COMP20008 - Elements of Data Processing (Semester 1, 2024)

Assignment 2 – Who else likes this book?

Group: W03G3

Group Members: Brian Liew We(1568328), Seah Ding Xuan(1562767), Trieu Khai Luu(1462027)

1. Executive Summary

This report will present the findings of an analysis conducted to predict book sales through book ratings given by users, based on features such as user demographics, book ratings, and author characteristics. The objective will be to develop machine learning models that can accurately classify whether a book will sell well or not, identifying the best performing model and using data visualisation to help bookstore managers understand the key factors they should look into to optimise their inventory and increase sales.

Key findings include that the random forest algorithm has outperformed the K-nearest-neighbours algorithm in terms of adjusted R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). Year of book publication, user age and publisher name with the word “book” were identified as the most influential features in predicting book sales. Recommendations for bookstore managers include methods such as

1. Analysing content from the high-rated era of 1930-1935 to identify popular themes, genres, and writing styles to inform future acquisitions and development
2. Paying closer attention to the preferences and feedback of readers within the age bracket of 32-37 years old
3. Exploring the publishers with the publisher name “book” to identify specific factors contributing to their success.

The report will conclude by identifying the possible areas for improvement and limitations of the project.

2. Introduction

The scope of the data analysis involves examining three datasets: 'BX-Books.csv', 'BX-Users.csv', and 'BX-Ratings.csv', to gain insights into factors influencing book sales and customer satisfaction in a bookstore. The objective is to identify key factors influencing book sales and user satisfaction in the online bookstore. This includes understanding the relationship between book characteristics, user demographics, and user ratings. By analysing these factors, the aim is to provide actionable insights to managers for making data-driven decisions on which books to purchase and avoid.

3. Methodology

The methodology employed for analysis involves the following steps:

Exploratory Data Analysis: Initial investigations on the data to discover patterns and spot anomalies/inconsistencies in the data for both the features as well as the target variable, Book_Rating. Distribution of the target variable is also analysed and found to have an imbalance in distribution. Data is cleaned, including the removal of trailing characters as well as missing values are not systematically related to any other variables in the dataset. The 3 csv files are also merged using an inner-join to ensure only records with matching ISBN and User-ID values in both dataframes are included in the resulting dataframe, merged_data. Identifiers such as User-ID were also removed as it does not add any predictive power to the models, and may instead introduce noise. Outliers for numerical variables were observed through boxplots and removed accordingly.

Data Preparation: The dataset was cleaned and preprocessed further, where User-City and User-State were dropped. The User-Country column offers a more generalised representation of the user's location, hence by focusing on just User-Country we eliminate redundancy while still preserving essential information about the user's geographical locations. ISBN was also dropped as they also only serve as identifiers, similar to User-ID. Rows for User-Country which does not provide any useful geographical input to the machine learning task was also removed. Train-test split of the dataset was done here, before any further preprocessing steps, to prevent data leakage. Text was also preprocessed to focus on only the essential content of the text, and techniques such as tokenization, removal of punctuation marks and stopwords, lemmatization were used. TF-IDF vectorisation is applied here, where the textual data for "Book-Title", "Book-Author", "Book-Publisher" is converted into numerical vectors to be used as input features into the machine learning model. Categorical variable User-Country is then encoded using one-hot encoding. Aggregation was also done to the User-Country column to reduce dimensionality while retaining the geographical information. Numerical variables were feature scaled using Min-Max scaling to prevent large units from dominating features with smaller units. Lastly, the TF-IDF matrices for train set and test set respectively are combined with the other columns in the dataframes.

Model Training: Two machine learning algorithms, K-nearest-neighbours and Random Forest, were trained on the dataset to predict book sales. The best model is selected here, based on model evaluation metrics used in the section below.

Model Evaluation: The performance of the models was evaluated using metrics such as adjusted R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). Feature importance analysis was also conducted on the best performing model (Random Forest) to identify the most influential features in predicting book sales. The top 3 features were picked and further analysed using data visualisations to further understand and provide insights for the strategy for the managers of the bookstore, of which findings will be discussed in detail below.

