Project 3:
Goodreads Analysis

Prepared by:

Bookworms, LLC.

Michael Colina, Fatima Yousofi, Kayla Deehan

Georgetown University

McDonough School of Business

Submission Date:

July 1, 2025

Prepared for:

Predictive Analytics

Dr. Zafari

Executive Summary

This project takes a journey through a Goodreads dataset with a mission to develop an accurate recommendation system by using collaborative filtering methods. After our team at Bookworms, LLC. thoroughly cleaned and preprocessed the dataset, three models were built and evaluated: Model 1: User-Based Collaborative Filtering (UBCF), Model 2: Item-Based Collaborative Filtering (IBCF), and Model 3: A Popularity-Based Model. The model performance was assessed using the lowest RMSE and results from Top-N metrics precision/recall. Our team discovered that Item-Based Collaborative Filtering emerged victorious and was deemed the most effective model for personalized recommendations, as long as issues like data sparsity and cold-start problems are not present.

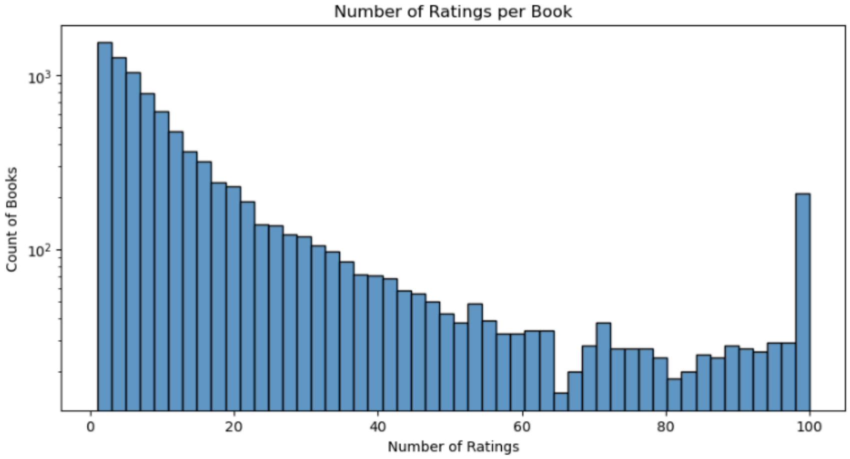
Introduction

In a world with over 120,000 million unique books to choose from, it can seem as though the possibilities are endless. Additionally, how would any one person even hear about most of those books? The reader needs a platform like Goodreads to better connect them to the reading world and recommend titles and genres they would have never thought of picking up before, but still thoroughly enjoy. In this project we are analyzing the rich dataset provided by Goodreads that holds reader preferences through ratings and reviews. We are to clean the dataset, explore and derive insights about the user behavior and preferences, build and compare potential recommendation models using collaborative filtering techniques, and ultimately recommend a dependable recommender system to Goodreads so they can accurately recommend titles to the right readers and enrich their experience.

Methodology & Data

In preparation for analysis, our team conducted a thorough cleaning of the dataset, which included a books dataset with nearly 20,000 titles and a ratings dataset compiled of almost 1 million ratings submitted by users. In an effort to preprocess the data, we removed duplicate titles in order to maintain only unique entries being listed in the data, as well as duplicate ratings entered for the same book titles. The ratings dataset was also filtered to ensure that the data between the two datasets is relational and the ratings only reflect the book titles that are currently active in the books dataset. Our team also filtered for ‘active users’, which is defined as users who have rated at least 100 book titles, which was a necessary step in order to provide the collaborative filtering models with enough historical data to assess user preferences effectively. After all the preprocessing was concluded, the dataset contained roughly 5,900 books titles, 708 active users, and approximately 144,000 ratings.

