Contents

1 Introduction

In 1880, the Directorate of Fine Arts commissioned French sculptor Auguste Rodin to create La Porte de l'Enfer (The Gates of Hell), depicting a scene from Dantes Divine Comedy. Famously, Rodin relentlessly worked on the project, perfecting each and every detail for 37 years until his death at the age of 77. Standing at a towering 6 metres high (19.7) and containing 180 sculpted figures, the grandiose sculpture is currently on display in Paris at the Muse Rodin. The task of constructing a statistical model is a similarly artistic endeavour. In just under a month – a fraction of Rodins 37 years – we have worked tirelessly to build a beautiful statistical model.

Our model strives to best predict IMDb scores for the slate of movies arriving this holiday season. Using a combination of continuous and categorical variables from 3,132 past movies, ranging from The Shawshank Redemption (9.3 on IMDb) to Foodfight (1.7), we sought to minimize our models Mean Squared Error (MSE), essentially a measure of a models ability to predict new observations. Models with a lower MSE (as close to zero as possible) are better at predicting new, out of sample observations, which is exactly what we are looking for when forecasting the IMDb score of upcoming films.

When any movie, from Gone with the Wind to The Godfather, is searched on Google, IMDb is always the first result. In a matter of seconds, without even clicking a link, searchers are presented with the movies IMDb rating. Last Sunday evening, for the team movie night, we couldnt decide between watching Django Unchained, Quentin Tarantinos revenge thriller, and Eat Pray Love, a romantic comedy drama starring Julia Roberts. After much discussion and debate, we turned to IMDb to solve our dilemma. Compared to Djangos 8.4 stars, Eat Pray Loves 5.7 didnt stand a chance.

In addition to helping our group with our Netflix movie night, IMDb scores greatly influence ticket sales and box office revenues. A model that can accurately predict IMDb scores would be of great value to movie studios and producers hoping to cash in. What director should they select? Should the movie be two hours, or only one and a half? How many faces should be in the poster? Hopefully, our model will be able to address all these questions and more, in determining the optimal variable selection and statistical methodology to predict the IMDb scores of an upcoming movie.

2 Data Description

We started with 3,132 observations and 69 predictive variables, some continuous and other categorical. Our first step was cleaning the data, and making it suitable for a proper statistical analysis. We conducted research to replace any N/A entries in the data with the proper information. We did this for a total of 81 entries across distributor, number of votes and other values. Some examples include inputting Peter Deluise as the director of Stargate SG-1 and Indian Paintbrush as the production company responsible for Like Crazy. Filling all of the holes in the data prior provides us with a more complete and comprehensive dataset, which will lead to a better model and more accurate predictions. In cleaning the data, we also set all character columns to numeric for continuous variables such as movie budget, and set factor columns to character, for example the movie actor 1 is known for. This guarantees that each predictor is measured in the correct way.

When looking through the data, we realized that many of the categorical variables had hundreds of different response values. Some values were recurring, showing up time and time again, whereas others only appeared once or twice in the entire dataset. For example, Martin Scorsese has directed 17 movies (almost all classics), compared to Jonathan Caouette who has only directed one. Naturally, we can expect to better predict the IMDb rating of Scorseses next movie than Caouettes, as we have a significantly larger sample to inform our predictions. To address this phenomenon, we decided to bundle together all categorical responses below a certain threshold into an other value. We kept Scorsese, Spielberg and all other directors with at least five movies directed, as this is what we deemed a large enough sample to serve as a strong predictor. All other directors would be group in the other bin, as they did not have a large enough body of work to be considered predictive as directors. This choice was backed up with the really low p-value that had those others directors. They were not statistically significant. We chose the number five because it yielded an acceptable r-squared while simplifying the model.

We found that an overly complex model yielded poor predictive results but had a high r-squared. We conducted a similar process for a number of other categorical variables, with varying thresholds depending on the nature of the data. For distributor, where the sample sizes are often much larger (Lionsgate has distributed 64 movies in the dataset), we chose a minimum threshold of 20. It was all about finding the right balance between a good enough r-squared and a simple enough model, one of the many arts involved in model building.

The next step after cleaning and organizing our data was to examine the distribution amongst and relationships between variables. Many of the variables were distributed in a manner we expected, with a few interesting exceptions. One instance is release year, which was skewed left because most of the movies in our dataset came out after 2000 (Appendix 1). This will be important to be aware of later on. Additionally, the IMDb scores of the movies are fairly normally distributed with a slight left skew and a mean score of 6.52 (Appendix 2). We found this to be surprising, as we expected the average score to be lower, and to have more variance.

We also found a few interesting relationships between our predictive and response variables. Notably, critic reviews number has a very uniform distribution when plotted again IMDb score (Appendix 3). We found this surprising as we expected to find a positive trend in the data, a higher number of critic reviews being correlated with a higher IMDb score. Number of votes however, shows a positive trend when plotted against IMDb score (Appendix 4). We would have expected that these two variables would have similar plots, and perhaps even collinearity, wherein two highly correlated variables prevent the impact of either one from being properly understood. However, the adjusted R-Squared for number of votes is significantly higher than critic reviews number, meaning it has much stronger correlation to a higher IMDb score. Upon further reflection, we concluded that critics are likely to review a movie regardless of its quality, because they have an obligation. On the other hand, number of votes is constructed by the public, and perhaps the public is more likely to vote on a movie that they enjoyed than one that they didnt enjoy or a movie that has earned a higher IMDB score gains more popularity and, in turn, more votes from users.

We were very disappointed when plotting the relationship between movie budget and IMDb score (Appendix 5). We predicted a strong positive correlation between these two variables, however there was practically no correlation between budget and IMDb score, with an r-score of 0.001. This certainly goes against the intuitive logic that a higher budget leads to a better movies.

A surprising relationship was found between release year and IMDb score (Appendix 6). Although the r-squared is not very significant due to the lack of data on older movies, we can see from the plot that older movies tend to have higher movies ratings. Does that mean that movies are getting worse? After some research, we realized that most of these movies were in fact released far before IMDb even existed. Technology journalist Paul Bischoff theorizes that new movies are receiving a full spectrum of criticism from the public, resulting in lower ratings. Old movies are generally only retroactively reviewed if they are notable classics, resulting in higher ratings. Who would go back and write a review for a mediocre movie from the 1940s!
For categorical variables, we generated bar charts to analyze the distributions of each variable. Although there were no major discoveries, this step helped us become more familiar with the categorical data. Release month was an interesting predictor as we expected there to be a spike in movie releases in the holiday season and the summer, however it was a relatively uniform distribution across all 12 months (Appendix 7). The bar chart for Content Rating was also surprising; R rated movies were the most common even though they are appropriate for the smallest portion of movie viewers (Appendix 8).

