Project: Book Rating Prediction Model

Diavila Rostaing Engandzi, Tanattiya Rungtham, Yohan Walter Jothipala
Data ScienceTech Institute, School of Engineering, Paris, France
{diavila-rostaing.engandzi, tanattiya.rungtham, yohan.jothipala}@edu.dsti.institute

Abstract: This report uses the dataset of Goodreads books based on real user information. The main objective of the project is to build a model that can predict book's rating successfully with good accuracy. This model provides the prediction of the book ratings based on the column values from dataset, for example; average_rating, num_pages, ratings_count, text_reviews_count etc. This report provides a flow of handling big data files, data engineering, building models and providing predictions. The model predicts book rating using Python libraries such as Numpy, Pandas, seaborn, missingno, smogn etc. Cross Validation is used to train the model using training set and make predictions on the validation set. Finally, several models are compared to identify the best model and the technique.

1. Introduction

This report uses Python libraries to predict the ratings for books using Goodreads dataset. Goodreads [1] is the world's largest site for readers and book recommendations. It was founded by Otis Chandler and Elizabeth in 2007. It is used to find good books to read and track the books that we are reading and want to read, provides book recommendations and suggestions and etc.

The dataset, books.csv contains the information on 200M book reviews for 45,000 unique books and user reviews for each and every book. This file consists of columns such as book title, authors, average ratings, isbn no., language, number of pages, rating count, text review count, publication date, publisher etc.

When a user posts a review on goodreads, they have the option to write a full review about the book they read. The full review allows for a detailed discussion of the book, often covering the characters, theme, and personal impressions, while the book rating provides a value between 1-5 that captures the essence of the review on a number; num 1 being the worst and num 5 being the best. In this report, full review is not been used, but the book rating.

2. Related Work

Karthic Guna (Sep,2019) [2] discusses a data mining approach for recommending books using the Kaggle's Goodreads-books dataset to explore data. It explains analyzing data through visualizations, data preparation (ordinal encoding, handling missing values), feature engineering, building multiple machine learning models, make predictions using each model and comparing models to gain a good accuracy by training data.

M. Abdel-Azim [3] explains the book rating prediction step by step on exploring and preparing the dataset in several ways. E.g.: grouping English languages together, extracting year from date column. It creates multiple new features for example, average_rating * rating_count, average_rating * num_pages, etc and identifies one of the generated features as the highest important feature. He shows that starting from 80s number of reviews is getting higher than before and gives his opinion that this is the effect of computer and internet. ADA boost has been selected as the best model with a training score of 93.6 and a testing score of 95.6.

3. Specification

The dataset comprises of one file, books.csv. The size of the dataset is 1.45MB. This goodreads dataset contains book reviews from goodreads site including 200 million reviews spanning over 45000 books.

Data Set	Size
Books.csv	1.45MB

All the libraries with their respective versions are available in the requirements.txt file in the GitHub.

4. Architecture

The project architecture is illustrated below (Figure 1). The data source being the first phase, the dataset is downloaded from the DSTI LMS, loaded to Jupyter notebook and cleaned appropriately. The second phase was Exploratory data analysis through visualization and making observations. The third phase involved selecting important features and transforming raw data into new features, which is called feature engineering. Finally, for model

building evaluation and tuning involved training and testing the models and fine tuning the parameters. The overall architecture is illustrated below in Figure 1.

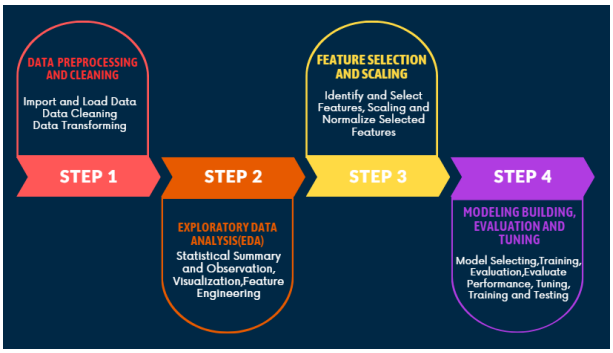

Figure 1: Architecture Diagram

5. Implementation Flowchart

To begin with, the raw dataset is downloaded from LMS. To perform prediction and build algorithms Machine Learning techniques were used. The machine Learning workflow has five steps as depicted in Figure 2.

