Predicting USA Share of Total Box Office Revenue

Maddy O'Brien Jones

The Question

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

Where should we focus our marketing?

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
  - Budget
  - Year
  - Runtime
  - User rating
  - Ratings count
- Categorical
  - MPAA rating
  - Director
  - Country
  - Genre

Linear and Polynomial Regression

- Best performing model was simple linear regression
- Overfitting problems
- Skeptical about linear regression for this problem

Model	Training Set	Test Set
Linear	.274	-30 septillion
2nd Degree	.636	-7 septillion
3rd Degree	1.0	-1818
Ridge (3rd Degree)	.935	-4.89
Lasso (3rd Degree)	.193	.066

Linear Model

- Country and director were very important

2nd Degree

- Directors very important

3rd Degree

- Prioritized runtime and genre

Ridge

- Prioritized runtime and genre

Lasso

- Prioritized genres and directors

Decision Tree Models

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

Model	Training Set	Test Set
Random Forest	.907	.277
Gradient-Boosted	.999	.109

Best Predictors

- Linear/polynomial
  - UK
  - USA
  - Steven Spielberg
- Decision tree models
  - USA gross
  - Budget
  - Ratings count
  - Year
  - Runtime
  - User rating
  - Comedy
  - USA

Conclusions

- Comedies and horror movies tend to perform better in the USA than abroad
- More data and features necessary to build better model
  - Marketing
  - Actors
  - Series