Summary

A Introduction

IMDb (Internet Movie Database) is a search-able online database that is designed to help users decide what to watch based on a rating system. Overtime, IMDb scores have become a significant factor for viewers deciding on a movie. On IMDb, registered users can cast a vote from one to ten for any title in the database. It takes all the individual votes for a given title and use them to calculate a single rating. This rating is not the arithmetic mean but rather the weighted average of votes. This means that all votes have a different impact on the final rating and each vote is assigned a different weight. The algorithm to calculate the weighted average is not publicly disclosed. The IMDb score takes into consideration many factors including film, cast and production characteristics. The primary goal of this project is to build an accurate model with a good out-of-sample performance which is able to predict the IMDb score of movies. To do so, a data-set of 2995 movies along with 51 characteristics of these movies was analysed. The data set was altered on the basis of characteristics' significance and variable type: numerical and binary variables were directly used while categorical variables were either manipulated or removed. This report outlines the rationale and methodology that was adopted to build an accurate prediction model and the insights that were gained from the results.

B Data Description

B.1 Data Preprocessing

In order to build the IMDb rating prediction model, the data-set obtained from IMDb was analyzed. The variables which were both continuous and categorical were either manipulated or removed, and their significance was evaluated. Variables including title, IMDb ID, IMDb URL, production country, main actor 3's name, main actor 3 is female were removed from the data-frame because they were deemed irrelevant in predicting the ratings. Then the categorical variables such as month of release was dummified. Since other categorical variables including main language, main actor 1, main actor 2, main director, main production company have over 100 or even 1000 unique categories,it would be computationally expensive and impractical to dummify all these categories. Hence, analysis was performed on these variables to come up with an approach to dummify them and the rationale and methodology to dummify the categories is as follows:

Main language: The languages were grouped to see if there were any popular languages. English, French, Deutsch and Español were the main languages for almost 94% of the movies; hence, these four categories of languages were dummified and the rest were classified as others.

Main actor 1, Main actor 2, Main director, Production Company : To clearly see the effect of each main actor, director and production company on the IMDb movie score, three different lists from the IMDb website were imported which represented the top 200 actors, top 25 directors and top 14 production companies. This helped dummifying the corresponding predictors into famous_main_actor1, famous_main_actor_2,famous_director and famous_production_company. For example, if the main actor of a specific movie is in the imported list, the value of the famous_main_actor1 predictor will be 1.

Year of release: The variable was dummified on the basis of decades since the trend of movies usually change in decade as opposed to yearly. After these variables were dummified, all the original variables were dropped from the data-frame.

B.2 Feature Selection

Feature selection is a very important aspect of modeling that is used for removing correlated variables, biases and unwanted noise. To create an accurate model, feature selection was done using two different packages, the first being Boruta package in R and the second being Earth package. Boruta was chosen as the feature selection algorithm because it captures all features which can be relevant to the target variable as opposed to traditional feature selection that mostly rely on a small subset of features that can yield a minimal error on a chosen classifier. The EARTH package uses Multi-variable Adaptive Regression Spline model which is based on Random Forest where the best predictors are chosen because they minimize the error in the model. It was evident from running both of these packages that some predictors have little to no effect on the IMDb score of the movie. As can be seen in figure 1 in Appendix, there are certain predictors including sports genre, short films, gender of director, and many more have negligible importance. Therefore, these predictors were dropped. It was also noted that Deutsch and Espanol languages have no significance on the model, so they were added to a third category of languages 'others'.

B.3 Distribution of Predictors

Prior to analyzing the relationship between each predictor and the target variable, it is important to analyze the distribution of all predictors to further understand their nature. To make it easier to visualize this description, box plots were created for both numerical and binary variables, and histograms were created for numerical variables only. For this data-set, histograms for binary variables are not as useful because there are only two categories. In addition to these visual presentations, the skewness test was also performed to have a more accurate understanding of the variables. Using a built in package in R called 'e107', the skewness of each predictor was determined along with a description of the intensity of the skewness and the results are shown in figure 6 in Appendix. Variables that are highly skewed have skewness value greater than **1** or less than **-1**. These results are further evident from histograms displayed in figures 2 & 3, which illustrates that all these variables are either right or left skewed.The high skewness can result in misleading coefficients and intercept values for regression models and can also lead to over-fitting and therefore it must be corrected. For this data-set, it was seen that even after correcting skewness of highly correlated variables by using log transformation, the mean squared error (MSE) was not reduced and therefore the variables were taken as it is. It is also evident from the histograms that in the case of high skewness, there are outliers present, which will be confirmed later by the outlier test.

