Factors Influencing Book Ratings

Group W{10}G{2}

Xingzhi Du
1346725
xingzhid@student.unimelb.edu.au

Luting Shen
1260562
lutings@student.unimelb.edu.au

Sofía Ramírez Martínez
1360630
sramirezmart@student.unimelb.edu.au

Executive Summary

This report uses the Kaggle repository "Book-Crossing: User review ratings" to explore whether there are factors that influence book ratings. This is done through the application of data pre-processing techniques, and machine learning models (linear regression, decision tree regression and decision tree classification). The objective of this analysis was to establish if there are factors that influence book ratings to provide insight into what books a bookstore should buy, and which books they should recommend to users. The key findings from this exploration were that the data used to train the machine learning models and the book reviews have an approximately linear relationship, which makes the linear regression model the best fit for analysis of this data, and that there are no factors that have a significant correlation with book ratings. Further research based on these findings could provide more insight into the relationship between book and user features and book ratings, given that the data used in this report had bad quality. User decisions are heavily influenced by user ratings, so conducting research on the determining factors is important to gain insight on book preferences, from which recommender models and other computing analysis can be performed.

Introduction

Opinions have always been an important aspect that influences decision making. In earlier times, people were limited to look to those around them to inform themselves, but with the rise of technology, a myriad of opinions is now available for anyone to access at any time (Rocklage et al., 2021). However, with such a large amount of information, it may be hard to narrow down and get a general idea of what the opinions are. Thus, one can look at the quantitative expression of these opinions to gauge what the overall sentiment towards something is: ratings.

Ratings have become a common standard to measure the quality of a product as a numerical measure, usually on a scale of 1-5 or 1-10, of the overall sentiment (positive or negative) towards it. However, studies have shown that although practical, ratings have an important caveat: they are heavily positively biassed (Rocklage et al., 2021). Sinan (2014) attributed this phenomenon to the Social Influence Bias, which they described as the susceptibility of feeling positively towards something and giving it a high rating as a result of seeing someone else have a positive sentiment towards something and awarding it with a high rating. Ultimately, this implies that ratings have unreliable predictive power of quality and success of a product, since they don't accurately represent reality (Rocklage et al., 2021).

Focusing these findings to books, it can be concluded that although ratings are influential in customer purchase decisions, they might not accurately reflect the quality of the book, the success it will have with the reader, or the true sentiment towards it. Therefore, since ratings are not directly related to the users' perceptions of the books, this evoked the question of whether there are factors, other than the social influence bias, that influence ratings.

To explore this question, we used the repository "Book-Crossing: User review ratings" obtained from Kaggle (Bhatia, 2021). This dataset contains three main datasets "BX-Books.csv", "BX-Ratings.csv" and "BX-users.csv". These datasets provide insight into books, ratings given by users and user details

which are useful in developing an analysis of the relationship between their components and book ratings.

Conducting this analysis is beneficial because it will provide insight into what are important factors to consider when making decisions regarding books other than book ratings, and it can be useful to understand why books have their given ratings. An example of a situation where this can be applied is to help managers make purchase decisions, and to know which books they should recommend to buyers, following the idea that they could recommend books that other users with similar demographics (such as age) have rated highly.

Methodology

To better address the research question regarding the factors that have a relatively significant impact on book ratings, our methodology incorporates the use of data pre-processing, exploratory data analysis and application of machine learning models (Linear Regression, Decision Tree Regression and Decision Tree Classification) to deduce insights.

Data Pre-Processing

The dataset used in the analysis comprises three main tables: Users, Books, and Ratings. The initial step involves importing and inspecting these datasets to understand their structure and contents to check and know what we should do next to do data pre-processing.

Before Data Pre-processing

Dataset name	Total rows
BX-Books.csv	18,185
BX-Ratings.csv	204,146
BX-Users.csv	48,299

Table 1: Table showing dataset names and total rows (entries) before the data pre-processing.

Users Dataset

1. Handling Missing Data:
   1.1. Initial inspection showed missing values in the 'User-Country' and 'Age' columns.
   1.2. Non-numeric age entries were converted to NaN to maintain numeric integrity.
2. Data Encoding:
   2.1. The 'User-Country' column was encoded using LabelEncoder from sklearn (Pedregosa et al., 2011) to transform country names into numerical labels. This step is crucial for modelling as it allows the inclusion of categorical geographic information as a numeric feature.
3. Feature Selection:
   3.1. After encoding, we retained only essential features for model training: 'User-ID', 'Age_Category', and 'User-Country_Encoded'.

