Book Rating Prediction Model Project

Project made by : **Samuel Bunicic** Professor : **Hanna Abi Akl**

1. Project Summary :

Goodreads dataset includes data on books and their ratings based on input from real users. It can be used for different kind of predictions which can later contribute to the development of quality products for the customers.

In the project, I tried to predict books rating using 2 different models. First is Linear regression and second is Random forest regression.

Goodreads dataset contains 12 different values:

1) **book ID**: A unique identification number for each book.
2) **title**: The name under which the book was published.
3) **authors**: The names of the authors of the book. Multiple authors are delimited by "/".
4) **average_rating**: The average rating of the book received in total.
5) **isbn**: Another unique number to identify the book, known as the International Standard Book Number.
6) **isbn13**: A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.
7) **language_code**: Indicates the primary language of the book. For instance, "eng" is standard for English.
8) **num_pages**: The number of pages the book contains.
9) **ratings_count**: The total number of ratings the book received.
10) **text_reviews_count**: The total number of written text reviews the book received.
11) **publication_date**: The date the book was published.
12) **publisher**: The name of the book publisher.

2. Project Objectives

The objective of the project is to use provided dataset in order to train a model that predicts a book's rating.
For making this prediction possible it is needed to make the following processes beforehand:

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes)
- Feature selection (feature engineering, feature pruning, choice justification)
- Model training (motivation for selected model, comparison of different models)
- Model evaluation (evaluation metric, results interpretation)
- Project report (short report explaining the approach and results)
- Project reproducibility (requirements file with necessary packages, README file for running the project)
- Project hosting on GitHub: https://github.com/koprivasam/BookRatingPredictionModel

3. Data Analysis

Cheeking for data composition

```
# Dataset composition (first and last 4 rows)
books.fillna(4)
```

```
# Statistical analysis of dataset
books.describe()
```

Cheeking for Datatype, Count and Column

```
#Description of dataser:Datatypes,Count,Column
books.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11127 entries, 0 to 11126
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   bookID              11127 non-null  int64  
 1   title               11127 non-null  object 
 2   authors             11127 non-null  object 
 3   average_rating      11127 non-null  float64
 4   isbn                11126 non-null  object 
 5   isbn13              11127 non-null  float64
 6   language_code       11123 non-null  object 
 7   num_pages           11127 non-null  int64  
 8   ratings_count       11127 non-null  int64  
 9   text_reviews_count  11123 non-null  float64
 10  publication_date    11125 non-null  object 
 11  publisher           11123 non-null  object 
dtypes: float64(3), int64(3), object(6)
memory usage: 1.0+ MB
```

We can see that the dataset contains 11127 entries with 12 columns

```
# Check for null values
books.isna().sum()
```

```
bookID              0
title               0
authors             0
average_rating      0
isbn                1
isbn13              0
language_code       4
num_pages           0
ratings_count       0
text_reviews_count  4
publication_date    2
publisher           4
dtype: int64
```

From this we can see that 5 columns (isbn, language_code, text_reviews_count, publication_date and publisher) contain some null values

4. Cleaning the data and feature selection

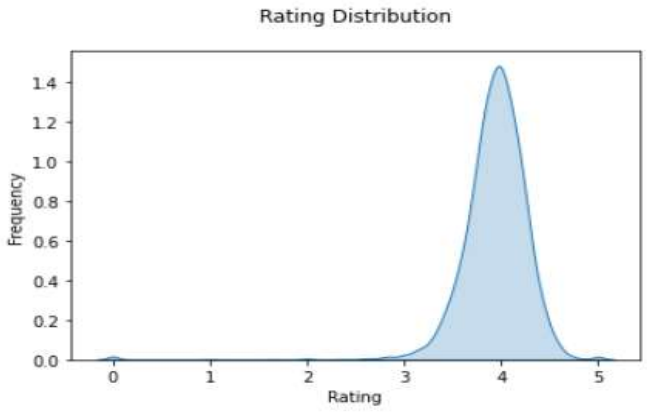
Objective of cleaning of data usually consists of formatting the text, removing special characters, removing extra whitespaces, clean parsing errors (publishers separated with comma).

