DESAUTELS FACULTY OF MANAGEMENT

MGSC 661 - Multivariate Statistics

THE 2021 IMDB PREDICTION CHALLENGE

Liu, Alice	261007356
Jiang, Yingxin	261007353
Hedley, Vahid	261026933
Liu, Juliana	261008240
Samuel, Nancy	260948517

November 16, 2021

Contents

1 Introduction

With the advent of the internet and the exponential growth of the movie industry across the globe, consumers and producers alike have drawn to ratings as a basis to evaluate the merit of a particular movie. These ratings typically are seen as being good predictors of the quality of the movie as they generally encapsulate any relevant aspects that would make a movie good. IMDB is one of the largest online movie databases with information pertaining to these movie scoring methods and serves as a central repository for many novice and experienced moviegoers. In this respect, IMDb offers a rating scale that allows users to rate films and the general scoring for the IMDB method is based out of 10, where the higher score is given for the better movie. IMDb indicates that submitted ratings are filtered and weighted in various ways to produce a weighted mean that is displayed for each film, series, and so on. However, the exact methodology for generating these scores remains elusive, and up until the present time, the public has no awareness how these scores are measured.

Since this score is oftentimes used by viewers as the ultimate source of validity for ascertaining the merit of a movie, it is imperative the actual calculation is accurate and correctly reflects all components of that movie. There are a range of features that impact the score of a movie, and some of them can be categorized and quantified into easily identifiable buckets, which can then be used to build an advanced statistical predictive model to forecast the IMDB score of a particular movie. The dataset that will be used to conduct this analysis includes predictors ranging from movie budget and number of actors to genre of movie and year of release. Each of these variables was examined meticulously to observe which ones have a statistically significant relationship with the overall IMDB rating of a movie. After selecting those features which have the greatest predictive power, a model was then constructed with the goal of generating insight into how future movies may be scored. To mitigate against potential overfitting issues on the training dataset, K-fold cross-validation was employed to test its out-of-sample performance. The findings of this report highlight our team's exploration into the distribution and relationships between variables, the overall methodology and approach to building the model, and the results of the final model.

2 Data Description

2.1 Data Preprocessing

The overall dataset used for this analytics project included 2,953 data points with 51 columns aggregated from the IMDB website. The features within this dataset ranged from broad movie characteristics, production characteristics, identifier elements, and cast attributes. There were several discretionary features that were dropped right off the bat, such as the imdb_url and the imdb_id, as it was determined through reason and logic that they would not generate any predictive power on the model. Moreover, it was concluded that most of the categorical features would be dropped as there would be many unique categories for every one of these attributes which would be computationally exhaustive if dummy variables were to be generated. The information captured by several of these dropped features would also be captured in binary categorical predictors, such as main_actor_name and main_actor_female, for example. The only two categorical features that were kept include main_lang and main_production_country as it was relatively easy to classify them into broader buckets based on their relative frequency.

2.2 Predictors Exploration

An initial scoping of the level of importance and distribution of each of the predictors was conducted relative to the IMDB score. Through examining the data through this lens, the skewness of the features can be explored to determine if the distribution is particularly uneven. Skewness in any of the variables, as indicated by a higher density or concentration of data points within a certain region, can be troublesome when developing a model as it has the propensity for overfitting if test dataset points fell outside of those ranges. To correct this, predictors which displayed signs of skewness were normalized into logarithmic scale. This process however did not generate any significant impact on our final MSE, and as such, it was established they do not warrant any further discussion.

The following points highlight at a high level the distribution and overall patterns observed within the selected features used for the final model.

Main Language: There were four buckets of languages that were classified into the same grouping as the overall proportion of languages dropped significantly after this point. 88 percent of observations fell into the 'English' grouping, with only 2 percent for 'French' and 1.3 percent falling into 'Dutch', respectively. Refer to image (Refer to Figure 2 and 12)

Budget in Millions: This variable had a right skewed distribution with a mean budget of 35.169 million dollars. The highest concentration, around 54 percent of observations, fell within the Imdb score region of between 6-8. (Refer to Figure 4)

Year of Release: Was left skewed with the median value for year of release being 2001. There were approximately 61 percent of observations falling between the years 1998-2020, accounting for a larger share of the noted data points. (Refer to Figure 5)

Duration in Hours: There was a very normal distribution observed for duration in hours, as is typically seen for time series datasets such as this. The mean time for movies was 1.854 hours, with a

median of 1.783 and with 75 percent of all observations falling below the 2.017 hour point. (Refer to Figure 11)

Total Number of Actors: The data points for this plot followed a normal distribution with 68 percent sitting within the scoring range between 6-8. The median for this variable sat at 18 actors per movie, with 25 percent of movies containing 14 actors or less. (Refer to Figure 14)

Total Number of Producers: There was a right skewed distribution with each step of the bucket of a histogram falling in a very consistent manner. The median number of producers was 2 per film, and 93 percent of movies had a score of 8 or lower. (Refer to Figure 13)

Main Production Country: For this variable, it was determined that we would generate several buckets with the highest frequency of movie count to include groupings to generate dummy variables. 70 percent of observations fell into the category of USA, 8.6 percent under the United Kingdom, 4.1 percent under France, 4 percent under Germany, 3.5 percent under Canada and the rest were classified as 'Other'. Initially, a model was created with only two binary groupings, 'USA' and 'Other', but after testing the model it was observed that the performance was superior when utilizing these five production country buckets as compared to the two binary ones. (Refer to Figure 3 and 15)

Genre categories:There are 23 genres in the form of dummy variables, all the genre_x columns have values of 0 or 1. From the plot in Figure 1, we see that there are no entries for 'realitytv' or 'shortfilm' genre. A vast majority of the movies are of the genre 'drama', 'comedy' and 'action'.

