Location, Location, Location: How Product Placement Impacts Clicks

DATASCI 203: Lab 2

Sophie Chance, Amy Zhang, Maureen Fromuth

Introduction

The importance of an effective online store has become increasingly critical in driving revenue with the growing relevance of e-commerce. The attention span of consumers is limited, as is the real estate on any given webpage. As such, it is important for e-commerce companies to understand the impact of various factors of product advertisement.

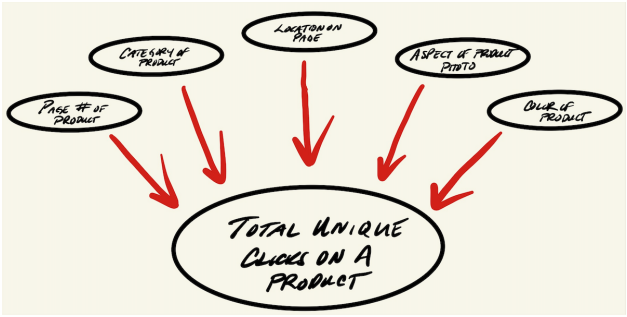

Figure 1: Causal Pathway for Clickstream Study

Investments have already been made in developing ranking algorithms to bring high conversion items to the top of each consumer's feed based on assumptions regarding the value of 'top of feed' or first seen items. However, there are still questions to be explored and answered more concretely in this space. As an online retail strategy analytics team, we believe that it is important for retail owners, such as our client maternity store, to understand the impact of location placement of a product on clicks generated. This information can be leveraged by retail owners and their supporting software team to leverage locations on webpages to increase clicks for priority products.

This study will use clickstream data with various features to evaluate whether or not the location of a product on a webpage has any statistical significance in altering the unique-per-session clicks the product receives off of a baseline location and page number. Applying a set of regression models, our team has determined the statistical significance and estimated the relative changes in number of clicks that result from the placement of an item on a webpage.

Given the languages of our client's customers are all right-to-left (RTL), the practical hypothesis of this test is that placing products in the top left position will result in greater number of total unique clicks per session. Statistically, the null hypothesis for this study is that the location of the product has no impact on the number of unique session clicks the product receives.

Description of Data

We will be leveraging a dataset containing observational information on clickstream from an online store offering clothing for pregnant women. The data is from five months of 2008. While this data is aged and e-commerce user experience has greatly improved in recent years, we believe that the general learnings around product placement and clicks generated will still hold.

The data represents 165,474 independent clicks on one of 217 different products. Each click identifies the month, day and year of the click as well as the online session ID for the user. The data provides details about the products to include category of the product, the specific color, and the price in U.S dollars. It also identifies details about the layout of the website such as the location of the product photo on the webpage, page number that the photo of the product, and whether the photo for the product is either profile or face on. The product information and website layout are static and do not change during the course of the data collection. As such, we selected the individual product as our unit of observation.

Operationalization of Key Concepts

To be able to analyze the number of clicks per product, we transformed the original dataset and grouped by products to return the total clicks per each of the 217 products. This study leveraged variables in this transformed of the primary dataset from which we selected variables. We chose the variable 'location of product' in the dataset to represent the treatment variable of location on the webpage. For the output

Table 1: Table 1: Estimated Regressions

	Output Variable: Unique Clicks per Product		
	(1)	(2)	(3)
Pos: Bottom Left	−0.005	−0.01	−0.01
	(0.02)	(0.02)	(0.02)
Pos: Bottom Middle	0.01	0.002	−0.002
	(0.02)	(0.01)	(0.02)
Pos: Bottom Right	−0.01	−0.02	−0.03
	(0.02)	(0.02)	(0.02)
Pos: Top Middle	−0.0003	−0.003	−0.02
	(0.02)	(0.02)	(0.02)
Pos: Top Right	−0.03	−0.03	−0.03
	(0.02)	(0.02)	(0.02)
Page 2		−0.11**	−0.12**
		(0.04)	(0.04)
Page 3		−0.09***	−0.09***
		(0.02)	(0.02)
Page 4		−0.08***	−0.07***
		(0.01)	(0.01)
Page 5		−0.07***	−0.07***
		(0.01)	(0.01)
Category: Blouse		0.04***	0.03*
		(0.01)	(0.02)
Category: Skirt		0.06**	0.07*
		(0.02)	(0.03)
Category: Trousers		0.07***	0.08***
		(0.01)	(0.02)
Price			−0.001
			(0.001)
Base: Top Left, Page 1, Sale	1.43***	1.45***	1.49***
	(0.01)	(0.01)	(0.03)
Product Color			✓
Model Aspect			✓
Price Higher than Average			✓
Observations	218	218	218
R^2	0.02	0.36	0.42
Residual Std. Error	0.09 (df = 212)	0.07 (df = 205)	0.07 (df = 189)

