PREDICTING IMDB RATINGS

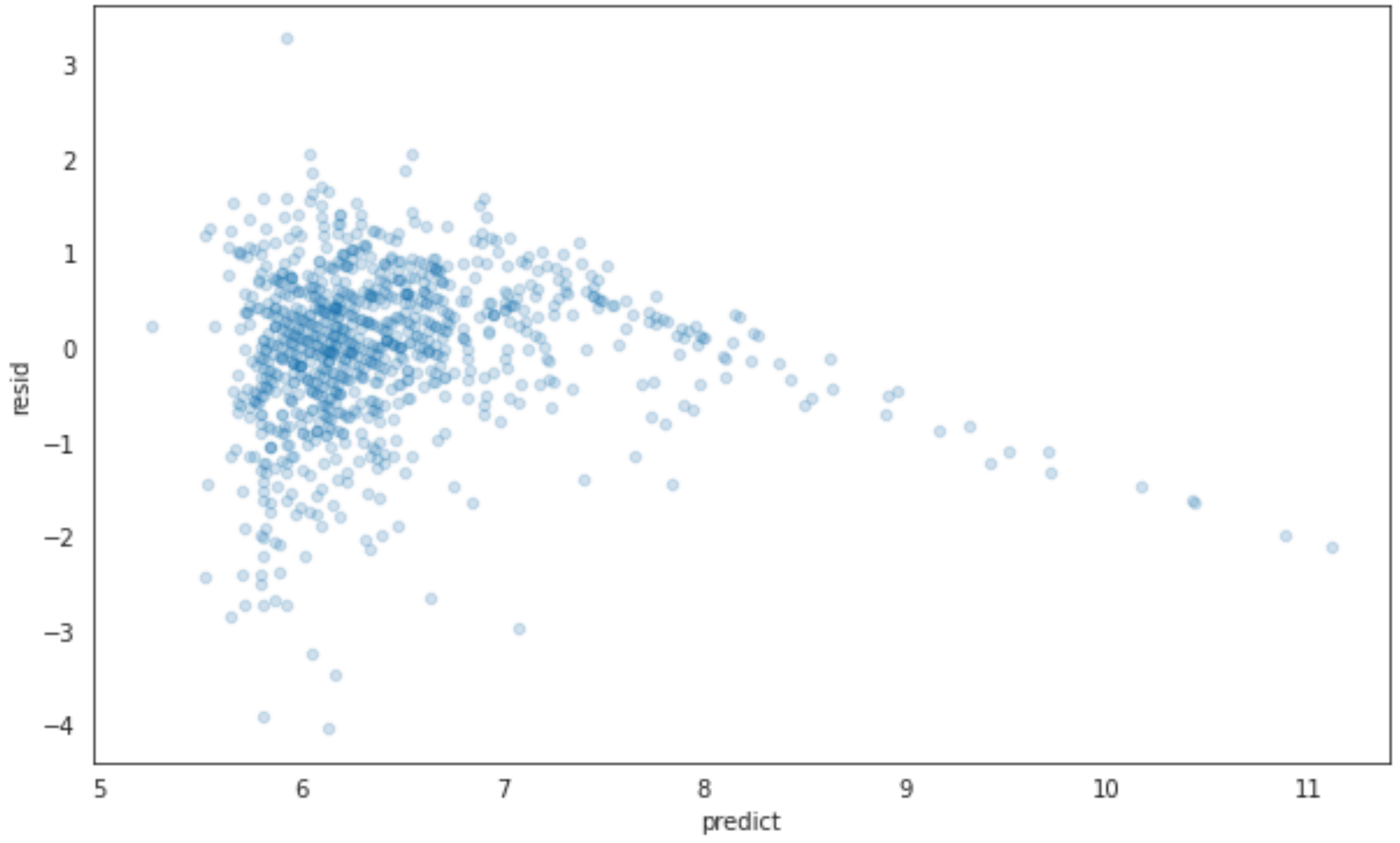
USING LINEAR REGRESSION

WEB SCRAPING FOR THE MOVIE DATA

- Scraped more than 1000 movies from the IMDB site using BeautifulSoup
- 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

