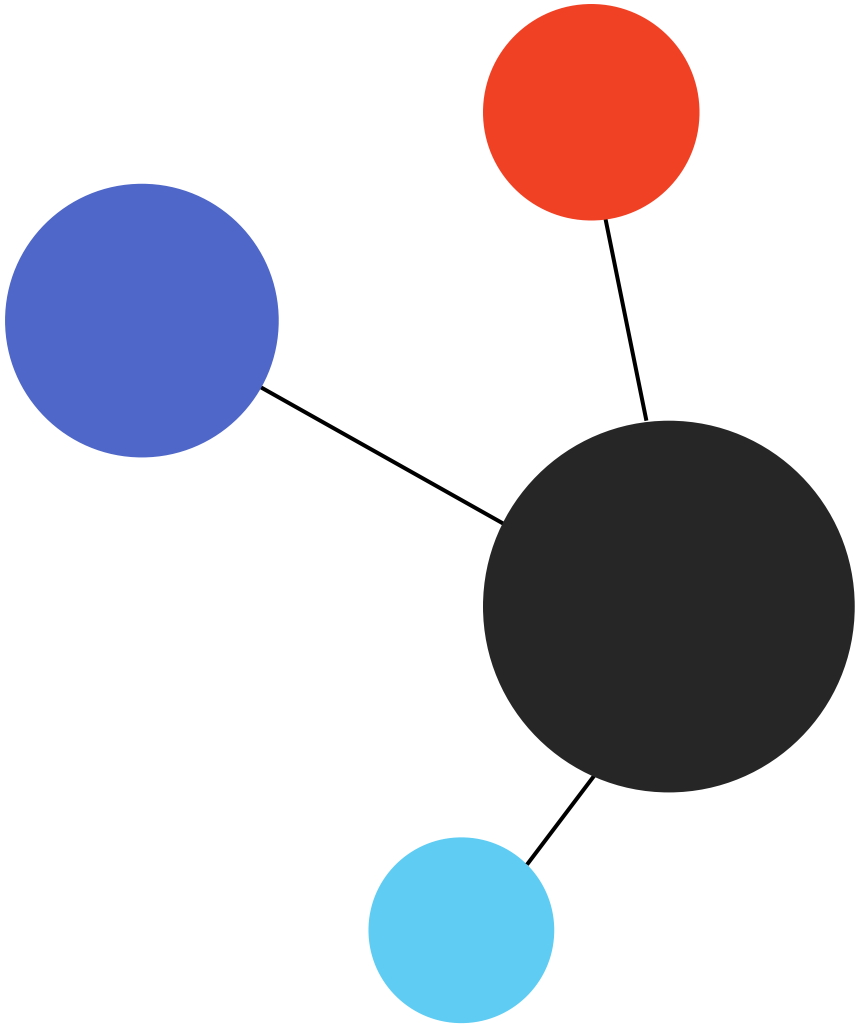
MovieLens Analysis and Prediction

By: Michelle Chekwoti, Joshua Karanja, Myrajoy Kiganane



Business Problem

- This project aims to build a **personalized movie recommendation** system using the MovieLens dataset based on user ratings of other movies.

- The core **business problem** is to improve content discovery and user engagement on a movie streaming platform by offering accurate, data-driven movie recommendations.

Data Understanding and Processing

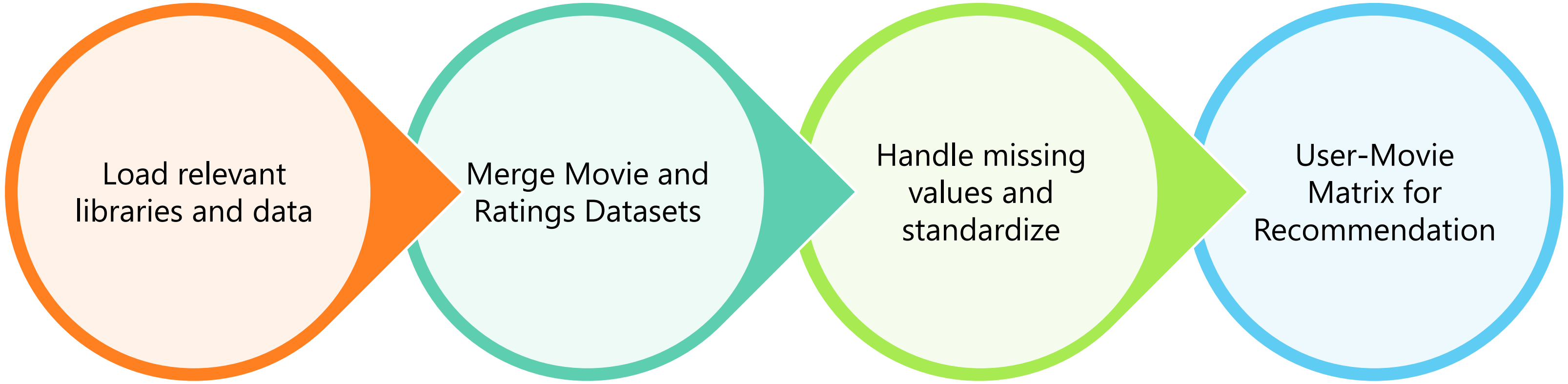
The dataset is from MovieLens. This is a movie recommendation platform that captures user activity in the form of 5-star ratings and free-text tags.

