Movie   Recommendation System Modeling

By Mercy Ayub

Agenda

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Exploratory Data Analysis
5. Modeling & Tuning
6. Model Evaluation , Model Selection & Testing
7. Conclusion
8. Recommendation

Business Understanding

Overview

- With the rapid expansion of internet streaming platforms, users are overwhelmed by the sheer volume of available movies. Providing personalized recommendations is critical for increasing user engagement, satisfaction, and retention.
- The MovieLens (ml-latest-small) dataset, from http://movielens.org - a movie recommendation service, includes user-generated 5-star ratings and free-text tags that can be used to create a powerful recommendation engine.

Problem Statement

- With thousands of movies accessible on streaming platforms, customers struggle to discover ones they'll like. This choice overload frequently results in dissatisfaction, decision fatigue, and low user engagement.

- Many systems rely on generic rankings or trending lists that do not consider individual preferences. This leads to irrelevant movie suggestions that do not match user preferences, longer search times lower user happiness, or low retention rates due to users potentially switching to competitors with better recommendations.

Objectives

1. Analyze user ratings and tags to find patterns and trends.
2. Create a recommendation model that incorporates collaborative filtering, content-based filtering, or hybrid approaches.
3. Address critical issues such as data scarcity, cold start issues, and bias in user ratings.
4. Assess model performance using relevant measures such as RMSE or MAE.

Proposed solution

The objective is to examine and use the dataset to boost user engagement by developing a movie recommendation engine. Potential applications include:

- Personalized ***Movie Recommendations*** - Predict user preferences based on previous ratings.
- Customize content by ***segmenting and clustering users*** based on similar preferences.
- ***Trend Analysis and Insights:*** Discover popular genres, top-rated movies, and viewing behaviors.
- Tag-based ***sentiment analysis*** provides insights into user perception of movies.

Challenges & Considerations

- ***Data Sparsity:*** Not all users have rated all movies, leading to gaps in the dataset.
- ***Cold Start Problem***: New users/movies lack enough data for accurate recommendations.
- ***Bias in Ratings***: Some users may consistently rate higher or lower than others.
- ***Scalability***: The model should be efficient enough to handle large datasets in real-world applications.

