PREDICTING IMDB RATINGS

USING LINEAR REGRESSION

INTRODUCTION

- The goal of the project is to predict the IMDB movie ratings which is the Target
- Numeric Features included in the model are:
  - Votes,
  - Budget,
  - Box Office Opening Weekend,
  - Box Office domestic,
  - Box Office Gross,
  - Runtime
- Categorical features included are:
  - MPAA,
  - Genre,
  - Actor,
  - Writer,
  - Director,
  - Production Company

Details

Release date October 14, 1994 (United States)

Country of origin United States

Official sites Official Facebook · Warner Bros. (United States)

Language English

Also known as Rita Hayworth and Shawshank Redemption

Filming locations Mansfield Reformatory - 100 Reformatory Road, Mansfield, Ohio, USA

Production company Castle Rock Entertainment

See more company credits at IMDbPro

Box office

Budget
$25,000,000 (estimated)

Gross US & Canada
$28,699,976

Opening weekend US & Canada
$727,327 · Sep 25, 1994

Gross worldwide
$28,817,291

Taglines Fear can hold you prisoner. Hope can set you free.

Genre Drama

Motion Picture Rating (MPAA) Rated R for language and prison violence

Parents guide

WEB SCRAPING THE IMDB SITE

INITIAL OLS MODEL

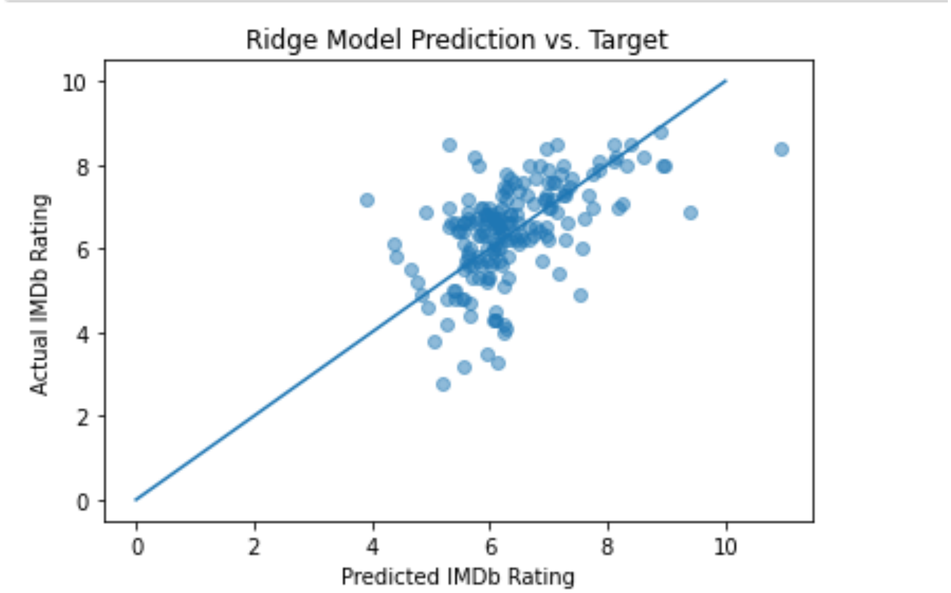
- Initial model with all numeric features: R-squared of 0.428
- Remove feature ‘box_office-gross’ which has high p-value, R-squared is still 0.428
- Trying to trim down the features based on VIF seems to reduce model accuracy
- Going with the features - runtime, budget, box office opening weekend, box office domestic and votes seems to give optimum R-squared value

BASELINE LINEAR MODEL

- Simple Linear Regression model with all the numeric features has R-squared = 0.4280
- Adding all the categorical features to the model brings up the R-squared to 0.983
- However the score on validation data drops to 0.122 which means the model is overfitting the training data

