WHAT MAKES A FUNNY MOVIE MONEY?

Linear Regression Interpretive Case Study

Movie attendance is at an all time low, prior to investment a producer might want to interpret which attributes might help in predicting high earnings at the box office.

1134 HIGHEST EARNING COMEDY FILMS OF ALL TIME

Source

IMDB.com (US Box Office)

Target

US Box Office Gross Earnings

=

$$$ raised by ticket sales

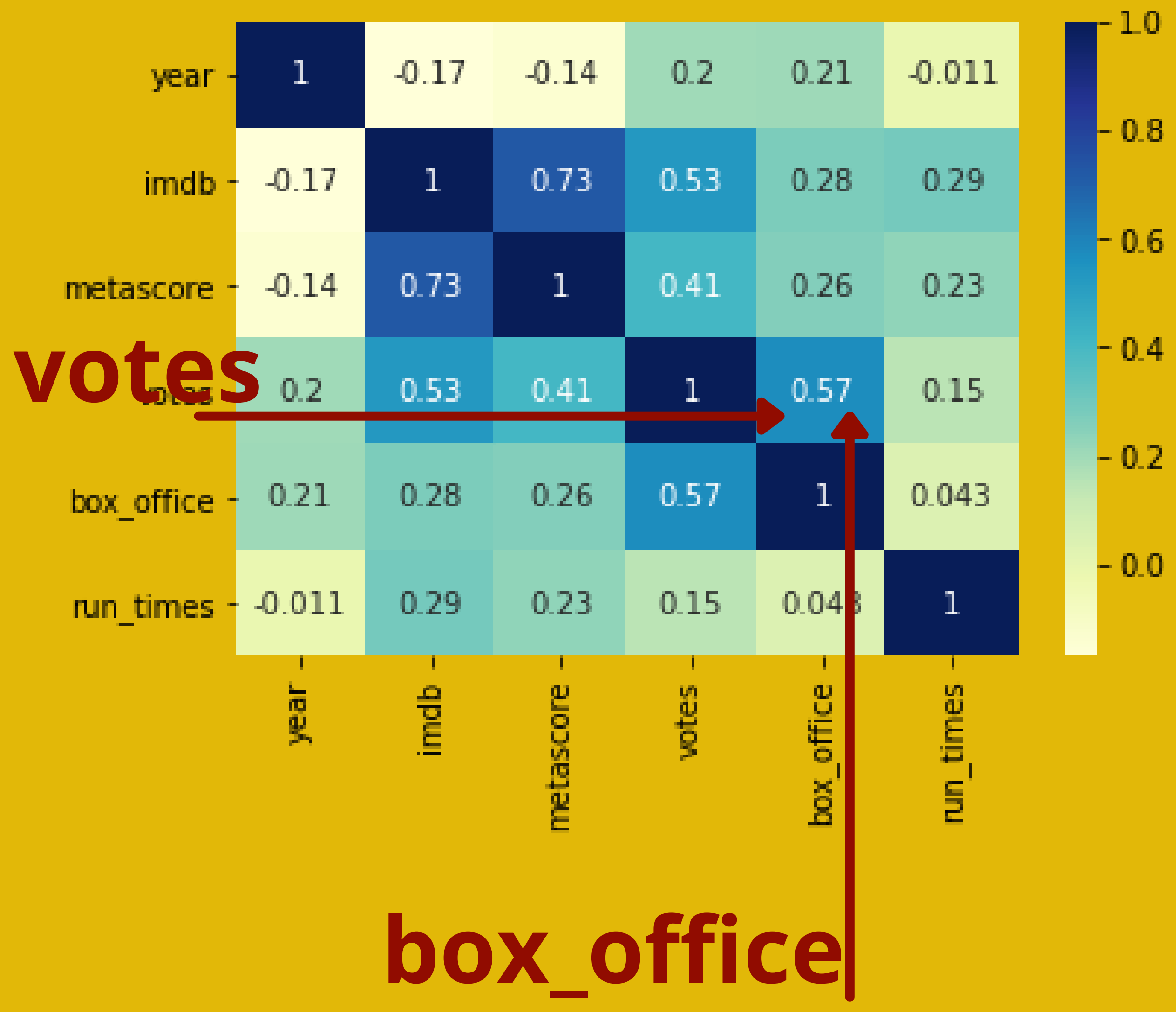
Features

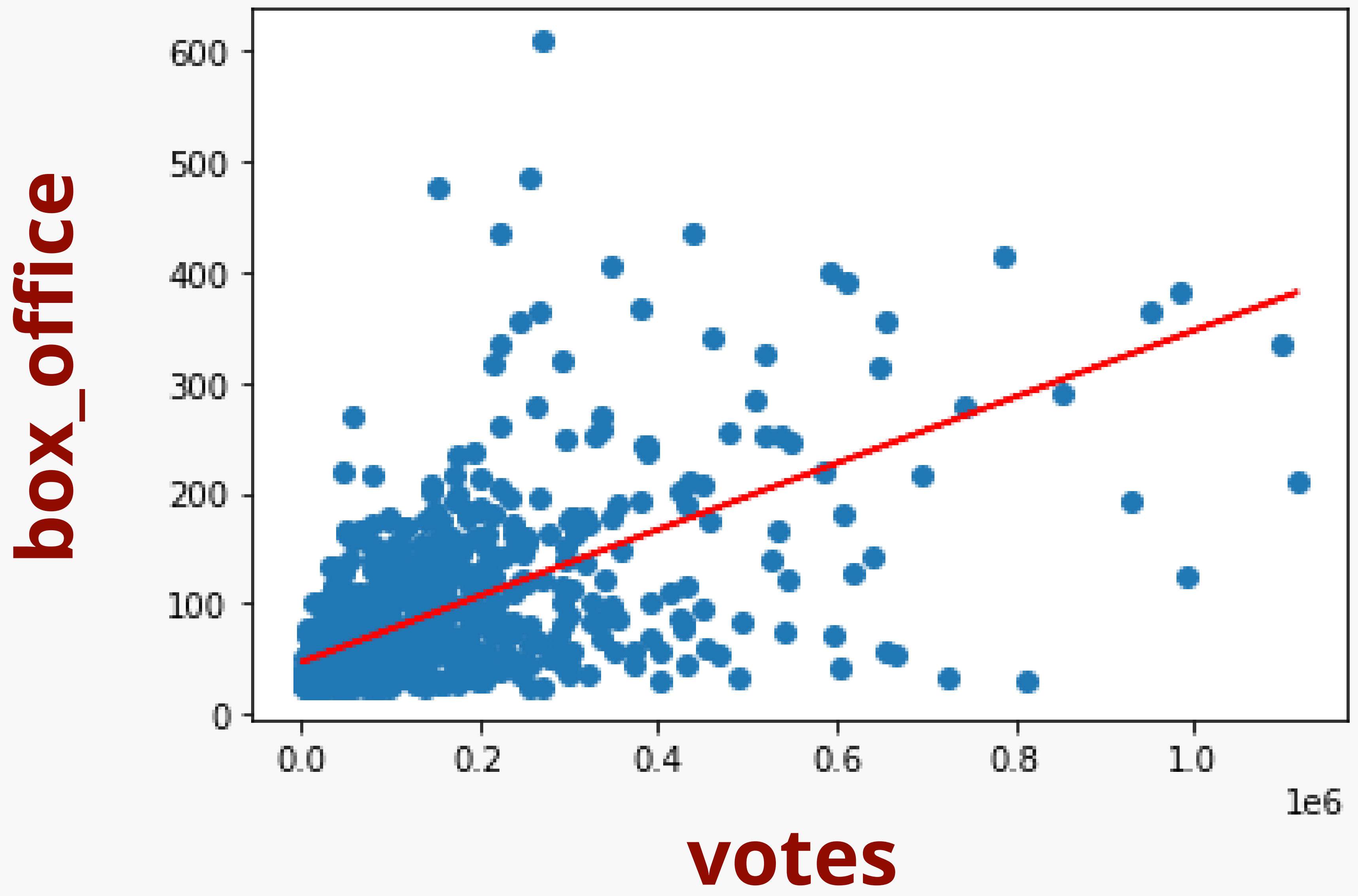
Year of release, IMDB rating, Metascore, Votes, Genre, Rating (PG, PG13, G, &R), Runtime, Director

Acquisition

Beautiful Soup Scraping

Baseline





- The coefficient for Votes is 0.003, and its corresponding p-value is very low, almost 0. That means the coefficient is statistically significant.
- R-squared value is 0.332, which theoretically means that 33.2% of the box_office variance can be explained by the votes column using this line.*
- Prob F-statistic has a very low p-value, so the model fit is statistically significant.

But is it a good model?

WE EXPECT THE NUMBER OF VOTES TO BE CORRELATED WITH US GROSS BOX OFFICE VALUES, BUT WHAT DOES THIS REALLY TELL US? NOT MUCH, SO LETS INCORPORATE OUR NON-NUMERICAL FEATURES AND BUILD A BETTER MODEL.

FEATURE ENGINEERING

(COOKS D ANALYSIS REVEALED NO OUTLIERS, MAX=

GENRE

- 90 Categories cleaned to 8
- Dummy variables created

RATINGS

9 Categories cleaned to 5

DIRECTORS

641 Unique Categories!!
unused in model 2.0

VOTES

standardized for ridge regression

EXPAND & REFINE MODEL

CROSS VALIDATION WITH K-FOLDS

The results were tied for the last random state tested but higher for ridge regression in initial test.

