PYTHON MACHINE LEARNING PROJECT

GOODREADS BOOK RATING PREDICTION MODEL

PYTHON MACHINE LEARNING LABS
Joyee Banerjee - 20210300
Applied MSc in Data Analytics 2021-23
Data ScienceTech Institute

Data ScienceTech Institute

Table of Contents

I. **Introduction** 3

Overview 3

II. **Methodology** 3

Csv File 3

Data Extraction and Processing 4

Data Cleaning 4

Exploratory Data Analysis 4

Feature Engineering 6

Multiple Machine Learning Models 6

Compare ML Models 6

III. **Conclusion** 6

Takeaway 6

Limitations and Areas of Improvement 7

Tools and Web Services used & Files Included 7

I. Introduction

OVERVIEW

'Books are a uniquely portable magic.' – Stephen King.

Books touch everyone's lives knowingly or unknowingly. They hold a plethora of knowledge or help us transport ourselves to live exciting adventures or travel back in time. Irrespective of the kind or genre of book that one prefers to enjoy, it is undeniable that a good recommendation or a good rating definitely helps direct us to the ones that have been loved for generations or the new arrivals in the market.

This report will help explain the various steps involved in processing, cleaning and exploring the database given to understand the key insights and trends observed. These observations will further help in selecting the relevant features to train and test the machine learning models applied to the database to help predict the book's rating.

II. Methodology

CSV FILE

I noticed that there was an error while loading the given books.csv dataset as there was data included in an additional column 'M'.

To solve this issue I used Excel to filter all the columns and I noticed for four of the rows they had data in the column 'M' as the author column was shifted to the average rating column and hence consequently all the following columns for these respective rows had also moved.

Then to fix the columns I shifted the additional authors to author column and separated them by '/' and copied the remaining data in their place.

The above steps could have also been done using python but since I'm new to the language, Excel was faster way to solve this issue. I renamed the corrected file books_corrected.csv.

DATA EXTRACTION AND PROCESSING

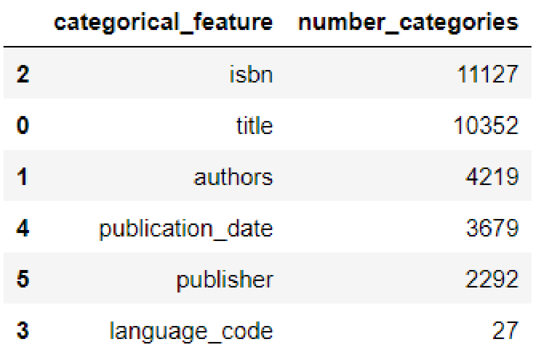
The file books_corrected.csv was used for this project.

The dataset was found to have 11127 rows and 12 columns.

DATA CLEANING

The dataset was found to have no missing or duplicate values.

The dataset was found to have six columns with categorical data type and the remaining six with numerical data type.



	categorical_feature	number_categories
2	isbn	11127
0	title	10352
1	authors	4219
4	publication_date	3679
5	publisher	2292
3	language_code	27

Figure 1: Columns with categorical values.

Two books were found to have missing publication dates which were then googled and inserted manually.

The correlation matrix found the relationship between the numerical columns like between the ratings_count and text_reviews_count columns.

EXPLORATORY DATA ANALYSIS

To understand the dataset better several visualizations were created to derive insights.

It was observed that the most number of books were written in English (eng) with variations of the language differentiating on country like en-US and en-GB.

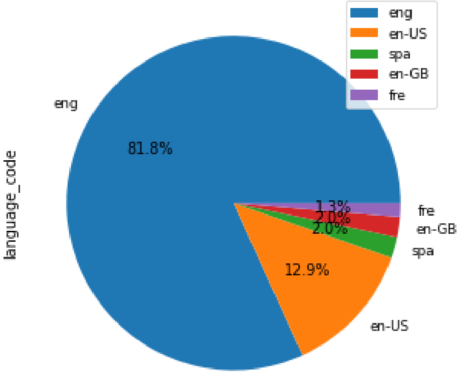


Figure 2: Top 5 languages

Then visualizations were created to answer the following questions:

1. Top 10 highest rated books
2. Top 10 highest reviewed books
3. Top 10 books under 200 pages
4. Top 10 longest books (with the most number of pages)
5. Top 15 published books
6. Top 5 authors with the highest rated books
7. Authors with the highest publications

It was also observed that for most books the ratings were between 3.7 to 4.3 and the books with rating higher than 5 were comparatively much lesser.

