Predicting Box Office Revenue For Movies

John Lee

Objective

- Predict the lifetime domestic gross revenue of movies **before** they are released
- Provide early insight to studio executives and producers who want **high predictive ability**, to help determine whether they should greenlight a project

Web Scraping

1979 ~ 2019
30 years of box office data

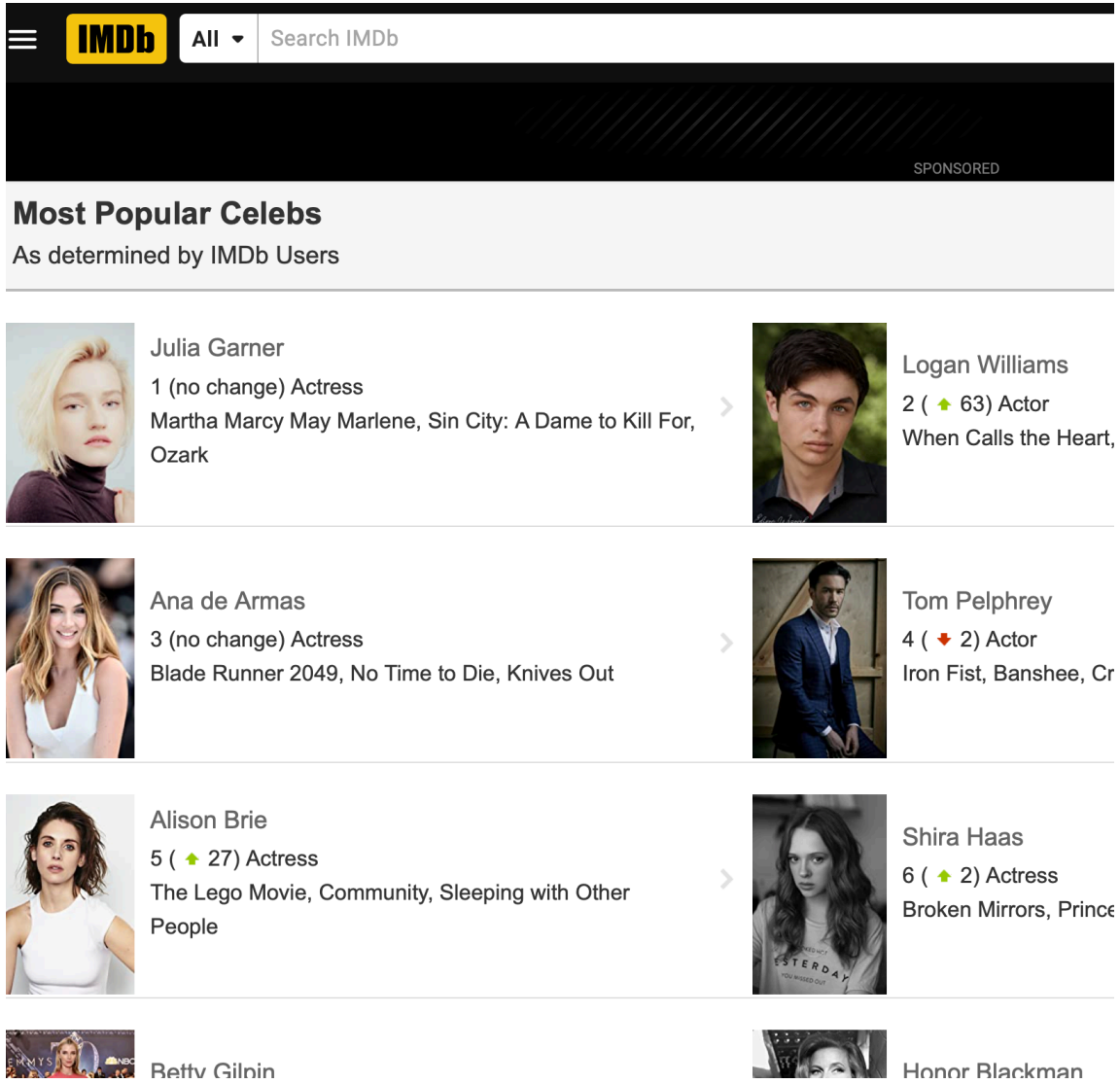
3,148
Movies with box office data

9
Features

Runtime, certificate, release date, genre, director, budget, cast, star popularity and domestic gross

Popular Celebs

- Scraped the top 1,000 most **popular celebrities** from IMDB
- Based on IMDB's STARmeter rankings which are based on **frequency** and **volume** of people visiting a person's page

