An Exploration into Content-Based Predictive Models and Recommendation Systems for Books

Group W06G2

Alice Kjar
akjar@student.unimelb.edu.au

Annisa Chyntia Yusup
ayusup@student.unimelb.edu.au

Yin-Xi Chloe Lin
yinxichloel@student.unimelb.edu.au

Executive Summary

The vast number of books available on the market today has presented a challenge for booksellers in marketing and increasing their sales. This study investigates a number of machine-learning techniques to aid bookstore managers in stock curation. The purpose of this study is to simplify decision-making processes to optimise sales in a bookstore and to provide additional book recommendations for the customers to purchase.

Technical applications within the study include data preprocessing (utilising imputation, text processing and discretisation), content-based recommendation systems built on TF-IDF (Term Frequency-Inverse Document Frequency) and Cosine Similarity, and supervised machine learning techniques, such as Correlation and k-Nearest Neighbours.

The findings of this study suggest that, based on data available, it is difficult to predict the average rating of a book based on its characteristics. However, an effective content-based recommendation system was constructed, which would allow booksellers to recommend new books for customers to buy based on previous purchases. It is recommended that this system be implemented by online booksellers to present customers with other books they might like after a book has been bought/reviewed.

1 Introduction

Recommender systems have been mainstreamed as an efficient personalisation tool in processing and filtering data online to enhance users' experiences (Roy & Dutta, 2022). In the context of bookstores, data about books, reviews, and customer demographics are salient factors that inform managers of the potential popularity of books and help them understand consumers' behaviours. Current literature has presented three categories of recommendation systems: content-based, collaborative and hybrid (Roy & Dutta, 2022). According to Knotzer (2008), content-based analyses involve generation of associative features based on the customer's interest and recommending books of similar characteristics, such as ratings, publishers, and authors. Collaborative filtering engages with the demographics of the user and suggests books based on the similarity of their personal information (Portugal et al., 2018). The final category would be the hybrid of the previous two methods, where both sets of information contribute to the recommendations.

In this investigation, we utilised several CSV files:

- 'BX-Books.csv' a dataset of details on 18,185 books with their ISBN, Title, Author, Year of publication and Publisher.
- 'BX-Ratings.csv' a dataset containing users' IDs, ISBN code of books, and the ratings given.

The aim of this investigation is to provide useful insight into the impact of book characteristics on the purchasing group and average rating. This information will assist booksellers to recommend other books for customers as possible additional purchases and predict the popularity of a book, when they are adding new books to their stock. The two questions that we aim to answer are:

1. Can we predict a new book's rating based on the book's characteristics?
2. Can we use book characteristics to recommend similar books for readers?

Hence, to apply the most apt recommendation system that aligned with our aims, we adopted the content-based approach. Due to the lack of user-based data relating to newly released books, content-based recommendation systems are more appropriate in the context of a bookstore, where a significant proportion of the sales are new releases.

2 Methodology

2.1 Preprocessing

Prior to analysis, the datasets underwent a number of preprocessing techniques to ensure accuracy, completeness and consistency.

To aid in vectorisation of titles and consistent grouping of publishers and authors, text processing was applied to all text entries. This included correcting mojibake caused by inconsistencies in decoding, case folding, and removing punctuation and whitespace inconsistencies. Additionally, all stopwords were removed from titles and publishing houses. This includes common stopwords found in English, as well as words which contextually provided little meaning, or caused unnecessary differentiation between entries (such as “Random House” and “Random House Inc”). Titles then underwent lemmatisation to remove morphemes and group together words with the same root. The languages of each of the books were extracted from the title, with confidence metrics needing to be over a threshold of 80%. This was to account for inaccurate readings generated from shorter book titles.

A number of values in the ‘Year-Of-Publication’ column were found to be beyond a plausible range for year values [1920 – 2024], so imputation was performed to correct this. Incorrect values were imputed using the median of correct values, rather than the mean, to account for the substantial skew in the distribution.

Several basic calculations were also performed on the data to aid in later analysis. Customer ratings (obtained from the BX-Ratings database) were used to obtain the average ratings for each book. The books were also grouped by author and publisher and the frequencies of each were used to obtain the number of books written by a book’s author (‘Author Count’) and published by its publisher (‘Publisher Count’).

