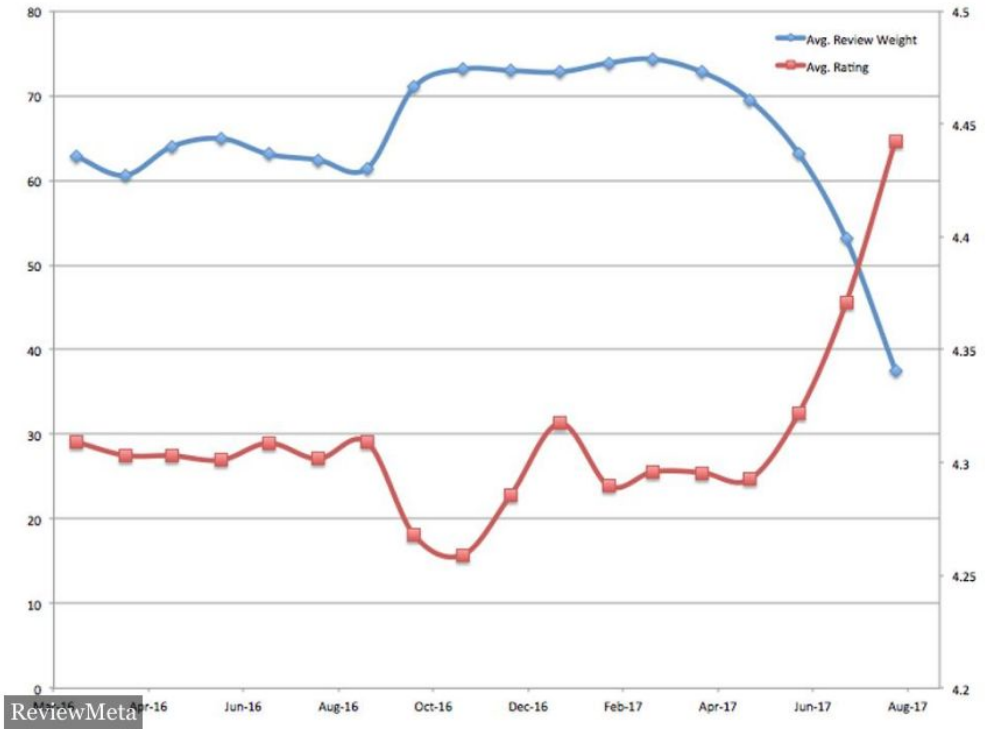
Huitong Pan (hp4zw)
Yingjie Liu (yl9qr)

Amazon Fine Food Reviews Analysis

Identify the problem

One way businesses generate and showcase good reputations is through their online reviews. With an extensive selection of shopping options, a helpful or unhelpful product review can have a huge impact on a potential customer's purchase decision. As a result, businesses are incentivized to create fake reviews, which not only mislead consumers but also lead to inefficient product purchasing decisions for e-commerce platform operators, particularly Amazon, who permitted 'incentivized reviews' until October 2016. Some reviewers were given free or discounted products in return for reviews, as long as the reviewer disclosed the arrangement. This strategy clearly incentivized fraudulent reviews, but even after disallowing them for most products, instances of fraudulent reviews have still been increasing. As reported by ReviewMeta, an Amazon review checker, there has been a drastic increase in the number of fraudulent reviews on Amazon, as shown in the graph below. The average review weight - the measure of a review's trustworthiness - has fallen in June, July and August 2017.[1]



Drastic increase in the number of fraudulent reviews on Amazon

[1] Woollacott, Emma. "Amazon's Fake Review Problem Is Now Worse Than Ever, Study Suggests." Forbes. September 13, 2017. Accessed December 07, 2017. https://www.forbes.com/

The desired alternative situation is for only honest reviews to appear on the website. Amazon has managed to ban many clearly fraudulent reviews for a vast majority of products and brought lawsuits against over 1,000 sellers that have benefitted from fraudulent product reviews.

