What Makes a Valuable or Fake Customer Review?

Springboard Data Science Career Track
Capstone 2 - Milestone Report 2
by Chantel Clark

The Problem

The online reputation of businesses is extremely important these days in gaining the trust of customers. One way that people determine whether or not to buy a product or service is through reading customer reviews. However, it can be difficult to know which reviews are useful because of the presence of 'fake' reviews. Products on Amazon with a 5-star review may not actually have a true 5-star rating because of a flood of incentivized or paid reviews. The inflation of reviews makes it very difficult for the average or ethical company to have their products show up in user searches when another product has over a thousand fake, glowing reviews. Since 2015, Amazon has been working to crack down on the users who write paid reviews, and companies that solicit them (Dwoskin and Timberg, 2018). However, 2019 ReviewMeta found that even products which receive Amazon choice badges contain many suspicious reviews. Fake reviews pose a problem for both businesses and potential customers. How can companies compete with those that use paid reviews, and how can customers get a true estimate of the quality of product or service?

A machine learning model that uses Natural Language Processing (NLP) and reviewer metrics can help customers to obtain the true quality of online products by removing suspicious reviews and finding the valuable and unique reviews. By sifting through the reviews more efficiently, a machine learning model can help customers to save time and money when shopping online. The aim of this capstone is to explore how NLP can be used in combination with reviewer metrics to determine what makes a review valuable or suspicious.

Data

Amazon has an open dataset of over 130 million customer reviews collected between 1995 and 2015, available as URL's at https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt. Each line in the dataset represents one product review. Reviews are grouped by product categories such as apparel, automotive, books, e-books, etc. The columns in the dataset include: 'marketplace' (country code), 'customer_id', 'review_id', 'product_id', 'product_parent' (random identifier for aggregate reviews for the same product), 'product_title', 'product_category', 'star_rating', 'helpful_votes', 'total_votes', 'vine', 'verified_purchase', 'review_headline', 'review_body', 'review_date'.
Complete an exploratory data analysis to answer the following questions:

1. *What is the mean and median number of reviews per customer?*
2. *What do the reviews of a highly active (>500 reviews) reviewer look like?*
3. *What do the reviews that a one review customer look like?*

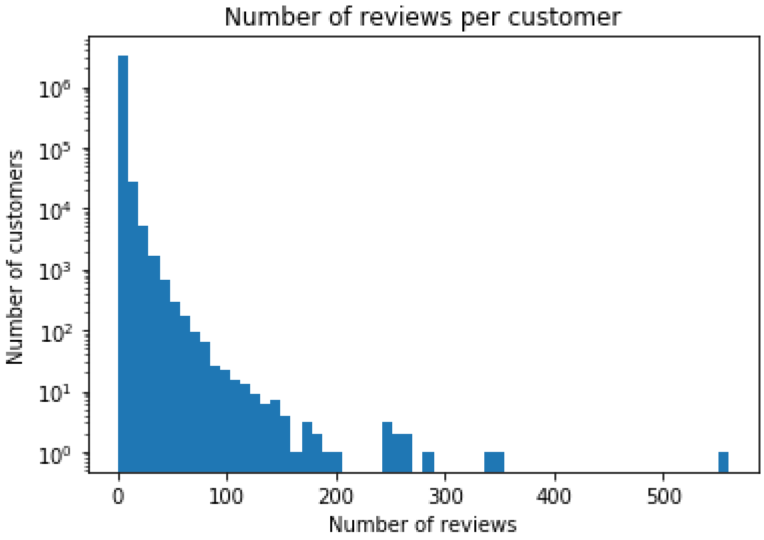
4. *Do customers who write different amounts of reviews give the same distribution of star ratings?*

