“Snap Judgment : Predicting Snapchat Impressions for Political Ads Based on Spend”

George Godinez | Lucy Herr | Dhyuti Ramadas | Elias Tavarez

2023-04-19

Introduction

In the competitive landscape of political campaigns, effective outreach efforts are crucial for success. However, organizing effective social media campaigns can be a fiscal sinkhole if managed incorrectly[1, 2]. As a political outreach agency, our mission is to deliver the most effective and targeted social media advertising products to clients by building models which predict this outreach, hereby synonymized with “impressions”. The client will be a campaign manager for a political candidate who chooses to advertise on the social media platform Snapchat. The following research question is presented:

What effect does political ad spend have on the impressions a political ad receives on Snapchat?

We hypothesized that an increase in ad spending will lead to an increase in Snapchat impressions received by political ads. We also expect ad runtime to positively impact impressions, but these may have less detectable impacts in our dataset. The following figure encapsulates our thoughts on these and other variables (explained later in the text) affect the number of ad impressions:

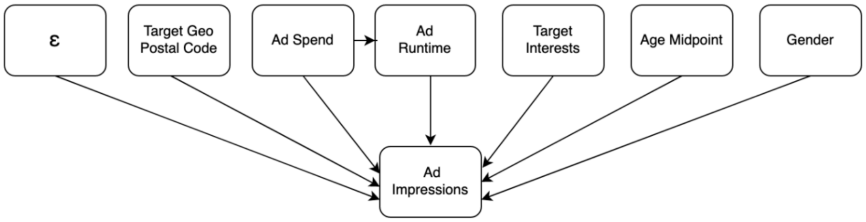

Figure 1: Initial Causal Pathway

Data and Methodology

The dataset comes from the Political Ads Library that was published by Snapchat and spans all geographic regions that used Snapchat for political advocacy in 2018 [3]. For each ad observation, the Snapchat dataset contains information about the total amount spent to date, the sponsoring organizations and paying advertisers, and the start and end dates of delivery on the platform. It also includes a number of variables describing Snapchat user targeting criteria, including users’ geographical region or interests.

The dataset was composed of 659 unique advertisements. Initially, the dataset was quite sparse on design. Missing values in the dataset indicated that no specific group was targeted and so we filled in all missing values with ‘ALL’ where specified (e.g., the gender column represented ads that targeted male, female, or all).

Since we decided to only focus on advertisements that were run in the US, we restricted the data to ads run in the US. Additionally, we chose to focus on the age range of 18-34 year olds because it makes up the majority of Snapchat users (roughly 82% in 2022)[4]. Lastly, we chose to remove any advertisements that did not have a listed end date. The following table describes the result of all these operations:

Table 1: Advertisements Removed in Data Cleaning

	Advertisements	Advertisements Removed
Raw Dataset	659	
US Only Filter	499	(160)
Ages 35+ Filter	494	(5)
Missing End Date Filter	437	(57)
Final Dataset	**437**	**(222)**

After having cleaned the data, exploratory data analysis was performed with 30% of the data which was randomly sampled from the dataset, while the final regression analysis was conducted with the remaining 70%.

Ad expenditure was operationalized using the Spend variable in the dataset, which indicates the amount (in U.S.D.) paid for the ad delivery on the Snapchat platform. To measure ads' reach among Snapchat viewers, we used the Impressions variable, which represents the total views received by an individual ad. Table 2 summarizes the variables from the dataset in our regression analysis:

Table 2: Variable Conceptualization

Raw Variable	Model Variable	Description
Impressions	Log (Impressions)	Total views received by an individual ad.
Spend	Log (Spend)	Amount (in U.S.D.) paid for ad delivery on Snapchat.
Run Time	Log (Run Time)	Period (measured in hours) which the ad ran for.
Gender	Gender	Indicates whether the ad targeted a gender - male, female, or all.
Age Bracket	Age Midpoint	The midpoint of the targeted age bracket.
Interests	Target Interests	Indicates whether the ad targeted specific interests or not.
Geo Postal Code	Target Geo Postal Code	Indicates whether the ad targeted geo postal codes or not.

We chose to take the natural log of Impressions, Spend, and Run Time since these variables were heavily right skewed. Interests and Geo Postal Code targeted by the ads ranged from long lists of values to empty fields indicating no target criteria used, so we chose to binarize these two for integration into our model. Finally, Age Bracket included overlapping and differently-sized age ranges, so we transformed these by taking the median (midpoint) year value.

