Project 3: COVID-19 and Movie Sentiment

Vivian Liu and Kate Marsh

Professor Lee
Applied Data Mining
19 April 2021

Abstract:

The initial motivation for this project was to compare the IMDb and NYT data to see differences in the two sources of movie reviews. After looking at the differences, we decided to narrow in on before and after the pandemic hit to find how sentiments have changed after the large cultural and societal changes that have come with the pandemic, especially with the closing of movie theaters and the increased shift towards streaming services distributing films (for example with Mulan's exclusive release to Disney+ in March 2020). The data-mining aspect of this project is that we combined the NYT and IMDb data, and created lasso models to find the largest coefficients compared to both rating systems, and found that there was a shift towards horror movies and away from thrillers after the strike of the pandemic in both the IMDb and NYT reviews. In addition, while IMDb rated animation and documentaries highly both before and after the pandemic, the New York Times had no such trends. Something we might have done differently is incorporate box office performance into the model to see if certain genres were getting more money before and after the pandemic. The link to our code is here: https://github.com/katelmarsh/3106finalproject/

Introduction:

Due to the highly disruptive nature of the COVID-19 pandemic, many die-hard moviegoers have been relegated to laptop screenings of new films. Since last March, popular movie theater chain AMC alone experienced $4.6 billion in loss[1], as other smaller chains suffered total bankruptcies. However, even as movie theaters continue to struggle, movie production companies have actually miraculously emerged relatively unscathed. By taking advantage of the booming streaming industry, companies like Warner Brothers have been operating, successfully, on a charge-per-stream or streaming service business model that conveniently bypasses the need for a theater.[2]

While nothing can really replace the excitement and ambience of a real movie theater, we wanted to explore how this new normal might influence how viewers react to the media that companies produce. After all, the intimacy of a one-laptop viewing may contribute to the poignancy of a drama film, while the surround-sound system of a movie theater may deliver more punch to an action movie. To do this, we probed the effects of customer sentiment on films released before and during the pandemic. We drew data from two different sources: the New York Times (NYT) movie reviews and IMDb crowd-sourced ratings. Ultimately, we wanted to see if the pandemic caused a change in movie tastes, which would be reflected in how critics review new movies, or how IMDb users rate films. Anyone who loves films can use this as a moment of self-reflection—have the events of the pandemic caused us to gravitate towards different films? Since we spend so much time with our computers, has our increased consumption of all online media caused us to become more or less critical of our films? Through mining, wrangling, and analysis, we were able to determine what genres of movies audiences

[1] https://variety.com/2021/film/news/amc-theatres-4-6-billion-loss-covid-1234927642/
[2] https://www.npr.org/2020/12/21/948697829/what-will-the-film-industry-look-like-post-covid-19-pandemic

loved most before and after the pandemic hit, and glean some insight into public sentiment through the lens of film consumption. We attempt to answer these questions through analysis of our data. The value of this project lies in the self-reflection that someone might have on their own movie consumption throughout the course of the pandemic in conjunction with public sentiment.

Data Importing, Cleaning and Processing:

We drew our data from two different sources: New York Times Movie Reviews[3], and IMDb[4]. To grab the NYT reviews, we worked with the NYT Developer API. Due to restrictions on the number of requests per minute, we staggered our calls with 10 seconds between each and setting a restriction in the loop so it would break when there were no more results. The API request asked for all movie reviews of movies released between March 15, 2019 and March 15, 2021. Our final pull from the NYT API resulted in a json file containing the title, whether it was a critic's pick, the MPAA rating, headline of the article, summary of the article, article publication date, movie opening data, a link to the article, and multimedia descriptions.

For the IMDb data, we downloaded a TSV file of basic information about the movies, including an identifier, Title, isAdult, release year, runtime, and genre. We also downloaded another TSV from IMDb that included the identifiers, average rating, and number of votes. We merged these datasets by the movie identifier to have a complete dataset, and subsetted it so that it only included movies from the same 2019-2021 time frame as the NYT reviews. To create the combined dataset with both IMDb average rating and NYT critic's pick, we merged the two

[3] https://developer.nytimes.com/docs/movie-reviews-api/1/overview
[4] https://www.imdb.com/interfaces/

based on film title. Luckily, these datasets were relatively clean and we were able to make this merge without any issues with matching the titles.

