Never judge a book by its cover?

DANIEL KELEMEN, Eötvös Lóránd University, Hungary

In this research, we take a look at a large collection of books' data acquired form Kaggle [1], called "Goodreads-books". We also build a model capable of predicting the rating the book received, based on the multiple features provided and the TF-IDF vectorization of the title. We find that it is possible to make a model with a R^2 value of around 2.1 that also mostly uses the title vectors as a base-point of predictions, with other attributes being less important in the contribution of making the prediction.

Additional Key Words and Phrases: Book, Machine learning, Predictions, Book title, Title

ACM Reference Format:
Daniel Kelemen. 2025. Never judge a book by its cover?. *J. ACM* 1, 1, Article 1 (November 2025), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The first thing we notice when coming across a new book is the title along with the cover, and so people tend to make first judgments based on them. Since the cover is rather difficult to analyze in this way and also usually contains the title, we stick to the analysis of only the title, focusing on how it affects a book's ratings.

We first go into detail on the dataset used to conduct this research, and the culling and transformations used on the features. We then proceed with explaining the model used, and the results obtained.

2 The dataset

As noted before, the dataset used is called "Goodreads-books", and was obtained from the website Kaggle[1]. The dataset contains the following features, which are relevant for this research:

- title
- author
- average rating
- language of the book
- page count
- rating count
- text review count

2.1 Cleaning

The dataset was cleaned for the sake of finding only relevant data. Originally, before any action done, the dataset contained 11127 books.

The first section included a filtering of books that don't have titles or ratings, or their values are not valid. Then the non-english books were filtered, followed by the ones whose title was

Author's Contact Information: Daniel Kelemen, ex9blv@inf.elte.hu, Eötvös Lóránd University, Budapest, Hungary.

ACM 1557-735X/2025/11-ART1
https://doi.org/10.1145/nnnnnnn.nnnnnnn

considered too long or too short. The final filter applied was for the number of ratings given, since ones with too few could have skewed or unreliable data. This brought the available book count down to 8198.

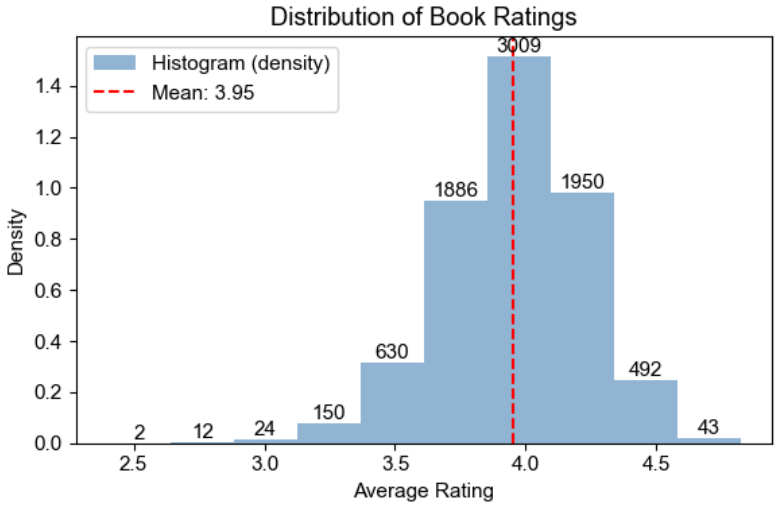

Fig. 1. Average ratings distribution of books

2.2 Feature engineering

The transformation of the title is done using TF-IDF vectorization [3] to get 5000 unique values with which our model can work.

3 Model

The model was created using the random tree regression method [2], which was trained on a randomly picked 80% of the total data, with the predictions obtained for the remaining 20%.

As shown, the model uses the title attributes vectorized features to obtain about 50% of the prediction.

3.1 Performance evaluation

The model's performance is acceptable, albeit leaves much to be desired. The primary aim of using mostly the title to predict the score is fulfilled, and the obtained recommendations are shown to be certainly better than using the baseline (mean) performance.

Table 1. Model's performance compared to baseline

	MSE	RMSE	R^2
Model	0.054	0.233	0.211
Baseline	0.068	0.262	0

The results show a 21% improvement in MSE score and an 11% improvement in RMSE score.

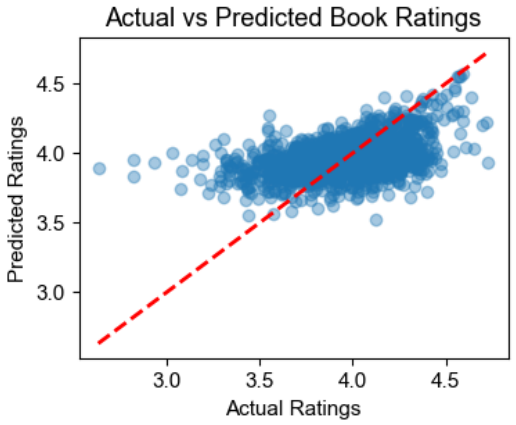

Fig. 2. Model's rating performance

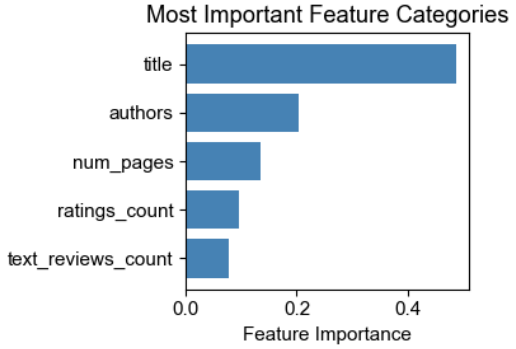

Fig. 3. Model's feature importance

4 Conclusion

The model is rather successful in showing that a book's title indeed correlates to how well it is perceived, although not as much as one might think at first. As such it completes the aim of disproving the claim to "Never judge a book by its cover".

References

[1] Soumik. 2019. Kaggle: Goodreads-books dataset. https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/data
[2] Wikipedia. 2025. Random forest regression. https://en.wikipedia.org/wiki/Random_forest Accessed: November 2025.
[3] Wikipedia. 2025. TF-IDF vectorization. https://en.wikipedia.org/wiki/Tf%E2%80%93idf Accessed: November 2025.