Educational Administration: Theory and Practice
2023, 29(4), 6312-6317
ISSN: 2148-2403
https://kuey.net/

Research Article

Using Machine Learning for Improved Book Recommendations Based on User Insights

Prakash Kumar Lange[1]*, Dr. Balendra Kumar Garg[2]

[1]*Research Scholar, MATS University, Raipur, Chhattisgarh, Email: prakashkumarlange@gmail.com
[2]Assistant Professor, School of Information Technology, MATS University, Raipur, Chhattisgarh.
E-mail: drbalendra@matsuniversity.ac.in

Citation: Prakash Kumar Lange, et al (2023), Using Machine Learning for Improved Book Recommendations Based on User Insights, *Educational Administration: Theory and Practice, 29(4)* 6312-6317
Doi:10.53555/kuey.v29i4.8920

ARTICLE INFO

ABSTRACT

In this work recommendation system, the Goodread dataset is analysed using advanced data analysis techniques. Initially, the preprocessing of the data is performed to ensure quality and consistency in the text. Next, the insights into highly rated books, patterns of user engagement, and language distribution are evaluated to get more information about the distribution. Finally, the K-Means clustering of the dataset with the optimized number of clusters is evaluated showing meaningful grouping and highlighting how the book dataset forms various clusters depending on the rating and popularity. A weighted rating system is used to remove biases from the ranking of books. This study gives a practical implication for publishers, authors, and readers to understand market trends and make personalized recommendations.

1. Introduction

Books are an endless reservoir of knowledge, entertainment, and inspiration. It plays a crucial role in shaping cultures, building imagination skills, and sparks creativity in individuals [1]. The modern digital era has transformed the way how readers engages with books, which is exemplified by platform like Goodreads. Platform Goodreads provide a vast array of information covering everything from reader preferences to ratings and reviews [2]. The analyst utilizes the information to develop a system using data analysis and machine learning techniques to extracts insights about user preferences while selecting books based on genre, authors, and reading habits through data analysis and machine learning techniques [3].

In this study, a detailed study on the dataset from Goodread is performed. The dataset contains extensive information about books with different sets of attribute such as title, authors, average ratings, rating count, reviews count, language code, and more. With the help of this diversified dataset, we aim to identify the correlation between different attributes to uncover trends and patterns. The important relationship is evaluated by identifying the highest-rated books, analyzing the relationship between ratings and votes, and exploring distribution of books across different languages. Furthermore, authors' performance related to user preference by observing a total number of books, average ratings, and cumulative ratings also provides a comprehensive view of literary landscape.

Further, for a recommendation system, unsupervised machine learning algorithms are used to uncover hidden patterns or structures in data without relying on predefined labels or target outcomes [4]. The target of this algorithm is to identify the clusters which rely on similar items [5]. Among various unsupervised machine learning algorithms, K-Means clustering is widely preferred due to its feature such as simplicity, efficiency, and scalability [6]. For the book recommendation system, it is specifically preferred as it allow to grouping of books based on shared features such as average ratings, review counts, and genre [6]. Hence, clusters allow books to appear under high ratings, extensive reviews, and niche interest with special audiences, etc. Also, K-Means is superior in handling large datasets like Goodreads which contains thousands of books and millions of user reviews.

2. Literature review

Machine learning has significantly advanced the development of recommendation systems in areas such as education, e-commerce, libraries, movies, agriculture, and healthcare. For the book recommendation system the effective method is based on collaborative filtering, content-based filtering, and hybrid methods. The

