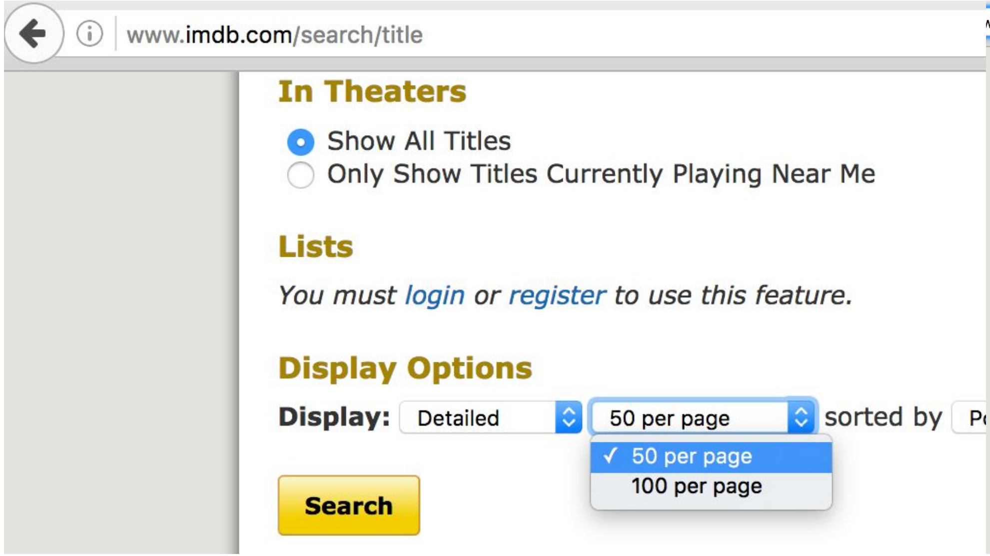
Predicting IMDb User Rating

Kevin Du

Objective

- Predict the IMDb user rating for movies
- User rating and reviews affect ticket sales, and thus revenue
- Netflix competition: $1 mil prize
    - Predict user ratings for movies and TV shows

Data scraping



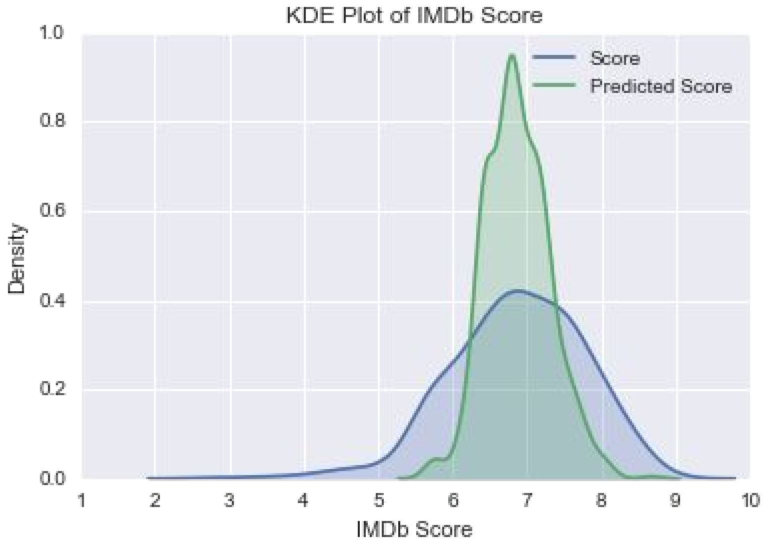
Works with any number.
Too high might crash your browser.

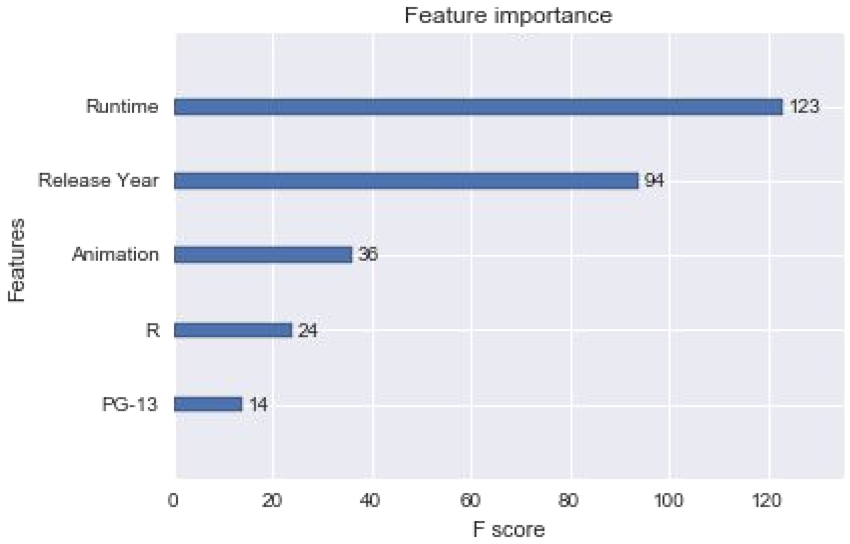
Dataset

- Scraped 2,332 movies from IMDb.com
- Only movies with over 50,000 ratings to preserve integrity
    - Avoids vote manipulation for less popular movies
- Features: actors, directors, genres, year, MPAA certification, runtime
- Will not use number of votes or gross revenue
    - Unknown before release
    - More suitable as label than feature

