Summary

- INTRODUCTION
- DATA ANALYSIS
- FEATURES SELECTION
- MODEL TRAINING

Introdiction

According to Ernest Hemindway, "there is no friend as faithful as a book". Today millions of people are venturing into the world of reading. Nowadays with the existence of so many books, it becomes difficult for readers to choose the best ones. Thus, by using Books data from a selection of Goodread books, we will analyze its data and predict the average rating of a book varying betsuite ween 0 and 5 in order to help readers choose the best-rated books.
Dans la

Data analyses | comment

In this dataset we have 11127 books to analyze (Fig 1). these books, whose languages vary (Fig2), have been written by different authors and published by different publishing houses. After publication, each book receives reviews and graded by the average of the scores obtained.

We can see that the average score is 3.96 (fig3). We can thus have an a priori on our prediction, that is to say expect that half of our predictions have a good average score. Moreover, as can be seen, there are years when the ratings are the lowest as in the 1930s (fig 4) or high as in the 1920s. eras is respectively badly and well noted.

To better see this, we will focus in the following on the one hand on the correlation of the variables and on the other to do the analysis of variance (anova) to observe the most influential variables.

DA : PLOT

average rate

0,00 5,00

authors

Tout

Year range

0 5

Distribution of books

Fig2: Number of language

27

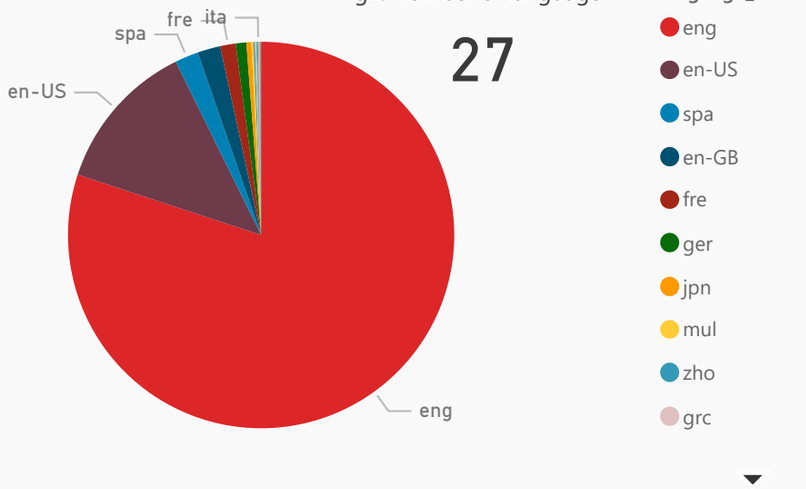


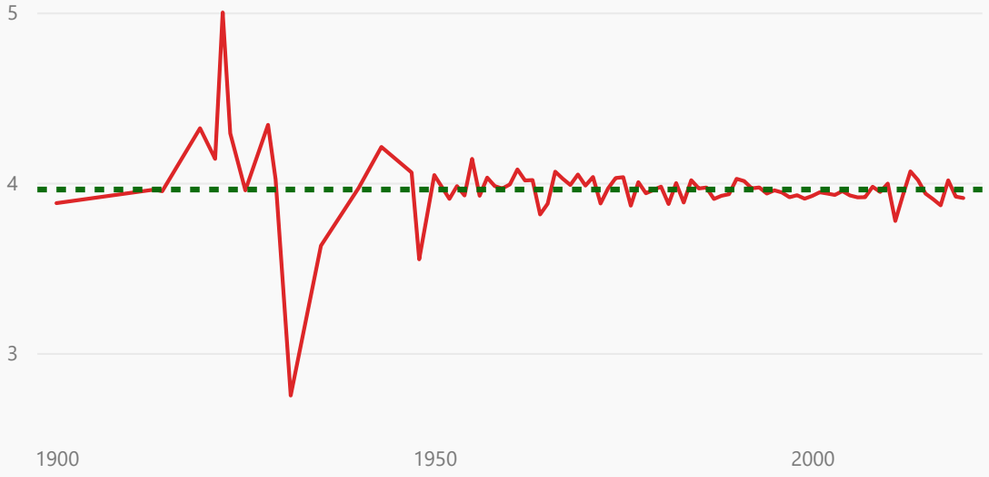
Fig1: Number of books

11127

Fig3: median of average_rating

3,96

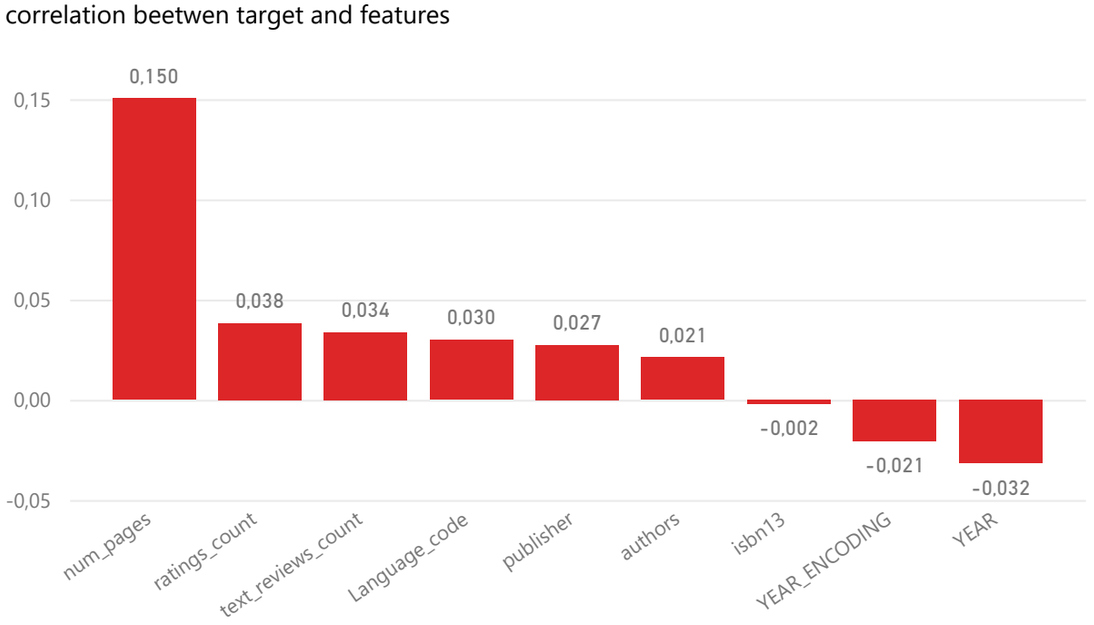
Fig4 : Average of rating_average by year

