



# Predicting IMDB Rating on Movies

Connie Xiao



# Introduction



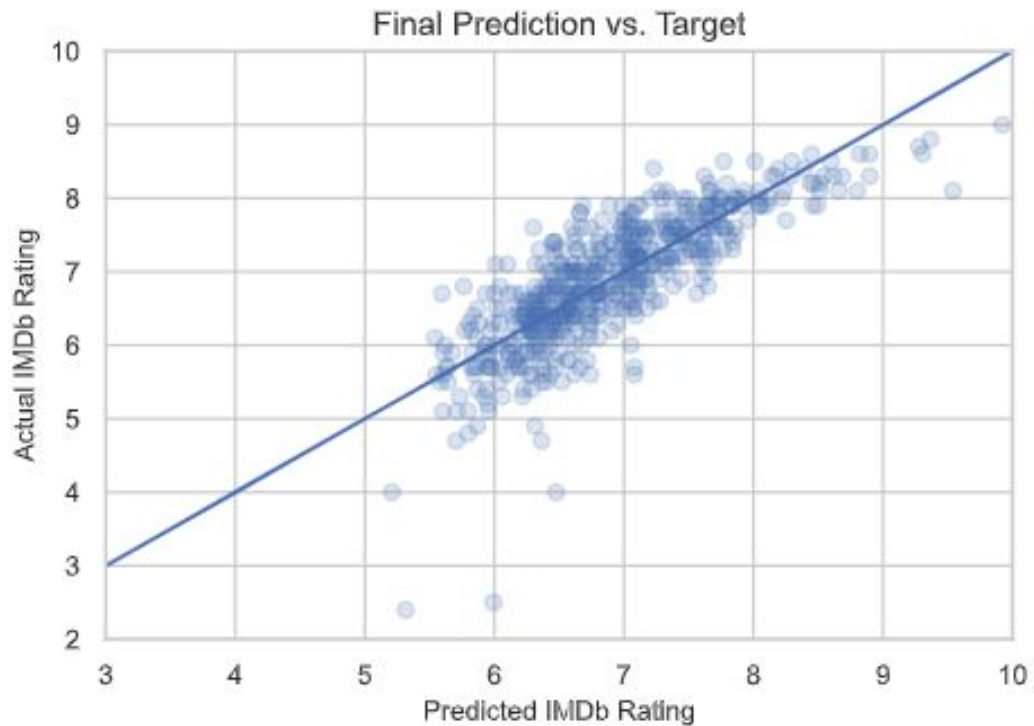
- IMDB is the largest movie database
- Goal of this project is to predict IMDB Ratings
- Visualize relationship between independent variables with dependent variable
  - Features include director, release year, metascore, MPAA, gross earnings, votes, genre
- Determine the best model to that accurately depicts IMDB ratings
- To provide film production companies an insight on what features to include in their production

# Methodology

- Web scraped data off IMDB.com using BeautifulSoup
  - Data cleaned/ simple EDA
- Added on features one at a time
  - Numerical
  - Categorical
- Modeling
  - Linear Regression
  - Polynomial Regression
  - Regularization (Ridge)

# Ridge Regression Model

Ridge R-squared: 0.60955  
Ridge MAE 0.38465



## Best Features

### Age

Movies between the ages 10-30  
tend to do better

### Metascore

Different rating system for movies

### Votes

More popular = better ratings?

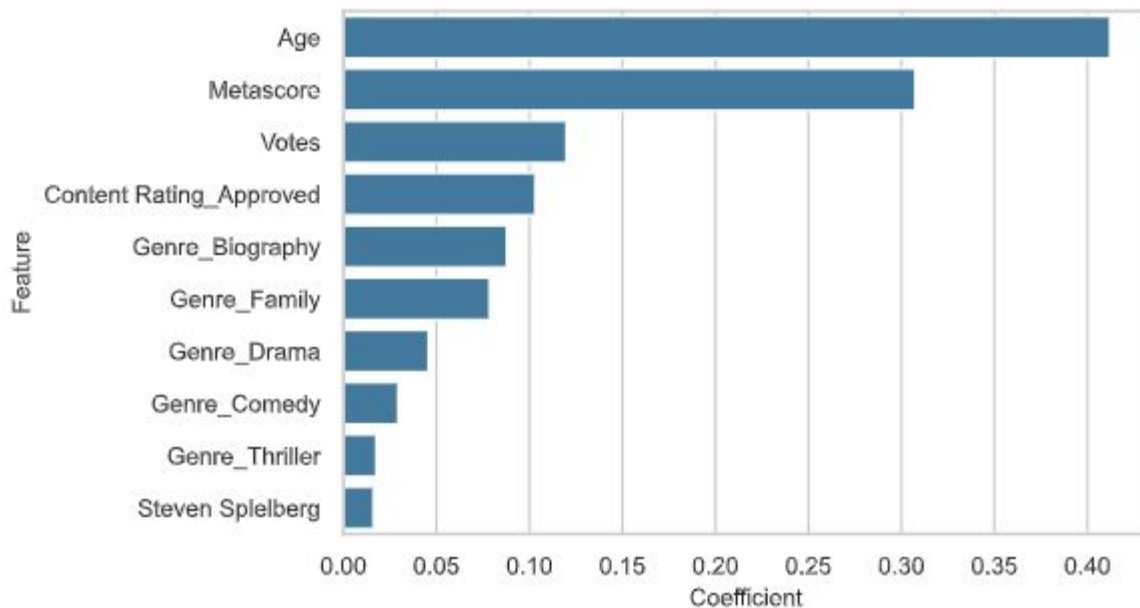
### Genres

Biography, Family, Drama

### Director

Movies directed by Steven  
Spielberg

Top 10 Features for IMDb Rating



# Conclusions/Results

- Ridge Regression gave me the best model out of Linear and Polynomial
- Linear and Polynomial Regression made my model overfit when testing on validation set
- Residuals showed that my model tend to over predict than under predict

# Future Work

- Determine if the profit made on release day can depict how well a movie's rating is
  - Do more popular films get better rating?
- Look into international films
  - Do movies produced in the US have better ratings than movies produced elsewhere?
- Extra features
  - Awards? Nominations?

# Appendix