3 Model Selection

3.1 Analysis of predictors

After going through and observing the relationships between each of the predictors and the Imdb rating, the next step would be to correct any model issues that may be latent within the dataset. Within this context, a simple linear model was first developed and a correlation heat map was created. Variance Inflation Factors tests (Refer to Figure 7) were conducted to analyze and measure the collinearity and correlation between the predictors. After running the tests, we had found out that no collinearity exists between any of the variables as all of the values for the Pearson correlation coefficients fell below 0.8 (Refer to Figure 6). Furthermore, the VIF scores (Refer to Figure 7) for all predictors was below 4, giving us more support to conclude that there exists no collinearity. To determine whether heteroskedasticity was present, a non-constant variance test was implemented and the values highlighted in orange in Figure 7 represent the values with a p-value less than 0.05. These were then corrected to acquire the accurate linear performance of those particular predictors. Residual plots were constructed for each of the predictors as depicted in Figure 8, and the overall model, and the results obtained from this inspection indicated that the specific predictors can't be treated as being linear because their p-values fell below 0.05 (Refer to Figure 7 and 8). The ones which surpassed this threshold were identified to further examine their predictive power during the later stages of building and testing the model. Finally, to find any inherent outliers in the dataset, a Bonferroni outlier test was performed and there were 6 outliers that identified and removed having a p-value falling under the threshold of 0.05 (Refer to Figure 16).

3.2 Model Building

We begin with running a linearity test for all numeric predictors (Refer to Figure 7). Based on the results from residual plots, we observe that only month of release and total number of production companies satisfy the linear assumption. For those non-linear predictors, we decide to perform polynomial regression to avoid underfitting. We first choose the most appropriate degree for each non-linear numeric predictor by using the validation set test and k-fold cross-validation (Refer to Figure 9). We use K=54 for all K-fold cross-validations since $\sqrt{2953}$=54. We adopt the polynomial degree from the k-fold cross-validation which is less time-consuming and utilizes more data for training. We do not have a strong intention to use spline regression for numeric variables since we don't see any clear spline shape in the scatter plot. However, the predictor duration_in_hours has a relatively high polynomial degree, hence spline regression is implemented to see if it provides a better fit. We select 1.5, 2, 2.5, 3 as knots based on the pattern we observe in the scatter plot of duration_in_hours. In fact, there is no big difference between using a spline with degree 6 and using a degree-6 polynomial since both regressions generate similar R^2. Therefore, we still choose to perform polynomial regression for duration_in_hours.

Here are the steps of building our model with all numeric predictors:

Step1: We started building up our model by adding continuous predictors one by one. Based on our intuitions, we first construct a rough ranking of the importance of each continuous predictor. The more important a predictor is to the IMDb score prediction, the earlier we added it to the model.

Step2: We start with budget_in_millions, then add each predictor in the order of their ranking to the model and compare the adjusted R^2 before and after adding to determine whether we should

include this predictor or not. The adjusted R-squared increases when the new predictor improves the model. In our case, we would only include the predictors that make significant contributions (greater than 0.01) to the increase of adjusted R^2.

Step3: Since we haven't tested the significance of the first predictor added to the model, after adding all numeric predictors into the model, we removed the first predictor(budget_in_millions), checked the adjusted R^2 and then confirmed that it should be included in the model.

Step4: At this stage, the numeric predictors we include (d=polynomial degree) are: budget_in_millions (d=4), year_of_release (d=3), duration_in_hours (d=6), and total_number_of_actors (d=2).

Step5: We are aware of the fact that the high polynomial degree of duration_in_hours may lead to overfitting. After checking the summary and observing high P stats for duration_in_hours at d=5 and d=6, we decided to run a Polynomial ANOVA test to obtain an optimal degree for this predictor. Results from the ANOVA test (Refer to Figure 10) showed that degree 5 and 6 might not have significant improvements on our model. Therefore, we changed the polynomial degree of duration_in_hours from 6 to 4.

For the categorical variables, we are considering dropping main_actor1_name, main_actor2_name, main_actor3_name, main_director_name, main_producer_name, editor_name, main_production_company. The rationale is that generating dummies for these variables would create a model very prone to overfitting because these actors, directors, producers, editors, and production companies may or may not be in movies in the test dataset that will be given to us to look at predictions. For predictor main languages, given that the 3 dominant categories are English (88%), Francais(2%), Deutsch(1%), therefore, we decide to create 4 dummy variables, English, Francais, Deutsch and Others. This approach is easier than creating dummies for all of the main languages falling into the Others category, which are fairly insignificant. The same logic applies to predictor main production country, given that 5 dominant categories are US(70%), UK(8%), France(4%), Germany (4%), Canada(3%), therefore, we decide to create 5 dummy variables, US, UK, France, German, Canada and Others.

