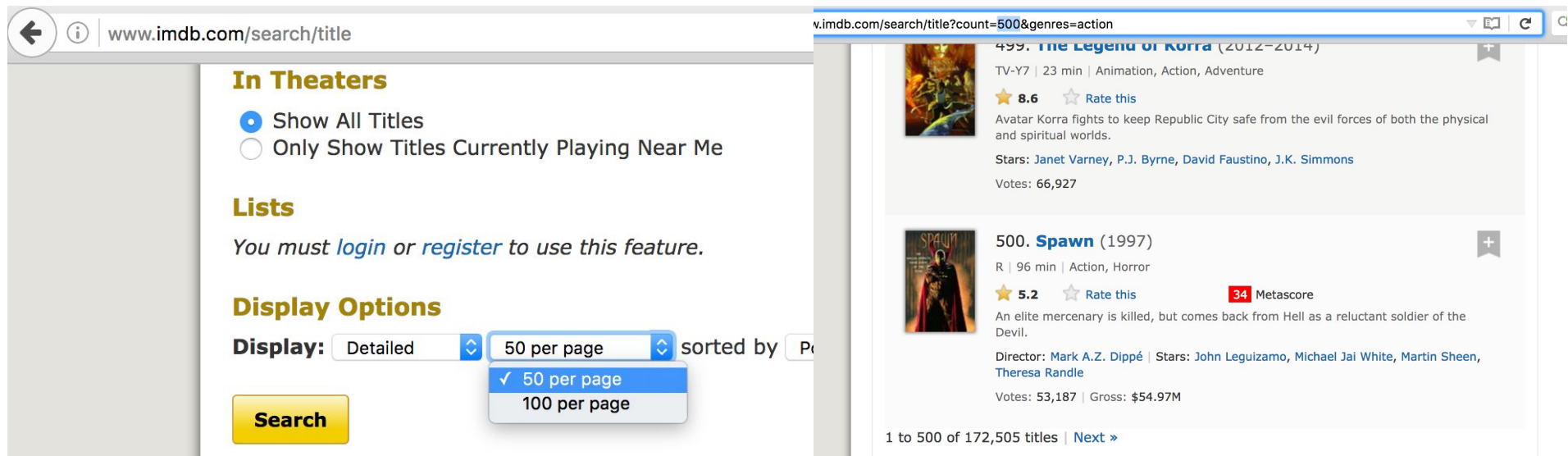

Predicting IMDb User Rating

— 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



The image shows a screenshot of the IMDb website's search results for action movies. The left sidebar contains filters and display options. The main content area shows a list of movies, with 'The Legend of Korra' and 'Spawn' visible.

Filters:

- In Theaters**
 - ☒ Show All Titles
 - ☐ Only Show Titles Currently Playing Near Me
- Lists**

You must [login](#) or [register](#) to use this feature.
- Display Options**

Display: Detailed | 50 per page | sorted by Popularity

50 per page is selected. Other options: 100 per page.

Search

Search Results:

499. **The Legend of Korra** (2012–2014)
TV-Y7 | 23 min | Animation, Action, Adventure
★ **8.6** [Rate this](#)
Avatar Korra fights to keep Republic City safe from the evil forces of both the physical and spiritual worlds.
Stars: [Janet Varney](#), [P.J. Byrne](#), [David Faustino](#), [J.K. Simmons](#)
Votes: 66,927

500. **Spawn** (1997)
R | 96 min | Action, Horror
★ **5.2** [Rate this](#) **34** Metascore
An elite mercenary is killed, but comes back from Hell as a reluctant soldier of the Devil.
Director: [Mark A.Z. Dippé](#) | Stars: [John Leguizamo](#), [Michael Jai White](#), [Martin Sheen](#), [Theresa Randle](#)
Votes: 53,187 | Gross: \$54.97M

1 to 500 of 172,505 titles | [Next](#) »

