STATISTICAL FOUNDATIONS OF DATA ANALYTICS
PROFESSOR: JUAN CAMILO SERPA

THE 2020 IMDb PREDICTION CHALLENGE

A REPORT BY

DATA RULERS

ZHIYU BAI	260769212
KEXIN CHEN	260769754
ALISON ZIQI LIU	260834371
SHUWEN LU	260792831
JUNQIU NI	260765732

PROJECT OVERVIEW

Executive Summary

Since the COVID-19 pandemic began in March, closed cinemas, canceled festivals and postponed film releases have had a severe impact on all movie businesses across the world. Facing detrimental revenue losses, film studios and producers are in need of a hit movie to survive after the lockdown is lifted. The motivation of this project is to identify which characteristics make a movie successful and popular. For our data analysis, we used IMDb rating as the target variable, as it is the most popular and authoritative source for public views on a movie. By creating a model that can predict a movie's IMDb score, we would like to help related practitioners create more outstanding movies.

First of all, we explored the data describing the characteristics of a movie by observing the distribution of dependent and independent variables. Then, we removed outliers by conducting outlier tests. After that, we explored the correlation relationships between each variable. We also used simple linear regression to overview the relationship between the dependent variable and independent variables. In the process of selecting the model with the best prediction power, we firstly selected the most relevant variables to build the base model. Then, to optimize the model, we tested multiple polynomial and spline regressions to check whether they improved our model's performance or not. With the final model defined, model issues were thoroughly considered to ensure an optimal model.

Descriptive Analysis

The dataset provided contains 2995 movie observations with 52 variables describing their characteristics such as budget and year of release. The dependent variable, 'imdb_score', represents the rating that a movie has received from the public on the IMDb website. People can rate a movie from 0-10 regarding how much they love the movie. Movie ratings provided in this dataset range from 1.9 to 9. The mean value is 6.7, with a majority of the movies having a rating between the range 6 to 8. There are

also three label variables that won't contribute to the regression, they were therefore removed from the dataset. The rest 48 independent variables were then separated into three categories: continuous variables, discrete variables and categorical variables.

Continuous and discrete variables were analyzed by boxplots and histograms in order to get an impression of their distributions. Most variables are positively skewed, meaning that most of the observations have a value smaller than the mean. Meanwhile, the variable 'year_of_release' is negatively skewed as the number of movies released has increased over the years. Therefore, skewness and the number of outliers that have a value larger than 1.5 times IQR(interquartile range) are calculated as numeric indicators (Table 2). The result proves our assumption, all numerical variables, except for 'month_of_release', are heavily skewed with their absolute skewnesses larger than 1.0. Highly skewed data often contain many outliers, which will seriously impact the accuracy of a linear predictive model. Logarithm transformation and removing outliers were considered during the regression process in order to correct the skewness, which will be furtherly explained in the model selection part.

In order to lower the interruption of outliers on model selection, we performed a basic outlier test for each continuous variable. Surprisingly, it turned out that observations 641 and 907 are outliers for most variables. Besides, two additional outliers (observation 582 and 2071) were detected in the test for variable 'total_number_of_actors'. These four entries were removed from the dataset before the regression process.

Other than those continuous and discrete variables, categorical variables also account for an important part of the dataset. Movie genres are already dummified into 25 different variables. Comedy and drama are the most popular genres, and it is noticeable that one movie could have several genres at the same time. Most of the movies are produced in the United States (2114), other movies included in the dataset come from 45

countries worldwide. As a result, almost 90% of the movies' main language is English. The whole dataset includes movies directed by 1349 different directors from 730 different production companies. Note that it is also possible for a movie to have more than one production company. Furthermore, by summarizing the entries "main_actor 1,2,3_is_female", the results showed that the movies that have three male main actors in this dataset nearly doubled the size of movies with three main female main actors (5975 observations vs 3010 observations).

