Machine Learning Techniques for Bookstore Inventory Management: Decision Trees for Book Rating Prediction and K-Means Clustering for Book Segmentation

Group Name: **W19G6** – COMP20008 – University of Melbourne
Workshop time: Thursday 12:00 p.m to 1:30 p.m.
Miles (Yueming) Li – 1450710 – yuemingl3@student.unimelb.edu.au
Skylar (Kyi Shin Khant) – 1450754 – kyishink@student.unimelb.edu.au
Ngoc Thanh Lam Nguyen (Lam) – 1450800 - ngocthanhlam@student.unimelb.edu.au

Abstraction

Machine learning techniques based on data analysis play a crucial role in enhancing business decision-making processes. This study focuses on the application of these techniques in the context of a bookstore, utilizing two main methodologies: decision tree analysis and K-Means clustering. In practice scenario, our decision tree makes prediction of a new book's rating possible for bookstores and their customers when making decision on purchases. Also, customers can retrieve a cluster of highly rated books which will be of their interest based on the title of the book with the use our K-Mean clustering system.

The primary objective is to demonstrate how these machine learning methods can be effectively utilized to support decision-making and improve business management practices, including customer recommendations. The study begins with a thorough exploration of data pre-processing techniques, which are essential for ensuring the quality and integrity of the data used in the analysis. Subsequently, the implementation of decision tree and K-Means clustering algorithms is detailed, illustrating how these methods are developed and utilized to achieve the desired outcomes in the context of the bookstore. Following the application of these methodologies, the collected data is evaluated and interpreted to derive meaningful insights.

Key findings from the analysis highlight the effectiveness of machine learning techniques in predicting customer preferences and optimizing inventory management strategies. However, the study also identifies limitations and areas for improvement based on the results obtained, shedding light on the potential and challenges of implementing machine learning in the context of a bookstore.

Overall, this study highlights the significant potential of machine learning techniques in enhancing decision-making processes in the bookstore industry. It also emphasizes the importance of thoughtful data pre-processing and the need for continued research and development to address the limitations and challenges associated with implementing machine learning in this context.

Introduction

Machine learning has proven to be a powerful tool for enhancing business strategies. By harnessing vast datasets, businesses can derive valuable insights to drive decision-making. In the context of bookstore management, where inventory optimization and customer segmentation are paramount, the application of machine learning holds immense promise for enhancing operational efficiency and driving business growth.

In this project, we aim to leverage machine learning techniques for rating prediction and customer segmentation, offering significant benefits for bookstore management and strategic decision-making.

The success of machine learning models hinges on the quality and reliability of the underlying data. To ensure the robustness of our analysis, we employ a rigorous data preprocessing pipeline encompassing imputation, text processing, and discretization. Imputation techniques are used to handle missing data in user ages and locations, ensuring the completeness of the dataset. Text processing methods standardize

country and author names, facilitating uniformity and consistency in the data. Discretization enables the grouping of data into meaningful categories, such as age groups and rating categories, simplifying the analysis and interpretation of results.

This project focuses on two machine learning techniques: decision tree and K-Means clustering. Decision tree prediction is employed to predict the rating category of books, aiding in inventory management decisions and customer targeting strategies. By selecting features with the highest information gain, including "User-City," "Book-Author," and "Book-Title," we aim to develop a robust predictive model capable of accurately categorizing books based on user preferences. Cross-validation is employed to evaluate the decision tree, with a particular focus on addressing challenges related to imbalanced data distribution in ratings.

The K-Means clustering is utilized to segment books based on their titles, enabling personalized recommendations for customers. Leveraging the Bag-of-Words (BoW) method, we transform textual information into numerical representations, facilitating clustering analysis. The elbow method is employed to determine the optimal number of clusters, ensuring the creation of cohesive and distinct book clusters. By recommending books with similar themes or content preferences, we aim to enhance the browsing and purchasing experience for customers, driving engagement.

