Introduction

This project aims to create a personalized movie recommendation system using the MovieLens 100k dataset. This report presents a detailed overview of the project, including data analysis, model implementation, training process, evaluation, and results.

Data analysis

A comprehensive data exploration phase was conducted on the dataset in order to understand the data and its characteristics. During the data exploration phase, a lot of interesting insights were discovered. The following are some of the insights that were discovered during this phase:

1. Drama and Comedy are the most popular genres among the users
2. The age distribution of the users is skewed towards the younger age groups, with the majority of the users being between 20 and 40 years old.
3. The distribution of the ratings is skewed towards the higher ratings, with the majority of the ratings being 4.

Model Implementation

The final recommendation System used is a merge of two types of recommendation systems which are as follow:

1. Content-Based Recommendation: This approach analyzes movie content, such as genre and keywords, to recommend movies similar to a user's past preferences. It involves creating a pivot table, calculating TF-IDF matrices for movies, and using cosine similarity scores to measure content similarity.
2. Demographic-Based Recommendation: User demographics, including age, gender, and occupation, are considered to provide tailored movie recommendations. Users are grouped by demographics, and movies highly rated within their specific demographic group are recommended.

This hybrid approach merges content and demographic-based recommendations, offering users personalized movie suggestions based on their content preferences and demographic characteristics.

Model Advantages and Disadvantages

The hybrid recommendation system employed in this project delivers personalized movie recommendations by considering both user preferences and demographic information. This dual approach enhances user satisfaction by suggesting movies aligned with individual tastes and demographics.

However, the system does have limitations. One notable drawback is the potential for recommendations to lack diversity, primarily due to content-based similarities.

This may result in users receiving suggestions that are too similar in content. Furthermore, the demographic-based recommendation component can introduce biases by reinforcing existing demographic preferences present in the dataset.

Training Process

The recommender system did not involve a traditional training process, as it relied on user ratings and content similarity measures to generate recommendations.

Evaluation

To evaluate the performance of the recommendation system, the RMSE metric was used. The RMSE metrics is a very common metric used to evaluate the performance of recommender systems. It is used to measure the difference between the predicted ratings and the actual ratings. The lower the RMSE value, the better the model performance.

Results

The model developed got an RMSE value of 1.56. The obtained RMSE value reflect the accuracy of the recommendation system. This metric help to assess the system's performance in predicting user preferences.