Do Amazon Reviews Matter?

An Empirical Analysis of Amazon Product Reviews and Product Prices

ECON 484
Machine Learning for Economists
Prof. Frandsen

Zachary Flynn
McGyver Clark
Treyce Watson
Kalten Toone

Introduction

E-commerce sales platforms have become a mainstay of consumer markets around the world. Amazon, starting as an online bookstore, has become the world leader in this industry accounting for 37.9% of all online retail sales in the United States of America for the year 2020 .[1] Amazon, being both a producer and an online storefront where private sellers and companies can sell products, has increased information and competition for almost every consumer product showing little inflation where national average inflation has steadily increased [2]. Therefore for both consumers and sellers in the Amazon market, understanding the dynamics and principal features that affect pricing on the market is crucial for making decisions around what to purchase and what to sell. In this analysis, we explore how previous customer ratings and review sentiments have a causal effect on current product prices. We utilize PCA for data cleaning to help categorize our items listed on Amazon, creating and for controlling for high dimensional covariates to create a robust and valid economic analysis. Causal regression forests are then used to estimate the causal impact that reviews, and review sentiment, have on prices.

Data

The data used in this project is both general product information and Amazon review information gathered by J. McAuley , and updated by Jianmo Ni containing 233.1 million unique Amazon reviews.[3,4] The Amazon review dataset contains a few control variables of interest, namely “Overall”, an overall star rating of the product, “Verified, which determines if the review is likely to be made by AI or not, and in the same vein, “Verified binary”, which shows if it was a verified purchase. We only use those observations that are non-AI and binary for our analysis to increase the relevance of our results for the consumer. As this is a large dataset, we filter this data to selectively include products that categories that can be considered close substitutes (e.g. “Appliances, Musical Instruments, etc.). We then further separated these categories into subcategories to ensure that we grouped like products (e.g. washing machines and washers are both in appliances but should not be compared in terms of pricing). This data was merged into panel data product and price, then categorized as product type by matching with review data. Review data was analyzed using rating information (measured by a rating scale from one to five stars). Furthermore, raw review text was analyzed by sentiment, creating a quantitative measure of the text being used. For our sentiment score, we utilized two different packages, Textblob and Natural Language Toolkit: Sentiment Intensity Analyzer to take the strings from each of our Amazon reviews and evaluate the sentiment of that review. It set each of our reviews on a range from -1 to 1, with -1 being the most negative, and 1 being the most positive. We then applied this to our dataset to receive a general sentiment score for each review.

1. *He, Leshui, Imke Reimers, and Benjamin Shiller. "Does Amazon Exercise Its Market Power? Evidence from Toys “R” Us." The Journal of Law and Economics 65, no. 4 (2022): 665-685.*
2. *Cavallo, Alberto. More Amazon effects: online competition and pricing behaviors. No. w25138. information retrieval, pp. 43-52. 2015.*
3. *McAuley, Julian, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. "Image-based recommendations on styles and substitutes." In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp. 43-52. 2015.*
4. *Ni, Jianmo, Jiacheng Li, and Julian McAuley. "Justifying recommendations using distantly-labeled reviews and fine-grained aspects." In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 188-197. 2019.National Bureau of Economic Research, 2018.*

Method

Traditional economic models that predict prices in markets such as use supply, demand, competition, and information to perform price prediction. In the case of the Amazon market, there are many individual sellers, small and large firms selling many heterogeneous goods. The market experiences almost perfect competition with almost perfect information. Therefore, we would expect that firms are price takers, prices are efficient, and demand would rely on individual consumer demand.[5] However, it would be safe to assume that product price would be heavily influenced by consumer-provided information via ratings and reviews. Ratings and reviews are subject to individual bias and when economic models include features of reputation, loyalty and quality-leveraging, the standard economic models can do a poor job of grasping the entire story of how prices are determined. Furthermore, due to the vast number of substitutes and the cost of research there are other factors to measure and tease out, especially regarding the behavioral side of psychological impacts of seeing low or high ratings and review. We will explore these ideas to extract inference into understanding the principal components of online e-commerce markets and the import of consumer ratings and reviews. Amazon itself is an interesting data playground to play on, with data readily available on price history, review analysis (including authenticity of reviews, review variances, or number of reviews' impact on price), or comparing its unique features as a storefront to other online marketplaces such as eBay.

A random causal forest allows us to use different parts of the tree to estimate the standard errors, while allowing us to make our estimates for our treatment effects honest, which gives us a more accurate manifestation of what our causal effects are.

