MGSC661 midterm IMDb Blockbuster Prediction

October 2023

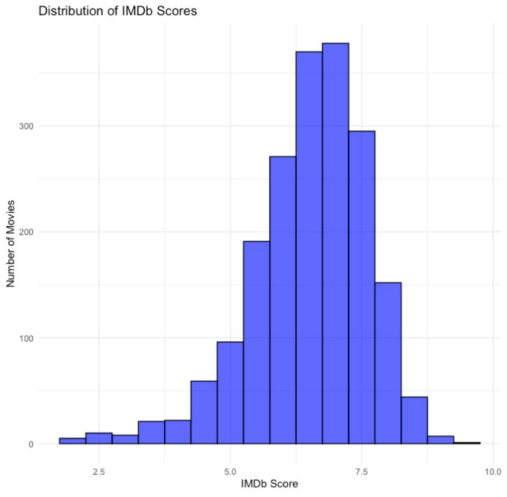
Introduction

In the film industry, making informed decisions on the success of potential movies before their release can offer valuable insight to filmmakers, distributors, and investors. A forecasted IMDb score can serve as a barometer to indicate a movie's potential reception among its audience, allowing stakeholders to tailor their marketing strategies to reach the best audience possible. This report will demonstrate the methodology behind the predictive model developed with R to predict the IMDb score of twelve upcoming blockbuster movies by leveraging an IMDb dataset spanning movies from the 1930s up to just before 2020. This project is centered around the development and evaluation of the predictive model, showcasing the process of identifying the significant predictors, modeling strategies, and prediction results. By analyzing the dataset collected from IMDb that encompasses important attributes such as film, cast, and production characteristics, this project aims to give stakeholders, such as movie fans and industry professionals insights into the potential successes of films before they hit the market.

Data Description

In the realm of film appreciation, an audience's decision to watch a movie is multifaceted and can be influenced by a combination of intrinsic movie attributes and the visual appeal of the marketing elements. Standing at the crossroads of choosing a film and identifying what are the core considerations that guide the choices is pivotal when analyzing the data.

The datasets provided for prediction encompass 42 attributes. Among these, the IMDb rating (`imdb_score`), representing the score given to the film by viewers on the IMDb website, is designated as the dependent variable for the predictive model. Historical scores of past movies will serve as the basis upon which the model forecasts ratings for the upcoming films. The movie title, ID, and URL are treated as labels and are excluded from the set of predictors of the model. From the histogram (see Appendix Figure 1), most movies received above-average to good IMDb ratings between 6.5 and 7.5. The slightly left-skewed distribution indicates there are few movies with particularly low or high ratings. Given the nature of the study, the extreme ratings also reflect strong audience sentiments and therefore will not be excluded from the datasets.



For all the remaining predictors, viewing through the lens of an audience, this section will systematically group them into four categories including movie details, marketing visibility, cast characteris-

tics, and production details. These categories resonate with the key considerations in an audience's decision-making process and will guide the process of understanding their impact on the prediction of the IMDb movie scores to assist in the selection of their inclusion or exclusion from the model.

Movie Details

When viewers browse new film releases, the "movie details" act as the primary filter and shape their initial perceptions. It includes release date (year, month, day), language, country, duration, maturity rating, genres, and the plot of keywords of the content from IMDb. The distribution of release date is evenly distributed in terms of month and date with the year left skewed due to a few old films released before the twenty-first century (see Appendix Figure 2). Both the Language and Country predictors display a lack of variability in the dataset, with 'English' and 'United States' represented as the dominant values, respectively. This near-unary distribution suggests that these predictors might not contribute significant information variance for the predictive model (See Appendix Figure 3 and 4). Most movies in this dataset have durations typical of standard feature films, falling between 1.5 to 2 hours. (see Appendix Figure 5) Few movies have extremely short or long runtimes. A preference in the film industry to produce movies for adult audiences rather than younger viewers can be seen through the fact that many movies in the dataset have "R", "PG", or "PG-13" ratings, which require parental guidance (see Appendix Figure 6). To understand the impact of these high-level details on the IMDb score using a simple linear regression of each predictor with IMDb scores, release day doesn't show statistical significance whereas release year has a positive relationship with IMDb scores (see Appendix Table 1). For every increase of one year, the scores decrease by 0.0182, suggesting newer movies have slightly lower scores than older ones. Only November and December demonstrate statistically significant p-values. However, the R-squared values for individual release months indicate a limited explanatory power for the IMDb score (see Appendix Table 4). Duration is statistically significant, and for every minute longer the score increases by 0.0213, which indicates that a longer movie may tend to have a higher IMDb score (see Appendix Table 1).

