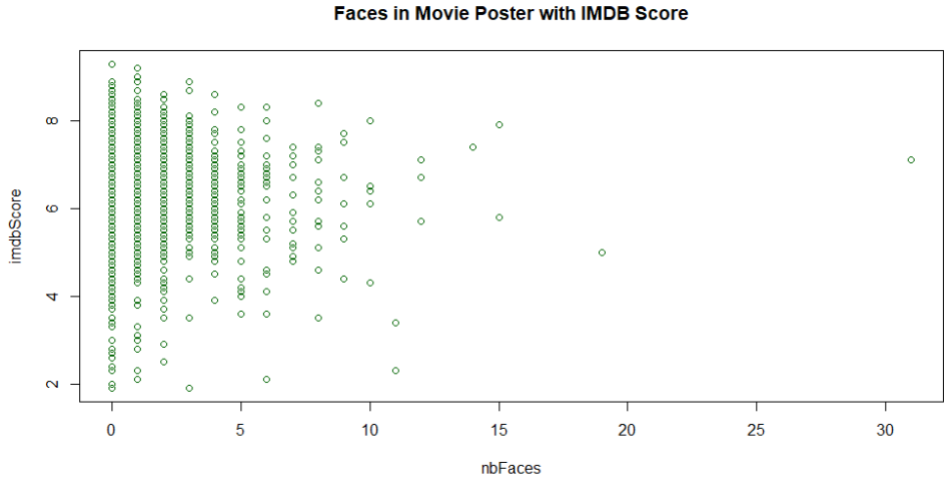


Figure 5: Scatter plot of # of faces in movie poster and IMDB score

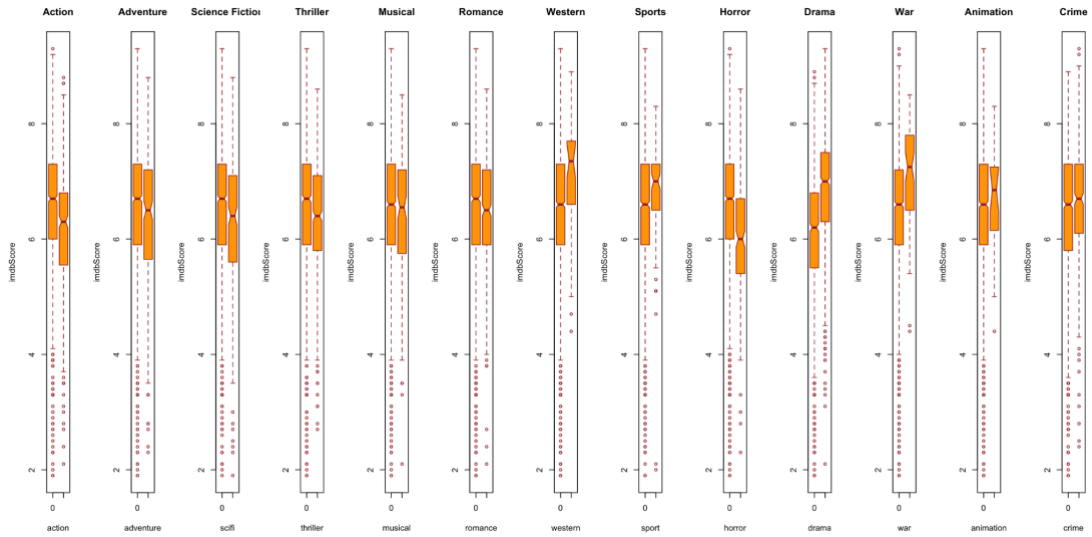


Figure 6: Boxplot of genres and IMDB score

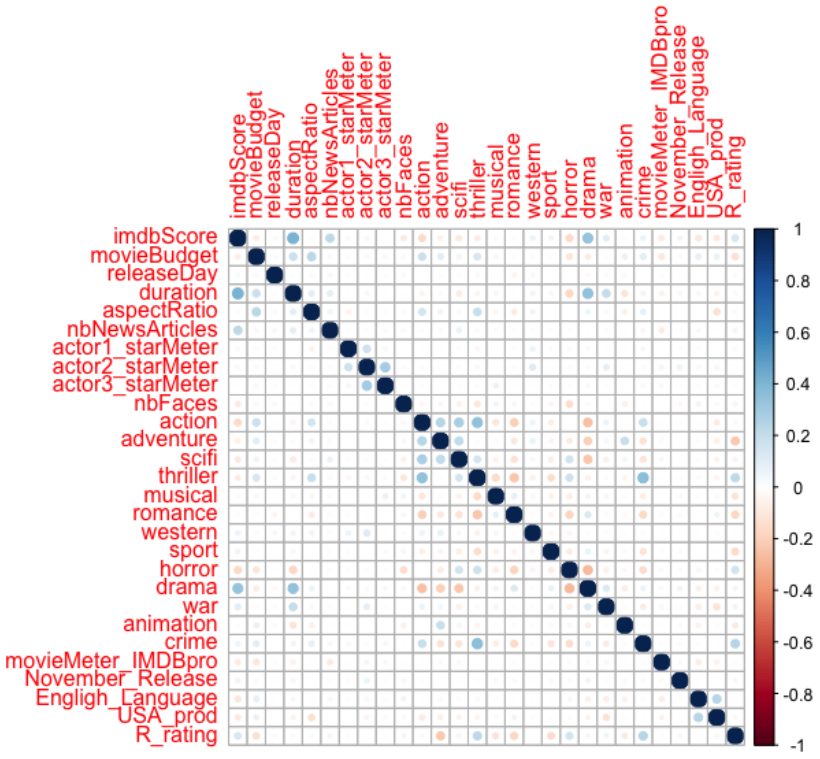


Figure 7: Correlation matrix of quantitative variables