Books Dataset

1. ISBN Validation:
   1.1. We enforced the ISBN integrity by ensuring each entry adheres to the ISBN format rules, specifically maintaining the character 'X' only if it appears as the last character.
2. Year of Publication Cleaning:

2.1. Entries outside the valid publication years (1920 to 2024) were removed, and we also handled missing values in this column by removing such entries.

3. Decade Categorization:

3.1. The publication years were categorised into decades to facilitate a historical analysis of publishing trends, which we visualised using a bar chart.

4. Text Pre-processing:

4.1. Titles, authors, and publishers underwent pre-processing to remove special characters and stopwords and were then tokenized and stemmed. This cleaned, simplified text facilitates more effective text analysis and machine learning modelling.

5. Encoding Authors and Publishers:

5.1. Similar to the 'User-Country', authors and publishers were encoded numerically to prepare these categorical text data for modelling.

6. TF-IDF Vectorization:

6.1. The 'Book-Title' was transformed into a numerical format using TF-IDF vectorization to capture the importance of words in titles relative to the dataset. This high-dimensional feature space will later be reduced using PCA for more manageable modelling.

Ratings Dataset

Mapping Ratings to Classes: In preparation for using decision tree classification models, ratings are mapped into categorical classes. This transformation is crucial for enabling the application of classification algorithms, which require categorical input to function effectively.

After Data Pre-processing

Dataset name	Total rows
BX-Books.csv	17,867
BX-Ratings.csv	204,146
BX-Users.csv	47,441

Table 2: Table showing dataset names and total rows (entries) after the data pre-processing.

Data Analysis and Machine Learning Models

Exploratory Data Analysis (EDA):

Implements matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for visualisations such as histograms and bar charts to explore data distributions and initial patterns.

Machine learning Models:

This analysis also takes advantage of the linear regression, decision tree regression and decision tree classification models to understand and interpret the impacts on book ratings. Each model is implemented using python's 'scikit-learn' (Virtanen et al., 2020) library and also is trained using a split of training and testing data, with k-fold cross validation employed to assess each model's performance and ensure robustness against overfitting.

Data Exploration and Analysis

The first thing we did was visualise the distribution of the Book Ratings, and the features used to train the Machine Learning models. These are Age, Country and Date Published. The book rating distribution graph below illustrates a noticeable skew towards higher ratings, indicating a prevalent tendency among users to assign positive ratings more frequently, which aligns with the discussions on

positive bias in ratings among people as highlighted in previous studies (Rocklage et al., 2021; Sinan, 2014).

