Movie Recommendation System

Group name: By a time zone united
Name: Olena Panchenko
Email: hel.panchenko@gmail.com
Country: United Kingdom
College/Company: Taras Shevchenko National University of Kyiv
Specialisation: Data Science
Link: https://github.com/helenapanchenko/DS_Internship_Glacier/tree/master/Week9

Problem Description

The video-on-demand streaming service is looking to develop a machine learning algorithm that can 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.

Data Preprocessing

Data Cleansing

Dealing with Missing Values

We use only genome-scores, movies and ratings files in our project. None of these files contain missing values.

Dealing with Outliers

We examined the ‘relevance’ column from the genome-scores file and the ‘rating’ column from ratings file.
‘relevance’: In our case, outliers for the ‘relevance’ column are values less than 0 and greater than 1. We don’t have such values and thus don’t have outliers in the ‘relevance’ column.
‘rating’: In our case, outliers for the ‘rating’ column are values less than 0.5 and greater than 5. We don’t have such values and thus don’t have outliers in the ‘rating’ column.

Dealing with Duplicates

We have no duplicates in any of our files.

Data Transformation

Dealing with Categorical and Text Data

Only movies file contains categorical and textual data.
‘title’: Column with textual data and has no value for our analysis. But each title contains a year the movie was released which can be useful for our model. Thus we extracted the release year from the ‘title’ column and kept it.
‘genres’: Column with categorical data which we encoded using Multi-label Binarisation.

Dealing with unnecessary columns

1. We dropped the **‘timestamp’** column from the ratings file because it is irrelevant to our analysis.
2. We dropped the '**(no genres listed)**' column from the movies file to avoid repetition(zeros in every genre column already explain this case).

Data Subsetting

For our analysis we use only useId of users who left at least 5 ratings.