It includes 100,836 ratings and 3,683 tag entries for a total of 9,742 movies.

The dataset is organized into four CSV files; however, we choose to focus on two: ratings.csv and movies.csv

This also involved:

- Loading and merging both datasets Movies and Ratings in order to connect movies and users.
- Handling missing values and standardizing columns to reduce the possibility of nulls and errors in prediction.
- Removing additional columns such as the rating Timestamp, which is not our focus in prediction.
- Ensured a clean dataset for generating user-item interaction matrices used in collaborative filtering



Load relevant libraries and data

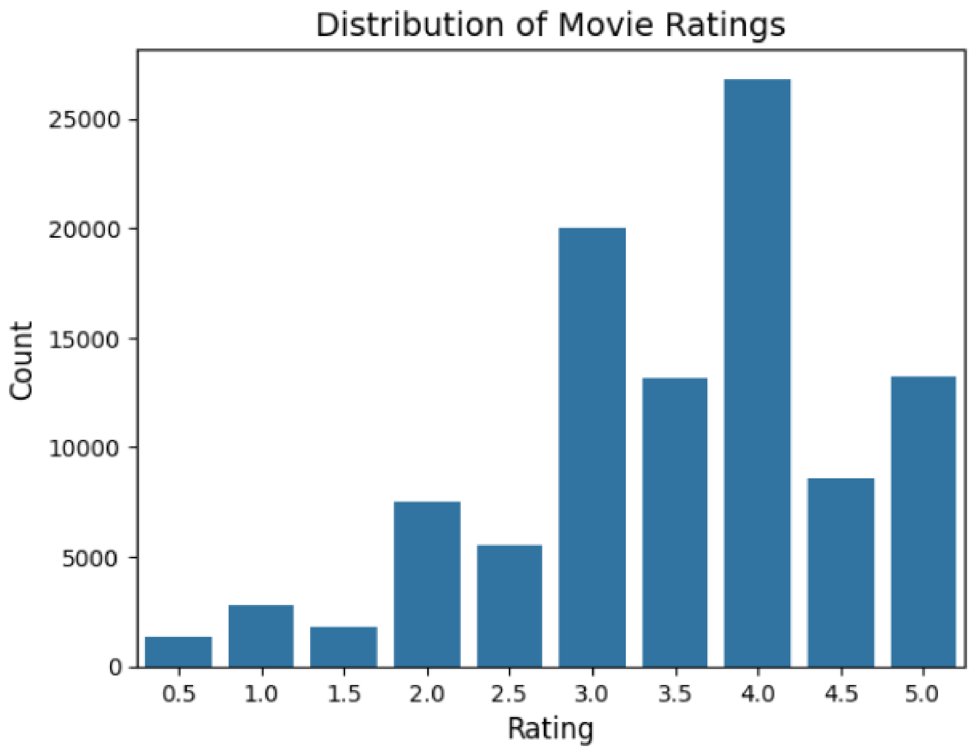
Merge Movie and Ratings Datasets

Handle missing values and standardize

User-Movie Matrix for Recommendation

Exploratory Data Analysis

What is the distribution of movie ratings?



- Mean global rating: 3.5.
- the distribution of the ratings tend to be positive ratings with most lying from 3.0 to 4.0

The top 10 highest rated movies

Title	Count
Forrest Gump (1994)	329
Shawshank Redemption, The (1994)	317
Pulp Fiction (1994)	307
Silence of the Lambs, The (1991)	279
Matrix, The (1999)	278
Star Wars: Episode IV - A New Hope (1977)	251
Jurassic Park (1993)	238
Braveheart (1995)	237
Terminator 2: Judgment Day (1991)	224
Schindler's List (1993)	220

- `Shawshank Redemption`, `The Godfather`, and `The Usual Suspects` are the most highly rated movies.
- `Speed 2: Cruise Control`, `Battlefield Earth`, and `Godzilla` are the worst rated movies

Recommendation Models

Using the **Surprise** library to build the recommendation system, we tested three collaborative filtering models.

- **SVD (Singular Value Decomposition)**

We used GridSearch to tune our SVD model and parameters such as learning rate (how fast the model learns), latent factors (number of hidden features) and more.

