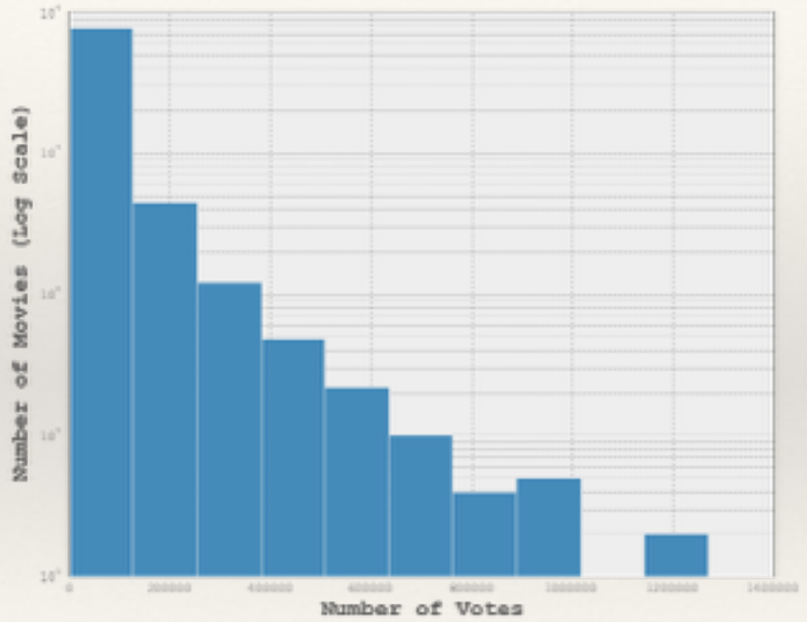
Movie Studio Presentation

Predicting Movie Opinions of the Crowd

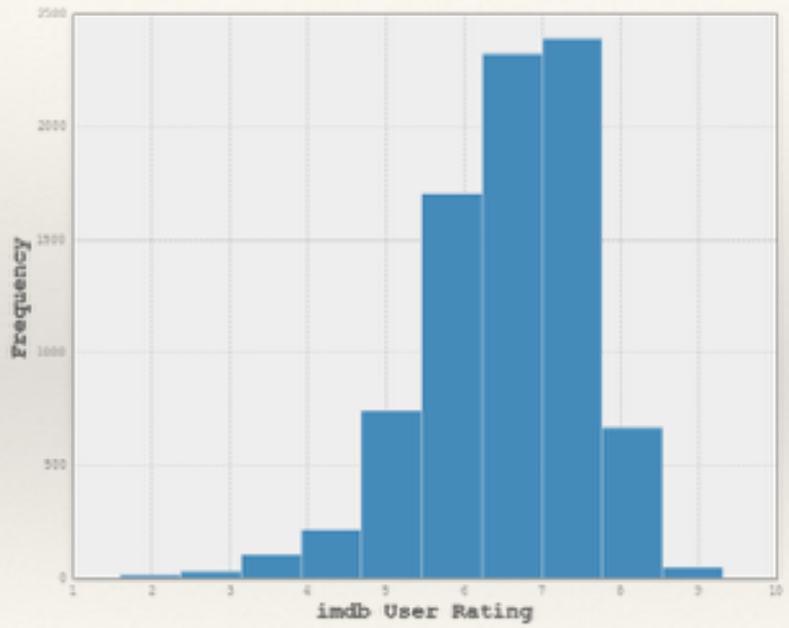
Gabriel Gluck

IMDb Votes as Proxy for public opinion



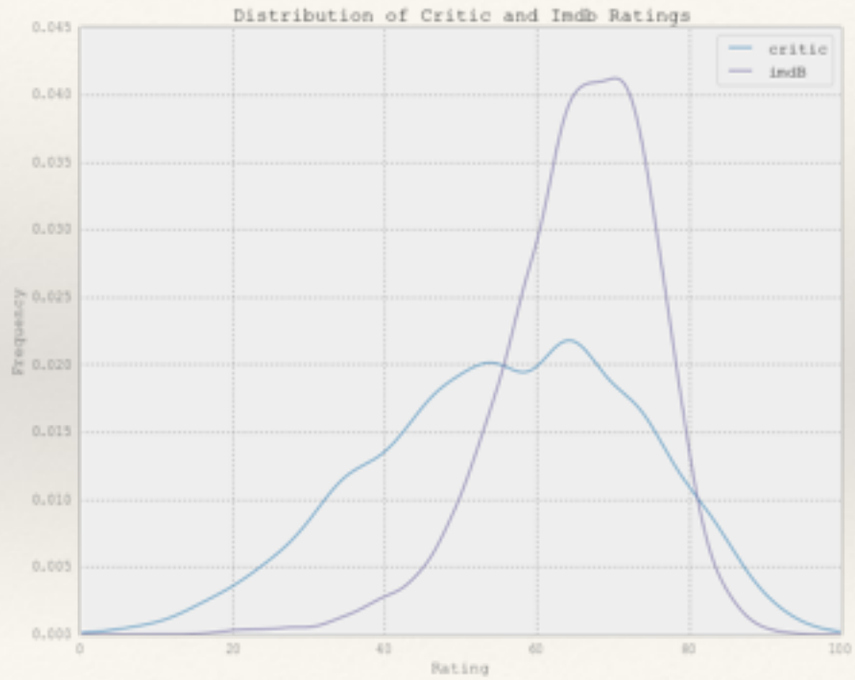
IMDb is biggest source of data on people's movie opinions. The average amount of reviews per movie is 36,000.

