

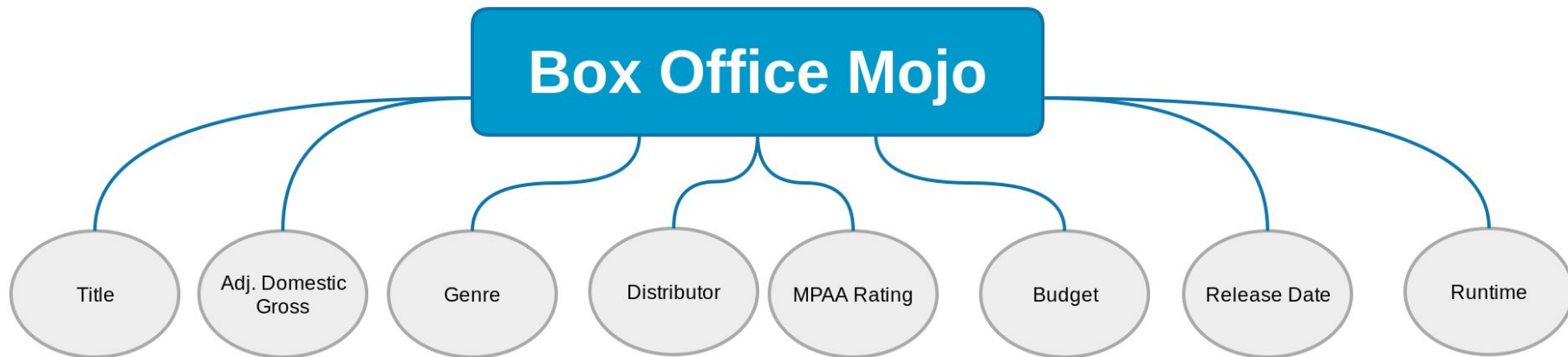


Predicting Movie Earnings

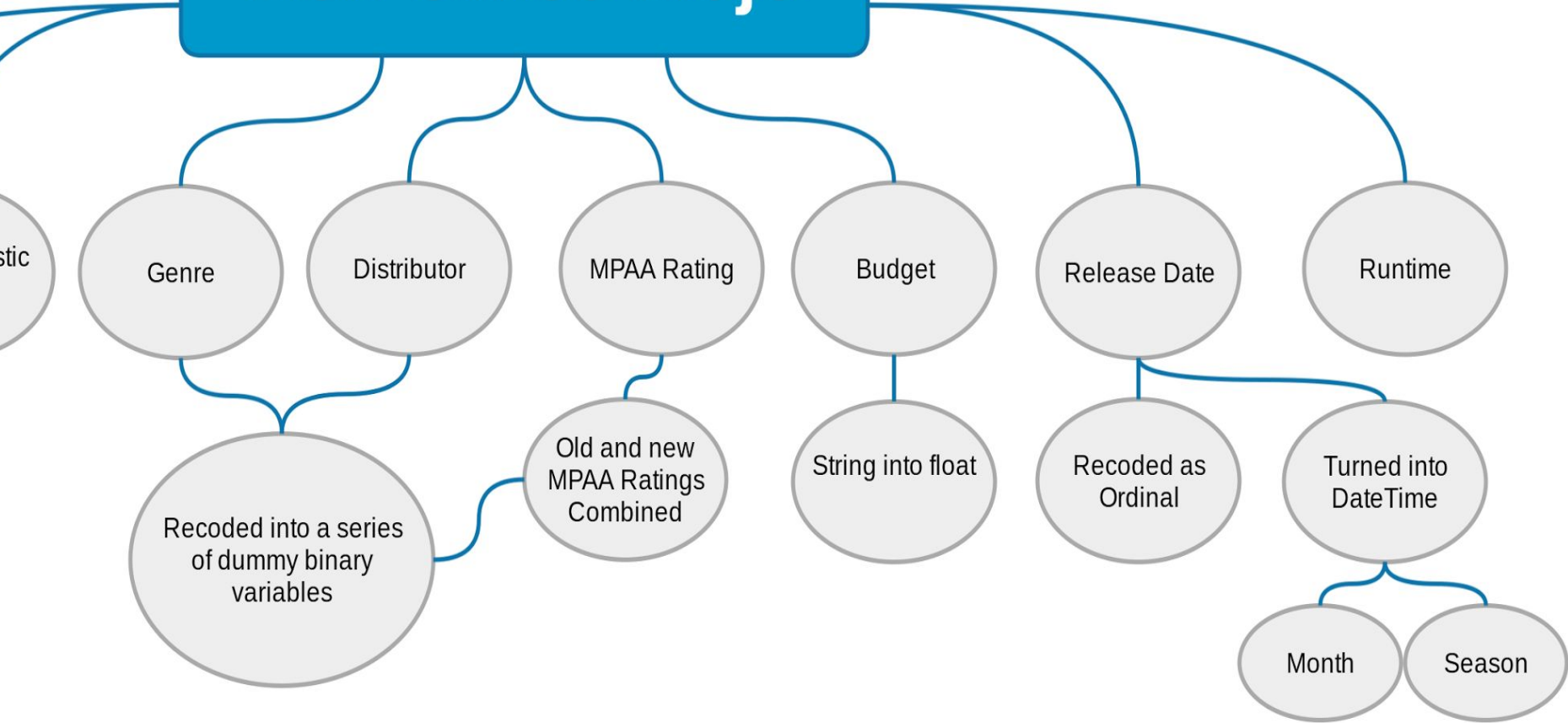
A successful series of failures

Catherine Henderson

Scraping and Munging

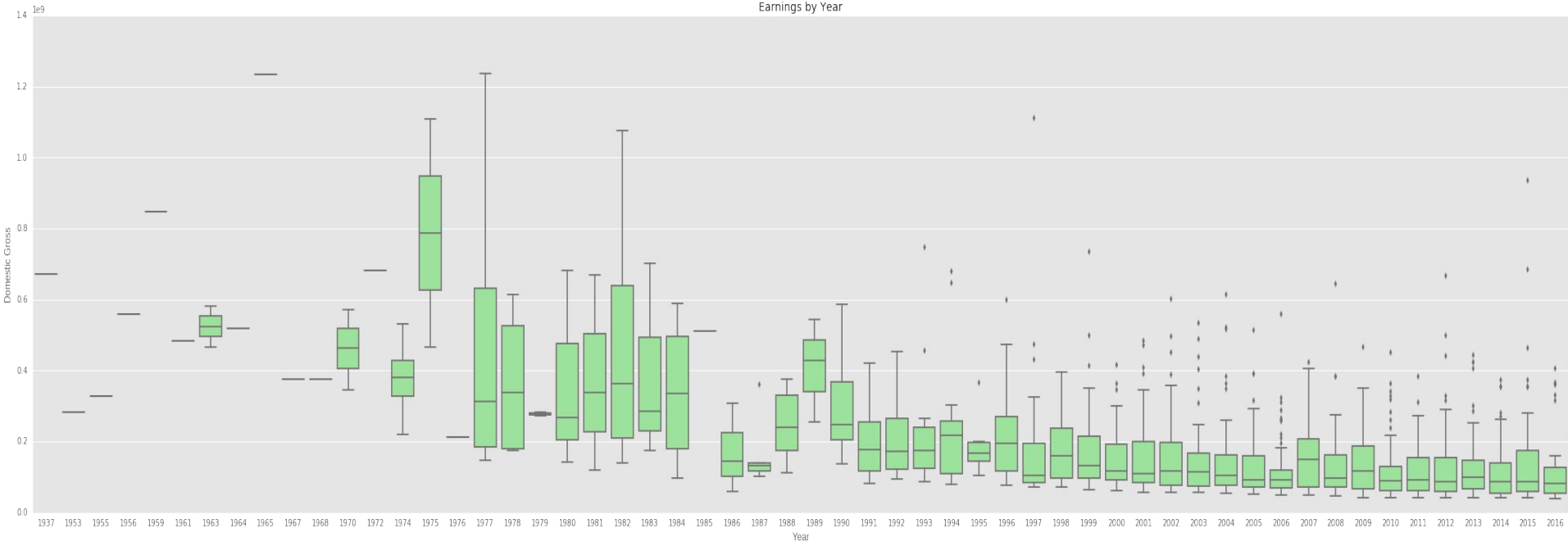


Box Office Mojo



Exploration

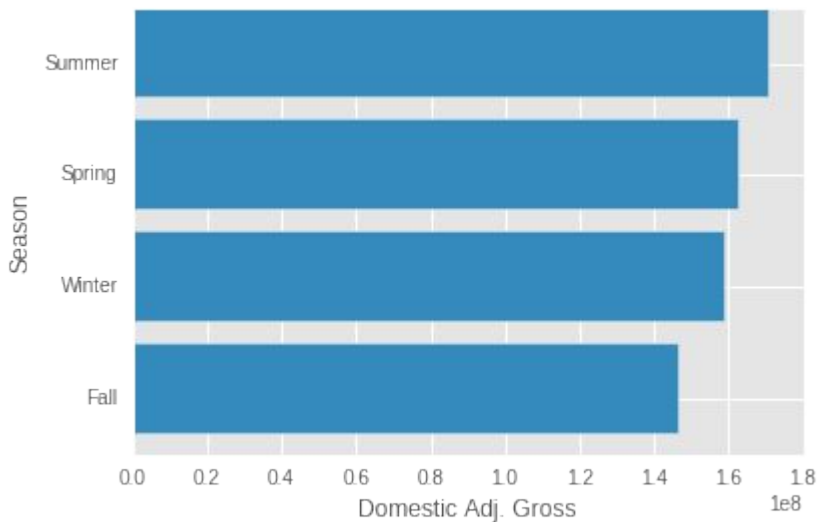
Earnings by Year



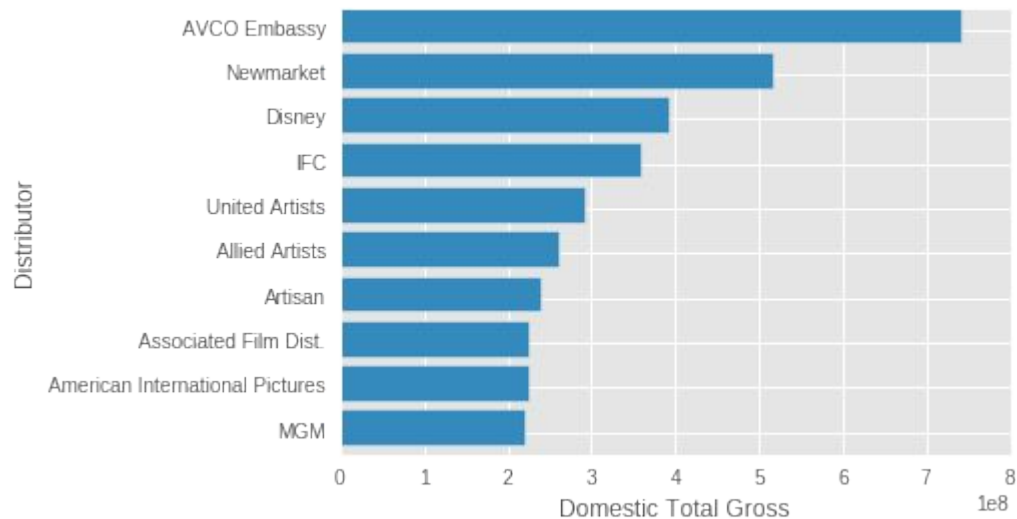
Exploration

— — —

Mean Gross by Season



Top 10 Distributors by Adj. Mean Gross