Afterwards, scatter plots were generated with the dependent variable as y and each numerical independent variable as x, in order to get a preliminary estimation on the relationship between different movie characteristics and its IMDb score received. As shown in Table 3, most independent variables display a funnel shape in their scatter plots with most observations gathering on the smaller-than-mean part of the plot. This may cause heteroscedasticity in the regression, as more observations usually have a bigger variance regarding their y values. Therefore, heteroscedasticity test and correction should be performed on the final model, which will be explained in detail at the end of the model selection part.

In the end, a correlation matrix was generated in order to find out the relationship between all the numerical independent variables and the dependent variable (Table 4). Among all the numerical variables, there is no strong collinearity presented as all the absolute values of correlations smaller than 0.5. Therefore, there is no concern that the predictors chosen in the model will influence each other. The VIF of the predictors were also calculated in order to ensure there is no collinearity between the predictors chosen. However, no predictor displayed a strong correlation with the dependent variable 'imdb_score' either. This represents that no single predictor in the dataset has a strong influence on the IMDb rating, which may result in a bad performance in the regression model.

Model Selection

To predict the IMDb ratings of movies, our group chose to use multiple linear regression to build our first model as a base model to build upon. The selection of independent predictors followed the following steps. First of all, we ran the correlation function to find the variables that are very likely to affect the movie scores and ranked the variables in descending order based on its degree of effect; Afterwards, we ran simple linear regressions on each independent variable with the dependent variable. Independent variables in the base model were chosen based on the correlation values and the adjusted R-squared values; Then, we conducted a linearity test on the model and tried different combinations of polynomial, spline and interaction variables on the non-linear predictors. The optimal model was determined by choosing the model that generated the lowest average and variance of MSE. In the end, model issues such as heteroscedasticity were tested and final trials and adjustments will be made in order to improve the model performance.

As for selecting independent variables in our model, we first ran the cor() function to calculate the correlations between each predictor and the dependent variable As mentioned earlier, all predictors have a relatively weak correlation with the target variable. Variable 'duration_in_hours' has the largest correlation with the dependent variable, which is only about 0.36. We were also surprised to find that the correlation between the budget of a movie and its rating was only -0.065. It is a negative correlation, which means that a higher budget generally leads to a lower rating. Therefore, it is necessary to include relatively more variables in our model to improve the accuracy of our prediction. After looking through all of the correlation coefficients, we decided to rank their correlation values in descending order and choose the variables that have a correlation greater than 0.1. Then, we ran simple linear regressions on the variables and chose the ones that also have the highest adjusted R-squared. Ten variables were selected for the base model, as shown in Table 5.

Afterwards, we tried to identify if any of them doesn't contribute to the model. We excluded the ten variables one by one to see whether the adjusted R-squared value decreases or not. Fortunately, all the variables improved the adjusted R-squared value to some extent. Although the increase is relatively small when we include the predictor 'genre_biography' and 'genre_history', we still decided to include them in our base model since they do improve the prediction power.

After we decided on the base model, we then looked for the optimal model that would best predict ratings of IMDb movies. We first conducted the linearity test by using the residualPlots() function. Unsurprisingly, the result showed that Tukey test has a P-value of 8.080x10-11, which is much smaller than 0.05. This indicates that our model so far does not satisfy the linearity assumption. Specifically, there are three non-linear predictors: 'year_of_release', 'duration_in_hours', and 'total_number_of_actors' (Table 6). We referred to the pattern of scatter plots for the three predictors to determine their relationship with 'imdb_score'. For variable 'year_of_release', we found that the spline function best applies to this predictor because the graph shows different patterns before and after the year 1980 (Table 7). Thus, we used 1980 as the knot for the spline function. For all the three predictors, we attempted numerous models by varying the degree of each variable and conducting the ANOVA test. In order to prevent the chosen model from overfitting, K-fold tests were conducted on every model and the model with the lowest average and variance of its MSE scores were chosen. The final model is shown as follows:

```
reg = lm(imdb_score~month_of_release +bs(year_of_release,knots=c(1980),degree=2)
+poly(duration_in_hours,4) +poly(total_number_of_actors,2) +genre_action
+genre_comedy +genre_drama +genre_horror +genre_history +genre_biography,
data=films)
```