Our data is not without limitation. While the data was relatively clean from the source, we had to further expand it by adding dummies for each of our categories, which introduced a large number of additional features, limiting us in our capability to cluster items using machine learning methods. Furthermore, clustering them based on our subcategories is fairly limited in scope, as items that match on all subcategories may not even be functionally the same, meaning that they are not relatively close substitutes. We also manually created our own sentiment scores, which we used to compare products which were binned in the same price range.

We cleaned and structured our dataset of product and reviews into panel data, filtering out duplicates and null data. Our sentiment analysis is the key feature that we explore in our analysis. We explore many methods of machine learning algorithms that predict the best accuracy for this use case such as Linear Regression, Random Forest, Gradient Boosting Regressor, Support Vector Machines, and Elastic Net.

In our analysis, we attempt to answer the following:

1. A forecasting model that uses product rating and reviews to determine the causal effect that review, and general review sentiment has on product prices.
2. Understand relationships between product features, consumer reviews, and economic factors on ecommerce pricing.
3. Find actionable insight for consumers and sellers based on analysis.

[5] *Chevalier, Judith, and Austan Goolsbee. "Measuring prices and price competition online: Amazon. com and BarnesandNoble. com." Quantitative marketing and Economics 1 (2003): 203-222.*

Due to the popularity and importance of reviews and ratings, a new market has emerged of companies incentivizing through gifts, discounts, and direct payments to users, consumers, and other companies to publish positive ratings and reviews. Therefore, we assume that larger firms would have greater probability of having positive reviews due to their resource availability to purchase reviews, introducing endogeneity into our model. Therefore, we use a bias score based on product type, and seller type to control for potential endogeneity and bias in review data to cancel out the recursive effect on price as companies will participate in buying positive sentiment.

Results

Using our random causal forest method, we determine that reviews do have a statistically significant causal effect on the prices of a product, particularly inexpensive products. When examining items at similar prices, and without considering the higher purchase volume of certain items among items that are comparable, reviews appear to make a greater impact than if the item were more expensive. As shown in Table 1, there is a treatment effect for each of our clusters, but our clusters produce results with decreasing statistical significance as our price range increases and our cluster size decreases.[6] In an effort in robustness, our results are the product of a random causal forest model iterated 100 times and then averaged as to avoid any model irregularities.

Figure 1 displays more clearly that higher priced items had sentiment scores which were centered on a neutral rating, with the treatment effect for items which had a combination of these clusters and buckets because of the larger disparity between high and low prices.[7] In figure 1 we omit the first bucket to avoid multicollinearity which would occur otherwise.

Unfortunately for our model, the conditional independence assumption likely does not hold, as we do not have access to enough control features to ensure that our data is as good as randomly distributed. Because of this, we can only concretely say that price and sentiment scores vary with each other by the average treatment effects listed above. Figure 1 shows that our treatment effects in each of our buckets are statistically significant.[7]

Conclusion

Our research shows that there is a statistically significant effect of overall review sentiment on product pricing, but it disproportionately affects those products that are on the cheaper end of the price spectrum. With more expensive items, the price tag itself seems to speak to the value of the item more than the marginal difference between, for example, a 4.2- and a 4.4-star review. Furthermore, reviews that contain more information tend to be more fair and neutral reviews, listing both pros and cons of the product, as opposed to a brief five-star review stating that the product was simply good. The presence of these neutral reviews suggests to the buyer that they can purchase with confidence, equipped with both the pros and the cons of the item.

Our model has some limitations, as we must introduce several sources of error to create our model. The products in each of our clusters are not perfect substitutes and sometimes may even be poor substitutes. Matching on Amazon subcategories and pricing as a proxy for substitutes is the closest, we could come without matching on a "similar item" value, but it is not perfect. Further error could have been introduced by our sentiment algorithm, as an endogenous source of data, given that our algorithm may have made mistakes.

[6] *Appendix:Table 1*

[7] *Appendix:Figure 1*

If we expanded our dataset by increasing our sample cluster size and decreasing price ranges, we could create more telling causal information. For further reading and analysis, *How Does the Variance of Product Rating* provides an analysis of review variance on sales.[8] This paper reports that those products that have a 4.1-star rating or lower have a higher-than-average standard deviation of sales. In terms of actual customer sentiment and review content, *"What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com"* analyzes what keywords, phrases, and other features of an Amazon review truly impacts the sales and price of a product, further expounding upon our finding that neutral reviews make for more informative reviews. [9]

[8] *Appendix: Figure 2*

[9] *Sun, Monic. "How does the variance of product ratings matter?." Management science 58, no. 4 (2012): 696-707.*