Areas of Improvement and Limitations: Areas of improvement for the entire project was looked into, and 2 key areas of improvement were identified to further enhance the predictive performance of the machine learning models - the usage of hyperparameter tuning and checking for over/underfitting using learning curves. Limitations of the project include the limited set of features which does not really capture the complexities of the book_rating, hence book sales as well as computational constraints which lead us to be unable to deploy high-computation resource models for text processing, hence potentially not fully capturing the semantic relationships within the data and potentially impacting the accuracy and depth of our analysis.

4. Data Exploration and Analysis

We start off data exploration by looking into patterns and spot anomalies/inconsistencies in our data, and we will first take a look at our target variable, Book_Rating in ratings_df.

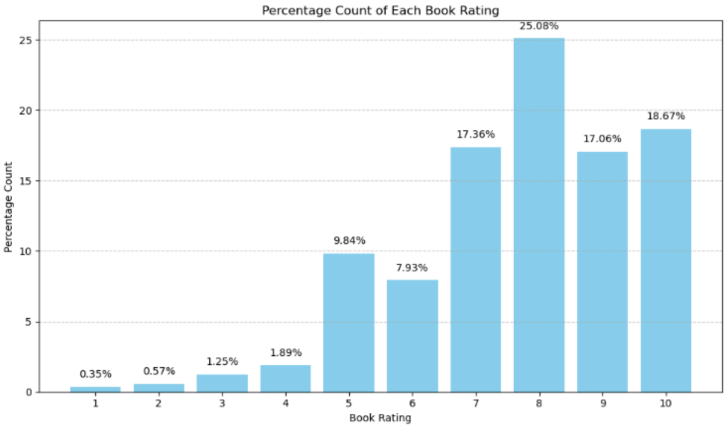
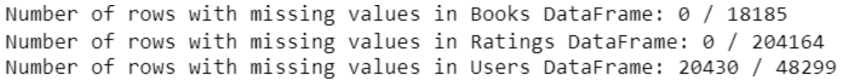


Figure 1: Distribution (Percentage Count) of Target Variable, Book_Rating

One notable observation is that the distribution of ratings across the different categories is largely imbalanced. For instance, Book-Ratings = "1", "2", "3" together account for less than 3% of the whole dataset, while ratings "8", "9", "10" together account for over 60% of the dataset. This imbalance is something notable which can affect the performance of the machine learning models (Aguiar, 2024).

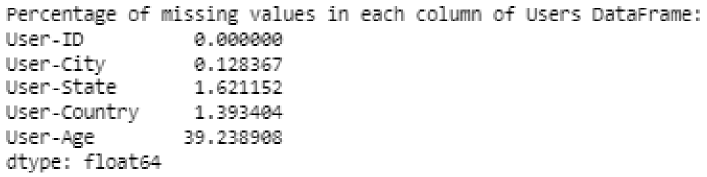
During the inspection of the dataset, trailing " were found in the users_df in the User-Country and User-Age columns. These inconsistencies could potentially affect the data analysis, hence a data preprocessing step to remove these extraneous characters was carried out. By removing the trailing " from these columns, we ensure that the data is clean and consistent, which is essential for accurate analysis and interpretation.



```
Number of rows with missing values in Books DataFrame: 0 / 18185
Number of rows with missing values in Ratings DataFrame: 0 / 204164
Number of rows with missing values in Users DataFrame: 20430 / 48299
```

Figure 2: Missing values in the dataframes

Checking for missing values (considering those which have empty fields as well as those which contain the string "n/a"), we can see that there are 0 missing values in books_df and ratings_df, but 20430 missing values in the users_df.



```
Percentage of missing values in each column of Users DataFrame:
User-ID          0.000000
User-City        0.128367
User-State       1.621152
User-Country     1.393404
User-Age        39.238908
dtype: float64
```

Figure 3: Missing Values in the Users Dataframe

Upon inspecting the users_df dataset, we can see that there are no missing values for the User-ID column. For the User-City, User-State, User-Country and User-Age columns, the percentage of missing values is 0.128367%, 1.621152%, 1.393404% and 39.238908% respectively. To handle these missing values, we first look into what type of missing data it is. From observation, the probability of the missing data is unrelated to any other variables and is unrelated to the variable with missing values itself, hence missing values are missing completely at random (MCAR).