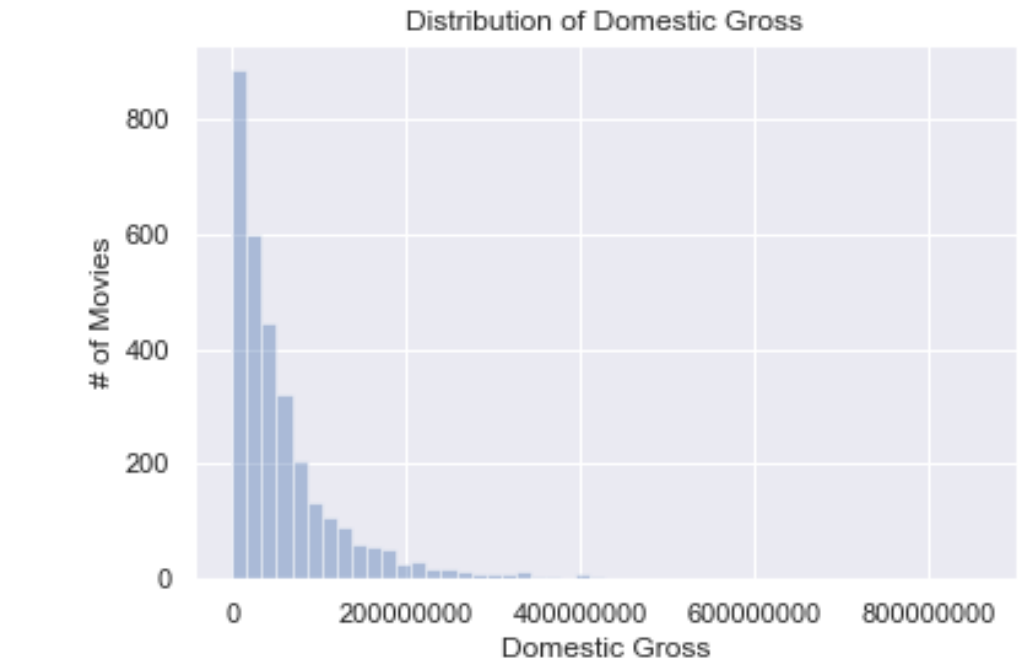


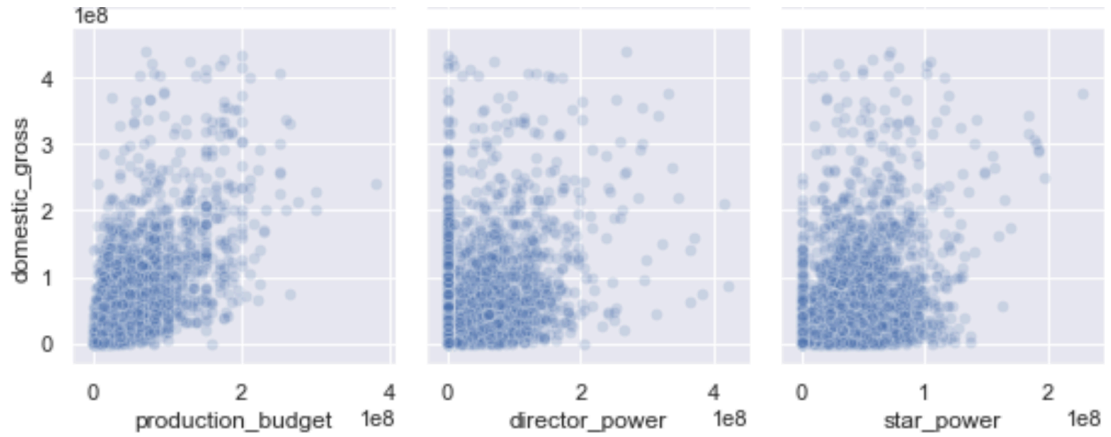
Engineered Features

- Director Power
  - Average domestic revenue of all prior movies from director
- Star Power
  - Average domestic revenue of all prior movies star starred in
- Star Points
- Star Appearances
- Genre Count