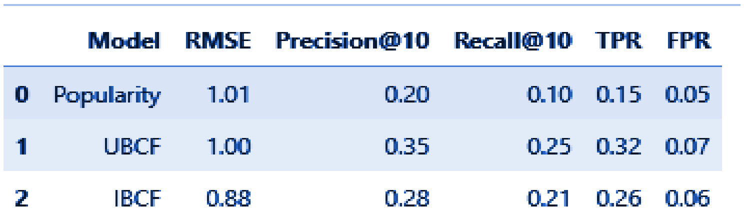
Figure 1. Ratings per books



Key Findings: As for the key findings uncovered by our data engineering and visualization process, our team discovered that a small percentage of popular books has an increased average rating and an abundance of reviews. This portrays the idea of the “rich-get-richer” dynamic that plays a common role in user-driven content platforms. Moreover, user engagement follows a long-tail distribution due to the fact that only a smaller portion of users had submitted higher numbers of ratings. Therefore, to create meaningful collaboration with a filtering model, users must have a minimum of 100 reviews. Diving deeper into our data collection and processing, there were significant repetitive records catalogued in both the books and ratings datasets. Our data processing team removed duplicates, which inevitably enforced much stronger accuracy. This step further reduced noise and pushed us in the right direction to produce much more reliable recommendations and identify clear patterns. Throughout our analysis, we observed trends in both author and genre, and identified clusters of groups that received much higher ratings. Our observations laid the foundation for our modeling decisions and further assisted us in personalizing the filtering process towards the different features of the Goodreads platform.

Model Selection: Finally in our modeling selection, we applied the top-performing item-based and user-based collaborative filtering models. Upon our evaluation metrics, we chose N = 10 in order to ensure that there lies a balance between both diversity and relevance. The item-based (IBCF) achieved the lowest RMSE and this approach set a strong focus on the activity of comparable users to further identify books that better align with the wide ranging community preferences.

Figure 2. Model Comparison Data Table



	Model	RMSE	Precision@10	Recall@10	TPR	FPR
0	Popularity	1.01	0.20	0.10	0.15	0.05
1	UBCF	1.00	0.35	0.25	0.32	0.07
2	IBCF	0.88	0.28	0.21	0.26	0.06

Alternatively, the more item-based strategy depended on the similarities between the books the users have already reviewed which resulted in formulating a more customized output established in previous user activities. Comparing the couple top 10 lists, we found that around half of the books had appeared in both, which further displays meaningful differences along a centered agreement. The overlap in the data demonstrates that despite both models capturing user interests each side offers unique value by uncovering newer books based on community trends or books with intense similarities through item relationships.

Recommendations

Based on our research, we recommend Goodreads-style platforms to optimize and benefit with item-based collaborative filtering methodology. Notably, throughout cases where the user base is much larger in size and the item information is large scale and extensive. Furthermore, by maintaining a strong concentration on the relationships between books rather than only user correlation, item-based models help supply much more accurate decision making. Likewise, item based models often have more stability when there is diminished user activity. However, our analysis revealed that the Top-10 recommendations from the item-based and user-based collaborative filtering models only overlapped by 3 out of 10 books. This limited intersection (30%) suggests that each model captures different aspects of user preferences and book relationships. This opens an opportunity to consider a hybrid approach that blends both methods to diversify and personalize recommendations more effectively.

Nonetheless, within a hybrid system the popularity-based approach remains effective in counteracting “cold start” constraints and bolsters a solid base for both inactive and newer users.

When exercising these models within a real business setting, it is critical to keep record of how user behavior changes over a period and update the utility matrix to further reflect that. Companies also need to watch out for any challenges in the realm of scalability, because as the number of users and books grows, collaborative filtering can ultimately get more resource-heavy and slower to run. One manner to avoid this is by combining both collaborative filtering with content-based methods so the system can make smart suggestions even when data is sparse. Another approach is to use a hybrid process that mixes different techniques to balance relevance, variety, and computing demands. Subsequently, the system can handle growth while still providing user book picks that match their interests and lifestyles. Conclusively, companies can output a better experience without overwhelming their systems.