Firstly, we look into rows with missing values in the User-Age field. Despite the large percentage of missing values, we choose to remove these rows because the missing values are not systematically related to any other variables in the dataset. Therefore, removing these rows is unlikely to bias our analysis, and it allows us to work with a cleaner dataset without imputing potentially unreliable data. Secondly, we look into the rows where User-City, User-State and User-Country are missing. Since the percentage of missing values for User-City, User-State, User-Country is relatively low with missing percentages of 0.085188%, 1.915017% and 2.283027% respectively, we decided to remove the corresponding rows. The missing values are negligible and removing them would not significantly impact the overall dataset. We will also be removing rows with Book-Publisher = 'Not Avail'. Retaining rows with incomplete location information could introduce bias or inaccuracies in subsequent analysis. Therefore, to ensure data quality and reliability, we opted to remove these rows with missing values in the columns.

In the initial preprocessing step, we merge the three respective data frames using inner joins. First, the ratings_df is merged with the books_df based on the common ISBN column. This combined dataframe is then merged with the users_df_cleaned dataframe based on the common User-ID column. By employing an inner join, we ensure that only the records with matching ISBN and User-ID values in both dataframes are included in the resulting dataframes.

Upon further inspection, we have also decided to remove the User-ID column because it is likely to not provide any valuable information for predicting book ratings. It serves as an identifier and including it as a feature would introduce noise to the model without adding any predictive power. Removing it simplifies the dataset and improves the quality of input data for the machine learning model, ultimately leading to better predictive performance.

Outlier analysis for numerical variables were also conducted, and visualised using boxplots.

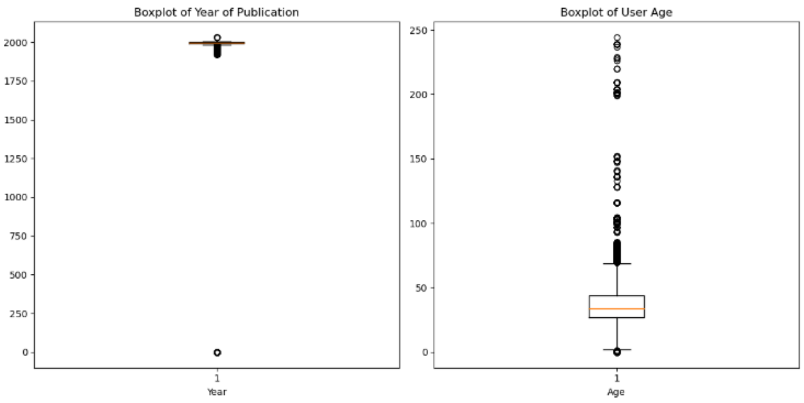


Figure 4: Boxplot of numerical variables (Year_Of_Publication, User_Age)

From the boxplot of Year of Publication, we can see that there are multiple points where the Year of Publication is 0, which is highly unlikely and suggests erroneous or missing data, since a book is likely to not have a publication year of 0. Therefore, we have decided to remove the rows associated with these erroneous data points to ensure the integrity of our analysis.

From the boxplot of User Age, we can see that there are multiple points where User Age is above 117, which is the oldest recorded person alive. (Martinez, 2024). Therefore, we conclude that these data points are likely erroneous or represent outliers. Therefore, to ensure the reliability of our analysis, we have decided to remove all data points where the User Age exceeds 117. Also, we have decided to remove all data points for User Age <= 6 because it is found that reading readiness age starts from 6-7 years old on average (Ward, 2022).

5. Data Pre-Processing

The data pre-processing section aims to transform raw data into a usable and understandable format.

For the first step of data-preprocessing, we will be looking at the 3 columns that provide geographical information about users - User-City, User-State, and User-Country. We have decided to remove the User-City and User-State column, and retain only the User-Country column, because the User-Country column offers a more generalised representation of users' locations compared to User-City and User-State, which provide more specific details. By focusing on User-Country, we simplify the dataset and eliminate redundancy while still preserving essential information about users' geographical locations.