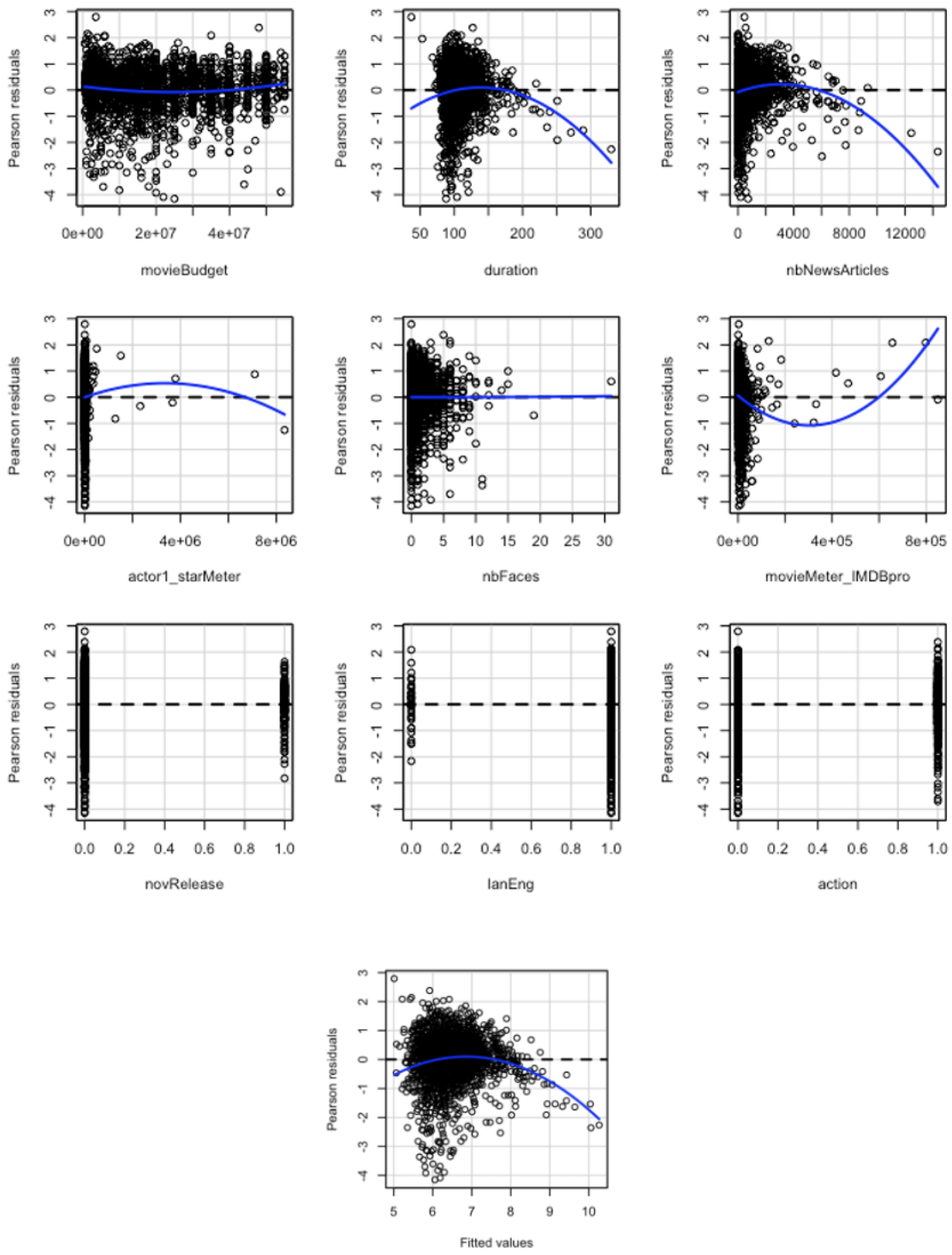


Figure 8: Non-Linear Residual plots of variables in mreg13

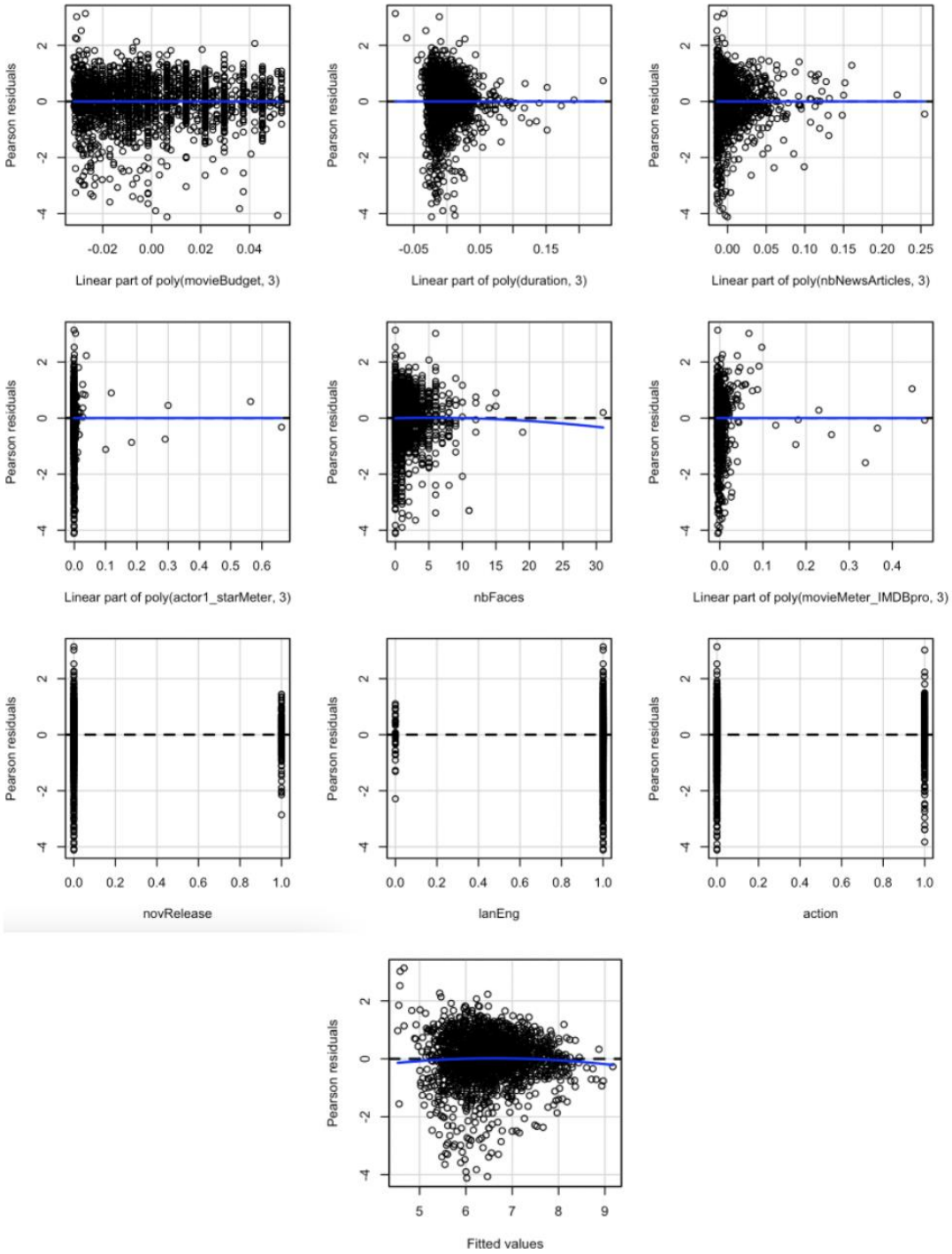


Figure 9: Linear Residual plots after polynomials added

Variables Excluded in Model Building:

Excluded predictors are for low P-Value; releaseDay P-value=0.35; actor1_starmeter 0.20; actor2_starmeter 0.09; actor3_starmeter 0.86; musical 0.32; romance 0.51.

Excluded predictors for no negative effect on R^2:adventure, scifi, war, distribution company, thriller, crime