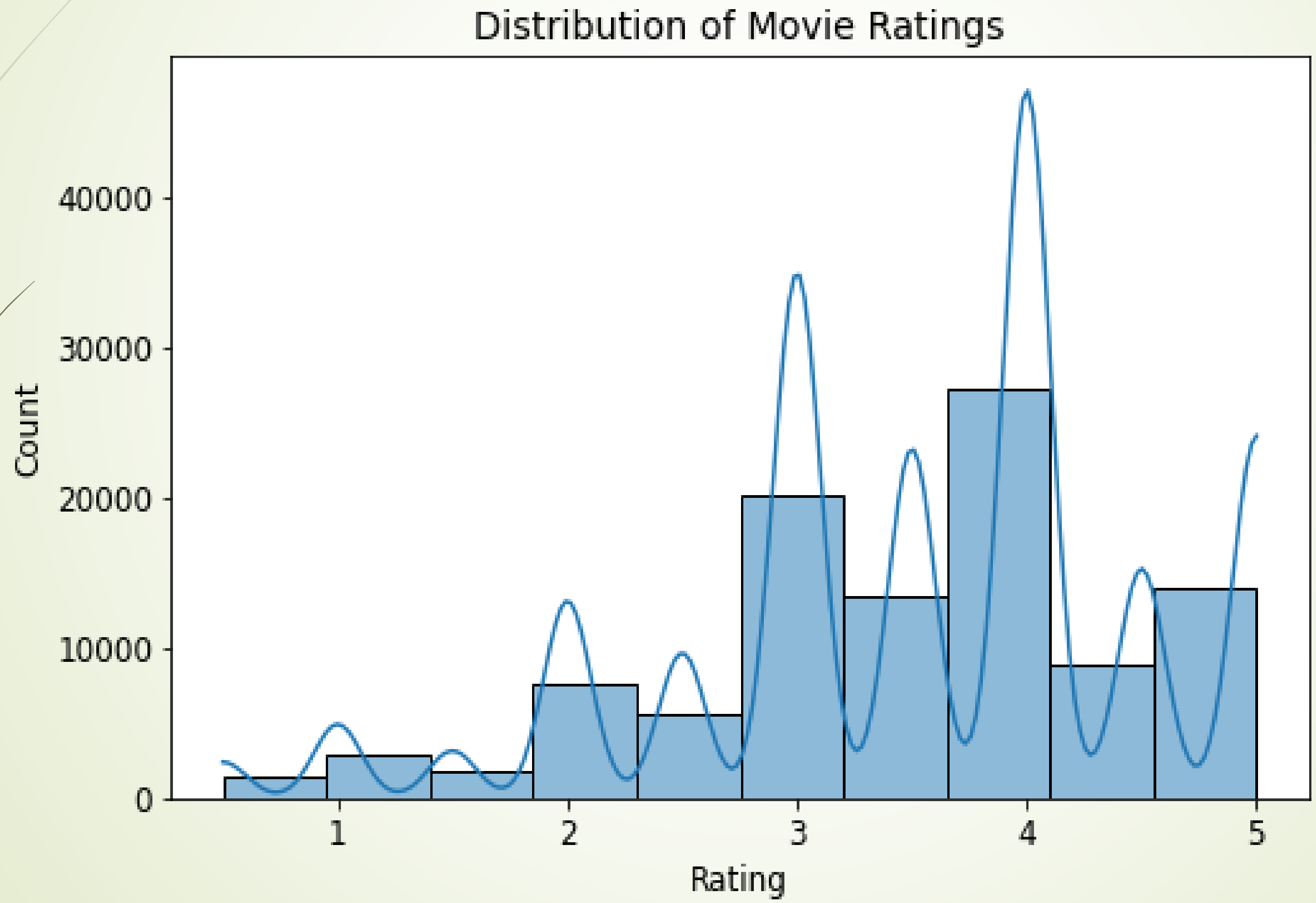
Data Understanding

- ratings.csv: Contains user ratings for movies (1-5 scale).
- movies.csv: Metadata including movie titles and genres.
- tags.csv: Free-text tags assigned by users to movies.