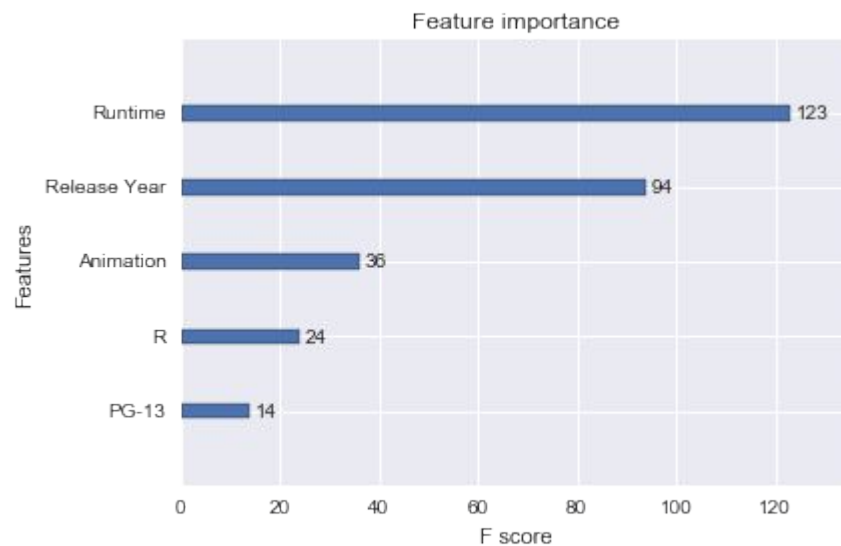
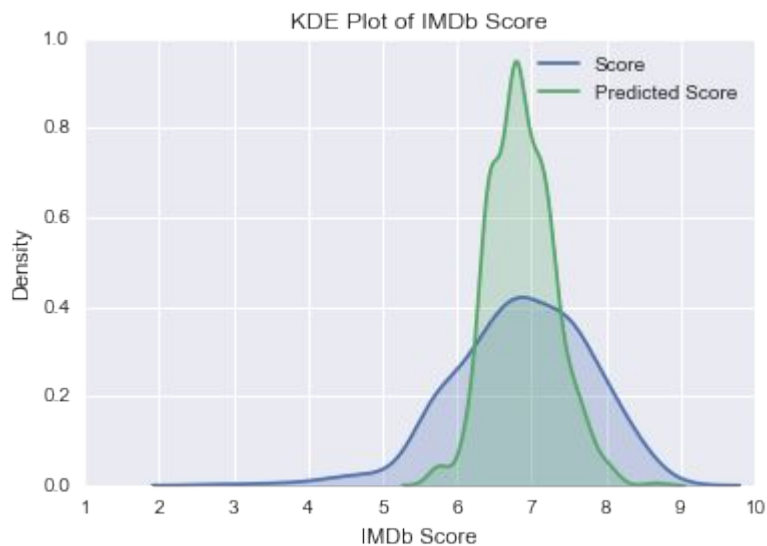
Works with any number.
Too high might crash your browser.