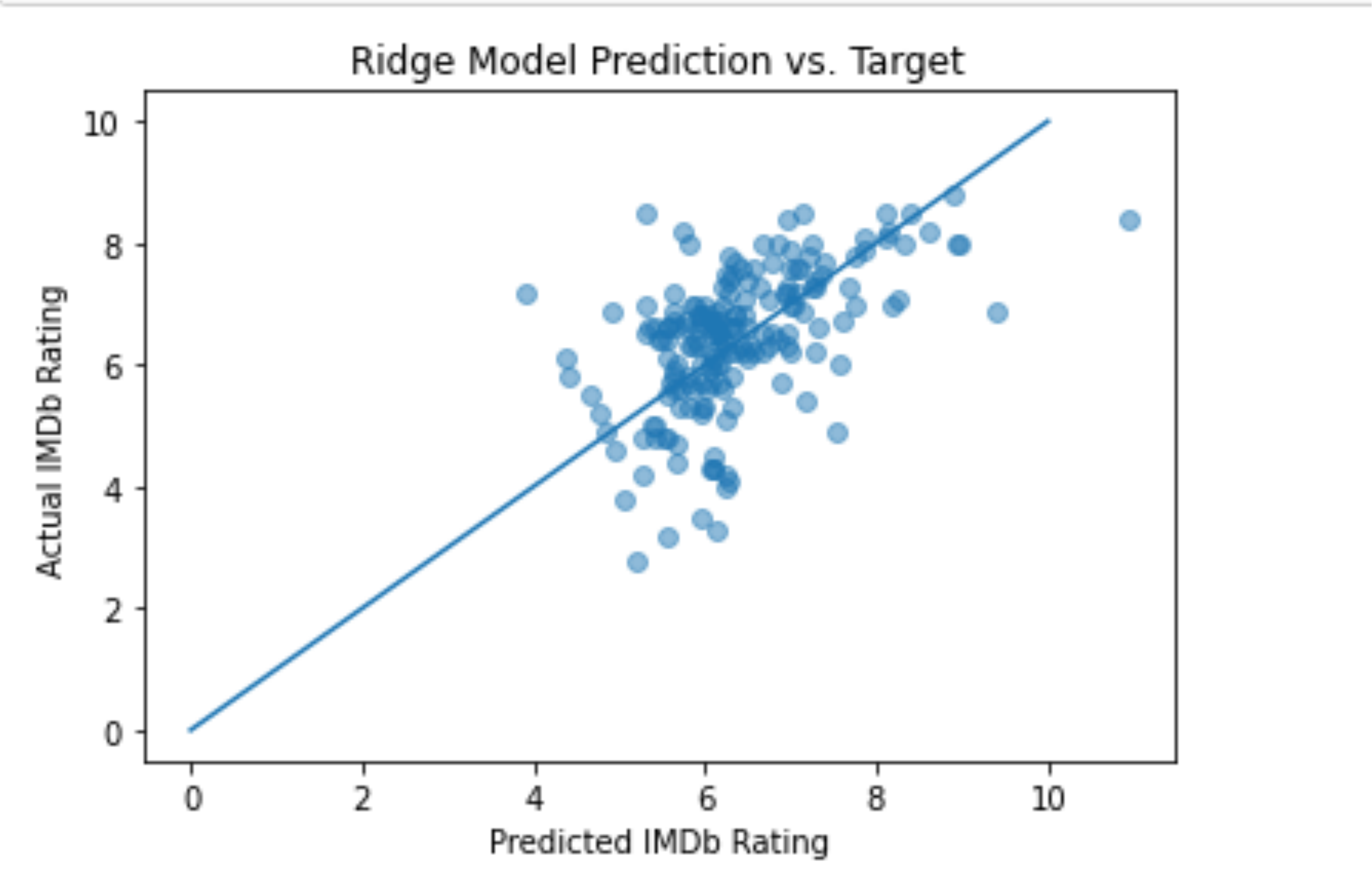
OLS MODEL

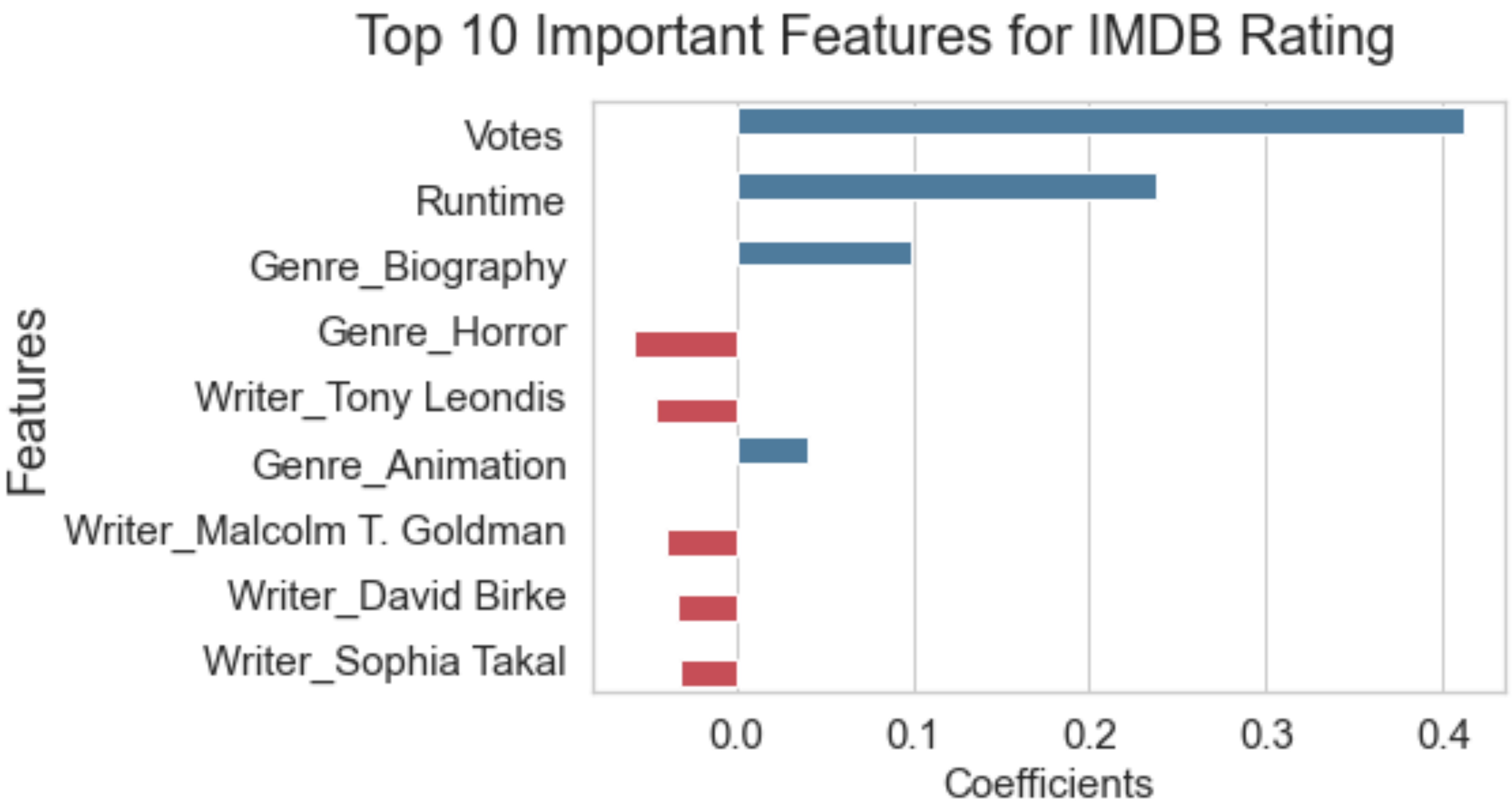
- Initial model with all numeric features: R-squared of 0.428
- Remove feature ‘box_office-gross’ which has high p-value, R-squared is still 0.428
- Trying to trim down the features based on VIF seems to reduce model accuracy
- Going with the features - runtime, budget, box office opening weekend, box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



