Walmart Final Report

[redacted]

[redacted]

Northeastern University

DS 3000: Foundations of Data Science

Dr. Eric Gerber

[redacted]

Abstract

This paper examines potential biases in search suggestions and product rankings on Walmart's website. It aims to review whether Walmart prioritizes its own brands over others by examining searches and product specific factors. The analysis was collected from the Walmart website and utilizes Selenium to scrape and extract the searches from inputted all individual letters and then storing the automated result filled with the name of products, price, stars, reviews, whether it was sponsored, and its ranking in the overall search. The findings reveal that there is a bias for Walmart-sponsored items such that sponsored Walmart products get placed higher than non-sponsored ones and that Walmart sponsored products are rated higher than non-sponsored ones. This was further confirmed from the logarithmic regression model, which depicted that being a Walmart brand and then the product's rating most affect the probability that one product will be ranked higher than another in a product comparison.

Introduction

With Walmart and other major e-commerce businesses being warned about false endorsements and ratings on their products[1] and previously being sued for "greenwashing" some of its products materials and details[2], consumers and government have raised concerns on business malpractice and ensuring a fair market and platform for sellers. Therefore, as a business that aims to increase its profits no matter the length, one must consider how the interface of the website aids this. For example, when one goes onto Walmart's website to shop, what products come up as the recommended searches and whether they are Walmart brands, sponsors, or

[1] https://www.businessinsider.com/ftc-warns-amazon-walmart-hundreds-companies-fake-reviews-face-fines-2021-10

[2] https://www.ftc.gov/business-guidance/blog/2022/04/55-million-total-ftc-settlements-kohls-and-walmart-challenge-bamboo-and-eco-claims-shed-light

associated ones should be considered. The factors that mostly come into question are the name of the product, its price, the rating that is determined in number of stars, number of reviews, whether it is sponsored, and its ranking in search placement. This research paper analyzes the factors that determine product rankings in a search and, more specifically, whether Walmart unfairly places its sponsored products above non-sponsored ones.

The study focuses on three fundamental research questions:

1) Can and how do we determine if Walmart is prioritizing its sponsored products?
2) Do certain factors affect a product's placement on its website?
3) How can we predict what factors will make a product ranked higher than another?

To address these research questions, extraction, visualizing, and logarithmic regression models were employed from the dataset that was taken directly from Walmart's website. The data taken produces searches that are focused to Saugus, Massachusetts and had automated 260 suggestions to a product based on a letter, which ultimately allowed over 6000 products to be analyzed. By examining one product to another based on its features, there are more prioritized considerations in which one product might be placed higher in a search. Furthermore, this analysis will argue that Walmart indeed does unfairly prioritize its sponsored and high rated products. Overall, the findings in this study convey the need for better internal changes that Walmart needs to make in order to ensure fair practices for its third party sellers as well as any support that the FTC and other regulatory bodies can assist in.

Data Description (Link to our dataset[3])

We chose the features - Numbers of stars, if a product is sponsored, if a product belongs to a Walmart brand, the number of reviews/ratings, the rank of the product on the page, i. e. the position of the product on the page (eg. first product would have a rank of 1). The way we chose to determine if a product was from a Walmart brand is by checking the title of the product and seeing if it contained the name of one of the brands. We chose these features to not overfit our model and since it captures the main features that a person would consider before buying a product and its consequent placement on the page.

Pipeline Overview

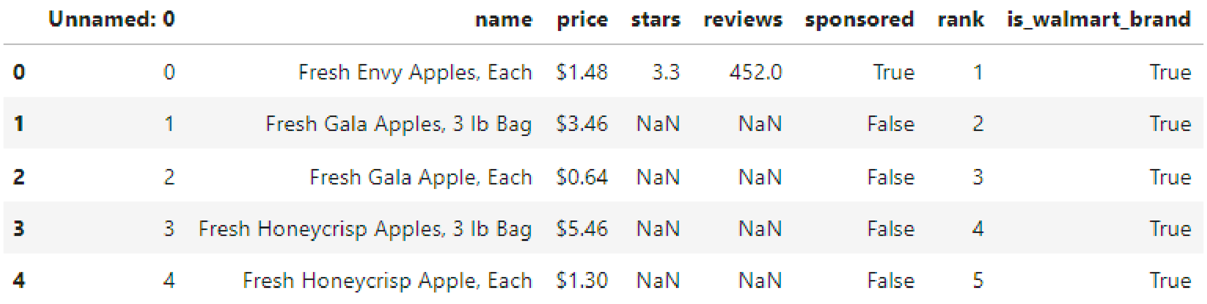
Suggestions Retrieval (Undocumented API)[4]

