Project Report

Subject – ML Predicting Books Rating

Student Name – Christian Azumah

Cohort – S22.

Date: March 26, 2023

Data Collection

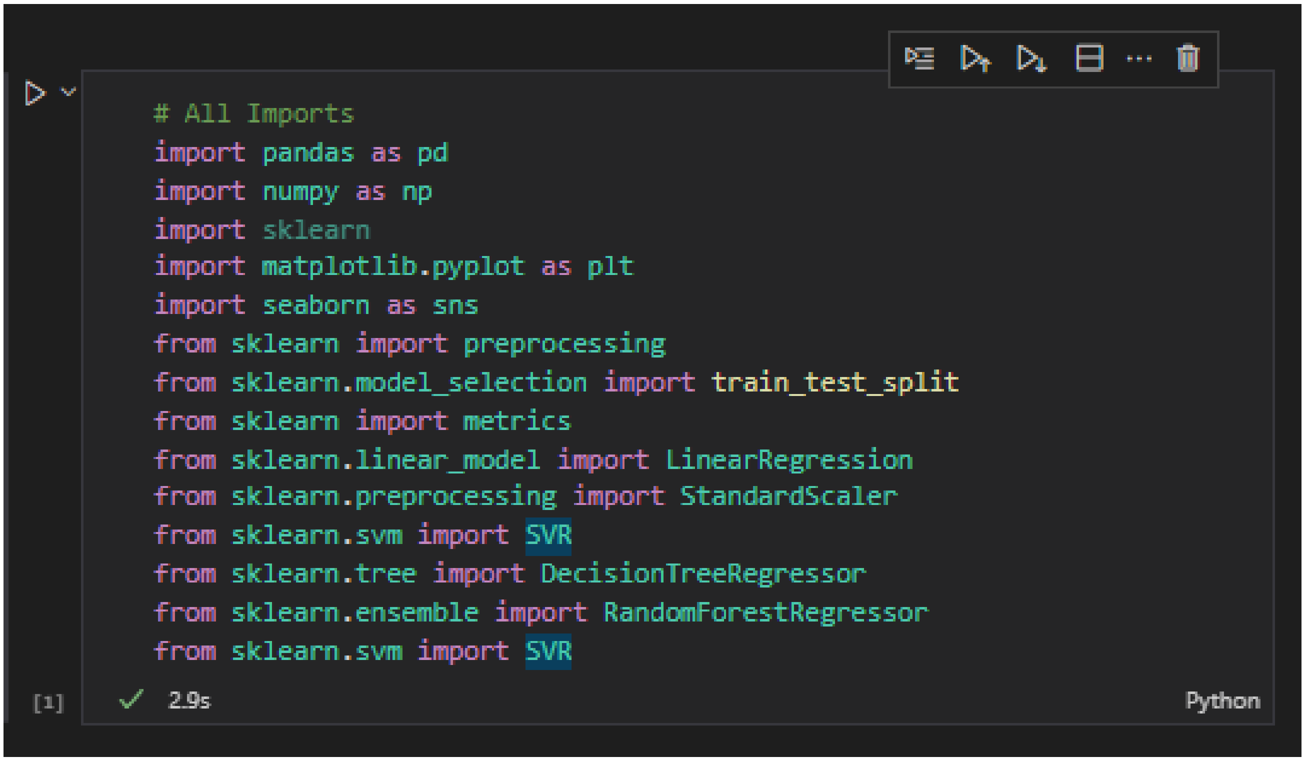
Books dataset CSV file was provided

Problem Statement - Using the provided dataset, you are asked to train a model that predicts a book's rating.

Packages and Libraries

All relevant packages and/or libraries to be used for the statistical analysis, plotting graphs, and building ML learning models.

List of Packages as below:



```
# All Imports
import pandas as pd
import numpy as np
import sklearn
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
```

Below table provides the information about dataset attributes. Data Variables

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

Data Cleaning, Wrangling and initial Analysis

Pandas pre-profiling methods were used to gain an initial understanding of the dataset, to clean the dataset and to get statistical insights of data.

1st Phase Data cleaning steps:

1. The Dataset was generally already clean and didn't require many alterations.

2. One Column variable was renamed from ' num_pages' to 'num_pages', to make it consistent with the other variables and reduce future errors.