Analysis

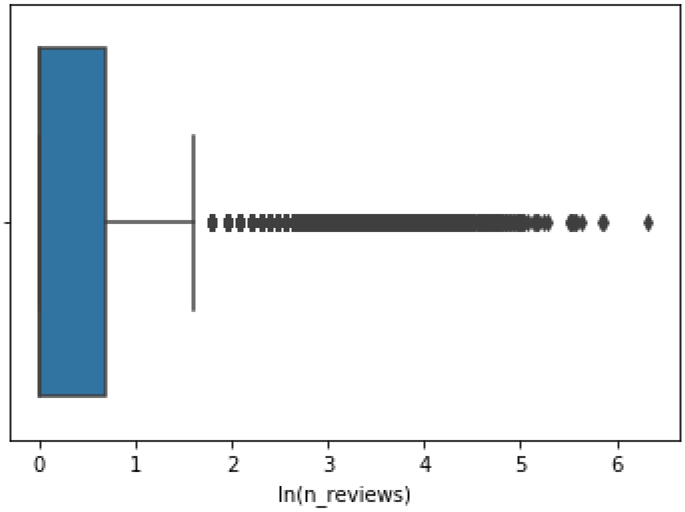
To analyze the dataset, an ‘apparel' subset with 5,881,873 reviews was used. About 10.1% of the reviews were unverified purchases, meaning that the product was not bought through Amazon. Browsing through the data, it was apparent that some reviews were misclassified in the apparel category. For example, a pineapple corer and beach hammock were contained within the apparel category. In order to get an idea about who the customer reviewers are, the following questions were addressed:

1) *What is the mean and median number of reviews per customer?*

The mean number of reviews that a customer writes is 1.83, while the median number of reviews per customer is 1. Because the average is larger than the median, it is evident that some customers are posting a very large number of reviews which is skewing the dataset to the right (long right tail). The median number of reviews per customer is low, and could be a result of fake accounts where only one review is posted.



There is an extremely large number of customers that post only one review (68.4% of all customers). As an attempt to identify outliers (highly active reviewers), the data was log transformed. The interquartile range (IQR) was computed and multiplied by 1.5 to find the length of the whisker in the box plot.

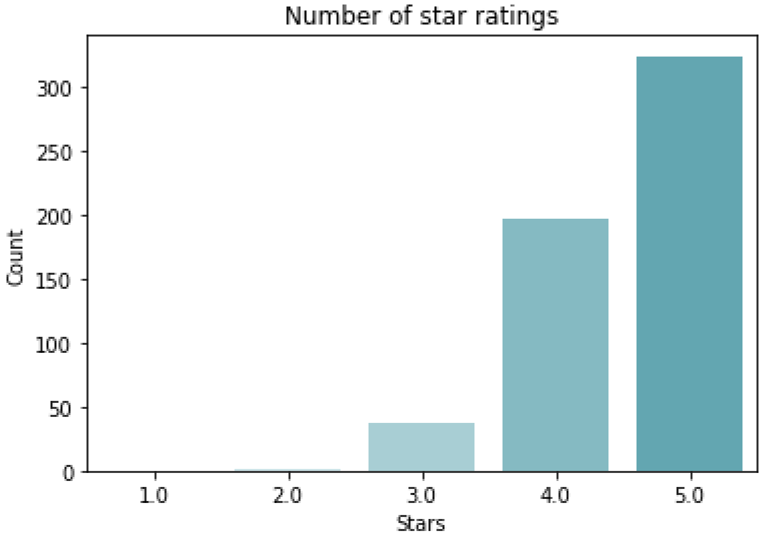


Using this method, customers who wrote more than 5 reviews would be considered as an outlier. In this dataset, there were 130,943 (4.1%) customers who would be classified as an outlier. This obviously does not seem right because on average, a 'real' (not fake) customer could write 6 reviews easily. The large amount of customers that write only one review is still skewing the data. In this case, the multiplier of 1.5 could be increased to 3 (or any desired multiple) to modify the threshold in which to identify customers that could potentially be a fake reviewer.

An alternative approach is to remove the reviewers who only wrote one review. When the ‘one hit wonders’ (people who only wrote one review) were removed, the threshold for the number of reviews that a customer writes to be considered as an outlier increased to 12. Less than 1% (0.89%) of all customers wrote 12 or more reviews.