Initial exploration showed that the average IMDb movie rating was a 6.38, with a range of 2.7 to 9.6. Since IMDb ratings of 10 are almost impossible to achieve due to the large amount of votes and variation in opinion, this range appeared reasonable[5]. We also found that around 20% of movies reviewed by the New York Times were "critics picks." Comparing the top 21% of IMDb ratings and the NYT critic's picks, we found that there was a 35% overlap between these top rated movies. Given that the reviewers on these websites are looking for different things that make a top movie (IMDb is crowd-based and leans male[6]), the difference in the top-rated movies was not surprising. Additionally, the opening dates of movies seemed to correspond to movie releasing trends, with notably less movies released in the summer of 2020, compared to the summer of 2019, which is backed up by the long list of movies postponed after the closure of most movie theaters around the country[7].

---

[5] For reference, the 2020 Best Picture Oscar Winner "Parasite" has an 8.6 rating on IMDb, and the "Top 250" List, the highest rating is "The Shawshank Redemption" with a 9.3 (https://www.imdb.com/search/title/?groups=top_250&sort=user_rating).

[6] https://www.wired.co.uk/article/which-film-ranking-site-should-i-trust-rotten-tomatoes-imdb-metacritic

[7] https://www.vulture.com/2021/04/here-are-all-the-movies-and-tv-shows-affected-by-coronavirus.html

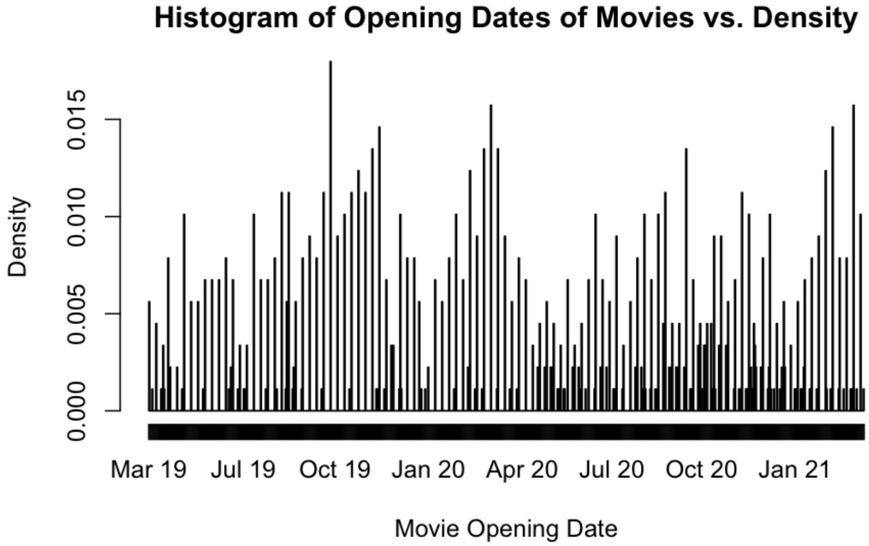

Figure A. Histogram of movie opening dates from March 15, 2019 to March 15, 2021 compared to the percentage of movies released that day.

In order to look into user sentiment based on movie genre, we introduced columns that binarized based on movie genre (e.g. Comedy, Drama, Thriller, Horror, etc.). Since most movies spanned multiple genres, some of the movies had "1" values in multiple columns. For example, the 2020 movie "Mulan" falls into the three categories Action, Adventure, and Drama so it has a 1 in those columns, but a 0 in all the other genre columns.

Our final dataframe contains 1114 unique movies, consisting of overlaps between IMDb and NYT reviewed movies, pulled between the dates of March 15, 2019 to March 15, 2021. Now that we have our dataframe with the binarized genres, we are ready to analyze our data.

