https://www.overleaf.com/project/5faae20a4d885282da101e03

MULTIVARIATE STATISTICAL ANALYSIS FOR MACHINE LEARNING

PROJECT:
THE 2020 IMDb PREDICTION CHALLENGE

Contents

1. Introduction

Movies are a form of art where its creators express their ideas and feelings through fiction or non-fiction, and as all expressive arts, their likeability depends, mostly, on their viewers' interpretations. However, apart from these deviations in likings, some movies are generally loved or hated unanimously which leads to the following questions:

- What really makes a movie a success or a disaster?
- Can I hack the system and tailor the perfect movie for my target audience? If so, what movie should I produce to get high ratings?

Using IMDb's movie database, the following report seeks to answer these questions through regression analysis on IMDb's user rating score. This score is completely based on user reviews which, combined with IMDb's high user engagement, leads to a score that should represent the true rating of these films.

Our analysis concludes that movie ratings are in fact correlated to specific factors like movie length, certain genres, and quality (directors, production companies, budget size) lead to better or worse movies from an audience perspective[1]. Given this, we have developed a predictive model with an MSE equal to 0.5403 points when tested.[2] This leads us to safely suggest that the specific mix of factors, when combined accordingly, may lead to higher ratings in movies.[3]

Finally, we believe the predictive accuracy of our model can be increased if additional data on movies is captured. We consider the movie plot is one of the key factors that make a movie successful and this type of information cannot be captured with the available IDMB database. For future analysis, we encourage using text analytics to extract specific details on the movie plot from its summary as successful patterns may be found.

[1] Refer to Section 2: Data Description

[2] Refer to Section 3: Model Selection

[3] Refer to Section 4: Managerial Implications

2. Data Description

2.1. Dataset Overview

The movie dataset contains 51 variables for 2995 movies, spanning across 100 years in 46 countries. The dataset includes a variety of variables providing information on the genre of movie, language, actors acted in the movie, details about director and editors.

In this analysis, “IMDb Score” is the target variable while others are the possible predictors. However, based on the model, we are focusing on only the predictors which have the highest predictive power in determining the IMDb score.

2.2. Target variable: IMDb score

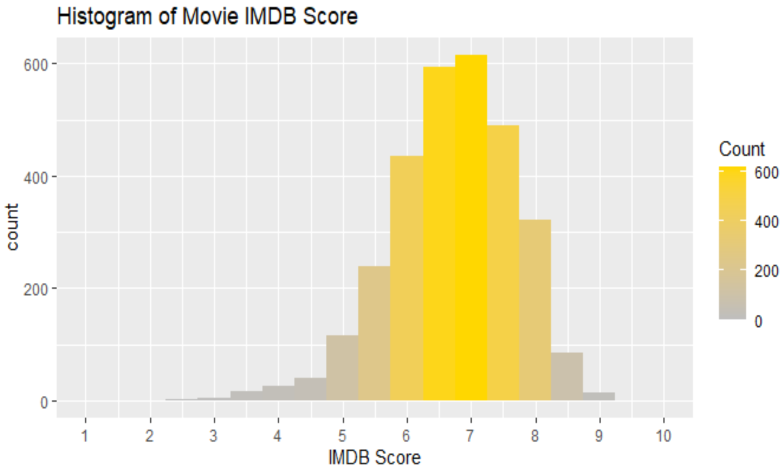
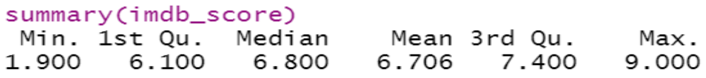


Figure 2: Histogram of Movie IMDb score



```
summary(imdb_score)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.900   6.100   6.800   6.706   7.400   9.000
```

Table 1: Summary of IMDb score

IMDb score is a measure of user rating of films. Here, the score varies from 1.9 to 9 and has a normal distribution with its peak around 7. The mean for the distribution is 6.7 and the median is 6.8 (that means that 50% of movies in the dataset have a score above 6.8).

2.3. Predictors:

The dataset consists of a variety of variables categorized as continuous, categorical, and binary variables.

2.3.1. Continuous Variables:

Among all the integer variables the most significant ones are:

Year of release:

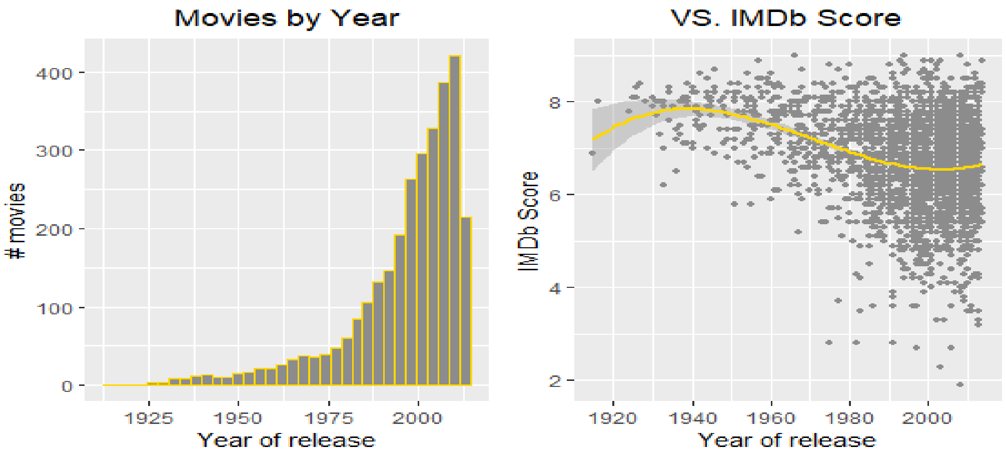


Figure 2: Histogram and Regression plot for Number of Year of Release

The data consists of movies from 1915 to 2014. As the distribution is left-skewed, we can observe that majority of movies in the dataset are released in the 2000s. Also, we can conclude that on average as the year increases, the IMDb score decreases.

