Movie Recommendation System

Sarah Zoeller, Joe Resis, Crissy Bruce

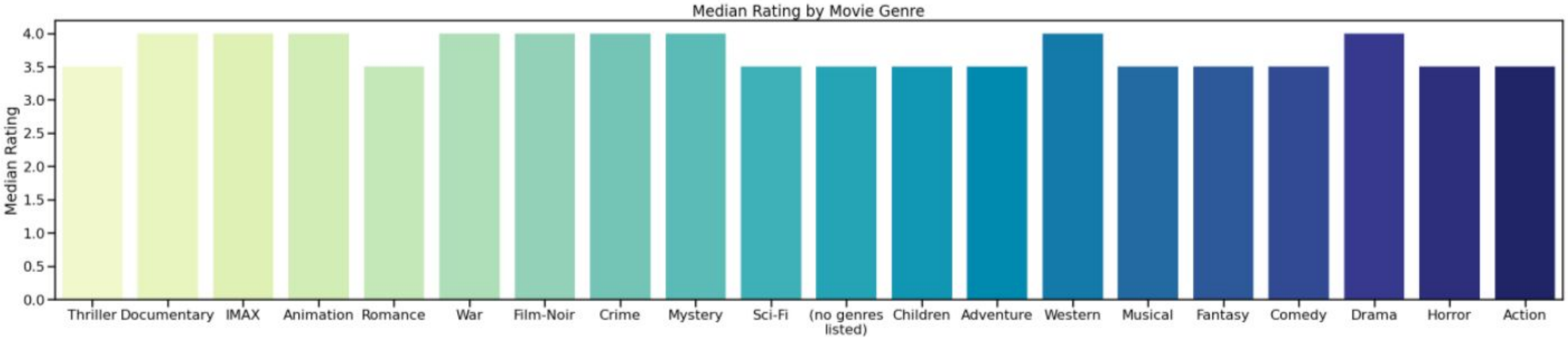
Business Case

Netflix has hired SCJ consulting to build a movie recommendation system to use for customers who first join the platform. The system asks users to rate five movies they have seen and returns five tailored movie recommendations. The system will use a collaborative filtering model to create the recommendations.

The Data

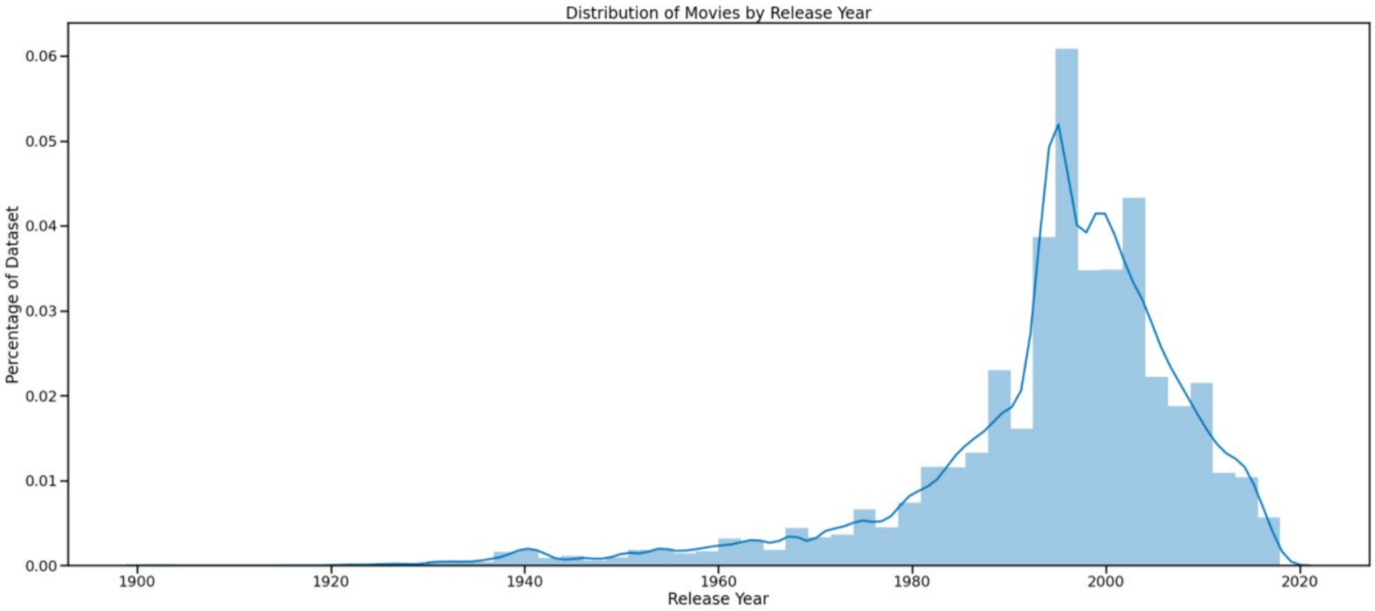
- Data was sourced from the Movielens dataset and IMDB
- Dataset included more than 100,000 ratings by 610 users for 9,700 movies
- Ratings are on a scale of 0.5 - 5, with 5 being the best rated
- Features of the dataset included User IDs, Movie IDs, Ratings, Release Year, Runtime, Director(s), and Genre(s)
- The mean rating is 3.50, and the standard deviation is 1.04.

