Mango | 芒果 DB

using machine learning to predict a movie's audience rating*

*uses 2016-2020 data

Roadmap

—

- Talk about IMDB, RottenTomatoes, Metacritic data, our assumptions
- EDA — what goes with what?
- Introduce linear regression model (degree =1) as a way of zeroing in on predictions (the “Liam-o-meter”)
- Visit 34.212.100.77/mangodb

Our Assumptions

- 7 is average, so scores will be normally distributed about μ = 7

Test-able, we can use this to determine normal or skewed review websites
Options? IMDB, RottenTomatoes, Metacritic

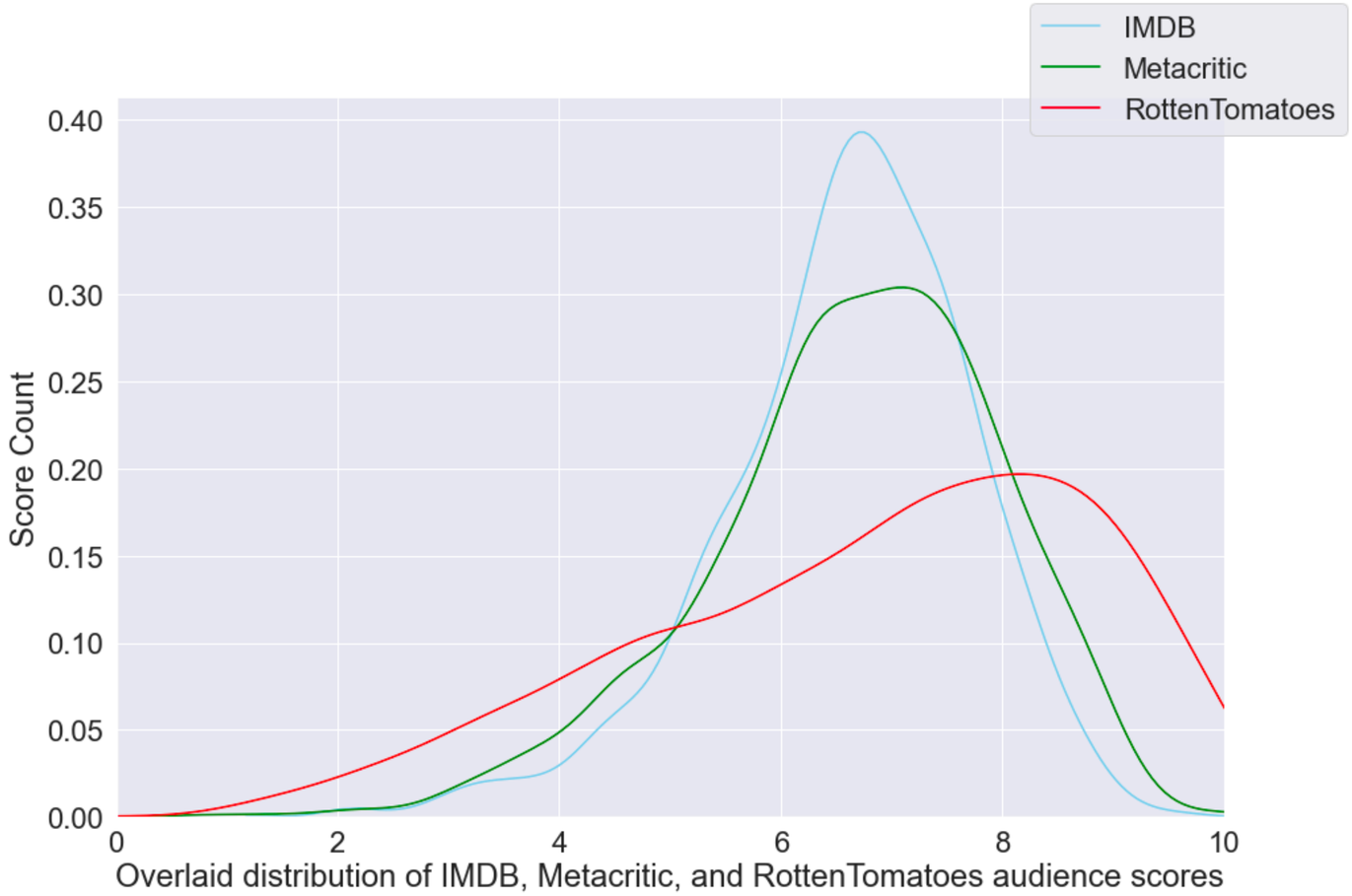
- Genre has to do with ratings and/or movie distributor (like Disney)