Other tests were also conducted on the model to detect model issues. We conducted a test on collinearity again using the VIF approach (Table 8). Our results illustrate that all variables, except for the one generated for spline regression, have a VIF smaller than 2, confirming that there is no multicollinearity presented in the model. In the end, we conducted an

NCV test on the final model. The p-value of the NCV test is 2.77e-13, which is way smaller than 0.05, representing the existence of heteroskedasticity in the model. As a result, the coeftest() function was used to correct the heteroskedasticity (Table 9). Several variables decreased in the significance level， but still at a favorable significant level. Therefore, we decided not to include additional transformations on the selected variables. As shown in Table 10, our corrected final model has an average MSE of 0.626, representing that the average squared residual of our predictive model. In other words, if the model is used to predict a new movie's IMDb score, the approximate error will be 0.39.

Managerial Implications

Looking over our results, several insights could be generated for movie directors and producers to make a distinguished blockbuster. Firstly, the correlation between the budget of a movie and its rating was really small, and the effect on the film rating can thereby be neglected. Therefore, the rating of the movie does not directly depend on the budget, the producers need not worry too much about their budget affecting the final film's score. Regarding the coefficients of the genre variables, the coefficients for 'genre_drama' and 'genre_biography' are both positive, respectively about 0.2 and 0.13. The numbers can be interpreted as: with all else being equal, theatrical and biographical films tend to get higher ratings. However, the coefficients of the remaining genre variables are all negative. This suggests that with all else being equal, movies in action, comedy, horror, and history categories, with coefficients -0.36, -0.24, -0.55, and -0.17 respectively, normally scored lower than those in other categories. Therefore, it is significant for the directors of these types of films to enhance other aspects of their films, such as duration and the total number of actors, in order to reach a high rating.

As for the month of release of the movie, the coefficient of the variable is positive, indicating that movies released at the end of the year (November and December) would normally receive a better rating. This may be due to

a higher volume of audiences visiting cinemas during the Christmas and the New Year's holidays. It is recommended that producers seize this opportunity to earn high box-office income by releasing high-quality films at that time since a higher rating movie always reaches a higher box-office income. However, the magnitude of the coefficient is pretty small, only about 0.012; as a result, the effect on the film rating can thereby be neglected. This makes sense since the month of releasing a movie does not contribute much to the quality of the movie; moreover, it is common that the directors cannot control the releasing time themselves. Besides, the coefficients of spline regressions for 'year_of_release' cannot be interpreted, we include it in our model only for prediction purposes.

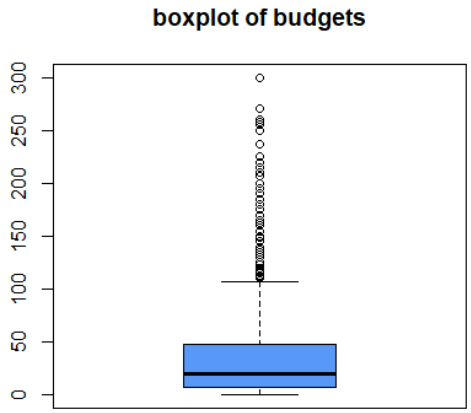
Regarding the significance levels of those characteristics, we determined them by looking at the magnitude of coefficients in the regression model. Our model reveals that duration of a film (degree=1, the rest of the degrees didn't have a significant contribution to the model) and the total number of actors of a movie (degree=1) are the two most significant predictors for the reason that they have the coefficient of 9.95 and 7.34 respectively. This is strong evidence proving that an increase in either of those two aspects would effectively improve the rating of a movie. A movie with a longer length might imply that more plots/stories are narrated. For instance, if a movie is adapted from fiction, longer durations would allow it to involve more details and hence increase the possibility to satisfy audiences' expectations. Thus, it is reasonable to get a higher rating from that. Similarly, the total number of actors could reflect the effort that a production team has put into the movie as well and therefore is significant for predicting the ratings.