The Walmart data was scraped by generating product queries using a custom get_suggestions function, tapping into an undocumented API for autocomplete results. The function would search each letter from A to Z and would store all the autocomplete suggested results in a list.

Web Scraping (Headless browser using Selenium)[5]

To bypass Walmart's anti-scraping measures, Selenium was leveraged for web navigation and data extraction. Custom functions like open_browser(), load_walmart(), and search() facilitated this process. A compiled list of Walmart company names helps categorize scraped results. The script seamlessly integrates query generation and web scraping, systematically extracting details for each product on search pages. The collected information is stored in dictionaries, ensuring efficient retrieval despite Walmart's anti-scraping measures.

[3] https://github.com/yashjay1/DS3000-project/blob/main/extendedWalmartData_with_brand_updated.csv
[4] https://github.com/yashjay1/DS3000-project/blob/main/WalmartScraper.ipynb
[5] https://github.com/yashjay1/DS3000-project/blob/main/WalmartScraper.ipynb

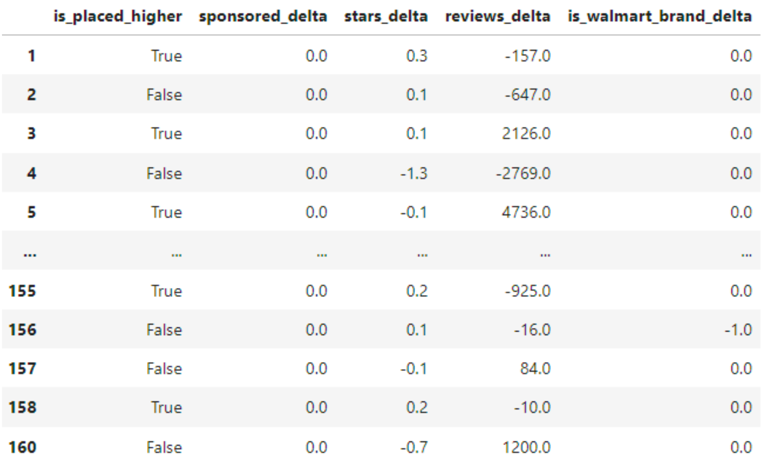


	Unnamed: 0	name	price	stars	reviews	sponsored	rank	is_walmart_brand
0	0	Fresh Envy Apples, Each	$1.48	3.3	452.0	True	1	True
1	1	Fresh Gala Apples, 3 lb Bag	$3.46	NaN	NaN	False	2	True
2	2	Fresh Gala Apple, Each	$0.64	NaN	NaN	False	3	True
3	3	Fresh Honeycrisp Apples, 3 lb Bag	$5.46	NaN	NaN	False	4	True
4	4	Fresh Honeycrisp Apple, Each	$1.30	NaN	NaN	False	5	True

Preview of the first 4 four rows of our dataset

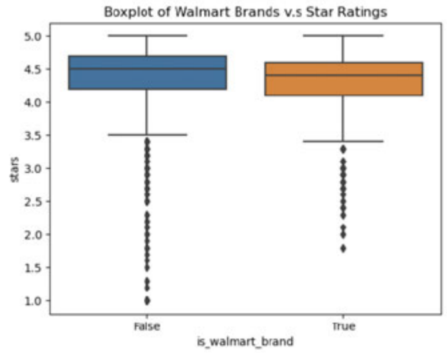
Preparing the data for our model

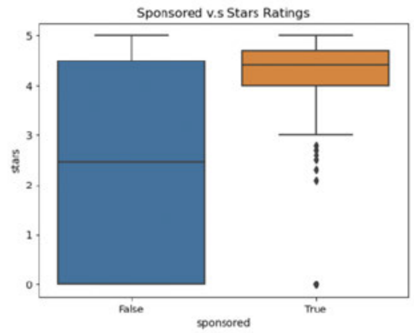
We created a new dataset from our old dataset that paired the top two products from each query together and contained the deltas for each feature as well as which product was placed higher. We trained the model using this dataset so we could use the model as a binary classifier.

