Predicting the Domestic Gross of Movies

By Alex Smith

Overview

- Problem and Background
- Data
- Models
- Findings and Future Direction

Problem and Background

Problem and Background

- 1980: domestic total gross = 2.8 billion dollars
- 2016: domestic total gross = 11.37 billion dollars
- Can we predict dtg of a movie before it’s released?
    - Movie production studio
    - Executive producers (investors)

https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/

Data

Data Collection

- BeautifulSoup to scrape Box Office Mojo
- Golden-globe winning directors: list scraped from Wikipedia
- Engineered features
- Final features: title, director, budget, runtime, genre(s), rating, release date, year, season, winning director, domestic total gross

Models

Models

- Scikit-learn
- Scoring metric: R-squared
- 2 models with 5-fold cross-validation:
  - 1. All features except genre, release date, director, title
  - 2. Logged budget, logged dtg ➜ higher R-squared

Lasso / L1 Regularization

- Baseline
- Grid search: normalize, tiling for alpha
- Coefficients
- Final model: better R-squared on test than on train

Findings and Future Direction

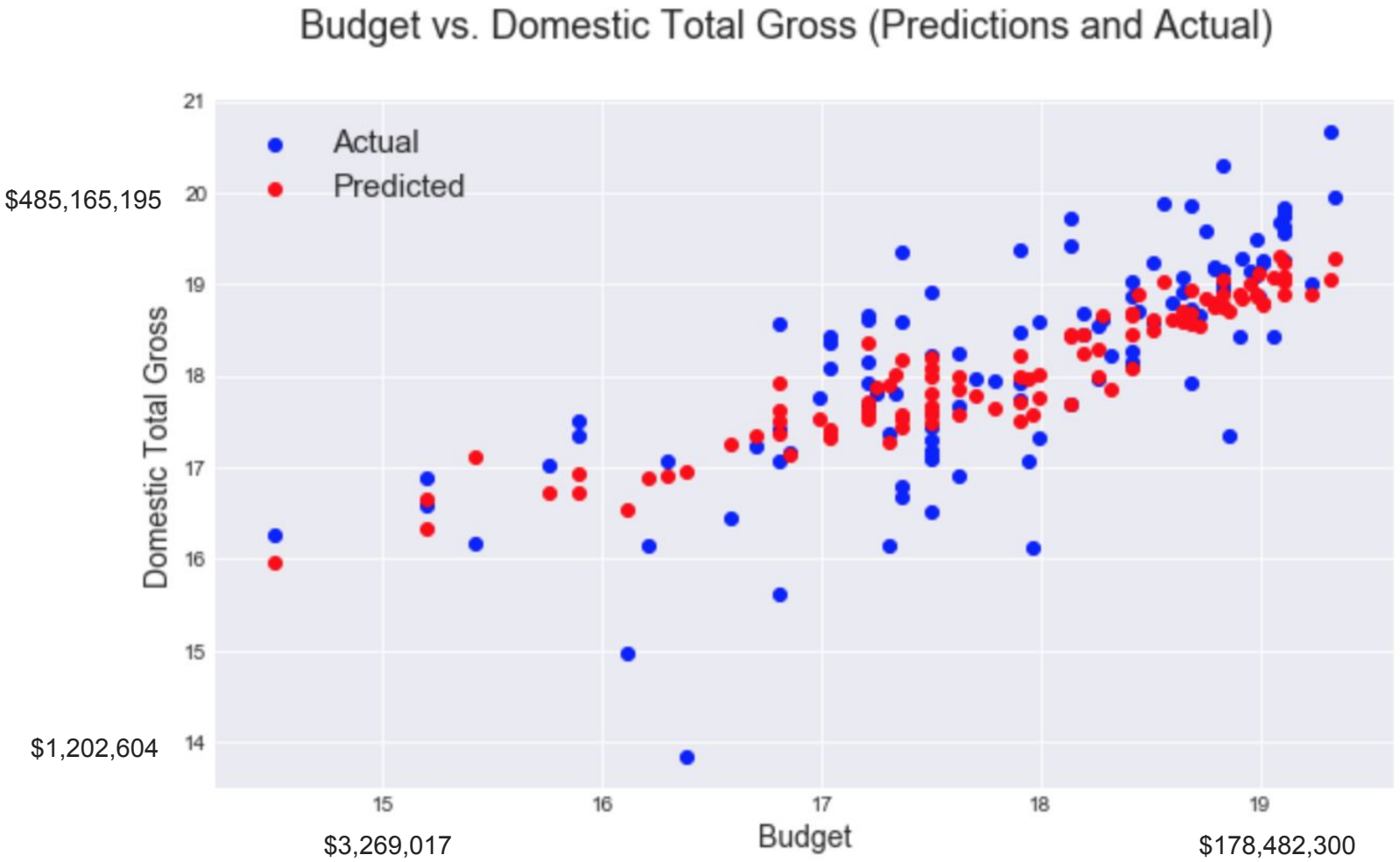

Residuals
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Residuals
16.0
16.5
17.0
17.5
18.0
18.5
19.0
Predictions of Domestic Total Gross

R-Squared Comparison

	Linear Regression	Lasso (Non-optimized)	Lasso (Optimized)
All features but genre	0.407	n/a	n/a
All features but genre, budget and dtg logged	0.411	0.435 0.615	0.447 0.609

TRAIN

TEST


Budget vs. Domestic Total Gross (Predictions and Actual)
Actual
Predicted
Domestic Total Gross
Budget
21
20
19
18
17
16
15
14
15
16
17
18
19
$485,165,195
$1,202,604
$3,269,017
$178,482,300

Findings and Future Direction

- RMSE: $14,850
- Underfitting
- Other models
- More data
- Other features: genre(s), composer, actor(s) or actress(es), and social media likes or mentions

Thank you for your time

Questions?