RIDGE REGRESSION MODEL

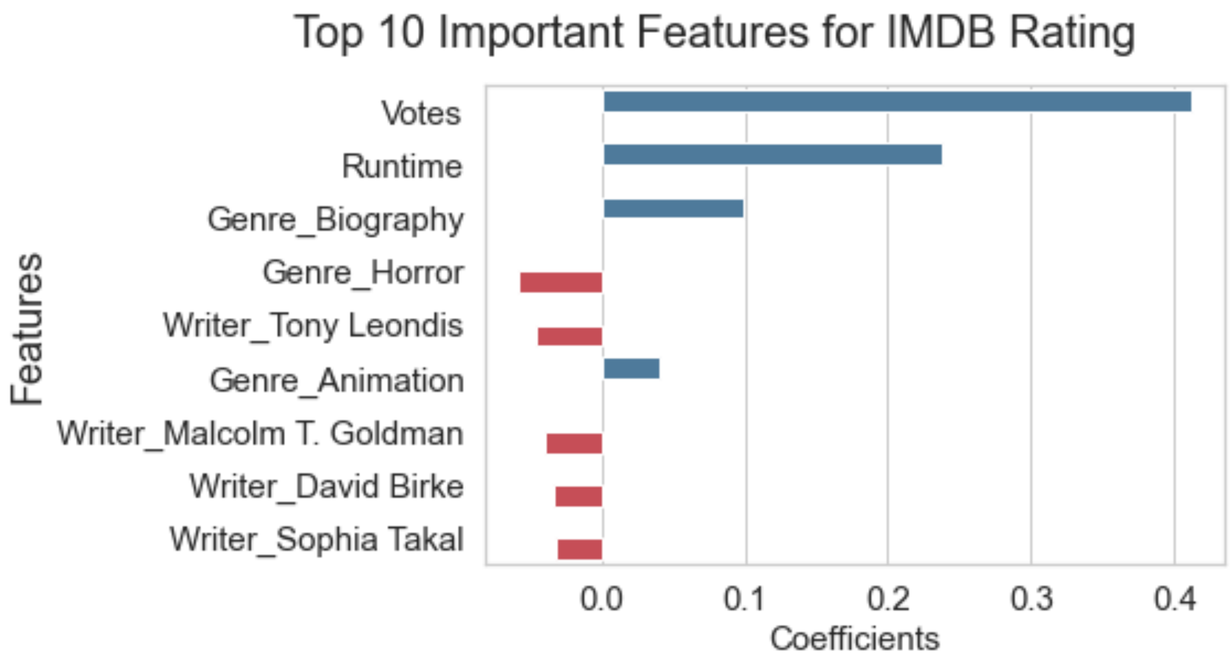
- Fitting a regression model and calculating the R-squared on test data = 0.2055 which is better than the Simple LR but there is still room for improvement



LASSO REGULARIZATION

- Tune the features by excluding movies with a very low number of votes
- Fit a Lasso regularization model on numeric and categorical features
- Validating the model on test data, gives an optimum value of:
  - R-squared: 0.6532
  - Mean Absolute Error: 0.4397