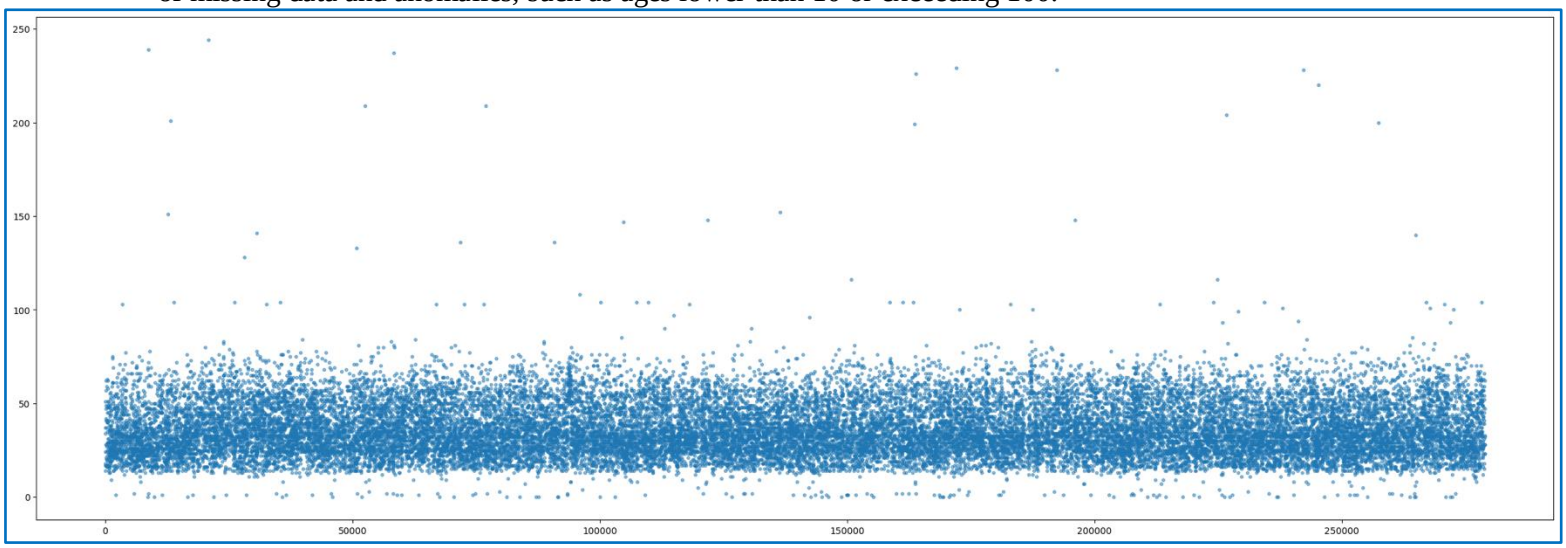
Methodology

1. Data Preprocessing

The raw dataset is extensive, containing numerous instances of missing and anomalous data, making it unsuitable for direct machine learning applications. To address these issues, a comprehensive data preprocessing strategy is employed, focusing on imputation, text processing, and data discretization. This approach aims to rectify missing values and anomalies in user ages and locations, standardize textual data like country names and author names, and discretize age groups and ratings. These preprocessing steps are vital for preparing the dataset for machine learning models, ensuring that the data is clean, consistent, and ready for analysis.

1.1. Missing Data Handling – Imputation

One of the critical datasets for our machine learning models in this project pertains to users' ages. Upon analyzing the distribution of users' ages in the raw data, it became evident that there were instances of missing data and anomalies, such as ages lower than 10 or exceeding 100.



To address this issue, abnormal data points will be removed, and the missing data will be randomly imputed with seed number 42 for consistency based on the distribution of the raw data.

1.1.1. User Age

For age data, after the removal of abnormal values, to impute missing or extreme age values, a list of random ages is generated based on the distribution of age groups in the dataset. The number of random ages generated is equal to the total number of rows with null or extreme age values. These random ages are then used to fill the missing or extreme age values in the dataset, ensuring that the imputed ages are realistic and in line with the overall age distribution of the users. This approach ensures that the imputed ages are realistic and representative of the overall age distribution in the dataset. Furthermore, to facilitate analysis, the ages will be discretized into age groups. This step categorizes the ages into meaningful intervals (10 years age as a bin), making it easier to interpret and analyze age-related patterns in the data.

1.1.2. User Location

Location data imputation involves several steps to ensure missing values are filled appropriately. First, all unique countries from the dataset are listed, and rows where the country is either "n/a" or null are identified. For each missing country, a random choice is made from the list of known countries, ensuring that only valid countries are selected. This process avoids introducing new, erroneous country names into the dataset.

Similarly, for city imputation, all unique cities from the dataset are listed, and rows where the city is null are identified. Random cities are then chosen to fill these gaps, ensuring that the selected cities are valid and maintaining the integrity of the dataset.

Lastly, for state imputation, all unique states from the dataset are listed, and rows where the state is either "nan" or an empty string are identified. Random states are chosen to fill these gaps, ensuring that the selected states are valid and maintaining the integrity of the dataset.

