Amazon Review Research

Ryan Erwin

December 2, 2015

Contents

Overview

This project will entail web scraping, text mining, and predictive models with the objective of predicting "Review" (y/n) and/or the rating (number of stars). This approach seeks to help sellers target the reviewers most likely to review their product with a high rating, which will also be seen as helpful to other shoppers.

Data

The information to be analyzed must be scraped from Amazon.com's list of Top Reviewers. For example, we'll need to identify the top reviewers then gather reviews. For each review, we'll want the review text, rating, percent and absolute value of helpful votes, product, product metadata, and any user (Reviewer) information available.

Methodology

As a first approximation, we'll apply the random forest algorithm (RF). I choose RF initially for out-of-box performance and relative ease of application. As the modeling progresses, the modelling approach will certainly evolve. The vast majority of programming will take place within the R language.

Conclusions

To work with such a large dataset, R, running on local machine becomes very limited in terms of analysis. This project is just as much about the technology to support the analysis as it is the analysis itself.

Next Steps

Use Amazon's Product Advertising API to get Product information which was not gathered during web scraping.