- links.csv: Provides mappings to external movie databases (IMDB, TMDb).

Data Cleaning

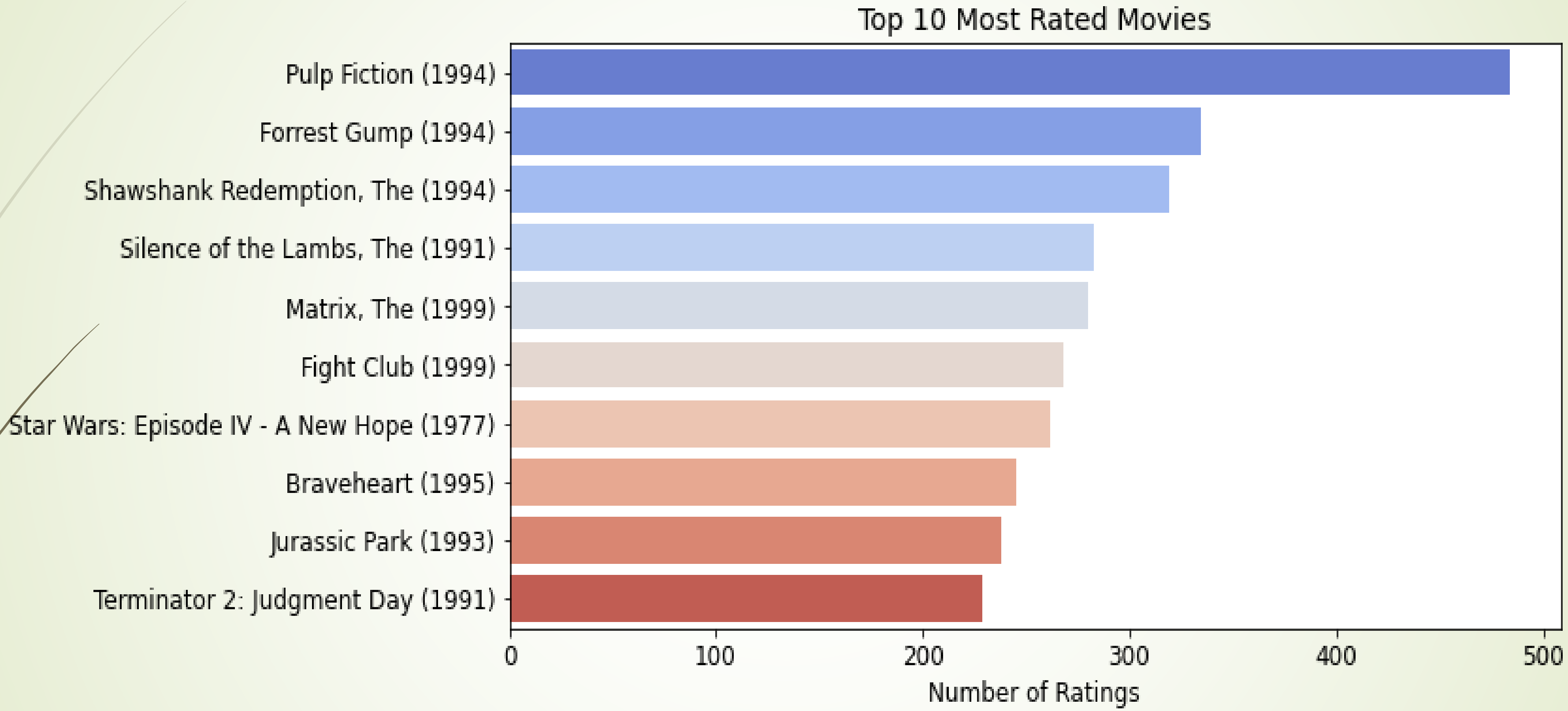
- The data was checked on the following:
  - Null values
  - Duplicate values
  - Dropped unnecessary columns
- The different datasets were merged for exploratory analysis