Although our study is based on Amazon reviews, it addresses the broader question of how to effectively detect and ban sellers who benefit from fraudulent reviews, which is a major concern for all e-commerce platforms. Moreover, consumers, sellers, and legislation are also affected by the problem of fraudulent reviews.

Transitioning from the current situation to the desired alternative, in which only honest reviews are published, would have various positive impacts on e-commerce platforms, consumers, sellers, and law enforcement.

Since honest reviews create a better online shopping experience, e-commerce platforms will be able to attract more customers and, in return, more sellers to their marketplaces. Consumers will be able to make better-informed shopping decisions based on honest reviews. Sellers will enjoy fairer competition that will encourage them to deliver better products and services. Together, this concert of positive impacts will help to create a better online shopping ecosystem.

Define objectives and metrics

Our objective was to predict whether a review was helpful or not through text mining. There were many possible performance metrics we could have used to evaluate our models, including ROC, Accuracy, the Gini index, and MSE. We recognized that our dataset had significantly more unhelpful than helpful reviews, so we chose ROC as our performance metric since it would be less affected by the bias that skewed data might introduce.[2]

Understand the State-of-the-Art

There are many Kaggle users exploring the same problem using Amazon's review dataset. There are also many commercial companies investigating review datasets from e-commerce platforms. It seems that several machine learning techniques have been applied toward solving this problem, including artificial intelligence, natural language processing, and neural networks.

What makes this problem difficult is the nature of human language, in particular, that common

[2] Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. "Facing Imbalanced Data Recommendations for the Use of Performance Metrics." International Conference on Affective Computing and Intelligent Interaction and workshops : [proceedings]. ACII (Conference). 2013. Accessed December 08, 2017. https://www.ncbi.nlm.nih.gov/.

constructs often convey mixed or ambiguous information. In addition, text analysis methodologies rely on hand-labeling, which is susceptible to the introduction of human biases and is scientifically unreliable.

Define hypotheses and approach

We are going to use the Amazon Fine Food Reviews dataset, which consists of 568,454 fine food reviews written by Amazon users up through October 2012. The dataset provides:

- Amazon user ID
- profile name
- product ID
- the number of people who indicated the review was helpful ("helpfulness numerator")
- the number of people who reviewed the review ("helpfulness denominator)
- rating of the product ("score"),
- timestamp of the review publication date ("time")
- summary of the review text ("summary")
- review text ("text")

Hypotheses:

1) The frequency of certain words can indicate whether a review is helpful or not with greater accuracy than random guesses.
2) The length of a review can indicate whether a review is helpful or not with greater accuracy than random guesses.
3) The product score can indicate whether a review is helpful or not with greater accuracy than random guesses.

Our approach consists of 7 steps:

Data:

1) Clean duplicates

Many duplicate reviews were present in our dataset, often written by the same users, and sometimes even posted contemporaneously. These duplicate reviews and the users who wrote them were very likely to be fraudulent. However, since our study focused on the helpfulness of reviews, we did not perform further analysis on the duplicates. As such, we deleted them from our dataset, reducing the total number of rows from 600,000 to 400,000.

2) Create indicator variable for helpfulness

We created an additional variable, *help_int*, to indicate whether a review was deemed to be helpful. If *help_int* is 0, the the review was considered unhelpful, and vice versa. We considered a review as helpful if it fulfilled the following criteria:

- It had more helpful than unhelpful counts (i.e. helpfulness numerator/helpfulness denominator > 0.5)

- The sum of the helpful and unhelpful counts was greater than 3. If fewer than 3 people evaluated the helpfulness of the review, the indication of helpfulness may not be reliable (many reviews had only 1 helpful count).