We have also decided to remove the ISBN column because while ISBNs are unique identifiers for books, the ISBN does not provide much useful information or features for the machine learning prediction target Book-Rating. By excluding the ISBN column, we reduce the dimensionality of our dataset and simplify our model, focusing on more relevant features that can better contribute to the prediction task. We will also be removing the rows for User-Country containing:

1. "far away..." because it is vague and does not provide any useful geographical input to our machine learning task
2. "quit" because it is not a valid country and does not provide any useful geographical input to our machine learning task
3. "ysa" because it is not a valid country and does not provide any useful geographical input to our machine learning task
4. "universe" because it is not a valid country and does not provide any useful geographical input to our machine learning task

Next, we will be conducting the train-test split. The train-test split is done before the text preprocessing, encoding to prevent data leakage, which could occur when the test set inadvertently influences the training process. Data leakage can happen if text preprocessing techniques or encoding methods are applied to the entire dataset because the model could be learning information from the test set that it should not have access to during the model training. If feature scaling is applied before the train-test split, information about the data distribution in the test set may also influence the scaling process, leading to biased results. Therefore by conducting the train-test split before these steps, we ensure that the model is only trained on training data and evaluated on unseen data in the test set.

Next, we will process the columns with text - "Book-Title", "Book-Author", "Book-Publisher". We first tokenize the text using NLTK's word_tokenize function, breaking it down into individual words or tokens. The tokens are then converted to lowercase to ensure consistency in text representation. Punctuation marks are then removed to focus on the essential content of the text. Afterwhich, we moved stopwords (eg. common words like "the," "and," "is," etc.) that don't add much meaning to the text using NLTK's English stopwords list. Lemmatization is then performed to reduce words to their base or dictionary form, which standardises text and helps capture word meanings more accurately. Finally, the preprocessed tokens are joined back together to form a single string output of the cleaned and normalised text. Before we continue on with the text processing, we will be using WordCloud to inspect the textual data of the "Book-Title", "Book-Author", "Book-Publisher" columns visually to identify the common words as well as spot any inconsistencies in the data.

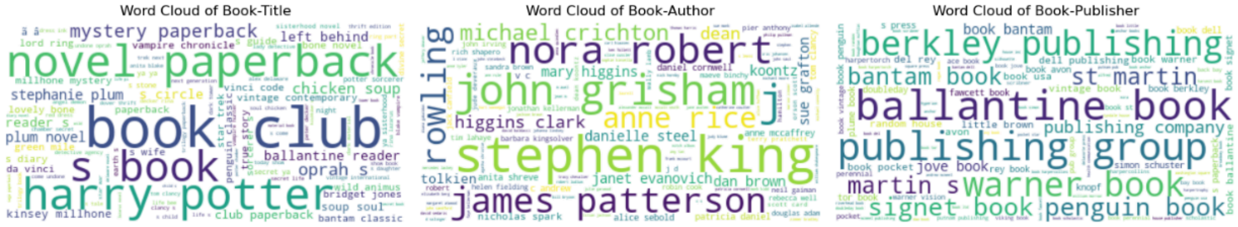


Figure 5: Word Cloud for Book-Title, Book-Author, Book-Publisher

The word cloud gives us a good visualisation and summary of what are the words in each respective column, which we will use for further analysis. TF-IDF vectorisation is applied here, where the textual data for “Book-Title”, “Book-Author”, “Book-Publisher” is converted into numerical vectors to be used as input features into the machine learning model. The resulting TF-IDF matrices will capture the importance of each word in contextual detail, enabling machine learning models to learn patterns and make predictions based on the contextual detail.

Before we look into encoding the remaining categorical variables, we will first check the number of unique values for each categorical variable to understand the best way to encode these categorical variables for input to machine learning models. Since "Book-Title", "Book-Author", "Book-Publisher" has already been converted into a TF-IDF Matrix, we will focus on encoding the remaining categorical variable - "User-Country".

```
Number of unique countries: 99
```

Figure 6: Number of unique categories for categorical variables

For the "User-Country" column, we will use one-hot encoding. One-hot encoding creates new dummy binary variables for every class in each categorical feature for categorical variables with no ordinal relationship. This method is suitable for categorical variables with a small number of unique categories without an ordinal relationship. Before we proceed with one hot encoding, we will first aggregate the "User-Country" column into the different continents instead as the current number of unique categories for the column "User-Country" is still quite high (count: 99). For better representation of the column "User-Country", we will rename it to "User-Continent" as we have aggregated and reclassified the countries into the respective continents

Feature scaling will be applied to numerical data because the features are commonly not on the same scale. Machine learning models, on the other hand, are largely euclidean distance-based. Therefore, to prevent the issue where features with large units dominate over features with smaller units during the distance calculation (causing features with smaller units to be neglected), we will be appling Min-Max scaling to scale the numerical features (Toprak, 2019). In this project, we will be using MinMaxScaler() to scale and preprocess the data through the Min-Max Scaling method to scale the numerical data values to a range of 0-1.