Finally, discretisation was applied to the average rating of each book to divide ratings into ‘Low’ [0-7.71] and ‘High’ (7.71 – 10] values and aid in k-Nearest Neighbours analysis. Equal-frequency binning was utilised here to again account for any skew in values and ensure balanced bins.

2.2 Exploratory Data Analysis

To further aid in the investigation, an analysis of variables of interest was done to identify the distributions and trends in the dataset. Retrieving the general details of the three independent variables (year, author and publisher) and one dependent variable (average rating) was done by generating descriptive statistics regarding their distributions.

Subsequently, data visualisation tools were then utilised to highlight patterns in the data. Histograms were used to display the distribution of each of our four numerical variables. However, the initial distribution plots for author and publisher counts were starkly skewed, which limited readability. Thus, to mitigate this problem, a logarithmic scale on the y-axis was applied to these two graphs.

Finally, to represent text-formatted book features (i.e. the title), a word cloud was generated from the distribution of words frequencies. These data visualisation tools were effective in enhancing the legibility of data that was previously challenging to interpret.

2.3 k-Nearest Neighbours

To predict the average rating of the books, based on their attributes, a k-Nearest Neighbours model was constructed. To do this, the dataset was randomly split into 3 subgroups; train (60% of the original dataset), validate (20%) and test (20%). The training data was then used to first find the correlation between each of the features of interest (number of books written by author, number of books published by publisher and year of publication) and the average rating of the book.

For a number of k values (k = 1, 3, 5, 7, 9), a k-NN model was then built using the features of interest to predict a high or low average rating. These models were built using the train dataset and the accuracy found using the validate dataset, observing the k value with the highest accuracy.

The validate and training subsets were then merged and used (along with the most accurate k value) to train a final k-NN model. The accuracy, precision, recall and F1 of this model were then evaluated using the test dataset and a confusion matrix was generated.

2.4 Content-Based Recommendation System

Content processing was utilised to generate book recommendations based on two features: title and author. Both features that had been pre-processed were combined into a single text, and duplications were omitted to prevent recommending reprints of the same book from different years (for example: "The Angel Maker" by Ridley Pearson was released in 1994 and 2001). This combined text was converted into a matrix by performing TF-IDF vectorisation, where rows correspond to the combined title and author, columns correspond to unique words in the corpus, and the values in the matrix represent the TF-IDF scores for each word in the combined text.

To measure the similarity between books, Cosine Similarity technique was applied in this analysis. The calculation was implemented by comparing vectors between books to determine whether they are pointing in the same or different directions (Januzaj & Luma, 2022). The output resulting from the calculation is a symmetric matrix, where each element represents the Cosine Similarity values between the TF-IDF vectors of the combined title and author. This value ranges from -1 to 1, which implies the higher the value, the more similar the books are.

By using the Cosine Similarity matrix, a recommendation system was built to provide a list of books for a specific item based on the extracted similarity values, sorting them in ascending order. Only books listed in the top 10 similarity rank with values between 0.5 to 0.9 were provided in this system. The upper bound of 0.9 was set to dismiss highly similar books to prevent duplications, and the lower bound of 0.5 was to ensure that suggested items hold strong similarities.

3 Results

3.1 Exploratory Data Analysis

Table 1: Descriptive Statistics of the book features and average rating

Variable	Count	Mean	StDev	Min	Q1	Median	Q3	Max
Year of Publication	18185	1996.02	7.00	1920	1993	1998	2001	2005
Num. books by author	5964	3.05	6.53	1	1	1	2	195
Num. books by publisher	1193	15.24	58.26	1	1	2	5	864
Average Rating	18185	7.66	0.89	3.00	7.14	7.71	8.25	10.00

The descriptive statistics can be used to determine the extent of skew present in the data via the comparison of mean and median. The year of publication demonstrated a moderate of negative skew about a median of 1998. Comparatively, the number of books by author and number of books by publisher had very strong positive skews with differences between mean and median of 2.05 and 13.24 respectively. The average rating had a roughly symmetric distribution about a mean of 7.66 (Table 1). These distributions can be visualised in Figures 1 and 2. The rescaling the y-axes by log in subplots 2 and 3 of Figure 1 help indicate the extent of this skew, while the outliers are readily identifiable.