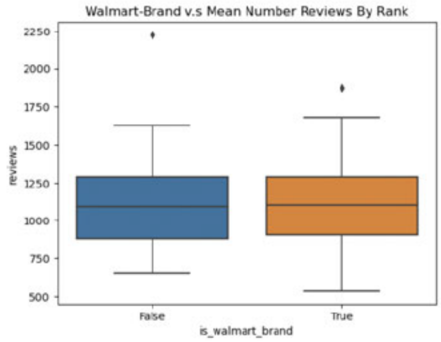


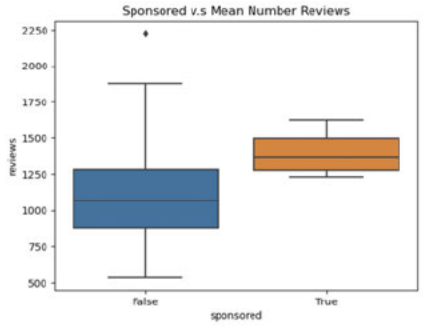
	is_placed_higher	sponsored_delta	stars_delta	reviews_delta	is_walmart_brand_delta
1	True	0.0	0.3	-157.0	0.0
2	False	0.0	0.1	-647.0	0.0
3	True	0.0	0.1	2126.0	0.0
4	False	0.0	-1.3	-2769.0	0.0
5	True	0.0	-0.1	4736.0	0.0
...	...	...	...	...	...
155	True	0.0	0.2	-925.0	0.0
156	False	0.0	0.1	-16.0	-1.0
157	False	0.0	-0.1	84.0	0.0
158	True	0.0	0.2	-10.0	0.0
160	False	0.0	-0.7	1200.0	0.0