LINEAR, RIDGE REGRESSION MODELS

- Simple Linear Regression model with all the numeric and categorical features seems to be overfitting
- Ridge Regression model R-squared seems to be better



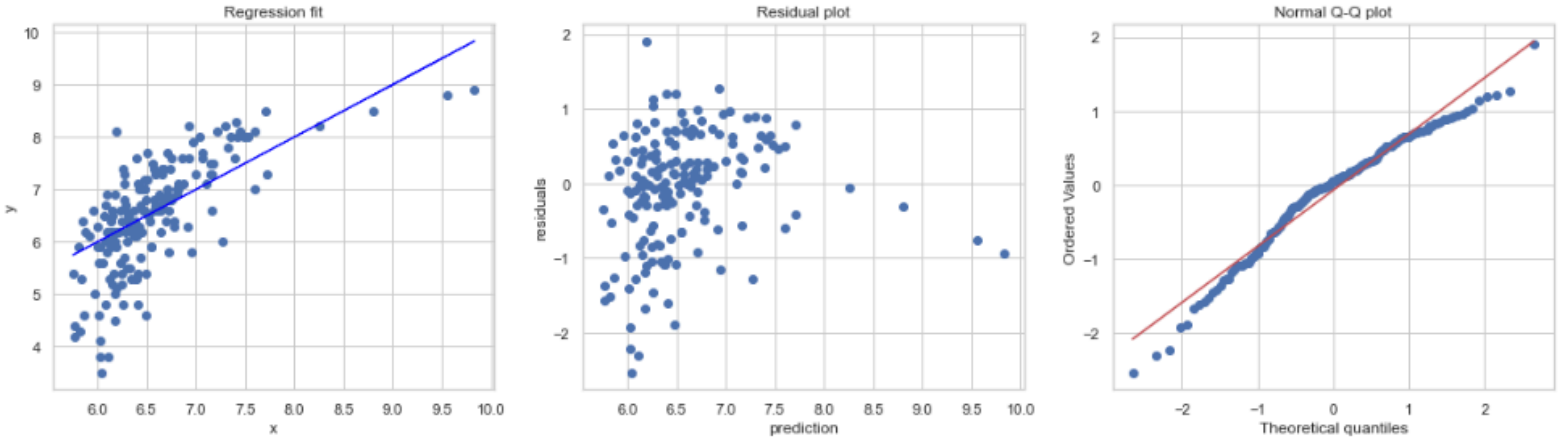


LASSO MODEL

- Lasso model seems to give the best R-squared value - 0.6532 and MAE = 0.4397

CONCLUSION

- 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



Appendix - Diagnostic plots for the final model