Report Project: Predicting a Book

The project consists in predicting the rating of a book through a set of data: *books.csv*. In this file we can find several characteristics of a book:

* **bookId**: The unique identifier of the book
* **title**: The title of the book
* **authors**: The author of the book
* **average\_rating**: The average grade for the book
* **isbn**: Unique number of the book
* **isbn13**: Unique number for the book
* **language\_code**: The language of the book in which it is written
* **num\_pages**: The number of pages of the book
* **ratings\_count**: The number of votes the book had
* **text\_reviews\_count**: The number of comments the book received
* **publication\_date**: The publication date of the book
* **publisher**: The publisher of the book

To meet our need, we must go through different steps:

* Data analysis: data processing, data cleaning, exploratory analysis, and relevant graphics.
* "Feature selection": Engineering of characteristics, pruning of characteristics and justification of choices.
* "Model Training": Model selection and justification and comparison with other models.
* "Model evaluation": Measurement and interpretation of results.

# Data Analysis

To have a more concrete analysis process, after importing the csv file, we need to clean the data to make our data set more "clean" and closer to reality. This process allows you to modify or delete incorrect, incomplete, irrelevant, corrupted, duplicated or formatted data.

In fact, our data set had some inconformities. We analysed several characteristics:

* **NULL or blank values**: This dataset had no null or empty values.
* **Useless columns**: we have deleted the ‘isbn’, ‘isbn13’ and ‘publisher’ columns. Indeed, these columns will be useless to our analysis.
* **Doubled values**: Indeed, in a large dataset like this we are confronted with duplicates, for this we must delete them. For our part, we chose to delete them based on title, language and authors.
* **Inconsistent data**: When we looked more closely, we noticed that books were 0 pages, so we also have to delete them.

# Classification model

The classification process searches for a function that helps divide the dataset into classes based on different parameters.

In our case, we cannot have a classification process because the value we want to predict is not a class but a continuous value.

If our data set included the genres of books, we could have performed a clustering model to determine the genre of a book.

# Regression model

The regression process involves finding correlations between dependent and independent variables. It helps to predict continuous variables as in our case with the rating of a book.

We had to analyze the correlation between the ‘average\_rating’ column and the other columns. The result was:

* ratings\_count and text\_reviews\_count: The number of notes depends on the number of comments. A reader will, most of the time, write down the book AND write a comment. We may think that adding a comment is mandatory to put a note.
* num\_pages et average\_rating : The average rating depends on the number of pages.

In order to correlate all our columns, we used the standardisation process. In fact, by converting the data into “standard” format, we tried to find new correlations between different columns. The result has not allowed us to discover any other addiction.

# Training Model and Evaluation Model

We had to choose the predictors from the columns in the dataset. Based on the previous analysis, we only kept the columns where the ‘average\_ratings’ is correlated:

* 'num\_pages'
* ‘text\_review\_counts’
* ‘ratings\_count’

To start, we have divided the dataset into two parts:

* Train is used to train the model.
* Test is used to evaluate the model.

We chose the Linear Regression and Random Forest models to perform the training and evaluation of the dataset. These regression models are known to perform well on unclassified data.

In model evaluations, we used several indicators:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean Squared Error | Mean Absolute Error | Max Error |
| Linear Regression | 2.64e-31 | 2.96e-16 | 3.12e-15 |
| Random Forest | 0.001 | 0.002 | 0.61 |

These indicators allow us to assess the efficiency and accuracy of a model predicting a value. Here, we compared the indicators of each model and we observe that the Linear Regression model will be more efficient because it gives results closer to the real than the Random Forest.

# Exploratory phases

During this project, we have several exploratory phases to test different paths in our analysis. The purpose of these phases is to try to find correlations within our data set.

## Number of votes by number of pages

In this section, we visualize the correlation between the number of pages of a book and its average.

We used the regression chart where we noticed an upward trend in the rating relative to the number of pages. The reader tends to put higher notes on books with more pages.

We also created a scatter and a boxplot of the page number distributions to see the offset points (outliers). This allowed us to see that values above 1000 pages had to be deleted for the dataset to be more accurate.

## Data analysis with the ‘language\_code’ column

We also analyzed our data with the language of books to see if it impacted our analysis. For this, we have shown, through the use of graphs, the language that has been rated the best and the least well.

Through this analysis, we have concluded that a reader will be more likely to rate a book in English as well as score between 3.5 and 4.5 out of 5.

# Conclusion

We can conclude that the regression model appropriate to our need and our dataset is the linear regression model. This one is more reliable than the Random Forest model. There is more chance to predict a result close to the real, it will generally make fewer mistakes. According to our analysis, a reader will rate a book between 3.5 and 4.5.