Final Capstone Project

Proposal

**Box Office Movie Prediction – Final Capstone**

**Problem:**

1. How much money will a movie make opening weekend?
2. How much money will a movie make in total after its opening weekend?

It is incredibly difficult to forecast the performance of a theatrical movie release. There are numerous examples of movies underperforming studio expectations [[1]](#footnote-1), and in some rare instances like Black Panther the movie far outperforms anyone’s expectations [[2]](#footnote-2). Traditional measurement methods conducted via movie exit surveys fail to correlate those metrics to box office performance [citation]. Additionally, on average, 15-20% [[3]](#footnote-3) of a movie’s total marketing budget goes toward digital marketing. So, assessing the reach and engagement of digital metrics should be an indicative measure of a movie’s success.

**Solution:**

I would like to deliver a pipeline that from the Wikipedia url outputs a box office opening weekend prediction and movie multiple (ratio of total box office to opening weekend box office).

The goal is to feed the model a movie (based on its Wikipedia link) and then collect relevant metrics such as movie info, pageview metrics, critic scores, audience feedback, trailer metrics, and social media reach and engagement to formulate a “digital footprint” of the movie. This digital footprint then would be fed into the linear regression model or neural network to predict opening weekend box office as well as the movie’s expected multiple to determine the final result. The theory is that critic scores and audience sentiment should be a good indicator for the movie’s staying power.

The ideal solution would be to provide an opening weekend forecast 3-4 weeks prior to release. The movie’s multiple is less critical this early. However, it might be interesting to see how accurate the prediction is that early, in which case you could have an idea of the film’s total performance as well as opening weekend.

Every feature variable of the model represents a potential business decision. The number of theaters the movie is released in is incredibly important. The linear regression model would provide a coefficient for just how important each incremental theater is, which is a negotiated item for a film release. There is a massive industry built around targeting and retargeting to increase trailer views and social media metrics. Again, understanding the coefficients for those features could impact whether or not to spend more marketing dollars. Ultimately, each feature represents a decision, and I would like the model to quantify how important each decision is to the movie’s opening weekend performance.

**Data sources / access:**

1. Wikipedia
2. Wikipedia pageview API
3. RottenTomatoes
4. Metacritic
5. Box Office Mojo
6. Youtube trailer views
7. TheMovieDB
8. Facebook / Instagram

**Techniques:**

1. Scraping
2. Clustering
3. PCA
4. Linear regression
5. Neural Network (NN)

**Challenges:**

One of the challenges I may face with my NN approach is to understand what variables impact box office performance. If the neural network outperforms the linear regression, that will be great from an output perspective, but it will be very difficult to determine what input variables are important to the NN model. It might be that the NN is super accurate, but since I won’t have coefficients like in the linear regression, it only gives me insight into the movie’s future performance a week or two before the film releases (or whenever I have all the data I am planning to use in the model). The linear regression model would be much more useful from that perspective, as I’d be able to know how important each feature is far in advance. This information could drive my marketing strategy for the film and lead to greater results.

Ultimately, the NN is an upper bound for the accuracy of the linear regression model, and should provide me with a roadmap to how good my data is at predicting box office. For this capstone, I will run the linear regression model many ways to try and get as close as possible to matching the NN accuracy.

