Executive Summary

Predictive Modeling for Marketing Effectiveness with Multiple Linear Regression

David Harvell

Master of Science, Data Analytics Program

Data Analytics Graduate Capstone – D214

January 2022

A. Statement of the Problem and Hypothesis

We will try to answer the question "How much does the frequency of ads impact the probability of purchase".

There is a staggering amount of money spent on advertising, and there are constant questions regarding the effectiveness and return on investment. Some of the world's largest websites (Facebook, Google, YouTube, etc.) primarily collect revenue through advertising. Because of this, our focus will be online/digital advertising. We are only now beginning to brand that have truly grown recognition online, whereas television advertising has been building brands for over 50 years.

Our assumption is that a higher frequency of ads will impact the probability of purchase with statistical significance. Because of this, our null hypothesis with be that total ad count has NO impact on the probability of conversion. We will test our hypothesis against a Marketing A/B test dataset.

B. Summary of the Data Analysis Process

A dataset of Marketing A/B results has been downloaded from Kaggle at the location https://www.kaggle.com/faviovaz/marketing-ab-testing. This data contains a record per user, whether they were shown ads or PSAs, how many ads, and whether they were converted to a purchase. There are over 500K records in a 22MB CSV file. This dataset was very clean – containing no null records or extreme outliers.

Our analysis began with visual inspections of the data, and then followed by dropping unnecessary identity columns. Next, we created a new variable that describes the rate of

conversion and charted it against the number of ads shown. This seemed to trend in a positive manner.

Before starting our model, we converted categorical text values into dummy integer values, to allow the values to be utilized in Multiple Linear Regression. Next, we began to investigate correlation. Slight correlation is shown between the number of ads and conversion, but other variables show negligible correlation.

Finally, we create the Multiple Linear Regression (MLR) model. After evaluating all variable, we removed inconsequential columns (day and hour with most ads shown) and rebuilt the MRL model.

C. Outline of Findings

OLS Regression Results

Dep. Variable:	converted_d	**R-squared:**	0.047
Model:	OLS	**Adj. R-squared:**	0.047
Method:	Least Squares	**F-statistic:**	1.462e+04
Date:	Thu, 06 Jan 2022	**Prob (F-statistic):**	0.00
Time:	12:34:54	**Log-Likelihood:**	2.6923e+05
No. Observations:	588101	**AIC:**	-5.384e+05
Df Residuals:	588098	**BIC:**	-5.384e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0007	0.001	-0.679	0.497	-0.003	0.001
total ads	1.6101	0.009	170.826	0.000	1.592	1.629
group_d	0.0076	0.001	7.503	0.000	0.006	0.010

The R-Squared value of 0.047 (explaining 4.7% of the variance in conversion) is very low and indicates that the model will not product accurate results. The F value and degrees of freedom result in a P-value of less than 0.001 that the F value exceeds 1.462e+04. This infers that we should reject our null hypothesis. This is further reinforced with the P-values of 0 (or essentially 0) for the independent variables of "total ads" and "group_d" (shown ads or PSAs).

Since we are rejecting our null hypothesis, that means we should accept the alternative hypothesis that the number of ads and being shown ads vs PSAs impact the success of conversion in a statistically significant manner.

D. Limitations of the Techniques and Tools Used

Our toolset consists of Python and a notebook system (Google Colab). One disadvantage/limitation of this toolset is the speed at which large amounts of data can be processed. If we intend to move our process into production, we will need to look for ways to improve the processing speed.

The model was built using Multiple Linear Regression. Although this is a trusted method, it can sometimes lead analysts to sometimes confuse correlation with causation. However, most problems with MLR are usually due to issues in the data rather than the method itself.

E. Summary of Proposed Actions

I would end the analysis with a suggestion to collect more data and expand the variable set for the model. There are likely other variables that could be inserted to potentially arrive at a stronger R-squared value, including targeted audiences.

After revieing the model summary, it does appear as though advertising statistically affects convers, so I would suggest some budget is allocated to advertising. Finding the accurate ROI and exact portion of budget that should be allocated will require a strong model and deeper analysis.

F. Expected Benefits of the Study

This study reinforced the impact of online advertising on sales. Allocated money towards advertising should result in an increase in sales – accounting for 4-5% of the increased sales. This study should also reinforce a proposal to investigate a deeper predictive model to investigate ROI. We might be able to prove a higher correlation if ads are targeted to a specific, relevant audience.