Genres are intuitively an important feature, and as many movies belong to multiple genres, looking at the histogram of the dummied genres for each type helps visualize the distribution of genres. There are 13 different genres in total, among these, Drama, Comedy, and Thriller are the top three common genres, while Western, Animation, and Documentary are more on the niche side. There is no collinearity problem (see Appendix Figure 13) among these genres based on the test despite having a relatively stronger correlation between Thriller and Action and Thriller and Crime. This is understandable because they are similar movie types and usually some movies can be classified as both. This would not have a huge impact on the model selected in the later stage. All movie genres, with the exceptions of musicals, animation, and romance, have a statistically significant relationship with IMDb scores. Drama stands out in both effect size and explanation power, where movies within the Drama genre tend to score 0.34 points higher on IMDb scores compared to movies outside this genre. While both the Action and Horror genres negatively impact IMDb scores, their influence, as indicated by their R2 values, is less than that of the Drama genre but still greater than most other genres. The plot of the keywords is extracted from the movie description, to visualize the frequency of the top keywords for each film, a bar chart was produced with the top 5 keywords as murder, love, friend, death, and high school, other keywords appearing often when utilizing a word cloud are wedding, vengeance, and cult film (see Appendix Figure 7).

Marketing Visibility

The second category comprises predictors illuminating the impact of marketing visibility and popularity on the audience's sentiments. These include the number of faces on the poster, the count of news articles discussing the movie, the distributor who plays a pivotal role in marketing the movie, and the 2023 IMDbPro ranking serving as an indicator of the

movie's "word-of-mouth popularity".

	Dependent variable:
	imdb_score
top_4_distributor	-0.005
	(0.055)
Constant	6.513***
	(0.030)
Observations	1,930
R^2	0.00000
Adjusted R^2	-0.001
Residual Std. Error	1.100 (df = 1928)
F Statistic	0.009 (df = 1; 1928)
Note:	*p<0.1; **p<0.05; ***p<0.01

The number of faces on the poster is right skewed, where most posters feature zero to two faces at most (see Appendix Figure 8). There exists a statistically significant negative relationship with the IMDb score and for each additional face on the poster, the score decreases by 0.05 points. No significant statistics show that the count of news articles about the movie will affect the score (see Appendix Table 1). There are 334 unique distributors, and the top distributors are Warner Bros, Universal Pictures, Paramount Pictures, and Twentieth Century Fox. After categorizing movies into two groups—1 for those distributed by the top 4 distributors and 0 otherwise and conducting a linear regression—it's evident that a movie's affiliation with the top 4 distributors doesn't statistically influence its IMDb score.

Cast Characteristics

The third category mainly focuses on the casting team to identify key personnel on a movie's success, the predictors including the director, cinematographer, main, second, and third main actors, as well as their respective actor star meters. The actor star meter consists of a ranking of actor popularity based on the traffic to their pages and the credits they are given, where a low number indicates a higher ranking. Given this, the high number of outliers showed how many actors are significantly less popular than the A-list celebrity actors (see Appendix Figure 10). While other variables' values are different names, which consist of more than 700 unique values, they reveal substantial variability among the names in the dataset. From observing the top names and the number of movies associated with them, it's evident that no single name overly dominates which suggests a diverse representation of individuals rather than a repetitive appearance of the same names.

Production Details

The last category is factors related to the making of the movies, the production company, the movie's budget, the aspect ratio of the image of the film, and the color of the film. There are 768 unique production companies, right skewed indicates the dominance of certain production companies with many others producing a smaller number of movies (see Appendix Figure 11). If this is performed similarly to distributors, the regression analysis shows whether a movie is produced by one of the top 4 production companies doesn't seem to influence its IMDb score. The aspect ratio was discovered as nearly binary with 1.85 and 2.35 for most of the film and has a weak relationship with IMDb scores (see Appendix Figure 12 and Table 1). The movie budget has a coefficient close to 0 with an IMDb score indicating a poor predictive value. Most of the movies are "Color" films and on average, have an IMDb score that's about 0.9953 points lower than "Black and White" films, this predictor is statistically strong (see Appendix Table 1).