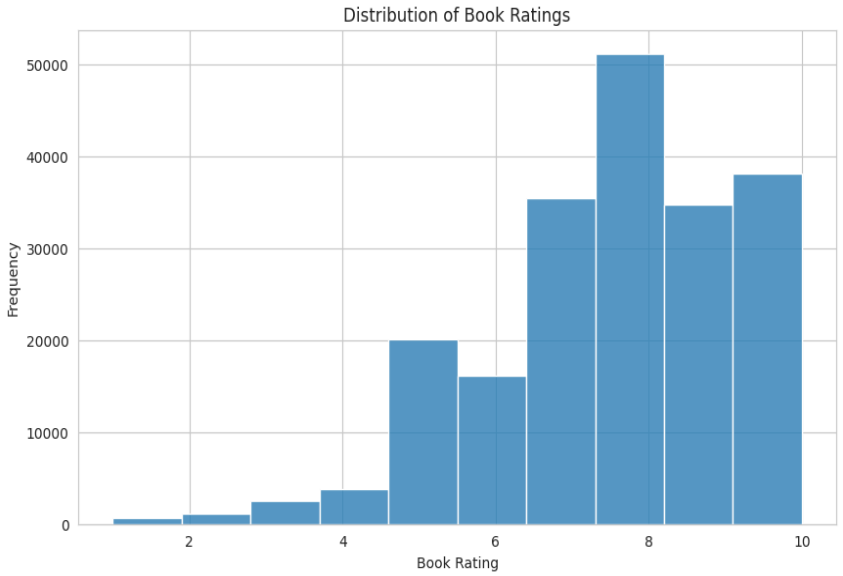

Figure 1: Distribution of Book Ratings

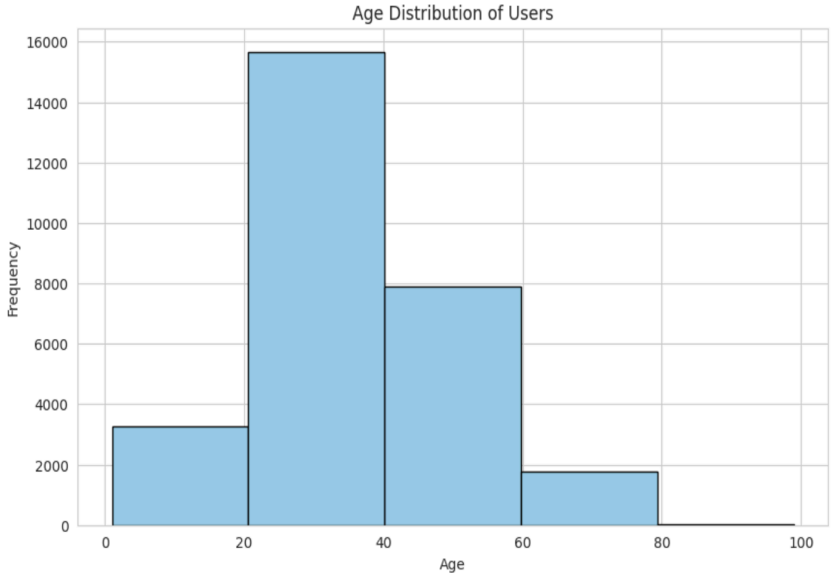

Figure 2: Age Distribution of Users

The Age Distribution of Users shows that most users are between 20 and 40 years old, with the minority being the group between 80 and 100 years old.

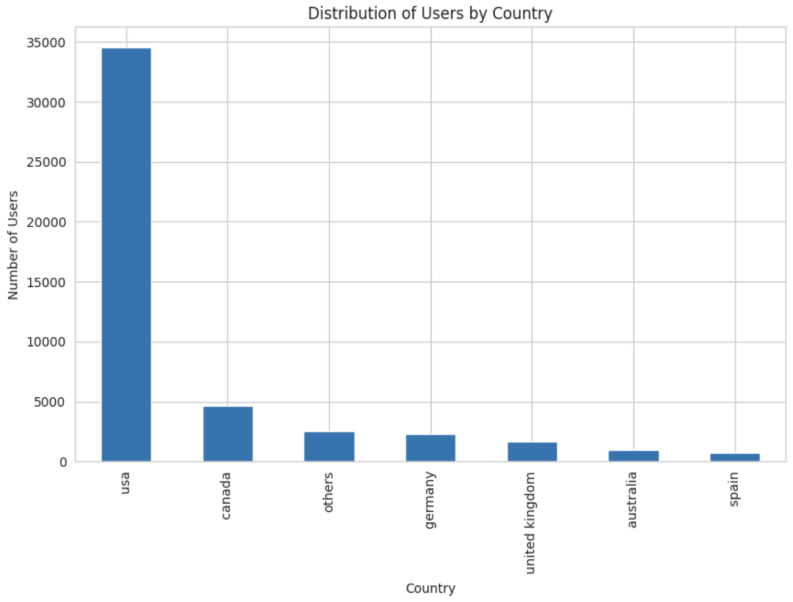

Figure 3: Distribution of Users by Country

The Distribution of Users by Country shows that most users are from the USA, followed by Canada, and then "other" countries.