Step1: Running a general multinomial model with all the numerical polynomial predictors (with their optimal combination of polynomial degrees from the k-fold validation set), numerical linear predictors and categorical predictors.

Step2: Based on the model summary from step1, removing genre dummy variables that are either not significantly affecting the IMDb score (p-value >0.5) or not a dominant category for all the movies.

Step3: Based on the model summary from step2, remove numerical predictors (both polynomial and linear) that do not significantly affect the imdb score (p-value >0.05).

Step4: Since most of the movies have more than one genre, so we add interaction terms between each genre to see the imdb rating difference in the effect of having a movie that belongs to one genre to another (i.e. how much an adventure genre movie score will change when a movie is also vs. isn't a sci fi genre movie).

Step5: Based on the model summary from step4, remove genre interaction terms that do not have an interaction value (i.e. interaction coefficient = 0).

Step6: Based on the model summary from step5, only keep interaction terms for genres that significantly affect the imdb score (p-value >0.1).

Step7: Adding interaction terms between predictor year of release and genre to see how the movie genres affect the IMDB score throughout the year (i.e. the positive slope coefficient, 0.62, of the interaction term between drama genre and year of release against the imdb score suggests that the quality of comedy genre movie has been increasing over the years)

Step8: Based on the model summary from step7, only keep interaction terms for genres and year of release that significantly affect the IMDB score (p-value >0.05). In addition, after trying all the other interaction terms between predictors, we consider four important ones to include in the model since they significantly affect the IMDB score (p-value >0.05), which are interaction terms between numerical predictors (total number of directors and budget in millions, total number of production countries and total number of actors), interaction terms between categorical predictors (main_actor1_is_female and main_actor2_is_female, main_actor2_is_female and main_actor3_is_female).

Step9: Based on the model created in step 8, we use resampling methods of K-fold cross-validation test with K=200 to test performance of the model to make sure that the model is not overfitting. The MSE we get from the K-fold cross-validation test is 0.55.

Step10: Conduct an outlier test for the model we created in step 8, and remove the 6 outliers (Refer to Figure 18) from the original dataset, and rerun the K-fold cross-validation test with K=200 on the new dataset without the outliers.

3.3 Model Performance

The final mean squared error (MSE) we get is **0.526**, which implies that on average, the final model will deviate from the real IMDB rating of a new movie by +/- 0.526 rating points. The adjusted r-squared we get from the final model is 0.4322, which implies that the final model will capture the variability of IMDB rating for 43.22% of the movies in our dataset. We choose not to use Validation Set Test and Leave One Out Cross-Validation because it is more computationally and time-consuming due to the large data set.

4 Managerial Implications

Based on the p-value in the summary report of our model, movies with a large number of actors, as well as movies with genre in action, adventure, drama-action, crime-fantasy and main actors being female tend to hurt the Imdb score. Meanwhile, movies with high budget, long duration, more producers and genre documentary, as well as movies released recently or at the end of the year tend to have high Imdb scores.

Movies with more producers, long duration and high budget require massive initial investment, which usually only happens with large production companies. Because of these companies' maturity in the film industry, the movies produced are generally of high quality, hence leading to a high IMDb score.

There is an increasing number of action/adventure/fantasy type movies due to audience' preferences these days. However, movies of these genres are usually costly and hard to produce. Therefore, although many companies are trying to attract the audience by producing movies of these genres, only a few of them actually have the capability of making a good one, leading to low ratings. This explains why movies of genre action/adventure/fantasy can hurt the IMDb score in our model.

The reason why movies released recently tend to have higher scores is due to the development of technology, which brings unique and new experiences to audiences. Hence, companies should consider investing in high-tech such as special effects, IMAX, etc.

As the end of the year approaches, so does Christmas break and New Year. People tend to go to the cinema with a happy mood, which positively affects the movie-watching experience. Moreover, since there is high competition among movies released around Christmas, companies are putting their full efforts in making the best ones. Hence, movies released at the end of the year tend to have high IMDb scores.

The movie industry is gender-biased, especially in Hollywood. Many movies from the past tend to have male as main characters and female characters as supporting roles. The two genders are usually connected with romantic relationships in the plot, with male characters superior to female characters. Hence, female leading roles used to be rare and not favored by the audience. However, this doesn't mean companies nowadays should still follow this path. In fact, there is a growing trend of female leading roles in movies due to feminism awakening. Therefore, investing in movies with feministic themes might be a better option.

To add another dimension of depth to this analysis, future model development could include integrating other features to further enhance the model. Other features could include data pertaining to amount spent on marketing campaigns, the frequency of movie premiers, or the total spending on sub-categories of a movie as a function of the movie type (ie. budget for special effects for science fiction movies). The only challenge with this was the consistency of data across all areas, which would limit the efficacy of such an approach for this type of model. To conclude, the data used within this project and model has been able to fairly accurately generate predictions with a low margin of error while minimizing potential for overfitting and reducing the impact of outliers.