While the correlation matrix (see Appendix Figure 14) revealed no strong associations among the predictors, several displayed heteroskedasticity when individually regressed against the dependent variable. These findings are detailed in Appendix Table 2 and will be considered when building the model.

Model Selection

The methodology used to build the prediction model was to first build regressions between the target variable IMDb score and each predictor to find their relationships, p-values, and R-squares. Based on the results of how strong the relationships were, each predictor was chosen to be either included or excluded

in the model.

First, all the release information for the films was excluded from the model including the information around the release such as the poster information. For release year, release month, release day, number of faces on the poster, and distributor, by looking into data rows and making dummy variables, it was found that they are all relatively complex categorical variables that are hard to convert and have fairly weak relationships with the IMDb score due to their low R-squared. The number of news articles, it was dropped for the same reason due to having a fairly low R-squared.

The genres were included in the model since there were strong relationships between them and the IMDb score while only implementing linear for each genre as their binary features. Other irrelevant movie details were dropped from the model such as movie budget, duration maturity rating, and aspect ratio for their low R-squared, language for its unary nature since most of them are in English, country, cinematographer, production company, and director for its complexity in categories. The color of the films was included because the score of black and white movies is significantly lower than colored movies.

The actor names and star meters were also excluded from the model. While the relationship between the actor names and IMDb score is strong when alone, at 64% of variance explained by the first actor's name, the impact on the predictive power of the final model during testing was low (see Appendix Figure 12). Additionally, the actor's IMDb star meter has a very weak relationship and correlation with the IMDb score. This is evidenced by the fact that only 0.08368% of the score is explained by the first actor's star meter (see Appendix Figure 13). When using the Tukey test to test if the star meter is non-linear, there was evidence of nonlinearity particularly for the third actor's star meter (see Appendix Figure 14). However, even when testing the results of these variables using both polynomial regressions and splines, the r-squared improved very minimally. As such, the actors' star meter variables were dropped as well.

Given the results of this initial exploratory analysis and selection process, the following predictors were kept for the final model building through ranking them by predictor strength: drama, horror, action, war, sci-fi, thriller, adventure, western, crime, sport, musical, animation, romance, genres, color of the film, and movie_meter_IMDBpro. The model was first tested with all the genres because genre is an important characteristic of the movie and nearly everyone can tell the difference between genres. Since the genres are all binary variables, multi-variable linear regression was the only model considered to model this variable. Then from the result, variables with a p-value larger than 0.05 were excluded because the hypothesis test shows that there is not enough evidence to prove that there is a relationship between those variables. As a result, adventure, sci-fi, sport, and animation were excluded after this second filtering process.

Since the colored film is a binary variable, dummy variables were created for it and included in our model. Then, the model was built for the movie meter, and in its scatter plot, most of the data are centered between 0 and 100000. Therefore, the movies with IMDBPro scores larger than 100000 can be regarded as outliers and discarded. From further observation, a lot of data was still clustered in certain areas, particularly between 0-12500, so the range was narrowed again to only look at data in this range. There was a curve trend in the scatter plot but without clear knots, therefore, polynomial regression was used to correlate score and ranking with the usage of ANOVA testing to find the best degree for polynomial regression of ranking. By doing the mentioned steps, the final model used was the following with the highest R-squared tested:

lm(imdb_score ~action + thriller + musical + romance + western + horror + drama + war + crime + poly(movie_meter_IMDBpro, 3) + black + color)

This can be translated to linear modeling of IMDb score using these predictors as linear: action, thriller, musical, romance, western, horror, drama, war, crime, black, and white, and this predictors as a third-degree polynomial or quadratic polynomial: IMDb Pro movie meter.

Results

After constructing the model with these predictors, the model was used on the test set which consisted of the 12 upcoming films. With the model, predictions were made to find the IMDb score of these 12 unreleased movies.

