Exploring Book Recommendations on Goodreads: Insights, Models, and Comparative Analysis

Group 40

Braden Donayre & Paul Sweda

June 24, 2024

1. Abstract

This paper explores the Goodreads Dataset to develop book recommendation models using user-based and item-based collaborative filtering. It aims to understand user behavior, enhance book preference insights, and evaluate recommendation strategies. The dataset undergoes rigorous preprocessing, removing duplicates and filtering ratings based on user activity.

A utility matrix is constructed from the filtered data to evaluate and compare model performance in predicting user preferences. Association rules analysis using the utility matrix identifies top book associations, offering deeper insights into user preferences and reading habits.

The paper discusses user-based and item-based collaborative filtering, alongside association rules, highlighting their business applications and challenges. These models aim to improve user experience and drive business value through personalized recommendations, addressing challenges like data sparsity and scalability specific to the Goodreads Dataset.

2. Introduction

In today's digital age, being able to sift through vast amounts of data collected daily, to understand user preferences and provide real time personalized recommendations has become pivotal for online platforms. Among these platforms, Goodreads stands out as a prominent community driven platform dedicated to readers and book enthusiasts. With millions of users and a comprehensive database of books, reviews, and ratings, Goodreads offers a perfect scenario for exploring the dynamics of reader behavior and developing an effective recommendation system.

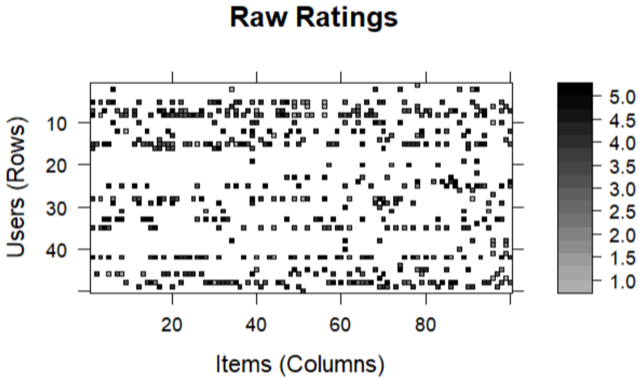
3. Exploratory Data Analysis & Construction of Utility Matrix

Exploring the dataset is crucial for gaining insights into the users behaviors, review patterns, and the books themselves. Understanding these patterns is essential in unsupervised learning projects, where the goal is to uncover hidden relationships. Through exploratory analysis, the top 10 books were identified based on average rating and total number of ratings *(ratings_count)*, highlighting popular and well-received titles among users.Similarly, the top 10 authors were identified based on the average rating of their books. This analysis helps shed light on authors whose works consistently receive high praise from readers, emphasizing their influence and reputation within the dataset.

A scatter plot with a linear trend line was also generated to help visualize the relationship between a book's average rating and its total number of reviews *(ratings_count)*. This exploration helps in identifying any trends or correlations between how well a book is rated and how frequently it is reviewed, offering additional insights into user behavior and book popularity.

To ensure the reliability and consistency of our analyses, rigorous preprocessing steps are undertaken. This includes removing duplicate records from the Books dataset and filtering out redundant reviews from the Ratings dataset. Only ratings associated with books listed in the Books dataset are retained, setting the stage for constructing an accurate utility matrix.

To facilitate recommendation systems and deeper analysis, a utility matrix was constructed where users are rows, books are columns, and ratings are the entries. This matrix, transformed into a ‘realRatingMatrix’ format compatible with recommender systems, serves as a foundational data structure for later modeling and algorithmic approaches. The rating matrix was further analyzed to understand its structure and content. A visual inspection was carried out to assess how ratings are distributed among users and books.



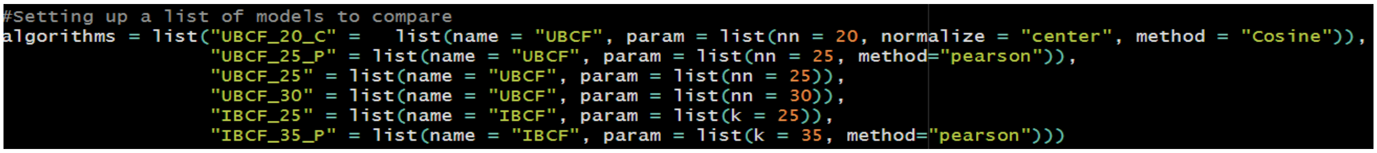
4. UBCF & IBCF Analysis

User-based Collaborative Filtering (UCF) is a recommendation technique that searches for similarities in a target user’s tastes, and offers suggestions based on those similarities with other users. The model assumes that users who have liked related things in the past will continue to like similar things in the future. Based on this assumption, the model identifies other users with similar preferences and will recommend items (in this case, books) to a target user that other similar users enjoyed.

