Regression Analysis

Features influencing high gross revenues in movie industry.

By: Archana Turumella

**ABSTRACT**

The objective of this analysis is to determine some of the features influencing in predicting a movie’s success in terms of high gross revenue. Movieland production company hired me to perform this analysis which will help them make informed decisions when investing in movies. I worked with data collected from [www.boxofficemojo.com](http://www.boxofficemojo.com) leveraging categorical feature engineering along with linear regression model to achieve an optimum fit to predict the gross values using training data. After fitting and analyzing several regression models on the data, the polynomial linear regression model is the best possible model that can be fit to the movies data since it has the best R squared value and lowest Mean absolute error value. Along with some EDA, I plotted graphs to visualize and communicate my results.

**DESIGN**

Data is scraped from the website [www.boxofficemojo.com](http://www.boxofficemojo.com) using Beautifulsoup library of Python. After performing initial cleaning and EDA, features are then transformed through feature engineering. Studying the correlation between features and the target variable can give us insights of all the features that make a significant impact in increasing the gross revenue. Using such features as the base to build a model that will best fits the given data is the primary objective of this analysis.

**DATA**

I collected top 200 movies from every year between 2015 and 2022 from boxofficemojo.com which resulted in 1400 datapoints and 11 features. After performing initial cleaning and EDA along with some feature engineering, the final dataset I worked, consisted of 1336 datapoints and 13 features. A few feature highlights include movie genre, MPAA rating (PG, R etc.), number of theatres a movie was released in and number of days a movie played in theatres.

**ALGORITHMS**

*Collecting Data*

* Using Beautifulsoup library in Python, I built a robust pipeline that scrapes data from the multiple pages in the website and stores in a pandas Dataframe.

*Feature Engineering*

* Converting categorical features in to binary dummy variable.
* Combining certain features under one single variable to reduce feature complexity.
* Movie released date feature is transformed to show if the movie was released on a weekend or not.

*Models*

* Using Beautifulsoup library in Python, I built a robust pipeline that scrapes data from multiple pages in the website and stores in a pandas Dataframe.
* After performing, Ordinary least squares from Statsmodel, Linear Regression and Lasso from Scikit, I settled with polynomial linear regression model from Scikit as this model showed the best R square value and lowest Mean Absolute Value.
* The entire dataset was split into 80/20 ratio where 20% was held out for testing. The training dataset was further divided as training and validation in the ratio 75/25 to achieve a thorough testing of the model.

**TOOLS**

* Beautifulsoup and requests for data scraping and parsing from website.
* Numpy and Pandas for data manipulation.
* Matplotlib and Seaborn for plotting.
* Statsmodel and Scikit learn for modeling.

**COMMUNICATION**

Slides and visuals for presentation. Work and code will be available in Github.