Optimising Bookstore Recommendation Systems Through User Demographics and Prediction Accuracy

Group W19G7

Hoang Nam Le
COMP20008
hoangnaml@student.unimelb.edu.au

Aden Wright
COMP20008
amwright@student.unimelb.edu.au

Daniel Nam
COMP20008
djnam@student.unimelb.edu.au

Executive Summary

Through a bookstore's need for user recommendations, this report addresses the problem of fitting a recommendation system to the demographics of real-world users. This is accomplished through the testing of two different recommendation implementations applied to a singular dataset.

At its outset, this report describes the preprocessing steps applied to its datasets, including, but not limited to, data imputation, data manipulation, and data scaling. These steps will lead to the development of two recommendation systems pitted against one another: collaborative filtering and matrix factorisation through SVD.

Our investigation found that the primary demographic of users at this bookstore are North American adults between 18 to 40 years old. Most books in this bookstore were written around the early 21th century. Moreover, we discovered that the SVD-based recommendation system is more effective than the Collaborative Filtering-based system in our use case.

Introduction

This report gives an insightful look into users' behaviour and different preferences in books in an online bookstore, alongside creating a recommendation system for users based on the results found by past users. The report will use three different datasets to analyse: BK-Users.csv, BK-Books.csv, and BK-Ratings.csv, alongside another three datasets to create and test the recommendation system: BK-NewBooksUsers.csv, BK-NewBooks.csv and BK-NewBooksRatings.csv.

This report aims to learn and unearth commonalities amongst a diverse user base whilst also comparing different recommendation systems to see which best exploits these similarities to recommend books accurately.

Methodology

We discovered that different datasets required different tools and methods for data preparation. These, for each dataset, were as follows:

For BX-Books.csv, scaling was used for the column Year-Of-Publication. This column mostly had data that was at a reasonable number, with the range of the typical data figures being from 1920 to 2005, but there were plenty of irregular years that were clearly out of place. One such year was 0, with 314 recorded Year-Of-Publication columns holding such data figures, possibly correctly inputted numbers. Still, we determined that this was unlikely, and it would be more appropriate to perform scaling on these columns so it would be easier to accurately represent the data and prevent making any abnormal correlations while analysing the data. Another year was 2030, which was immediately recognised as an abnormal figure, as the current year is 2024. In replacing these two years determined as errors, scaling was implemented, with both 0 and 2030 data figures replaced by the year mode.

Case-folding was used for the column Book-Publisher. We determined that since the data in Book-Publisher is of names and not a generic group of words, we would leave in any punctuation/noise as they could be a part of the official name of the publisher, such as a hyphen "-" in Amber-Allen Publishing. All data in this column was put in lowercase, and the data of publishers in a foreign language were left as is to preserve the accuracy of the data, despite a seemingly corrupted string representing said data.

As for the column Book-Author, case-folding and punctuation/noise removal were used. We determined that since the data in Book-Author is a name and not a generic group of words. Still, there are also legalities behind what punctuation can be in a name, so we removed all punctuation in this column except for hyphens, apostrophes, and full stops. All data in this column was put in lowercase, and the data of author names in a foreign language were left as is to preserve the accuracy of the data, despite a seemingly corrupted string representing said data.

Case-folding was also used for the column Book-Title. We determined that since the data in Book-Title is of names and not a generic group of words just like with Book-Publisher, we would leave in any punctuation/noise as they could be a part of the official name of the book. All data in this column was put in lowercase, and the data of titles in a foreign language were left as is to preserve the accuracy of the data, despite a seemingly corrupted string representing said data.

As for the column ISBN, case-folding and scaling were used. After checking that every ISBN was in the 10-digit format, checks for any random symbols or letters were made. A final check was made to see if any Xs were in lowercase, as an uppercase X is usually used to represent a 10, to keep the ISBN number in the 10-digit format while making the ISBNs with an X consistently have it in uppercase.