This method is intuitive and straightforward to implement. However, it can suffer from scalability issues as the number of users and items grows. Despite this challenge, UCF remains popular in applications where user preferences are stable and well-defined, such as movie recommendations on streaming platforms like Netflix or music recommendations on Spotify.

Item-based Collaborative Filtering looks for similar items based on the items users have already liked or positively interacted with based on the features of the items that the user shower preference for. By using Item-based Collaborative Filtering recommendation methods we attempt to make recommendations based on the characteristics of each item rather than using other users that have similar preferences.

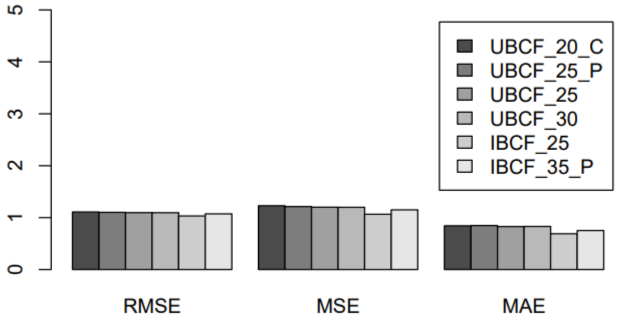
In this report, we explored 6 different models to compare their performance and get a better idea of which method would provide the best possible recommendations for Goodread’s user base. Specifically, we compared 4 different UBCF methods with different parameters and 2 different ICBF methods with different parameters. Before settling on the 6 models being presented in this report, we attempted many different parameters to identify the most interesting ones. For the UBCF we evaluated models with the following parameters:



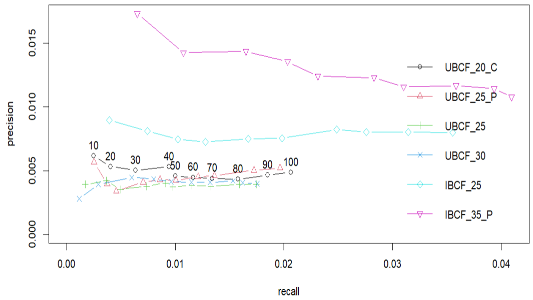
```
#Setting up a list of models to compare
algorithms = list("UBCF_20_C" =   list(name = "UBCF", param = list(nn = 20, normalize = "center", method = "Cosine")),
                  "UBCF_25_P" = list(name = "UBCF", param = list(nn = 25, method="pearson")),
                  "UBCF_25" = list(name = "UBCF", param = list(nn = 25)),
                  "UBCF_30" = list(name = "UBCF", param = list(nn = 30)),
                  "IBCF_25" = list(name = "IBCF", param = list(k = 25)),
                  "IBCF_35_P" = list(name = "IBCF", param = list(k = 35, method="pearson")))
```

Next, we evaluated the different recommendation techniques we tried by using an evaluation scheme we set-up. For the initial evaluation, we elected to use ratings and a train and test split method. Although the cross validation method may be more accurate for evaluating recommendation techniques, we used the 'split' method because it is less computationally demanding when evaluating multiple models. We used 85% of the data to train and 15% to test the model for evaluation. We set the 'given' parameter equal to 100, because we already filtered the dataset to only include users with at least 100 ratings. Lastly, we defined a good rating as a book with a rating of 4 or higher because when exploring the data we found that the average rating was about 3.8. This means good ratings are defined by being above the average for the data set.

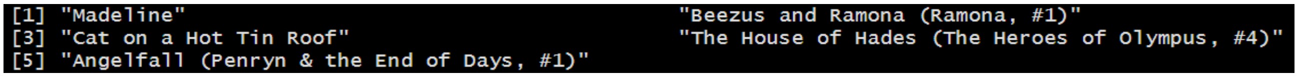
Using the evaluation criteria explained above, we compared the 6 models we created and found that the Item-based content filtering with 25 nearest neighbors had the lowest value for RMSE, MSE, and MAE across the board for all the models we evaluated. The user based content filtering with 30 nearest neighbors performed best of the UBCF models. You can see the evaluation metrics for all 6 models plotted to the right.