5 Appendices

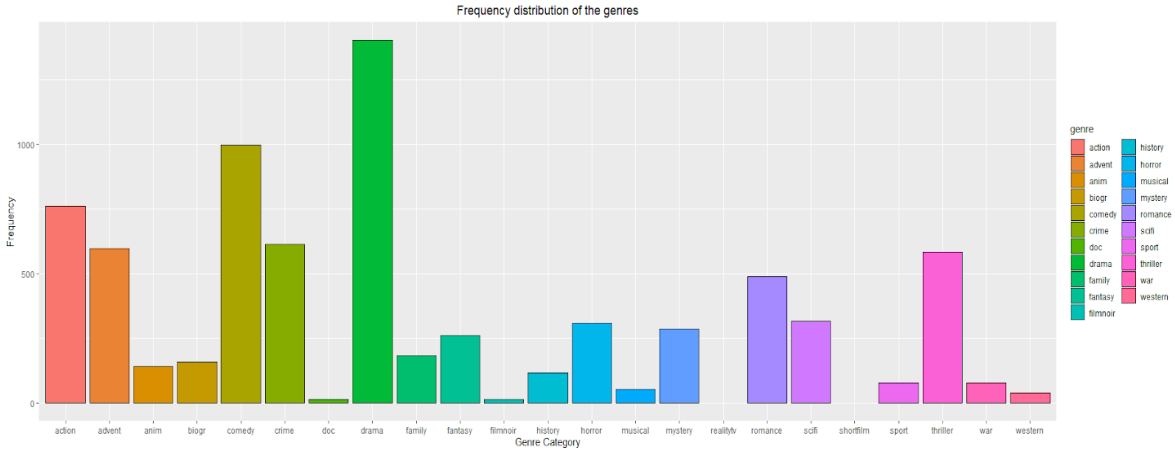

Figure 1: Frequency distribution of the genres.

