Lab2 - Team Turquoise

Samantha Williams, Kate Kostelc, Mick Rejniak, Sean Baughman

3/20/2022

Contents

1 Introduction

As researchers for a start-up video streaming service company, we want to maximize viewer subscriptions to our new platform by offering the highest-rated films. Our business model is based on the idea that the more desirable our inventory (i.e., highest-rated), the greater our subscriber base, the greater our subscription revenue. However, by the time a film falls into the highest-rated category, licensing fees skyrocket. To maximize profit and minimize licensing costs, we want to predict which early releases will eventually become the highest-rated.

To this end, we will analyze the Internet Movie Database (IMDb) movie data to understand how different movie attributes affect the *meta score* of a movie. The meta score represents the overall, average rating of the movie. Our examination will include factors such as gross receipts, IMDb rating, run time, etc. using a selection of the top 1,000 movies.

2 Data Set and Research Design

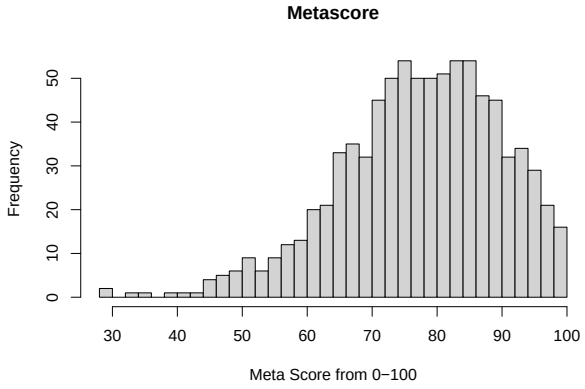
To build a model that determines the titles that should be included in our online streaming platform's inventory, we will utilize Kaggle's IMDb Movies data set that contains 1,000 records with 16 columns that describe the top 1,000 highest IMDb-rated movies from 1920 through 2020. In 1990, IMDb began as fan-operated database of mostly films and television series and has grown to include information related to video games and streaming content online. It is now a subsidiary of Amazon. As of March 2022, the database contained approximately 10 million titles (605,284 movie titles) with more than 6.7 million user reviews. According to IMDb, the site has 83 million registered users that use its free and pro versions of the site.

2.1 Data Format and Access

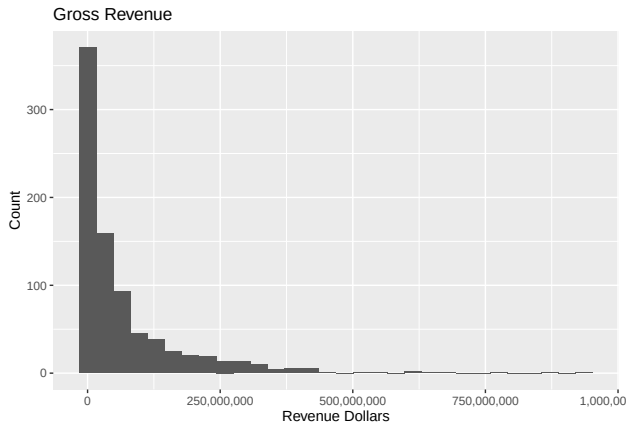
IMDb does not have an API for automated queries so this data is scrapped directly from the IMDb website and then provided by Kaggle with movie-related data provided in English. For our purposes we will be interested in the following columns provided by the Kaggle data set:

- `Series_Title`: Name of the movie.
- `Released_Year`: The year in which that movie released.
- `Runtime`: Total run time of the movie in minutes.
- `Genre`: Indicates the genre of the movie. A film may fall into more than one category of film genre. We have broken this out into single variables to analyze genre as it relates to our variables of interest.
- `IMDB_Rating`: Registered users are invited to rate titles on a scale of 1 to 10 on any released title within the database. A title is considered released if it was shown publicly at least once. IMDb applies a weighted mean to determine the final rating. Prior formulas were published and used the equivalent of a Bayesian posterior mean. However, their current formula is unknown, but IMDb does state that their current formula will prevent ballot-stuffing. Each user may vote as often as they'd like, but any new vote on the same title will overwrite their previous vote. The film with the highest IMDb score (9.3) is The Shawshank Redemption. There are no missing values for this section.
- `Meta_score`: Metascore is the rating of a film ranging from 0 to 100 based on the weighted average of at least four professional critics' reviews from newspapers, magazines, or other publications. The higher the score, the more positive the movie review. After examining this variable in our data set,

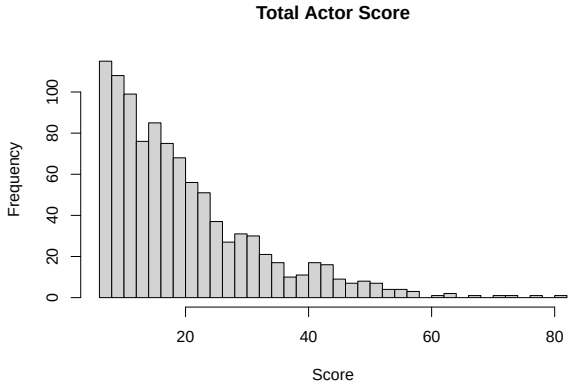
there are 157 missing values. This data set contains a film with the lowest score of 28 and 12 films with the maximum score of 100.



- `Director`: Name of the Director.
- `No_of_votes`: The total number of users who have voted for the film.
- `Genre`: Indicates the genre of the movie. A film may fall into more than one category of film genre. We have broken this out into single variables to analyze genre as it relates to our variables of interest.
- `Gross`: This is money earned by the movie. This is not adjusted for inflation or changes in movie ticket prices. These are rough estimates and are based on cumulative figures as compared to weekend gross. It is not clear from this data set if the gross is referring to worldwide, Non-USA, or USA only figures. We will assume that they are USA figures since that is the default displayed. Gross has a range of \$1,305 to \$936,662,225 with 169 missing data points. It does appear that franchise films seem to be the most financially successful based on our initial EDA.



- `Star1`,`Star2`,`Star3`,`Star4`: There are four actors listed for each film. We have created a dummy variable called `Total_Actor_Score` to identify any advantage of having a certain actor or a combo of actors as part of the cast of the film.



2.1.1 Calculating Total Actor Score

Each movie has four actors that starred in the movie listed in the variables `Star1`, `Star2`, `Star3`, `Star4`. Given the number of unique actors (2709) and the size of our dataset (1000 movies), it wouldn't be feasible to produce dummy variables to indicate whether or not an actor starred in a movie. Instead, we formulated a score for each individual actor as described below:

$$Actor's\ Score\ Per\ Movie = IMDb\ Rating\ of\ the\ Movie \div Number\ of\ Actors\ in\ the\ Movie$$

An actor will receive points for the score of the movie (`IMDB_rating`), and these points are shared equally amongst the actors in the movie. Once a score for a single movie is calculated, the actor's respective points are then added to the actor's individual score.

$$Individual\ Actor's\ Score = \Sigma\ Actor's\ Score_{\ Each\ Movie}$$

Once all of the actor's individual scores have been tabulated, then each movie is given a score that is generated by summing each actor's individual score.

$$Movie's\ Total\_Actor\_Score = Actor\ 1_{Score} + Actor\ 2_{Score} + Actor\ 3_{Score} + Actor\ 4_{Score}$$

For example:

Interstellar's `IMDB_Rating` was 8.6, so each Actor in Interstellar receives 2.15 points for this movie, and that score is added up in their Actors Total Score. In this case, 2.15 points are added to Jessica Chastain's total score, which ends up being 4.150.