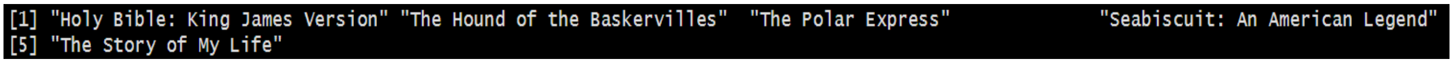
Next, we evaluated the different recommendation system methods we tried by using the same criteria defined above, but using the top 10-100 (with increments of 10) recommended books instead of the predicted ratings. Below we can see the precision vs recall graph for the 6 models that we evaluated. Looking at this graph we can see that as you increase the threshold (i.e., recommend more items), recall tends to improve because more relevant items are included. However, higher recall often comes at the cost of lower precision. Recommending more items may introduce some irrelevant ones.



Finally, we looked at the top 5 recommendations for the first user in the dataset using both the best UBCF and IBCF models. For the best UBCF model, the first user received these top 5 recommendations:


```
[1] "Madeline"                                 "Beezus and Ramona (Ramona, #1)"
[3] "Cat on a Hot Tin Roof"                    "The House of Hades (The Heroes of Olympus, #4)"
[5] "Angelfall (Penryn & the End of Days, #1)"
```

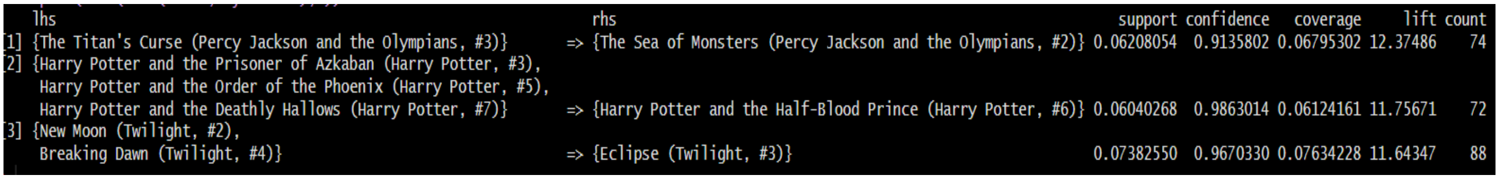
For the best IBCF model, the first user received these top 5 recommendations:


```
[1] "Holy Bible: King James Version" "The Hound of the Baskervilles"  "The Polar Express"            "Seabiscuit: An American Legend"
[5] "The Story of My Life"
```

The team found that there is no overlap between the top 5 recommendations for the first user between the two model types. The UBCF model recommended "Madelin" as the top recommendation based on other users with similar interest. The IBCF model recommended "Holy Bible: King James Version" based on the similarity in the content the first user has shown a preference for.

5. Association Rules Analysis

Association Rules are the relationships between items based on transactional data. It is useful for identifying frequently co-occurring items but may produce numerous rules that require interpretation and filtering. Using the Apropri Code, we were able to identify the top 3 most useful rules for association between books. In order to identify these important association rules in the matrix, we plugged in a few different support and confidence values. Using the support value of 0.06 and 0.90 we found the most interesting results. A support value of 0.06 may seem low, but we believe this is a good threshold for the support value because the application of these rules may be limited, as not all users read and rated all of the books. A lower support value limit would allow for rules that are not as common in the data set and may not be as valuable. We used 0.90 for the lower limit for confidence level because it indicates a strong association between the antecedent and the consequent without eliminating too many of the rules. Lastly, we identified the 3 most important rules by using the Lift metric. Using the lift metric is the best metric to identify the most impactful rules because it indicates that the association between the antecedent and consequent is significant and not just random. Lift helps us identify the rules that have the strongest predictive power. When using this criteria we identified the following rules as the top 3 most important rules:



```
    lhs                                                          rhs                                                          support confidence   coverage     lift count
[1] {The Titan's Curse (Percy Jackson and the Olympians, #3)}    => {The Sea of Monsters (Percy Jackson and the Olympians, #2)} 0.06208054  0.9135802 0.06795302 12.37486    74
[2] {Harry Potter and the Prisoner of Azkaban (Harry Potter, #3),
     Harry Potter and the Order of the Phoenix (Harry Potter, #5),
     Harry Potter and the Deathly Hallows (Harry Potter, #7)}    => {Harry Potter and the Half-Blood Prince (Harry Potter, #6)} 0.06040268  0.9863014 0.06124161 11.75671    72
[3] {New Moon (Twilight, #2),
     Breaking Dawn (Twilight, #4)}                               => {Eclipse (Twilight, #3)}                                    0.07382550  0.9670330 0.07634228 11.64347    88
```

