Predict Star Ratings using Amazon Movie Reviews Dataset

A. Problem Description

The goal of this assignment is to predict star rating associated with user reviews from Amazon Movie Reviews using the available features.

You are allowed to use any technique for your predictions, as well as classical machine learning algorithms, like random forests, regression trees, etc. Using deep learning models, or any other related technique, that lies far from the syllabus of this class is prohibited. What we mainly seek from this competition - besides performance- is smart ways to make sense of the data, construct new features from the available metadata, and understand your thought procedure.

In addition to the solution, you need to provide a 2-page writeup that describes the algorithm you have implemented and the special tricks you used in order to make it work (or improve). It is important that you show your thought procedure.

B. Data Description

You have been provided with the following files.
File Descriptions:

- train.csv - 1,697,533 unique reviews from Amazon Movie Reviews, with their associated star ratings and metadata. It is not necessary to use all reviews, or metadata for training. Some reviews will be missing a value in the 'Score' column. That is because, these are the scores you want to predict.
- test.csv - Contains a table with 300,000 unique reviews. The format of the table has two columns; i) 'Id': contains an id that corresponds to a review in train.csv for which you predict a score ii) 'Score': the values for this column are missing since it will include the score predictions. You are required to predict the star ratings of these Id using the metadata in train.csv.
- sample.csv - a sample submission file. The 'Id' field is populated with values from test.csv. Kaggle will only evaluate submission files in this exact same format.

Data Fields:

- ProductId - unique identifier for the product
- UserId - unique identifier for the user
- HelpfulnessNumerator - number of users who found the review helpful
- HelpfulnessDenominator - number of users who indicated whether they found the review helpful
- Score - rating between 1 and 5
- Time - timestamp for the review
- Summary - brief summary of the review
- Text - text of the review
- Id - a unique identifier associated with a review

Note: Some of the rows of the table may have some of these values missing.

Dataset Citation

J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013

Data Download:
https://www.kaggle.com/c/bu-cs506-spring-2020-midterm/data