Exploratory Analysis

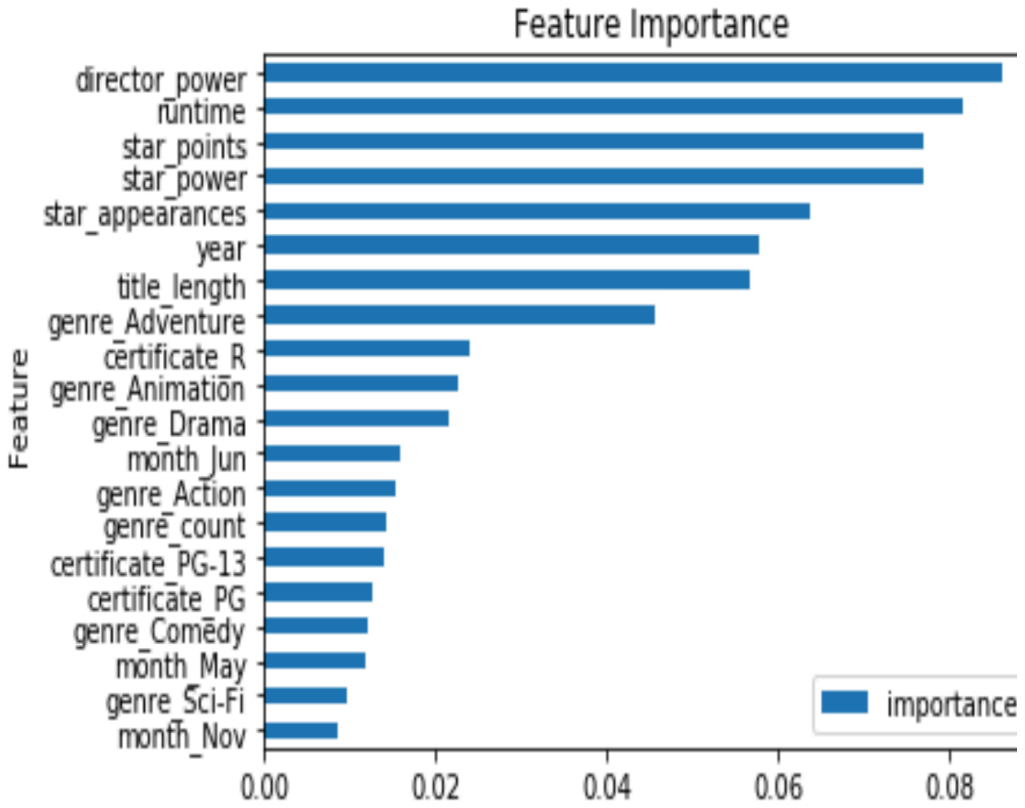
- Largest concentration within the lower grossing movies
- Long, right-reaching tail due to a few blockbuster hits
- Some relationship exists with budget, director power and star power





Modeling

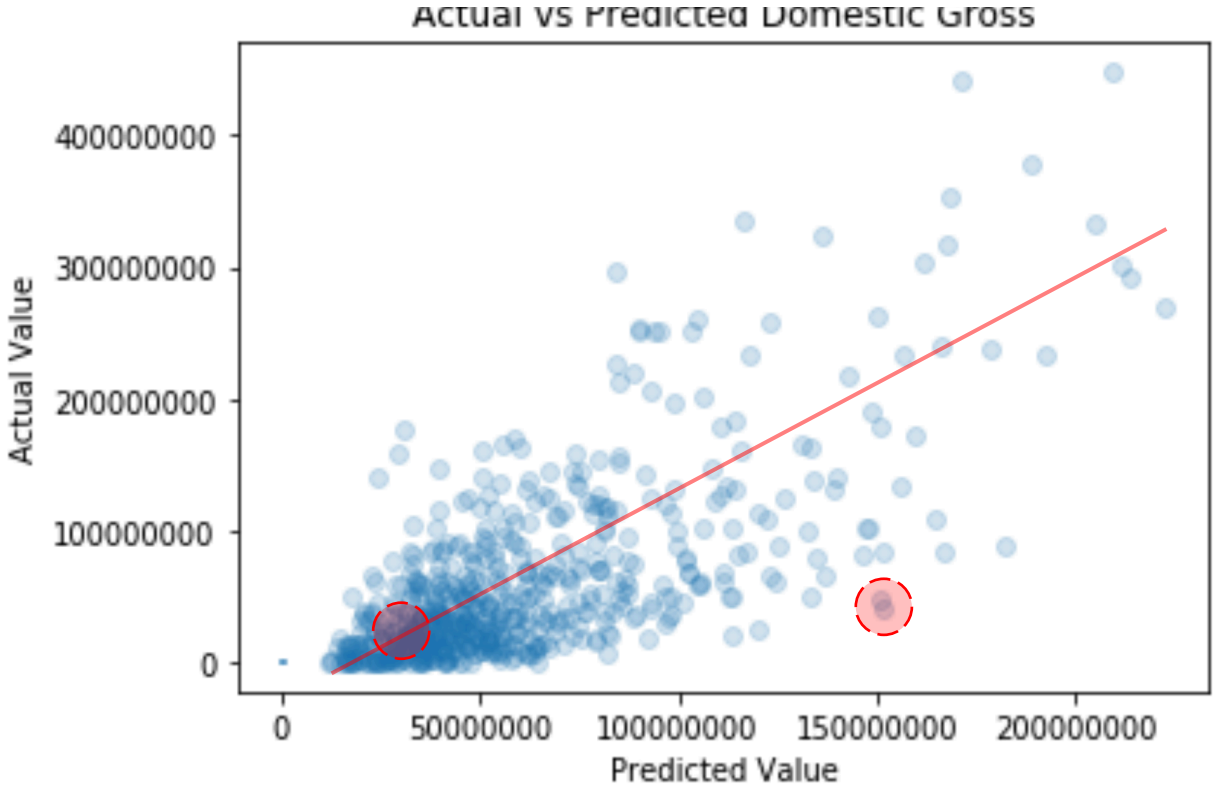
- Removed several **outliers**
- Model trained on **all** features
- **Categorical** features encoding
- **Polynomial** and **interaction** terms
- 80/20 train-test split, using **5-fold cross validation** on polynomial regression with LASSO ~ R^2 score of 0.371
- **Random Forest** ~ R^2 score of **0.458**