Then, to compute Interstellar's Total Actor Score, the following values were referenced and summed:

Actor	Individual Actor's Score
Matthew McConaughey	12.000
Anne Hathaway	6.150
Jessica Chastain	4.150
Mackenzie Foy	4.075
—	—
Interstellar's Total_Actor_Score	26.375

2.2 Research Design

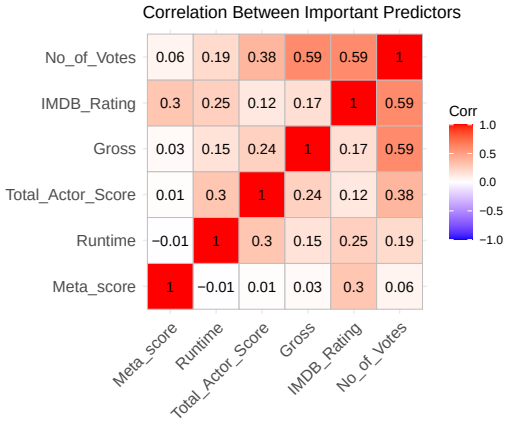
Our goal in conducting this study is to understand the relationship between a film's `Meta_score` and its `gross` and help us answer the question: "Does a commercially successful (`Gross`) movie become a highly rated movie (`Meta_Score`)?" Our investigation will focus on whether or not there's a positive correlation between a movie's ability to make money and its critical acclaim.

To help us dive into this question we will use the variable `Meta_score` as our outcome variable of interest. This was an ideal choice as most professional critics films are permitted to watch the film well before they are released publicly in order to write their reviews in time to generate buzz. It is relatively normally distributed, with a slight positive skew. The variable `Gross` will be our primary beta which is heavily skewed. We will examine the impact of additional variables `IMDB_Rating` and our calculated `Total_Actor_Score` in subsequent models. Our study takes care to focus on variables of movies that exist early in a movie's lifecycle to ensure that we are able to make decisions on a film as early as possible.

Ultimately, we will reject or fail to reject the null hypothesis that there is no impact of gross receipts on meta score.

2.3 Models

Based on the correlation plot, we will be evaluating three models.



2.3.1 Model One

For the first model we are examining only the primary beta without any covariates, the `Gross` variable:

$$Meta\_score = \beta_0 + \beta_1(Gross)$$

2.3.2 Model Two

For the second model we are examining the primary beta (`Gross`) with the covariate of `IMDB_Rating`:

$$Metascore = \beta_0 + \beta_1(Gross) + \beta_2(IMDBRating)$$

2.3.3 Model three

For the second model we are examining the primary beta (`Gross`) with the covariates of `IMDB_Rating` and `Total_Actor_Rating`:

$$Metascore = \beta_0 + \beta_1(Gross) + \beta_2(IMDBRating) + \beta_3(TotalActorScore)$$

3 Results

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Apr 15, 2022 - 14:32:07

Table 2: Movie OLS Model Results

	Dependent variable:		
	Meta_score		
	(1)	(2)	(3)
Gross	−0.000	−0.000*	−0.000
	(0.000)	(0.000)	(0.000)
IMDB_Rating		12.592***	13.019***
		(1.525)	(1.541)
Total_Actor_Score			−0.066*
			(0.036)
Constant	77.713***	−21.891*	−24.024**
	(0.547)	(12.079)	(12.116)
Observations	750	750	750
R^2	0.001	0.084	0.089
Adjusted R^2	−0.0004	0.082	0.085
Residual Std. Error	12.496 (df = 748)	11.971 (df = 747)	11.952 (df = 746)
F Statistic	0.700 (df = 1; 748)	34.447*** (df = 2; 747)	24.144*** (df = 3; 746)
Note:			*p<0.1; **p<0.05; ***p<0.01

Our alternate hypothesis tested "Does a commercially successful (Gross) movie become a highly rated movie (Meta_Score)?", and we found that Gross revenue does not have a statistically significant effect on a movie's Meta_Score, and therefore, failed to rejet our null hypothesis.

