PREDICTING WORLDWIDE MOVIE GROSS FROM "THE-NUMBERS" DATA

Sarick Shah

MOTIVATION

- See if I could accurately predict worldwide gross
- Movies generally screen in America first
- Domestic Factors:
  - Percent change in gross each weekend
  - Gross per theater each weekend
  - Number of theaters the movie is shown in each weekend

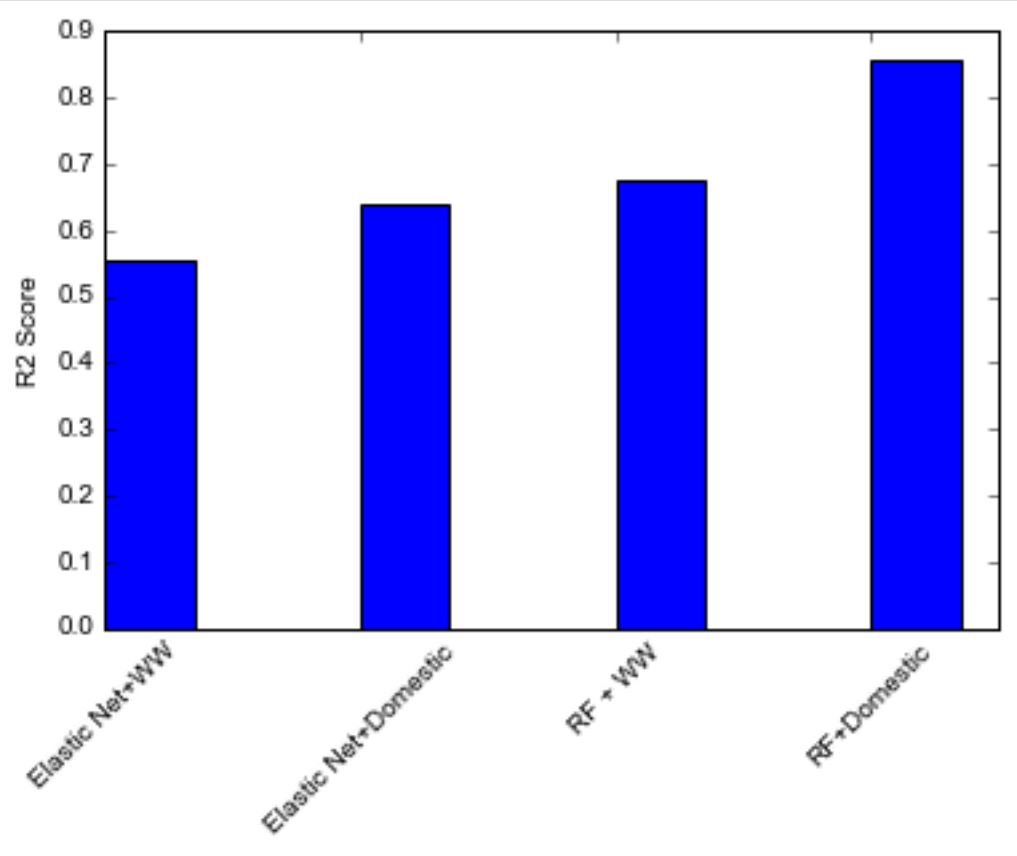
DATA

- ~3900 Movies
- ~500 cells contain averages
- All box office data from the US only
- Only looked up till weekend 3 due to lack of data
- 1 genre rating per movie

Features
Budget
%Δ Theater Wknd 2-3
Gross/Theater Wknd 1-3
Theaters Wknd 1-3
Critic Rating
Audience Rating
Genre
MPAA Rating

