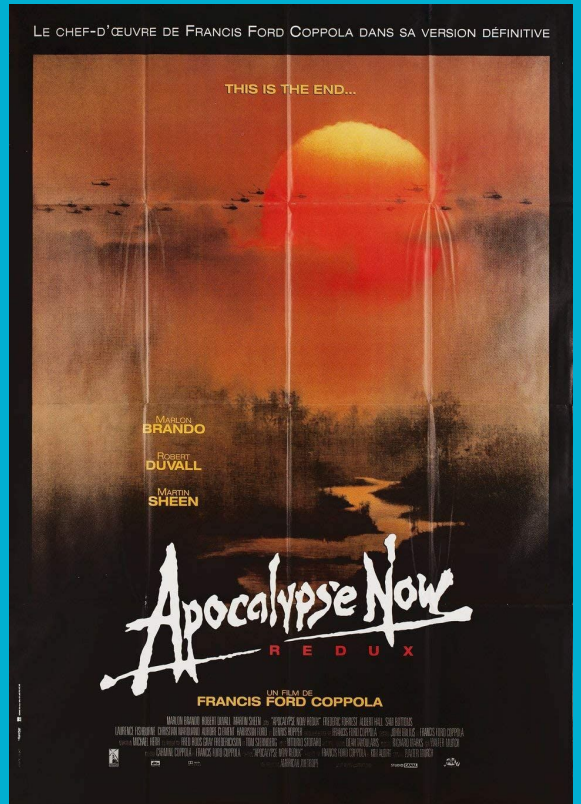
IMBD:
Gaming the Rating System

—

Diego Wahl


PALME D'OR
FESTIVAL DE CANNES
"ONE OF THE BEST FILMS OF THE DECADE."
AWARDS CIRCUIT
PARASITE
A FILM BY BONG JOON HO
ACT LIKE YOU OWN THE PLACE
SONG KANG HO LEE SUN KYUN CHO YEO JEONG CHOI WOO SHIK PARK SO DAM


WALT DISNEY STUDIOS PRESENTS
A STUDIO GHIBLI FILM
MIYAZAKI'S
SPIRITED AWAY


LE CHEF-D'ŒUVRE DE FRANCIS FORD COPPOLA DANS SA VERSION DÉFINITIVE
THIS IS THE END...
MARLON BRANDO
ROBERT DUVALL
MARTIN SHEEN
Apocalypse Now
REDUX
UN FILM DE
FRANCIS FORD COPPOLA

Scraping from IMDb

(and Box Office Mojo)

- Top 1000 ***movies*** by rating
  - No TV- shows were included in this analysis
- Features composed of information known prior to the release of a movie
  - Hence information such as Box Office Open not included
- Scraping can be a bit inconsistent, as not all data are available in all cases
  - MPAA Rating in particular for international and older movies
- Managed not to get myself IP-blocked

Feature Selection and Features Selected

Features should be know ahead of release

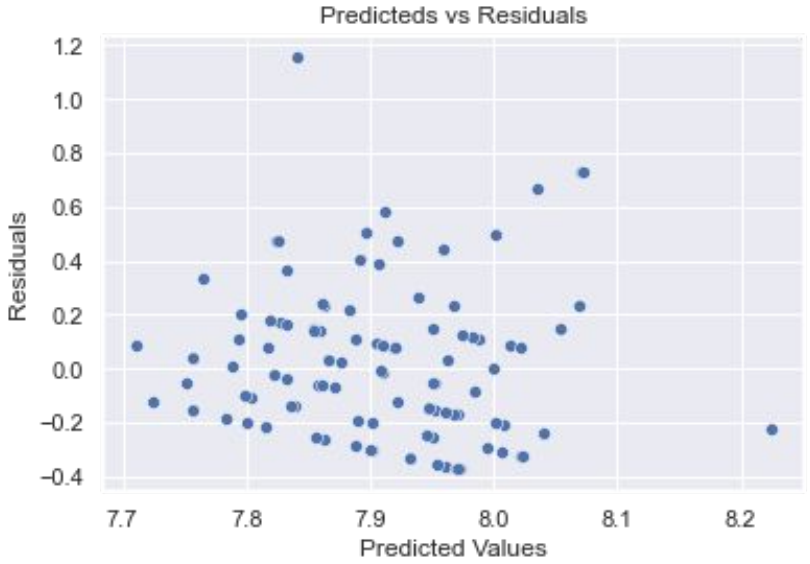
Features were selected on the basis of them being data available prior to the release of a movie, such as Budget, Runtime, MPAA rating, etc.

Lots of dummies

There's a fair amount of categorical data associated with movies, such as Genres and MPAA ratings.
In total- IMDB categorizes movies into 21, non-mutually exclusive genres.

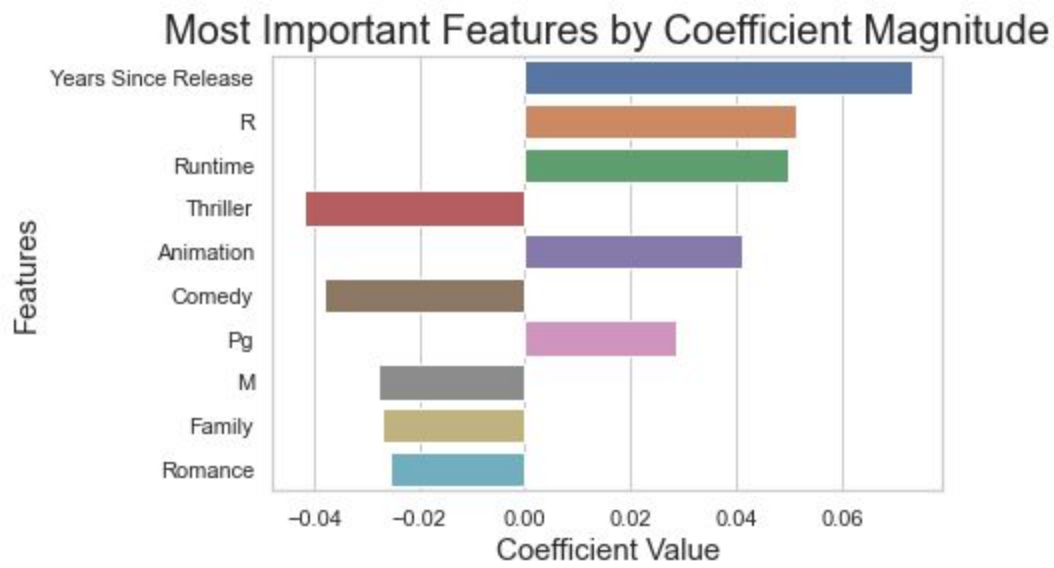
Room for Growth

One set of features considered, but not included, were the *people* associated with particular films. I would venture a guess and say that likely specific actors and directors are good predictors of an IMDB score.



Ridge Model

$R^2 = 0.049$
MAE = 0.221



Ridge Model Coefficients

- Top 10 selected by absolute value of coefficient
- Thrillers and Comedies to be avoided!

Lasso Model

Lambda = 0.023

(Lambda obtained via simple validation)
R^2 = 0.0425
MAE = 0.225