Test-able, we can use one-hot-enconding & linear regression
Options? LASSO

Our Assumptions

IMDB, Metacritic

RottenTomatoes



2601

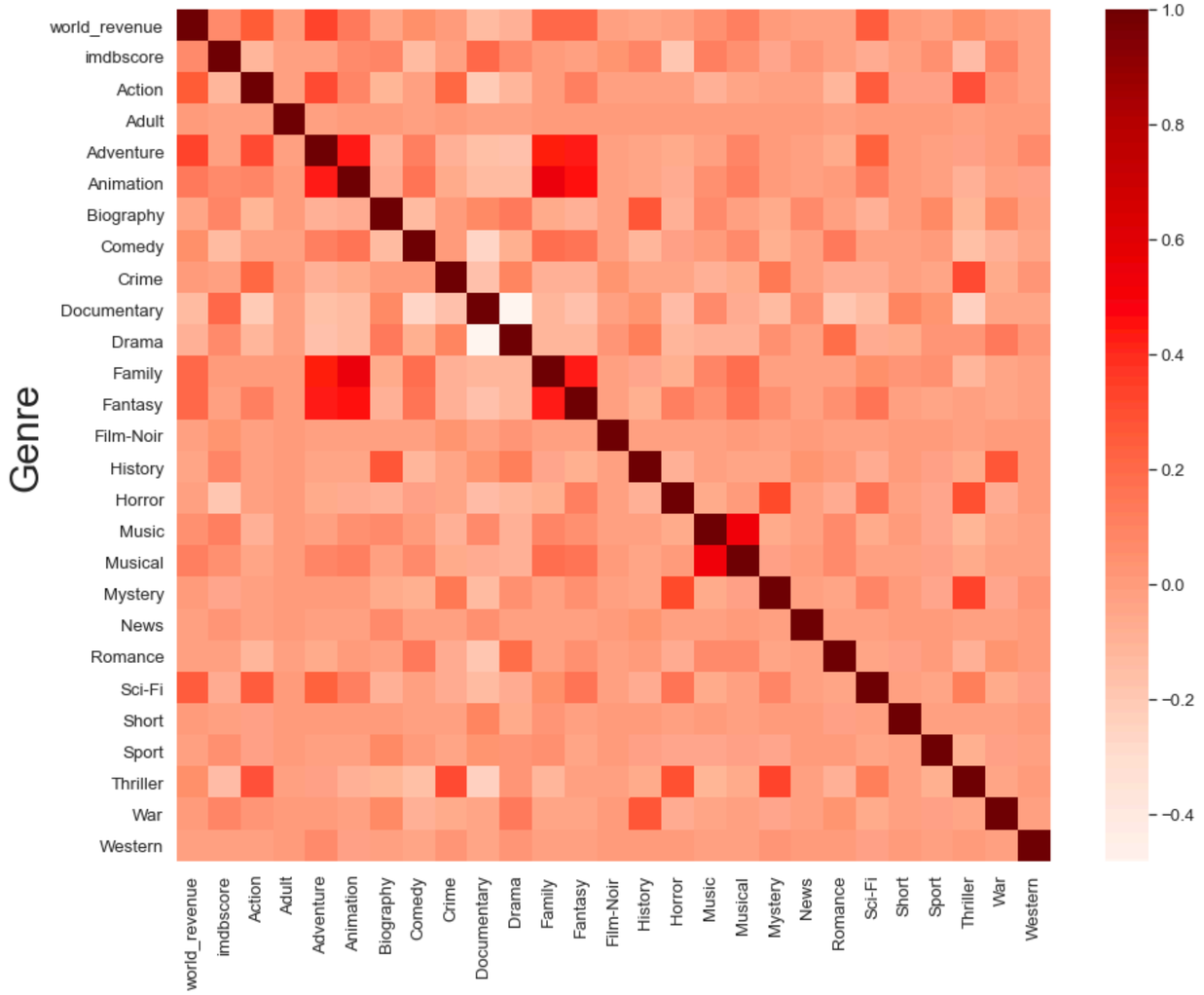
rows prepped for analysis 🧐

Uses 2016 to 2020 data!

At a glance

—

