HW2 Write-up

Mandy Guo 48340673, Xiaoyu Ji 48639648, Yiqing Sha 44691573, Qian Zhao 48666701

11/25/2021

Introdiction

In this report, we aim to tease apart what factors contribute to the classification of "good" and "bad" movies based on their ratings. The variables we use for our prediction includes (director, gender of the director, cast, budget and revenue). Understanding the relationship between movie quality and contributing factors can help production companies to predict movie ratings and maximize a movie's chance of success during the creation phase.

To conduct the analysis, we utilized the metadata and credits dataset. There are 45,186 observations in the metadata including characteristics such as budget, genre, revenue, vote of the movies. For the credits dataset, there are 45,476 observations including the cast and crew JSON columns of the movies. We have a relatively large dataset and only small portions of null values. We dropped some columns before joining the data for generating the logistic regression model to give our data simplicity and clarity.

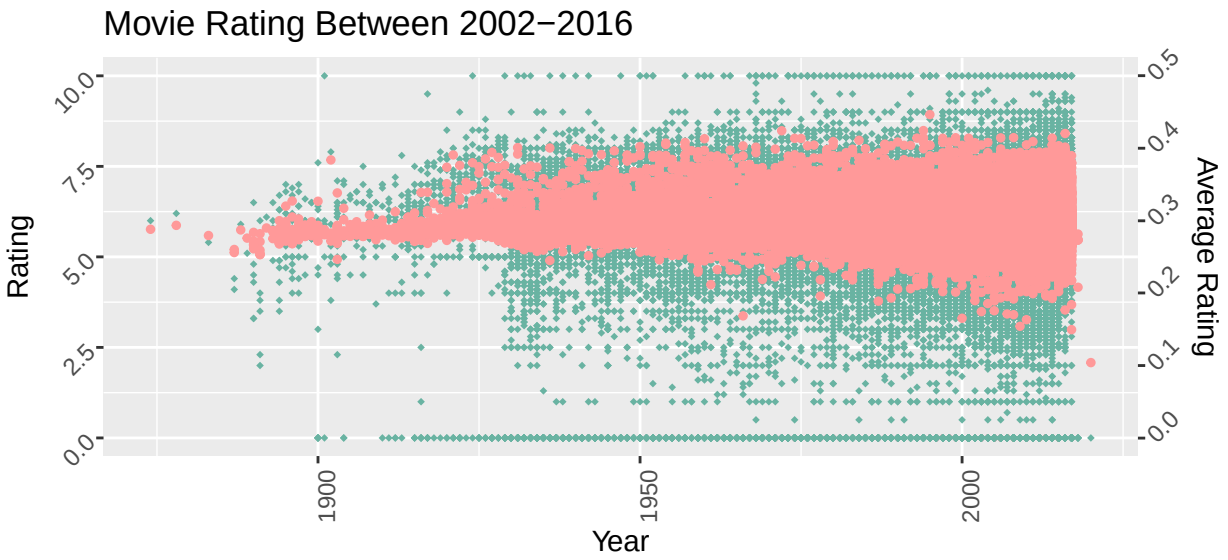
Cast and Directors

To get cast and crew information for analysis, we parsed them from the credits dataset. We eliminated our cast variable to only the top two actors in each movie. We also include the director and director_gender variables in our model. To include these variables in our analysis, we defined top cast as the actors / actress who appears in more than 10 movies in the dataset. Similarly, we defined top director as the ones who produced more than 5 movies in the dataset. 2 Dummy variables are created indicating the presence of top cast and top directors.

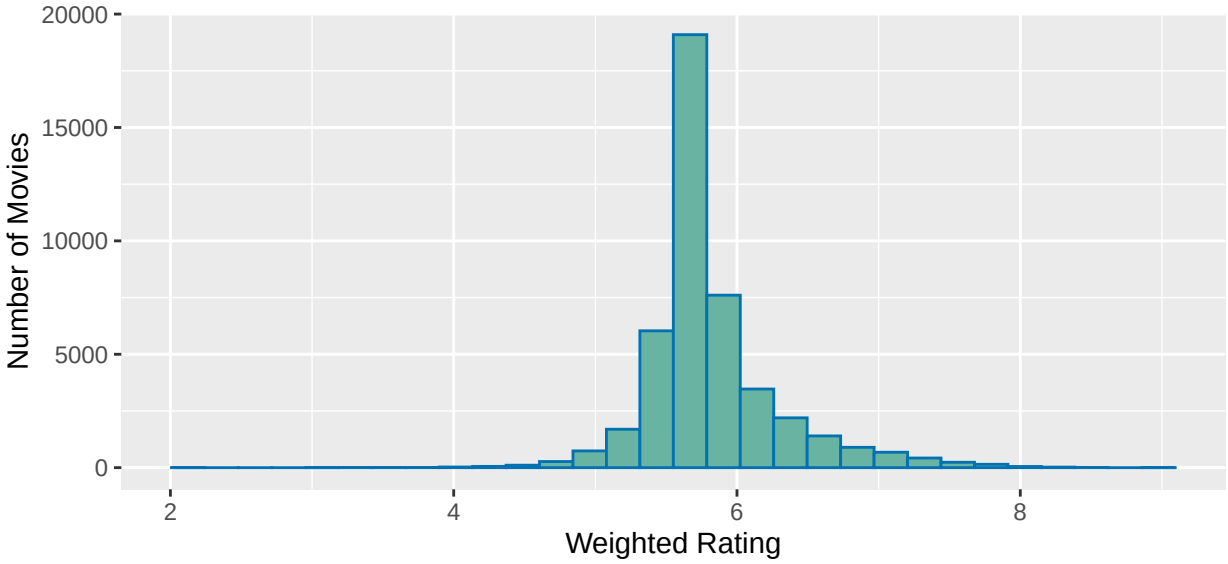
Movie Ratings

There are two sets of ratings in our dataset, vote_average (on a scale of 0-10) in the metadata and the rating (on a scale of 0-5) in the ratings data. In general, very high or very low vote_averages are based on small numbers of votes per movie. So we generated a new variable of weighted_rating with both vote_average and vote_count taken into consideration.