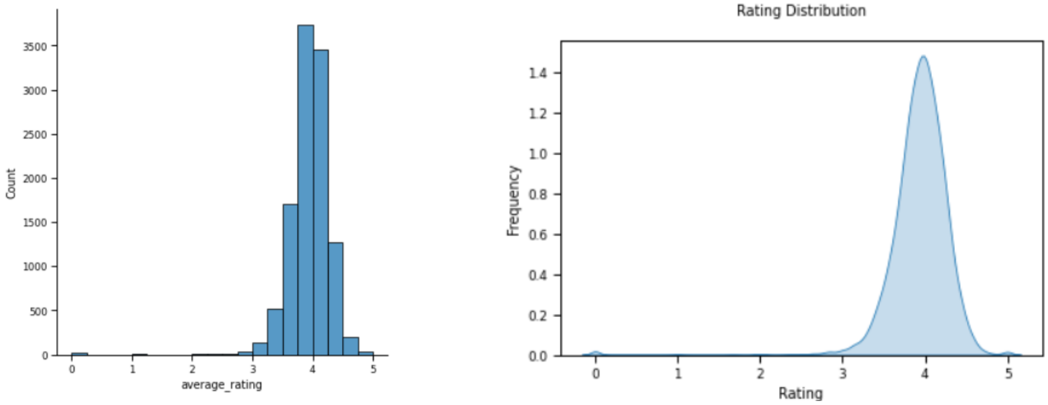


Figure 3: Book rating distribution

It was also observed that majority of the ratings were between 3 and 4 followed by ratings between 4 and 5.

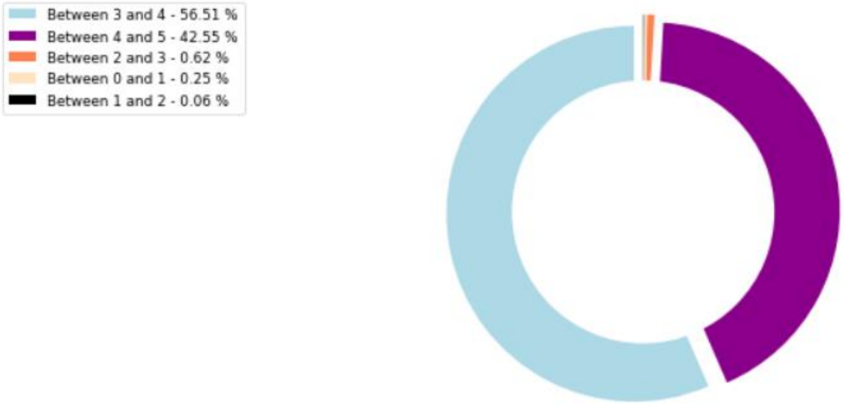


Figure 4: Book rating distribution split

Then the following were explored:

1. The relationship between ratings and review counts
2. The relationship between number of pages and ratings
3. The relationship between text reviews and ratings

FEATURE ENGINEERING

Based on the above analysis the outliers from the num_pages, ratings_count and text_reviews_count were removed.

The categorical columns such as language_code, title and authors were encoded.

MULTIPLE MACHINE LEARNING MODELS

The dataset was split into 80: 20::Train data: Test data for applying the machine learning models.

The linear regression, random forest regression and decision tree regression with AdaBoost were used. The language_code, isbn and average_ratings columns were used as attributes and the average_ratings column was used for the label.

COMPARE ML MODELS

The predicted percentages of all three models namely the linear regression, random forest regression and decision tree regression were close to the actual ones but none of them were precise.

The random forest regression was observed to be the best performing model.

III. Conclusion

TAKEAWAY

This machine learning project using python helped to understand the way to get started with a dataset, to clean and manipulate it to get it ready to train with machine learning models. The various steps involved from start to finish, highlighted the importance of exploring and understanding the data and using visualizations for better analysis and comprehension of the important features.

LIMITATIONS AND AREAS OF IMPROVEMENT

This being the first machine learning project carried out with python, it too some time to understand the dataset and selecting the key features.

All the ML models selected for this project exhibited results that were close but not precise. This indicates that the features could have been better pruned and selected and perhaps use of other ML models might have yielded better results.

TOOLS AND WEB SERVICES USED & FILES INCLUDED:

Following are the tools and web resources used for the project:

i) Anaconda to launch Juptyer Notebook using Python 3.7
ii) The environment 'classroom' created during the lab was used.

The files created for this project are as follows:

i) books_corrected.csv: manipulated file in MS Excel to correct the displaced columns.
ii) GoodreadsBooksPrediction_PythonML_JoyeeBanerjee.ipynb: Jupyter notebook.
iii) GoodreadsBooksPrediction_PythonML_JoyeeBanerjee.html: HTML version of Jupyter notebook.
iv) ReadMe.txt: Necessary packages and Installations instructions.
v) GitHub link: https://github.com/jybrj/PythonML