Duration in hours:

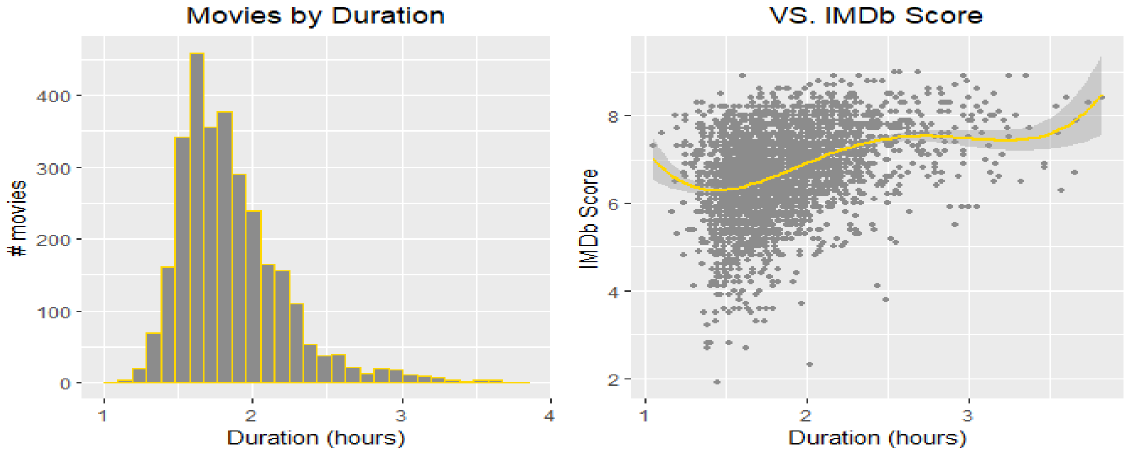


Figure 3: Histogram and Regression plot for Number of Duration of Movies

The distribution for the duration is almost normal ranging between 1.050 and 3.817. The average duration is 1.85 hours (∼ 111 minutes) and 50% of movies have a duration of more than 1.78 hours. On average as the duration of a movie increases, the IMDb score increases.

Total number of actors:

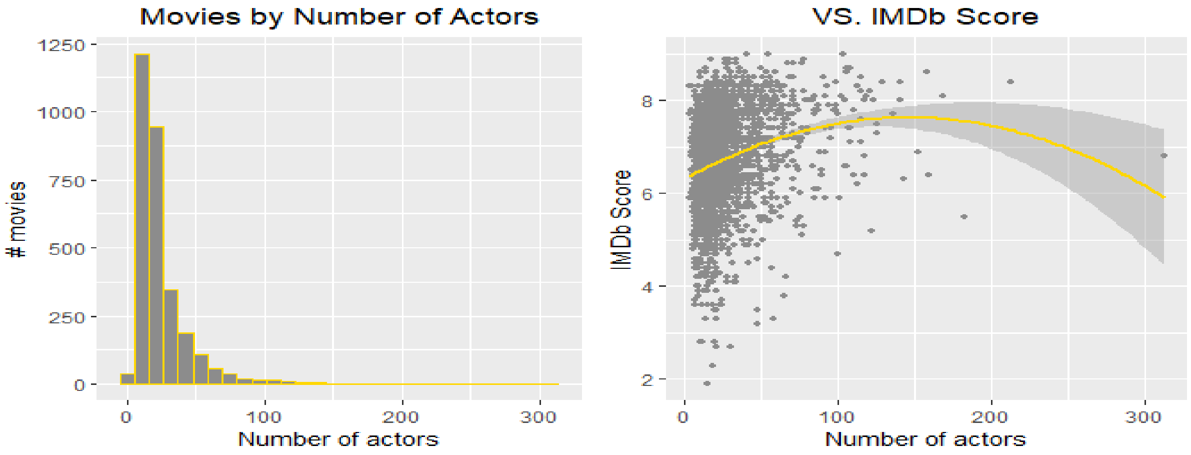


Figure 4: Histogram and Regression plot for Number of Actors

The distribution for the number of actors is right-skewed. On average, most movies have a minimum of 10 actors. From the trend of its relationship we can infer that a higher number of actors lead to better review score, but there is a saturation point were score decreases.

Budget (in Million):