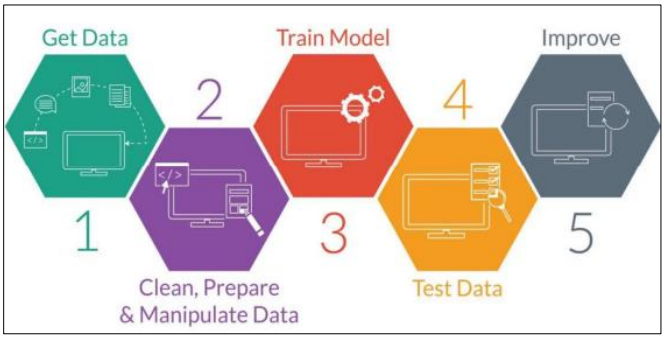

Figure 2: Machine Learning Workflow

The first step is retrieval of the dataset from DSTI LMS project1 file. After that, it has been prepared and manipulated in a way that it gives a more accurate result. The model is trained and tested and had multiple iterations to improve.

The implementation flow chart is shown below in Figure 3. The first and foremost step was to understand the project's objective and do research on similar work. The next stage was Data & Feature Engineering which involved preparing the data for modeling. Data Split is done to get two sets of data: train and test. In the next stage, which is Train, Test and Validate, data was trained and tested while validating the model. Finally, in the Evaluation stage the models were evaluated using the measures for accuracy.

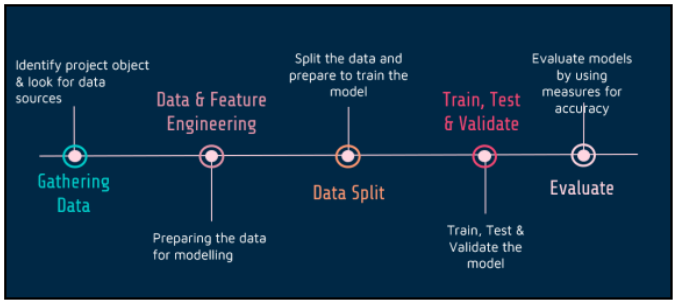

Figure 3: Implementation Flowchart

6. Data Engineering

The dataset's column summary is displayed below (Table 1).

Table 1: Column summary

Column Name	Data Type
title	object
authors	object
average_rating	float64
isbn	object
isbn13	int64
language_code	object
num pages	int64
ratings_count	int64
text_reviews_count	int64
publication_date	object
publisher	object

The following steps were taken to reach the goal during data engineering.

i. Data cleaning:

- Load the dataset while correcting the four rows with misaligned values
- Treat null values
- Remove unnecessary columns; title, isbn, isbn13
- Unpopular languages are grouped together
- Convert object columns having numbers into numerical columns
- Visualize the behavior of the features in different ways
- Study and check the correlations between the features

ii. Feature engineering

- Extracting year from publication_year
- Generate new columns with existing columns

Author average rating:

Calculated the average rating for each publisher by looking at all the books from that publisher.

Author average rating:

Calculated the average rating for each author by looking at all the books from that author.

Number occurrence:

Calculated the number of times each book title appears in the dataset by counting all occurrences of that title.

Author average page:

Calculated the average number of pages for each author by looking at all the books written by that author.

Author book count:

Calculated the total number of books written by each author by counting all the books associated with that author in the dataset.

- Measure the feature importance to check if the newly generated features are useful

iii. Data storing

- Download the previously processed data for the future steps

7. Machine Learning

The objective is to predict the rating scores of Goodreads books by utilizing various numerical features like language_code and ratings_count, along with categorical features such as publisher and author. Initially, Tree Based and Ensembled Model algorithms like Decision Tree Regressor, Random Forest were analyzed, but the accuracy was found to be unsatisfactory. After investigating, we found that the issue has occurred since the model prediction is done using only using the numerical features by dropping the categorical features.