After understanding the relationships between each predictor variable and IMDb score, it was essential to detect and remove outliers from our dataset. By removing outliers, we eliminated the least reliable data points, and improved the predictive power of our model. Given the massive sample of data, we decided that removing outliers numerically, through a Bonferroni test, would be the most accurate method. The majority of predictors had between one and three outliers, with a couple having none at all, insinuating that only the most extreme examples were completely inconsistent with the rest of the data, and that others may have been entry errors. Interestingly, every single outlier deviated negatively from the dataset. This may be because, with a mean IMDb rating of 6.52 it is far more likely for a movie to deviate negatively than positively from the average. Furthermore, the same group of low-scoring terrible movies appeared again and again, as outliers for many of our continuous variables. Foodfight, for example, appeared 24 times as an outlier. We dont plan on watching that for our next movie night.

After removing outliers, we wanted to examine the relationship amongst predictive variables to see if any were highly correlated. This is important because if some of the predictive variables do have collinearity, including them both in our model we would essentially be double counting, resulting in a biased model. Many of the results from our correlation matrix (Appendix 9) are intuitive, for example actor 1 Facebook likes and cast total Facebook likes are highly and positively correlated as expected. We would have expected an actors Facebook likes to be highly correlated with their star meter, as both are a measure of popularity. We were surprised to find that these two variables are not correlated at all. Now that collinearity has been identified, we can ensure we do not include collinear variables in our model.

The last step before constructing our model was to identify and corrected for any variables with instances of heteroskedasticity; found in many of our relationships. It is crucial to correct for heteroskedasticity to ensure that our correlations are not under or over estimated. For example, when examining user votes number, our linear regression is not very good at making predictions when user votes number is low, but our predictions get more accurate as user votes number increases. Using coefficient tests, we transformed the regression standard errors to eliminate heteroscedasticity. Correcting heteroskedasticity in each of our variables provides more accurate predictors in our model, leading to more authentic results.

3 Model Selection

After testing for linearity, collinearity and outliers, we were ready to begin crafting our statistical model. First, we looked at modeling continuous and categorical variables. Using residual plotting, we found the most accurate relationship between continuous variables and IMDb score (Appendix 10). Whether it was polynomial, spline, or linear, we selected the method which yielded the highest r-squared to maximize the predictive power of our variables. We immediately saw duration and number of votes as two of the most important continuous variables, due to their high r-squares, while also identifying other variables of note to include.

Hold on, how could we use the number of user votes to predict the IMDb score, before the movie has even come out? We started to give some thought to our predictors. By using logical reasoning, we eliminated the number of user votes as the data would not be available prior to a movies release and therefore can not be used to predict IMDb score. We went through every variable and used logical reasoning to determine if it would suitable to include it in our model.

We also identified a problem with using any of the variables involving Facebook likes, because after a movie is released, the actors or directors Facebook likes is likely to rise. For example, before Harry Potter, Daniel Radcliffe was not a popular actor. Now, he has over 400,000 Facebook likes. A movie with a lead actor with few Facebook likes prior to movie release should not be measured against an actor that has already had a major movie released. Critic reviews and user reviews number must be excluded as well, as they would not be available prior to release.

In the end, we only selected duration and number of faces in a movie poster as plausible continuous variables, because no other continuous variables yielded results that lowered our MSE and made logical sense. As for the categorical variables, our two variables that improved the model most dramatic were directors and distributors. Once we included them in our model, our MSE became drastically lower. Surprisingly, none of the variables involving actors helped lower our MSE. Further, we were convinced that the movies genre would have an impact on our MSE, so we crafted the best combination of genres. Some were not statistically significant enough to be included, and the remainder were selected via a trial and error process. An interesting point worth noting is that we first built the model trying to reach the highest possible r-squared. We were successful and found an r-squared of almost 0.99. But, when we tested its MSE, it yielded 1.5 and we were very disappointed. We were obviously overfitting, and started to understand that the relationship between r-squared and the MSE was not as straightforward as we thought. We then made the decision to focus on the MSE rather than the r-squared, as MSE is the best indicator of the accuracy of our a predictive model.

4 Results

The first model we ran had an MSE 1.51, which wasnt bad, but we knew we could do better. After experimenting with different predictors and regression techniques, we managed to construct a model with an MSE of 0.78, that had a small amount of error, and utilized variables that made it capable of predicting a movies IMDb score before it is released. To determine the various MSEs of our models and test the out of sample performance, we used the k-fold cross-validation technique. This technique was significantly quicker than LOOCV, and allowed us to train the model on a larger data set than if we employed a validation set test, making it a happy medium that could satisfy the criteria of accuracy and simplicity. We decided on using k=46, using the baseline square root of our number of observations and ensuring that the data was split enough times to get an accurate average test error.

We then tested our model for heteroskedasticity and collinearity. We discovered that the sum of total likes and cast toal likes are collinear, so we removed cast total likes in order to yield the best results. We also found evidence of heteroskedasticity in our model, so we removed it with an ncvTest. After taking these corrective measures, we were pleased with the results and integrity of our model.

Our final model (Appendix 11-13) included a combination of continuous and categorical variables. Of the variables we included in our model, the genre variables action, thriller, romance, horror, drama, animation and musical were incredibly statistically significant. In terms of effect, drama, and especially animation, caused movies to have higher IMDb scores. The other genres mentioned negatively affected IMDb scores, with horror having the strongest negative relationship. This makes sense, as horror movies are often too scary, and not appealing to the general public. Duration was an especially interesting predictor; it improved IMDb score, but eventually started to have a negative relationship, staying statistically significant throughout. This makes a lot of sense, as viewers dont like movies that are too long or too short.