These imputation processes help ensure that missing location data is filled with valid and realistic values, enhancing the quality of the dataset for further analysis and modeling.

1.2. Text Processing

Text processing is employed to standardize country names and author information in the dataset. Given the dataset's size, manual correction of all variations in country names and author spellings would be labor-intensive. Therefore, an automated approach is taken. Popular countries, such as the United States and the United Kingdom, which are often written in various forms, are mapped to their standard representations.

For countries with missing or inappropriate names, they are replaced with valid country names from the dataset or a predefined list available in the “pycountry” library. This replacement strategy ensures that all country names are standardized and appropriate for further analysis. Additionally, all country names are converted to lowercase to remove case sensitivity and ensure uniformity in the dataset.

Similarly, for author names, characters like periods, commas, single quotes, and extra spaces are removed to standardize their representation and avoid case sensitivity issues. This preprocessing step ensures that the author names are clean and consistent, facilitating easier analysis and comparison across the dataset.

1.3. Data integration and Discretization

Ages and ratings are discretized into meaningful categories to enhance the interpretability, robustness, and effectiveness of the predictive model based on user and book attributes. This preprocessing step is crucial before constructing a decision tree, as it simplifies the relationship between the features and the target variable.

For ages, the dataset is divided into age groups using predefined bins and corresponding labels. This discretization allows for easier interpretation of age-related patterns in the data. The histogram visualizes the distribution of age groups, providing insights into the age demographics of the users. Similarly, ratings are discretized into rating categories to simplify the analysis of user ratings.

After discretizing ages and potentially ratings, the datasets are merged and imputed for missing city, state, and country values. The final dataset is then checked for any remaining missing values to ensure that it is ready for further analysis or modeling.

2. Decision Tree Prediction System for Book Store

2.1. Feature Selection

Initially, we computed the Information Gain for each feature to assess its correlation regarding the target variable (Rating Category). By Calculating the entropy of each feature and utilizing Information Gain as a metric, it is identified the features which are most discriminative for the training of the decision tree.

2.2. Data Preparation

From the computation of information gain, we selected three features with the highest information gain as input features for the decision tree model. These features include the users' cities ("User-City", or bookstores' cities that we assume these bookstores are physical stores), books' authors ("Book-Author"), and books' titles ("Book-Title"). We utilized "OrdinalEncoder" to encode these features, enabling their transformation into a numerical format that the model could process.

2.3. Model Training

To train the model, we employed an entropy-based decision tree model to conduct the training process. This model selects the optimal splitting features at each node to signify information gain and generates a tree structure for classifying samples.

2.4. Cross Validation

To validate the model's training process and evaluate its performance and accuracy, we employed a ten-fold cross-validation. This approach partitions the dataset into ten subsets, utilizing nine subsets for training in each iteration and evaluating on the remaining subset. We computed the score from each cross-validation iteration and derived its average as the final model score.

2.5. Evaluation

We also assessed the training model using a test set from new datasets and computed its accuracy. Additionally, we constructed a confusion matrix to future evaluation for the model's performance, including metrics such as true positive rate, true negative rate, false positive rate, and false negative rate.

3. K-Means Clustering for Customer Segmentation

3.1. Selection of K-means Clustering

K-means clustering is chosen as the clustering algorithm due to its computational efficiency and its ability to generate well-defined clusters with clear boundaries, making interpretation straightforward.

3.2. Encoding of Book Titles

To apply K-means clustering, book titles need to be encoded into numerical representations. The Bag-of-Words (BoW) technique, a fundamental method in Natural Language Processing (NLP), is utilized for this purpose. Given that only book titles, rather than the entirety of the book content, will be embedded, there is a deliberate emphasis on frequency count over semantic meaning or contextual information. Consequently, the Bag-of-Words (BoW) methodology is perceived as a simpler and more time-efficient approach, prioritizing computational speed over semantic depth.

3.3. Preprocessing of Book Titles

Prior to vectorization, preprocessing of book titles is conducted. This involves removing punctuation, stop-words, and converting all text to lowercase, ensuring uniformity and consistency in the data.

3.4. Execution of Elbow method and K-means Clustering

The optimal number of clusters (K) is determined using the elbow method. Given the presence of duplicate entries in the dataset, only unique books are considered for this analysis. Once the optimal K value is obtained, K-means clustering is executed. This step partitions the dataset into K clusters, with each cluster containing books with similar titles.

3.5. Recommendation Strategy: Selection of High-Rated Books within Clusters

While clustering provides a means to group similar books, recommending an entire cluster may overwhelm the user. Hence, a refinement is made by filtering out highly rated books within each cluster.

