Predicting Domestic Box Office Revenues

Chris Chung

Project Overview

Challenge

Better forecast movie success for studios

Sources

IMDb (imdb.com)

- IMDb search filtered to *Feature Film, Released between 2010-01-01 and 2019-12-31, Rating Count at least 25,000 (Sorted by Popularity Ascending) Exclude Adult Films*
- **Dimensions**: Title, Year Released, Genre, MPAA Rating, IMDb Rating, # of IMDb Ratings, Metascore (from Metacritic), # of User Reviews, # of Critic Reviews, Runtime, Director

The Numbers (the-numbers.com)

- **Dimensions**: Budget, Domestic Gross Revenue

Extract and Transform

Extract

- In Python, used BeautifulSoup4 to scrape from IMDb and The Numbers
    - Roughly 1.7K movies from IMDb and roughly 5K from The Numbers

Transform

- Grouped tables for IMDb and The Numbers by Title and Year to avoid matching with duplicate movie titles
    - Yielded roughly 1K movies in the final dataset
- The majority of null values in the dataset were missing Budget, so those rows were removed since Budget has the highest correlation to Domestic Gross
- Split the data to train/test/validate (60/20/20)

Project Model

The initial data modeled with OLS yielded an R^2 of .59

- Ran the initial model with numerical values only to set a baseline, which yielded R^2 values of .63 and .59 against the training and testing set, respectively

The next step was to include dummy variables to quantify the impact of non-numerical features

- Genre and MPAA Ratings
- Movies directed by someone that had a mean domestic gross in the 75th percentile of directors and had directed more than one movie in the data set

DIRECTOR	MEAN_GROSS	MOVIE_COUNT
Anthony Russo	551254947.250	4
Joss Whedon	541181889.000	2
Josh Cooley	434038008.000	1
Chris Buck	433607387.500	2
Patty Jenkins	412563408.000	1

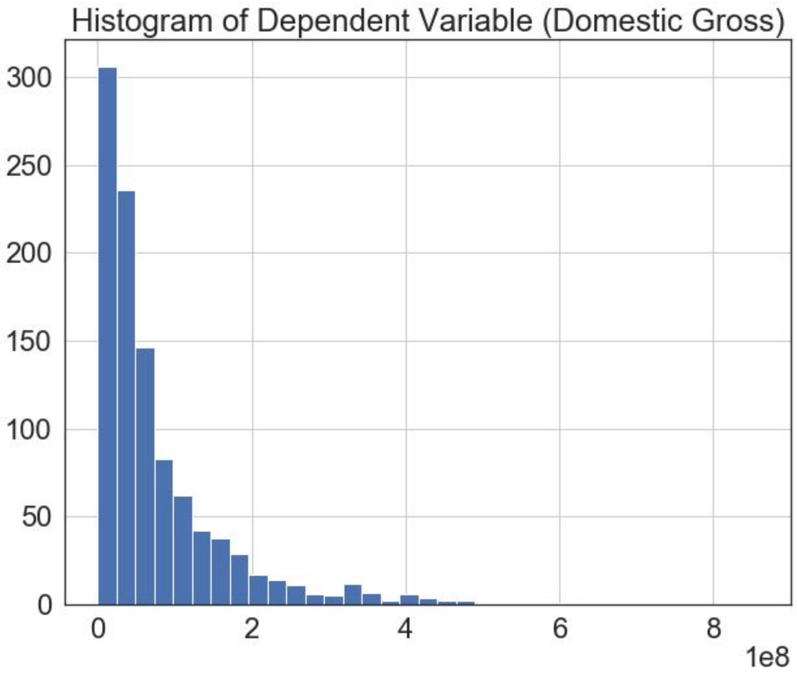
Additional features resulted in small improvements

OLS

- The R^2 improved to .65 after including the dummy variables

Polynomial

- Polynomial Features with degree 2 yielded a score of .44 against the test set vs. .68 for the training set, indicating the model was overfit
- Lasso Regularization limited the model to 28 features from 378



.59 → .72

R^2 improved significantly by following the steps to awesomeness

The budget, having a big name director, and the count of reviews from users and critics were most prominent amongst the 28 remaining features

** Full list of coefficients in the appendix*

Areas to Expand

Features

- Actors
- Sequels
- Marketing budgets
- Seasonality
- Widest release
- Remove any rating features to make model purely predictive

Other Model Ideas

- Assess movie value for streaming platforms

Questions?