Of the directors, Woody Allen and Steven Spielberg were the most statistically significant, likely due to their large bodies of work. The two of them, along with Tim Burton, were the directors who the model was most favorable to; if they direct a movie, the model predictors it will have an IMDb score that is roughly one point higher than the excluded category. Only two directors, Michael Bay and Spike Lee, had a negative effect on movie ratings, but both observations were deemed to be statistically insignificant. Interestingly, the number of faces in a movie poster has a notable negative impact on the predicted IMDb score of a movie, as does colour, with black-and-white movies receiving higher scores. Both of these predictors were very statistically significant. Most of the movie distributors are not particularly statistically significant, with the exception of Dimension Films and Lionsgate which are, and have a negative impact on predicted IMDb score relative to the excluded category.

Overall, we are pleased with our models predictive power. If we had more time, we would have improved our model by dummifying and including all of the genres, including comedy, history, sci-fi, and sports. Given the statistical significance and predictive power of the genre variables we did include, we believe our model would have made more accurate predictions if we had we included these other genres of movie. As a result, we are far more confident in predicting movies where our model includes all of their listed genres. Two examples are Postcards From London (drama) and The Long Dumb Road (comedy), which we expected to predict quite accurately.

Now that we have confirmed our final model and its ability to make accurate predictions, we can begin predicting the IMDb score for upcoming holiday movies. For the predicted IMDb scores, see Table 1 (Appendix 14). The mean of our 12 predictions was 6.61, which is only marginally higher than the mean IMDb score of 6.52. Interestingly, all of our films were predicted to have a score between 5.88 and 7.51, which makes sense for our normally distributed response variable.

Upon examining our predictions of upcoming movies, we found some problems in our model that we would want to account for in the future. Firstly, having many faces in a movie poster, for example 78 in the case of Ralph Breaks the Internet (if you include the minions), leads to an artificially low IMDb score. We adjusted and included only the faces of humans, but nonetheless, this revealed a problem with our model. If a movie poster has over 100 faces, our model is likely to predict an extremely low IMDb score. Another issues with our model is that to ensure proper consistency, ideally all of the data should have been gathered at the time of release for many of the variables. That way, movies that have been out longer will not have higher results and an advantage in areas like Facebook likes and movie reviews.

Additionally, we would have liked to include information on sequels in our model. For example, our prediction for the upcoming movie Creed II does not take into account the IMDb score of the first Creed film. We would be interested to see how movies in sequels are related, and if they can help predict the IMDb scores for upcoming sequels.

Although we trust our beautifully crafted model, we wanted to do some more external research into what effects an IMDb score. We retrieved a peculiar result: men. Yes thats right, men have one of the most significant impacts on the IMDb score a movie receives. This is because almost all movie reviewers and critics are male, resulting in movies with masculine themes and male lead actors being more likely to receive a high score. For this reason, we would in the future want to look particularly at the masculinity of plots and genders of lead actors, as we believe it would be a strong factor in predicting IMDb score's of movies released in the future.

5 Appendix

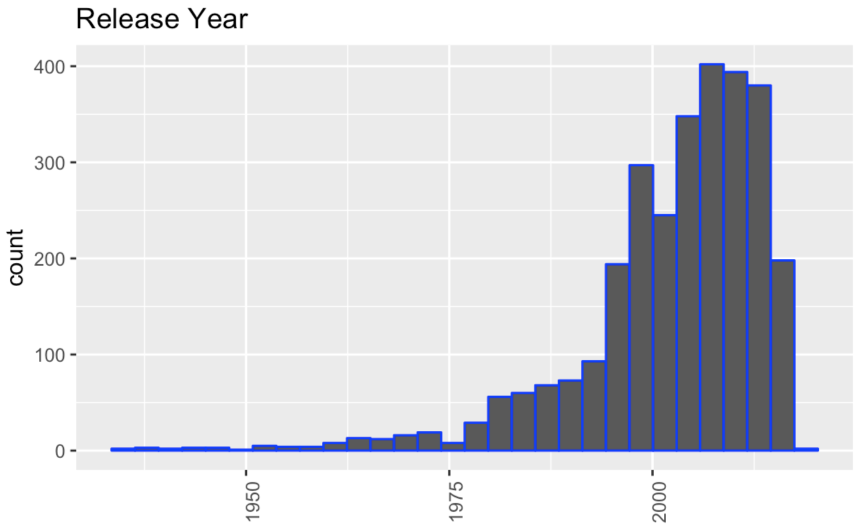


Figure 1: A histogram showing the frequency of movie releases by year; increased significantly in the past 20 years

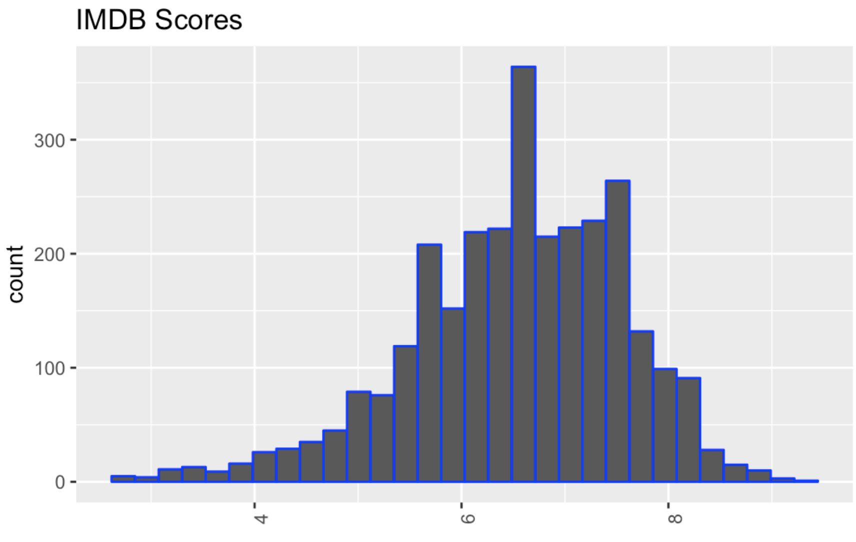


Figure 2: A histogram showing that the distribution of IMDb scores is fairly normally distributed