Distribution of IMDb user ratings



Although ratings are not symmetrical, there is a fair amount of variation in people's opinions. Since this is not a random sample there are issues with selection bias. What types of people post their reviews on IMDb? Of the movies they've seen how do they choose which to post their reviews on?

Critic Reviews versus IMDb Ratings



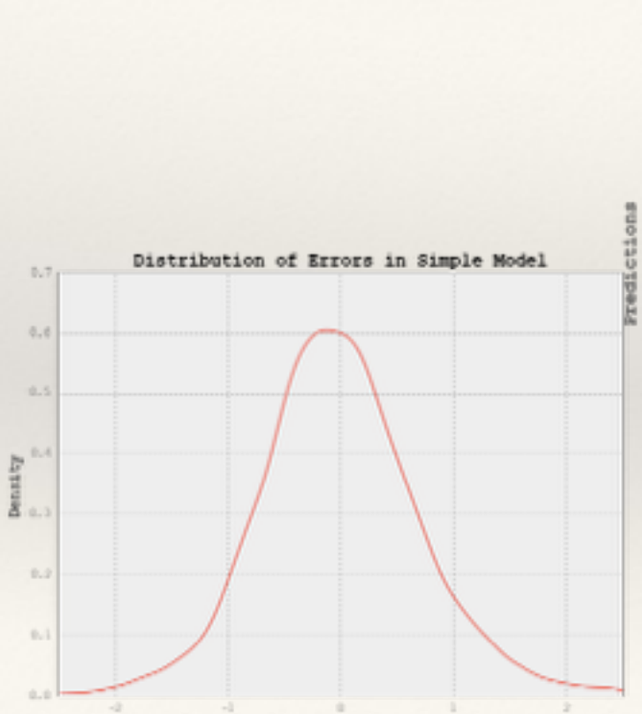
Critic reviews were obtained from metacritic.com. In comparison to overall IMDb ratings they are more negative and have a larger variation.

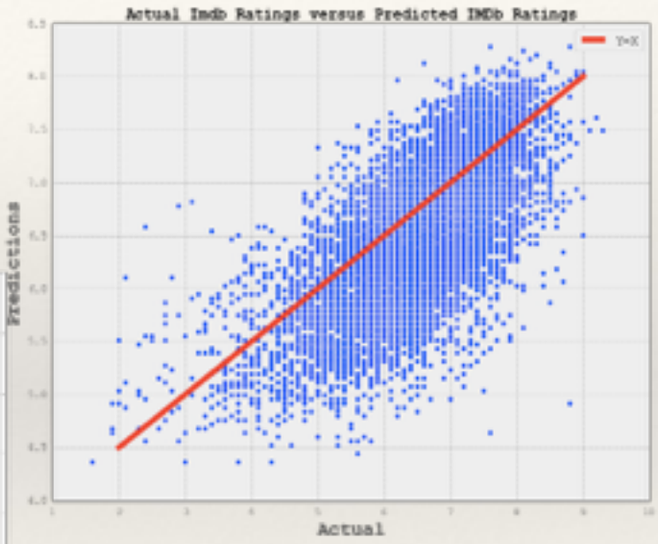
Simple model (just critic rating)

Linear Regression: IMDb Rating= 4.32 +.04* Critic Rating

Both the coefficients for the intercept and for critic rating are statistically significant. R-Squared for the simple model is .47.

