Success Factor in Social Media Advertisements

Nicole Aguilera

Bellevue University

Introduction

Advertisements are a very controversial subject when it comes to social media. This goes along with who is benefitting from the ads and how often they come up on a feed. Businesses use data from social media sites to correctly target their audiences for their sales tactics. For this reason, I am researching how much of these advertisements are getting clicked on/purchases. These datasets look at the advertisements that arise on different social media sites; some look at only Facebook, while others do not specify which sites are being looked at. I would like to see how this data on advertisements affect the public on social media sites, and what it takes for engagement to happen. Some main questions that are being looked into are: What percentage of ads get clicked on/purchases? What factor(s) contributes to an advertisement getting more views/clicks? Do demographics indicate whether a person will choose to click the ad/buy the product? The analysis will mainly look at how an advertisement gets clicked on or items get purchased based on that ad. It will look at the relationship between variables and the variables with the factor of whether the user continues further with the ad. Some of the packages that I intend to use include but are not limited to ggplot2 for visuals, and readcsv to obtain csv files. Some of the plots and tables needed include but are not limited to scatterplots, regressions, and residuals.

Analysis

I have imported my data off of Kaggle to my local computer and have only imported the csv versions in order to keep consistency when downloading them to Rstudio. I know that I must make sure that there are no holes in the data, because errors would be produced, and data would not be constructed correctly if there are holes. I first started off by cleaning the data. I got rid of some variables in the datasets that I determined were irrelevant in conducting my research. One

thing that I regret to have done but was necessary in continuing with my project in a timely and

efficient fashion, was that I had to eliminate one of my data sets. This was the dataset that

contained data as to whether or not a Facebook ad was political or not. It had over 350,000

entries and was ultimately slowing down Rstudio so much, that I was not able to use it. Because I

will be looking at all of data sets individually due to the differing variables, this will not impact

any of my other data sets nor their significance in my research. Both data sets had a gender

variable, which needed to be converted to binary numbers in order to be measured in R. The final

outcome of the remaining two datasets after cleaning resulted in five variables in the dataset that

gives some demographics about the user and then indicates if they have purchased anything from

the site (df_ad_interaction), and six variables in the data set that looks at a Facebook ad

campaign (df_fb_campaign).

I began running some basic exploratory analysis on the data sets to gain a bit of structure

to see where I need to go with this project. The df_ad_interaction dataset had about equal

genders in their study (found via histogram), compared to those of df_fb_campaign, which is a

more accurate representation of a sample size, creating less bias. In answering one of my

questions of the percentage of ads get clicked on or purchases, I did some simple arithmetic

using Rstudio. I found that the total percent of advertisements that subsequently received

purchases in the df_ad_interaction data set was about 35.75%. The percentage of ads that get

clicked on in the df_fb_campaign data set was a bit tricky in that some ads were clicked on more

than once by a person, and some people did not click on any ads at all. With keeping that in

mind, I calculated the percentage of ads that were clicked on and it came out to be a large

percentage of over 1000%, because of how often they were clicked on compared to the number

of participants there were. The average amount of money spent by each person in the

df_fb_campaign data set was about $17.60.

I then moved on to running regressions on each of the data sets in order to address the question of looking at which factor(s) contributes to an advertisement getting more views, clicks or purchases. Based on the regression of the df_ad_interaction data set, all of the variables came up to be significant and strongly contributing to the reasons as to why an advertisement gets more purchases, in this case, except the GenderBinary variable (a new variable I created after converting the gender variable to binary). This is interesting to me because I would have assumed that there would be some significance, due to the somewhat stereotype of females online shopping, but in this regression, that does not seem to be the case. The R2 value for this data set is 0.4599, which seems to be pretty good in this scenario. When looking at the regression of the df_fb_campaign regression, it shows that all of the variables come up to be very significant in the impact that it has on the dependent variable of clicks. So, with this regression, it seems like all of the variables impact the number of clicks an advertisement gets. The R2 number is also very high at 0.9915. Because this value is so high, it may not necessarily be a great thing.

The only limitations I encountered while conducting my analysis of the data in the sets that I have acquired is that I was unable to condense the age variable in the df_fb_campaign data set because of how many increments of age brackets there were. I do not know how exactly this variable was measured, seeing that there were some age values upwards of 110. Because of this, I decided to exclude this variable from my regression, due to the fact that it was not going to be a significant factor towards my end product. I am unsure if this produced any bias or skewed the results of my other variables, but it seemed better to leave out.

Conclusions

Working through the analysis of these data sets proved some real results to the questions that I had originally called attention to. It was found through one of the data sets that about 40%

of advertisements on social media sites had the interaction of purchases, while about 60% of

advertisements did not receive purchases. The demographics of the estimated salary and age

through the df_ad_interaction data set were significant factors that contributed to the interaction

of the ad, while the gender variable was not. All of the variables contained in the

df_fb_campaign data set proved to be significant factors in contributing to the number of clicks

that an advertisement receives. Because the adjusted R2 came out to be around 99%, I am not

sure how accurate this data set was in telling me the results that I was looking for; however, for

the purposes of this project, I will accept this as a positive result for the time being.

References

Kaggle. (n.d.). Facebook Ad Campaign. Retrieved from Kaggle:

      https://www.kaggle.com/madislemsalu/facebook-ad-campaign/data

Kaggle. (n.d.). Social Network Ads. Retrieved from Kaggle:

      https://www.kaggle.com/rishabhsingh98/social-network-ads