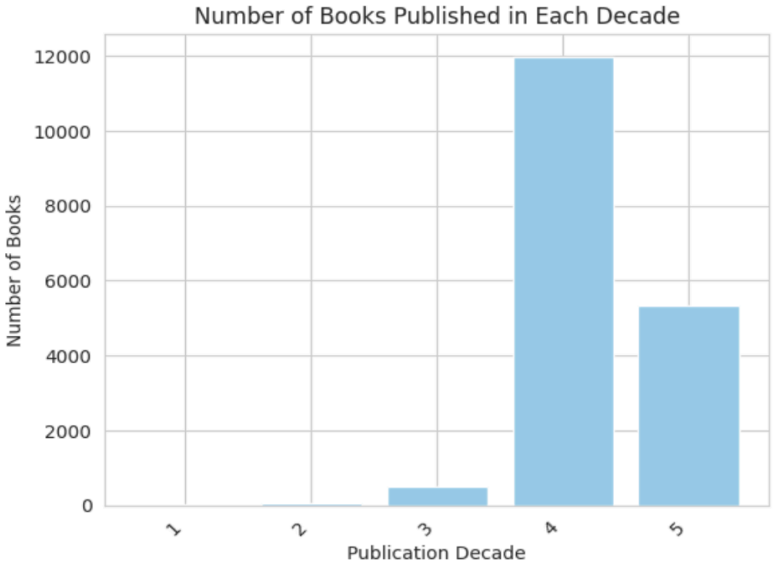

Figure 4: Distribution of Books by Date Published

The Distribution of Books by Date Published shows that the majority of books were published in bucket 4, corresponding to the years between 1981 and 2000, where the buckets represent the following intervals: bucket 1 (1920-1940], bucket 2 (1240-1960], bucket 3 (1960-1980], bucket 4 (1980-2000], and bucket 5 (2000-2024].

Before using machine learning models to do analysis, we used a correlation matrix and a heatmap to help identify multicollinearity – which is the presence of highly correlated independent variables in a regression analysis. This is useful because when predictors are highly correlated, it becomes difficult for the model to estimate the individual effect of each predictor accurately. Multicollinearity can cause coefficients to have high variability, making them sensitive to small changes in the data, which makes it challenging to interpret the impact of each predictor on the target variable consistently.

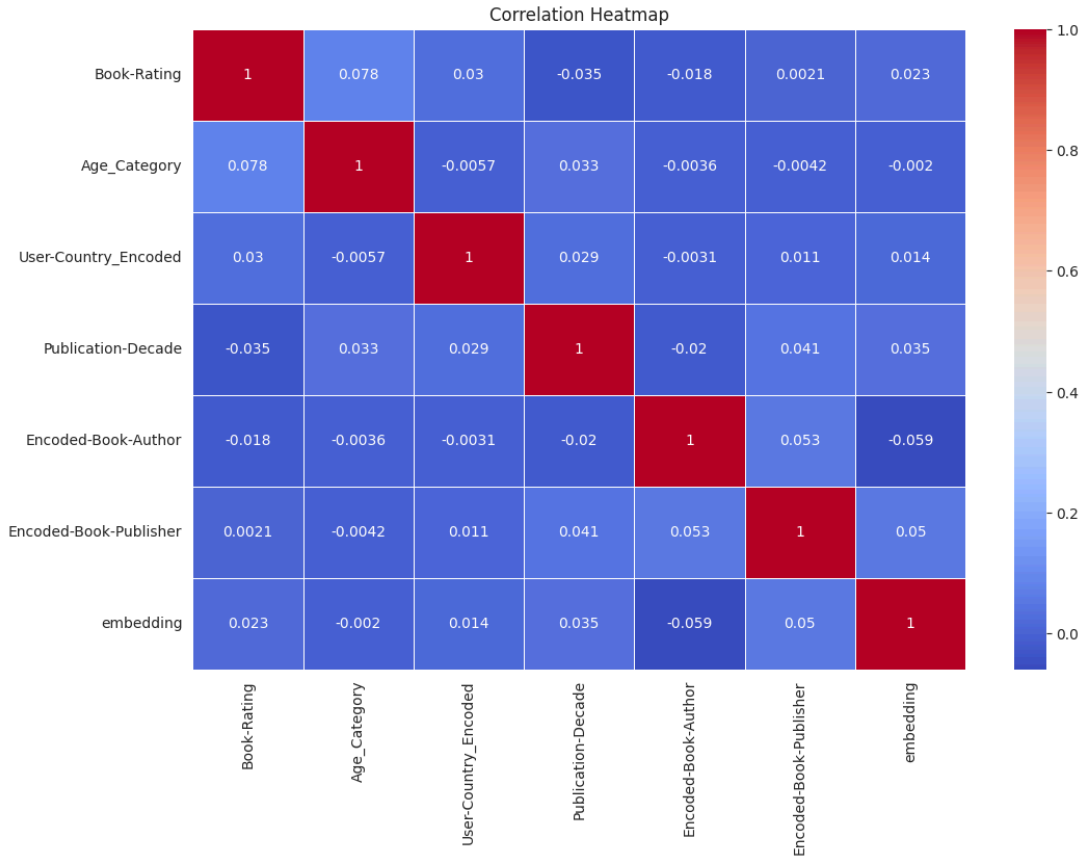

Figure 5: Heap Map

The heatmap shows that there is no significant multicollinearity between the features, which facilitates the estimation of the individual effect of each predictor.