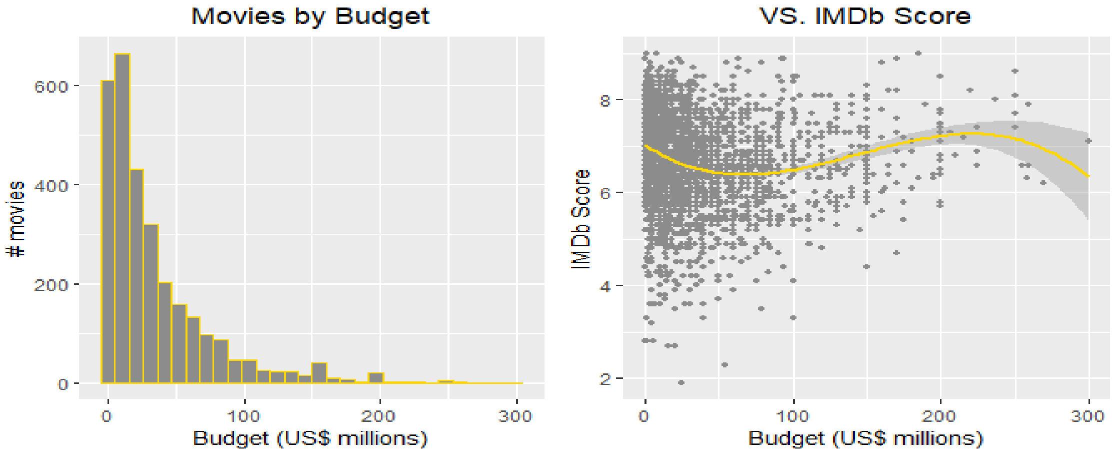


Figure 5: Histogram and Regression plot for Budget

The distribution for the budget is right-skewed. The budget of the movies varies from a few thousand to millions, the majority of movies have a budget of less than 100 $ million. From the trend we can infer that budget has an oscillating effect on the IMDB score as it increases or decrease IMDb score depending on the range.

2.3.2. Binary Variables:

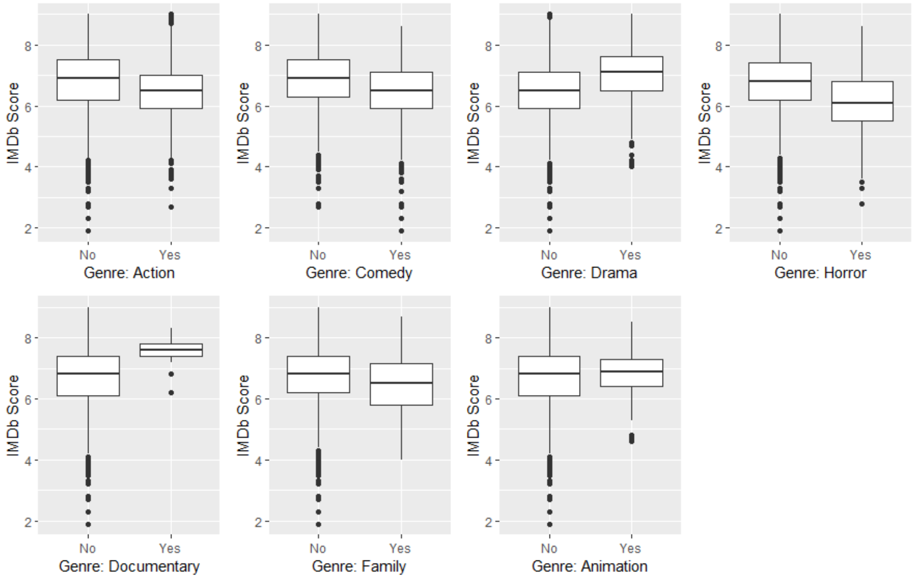


Figure 6: Box-plot for Genres

Another aspect of the data is the Genre that each movie is related to. From all the genres, the most significant genres for prediction are selected and analyzed using box-plot.[4] From the box-plots, we can observe that almost all genres are symmetric i.e. the data for genres are normally distributed and have low variation. We can infer this from the fact that our dataset has a limited number of movies. The box-plot shows that genres such as drama and documentary have a relatively higher score than other genres. While, genres like comedy and horror have relatively lower scores than other genres.

2.3.3. Categorical Variables:

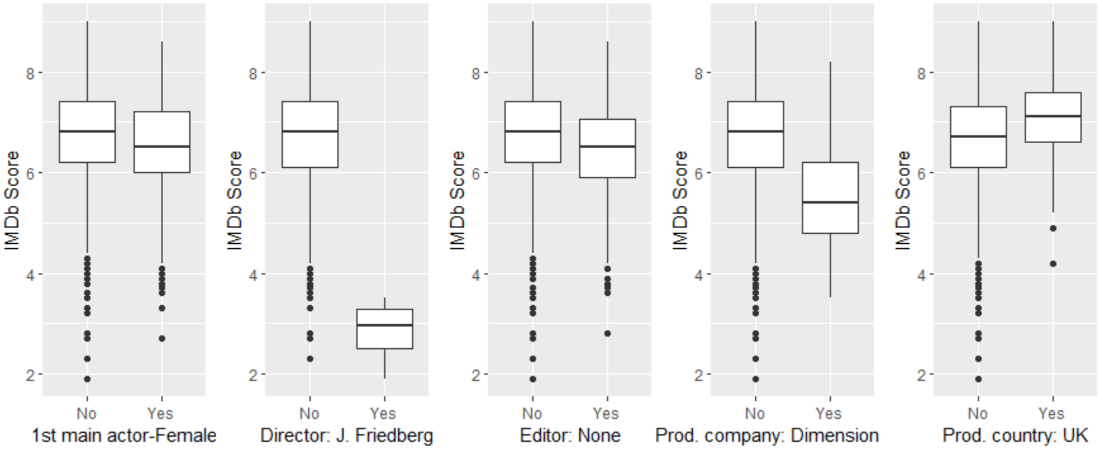


Figure 7: Box-plot for Categorical Variables

[4] Refer to Section 3: Model Selection

The above graph shows the significant categorial variables[5] for our model. It can be observed that the movies directed by J. Friedberg, and movies produced by Dimension Films have a negative impact on the IMDb score. On the contrary, if the movie is produced in the UK, it increases the IMDb score.

