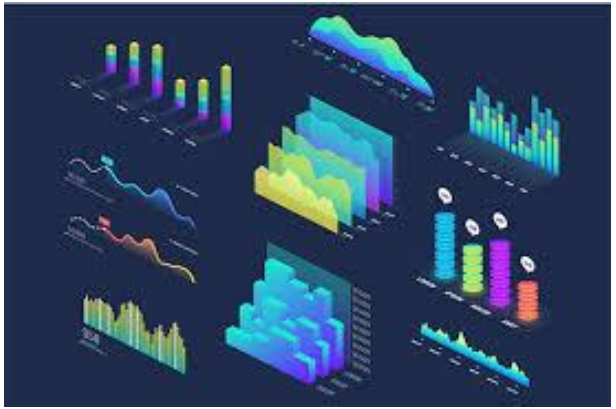
Factors Influencing Book Ratings

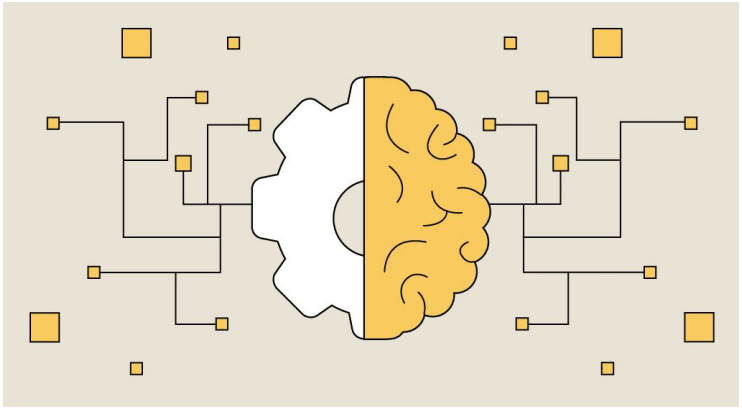
Xingzhi Du, Luting Shen, Sofía Ramírez Martínez

Task and Data Introduction

What are the factors that influence book ratings?

Methods



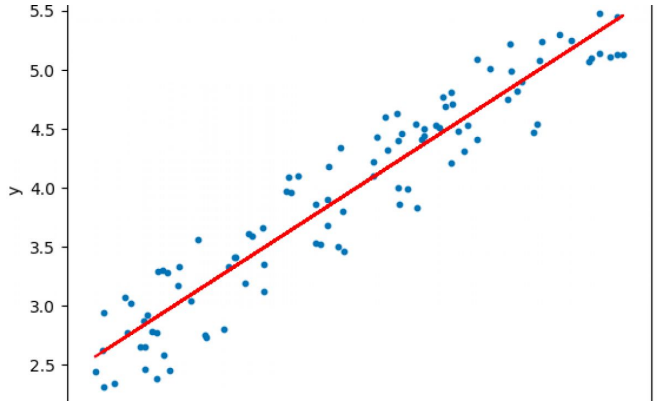


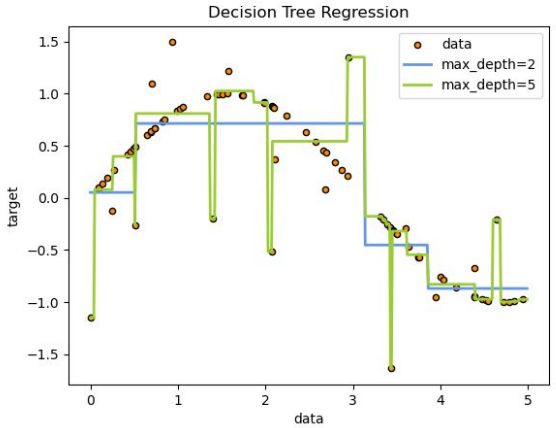
1. Pre-processing

- Users
  - Remove non-numerical
  - Encode countries
  - Feature selection
- Books
- Ratings