Results

Our analysis incorporates both supervised and unsupervised learning models to address questions related to user demographics and book ratings. The primary focus is on determining the extent to which different features influence book ratings, and the effectiveness of different machine learning models in predicting book ratings.

Linear Regression Model

Using a linear regression model, the figure below shows the importance of each feature in predicting book ratings. ‘Age-Category’, ‘Publication-Decade’ and ‘User-Country-Encoded’ are the three most important features to determine the book ratings. ‘Age-Category’ has the highest importance score, which suggests that the age group of the users is a strong predictor of how they might rate a book.

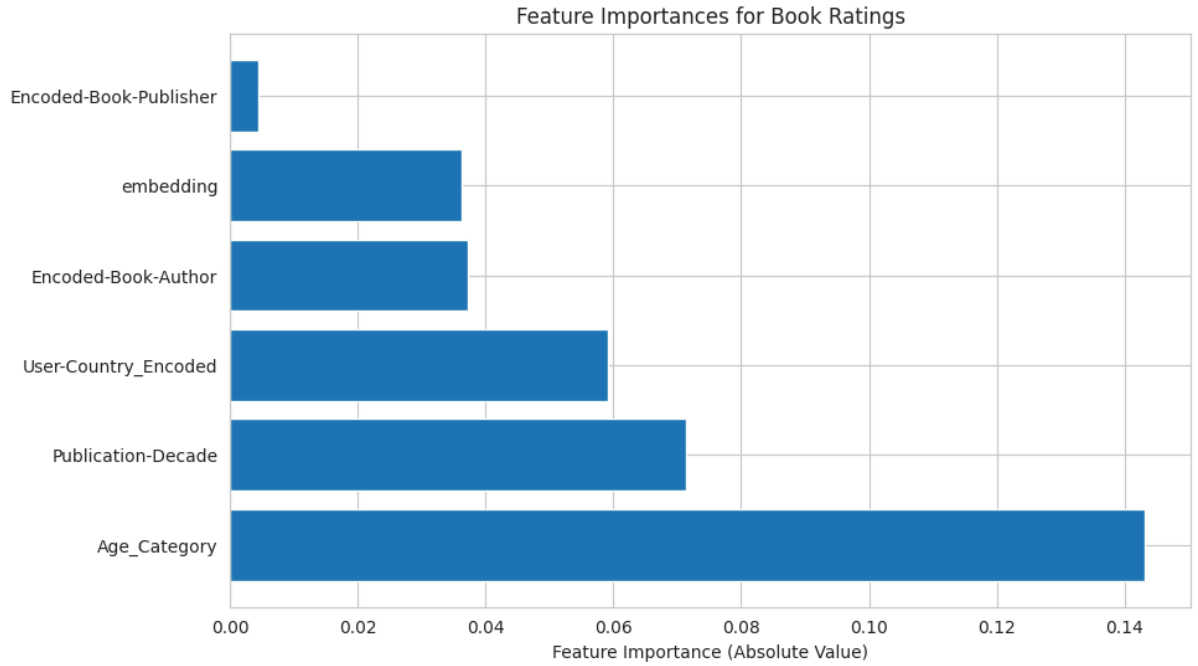

Figure 6: Feature Importance for Book Ratings, Linear Regression

Decision Tree Regression Model

The Cross Validation Mean Standard Error score for the decision tree regression model is about 5.08, which compared to the 3.17 of the linear regression model, is notably higher. When evaluating regression models, a low MSE score means that there is small error, and thus, regression models aim to have this. Thus, the high CV MSE score of the decision tree regression model suggests that the model could be overfitting the data, leading to inaccurate predictions, or that it struggles to capture all the underlying patterns in the data. Compared with the linear regression model, this model resulted in worse performance.

Decision Tree Classification Model

Using a different machine learning model resulted in a different ranking of features than the one calculated through linear regression in figure 6. In this model, the embedding (another way to represent the book title information), 'Encoded-Book-Author', 'Encoded-Book-Publisher' and 'Age-Category' were identified as the main features that determine the book rating, indicating the varied nature of their influence on book ratings when using a decision tree model. In this model, 'embedding' has the highest importance, significantly higher than the other features.