The last step for pre-processing will be to combine the TF-IDF matrices with the respective train and test dataframes.

6. Model Selection

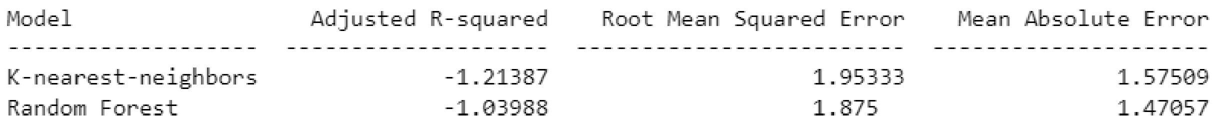
We will now run 2 machine learning models to test and evaluate which one performs the best. Selecting the best-performing model is crucial because ensure that the predictive model will generalise well to unseen data, and hence make accurate predictions in real-world scenarios.

The models we will be using are: K-nearest-neighbours, Random Forest. After which, we will be evaluating each respective model with the test set.

K-nearest neighbours is a non-parametric machine learning algorithm that classifies new data points based on the majority class of its K nearest neighbours within the feature space. Our group chose K-nearest neighbours because of its interpretability and performance. Unlike more complex models like neural networks, k-nearest neighbours can be easily understood and interpreted. This interpretability is crucial for our project, as it allows us to gain insights into which features contribute most to the predictions. The K-nearest neighbours algorithm also memorises the training dataset instead of learning explicit models, making it computationally efficient during training. It also tends to perform well when the dataset is not too large. Moreover, K nearest neighbours is robust to noisy data and outliers, making it suitable for real-world application (Subramanian, 2019).

Random Forest is a machine learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or the average prediction (regression) of the individual trees. The output of each respective decision tree will then be combined to give the final prediction of the target variable. Our group chose random forest because of its interpretability and performance. Random forest allows for model interpretability through feature importance, as compared to more complex models such as neural networks which are not interpretable. Model interpretability is key for us in this project because we need to understand which features are driving the predictions. Random forest also tends to perform well over a large range of datasets, and can handle outliers and noisy data well, making it highly suited for real world applications (Sruthi, 2024).

7. Model Evaluation & Results



```
Model                  Adjusted R-squared    Root Mean Squared Error    Mean Absolute Error
-------------------  --------------------  -------------------------  ---------------------
K-nearest-neighbors              -1.21387                    1.95333                1.57509
Random Forest                    -1.03988                    1.875                  1.47057
```

Figure 7: Model Performance of the models

The Adjusted R-squared metric measures the proportion of the variance explained by the machine learning model, while taking into account the number of predictor features, and can range from negative infinity to 1. This is done by penalising the addition of predictors which are unnecessary and do not improve the machine learning model performance. In this project, adjusted R-Squared hence explains the variability of the target variable book_rating, adjusted for the number of predictor features in the model. An Adjusted R-squared value close to 1 indicates a good fit of the regression model to the data. For instance, an Adjusted R-squared of 0.80 suggests that 80% of the variability in book_rating is explained by the independent variables in the model, adjusted for the number of predictors.

Mean Squared Error (MSE) is the mean of the squared differences between the actual target values and predicted target values, and can range from 0 to infinity. Root Mean Square Error (RMSE) is the square root of the Mean Square Error, and in this project is used to quantify the magnitude of the mean squared deviation of the predicted book_rating and actual book_rating values. RMSE is also particularly useful as it will be in the same unit as the target variable book_rating, making it easier to interpret. Lower RMSE will suggest that the machine learning model's predictions are close to the actual values, suggesting that the regression model is a good fit to the data and is a good estimation for book_rating. For instance, an RMSE of 0.23 suggests that on average, the deviation of predicted book_rating from actual book_rating is 0.23.