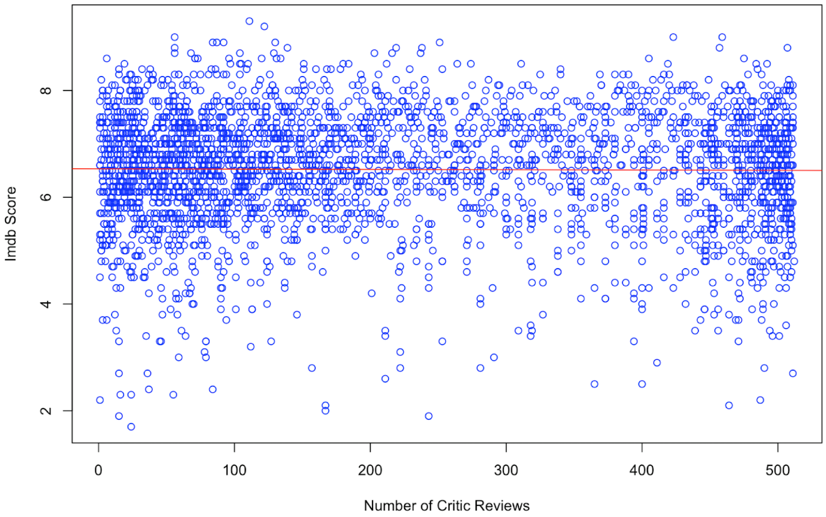


Figure 3: A plot showing the lack of correlation between number of critic reviews and IMDb Score

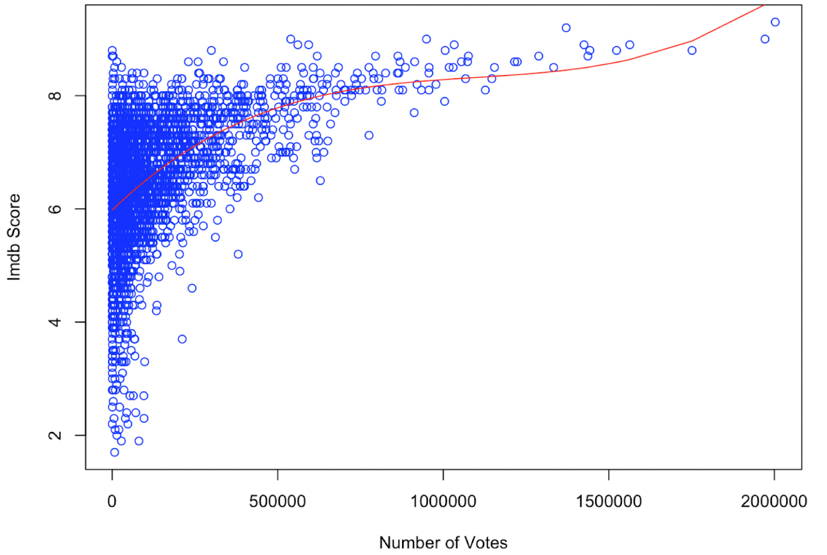


Figure 4: A plot showing the positive correlation between number of votes and IMDb score

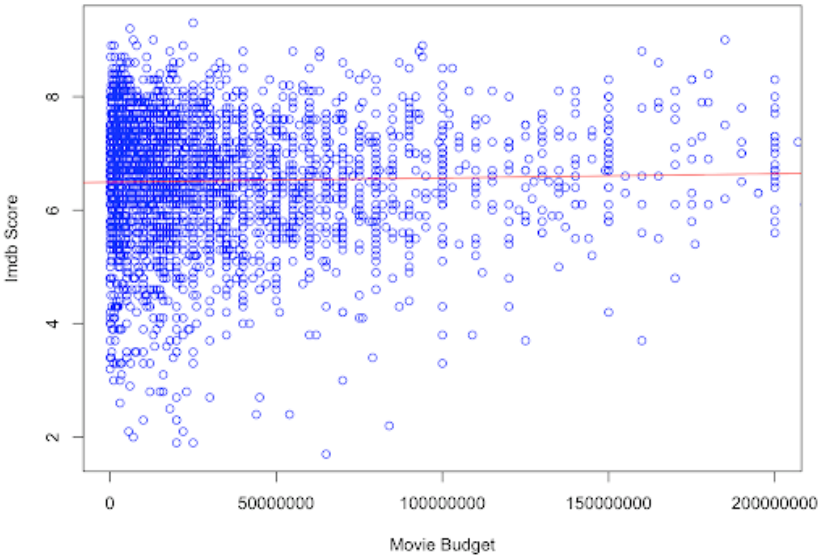


Figure 5: A plot showing the lack of correlation between movie budget and IMDb score

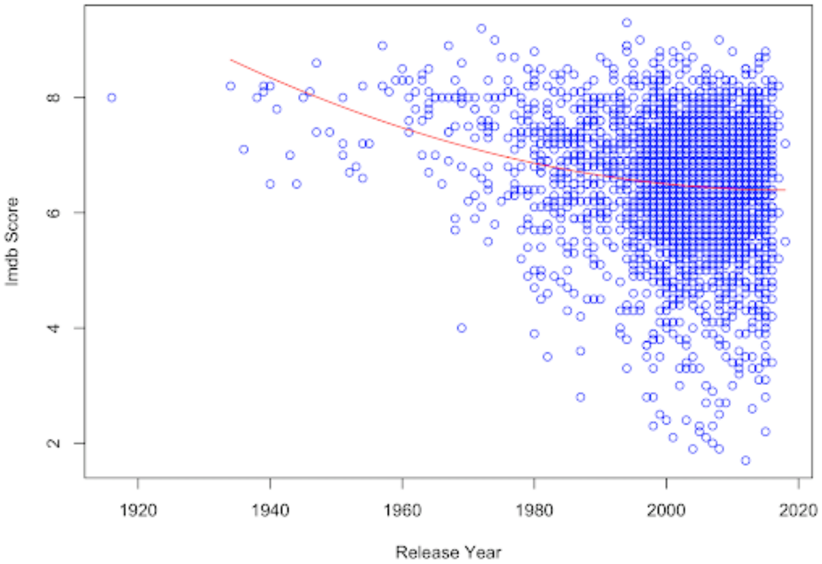


Figure 6: A plot showing the negative correlation between release year and IMDb score