Table 1: Predicted IMDb Scores

Movie ID	Predicted_IMDB_score
1	4.669633
2	6.325256
3	7.215588
4	7.655658
5	7.735894
6	6.961345
7	7.892079
8	7.349640
9	6.138990
10	7.443147
11	7.558167
12	7.271077

The lowest predicted score of the 12 is Movie 1, *Pencils vs Pixels*, with an IMDb score of 4.67. The highest score is 7.89 for Movie 7, *The Hunger Games: The Ballad of Songbirds and Snakes*. When observing the results, the majority of the scores are centered around a score of 6 to 7. The overall R^2 of the model is 0.368, meaning that there is room for improvement. For the significance of each predictor, *action*, *drama*, *movie_meter_IMDBPro*, and *horror* all demonstrate extremely low p-values (approximately 0), which indicates a very high significance for these predictors. The significance of *romance* and *black* are also high because they have low p-values as well. However, they are not as significant as the predictors with a p-value close to 0. When it comes to *War* and *Western*, these genres have a significant impact on the regression model and both of them have a p-value < 0.01, though they are still less significant than previously mentioned predictors. Compared with the significant indicators, there are several variables, including *Thriller*, *Musical*, and *Crime*, that have p-values that are not significant (p-value > 0.05). These predictors are still included in the regression model because when testing, it was found that the R^2 decreases after dropping them.

Given that the model included the color of the film, it included the dummy variable created for *film_color*. This indicates that if the film is in color, it will be assigned 1 and if the color is black and white, it will be assigned 0. As such, the value for the color films is not applicable in the regression model because there is another variable for black and white films which has a linear relationship. Because the variables are dummified and show a linear relationship, the model doesn't have any values for the "color" values because it is the comparison of the impact of black and white films vs. color films, where color films are the baseline.

R Estimates Results is:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.45973	0.04530	142.594	< 2e-16
action	-0.25017	0.06000	-4.169	3.22e-05
thriller	-0.10101	0.05457	-1.851	0.06434
musical	-0.13035	0.08649	-1.507	0.13197
romance	-0.21219	0.05251	-4.041	5.59e-05
western	0.45585	0.15665	2.910	0.00366
horror	-0.48611	0.07210	-6.742	2.20e-11
drama	0.59253	0.04701	12.604	< 2e-16
war	0.39070	0.12320	3.171	0.00155
crime	0.10117	0.05729	1.766	0.07761
poly(movie_meter_IMDBpro, 3)1	-16.07230	0.83287	-19.298	< 2e-16
poly(movie_meter_IMDBpro, 3)2	6.42499	0.83481	7.696	2.48e-14
poly(movie_meter_IMDBpro, 3)3	-1.45161	0.83203	-1.745	0.08124
black	0.69273	0.11584	5.980	2.76e-09
color	NA	NA	NA	NA

The R-squared for the predicted model based on test data is 0.709, which was calculated using the mathematical definition of R-squared. The resulting R-squared of the model is considered high, indicating good model performance. However, the result is only tested on a very small test dataset, and the score stems from a very small number of users outside North America. Therefore, we cannot rely on those scores to test our prediction model. Moreover, K-Fold Cross-Validation (K-Fold CV) was used with k=10 to test the out-of-sample performance of our predicted model. The MSE for the model was 0.6934, which is considerably low. It is important to note however, that the model is built on the dataset up until the pro-score of 12500, which means that the IMDB-pro score for the model cannot exceed 12500, since the model considered an IMDB-pro score greater than 12500 as outliers.