2.3.4. Correlation of Predictors variables:

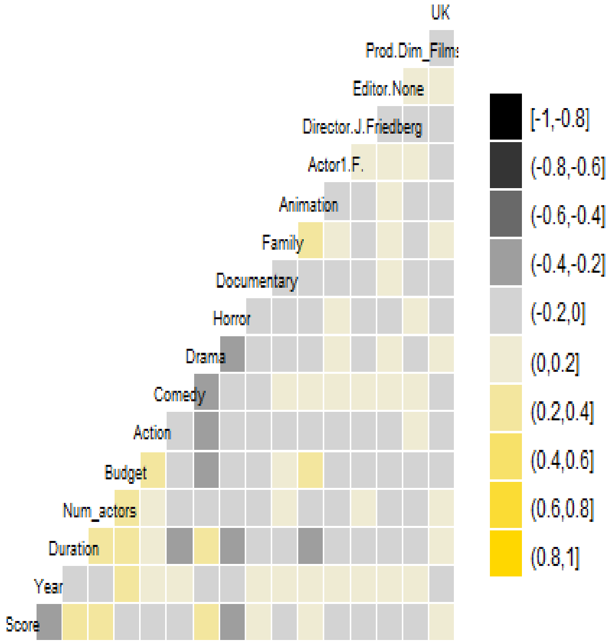


Figure 8: Correlation Matrix

Based on the heatmap, we observe that the predictors chosen are not highly correlated and thus, we are avoiding the effect of double-counting of predictors in the model. Hence, choosing the variables that affect the target variable significantly.

[5] Refer to Section 3: Model Section

3. Model Selection

To develop the final model that modelling process underwent three different phases.

3.1. Phase 1: Data Preparation

Since IMDb ID, IMDb URL, and Title of the movies were not valuable for the model, we decided not to include them in our model. Also, variables like Genre: Reality-TV and Genre: Short Films had only one value (equals to 0) for every movie that would have created noise. Hence, we decided to remove Genre: Reality-TV and Genre: Short Films from our model.

Categorical text variables were then dummified to perform LASSO and correlations. To convert these to dummy variables we used the fastDummies library. Using this library, 9392 dummy variables were created for the predictors- Main Language, Main Actor 1 Name, Main Actor 2 Name, Main Actor 3 Name, Main Director Name, Main Producer Name, Editor Name, Main Production Company and, Main Production Country.

3.2. Phase 2: Feature Selection

Correlation and LASSO were the two main techniques, followed by the hit and trial method for a few variables that lowered the MSE and P-Value of the model.

Firstly, LASSO was applied to the numerical predictors including the numeric categorical variables by keeping alpha(penalty)= 0.55. As a result, it gave 11 variables with a coefficient not equal to zero, hence significant for our model. For the categorical text variables, at alpha=0.09, the results suggested to include 4 variables in the model.

Secondly, Correlation was applied to the same numerical predictors (Continuous + Binary) at corr>=20 that gave 7 significant variables for the model [Table 6][6]. For the categorical text variables, at corr>= 10, it gave 6 significant variables for the model [Table 7][6].

With these variables, new models were tested aimed at bringing the P-Value less than 0.05 for each predictor, decreased MSE and increased adjusted R^2 value with the addition of each variable in the model.

3.2.1 Removing Outliers

For the selected variables, we developed a multiple linear regression model to identify and remove the outliers in the data set if any. With the help of the Bonferroni test, it was found that the data set has 4 outliers (Table 5, Figure 5)[5]. These 4 outliers were then removed resulting in an increased R^2 value of the model.

3.3. Phase 3: Model Build Up

To select which model would make the best prediction, we needed to evaluate the feasibility of a predictive learning model that best fits our data and lowers our Mean Squared Error. Therefore, with the variables chosen in the feature selection phase, we developed a multiple linear regression to test our hypothesis. We decided to start with this model because of its ease of interpretability and the effective ability to analyze how the independent variables impact the dependent variable.

[6] Refer to Section 6: Appendices

To determine if our initial model was accurate and not biased, we perform several tests to evaluate it:

3.3.1 Linearity

Once we came up with our multiple linear models, we tested it in terms of linearity. In this part, we performed a residual plots test on our numerical variables, where we found out that our initial model was not linear. This can be seen in figure 9.