Note: HC_1 robust standard errors in parentheses.

variable, we selected the variable 'total clicks per product' to represent the total unique session clicks per product.

These variables match the concepts well in general, but some transformations were necessary to minimize the gap between the conceptual and operational definitions. Specialically, we renamed columns to enable readability and we also transformed numerical values to the named value (e.g. we transformed the value 1 in 'location on page' to Additionally, for 'total clicks per product', this study wanted to mitigate the potential impact of certain sessions where users clicked repeatedly on the same product and biased our results. Therefore, to control for potential dependence between clicks in a single session, we aggregated the sum of clicks for each individual product, but only counted one click per product per session. This control resulted in a removal of 13,058 total clicks. However, there may still be a gap to the ideal operationalization here, as the dataset does not allow us to ensure independence between sessions and clicks.

X or Y	Concept	Actual Feature Used
X	Location of Product	Location of Product - designated as a number 1-6
Y	Total Number of Clicks per Product	Summing clicks (count of rows) by product, counting only 1 click per session

While we considered other variables as our treatment variable, such as page number or price, we believed that the impact of the specific location on the page was less studied in this field, and therefore more valuable for our client to understand.

Explaining Key Modeling Decisions

We have split our data into training (30%) and confirmation (70%) sets at the individual session level.

To mitigate potential confounding variables and observations, we removed a portion of observations, applied transformations where helpful, and left out covariates that were not incremental to our model to prevent overfitting.

Recognizing that outliers are particularly impactful for linear models, and there may exist products that are particularly popular/not, we removed a total of 7 outliers from the exploratory dataset based on each product observation's Cook's distance.

Dataset	Element Changed	Amount Removed
Original Dataset (rows = clicks)	Repeat clicks counted on the same product within the same session	13,065 clicks removed
Exploratory Dataset, Grouped by Product (rows = products)	Outlier Products based on Total Unique Session Clicks per Product	7 products removed

During the initial EDA, we also observed skew in the data for the number of clicks. Using the Shapiro-Wilks Normality Test, we identified and applied a lambda transformation using Box Cox to the outcome variable to normalize the data.

Ultimately, the model will be based on the variables location on the page, page number, and category of the product. The baseline position for our model will be for a sale product that is located in the top-left corner of the page and on the first page of the website.

The other covariates were left out for a variety of reasons. We did not include the price of the product as the price is not displayed on the page, and therefore the consumer decision to click would be not impacted by the price of the product. We did not include the aspect of the photo as there is a negligible difference between the two options, and was therefore not incremental to include in the model. Lastly, we did not include color as there were 10+ unique color options, and including this variable would greatly increase the degrees of freedom for the model and contribute to potential overfitting.

Results

- Regression Table
- Demonstrate your specifications in a stargazer regression table
- Make it easy for the reader to find the coefficients that represent key effects near the top of the table
- Share the most appropriate standard errors
- Need to estimate at least three model specifications - first is the only the key variable you want to measure
- *Run this table on the confirmation data
- *Model 1-treatment & output, Model 2-model 1 plus covariates you think most important but no more than 5 covariate concepts (dummies not included), Model 3-kitchen sink
- Comment on statistical and practical significance, may want to look at something other than standard t-test
- Make clear to your audience the practical significance - how will Y/product change as a result of what discovered
- Answer if there are limits to the change we are proposing
- Answer what are the most important results and least important results discovered
- *On practical significance, refer back to your client/thesis
- *How does the results/significance compare to your thesis
- *Discuss the relevant material from the stargazer table - what surprised you, what did not
- *Interpret the coefficient of your treatment variable - do you think the client could/should use this to manipulate the process