For BX-Users.csv, data imputation, discretisation, and data manipulation for the column User-Age. This column has data that is missing completely at random, and to handle this problem, the age mode is used to fill in the missing data. Furthermore, data manipulation is used to strip off the "-" at the end of every data. This will facilitate finding the mode of the users' age and discretise the age into different categories. Since the range of age is vast, it would be difficult to analyse and use it to predict a user's preferences. Therefore, to solve this problem, it will be categorised into five categories (under 10 is kids, between 11 and 17 is adolescents, between 18 to 39 is adults, between 40 and 59 is middle-aged adults, and from 60 is old adults). Discretising this column would make finding correlations between users and their books' preferences easier.

As for the column User-Country, case-folding, data manipulation, and punctuation/noise removal were used. Since an external dictionary is used to check the validity of an input, case-folding and punctuation removal facilitate the validation process. All inputs in this column are put in lowercase, and any punctuation is removed (for example, u.s.a will become usa). Afterward, some of the supposedly invalid data, due to wrong spelling or technical mistakes in different languages, were manipulated back so that it would be in the correct format. After the validation test, all input that did not pass it would be manipulated back to the form "-". Finally, data manipulation and imputation were used for User-States and User-City.

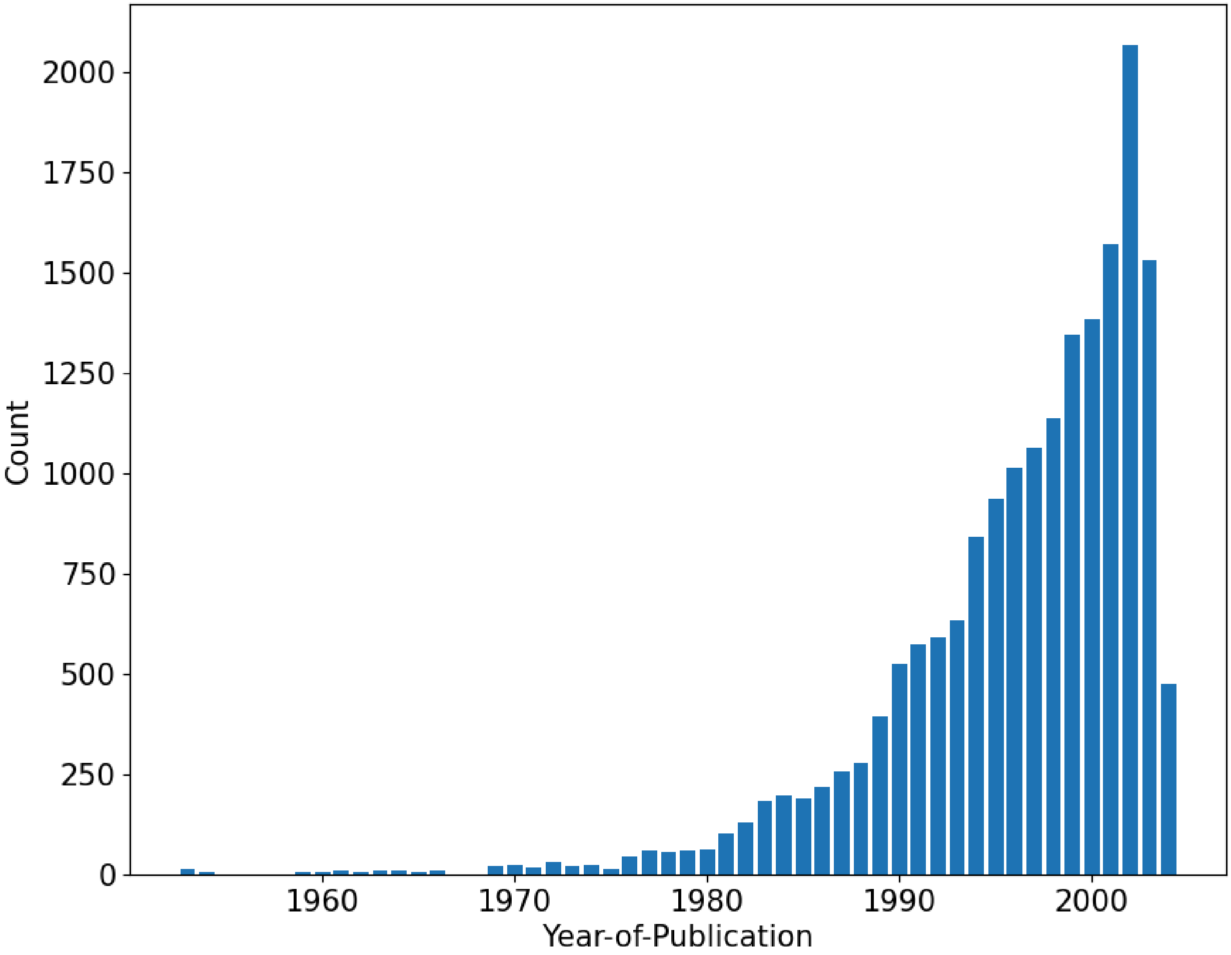
For User-States, all values that are invalid due to the User-Country's input of that row is either "-" or is a country that does not have a state or that value is left blank would become "-", and similarly, all values in User-City that have an invalid User-Country's value or the value is left blank would become "-".

