CIS 5570 Big Data Final Project
Recommendation Systems on Amazon

Project Goal:

The agenda of the project is to portray how recommendation systems work on the Amazon dataset. For this project, we have chosen to perform two different recommendation systems i.e. Popularity-Based recommendation and Collaborative-Filtering. The output of this project is to show how well recommendation systems can provide ratings for various products by users.

Dataset:

The dataset we have chosen is based on Amazon. The dataset contains the following features:

1. "User_ID": ID representing users in the data.
2. Product_ID: ID representing products in the data.
3. Rating: The rating given for a product by a user.
4. Timestamp: The time recorded when the rating was given.

The data contains a total of 7.82 million ratings. The size of the dataset is approximately 318 MB.

```
+--------------+----------+------+----------+
|       User_ID|Product_ID|Rating| Timestamp|
+--------------+----------+------+----------+
| AKM1MP6P0OYPR|0132793040|   5.0|1365811200|
|A2CX7LUOHB2NDG|0321732944|   5.0|1341100800|
|A2NWSAGRHCP8N5|0439886341|   1.0|1367193600|
|A2WNBOD3WNDNKT|0439886341|   3.0|1374451200|
|A1GI0U4ZRJA8WN|0439886341|   1.0|1334707200|
|A1QGNMC6O1VW39|0511189877|   5.0|1397433600|
|A3J3BRHTDRFJ2G|0511189877|   2.0|1397433600|
|A2TY0BTJOTENPG|0511189877|   5.0|1395878400|
|A34ATBPOK6HCHY|0511189877|   5.0|1395532800|
| A89DO69P0XZ27|0511189877|   5.0|1395446400|
+--------------+----------+------+----------+
only showing top 10 rows
```

Methodology:

The project is executed in Google Colab and certain operations have been performed to achieve the result for both recommendation approaches. They are as follows:

1. Installed Spark and configured the session on Google Colab.
2. Uploaded the dataset using Google Colab built-in functionality.
3. The dataset is in CSV format hence we have read it in the same format and loaded it as a data frame using Spark.
4. Performed a few pre-processing steps on the dataset to maintain the correctness of the data before performing recommendations. The following are the steps:
    a. The data frame was unnamed hence we have renamed the columns for proper identification.
    b. Checked for null values. We figured out there were no null values.
    c. Converted the data from strings to numerical values for modeling and prediction by the algorithms.
    d. Dropped the timestamp column as it is not useful for our design and model.
5. Implemented the Global Baseline Recommendation method for the data frame. The following are the steps performed:
    a. Splitting the total data into train and test data with training data being 70% and testing data being 30% of the complete dataset.
    b. Calculated the total/global mean rating for the whole data.
    c. Calculated the mean rating for every user and product and computed the deviations for each user rating.
    d. Using the global baseline approach implemented the prediction of ratings for the test data.
    e. We chose RMSE as our evaluation metric to evaluate the correctness of the predictions by comparing them with the actual values of the testing data.
6. For our second approach, we have used Item-Item collaborative filtering. To perform this recommendation, we have used the Alternate Least Square (ALS) approach. The following are the steps performed:
    a. Splitting the total data into train and test data with training data being 70% and testing data being 30% of the complete dataset.
    b. Creating an ALS model using the Spark built-in library and setting the required parameters for the model such as maximum iterations, strategy, user, item, and the values i.e. the rating of the products.
    c. We fit the training data into this model. Further predictions have been performed on the testing data.
    d. We chose RMSE as our evaluation metric to evaluate the correctness of the predictions by comparing them with the actual values of the testing data.

e. Retrieved a few users from the data and performed two product recommendations as per the model.

Results and Performance:

Both the recommendation approaches worked very efficiently on the huge data frame of 7.82 million records. The following are the evaluation results for each type of recommendation approach:

1. **Global Baseline Recommendation:**
   a. **RMSE:** 1.42
   b. **Performance:** TBC
2. **Collaborative Filtering Recommendation (ALS Method):**
   a. **RMSE:** 0.19
   b. **Performance:** TBC

Contributions and Responsibilities:

1. Pranitha Velusamy Sundararaj:
   a. Initialized the Google Colab environment with the appropriate setup.
   b. Installed necessary packages and configured the session for code execution.
   c. Involved in checking the correctness of data and pre-processing.
2. Devyani Deore:
   a. Involved in pre-processing of data and converting the data into numeric data for allowing the recommendation models to execute.
   b. Performed Exploratory Data Analysis on the dataset.
   c. Involved in preparing the final report for the project.
3. Sai Sanjith Sivapuram:
   a. Implemented the Global Baseline Algorithm.
   b. Evaluated the model using the Root Mean Square Error approach.
   c. Involved in the final report and presentation of the project.
4. Anand Jha:
   a. Implemented the Collaborative Filtering using the ALS approach for the data.
   b. Evaluated the model using the Root Mean Square Error approach.
   c. Generated two new recommendations for a few users using the ALS model designed.
   d. Involved in the final presentation of the project.