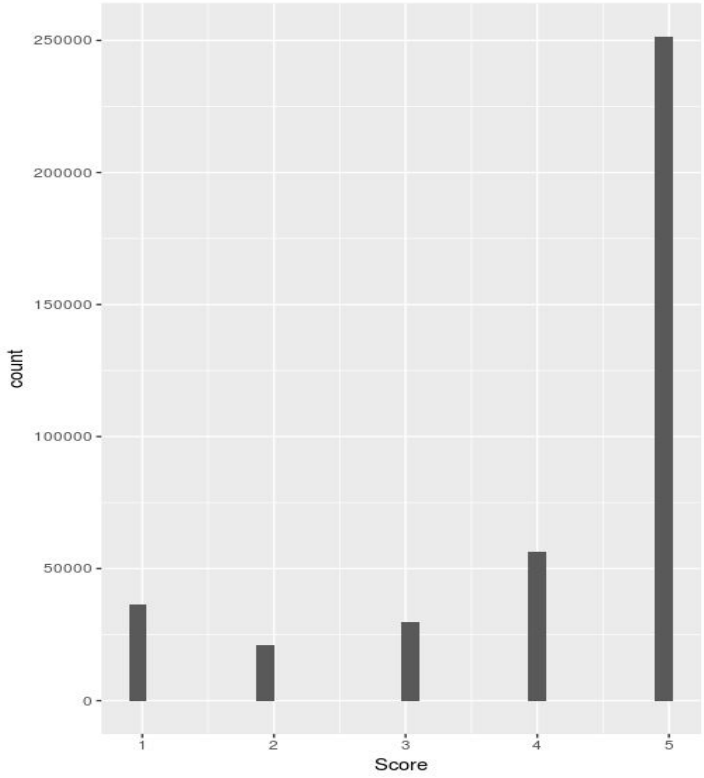
3) Calculate review and summary length

After reading some of the reviews, we discovered that the helpful ones were usually longer. Thus we created new variables, *text_length* and *summary_length*, and used them to test our second hypothesis.

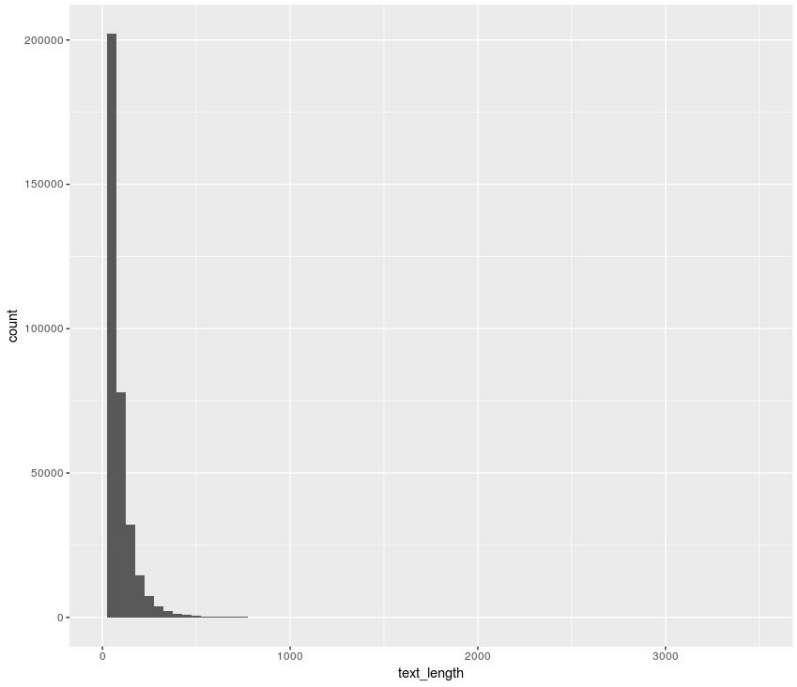
4) Data exploration

We performed some data exploration to better understand our dataset. Below are some of the conclusions we drew from the data exploration.

