What factors lead the interaction between books and readers?

Andra Condurache
Departament of Computer Science
City, University of London

Abstract

This study is based on a dataset containing information about books regarding authors, ratings, reviews and genres. The main focus of interest is the relation between reviews and ratings. The study includes an analysis of the others features. The investigation has been conducted by using visual analytics and regression models.

Introduction

Books have been one of the greatest source of information and reading has played an important role in the process of education. Over the years, genres, topics and even writing styles have been changing and adapting to our needs. Nowadays, more and more books are being written and this may create just another form of consumption.

Every book provider has their own charts promoting subjectively specific authors, genres or subjects. Online providers, like Amazon, have best seller lists updating very frequently, suggesting specific books that may influence the charts. As our life has been overwhelmed by a considerable amount of genres and topics, reading a review may be sometimes the decisive factor which will determine choosing a book. Popularity is another causative factor when it comes to choosing a book. Like in any other domains, the desire to be on trend is also present.

The aim of this investigation is to offer an overview of the factors that influence writing reviews for books and what makes a book more appealing to a large public. This study will focus on outlining the features that a book awakes the desire of people to engage in interaction with the book itself and to emphasise with the characters.

1.Data, research questions and plan

1.1 Data

A dataset from Kaggle was used for this investigation. It contains information regarding authors, number of pages, ratings, numbers of ratings, numbers of reviews, genres and more. The dataset contains 68412 rows and 12 columns. Since the source contains errors or missing values, some of the values will be removed or replaced with suitable values.

As the dataset contains information regarding the cover of the books which is irrelevant to this investigation, the 'image_url' column was removed. The main purpose of this study is to investigate the values related to ratings, reviews and genres. The main focus of this investigation will be numerical values and correlations between them. Some of the columns will be transformed for the analysis. Few values may be biased as in any other artistic domain.

1.2 Research questions

The research questions are:

- What genres are most likely to have higher ratings and appeal more to readers? Does genre influence writing a review?
- Do people write a review if they find a book particularly bad or particularly good? Does the rating influence the number of reviews?
- What are the most important features in predicting the way a book is received?

The results aim is to offer a better understanding of what factors are usually likely to contribute to higher ratings, and implicit popularity. Furthermore, one interesting aspect is studying the relation between number of reviews and readers' feedback.

1.3 Analysis plan

The main objective is to predict a book popularity by analysing the behaviour of book ratings and number book reviews. The results are presented using suggestive histograms and models.

1. Import the .csv file
2. Clean the data: remove the outliners, handle the missing value, reshape columns.
3. Investigate various scatter plots and histograms related to rating and reviews in order to extract relevant characteristics.
4. Explain the initial analysis.
5. Develop a model.
6. Compare results and discuss factors.

2.Findings and reflections

2.1 Overview of genres

The first step is checking how the values are distributed. Some of the book ratings values were considered outliners and they were removed. The interval that reflect representative values is considered for future analysis. As the histogram shows, book ratings follow a normal distribution.

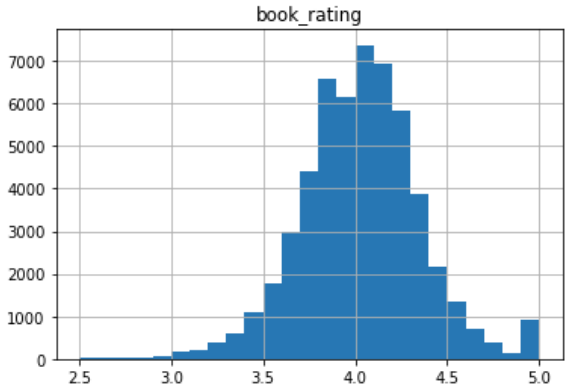

Figure 1 - Book rating

There are 867 unique genres identified in total. In order to reflect the most popular choices, the most frequent 20 genres were chosen for future analysis. These are 'Fiction', 'Fantasy', 'Romance', 'Young Adult', 'Historical', 'Paranormal', 'Mystery', 'Nonfiction', 'Science Fiction', 'Historical Fiction', 'Classics', 'Contemporary', 'Childrens', 'Cultural', 'Literature', 'Sequential Art', 'Thriller', 'European Literature', 'Religion', 'History' .

