Linear regression analysis of IMDb Movie Ratings as an interpretive model for audience score

Liam Isaacs

Abstract: Movie ratings on popular sites such as IMDb, RottenTomatoes, and Metacritic are often used by movie-goers to decide whether or not to see a movie. We presume that a 1-10 rating is not evenly distributed over all movies — in other words, that these movie-goer reviews tend to prefer certain types of movies. For any person trusting reviews on the aggregate, it can be useful to understand what underlying trends occur in the data as to better inform their decision-making process. We scrape and cleaned 2316 records for IMDb movies premiered during the 2016-2020 period and choose to focus only on studio and genre. We then use k- (5-) fold cross-validated lasso multiple linear regression results to feature engineer an interpretive model of audience scores, and conclude that of genre and studio, the two greatest factors influencing an IMDb review are Animation and well-known (>=7 movies/year) studio distributor. We create a web application that movie-goers can use to gauge a movie with respect to its interpreted score from this model (the "Liam-o-meter") based off these biases to answer the question: how well does this movie perform in comparison to if I were to only account for these biasing factors influencing its score?

Index-terms: *lasso regression, linear regression, k-fold cross validation, one-hot-encoding, web scraping, web application.*

Introduction:

Movie-goers, referred to here as "reviewers", use online websites to give an index of what movies to watch. Reviews tend to be scaled 1-10 or 1-100 and tend to be accepted as a trusted source of opinion on a given film. What is conveyed to users indirectly, as with any online voting system, is that ratings are independent of each other, since we are not inscribed to a given reviewer's taste in film. Of course, this seems to be the advantage of public opinion, that it reflects what "the people" are thinking; however, things can get more complicated than that. It could be true this system prevails for "snap-judgements" made about movies, that reflect only peoples' initial reactions, or that in general these reviews suffer from being bias by mainstream Western culture's beliefs.

For our experimental evaluation, genre and distributor stood out as potential ways in which we make snap-judgements; for instance, "I don't like horror" or "I like Disney movies". Our hypothesis is that these sorts of baseline assumptions will prevail in a review system designed on initial reactions to movies.

H1: the genre of a movie will influence its rating;
H2: the distributor of a movie will influence its rating.

We scraped data from three websites, IMDb, RottenTomatoes, and Metacritic, and used p-tests to select the website with most normality, IMDb. We then used linear

regression on a volume of 2601 movie ratings from IMDb released in the 2016-2020 period to test our hypothesis. We make the distinction of this model being *interpretive*, that is, used to quantify the relationship between genre/distributor (our "features") with IMDb score (our "target"). To counter-weight overfitting of our model, we use lasso regularization (L^1-norm penalty) to penalize the least squares cost function. We arrive at a correlation coefficient highest for Animation and Disney.

Results:

I. Data collection

Data was scraped from BoxOfficeMojo to obtain a comprehensive list of 3107 movies from 2016-2020 (seemingly all movies released in this period). Movies were then searched on IMDb, RottenTomatoes, and Metacritic using the Scrapy web crawling framework (1). Any fields of data for 'rating' or 'genre' were auto-populated as 'N/A' as to avoid under-sampling under-reported movies. All crawlers resulted in upwards of 90% of movies being returned as data without 504 errors (interpreted as the movie not existing on IMDb, RottenTomatoes, or Metacritic), with 2777 for IMDB, 2658 for RottenTomatoes, and 1971 for Metacritic. Any ratings that could not be found were auto-populated to the mean of the data.

II. Data selection

A primary assumption we use is that ratings should be normally distributed about a mean of 7, as we assume movie ratings will center around a 7 is average, as opposed to a 5 is average.

Using this assumption, we can plot an overlaid histogram of IMDb, RottenTomatoes, and Metacritic to evaluate the validity of the reviews (Figure 1).