Limitations

- Describe statistical consequences for any violations of the assumptions & strategies to mitigate consequences
- *Evaluate CLM assumptions even if claiming large sample; highlight which ones may cause problems and demonstrate you know where to start to improve your model
- Identify if there are any outcome variables on the right hand side - if so, provide the direction of bias this causes

This model is based on a large sample of data.

IID Data *probably most problematic* Currently, the dataset does not provide information as to the customer who is linked to each session and click. We partially accounted for independence concerns by deduping clicks that occur in the same session. However, we cannot guarantee that the observations are therefore independent as there is no tracking possibility across sessions. There is no evidence of autocorrelation in the data as shown by the Durbin-Watson test.

Linear Conditional Expectation *evidence of a non-linear relationship* The assumption of linearity holds; a graph of the residuals v fitted values indicates a random and consistent spread of data points around 0 on the x-axis.

No Perfect Collinearity *no strong evidence of collinearity* Variables are automatically dropped to avoid perfect collinearity

Homoskedastic Errors *no evidence of heteroscedastic errors* We performed a Breusch-Pagan test on our confirmation model. The p-value for this test was .71, and as such the model fails to reject the null hypothesis that there is no evidence for heteroscedastic error.

Normally distributed errors *evidence of non-normality in residuals* We performed several tests to include the Shapiro-Wilk normality test, which resulted in a p-value less than 0.05. This resulted in a rejection of the null hypothesis of normality in residuals. Similarly, further evaluation of the distribution of the residuals demonstrated in a left skew and a significantly high kurtosis at 16 signifying a Leptokurtic distribution.

Reverse causality There is low possibility for reverse causality for this study's model, as the outcome clicks does not affect the location of the product. This may however be different for e-commerce websites today, where sorting algorithms based on best-selling products may change the future location a product appears on the webpage (i.e. higher clicks would result in a higher position or page number for that product).

For structural limitations, our model may be biased by several omitted variables. In particular, this dataset does not provide the products' relative popularity or dates of availability on the website. For relative popularity, we are in particular concerned with trendy items that may see higher clicks regardless of its placement on page, page number, etc. We expect a positive correlation between product popularity and total clicks. Therefore expect a negative omitted variable bias on the key variables. The main effect is therefore driven towards zero, making our hypothesis tests underconfident. This is similarly true for the amount of time that a product is available to customers. It is unclear how much time each product is listed on the website for resulting in a positively correlated bias with total clicks. However, we do not believe this OVB calls the results of this study into question, as we have accounted for outliers that may be unduly influenced by variables such as popularity.

Updated causal path diagram: https://github.com/mids-w203/lab-1-stat-pack/blob/main/Lab%202/Lab2%20Causal%20Path%20Updated.jpg

Conclusion

In conclusion, this study found that product location on page had no statistically significant impact to total clicks a product received. However, the top-right location performed the worst, seeing X% fewer clicks. Rather, the page of the product was statistically relevant, with a negative correlation between page number and number of clicks. The category of the product was also significant, and sales products saw increased clicks whereas blouses saw lower clicks. The client should leverage these findings as they consider the page of the product as an actionable insight when determining ordering of priority products, and pay less attention to the specific location on the page.

For future research, it would be helpful to design an experiment or leverage a new dataset with more recent data, which should give a larger sample size and depth of data that will allow the research team greater flexibility in model building. For example, it would be helpful to have a dataset where the same product exists in multiple locations on a webpage to allow researchers to better control for product feature covariates. It would also be helpful to have additional data on the potential omitted variables mentioned, including popularity of products and amount of time available on the website. Lastly, it would be helpful to have a customer ID included in this new dataset to augment our model based on repeat clickers and even potentially customer demographics.

Sources

1.

2.

3.

4.

5.