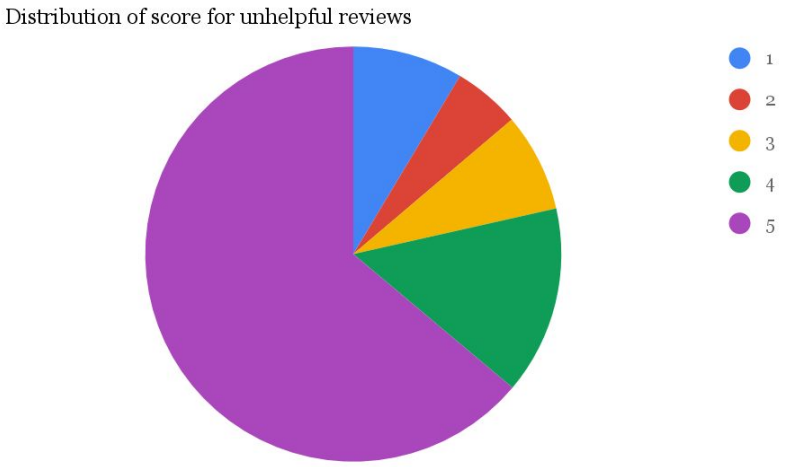
a. Most of the reviews were associated with score of 5.

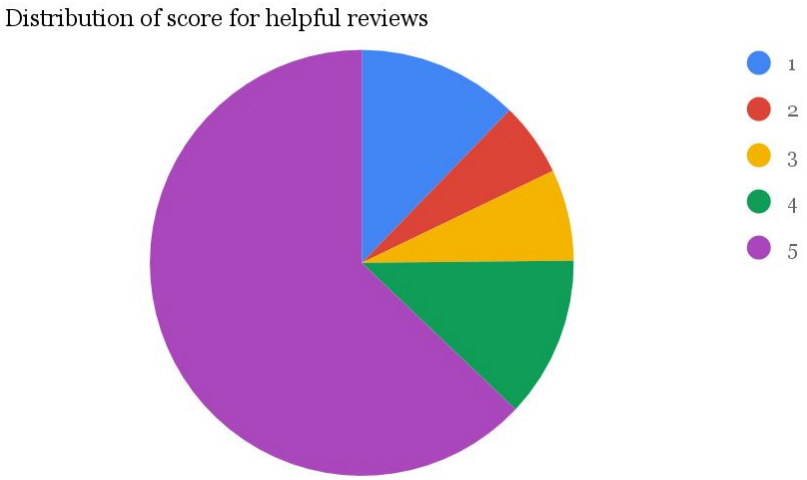


b. A majority of the reviews were relatively short, but some were longer than 3,000 words.

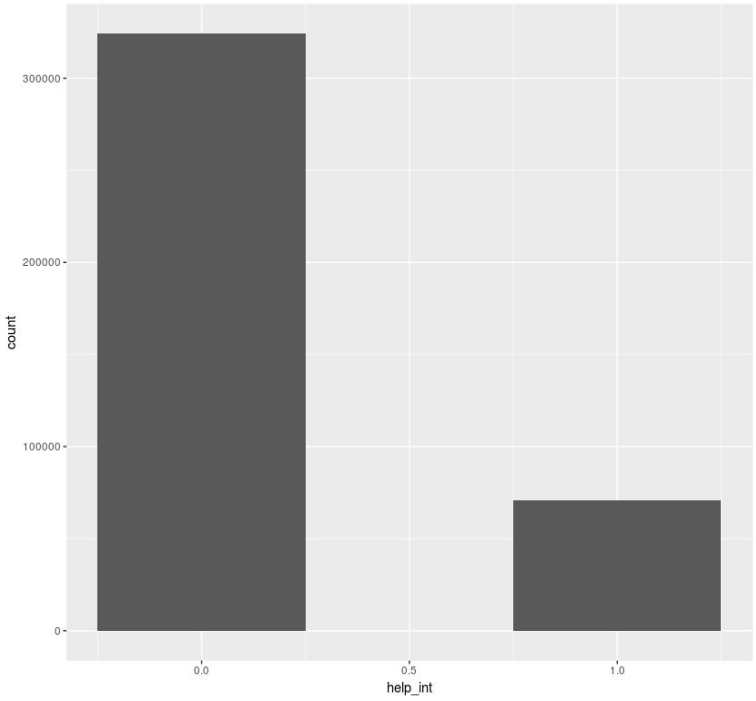


c. Most of the helpful *and* unhelpful reviews had a score of 5.





d. Reviews were more often marked as unhelpful



We also checked the pairwise correlation between helpfulness and review rating, the length of the review summary, and the length of the review text. There seemed to be no significant correlation between those variables, with the highest correlation of 0.15 being between helpfulness and the length of the review text.

