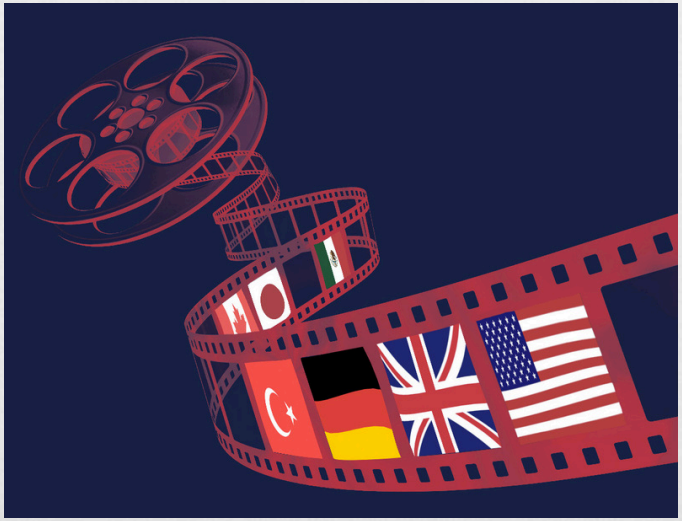
Predicting USA Share of Total Box Office Revenue

Maddy O'Brien Jones

IMDb

The Question

What influences whether a movie is more successful in the USA or abroad?

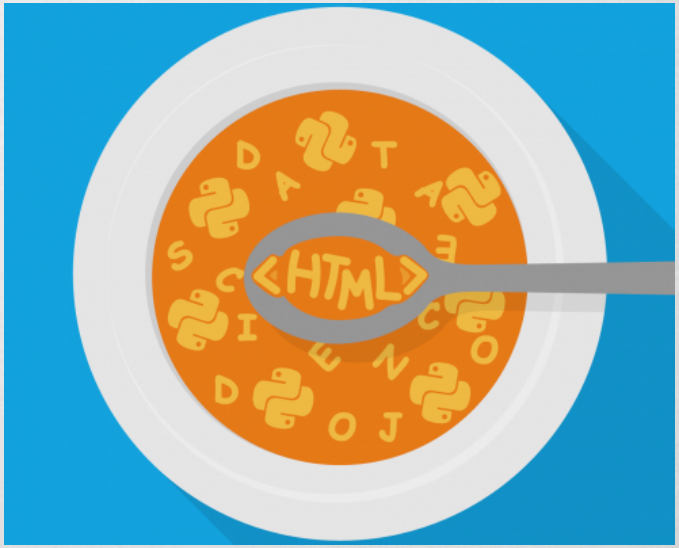


Selenium

- Scraping genre pages of IMDB
- Collecting movie-specific codes

BeautifulSoup

- IMDB
  - Basic information
  - Budget and revenue
- The Numbers (supplementary)
  - Budget and revenue

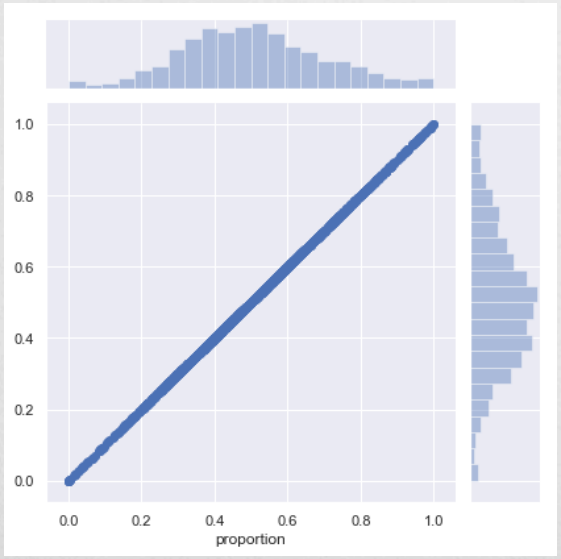


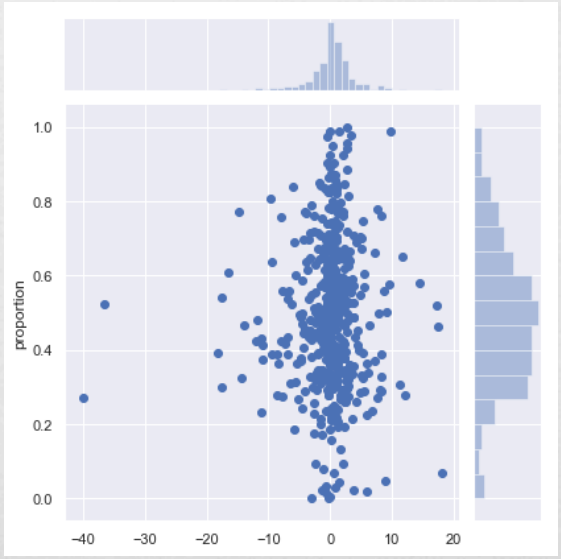
Feature Selection

- Numerical
  - USA gross box office revenue
  - International box office revenue
  - Budget
  - Year
  - Runtime
  - User rating
  - Ratings count
- Categorical
  - MPAA rating
  - Director
  - Country
  - Language
  - Genre

Linear and Polynomial Regression

- Best performing model was simple linear regression
- Overfitting problems
- Residuals for each were normally distributed





Decision Tree Models

- Best performing model was Random Forest
- Issues with overfitting but performed relatively better
- Residuals normally distributed

Predictors

- Linear/polynomial
  - Runtime
  - Crime, romance, comedy
  - Ratings count
  - User rating
- Decision tree models
  - USA gross
  - Budget
  - Ratings count
  - Year
  - Runtime
  - User rating
  - Comedy, horror
  - USA

Conclusions

- Year and genre relatively important
- Linear models relied more on categorical features
- Decision tree models relied more on numerical features
- More data and features necessary to build optimal model