For BX-Ratings.csv, case-folding, and scaling were used for the column ISBN. After checking that every ISBN was in the 10-digit format, just like in BX-Books.csv, checks for any random symbols or letters were made. A final check was made to see if any Xs were in lowercase, as an uppercase X is usually used to represent a 10, to keep the ISBN number in the 10-digit format while making the ISBNs with an X consistently have it in uppercase.

Scaling was used for the column Book-Rating. It was made sure that none of the ratings were outside of the ranges 0-10, as this was a range that made sense for a rating system. After checking, there were no values that seemed to be abnormal.

Scaling was used for the column User-ID. It was made sure that there were no abnormal data entries, with all of the entries being checked to see if it was a positive integer, as well as consisting of a reasonable amount of digits, in this case the maximum being 6.

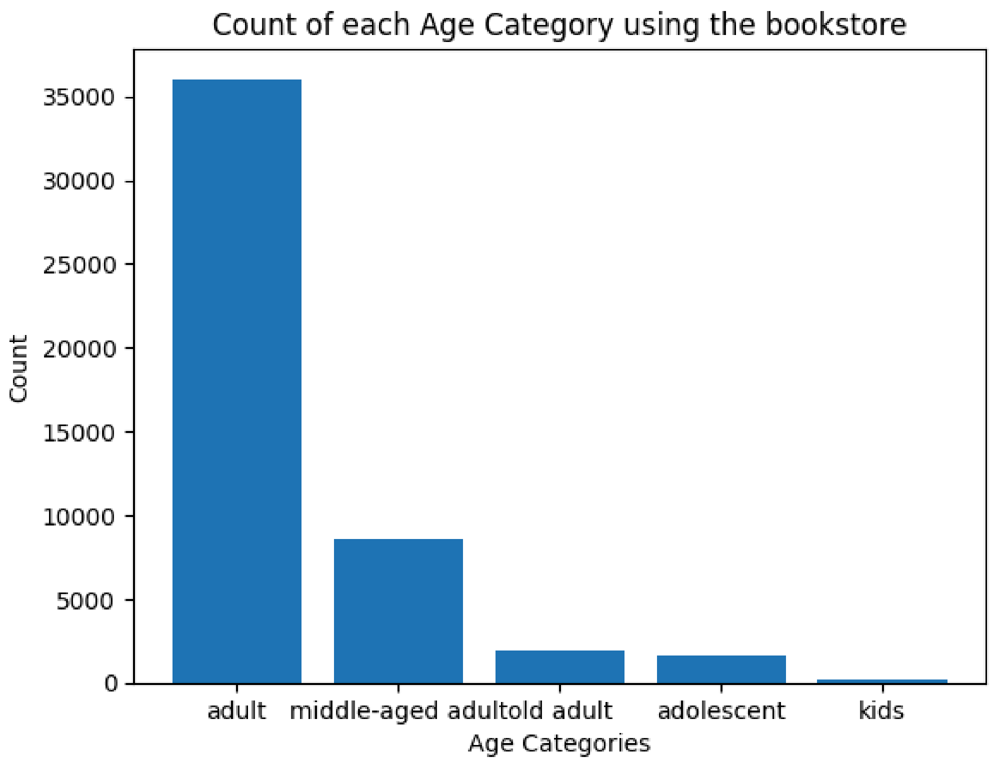
Data Exploration and Analysis



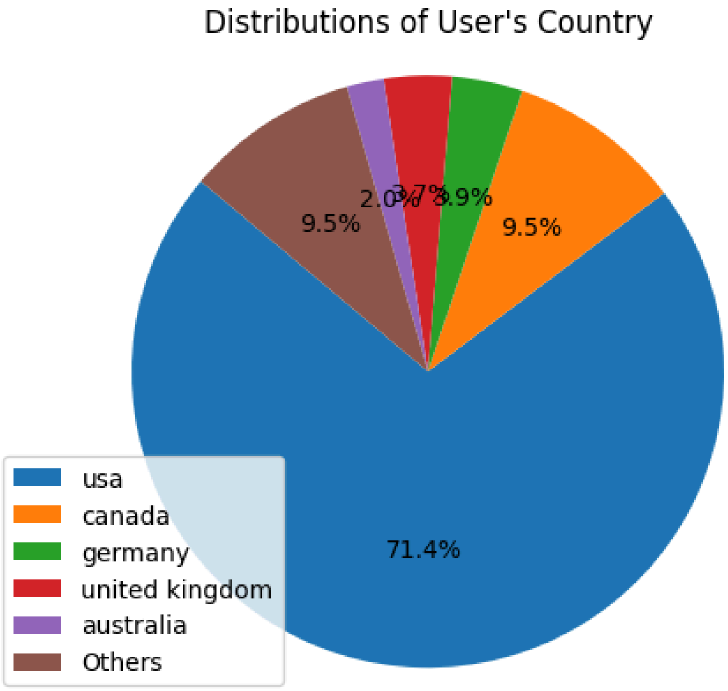
Graph 1: Chart showing the distribution of publication years

From preprocessing data for BX-Books.csv, it could be seen that the mode year of publication of books was 2002, with 2067 books counted that were published this year, out of a total of 18185, which meant that 11.4% of the books in this data set were from 2002. As shown in the bar graph above, most of the books in the data set were published in the late 1990s and the early 2000s.

From preprocessing data for BX-Users.csv, the graph below shows the distribution of users' age and nationalities:

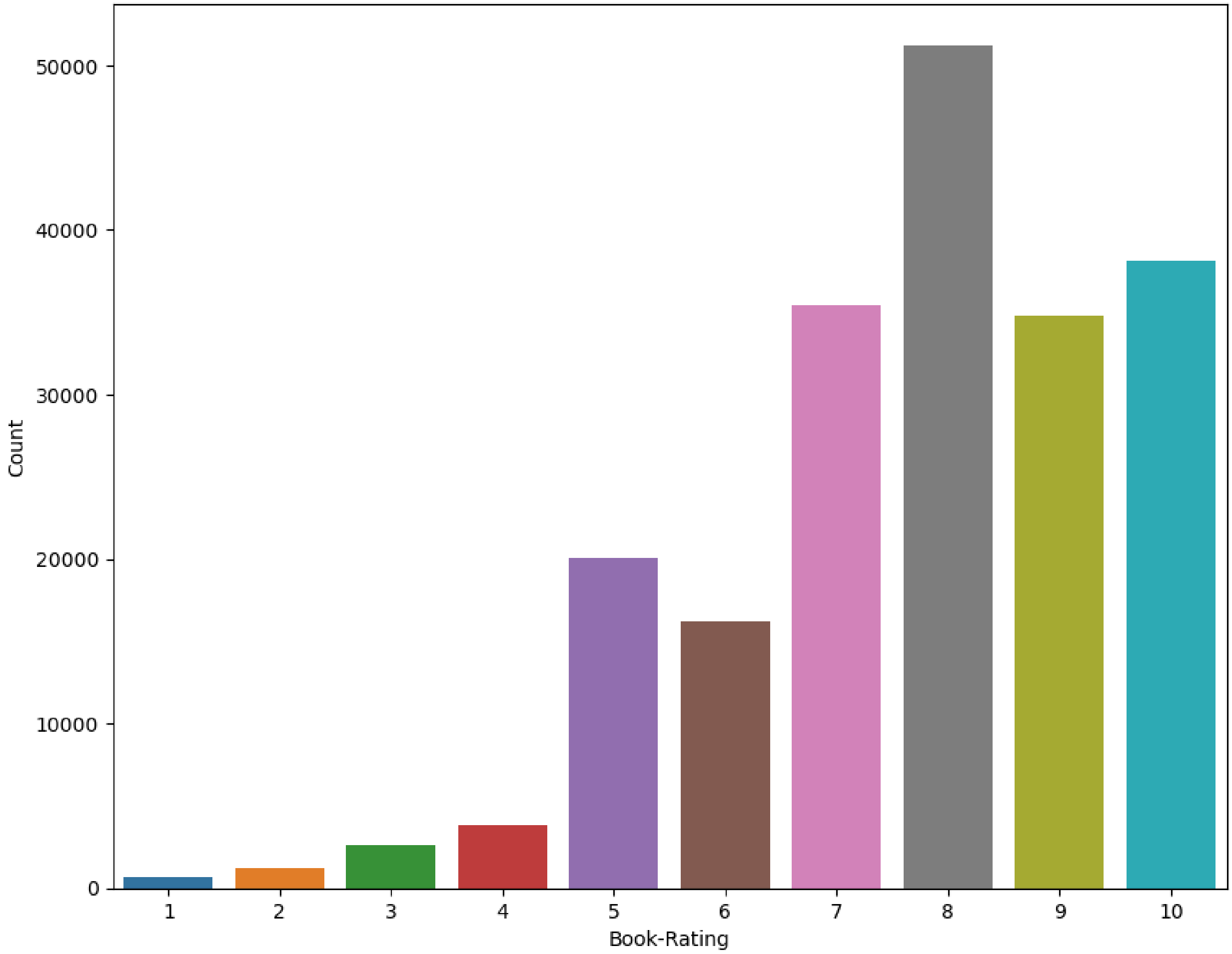


Graph 2: Bar Chart showing the age distribution of users



Graph 3: Pie Chart showing the distribution of users' countries

The graph above further shows that the majority of users are adults with more than 35,000 users, which is about 75% of the total users. It is also worth noticing that adolescents and kids did not contribute a lot to the total number of users of this bookstore. Moreover, nationalities-wise, nearly 92% of the users are from Canada or the USA.



Graph 4: Bar Chart showing the distribution of book ratings

From preprocessing data for BX-Ratings.csv, it could be seen that the mode ratings for the books was 8, with 51206 books being rated an 8 by users, which was about 25.1% of the total number of books rated at 204164 total. The numerically largest ID was recorded to be 278854, which indicates that on average every user in the database rated 0.732 books per person. As seen in the graph above, there were much more books that were rated positively, with 7, 8, 9 and 10 all having more than 30000 counted, while 1, 2, 3 and 4 all didn’t reach 5000.

Results

We first utilised a collaborative filtering approach to the recommendation. To do so, we constructed sparse user-item interaction matrices to form the basis for calculating cosine similarity. This process was done for both user-user and item-item similarities as we aimed to investigate the effectiveness of different methods in solving our research question. Cosine similarity is especially useful here as it is most affected by the overall pattern of ratings as opposed to their numerical magnitude. By nature, a 1-10 book rating scale is a subjective measurement. Investigating user-based predictions and item-based predictions further allowed us to capture the plurality of our dataset's demographics.