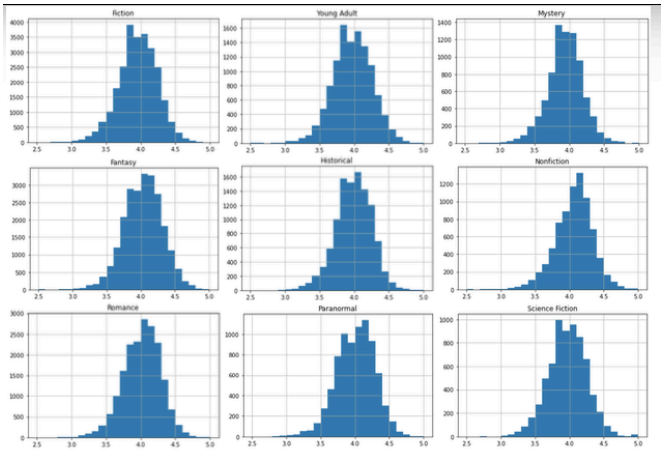

Figure 2 - Histograms of genres

There are similarities between genres and all of them follow a normal distribution. Figure 2 shows some of genres have higher ratings. Looking at the first histograms we can see 'Fantasy', 'Romance', 'Paranormal' and 'Nonfiction' genres have relatively high ratings, comparing them to the rest of the most popular genres. 'Sequential Art', 'Religion', 'History' genres have even higher ratings. This could suggest splitting genres with ratings above average values into 2 categories. One category is constituted of the most

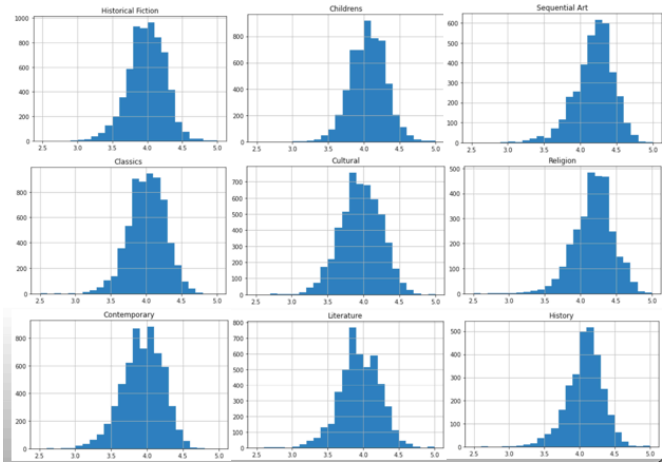

Figure 3 - Histograms of genres

appealing books that are easy to read and suitable for any reader regardless their background. The second category is constituted of genres suitable for specific readers with a background in a niche area. The high ratings of the second category could suggest these are more valued and well received by readers or could be a consequence of biased opinions.

2.2 Scatter plot analysis

For this investigation one feature was introduced to the current data. Variables 'book_review_count' and 'book rating_count' were used to generate an additional feature correlated to a book. Rows containg missing values or zeros were removed, for both of the columns, 'book_review_count' and 'book rating_count'. The feedback percentage was calculated to expressed the proportion of reviews.

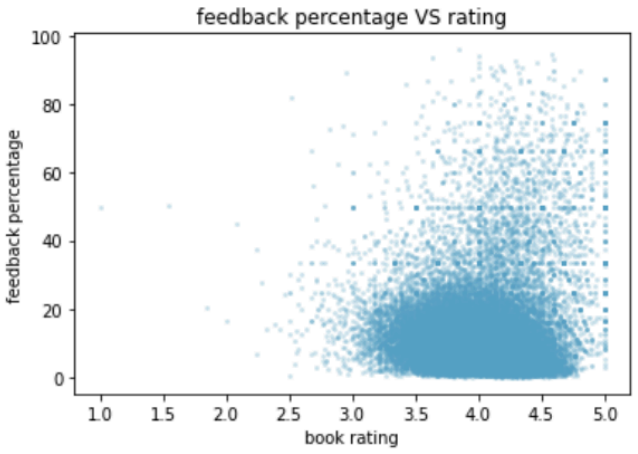

Figure 4 - Scatter plot of Feedback percentage versus rating

As the aim of this investigation is to determine the reason of writing a review, we will analyse the high percentages.

Feedback percentage	rating less than 4	rating greater than 4	rating greater than 4.5	Total books
>90%	2	12	2	14
>80%	13	52	20	65
>75%	19	94	37	113
>50%	146	623	274	769

Tabel 1 - Distribution of ratings for high feedback percentages

The greatest density shows that the most of the books have up to 30% reviews out of ratings and the they are rated between 3.0 and 4.5. This is an obvious behaviour. The previous table shows that the books with a considerable amount of reviews tend to have high ratings. This behaviour is observed for every interval set for previous analysis. It was concluded that people are mostly like to write a review when they find a book particularly good, usually rated above average.