Mean Absolute Error (MAE) is the mean of the absolute difference between the actual values of book_rating and the predicted values of book_rating, and can range from 0 to infinity. In this project, MAE hence measures the average absolute deviation of predicted book rating from the actual book rating values. MAE provides a robust measure of the machine learning model's performance, as it is less sensitive to outliers compared to RMSE. A lower MAE value will suggest that the model's predictions have smaller errors, providing a more accurate estimate of the target variable book_rating. For instance, an MAE of 0.23 suggests that on average, the absolute deviation of predicted book_rating from actual book_rating is 0.23.

As a whole, using our selected evaluation metrics of adjusted R-squared, MSE, MAE we can conclude that the random forest algorithm is the best performing model, which we will be using for further analysis below.

8. Discussion and Interpretation

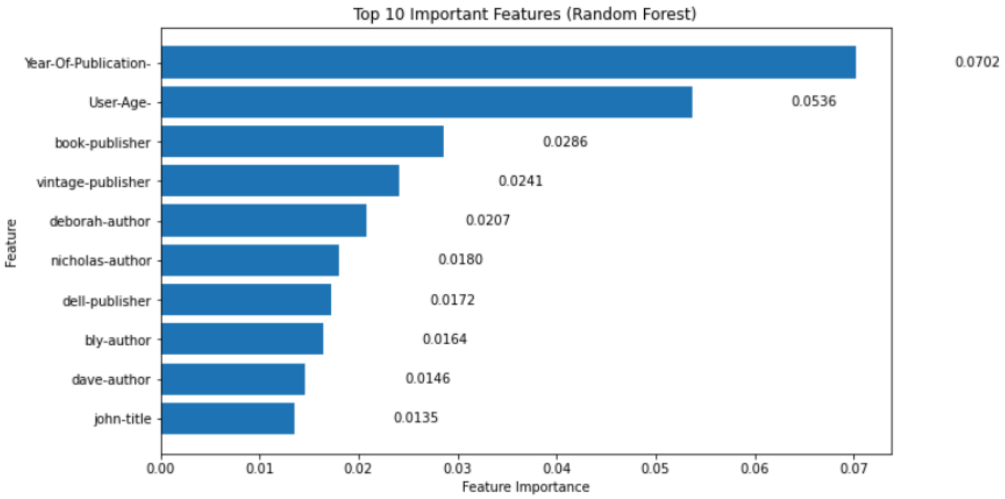


Figure 8: Feature Importance of the random forest algorithm

Considering the top 3 most important features: Year-Of-Publication, User-Age, book-publisher

