Liner Regression: Machine Learning Algorithm for Thriller Movies

Abstract

The business situation for this project is that there's a newly created production studio that plans to make movies in the thriller genre. They'd like to know which characteristics of thrillers are predictors of US Box Office Gross.

A linear regression model was created to answer the following questions:

- Does a set of features do a good job in predicting US Gross for Thriller Genre?
- Which features are significant predictors of US Gross for thrillers?

Design + Data + Tools

Design + Data	Tools
Webscraping •Scraped IMDB Thrillers for target and feature data •1100 thriller titles, 16 potential predictor variables	Request Module, BeautifulSoup Library
EDA + Regression Viability •Ensure data correct and appears as expected. •Data cleanup, address missing values, etc •Correlation matrix, reg plots, R^2 score •Feature engineering	Pandas, Seaborn, Statsmodels cpi library (to apply inflation to budget based on year)
Determine Baseline Model •Filtered to small set of features that had strongest correlation with US Box Office Gross •Tested log transform vs no transform	Pandas, Sklearn

Train - Validate - Test •Utilized cross validation •Tested two models	Sklearn

Algorithms

Feature Engineering

- Converting categorical features to binary dummy variables
- Combined Categorical Features

Model

- Linear Regression with Cross Validation