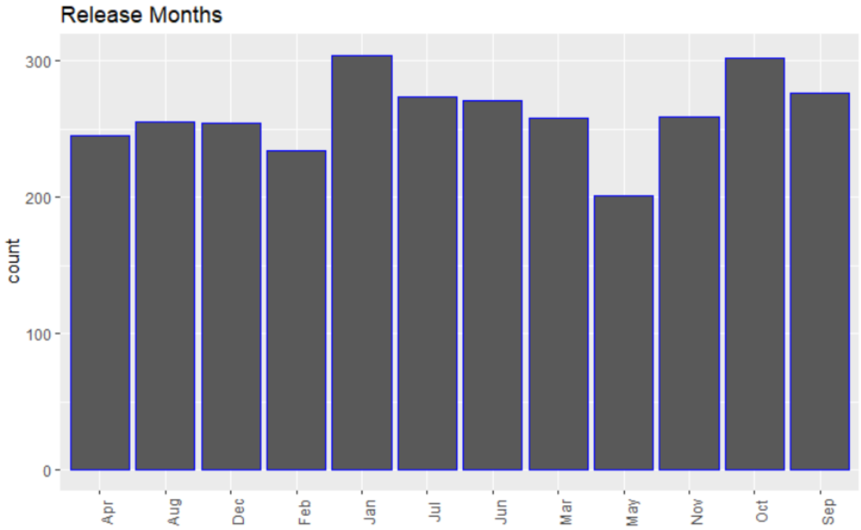


Figure 7: A histogram showing movie releases distribution across the months of the year

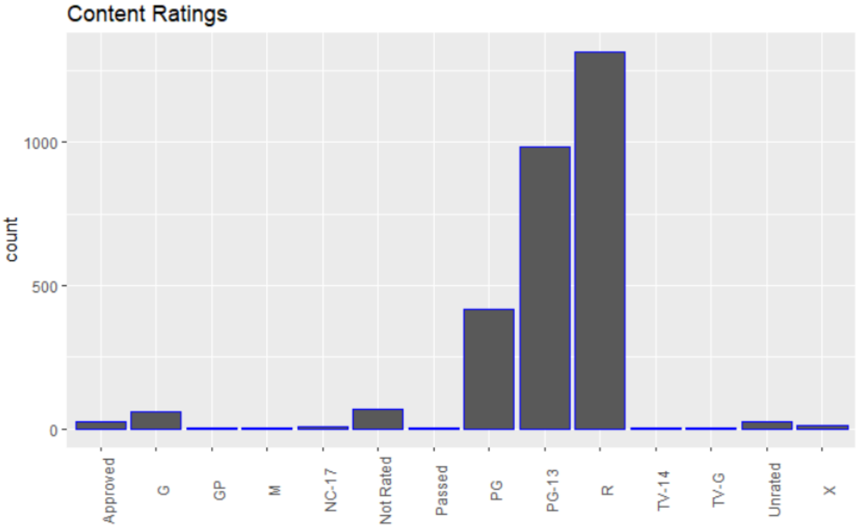


Figure 8: A histogram showing movie releases distribution based on content rating