Data Exploration and Analysis

1. Decision Tree Prediction System for Book Store

1.1. Feature Importance

We initiated our analysis by computing the information gain (IG) for each feature to discern their significance in predicting the rating category and correlation between them and rating category. The calculated IG values for each feature are as follows (in 4 decimal places):

User-City	0.1038
User-State	0.0166
User-Country	0.0022
Age-Group	0.0008
Book-Author	0.0751
Year-Of-Publication	0.0027
Book-Publisher	0.0231
Book-Title	0.1435

Based on these IG values, we selected the top three features with the highest IG scores for model training: “User-City”, “Book-Author”, and “Book-Title”, which could help to increase the accuracy of our model (Bailey, 2024).

1.2. Model Validation

To evaluate the performance of the decision tree model, we conducted ten-fold cross-validation. The cross-validation scores obtained as follows (4 decimal places):

0.6655
0.5969
0.5639
0.6198
0.6319
0.6361
0.6506
0.6432
0.6497
0.6536

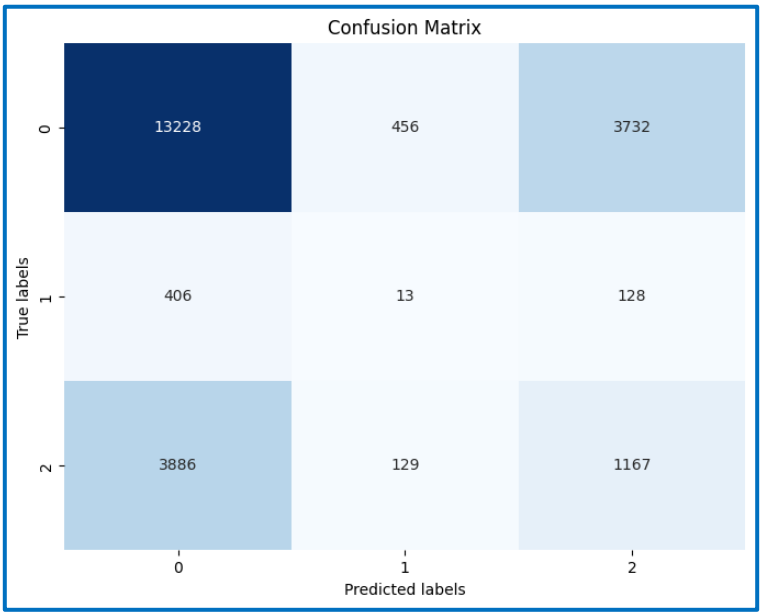
The average score was 0.6311.

The cross-validation scores demonstrate stability across folds in performance. However, the observed average score suggests room for improvement in the model’s accuracy.

1.3. Confusion Matrix Analysis

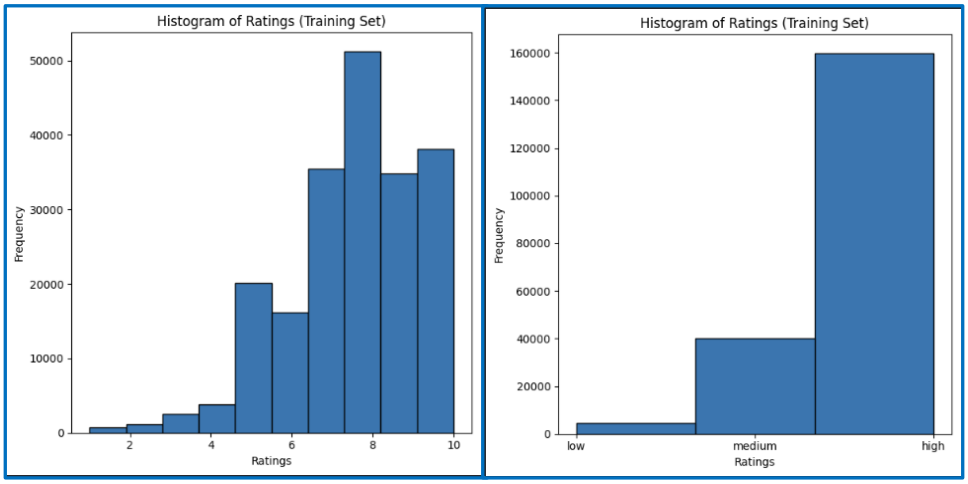
Upon scrutinizing the confusion matrix, we observed the following results (Labels represent the rating categories, 0: high, 1: medium, 2: low).

