Movie Recommendation System

Name: Olena Panchenko
Email: hel.panchenko@gmail.com
Country: United Kingdom
College/Company: Taras Shevchenko National University of Kyiv
Specialisation: Data Science
Github repo:
https://github.com/helenapanchenko/DS_Internship_Glacier/tree/master/Week10

Problem Description

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.

EDA

Introduction

This document outlines 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.

Data Visualisation

Distribution of Ratings

The distribution of ratings was visualised using a histogram. This helps understand the commonality of each rating value:

- Rating Scale: Ratings are predominantly between 1 and 5.
- Frequency: The histogram shows the count of each rating value, providing insights into user rating behaviour.

Number of Ratings per User

The distribution of the number of ratings per user was analysed:

- Histogram: A histogram with a logarithmic y-scale was used to visualise the number of ratings per user.
- Quantiles and Mean: The 99th percentile and the mean number of ratings per user were calculated to understand user engagement.

Most Active Users

The top 10 most active users were identified and visualised:

- Bar Chart: A horizontal bar chart was used to display the number of ratings by the most active users.
- User Engagement: This analysis highlights the users who contribute the most ratings, which could be useful for targeted analysis or recommendations.

Most Popular Movies

The top 10 most popular movies based on the number of ratings were identified:

- Movie Titles: By merging the ratings data with a movie titles dataset, the titles of the most popular movies were obtained.
- Bar Chart: A horizontal bar chart was used to visualize the number of ratings for these movies, providing insights into the movies that receive the most attention from users.

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.

Why Correlation and Multicollinearity Analysis are Unnecessary

Correlation analysis and multicollinearity analysis are typically used in traditional regression models to understand relationships between variables and to avoid redundancy in predictors. However, in the context of matrix factorization 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.

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.