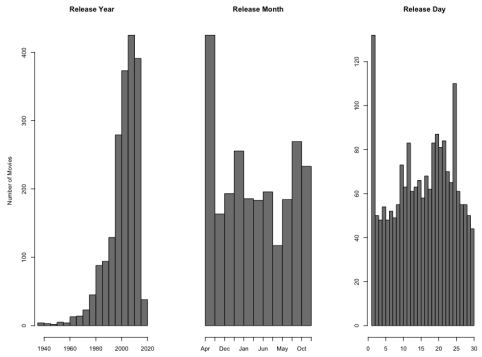


Figure 2: Release Time

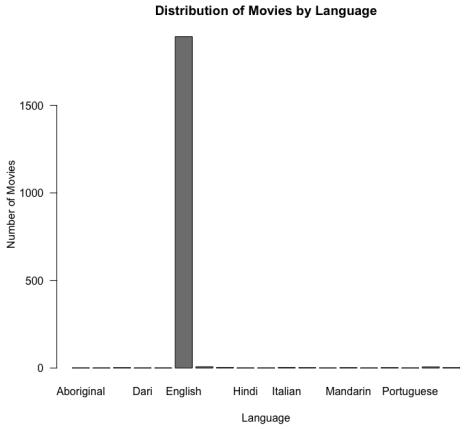


Figure 3: Distribution of Movies by Language

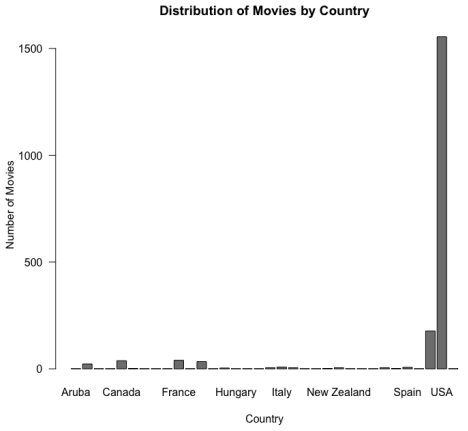


Figure 4: Distribution of Movies by Country

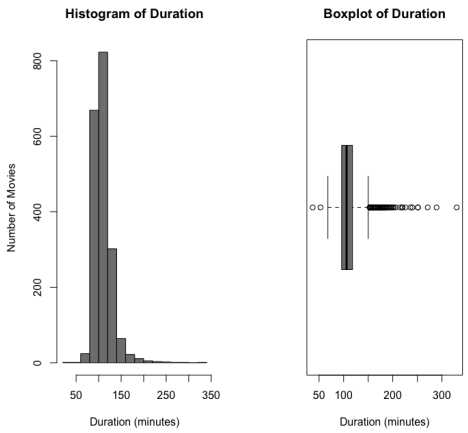


Figure 5: Duration

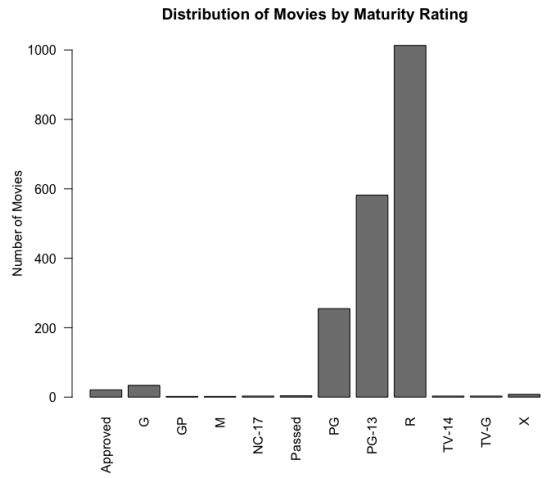


Figure 6: Distribution of Movies by Maturity Rating

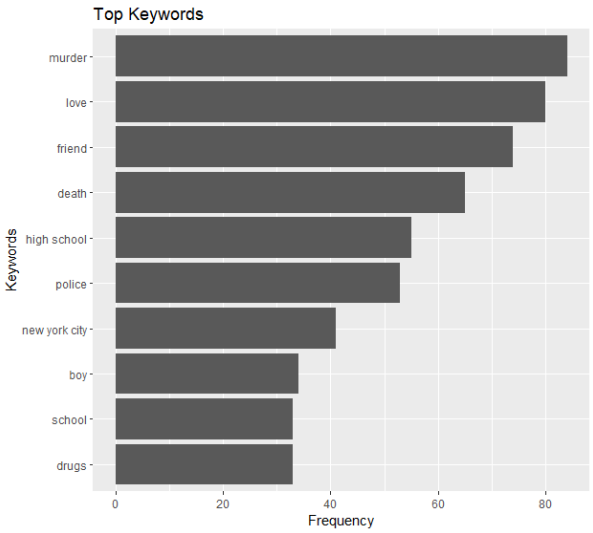


Figure 7: Top 10 Key Words

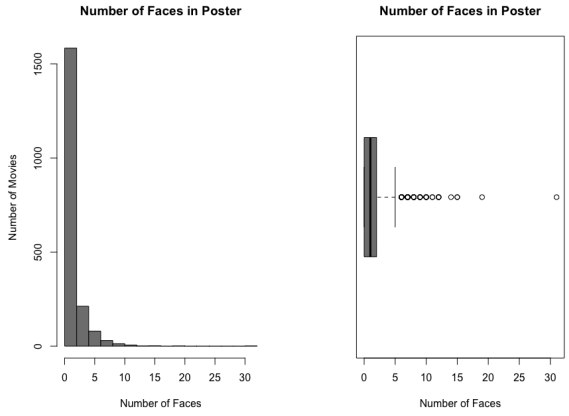


Figure 8: Number of Faces in Poster

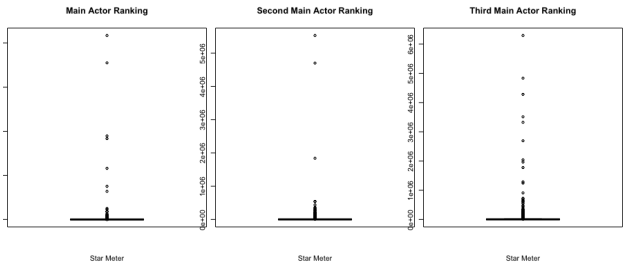


Figure 9: Actor Meter

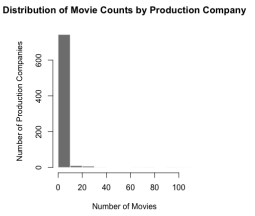


Figure 10: Production Company

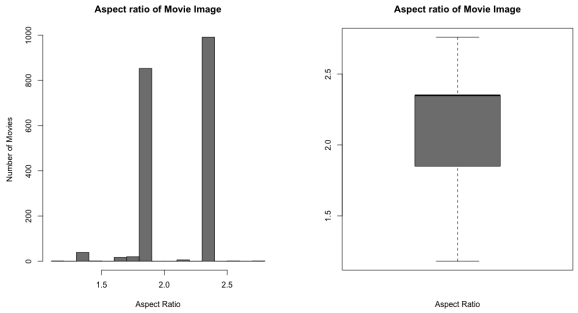


Figure 11: Aspect Ratio

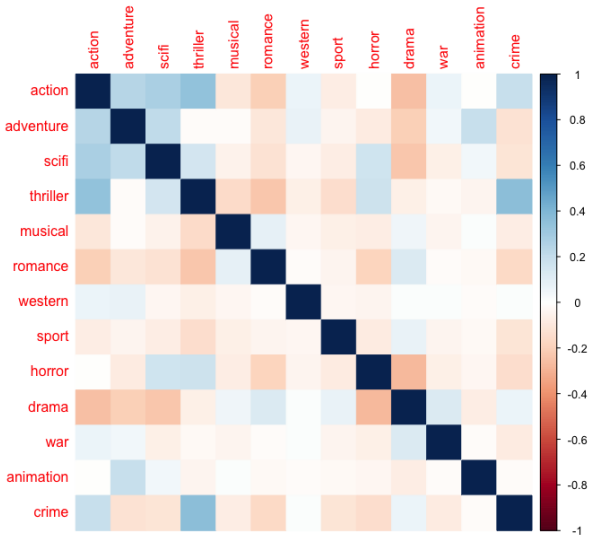


Figure 12: Correlation Matrix for Genres

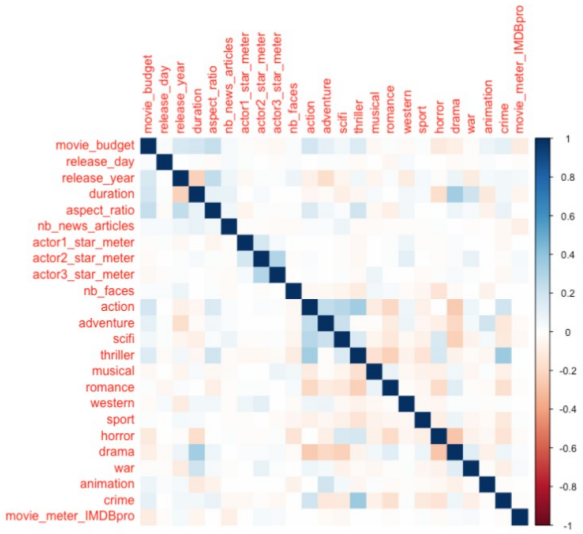


Figure 13: Correlation Matrix for All

Table 1: Linear Regression Results For Numerical Variables

Dependent Variable: imdb_score

Predictors	Coefficient	p-value	R-square
release_day	0.002827	0.349	0.0004553
release_year	-0.018199	<2e-16	0.03796
duration	0.021261	<2e-16	0.1686
action	-0.43688	2.1e-12	0.0253
adventure	-0.22136	0.00328	0.004474
scifi	-0.33197	3.67e-05	0.008799
thriller	-0.19246	0.000432	0.006406
romance	-0.03805	0.513	0.0002215
western	0.54784	0.00397	0.004295
