PREDICTING IMDB MOVIE REVIEW SENTIMENT

BY TRAVIS JAMES

MOVIE REVIEW SENTIMENT: WHY IS IT IMPORTANT?

- The general public's opinion on how enjoyable a film was
- Important for driving current and future revenue streams
- Advances actor's and director's reputation
- Forecasting general success of a film (franchise?)

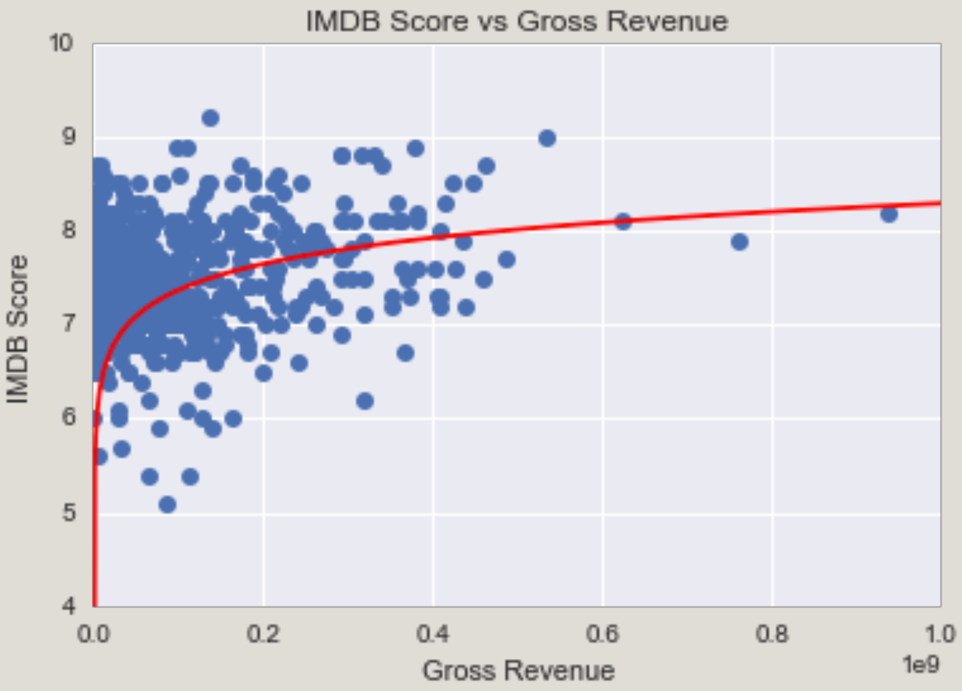
IMDB SCORE

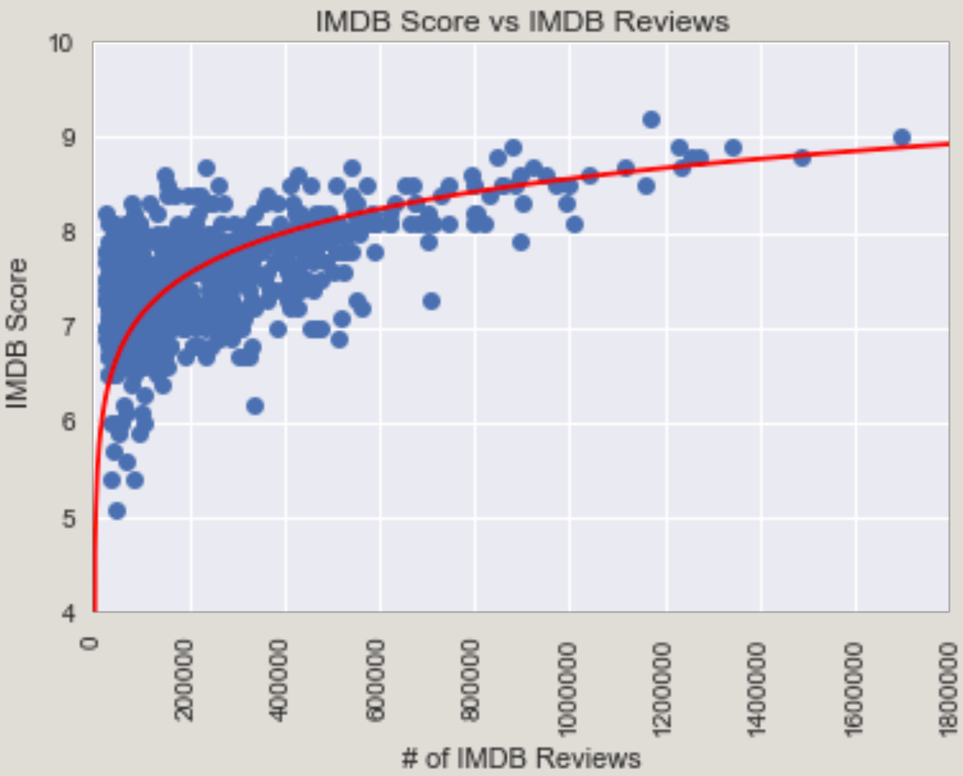
- Decided by IMDB registered users and critics alike
- Number of reviews for a single film range from 25,000 to over 1.6 million!
- Allows for a more unbiased public opinion on a film
- Good proxy for public sentiment
- Scraped data on 720 IMDB movie scores and relevant features