- The darker the red, the stronger the correlation
- There does seem to be some variance, maybe it’s worth looking into
- Onto **linear regression**



By 0.1

Our regression overfits to our training.

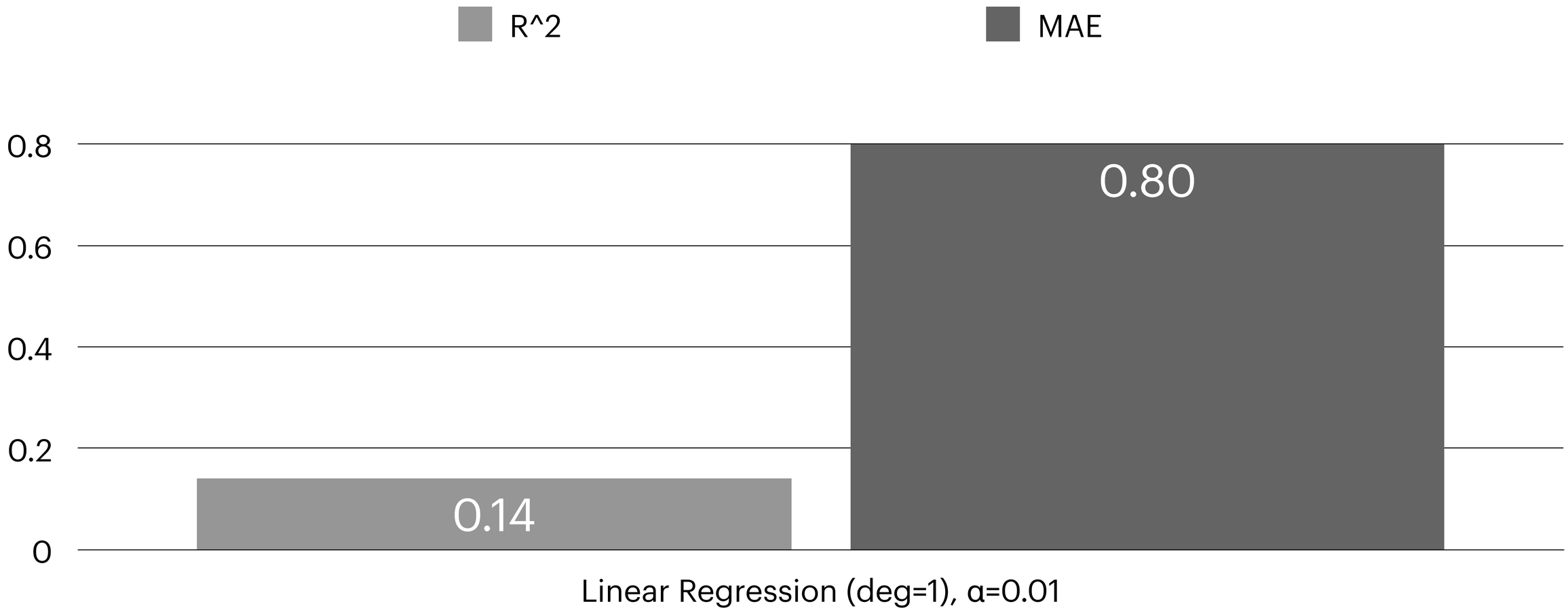
Lasso Cross-Validation

what matters?

