Final Capstone Project

Proposal

**Box Office Movie Prediction – Final Capstone**

* Proposal clearly outlines the problem it’s attempting to solve (0/1 point)
* Student discusses the biggest challenges they anticipate facing in the project (1/1 point)
* Student shows how the project solution is valuable (0/1 point)
* Project goals are clearly articulated and achievable (0/1 point)
* The proposal talks about the project in terms of a product that end-users can use. Even if the student doesn’t have a user facing product at the end, they should discuss how a user could use their “product”. (0/1 point) *Clarify what your goals are. State the problem, what you plan to deliver, and how it solves the problem and how it will be used.*

Specifically, what are your deliverables and how will they be used? Think about your audience and describe how you will solve the problem. For example, how will you use your learnings to make a business decision? What are you trying to predict? What are some candidate input variables? For example, what are the "search metrics" you plan to use PCA on? Are theses engineered features? Are these available from your sources? And how will you evaluate your models?

**Problem:**

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

It is incredibly difficult to forecast the performance of a theatrical movie release. There are numerous examples of movies underperforming studio expectations, and in some rare instances like Black Panther the movie far outperforms anyone’s expectations. Traditional measurement methods conducted via movie exit surveys fail to correlate those metrics to box office performance. Additionally, on average, 15-20% 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, potentially even an API that takes the Wikipedia url and then outputs a box office opening weekend prediction and movie multiple.

The goal is to feed a 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 (opening weekend box office / total box office) to determine the final result. The theory is that critic scores and audience sentiment should be a good indicator to 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.

It might be difficult to see what business decision could be made from this information. However, every feature variable of the model represents a potential business decision. As I’ve identified in prior iterations of this model (I have run linear regressions on some, but not all, of the data sources listed below in prior capstones), 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 budget pushing the movie. Ultimately, each feature represents a decision, and the model will quantify how important each decision is to the movie’s opening weekend or total 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. CNN

**Challenges:**

Most of the sources will be fairly easy, mainly because I’ve done the work on scraping and cleaning those sources in prior projects. Also, finding TheMovieDB API is a lifesaver to make getting the trailer views much more manageable. However, getting the right movies from TheMovieDB to match will require me to have the IMDB link, so I need to make sure I’m grabbing enough of those from Wikipedia articles. Facebook scraping might be a bit of an issue, just because I’m unfamiliar with it.

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 all of the search metrics I come up with, 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 to keep all of the output PCA features.

Finally, related to techniques, I’m going to compare my linear regression model to a neural network model for box office prediction. 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 CNN model. I will probably try to run the linear regression a few different ways, leaving out different input variables to try and back into feature importance, but that will be a challenge in the model evaluation phase.

My theory is that the additional data I am bringing into this challenge (trailer views, Instagram followers and content metrics, using both RT audience reviews and Wikipedia pageviews together), will be more accurate than prior iterations. I think both the CNN and linear regression with clustering techniques will be able to account for the different release patterns (greater than or less than 1k theater releases). Unfortunately, CNN is a little bit of a black box, and won’t provide me with coefficients to just how important each feature is to the movie’s performance. Hence why I want to compare both models. If I am able to be super accurate with the CNN, it might justify using that model even though it will be much more difficult to explain why the movie’s performance is a certain number.

**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:**

