Helpful Reviewers

An Amazon Book Reviews Dataset

Inferential Statistics Report

Corey Wade

How helpful is a given review? I analyzed an Amazon Book Reviews Dataset to answer this question. Helpful Votes and Total Votes were tallied into a list. I combined them to determine a Helpful Percentage. But Helpful Percentage does not tell the entire story.

A very strong correlation between Total Votes and Helpful Votes (0.985) indicated that users are more inclined to click if they find a particular vote helpful. Thus Helpful Votes is equally, if not more, significant than Helpful Percentage.

In addition to Helpful Percentage, and Helpful Votes, I considered a third factor, the total percentage of helpful votes for a particular book given to one reviewer. My rationale is that reviewers of popular books will likely receive more votes than reviewers of less known books, but that does not necessarily make them a better reviewer.

The quartiles for Helpful Book Percentage were extremely low, even in a limited sample of 10,000 plus rows, and strongly skewed to the right. I computed the log of the percentage and added 1 - e to obtain a maximum of 0.999. This left a minimum of just over 0.5.

I combined Helpful Percentage, Helpful Votes and Helpful Book Percentage into one weighted column, splitting 0.96 among Helpful Percentage, and Helpful Votes, while leaving 0.04 for Helpful Book Percentage. The Helpful Score column had a range of 0.02 to 0.99, and median of 0.54.

I created additional columns related to the text of the review, including Review Length, Average Sentence Length, and Average Word Length. A correlation matrix comparing these columns to Helpful Score revealed positive correlation coefficients ranging from 0.20 to 0.28. A negative correlation between Overall Rating, and Helpful Score, at -0.07 was also computed.

I verified the correlations by computing their p-values. The highest p-value had 15 zeroes after the decimal point. The significance of the correlations is due to the fact that over 10,000 points were sampled to determine the coefficients.

Histograms of all columns revealed a wide range of data. Overall Rating was strongly skewed to the left with a medium rating of 5 and a mean rating of just over 4. Word Length, by contrast, nicely approximated a normal distribution, and Helpful Percentage

was somewhat U-Shaped. All histograms summing votes were strongly skewed to the right.

In addition to histograms and correlations, I produced a scatter matrix and subsequent scatter plots to glean more insight. The scatter plots revealed crowding near the origin due to an imbalance of reviews that received very few votes, with patterns becoming more evident as the number of votes increased. A violin plot confirmed an overall pattern of Helpful Votes increasing, and Unhelpful Votes decreasing, as Total Votes increased.

My analysis reveals significant underlying patterns within the data. Further research is needed on the review texts to create more categories to supplement Review Length, Average Sentence Length, and Average Word Length. Creating numerous categories, however, may not prove as effective in the long run as an unsupervised approach to machine learning.

My entire Exploratory Data Analysis is available as a Jupyter Notebook via the link provided below. The project will continue with more research implementing a variety of machine learning methods to make further progress toward answering the question, “How helpful is a given review?”

https://github.com/coreyjwade/Ranking_Reviewers_EDA

May 27, 2018