The distribution and statistics of the predictors are presented in figures 2 & 3 in Appendix and are summarized below:

B.3.1 Continuous Variables:

- **Budget (in millions):** The budget of the movies is right skewed with the average budget of a movie being 35 Millions. Additionally, 75% of the movies have budget of less than 48 Millions.
- **Duration (in Hours):** The variable is right skewed with 75% of movies having duration of less than 2.1 hours. The average duration of a movie is 1.8 hours.
- **Number of Actors:** The total number of actors is highly skewed towards the right. The movies has an average of 24 total number of actors while 75% of the movies have 28 total actors.
- **Number of Production Companies:** On an average, there are three total number of production companies for movies. 75% of the movies have less than four total number of production companies and the variable is right skewed.
- **Number of Production Countries:** On an average, the movies are shot in one country only while 75% of the movies have been shot in less than two countries. The variable is also highly skewed towards the right.
- **Number of Producers:** The variable is highly skewed towards the right with movies having an average of two producers and 75% of the movies having less than three producers.
- **Number of Directors:** The variable is highly skewed towards the right with movies having an average of one director and 75% of the movies having less than one director.
- **Number of Languages:**On an average, the movies have around two total number of languages and the variable is right skewed.

  Additionally, it can be seen from the histograms that the average IMDB score is between 6 and 8.

B.3.2 Categorical Variables:

- **Genre:** There are total 24 different genres with approximately 47% of movies falling under drama genre, 34% under comedy, 26% under action, 21% under crime, and 20% under thriller and adventure.
- **Main Actor 1 and 2 and Director being male or female:** There are approximately 22% of the movies which have female as main actor 1 while there are 41% of movies with female main actor 2. Around 4% of the movies have female director.

C Model Selection

C.1 Model Issues

C.1.1 Correlation and Multicollinearity

After data preprocessing, the linear model of all the predictors and IMDB rating (excluding the dummified variables) was created. To ensure that the model predicts accurately, it is essential to check that the predictors are not biasing the results. Hence, four key problems including non-linearity of predictors, heteroskedasticity, presence of outliers and collinearity were tested for each model.

- **Collinearity:** When two or more quantitative predictors are highly correlated to each other, they can lead to collinearity, which can impact the significance of these predictors individually. Hence, a correlation heat map displayed in figure 4 in Appendix and Variance Inflation Factors (VIF) matrix between the predictors were created to test for collinearity. As a rule of thumb, collinearity is problematic if the absolute value of correlation coefficient is greater than 0.8 or VIF is greater than 4. None of the predictors met this criteria ; hence, collinarity was ruled out as one of the model issues. Next, the correlation between the IMDb score and all numerical predictors was determined to ensure the absence of high correlation / multi-collinearity. Based on the results, no high correlation was found.

- **Outliers:** Outliers are observations that lie in an abnormal distance from other observations, which can significantly impact the coefficients and significance of results. Bonferroni test was conducted on a simple multiple linear regression between the IMDb score and all the predictors to detect the outliers. The test results showed six outliers, which were removed from the sample.

- **Heteroskedasticity:** One of the other model issues include heteroskedasticity, which implies non-constant variance and can result in misleading values for t-statistic and p-value. Non-Constant-Variance(NCV) test was performed for each numerical variable to observe if the p-values were less than 0.05, which implies the presence of heteroskedasticity in the model. This was also evident from residual plots where some of the predictors displayed a funnel shape as shown in figure 5 in the Appendix. The predictors that showed heteroskedasticity:, such as duration (hours) and total number of actors were then corrected using Sandwich Correction.