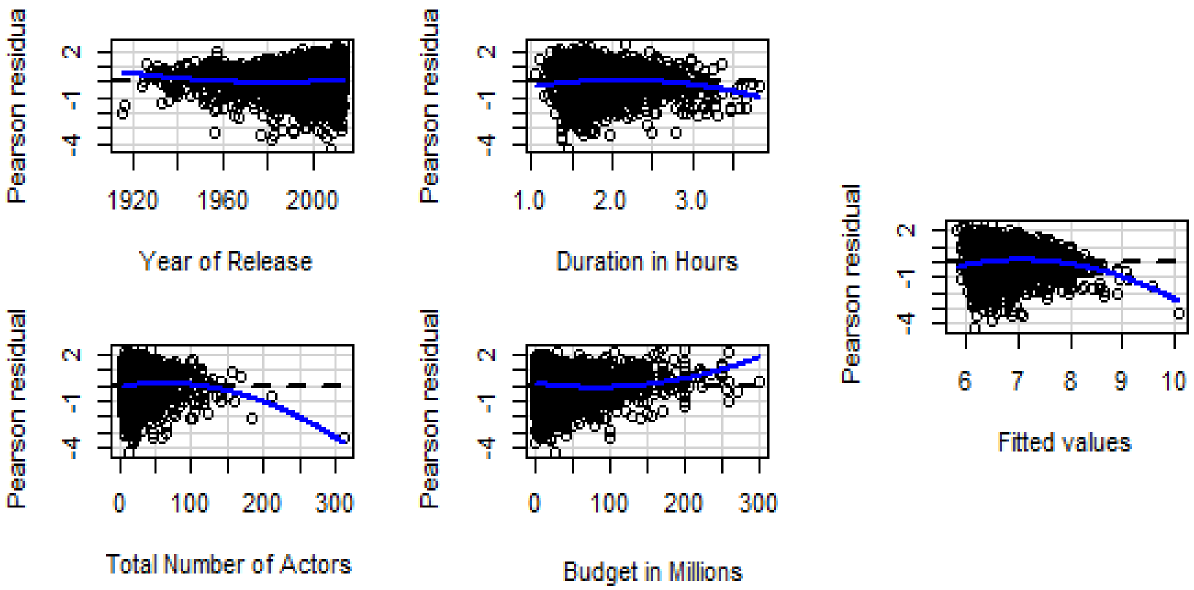


Figure 9: Residual Plots Test

Also, according to our test we found out that our four continuous variables are non-linear as our p-values for each variable are less than .05, this can be seen in Table 2.

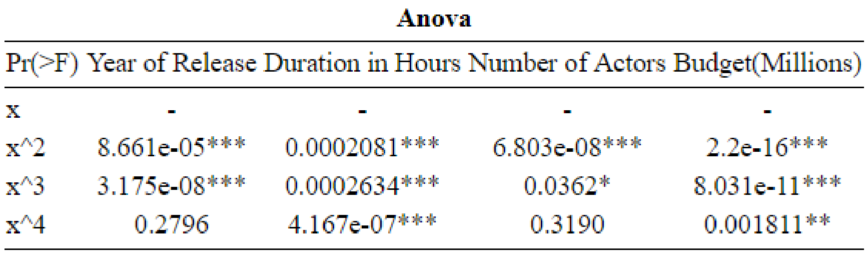
Residual Plots Results

	Test stat	Pr(> | Test stat|)
Year of Release	3.90	0
Duration in Hours	-5.44	0
Total Number of Actors	-5.98	0
Budget in Millions	8.29	0
Tukey test	-7.43	0

Table 2: Tukey Test Results

3.3.2 Polynomial Degree Selection

To determine if a polynomial was the best fit for our model, we performed an ANOVA test to find which degree was optimal for our model. We found out that for the year of release the best degree is a third degree, Duration in Hours: fourth-degree, Number of actors: second degree and Budget: third degree. To come up with this selection we looked for statistically significant P-values, with a threshold of 5% , that can be seen in Table 3.



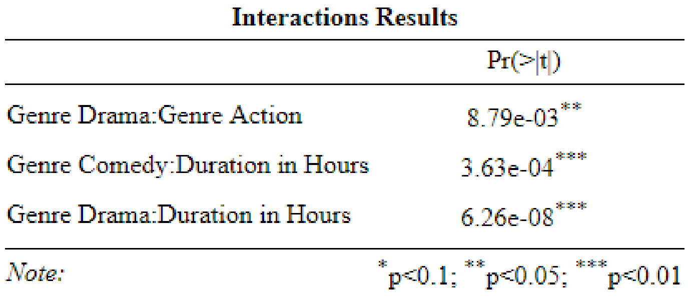
Anova

Pr(>F)	Year of Release	Duration in Hours	Number of Actors	Budget(Millions)
x	-	-	-	-
x^2	8.661e-05***	0.0002081***	6.803e-08***	2.2e-16***
x^3	3.175e-08***	0.0002634***	0.0362*	8.031e-11***
x^4	0.2796	4.167e-07***	0.3190	0.001811**

Table 3: Annova Test on Different Continuous Variables

3.3.3 Interactions

Finally, to tune our model and decrease the MSE, we tested different interactions that the variables could have between each other. During our test, we looked for an interaction that lowered the MSE of the model and its variables were statistically significant in our final regression. We tested all feasible combinations to determine which grouping were important to our model, therefore we found out that four interactions met our hypothesis. These interactions can be seen in table 4. With these interactions our MSE lowered from .5448 to .5386.



Interactions Results

	Pr(>|t|)
Genre Drama:Genre Action	8.79e-03**
Genre Comedy:Duration in Hours	3.63e-04***
Genre Drama:Duration in Hours	6.26e-08***
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Interaction Variables Significance Test

3.3.4 Modelling Conclusions

Our final model is made up of 27 variables and a constant; where 11 variables are continuous variables, 13 binary variables and 3 interactions, with a MSE of .5386. To see the final model with p-values and MSE [Table 8][7]. It is important to highlight that, although there are ways of obtaining a MSE much lower with different methodologies and more advance algorithms, we decided that polynomial regressions with interactions are the best approach to interpret results to decision makers. This allows us to explain each variable and its impact on the score of each film.

[7] Refer to Section 6: Appendix

4. Managerial Implications

According to Watson (2020), the global film industry amounted to 42 billion USD in 2019; the pandemic had a tremendous effect on the industry as many theatres were forced to shut off globally. With the growing complexity and increase in competition, it is essential to use an efficient data model that can help access the factors that increase the movie's likeability by the viewers. IMDb has been collecting the data since 1990 and has around 12M movie data collected over time (IMDb, 2020). The prediction model developed by the team will help in the prediction of the IMDb score, where the error of approx. 0.70 in scores can be due to the modeling error. Therefore, it is crucial to understand the various factors that positively and negatively affect the movie's IMDb scores.