2. Visualisation

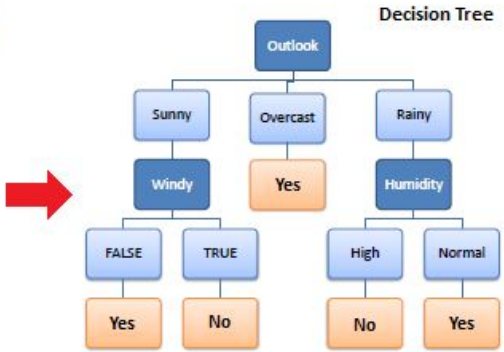
3. Machine Learning



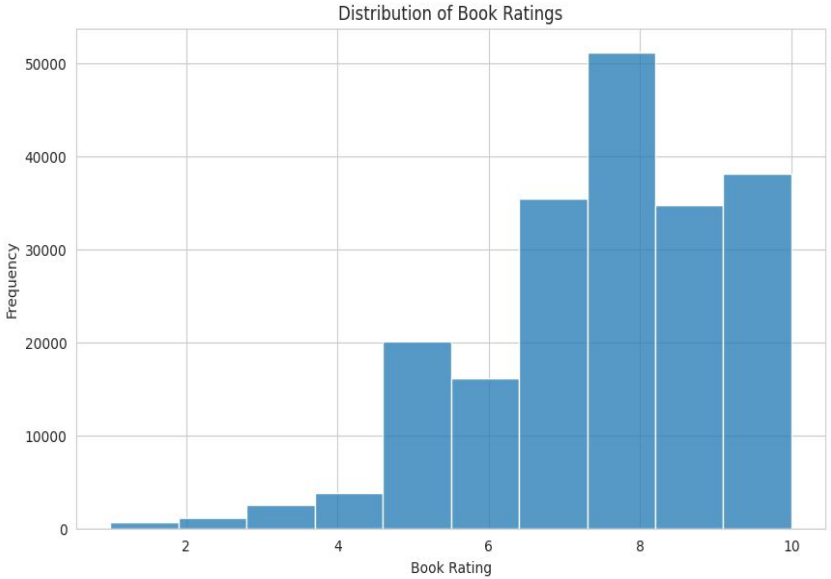


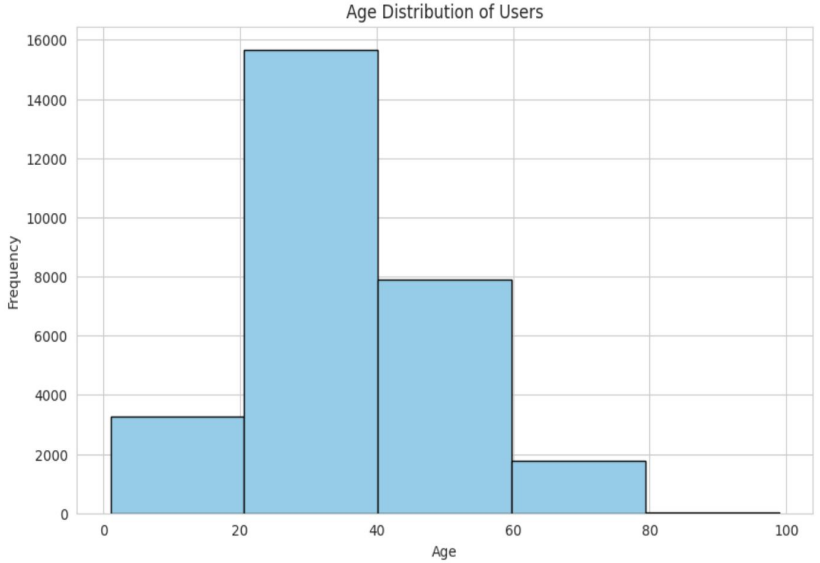
Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