EDA



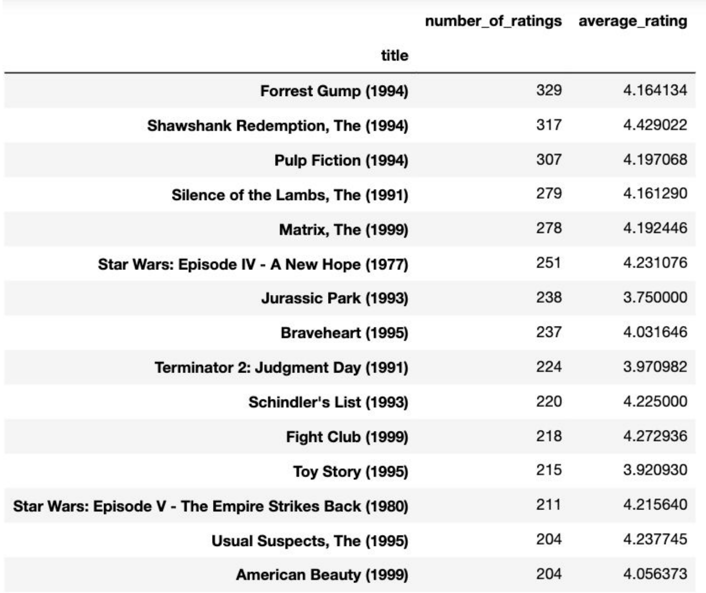
- Median rating is mostly consistent across all genres (3.5-4)
- Minimal genre bias in model

EDA

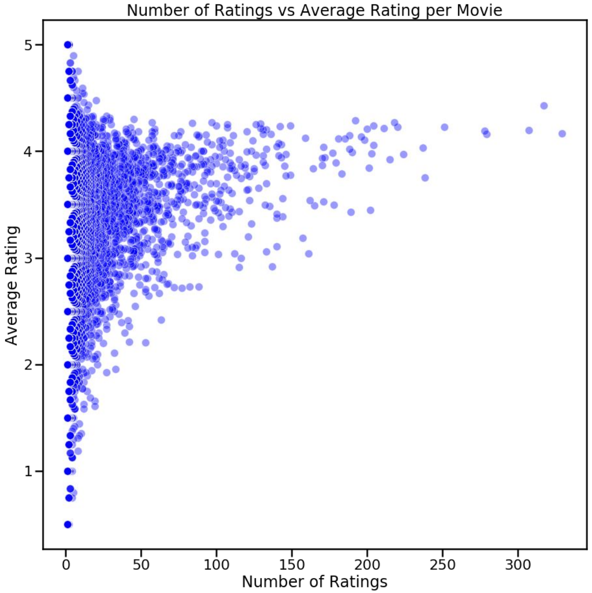


- Most ratings from movies between 1990 and 2010

EDA



title	number_of_ratings	average_rating
Forrest Gump (1994)	329	4.164134
Shawshank Redemption, The (1994)	317	4.429022
Pulp Fiction (1994)	307	4.197068
Silence of the Lambs, The (1991)	279	4.161290
Matrix, The (1999)	278	4.192446
Star Wars: Episode IV - A New Hope (1977)	251	4.231076
Jurassic Park (1993)	238	3.750000
Braveheart (1995)	237	4.031646
Terminator 2: Judgment Day (1991)	224	3.970982
Schindler's List (1993)	220	4.225000
Fight Club (1999)	218	4.272936
Toy Story (1995)	215	3.920930
Star Wars: Episode V - The Empire Strikes Back (1980)	211	4.215640
Usual Suspects, The (1995)	204	4.237745
American Beauty (1999)	204	4.056373



- Frequently rated movies were rated higher on average.
- Over 6000 movies had less than 5 ratings, potentially resulting in popularity bias

Modeling

Baseline Model: KNN Basic

RMSE: 0.97

On average, the model's predictions for user ratings are approximately 1 point off (on a scale of 0.5-5).