Methods:

5) Text Mining

We combined each review's text and summary into one column and performed text mining on the data. First we removed numbers, whitespace, punctuation, and stopwords. Then we stemmed it and converted it to a term-frequency matrix. We removed sparse terms with the threshold at 0.99, which reduced the number of terms from 291,000 to 657.

We also discovered that the frequent bigrams added 44 terms to our term frequency matrix. The trigrams did not add any terms to our term frequency matrix.

6) Modeling

We used *help_int* as our output variable and experimented with three scenarios:

1. included only term frequencies as feature vectors,
2. included both term frequencies and the combined review text length as feature vectors, and
3. included term frequencies, combined review text length, and score as feature vectors.

Since modeling with all the data was computationally infeasible for us, we sampled 10,000 rows to tune our models. Most of our models were built with the Caret package. The models we built include:

a) Logistic regression
We chose logistic regression because our response variable, *helpful_int*, was binary. Logistic regression also provides high interpretability, which would allow us to easily tell whether longer text positively contributed to the helpfulness of a review.

b) Boosting
We suspected that our inputs would only slightly correlate with the true classifier, so we chose to use a boosting algorithm since it would iteratively learn weak classifiers and convert them to strong ones, boosting the overall performance of the classification.

c) K-NN
We hypothesized that helpful reviews would use similar words, so we chose K-NN to see whether reviews formed predictive clusters or not.

d) Topic modeling through K-means and SVM
We used K-means to form clusters with similar topics, and used that topic information to predict review helpfulness through SVM.

e) Recursive partitioning.
Recursive partitioning is a statistical method for multivariable analysis, which was suitable for our 700-variable prediction problem. Our recursive partitioning model created a decision tree that strove to correctly classify reviews by splitting

the data into subpopulations based on several dichotomous independent variables.

Evaluation setup:

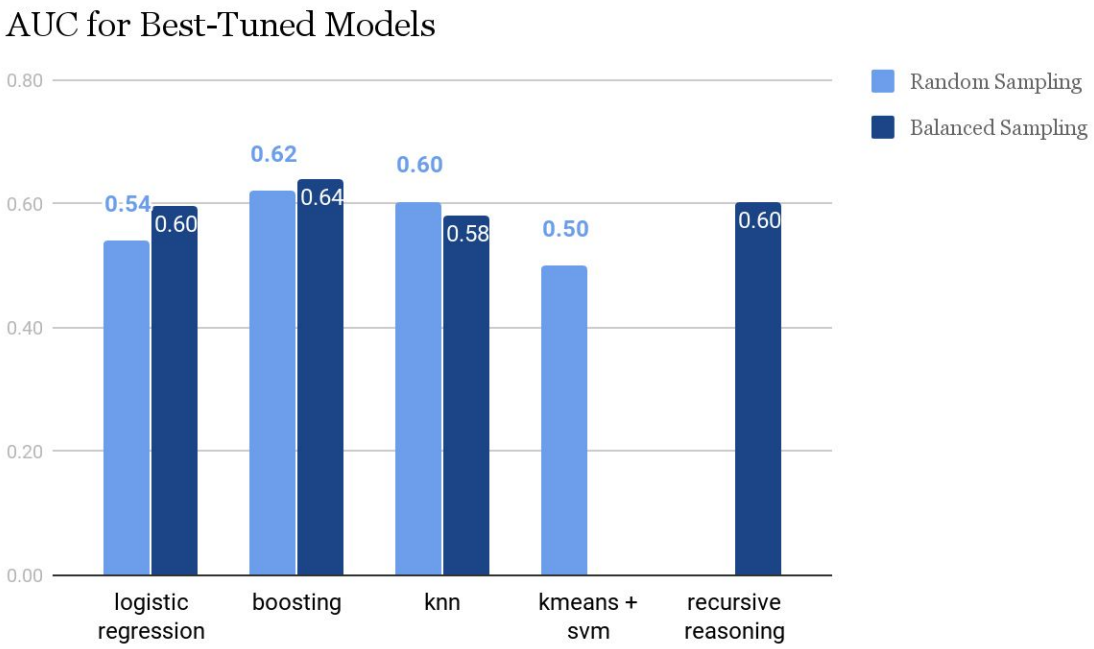
7) Cross-validate

We implemented two-folds cross-validation to determine the best set of parameters and to evaluate each of our models. Due to the size of the dataset, cross-validation with more than two folds would be computationally infeasible for us. We compared the average AUC results generated after cross-validation for each model using random and balanced (same number of helpful and unhelpful reviews) samplings to discover the best-performing model.

Execute approach and report results

We compared the results of the three scenarios identified in the Modeling section. Scenario 2, wherein both term frequency and the combined review text length were included as feature vectors, achieved the best results across all models.

Since most of the reviews in our dataset were deemed unhelpful, we also performed balanced sampling. The AUC results from random sampling and balanced samplings are shown below.



What we learned about the problem was that review text and length provide only limited information about the helpfulness of a review. Our best-performing model was boosting, which yielded an AUC of 0.64 with balanced sampling. This result is only slightly better than a random guess. According to ReviewMeta, "...along with their five-star ratings, fake reviews tend to be longer than average and written like a product brochure, often incorporating pictures and

video." Due to the nature of the dataset, wherein five-star reviews drew in the lion's share of both helpful *and* unhelpful ratings, we are not able to extract much information from score. However, we did find that text length could be a useful indicator of the helpfulness of a review.

We believe we could perform additional analyses to extract more value from the review text, such as part-of-speech (POS) tagging. The POS tags could provide additional information on how words are used in review texts. For example, supposing we discovered that helpful reviews contained fewer adjectives and more nouns, we could include the frequency of adjectives in the review text as an indicator of the helpfulness of a review. Due to the limitations of our computing power, we only provided demonstration code for POS tagging as a proof of concept.

Regarding the data, we learned that the uneven distribution of helpful and unhelpful reviews created a bias that reduced the performance of our models. When we trained our models on a balanced sampling, they performed better than on an imbalanced, random sampling of data.

Regarding the methods, we learned that boosting was the best-performing model for our situation. This may result from boosting's slow learning process that tends to perform well.

Regarding the evaluation setup, we learned that even though ROC is relatively insusceptible to imbalanced data, our models achieved better AUC results when the data was balanced. Note that the high ratio of unhelpful to helpful reviews in our dataset was the opposite of what generally occurs in the real world, where the majority of reviews are useful. In our case, our model was less likely to show bias toward the costly false-positive: unhelpful reviews labeled as helpful. As such, when implementing our model in real-world situations, care should be taken to balance the data and use several performance metrics to select the best-performing model, such as the general F1-score.

Overall, our results partly supported our hypotheses. Although our models only slightly outperformed random guesses, we were able to conclude that the use-frequency of certain words, together with review text length, can indicate whether a review is helpful or not with greater accuracy than a random guess. However, the product score associated with a review does not provide much value in evaluating whether the review is helpful or not.

While performing our analyses, we began to sense that particular pieces of unstructured data in the review text would be worthy of evaluation and perhaps bring us even closer to the desired alternative situation. For instance, reviews indicating that the user reviewed the product after a certain length of usage (e.g. 2 weeks) are oftentimes helpful. Thus, an indicator for the length of usage might be predictive. Another possible indicator is the occurrence of comparisons to other brands or products. There are many additional features we could consider for building a stronger model, but they require a significantly more granular approach.