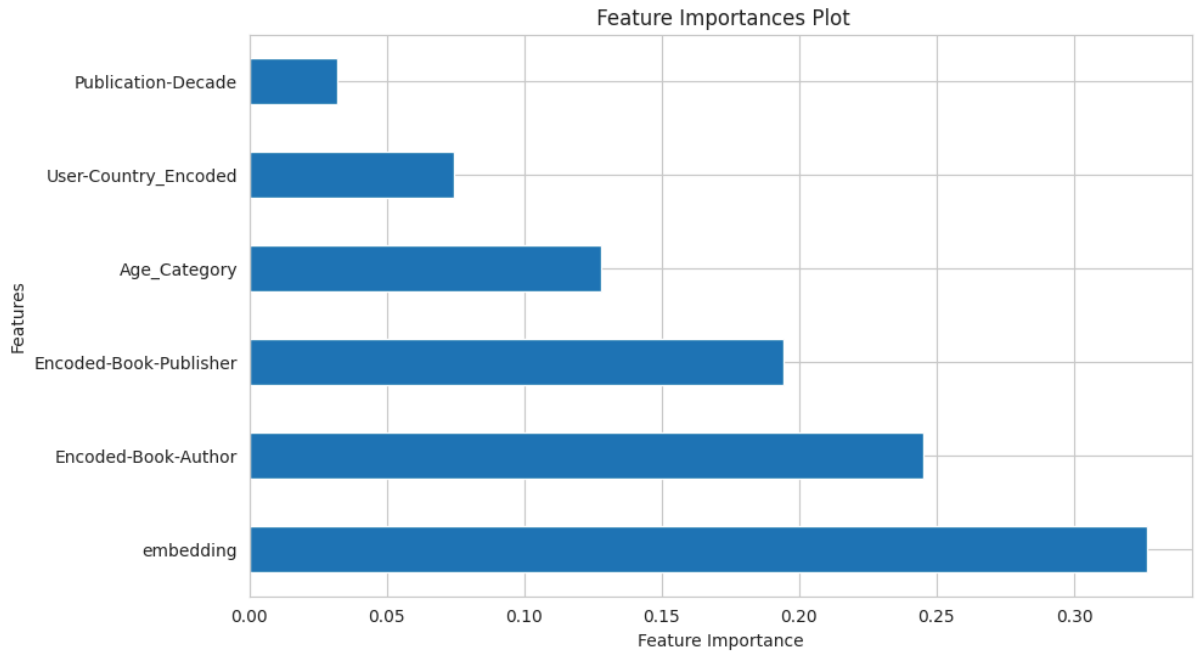

Figure 7: Feature Importance for Book Ratings, Decision Tree Classification

Discussion and Interpretation

```
Accuracy: 0.5427709765634984
Classification Report:
              precision    recall  f1-score   support

           0       0.10      0.09      0.10      1546
           1       0.40      0.39      0.40     13696
           2       0.65      0.66      0.65     23885

    accuracy                           0.54     39127
   macro avg       0.38      0.38      0.38     39127
weighted avg       0.54      0.54      0.54     39127
```

Figure 8: Decision tree classification model result

From the figure 8, we can see that the decision tree classification shows moderate accuracy (0.54) And, this model performs relatively well for predicting higher ratings (class 2), with a higher precision, recall, and F1-score compared to the other classes.

However, it struggles with lower ratings (class 0) and shows moderate performance for mid-range ratings (class 1). The low accuracy on predicting the low ratings may be caused by imbalanced data which has relatively few samples with a class of 0 compared to other rating categories, it could result in poorer performance in predicting 0 class (low ratings). This imbalance in the data distribution can affect the model's learning and generalisation capabilities.

Overall, while the model can distinguish between the classes to some extent, there is room for improvement, especially in correctly identifying lower ratings.

```
Cross-validated Mean Squared Error (CV MSE): 3.172241909947137
Mean Squared Error (MSE): 3.138222792100308
R-squared (R2): 0.00859234405454845
Feature Importances:
Age_Category: 0.1432257028065067
Publication-Decade: -0.07129535602106873
User-Country_Encoded: 0.0592116047461171
Encoded-Book-Author: -0.037223152143750604
embedding: 0.03621658961017343
Encoded-Book-Publisher: 0.004405472673894462
```

Figure 9: Linear Regression Model Performance Result

From the figure 9, The Linear Regression model exhibited a very low R-squared value of approximately 0.009, indicating that only a small fraction of the variance in book ratings could be explained by the model. This suggests the model's limited capacity in this context due to the complexity and variability of user rating behaviours. We can also see that the negative coefficient for Publication-Decade suggests newer books receive slightly lower ratings, potentially reflecting evolving reader preferences or critical standards.