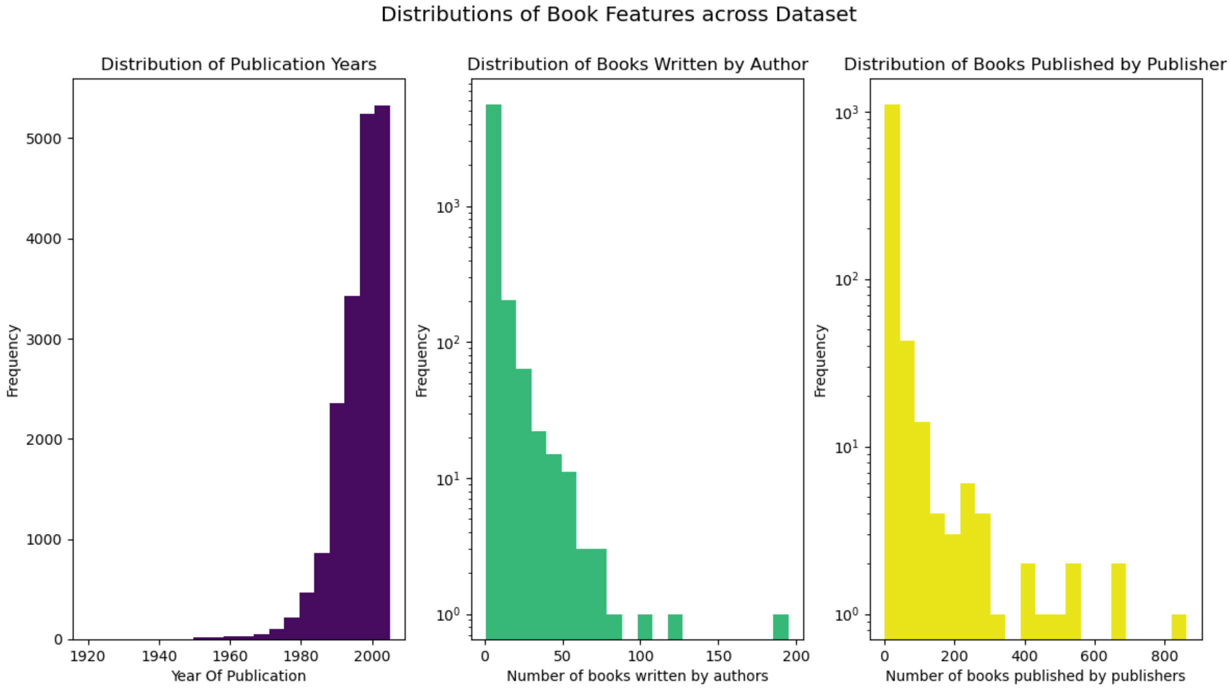

Figure 1: Histograms of distributions of book features (Note that subplots 2 and 3 have a logarithmic y-axis)