- Drop the columns **isbn** and **isbn13**, because it has no impact on the target variable "**Average rating**"
- Pick the language_code as series and then convert it into a Set
- Join the Language codes: en-US, en-GB,en-CA to eng
- Looking for distinct values for language code, and when there are different codes with the same meaning we merge in one.
- Explore the distinct values for language_code after merging same Language code
- Clean some column

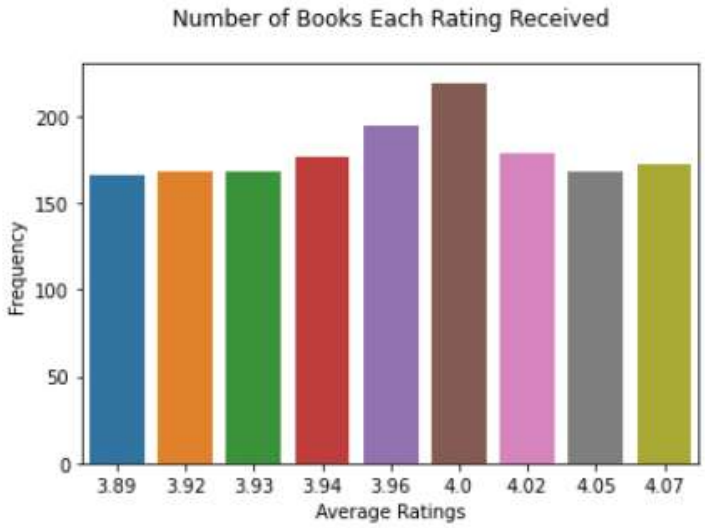
5. Visualization of reverent attributes

As the aim is to predict books' rating, it is beneficial to explore more data regarding average rating and identity which books were most reviewed.

```
Text(0, 0.5, 'Frequency')
```



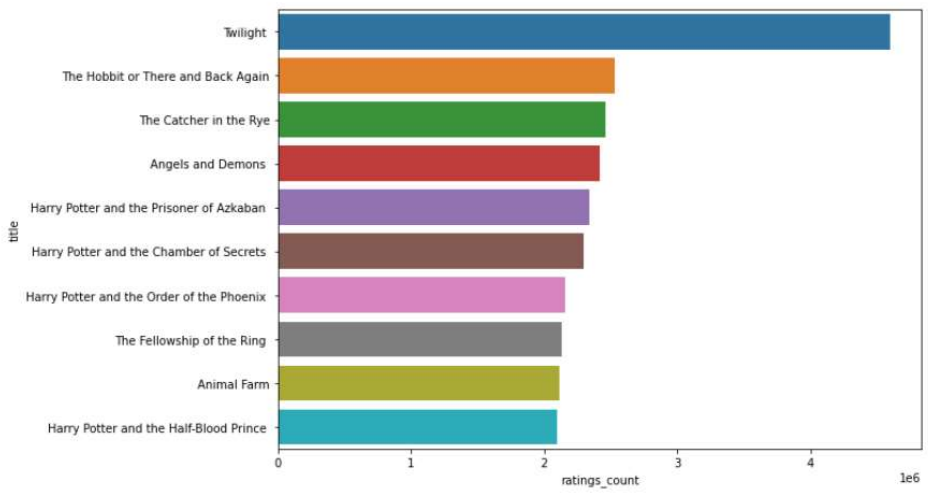

```
Text(0, 0.5, 'Frequency')
```



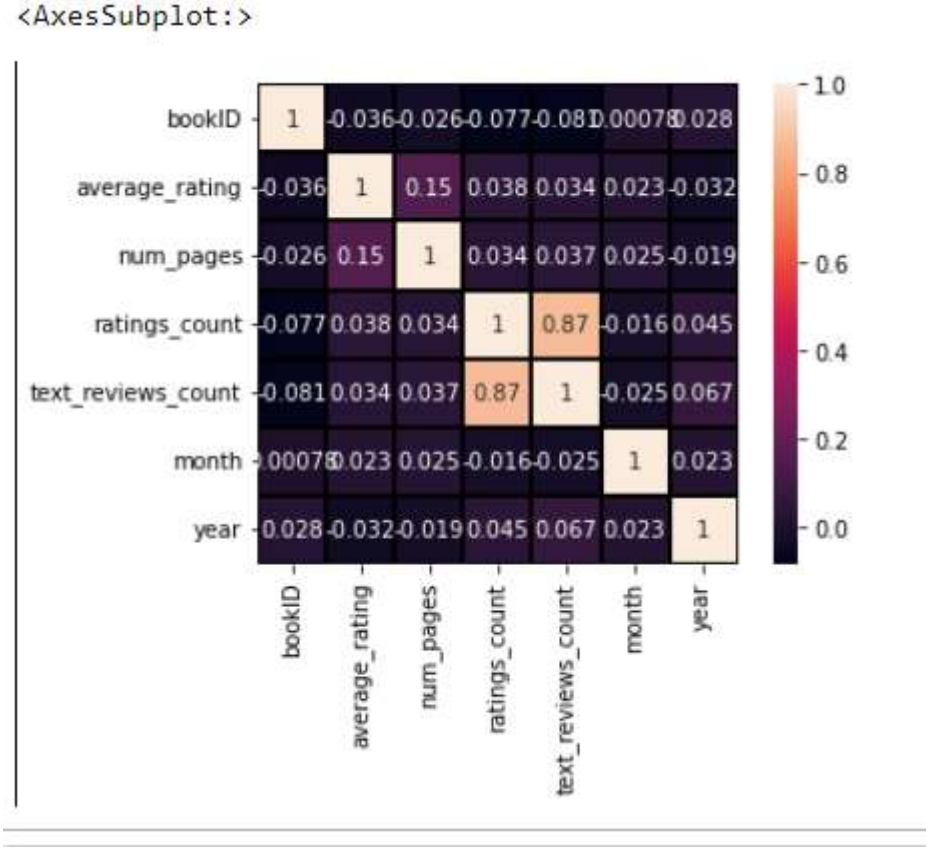
It can be seen that the majority of books have average rating close to 4.0

