|  |
| --- |
| **Machine Learning with Python Labs – A22**  *Date*: March 26th, 2023  *Professor*: Hanna Abi Akl  *Project contributors*:  Diego Carriel Lopez : [diego.carriel@edu.dsti.institute](mailto:diego.carriel@edu.dsti.institute)  Romain Duhomme : [romain.duhomme@edu.dsti.institute](mailto:romain.duhomme@edu.dsti.institute)  Lan Le : [le.lan@edu.dsti.institute](mailto:le.lan@edu.dsti.institute)  Teresa Graffi : [teresa.graffi@edu.dsti.institute](mailto:teresa.graffi@edu.dsti.institute) |

**PROJECT REPORT**

**Contents**

[1. Overview of the targeted problem 1](#_Toc128036578)

[2. Exploratory Data Analysis and Data Preprocessing 2](#_Toc128036579)

[3. Feature engineering and Model selection 6](#_Toc128036580)

[4. Model training and evaluation 7](#_Toc128036581)

[5. Deployment 7](#_Toc128036582)

[6. Results and recommendations 7](#_Toc128036583)

\*\*\*\*\*

1. **Overview of the targeted problem**

The aim of this machine learning project is to build a model that can help Goodreads users get book recommendations based on authors, book types, and other relevant factors.

We believe that the problem at hand has an average complexity and will require a manageable number of features (independent variables) for the machine learning model.

*Questions to consider before starting the project?* (to be completed)

*How we intend to address the problem?* (to be completed)

….

*Methodology*

To begin, we conducted research to gain a general understanding of the Goodreads website and its user behaviors. We found that the rating scale is one of the most indicative factors of a book's popularity and sales. Therefore, we will use book rating score as the primary dependent variable in this predictive model. The quantitative analysis will be divided into the following steps:

* **Exploratory data analysis** to understand the dataset and **data pre-processing** to clean and format the dataset to ensure it is in a suitable format for machine learning algorithms.
* **Feature engineering** by creating new features or extracting relevant features from the data to improve the model's predictive performance, and **Model selection** which will depend on the nature of the data and the required predictive accuracy for predicting book ratings.
* **Model training and evaluation**: ………….(to be completed)
* ………..

1. **Exploratory Data Analysis and Data Preprocessing**

*About the dataset*

To achieve our objectives, we used a dataset containing approximately 11,000 rows of data scraped from the Goodreads API. This dataset includes customers' ratings of books, as well as other information such as author, book code, number of pages, count of text reviews, publication date, and publisher obtained from the Goodreads website[[1]](#footnote-1).

Here, questions regarding the volume and accuracy of the data arise. While we believe that the collected dataset may be sufficient to build an ML model, it is crucial to ensure that the dataset is representative of the target population and contains a diverse range of books and ratings. The first step in determining the sufficiency of a dataset is to experiment with exploratory data analysis and data preprocessing techniques to validate the dataset.

*Observations about data*

In this stage, we will be using Excel and Python to conduct EDA and DP processes.

1. ***Excel***

In Excel, the first step was to analyze the original dataset, which was a CSV file separated by the delimiter *";"*. We noticed that there were some blank spaces before the title *"num\_pages"*, which we removed to ensure the file could be imported correctly in Excel.

The next steps involved correcting the data types. We deleted non-numerical values in the *"rating"* and *"num\_pages"* columns and transformed some numerical and non-text data in the *"language\_code"* and *"publisher"* columns into the appropriate data types.

We noticed that some book *“titles”* appeared multiple times across different rows in the dataset (1325 rows in total). To address this, we decided to keep the maximum value and add the number of votes for each title.

We also eliminated columns that were not relevant to the project (such as *“idbook”*).

We corrected the discrepancies between book titles and their corresponding data in some rows by matching them with the correct column.

To address *“num\_pages”* rows with a value of 0 (76 rows), we either used the data from the corresponding book if it was available. For the remaining rows, we calculated the average number of pages by publisher.

We decided to address the issue of *“author”* rows containing more than 51 authors by selecting the first author, who is typically the principal writer, and separating them from the other authors.

1. ***Python***

In Python, we installed all necessary libraries and imported the dataset using Pandas.

Text

Description automatically generated

Then, we checked the data types using the code *"data.info()"* to get an overview of the dataset's dimensions and description. We identified two types of variables:

* + Categorical variables such as “*title”, “authors”, “average\_rating”, “isbn”, “isbn13”, “language\_code”, “publication\_date”,* and *“publisher”* with object data types, and
  + Numerical variables such as “*bookid”, “num\_pages”, “ratings\_count”,* and *“text\_reviews\_count”* with integer and float data types.

Graphical user interface, table

Description automatically generated with medium confidence

After that, we conducted a descriptive statistics analysis using the code *"data.describe()"*.

Table

Description automatically generated

We installed Matplotlib to create static and interactive visualizations in Python.

Chart

Description automatically generated with medium confidence

Our aim was to verify if there were any relationships between different variables in the dataset. To do this, we computed the correlation matrix using the following code: *corr = df.corr()*

Then, we generated a heatmap from the correlation matrix with the following code:

*Graphical user interface, text

Description automatically generated*

From the heatmap, we observed a strong correlation between *"ratings\_count"* and *"text\_reviews\_count"*.

Completing this step helped us gain a deeper understanding of the dataset and prepared it for the next steps of feature engineering and model selection.

1. **Feature engineering and Model selection**

We decided to create new features as follows:

* For the *“rating count”* column, we transformed this into categories - a little, medium, and a lot - using a logarithmic function. Before that, for some rows where the number of votes is indicated as 0 but the rating value is not 0, we decided to replace the number of votes with 1.
* Another new feature we created was to transform *“num\_pages”* into categories - low, medium, and high - by using a classic splitting method.
* We transformed also the *“rating”* score into categories - bad, medium, and good book - using a logarithmic function. At this stage, we found out that 91% of the data was between 3.5 and 4.4, indicating a normal distribution.
* For *“publication\_date”*, we opted to split the data by year and by season with corresponding numerical indications.
* We also created a column for the *“average\_rating”* vote, calculated through the average by author and ponderate average by the number of votes.

1. **Model training and evaluation**

(to be completed)

1. **Deployment**

(to be completed)

1. **Results and recommendations**

(to be completed)

1. Goodreads is an American social cataloging website and a subsidiary of Amazon that allows individuals to search its database of books, annotations, quotes, and reviews. [↑](#footnote-ref-1)