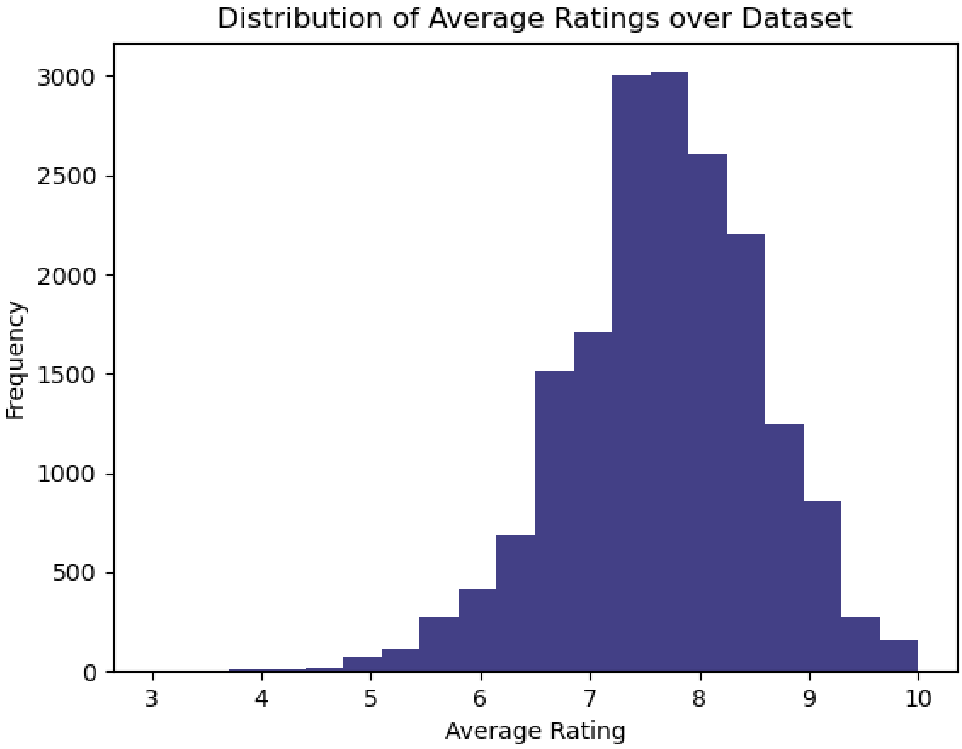

Figure 2: Histogram of distribution of average ratings

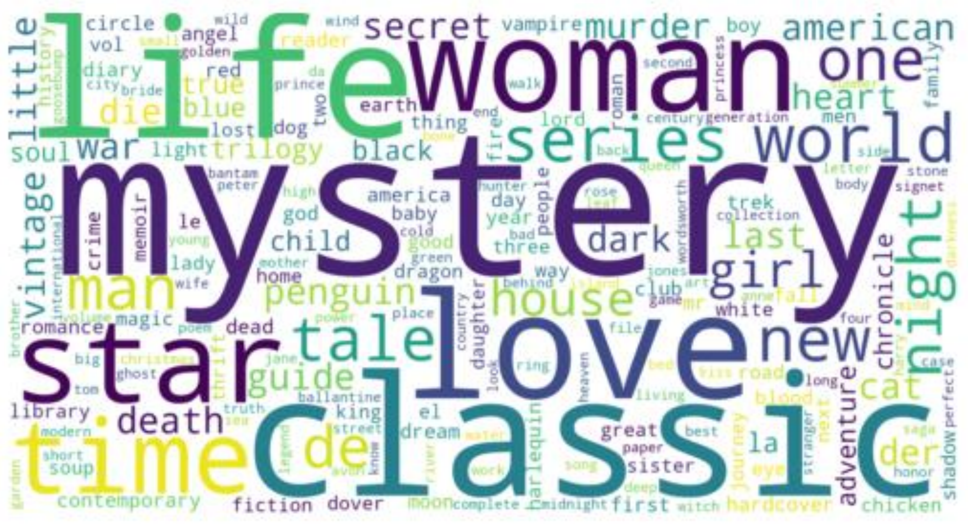

Figure 3: Word Cloud of popular words in book titles

The word cloud shows particular words extracted from book titles. Font size represents total number of words, so the larger the word appears in the cloud, the more frequently it occurs in the titles. After pre-processing titles, including the removal of English stopwords and other insignificant phrases that may appear in repeat like “novel”, “book” and “story”, there were numerous of words displayed in the word cloud with different sizes. Some of them were noticed to have more prominent sizes than others, which include “mystery”, “life”, “classic”, “love”, “star”, and “woman” (Figure 3), and it can be said these words are the most frequent words found in the collection of book titles. While the majority terms are found in English, there were some unusual words captured such as “*de*”, “*la*”, and “*der*”, which were extracted from titles of non-English books.

3.2 k-Nearest Neighbours

The correlations between each of the features and the average rating were quite small (Table 2). The year of publication had the strongest observed effect on the ratings; however, this was still quite a weak correlation of r = -0.1213. The correlations for number of books written by author and number of books published by publisher were negligible.

Table 2: Pearson correlation between each of the book features and the average book rating.

Book Feature	Correlation
Year of Publication	-0.1213
Number of books written by author	0.0283
Number of books published by publisher	-0.0706

As a result of these weak correlations, the accuracy of the k-NN model (k = 7), was low and, as can be seen in the confusion matrix (Figure 4), there was not a great difference between the number of correct and incorrect predictions. The accuracy for the model was found to be 0.5648 and the recall, precision and F1-score were similar with values of 0.5648, 0.5647 and 0.5647 respectively. While these results are a slight improvement from if the predictions were just made randomly, irrespective of book attributes, the accuracy is likely not high enough to be of significant use to the bookseller.