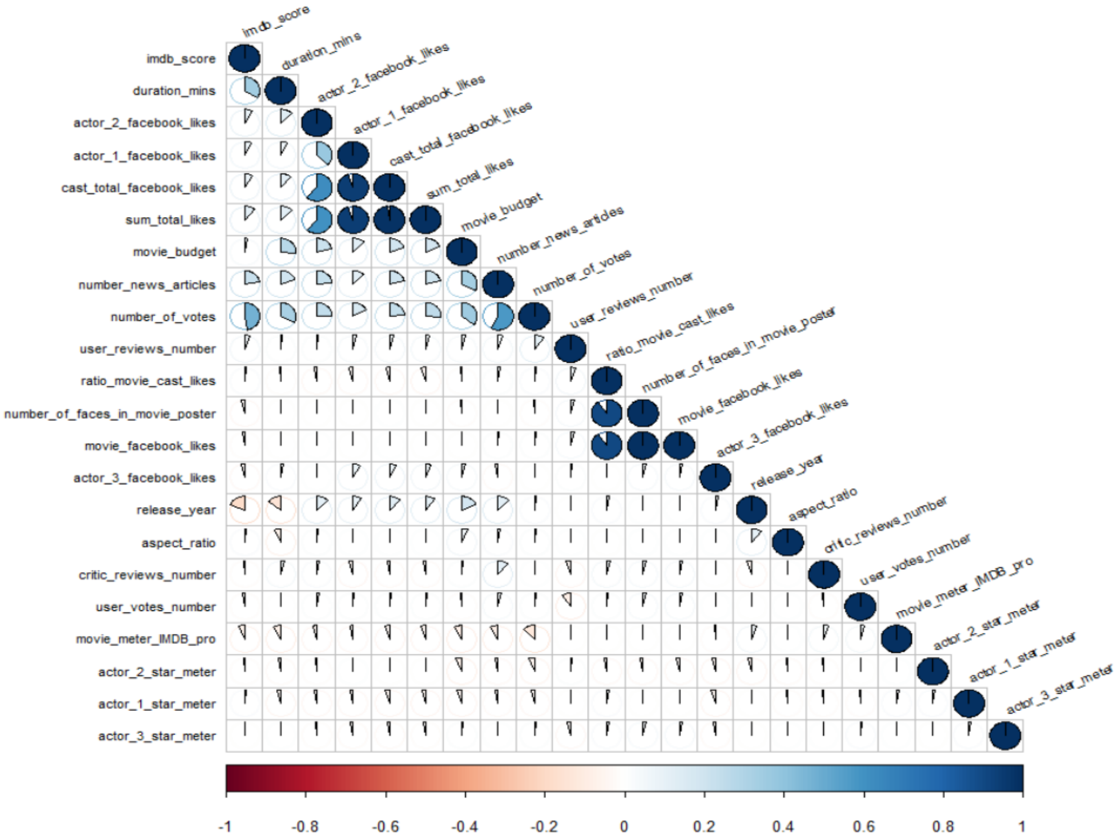


Figure 9: A correlation matrix showing collinearity between continuous variables

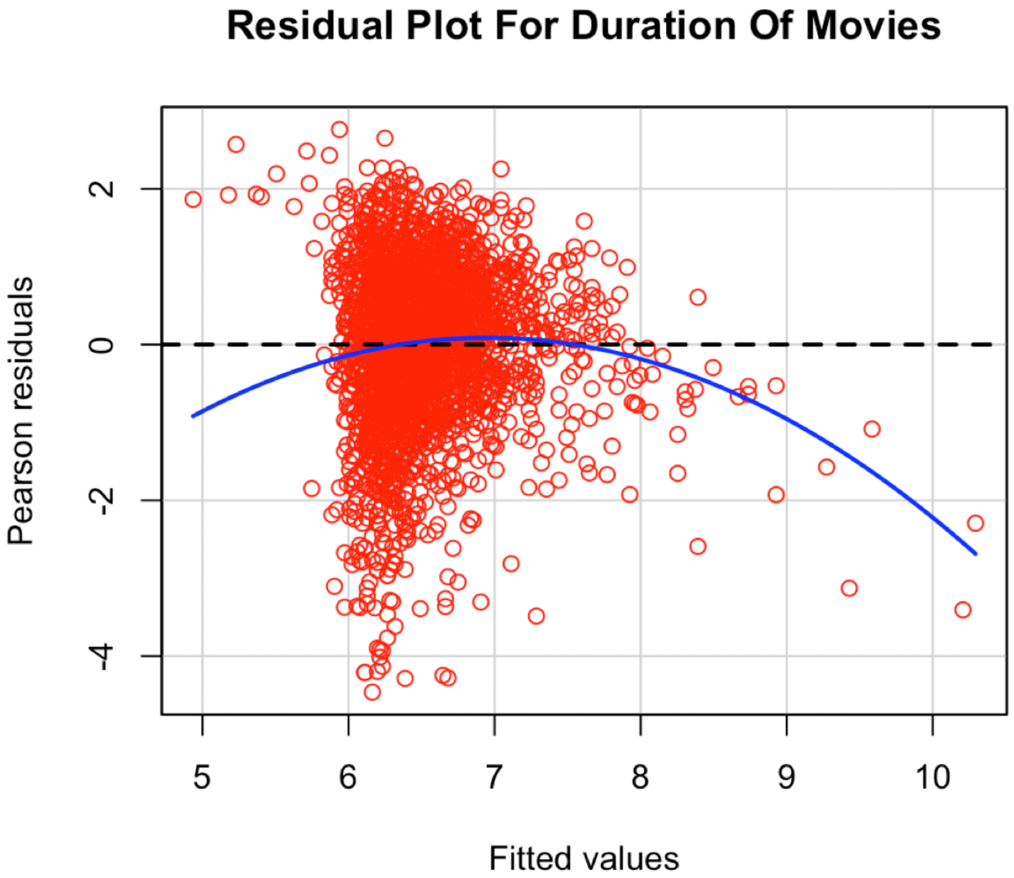
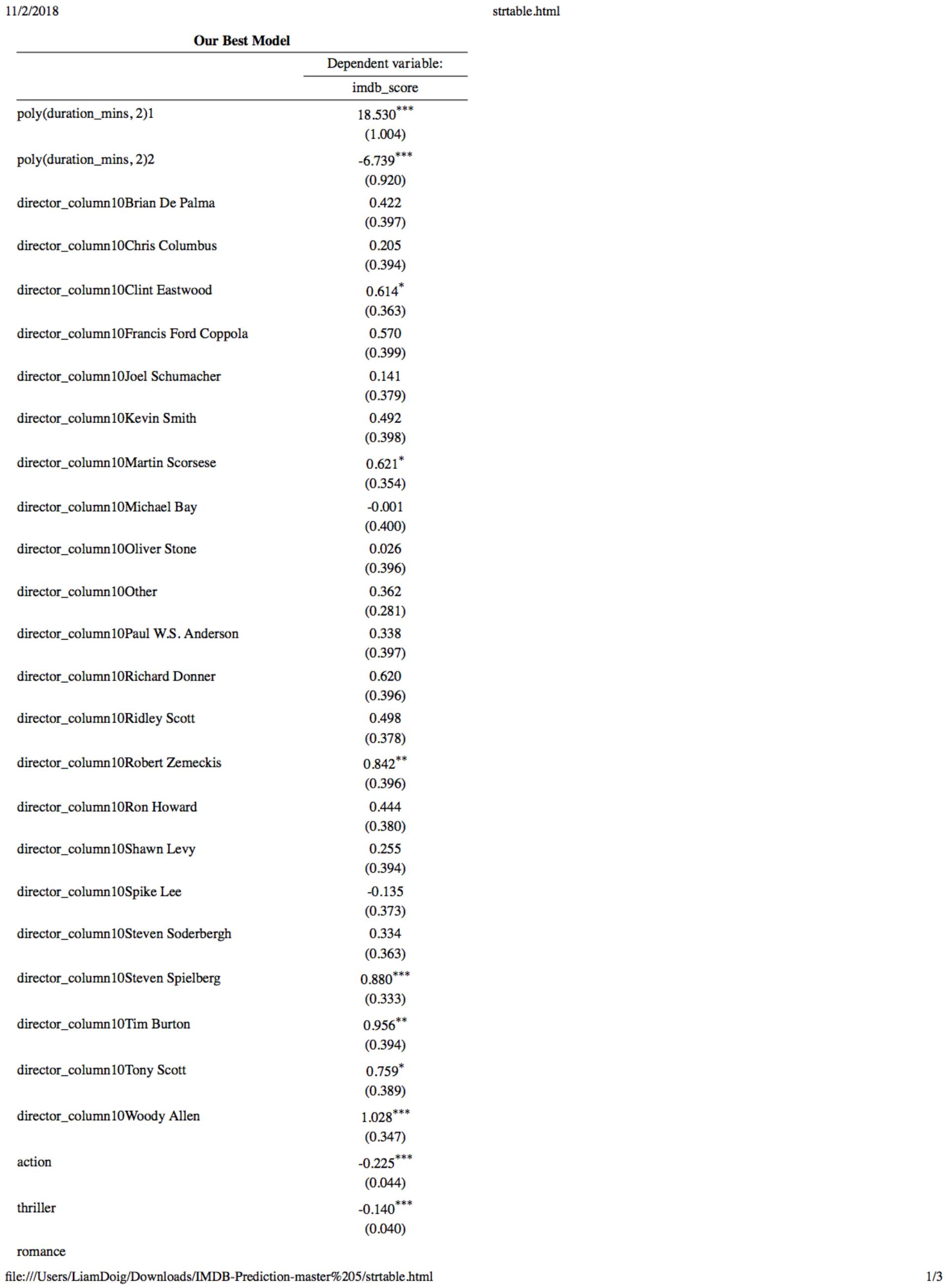


