Building a Movie Recommendation System

Miriam Semmar

Overview

- We'll be working to create a movie recommendation system on behalf of Moviefy, a new streaming video competitor. A solution to boredom and overwhelming choice, this new platform will be entirely based on our recommender system, eliminating the ability to search for titles at all.
- For this analysis, we leveraged a python library (surprise) and Spark ALS models to train recommender systems on the data provided*.
    - Data included movie titles, user ids and user ratings of movies
- Our primary KPI was RMSE

Approach

- We started out by doing a brief exploration of our data
- Then, we tested multiple prediction algorithms from the surprise library and narrowed our focus to three. We then used these three algorithms to perform GridSearchCV in order to improve the model performance.
    - While GridSearchCV did yield RMSE improvements, the improvements were minimal.
    - Moreover, the winning surprise model, however, was very slow.
- Instead, we built an ALS based recommendation model using Spark which resulted in a similar RMSE score. The ALS model predicts movie ratings for users within +/- 0.9 stars of their actual rating.

Model	RMSE
SVDpp	0.861879
Baseline	0.876181
SVD	0.878338
KNN Baseline	0.878345
KNN with Z-Score	0.899625
KNN with Means	0.900986
Slope One	0.905059
NMF	0.923854
Co-Clustering	0.948221
KNN Basic	0.949513
Normal Predictor	1.430104

How Our Recommendation System Works

- Our recommendation system works by first determining if someone is a new or existing user by prompting the user for an ID.

- If this is an **existing user**, the system generates 5 recommendations based on their viewing and rating history using the ALS model.

- If this is a **new user**, the system asks the user to rate 5 movies. This helps to avoid the cold start problem, as the recommendation system is dependent on having existing ratings for users.
    - Based on these responses, the new user received 5 movies recommendations.

Next Steps

- Ask new users to rate some of the most popular movies (see Appendix), to avoid drawing out the initial survey. We cannot eliminate the initial survey is necessary to resolve the cold start problem.

- Create a more sophisticated model that can generate recommendations based on genres or even age group (children only, for example).

- We should be updating the dataset as new movies become available on our platform.

- Our dataset has a high number of comedy, fantasy and thriller movies (see Appendix). It's possible that our recommendations won't be as strong for movies that fall into less popular genres. We could consider a larger dataset.

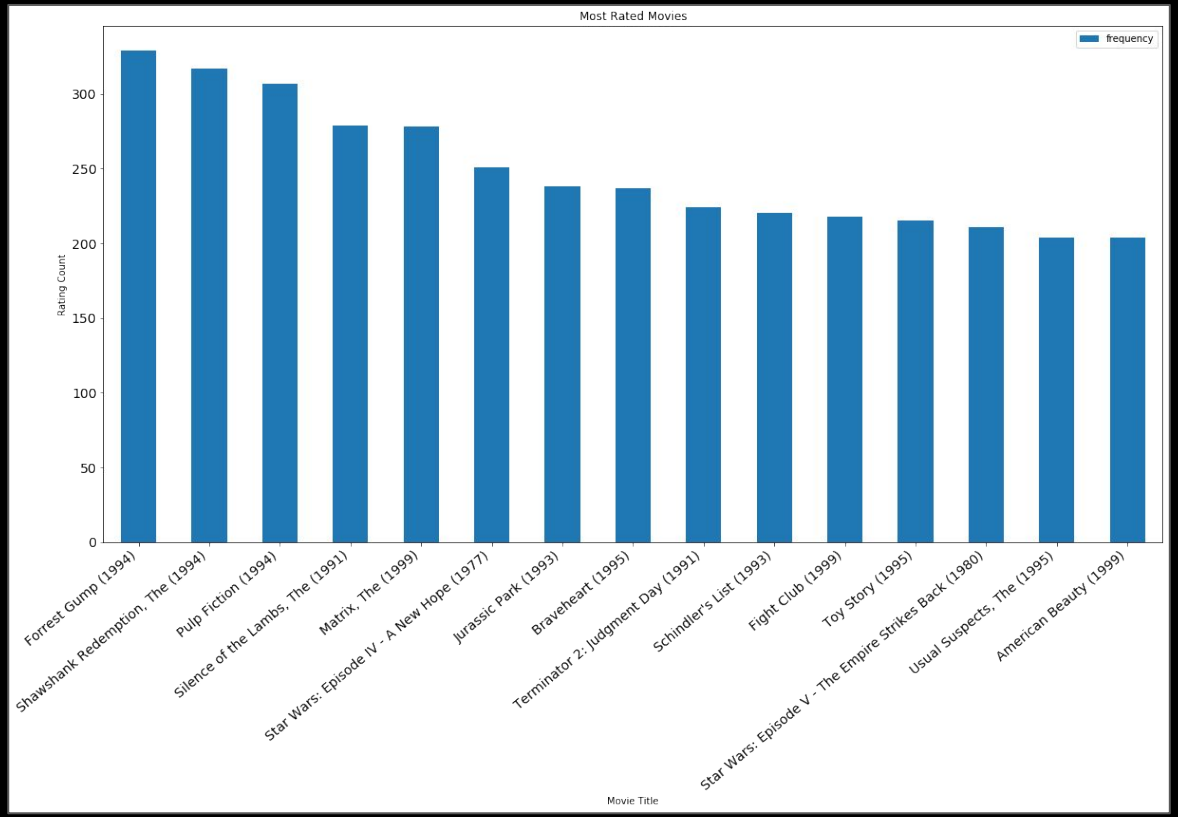
Thank You!