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

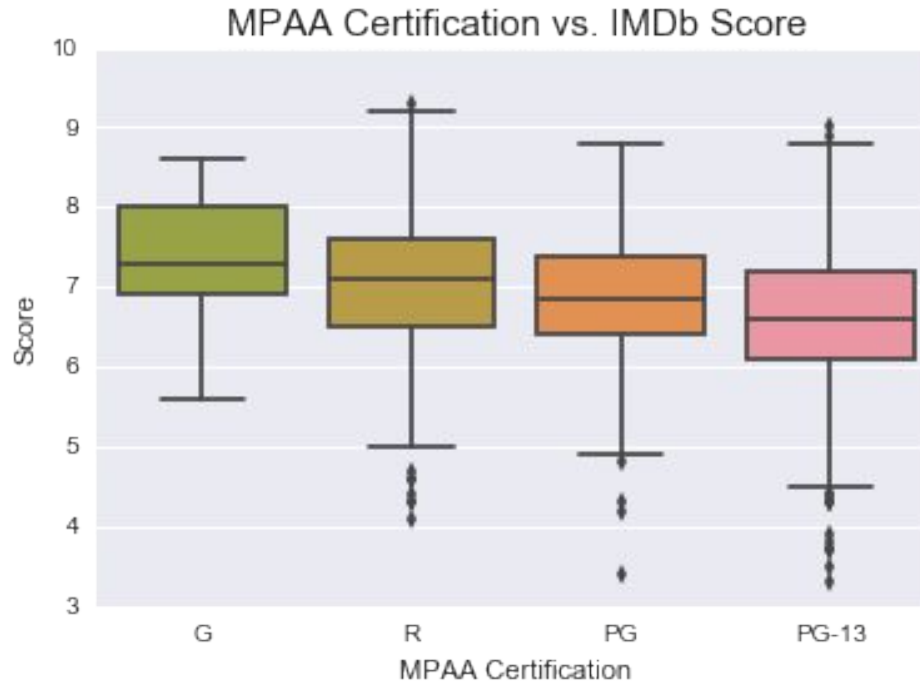
Results

XGBoost R^2 score = 0.275

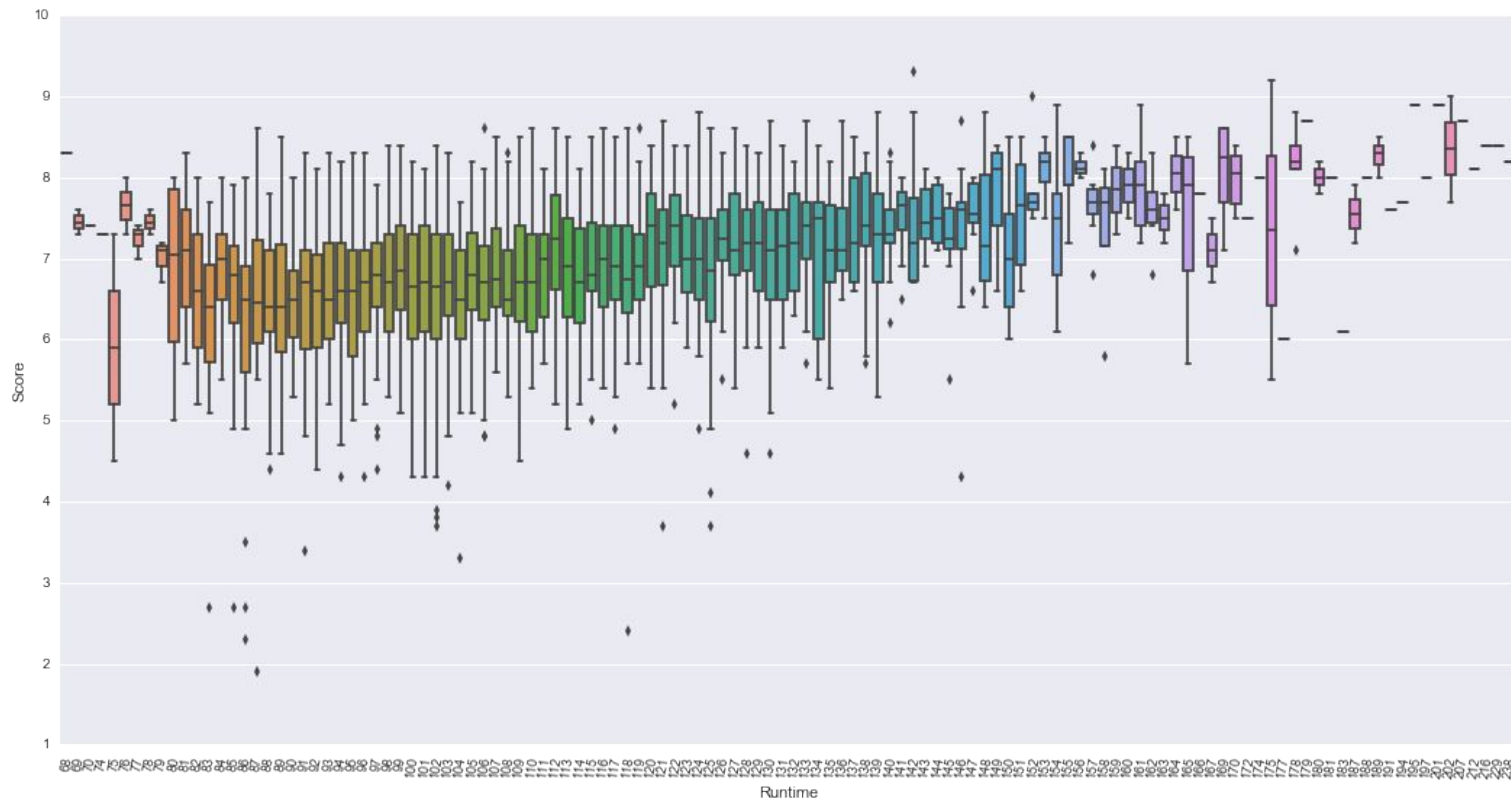