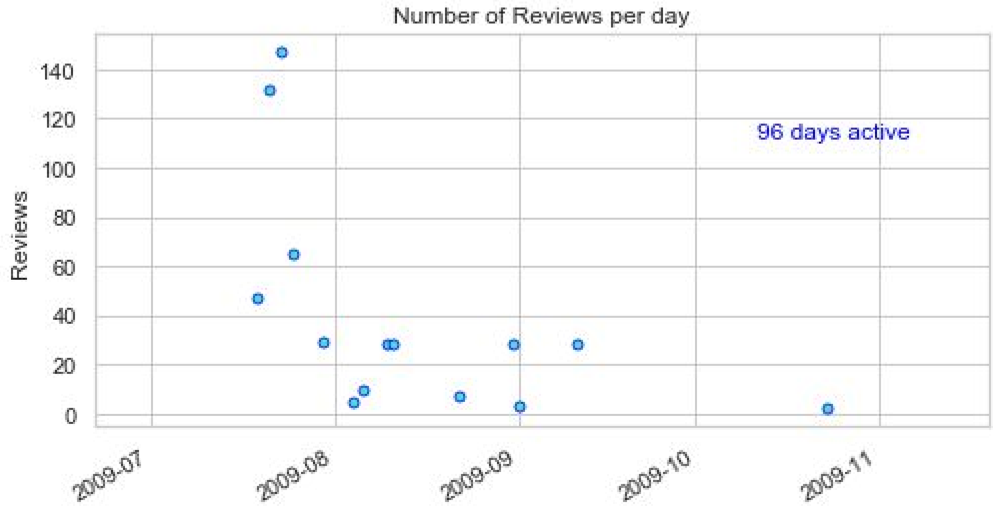
2) *What do the reviews of a highly active (>500 reviews) reviewer look like?*

None of the 559 purchases were verified. The average star rating was 4.5, with just 300 5-star ratings and over 200 4-star ratings. It’s actually quite surprising that there were 3-star ratings, and one 2-star rating. Perhaps this is a strategy from getting caught.



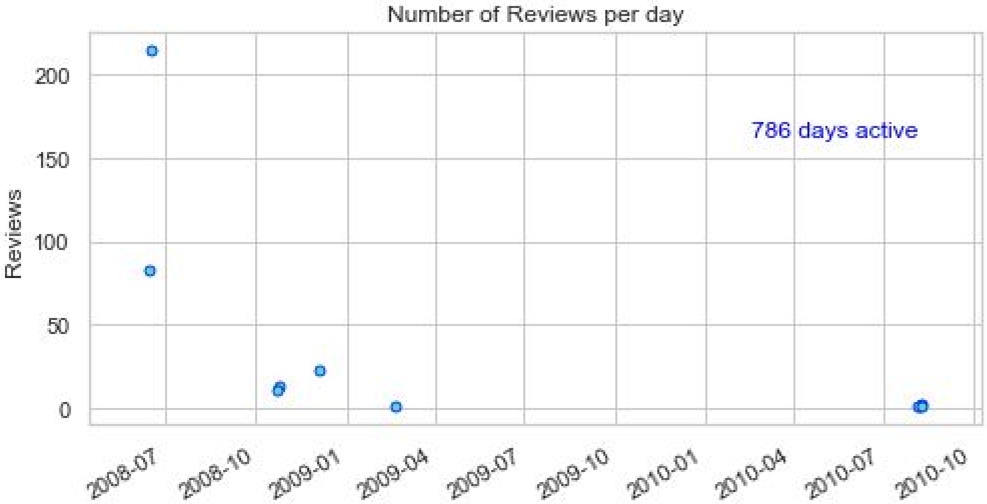
There was only one 2-star rating from the reviewer that is clearly a positive review which does not match the 2-star rating and is obviously a promotion for the store:

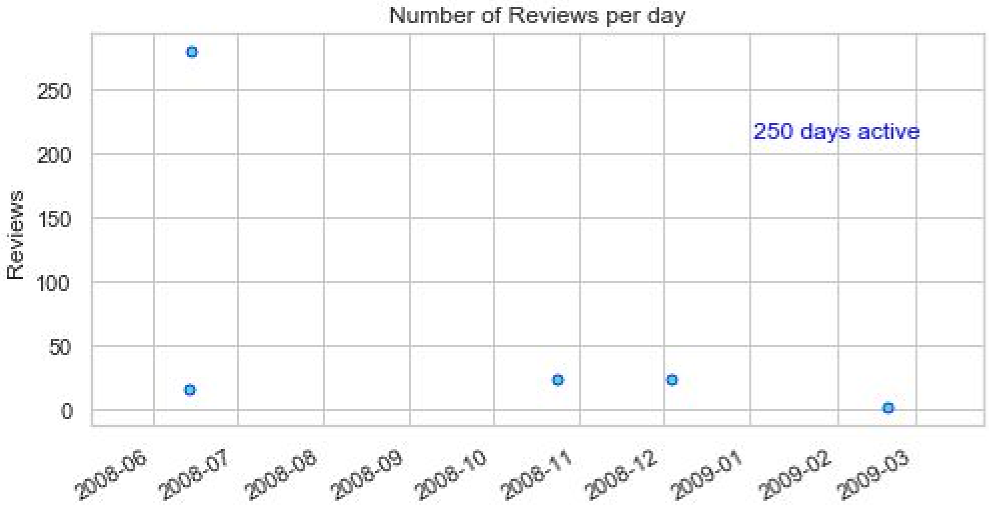
> *“Bollywood Style Designer Indian Kurti - Unique Indian Kurtis for Women -Summer Dress lucknowi-chikan-- It's the latest in Indian fashion! If you are looking to buy the designer kurtis (tunic) to match with your jeans, trousers or salwars, simply browse this store.............”*



The timeline above shows the frequency and number of reviews posted over time. This customer’s first review was written on July 19, 2009, and last review on October 23, 2009. The total time that this customer was actively writing reviews was 96 days. On July 21 and 23 of 2009, there were over 100 reviews posted, clearly an anomaly. Number of reviews drastically decreased through October 2009. One way to flag suspicious review activity is to identify a maximum limit of reviews written per day for each reviewer. In this case, it might be reasonable to say that writing over 30 reviews is suspicious. Regardless of what threshold is chosen to identify the number of reviews posted in a day, it can be adjusted in the model.

Below are graphs displaying the frequency of posts per day for other reviewers who posted 30 or more reviews in a day. The distributions similarly start very high and taper off quickly within a month.





Many reviews from the most active customer reviewer are exactly the same. Index of the review followed by the review body are below:

> *5027866 Stunning Cotton Kurti with gorgeous colored print. This ethnic kurti is master piece of Indian Ethnic Art.*
> *5064403 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*
> *5064413 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This crinkle skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*
> *5064414 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*

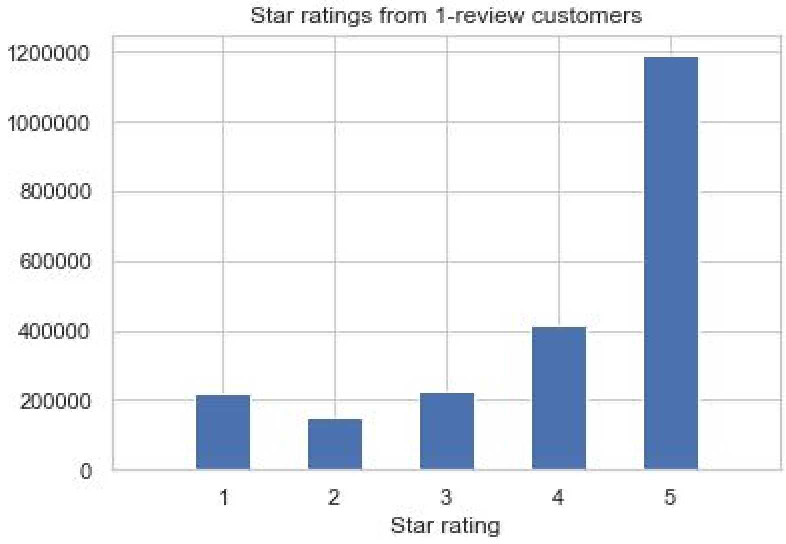
It is quite obvious that these reviews look like fake or unreliable reviews because of the repeated text for different items. Extremely large numbers of similar reviews can be an indicator of unreliable reviews, which could be identified through Natural Language Processing.