FEATURES

- Metacritic Score
- Budget and Revenue
- # of Oscars awarded to film
- Run Time
- Number of IMDB User Reviews for Film
- Categorical Controls:
  - Year, Month, Primary Genre, Secondary Genre, MPAA Rating

FEATURE ENGINEERING





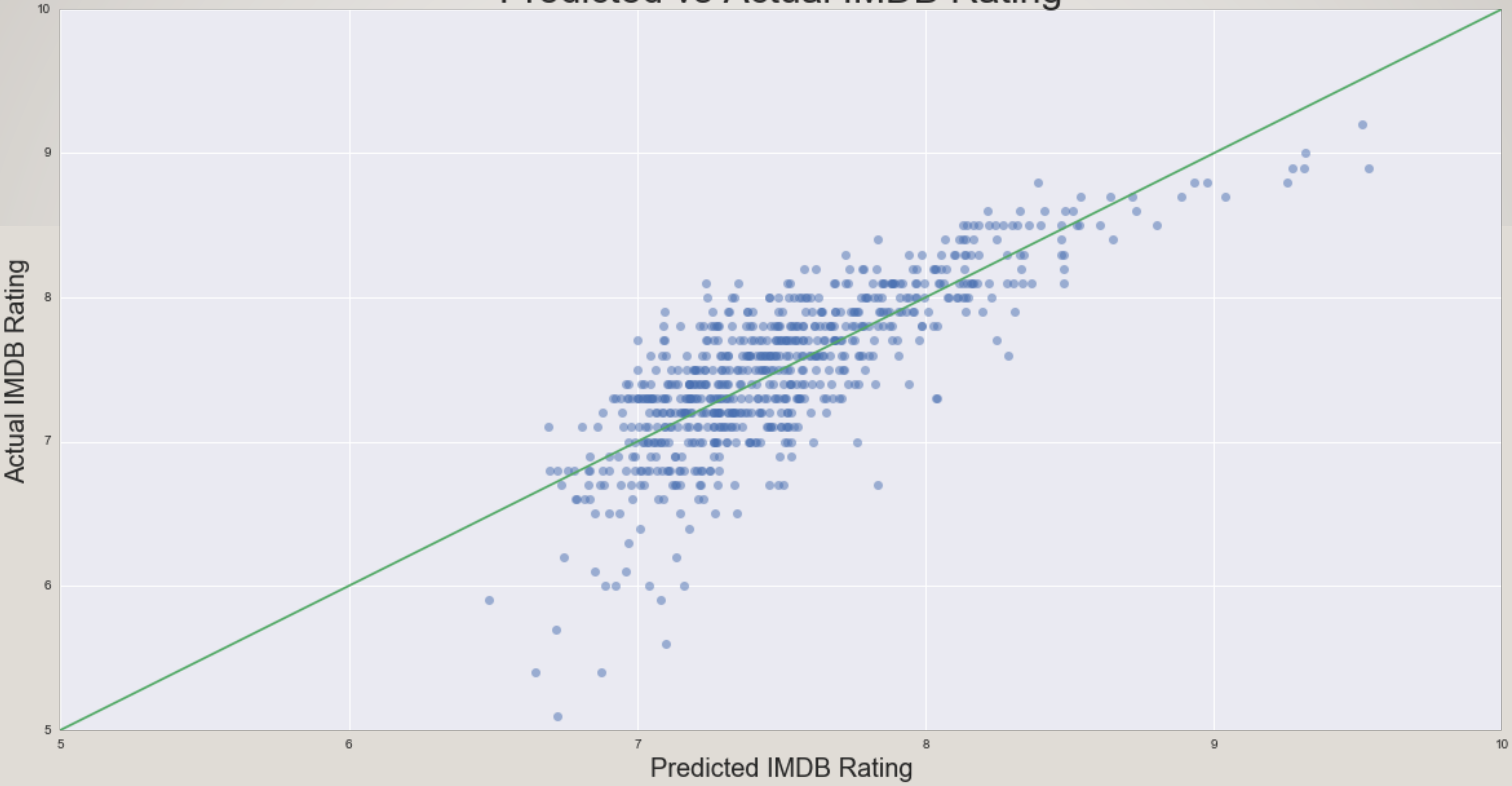
MODEL SELECTION

- 5-Fold Cross Validation on OLS
- Grid Search 5-Fold CV:
  - Ridge ($\alpha = 0.0001$)
  - Elastic Net ($\alpha = 0.0001$)
  - Lasso ($\alpha = 1$)
- Gradient Boosting
- Random Forest