Results

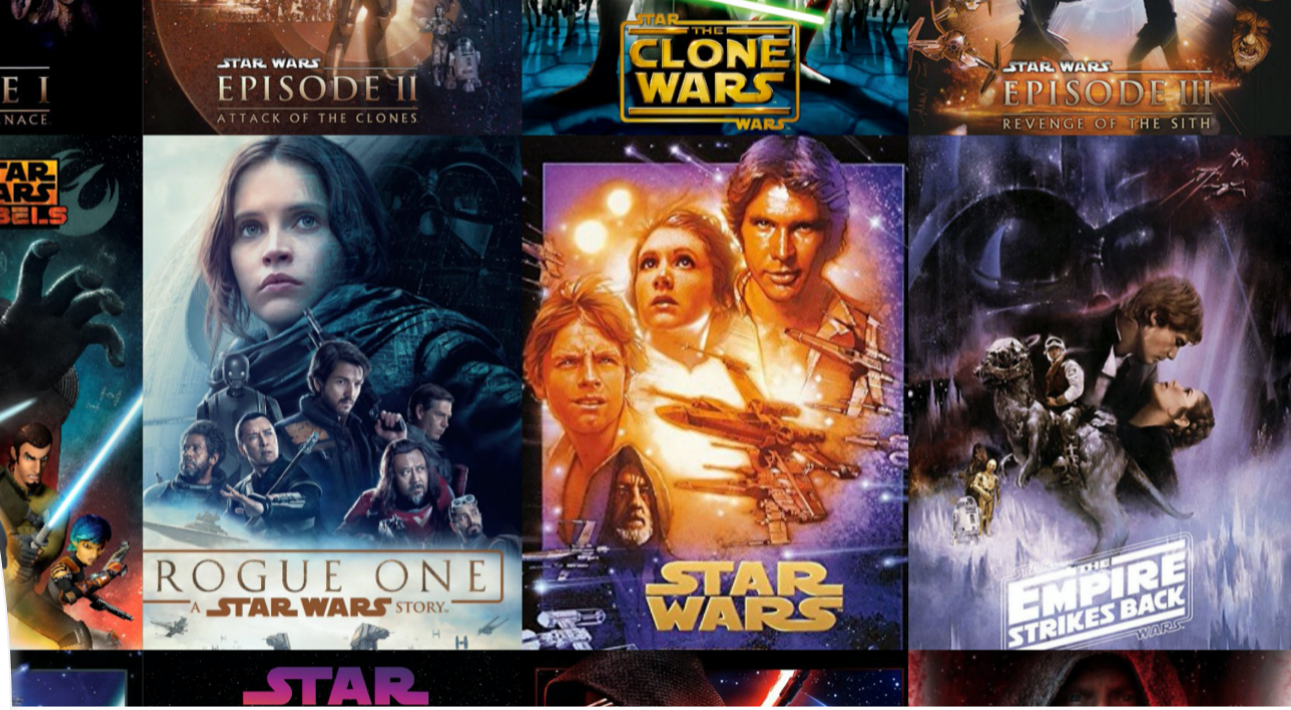
- R^2 of **0.498** on test data
- RMSE of $47,637,781
- The predicted values, **on average**, were off from actual values by **$33,338,138** (MAE)

Movie	Release Year	Predicted Gross	Actual Gross	% Error
Johnny English Strikes Again	2018	154,828,300	4,412,170	+97.1%
Ready or Not	2019	32,711,070	28,714,231	+12.2%



Future Work

- More features including:
  - Franchises / major production companies
  - Cast / crew popularity on social media
  - Is it a sequel?
  - Trends via clustering and times series
  - Award nominations and wins
  - Streaming regime impact



Thank you!