CMPE256 - S23 - Book recommendations

Vinit Kanani

016651323

Kaggle: [@iamvinitk](https://www.kaggle.com/iamvinitk)

1. Scoring - RMSE
   1. SVD - 1.56784
   2. XGB - 1.64211
2. Data Preprocessing:
   1. Books Data

The books data had issues with CSV parsing due to escape characters. I used preprocessing techniques to format the data correctly by cleaning up the escape characters and converting the data into a more readable format. This involved removing unnecessary columns and creating new columns that would be more useful in the prediction process.

* 1. Users Data

The user’s data had inconsistencies in the format of the location attribute. To address this, I extracted the country from the location attribute and added a new column for the region. Additionally, there were around 40% missing values for the age attribute, so I replaced these values with the average age of the country.

* 1. Train Data

The training data had some ratings above 10, which were replaced with 10 to ensure consistency and fairness in the prediction process.

1. Algorithms
   1. **SVD**

SVD (Singular Value Decomposition) is a matrix factorization technique commonly used for collaborative filtering in recommender systems. It decomposes a large user-item rating matrix into three smaller matrices: U, S, and V. The U matrix represents user preferences, the S matrix contains the singular values, and the V matrix represents item attributes.   
I used *GridSearchCV* to find the optimal values for the hyperparameters of the SVD model, which included the number of epochs, the learning rate, and the regularization term.

* 1. **XGB**

XGB (Extreme Gradient Boosting) is a powerful algorithm used for both classification and regression problems. It is based on decision trees and is known for its speed and accuracy. The *MinMax* Scaler transformed the data so that the values were between 0 and 1, which helped to improve the performance of the model. *Label encoding* was used to transform the categorical features into numerical values, which allowed the XGB model to use them in its calculations. I used XGB for regression in this project and tuned the hyperparameters using *GridSearchCV*. The hyperparameters included the maximum depth of the tree, the learning rate, and the number of estimators. The best parameters were *learning\_rate=01, max\_depth=7, n\_estimators=1000.*

1. Reference
   1. [https://www.kaggle.com](https://www.kaggle.com/)
   2. [https://github.com](https://github.com/)
   3. [https://scikit-learn.org](https://scikit-learn.org/)
   4. [https://www.tensorflow.org](https://www.tensorflow.org/)
   5. <https://towardsdatascience.com/>
   6. <https://xgboost.readthedocs.io/en/latest>