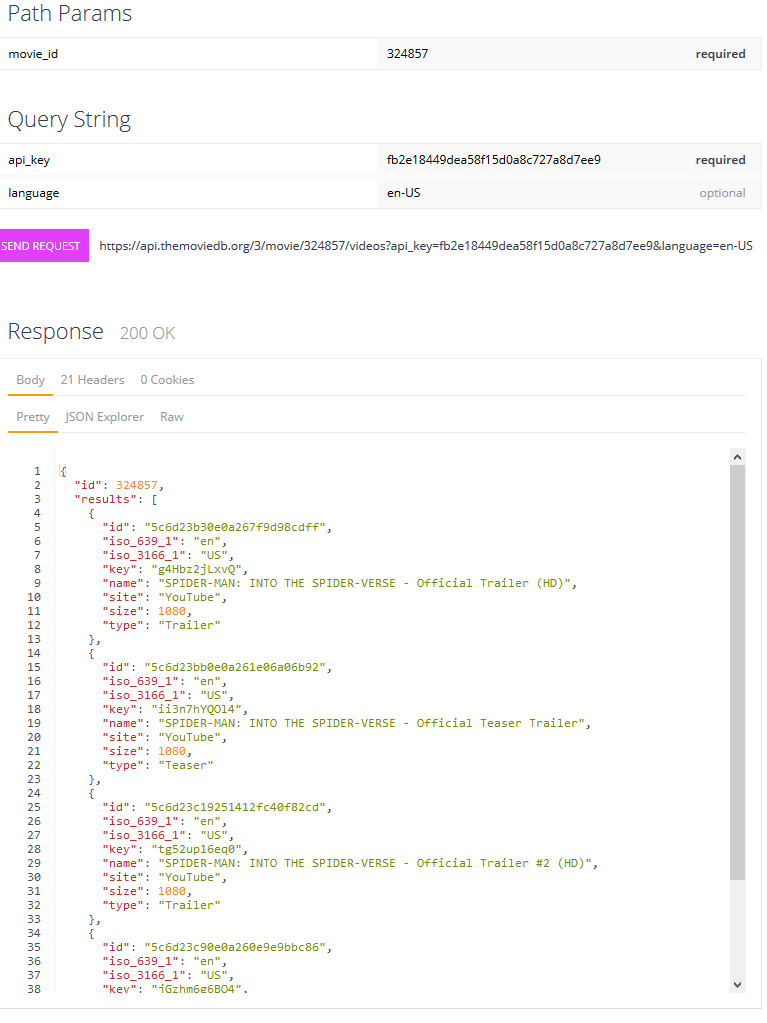
Most of the sources will be fairly easy to get the data. I’ve done the work on scraping and cleaning most of the sources from prior projects and the new sources will provide data neatly in the form of APIs. TheMovieDB API is a lifesaver to getting movie info easier as well as making it possible to get all the trailer views for a movie. However, I need the IMDB link for theMovieDB API, so I need to make sure I’m grabbing enough of those from Wikipedia to have a big enough sample. Facebook scraping is going to be a bit of an issue, and the API is difficult to get access to. It may be impossible for me to get this data.

The Wikipedia pageview data will be a challenge. In a prior project I didn’t find the information to be highly correlated to box office success. Plus it limits my dataset to only movies after 2015 when the API was made available. A few things I’m going to try differently; 1) is PCA on the search metrics, and 2) try some clustering techniques to identify similarities in pageviews that might help box office correlation. Also, I had tried PCA in the unit 3 capstone, but I only pulled out the most highly correlated variable, which I’ve since learned doesn’t make as much sense as keeping all of the output PCA features.

**Additional detail**:

* RT reviews – scores calculated from top critics
* RT audience reviews – text from first 1000 audience reviews
  + Clustering technique applied to reviews to group similar films
* Metacritic scores – scores calculated
* Director, actor, distributor clustering from BOMojo
* Number of theaters information from BOMojo
* Output variables – total box office and opening weekend box office from BOMojo
* Film trailer views from YouTube
  + Get as many as possible: calculate min, max, average
* Wikipedia pageview data – use metrics from unit 3 capstone to identify time before release
  + Clustering? PCA? Try both techniques to come up with higher correlated variable to box office
* Crawl Facebook for posts and followers
  + Film-specific hashtag from theMovieDB API
  + How to get historical or time-specific mentions from hashtag through API

**Example of TheMovieDB API /movies request:**



1. [LEGO Movie 2](https://variety.com/2019/film/box-office/lego-movie-2-box-office-fell-short-1203134695/), [Solo](https://www.cinemablend.com/news/2430760/why-solo-underperformed-at-the-box-office), [Pirates and 2017 in general](https://www.independent.co.uk/arts-entertainment/films/features/franchise-fatigue-blockbuster-sequels-hollywood-obsession-marvel-universal-warner-bros-despicable-me-a7829546.html), [not a new phenomenon either (2010)](https://www.denofgeek.com/movies/15904/why-are-so-many-films-underperforming-this-summer) [↑](#footnote-ref-1)
2. https://comicbook.com/marvel/2018/02/20/disney-ceo-black-panther-obliterated-expectations/ [↑](#footnote-ref-2)
3. https://variety.com/2019/digital/news/10-things-we-learned-at-varietys-silicon-valleywood-summit-1203196318/ [↑](#footnote-ref-3)