Data ScienceTech Institute

Applied MSc in Data Analytics

Course: Python Machine Learning Labs

Project: Book Rating Prediction Model

Instructor: Hanna Abi Akl

Project Summary:

"There is no friend as loyal as a book." - Ernest Hemingway

Nowadays with so many books available, it can be hard to select the best ones to read. The dataset provided is a curation of Goodreads books based on real user information. It can be used for many tasks like predicting a book's rating or recommending new books.

Below is the information you have regarding the dataset attributes:

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

Project Objectives:

Using the provided dataset, you are asked to train a model that predicts a book's rating. The project can be submitted as a Jupyter Notebook and should include exploratory analysis of the data, feature engineering and selection, model training and evaluation.

You may use additional resources from those that are suggested in the "Project Resources" section or others as you see fit (provided you can justify how they can serve your solution). You can even consult similar solutions from the Internet. **However, this comes with a big responsibility: any submission that is over-plagiarised or does not reflect personal work will not be accepted**.

Project Resources:

Here are additional resources that may be helpful for the project. These resources are not mandatory to use but are meant to give you ideas on enriching the data or analysing the attributes in the dataset.

- Goodreads Datasets
- Recommending Goodreads Books using Data Mining

Project Evaluation:

The project will be evaluated using the following rubric. It contains the required items for a complete submission as well as bonus elements. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes) **[1 point]**
- Feature selection (feature engineering, feature pruning, choice justification) **[1 point]**
- Model training (motivation for selected model, comparison of different models) **[1 point]**
- Model evaluation (evaluation metric, results interpretation) **[1 point]**
- Project report (short report explaining the approach and results) **[1 point]**
- **BONUS:** Project reproducibility (requirements file with necessary packages, README file for running the project) **[1/2 point]**
- **BONUS:** Project hosting (Github, Docker, AWS, Heroku or any other method) **[1/2 point]**

Project Timeline:

The deadline for the project is **25 June 2022**. Additionally, you are free to set a meeting with the instructor to discuss possible approaches, problems or other points pertaining to the project.