The Prediction and Features of Movies Gross Earnings

Abstract

This project aims to help investors decide which and what kind of movies are worth investing in. Several factors influence the gross earning of the movies. Among the factors we can get before the realease of movies, the result shows that the budget, director, production company, and runtime greatly influence gross earning in movies.

Design

What kind of movies have the potential highest gross earning and are the best choice for the investors. This project aims to build the regression model and use some feature analysis to understand the relation among movies' features better. In this way, the new investors can use the result of this project as a reference to decide what kind of movie are the most worthwhile for them to invest in. Also, the film companies can know which features of the movies influence their gross earning the most and adjust their filming plan to increase the revenue.

Data

The IMDb website is the source of data in this project. We gathered movie data from 2015 to 2020 using BeautifulSoup. We scrapped multiple pages, and each page has 100 movies. The raw database has 7856 data, and after cleaning, 856 data points were left. The factors, runtime, budget, MPAA Rating, genre, director, writer, stars, production, country, language, and release date (release month and years since release) are used in the prediction model.

Algorithms

- Feature engineering- variance inflation factor, adding polynomial terms, adding interaction terms, transferring categorical data into dummies, and choose the top ten row
- Cross validation train/validation/test
- Linear regression, polynomial regression, regularization (RidgeCV, find the optimal penalty), LASSO (LassoCV, find the optimal lambda), elastic net (ElasticNetCV), random forest regression, and gradient boosted regression

Tools

We used beautiful soup to scrap the data from the IMDb website. Also, the pandas, NumPy, and sklearn are used in organizing the data and doing the regression analysis. Finally, the visualization and figures are displayed by the matplotlib and seaborn modules.

Communication