Predicting what features determine a movie’s success

Abstract –

For this project, I wanted to find out what features have impact making a movie successful by predicting total worldwide gross using linear regression. Some features I’ve added were – budget, opening, runtime, domestic opening, rating, total # of theaters, and genres. Data I used were from BOXOFFICEMOJO.com and scrapped all-time best movies. Finding out which features have the best impact in making most gross revenue. After refining a model, I built linear regression and compared different models with different features to show how they are related and can affect movie’s success.

Design –

The data was provided from BOXOFFICEMOJO.com and used classified each features based on movie’s success. There are many thing that can make movie’s successful, but I’ve cut down to only three features – budget, runtime, and season. I would gather data, make a model based on selected features, make predictions on those features, and make a conclusion how and which features have the most impact making a movie successful.

Data –

The initial dataset I started with was over 1,000, but I’ve cut down to about 600 after making cleaning and EDA. A few feature highlights to include was budget – more budget generally means better actors, directors, and stage setting, runtime – more runtime means better storyline with clear climax and resolutions, and season – releasing a movie based on different time of the year really have impact on bringing revenues. Blockbuster movies debuts generally summer and winter.

Algorithms-

Selected subsets of features against worldwide total gross to see the relationship. After making initial model, I tested on test data to predict how much revenue it can bring.

Tools –

Numpy and Pandas for data manipulations. Scikit-learn for modeling. Seaborn for plotting and PowerPoint to present to the class.