- **Non-linearity test:** Residual Plots were plotted to test the linearity assumption of each numerical variable and the model. If p value was less than 0.05 for the variable or model, it implied that it did not satisfy the linearity assumption. For the variables that did not pass this linearity test, different degrees were tested to see which one gave the least mean squared error. The results showed that the only predictor that could be assumed linear is the total number of languages, whereas all other predictors including budget, total number of actors and duration failed the linearity assumption since they had p value of less than 0.05. The non linearity of these predictors in evident in Figure 5 of Appendix since there is curved pattern to the residual plots.

C.2 Model Building

After addressing all the issues, different models were created using different approaches to build the optimal model. The first step in building the model was to explore the relationship between the target variable (IMDb score) and eight numerical variables. After running a regression of each predictor versus the target variable, the predictors which had a p-value of greater than 0.05 were removed as they were considered insignificant. These predictors included total number of directors, total number of producers, total number of production companies and total numbers of production countries. Following that, numerical variables with a nonlinear relationship with the target variable, which included budget(millions), duration(hours) and total number of actors were analysed. To find the right polynomial degrees for these variables, three different methods were used:

- First method was to find the best polynomial degree by running the ANOVA test on each predictor individually while varying the degree.
- Second method was through running nested loops on different polynomial degrees and applying cross validation. For each model, 50 different cross validation tests were performed in order to obtain an accurate performance score, which is the average of the fifty MSE values that were obtained. For each iteration, the dataset was split into training and test data sets at a test ratio of 0.3.
- Final method was through running nested loops over the polynomial degrees of the predictors using K-fold test with 10 folds. A similar process as the previous method was followed where the process was repeated 50 times and the results provided the optimal combination of polynomial degrees with the lowest MSE.

The methods included all significant binary variables, along with the variable, total number of languages that satisfied the linearity assumption. The mean squared error of the best polynomial degrees resulting from each method was found through both cross-validation and K-fold tests. After comparing the results, both methods had the same polynomial degrees. The optimal combination of polynomial degrees obtained is as follows:

- **Budget:** 4^{th} degree
- **Total number of languages:** 1^{st} degree
- **Total number of actors:** 2^{nd} degree
- **Duration:** 4^{th} degree

Upon further research, it was concluded that despite giving the same result as cross-validation loop, using K-fold to find the optimal degree was not the best method, but rather an effective method to find MSE. K-fold method generally gives more accurate results as compared to cross validation. Leave one out cross validation (LOOCV) test was not performed for computational reasons. The best fit model was found to have an MSE equal to **0.5476347**. In order to further expand the scope in finding the best model, a spline regression model was also created using for loops that iterate between the number of knots (two, three) and the corresponding polynomial degrees (2 to 5), and determine the optimal combination of knots and degree using K-fold test. The spline model had an MSE of **0.571**.

Since the previously found optimal model using k-fold test has better performance than spline, the spline model was not considered as the final model. To summarize, the final model has the following set of predictors and degrees:

```
Budget (4th degree), Number of actors( 2nd degree), Duration (4th Degree),
total_number_of_languages, genre_action, genre_adventure, genre_animation, genre_comedy,
genre_documentary, genre_drama, genre_family, genre_history, genre_horror,
genre_romance, main_actor1_is_female, main_actor2_is_female,
famous_main_actor_2, famous_director, English,French,
years_of_release( all decades from 1910s to 2010s), months_of_release(all months from 1 to 11)
```

D Managerial Implications

Based on the developed model with the given data set, it can be inferred that the most important predictors for the IMDB score include movie budget, duration of the movie, genres, total number of actors and the popularity of the main actors and director. In addition, the gender of the main actors and the movie era also have significant impact on the IMDB score. Out of the the important predictors, the following conclusions can be drawn:

- **Movie Budget:** It was observed that movies with the budget of $0.58 million tend to have a lower IMDB score by almost 1.79 than average. The break-even point for movie budget occurs at $1.14 million i.e. At this point, the movie budget has no impact on average IMDB score. However as the budget increases more than $1.14 million, the IMDB score tends to increase considerably in a non-linear fashion.

- **Total number of actors:** It was observed that as the total number of actors increase, the IMDB score tends to go down in a quadratic fashion. Thus, it can be inferred with almost 100% probability that movies with lower number of actors, tend to have a better IMDB score.