Preparing

- Standardized all Variables



- Used 10x K-fold for cross validation of models

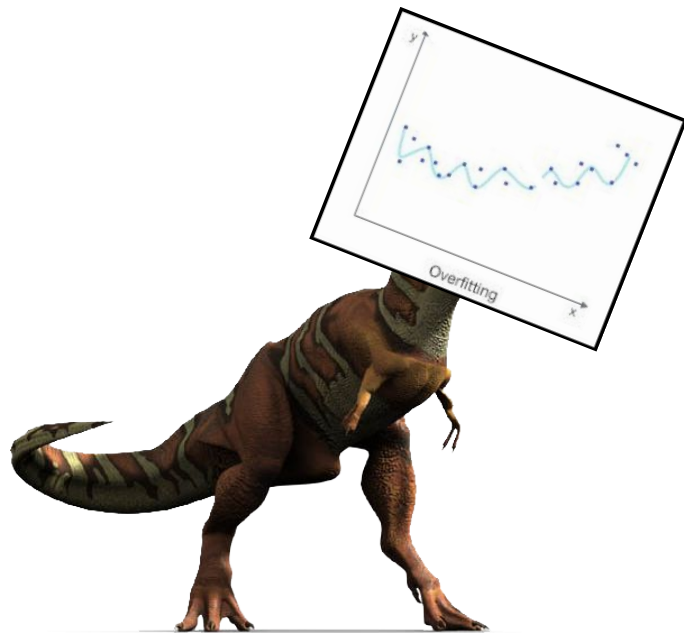
Mega-Regression

Everything including the kitchen sink

$$R^2 = .482$$

But...

$$\text{Cross validation score} = -9.02765505069e+27$$

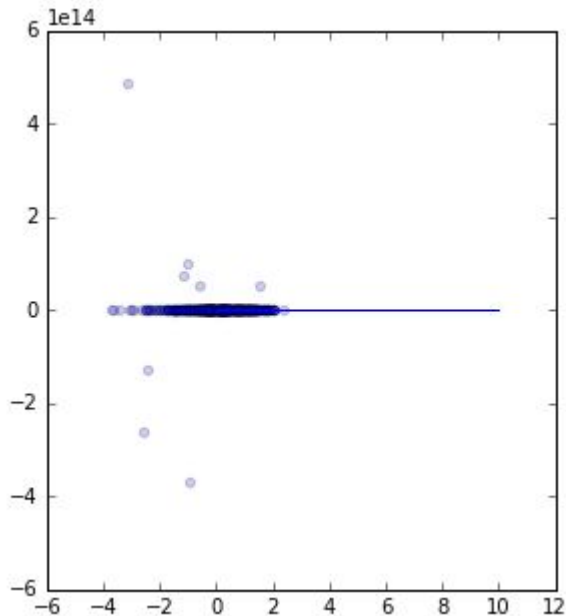


A more considered approach

- Selected most predictive features from the mega-regression

$$-R^2 = .40$$

-Graphing train vs test looks good!



Observations

- Almost all distributors did not predict earnings once genre was controlled for.
- R rating predicted lower earnings, PG-13 predicted slightly higher earnings
- Some genres (e.g. Action/Adventure) almost always predicted high earnings while others had a lot more variability (e.g. Horror).

Luther is DONE!

Takeaways:

- Never assume you're close
- Try everything
- ...but know what you're trying and why