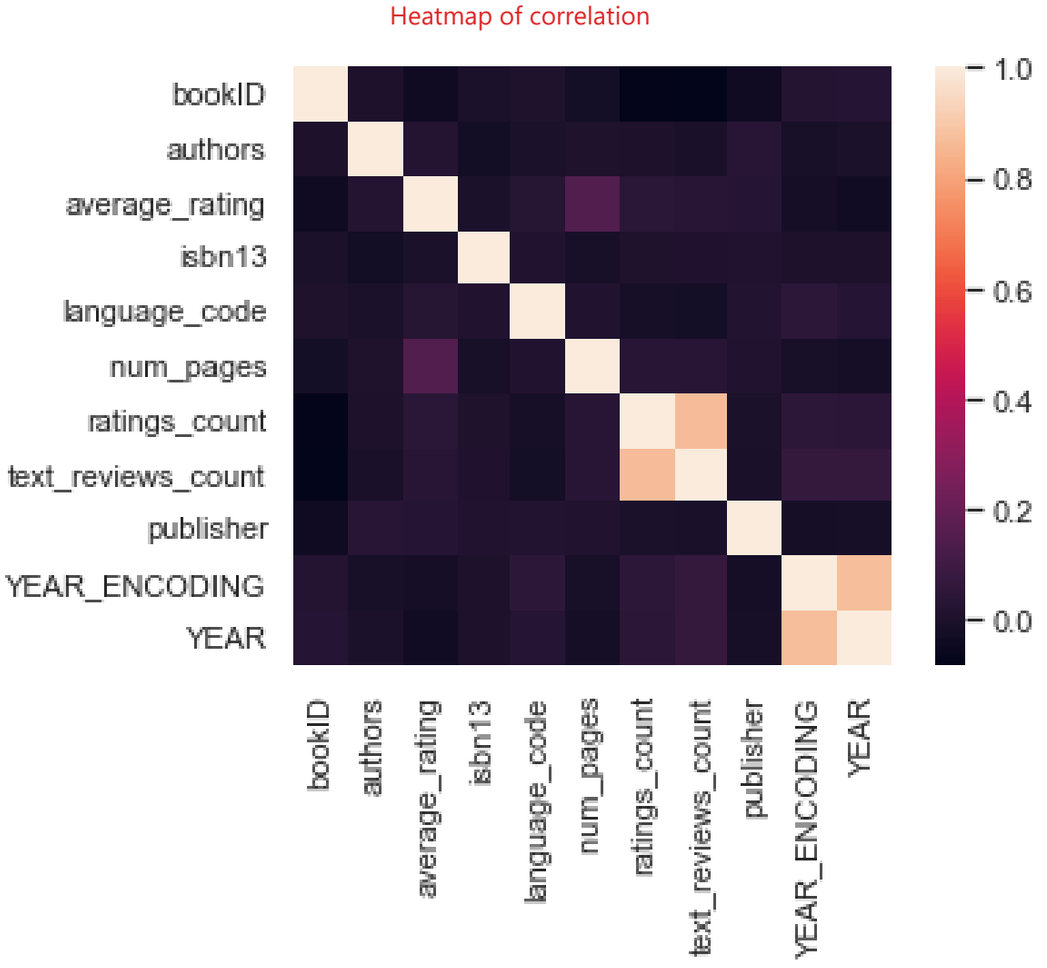


authors	Nombre de authors
P.G. Wodehouse	40
Stephen King	40
Rumiko Takahashi	39
Orson Scott Card	35
Agatha Christie	33
Piers Anthony	30
Mercedes Lackey	29
Sandra Brown	29
Dick Francis	28
James Patterson	23
Laurell K. Hamilton	23
Margaret Weis/Tracy Hickman	23
Terry Pratchett	23
Gordon Korman	22
Alan Dean Foster	21
Bill Bryson	21
Dan Simmons	21
Janet Evanovich	21
Total	**11123**

Feature selection | Correlation

In this section we seek to find the most correlated variables with our target variable (average rating). The results presented below are the outputs of the python code (importance_variable.py). We can see that the variables are weakly correlated with our target variable. We can see all the same that there are very weakly correlated variables such as for example the variable isbn1" compared to the other variables.