Highest accuracy of 99% with NuSVR could be reached upon the generation of new features. These features used the previously eliminated columns and showed importance greater than 0.05% of the prediction model in Table 2.

Table 2 : Feature Importance greater than 0.05%

Feature	Importance (%)
rate_occurrence	75.15
number_occurrence	14.89
author_average_rating	9.65
rate_per_pages	0.07
rating_weight_	0.06
author_book_count	0.05
publisher_average_rating	0.05
author_average_page	0.05

Multiple machine learning algorithms are used to identify the best model.

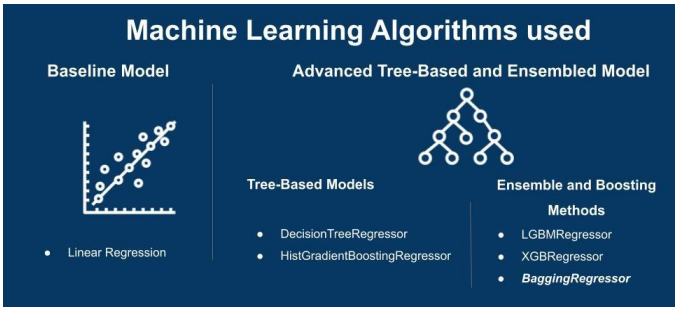

Figure 4: Machine Learning Algorithms and Recommendation Model details

The module, lazyregressor was used to automatically compare several models based on their predictive performance. This allowed us to select the most suitable model for our dataset. After identifying the best model with highest R-squared values, six models are selected and evaluated the performance using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 score and execution time.

Table 3: Evaluation of Baseline Model

Model Name	MAE	MSE	RMSE	R2-score	Time (s)
Linear Regression	0.057	0.0117	0.1082	0.9030	0.02

A comparison of top advanced tree-based algorithms are shown in Table 4.

Table 4: Comparison of Models for Advanced Tree-Based and Ensembled Model

Model Name	MAE	MSE	RMSE	R^2-score	Time (s)
DecisionTreeRegressor	0.0046	0.0009	0.0309	0.9921	0.19
BaggingRegressor	0.0045	0.0004	0.0211	0.9963	1.35

HistGradient BoostingRegressor	0.0084	0.0010	0.0327	0.9911	1.75
LGBM Regressor	0.0070	0.0006	0.0259	0.9944	0.24
XGBRegressor	0.0078	0.0007	0.0270	0.9939	0.12

8. Deployment

After choosing the best model, the trained model is downloaded as a PICKLE file.

Conclusion

We aimed to predict Goodreads book rating prediction by considering features like author, ratings_count, num_pages, and text_review_count, etc. To conclude the step by step process; dataset was loaded, cleaned, processed, measured the correlation between features, analyzed the behavior of the features and processed them accordingly. New features were introduced and downloaded the dataset as processed dataset. It is loaded back for model training, separated X, y and split into two sets: training 70% and testing 30%. Initially, we used only numerical features and tried multiple machine learning algorithms like Random Forest and linear regression leading us to an unacceptable result with a high mean squared error. This result revealed to us the importance of categorical features in our dataset. Several new features were produced using the existing features to improve the model. During this process categorical features like author and publisher were used. Consuming these columns helped a lot to increase the accuracy of the models. As per the evaluation comparison showed in the above section, Bagging regressor gives us the highest R2 squared value of 0.995 with MSE value of 0.000. Its execution time of 0.64 seconds is also satisfactory. It was chosen as the best model suitable for us. All the scripts with input dataset are available on GitHub.

References

[1] "Goodreads," 2024. [Online]. Available: https://www.goodreads.com/. [Accessed 2024].

[2] M. Wan, "Goodreads Book Graph Datasets," May 2023. [Online]. Available: https://mengtingwan.github.io/data/goodreads.html. [Accessed 16 August 2024].

[3] M. Abdel-Azim, "Goodreads Book Ratings Predictions," 01 August 2020. [Online]. Available: https://github.com/MOHED1224/Goodreads-Book-Ratings-Predictions/commits?author=MOHED1224.