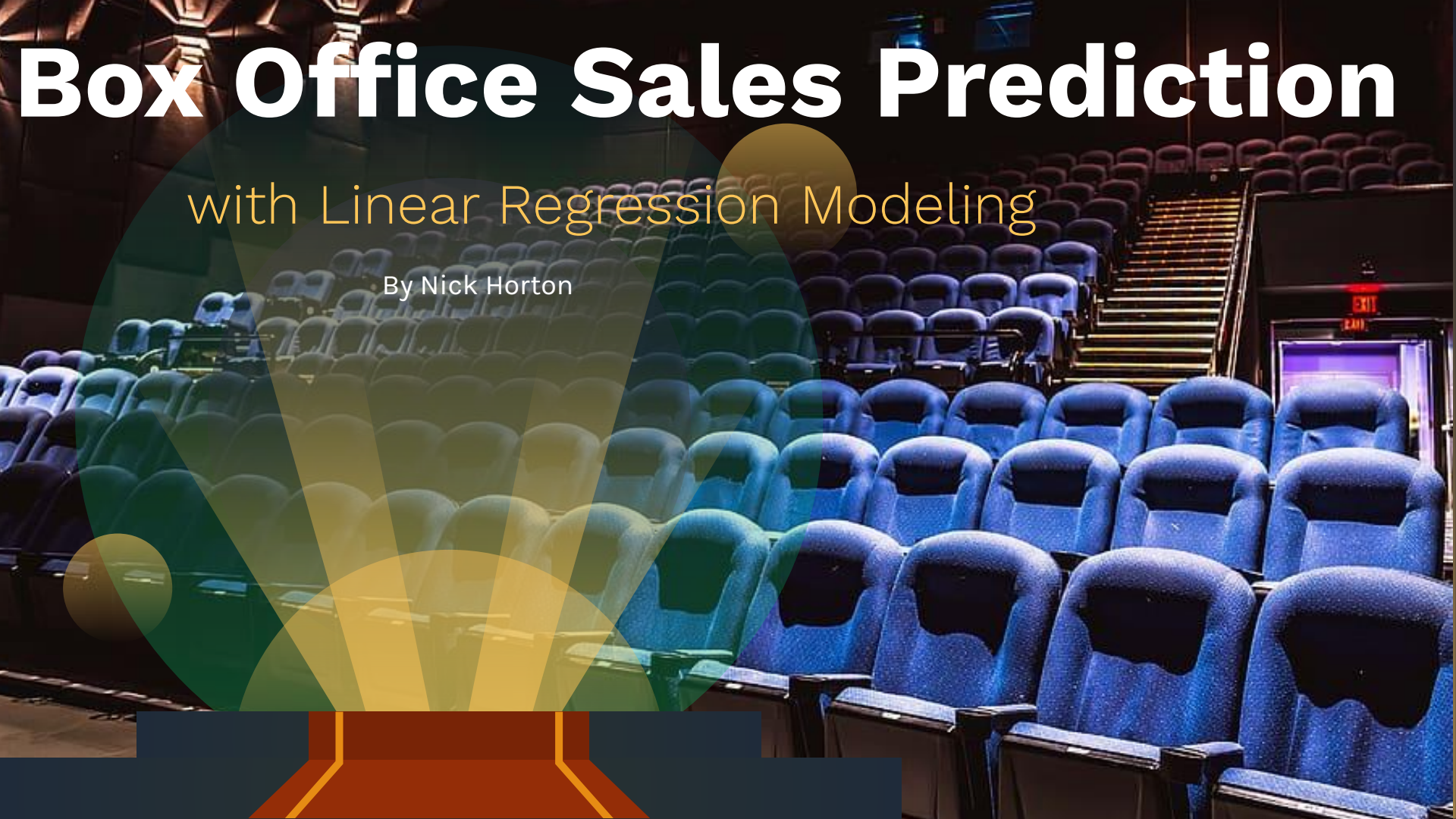


Box Office Sales Prediction

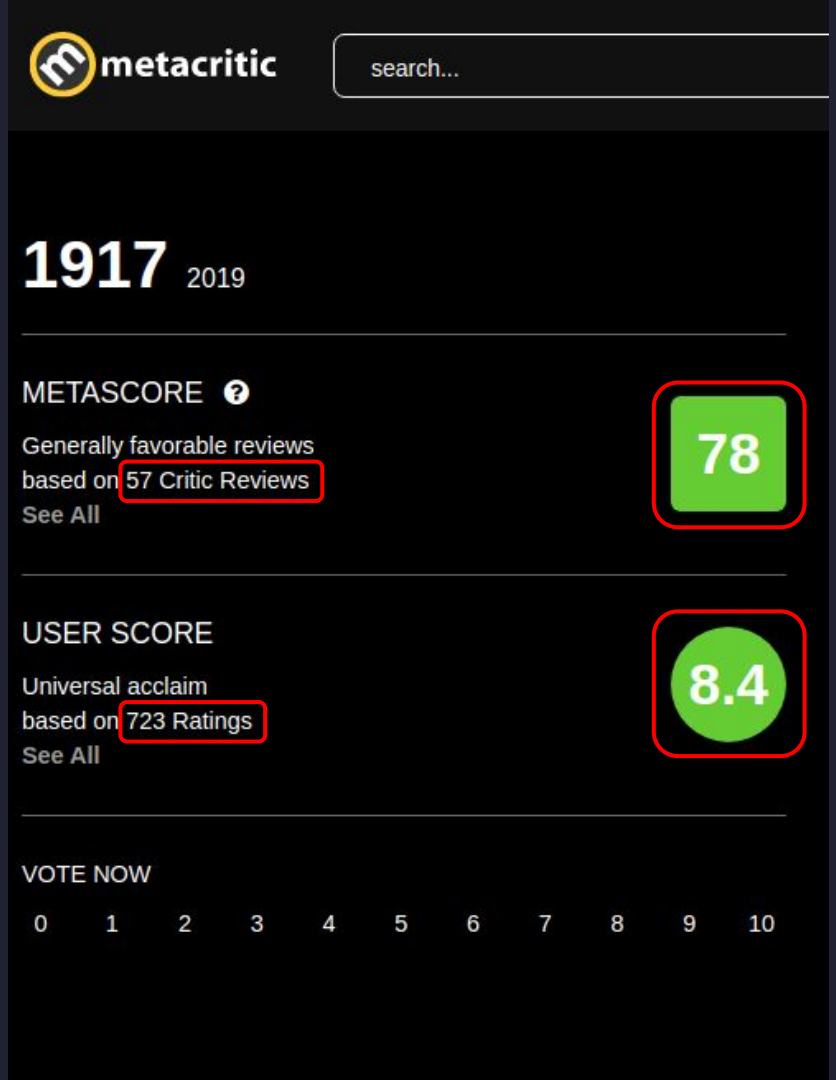
with Linear Regression Modeling

By Nick Horton

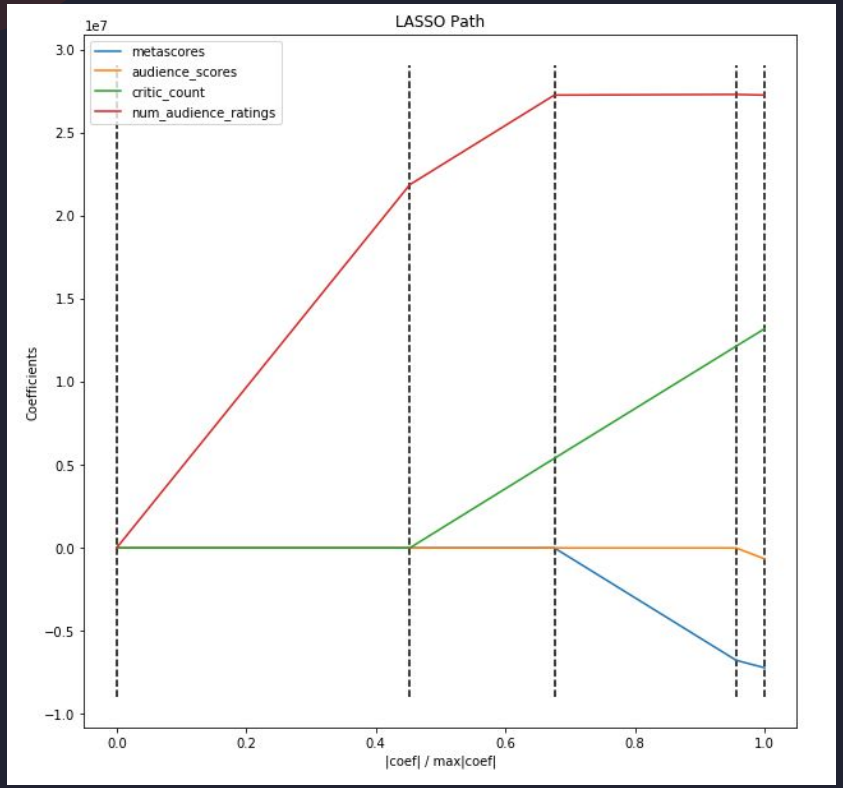
Table of Contents

Question:

Is the number of online reviews for a movie predictive of its box office sales?


metacritic
search...
1917 2019
METASCORE
Generally favorable reviews
based on 57 Critic Reviews
See All
78
USER SCORE
Universal acclaim
based on 723 Ratings
See All
8.4
VOTE NOW
0 1 2 3 4 5 6 7 8 9 10

Scraping



Number of Audience Ratings

The number of audience ratings was by far the most influential feature on box office sales.