In the model, the movie's IMDb score shares the nonlinear relationship with the year of release of the movie, duration (running time) of the movie, number of actors in the movie, and budget of the movie.

The genres, such as drama, action, comedy, horror, documentary, family, and animation, significantly impact the IMDb score.

1) Using the results, it can be inferred that drama, action, and family movies tend to have low scores than movies with other genres. For example, the movie with genre action has 0.26 less IMDb score as compared to movies categorized with other genres, provided all other conditions are the same. Similarly, we can compare how the genre mentioned above impacts the IMDb scores using Table 7[8] and corresponding coefficient values.

2) Movies with genre comedy, documentary, and animation has a positive impact on IMDb scores. So, the movies with the genre mentioned above tend to get more scores than other genres, keeping other factors such as budget, duration, etc. constant.

    a) The animated movie will have 0.52 more IMDb score than any other non-animated movie, provided all other conditions are the same.

    b) The drama movies also attract more scores from the reviewers. On an average the drama movies have 1.16 more scores than movies that are not from drama genre.

    Similarly, we can compare how the genre, as mentioned earlier, impacts the IMDb scores using Table 7[8] and corresponding coefficient values. It is recommended to the producer and directors to make movies of genre Comedy, Drama, Animation and Documentary since they are more liked by the viewer. Furthermore, we also suggest that it is better to avoid Action movies, Horror movies and Family movie as they strike low IMDb scores.

The movie genre impacts the IMDb score to some extent; however, the other combination of factors such as runtime of the drama movie; runtime of the comedy movie and a drama & action movie also plays the crucial role in predicting the IMDb score.

[8] Refer to Section 6: Appendix

1) The comedy and drama movie with longer run time tends to have a lower rating as compared to movies which does not belongs to comedy or drama genre. In other words, it is advisable to have a short duration of comedy and drama movies to have a higher IMDb Score.

2) The action-drama movie is also not a good option to invest movie in, as the combination generally have lower by 0.18 IMDb scores as compared to a drama or an action movie separately.

Some other factors tend to affect the IMDb score of the movie. Please note that the analysis is done when one parameter is changed, and the other parameter remains constant.

1) Movies directed by Jason Friedberg have a lower IDMB score compared to movies directed by other directors.

2) Movies that were produced by Dimension Films tend to have a lower IMDb score by 0.53 than movies produced by any other production company.

3) Movies produced in the United Kingdom have higher IMDb scores by 0.18 compared to movies produced in any other country.

4) Movies with the main actor as females have lower IMDb scores by 0.20 than the movies in which the main role is played by a male.

So, the factors that account for the movie with a high rating, which the producers and directors should also keep in mind while making a movie. Just to summarize, the factors that affect the scores in a positive fashion and negative fashion, respectively [Table 7][9]

Positive Factors- Comedy Movie, Drama Movie, Documentary Movie, Animated Movie, Movie Produced in UK

Negative Factors – Action Movie, Main Actor as Female, Horror Movie, Family Movie, Movie Directed by Jason Friedberg, Movie Produced by Dimension Films, Drama-Action Movie.

[9] Refer to Section 6: Appendix

5. References

IMDb. (2020, October). IMDb Statistics. Retrieved from IMDb: https://www.imdb.com/pressroom/stats

Watson, A. (2020, November 10). Film Industry - statistics & facts. Retrieved from Statista: https://www.statista.com/topics/964/film/

6. Appendix

6.1 Code

Click on the link to view the code: Code

6.2 Other Figures and Tables

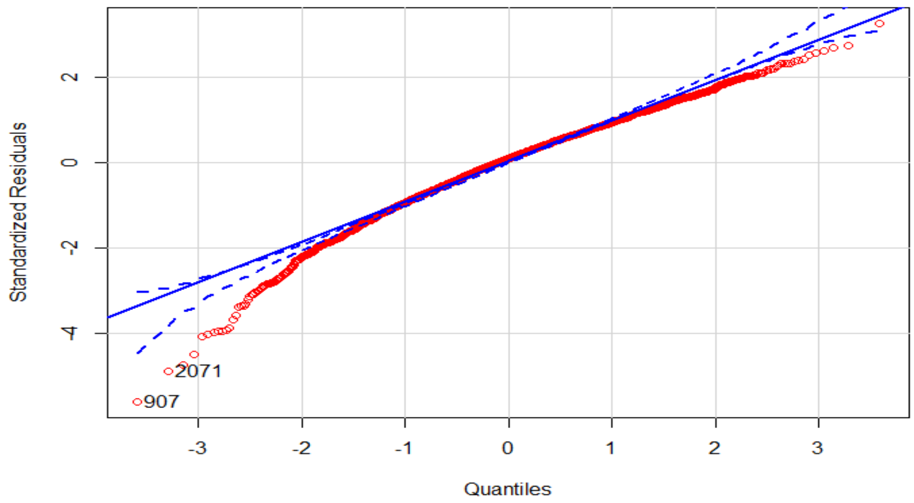
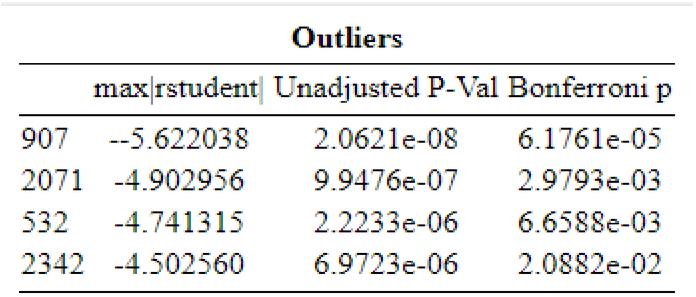