Model 1, which solely looked at `Gross` related to `Meta_Score`, had zero (0.0000) effect as seen in Table X and produced a negligible R^2 of only 0.001.

In Model 2, we added in `IMDB_Rating` as a new beta, and we did find it to be statistically significant with a p-value of < 0.01. `IMDB_Rating` correlated positively with the `Meta_Score`, so an increase in an IMDB rating would increase a `Meta_Score.` The intercept of the model was negative, and this should be ignored because it's impossible to get a negative rating. The R^2 of Model 2 was only 0.084, so this still fails to account for even 10% of the overall variance in the dataset.

In Model 3, we added in `Total_Actor_Score` as a new beta, and it also affected `Meta_Score` with a statistically significant effect and a p-value of < 0.05. `Total_Actor_Score` was negatively sloped so it will draw the regression model slightly closer back to zero. Again, the R^2 of Model 3 was 0.089, so ~9% of the variance in this model is explained by these betas. The intercept was also negative, which should be ignored because it is not practically possible to get a negative rating.

In practical terms, this means that a movie's gross revenue has little or nothing to do with how successful it is rated according to its `Meta_Score`. Other factors such as `IMDB_Rating` and `Total_Actor` score did explain 8.9% of the variance in our dataset, but this R^2 score is still very low which means that are other factors that influence `Meta_Score` that this dataset was not able to explain.

In terms of our business problem, we should look to evaluate data elements like our customer's views of these

movies in our service and other indicators that might be able to predict outcomes with greater certainty. It is clear that Gross revenue is not a valid predictor. Just because people pay to see a movie, it doesn't mean that it's good. It's also hard to predict a movie will be successful in terms of a `Meta_Score`, as most of the data points are lagging indicators. We recommend further study and analysis in this area.

From this model, we can see that for a score of 77.71, the film's `Gross` will not have a perceived impact with a change of -0.000000003372 for every point increase.

3.1 Model Limitations

3.1.1 Statistical limitations of Model

A statistical concern, or limitation, of our model is the data's lack of independence. The movies in our data are produced through time. In other words, a movie that follows a preceding movie is very likely influenced by its predecessors. In addition, the production of high-grossing and popular movies tends to come from the movie industry in Hollywood, California. This insular production phenomenon can result in competitive and creative interactions, otherwise known as strategic interactions, as well as clustering risks given the movies limited geographic origins. These potential violations of the large sample assumption of independence must be recognized as potential limitations in our model's validity.

3.1.2 Structural limitations of Model

The relationship between variables explaining human behavior is complex and interrelated. The overall outcome is to generate a predictor of human behavior, a predictor of what movies will generate the most views on the new streaming platform. The models include the variables of the film's `Gross`, `IMDB_Rating`, and `Total_Actor_Score.` Many other factors may contribute to the number of views on a streaming platform. Some of these factors were omitted by choice, while others were not available at the analysis time.

3.1.3 Intentionally Omitted Variables

The movie title and the year the movie was released were both intentionally omitted from the models. This data had been collected and was available at the time of analysis but was omitted as it was determined it had little impact on the viewability of the movie on the new streaming platform.

3.1.4 Further Examination

Two factors that could impact Meta scores were runtime and genre of movie. Runtime ranged from 45 minutes to 321 minutes with an expected value of 123 minutes. The amount of time a movie takes to run may influence its attractiveness to viewers. Further examination of this variable is suggested.

The genre of the movie may also influence its attractiveness to viewers. Popular genres of the times could guide more viewers to stream a movie. Many movies are classified into multiple genres, such as "Action, Adventure" or "Drama, Action, Biography, Crime". The double or triple classification can make selecting a genre that will draw viewers to the streaming platform challenging and unpredictable. While further examination of the relationship between genre and the number of viewers streaming the movie is suggested, the current models omitted genre to maximize the models' reliable predictive power.