3) *What do the reviews that a one review customer look like?*

There were 2,201,632 customers who left only one review (37.4% of all reviews). A random sample of 50 reviews were chosen to inspect the reviews from customers that only posted one review. Most of the reviews in this subset were surprisingly thoughtful and unique, and did not seem to be computer generated. Even the reviews that were unverified and were not voted as ‘helpful’ were surprisingly well-written:

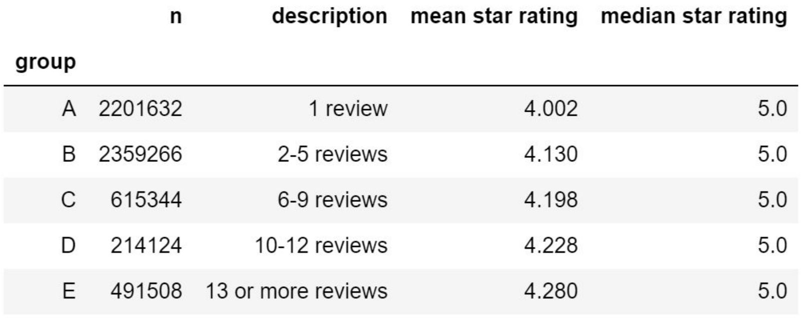
> *Fits true to size BUT don't be alarmed when you put them on the hips are a little tight. Trust me that they will loosen to a comfortable fit.*

The distribution of star ratings for the one-review customer were mostly 5-star ratings (54.1%) followed by 4-star ratings (18.9%). While most of these reviews were 5-star ratings, the amount of 1 through 4-star ratings was reasonable, within 1 order of magnitude.



4) *Do customers who write different amounts of reviews give the same distribution of star ratings?*

If customers were grouped by the number of reviews that they have written, are the star distributions different? To compare, customers were segmented into group A if they wrote 1 review, group B for 2-5 reviews, group C for 6-9 reviews, group D for 10-12 reviews, and group E for 13 or more reviews.



	n	description	mean star rating	median star rating
group				
A	2201632	1 review	4.002	5.0
B	2359266	2-5 reviews	4.130	5.0
C	615344	6-9 reviews	4.198	5.0
D	214124	10-12 reviews	4.228	5.0
E	491508	13 or more reviews	4.280	5.0

Because the data is ordinal, and observations (reviews) can come from the same reviewer, a Kruskal Wallis test was used to determine if the star ratings have the same median across groups, and come from the same distribution.

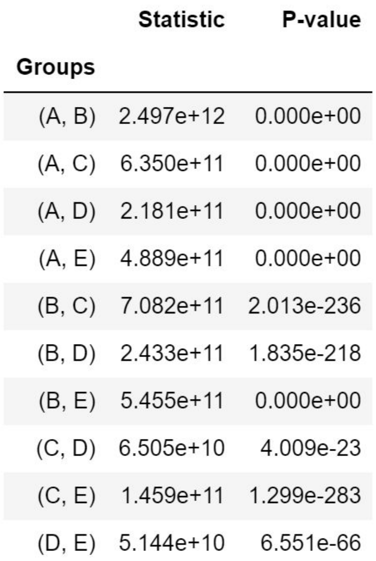
H_0 : $median_A = median_B = median_C = median_D = median_E$
H_1 : Two or more of the distributions do not have the same median

The Kruskal Wallis test was run for groups C, D, and E, and returned a test statistic of 1298.9, and a p-value of 8.706e-283. The p-value is very close to zero, therefore the null hypothesis can be rejected and we conclude that the medians of groups C, D, and E are different. This answers the question at hand - customers from different groups give different distributions of star ratings.

To further analyze the differences between each pair of groups, a Mann-Whitney U test was used with the following hypotheses.