In conclusion, the variables in our model contribute a lot to predict the movies' rating, however, only looking at those predictors might not lead to an extremely accurate estimation. Many of the reasons people like or dislike a movie are subjective, such as how intriguing a plotline is. Also, one common characteristic of award-winning and highly related movies is their uniqueness: a discussion on new topics, the inclusion of a new

technology, or the introduction of a new style. Moreover, people who tend to give a rating on IMDb tend to be the ones that have a strong opinion, either positive or negative, adding an extreme-responding bias in the dataset. Therefore, more rigorous studies should proceed in order to find out the core of a movie's success.

Appendix

Table 1: Examples of Data Exploration - Numerical Variables



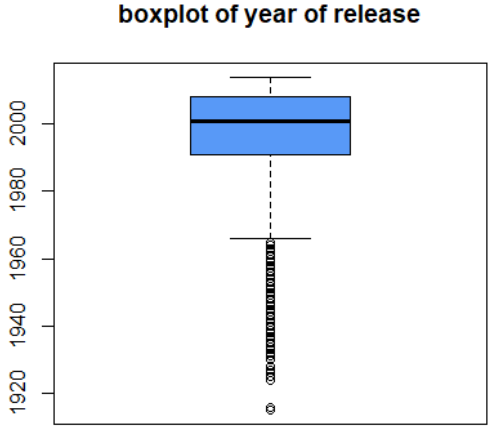
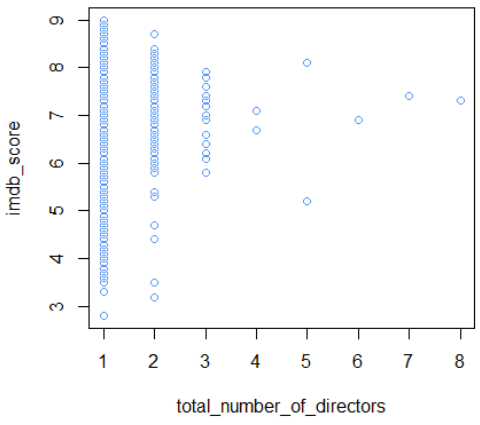


Table 2: Summary of Skewness - Numerical Variables

Variables	Mean	Skewness	Outlier
Budget (in millions)	35.16	2.22	192
Month of Release	7.04	-0.16	0
Year of Release	1996	-1.75	194
Duration (in hours)	1.85	1.44	120
#Languages	1.54	2.37	145
#Directors	1.07	8.83	158
#Actors	24.45	3.71	254
#Producers	2.256	1.19	66
#Production Companies	2.95	2.11	94
#Production Countries	1.38	2.66	78