- **Duration of the movie:** It was observed that movies having a time duration of 1 hour 41 minutes, tend to get a lower IMDB score while at 2 hrs 33 min the duration of movie has no effect on the average IMDB score. However, as the duration of movie increases more than 2 hrs 33 min, the IMDB score tends to increase considerably in a non-linear fashion.

- **Genre:** The most significant genres which contributed to high IMDB score were found to be Documentary, Animation and Drama which accounted for an increase in the IMDB score by 0.763, 0.746 and 0.198 higher than the average, respectively. Genres such as Horror, Action, Comedy and Romance tend to bring down the IMDB score by almost 0.47, 0.28, 0.21 and 0.11 respectively.

- **Popularity of director, actors and their gender:** While it can be said with almost 100% probability that a movie directed by a famous director will have a higher IMDB score by 0.395, the positive effect of top 2 main actors' popularity of around +0.1 can be claimed with roughly 95%

probability. A possible explanation could be that famous directors tend to have a unique style of direction and art, while popular actors tend to play main roles in a wide range of movie genres. The IMDB score also tends to be higher by 0.24 when the first main actor is male and by 0.11 when the second main actor is male.

- **Movie era:** It was observed that movies from the 1920s had a higher rating of 1.2 than average, followed by 1930s (+0.65) and 1950s (+0.44). However, newer movies from the 1990s tend to have a lower rating as compared to older era movies. A possible explanation could be the "classic" status of older movies, as compared to newer movies.

Thus, it can be inferred that movies with higher IMDB score usually have:

1. A budget greater than $1.14 million
2. Fewer number of total actors
3. A duration of over 2hrs 33 min
4. Belong to the Documentary, Animation and Drama genres
5. Directed by a famous director
6. Includes famous male actors as top 2 main actors
7. An older era movie, typically older than 1990s era

Therefore, if a producer is looking to produce a movie with a higher IMDB score, it is recommended to produce an animated version of a 1920s documentary/drama genre movie which has a budget of over $1.14 million, is 2.5 hours long, has 2 famous male actors in main roles and is directed by a famous director. Coincidentally, a documentary called "World War 1 in Color" satisfies the above and has an impressive IMDB score of 8.2.

D.1 Conclusion

The goal of this project was to build a model that can accurately predict the IMDb score of movies. Different models were tested to finalize a model that gave the best out-of-sample performance with the lowest mean squared error. The mean squared error of the final model was determined to be 0.5476347. The final model took into consideration the issue of both overfitting and underfitting. Additionally, the analysis was able to provide meaningful insights on the characteristics that play a significant role in predicting the IMDB score of a movie.

Appendices

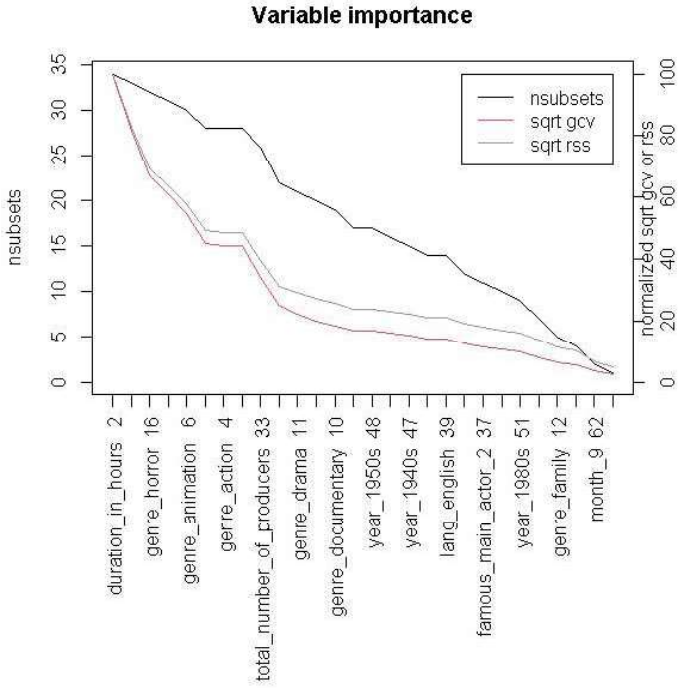

Figure 1: Significance Of Predictors (1)

