Python Machine Learning Labs

Develop an end-to-end Machine Learning Pipeline

Book Rating Prediction Model

**Group 5**

Marius Akre

Vaishnavi Jagtap

Patrick Githendu

Stefanny Peraza

**Instructor**

Hanna Abi Akl

Abstract

#TODO

Keywords: Add keywords here.

Book Rating Prediction Model

The following analysis aims to predict book ratings using different machine learning methods. A model of this type could help predict whether a book would be well or poorly evaluated by the reading community, based on certain characteristics, and in such a way the reader could focus on those that would have better ratings. Although it is not the objective of this analysis, the available data could also be used to make recommendations not only based on the estimated score, but rather based on the characteristics that are most important to the reader. However, it should be noted that the characteristics or variables available for this study are quite limited and correspond only to the books that have been evaluated in the database of Goodreads[[1]](#footnote-2).

## Data

The database contains 12 columns and 11,127 registers. However, of these 12 columns there are four (*bookID, isbn, isbn13, title*), which correspond to book identifiers and are therefore discarded as predictors variables; the *authors* and *publishers* are also not taken as variables on this occasion. Finally, the *average\_rating* variable is the one to be predicted, therefore only 5 possible predictors are available: language\_code, num\_pages, ratings\_count, text\_reviews\_count and publication\_date.

## Data cleaning

During data loading, problems were detected for the correct identification of the columns due to the presence of "," in the names of the authors, and the use of the "," at the same time as separator character of the fields in the file. To solve this problem, the "," in the authors field was replaced by a " -"; this was applied specifically to 4 records that were the only ones showing this problem, thanks to that change it was possible to keep them.

## Data processing

## Exploratory analysis

plots of relevant attributes) and feature selection (feature engineering, feature pruning, choice justification)

## Methodology

Model training (motivation for selected model, comparison of different models) and evaluation (evaluation metric, results interpretation)

## Results

References

Last Name, A. B. (Year). Article Title. Journal Title, Pages #-#. URL. URL.

Tables

Table 1

Table Title

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Head | Column Head | Column Head | Column Head | Column Head |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |
| Row Head | 123 | 123 | 123 | 123 |
| Row Head | 456 | 456 | 456 | 456 |
| Row Head | 789 | 789 | 789 | 789 |

Note: Place all tables for your paper in a tables section, following references and footnotes. Start a new page for each table, include a table number and table title for each, as shown. All explanatory text appears in a table note that follows the table, like this one. Use the **Table/Figure** style, available on the **Home** tab, in the **Styles** gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been set up for this template that fits APA guidelines. To insert a table, on the **Insert** tab, click **Table**.

Figures Title

Figure 1.

Include all figures in their own section, following references, footnotes, and tables. Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.



*For additional information on APA Style formatting, please consult the* [*APA Style Manual, 7th Edition*](https://apastyle.apa.org/style-grammar-guidelines)*.*

1. Goodreads: the world’s largest site for readers and book recommendations, launched in January 2007. [↑](#footnote-ref-2)