- **KNNBasic (k-Nearest Neighbors)**

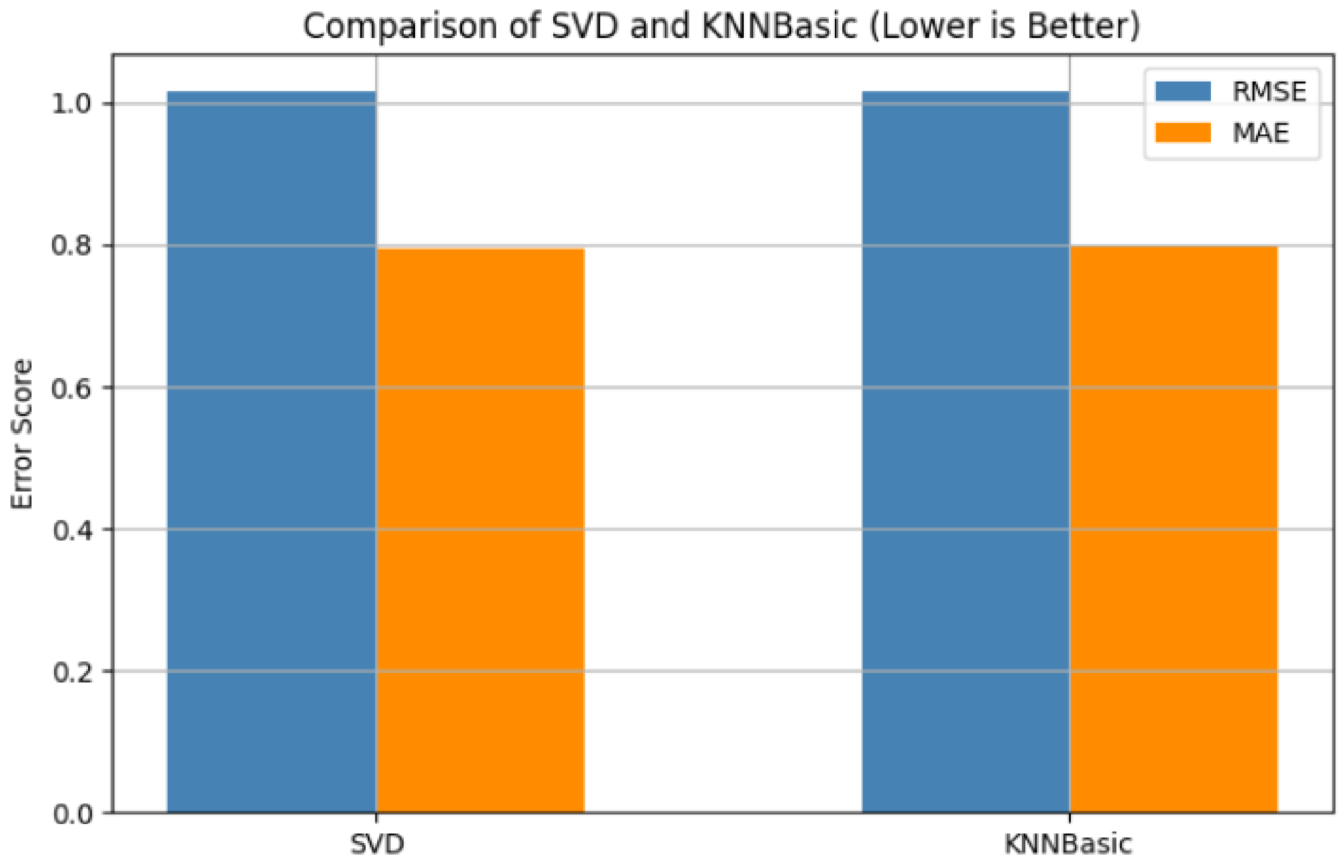
This recommends movies based on similarity and after GridSearch to test

K (number of neighbours), similarity method (Cosine or Pearson), User vrs Items etc.