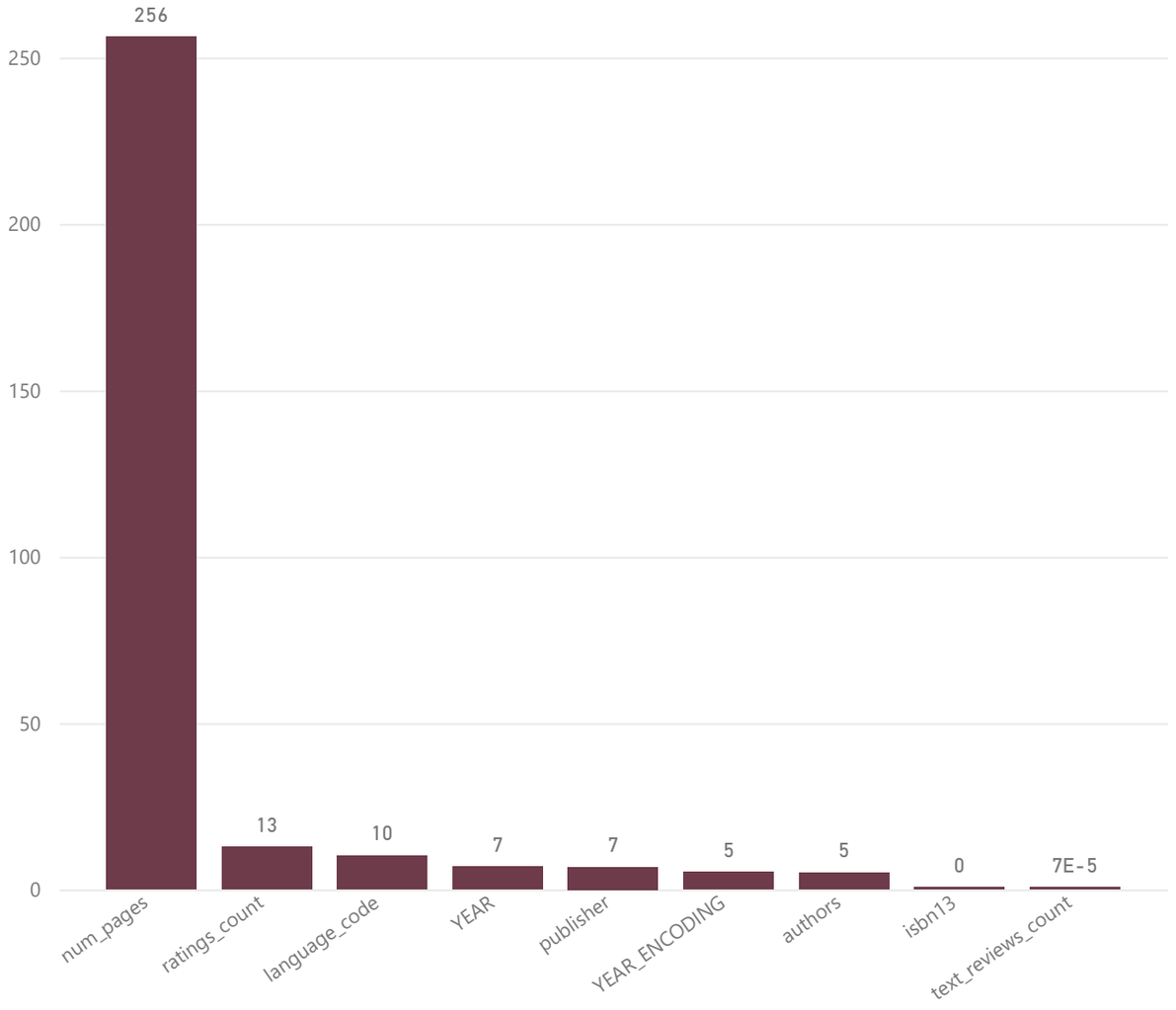


Feature selection | ANOVA

As can be seen in the table below the hypothesis "the variable ibn13 influences the average score is rejected" (rejection level 88%), which is in line with the correlation study presented in the previous page According to our anova study, the other variables have almost the same level of influence on the target variable except for the number of pages variable which is more influential as can be seen in the figure opposite.

ANOVA RESULT

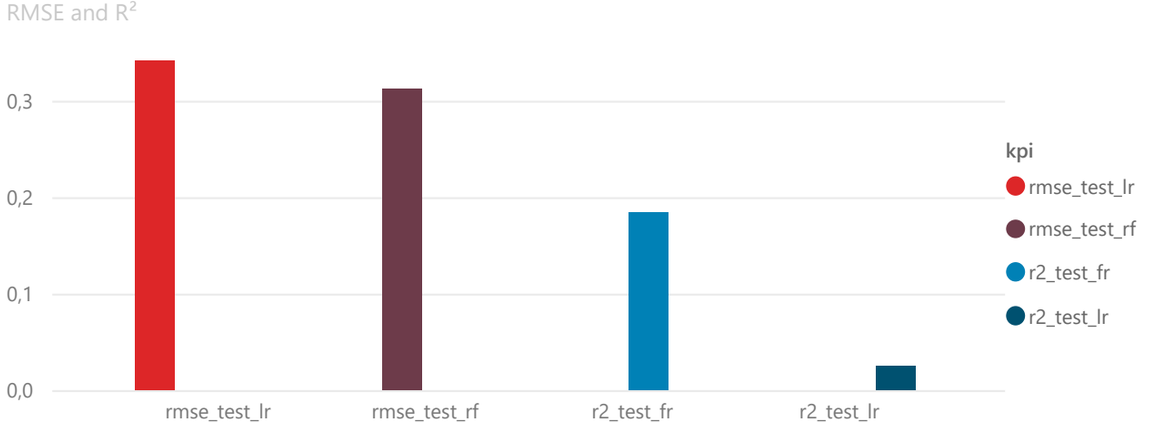
Column1	df	mean_sq	sum_sq	F	PR(>F)
authors	1	0,61	0,61	5,06	0,02
isbn13	1	0,00	0,00	0,02	0,88
language_code	1	1,23	1,23	10,24	0,00
num_pages	1	30,74	30,74	256,28	0,00
publisher	1	0,81	0,81	6,76	0,01
ratings_count	1	1,53	1,53	12,79	0,00
Residual	11114	0,12	1 332,92		
text_reviews_count	1	0,00	0,00	0,00	0,99
YEAR	1	0,82	0,82	6,83	0,01



Modele training | result & comment

The results obtained (modele_training.py) are presented opposite.

The first model formed is the linear regression model (LR) and the second is the random forest (RF). We have chosen the most correlated variables which are also the most influential as listed here (*Language code , authors, publisher, text_reviews_count, num_pages, average_rating, YEAR, text_reviews_count*). Looking at the results in the prediction table, we can see that the predicted scores are close to the actual scores. However, looking at the RMSE and R^2, we can conclude that even if the models are quite robust, the random forest model better predicts book scores. Note also that the models poorly predict very high (5) or very low (2) scores. This is due to the fact that these scores are not very representative in the datasets.
A possible solution would be, for example, to generate new data for a better distribution of scores or to delete these lines in the training data by considering them as outliers.



140 160

Prediction

Column1	average rating	average rating predicted LR	average rating predicted RF
140	3,59	3,93	3,90
141	3,64	3,93	3,87
142	3,94	3,97	4,07
143	4,03	3,97	4,25
144	3,93	3,92	3,81
145	3,97	3,97	3,99
146	3,64	3,90	3,85
147	4,01	3,98	4,21
148	4,18	4,05	4,10
149	3,92	3,93	3,88
150	3,80	3,89	3,90
151	3,42	3,91	3,82
152	4,03	3,93	3,93
153	3,96	3,94	3,89
154	3,82	3,92	3,95
155	3,97	3,90	3,88
156	3,46	3,92	3,66
157	4,50	4,08	3,76
158	3,50	3,98	4,03
159	3,95	3,94	3,86
160	4,44	3,90	4,29