Algorithm

To test which genres were most highly correlated with the NYT and IMDb reviews before and after COVID, we regressed with either Critic's Pick or IMDb rating as the y-variable, and the different genres as the x-variables. To start, we split the dataset into two parts, pre- and post-COVID-19 pandemic in the United States, which was determined by the engineered feature "COVID." This determines whether the opening date of each movie in the dataset was before or after the beginning of the COVID-19 pandemic, giving movies released after the onset of the pandemic (March 15, 2020) started a score of 1 and movies released before the pandemic a score of 0.

For each specific group, we tested four different types of regression models: ordinary least-squares, stepwise, lasso, and ridge. After splitting the data into 10 folds, we performed 10 iterations of all four models and found the average of the MSE in order to determine the best model to run. Because Critic's Pick is a vector of binary variables, we used logistic regression in the metric tests and in the final models for these, while we used a linear model in the metric tests and in the final models for the IMDb rating since it is on a continuous scale of 1-10. Figure B below shows an example of analysis of the best model to use based on MSE for IMDb post-COVID.

	ols <dbl>	stepwise <dbl>	lasso <dbl>	ridge <dbl>	min <chr>
1	0.5424485	0.5675589	0.5299252	0.5418050	lasso
2	0.5808432	0.5367537	0.5523774	0.5787731	stepwise
3	0.6151336	0.6301242	0.5904785	0.6015896	lasso
4	0.7206515	0.7212437	0.7283408	0.7192630	ridge
5	0.7223145	0.6549913	0.7364463	0.7309174	stepwise
6	0.8428200	0.8635036	0.8423268	0.8421087	ridge
7	0.7588944	0.7212265	0.7381115	0.7605804	stepwise
8	0.4877455	0.4787682	0.4806156	0.4905490	stepwise
9	0.8113157	0.8270342	0.8177163	0.7951596	ridge
10	0.4142017	0.4897392	0.4645063	0.4300154	ols

Figure B. Summary of MSE for Different Models, IMDb Rating Post-COVID

For Critic's Pick, since we used a logistic regression, we decided to use a log-loss metric to determine the best model, and for IMDb reviews, we used MSE since we used a linear model. For the models testing the Critic's Pick against pre- and post-covid, lasso had the best performance each time. To keep everything uniform, we decided to model with lasso every time.

From the lasso models, we ran each model twice — once normalizing the features and once without normalization — and then found the coefficients that overlapped in both the normalized and non-normalized version of each model. For each model, the overlap was 4 genres.

For the models testing the IMDb rating, our analysis was split evenly between stepwise and lasso with the lowest MSE between the ten-fold test/train sets that we ran for both the pre and post-covid datasets. Ultimately, we decided to use the lasso model to stay consistent with the critic's pick analysis. Again, we split based on pre- and post- COVID. We ran linear models using the rating (1-10) as the y-variable, regressed against all of the binarized genres in order to see which genres contributed the most to higher IMDb ratings.

In summary, we regressed on four groups: pre-COVID critic's pick, post-COVID critic's pick, pre-COVID IMDb ratings, and post-COVID IMDb ratings.

Results of Data Mining

The results of the four lasso models are as follows. Best-rated in this paragraph refers to the genres with the highest coefficients in the lasso model. The best rated genres pre-covid for the New York Times were Western, Comedy, Thriller, and Music. The best rated genres post-covid for the New York Times were Crime, Drama, Horror, and Action. The best rated genres for IMDb pre-covid were Animation, Documentary, Thriller, and War. The best rated genres for IMDb post-covid were Animation, Documentary, Horror, and Western. These results are summarized in figure B below. The coefficients from these lasso models are shown in the appendix.

Top Results for Lasso Models of Reviews for IMDB and NYT Against Before/After COVID-19 Pandemic

imdb_pre_covid	nyt_pre_covid	imdb_post_covid	nyt_post_covid
animation	western	animation	crime
documentary	comedy	documentary	drama
thriller	thriller	horror	horror
war	music	western	action

Figure C. Table of the Largest Non-Zero Normalized Coefficients for Each Model. Cells in light pink are ones that are not repeated, while cells in hot pink are repeated in another column.