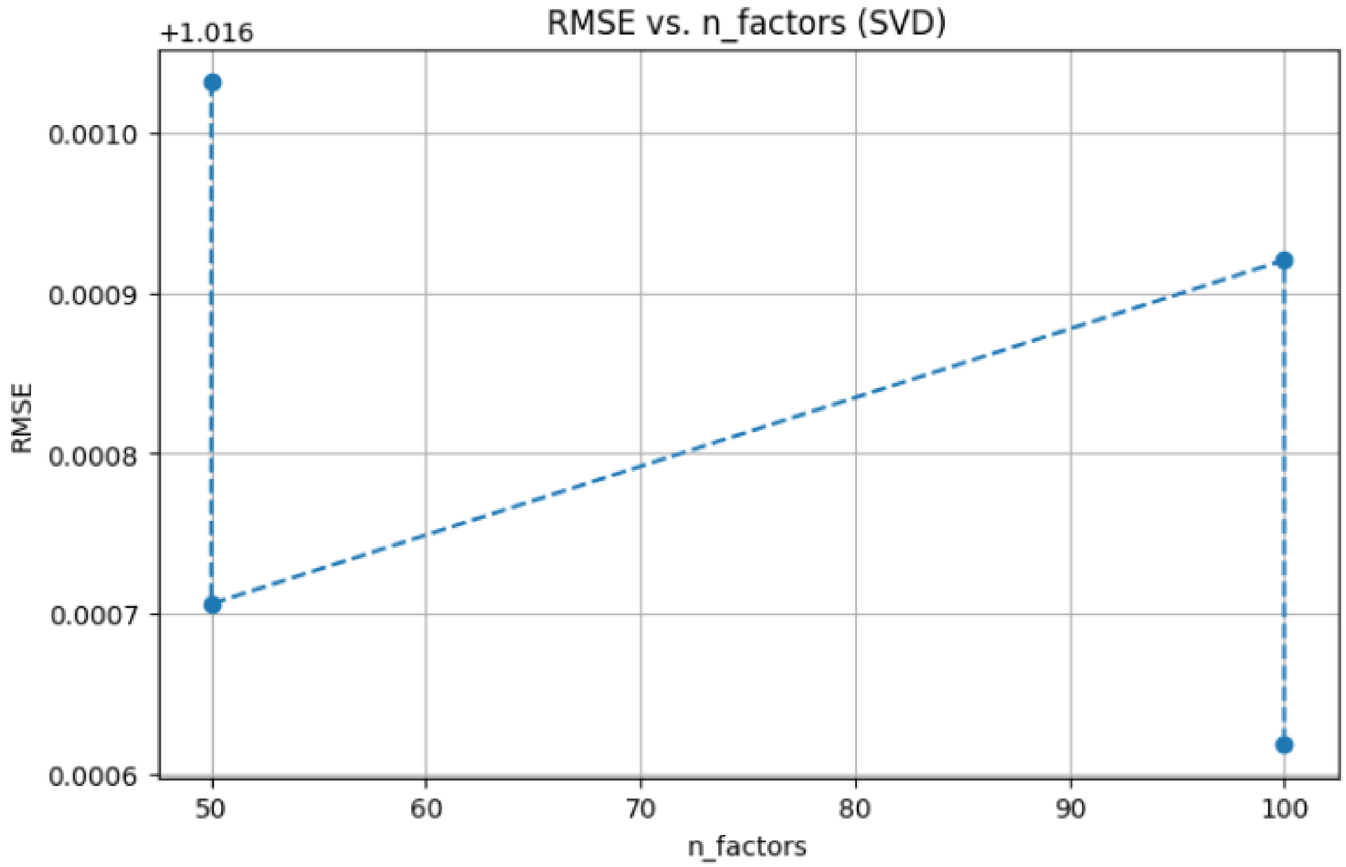
- **KNNBaseline**

This is an advanced version of KNNBasic with a fairer comparison in calculating similarities by taking in Bias.

We used 5 fold cross validation and looked at the RMSE and MAE.



A comparison of SVD and KNNBasic shows that while both models perform similarly, SVD is slightly more accurate overall. It achieved lower error scores on both RMSE and MAE, making it the preferred model for predicting movie ratings.



The graph shows that the SVD model performs best with 50 latent factors, achieving the lowest RMSE. Increasing the number of factors to 100 slightly worsens performance, suggesting that more complexity doesn’t help and may lead to overfitting.

Evaluation

Model	RMSE	MAE
SVD	1.016 (with GridSearch)	0.795 (with GridSearch)
KNNBasic	1.017	0.799
KNNBaseline	1.018	0.797

The SVD model, after tuning, had the lowest RMSE and MAE, both on the test set and in cross-validation. Although KNNBasic and KNNBaseline were close in performance, they did not outperform SVD. Since lower scores mean better predictions, SVD was chosen as the final model.

Recommendations

- Use the SVD model as it outperforms memory-based KNN approaches and learns features like genre preference, hidden user tastes, and rating habits.
- We chose Singular Value Decomposition because it doesn’t just memorize neighbor ratings. Further, by tuning to 50 latent factors, we get strong accuracy without overwhelming computational resources.
- Top Films like Forrest Gump, and Shawshank Redemption should be featured by the streaming platform . They should be unpersonalised recommendations because they are classic high rated hits and boost overall platform engagement which helps onboard new users with guaranteed hits.

Thank you!

Any Questions?

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

By Group 5:

- **Michelle Chekwoti,**
- **Joshua Karanja,**
- **Myrajoy Kiganane**