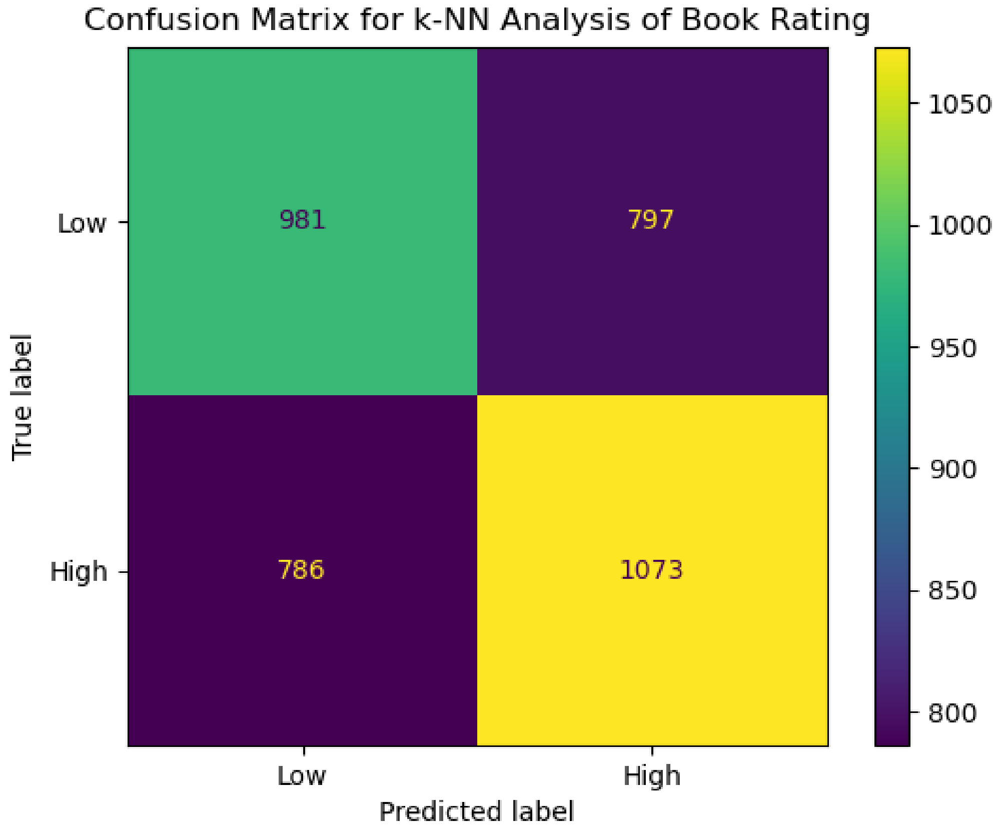

Figure 4: Confusion matrix evaluating the accuracy of predictions made by the k-NN model.

3.3 Content-Based Recommendation System

After dropping duplicates in title and author, the number of books was reduced by 13.6% from the original dataset. All books were taken into TF-IDF vectorisation and Cosine Similarity computation, and the output from this analysis is a matrix of similarity values between all pairs of books based on title and author. Based on the attained values, a recommendation function was formed to extract a list of items that hold high similarities with one specific book and provided them as recommendations. For example, given a book entitled "Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))" written by J.K. Rowling, some other related books from the dataset were suggested in the table below (Table 3).