Results

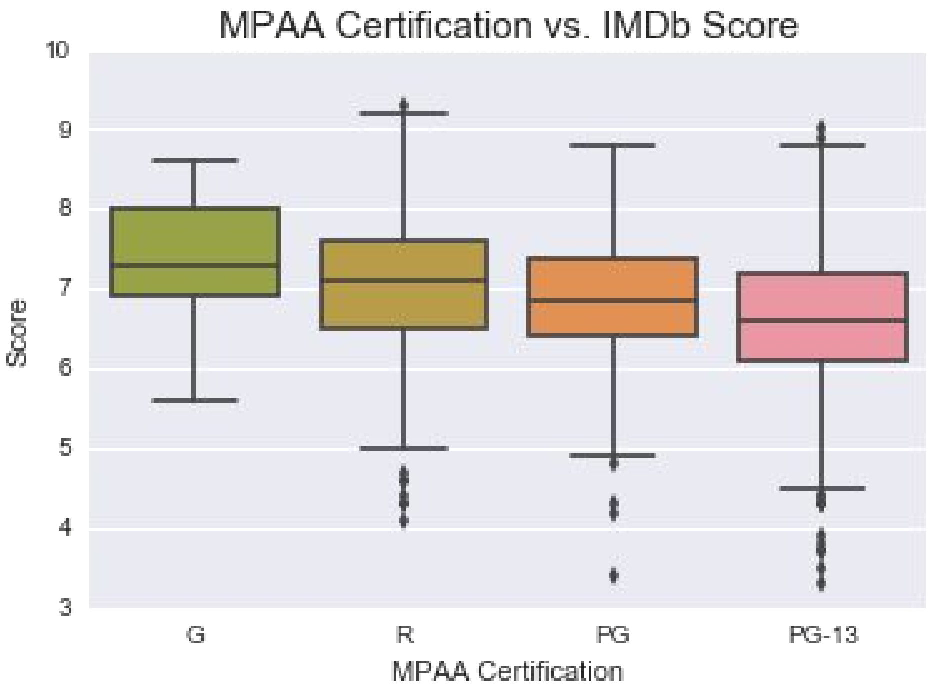
XGBoost R^2 score = 0.275



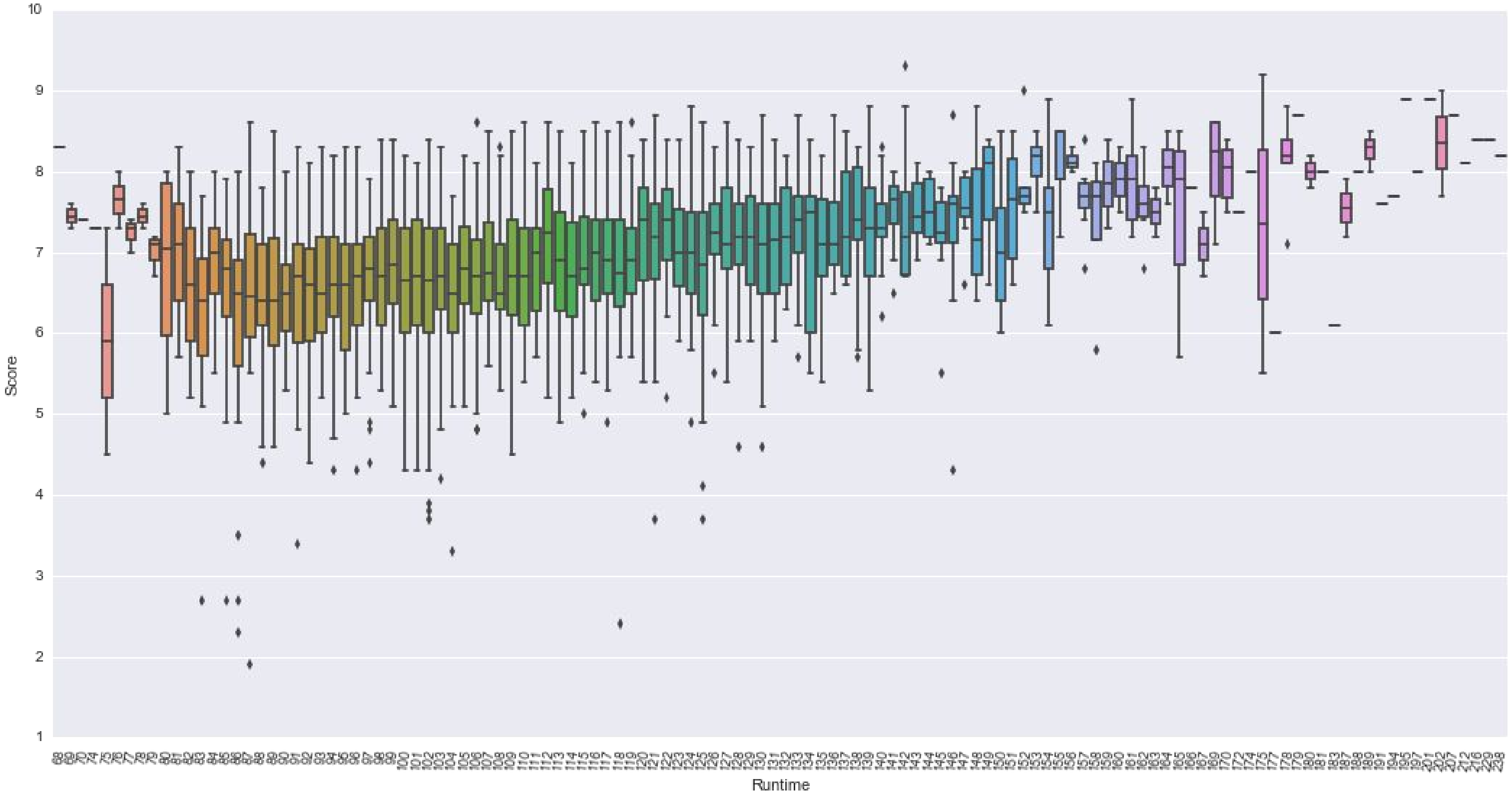


Predictions tend to be safe, near the mean

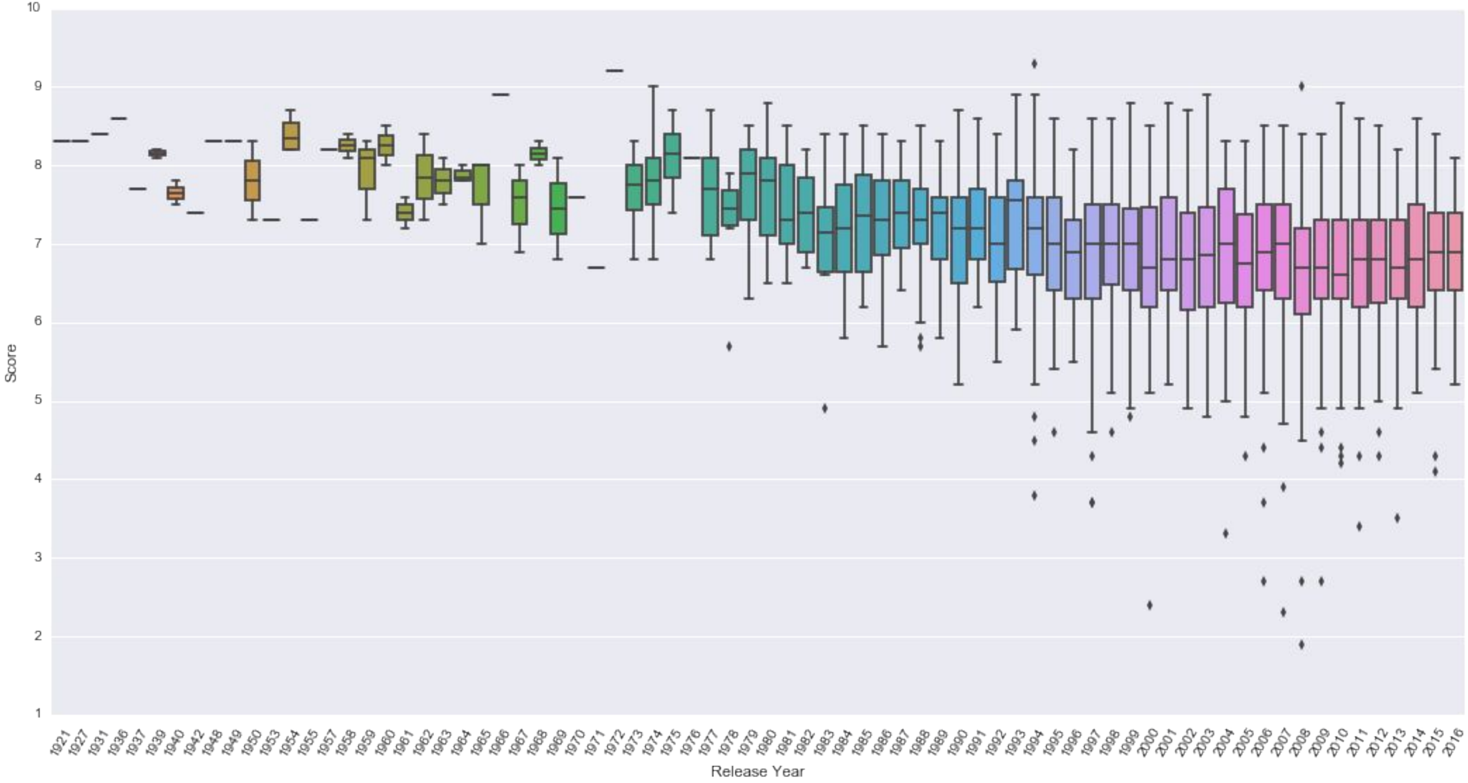
MPAA Certification



Runtime



Release Year



Feature Engineering

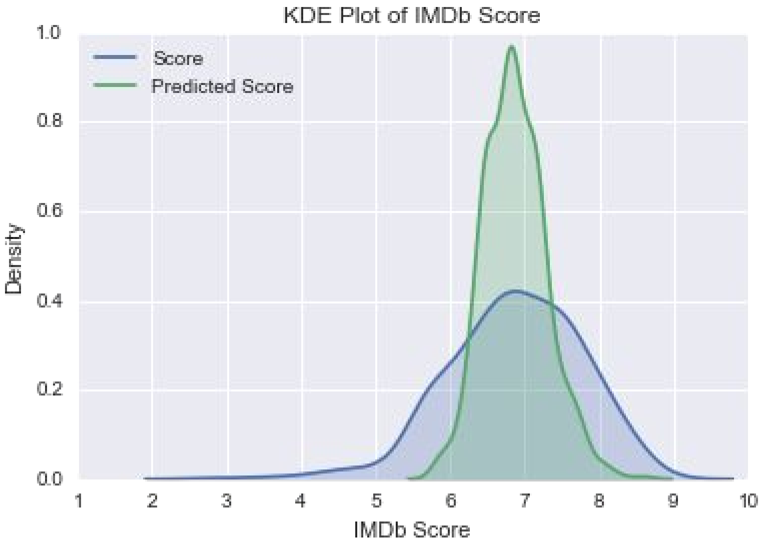
- Title contains ‘2’, ‘3’, or ‘4’?
- Title contains a colon?