$H_0 :\ median_A = median_B$
$H_1 :\ median_A \neq median_B$

Results of the Mann-Whitney U tests are in the table below. P-values are all either equal to zero or very near to zero, therefore we reject the null hypothesis and conclude that each group has a median and distribution that is different from the other groups, and there are no two groups that are alike.



Groups	Statistic	P-value
(A, B)	2.497e+12	0.000e+00
(A, C)	6.350e+11	0.000e+00
(A, D)	2.181e+11	0.000e+00
(A, E)	4.889e+11	0.000e+00
(B, C)	7.082e+11	2.013e-236
(B, D)	2.433e+11	1.835e-218
(B, E)	5.455e+11	0.000e+00
(C, D)	6.505e+10	4.009e-23
(C, E)	1.459e+11	1.299e-283
(D, E)	5.144e+10	6.551e-66

Building a Machine Learning Model

Preparing the data

785 reviews without text were removed for NLP analysis.

Labeling ‘suspicious’ reviews

A threshold of 30 or more reviews posted within a day was chosen as an identifier of a ‘suspicious’ reviewer. A boolean array was added to the dataset to identify whether or not a review was written by a suspicious reviewer. The labels were highly imbalanced, with 15,360 reviews from suspicious reviewers, and 5,865,729 reviews from non-suspicious reviewers. Based on the threshold, only 0.26% of reviews were labelled as ‘suspicious’.

Splitting data into training and test sets

One thing to keep in mind is that reviews are not independent - rows may represent reviews that are written by the same person, or for the same product. From the exploratory data analysis, it was apparent that the suspicious reviews are repetitive and oftentimes exactly the same for various products. Having reviews from suspicious reviewers in both training and test sets would adversely affect model outcomes by increasing bias, therefore the data was first split by reviewers/customer ID. Suspect and non-suspect reviews were separated. There were a total of 268 unique customers labeled as 'suspect'. The reviews from 80% of the suspects were used for training (12,143 reviews), and the reviews from the remaining 20% of suspects were used for model testing (3,217 reviews). The non-suspect reviews were also split, with 80% of the reviews (4,692,583 reviews) added to the training set, and 20% (1,173,146 reviews) added to the testing set.

Natural Language Processing

Tokenization was used to break up each review by unique words, and build a vocabulary for the entire corpus (collection of reviews). Words that have little meaning such as 'a', 'the', 'she' or 'he' are called stop words, and these were removed with the 'english' dictionary provided by TfidfVectorizer. To account for the varying forms of a word with the same general meaning, lemmatization was implemented (for example, 'bad' and 'badly' would both be converted to the word 'bad').

Each review was vectorized with TF-IDF (term frequency - inverse document frequency). The TF-IDF score is directly proportional to the word frequency in a specific review, and inversely related to the word frequency across all documents. As a result, words that are very common in other reviews receive a lower score than the words that are used infrequently. The length of each vector is the total number of unique words from the corpus, so as you can imagine, each vector contains many zeros because there are thousands of words in a corpus that are not in any given review.

Pipeline

A pipeline was created in order to efficiently test different classifiers. The classifiers that were trained and tested include Naive-Bayes, Random Forests, and Linear SVC. The pipeline first used TfidfVectorizer to vectorize each review in the training set. This process included tokenizing the review, lemmatization, and removal of stop words and words that appeared in less than 5 reviews. After reviews were vectorized, the testing set was vectorized and used to predict the outcomes of the test set.

Outcomes

Results had high accuracy (above 99% for all classifiers) but really bad precision and recall (mostly 0%), which was a result of highly imbalanced data (15,360 reviews from suspicious reviewers, and 5,865,729 reviews from non-suspicious reviewers). Based on the threshold, only 0.26% of reviews were labelled as 'suspicious'. Another reason for low precision and recall was the difficulty with using TF-IDF vectors alone to classify into suspicious or non-suspicious reviews. To improve the identification of suspicious reviews, other review metrics such as number of words, verified purchase, star ratings, and ‘helpful’ votes will be incorporated into the model.