**PYTHON PROJECT**

**PRESENTATION**

We have been provided with a dataset of books from the online library Goodreads.

The purpose of this project is to build a model that can predict a book’s rating based on the data we have on hands, using the Python programming language.

**STEPS OF THE PROJECT**

1. Data discovery: This step is very useful to understand what the type of data is we have, the type (numerical of categorical) and if there are missing and / or erroneous values.

Here, we will make some basic data exploration with statistical functions have a basic yet useful overview of our data. At the end of this step, we should have a cleaned and complete dataset.

1. Data analysis/Feature engineering: We will go a little bit deeper in the analysis phase and try to find correlation between different columns (features). From this step we should end up selecting the features that will be used in our prediction model.
2. Building and testing the model: Now that we have the workable dataset with only the features we have Identified as the one influencing a books rating, we can now build and create the model to use it and get the results we are looking for.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. **DATA DISCOVERY**

Here are some basics information we want to first get from our dataset:

* What is the size? (ie. How many rows/ columns)
* What are the column names and their type?
* Have a look at the data we actually have inside the columns
* How many missing values in each column?
* How is the data distributed in each column?
* Could the table be enriched with more additional data to make the model more precise?

The first issue we run into is when we try to import the data set and we are faced with a parsing error. Looking closely into the dataset, we realize that there is a column containing data (4 lines) but it has no header.

Filtering the data in Excel, we can see that it’s because, some authors name have “spilled over” to the next available column, causing the whole data to be spilled on the right.

Solution: we chose to correct this manually inside the file, since it’s only 4 lines. It would have been a different problem with thousands of shifted lines.

**A short summary of the structure of our data**

11 127 rows x 11 columns

|  |  |
| --- | --- |
| Categorical | Numerical |
| Title | Average\_rating |
| Authors | Isbn\_13 |
| isbn | Num\_pages |
| Language\_code | Ratings\_count |
| Publication\_date | Text\_review\_count |
| Publisher |  |

The ISBN is a book identifier. Although the ISBN 13 is stored as a numerical instead of a text data, we do not see it as critical information to predict the rating of a book, so we will probably not keep it to build our model.

However, looking closely into our dataset with the *df.head()* function we can make a few observations:

Categorical data

* The authors column contains more than one name.
* The language code can have more than one label to say the book is in English: eng or eng-US
* ISBN 13 is stored as a numerical and not a text

Numerical data

* Average rating goes from 0 to 5, so we can assume all ratings are over 5 (ex. 3.9/5 is the average book rating of the whole dataset)
* Number of pages go from 0 to 6576. We should identify which book(s) have zero pages and decide whether it can be considered as a book, since there are no physical pages or if this record is an error.
* Some books have no ratings, that must be related to the book(s) having a rating of zero. Is it an absence of rating or a real rating set at zero by the customer?
* The review counts go from no review at all to 94265 reviews for one book.

1. **DATA ANALYSIS**

Categorical data

Authors: let’s see how many authors we have

<https://thats-it-code.com/pandas/how-to-convert-multivalue-column-to-multiple-rows/>

We used this technique of explode to separate each author on his individual line, the problem is that it will modify the rating for each book

Also, looking at this snippet, there seems to be other issues appearing

A picture containing calendar

Description automatically generated

* The same title has different books ID but the same ISBN 13
* Their publishing dates are all different, but are the same for each publisher, except when there is a second (later on) publishing date. We are looking for the book rating