- Title starts with ‘The’?
- Length of movie title
- Number of directors
- Number of genres
- Various actors: Leo, Christian Bale, Matt Damon, Brad Pitt, Adam Sandler
- Various directors: Christopher Nolan, Quentin Tarantino, Clint Eastwood

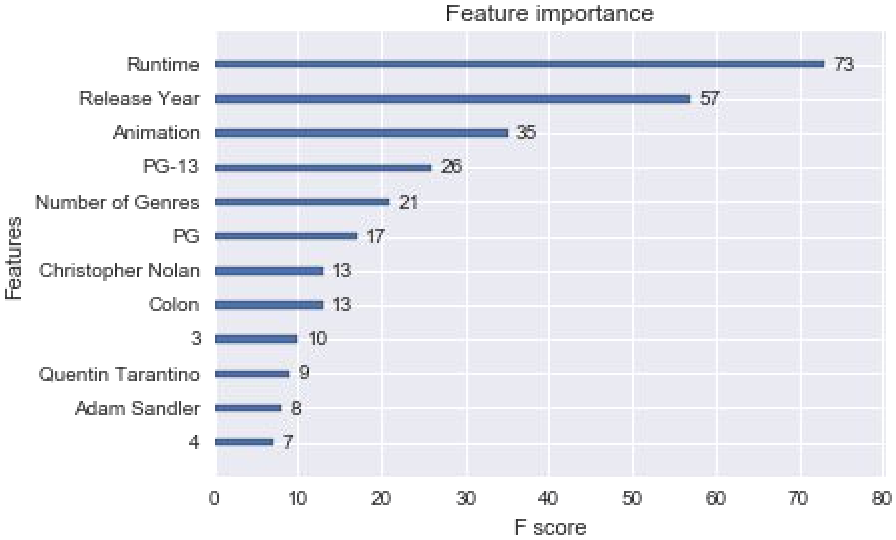
Positive effect
Insignificant
Negative effect

(Assuming α = 0.05)

Results Redux

XGBoost R^2 score = 0.306





Conclusion

- Best features were runtime and release year
- Animations and G movies tend to be rated higher
- Engineered features improved the model
- Certain actors and directors have an impact on ratings
- Other possible features:
    - Facebook likes or tweets for movie, actors, directors
    - Interaction effects between certain actors and directors