BOOK RATING PREDICTION MODEL

by Luis Monteiro

Using the dataset from Goodread we are asked to train a model that predicts a book's rating. First step with a data analysis,

and prediction.

DATA ANALYSIS

First off all, i changed the original file, because there was 4 line off data that wasn't separeted by "," (lignes 3350, 4704, 5879, 8981).

Had to fix it otherwise an error occurred.

After that i took a look to see if any cleaning was necessary (Overview of data).

I found some rows with onlu the author name and the rest was empty. I deleted them.

Data Cleaning:

- num\_pages": correct the column name, cause there was 2 spaces before num\_pages (" num\_pages"), and transform the

data float to int64 , cause there is no float pages.

- isbn13: transform to a string

- publisher: raname the column name, cause there was ";;;" at the end, and at every data in this column.

- language\_code: transforme en-US, en-GB, en-CA, enm into "eng"

- publication\_data: transform into DateTime and add year separeted

- rationg\_count: transform into int64 and delete the 0 values

- text\_reviews\_count: transforme into int64, cause there is no half counts

DATA VISUALIZATION

3 Questions:

- How much books per language ?

- What's the average rating per year ?

- Top 3 highest rated authors who have published at least 35 books ?

I saw that the books were distributed by language, with the domination of English with more than 10k.

And the closest after that is the Spain and France with 212 and 139.

The top 3 of published books by author are P.G. Wodehouse, Stephen King and Rumiko Takahashi with 40, 39 and 38 published books,

with an average rating 4.14, 3.97 and 4.19.

CLASSIFICATION

3 categores: "Less good", "Good" and "Very good" with the domination of 56.63% Good, 42.70% Very good and only 0.67% Less good,

representative in the dataset. Less good books are <= 3, Good books are >3 and <=4 and Very good > 4.

In Features Engineering, i dropped some unusable columns to predict the group, cause they was unusable and can't be transformed

into categories.

Bellow, we can see the classification model, and see that the recall is much bigger in Good that Very good.

Table

Description automatically generated

The f1-score, is the harmonic mean of the recall and the precision, it is 0.72 and 0.32. 0.72 means that 7 times per 10,

the model will predict correctly the group good, but only 0.32 for very good. 0.72 is almost enough to be confident.

Model 85% / 15%