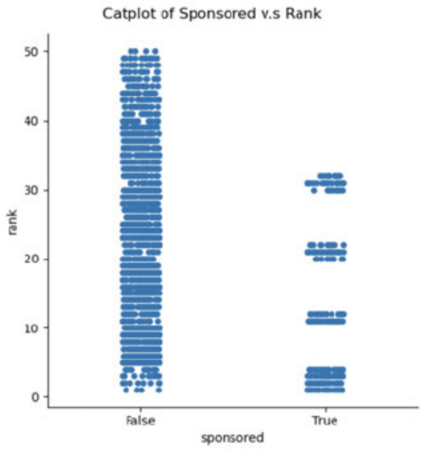
The dataset the model was trained on

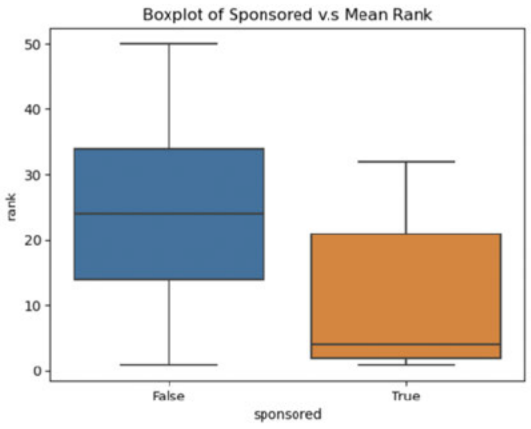












Data Visualization of the Collected Statistics

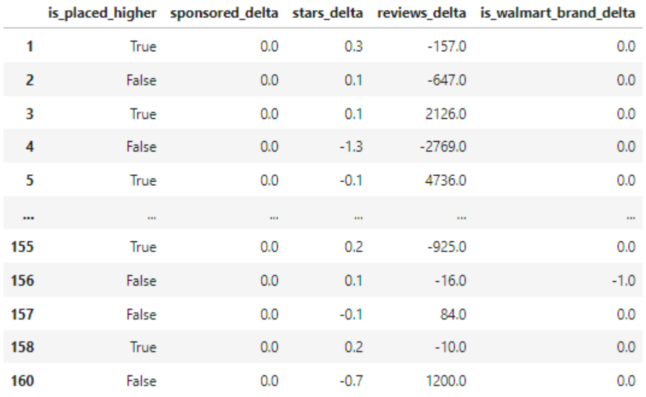
From the data collected, several plots were made to compare the statistics of Walmart-sponsored products to non-sponsored products and Walmart-branded to non-Walmart brands. We decided to look at the overall ranking comparison, star ratings, reviews, and pricing. It was found that sponsored items were placed significantly higher considering both the range and mean of the ranks on the webpage than non-sponsored ones, and specifically at the top of a new page. With similar popularity, sponsored products showed a higher mean rating than non-sponsored ones, which could have also been affected by the significant difference in the range between the two categories, however, Walmart-brand showed similar mean star ratings. The analysis supported that there is a bias for sponsored or Walmart items on the web page by putting them higher than non-sponsored or non-Walmart-brand ones, which becomes the basis for our machine learning predictions.

However, when drawing business insights from this graph, one should be cognizant of the underlying factors that could sway the mean number of reviews. A strategic analysis would benefit from cross-referencing these figures with sales data, product launches, and promotional campaigns to understand the full context. The graph underscores the importance of monitoring customer feedback across all ranks, as it plays a crucial role in shaping product reputation and influencing potential buyers.

Methodology

The chosen analytical approach used graphical representations and logistic regression, is appropriate for several reasons. Graphical representations, such as line graphs, provide a visual overview of data trends, helping in the identification of patterns and relationships. The inclusion of mean ratings corresponding to each rank further refines the analysis by incorporating a quantitative measure of product performance.

Logistic regression, as a binary classifier, is appropriate for this problem as it allows us to assess the impact of various features on a product's rank. The decision to focus on the top 2 products for logistic regression enhances the precision of the analysis, as it directly compares the most relevant products in terms of ranking. By organizing the data into pairs and introducing a new column indicating which product ranked higher, the model is trained to understand the factors influencing the product placement in Walmart's search results.



	is_placed_higher	sponsored_delta	stars_delta	reviews_delta	is_walmart_brand_delta
1	True	0.0	0.3	-157.0	0.0
2	False	0.0	0.1	-647.0	0.0
3	True	0.0	0.1	2126.0	0.0
4	False	0.0	-1.3	-2769.0	0.0
5	True	0.0	-0.1	4736.0	0.0
...	...	...	...	...	...
155	True	0.0	0.2	-925.0	0.0
156	False	0.0	0.1	-16.0	-1.0
157	False	0.0	-0.1	84.0	0.0
158	True	0.0	0.2	-10.0	0.0
160	False	0.0	-0.7	1200.0	0.0

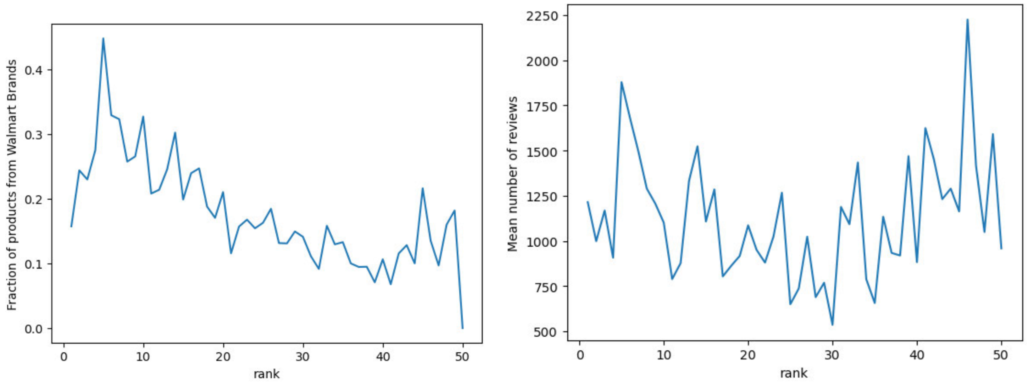
The dataset used for training the model that contains the pairs of products

The method's appropriateness lies in its ability to quantitatively analyze the impact of features on the probability of a product outranking another. Logistic regression provides a systematic and interpretable framework for understanding the relationships between features, making it suitable for uncovering the factors influencing product placement. We also used 80% of our dataset to train the model to give us accurate predictions.