Table 3: List of Books Recommendation

ISBN	Book-Title	Book Author	Detect Language	Similarity Score
043965548X	Harry Potter and the Prisoner of Azkaban (Harry Potter)	j k rowling	en	0.7336
807281956	Harry Potter and the Sorcerer's Stone (Book 1 Audio CD)	j k rowling	en	0.6876
2070556859	Harry Potter et l'Ordre du Phénix (Harry Potter, tome 5)	j k rowling	fr	0.6145
767908473	The Sorcerer's Companion: A Guide to the Magical World of Harry Potter	allan zola kronzek	en	0.6017
439064872	Harry Potter and the Chamber of Secrets (Book 2)	j k rowling	en	0.5640
043935806X	Harry Potter and the Order of the Phoenix (Book 5)	j k rowling	en	0.5556

439139597	Harry Potter and the Goblet of Fire (Book 4)	j k rowling	en	0.5530
439136350	Harry Potter and the Prisoner of Azkaban (Book 3)	j k rowling	en	0.5312
3551551936	Harry Potter Und Der Feuerkelch	joanne k rowling	de	0.5112
439425220	Harry Potter and the Chamber of Secrets Postcard Book	j k rowling	en	0.5051

These ten recommendations were at the top of the list of highly similar books with the given title based on the Cosine Similarity score. All titles shared the words "Harry" and "Potter", and almost all books were written by the same author; J.K. Rowling, except one book by Allan Zola Kronzek. Notice that two books were detected to have languages other than English (de: German and fr: French) as the titles contain non-English words such as "*Feurerkelch*" and "*l'Ordre*".

4 Discussion and Interpretation

4.1 Preprocessing

Preprocessing the datasets was a relatively simple process, due to the largely consistent structure of the data, however there were some observations of note throughout the process. To fix the mojibake (incorrectly decoded non-ASCII characters) in the book titles, we initially attempted to use the 'fix_text' function from the 'ftfy' library directly on the text. However, this was unsuccessful, and it was found that all question mark characters had to be first removed from the text for this function to accurately correct the broken characters. This extra step is of note as it suggests that potentially this data was incorrectly decoded multiple times.

Additionally, when processing the publishers, many were found to have slight variations in their names, for instance "Random House" and "Random House Inc". Initially, it was considered that this be managed this taking the first *n* characters of each of the publishers' names, however this raised problems for similarly named publishers (such as "University of Chicago Press" and "University of California Press"). Instead, to manage this issue, a number of common 'filler' words (such as "books", "publishers" and "inc") were removed from the publishers' names. This technique was not completely effective but was much less risky than the aforementioned process. Similarly, there were some authors which had different formats of the same name, for instance, "Walter M. Miller Jr.", "Walter M., Jr. Miller" and "Walter M Miller". While casefolding and punctuation removal fixed some of these inconsistencies, there was a very small number which did not follow a clear pattern. For the sake of simplicity, as there were very few of these occurrences, they were not further corrected.

As the titles of the texts were in many different languages, stopword removal was difficult, as every language has different stopwords. It was considered that this be managed this by removing stopwords from all of the languages offered by the 'nltk.corpus' library, however it was quickly discovered that this was problematic, as a number of stopwords in one language have important meaning in another. One such example of this is "*war*" which is a German stopword meaning "was", but in English holds crucial meaning about the book's genre or themes. Eliminating this could prevent books about military history, for instance, from being grouped together. Instead, just English stopwords were removed from the titles as this was the most common language amongst the books and hence had the smallest margin of error.

4.2 Exploratory Data Analysis

Upon the production of the histogram of the frequency in books sold with respect to year of publication, a left-skewed, unimodal data plot was observed **(**Figure 1**).** This pattern suggests that recently released books are being stocked, and therefore purchased more frequently. Hence, the recommendation system will have an increased likelihood of promoting recent books in comparison to earlier published books, due to the underlying distribution.

The strong positive skew found in both the distributions of the number of books per author or publisher suggests that there are very few authors and publishers that are releasing a large number of books. This makes sense as the popularity of the authors and publishers creates competition and more successful individuals/groups control market (Phillips et al., 2019, p.13). It is also expected that there is such a high concentration of values at the lower tails of both graphs, as the number of smaller, self-published authors is much higher on online bookstores than in conventional bookstores (Sheelam, 2020). However, these frequencies may be slightly inflated by the inconsistencies in grouping mentioned in Section 4.1.

From the descriptive statistics, it was observed that the ratings were centred around the upper half of the scale (mean = 7.66). This information would allow the bookstore manager to establish a benchmark while selecting the books, as well determining the extent to which they would be willing to invest.