Movie Rating Distribution



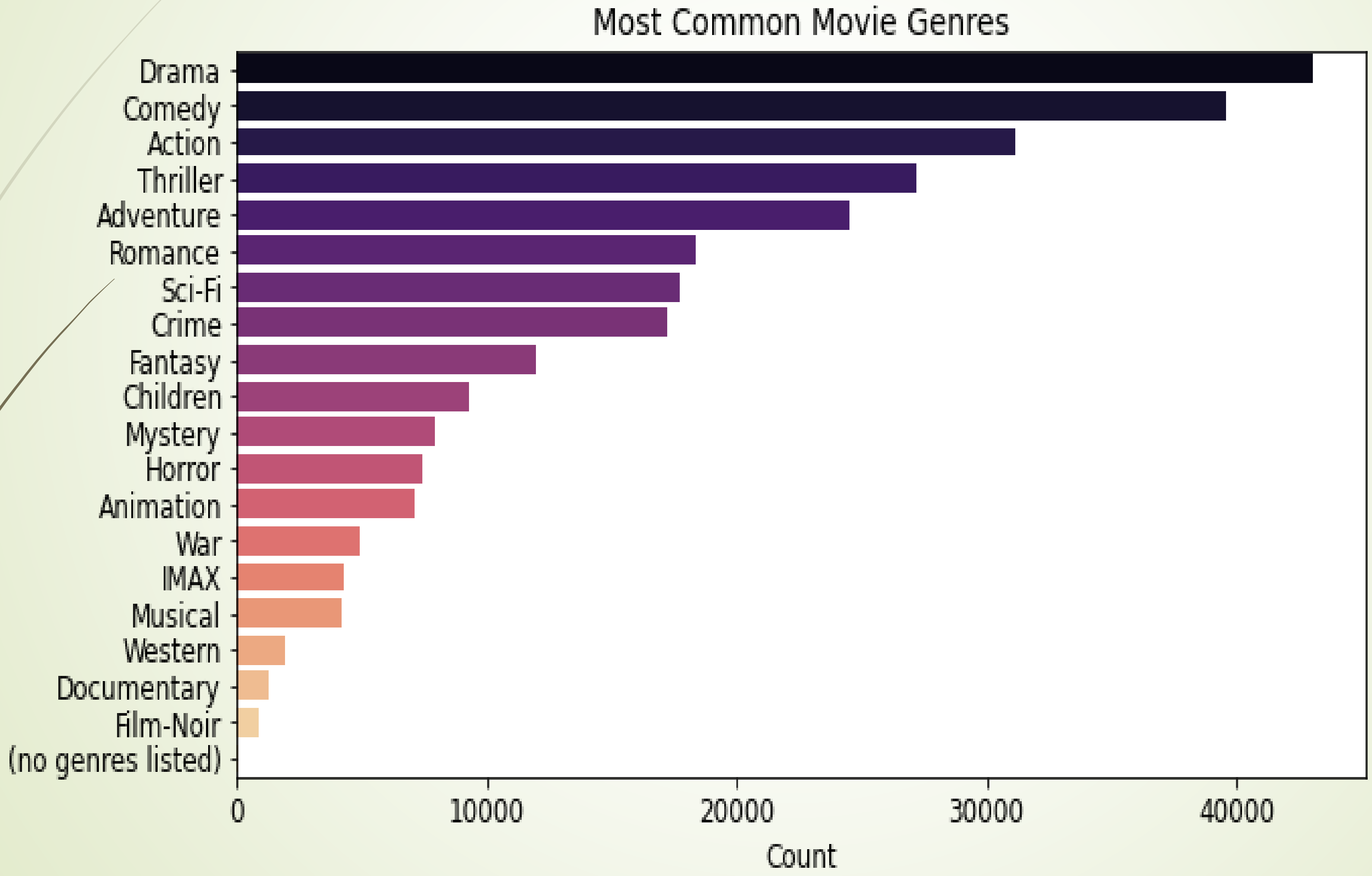
- Most movies are rated between 3 and 4 with 0.5 being the least.