This barplot shows the most reviewed books in the dataset:

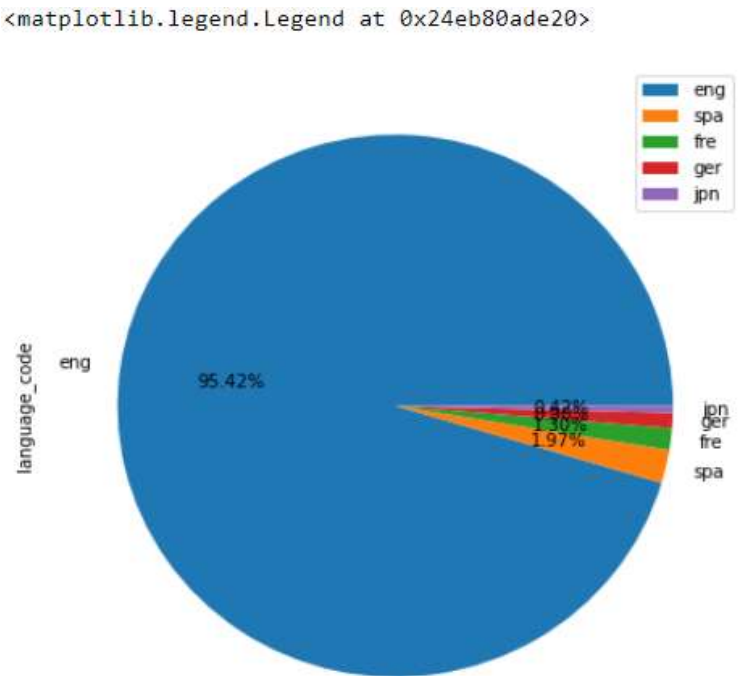
```
<AxesSubplot:xlabel='ratings_count', ylabel='title'>
```



There is a high correlation between the ratings_count and the text_reviews_count (~ 81%)



This visualisation shows top 5 languages used and we can see that English accounts for 95.42%



For the STA Packages, data from different available data sources were imported.

Project Properties had to be changed, in particular the change from 64 to 32 bits, to allow for importing the Excel file. Sometimes, it was necessary to open the task manager to finish the tasks in progress (Debug) to be able to lunch the Debugging, similarly to what was done in the course.

6. Training model

During model training I used cleaned datasets which were fed into the model. Scikit-Icarn libraries were used.
I used 2 models: first, Linear regression was used and afterwards random Forest regression.
For performing evaluation of the models I used :R-square, mean_sqrd_error and
root_mean_squared error is

7. Results

Data of Linear regresion model

R-square is 0.01894182103542874
mean_squard_error is 0.11261818181818181
root_mean_squared error is 0.33558632543383204

Data of Random Forest regression

R-square is 0.17614237683496647
mean_sqrd_error is 0.09457272727272728
root_mean_squared error is 0.3075267911462793

8. Conclusion

The lower value of root_mean_squared error (RMSE) closer to 0 preferred a better model performance. On the other hand, the higher value of R-square (R2) closer to 1 shows that the regression line fits the data and the model performance is better. Mean square error (MSE) is the average of the square of the errors. The larger the number the larger the error. This in turn means there is bigger variation between the test result and the predicted result. Therefore, the Random Forest Regressor model is much better than Liner Regression model.