Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Movie Recommendation System

16/04/2024

Agenda

Problem Statement

Introduction

EDA

Model Recommendation

Conclusion

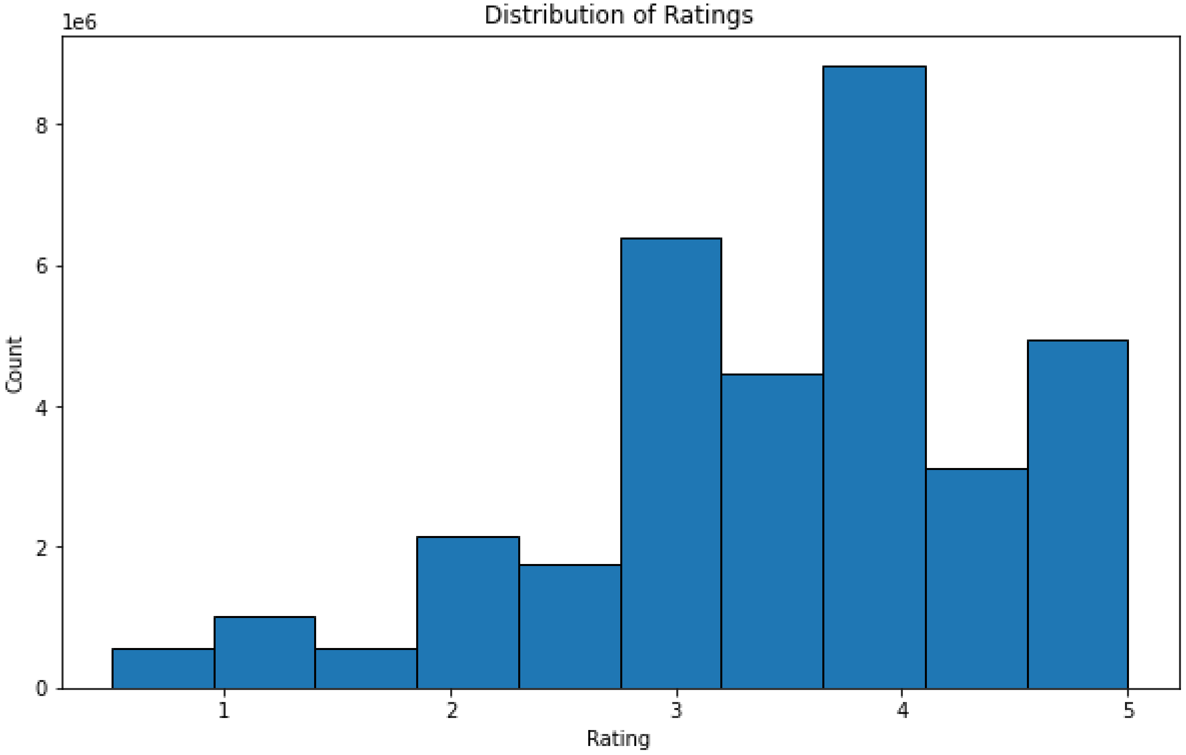
Problem Statement

The video-on-demand streaming service is looking to develop a machine learning algorithm to predict which movies a user will enjoy based on various factors such as genre, online ratings, and previous decisions. The primary objective is to create a system for movie recommendations.

Introduction

The Exploratory Data Analysis (EDA) performed on a movie ratings dataset and recommends a final model for predicting user ratings. The model selected is Matrix Factorization, which is a supervised algorithm. This method was chosen to allow for the evaluation and validation of the model's performance. It is important to note that correlation and multicollinearity analyses are not applicable in this context, given the nature of the data and the modelling approach. Data Overview and Data Description steps were performed earlier.