The confusion matrix revealed a notable discrepancy in accuracy of predicting high ratings compared to medium and low ratings. This imbalance may have contributed to the diminished precision for medium and low rating categories.



1.4. Rating Distribution Analysis

Further analysis of the distribution of Book-Rating and Rating-Category covered a skew towards high ratings, indicating data imbalance. This imbalance impacted the model's performance, leading to a suboptimal average cross-validation score and potentially affecting accuracy on the formal test set evaluation.



2. K-Means Clustering for Customer Segmentation

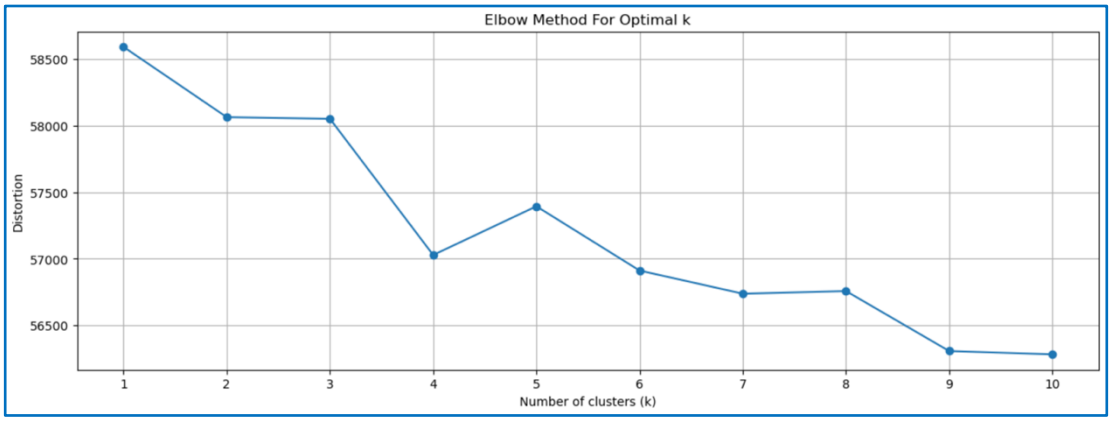
The foundation of the book recommendation system relies heavily on the examination and analysis of book titles. In this section, we delve into the exploration and analysis of the dataset, focusing primarily on the "Book-Titles" column within our merged data frame.

2.1. Duplicate Data

Initially, our dataset contained many duplicate book titles, totaling 204,164 entries. To streamline our analysis and ensure the integrity of our results, these duplicates were removed, resulting in a reduced dataset comprising 15,976 unique book titles. Subsequently, the dataset was prepared for ingestion into the Bag-of-Words (BoW) method.

2.2. Elbow Method and Clustered Data Distortion

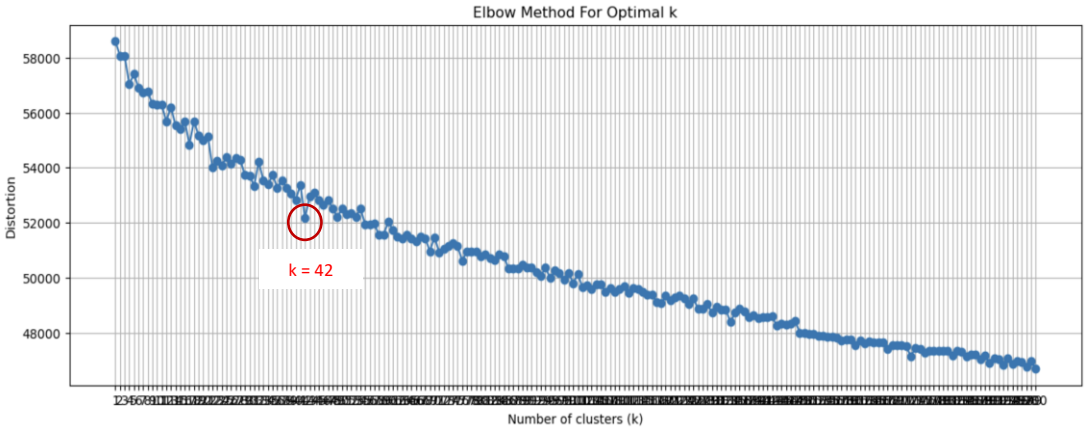
To determine the optimal number of clusters (K) for k-means clustering, initial exploration involved plotting the distortion for K clusters ranging from 1 to 10. However, the resulting plot did not exhibit a distinct "elbow," complicating the application of the elbow method to identify the optimal K value.