Actual opinions versus predictions





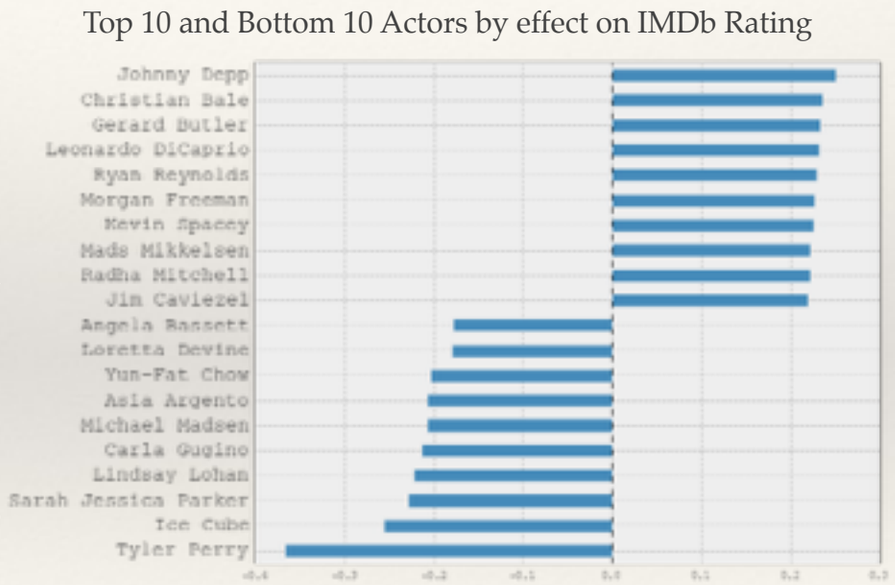
The distribution of errors for the simple model is not symmetrical. The actual IMDb Ratings are likely to be more negative than the predictions.

Building a more complex model

Added many more features:
The 500 most frequently appearing actors in data set.
The 50 most frequently appearing directors in data set
All 25 Genres
Switched method to ridge regression because of the addition of 575 more features

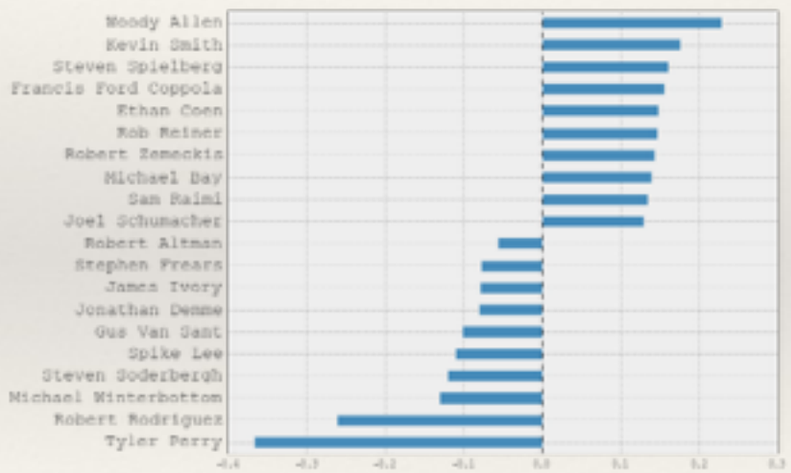
Most impactful actors

Interpretation:
Holding critic rating, director, genre, and all other actors constant, Johnny Depp provides a movie that he appears in a boost of .25 IMDb ratings points. Conversely, holding all else constant Tyler Perry subtracts .36 IMDb rating points for a movie that he appears in.