We created three models which highlighted the variables we were most interested in. The first model examined only the effect of Spend on impressions. The second model incorporated the covariates we expected to affect Impressions the most: Run Time and Gender. We expected ads with greater Run Time and ads targeted to female voters to garner more impressions since female voters historically have a greater turnout in elections[5]. Lastly, the third model incorporated all covariates mentioned in Table 2. Our specifications are represented as follows:

$$ln(\widehat{impressions}) = \beta_0 + \beta_1 \cdot ln(spend) + \mathbf{Z} \; + \epsilon$$

where $\mathbf{Z}$ is a row vector of the additional covariates, γ is a column vector of coefficients and ϵ is the error rate in the model.

Results

The results of the three regression models can be found in Table 3. Throughout all three of our models, it is clear that Spend is the most statistically and practically significant variable in determining the number of impressions an ad receives. We arrived at this conclusion because as we add additional covariates, the Spend coefficient has only a small shift from 0.972 to 1.001, and the model's variance explainability (represented by R^2) only slightly increases.

Furthermore, we found that Run Time, Target - Geo Postal Code and Target - Interests were all statistically significant, while Gender (both male and female) and Age Midpoint were not. The lack of significance in Gender and Age Midpoint can most likely be attributed to the largely homogenous data we noticed in our exploratory data analysis.

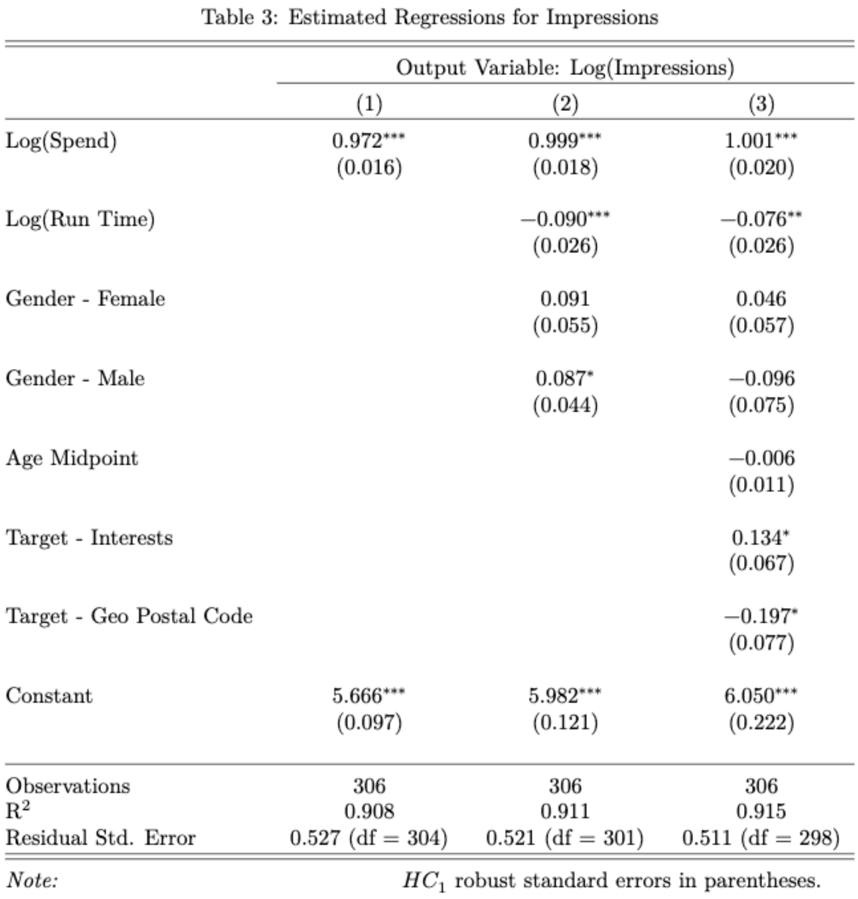

Table 3: Estimated Regressions for Impressions

	Output Variable: Log(Impressions)		
	(1)	(2)	(3)
Log(Spend)	0.972***	0.999***	1.001***
	(0.016)	(0.018)	(0.020)
Log(Run Time)		−0.090***	−0.076**
		(0.026)	(0.026)
Gender - Female		0.091	0.046
		(0.055)	(0.057)
Gender - Male		0.087*	−0.096
		(0.044)	(0.075)
Age Midpoint			−0.006
			(0.011)
Target - Interests			0.134*
			(0.067)
Target - Geo Postal Code			−0.197*
			(0.077)
Constant	5.666***	5.982***	6.050***
	(0.097)	(0.121)	(0.222)
Observations	306	306	306
R^2	0.908	0.911	0.915
Residual Std. Error	0.527 (df = 304)	0.521 (df = 301)	0.511 (df = 298)

Note: HC_1 robust standard errors in parentheses.