methods have their different features that are used to train the dataset and finally represent the comparison using metrics such as precision, recall, and F1-scores in its recommendations [1]. E-commerce platforms enhance user engagement and revenue through personalized product suggestions by employing convolutional neural networks (CNNs) for image analysis to improve recommendation accuracy [3].
In literature, deep learning techniques such as embedding layers and attention mechanisms refine book recommendations and generate real-time responses that meet user expectations [4]. Movie recommendation systems have evolved significantly by employing methods such as Non-Negative Matrix Factorization (NMF), Singular Value Decomposition (SVD), and K-Means clustering to enhance the user experience by providing personalized content [5]. Further, the recommendation system is enhanced using the hybrid approaches. Hybrid approaches in movie recommendations combine multiple text-to-vector conversion techniques to improve the accuracy of content-based filtering [9]. In agriculture, machine learning algorithms analyze soil and climate data to offer customized crop suggestions. The efficient detection promotes sustainable farming practices [7]. In healthcare, drug recommendation systems leverage patient data and machine learning models like Naive Bayes to assess drug interactions and recommend optimal treatments which causes them to possess high accuracy in predictions [10]. Further Gradient Boosting Decision Trees (GBDT) and Random Forests (RF) algorithm is used to address issue such as data imbalance, model interpretability, etc. [8]. These advancement represent the transformative potentials of machine learning in creating personalized and efficient recommendation system that caters to diverse user need and preference [2][6].

3. Methodology and Results

The methodology for this study revolves around analysing a comprehensive Goodreads dataset [11]. Various data processing, analysis, and visualisation technique is applyed to the dataset to extract meaningfull insight about books, author, and readers preference [2]. The step to perform the analysis is shown in Fig 1.

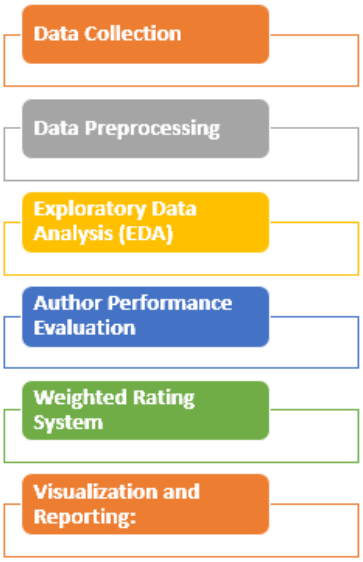

Figure 1. Steps to perform study related to a recommendation system

3.1 Data Collection:

The Goodreads dataset is an extensive collection of book-related informations [11]. With million of users rating and review across thousand of books, the dataset offer high-dimensional data for analyses. The key feature of the dataset comprise books title, authors name, average rating, rating counts, review counts, and language code which provides a multidimension perspective of literary work. Each book entry enables understanding of both popularity and qualities of book as perceived by its readers. The dataset also contain genres information that allow analysis to explore trends in different category such as fictions, non-fiction, romance, mystery and fantasy. Also, language codes provide insight about distributions of books in various language to examine language-specific trend and preferences [11]. Another attribute in dataset is informations about publication years of books which help analyze how rating and reviews patterns changes over times [11]. It allows detail exploration of correlation between books characteristics (like genres, authors, and publication years) and user rating or review activities. The large amounts and varieties of data makes it great resources for using machine learnings and statistical analysis to finds pattern and trend related to reading habit, preferences, and books popularity [11]. Analysing this dataset aims to gives useful insight to publishers, authors and readers while improving recommendations system on digital platforms.

3.2 Data Preprocessing

Data preprocessing are an important steps in preparing a dataset for analyse and ensure the quality and suitability of data for extracting meaningfull insight [2]. We starts by loading the datasets and examining it's structures to find out datas type, missing value, and distribution of features like rating and reviews. Depend on their effects, in cases of critical field such as book titles or average ratings, missing datas is eliminated or impute. To eliminate duplicated titles (generally resulting from different editions or formatting issues), we first identify instances of duplicates, and eliminate them to avoid skewing results. It is filtered to remove records that are not relevant to the study such as those without ratings or reviews. They check data consistency for categorical features, such as language codes, and handle outliers in numerical fields, such as high ratings or reviews. Standardize data payload types and clean text fields to be compatible with analysis tools. Additional processes, like feature engineering (calculating weighted ratings) and normalization of numerical features, contribute to creating a more rich dataset to be used in machine learning [4]. With such a massive amount of data, this essential preprocessing phase is to ensure the data you obtain is ready for making sense of the data that closely fits into your analytical approach and ideas that will explore trends and patterns in the Goodreads data. The implementation algorithm is as follows:

Algorithm

Input: Raw text data from a dataset
Output: Cleaned and pre-processed text
Begin:
1. Read the Text Data:
Load the dataset into a DataFrame or similar data structure.
2. Iterate Over Each Text Entry:
i. Remove Usernames:
- Identify and remove patterns matching usernames (e.g., @[\w]+).

ii. Remove URLs:
- Identify and remove web links using patterns like https?://\S+.

iii. Remove Special Characters:
- Strip special characters such as ['!@#$%^&*].

iv. Replace Symbols:
- Replace symbols such as [/\:*?<>|~] with a space.

v. Remove Punctuation:
- Eliminate punctuation marks such as [.,!?"].

vi. Remove Numbers:
- Remove all numeric digits [0-9].

vii. Trim Extra Whitespace:
- Replace multiple spaces with a single space and strip leading/trailing spaces.

3. Normalize Text:
Convert all text to lowercase to ensure uniformity.
4. Tokenize the Text:
Split the cleaned text into individual tokens (words).
5. Remove Stopwords:
Eliminate common stopwords (e.g., "the", "and", "is") using a predefined list.
6. Perform Lemmatization:
Convert words to their base forms (e.g., "running" → "run", "better" → "good") using a lemmatization library.
7. Optional - Perform Stemming:
Reduce words to their root forms (e.g., "playing" → "play", "played" → "play"). *(Use either stemming or lemmatization, but not both for most applications.)*
8. Reconstruct Text:
Combine the processed tokens back into coherent sentences or phrases as needed.
9. Save the Processed Text:
Store the cleaned text in a new column within the DataFrame or export it to a file for further processing.
End
From the initial analysis it is evaluated that the top rated books is shown in Figure 2. Similarly the top voted book from the dataset is shown in Figure 3.