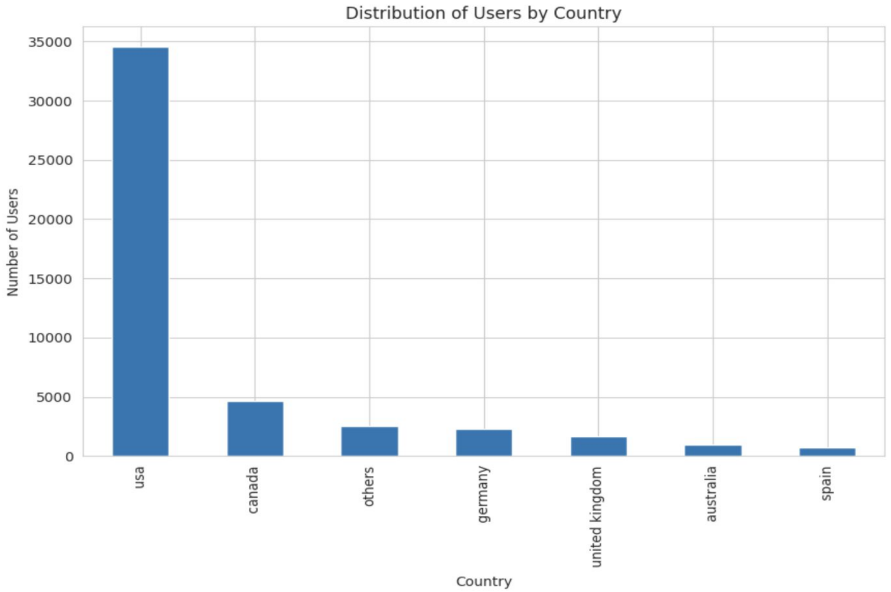


Data Exploration and Analysis

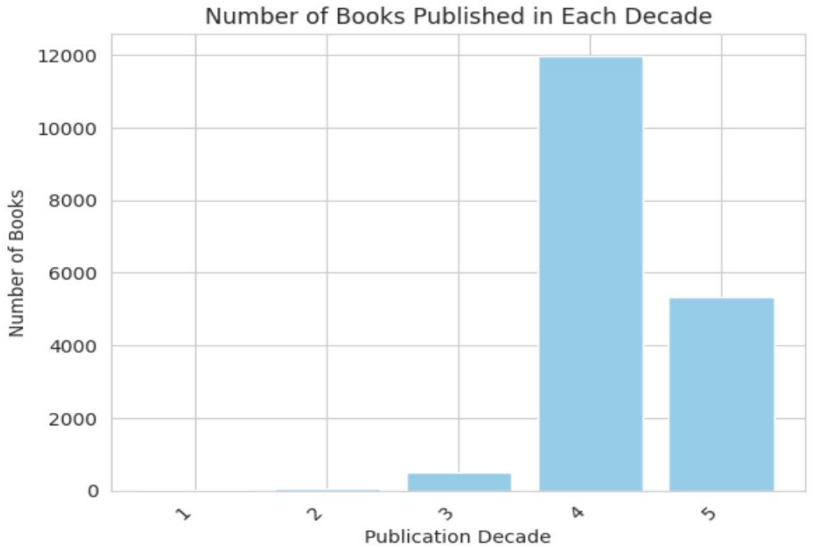




Data Exploration and Analysis

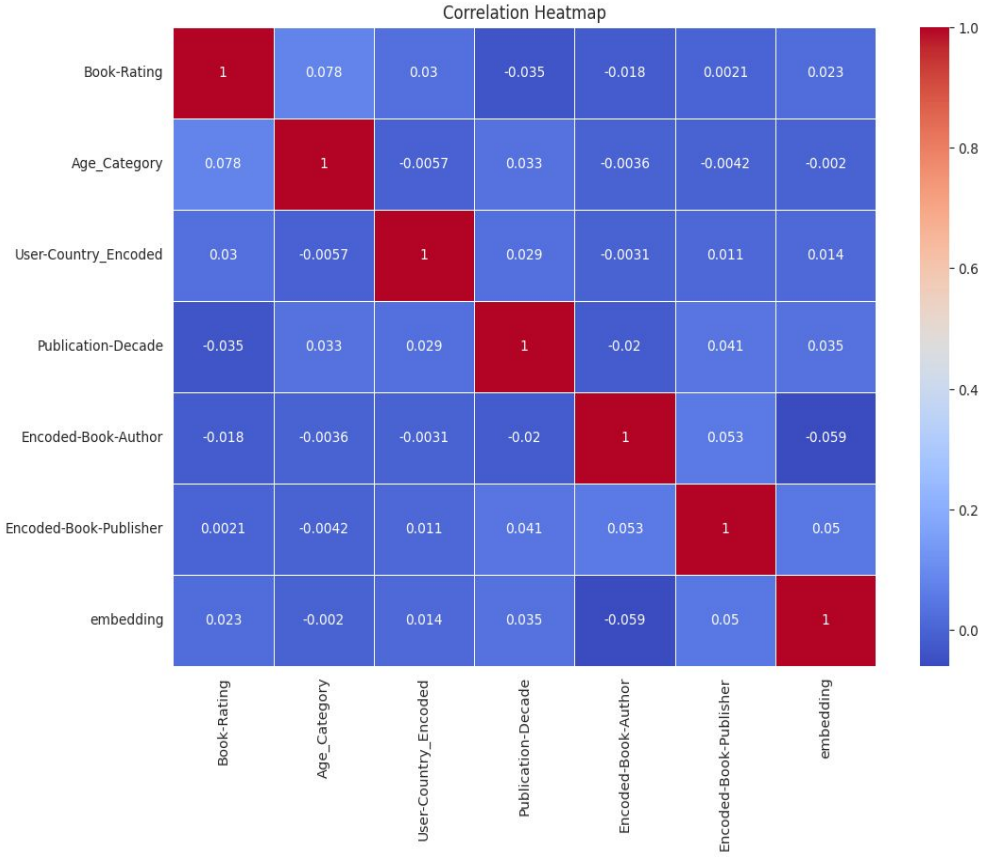


the buckets represent the following intervals: bucket 1 (1920-1940], bucket 2 (1240-1960], bucket 