Top 10 Most Rated movies



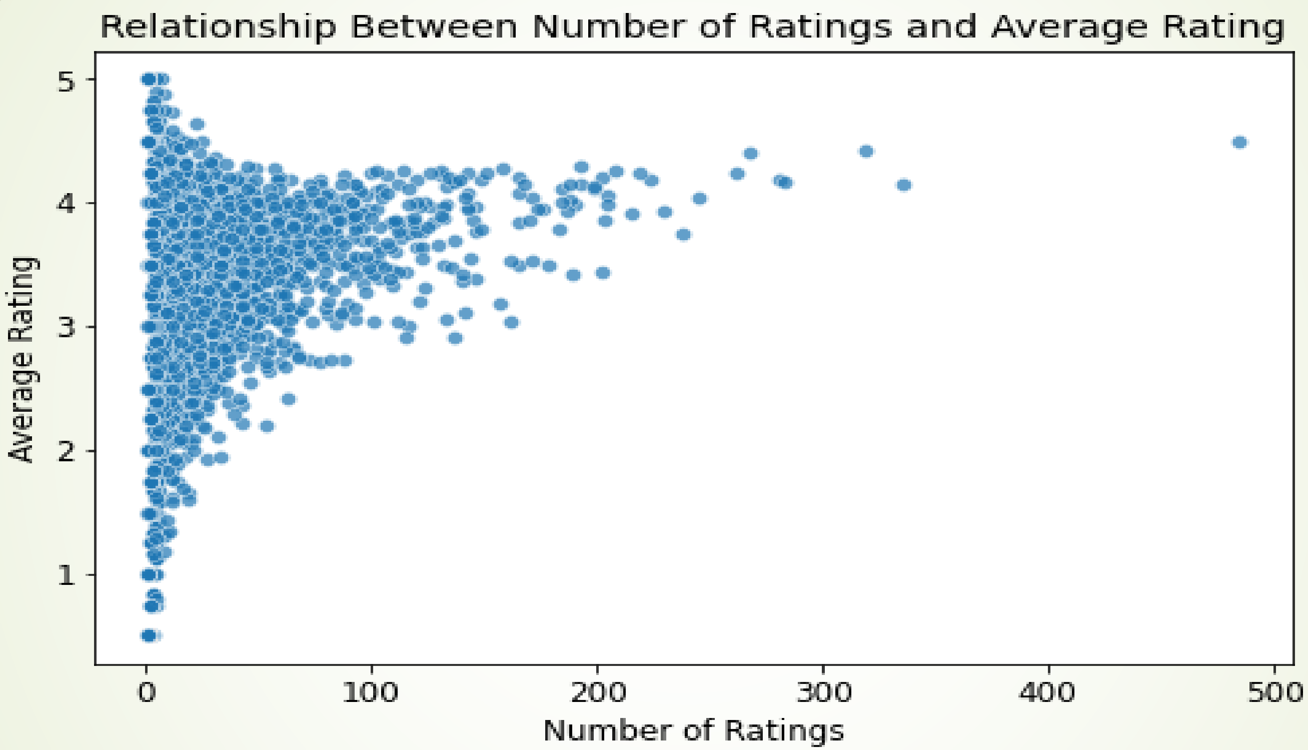
- The pulp fiction has the highest number of ratings
- Majority of the top 10 movies are rated between 200-300 times

Genre popularity using ratings



- The most common genres is Drama and Comedy
- Film-Noir is the least common genre