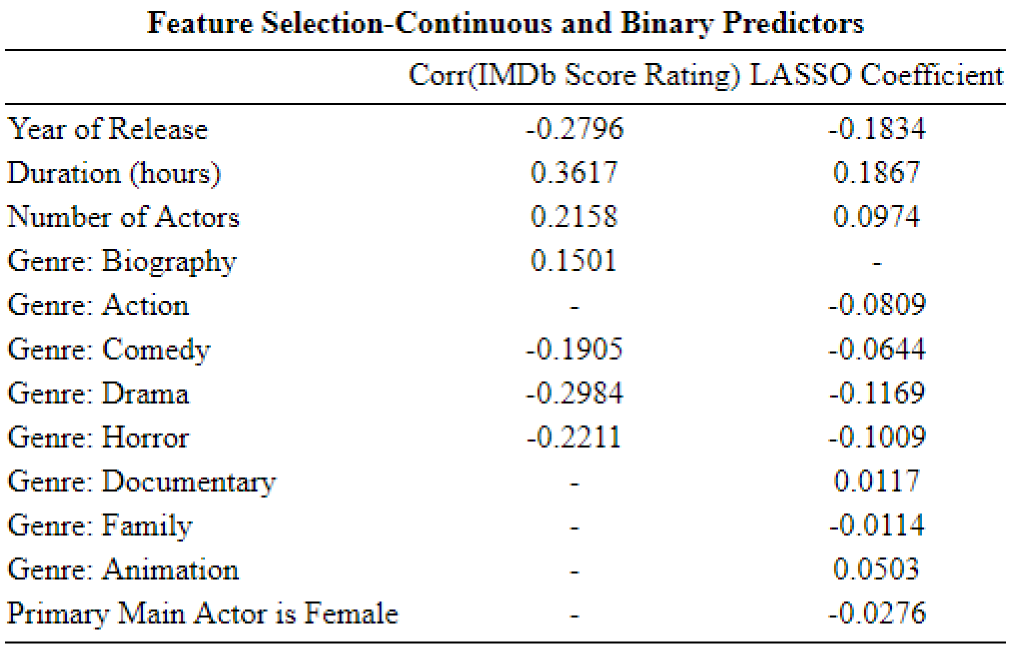
Figure 10: Graphical Notation of Outliers



Outliers

	max|rstudent|	Unadjusted P-Val	Bonferroni p
907	--5.622038	2.0621e-08	6.1761e-05
2071	-4.902956	9.9476e-07	2.9793e-03
532	-4.741315	2.2233e-06	6.6588e-03
2342	-4.502560	6.9723e-06	2.0882e-02

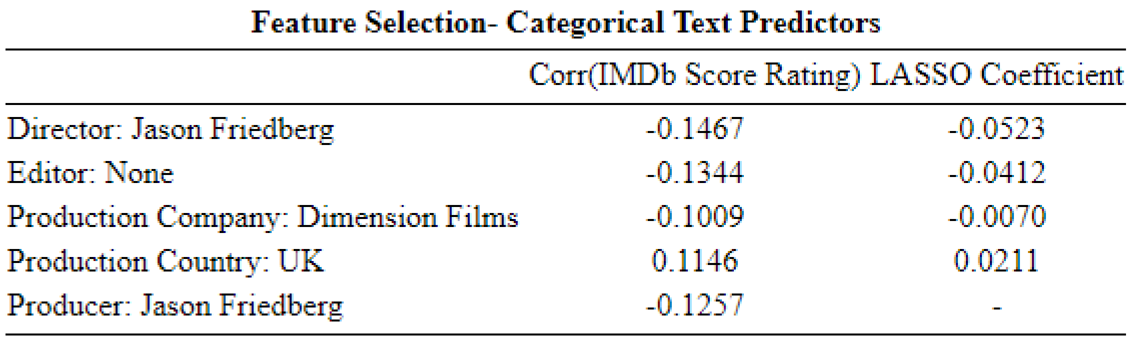
Table 5: Outliers in the Data



Feature Selection-Continuous and Binary Predictors

	Corr(IMDb Score Rating)	LASSO Coefficient
Year of Release	-0.2796	-0.1834
Duration (hours)	0.3617	0.1867
Number of Actors	0.2158	0.0974
Genre: Biography	0.1501	-
Genre: Action	-	-0.0809
Genre: Comedy	-0.1905	-0.0644
Genre: Drama	-0.2984	-0.1169
Genre: Horror	-0.2211	-0.1009
Genre: Documentary	-	0.0117
Genre: Family	-	-0.0114
Genre: Animation	-	0.0503
Primary Main Actor is Female	-	-0.0276