In the scatter plot of Movie rating between 2002-2016, we have more ratings after 2000 and more disperse ratings before 1950 for both two types of ratings. The ratings on a scale of 0-5 have more extreme values with ratings of 0 and 5, while the weighted_rating is comparatively normally distributed and has less extreme values. So we choose to use our new variable of weighted_rating for determining whether a movie is good or not.



From the histogram below, we could get an overview of the distribution of the weighted ratings. The most frequent ratings in the dataset are between 5 and 6. We choose a separate point of 6 or more for "good" movies, which include roughly 22.19% of all movies in the dataset.



We found that the top 10 highly rated movies vary in different aspects (see table below). Some are well-known great movies like *The Shawshank Redemption* and *The Godfather*. There are award-winning Bollywood romantic movie and BBC documentary, as well as Japanese animated fantasy films. This makes our analysis more interesting for production companies to understand the potential contribution factors of a highly rated 'good' movie.

```
##                             title weighted_rating
## 8860  Dilwale Dulhania Le Jayenge            8.93
## 224       The Shawshank Redemption            8.49
## 187                   The Godfather            8.48
## 43449                    Your Name.            8.41
## 113                The Dark Knight            8.29
## 415                     Fight Club            8.29
```

```
## 536                  Pulp Fiction            8.29
## 35876                Planet Earth            8.29
## 90                  Spirited Away            8.28
## 323              Schindler's List            8.28
```

The Logistic Regression Model

In the final logistic regression model, we included budget, revenue, presence of top cast (who appears in more than 10 movies), presence of top director (who produced more than 5 movies) and the director's gender to make the prediction.

In our final dataset, the percentage of good movies is 24.08%. To understand the accuracy of the model, we produced a classification matrix as showing below to compare the predicted and actual values. Based on the matrix, the prediction accuracy is at 77.35%. We are quite confident about model performance.

```
##
## pred            0     1
##   badMovie  65267 18926
##   goodMovie   776  2026
```

The coefficient table of the model is shown below. Based on an intercept of -1.73 we could calculate the probability of good movie on the 'baseline' situation when budget and revenue are 0, and no presence of top cast and top director, and the gender of director is unknown is 15.07%. The coefficient table also indicates that the variables in the model had significant impacts on predicting a good movie. As we looked further into each of these variables, specific conditions were not as impactful as others. Budget and revenue are both significant, but the coefficient is 0, which indicates that both budget and revenue of the movie do not contribute much to the movie rating. Top_cast, top_director along with director's gender are all positively contributing to the prediction of a good movie.

To quantify the effect of these factors, we calculated probability based on different conditions. When holding other factors same as the 'baseline' situation, presence of top director will increase the probability of good movies to 21.27%. While the presence of top cast will increase the probability to 18.04% only. Comparing to the situation when director's gender is unknown, knowing the gender of the director will increase the probability of good movies to around 18% as well. While the probability is slightly higher for male directors than female directors. We then further assume a scenario that a production company is investing 5 million dollars in producing a movie with a top male director and top cast and expecting a revenue of 10 million. The probability of this movie to be good is 32.07% based on our model. The probability under this scenario is doubled compared to the 'baseline' situation.

Due the non-linear nature of logistic models, when we covert coefficients to probabilities, generally we should consider that the impact on the probability should depends on where the change occurred. However, in our case, the most impactful variables (i.e. presence of top casts and top directors) are binary. We only consider the change between 0 and 1.

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.73       0.02 -106.90        0
## budget               0.00       0.00    3.13        0
## revenue              0.00       0.00   29.09        0
## Top_cast             0.21       0.02   12.11        0
## Top_director         0.42       0.02   23.44        0
## director_gender1     0.22       0.04    4.99        0
## director_gender2     0.25       0.02   12.74        0
```

From our results, we suggest movie production companies to focus more on the selection of casts and directors when producing a movie to make it more likely to be rated well. It might be explained by the fact that top

directors are more experienced in the production, and they tend to be more aware of the characteristics of a good movie. While top casts may act better than the other casts, and potentially bring a superstar effect to viewer's perception of the movie.

Conclusion and Limitations

Our analysis on prediction of good movies has given production companies a few factors to consider when they decide to produce a highly rated movie. Based on our analysis, we suggest that top casts and top directors will contribute the most to a good movie. Although our model suggests that budget and revenue only have minimal effect in predicting a good movie. They should also be considered by the movie production companies. This is because that for many situations, budget of the movie is correlated to the presence of top directors and top casts. It could be explained by the fact that for most commercial movies, top directors and top casts cost more money to hire. These factors should be considered simultaneously in practice since the actual situation tends to be more complicated.

Apart from the factors mentioned above, we should also consider the possibility of endogeneity. In our case, there might be a potential omitted variable bias. For example, good movie script might be correlated to both the movie rating and independent variables such as budget, revenue, and presence of top director. There is no evidence showing this bias is present in our analysis, but production companies should take it into consideration when using the model results to make decisions.