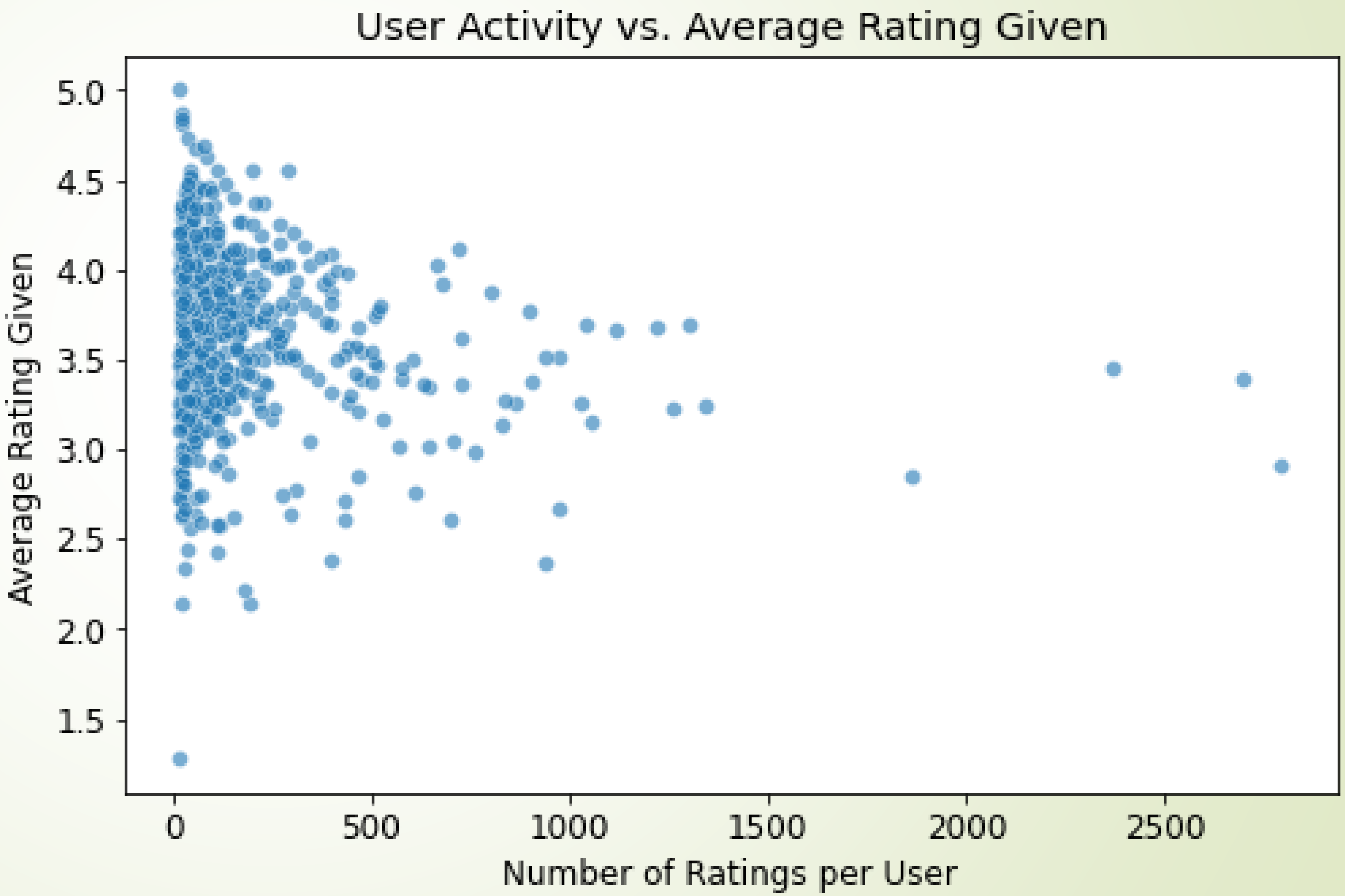
Number of Movie Ratings VS Average Rating



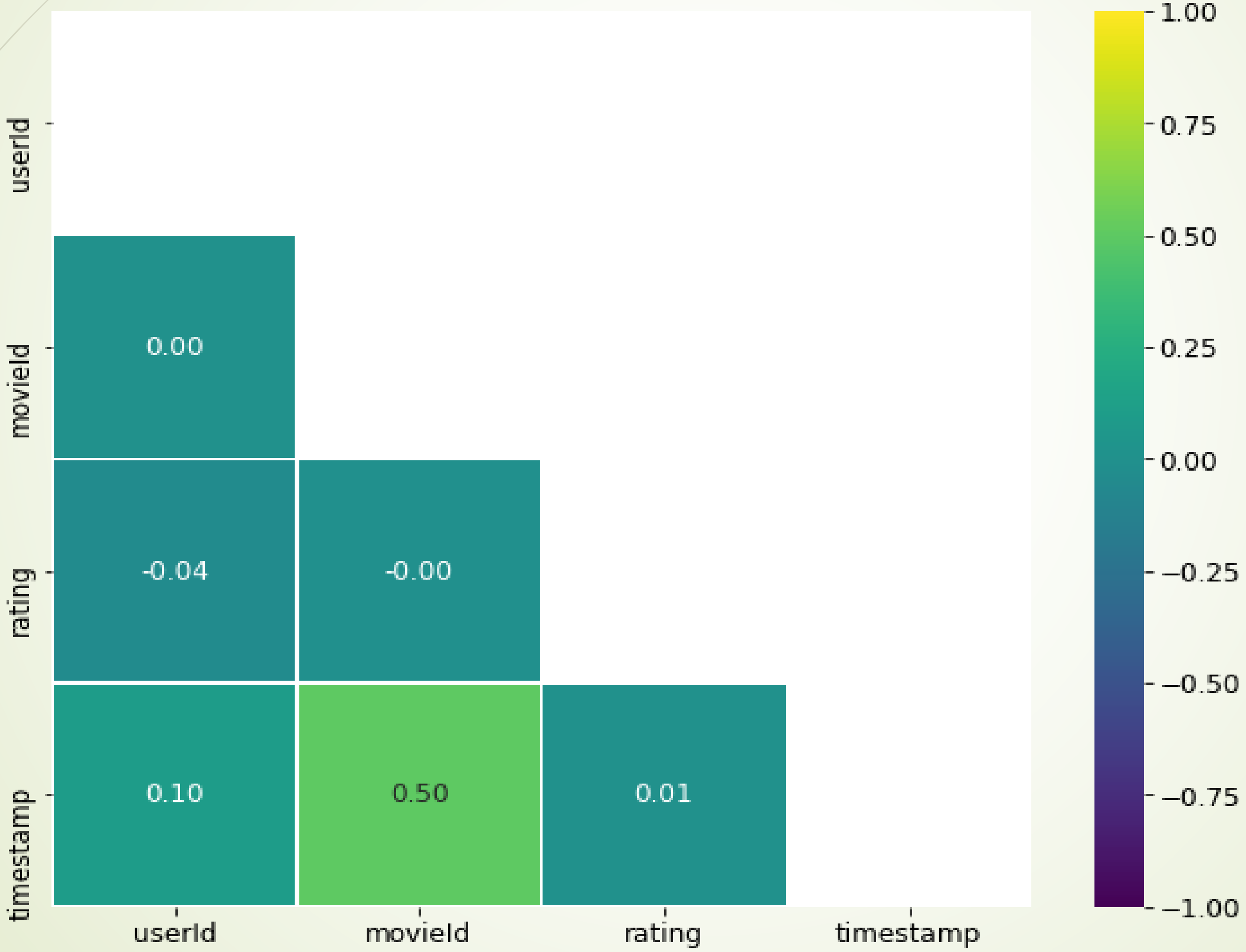
- Most move average ratings seem to be between 2.5 and 4 average rating
- The total number of rating is 200 and below