The most frequent words identified in the word cloud can be associated with popular themes or genres of books. For instance, “mystery”, “death” and “murder” can clearly be related to crime fiction. This suggests that book title is often representative of book content, and hence it is valid to make recommendations based on the content of a book’s title. Bookstore managers are able to leverage this information in their marketing strategy by providing books with those genres in the collection. However, as the result of not removing stopwords in non-English books, some meaningless words did appear somewhat frequently, giving no information about the genres.

4.3 k-Nearest Neighbours

Prior to the construction of the k-NN model, initial analysis showed that there was a limited relationship between the books’ features and their rating. There was a weak, negative correlation between year of publication and average rating, which is somewhat expected as people are much more likely to only buy books from many decades ago if they are popular and have a high rating (i.e. classics). Conversely, there was negligible correlation between both the number of books written by the author and the number published by the publisher, and the average rating. This is surprising, as one would expect authors who have written more to be more successful and therefore get better ratings. Similarly, it would make sense for larger publishing houses to get better ratings as they are able to afford more expensive contracts with more popular authors.

The initial plan for analysis was to build a linear regression model from these numerical features to predict what the average rating for a book would be. However, after generating these correlations, and plotting the individual linear regression models for each feature, it was quickly observed that this model would have an extremely low accuracy, and so a k-NN model was adopted instead.

Due to these low correlations, different combinations of features were trialled to construct the k-NN model, however the highest accuracy came from the model which included all three features, which is interesting due to the low correlation of the author and publisher counts. Expectedly, this accuracy was still quite low because of the limited correlations. This accuracy metric (0.5648) suggests that while this model may have some success, it would be quite risky to make any real-world decisions based on its predictions, particularly since there would likely be financial repercussions for the bookseller. However, there is the potential for this predictive model to be improved, as will be discussed in Section 5.

It was expected that the accuracy metric be quite similar to the F1 metric, as the bins were intentionally kept balanced in the discretisation process by using equal-frequency binning. One further point of interest in this model is that the k value found to get the most accurate predictions was quite high (k = 7), as usually k values range from 1 to 5. However, this is likely just due to chance as a result of the accuracy being so low.

4.4 Content-Based Recommendation System

Implementing a recommendation system based on Cosine Similarity metrics in this study depended on the content of books, represented by their titles and authors. At the beginning of the analysis, some books with the same exact content but different ISBNs and publication years existed. To mitigate this, those items were treated as duplications and were omitted to avoid suggesting the same books. Another similar issue was found in books which had identical titles but had some differing additional information after them, such as the series and editions. Setting an upper limit of similarity value (0.9) was quite effective in anticipating such instances. However, the bound might need to be adjusted differently in certain conditions because some books in the suggestion lists that held a 0.8 score, for instance, could either be considered equal or different items. The variation was also discovered in authors where several books share the same author, yet their names are written in two versions on the dataset, like “Joanne K.

Rowling” and “J.K. Rowling”. This might affect the vectorisation of book contents as both names will be regarded to be distinct.

The list of recommended books from the Cosine Similarity score analysis was obtained based on content processing of book titles and authors. This means that only books containing similar words in their titles or similar authors' names with the books that had been purchased will be given as recommendations. Word similarity in titles can be assumed to often indicate the same topic or genre (see popular words in Figure 3) between books, and matching authors might ensure the same writing style in their works. This recommendation system works by grouping books with similar content to match customers’ preference based on one book they have already read. Booksellers can utilise this tool as a part of marketing and to increase their sales by suggesting both old and new collections of books to customers. Two specific potential implementations of this recommendation are;

- Inputting new releases into the system to find similar older books and then recommending the new release to customers who have already purchased these books.
- Inputting a book that a customer has just purchased into the system and using the recommendations to suggest additional items they might like to buy.

A more targeted approach to advertising would mean that a higher proportion of individuals who see these advertisements would be interested in them, and hence they are much more likely to actually make a purchase. This means that the bookseller is able to market more effectively and hence generate more income.

Information regarding book language is worthwhile for offering customers the option for only being shown books in the language they know. This means that advertising space is not being wasted on books that a customer would never buy because they can’t read them. A language detector was used in this study, and it performed reasonably well in distinguishing the language of titles by identifying character frequencies. Even so, it might not be powerful enough to rely solely on book titles to predict the language as they hold limited information and possibly have contradictive language with their content (e.g. a book that was written in Spanish but has English title, or a book has a person’s name as the title).