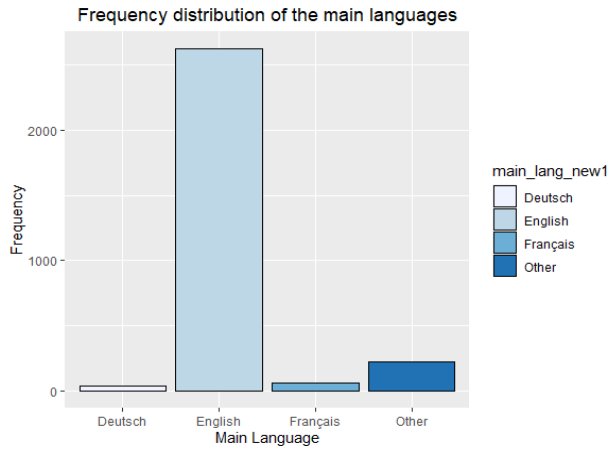

Figure 2: Frequency distribution of the Main Language

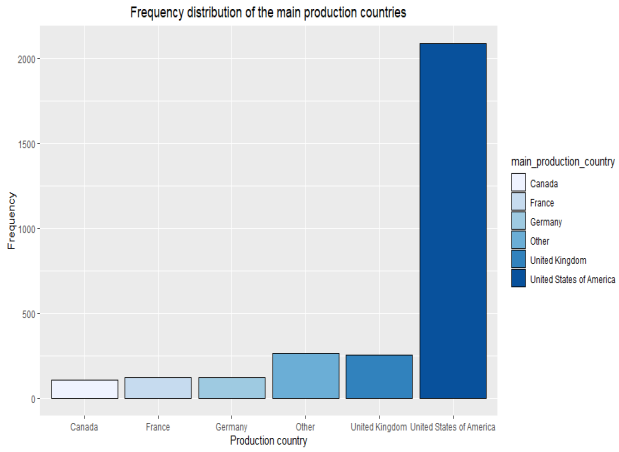

Figure 3: Frequency distribution of the production country

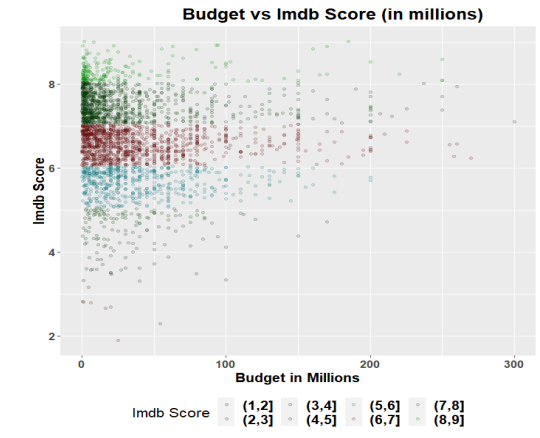

Figure 4: Frequency distribution of Budget (millions) wrt Rating

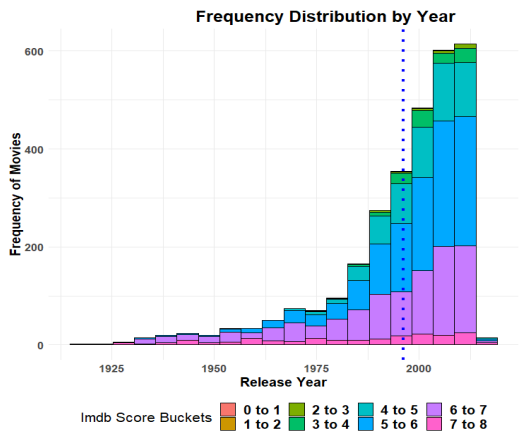

Figure 5: Frequency distribution of Release Year wrt Rating

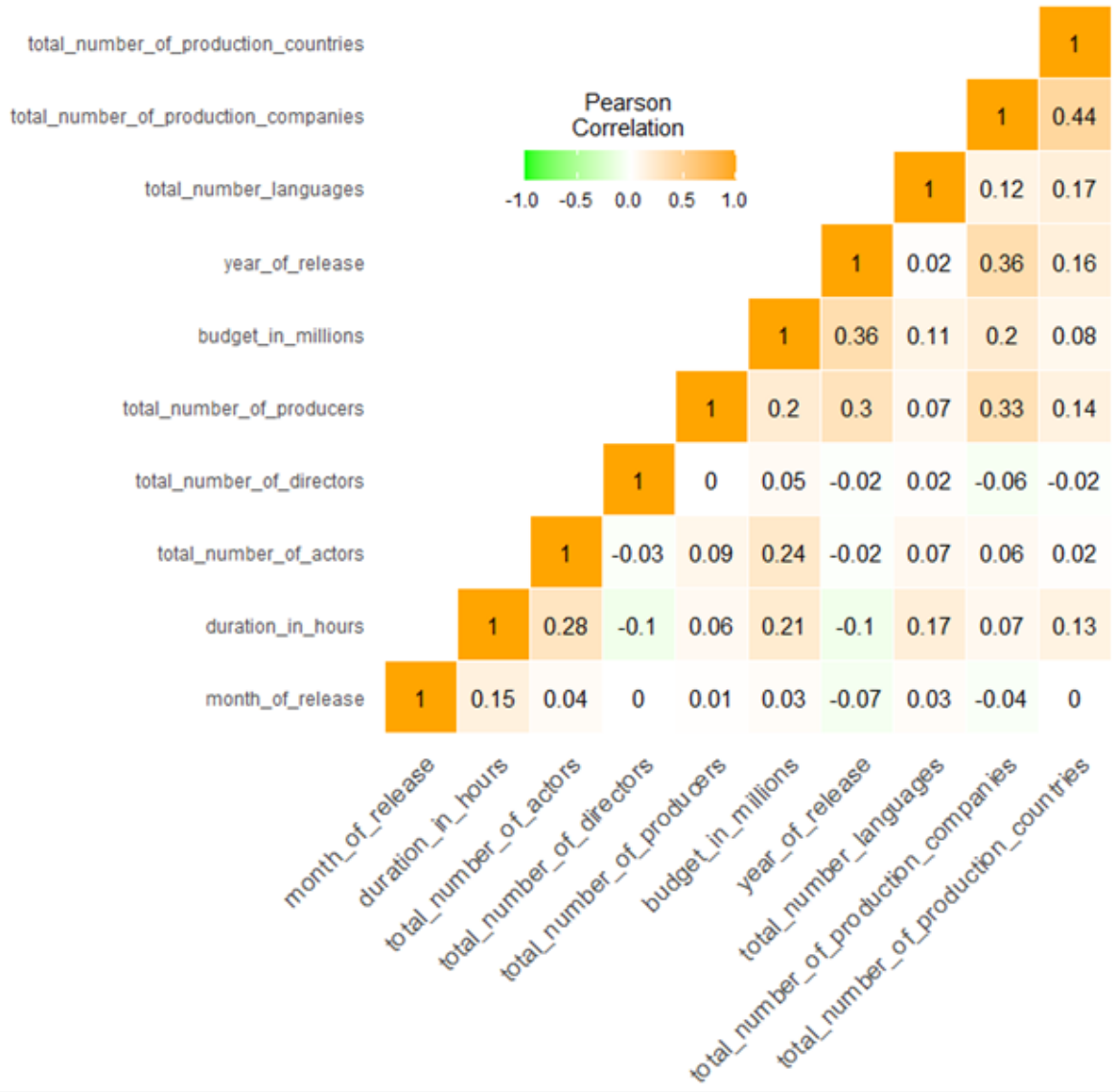

Figure 6: Correlation Heatmap for predictors

Model Issues Identification			
Predictor	**Collinearity: Variance Inflation Factors**	**Heteroskedasticity: Non-constant Variance Test**	**Non-linearity: Residual Plot P-value**
Budget in Millions	1.315295	0.58678	2.20E-16
Month of Release	1.027988	0.8597	0.8507123
Year of Release	1.393294	1.92E-06	0.0001801
Duration in Hours	1.211702	4.83E-10	1.70E-08
Total Number of Languages	1.050183	0.068304	0.4351951
Total Number of Actors	1.134327	0.00012709	2.47E-08
Total Number of Directors	1.025663	0.76973	0.0338293
Total Number of Producers	1.188529	0.014396	4.68E-03
Total Number of Production Companies	1.259946	0.7394	0.3068942
Total Number of Production Countries	1.274396	0.97452	1.36E-02
			Tukey Test: 9.39E-16