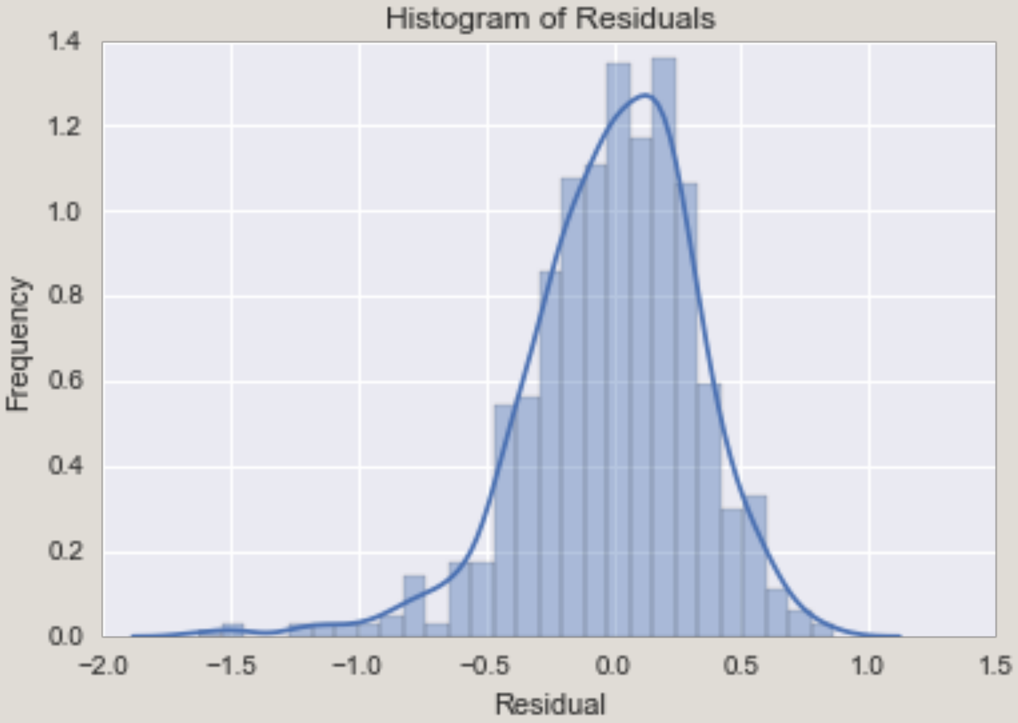
TAKEAWAYS

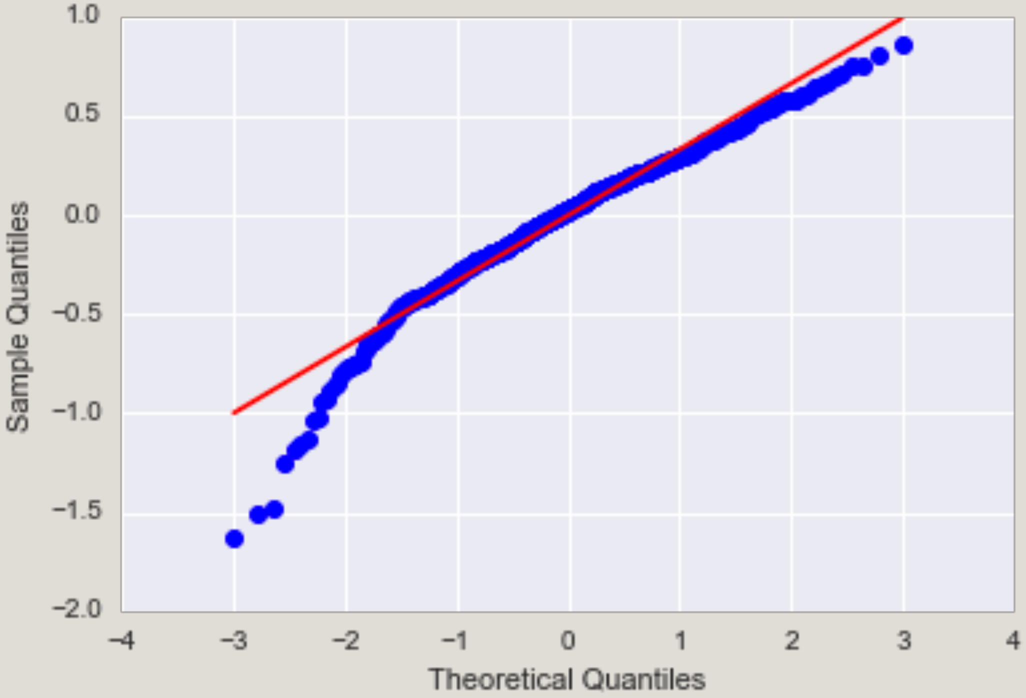
- Ridge regression performed best
  - α chosen from Grid Search 5-fold Cross Validation
- Test Set R-squared- 0.6879
- Run Time and Metascore positively influence public sentiment
- Oscar Wins and Budget negatively influence public sentiment

Predicted vs Actual IMDB Rating



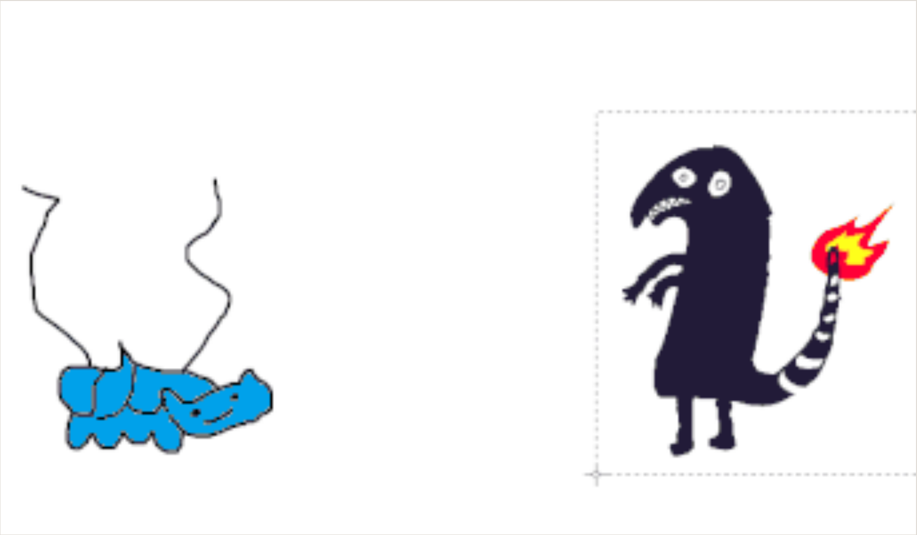
NORMALITY OF RESIDUAL ERRORS





LIMITATIONS

- Simultaneity
- Robust Standard Errors and Weighted Least Squares
- More Data!



MODEL SELECTION AND TEST PERFORMANCE

- 5-Fold Cross Validation on OLS - 0.4887 R-squared
- Grid Search 5-Fold CV:
  - Ridge (α = 0.0001) – **0.6878**
  - Elastic Net (α = 0.0001) – 0.6797
  - Lasso (α = 1) – 0.6802
- Gradient Boosting - 0.5561
- Random Forest - 0.5517