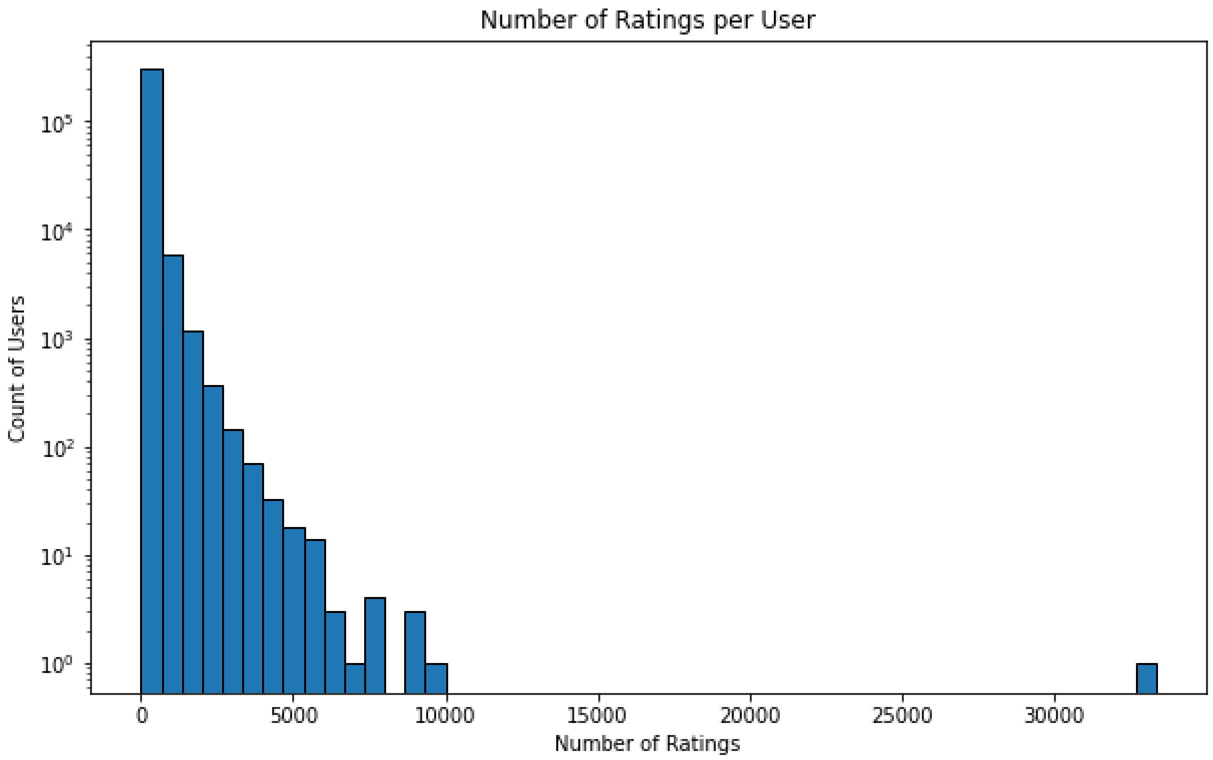
EDA. Distribution of Ratings



The histogram shows the count of each rating value, providing insights into user rating behaviour.

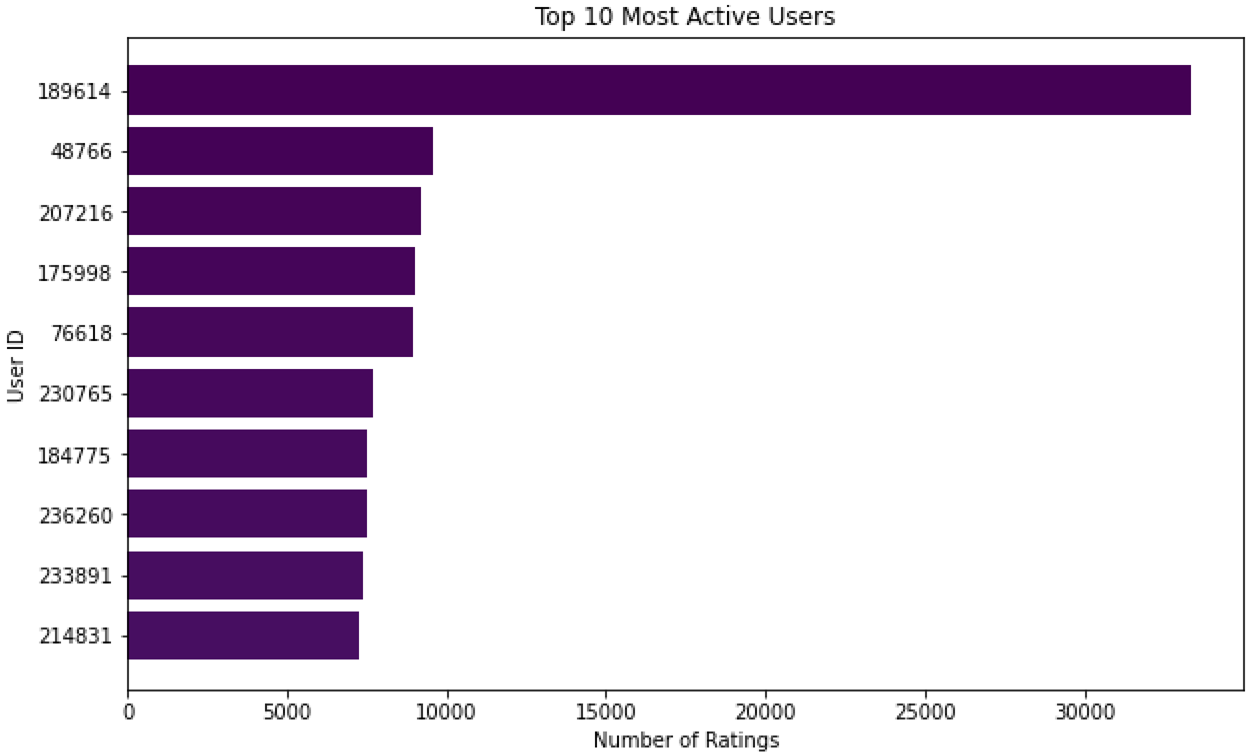
People are more likely to rate the movie using integer number rather than floating point number. 4 is the most popular rating with 3 and 5 on the second and third places respectively by popularity.