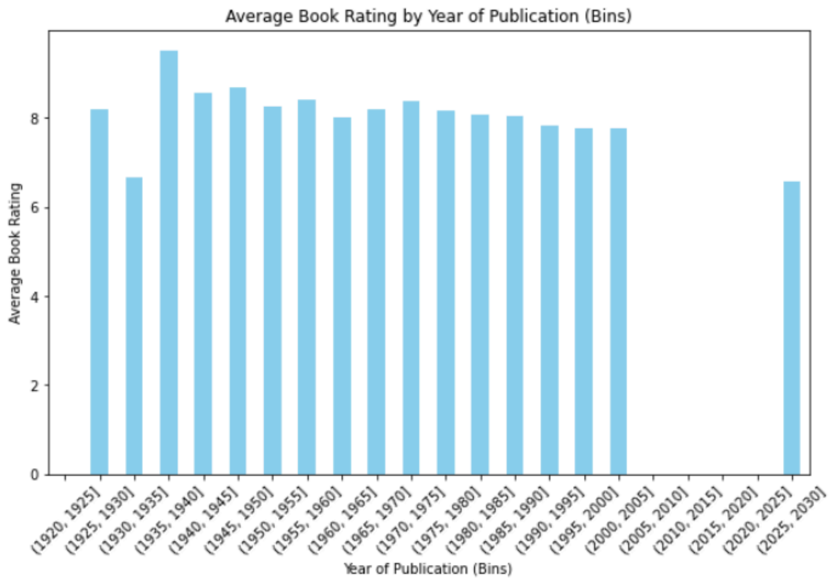


Figure 9: Bar Plot of Average Book Rating by Year Of Publication

In this particular case, the feature importance analysis has revealed that the "year of publication" attribute holds the highest importance score, with a value of 0.072. An analysis of book ratings grouped by publication year (5-year bins) revealed a trend in reader preference. Books published between 1930 and 1935 boasted the highest average rating, exceeding 9, while those from 1925 to 1930 fell short at just above 6. This suggests a potential shift in reader tastes during this period. Capitalising on this data, publishers can analyse content from the high-rated era to identify popular themes, genres, and writing styles to inform future acquisitions and development. Similarly, revisiting publishing approaches from the low-rated period may expose areas for improvement in editing, cover design, or marketing efforts. By promoting hidden gems and fostering collaboration with successful authors, publishers can curate a collection that resonates with readers, ultimately driving engagement in the literary market.

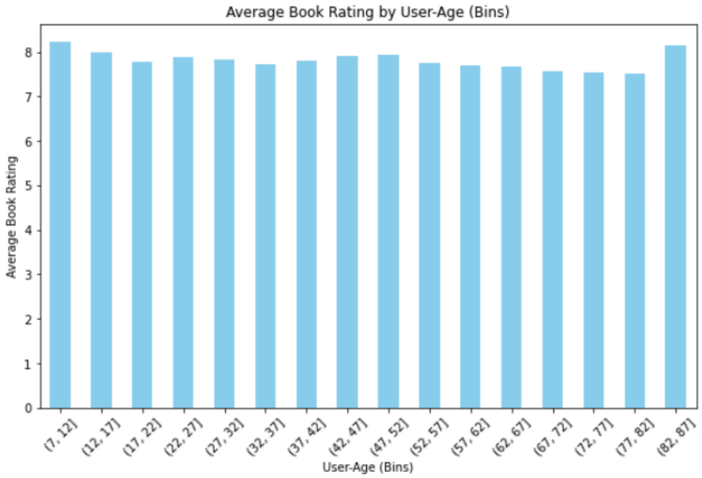
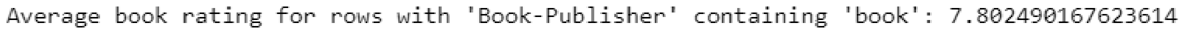


Figure 10: Bar Plot of Average Book Rating by User Age

The analysis of the average book ratings against user age reveals insightful trends for book managers. Notably, individuals aged 82-87 consistently provide the highest book ratings, indicating a strong satisfaction with the available literature. Conversely, the age group spanning 32-37 years appears to be the most critical, although the difference in average ratings compared to other age groups is not significant. Despite this, it may be beneficial for book managers to pay closer attention to the preferences and feedback of readers within this age bracket. Engaging with this demographic through targeted marketing campaigns, personalised recommendations, or curated book selections could help enhance their overall satisfaction and foster a deeper connection with the literary offerings.



```
Average book rating for rows with 'Book-Publisher' containing 'book': 7.802490167623614
```

Figure 11: Average book rating for rows with publisher name containing 'book'

book-publisher also holds significant importance in predicting book ratings above other features, with a feature importance score of 0.0286. The analysis reveals that among the entries where the 'Book-Publisher' field contains the term 'book', the average book rating stands at approximately 7.80. This suggests a generally positive reception for books associated with publishers containing 'book' in their name. Notably, while exploring these publishers, it's essential to identify specific factors contributing to their success. This could involve examining the genres, themes, or authors prevalent in their publications, as well as any distinctive marketing or distribution strategies they employ. By delving deeper into these aspects, book managers can gain valuable insights into what resonates most with readers and leverage this knowledge to optimise their own publishing endeavours. Additionally, fostering collaboration or partnerships with successful publishers in this category may present opportunities for mutual growth and knowledge sharing. Ultimately, by leveraging these insights and relationships, publishers can strive to enhance their offerings and strengthen their position in the market, ultimately leading to increased reader satisfaction and engagement.

By leveraging the above mentioned insights, bookstore managers can make informed decisions regarding book procurement and customer recommendations.