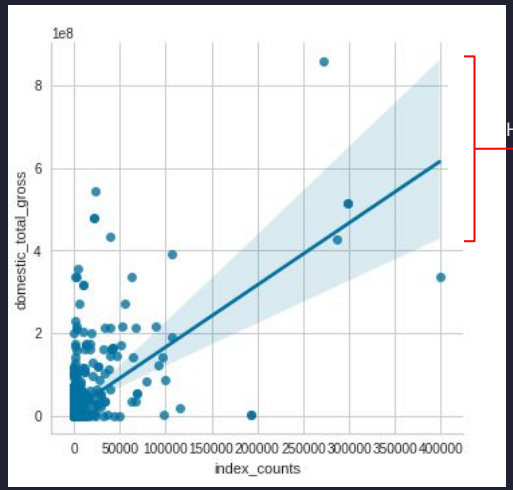
Metascores

The feature I had originally thought would be most important actually looks to negatively correlate with box office sales.

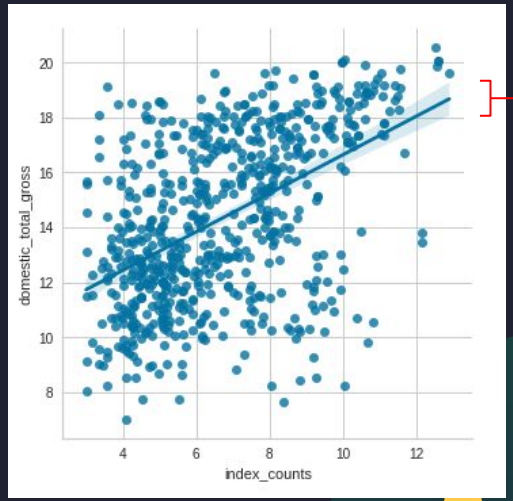
Interaction terms?

Based on the similarity between number of critic ratings and number of audience ratings, I combined them into an interaction term.

Before / After Log Transform



Heteroskedasticity



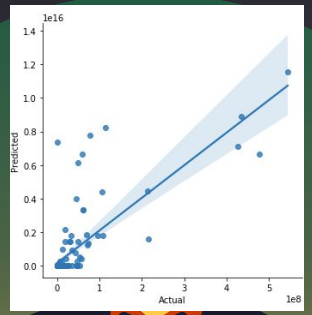
Less Heteroskedasticity

index counts = num audience ratings * num critic ratings

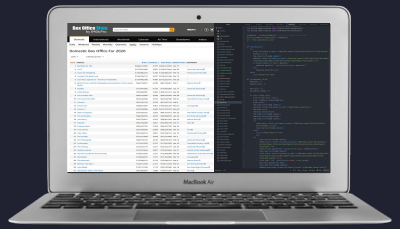
And the Oscar goes to...

LASSO Regression!

R^2 = .734
Mean Absolute Error = 16,051,348.38
Mean Squared Error = 1,607,491,498,383,146.8

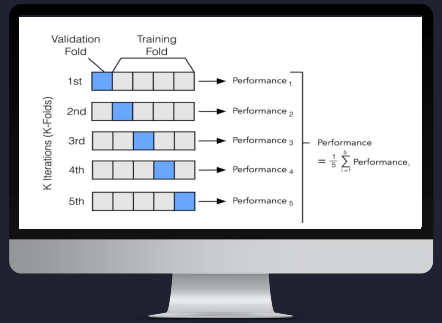


Future Work



Improve Data Collection

Metacritic didn't make it easy for me to scrape data. In next steps for this project, I would try to find the same type of metrics, but from a more accessible location. I'd also try to increase my sample size.



Enhanced Validation

Other steps to take would be improved validation through K-Fold cross validation. I'm think the extra validation is needed because although my R^2 seemed good - .734 - the plot of actual vs. predicted values was heteroskedastic and the condition number was high.