Predictions tend to be safe, near the mean

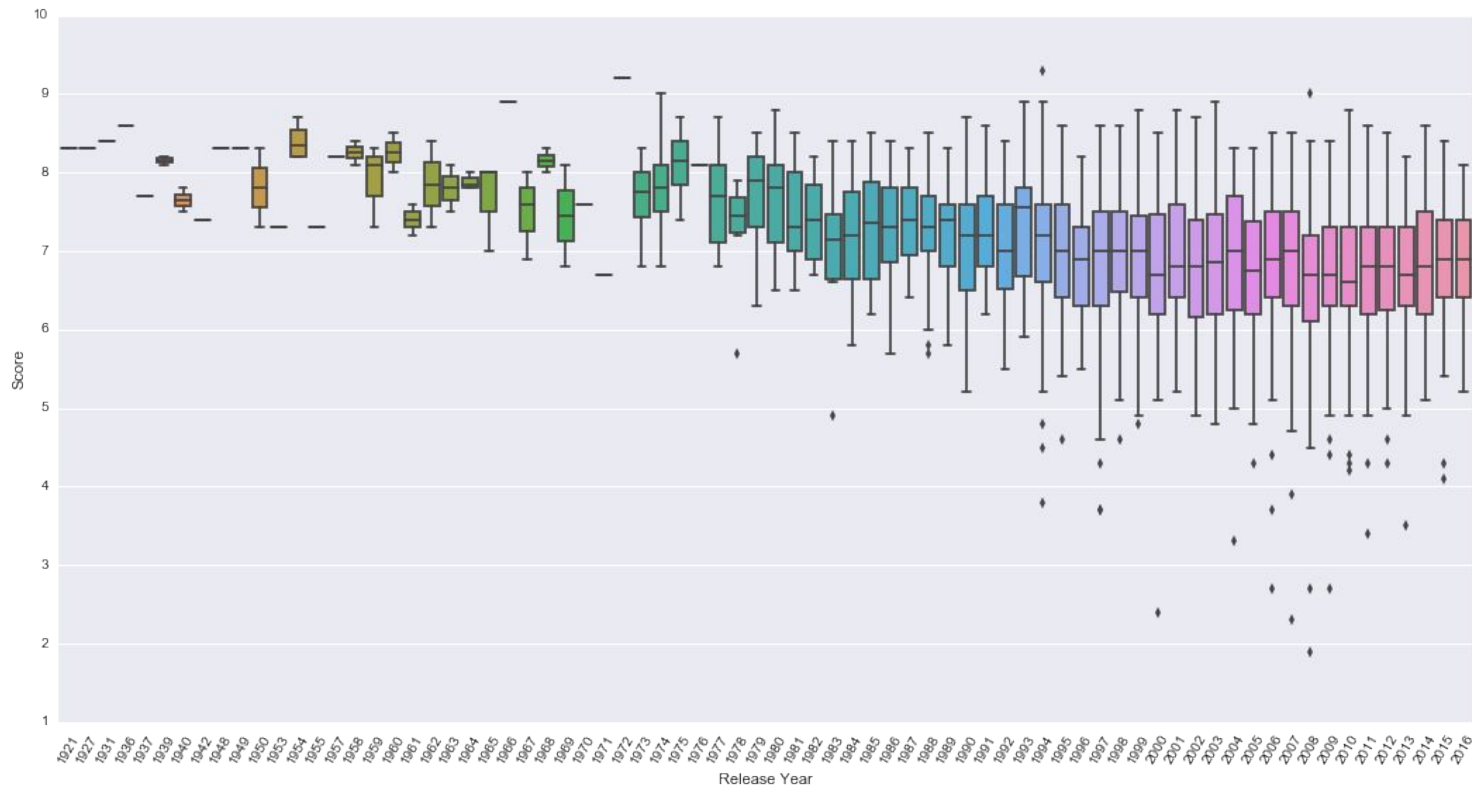
MPAA Certification



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