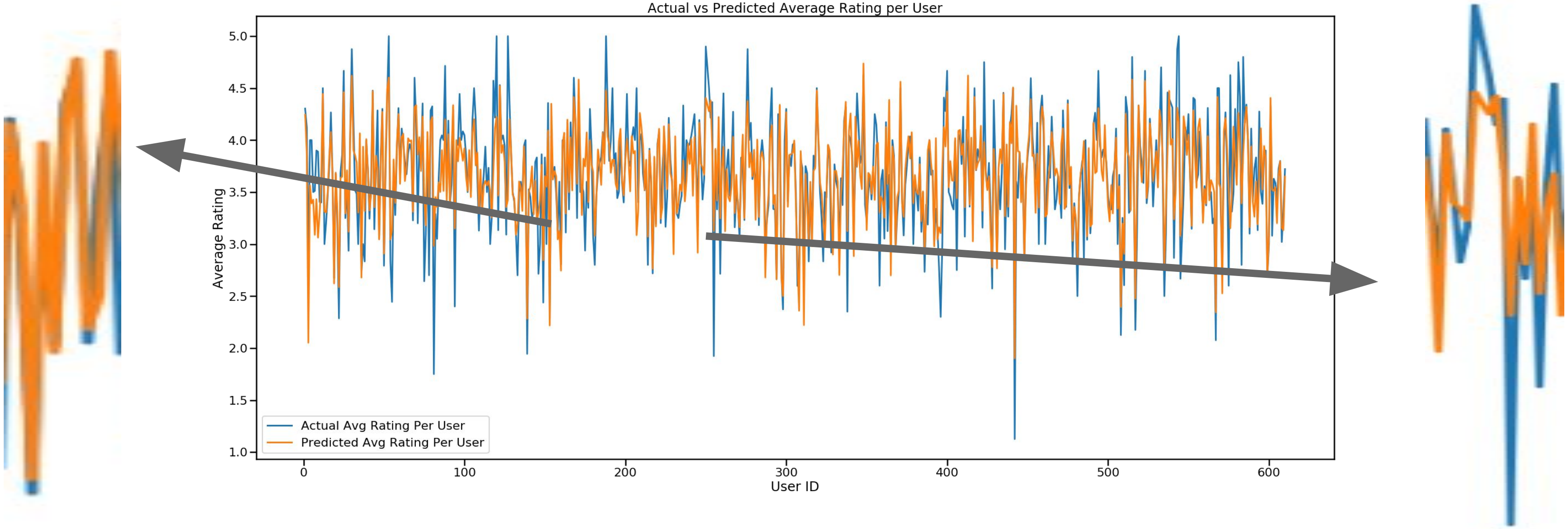
Final Model: SVD

RMSE: 0.87

The grid search returned an SVD model that is **marginally better** than the baseline model. This is the model we will use for our recommendation function.