We can see that the 'Age_Category' feature has a significant positive impact, indicating that older age categories tend to give higher ratings. Because of this, we produced a regression line (shown in figure 12) to visualise their relationship.

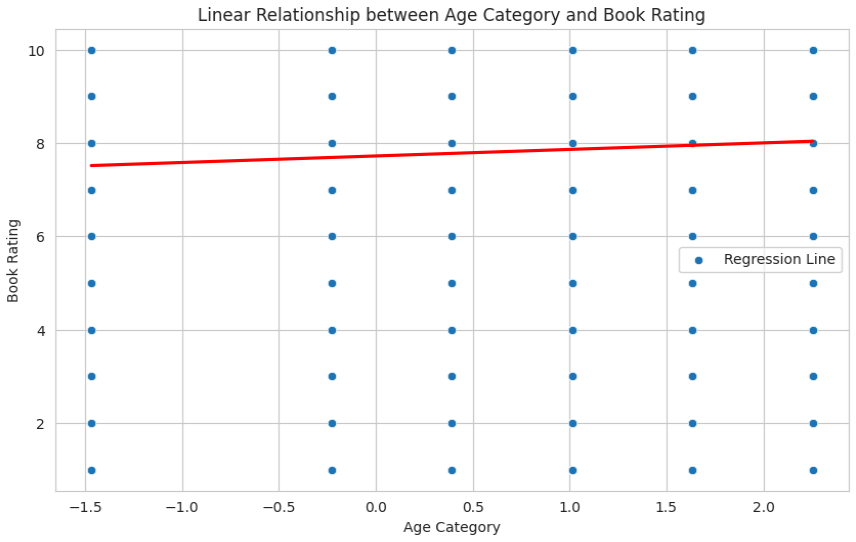

Figure 10: Stable Book Rating Across Age Categories

The regression analysis shows a positive correlation between age and book ratings, which implies that as users get older, they tend to give higher ratings. Looking at this result in isolation provides little explanation, however, considering the type of books older people are reading could give insight into this phenomenon. A study investigating consumer ratings of popular films (Simonton et al., 2012) theorised that older people prefer older content, referred to as classics, which have an intrinsic value due to their complexity. Thus, older people might be giving books higher ratings because their book choices are perceived to be of higher quality, and thus, merit a higher rating. To test this hypothesis, an experimental design would need to be conducted.

```
Cross-validated Mean Squared Error (CV MSE): 5.08461299762938
```

Figure 11: Decision Tree Regression Model Result.

From figure 11, we can see that the higher CV MSE suggests that using decision tree regression models may be less effective at predicting book ratings or possibly overfitting, despite Decision Trees typically being better at handling non-linear relationships. This can affect the reliability of the feature importance values obtained from the Decision Tree Regressor, as the model itself is not performing well in terms of prediction accuracy, so we don't visualise feature importance in this model.

Comparing the Linear Regression and Decision Tree Regressor models, the cross-validated Mean Squared Error (CV MSE) for the Linear Regression was lower than that of the Decision Tree Regressor. This indicates that, on average, the Linear Regression model had smaller errors in predicting book ratings compared to the Decision Tree model. The superior performance of Linear Regression suggests that the relationships between features and book ratings might be more linear in nature. Linear Regression's assumption of a linear relationship between predictors and the target variable seems to be more appropriate in this context, leading to better predictive performance.

An unexpected finding of this analysis was the Decision Tree Regressor's higher CV MSE, as decision trees are known for their ability to capture complex relationships in data. However, in this case, the linear nature of relationships in the data might have favoured the Linear Regression model. This unexpected result highlights the importance of choosing the appropriate model based on the underlying patterns in the data. While decision trees can handle nonlinear relationships well, they might not always outperform simpler models like Linear Regression if the data follows more linear patterns.

