Machine Learnings on books data project report

1. Define the Problem:

We have books.csv data from Kaggle. It has many columns which we use as features and finally try to predict value of rating of the book using prediction using Machine learning in Jupiter. We use packages mentioned in requirement files for the functionality.

2. Collect and Prepare Data:

We read data in Jupiter and then delete all unnecessary columns like "isbn13", "isbn", "title", "publication_date“ and also modify some columns for numpages. Then we check null values and manage them by averaging if needed.

3. Exploratory Data Analysis (EDA):

We generate various graphs to see relation of selected features like numpafes, Author, publisher,publisher, languagecode,rating count with each other and most importantly with rating

4. Feature Engineering:

Corellation Matrix is best to select features for Machine learning

5.Data Splitting: Divide your dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used to tune hyperparameters, and the testing set is used to evaluate the model's performance.

6.Select a Model: Choose an appropriate machine learning algorithm or model architecture based on the nature of your problem (classification, regression, clustering, etc.). Consider factors like interpretability, complexity, and performance.

7.Model Training: Use the training data to train the chosen model. This involves adjusting the model's parameters to minimize the difference between its predictions and the actual target values.

Models we used:

XGBClassifier

RandomForestClassifier

MLPClassifier

DecisionTreeClassifier

8.Hyperparameter Tuning: We do encoding indata by converting to numeric values of language, Author, publisher .

9. **Model Evaluation:** Evaluate the model's performance on the testing set using appropriate metrics. Common metrics include accuracy, precision, recall, F1-score, and Mean Squared Error (MSE), depending on the problem type.

Our score:

Mean Absolute Error: 0.4314606741573034, Mean Squared Error: 0.4350561797752809

Root Mean Squared Error: 0.659587886316358

R-squared: -0.719836694897528

10.Model Interpretation (Optional): Depending on the complexity of the model, interpret its decisions. Techniques like feature importance analysis, SHAP values, and model visualization can help in understanding how the model arrives at its predictions.

11Monitoring and Maintenance: Continuously monitor the model's performance in the real-world environment. Drift in data distribution or degradation in performance over time might require retraining or fine-tuning the model.

12.Iterate and Improve: Machine learning is an iterative process. Use insights from monitoring and user feedback to make improvements to the model and the overall process.

13.Deployment: We finally deploy on Github. Follow readme file to run the project