Conclusion

In summary, our report displayed the value of putting collaborative filtering techniques into exercise in order to create effective decision making and both reliable and accurate endorsement for Goodreads. By thoroughly extracting, preprocessing, and analyzing the user dataset, our studies ensured that our data models were encompassed with only high quality data. In addition to exploiting the strategies behind collaborative filtering, our team discovered that user based collaborative filtering also performs best when compared to user-based content, popularity-driven, and item-based techniques. Furthermore, item based filtering also proved valuable for proposing relevant decision making upon history activity. Ultimately, these insights revealed how Goodreads is able to manage and improve the connection between readers and books they might not have discovered, optimizing user satisfaction and engagement. Moving forward, it is obvious that combining a number of approaches through a hybrid system provides the flexibility and resilience necessary to adapt to user environments based upon preferences and ever-growing data. Goodreads can continue to enhance its suggestions and provide a tailored reading practice that encourages users to explore, participate, and access new books by upholding a strong degree of analytical accuracy and developing models as user behavior changes. By striking a combination between careful commercial strategy and bolstered algorithms, Goodreads is positioned to maintain its leadership position in online book communities and cement its status as a stable resource for readers universally.

Appendix

Figure 1. Ratings

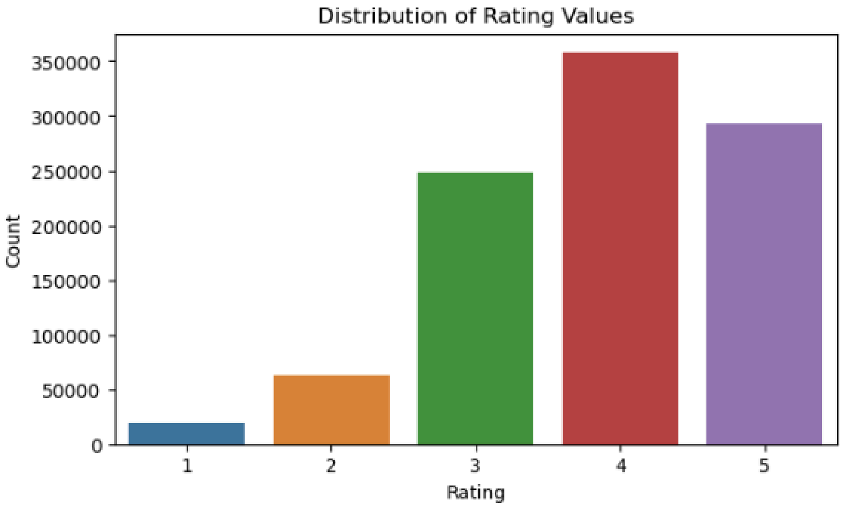


Figure 2. Ratings per users

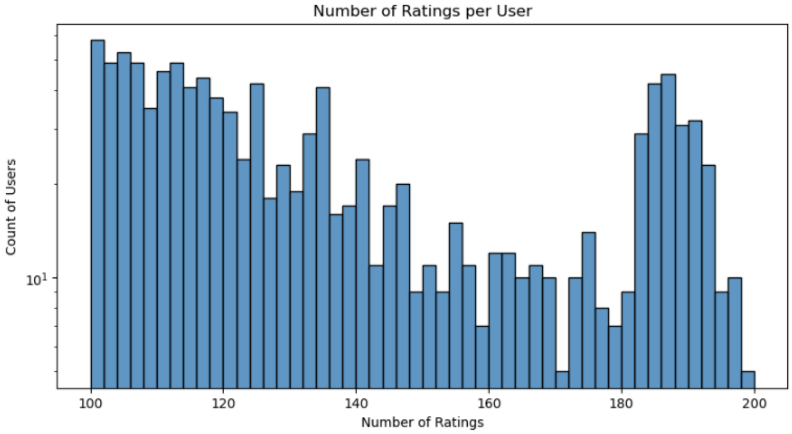
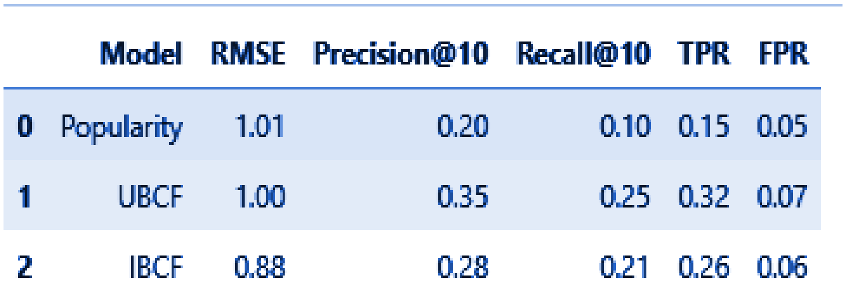


