Helpful Reviews
Milestone Report
Data Wrangling & EDA

Amazon has been under scrutiny lately for false reviews and fraudulent products. They are working with multiple teams to resolve the problem. I am part of independent review team developing machine learning algorithms to determine the reliability of reviews and reviewers.

We have finished stage three of the following four-step process:

1) Wrangle dataset
2) Create new metric
3) Analyze data
4) Apply machine learning techniques

We wrangled an 8.9 million row dataset of Amazon Book Reviews (Julian McCauley, USC). The original data was fairly clean. In my 15,000 sample, one review was eliminated due to the number of helpful reviews exceeding the number of total reviews. Two additional reviews were eliminated because there was no text in their reviews. All reviews with not total votes, helpful or unhelpful, were also eliminated.

I developed a new target column, Helpful Rating, weighted primarily by the percentage of helpful reviews and the total number of helpful reviews. The goal of the project is to predict Helpful Rating based on the Review column.

Exploratory data analysis was conducted on all columns to detect patterns within the data. Additional columns were added including Review Length, Average Sentence Length, and Average Word Length to get a better feel for the Review column.

Our initial findings are as follows:

1) The median review is 5.0 and the mean review is 4.25.
2) Users are much more likely to click (thumbs up or thumbs down) if they like a given review.
3) Review Length, Average Sentence Length, and Average Word Length were all positively correlated with Helpful Rating between 0.20 - 0.28.
4) Overall Rating and Helpful Rating were negatively correlated at -0.07.
5) There are a disproportionate number of reviews with Helpful Votes and Total Votes close to 0. Theses distributions are strongly right skewed.

The technical components of our four-step process may be utilized for other datasets including other Amazon products, book review sites such as GoodReads, and any other dataset that includes reviews in the column space. Our work can be of value to any company interested in creating a Helpful Rating for reviews, and furthermore, to implement machine learning algorithms to rate the reviews immediately upon being written.