Feature	Correlation coefficient
Animation	0.3
Drama	0.32
<30 films/yr. distributor.	-0.12
Music	0.21

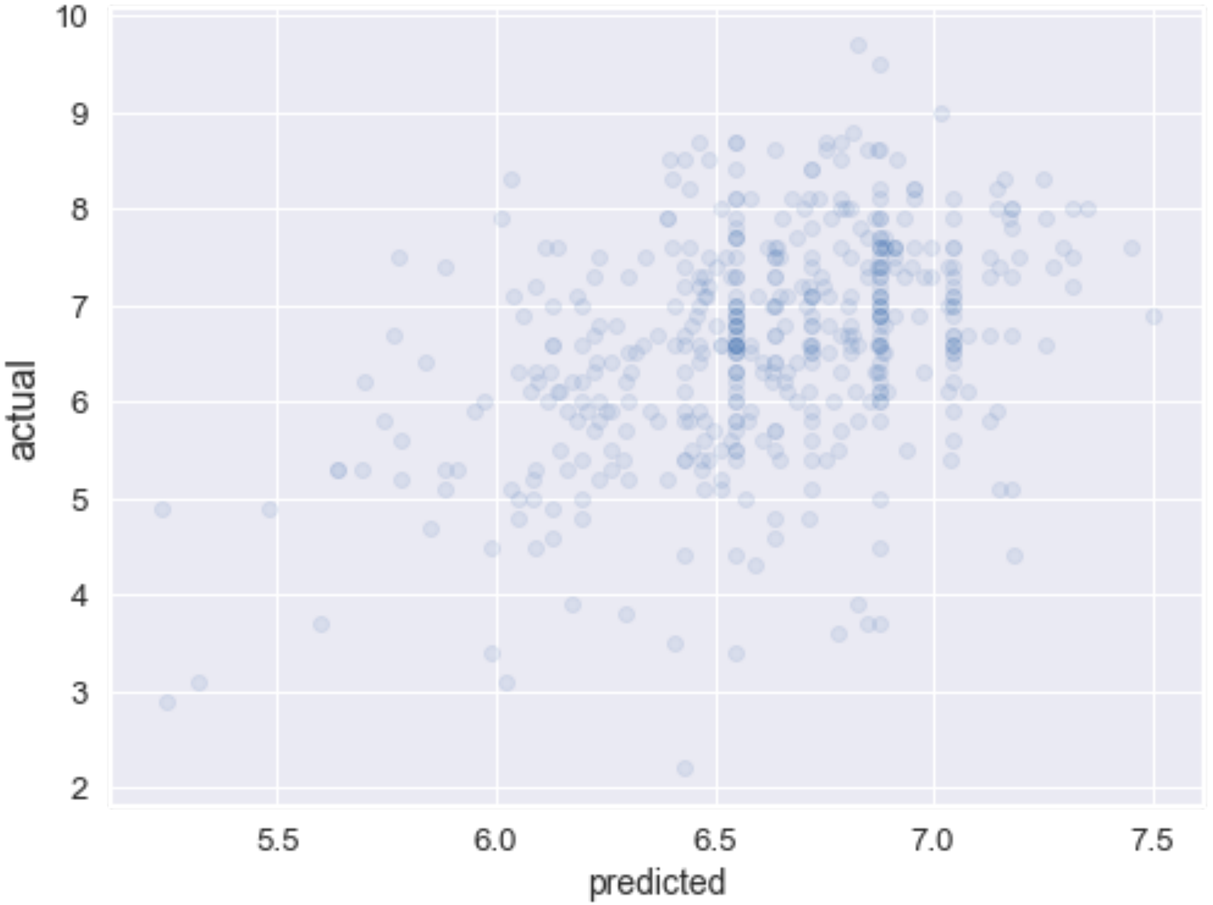
13 features left

Post-LASSO regression results



Visualization

http://34.212.100.77/mangodb



Linear regression (deg =1) residual analysis