Goodreads Book Ratings Predictions

KOUAMEN NKOUIDJA GILLES CEDRIC

29/03/2023

Contents

1 Introduction

We all know that before we start reading a book, we like to hear from others first. This is why we are always encouraged to consult the ratings of former readers or seek the advice of some friends who have already read the book in order to have an opinion on the quality of the work. The goal of our duty is therefore to design a prediction model that will help us to have an opinion on the rating of a book without having to refer to someone. In order to develop our model, we will use the data set provided by Goodreads, the latter based on real user information.

2 Data set information

See figure Data set information

3 Data set Description

This data set concerns differents book ratings, there are real information gatter from different person. It is a regression problem, where based on the given set of attributes each book is given an average rating. There are a total of 11125 instances and 12 columns as attributes in which one will be considered as target attribute.

See figure Data set description

Missing attributes values

29 rows had all missing values in each of the 12 columns

Abnormal row

2 rows had more thah 12 columns, i manually remove them

Abnormal information

I remove row with average_rating equal to zero because its give no prior information I also decided to remove row with ratings_counts equal to zero but with average_rating different form zero because it has no sense and they were not enough .

The total amount of row deleted is 130 which represent 1 per cent of the entire data set

4 Attribute description

- bookID: A unique identification number for each book.
- title: The name under which the book was published.
- authors: The names of the authors of the book. Multiple authors are delimited by "/".
- average_rating: The average rating of the book received in total.
- isbn: Another unique number to identify the book, known as the International Standard Book Number.
- isbn13: A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.
- language_code: Indicates the primary language of the book. For instance, "eng" is standard for English.
- num_pages: The number of pages the book contains.
- ratings_count: The total number of ratings the book received.
- text_reviews_count: The total number of written text reviews the book received.
- publication_date: The date the book was published.
- publisher: The name of the book publisher.

5 Method

5.1 Feature selection

I decided to remove columns "Title", "authors", "bookID", "isbn", "isbn13", "publisher". I remove bookID, isbn and isbn13 because they are unique idenfier of each book and keeping them will force our model to learn information from each specific book, and trying to predict the rating of an unknown book will fail . I remove authors and publisher because i wanted my model to be able to predict the rating of a book with authors and publisher different than the one we had for our model training.

5.2 Feature engineering

-I grouped my data set in two kind of language_code < English and other > because most of the books where written in English as i considered American English, UK English, Canadian English, all of them as English because there is no prior difference, and the other books as because for each different language we did not have enough representation .

See figure language code

-I also decided to group my data set in two group based on the year of publication. I choose the median of the column publication_date in other to divided in two.

5.3 Data splitting

I used 80-20% train-test split data

5.4 Machine learning Techinques:

- Linear regression

- Random Forest Regressor
- Decision Tree regressor
- Random Forest Regressor with Gread Search

See figure report

6 Conclusion

The best model is the model obtained with Random Forest Regressor Gread Search

	Model	Mean Squared Error (MSE)
0	Linear Regression	0.091908
1	Random Forest	0.101020
2	Decision tree regressor	0.158888
3	Random Forest Gread Search	0.084359

Figure 1: Report.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11123 entries, 0 to 11122
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   bookID              11123 non-null  object
 1   title               11094 non-null  object
 2   authors             11094 non-null  object
 3   average_rating      11094 non-null  float64
 4   isbn                11094 non-null  object
 5   isbn13              11094 non-null  float64
 6   language_code       11094 non-null  object
 7     num_pages         11094 non-null  float64
 8   ratings_count       11094 non-null  float64
 9   text_reviews_count  11094 non-null  float64
 10  publication_date    11094 non-null  object
 11  publisher;;;        11094 non-null  object
dtypes: float64(5), object(7)
memory usage: 1.0+ MB
```

Figure 2: Data set information.

	average_rating	isbn13	num_pages	ratings_count	text_reviews_count
count	11094.000000	1.109400e+04	11094.000000	1.109400e+04	11094.000000
mean	3.935026	9.759826e+12	336.543537	1.798750e+04	543.304309
std	0.346458	4.435532e+11	241.313733	1.126427e+05	2579.856004
min	0.000000	8.987060e+09	0.000000	0.000000e+00	0.000000
25%	3.770000	9.780345e+12	192.000000	1.050000e+02	9.000000
50%	3.960000	9.780582e+12	299.000000	7.490000e+02	47.000000
75%	4.140000	9.780872e+12	416.000000	5.018750e+03	238.000000
max	5.000000	9.790008e+12	6576.000000	4.597666e+06	94265.000000

Figure 3: Data set description.

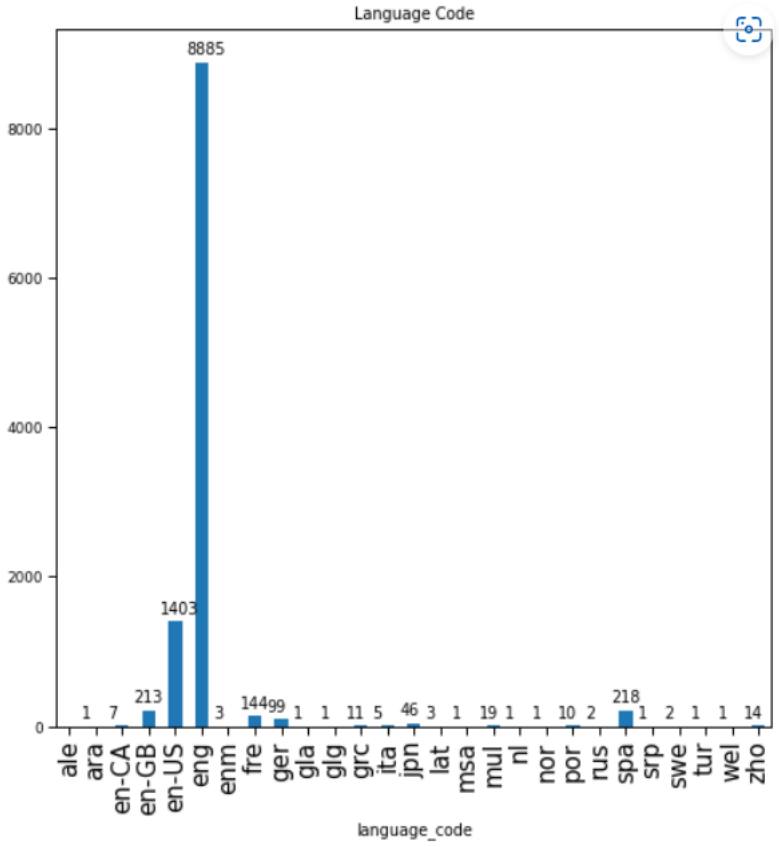

Figure 4: Language code.