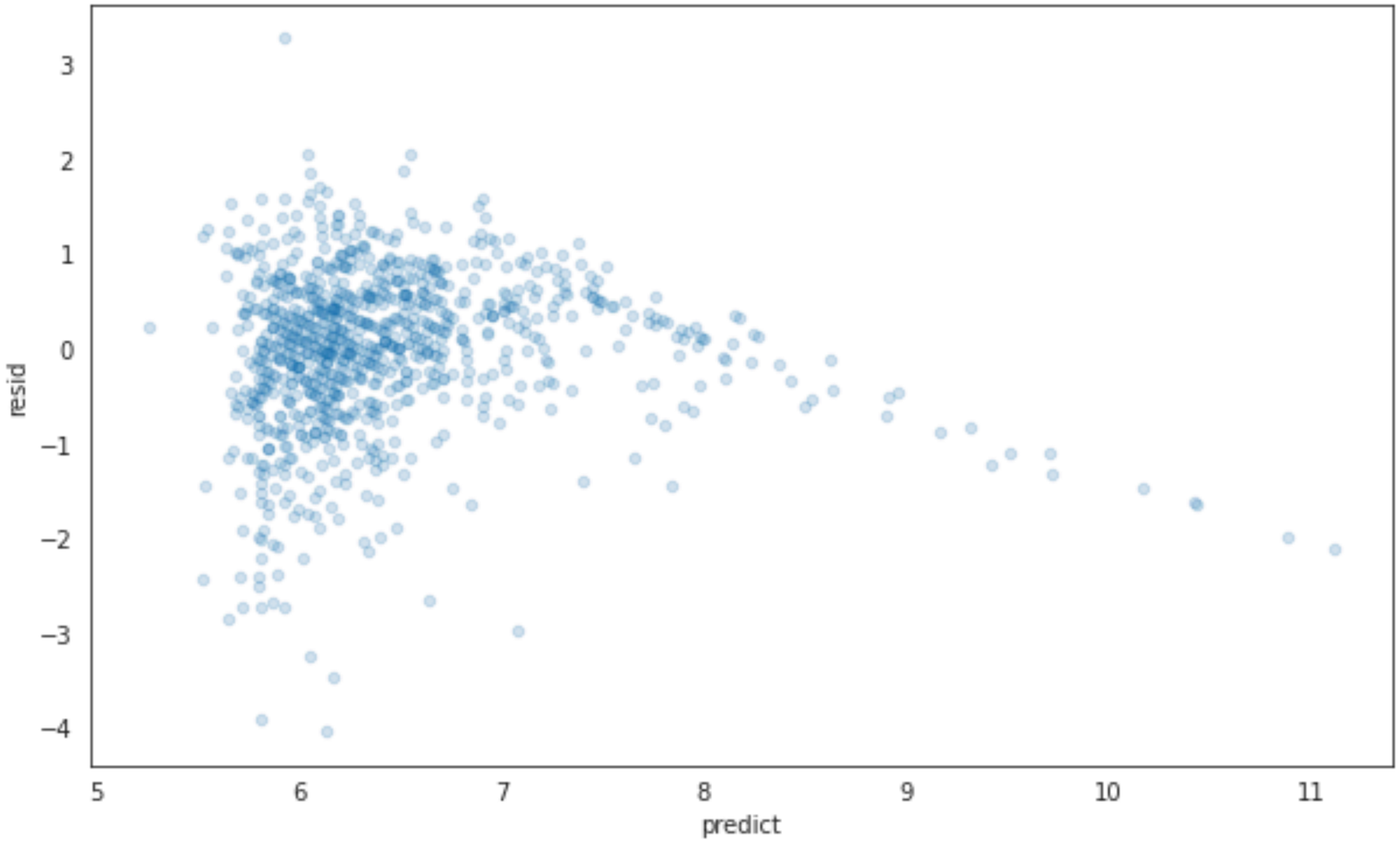
MODEL INTERPRETATION



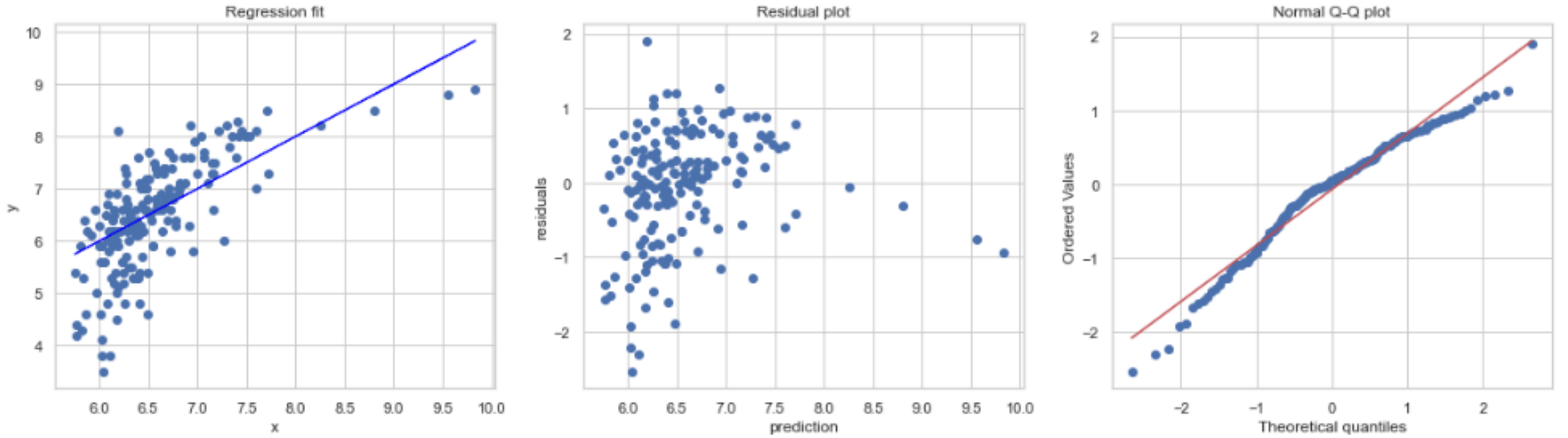
- Main features that impacts the IMDB rating for the chosen set of movies are:
  - Votes
  - Runtime
  - Genre: Biography and Animation
  - Horror movies seem to have a negative impact on the IMDB rating score

FUTURE WORK

- Include movies from more years to expand the scope of values
- Add data from websites other than IMDB



Appendix: Residual Plot of the initial OLS model



Appendix - Diagnostic plots for the final model