However, there are assumptions and potential pitfalls associated with this method. One assumption is that the selected features adequately capture the factors influencing product ranking, and that these features exhibit a linear relationship with the log odds of a product outranking another. Additionally, logistic regression assumes independence of observations, which might be compromised if there are dependencies between products.

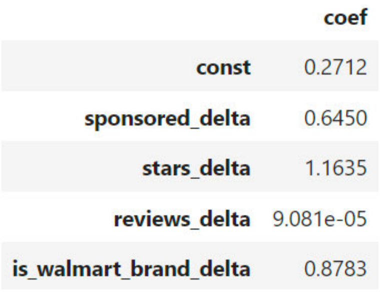
The choice to focus on the top 2 products may overlook nuances in the ranking of lower-positioned products, potentially missing insights into their performance and factors affecting their placement. Furthermore, the effectiveness of the logistic regression model is contingent on the quality and relevance of the features selected. Inaccurate or insufficiently relevant features could lead to biased model predictions.

Results



Graphs that show the relationship of features to ranks

After training our logistic regression model, we were able to extract the coefficients associated with each feature, shedding light on their respective impacts on product placement. The analysis revealed that the number of stars assigned to a product had the most significant influence on its placement, followed by whether the product belonged to a Walmart brand. Notably, being a sponsored product emerged as the third most influential factor, and the number of reviews came in as the fourth.



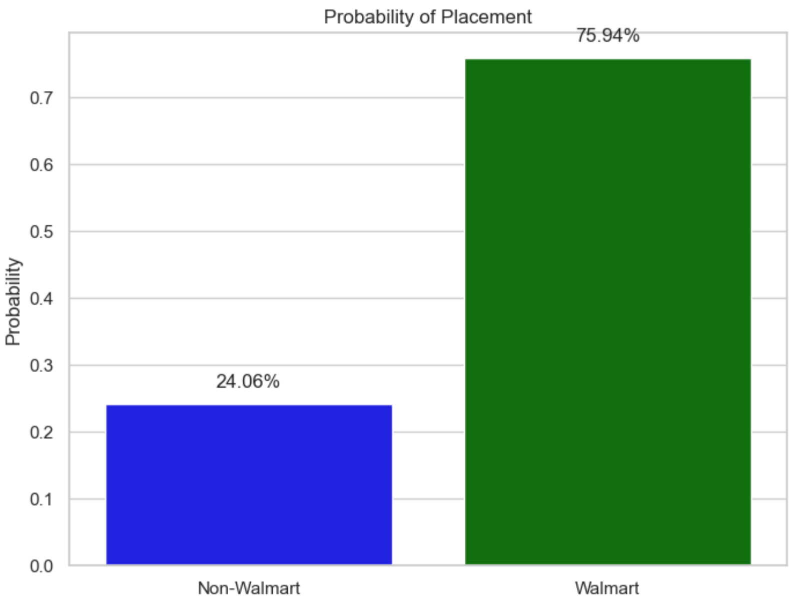
	coef
const	0.2712
sponsored_delta	0.6450
stars_delta	1.1635
reviews_delta	9.081e-05
is_walmart_brand_delta	0.8783

