Netflix Project

Group Members

1. Marcos Diaz Puerta
2. Fiyinfoluwa Olajide.

Completed Task

The following are completed tasks:

1. Data preprocessing
2. Data cleaning and visualization
3. Cross Validation

Data preprocessing

We transformed the dataset into a standardized Pandas data frame to initiate the cleaning process. Given the large amount of data and the compute resource limitation of Google Colab, we focused on one of the four training data files. The pre-processed data was transferred into another output CSV file.

Data cleaning and visualization

We carried out the data cleaning by checking for missing values, duplicates, and null values. We dropped empty rows if found in the training dataset outputted from the data preprocessing task.
After cleaning the data, we wanted to view the ratings of the users; therefore, we created two histograms using the matplotlib library, which showed how ratings were distributed.

Cross Validation

We discovered a library called Surprise, which is very focused on recommendation systems. We used it to perform cross-validation on our given dataset. RMSE (Root Mean Square Error) and MAE (mean absolute error) were the measures used in the cross-validation process, and SVD (Singular Value Decomposition), NMF (non-negative matrix factorization) and SVDPP (an extension of SVD) were candidates. After the cross-validation, we discovered that SVD was the best performer.

Data

This https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data is the dataset used in the project.

Challenges

1. Google Colab: The free resources from Google Colab were limited and not enough, and neither is it designed to handle large amounts of data. Hence, we had to work on only one of the training datasets.
2. Data: in other to fully utilize the data, we had to organize the movie ID by transposing it from a row to a column field and for the movie title table, we had to merge all the fields with the title into one. We used some programming in Python to achieve it.

Collaboration

In other to meet the project deadline, we meet twice weekly at the health campus. In the group, we are fine collaborating.

Next steps

1. We will be training the algorithm with the data.
2. We will be measuring the predictions and making adjustments where necessary.
3. We will be measuring the accuracy of the algorithms.