From a practical perspective for the clients who advertise on social media, we can describe the impact as follows: if we increase our spend by 20%, we expect that Impressions will also increase by roughly 20%*; but if we increase Run Time by 20%, we expect Impressions will decrease by 1.38%*. If we chose to target specific Geo Postal Codes, we may expect an increase of 14.34%* and if we target interests, we expect a decrease of 17.88%* in Impressions.

*ceteris paribus

Limitations

One possible statistical limitation in our model is the potential for clusters by ad sponsor or geographical region. While the majority of organizations and advertisers had single ads in the dataset, a few had 30 or more. Because of these clusters, we cannot conclude that the data points have been drawn independently of one another. We also lack sufficient information to determine whether subsets of ads from sponsors with certain agendas or budgets represent distinct populations. Regarding independence, we also note that the dates of the ads' delivery represent a possible issue; although our EDA indicated that the vast majority were delivered within a comparable time frame.

In addition, results of a Breush-Pagan test of the model were significant ($p<0.010$), indicating heteroskedasticity. As the assumption of constant error variance has not been met, the standard errors may be biased downwards.

Furthermore, visual inspection suggested deviations from normality in the model residuals, especially in its tails. As this distribution is highly leptokurtic (kurtosis=4.87) and moderately negatively skewed (skewness= -0.56), it's possible that our results are misleading since non-normality may lead to inaccurate test statistics.

We also considered the bias of two omitted variables on our model. The first is a binary variable for whether the ad ran during peak hours for Snapchat usage. The relationship between this variable, Peak Hours, and the measured variables is likely to be positive because ad agencies may increase Spend to advertise during a peak period. The relationship between Peak Hours and Impressions is likely positive as more users will be on Snapchat during this time. This results in a positive bias and βr > βc. Since βc is positive, the bias will move away from 0.

The relationship between the second omitted variable, Ad Length, and the measured variables is likely to be positive because ad agencies may increase Spend for longer ads. The relationship between Ad Length and Impressions is negative because it is likely that the ad will not be displayed as frequently to users if it is lengthy. This results in a negative bias and βr < c. Since βc is positive, the bias will move towards 0.

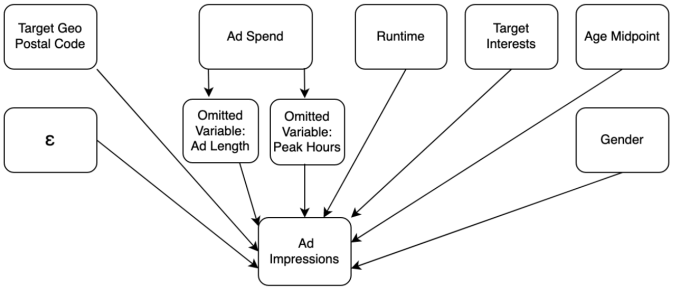

Figure 2: Final Causal Pathway

Conclusion

This study demonstrated that Spend has a strong positive linear relationship with Impressions on Snapchat. Incorporating more variables led to more accurate predictions of ad impressions, despite some factors such as Gender and Age Midpoint not showing significant influence on the outcome. An important consideration is that these models are not infallible and may require updates and adjustments as Snapchat's algorithms evolve to maintain their relevance and usefulness. Furthermore, it is important to recognize that Impressions may not be a valid metric for predicting votes. Our models however may be used as a foundation for further research an ad agency could pursue.

References

1. Ifeanyi, K. C. (2022, June 24). Inside the good, bad, and very ugly of social media algorithms. Fast Company. https://www.fastcompany.com/90761087/inside-the-good-bad-and-very-ugly-of-social-media-algorithms

2. Stoffer, M. (2018). The ROI of social media [infographic]. mdgSolutions. https://www.mdgsolutions.com/learn-about-multi-location-marketing/infographic-the-roi-of-social-media-2/#:~:text=You're%20Not%20Alone.,to%20quantify%20social%20media's%20success

3. Snap Inc. (n.d.). Snap Political Ads Library. Retrieved April 11, 2023 from https://www.snap.com/en-US/political-ads

4. Dixon, S. (2022, August 2). Distribution of Snapchat users worldwide as of January 2022, by age and gender. Statista. https://www.statista.com/statistics/933948/snapchat-global-user-age-distribution/

5. Rutgers Eagleton Institute of Politics. (n.d.). Gender Differences In Voter Turnout. Retrieved April 19, 2023 from https://cawp.rutgers.edu/facts/voters/gender-differences-voter-turnout