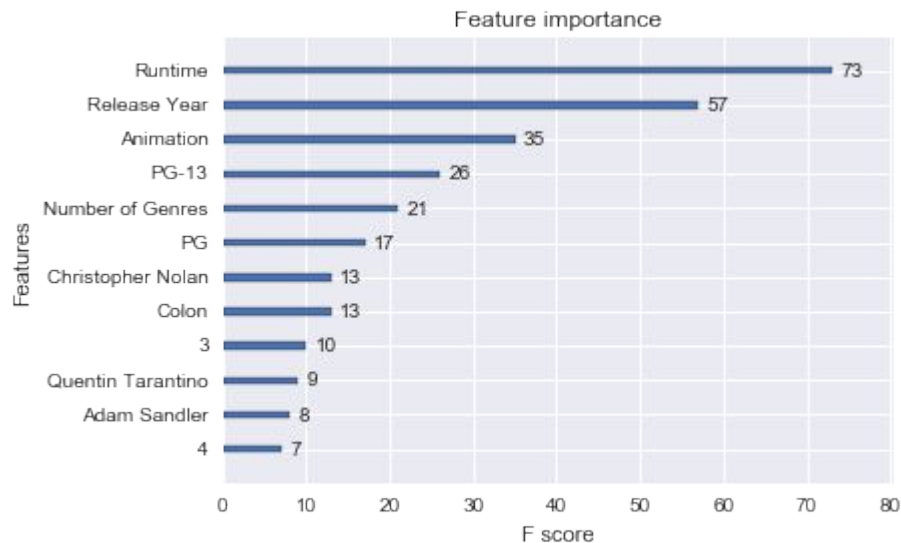
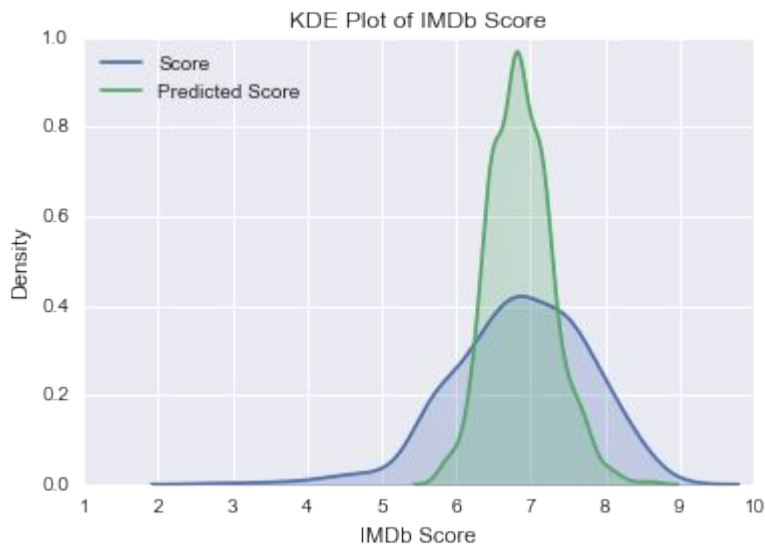
Insignificant

Negative effect

(Assuming $\alpha = 0.05$)

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