We used an SVD-based model in the Surprise library to benchmark against our collaborative filtering implementation. The performance of both models is scored through the same RMSE metric, allowing us to draw conclusions about which model best allows us to explore the user's demographics and predict ratings accurately.

Discussion and Interpretation

Findings from the dataset

- The dataset shows that most books read by readers were published in the late 1990s and the early 2000s. In contrast, there were barely any books published in the 1920s up to the 1980s that were read compared to the more modern decades, which indicates either that the readers as a group preferred much newer books than older ones, or that there were much more books published in recent years compared to 50 plus years ago, or both.
- The dataset shows that most users are adults, followed by middle-aged old adults. This conveys that these age categories should be the target audience for this bookstore. However, there are significantly fewer adolescents that use this bookstore, which shows that the functionality of this bookstore does not fit well with young people. These could be some of the features that the manager would like to look into and improve upon to accommodate and hence attract more young users.
- The dataset shows that most ratings made were of the higher numbers, namely 7, 8, 9 and 10. In contrast, there were barely any ratings made for 1, 2, 3 or 4, with there being a large disparity between these two groups of ratings. This indicates either that the books were generally very well written and readers considered them to be of high quality, or that the readers were

very avid readers and were happy reading these books despite their quality, or both.

Recommendation system from Collaborative Filtering:

- The result shows that the root mean square errors' value by both item-based and user-based are almost 8, which shows that the difference between the ratings predicted by the recommendation system and the actual ratings by the users are off by 8 values. This is a huge discrepancy as the rating is generally just from 1 to 10, which shows that this recommendation system is not viable for users.
- If this recommendation system is used, then there would be a high chance that users won't get their desired books in their recommendation section, which could lead to a plummet in the number of users of the bookstore. Ideally, for a good recommendation system, the ratings predicted should not be too far off from the actual test set, ideally, the root mean squared value should be around 1 or 2.

Recommendation system created by SVD algorithm

- The result shows that the root mean square errors' value from the SVD system is 1.62, which as explained above is the ideal value for a good recommendation system. This means that if the actual rating of a user to a book is 4, then this recommendation system should predict that rating to be in a range from 3 to 5. These results would be helpful for users using this system as they would now be recommended books closer to their interest as opposed to if the recommendation system is based on Collaborative Filtering.
- To further prove that this recommendation system is ideal, we also did a test rating on user 276744 on a book with the ISBN 0358550120X. The actual rating from BX-NewBooksRatings.csv of that user reading that book is 7, and the recommendation system predicts that value to be 7.32, which is very close.

Limitations and Improvement Opportunities

Initially, our research question attempted to explore the confluence of age and user-item similarity. However, this investigation soon became infeasible when we were confronted with the complexity of isolating age, or any single variable, as a significant determinant of user preference. Data sparsity limited our potential understanding of the user base. Our response was to broaden our exploration of the data and, instead, undertake a holistic analysis of user demographics and behaviour. The comparison of different recommendation systems thus became our primary concern as it gave a means to diversify our recommendations to users.

Conclusion

This report aims to look at the members of an online bookstore's book preferences and reading styles and provide recommendation systems using Python and its libraries - such as Pandas - data preprocessing, and recommendation system guided data analysis. We believe this aim was accomplished. Through preprocessing, we found that the majority of the books read were published in the late 1990s and the early 2000s, suggesting that there was a trend towards modern books, and most users of this bookstore were in the age group of 18 to 40, mostly living in the United States. Users would generally also have a preference for rating these books higher. The recommendation systems also provided helpful information. While the collaborative filtering method showed differences between predicted and actual ratings, the SVD model proved to be more accurate, with it having a lower root mean square error, where ideally, the root mean squared value should be around 1 or 2 and the SVD model had 1.62.

References