Goodreads: Book rating prediction report

Author: Luciana Zeferino

Project objective

The objective of this project was to train a model that predicts the rating of a book. To achieve this, I followed a standard approach consisting of data preparation, exploratory data analysis, feature engineering, modeling, and evaluation.

Data Preparation

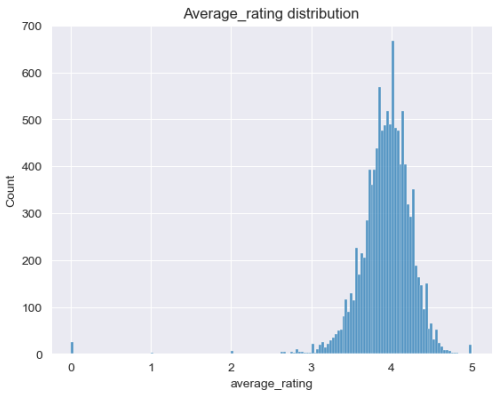
In this section, I imported libraries, imported the dataset, and performed exploratory analysis to better understand the data. Then, I cleaned the data by:

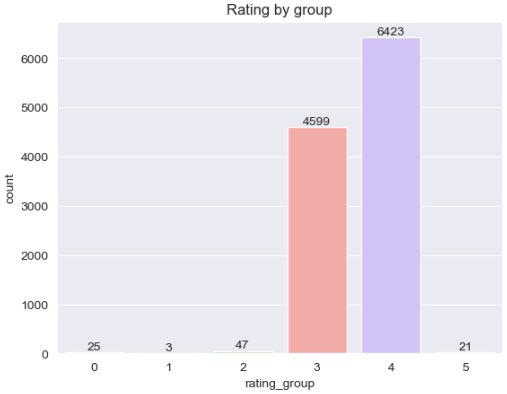
- changing the data type of "publication_date" to datetime,
- correcting some inconsistencies in the "language_code" column for the English language,
- removing the empty space at the beginning of the "num_pages" column,
- replacing books that have 0 pages by the mean of "num_pages" by language code, and
- dropping the "Not a book" rows.

Data exploration

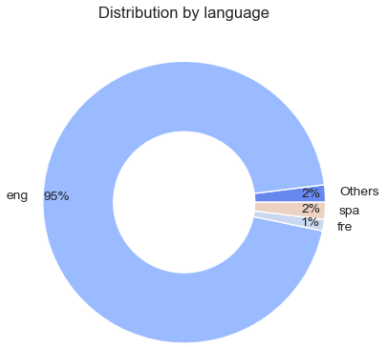
In this section, I explored the data in more depth, created visualizations, and examined the correlation between the existing features. I found that:

- Most of the average ratings follow around 4

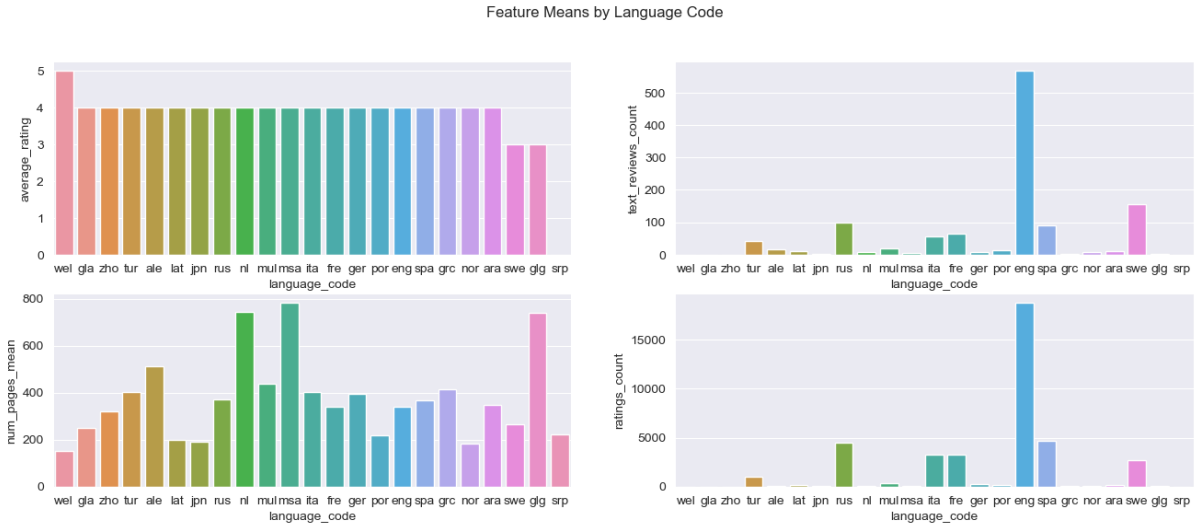




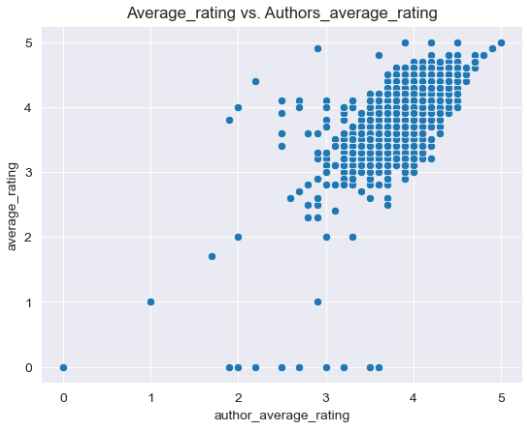
- There are significantly more books in English.



- There is no significant difference between the average rating of books across different languages



- The average rating of books is very similar to the average rating of their respective authors, and the two are positively correlated



- Books with a very large number of pages tend to be collections of multiple works

bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages
24520	The Complete Aubrey/Maturin Novels (5 Volumes)	Patrick O'Brian	4.70	039306011X	9780393060119	eng	6576.0
25587	The Second World War	Winston S. Churchill/John	4.45	039541685X	9780395416853	eng	4736.0

- The books with the highest ratings have a relatively small number of ratings count

	title	average_rating	ratings_count	text_reviews_count	num_pages
0	His Princess Devotional: A Royal Encounter Wit...	5.0	2.0	0.0	240.0
1	Willem de Kooning: Late Paintings	5.0	1.0	0.0	83.0
2	The New Big Book of America	5.0	2.0	1.0	56.0
3	Taxation of Mineral Rents	5.0	1.0	0.0	350.0
4	The Complete Theory Fun Factory: Music Theory ...	5.0	1.0	0.0	96.0
5	Oliver Wendell Holmes in Paris: Medicine Theo...	5.0	1.0	1.0	179.0

Feature engineering

Based on the insights gained from the exploratory data analysis, I performed feature engineering to select and transform the most important features for the models.

- Transform the feature language_code from categorical to numbers
- Create new features
    - **Author popularity score** : calculated by looking at the average rating of the author previous books, the number of ratings they've received, and the number of reviews, weighted by the number of books they've published

```
# Calculate the author popularity score
grouped['author_score'] = ((grouped['average_rating'] *
grouped['ratings_count'] * grouped['text_reviews_count'])
                           / grouped['title'])  # weighted by
number of books published
```

    - **Authors average rating**: calculated grouping the data by author, and taking the mean of ‘average_rating’
    - **Language average rating**: calculated grouping the data by language, and taking the mean of ‘average_rating’
    - **Publication year**: Keeping only the year of date
- Dropping columns: 'title', 'authors', 'isbn', 'isbn13', 'publication_date', 'publisher'

Modeling

To predict the continuous target feature, I trained and evaluated three models: linear regression, decision tree, and random forest. I chose these models based on the data type of the target feature, which required a regression model, as well as their popularity for this type of problem. I evaluated the models using accuracy and error metrics such as mean absolute error, mean squared error, and root mean squared error.

Results

1. Accuracy assessment: This metric measures how well the model predicts the target variable, on a scale from 0 to 1. An accuracy of 1 indicates a perfect prediction, while a lower accuracy score means that the model's predictions are less accurate. In this case, the Decision Tree model has the highest accuracy of the three models (0.9999999078694201), followed by the random forest model (0.9796798714221829), and the linear regression model (0.875889358098295).

Linear Regression	Decision Tree	Random Forest
0.8758	0.9999	0.9796

2. Mean absolute error (MAE): This metric measures the average absolute difference between the model's predicted values and the actual values. The lower the MAE, the better the model's predictions. In this case, both linear regression and random forest models have the same MAE (0.07), while the decision tree model has a slightly higher MAE of 0.09.

Linear Regression	Decision Tree	Random Forest
0.07	0.09	0.07

3. Mean squared error (MSE): This metric measures the average squared difference between the model's predicted values and the actual values. The lower the MSE, the better the model's predictions. In this case, all three models have the same MSE value of 0.02.

Linear Regression	Decision Tree	Random Forest
0.02	0.02	0.02

4. Root mean squared error (RMSE): This metric is the square root of the MSE, and is used to measure the average difference between the model's predicted values and the actual values. The lower the RMSE, the better the model's predictions. In this case, the decision tree model has the highest RMSE (0.17), followed by the random forest model (0.13), and the linear regression model has the lowest RMSE (0.12).

Linear Regression	Decision Tree	Random Forest
0.12	0.17	0.13

In summary, the results suggest that the decision tree model has the highest accuracy but also has the highest RMSE, while the linear regression model has the lowest RMSE, but lower accuracy. The random forest model appears to have a good balance between accuracy and RMSE, and has the same MAE as the linear regression model.