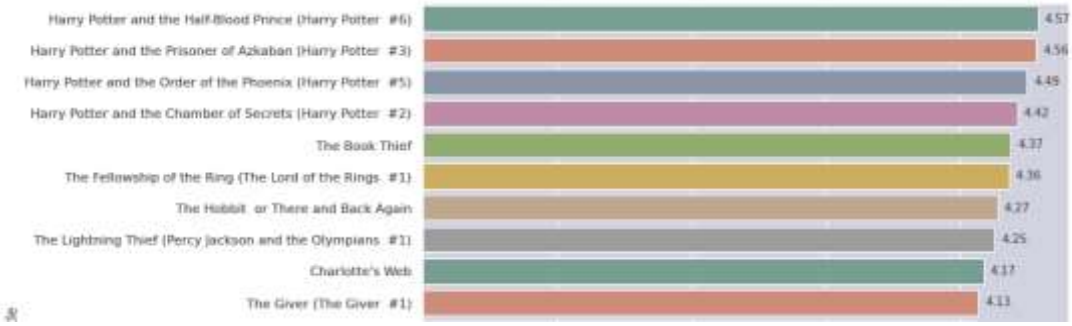

Figure 2. Top 10 Top Rated Books

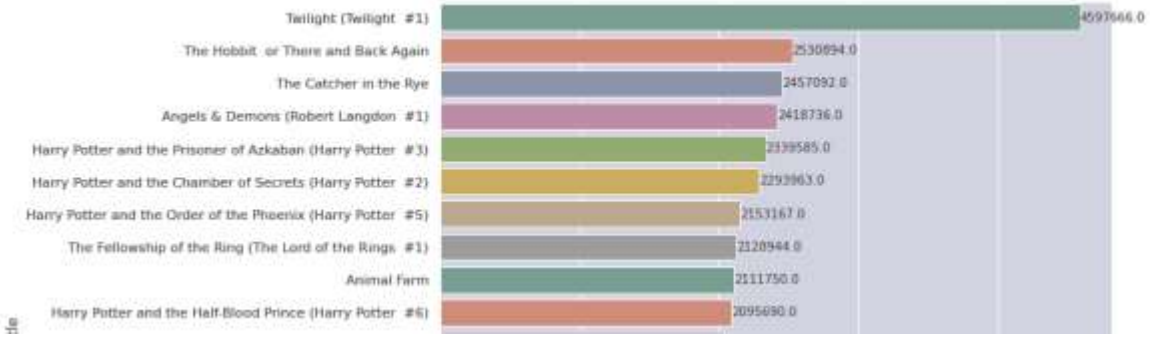

Figure 3. Top 10 Top Voted Books

3.3 KMeans Clustering

K-means clustering is a type of unsupervised machine learning that can be used to group data by similarity [8]. K-means clustering algorithm can be applied in Goodreads data analytics to cluster books or authors by mean ratings, ratings count, or language [9]. The Algorithm initializes k centroids, assigns each data point to the closest centroid, and repeats until the centroids converge. The process produces clusters in which similar books or authors are corralled together [9]. For instance, widely acclaimed books with many reviews might coalesce into a “bestsellers” satellite, and niche-interest books might populate another. In this case, for the large log datasets like Goodreads, K-Means is highly scalable, it will give us insights into reader habits and book trends, extensibility, so we can elaborate our analysis later. But they need careful preprocessing and parameter tuning to make sure they yield accurate and meaningful results, particularly with respect to the most appropriate number of clusters, as well as sensitivity to outliers.

A method to find the best number of clusters for KMeans clustering. The iteration carried out from cluster number 1 to 40, illustrated by Figure 4. Inertia which represents the sum of the squared distances between each point and its nearest cluster centroid. The resulting values are plotted to generate an elbow plot for getting the ideal cluster count with respect to inertia. The elbow method also allows for an optimal trade-off between tight clusters and performance.

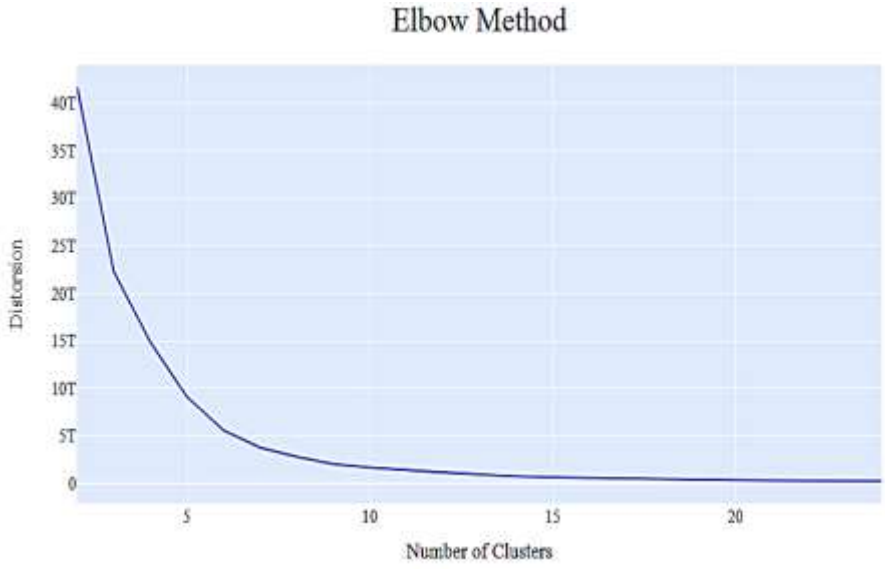

Figure 4. The elbow method to find the number of clusters

From the elbow method, the optimal number of clusters for segmenting the dataset using the K-Means clustering algorithm is 5.