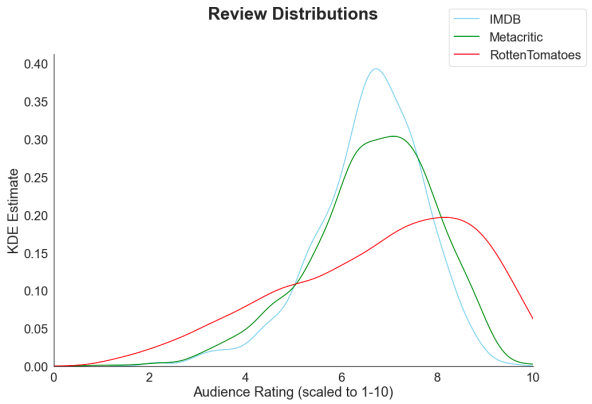

Figure 1: a plot of audience score (scaled 1-10) against the Kernel Density Estimate (KDE) of the probability distribution for any given score.

A KDE plot of RottenTomatoes audience ratings shows clear inflation and is skewed left. This is not unsurprising, as RottenTomatoes reviews are just a fraction of a binomial distribution, e.g. "90% positive", which will, in a culture where 7 is average and positive/negative is seen as <5 or >5, reviews are bound to skew left, favoring more positive reviews and not actually representing a movie rating.

We chose IMDb ratings as there was a greater volume of data (n=2777).

III. Data cleaning

'N/A' values were replaced with the average for rating; any null values for "genre" were dropped. We were left with 2316 rows.

IV. Linear Regression

We used one-hot-encoding to encode genre as a binary variable. The data was split into train (80%) and test (20%). A baseline linear

regression on the training data gave an R^2 value of 0.15. Distributor was split into the top 23 studios, with the latter 248 being replaced with "Other", as <7 movies/yr. were released. This gave an R^2 value of 0.17. Test data was linearly regressed to give an 0.16, indicating that the model was overfit. Lasso regularization was performed to minimize coefficients. We use k-fold cross-validation to arrive at a regularization strength of α=0.01. Post-lasso regression gave a R^2 value of 0.14 and mean absolute error (MAE) of 0.80 across both our training and testing data.

From the coefficients of our linear model, there are 13 influential features, with only six above a coefficient of 0.1:

Feature	Coefficient
Indican Pictures	-0.46
Drama	0.32
Animation	0.31
Music	0.21
War	0.14
Other distributor	-0.13

Figure 2: A table displaying the top 50% of factors between genre and distributor that influence IMDb ratings.

Discussion:

A central way of seeing if our results are valid is a residual analysis. Central assumptions of linear regression are that our error terms are normally distributed in a zero-population mean, that error terms have constant variance as we move through our distribution space, and that errors are uncorrelated. A residual plot of our data (Figure 3) shows that errors are not normally distributed.

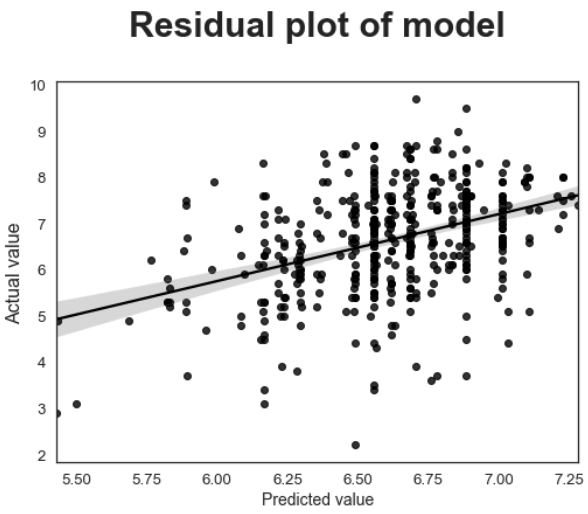

Figure 3: a residual plot of the linear model.

This suggests that our data may be not accurately predicting some values, in addition to only predicting values with a 5.5-7.25 range.

Web application

The goal of this study is to provide movie-goers a reference score to use in conjunction with IMDb ratings in order to make more informed decisions on which movies to watch. As ratings are partially dictated by underlying trends in mainstream culture, "outstanding" movies may not just be those with high-ratings, but those with *surprisingly* high ratings, that defy expectation; in other words, the outliers.

We can think to visualize this in a way that makes sense to users: non-statistical and highly visual. To do this, we elected to use Amazon Lightsail to host a FlaskApp webpage, and referred to the predictions of our linear model as the "Liam-o-meter", which can be viewed ***here***.