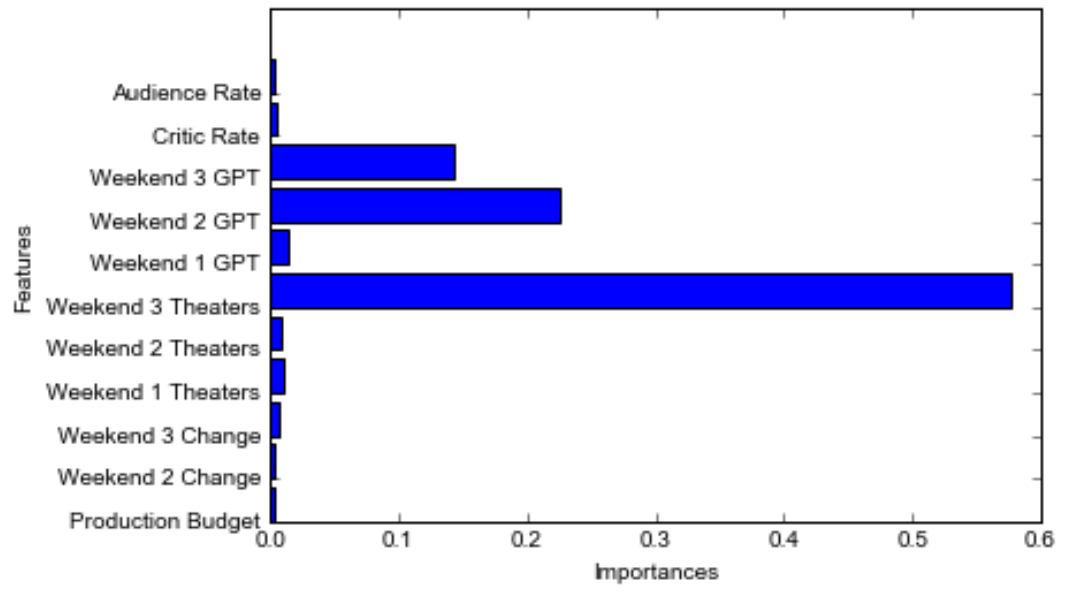
MODELS AND PERFORMANCE

- Elastic Net + WW
  - $R^2 = 0.555$
- Elastic Net + Domestic
  - $R^2 = 0.641$
- Random Forest Regressor + WW
  - $R^2 = 0.675$
- Random Forest Regressor + Domestic
  - $R^2 = 0.856$