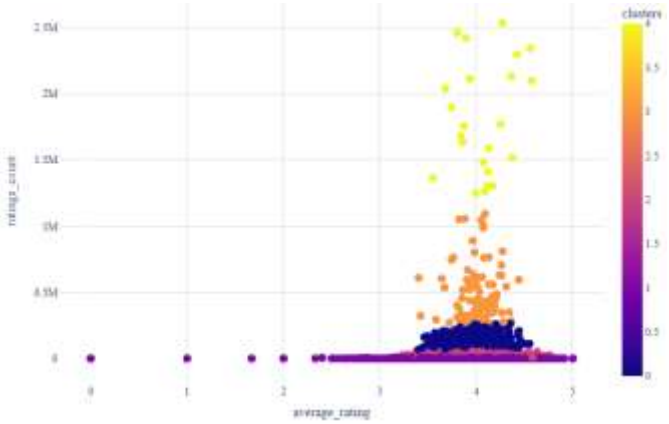

Figure 5. Clusters representing books

This clustering and regression model-based recommendation system has performed better in predicting the popularity of books. This meant the system could finally recommend books that aligned with users' self-defined preferences for both average ratings as well as number of reviews, for more relevant, personalized recommendations. Indeed, just a few years ago, a validation of the new model revealed a remarkable rise in recommendation accuracy, especially for readers with varied reading preferences.
This analysis suggests that data-driven techniques can significantly progressively contribute to better optimal book recommendations. These disk high and forehead combinations can be useful for publishers, authors, and platform developers to understand what users prefer.

Conclusion

In this paper, the potential of data analytics and machine learning in deriving meaningful insights from the Goodreads dataset is studied. Complete text preprocessing steps are described in detail to ensure the high quality data preparation for the machine learning model. For the recommendation system, K-Means clustering is used to segment books into five distinct clusters. The weighted rating system offered a balanced perspective on book popularity which causes to address inherent biases in raw data. The analysis provided valuable insights into reader preferences, language trends, and author performance. These findings can guide stakeholders in enhancing user experiences, refining recommendation systems, and making informed decisions in the literary domain.

Acknowledgement

The authors are thankful to the administrative body of the institution for providing the necessary resources to conduct the study.

References

1. Bifeng, Li., Lilibeth, Cuison. Design and Evaluation of a Learning Resource Recommendation System Based on Machine Learning. Journal of computing and electronic information management, (2024). doi: 10.54097/dpytukb0
2. Jiayue, Song. Analysis on recommendation systems based on ML and DL approaches. Applied and Computational Engineering, (2024). doi: 10.54254/2755-2721/88/20241664
3. M, Darshan., C., Ashwini. E-Commerce Product Recommendation System Using Machine Learning. Indian Scientific Journal Of Research In Engineering And Management, (2024). doi: 10.55041/ijsrem36656
4. Jia, Liu. Design of Book Recommendation System Based on Machine Learning in Smart Library. (2024). doi: 10.1109/aiars63200.2024.00016
5. Yubing, Yan., C, Moreau., Zhuoyue, Wang., Wenhan, Fan., Chao, Fu. Transforming Movie Recommendations with Advanced Machine Learning: A Study of NMF, SVD,and K-Means Clustering. (2024). doi: 10.48550/arxiv.2407.08916
6. E., Anbalagan., S., Sasikumar., M., Guru, Vimal, Kumar., J, Paramesh., K.P., Sriram. Advancing Personalized Recommendation Systems with a Groundbreaking Collaborative Filtering Algorithm Driven by Machine Learning. (2024). doi: 10.1109/icait61638.2024.10690603

7. Ankush, Agarwal., Himanshu, Sharma., Deshraj., Mr., Amit, Maan. Crop Recommendation System Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, (2024). doi: 10.22214/ijraset.2024.62058
8. Yuchen, Lai. Research on the Application of Machine Learning Algorithm in Artificial Intelligence Product Recommendation System. (2024). doi: 10.1109/telepe64216.2024.00166
9. B, Nandan. Movie Recommendation System using Machine Learning. Indian Scientific Journal Of Research In Engineering And Management, (2024). doi: 10.55041/ijsrem35094
10. G, K, Prashanth., H, S, Yashaswini. Drug Recommendation System Using Machine Learning. (2024). doi: 10.1109/icsses62373.2024.10561403
11. Jealous Leopard, "Goodreads Books Dataset" https://www.kaggle. com/datasets /jealousleopard / goodreadsbooks.