Data Science Assignment

Objective

- Develop a model to predict the 'Impact' of a book, a composite score achieved post-publication, using its attributes and metadata.

Dataset

- Provided dataset contains book descriptions and metadata. [LINK]
- Given the books' attributes, model the 'Impact'. You can filter/subsample/oversample and make necessary assumptions to enable modeling on the given data.

Tasks

- Set up and run Spark locally: Simulate different worker configurations.
- Develop an Application:
    - Load the dataset from CSV.
    - Implement preprocessing and feature engineering pipelines.
    - Create and train a regression model.
    - Track and save cross val MAPE and Total Training Time for different worker configs.
    - Run the app with 1, 2, and 4 simulated workers.

What We're Looking For

- **Code Quality**: Use of best practices. (e.g. Use pre-processing pipelines, etc.)
- **EDA**: Using statistical methods, tell us a few things about the data.
- **Experimentation and Tracking**: Eg. Usage of tools for tracking multiple experiments while engineering features/modeling.
- **Performance Analysis**: Differences in cross val MAPE and Training Time with varying worker counts.

Additional Requirements

- **Programming Language**: Use high-level languages like Python or Spark (We don't work with statistical languages like R and SAS).
- **Model Building**: You can select one or more model building algorithms for building the model at this stage. Please explain your reasoning for choosing a particular algorithm. Also, any suggestions on how to improve the model you have built here fetches brownie points.
- **Documentation**: Clearly document your approach and rationale. Focus of the evaluation of assignment is on the approach you take for different steps and not on the actual model results.

Submission

- A one-page write-up explaining your solution and thought process.
- Well-documented code.
- Share results and code via GitHub.