TITLE: PROJECT #3 – RECOMMENDER SYSTEMS

AUTHORS: KASSANDRA SELLERS, CLARK NECCIAI

GROUP 46 – PREDICTIVE ANALYTICS

DATE: JUNE 24TH, 2024

**Executive Summary**

For this analysis, we were tasked with evaluating the capabilities of various Recommendation Systems such as user-based collaborative filtering (UCBF), item-based collaborative filtering (IBCF), and association analysis. We evaluated a variety of UCBF and IBCF models in recommending book suggestions from data found in the Goodreads Dataset. Ultimately, the IBCF model which used pearson correlation and the twenty-five (25) most similar items yielded the best overall performance. Utilizing association analysis, we detail three rules describing book sets which tend to be frequently purchased together. We conclude with a brief discussion on UBCF, IBCF, and association rules where we detail their differences, applications, implementation challenges, and our team’s preferred approach for this dataset.

**Problem Statement and Approach**

Our primary tasks in this analysis were fourfold:

* Perform exploratory analysis and provide notable insights on users, reviews and books
* Preprocess our Goodreads dataset for running the recommender system models
* Determine the best UBCF and IBCF models and compare their prediction overlap
* Generate three association rules detailing subsets of frequently read books

We began our approach by first evaluating for data missingness and duplicated observations. After we performed data preprocessing over our original dataset, we explored our remaining data and discovered notable aspects of the Goodreads datasets. We then appropriately formatted and evaluated our data for Recommender System modeling. After assessing multiple UBCF and IBCF models, we compared our best performing model’s top five predicted books. Lastly, we generated association rules describing groups of books that tend to be read together.

**Data Preprocessing**

The Goodreads dataset consists of two subsets of data containing information on 10,000 books (Books dataset) and their corresponding 981,756 ratings (Ratings dataset). We evaluated the magnitude of missing values for variables. The variables and their respective percentage of missing values are as follows: *isbn (7.0%), original\_publication\_year (0.21%), original\_title (5.85%), language\_code (10.84%), mths\_since\_last\_record (90.88%), mths\_since\_recent\_inq (9.98%).*  Given that these variables will not be used in modeling, nor a reliable means to infer an imputable value readily available, we reasonably concluded that it would be safe to ignore these missing values.

We then proceeded to remove duplicate books from the Books dataset based on their title, finding that some observations had been duplicated two, three, and even four times over. In order to ensure no duplicates remained in the Books dataset, a total of 36 repeat observations had to be removed. Similarly, we removed 2,278 duplicate ratings from the same users within the Ratings dataset which was done by grouping *user\_id* and *book\_id* and then performing duplication removal. Within the Ratings dataset, a total of 3,511 observations which had no reference within the Books dataset were also removed. Lastly, as a means of attaining computational feasibility, we retained only those observations which belonged to users with at least 100 ratings, bringing our total number of ratings down to 164,733. For the remaining observations in both datasets, we examined summary statistics, finding no variables which contained illogical values.

**Exploratory Data Analysis (EDA)**

Given the unsupervised nature of our analysis, it is paramount to engage in a rigorous discovery phrase. Doing so grants us insight into deeper understanding of the inherent relationships in our dataset which may contain useful, explanatory information. During our results generation in the post-modeling phase, we may be able to draw from these insights and reason as to how we arrived at these outcomes. We began our exploratory analysis after taking the aforementioned preprocessing steps.

***Oldest Recorded Books***-In an effort to determine the validity of our negative values of *original\_publication\_year,* we recognized these values as representing texts written before the common era (B.C.E) due to prior knowledge. A few of these books include *The Epic of Gilgamesh, The Illiad, The I Ching or Book of Changes* and *The Odyssey*. Of all books written during this era, the author who had written the most books was Plato with a total of five books.

***Distribution of Ratings******-***Our team speculated that there may be a relationship between the number of ratings a book receives and the quality of that book. We examined the distributions of *ratings\_1, ratings\_2, ratings\_3, ratings\_4,* and *ratings\_5*, finding that indeed, we see from our variables’ distributions that books that had ratings of 1 or 2 had very few ratings. We speculated this to mean that as individuals spread word that a book is poorly written or low quality, that other people tend to not want to read and rate these books, leading to the distribution we see in the chart. For books that had ratings of 3, 4, or 5, their distributions appeared similarly flat but spanning into higher values of the number of ratings. We take this to mean the antithesis; as books are found to be of higher quality, more and more individuals become aware and wish to read and provide ratings.

***Highest Rated Books******-***

***Highest Rated Authors******-***

**UBCF and IBCF Modeling**

**Association Rule Modeling**

**Conclusion**

**Appendix**