Figure 10: A residual plot of movie duration and IMDb score, showing a non-linear relationship



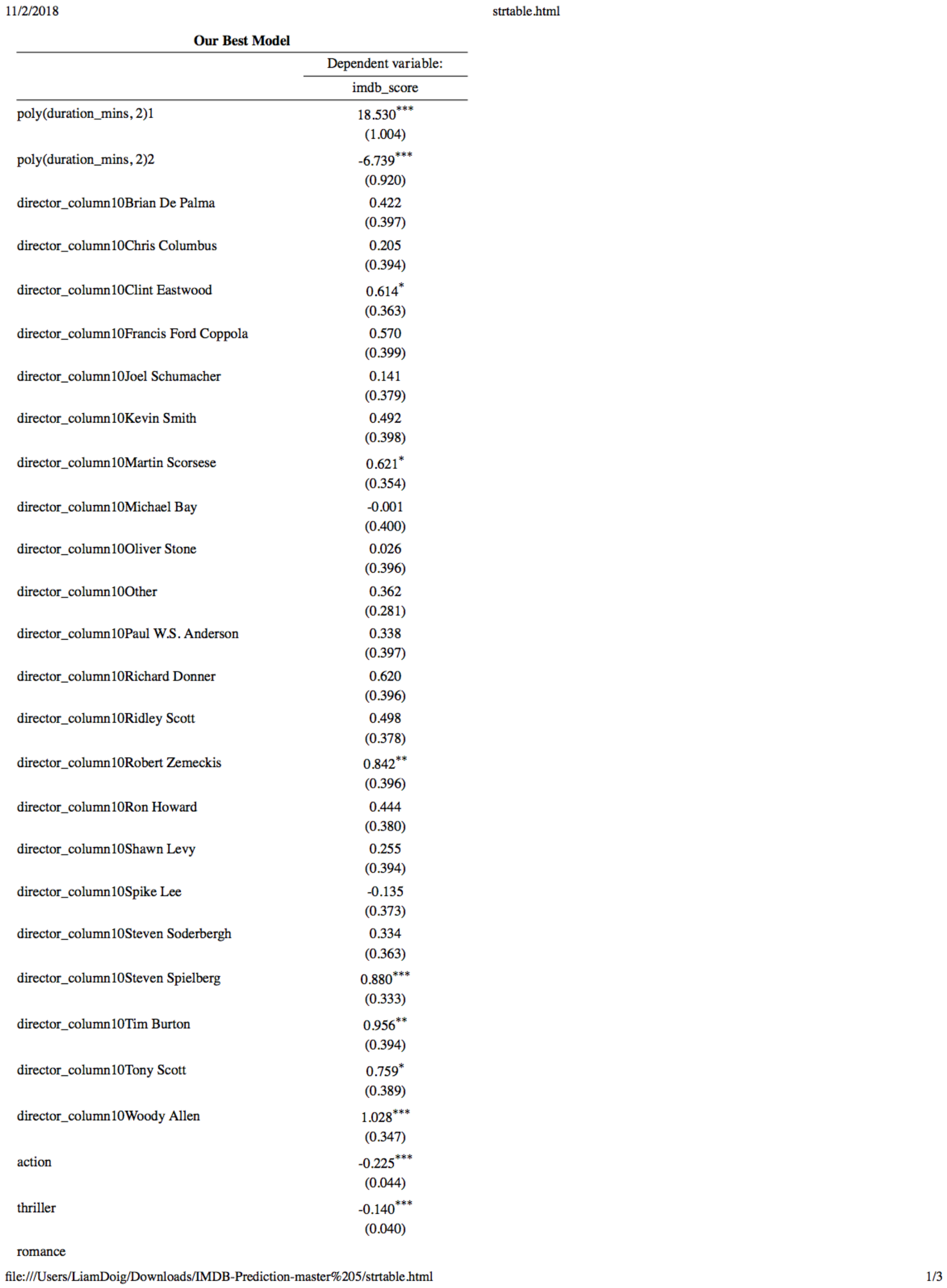
11/2/2018 strtable.html

Our Best Model

	Dependent variable:
	imdb_score
poly(duration_mins, 2)1	18.530*** (1.004)
poly(duration_mins, 2)2	-6.739*** (0.920)
director_column10Brian De Palma	0.422 (0.397)
director_column10Chris Columbus	0.205 (0.394)
director_column10Clint Eastwood	0.614* (0.363)
director_column10Francis Ford Coppola	0.570 (0.399)
director_column10Joel Schumacher	0.141 (0.379)
director_column10Kevin Smith	0.492 (0.398)
director_column10Martin Scorsese	0.621* (0.354)
director_column10Michael Bay	-0.001 (0.400)
director_column10Oliver Stone	0.026 (0.396)
director_column10Other	0.362 (0.281)
director_column10Paul W.S. Anderson	0.338 (0.397)
director_column10Richard Donner	0.620 (0.396)
director_column10Ridley Scott	0.498 (0.378)
director_column10Robert Zemeckis	0.842** (0.396)
director_column10Ron Howard	0.444 (0.380)
director_column10Shawn Levy	0.255 (0.394)
director_column10Spike Lee	-0.135 (0.373)
director_column10Steven Soderbergh	0.334 (0.363)
director_column10Steven Spielberg	0.880*** (0.333)
director_column10Tim Burton	0.956** (0.394)
director_column10Tony Scott	0.759* (0.389)
director_column10Woody Allen	1.028*** (0.347)
action	-0.225*** (0.044)
thriller	-0.140*** (0.040)
romance	

file:///Users/LiamDoig/Downloads/IMDB-Prediction-master%205/strtable.html 1/3

Figure 11: A summary of our final model's results (part1)



