Draft Capstone Project Proposal - Predicting RottenTomato.com Movie Rating for Fresh or Rotten

Hypothesis and Background.

Can the Tomato Meter rating of a theatrically released feature length film on Rotten Tomatoes be predicted? Is it possible to use the rating history of Movie Critics reviews that are published in Rotten Tomatoes to determine the Fresh or Rotten status of a new release? Or predicting chances of being nominated for Oscars? Can we formulate a model of what constitutes a critically acclaimed movie?

Method & Data Sources: EDA and graphing and modeling

RottenTomatoes.com has a section specifically of Movie Critics reviews that I was able to web scrape to collect each critic's movie review scores. I chose to limit the analysis to certified critics section since they are more likely to see a broad range of films, as opposed to audience members who self-select the movies they see and possibly only rate ones they love or hate. I also downloaded the IMDB 5000 movie title database that contains information about the actors, director and other elements of each title. From the official Oscars website I downloaded the data of the Oscar nominations 1996-2016.

Rottentomatoes.com Critic's pages
~1200 movie critics with 25 or more reviews on Rotten Tomatoes
~38,000 movie titles where reviewed – dating from 1898 thru Feb 2017
~611,000 reviews were collected

Oscars:
21 years of academy award nominations (1996-2016)
420 nominations for 253 actors
105 nominations for 74 directors
1187 total nominations when counting additional categories (writing, producing, editing, art, design, and cinematography)

IMDB:
The IMDB dataset of 5000 movies lists genres, directors and top three actors, budget and film length.

Collating and cleaning these 3 sources produced a dataset of:
3,320 titles (1,439 are rated "Fresh")
202,011 movie critic reviews

The important variables I found here are:
Movie title, Movie year, Actors, Director, Budget, Genres (21 types), Oscar nominations for a title and for each actor and director, the tMeter score from Rotten Tomatoes, the critic's reviews for each movie.

Data Limitations:
This data does not contain information about the remainder of the crew that received nominations nor does it identify the specific main genre for a film. The data does not account for accolades a movie, cast member or crew received from other sources such as BAFTA.

Data Cleaning and Wrangling:
From the Rotten Tomatoes database I had to standardize to critic names to account for Spanish characters to enable better web scraping. For the reviewer's scores I had to standardize them where some critic's gave letter grades A+ to F- and others gave scores out of 4 or 5 or 10 stars. I chose to standardize the scores to 100 (since the tMeter score is in that format). For each letter grade I used the % scores of what other critics gave movies rated A+ and used the median. I had to remove data where it was not possible to determine when reviewer's score was missing or incomplete.

For the Oscars data I did a lot of the cleaning in Excel to associate each actor and director for each film regardless of whether an actor played a lead or supportive role. More work could be done for the writers and different departments involved that were nominated but the IMDB data does not contain that information and I would need to collect that from somewhere else.

Training and testing

My initial hypothesis was to determine if there was bias in Critics' reviews towards genres, actors, or a hidden bias towards less known element such as writer or producers.

Based on my analysis of the data available this evolved to predicting the Fresh/Rotten chance of a movie, using Linear Regression and Logistic regression models.

From the dataset I will be looking to see if a Movies score shows any correlation to the number of Oscars the cast and crew have had in their career(limited to the last 20 years) or to the number of movie critics reviewing the movie. Do movies with higher scores have a narrower band and hence a consensus on quality?

I also want to investigate the effect on Oscar nominations from past Oscar nominations, and see if the critic's dataset of movie review medians, IQRs or Standard deviation have any influence.

Are some critics better predictors than others? Are some critics a bellwether for Fresh or for Oscar nominations? If so what weight should be attributed to how far a critic differs from the median score of a movie?

Findings to Date

So far my linear models are producing very little correlation based on genres. There is some evidence that the more Oscars cast members have accumulated the higher chance the film with be rated “Fresh”.

Learned info, Next steps

Some next steps could be to see if any critics are “super critics” and are of themselves able to predict Fresh/Rotten. Do they have a genre speciality?

The dataset from IMDB lists all genres that the film falls into, but this is often seeming too broad. It may be better to reduce this to the 2-3 main genres, if possible.

The dataset I used for the cast and director did not provide additional cast and crew information for the other films in the time period of the dataset. There are a great number of good movies that do not make it to the finalist for Oscars. A next step would be to incorporate more screen credit nominations to include writers, producers, art directors and cinematography for the films in the time period, if possible.