3 (1960-1980], bucket 4 (1980-2000], and bucket 5 (2000-2024].



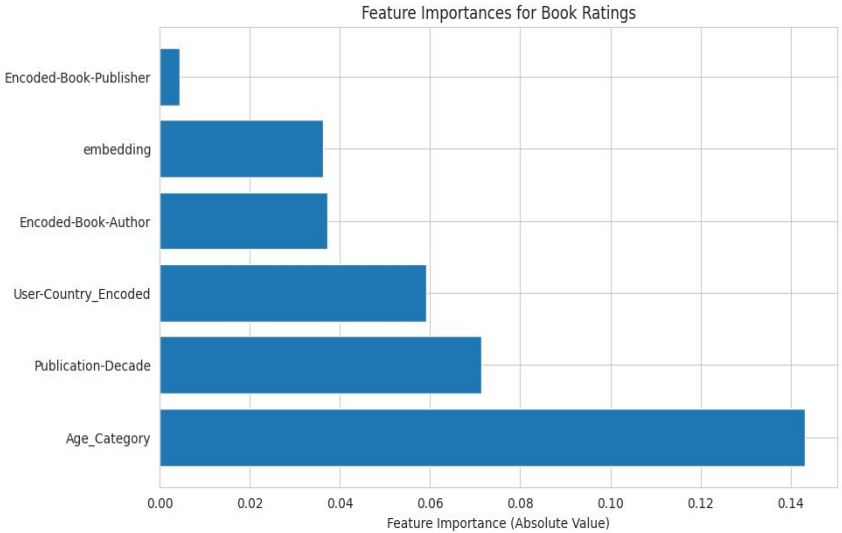
Correlation matrix

Correlation matrix and a heatmap to help identify multicollinearity – which is the presence of highly correlated independent variables in a regression analysis.