Table 3: Example of Scatter Plot



Appendix

Table 4: Correlation matrix - Y and Numerical Variables

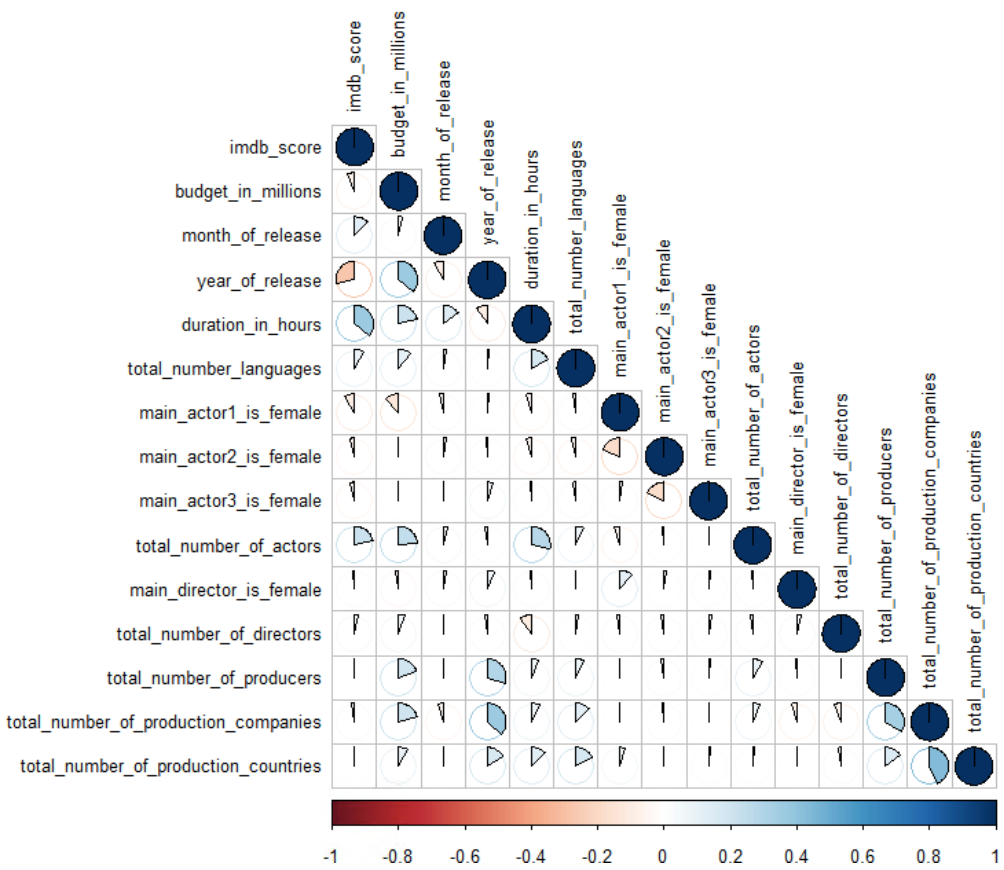


Table 5: Top 10 Predictors - Correlation & Adjusted R-squared

Variables	Correlation	Adjusted R^2
Duration (in hours)	**0.36**	**0.13**
Genre_drama	**0.30**	**0.09**
Year of Release	**-0.28**	**0.08**
Genre_horror	**-0.22**	**0.05**
#Actors	**0.22**	**0.05**
Genre_comedy	**-0.19**	**0.04**
Genre_biography	**0.15**	**0.02**
Genre_action	**-0.15**	**0.02**
Month of Release	**0.12**	**0.02**
Genre_history	**0.12**	**0.01**

Appendix

Table 6: Linearity test - residualPlots( )

	Test stat	Pr(>|Test stat|)	
month_of_release	-1.0266	0.304706	
year_of_release	2.1818	0.029199	*
duration_in_hours	-2.9139	0.003596	**
total_number_of_actors	-6.0370	1.764e-09	***
genre_action	1.5487	0.121550	
genre_biography	-0.9540	0.340166	
genre_comedy	1.4546	0.145874	
genre_drama	-1.6089	0.107748	
genre_history	0.9568	0.338728	
genre_horror	1.9348	0.053113	.
Tukey test	-6.5448	5.957e-11	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 7: Spline function - year_of_release (knot=1980, degree=2)

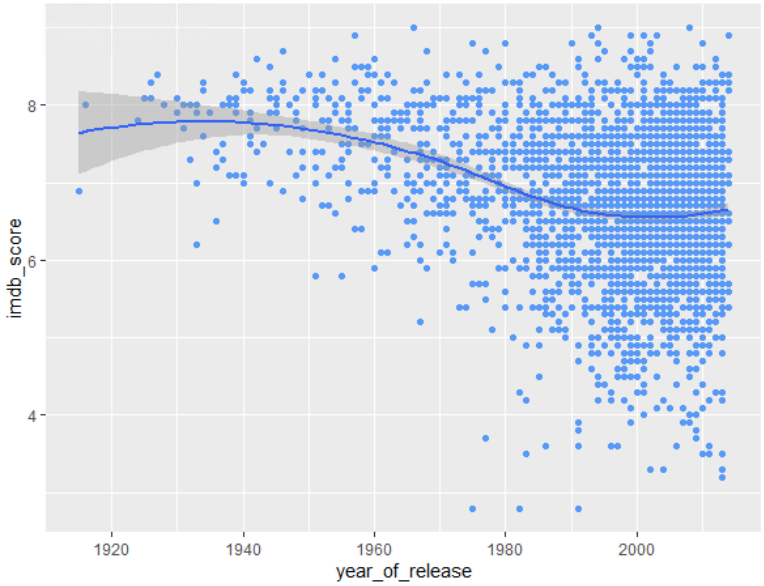


Table 8: Collinearity Test

```
> ols_coll_diag(reg4)
Tolerance and Variance Inflation Factor
---------------------------------------
```