As you can see above, the three most important rules involve books that are part of a larger series. For each rule, the strongest association is between a book from a larger series and other books that are part of that same series. This makes logical sense as the best association rules are likely to be between books of a series and other books within that series. For example, the first rule with a lift of 12.37 suggests a strong association between someone reading The Titan's Curse (Percy Jackson and the Olympians #3) and The Sea of Monsters (Percy Jackson and the Olympians #2) because an individual that read the 3rd book in a series is likely to have read the 2nd book. While this is very logical, it may not be very useful in providing recommendations unless you are suggesting the next book in a series. Association rules are very good for identifying common associations between books being read across the whole user base, but may lack the specificity and nuance that customer-based and item-based content filtering allows.

6. Recommender Strategy Comparison (UBCF, IBCF, Association Rules)

User based collaborative filtering focuses on providing users recommendations by identifying other users with similar preferences and providing recommendations based on their preferences. This strategy allows personalization of the recommendations for each user that are relevant to their personal preferences. User based collaborative filtering however is limited when dealing with large data sets that have a lot of users and items, and sparse ratings of those items. This was a challenge for Goodreads' data set because it is very large and many users did not have ratings for large numbers of the items. In addition, when making recommendations, UBCF is limited in providing recommendations for users with limited numbers of previous ratings. If used by Goodreads, this means this strategy would struggle to provide valuable recommendations to many of the users.

In contrast, Item-based collaborative filtering focuses on providing user recommendations based on similarities in factors between books that the user has shown a preference for and recommended books. This method provides personalized recommendations by determining what factors are important to the user based on previous preferences and finding recommendations that have similar characteristics. IBCF strategies are limited by the number of factors or characteristics available in the data to determine user preferences. Since there is a limited amount of valuable information to distinguish the books from one another outside of author names and language, IBCF methods are limited in providing recommendations with the given dataset. If more data was present on the characteristics of the books such as genre, length, year of publishing, etc, IBCF would be more precise in providing recommendations. However, including this additional data available on each book to the data set to create more accurate recommendations would come at a cost, as this would increase the computational requirements of providing the recommendations. As IBCF strategy has higher computational needs to run for providing recommendations, this strategy would ultimately mean recommendation would take a long time or require expensive computing for Goodreads.

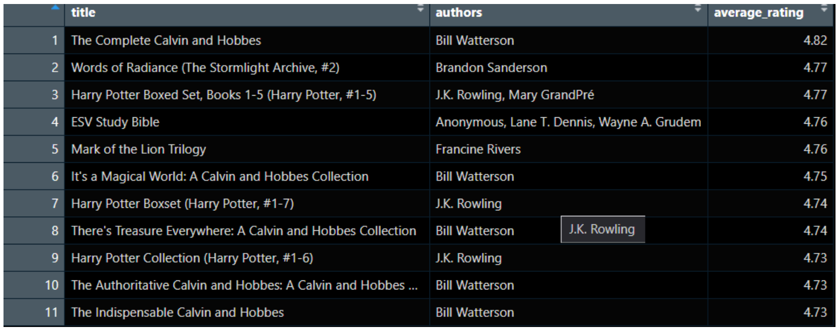
Lastly, the association rules provide valuable insights into item co-occurrence patterns. For the Goodreads data set, this means the association rules identify books that a user may be interested in reading based on how often other users that read that same book have read other books in the data set. Using this strategy, we found some interesting rules, but they were not particularly surprising as they mostly pertained to recommending books within a series to users who read other books in the series. This could be valuable in providing recommendations but is not very personalized as it does not rely on each individual user's personal preferences, rather the entire data set as a whole.

7. Conclusion

This paper set out to explore the Goodreads Dataset and uncover insights while developing robust recommendation models for books. Leveraging both user-based and item-based collaborative filtering approaches, our objective was to enhance understanding of user behavior, preferences, and reading habits through data-driven methodologies, as well as determine the best method for providing recommendations to users. Based on our analysis layed out in the report above, we have determined that user-based content filtering would be the best strategy for providing recommendations for Goodread’s users because it would provide personalized recommendations based on user specific preferences. In addition, it would not have computational requirements that may limit the required dynamicness of the recommender system to accept new inputs (like IBCF would).