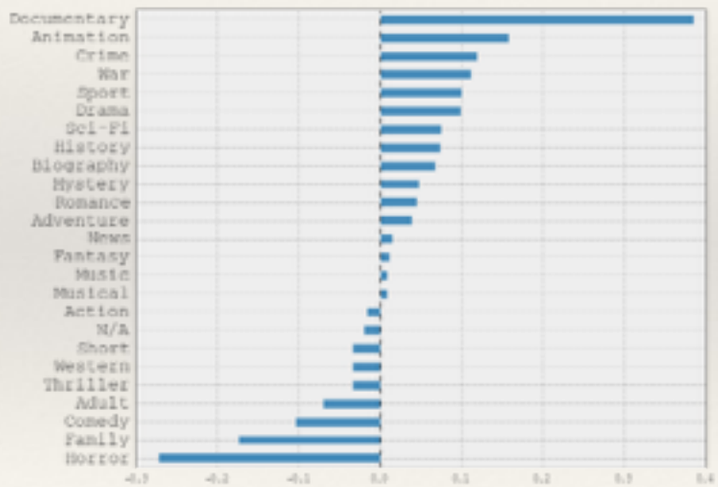
Most impactful directors

Top 10 and Bottom 10 Directors by their effect on IMDb Rating



Most impactful genres

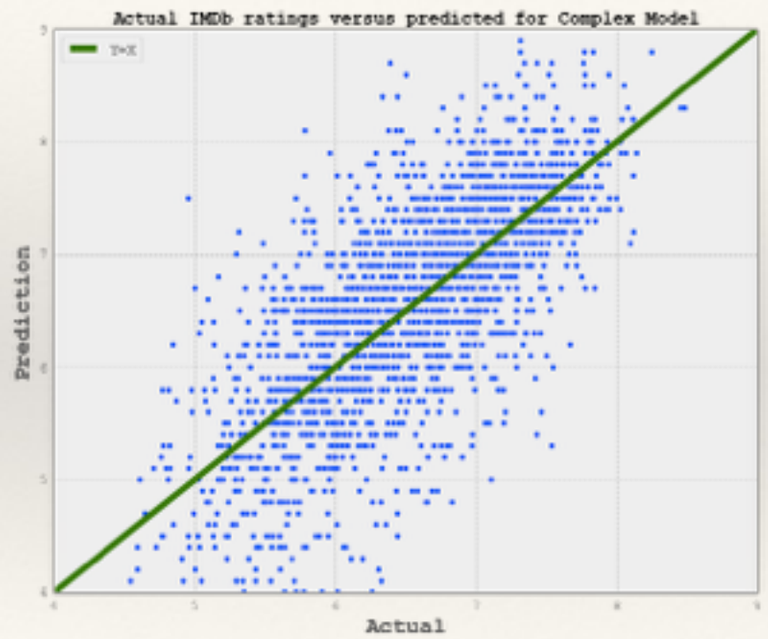
Effect of Genre on IMDb Rating



Distribution of coefficients

The distribution of coefficients is centered at zero.

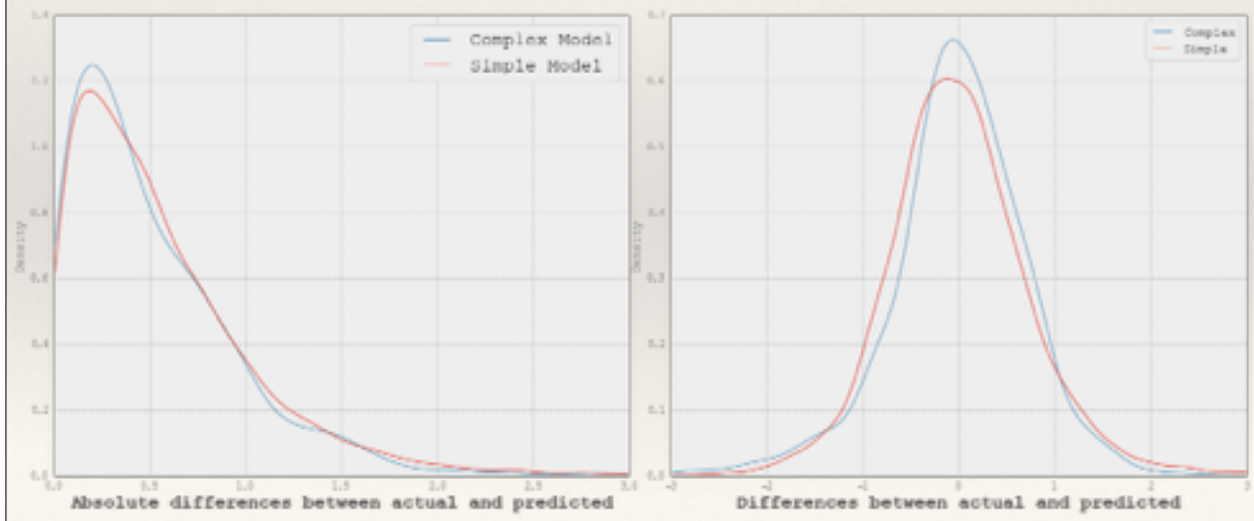
Comparing predictions versus actual values



Errors are a lot more symmetrical. R-Squared has improved to .56.

Comparison of models

Complex Model does a better job of predicting IMDb ratings. In addition its errors are much more symmetrical.



Future work

Predicting peoples' movie opinions is an inherently difficult subject. How does one quantify a good script, or chemistry between the lead actors?

A future direction would be to look at interactions between features. Are there combinations of actors or actor/director pairs which people rate highly?

A more in in-depth look at the type of people who post on IMDB or how they rate movies, would give a better model of how these reviews are generated. Otherwise looking for a better proxy of people's opinions than IMDb ratings or combining with other databases to give me a more representative sample of peoples' opinions.