User behaviour on Ratings

- Users have given less 500 ratings with majority ratings between 3-4

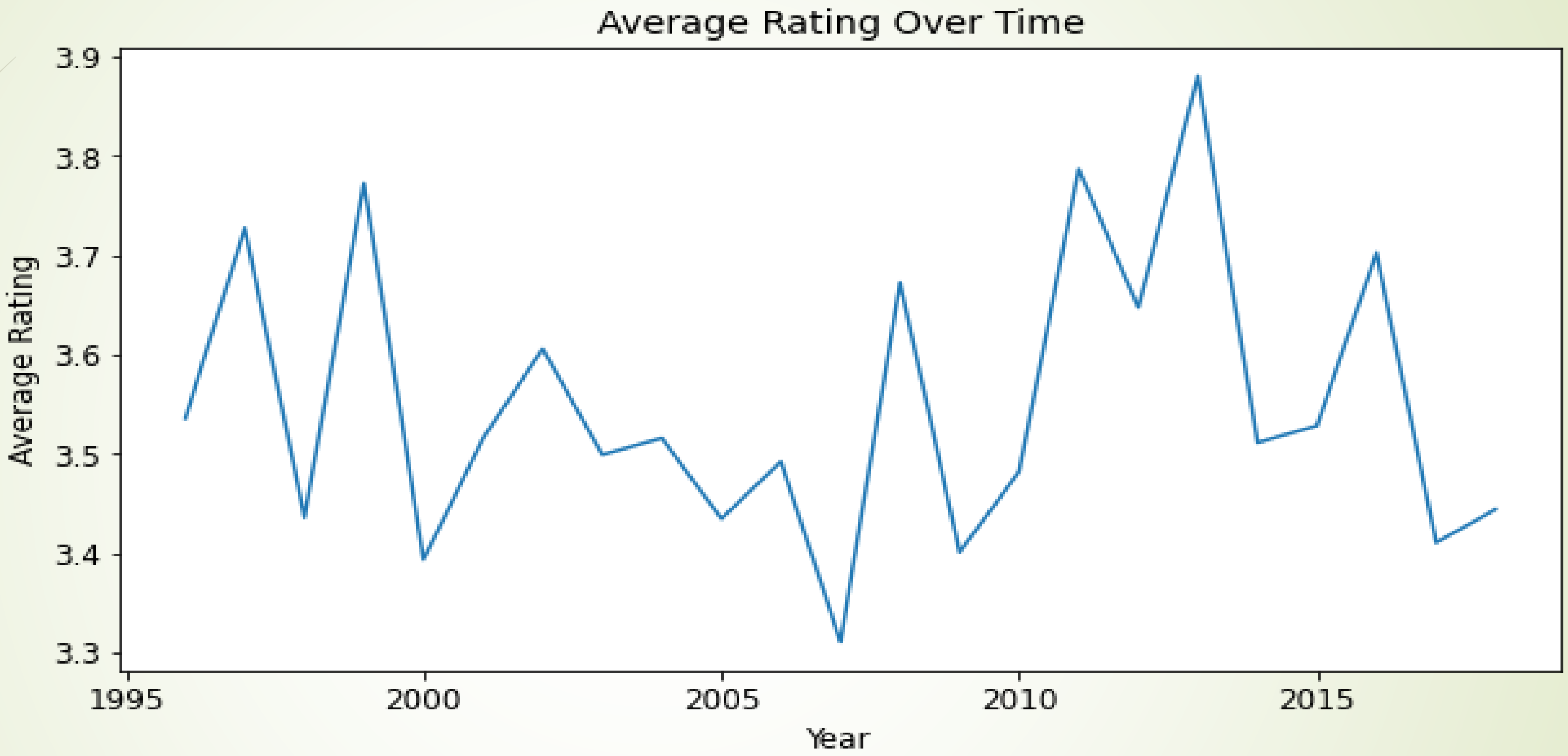


Feature Correlations



- There were no substantive/ meaningful relationships between features of the dataset

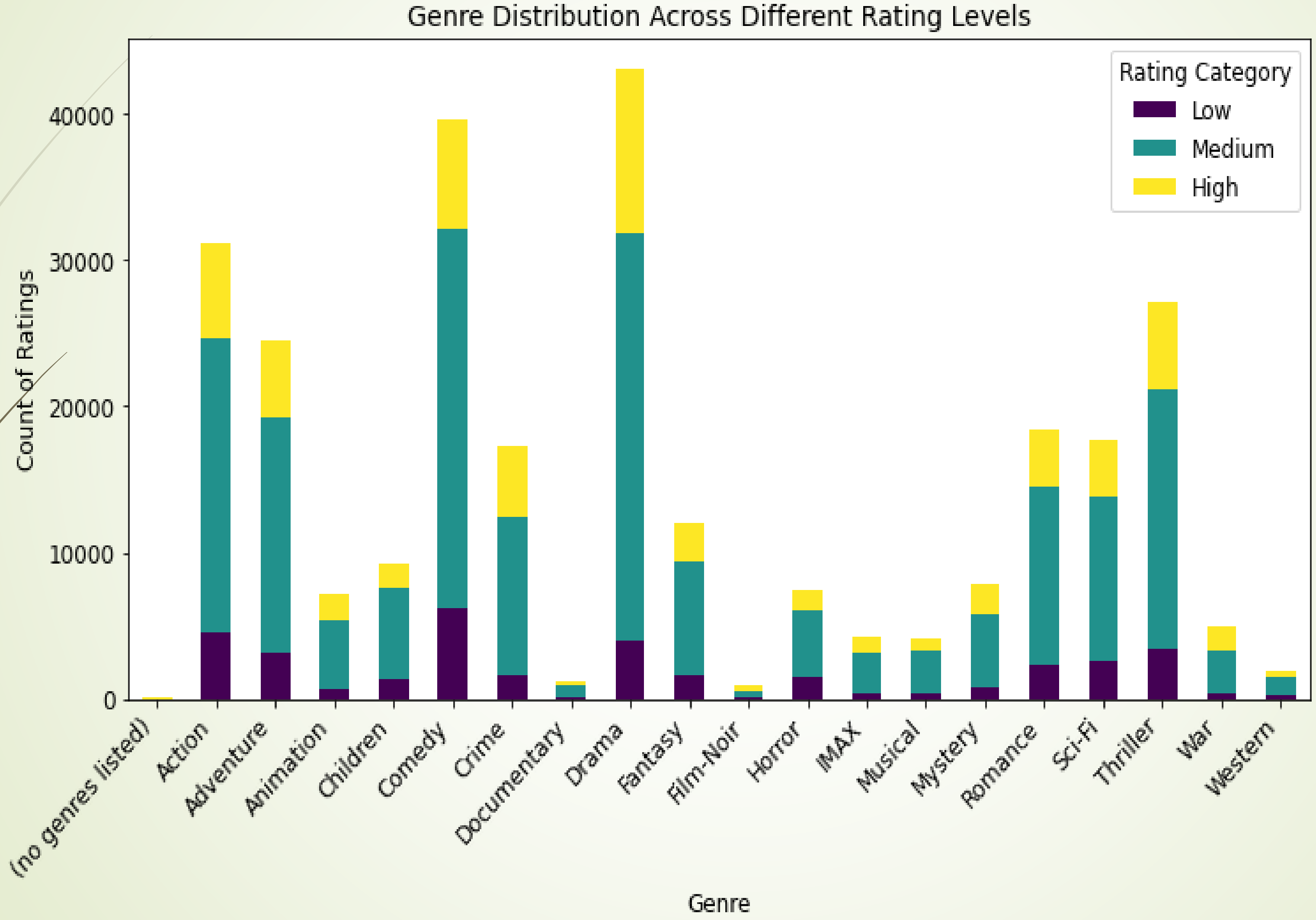
Average Ration Over time (1995-2020)