Analyzing these results, we can find that certain genres stayed popular before and after the COVID-19 pandemic began. On the IMDb website, both the documentary and animation genres stayed popular before and after COVID-19. For the New York Times, no genres were the highest correlation coefficients before and after the COVID-19 pandemic. Interestingly, there is some overlap between the genres rated highly before the pandemic hit for both sources. Before the start of the pandemic, thrillers were rated very highly by both sources. After the start of the pandemic, horror movies were rated very highly by both sources. Since these are very different sources of movie reviews and previously we found that these groups generally do not rate movies in a similar way (the overlap between titles in the top ~20% of IMDb and Critics Picks is only 35%), the fact that both these groups highly rated thrillers before the pandemic and horror movies after the pandemic is notable. One could argue that professional movie critics from the NYT preferred darker movie material (e.g comedy to drama, music to crime) after the onset of the pandemic, perhaps pointing to a reflection of general public sentiment.

This finding was unexpected; however, these results are actually corroborated in a 2020 University of Chicago study that found that adults that like horror films “exhibit less psychological distress” during the COVID-19 pandemic[8]. Additionally, pandemic-themed movies, which typically fall into the horror category, are more relevant to the general public after the pandemic started and thus might be rated higher by critics and IMDb users alike. On the contrary, another reason we have hypothesized is that fans of thriller movies might be already predisposed towards movies and, therefore, it is more dependent on the movies that have come out that year, rather than an actual shift in preferences.

We have a few next steps in mind for this project. The first is performing cross-genre analysis using Principal Component Analysis to see if different genres are reviewed similarly. A possible hypothesis is that drama films reviews (typically reviewed higher by movie critics) would be greatly different from action films (typically reviewed higher by the general audience). In addition, we wanted to take a closer look at the movie critics in the NYT reviews data—no data on the demographics of NYT movie reviewers exists online, so could use a Python NLP package to determine the gender of the reviewers and extract some insights into the critics who influence movie viewership. With this gender data, we could look into what genre of film female reviewers are typically assigned to, and whether or not they tend to leave higher or lower proportions of critic’s picks. Finally, we would like to use k-means clustering or tree structures to group similarly rated movies, and see if genre is a significant contributing factor to similar reviews. Finally, it would have been interesting to incorporate “box office” performances (including traditional theater revenue and streaming service revenue) into our analysis, and explore the relationship between commercial performance and perceived artistic value.

[8] https://doi.org/10.1016/j.paid.2020.110397

For our algorithm, we thought there was not much opportunity for data snooping. Since the linear models were pretty straightforward, we did not have the chance to tune many parameters. IHowever, if we were to implement the k-means clustering algorithm for grouping movies, a potential source of data dredging could come from cherry-picking an optimal number of clusters, and only reporting those results.

Critique of Bennett and Carlos' Project

We critiqued Bennett and Carlos' project on NBA G League Players. It was interesting to see how much insight could be gained from only 115 players worth of data—although the amount of data on up-and-coming players is not very extensive. We thought the generated features were very interesting, though it would have been nice to see (perhaps in the appendix) how the equations for each position were created. The scree plot was convincing in justifying the number of clusters used in k-means. Since cosine similarity is typically used in NLP contexts (as with our Twitter and Indeed.com cases), it was surprising to see in the context of basketball player classification. It would have also been nice to see how hierarchical clustering or using trees could apply to this data.

Overall, we were quite convinced of their findings. Not only were the visualizations easy to understand and the algorithm clear and understandable, the verification of the robustness of their model helped nail their point. On a final note, it was very interesting how Bennett and Carlos took the methodology from Project 2, and applied it to a completely new context—having done Project 2, we were able to follow quite well the justifications behind the algorithm.

Github:
https://github.com/katelmarsh/3106finalproject

Appendix

Figures C-F. Histograms of the Lasso Non-Normalized Lasso Coefficients.