In summary, no significant correlation could be established between any of the parameters. However, if regression models need to be used to do analysis, the linear regression model is better than using decision tree regression for book ratings given that the relationship seems to be approximately linear. On the other hand, if the goal is prediction accuracy and the problem is classification (such as predicting classes like ratings), the decision tree classifier with an accuracy of 54% is comparatively better than the linear regression model, which has limited explanatory power (low R-squared) for the given data.

Limitations and Improvement Opportunities

The analysis of the BookCrossing dataset, while insightful, faces several limitations that could impact the robustness and applicability of the findings. One major limitation is the quality and completeness of the data. Missing values, especially in user demographics and specific book details, might have introduced biases, affecting the accuracy and reliability of the results. Additionally, the models used, including linear and decision tree regressions, might not fully capture the complex, non-linear relationships that exist in the data. If we were to conduct this analysis again, it would be beneficial to drop all missing data instead of filling it to try and get a more accurate portrayal of the relationship between the different variables and book ratings.

For improvement, enhancing data collection efforts to cover a broader range of user demographics and book genres could significantly enrich the dataset, leading to more comprehensive analyses. In the future we will learn more precise machine learning techniques, so we can apply more sophisticated machine learning techniques such as ensemble methods or neural networks could overcome some of the limitations of the current models by better handling the data's complexities. Another way for enhancement is the integration of temporal analysis to examine trends over time, which could provide insights into how external factors like social media trends influence book ratings.

Moreover, as aforementioned, the results of this analysis only provide a correlational relationship between book ratings and the dataset features, most significantly, the user's age. An improvement opportunity of this exploration would be to consider different analysis techniques that could provide a causational relationship, which would be more beneficial in establishing the factors that influence book ratings. Focusing on age since it had the strongest correlation to book ratings, this could be done through an experiment where different age groups are asked to read and review books without looking at ratings beforehand and comparing the results. The results of this extension to our investigation would be greatly beneficial in aiding bookstore managers in making purchasing decisions, since they could tailor their purchases to the age group with the highest number of users, and they could also recommend highly rated books to people based on their age, which could significantly increase the probability of the user liking the books if a causational relationship is found. On the other hand, if no causation is found, they could ignore age when buying books and making recommendations, and they could rely on other methods to guide their purchases and recommendations. For example, user-user collaborative filtering.

Conclusion

The main findings of this analysis were that the relationship between the tested parameters and book ratings is best modelled by a linear regression, and that user country, user age, publishing date and book title are not significant predictors of book ratings, with all features having $|r| < 0.2$. Usually, a reason for a low correlation is a small sample size, but this analysis had a sample size of 48,299 book ratings, which can be considered representative. Another reason for the low correlation could be bad data quality. This reason could be more sustained given that in the data pre-processing stages, there were a lot of missing values, such that we had to create categories "others" or "not specified" for user countries and age respectively. Therefore, the missing values interfered with the model's ability to establish a correlational relationship between given variables. Through our exploration, we were able to establish that different machine learning models are useful in different situations, and that the best model for establishing correlation in this case was linear regression, whilst the best model for prediction accuracy was a Decision Tree Classifier. The best strategy moving forward with this exploration would be to implement an experimental research design, which would allow us to establish a causal relationship between our factors. Altogether, the objective of this exploration was to determine whether there were any factors that influenced the rating people gave to books, which was accomplished since our analysis showed that there are no factors that significantly affect book ratings.

References

Bhatia, R. (2021, February). *Book-Crossing: User review ratings*. https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset/version/3

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2021). Mass-scale emotionality reveals human behaviour and marketplace success. *Nature Human Behaviour*, *5*(10), 1323–1329. https://doi.org/10.1038/s41562-021-01098-5

Simonton, D. K., Graham, J. J., & Kaufman, J. C. (2012). Consensus and Contrasts in Consumers' Cinematic Assessments: Gender, Age, and Nationality in Rating the Top-250 Films. *Psychology of Popular Media Culture*, *1*(2), 87–96. Journals@OVID. https://doi.org/10.1037/a0027153

Sinan, A. (2014, January). *The Problem With Online Ratings*. 6.

https://learning.oreilly.com/library/view/the-problem-with/53863MIT55224/

The pandas development team. (2024). *pandas-dev/pandas: Pandas* (v2.0.3) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10957263

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P.,

Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

We acknowledge the use of ChatGPT [https://chatgpt.com/] to aid in this technical report. Various prompts were used, all with informative purposes, and all work included is our own. The use of ChatGPT is in accordance with the university's Assessment and Results Policy.