5 Limitations and Improvement Opportunities

While the preprocessing significantly improved the quality of the data, there were still some limitations. As mentioned, the grouping methods used for both the publishers and the authors were not completely comprehensive, and there were some discrepancies. One way in which this could be addressed in further analysis is by using techniques such as document clustering or approximate string matching to group together similar names which are not exact matches. The language detection process was also quite limited in this study, with some titles being labelled incorrectly or having a low confidence. This occurred mostly with very short titles as there was not much text to base the prediction on, and therefore could be improved by also using a short excerpt or description of each book. This would also aid in stopword removal, as we would be able to remove stopwords specific to the language of the book.

The k-NN model’s low accuracy meant that it would have limited practical use for a bookseller looking to identify the best books to stock their store with. However, with more data, we could find features which have a stronger correlation with book rating, and therefore improve the accuracy. In a similar, more extensive analysis Wang et al. (2019) found that they were able to generate a much more accurate k–NN model to predict book ratings, based on a wide range of characteristics. Some examples of potential features that would be of interest in future analysis include:

- Book pre-order revenue
- Historic sales performance for the author
- Author popularity (obtained from observing internet search traffic)
- Social media trends (for instance, hashtag usage)

The latter is of particular interest as recent ‘BookTok’ trends have had a significant impact on the book buying industry (Stewart, 2021).

Building a content-based recommendation system by analysing book titles and authors to provide book suggestions for customers is useful for personalising their preferences for specific topics or writers. However, there was a limitation in that the suggestion of books provided by this system could not provide recommendations of other topics, genres, or other writers beyond customer preferences. Therefore, it cannot explore any other items that might also capture their interest, and booksellers will lose the opportunity to attract customers on purchasing those books.

Another limitation in this analysis was the assumption that the title provides sufficient information about the content of books. This is not always the case, resulting in a failure to find appropriately similar items for books with very short or non-descriptive titles like "Waiting" or "Proof". To improve this matter, some other essential features could be used to increase the volume of available information (Niupian & Chuaykhun, 2023) such as:

- Book description
- Introduction or explanation
- Table of contents
- Topics covered in the book

6 Conclusion

This investigation found that some information useful to booksellers can be obtained from book characteristics. The recommendation system based on book title and author was successful in identifying books of similar genres and themes depicted from the title and books by the same author. A bookseller would be able to use this recommendation system to suggest new books to a user, based on one specific title they have read resulting in boosted sales. However, the results of this research also found that, based on the book characteristics, assertions about book ratings can be made with limited confidence. While this means that highly accurate predictions cannot be supplied to booksellers regarding ratings, there is the potential for more rigorous models which produce more accurate predictions to be built on a wider array of book features in future analysis.

7 References

Januzaj, Y., Luma, A. (2022). Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Word. International Journal of Emerging Technologies in Learning (iJET), 17, 258-268.

Knotzer, N. (2008). Recommender Systems: Functional Perspectives. In Product Recommendations in E-Commerce Retailing Applications (NED-New edition, pp. 47–78). Peter Lang AG. http://www.jstor.org/stable/j.ctv9hj934.5

Niupian, V., Chuaykhun, J. (2023). Book Recommendation System based on Course Descriptions using Cosine Similarity. ACM, New York, NY, USA, 273. https://doi.org/10.1145/3639233.3639335

Phillips, A., Clark, G., & Clark, G. (2019). Inside Book Publishing (6th ed.). Routledge. https://doi.org/10.4324/9781351265720

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, 205-227. https://doi.org/10.1016/j.eswa.2017.12.020

Roy, D., Dutta, M. (2022). A systematic review and research perspective on recommender systems. J Big Data (9), 59. https://doi.org/10.1186/s40537-022-00592-5

Sheelam, H. (2020). A Study on Marketing Strategies for Self-Published Authors through Online Platforms. Journal of Marketing Vistas, 10(2), 60-81.

Stewart, S. (2021). TikTok Booms: Books championed on BookTok have seen huge sales spikes. Publishers Weekly, 268(36).

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., & Barabási, A. (2019). Success in books: predicting book sales before publication. EPJ Data Science, 8. https://doi.org/10.1140/epjds/s13688-019-0208-6