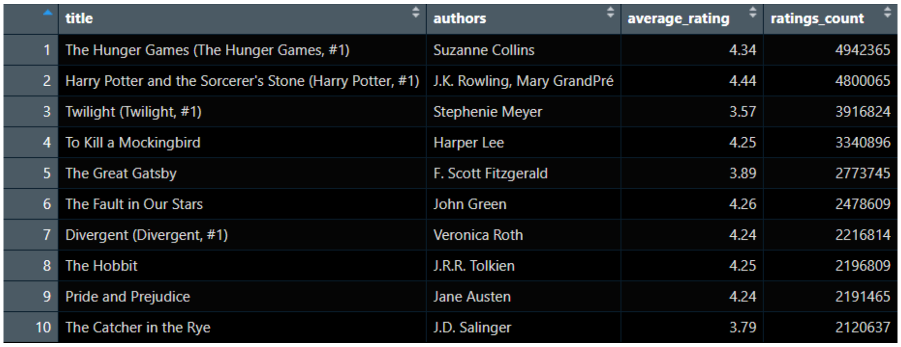
Appendix

Appendix A: Top 10 Books by Avg. Rating



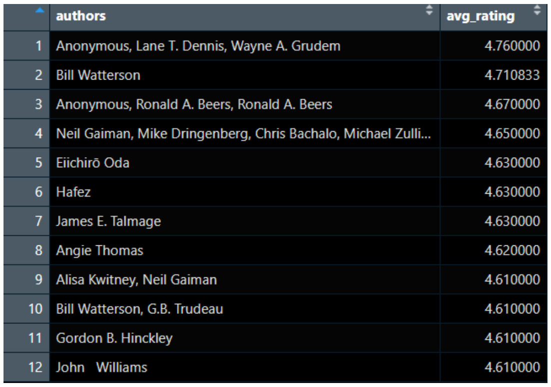
	title	authors	average_rating
1	The Complete Calvin and Hobbes	Bill Watterson	4.82
2	Words of Radiance (The Stormlight Archive, #2)	Brandon Sanderson	4.77
3	Harry Potter Boxed Set, Books 1-5 (Harry Potter, #1-5)	J.K. Rowling, Mary GrandPré	4.77
4	ESV Study Bible	Anonymous, Lane T. Dennis, Wayne A. Grudem	4.76
5	Mark of the Lion Trilogy	Francine Rivers	4.76
6	It's a Magical World: A Calvin and Hobbes Collection	Bill Watterson	4.75
7	Harry Potter Boxset (Harry Potter, #1-7)	J.K. Rowling	4.74
8	There's Treasure Everywhere: A Calvin and Hobbes Collection	Bill Watterson	4.74
9	Harry Potter Collection (Harry Potter, #1-6)	J.K. Rowling	4.73
10	The Authoritative Calvin and Hobbes: A Calvin and Hobbes ...	Bill Watterson	4.73
11	The Indispensable Calvin and Hobbes	Bill Watterson	4.73

Appendix B: Top 10 Books by # of Reviews/Ratings



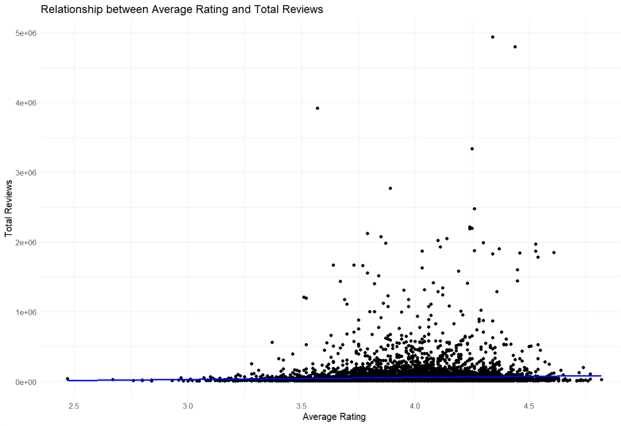
	title	authors	average_rating	ratings_count
1	The Hunger Games (The Hunger Games, #1)	Suzanne Collins	4.34	4942365
2	Harry Potter and the Sorcerer's Stone (Harry Potter, #1)	J.K. Rowling, Mary GrandPré	4.44	4800065
3	Twilight (Twilight, #1)	Stephenie Meyer	3.57	3916824
4	To Kill a Mockingbird	Harper Lee	4.25	3340896
5	The Great Gatsby	F. Scott Fitzgerald	3.89	2773745
6	The Fault in Our Stars	John Green	4.26	2478609
7	Divergent (Divergent, #1)	Veronica Roth	4.24	2216814
8	The Hobbit	J.R.R. Tolkien	4.25	2196809
9	Pride and Prejudice	Jane Austen	4.24	2191465
10	The Catcher in the Rye	J.D. Salinger	3.79	2120637