The coefficient of each feature

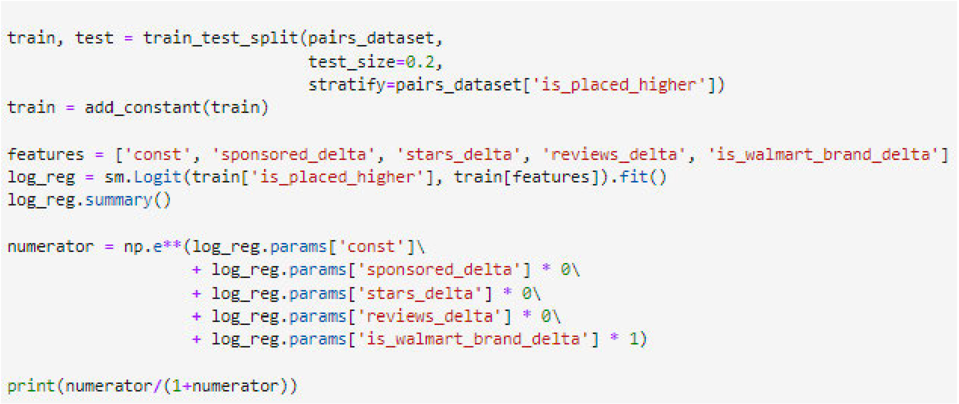
The hierarchy of effects, with the number of stars taking precedence, suggests that customer ratings play a pivotal role in determining product placement. This aligns with the common understanding that high-rated products are more likely to be prioritized in search results, emphasizing the importance of customer satisfaction in influencing placement.

The second most influential factor being a product's association with a Walmart brand underscores the advantage that such products enjoy in terms of placement. This finding implies that, even in the competitive landscape where companies might pay for sponsored positions, the intrinsic advantage of being a Walmart-branded product surpasses the impact of sponsorship.The revelation that being a sponsored product is the third most influential factor indicates that while sponsorship has a measurable impact, it does not surpass the advantages conferred by high customer ratings and Walmart brand affiliation. This insight is valuable for companies considering sponsored placements, as it suggests that the authenticity and quality reflected in ratings and brand association are paramount.

Lastly, the number of reviews was identified as the least influential among the factors considered. This suggests that, while reviews contribute to the overall assessment, they do not weigh as heavily in the product placement algorithm compared to factors like star ratings, brand affiliation, and sponsorship status.



The model's prediction that there is a 75.94% chance that a Walmart-branded product (product A) would be displayed higher than a non-Walmart product (product B) provides a quantitative estimate of the competitive advantage associated with being a Walmart-branded product. This probability serves as a useful metric for decision-makers aiming to understand the likelihood of product placement based on the identified features.



```
train, test = train_test_split(pairs_dataset,
                               test_size=0.2,
                               stratify=pairs_dataset['is_placed_higher'])
train = add_constant(train)

features = ['const', 'sponsored_delta', 'stars_delta', 'reviews_delta', 'is_walmart_brand_delta']
log_reg = sm.Logit(train['is_placed_higher'], train[features]).fit()
log_reg.summary()

numerator = np.e**(log_reg.params['const']\
                   + log_reg.params['sponsored_delta'] * 0\
                   + log_reg.params['stars_delta'] * 0\
                   + log_reg.params['reviews_delta'] * 0\
                   + log_reg.params['is_walmart_brand_delta'] * 1)

print(numerator/(1+numerator))
```

Code we used to generate the probability

The findings emphasize the dominance of customer ratings and Walmart brand affiliation in determining search result rankings, reinforcing the idea that these elements significantly impact consumers' online shopping experiences.

Discussion

The data visualization and ML analysis shows that aside from the stars rating of the products, there is bias for Walmart and sponsored items in their search webpage, showing the sponsored and Walmart items before the non-sponsored or non-branded and placing them in the beginning of each page. This is a marketing strategy from Walmart to push the customers to buy more from their brand, which presents unfair market competition that pushes brands to spend more money to sponsor Walmart. It's also important to note the location limitations as this dataset focused primarily on Saugus, Massachusetts, so any conclusions built are based off of that. Therefore, further research can be done to expand searches across the United States, which could be improved by employing several location services for Walmart. Other ways to improve on the model include using more complex observations instead of the binary classifier for the nuanced trends, examining more variables such as delivery speed or special offers that could affect the placement of the product on the web page. The overall context of unfairly placing products higher based on their sponsorship forces independent sellers into a position where it is much harder to compete. Therefore, it can be concluded that internal changes by Walmart need to be made such that there isn't an undue burden or an over-monopolized ranking where consumers are also forced into a position where they have to buy sponsored products.