K range: 1-10

2.3. Expanded K Range Analysis

In response to the inconclusive results from the initial exploration, further analysis was conducted by extending the range of K clusters considered. Ranges from 1 to 50, 1 to 100, 1 to 150, and 1 to 200



were explored successively. Each range was associated with execution times of approximately 1, 3, 6, and 8 minutes, respectively.

K range: 1-200

2.4. Identification of Optimal K for Data Clustering

Despite the smooth downward trend observed in the distortion plot, a notable dip was identified at K=42, suggesting a potential candidate for the optimal number of clusters. Given this observation, subsequent iterations of the K-Means method were executed with a fixed “random_state” parameter set to 42. This approach ensures consistency in the positioning of K centroids across different runs of the clustering algorithm.

Results

1. Decision Tree Prediction System for Book Store

According to evaluation with the test set comprised of data relevant to new books, our decision tree model exhibited a final accuracy of 0.6215, aligning closely with our initial expectations.

1.1. Implications and Interpretation

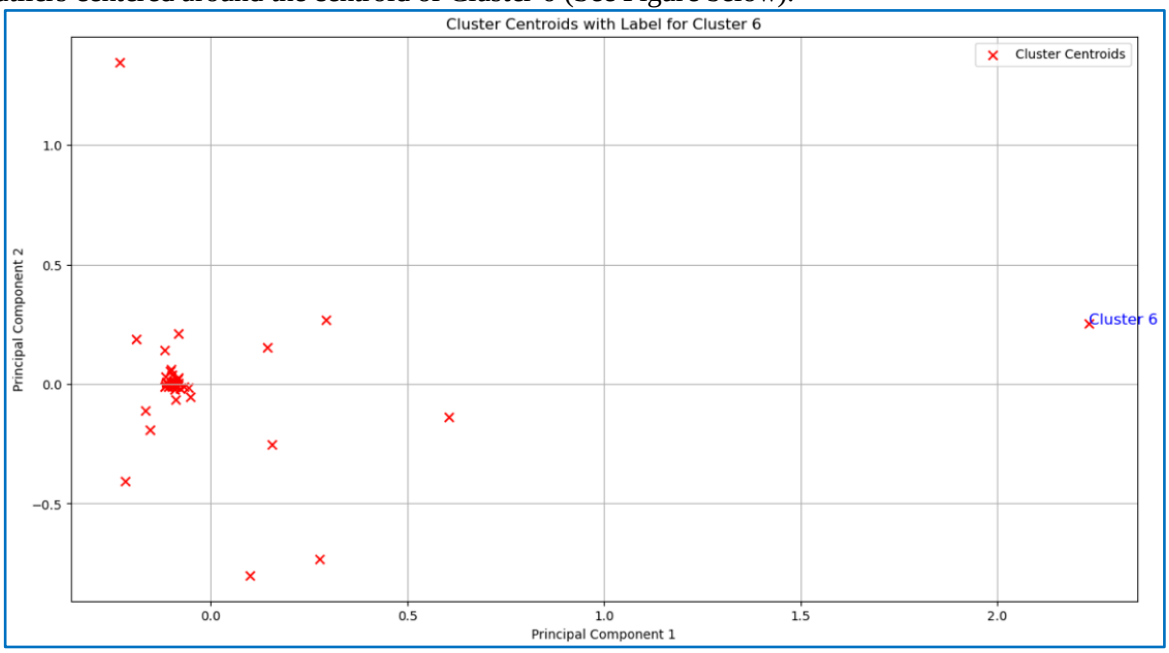
The obtained accuracy rate of 0.6215 highlights the challenges inherent in predicting book ratings accurately. The model's performance suggests potential limitations in its ability to generalize unseen data, indicating areas for future investigation and refinement.

2. K-Means Clustering for Customer Segmentation

Upon conducting the clustering analysis, it was evident that our dataset was partitioned into 42 distinct clusters, each containing an appropriate number of books. Furthermore, the process of filtering out highly rated books from each cluster yields a significant reduction, approximately 10 percent, in the overall number of books within the cluster.

2.1. Noise or Outlier

Notably, Cluster 6 emerged as the cluster with the largest number of books compared to the other clusters. As illustrated in the Figure below, this abundance of books could be attributed to the presence of outliers centered around the centroid of Cluster 6 (See Figure below).