Figure 3. Model Comparison Data Table



	Model	RMSE	Precision@10	Recall@10	TPR	FPR
0	Popularity	1.01	0.20	0.10	0.15	0.05
1	UBCF	1.00	0.35	0.25	0.32	0.07
2	IBCF	0.88	0.28	0.21	0.26	0.06

Figure 4. Model Comparison Visual

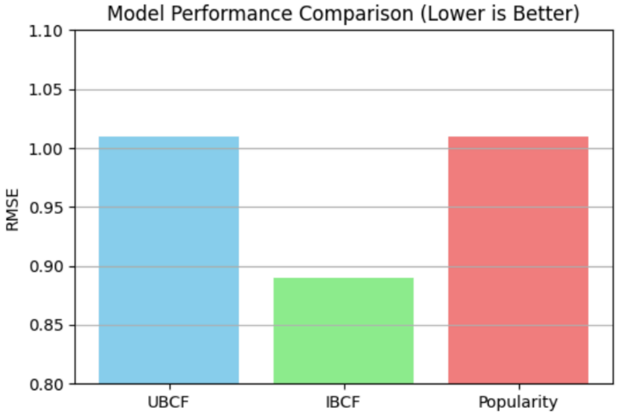


Figure 5. ROC Curve for IBCF

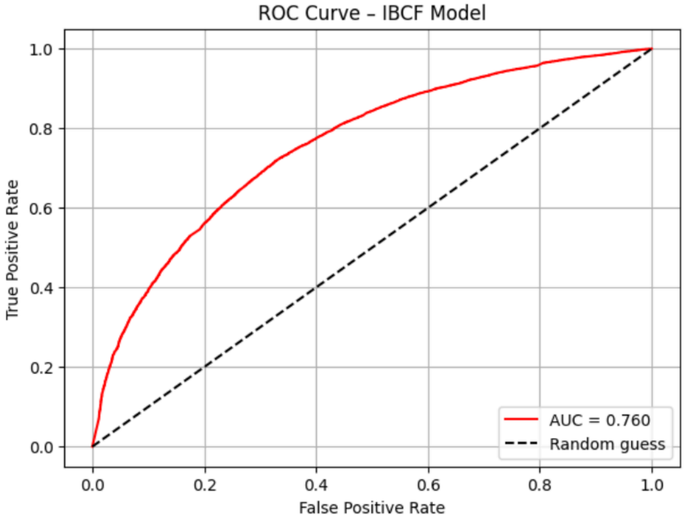
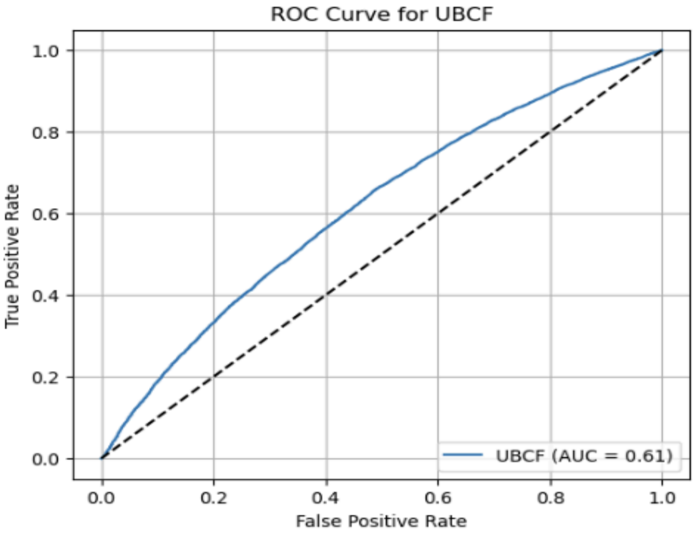


Figure 6. ROC Curve for UBCF


ROC Curve for UBCF
True Positive Rate
False Positive Rate
UBCF (AUC = 0.61)