0.408 +- 0.026
SIMPLE LINEAR REGRESSION MEAN

0.408 +- 0.025
RIDGE REGRESSION MEAN

R^2: 0.464
RIDGE REGRESSION TEST

But is it the best model?

LETS DO A LITTLE MORE FEATURE ENGINEERING AND FIND OUT

Directors

Added 6 categories based one the number of films directed, the other category contains mostly 1 & 2 time directors

Animation

Grouped all the animation related films into one category, this group now holds a the highest grossing film which had been grouped in 'other'

EXPAND & REFINE MODEL

CROSS VALIDATION WITH K-FOLDS

Slight improvement, but still tied between linear and ridge, this time we choose simple linear for the test.

0.417 +- 0.074

SIMPLE LINEAR REGRESSION MEAN

0.417 +- 0.074

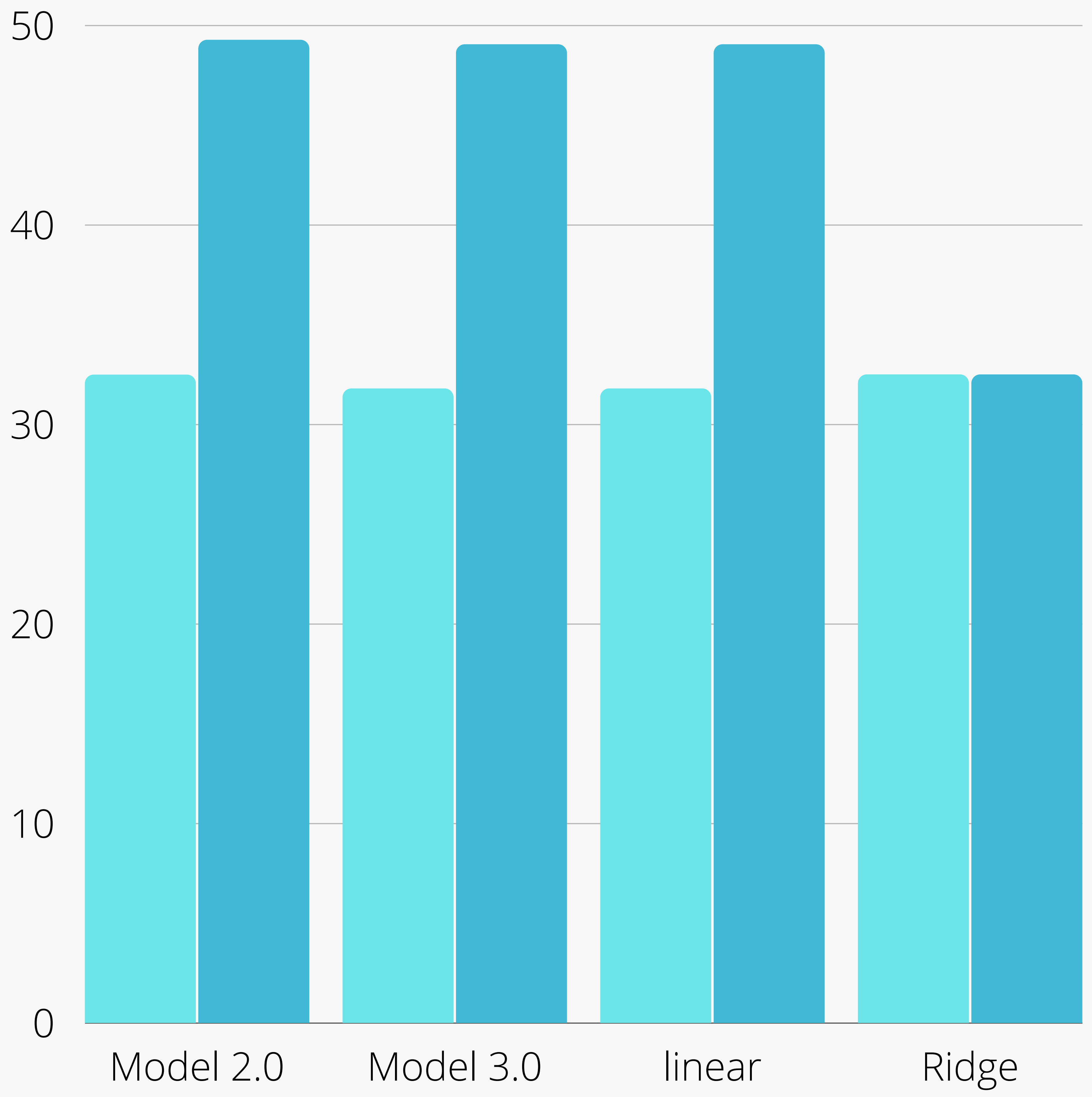
RIDGE REGRESSION MEAN

R^2: 0.486

LINEAR REGRESSION TEST

model 2.0 vs model 3.0

- MAE
- RSME


Model 2.0
Model 3.0
linear
Ridge