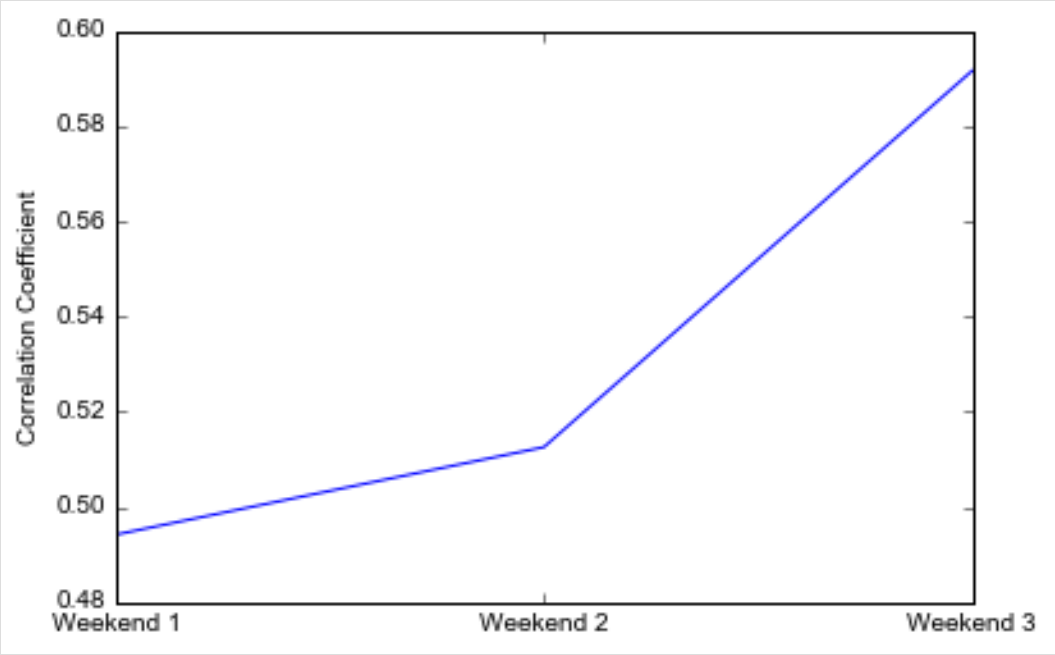


FEATURE IMPORTANCES

- Why is Weekend 3 Theaters so important?
- Momentum, Ratings, don't seem to matter much in random forest feature selection

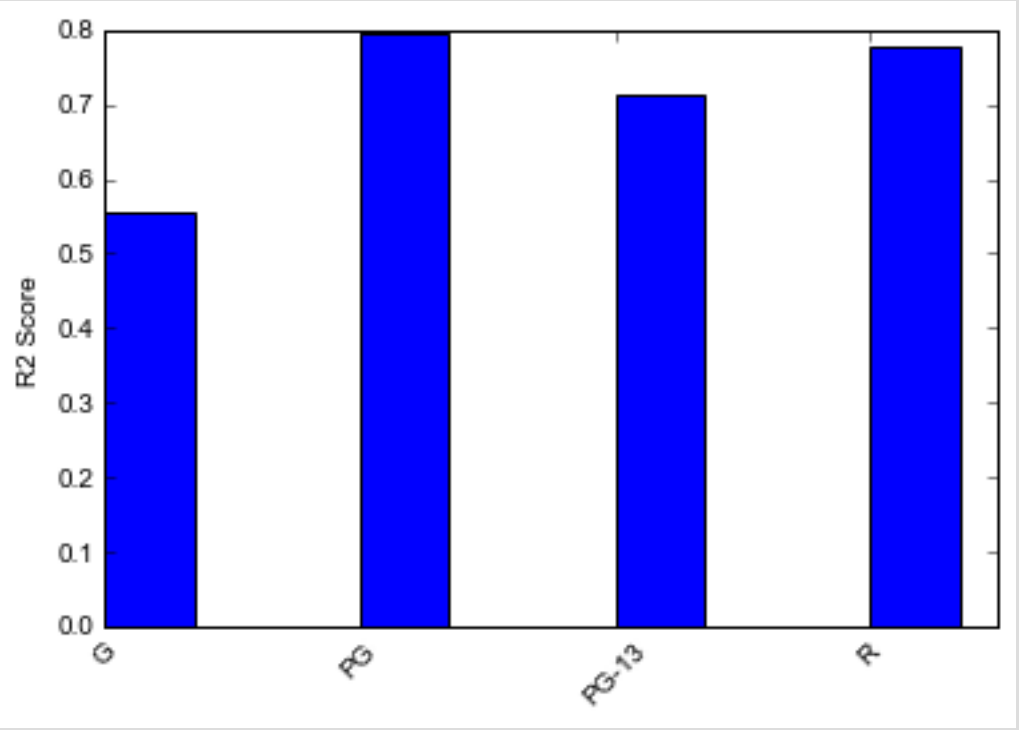


CLOSER LOOK AT THEATER NUMBER



IS MPAA RATING A BETTER MEASURE FOR WORLDWIDE PERFORMANCE?

- G: $R^2 = 0.513$
- PG: $R^2 = 0.793$
- PG-13: $R^2 = 0.712$
- R: $R^2 = .777$



* Low score for G is probably due to the fact that there were only 96 movies in the G rating.

FUTURE

- Incorporate seasonality
- Get data for major countries
- Number of A-list actors
- Make plots prettier

SUPPLEMENT

- Random forest doesn't require data normalization
- One feature is never compared in magnitude to other features
- Elastic Net:
  - L1 = .75
  - L2 = .25
  - Alpha = .1
- Random Forest:
  - 100 Estimators
  - Max Depth = 9
  - Features = 7