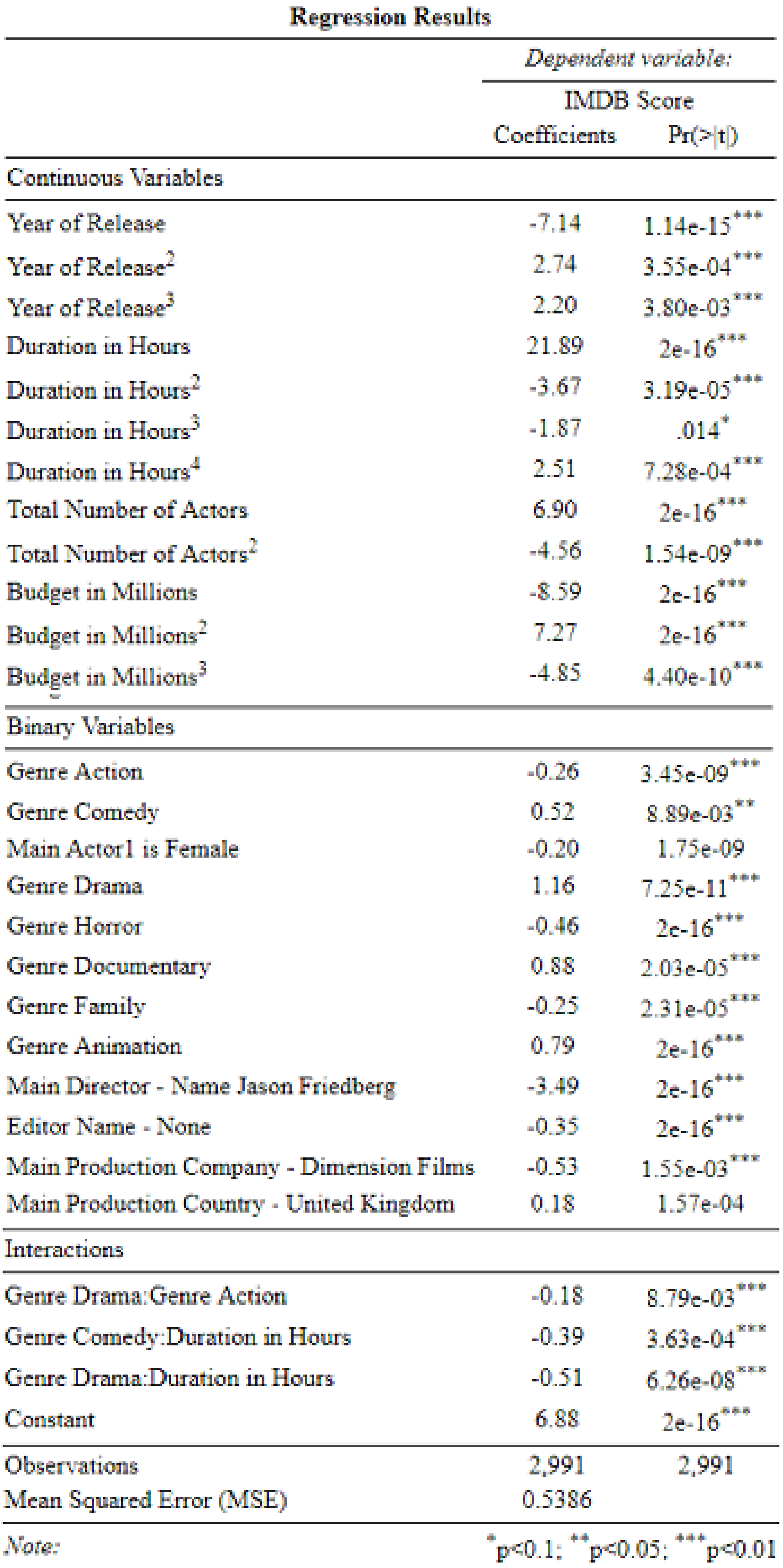
Table 6: Feature Selection Continuous and Categorical Variable



Feature Selection- Categorical Text Predictors

	Corr(IMDb Score Rating)	LASSO Coefficient
Director: Jason Friedberg	-0.1467	-0.0523
Editor: None	-0.1344	-0.0412
Production Company: Dimension Films	-0.1009	-0.0070
Production Country: UK	0.1146	0.0211
Producer: Jason Friedberg	-0.1257	-

Table 7: Feature Selection – Categorical Variable



Regression Results

	Dependent variable:	
	IMDB Score	
	Coefficients	Pr(>|t|)
Continuous Variables		
Year of Release	-7.14	1.14e-15***
Year of Release^2	2.74	3.55e-04***
Year of Release^3	2.20	3.80e-03***
Duration in Hours	21.89	2e-16***
Duration in Hours^2	-3.67	3.19e-05***
Duration in Hours^3	-1.87	.014*
Duration in Hours^4	2.51	7.28e-04***
Total Number of Actors	6.90	2e-16***
Total Number of Actors^2	-4.56	1.54e-09***
Budget in Millions	-8.59	2e-16***
Budget in Millions^2	7.27	2e-16***
Budget in Millions^3	-4.85	4.40e-10***
Binary Variables		
Genre Action	-0.26	3.45e-09***
Genre Comedy	0.52	8.89e-03**
Main Actor1 is Female	-0.20	1.75e-09
Genre Drama	1.16	7.25e-11***
Genre Horror	-0.46	2e-16***
Genre Documentary	0.88	2.03e-05***
Genre Family	-0.25	2.31e-05***
Genre Animation	0.79	2e-16***
Main Director - Name Jason Friedberg	-3.49	2e-16***
Editor Name - None	-0.35	2e-16***
Main Production Company - Dimension Films	-0.53	1.55e-03***
Main Production Country - United Kingdom	0.18	1.57e-04
Interactions		
Genre Drama:Genre Action	-0.18	8.79e-03***
Genre Comedy:Duration in Hours	-0.39	3.63e-04***
Genre Drama:Duration in Hours	-0.51	6.26e-08***
Constant	6.88	2e-16***
Observations	2,991	2,991
Mean Squared Error (MSE)	0.5386	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 8: Regression Results