Comment: Average error lower than standard deviation of 1.04

Post Modeling EDA

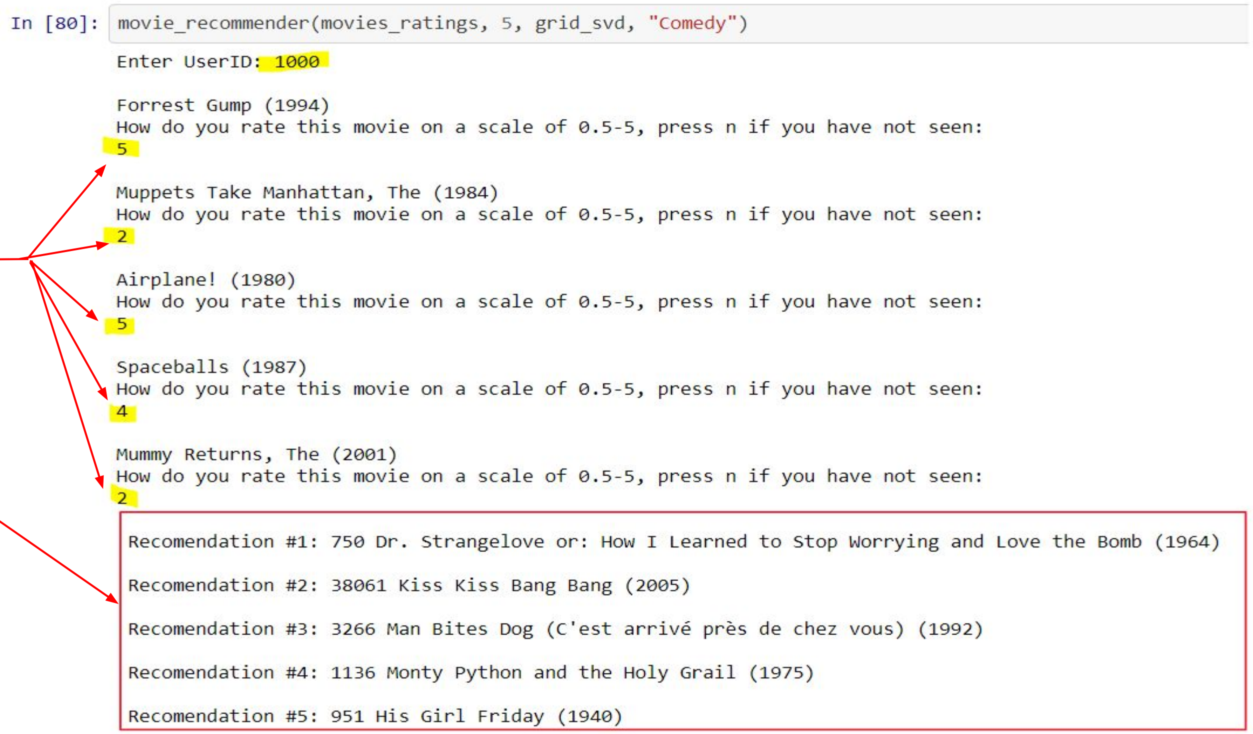


- Model predictions had less variance than actual ratings

Movie Recommendation Function

Function that presents movies from specific **genres** for **new users** to **rate.**

Then it outputs a list of **recommended movies** in the same genre based on the ratings provided by the user.



```
In [80]: movie_recommender(movies_ratings, 5, grid_svd, "Comedy")

         Enter UserID: 1000

         Forrest Gump (1994)
         How do you rate this movie on a scale of 0.5-5, press n if you have not seen:
         5

         Muppets Take Manhattan, The (1984)
         How do you rate this movie on a scale of 0.5-5, press n if you have not seen:
         2

         Airplane! (1980)
         How do you rate this movie on a scale of 0.5-5, press n if you have not seen:
         5

         Spaceballs (1987)
         How do you rate this movie on a scale of 0.5-5, press n if you have not seen:
         4

         Mummy Returns, The (2001)
         How do you rate this movie on a scale of 0.5-5, press n if you have not seen:
         2

          Recomendation #1: 750 Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)

          Recomendation #2: 38061 Kiss Kiss Bang Bang (2005)

          Recomendation #3: 3266 Man Bites Dog (C'est arrivé près de chez vous) (1992)

          Recomendation #4: 1136 Monty Python and the Holy Grail (1975)

          Recomendation #5: 951 His Girl Friday (1940)
```

Conclusions

- The final model is not a perfect fit based on the RMSE (0.87), but is less than the standard deviation of the original ratings dataset
- Recommendations appear to be more accurate when genre is specified

Future Steps and Limitations

- Obtain reviews for movies that do not have a sufficient amount of reviews to be deemed reliable to the dataset or consider removing from model
- Investigate approaches to deal with popularity bias so to increase the representation of less popular movies (considering including weights in model)
- Create a more robust model with LightFM by incorporating movie features into weighting
- Calculate similarity metric between recommended movies and highest rated movies to better validate recommendations

Thank you for listening!

Sarah Zoeller (swzoeller@gmail.com)

Crissy Bruce (crissybruce@gmail.com)

Joe Resis (jresis10@gmail.com)

https://github.com/swzoeller/Movie-Recommendation-System