EDA. Number of Ratings per User



- 99% of users left less than 1100 ratings.
- An average number of reviews left by a single user 110.

EDA. Most Active Users

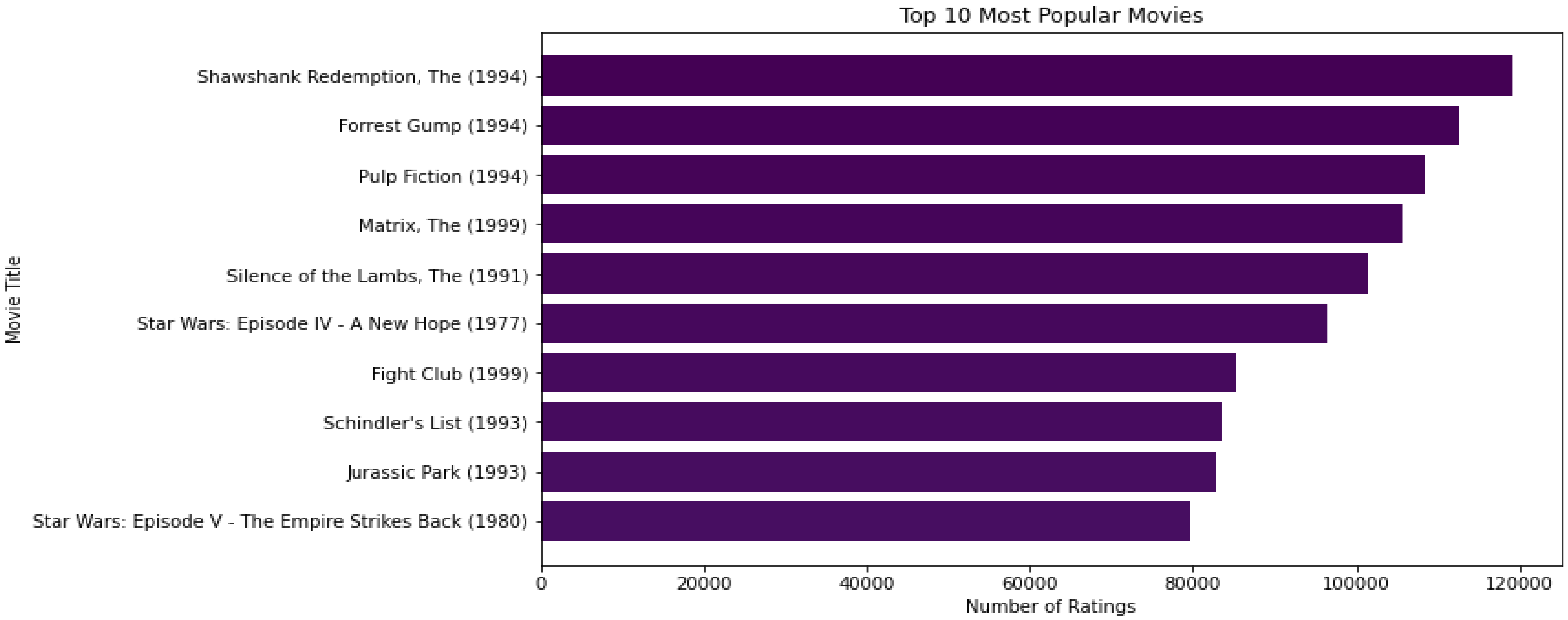


This analysis highlights the users who contribute the most ratings, which could be useful for targeted analysis or recommendations.