1. The ratings were higher during the year of 2010 to 2015
2. Between 2000 and 2007, was the lowest rating period.

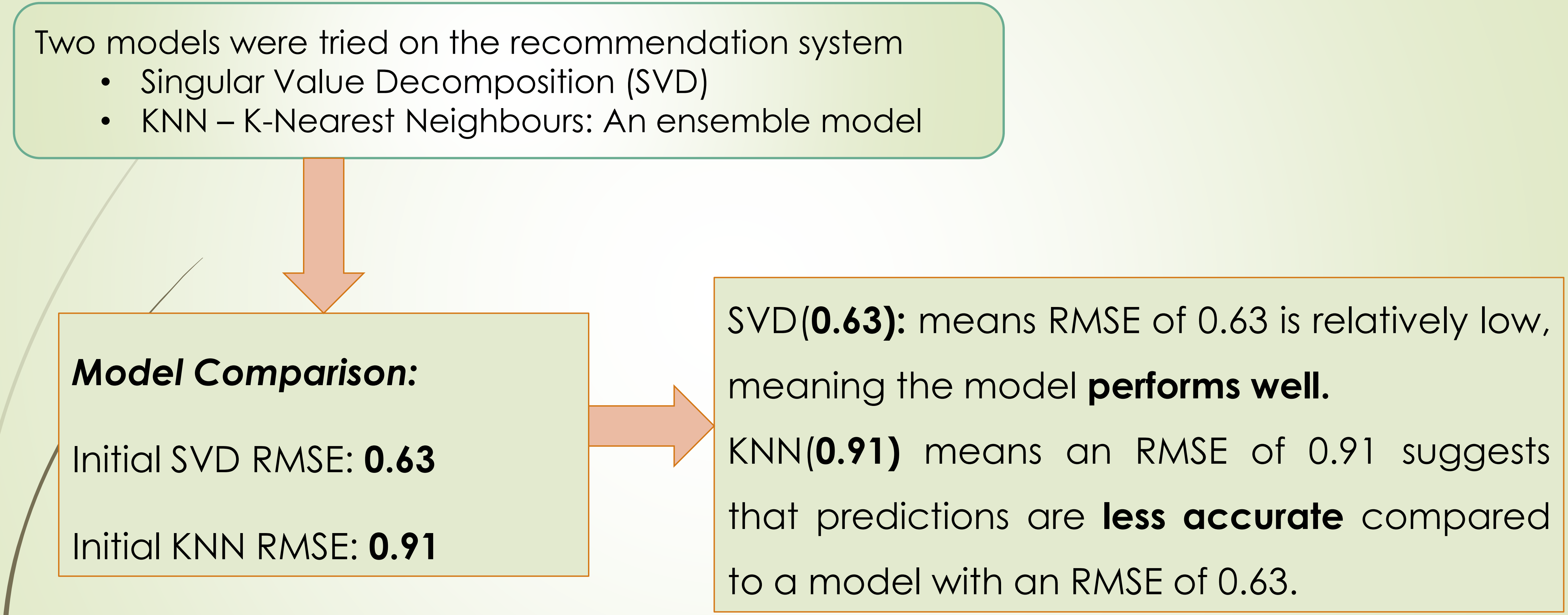
- This may indicate user preference shifts, or something caused low ratings during that time

Distribution of Genres across Different Ratings



- Most of the genres are medium rated, followed by the high rating.
- Minimum ratings are quite low in numbers
- Drama is the highest rated genre.
- Documentary, Film-Noir and Western are low in number as seen before

Modeling & Model tuning



Two models were tried on the recommendation system

- Singular Value Decomposition (SVD)
- KNN – K-Nearest Neighbours: An ensemble model

Model Comparison:

Initial SVD RMSE: **0.63**

Initial KNN RMSE: **0.91**

SVD(**0.63):** means RMSE of 0.63 is relatively low, meaning the model **performs well.**

KNN(**0.91)** means an RMSE of 0.91 suggests that predictions are **less accurate** compared to a model with an RMSE of 0.63.

Model Testing

- Collaborative filtering: Used rating to recommend movies
- Content–based Filtering: used movie embeddings (SVD Latent Factors) to recommend movies to a user

Users get different suggestions from these 2 methods. But contenet-based filtering is more precise to customer preference compared to Collaborative Filtering

Conclusion

1. A Singular Value Decomposition (SVD) model performs much better than KNN model for such a recommendation system.

2. A collaborative-filtering method of recommendation gives different results compared to content-Based filtering.

3. A user-specific model should use content-based as they are coser to user's preference. Though it does not automatically mean that different users will have same preferences despite similarity.

Recommendation

1. Modeling with SVD is much better when using recommendation system.

2. Other models should be tried to compare performances.

3. When using KNN, consider the resources to use especially when fine-tuning the model.

4. SVD model is recommended as it is generally better at capturing latent patterns in user-item interactions, leading to better prediction.