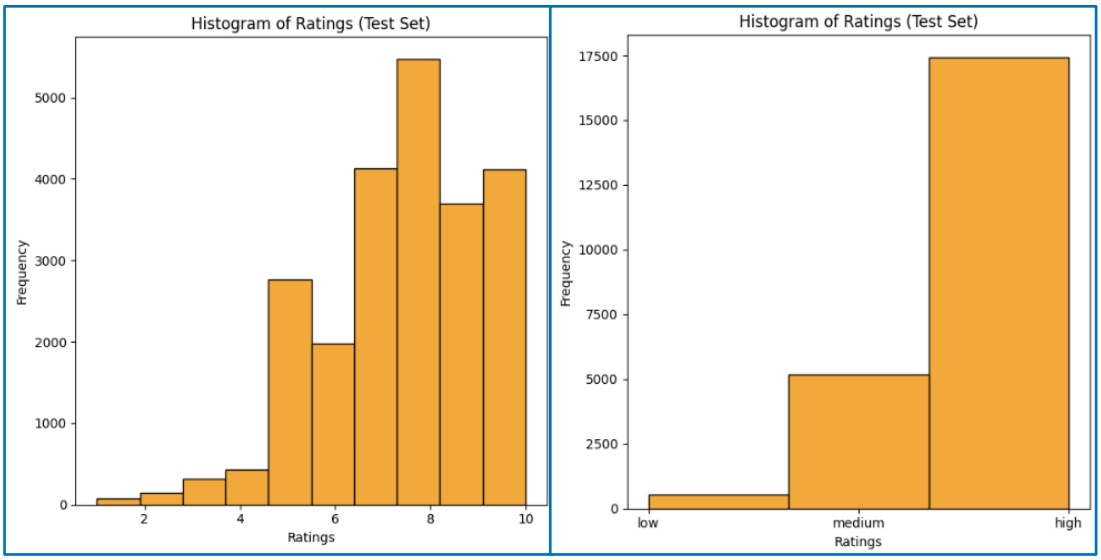


Discuss and Interpretation

1. Decision Tree Prediction System for Book Store

1.1. Distribution Analysis

We conducted a visualisation of the distribution of ratings and rating categories in the test set to understand the discrepancies in the decision tree model's accuracy. The analysis revealed that the distribution of rating and rating categories in the test set mirrored that of the training set.



This imbalance in data distribution contributed to the observed similarity between final accuracy and cross-validation scores, both of which were relatively low.

1.2. Feature Selection

The rationale behind selecting features with higher information gain was to priorities those that exhibited strong predictive power (Bailey, 2024) for the rating categories. By considering features such as "User-City", "Book-Author", and "Book-Title" with higher information gain values, we aimed to enhance the model's ability to discern patterns and make accurate predictions.

1.3. Real-world Application

In practice scenarios, where bookstores operate as physical entities, our decision tree model could serve as a valuable tool for inventory management. Bookstores can input information such as the title and author of a new book, and location of buyers into the system to predict their ratings, adding in the decision-making process regarding book acquisition. This application demonstrates the potential utility of our system in real-world settings, empowering bookstores to make informed decisions based on predictive analytics.

2. K-Means Clustering for Customer Segmentation

2.1. Cluster Composition Analysis

The analysis of cluster composition reveals a significant degree of cohesion within each cluster, with books sharing common themes or genres being grouped together. Notably, the clustering algorithm successfully groups together books with similar characteristics, such as the consolidation of "Star Wars" books within a single cluster. This clustering approach enhances the alignment of recommendations with user preferences, as users interested in specific genres or themes are more likely to find relevant suggestions within their respective clusters.

2.2. Exploration of Outlier Cluster

Cluster 6 emerges as a notable outlier within the dataset, characterized by a diverse array of book titles exhibiting considerable heterogeneity. Upon closer examination, it becomes evident that these titles encompass a multitude of distinct and sometimes nonsensical words, contributing to the uniqueness of this cluster. Consequently, the resulting feature space may exhibit sparsity, with numerous dimensions containing zero values. K-means clustering algorithm, however, identifies subtle similarities within this sparse feature space, facilitating the grouping of books within Cluster 6.

2.3. Impact of Highly Rated Book Filtering

The filtering mechanism to only include highly rated books ensures that recommended books presented to users are of exceptional quality and are finely tailored to their preferences. By prioritizing highly rated content, the recommendation system upholds standards of excellence in delivering suggestions.

Limitations and Improvements

1. Data Pre-Processing

It is essential to note that the missing data might not be random or insignificant. Therefore, simple imputation methods such as using the mode, median, or mean may not be appropriate. In such cases, more sophisticated imputation techniques, such as multiple imputation or machine learning-based imputation, may be more suitable for handling missing data. These methods can better capture the underlying patterns in the data and provide more accurate imputations for missing values, ultimately leading to a more reliable and robust analysis.

