Predicting Engagement in Online Advertisement from Consumer Features

Introduction to Research Question

The purpose of this analysis is to identify what consumer features predict engagement in online advertisements through ad-clicks. Consumer features may include time spent on site, time spent on the internet as well as descriptive consumer information.

Sales and marketing effectiveness can help businesses consider where to allocate resources in order to create and streamline strategies that may allow for consumer engagement in advertisements and increases product sales and market share.

The analysis of effective marketing can allow for an advertiser to understand the demographics and behavior of consumers interested in their advertisement and allow for targeted marketing strategies.

Methods

Sample

The sample included N=1000 observations of different demographics and from varying countries. The data set was obtained from Kaggle as a project to predict who is likely to click on the advertisement.

Measures

The ad-clicks were measured through a binary response with either yes or no. The consumer features involved descriptors including 1) daily time spent on a site in minutes 2) age of the customer in years 3) average geographical area income of the customer 4) daily internet usage in minutes, and 5) Whether or not the consumer was male.

Analyses

The variables were examined through descriptive statistics that were provided by models which included means, standard deviations and maximums and minimums. Each variable feature was measured against the clicked response in order to determine statistical significance through logistical regressions and Chi-Squared tests of independence, including controlling for confounding variables.

A random forest was generated with 60% (n=600) selected for the bagging process and used to model the most important features that caused an ad-click by consumers. The Gini index was used as the split criterion with a maximum number of 100 trees. Fit statistics were assessed to identify any misclassification rates and average square errors. The variable importance tables produced rankings by split criterion with both training and test indexes given.

A LASSO regression was used in comparison to the random forest to measure how each analysis would rank variables. The LASSO regression was used for an explanation of the event against the random forest which was used for predictions.

Descriptive Statistics

The following tables are descriptive statistics of consumer features with their respective quantitative figures. The average number of ad clicks was 500 of 1000 observations (sd= 50%).

Figure1. Descriptive Statistics

Variable: DTSS (Daily time spent on site)

Moments			
N	1000	Sum Weights	1000
Mean	65.0002	Sum Observations	65000.2
Std Deviation	15.8536146	Variance	251.337095

Variable: Internet_Usage (Daily internet usage)

Moments			
N	1000	Sum Weights	1000
Mean	180.0001	Sum Observations	180000.1
Std Deviation	43.9023393	Variance	1927.4154

Variable: AreaIncome

Moments			
N	1000	Sum Weights	1000
Mean	55000.0001	Sum Observations	55000000.1
Std Deviation	13414.634	Variance	179952406

Variable: Male

Moments			
N	1000	Sum Weights	1000
Mean	0.481	Sum Observations	481
Std Deviation	0.49988888	Variance	0.24988889

Variable: Age

Moments			
N	1000	Sum Weights	1000
Mean	36.009	Sum Observations	36009
Std Deviation	8.78556231	Variance	77.1861051

Bivariate Analyses

Individual Chi- squared testing revealed significant p-values signifying statistical significance for all variables except gender. In order to keep the Chi-Squared test valid, quantitative variables were grouped into categories of no more than five. Logistical regressions confirm that all variables remain significant when individually analyzed and when each feature was controlled for each other (table 1). Gender was included and found to not be significant.

Table 1.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	27.3606	2.7365	99.9711	<.0001
Internet_Usage	1	-0.0635	0.00676	88.1732	<.0001
Male	1	-0.4217	0.4043	1.0876	0.2970
Age	1	0.1709	0.0259	43.6585	<.0001
AreaIncome	1	-0.00014	0.000019	52.4875	<.0001
DTSS	1	-0.1927	0.0208	86.2388	<.0001

For predictive statistics, a random forest was created to generate a list of variable importance (table 2.) Baseline fit statistics showed an average square error of 25% a misclassification rate of 50% and a Log Loss of 69%.

Table 2.

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
DTSS	920	0.176682	0.15024	0.353364	0.326631
Internet_Usage	763	0.142460	0.12515	0.284920	0.268201
Age	875	0.072847	0.04455	0.145695	0.117373
AreaIncome	2313	0.107055	0.02598	0.214110	0.132818
Male	39	0.000536	-0.00038	0.001071	0.000087

For descriptive statistics a LASSO regression was used although its methods are used less at predicting and more so for explaining ad-clicks. The standardized coefficients for the regression reveal a different order of variable importance.

Figure 2.

