Project Report

Subject – ML Predicting Books Rating
Student Name – Maria F. Parisi
Cohort – S22.

Problem Statement

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

Tool Verification and collection

Downloaded relevant packages or libraries for statistical analysis, plotting graphs, and building ML learning models.

List of Packages as below:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn as sk
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.ensemble import RandomForestRegressor
from sklearn.svm import SVR
import seaborn as sns
```

Data Collection:

Books dataset CSV file is shared to undertake the given task.

Below table provides the information about dataset attributes.

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, is 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.

Pre-Profiling and Data Cleaning:

We used the pandas pre-profiling method to understand our data distribution and to get statistical insights of data.

1st Phase Data cleaning steps:

1. We observed in 4 rows data point values are mismatched with column values. We removed those 4 observations from our dataset.
2. We then perform pandas pre-profiling report to understand our data step by step and note the observations to take our data cleaning process ahead.

As per observation from pre-profiling, we took the below steps:

➢ **BookID** is set as index.

➢ **Average_rating** variable data type has changed to float

➢ **Isbn and isbn13** valuables are removed from the data frame.

➢ **Num_pages** datatype has changed to into integer.

➢ **Publication_date** variable is removed from the date frame.

➢ **authors –** we kept only the first author’s name and rest other authors we removed from the data set

➢ **totalNumOfAuthor**- we created this new variable where we have taken the count of the total author’s list.

With this first phase of the cleaning, we started with the Visualization part to understand data behavior to pick up input variables to train the model on.

visualize below scenarios:

➢ Top 10 Authors wrote most number books

➢ Which are the highest-rated authors

➢ Which are publishers published most number of books (Top 10)

➢ Top 10 Books received the highest ratings.

➢ Which Author has received highest text reviews?

2nd Phase of cleaning and handling outliers.

1. At EDA stage we observed that **publisher** information does not contribute to the response variable, so we decided to remove the attribute from our data set.
2. Handling Outlier

➢ **num_pages** the lower limit have been considered based on data central behaviour and respective observations were removed from the data set.

In the case of and **test_reviews_count** outliers are removed. This method has been decided to arrive at an upper limit and lower limit with some scientific approach.

Post profiling

At this stage, we again took the data profile report to reassure we have taken all the necessary steps to clean data and prepare it for ML modeling.

Data preparation for Machine Learning Modelling

Below steps are taken for data preparation

➢ Label encoding for categorical data.

➢ **title** and **author** label encoded with labelEncoder method.

➢ We split the data into the train (80%) and test (20%)\

Model Selection

➢ Linear Regression Model

➢ Linear Regression Model

➢ Random Forest Model

After splitting data into train and test we fit our Random Forest Regression model on it.
And make not of the
Mean Absolute Error (MAE)
Mean Squared Error (MSE)
Root Mean Squared Error (RMSE)
R squared value (R2)
We plot a bar chat to evaluate predicted values against actual value.

Conclusion

We studied briefly the data, its characteristics, and its distribution.
We investigated the features to retain and which to discard.
With the above model testing, we can say Linear Regression Model 2 is the best model to train data
In this way, we can say our Linear Regression model2 is performing well.