2.3 Correlation matrix

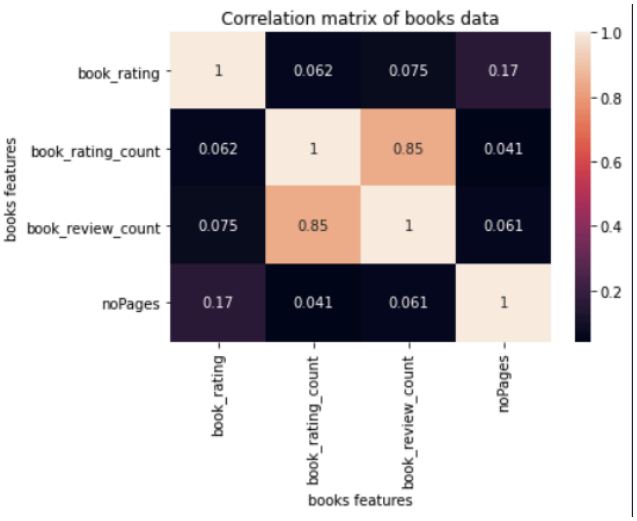

Figure 5 - Correlation matrix

The correlation matrix is used to select the features that are going to be used. The previous figure shows a Pearson's correlation indicating the association between features. This uses only numerical values, as the categorical values and numerical values are treated differently. The 'book_pages' column was reshaped to be included in the matrix. For this investigation we will use 'book_rating_count' and 'book_review_count' features because this study is oriented towards interaction between readers and books. 'book_rating_count' and 'book_review_count' have a positive correlation that means both variables change in the same direction. There is a low correlation between the others features.

2.4 Regression Analysis

The previous figure shows the rating count and the review count. The regression was used to predict the number of reviews using the number of ratings.

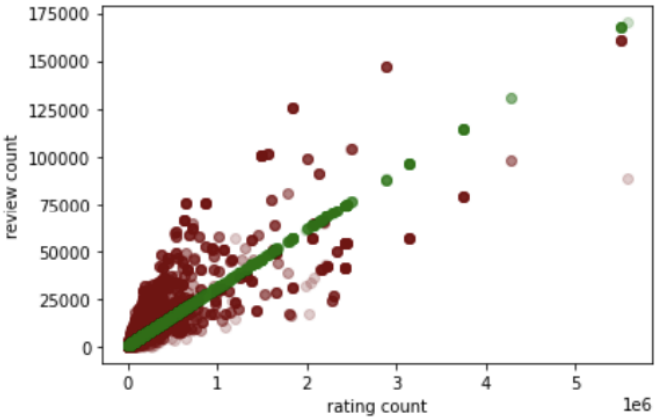

Figure 5 - Linear regression

Chracteristic	Value
Intercept	711.68786526
slope	0.03031554
R2	0.714997732905348
Standard deviation	4123.11
Mean review count	2066.59
Median review count	202.00

Table 2 - Values of the model

The slope shows 1 rating increases the number of reviews by only 0.03. The R-squared value is the coefficient of determination. For this investigation, the R-squared is 71% and it is relatively high, indicating that the model fits the data.

Figure 6 shows that the residuals are normally distributed with the mean of zero which is a good result illustrating no violations of the assumptions.

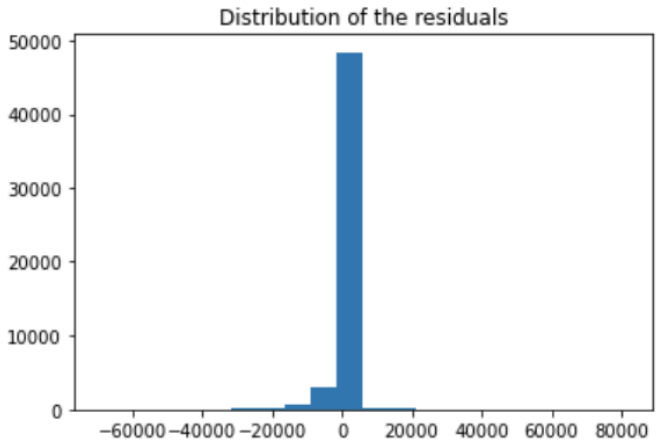

Figure 7 - Distribution of the residuals

2.5 Conclusion

All genres have similar values in terms of ratings and number of reviews. Few of them have ratings above averege that can conclude that we can split high rated books into two categories: books appealing to every kind of reader who can enjoy, mainly fantasy and romance books, and books designed for a niche type of readers who find books more valuable.

People tend to write reviews for books that are particularly good. There are considerable more books with an extremely high rating receiving reviews in comparison with particularly bad books. To conclude, readers express opinions and emphasise with high valued books.

References

[1] https://www.kaggle.com/arturgor/don-t-judge-a-book-by-its-cover-cnn/data?select=book_data.csv
[2] https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
[3] https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
[4] IN3061/INM430 Principles of Data Science (PRD1 A 2020/21) - Lab Feedback 05

Section	Number of words
Abstract	50
Introduction	204
Analytical questions and data	228
Analysis	470
Findings, reflection and further work	297