[10] *Mudambi, Susan M., and David Schuff. "Research note: What makes a helpful online review? A study of customer reviews on Amazon. com." MIS quarterly (2010): 185-200.*

Appendix

Table 1

Product Cluster	Average Treatment Effect (ATE)	Price Range ($)	Cluster Size (obs)
Cluster 1	0.0204	0.01 - 25.71	9,348
Cluster 2	0.0104	25.72 - 45.42	6,401
Cluster 3	0.012	45.43 - 83.59	2,987
Cluster 4	-0.0152	83.6 - 117.64	827
Cluster 5	0.0265	117.65 - 230.80	223
Cluster 6	-0.0299	230.81 - 864.78	75

Figure 1

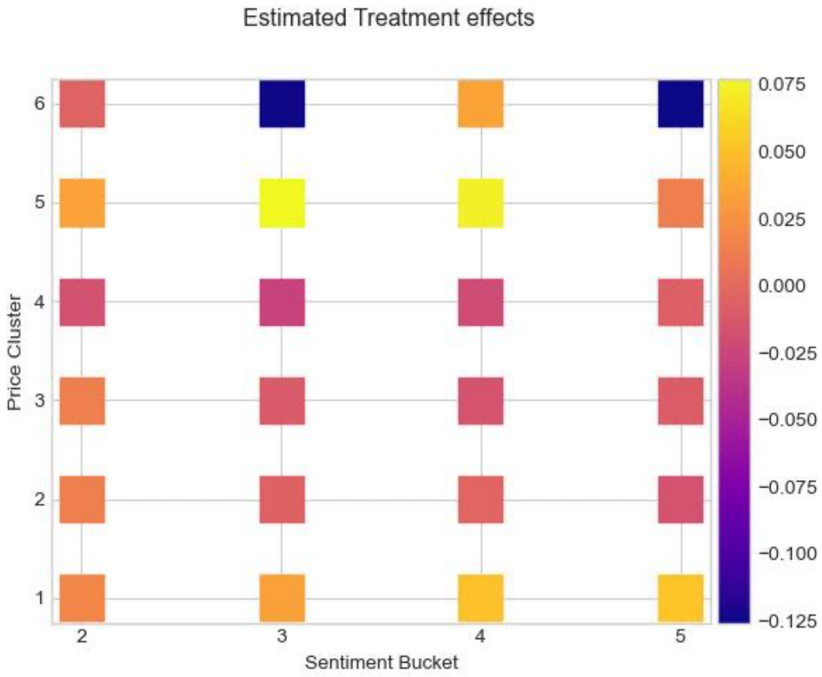
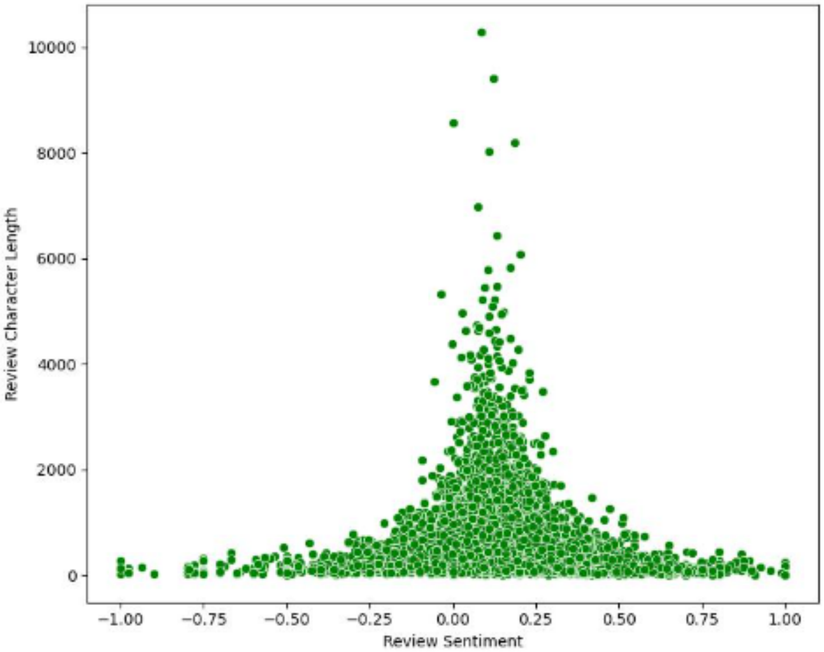


Figure 2


Review Character Length
10000
8000
6000
4000
2000
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Review Sentiment