9. Areas Of Improvement & Limitations

One way to improve is to conduct hyperparameter tuning for the machine learning models, such as the number of trees in Random Forest or the number of neighbours in K-nearest-neighbours. Hyperparameter tuning allows us to optimise the model's performance metrics like accuracy, precision, recall, and F1-score. This optimization ensures that our models are better suited to capture the underlying patterns in the data, leading to improved predictive accuracy (Jain, 2023)

Another way to improve is to check for Over/Underfitting through learning curves. It is important to assess if the machine learning models are overfitting (learning noise in the training data) or underfitting (too simplistic to capture patterns), and can be done by plotting learning curves which depict the model's performance (eg. accuracy/error) as a function of the training set size. Overfitting tends to show a large gap between the training and validation curves, while underfitting typically results in low performance on both sets (Bnomial, n.d.)

The first limitation is that the dataset provided has a limited set of features that captures the complexities of the book_rating, hence book sales. While features like User-Age, Title-Mean-Rating, and Author-Mean-Rating provide insights, they may not fully encompass all factors influencing book rating and sales. For example, variables such as marketing efforts, book genre, seasonal trends and socioeconomic factors could also play a role in book_rating but are not included in the dataset. The absence of these variables may limit the predictive power of the machine learning models and hence the comprehensiveness of the analysis.

The second limitation would be computational constraints which lead us to be unable to deploy high-computation resource models for text processing. As the textual data columns "'Book-Title', 'Book-Author', and 'Book-Publisher' are very expansive, there are some challenges in effectively capturing the semantic, intricate relationships between the texts. Advanced models such as BERT, GPT which are built on transformer-based architectures, or deep semantic similarity models such as Siamese networks will offer much superior capabilities in understanding the nuances in our text semantics (uncovering subtle contextual cues, synonyms and underlying meanings). However, due to resource limitations we were constrained to employ simpler techniques like TF-IDF, which may not fully capture the semantic relationships within the data, hence potentially impacting the accuracy and depth of our analysis. To mitigate this, we can focus on securing additional computation resources or cloud-based solutions in future to harness more sophisticated models so as to be able to extract more meaningful insights from our textual data.

10. References

Aguiar, H. (2024). *What Is Imbalanced Data and How to Handle It?* TurinTech AI. Retrieved May 1, 2024, from https://www.turintech.ai/what-is-imbalanced-data-and-how-to-handle-it/

Bnomial. (n.d.). *Overfitting and Underfitting with Learning Curves | Bnomial*. Machine learning articles - Bnomial. Retrieved May 1, 2024, from https://articles.bnomial.com/overfitting-underfitting-learning-curves

Jain, S. (2023, December 7). *Hyperparameter tuning*. GeeksforGeeks. Retrieved May 1, 2024, from https://www.geeksforgeeks.org/hyperparameter-tuning/

Martinez, J. (2024, March 4). *California-born woman, Maria Branyas Morera, is world's oldest person at 117*. FOX40. Retrieved May 1, 2024, from https://thehill.com/changing-america/well-being/longevity/4507292-california-born-woman-is-the-oldest-person-in-the-world/

Peebles, W. (2022, November 11). *Introduction to Balanced and Imbalanced Datasets in Machine Learning*. Encord. Retrieved May 1, 2024, from https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/

Sruthi, E. R. (2024). *Building a Random Forest Model: A Step-by-Step Guide*. Analytics Vidhya. Retrieved May 1, 2024, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Subramanian, D. (2019). *A Simple Introduction to K-Nearest Neighbors Algorithm*. Towards Data Science. Retrieved May 1, 2024, from https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e

Toprak, M. (2019, December 3). *Why ,When to Apply Feature Scaling(Normalization) and How to Apply it with SciKit Learn.* Medium. Retrieved May 1, 2024, from https://medium.com/@toprak.mhmt/why-when-to-apply-feature-scaling-normalization-and-how-to-apply-it-with-scikit-learn-fa659aa5fca8#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImUxYjkzYzY0MDE0NGI4NGJkMDViZjI5NmQ2NzI2MmI2YmM2MWE0ODciLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJod

Ward, M. (2022, June 3). *Best Age to Start Teaching Kids to Read: 5 Signs They're Ready*. iCode School. Retrieved May 6, 2024, from https://icodeschool.com.au/best-age-to-start-teaching-kids-to-read-5-signs-theyre-ready/