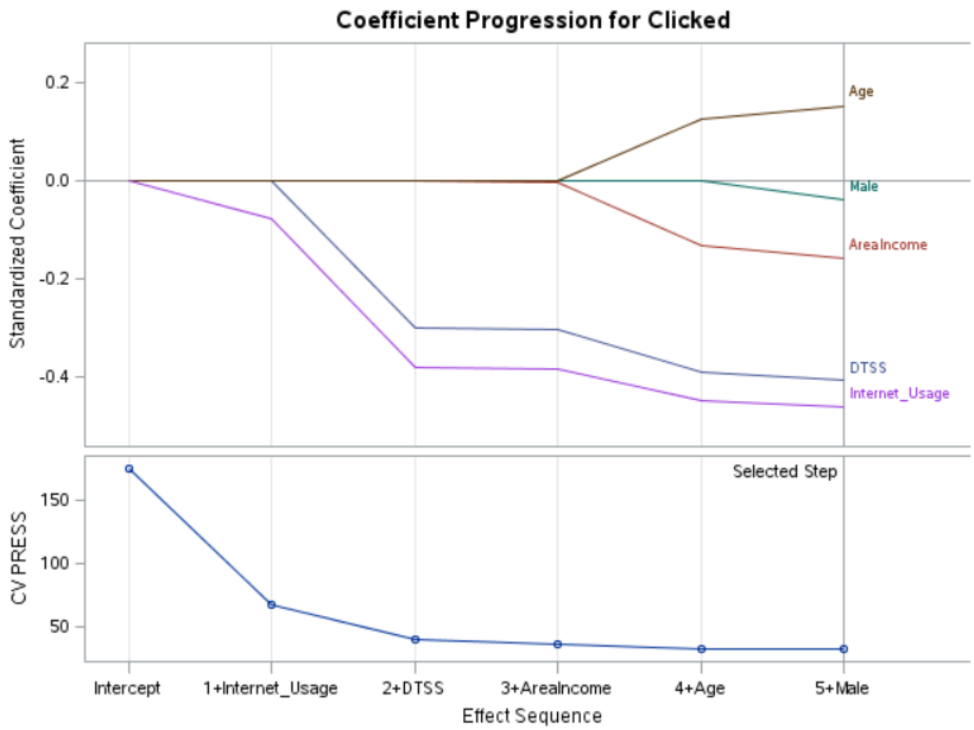
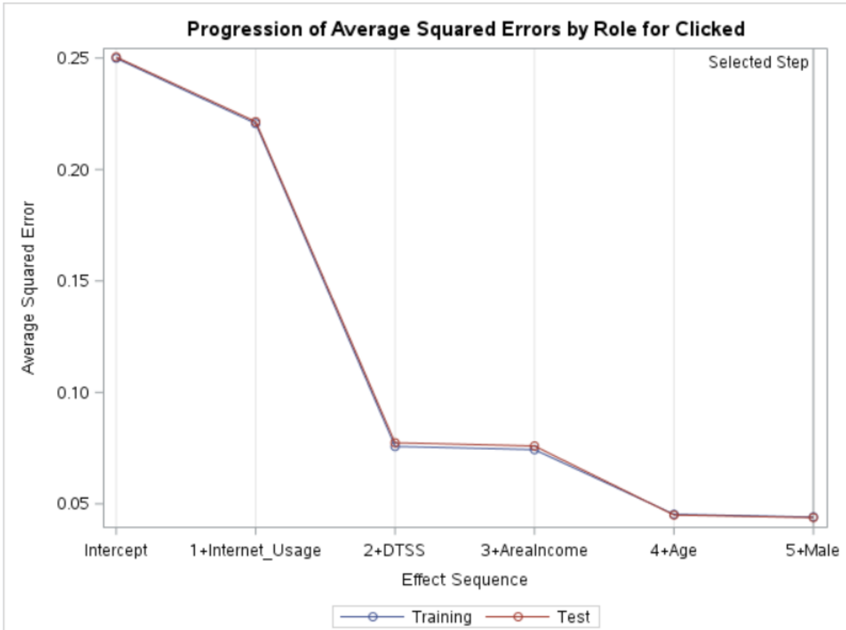


Table 3.

LAR Selection Summary					
Step	Effect Entered	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept	1	0.2499	0.2504	175.5740
1	Internet_Usage	2	0.2206	0.2215	66.7319
2	DTSS	3	0.0758	0.0774	38.8114
3	AreaIncome	4	0.0743	0.0760	35.0163
4	Age	5	0.0453	0.0450	31.7946
5	Male	6	0.0441	0.0438	31.6507*
* Optimal Value of Criterion					

The order of variable importance is distinct for each test. The top two variables are switched (internet usage and time on site) as well as third and fourth for each (age and area income). The final variable is gender for both tests as it was included but found to be not significant in earlier analysis. The LASSO criterion was 70% for training and 30% for testing.

Figure 3.



Conclusion/ Limitations

The project used random forest machine learning to predict the best variable indicator of engagement with an online advertisement. The engagement was measured in clicks where the total number of observations N=1000 were either a click or not. The variability was split evenly with half the observations showing users to have clicked and half showing no engagement with the advertisement.

The random forest generated a list of variable importance that placed the time spent on the website with the advertisement to be of greatest significance for predictive purposes. As part of the model gender was included with the other 4 variables although in preliminary analysis it was deemed not statistically significance. The internet usage variable, measured in minutes was given by the model to be the second most important variable. Then list of variable importance was finalized with age and area income respectively.

A lasso regression model was used to compare how variable selection was similar. The selection showed that as a descriptor of an event compared to a predictor of a future event, the lasso regression listed the internet usage as the best descriptor of the clicked event. Listed second was the time spent on site and finally area income with age. The models indicate that the time invested online more so than any descriptor of the user would impact the likelihood of engagement. The implications are that advertisements placed on sites that users are spending the most time on should be a target for marketers and companies. Survey data drawn from the observations would allow for a more in depth look of consumer behavior to single out other factors involved in ad engagement.

The random forest model predicted the variables by importance with a misclassification rate of 50% for baseline fit statistics. The forest model correctly classified the other 50% of the sample. There was a significant drop of in the out of bag estimates with regards to the lower ranking variables which was also seen in the lasso regression model's average square errors by role for clicks. To lower the misclassification rate in future models other variables may be necessary to predict the best variable for engagement. It may be that intangible variables or other confounding variables may have contributed to a weaker misclassification rate than desired as well. Finally, the advertisements in themselves and their effectiveness may play a role in consumer engagement. Given that users are interested in a product that they may find worth engaging with may contribute to the consideration of an ad-click.