PRODUCTION Box Office Prediction

STARRING Okechukwu Ofili, Andy Roberts, Aida Rahim, Matt Bildzok

DATE	SCENE	TAKE
Mar. 8, 2021	225	#1

Can we predict a movie's worldwide box office revenue?

Dataset: 3000 movies

Year released: 1924 - 2019

Budget range: Maximum $380M

Correlates with: budget, popularity, runtime

Dataset Features

Language:

English 88% (French, Spanish)

Largest Budget:

Pirates of the Caribbean: On Stranger Tides ($380M)

Highest Revenue:

The Avengers (2012 - $1.5B) on a budget of $220M

Dictionaries

- keyword, genre, production company, production country, cast, crew
- Extracted and converted to a usable format
- Director and director of photography

Methodology

- Model based only on numerical features
  - ○ Linear Regression
  - ○ Random Forest
- Incorporate text data and rerun model
  - ○ Linear Regression

Model Performance

	Train R^2	Test R^2	Test RMSE
Linear Regression (numeric)	0.61	0.5	$78.8M
Random Forest (numeric)	0.91	0.56	$74.2M
Linear Regression (includes text)	0.99	0.36	$89.1M

Random Forest (includes text) had to be interrupted

Baseline RMSE: $112M

Conclusions / Recommendations

- $1 of budget -> $2.46 revenue (all else being equal)
- Modeled box office revenue by 3 methods - all overfit
- Simple linear regression (Model 1), proved to have the best fit
- Random forest model (Model 2) had the lowest RMSE.

- Model refinement: incorporate a Ridge or Lasso regularization to reduce our variance
- Manually input missing data - could have added to our data for improved performance

QUESTIONS?