We observe a huge gap between the 1st and the 2nd place which is very surprising.

EDA. Most Popular Movies

The most popular movies in our dataset are:



EDA. Why Correlation and Multicollinearity Analysis are Unnecessary

Correlation and multicollinearity analyses are typically used in traditional regression models to understand relationships between variables and avoid redundancy in predictors. However, in the context of matrix factorisation for recommendation systems, these analyses are not applicable because:

- **Nature of Data**: The data structure in recommendation systems (user-item interactions) does not lend itself to traditional correlation or multicollinearity analysis.

- **Model Approach**: Matrix factorization focuses on latent features rather than direct variable interactions, making these analyses irrelevant.

Model Recommendation

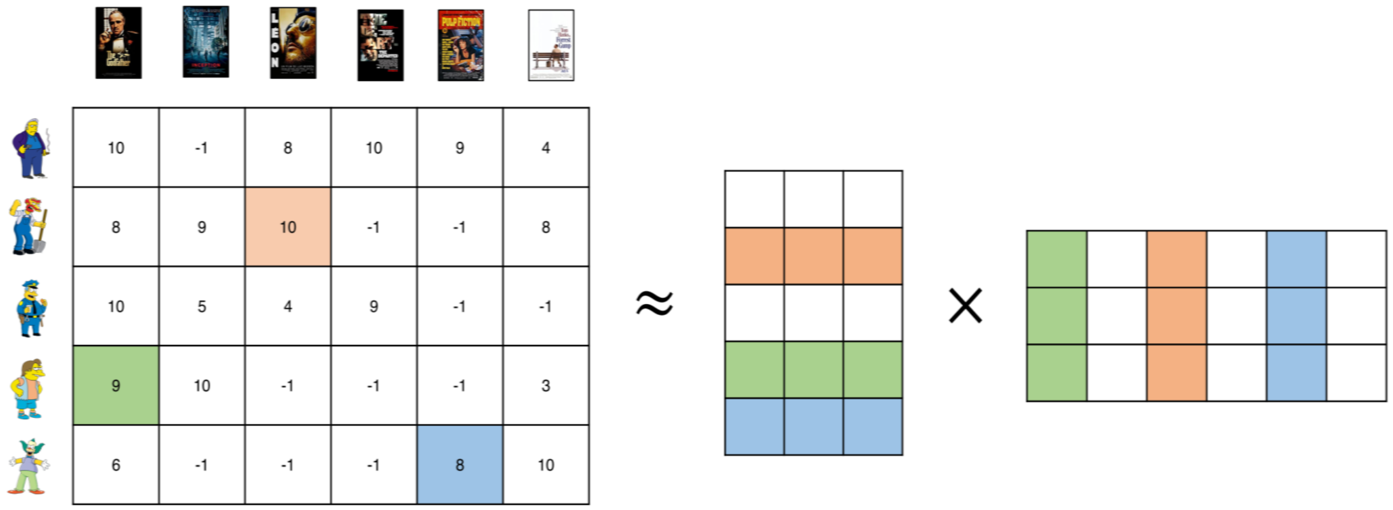
Matrix Factorization is recommended as the final model for predicting user ratings. This technique is suitable for collaborative filtering in recommendation systems. It works by decomposing the user-item interaction matrix into the product of two lower-dimensional matrices, capturing the latent features of users and items.

Advantages

Supervised Learning: Allows for evaluation using metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Scalability: Suitable for large datasets.

Latent Factors: Captures hidden features influencing user preferences.



10	-1	8	10	9	4
8	9	10	-1	-1	8
10	5	4	9	-1	-1
9	10	-1	-1	-1	3
6	-1	-1	-1	8	10

User-item Interaction Matrix (R)

User Matrix (Q)

Item Matrix (P)

Conclusion

The EDA provided valuable insights into user behaviour and movie popularity within the dataset. The final model recommendation, Matrix Factorization, is well-suited for building a recommendation system that can be evaluated using supervised learning techniques. The choice of model and the reasoning for not performing correlation and multicollinearity analyses are based on the specific characteristics of the data and the goals of the recommendation system.

Olena Panchenko
hel.panchenko@gmail.com
Taras Shevchenko National University of Kyiv
Data Science

Data Glacier
Your Deep Learning Partner

Thank You!