sport	0.28244	0.0157	0.003025
horror	-0.57699	2.1e-13	0.02758
drama	0.74751	<2e-16	0.1144
war	0.63849	1.76e-06	0.01178
animation	0.18005	0.467	0.0002749
crime	0.16419	0.00693	0.003775
nb_faces	-0.04773	8.39e-05	0.007992
nb_news_articles	1.333e-04	<2e-16	0.05083
movie_meter_IMDBpro	-2.457e-06	7.9e-05	0.008052
actor1_star_meter	1.113e-07	0.204	0.0008368
actor2_star_meter	2.430e-07	0.0928	0.001465
actor3_star_meter	-1.720e-08	0.858	1.657e-05
movie_budget	-5.916e-09	0.000542	0.006189
colour_film	-0.9953	1.19e-12	0.02586
aspect_ratio	0.04485	0.625	0.0001242

Figure 14: line Regression Result

Table 2: Non-Constant Variance and Linearity Test Results for all Predictors

Dependent Variable: imdb_score

Predictors	Non-Constant Variance Test		Linearity Test	
	p-value	Heteroskedasticity	p-value	Non-linearity
release_day	0.62425	N	0.2940	N
release_month	0.0050337	Y	NA	NA
release_year	0.52449	N	0.001764	Y
duration	4.3112e-07	Y	2.094e-11	Y
language	0.0026813	Y	NA	NA
country	5.1818e-06	Y	NA	NA
maturity_rating	4.1509e-07	Y	NA	NA
action	0.37102	N	0.7937	N
adventure	0.0015592	Y	0.7524	N
scifi	4.8163e-05	Y	0.7597	N
thriller	0.0010897	Y	0.7746	N
romance	2.4455e-05	Y	0.7441	N
western	0.48585	N	0.7293	N
sport	0.4815	N	0.9261	N
horror	0.56797	N	0.7771	N
drama	7.6872e-12	Y	0.9683	N
war	0.03503	Y	0.7192	N
animation	0.41475	N	0.7349	N
crime	0.001752	Y	0.7099	N