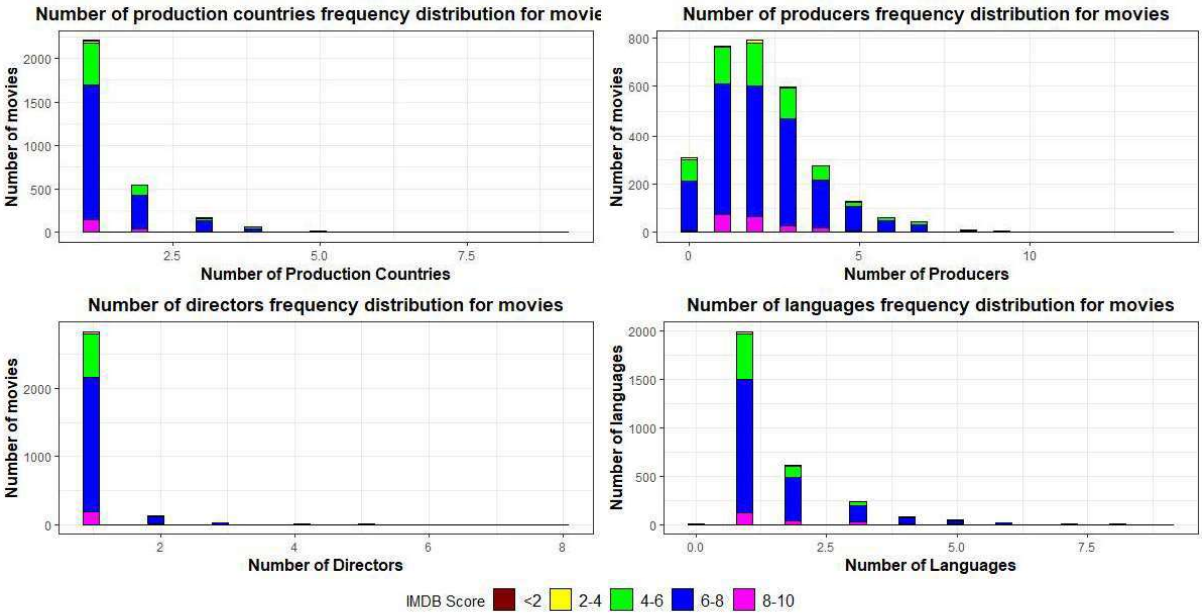

Figure 2: Histogram for total number of numerical variables - Part 1

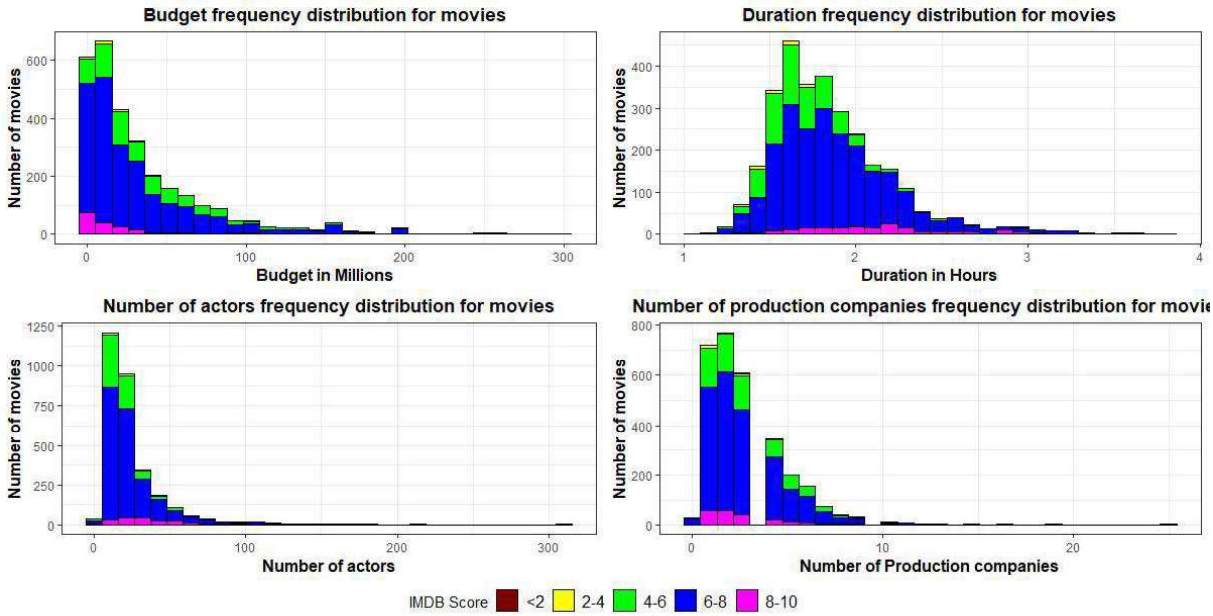

Figure 3: Histogram for total number of numerical variables - Part 2

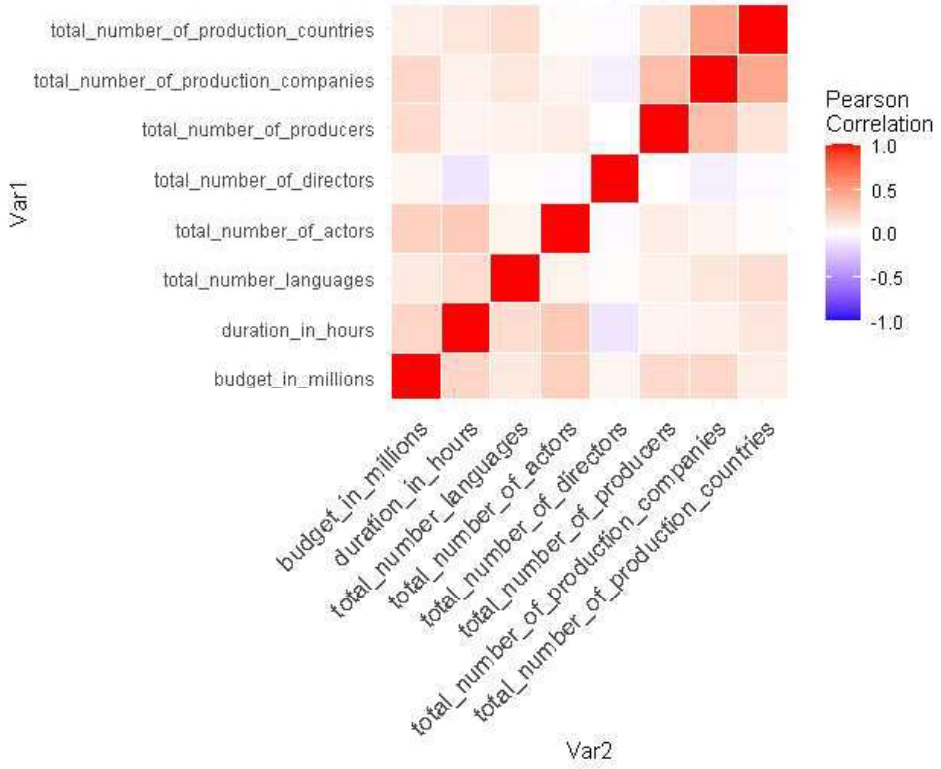