2. Decision Tree Prediction System for Book Store

2.1. Limitations of Our Decision Tree Model

One notable limitation of our decision tree model stems from the imbalanced distribution of ratings in the training set, which predominantly favors high-rated books. This imbalance negatively impacts the model's accuracy, particularly in predicting medium and low-rated books.

2.2. Potential Improvements

2.2.1. Data Augmentation

To reduce the impact of imbalanced training data, one approach is to collect more additional data on medium and low-rated books. By augmenting the dataset with more diverse samples, the model can learn from a broader range of instances and improve its ability to generalize across different rating categories.

2.2.2. Resampling Techniques

Alternatively, resampling techniques such as under sampling can be employed to balance the distribution of rating categories in the training data. Under sampling involves reducing the number of instances in the majority class (high-rated books). This technique aims to create a more equitable distribution of samples across rating categories to improve the model's performance.

2.3. Consideration of Alternative Models

2.3.1. Model Comparison

While decision trees offer interpretability and ease of implementation, they may not always yield the highest accuracy, particularly in the presence of imbalanced data. Therefore, it is shrewd to consider alternative modeling approaches that may offer improved performance in predicting book ratings.

2.3.2. Ensemble Methods

Ensemble methods such as Random Forests. They combine the predictions of multiple individual models to produce a more robust and accurate prediction (Ao et al., 2019, p. 777). Ensemble methods are known to handle imbalanced data more effectively and often outperform decision trees in terms of predictive accuracy (Ali et al., 2012, pp. 277-278).

3. K-Means Clustering for Customer Segmentation

3.1. Limitation: Sole Reliance on Book Titles

A notable limitation of our project lies in the exclusive utilization of book titles for embedding and clustering purposes. While this approach serves as a convenient starting point, it may not yield highly accurate results due to the inherent complexity and subtlety of book titles. The nuances embedded within book titles can pose challenges in accurately capturing the underlying themes or content of each book.

3.2. Recommendation for Improvement: Incorporation of Book Content Analysis

To address this limitation and enhance the accuracy of clustering, we propose a shift towards analyzing the content of each book in addition to its title. By employing vectorization on the content, the model can detect more comprehensive and conceptual similarities between books that may not be apparent from book titles alone, thereby improving the quality and precision of the clustering outcomes.

3.3. Consideration of Encoding Methods

While the Bag-of-Words (BoW) technique serves as a viable method for converting text into numerical values, it has its limitations, particularly in capturing the significance of rare words. As an alternative, we recommend exploring the Term Frequency-Inverse Document Frequency (tf-idf) encoding method, which offers a more nuanced representation of textual data. Despite its slightly higher complexity, tf-idf encoding is also computationally efficient and provides a more accurate reflection of word importance within the document corpus.

3.4. Consideration of Clustering Methods

Despite the effectiveness of K-means clustering in our analysis, a challenge persists in determining the optimal number of clusters (K). The process of selecting K values can be somewhat ambiguous and may not always result in the most suitable clustering outcome. To mitigate this issue, an alternative approach worth considering is hierarchical clustering. Unlike K-means, hierarchical clustering excels in handling outliers and noise within the dataset. However, it comes at the expense of computational efficiency, as it requires more extensive computations to construct the hierarchical structure of clusters.

Conclusion

In summary, this research showcases the practical applications of machine learning, specifically through decision trees and K-means clustering, in revolutionizing bookstore management strategies and customer interactions. The study emphasizes the critical role of meticulous data preprocessing in ensuring the success of these machine learning algorithms. While the decision tree model exhibits promising outcomes, there are evident areas for enhancement, particularly in its ability to handle novel data instances effectively. Similarly, the K-means clustering approach, while effective, encounters challenges related to outliers and noisy data points.

Moving forward, future advancements in machine learning for bookstore management could focus on addressing data imbalances, refining outlier management strategies, and exploring novel techniques such as ensemble learning and deep learning architectures. By proactively addressing these challenges and embracing innovative methodologies, we can further elevate the efficacy and accuracy of machine learning models in the bookstore industry.

Reference List

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). *Random forests and decision trees*. International Journal of Computer Science Issues (IJCSI), *9*(5), 272.

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). *The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling*. Journal of Petroleum Science and Engineering, *174*, 776-789.

Bailey, J. "*Supervised Learning Cont. Linear Regression and Experimental Design*". Lecture Slides. University of Melbourne, 2024.