Figure 7: Variance Inflation factor scores to detect collinearity of predictors and Residual plot p-values to detect non-linearity of predictors and the entire model

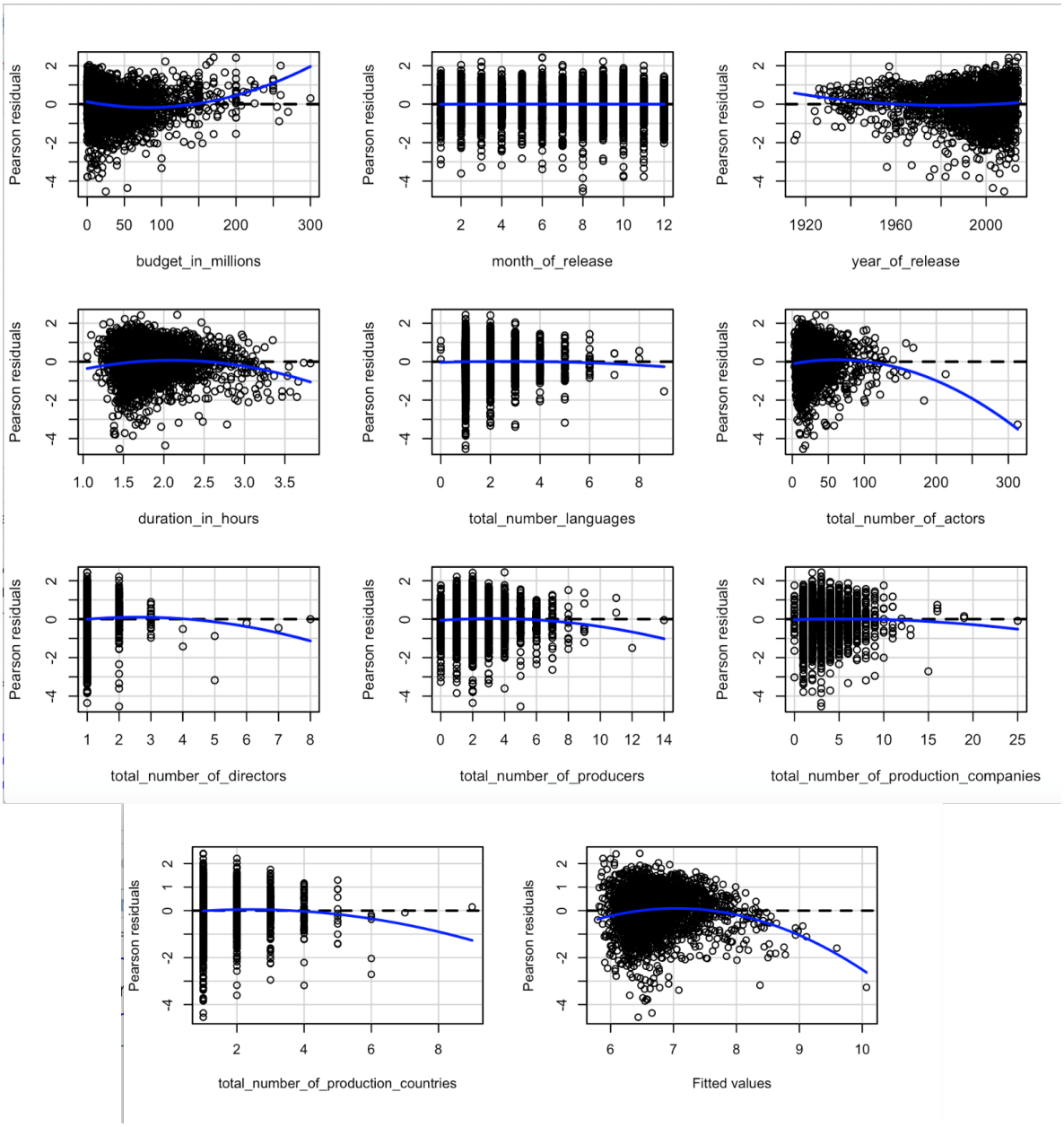

Figure 8: Residual plots to test Non-Linearity of predictors

Predictor	Highest R^2	Validation set test	K-fold Cross-Validation (k=54)	Choice
Budget (millions)	degree: 16 R^2: 0.05943	degree: 10 min MSE: 0.8463	degree: 4 min MSE: 0.8853	polynomial d=4
Release Year	degree: 16 R^2: 0.09559	degree: 16 min MSE: 0.8131	degree: 3 min MSE: 0.8489	polynomial d=3
Duration (hours)	degree: 16 R^2: 0.1505	degree: 10 min MSE: 0.7630	degree: 6 min MSE: 0.7976	polynomial d=6
Total # Actors	degree: 13 R^2: 0.05940	degree: 10 min MSE: 0.8456	degree: 2 min MSE:0.8849	polynomial d=2
Total # Directors	degree: 7 R^2: 0.00118	degree: 3 min MSE: 0.9010	degree: 2 min MSE: 0.9328	polynomial d=2
Total # Producers	degree: 12 R^2: 0.01609	degree: 11 min MSE: 0.8867	degree: 3 min MSE: 0.9293	polynomial d=3
Total # Production Companies	degree: 15 R^2: 0.00958	degree: 12 min MSE: 0.8955	degree: 2 min MSE: 0.9328	polynomial d=2
Total # Production Countries	degree: 7 R^2: 0.00384	degree: 7 min MSE: 0.8996	degree: 3 min MSE: 0.9324	polynomial d=3

Figure 9: Exploration of polynomial degrees for non-linear predictors

ANOVA TEST	
Degree 3 compared with Degree 2	0.004398 **
Degree 4 compared with Degree 3	1.604e-06 ***
Degree 5 compared with Degree 4	0.076095
Degree 6 compared with Degree 5	0.095478

Figure 10: ANOVA test output of comparing polynomial degrees of 'Duration of hours' predictor.