Figure 4: Correlation Matrix

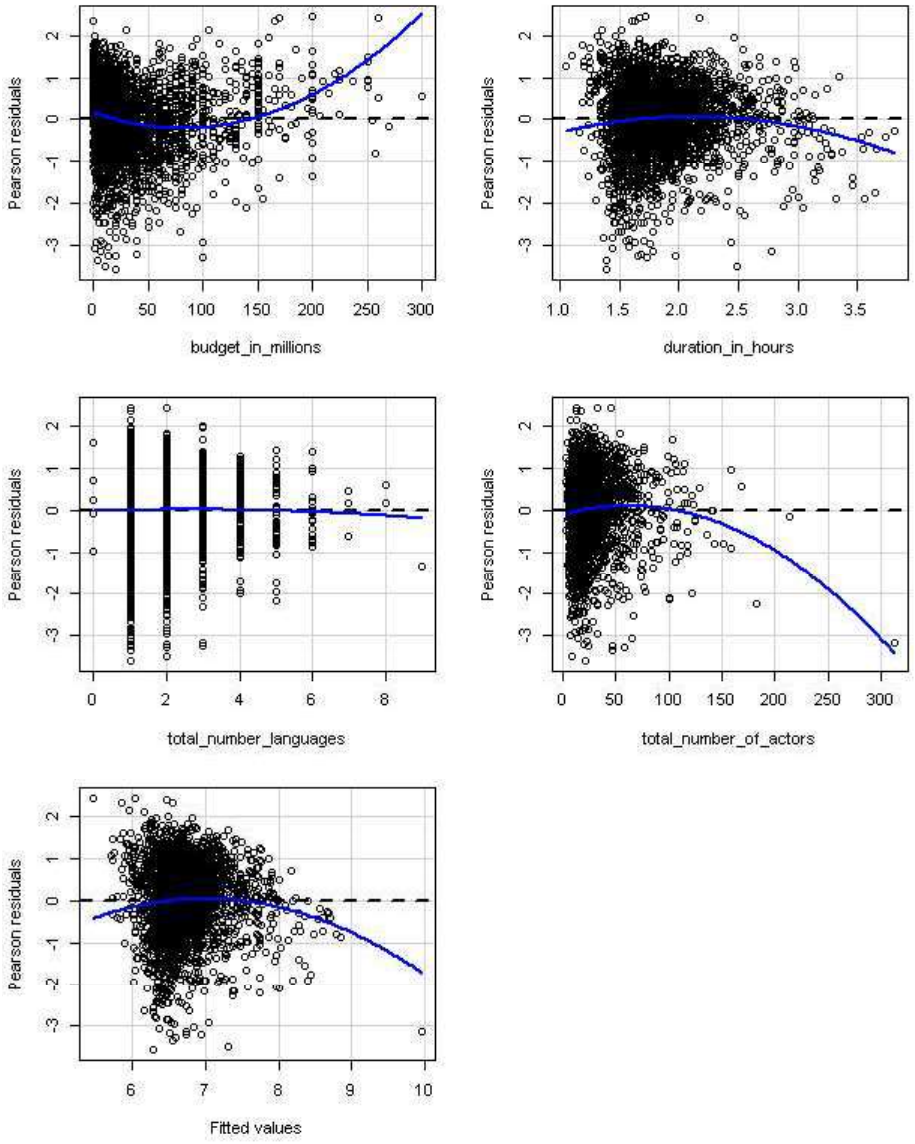

Figure 5: Residual Plots

Variables	Skewness	Type
budget_in_millions	2.221807	Highly Skewed
duration_in_hours	1.438646	Highly Skewed
total_number_of_languages	2.366989	Highly Skewed
total_number_of_actors	3.706510	Highly Skewed
total_number_of_directors	8.833900	Highly Skewed
total_number_of_producers	1.193303	Highly Skewed
total_number_of_production_companies	2.112489	Highly Skewed
total_number_of_production_countries	2.657892	Highly Skewed