3. Informed reasoning was applied to drop 'bookID' and 'isbn13' as they were expected to have no significant on the model due to the nature and purpose of their existence.

Data Exploration and Further Exploration

Once initial cleaning and data preparation was concluded, the next step was to analyze and understand the data characteristics in order to pick the appropriate input variables to train the model on.

This was done through a series of plots (Histograms, Boxplots, density curves, bar plot and more) with the aim of visualizing the distributions in the dataset and observing the relationship between the variables.

A correlation matrix and pairgrid of all the ‘relevant’ variables was used to identify the significant variables which would be used to train the Machine Learning models.

The Distribution of Language spread amongst the books was also investigated.

‘text_reviews_count’ & ‘ratings_count’ were found to be the relevant variables when trying to predict average rating which is the aim of our model.

The correlation matrix shows that the relationships between ‘text_reviews_count’ & ‘average_rating’ were weak at best with a correlation score of 0.04 & 0.03.

This led to the next investigation which then found a stronger correlation between books that received a high number of text_reviews (count number greater than 5,000) and average rating, as well as books that received a high number of ratings (count number greater than 100,000).

Feature Engineering

Before beginning the feature Engineering, variables shown to have a weak correlation with the significant variables were dropped.

Dropped variables;

- bookID
- isbn
- isbn13
- publication_date
- publisher

Certain languages were also chosen to be the focus of the dataset moving forward since they contained 97% of the entire dataset. The Chosen languages were 'eng', 'en-US', 'spa', 'en-GB' & 'fre'.

After that, the Dataset was split into two. One half of the dataset contained books that were highly rated (i.e. they were books that had received over 100,000 different ratings) and the other half contained book received less than 100,000 different ratings.

This was done because of the correlation relationships that were observed between ratings_count and average_rating for books that were rated highly.

Next the variables that contained non-numerical values were encoded in order to prepare the data frame for Machine Learning model.

Building the Linear Regression Model, Alternatives and Model Evaluation.

Before the data is used to train the Machine Learning model it is split into a training and testing set in an 80:20 ratio. 80% of the dataset shall be used to train the model and the remaining 20% would serve as a reference to assess the performance of the model by comparing how close the model's predictions are to the actual values in the testing set.

Once this was completed the data was fed into a Linear Regression Model for training. Since the Data was split into HighlyRatedBooks and Non-Highly rated books the Linear Regression model was trained on both Datasets.

The Linear Model Regression performed better on the HighlyRatedBooks dataset than on the Non-Highly rated dataset, as expected.

Alternate Machine Learning models were also trained on the HighlyRatedBooks dataset in an effort to see which other Regressors were more suited for use as the Machine Learning model.

the Algorithms evaluated were;

- Support Vector Regression (SVR)

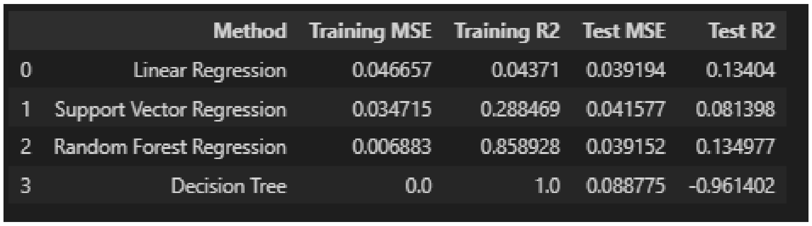
- Random Forest Regression (RDF)

- Decision Tree (CART).

The Models were evaluated by calculating a number of metrics;

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R squared value (R2)

The Predicted and actual values from each Regression model was plotted as a means of visualizing the results and comparing the alternative models.



	Method	Training MSE	Training R2	Test MSE	Test R2
0	Linear Regression	0.046657	0.04371	0.039194	0.13404
1	Support Vector Regression	0.034715	0.288469	0.041577	0.081398
2	Random Forest Regression	0.006883	0.858928	0.039152	0.134977
3	Decision Tree	0.0	1.0	0.088775	-0.961402

To Conclude, An evaluation of the 4 models shows that Linear Regression Model performs the best.

The comparison of the plots of the ML Models show that a Linear Regression Algorithm is a good model choice for this Particular Book Classification Project. The Random Forest Model is also a suitable alternative model.