3.1.5 Unintentionally Omitted Variables

Factors not available at the time of analysis include, but are not limited to:

- Total Movie Budgets

- Number and Types of Awards and Honors the Movie received
- Number and Locations of Countries where the movie was released
- Genre Fan Base Size
- Pre-existing Fan Base Size (from books or other media types)
- Number of languages the movie was translated into
- Adjusted Gross (movies over five years old)
- The number of times the film has been released

Omitted variables, both intentionally and unintentionally omitted variables, can potentially impact and influence the bias of the models. These factors may change the model's overall fit to reality and sway the model's predictive power. Three unintentionally omitted factors that were not available at the time of analysis that could impact the fit of the models are total movie budget, number and types of awards and honors the movie received, and number and locations of countries where the film was released.

Total Movie Budget A movie may have a large budget for many reasons, high profile actors or directors, large casts, intricate sets and costumes, complicated stunts, large advertising campaigns, or complications with production. Some of these factors may positively influence the outcome of viewership, but others may negatively affect the streaming of a movie. Therefore, the influence of budget on the current coefficients is unknown and needs further exploration. The 1000 movies in the IMDB dataset would need to be coordinated with other datasets to fill the budget amounts for every 1000 movies on the list. Compiling the data may take multiple datasets since the range of film spans numerous decades, different movie studios, and a wide range of genres.

Number and Types of Awards and Honors the Movie Received From Oscars to BAFTA, a movie can gain many awards and honors. Awards and honors may increase the public's desire to watch the film. Awards can bring visibility to a movie that did not have a large advertising budget, showcase a new rising actor or celebrate the life of a well-known one, and get a movie to a country or audience that has never been exposed to the film. Due to all of these interactions, the number and types of awards may positively affect the streaming viewership. This positive relationship may alter the coefficients and result in a better fit model to reality. The number and type of awards may have a positive bias and pull the model away from zero.

Number and Locations of Countries where the Movie was Released Movies are released all over the world to a variety of audiences. Often, movies are even dubbed into a wide range of languages to reach a greater audience better. Films with the budget and the audience base to be released in more countries are more likely to have a larger audience. Therefore, the effect of the number of countries the movie is released in, and viewership of a movie on the new streaming platform is predicted as a positive relationship. Further exploration should be completed to fully realize the impact of the number of countries the film is released in and the rating. If the relationship is positive, the bias will draw the model coefficients away from the zero of reality. Multiple data sources will need to be mined to generate information about each of the 1000 movies in the IMDB dataset.

4 Conclusion

4.1 Overall Conclusion

This exploration aimed to investigate the predictive factors of a movie's streaming success before it was released to the public in the hope of biding for films for the new streaming platform. The relationship between Metascore, the calculated combined reviews scores published by at least four different critics, and the total movie gross.

Metascore is not impacted by the total movie gross, and therefore gross is not a predictive measure of how well the movie will be rated. Other evaluated factors affecting Metascore were IMDb rating and Total Actor Score, derived from the ranking of actors and number of appearances in the top 1000 movies on IMDb. IMDb rating is a post-hoc measure of a film and would not help to determine which movies to bid for but

may be used to bolster predictive models. Total Actor Score could be used as a predictive measure of a movie's success since the listing of actors precedes the film's release to the public.

Observationally, Metascore and IMDb ratings did not match. Therefore, the experts' opinions did not correspond to the general public's ranking of movies on IMDb. The objective is to acquire films that the public will want to watch and incentivize them to join the new platform.

4.2 Further Study

As the dataset from IMDb was explored, other factors of a predictive nature revealed themselves, such as genre, total budget, total runtime, number of countries the film was released in, awards earned by the film, and the pre-existing fan base from other media. While some of these factors were available in the dataset, some were not and would need to be added in the future. Further study is recommended for the genre, pre-existing fan base films, and total budget of these factors. These factors' effect on the public's movie rating could help determine which films to bid on for the new streaming platform.