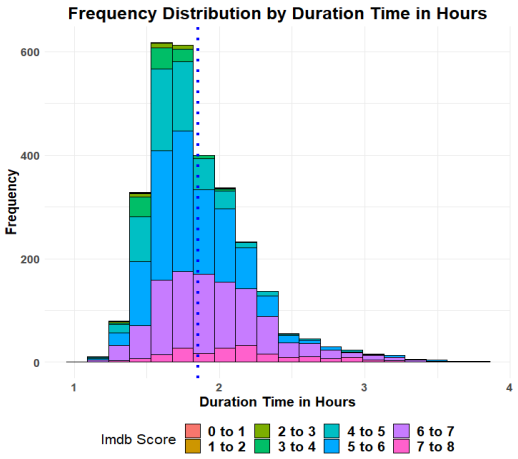

Figure 11: Frequency distribution of Duration (hours) wrt Rating

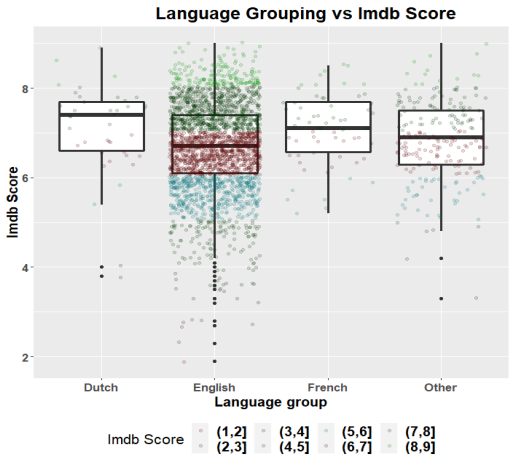

Figure 12: Rating vs Language group

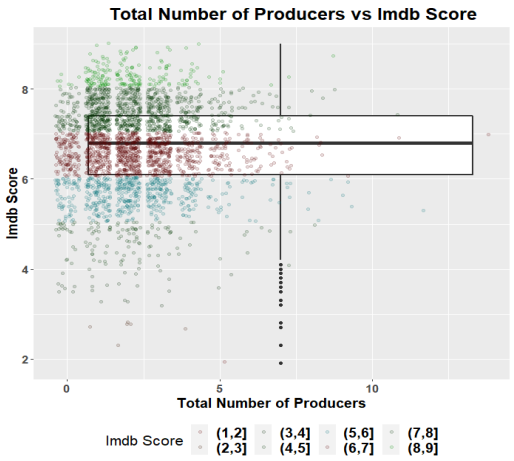

Figure 13: Rating vs Total No of Producers

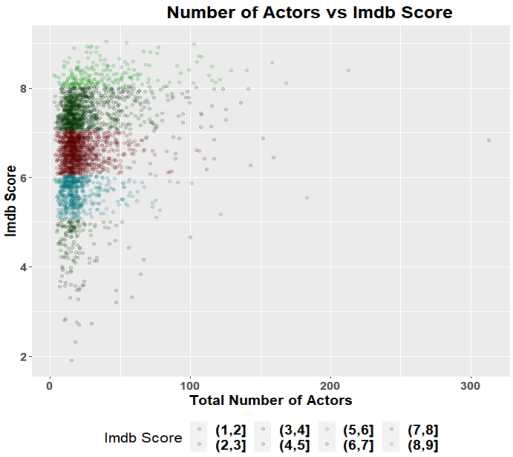

Figure 14: Rating vs Total No of Actors

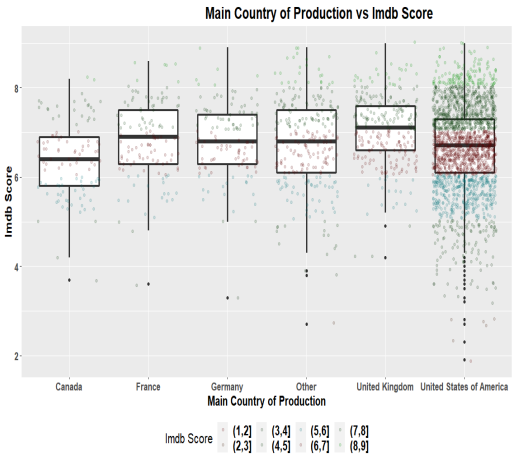

Figure 15: Rating vs Main Country of Production

BONFERRONI OUTLIER TEST	
Observation Number	Bonferroni P-value
633	8.21E-05
895	1.11E-04