Figure 6: Skewness

Results

	Dependent variable:
	imdb_score
poly(budget_in_millions, 4)1	-6.734***
	(1.123)
poly(budget_in_millions, 4)2	7.696***
	(0.856)
poly(budget_in_millions, 4)3	-3.355***
	(0.801)
poly(budget_in_millions, 4)4	1.565**
	(0.784)
poly(total_number_of_actors, 2)1	6.532***
	(0.817)
poly(total_number_of_actors, 2)2	-4.784***
	(0.763)
total_number_languages	0.017
	(0.019)
poly(duration_in_hours, 4)1	10.812***
	(1.020)
poly(duration_in_hours, 4)2	-3.731***
	(0.815)
poly(duration_in_hours, 4)3	-0.301
	(0.779)
poly(duration_in_hours, 4)4	1.337*
	(0.766)
genre_action	-0.281***
	(0.047)
genre_adventure	-0.062
	(0.052)
genre_animation	0.746***
	(0.093)
genre_biography	0.057
	(0.082)
genre_comedy	-0.207***
	(0.048)
genre_crime	-0.018
	(0.047)
genre_documentary	0.763***
	(0.272)
genre_drama	0.198***
	(0.046)
genre_family	-0.125*
	(0.076)
genre_fantasy	-0.063
	(0.065)
genre_filmnoir	0.197
	(0.243)
genre_history	-0.155
	(0.094)
genre_horror	-0.470***
	(0.067)
genre_musical	0.218*
	(0.124)
genre_romance	-0.108**
	(0.051)
genre_scifi	0.132**
	(0.059)
genre_thriller	0.002
	(0.051)
main_actor1_is_female	-0.247***
	(0.043)
main_actor2_is_female	-0.117***
	(0.035)
famous_main_actor_1	0.098**
	(0.038)
famous_main_actor_2	0.089**
	(0.043)
famous_director	0.395***
	(0.064)
famous_production_company	0.088**
	(0.042)
lang_english	-0.041
	(0.062)
lang_french	0.362***
	(0.129)
year_1920s	1.207***
	(0.380)
year_1930s	0.653***
	(0.164)
year_1940s	0.333**
	(0.162)
year_1950s	0.441***
	(0.135)
month_2	-0.107
	(0.081)
month_3	-0.100
	(0.077)
month_4	-0.096
	(0.081)
month_5	-0.033
	(0.077)
month_6	-0.038
	(0.073)
month_7	-0.007
	(0.077)
month_8	-0.066
	(0.077)
month_9	0.014
	(0.075)
month_10	0.014
	(0.074)
month_11	-0.047
	(0.076)
Constant	6.781***
	(0.118)
Observations	2,093
R^2	0.411
Adjusted R^2	0.395
Residual Std. Error	0.739 (df = 2035)
F Statistic	24.963*** (df = 57; 2035)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 7: Summary of Best Fit Model

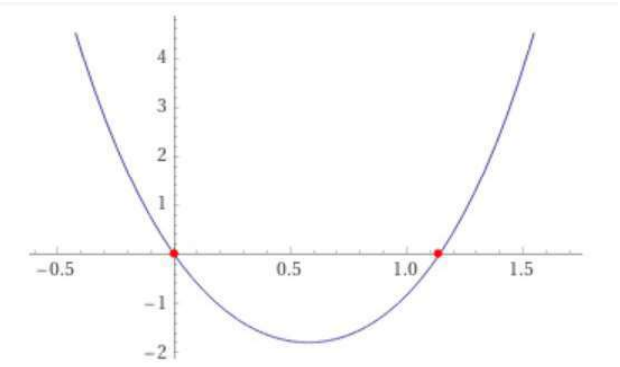

Figure 8: Equation of budget variable

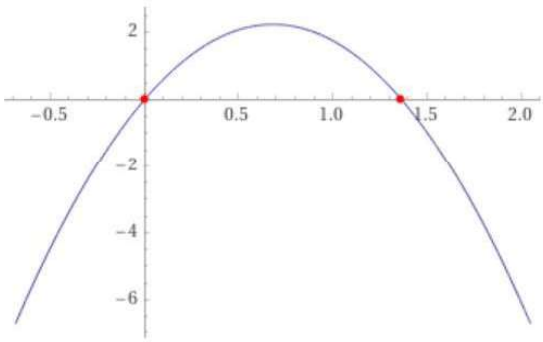

Figure 9: Equation of number of actors variable

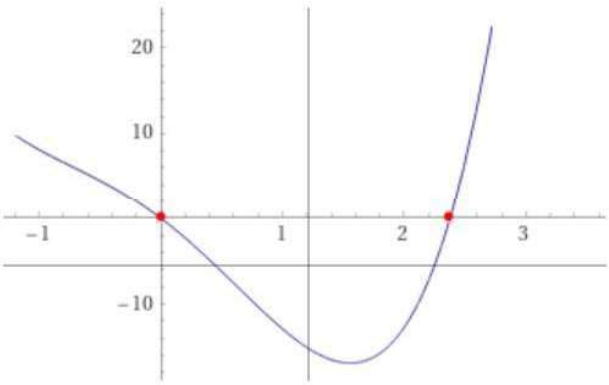

Figure 10: Equation of duration variable