Appendix

Most Popular Movies

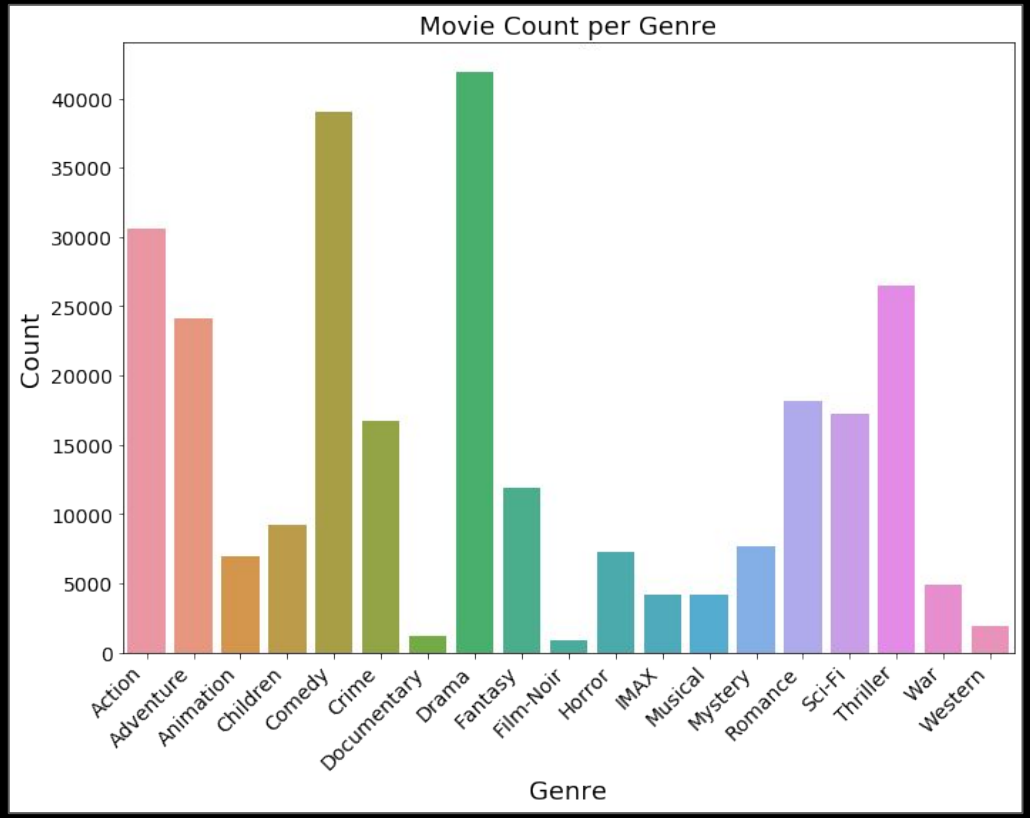
- We can consider limiting the initial survey for new users to the most rated movies

- Given that these movies have the most ratings, we can also infer that they are quite popular and likely to have been seen by new users.



Number of Movies by Genre

- The most common genres in this dataset are comedy, drama, action and adventure.
- Certain genres, like documentary films, are underrepresented.



Sources

- Movie Dataset (https://grouplens.org/datasets/movielens/latest/)