DSTI: Python Machine Learning Labs

-Aritra MONDAL : https://github.com/deadlocked01/PythonLabs
-EL Mehdi EL MAHYAOUI : https://github.com/Dasty96/Python-labs
-Melissa IDRAC : https://github.com/MelissaIdrac/Pythonlabs
-Hope AKONDENG : https://github.com/HopeAkondengnotAkon/PythonLabs

1) Abstract

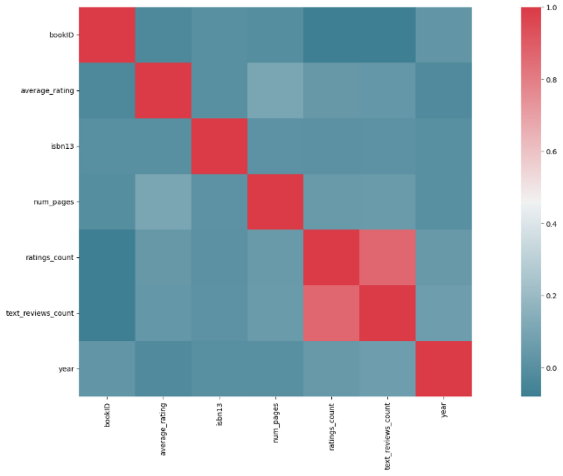
This project involves building a predictive model to estimate the rating of books based on their attributes such as author, title, genre, number of pages, publication date, and language. The dataset used in this project is obtained from the books.csv file provided and includes information on more than 10,000 books. The approach involves data cleaning, exploratory data analysis, feature engineering and selection, and model training using linear regression, and random forest algorithms. The project aims to provide insights into the factors that influence book ratings and to develop a reliable model that can be used by publishers, booksellers, and readers to estimate the value of books.

2) Data Cleaning:

- Introduction to Data Cleaning
- Removal of spaces in "num_pages"
- Identification and merging of books with same titles
- Handling of missing values for "publication_date"
- Conversion of "language_code" to a numerical column
- Addition of missing entries to "publication_date" from external sources
- Removing outlayers: Dropping rows with books that has more than 3000 pages

3) EDA: Exploratory Data Analysis

- Visualized the distribution of numerical variables using Bar graphs, Box plots, Scatter plots, and Heat maps.
- Used bar graphs to examine the distribution of book ratings based on number of books and to identify the highest count of ratings.
- Analyzed the correlation between numerical features and the target variable by creating scatter plots.
- Examined the relationship between multiple variables using heat maps (example below).



4) Feature Selection:

- Converted "language_code" to a numerical column to use it in the model training.
- Added missing entries to "publication_date" using data from Google search results.
- Considered dropping unnecessary columns like "book_id" and "image_url".
- Removed spaces before "num_pages".
- Merged duplicate entries of books with the same title.

5) Model Building:

Model building involves selecting appropriate algorithms and techniques to create a predictive model using the available data. The goal is to create a model that accurately predicts the target variable based on the input variables.

a) Algorithms Used:

- Linear Regression: used to model the relationship between the numerical features and the target variable, aiming to find a linear equation that best fits the data.
- Random Forest Regression: used to model the relationship between the features and the target variable using decision trees. The random forest model combines multiple decision trees to make predictions that are more accurate.

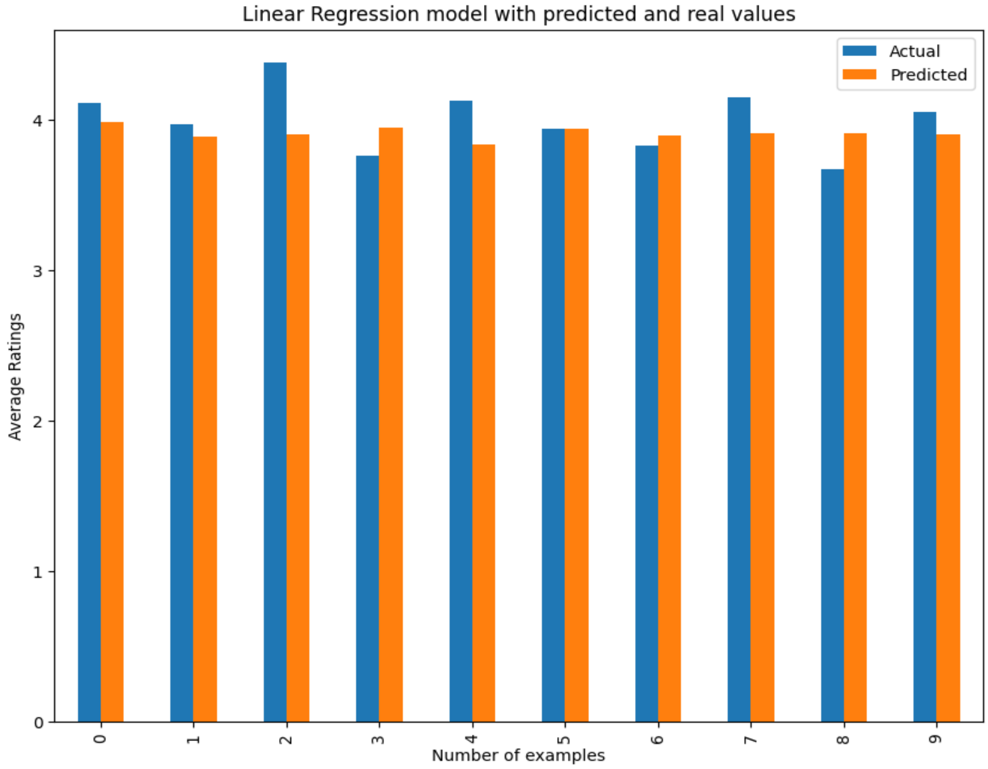
b) Model Performance Evaluation:

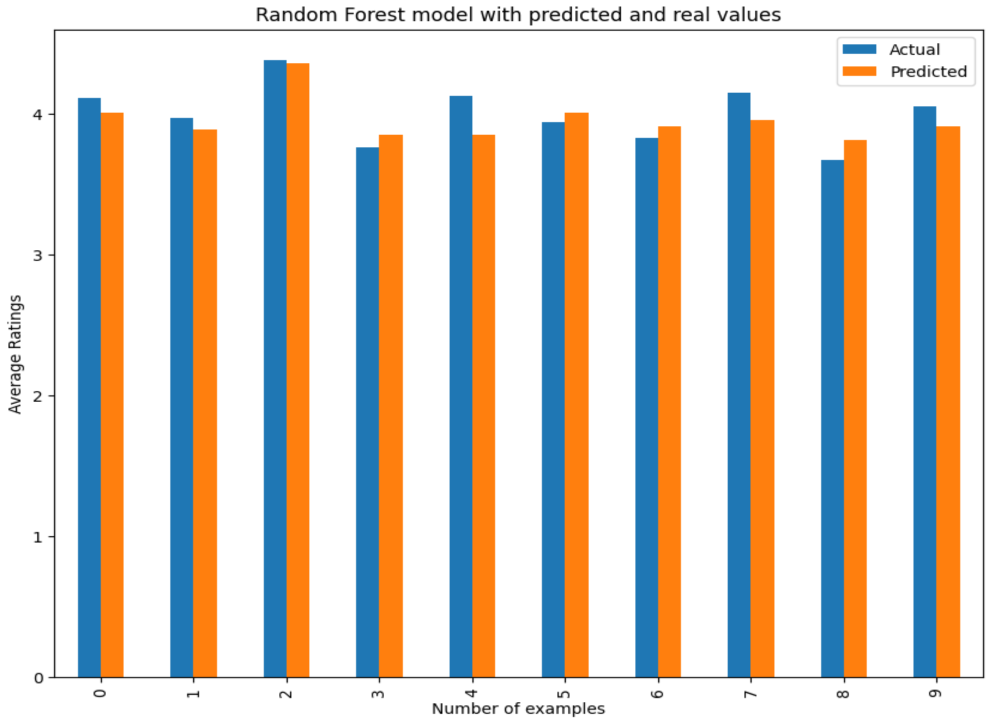
- The model performance was evaluated using metrics such as mean squared error (MSE), R-squared (R^2), mean absolute error (MAE), and root mean squared error (RMSE). For the linear regression model, the MSE was 0.126 for the training data and 0.099 for the testing data, while the R^2 was 0.016 for the training data and 0.003 for the testing data. The MAE was 0.231 for the training data and 0.225 for the testing data, while the RMSE was 0.355 for the training data and 0.314 for the testing data.

- The Random Forest model had a mean squared error (MSE) of 0.014 for the training set and 0.102 for the test set, while the R-squared (R^2) value was 0.888 for the training set and -0.035 for the test set. The mean absolute error (MAE) was 0.079 for the training set and 0.213 for the test set. Additionally, the root mean squared error (RMSE) was 0.120 for the training set and 0.320 for the test set.

6) Conclusion:

Based on the model performance evaluation, the Random Forest algorithm outperformed the Linear Regression algorithm in terms of mean squared error, mean absolute error, and root mean squared error. However, the R-squared value was higher for the Linear Regression model in the training set. Therefore, considering all evaluation metrics, the Random Forest model could be considered the better performing model.


Linear Regression model with predicted and real values
Actual
Predicted
Average Ratings
Number of examples


Random Forest model with predicted and real values
Actual
Predicted
Average Ratings
Number of examples