Appendix

Final model coefficients

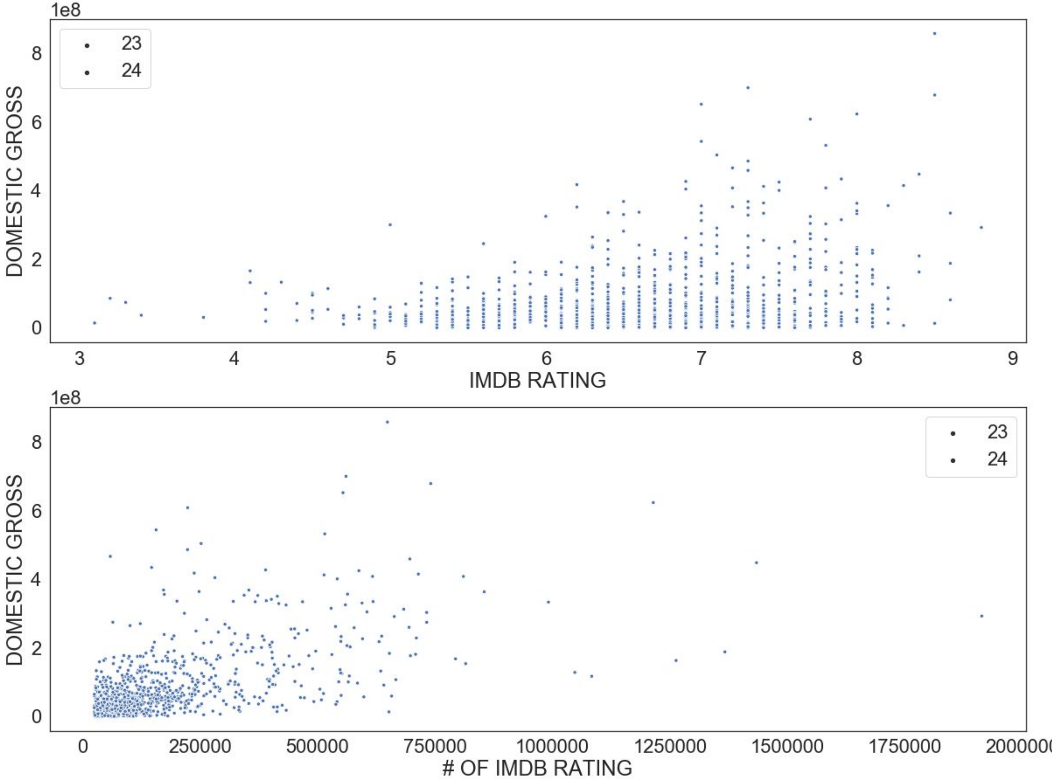
VARIABLE	COEF
NUM_IMDB_RATINGS	10,477,329
IMDB_RATING MPAA_R	-926,081
NUM_IMDB_RATINGS GEN_COMEDY	3,087,669
NUM_IMDB_RATINGS GEN_FAMILY	1,359,286
METASCORE COUNT_USER_REVIEWS	10,226,553
METASCORE BUDGET_TN	22,604,658
METASCORE GEN_DRAMA	-1,804,350
METASCORE GEN_FAMILY	294,078
COUNT_USER_REVIEWS BUDGET_TN	5,758,713
COUNT_USER_REVIEWS GEN_ANIMATION	13,235,828
COUNT_USER_REVIEWS GEN_COMEDY	304,852
COUNT_USER_REVIEWS GEN_FAMILY	1,426,252
COUNT_USER_REVIEWS MPAA_PG	4,283,876
COUNT_USER_REVIEWS MPAA_PG-13	2,852,021

VARIABLE	COEF
COUNT_USER_REVIEWS TOP25_DIRECTOR	7,175,658
COUNT_CRITIC_REVIEWS BUDGET_TN	7,258,438
COUNT_CRITIC_REVIEWS GEN_ADVENTURE	1,484,733
COUNT_CRITIC_REVIEWS GEN_DRAMA	-4,152,781
COUNT_CRITIC_REVIEWS TOP25_DIRECTOR	14,464,046
BUDGET_TN GEN_COMEDY	2,695,260
BUDGET_TN GEN_HORROR	-1,054,773
BUDGET_TN GEN_MYSTERY	-1,212,497
BUDGET_TN MPAA_R	-811,959
BUDGET_TN TOP25_DIRECTOR	3,666,461
GEN_COMEDY TOP25_DIRECTOR	3,483,111
GEN_FAMILY TOP25_DIRECTOR	1,945,478
GEN_ROMANCE GEN_THRILLER	483,122
GEN_SCIFI TOP25_DIRECTOR	1,937,612

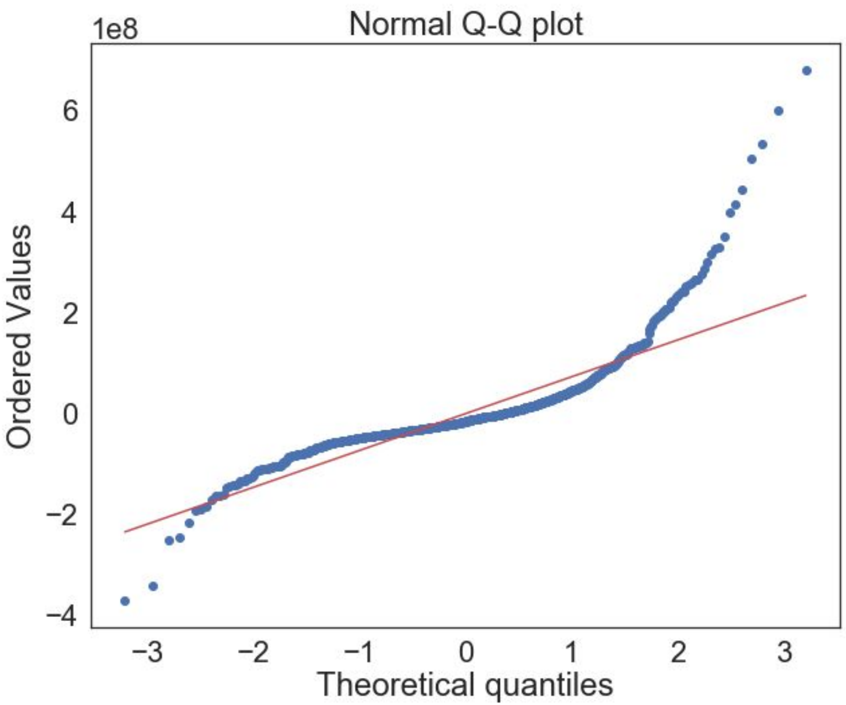
Critical success does not equate to commercial success

Highlight from EDA

Plotting IMDb Rating against Domestic Gross, the figure suggests that how well a movie does critically does not have a large impact on the revenue



Q-Q Plot



Normal Q-Q plot
1e8
6
4
2
0
−2
−4
Ordered Values
−3
−2
−1
0
1
2
3
Theoretical quantiles