Appendix C: Top 10 Authors by Avg. Rating

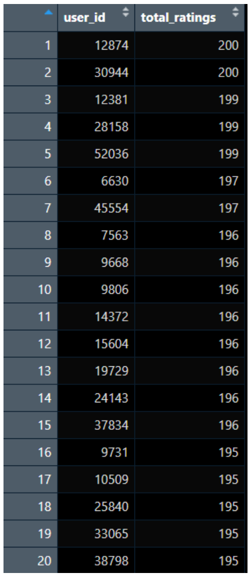


	authors	avg_rating
1	Anonymous, Lane T. Dennis, Wayne A. Grudem	4.760000
2	Bill Watterson	4.710833
3	Anonymous, Ronald A. Beers, Ronald A. Beers	4.670000
4	Neil Gaiman, Mike Dringenberg, Chris Bachalo, Michael Zulli...	4.650000
5	Eiichirō Oda	4.630000
6	Hafez	4.630000
7	James E. Talmage	4.630000
8	Angie Thomas	4.620000
9	Alisa Kwitney, Neil Gaiman	4.610000
10	Bill Watterson, G.B. Trudeau	4.610000
11	Gordon B. Hinckley	4.610000
12	John Williams	4.610000

Appendix D: Plot of Relationship between Avg. Rating and Total Reviews for Books

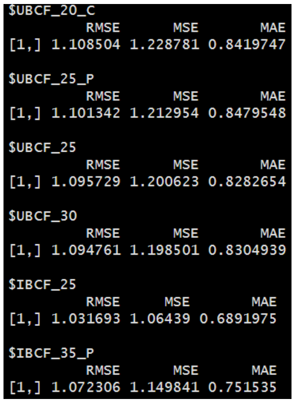


Appendix E: Top 20 Users by # of Ratings



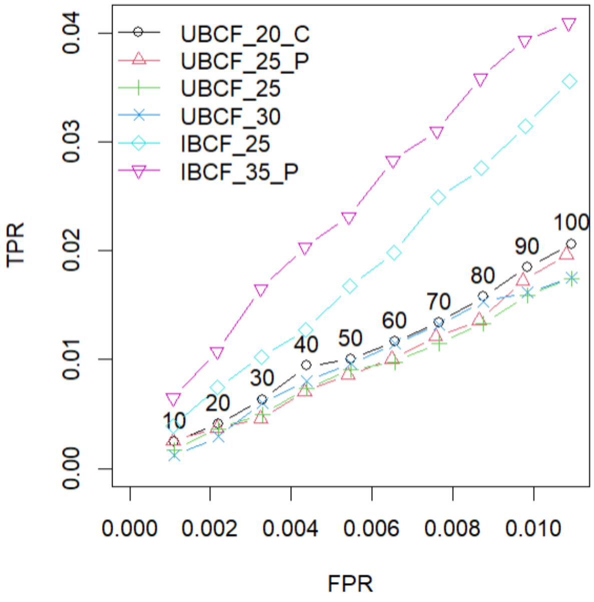
	user_id	total_ratings
1	12874	200
2	30944	200
3	12381	199
4	28158	199
5	52036	199
6	6630	197
7	45554	197
8	7563	196
9	9668	196
10	9806	196
11	14372	196
12	15604	196
13	19729	196
14	24143	196
15	37834	196
16	9731	195
17	10509	195
18	25840	195
19	33065	195
20	38798	195

Appendix F: Evaluation Results for 6 Models Using 1st Evaluation Approach



```
$UBCF_20_C
         RMSE      MSE       MAE
[1,] 1.108504 1.228781 0.8419747

$UBCF_25_P
         RMSE      MSE       MAE
[1,] 1.101342 1.212954 0.8479548

$UBCF_25
         RMSE      MSE       MAE
[1,] 1.095729 1.200623 0.8282654

$UBCF_30
         RMSE      MSE       MAE
[1,] 1.094761 1.198501 0.8304939

$IBCF_25
         RMSE     MSE       MAE
[1,] 1.031693 1.06439 0.6891975

$IBCF_35_P
         RMSE      MSE      MAE
[1,] 1.072306 1.149841 0.751535
```

Appendix G: ROC Plot for 6 Models (Top N[10-100] Recommendations)


TPR
FPR
UBCF_20_C
UBCF_25_P
UBCF_25
UBCF_30
IBCF_25
IBCF_35_P