2310	3.46E-03
2045	4.20E-03
526	5.12E-03
2718	7.06E-03

Figure 16: Bonferroni Outlier Test p-values to detect outliers of predictors

Observations	2,947
R^2	0.45
Adjusted R^2	0.43
Residual Std. Error	0.71 (df = 2879)
F Statistic	34.47^{***} (df = 67; 2879)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 17: Final Regression results

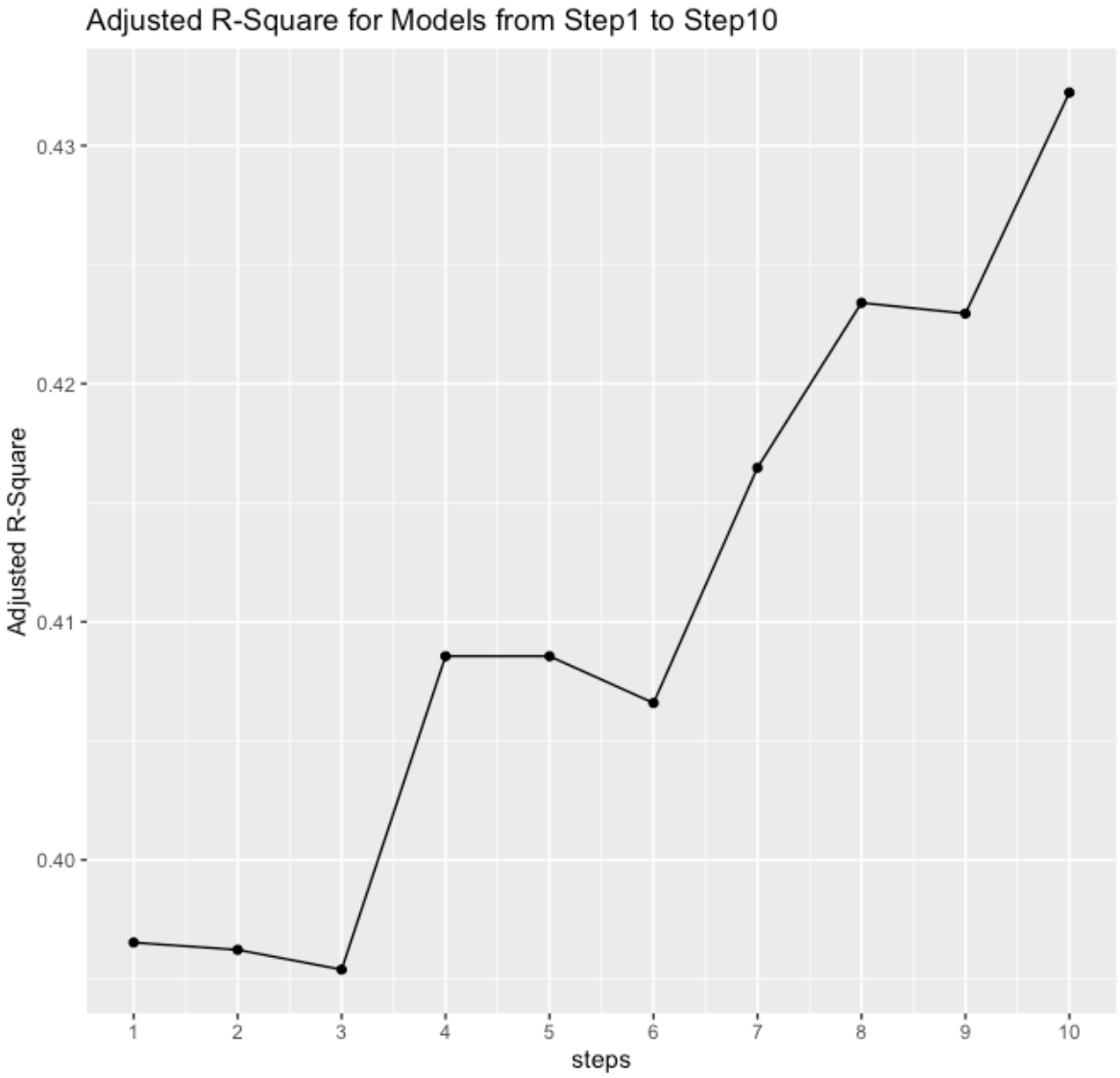

Figure 18: Feature selection graph demonstrating how the adjusted r-squared increases as number of effective predictors increases from step1 to step10