	Variables	Tolerance	VIF
1	month_of_release	0.95605877	1.045961
2	bs(year_of_release, knots = c(1980), degree = 2)1	0.07264322	13.765910
3	bs(year_of_release, knots = c(1980), degree = 2)2	0.08870233	11.273662
4	bs(year_of_release, knots = c(1980), degree = 2)3	0.03628479	27.559755
5	poly(duration_in_hours, 4)1	0.69695267	1.434818
6	poly(duration_in_hours, 4)2	0.93958496	1.064300
7	poly(duration_in_hours, 4)3	0.98198860	1.018342
8	poly(duration_in_hours, 4)4	0.98256979	1.017739
9	poly(total_number_of_actors, 2)1	0.86985263	1.149620
10	poly(total_number_of_actors, 2)2	0.97901033	1.021440
11	genre_action	0.76351668	1.309729
12	genre_comedy	0.71114163	1.406190
13	genre_drama	0.67238624	1.487240
14	genre_horror	0.78560038	1.272912
15	genre_history	0.85898362	1.164167
16	genre_biography	0.88003524	1.136318

Appendix

Table 9: Model Summary - Correct Heteroskedasticity

```
                                                 Pr(>|t|)
(Intercept)                                     < 2.2e-16 ***
month_of_release                                 0.005387 **
bs(year_of_release, knots = c(1980), degree = 2)1  0.012081 *
bs(year_of_release, knots = c(1980), degree = 2)2 1.798e-05 ***
bs(year_of_release, knots = c(1980), degree = 2)3  0.003842 **
poly(duration_in_hours, 4)1                     < 2.2e-16 ***
poly(duration_in_hours, 4)2                      0.009265 **
poly(duration_in_hours, 4)3                      0.001523 **
poly(duration_in_hours, 4)4                     7.858e-05 ***
poly(total_number_of_actors, 2)1                < 2.2e-16 ***
poly(total_number_of_actors, 2)2                4.108e-11 ***
genre_action                                    < 2.2e-16 ***
genre_comedy                                    5.802e-11 ***
genre_drama                                     2.519e-09 ***
genre_horror                                    < 2.2e-16 ***
genre_history                                    0.011962 *
genre_biography                                  0.015412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 10: Model Summary - Final Model

	Dependent variable:
	imdb_score
month_of_release	0.012^{***} (0.004)
bs(year_of_release, knots = c(1980), degree = 2)1	0.696^{**} (0.321)
bs(year_of_release, knots = c(1980), degree = 2)2	-0.811^{***} (0.226)
bs(year_of_release, knots = c(1980), degree = 2)3	-0.613^{**} (0.249)
duration_in_hours	9.954^{***} (0.945)
duration_in_hours^2	-2.199^{***} (0.814)
duration_in_hours^3	-2.471^{***} (0.796)
duration_in_hours^4	2.981^{***} (0.796)
total_number_of_actors	7.388^{***} (0.846)
total_number_of_actors^2	-4.788^{***} (0.797)
genre_action	-0.356^{***} (0.038)
genre_comedy	-0.242^{***} (0.036)
genre_drama	0.202^{***} (0.035)
genre_horror	-0.550^{***} (0.053)
genre_history	-0.172^{**} (0.080)
genre_biography	0.129^{*} (0.068)
Constant	7.292^{***} (0.244)
Observations	2,991
R^2	0.322
Adjusted R^2	0.318
Residual Std. Error	0.789 (df = 2974)
F Statistic	88.078^{***} (df = 16; 2974)
Note:	*p<0.1; **p<0.05; ***p<0.01