nb_faces	0.7731	N	0.08044	Y
nb_news_articles	< 2.22e-16	Y	< 2.2e-16	Y
distributor	9.2711e-07	Y	NA	NA
movie_meter_IMDBpro	0.00013063	Y	3.824e-15	Y
actor1	3.8245e-06	Y	NA	NA
actor1_star_meter	0.73173	N	0.3410	N
actor2	< 2.22e-16	Y	NA	NA
actor2_star_meter	0.85605	N	0.5531	N
actor3	6.0001e-07	Y	NA	NA
actor3_star_meter	0.026514	Y	0.03687	Y
director	1.4176e-05	Y	NA	NA
cinematographer	1.4815e-10	Y	NA	NA
movie_budget	2.9403e-06	Y	0.2960	N
colour_film	0.0025001	Y	NA	NA
aspect_ratio	1.3759e-06	Y	0.004182	Y
production_company	0.0012032	Y	NA	NA

Figure 15: Dependent Variable

Table 3: Linear Regression R-Square for Categorical Virables
Dependent Variable: imdb_score

Predictors	R-square
release_month	0.02606
language	0.01647
country	0.03768
maturity_rating	0.0474
plot_keywords	
distributor	0.2567
actor1	0.636
actor2	0.7678
actor3	0.8562
director	0.7822
cinematographer	0.5488
production_company	0.5158

Figure 16

	Dependent variable:
	imdb_score
Aug	-0.125
	(0.118)
Dec	0.490***
	(0.127)
Nov	0.326***
	(0.123)
Oct	0.206*
	(0.112)
Sep	-0.043
	(0.116)
Jul	0.041
	(0.122)
Jun	0.247**
	(0.123)
May	-0.192
	(0.140)
Mar	-0.061
	(0.121)
Feb	0.111
	(0.121)
Jan	0.169
	(0.113)
Constant	6.411***
	(0.084)
Observations	1,930
R^2	0.026
Adjusted R^2	0.020
Residual Std. Error	1.089 (df = 1918)
F Statistic	4.666*** (df = 11; 1918)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 17