MOVIE-PEDIA

Motivation

I started watching academy awards since 1998 when one of my favorite movies Titanic won many of them. I became a big Hollywood and academy awards fan since then. I used to love watching movie critics – their logical acumen and attention to detail used to enthrall me. At some point, I started taking their opinion with a grain of salt. There were bunch of movies that I liked that were miserably reviewed by critics – Transformers, Pirates of the Caribbean: At World’s End, Bruce Almighty, Meet the Fockers, The day After Tomorrow etc. These movies also ended up being huge commercial success. Not all of them that I liked won academy awards though. Some movies that I really liked like 127 hours, letters From Iowa Jima etc. had stellar critic reviews but they did really poor in box office.

I started thinking whether critics really matter for movies’ success. Whether I should really bother about their reviews before going to see a movie.

Then I also started bothering about what other factors that could affect the movie’s success? I decided to undertake a data science project to answer following questions:

* Whether critics’ review really matter for box office success?
* How important are other factors for a movie’s top box office success – time of showing, genre and audience’ review?
* What could be the general audience’s favorite genres of movie?
* What genre of movies people are likely to see at a specific time of the year?
* How are director’s and cast’s reputation going to affect a movie’s success?

Methodology

Web scrapped from Rottentomato, Twitter and TMDb. Cleaned and Wrangled all the data so that it is in a clean and ready to use form. Listed all of my independent variables using my gut feeling and common sense – Release Year, Audience ratings, Critic ratings, Director’s Past Movies’ Box Office Average, Genre, Actor’s Past Movies’ Box Office, Genre, Critic Ratings of the movie itself etc. Did the preliminary analysis to see the relationships between the factors of interest, especially checked if one independent variable is dependent on the other to prevent multicollinearity.

Exploratory analysis included box plotting of the various independent variables with time to check fluctuations. Various linear regression models were constructed to check independent variables’ interdependence. A final regression model was obtained that could predict a movie’s box office success.

Results

The final regression model was obtained.

Box office = 8.9203 + 0.0172\*Audience\_Rating + 1.8643e-08\*Director\_BoxOffice\_Avg + 0.0917\*Director\_Movies\_Count - 0.0204\*Director\_Critic\_Rating + 2.791e-08\*Genre\_BoxOffice + 2.212e-10\*Actor\_BoxOffice\_Sum + 0.0021\*Release\_year

Critics’ ratings for the movie and Box office are indeed not correlated! Critics’ review is inconsequential when it comes to movies’ box office success.

Future Work

Scrape more data from other sources and refine the model.

Test the model with recent data from 2017 and check the fidelity of the model.

In addition to linear regression, I intend to use Bayesian Inference to predict the general audience’s reviews and the box office of a movie. I can aggregate movie performance data in recent years to provide an informative prior. Then, I will use kNN and Random forest algorithm to classify critics, actors and movies. This result may be used to determine the likelihood function in the Bayesian inference model.