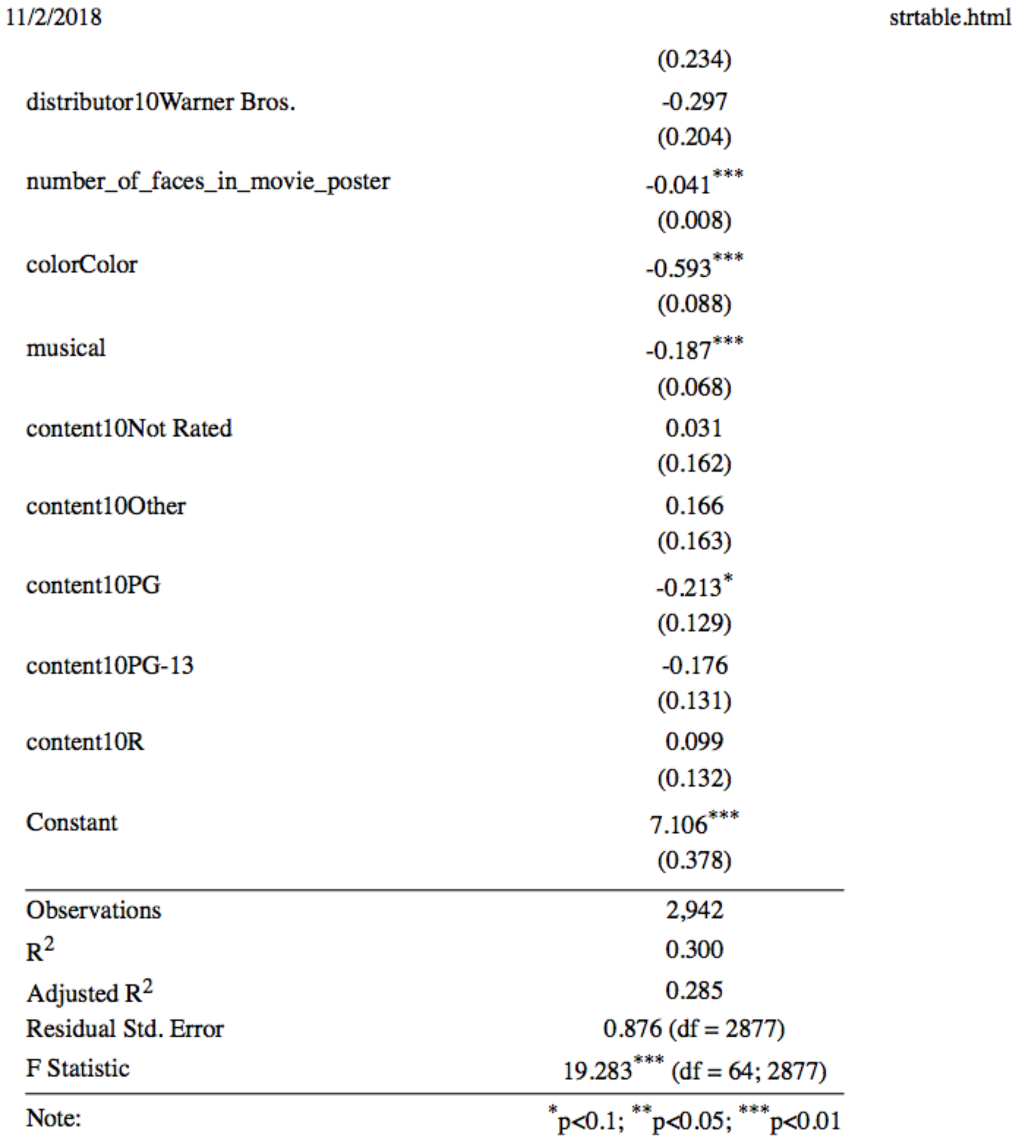
11/2/2018 strtable.html

Our Best Model

	Dependent variable:
	imdb_score
poly(duration_mins, 2)1	18.530***
	(1.004)
poly(duration_mins, 2)2	-6.739***
	(0.920)
director_column10Brian De Palma	0.422
	(0.397)
director_column10Chris Columbus	0.205
	(0.394)
director_column10Clint Eastwood	0.614*
	(0.363)
director_column10Francis Ford Coppola	0.570
	(0.399)
director_column10Joel Schumacher	0.141
	(0.379)
director_column10Kevin Smith	0.492
	(0.398)
director_column10Martin Scorsese	0.621*
	(0.354)
director_column10Michael Bay	-0.001
	(0.400)
director_column10Oliver Stone	0.026
	(0.396)
director_column10Other	0.362
	(0.281)
director_column10Paul W.S. Anderson	0.338
	(0.397)
director_column10Richard Donner	0.620
	(0.396)
director_column10Ridley Scott	0.498
	(0.378)
director_column10Robert Zemeckis	0.842**
	(0.396)
director_column10Ron Howard	0.444
	(0.380)
director_column10Shawn Levy	0.255
	(0.394)
director_column10Spike Lee	-0.135
	(0.373)
director_column10Steven Soderbergh	0.334
	(0.363)
director_column10Steven Spielberg	0.880***
	(0.333)
director_column10Tim Burton	0.956**
	(0.394)
director_column10Tony Scott	0.759*
	(0.389)
director_column10Woody Allen	1.028***
	(0.347)
action	-0.225***
	(0.044)
thriller	-0.140***
	(0.040)
romance	

file:///Users/LiamDoig/Downloads/IMDB-Prediction-master%205/strtable.html 1/3

Figure 12: A summary of our final model's results (part2)



11/2/2018 strtable.html

	(0.234)
distributor10Warner Bros.	-0.297
	(0.204)
number_of_faces_in_movie_poster	-0.041***
	(0.008)
colorColor	-0.593***
	(0.088)
musical	-0.187***
	(0.068)
content10Not Rated	0.031
	(0.162)
content10Other	0.166
	(0.163)
content10PG	-0.213*
	(0.129)
content10PG-13	-0.176
	(0.131)
content10R	0.099
	(0.132)
Constant	7.106***
	(0.378)
Observations	2,942
R^2	0.300
Adjusted R^2	0.285
Residual Std. Error	0.876 (df = 2877)
F Statistic	19.283*** (df = 64; 2877)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 13: A summary of our final model's results (part3)

Movie Title	**Predicted IMDb Score**
The Grinch	6.93
Postcards from London	6.44
The Long Dumb Road	6.25
Instant Family	5.88
Crimes of Grindelwald	6.68
The Clovehitch Killer	6.66
Robin Hood	5.88
Creed II	6.77
Welcome to Marwen	7.04
Ralph Breaks the Internet	7.51
Second Act	6.11
Becoming Astrid	7.22

Table 1: A table showing our predicted IMDb scores for the movies coming out this season