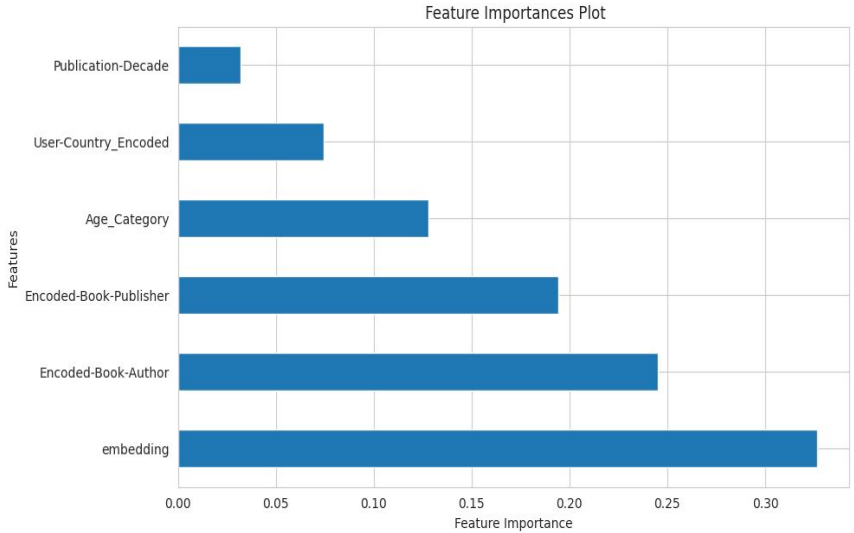


Result

- **Linear regression model**
- Mean Standard Error score is about 3.17



- **Decision tree classification model**
- Mean Standard Error score is about 5.08



Interpretation

Decision Tree Classification

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

1:moderate accuracy (0.54)

2: class 0 : rating from 0-4
Class 1: rating from 5-7
Class 2 : rating from 8-10

3:predicting higher ratings (class 2), with a higher precision, recall, and F1-score compared to the other classes

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

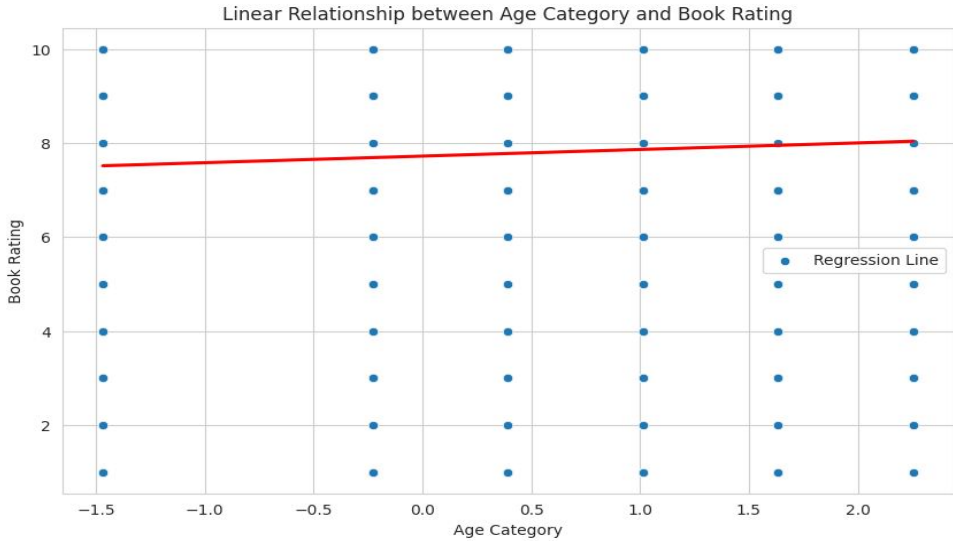
Linear Regression

Cross-validated Mean Squared Error (CV MSE): 3.172241909947137
Mean Squared Error (MSE): 3.138222792100308
R-squared (R2): 0.0085923440545484 5
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

a very low R-squared value of approximately 0.009

negative coefficient for Publication-Decade -0.071

Age Category vs Book Ratings



a positive correlation between age and book ratings

Comparisons

Decision Tree Regression

CV MSE Comparisons:
5.0846 > 3.1722

Linear Regression

```
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
```

- Decision Tree Regression ‘s higher CV MSE is unexpected as decision trees are known for their ability to capture complex relationships in data.
- The superior performance of Linear Regression suggests that the relationships between features and book ratings might be more linear in nature.
- choosing the appropriate model based on the underlying patterns in the data is important

Discussion

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

Considerations of limitations and opportunities for improvement