Star Digital Team Assignment2_Team14_SecA

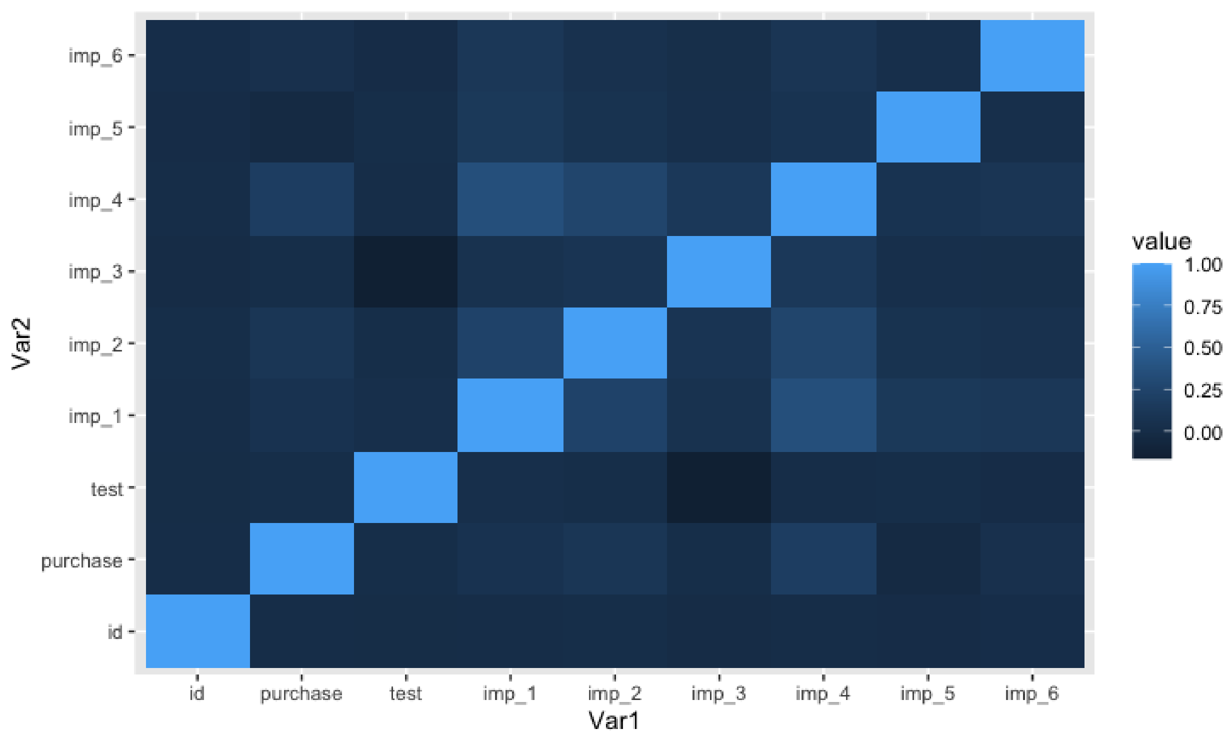
Garrett Chaffey, Chih Han Chi, Sammi Chueh, Pratheek Kumar

1. Is online advertising effective for Star Digital? In other words, is there a difference in conversion rate between the treatment and control groups?

Before we do analysis, we do a t-test to know the difference between the control group and the test group. First of all, we do the t-test of total impression, and the result shows that there are no significant differences and the p-value is large, which implies that this is not statistically significant.

Secondly, we do the t-test of purchase. As a result, the p value is 0.06, which is marginally significant. This shows us that there might be a difference between the control and the test. Therefore, we are calculating effect size and lift. The effect size is 0.383, which shows us that there is moderate difference between control and test. Moreover, the lift ratio also tells us that there is an approximately 3.95% increase.

Additionally, we conducted a correlation check for each variable before we do the regression, and there are no variables that are highly correlated to each other.



When looking at the outputs of a model examining the relationship between purchase (dependent) and test (independent) we can see from the p-value of “test” that the variable itself is not significant or not <u>strongly</u> correlated with the dependent variable (ie the p-value is greater than 0.05) however when looking at the slope coefficient we do see that it has a positive slope; meaning that for every unit increase of “test” we see an increase on the dependent variable (ie purchase).

```
> summary(ad.effect)

Call:
glm(formula = purchase ~ test, family = binomial(), data = dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.186  -1.186   1.169   1.169   1.202

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05724    0.03882  -1.474   0.1404
test         0.07676    0.04104   1.871   0.0614 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Is there a frequency effect of advertising on purchase? In particular, the question is whether increasing the frequency of advertising (number of impressions) increases the probability of purchase?

Yes there is, we can look at both the relationship between total number of impressions across all sites and we can look at the individual site impressions to examine the relationship between number of impressions and purchasing behavior.

```
> summary(fit.full)

Call:
glm(formula = dat$purchase ~ dat$Total_imp + dat$Total_imp_test,
    family = binomial(), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9091  -1.1272   0.1306   1.2150   1.2485

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -0.181875   0.014584 -12.471 < 0.0000000000000002 ***
dat$Total_imp       0.016228   0.002676   6.065        0.00000000132 ***
dat$Total_imp_test  0.015055   0.002930   5.139        0.00000027632 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(coef(fit.full))
       (Intercept)       dat$Total_imp dat$Total_imp_test
         0.8337053           1.0163605          1.0151687
```

(above is a picture of the model output modeling the combined number of impressions against purchasing likelihood)

```
> summary(full_imp_model)

Call:
glm(formula = purchase ~ imp_1 + imp_2 + imp_3 + imp_4 + imp_5 +
    imp_6, family = binomial(), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.2575  -1.1002   0.0125   1.2381   3.8323

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.200420   0.014696 -13.638 < 0.0000000000000002 ***
imp_1       -0.006101   0.003738  -1.632               0.1026
imp_2        0.014727   0.001760   8.366 < 0.0000000000000002 ***
imp_3       -0.040928   0.016002  -2.558               0.0105 *
imp_4        0.179939   0.006986  25.757 < 0.0000000000000002 ***
imp_5       -0.449888   0.048144  -9.345 < 0.0000000000000002 ***
imp_6        0.015993   0.003013   5.308          0.000000111 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(coef(full_imp_model))
(Intercept)       imp_1       imp_2       imp_3       imp_4       imp_5       imp_6
  0.8183872   0.9939171   1.0148357   0.9598983   1.1971444   0.6376999   1.0161214
```

(the picture above goes into more detail regarding the relationship each site has with impressions and likelihood of purchase)

In the first picture we can see that there are all "general" positive correlations. The "sum_imp" variable represents the row sums of impressions for sites one through six and is both significant (being less than 0.05) and has a positive slope meaning that overall, for every impression increase there is an increase of approximately 1.64% in the odds of purchase. The "sumimp_test" variable shows the sum of impressions for the test group. It shows that being a test group member will increase approximately 1.52% in the odds of purchase.
Taking a look at the second picture and going into more detail regarding each individual site, we can see that imp_1, imp_3, and imp_5 all have negative slopes while imp_2,

imp_4, and imp_6 have positive slopes. This means that imps 1,3, and 5 all have a negative correlation regarding number of impressions and increases in purchasing odds (ie for every increase of impression purchasing odds go down) while imps 2, 4, and 6 have positive slopes indicating that for every impression increase purchasing odds go up by 1.48%, 19.71%, and 1.61% respectively.

3. How does the conversion effectiveness of Sites 1-5 compare with that of Site 6?

We looked at conversion effectiveness between sites one to five and site six in two ways. We looked again at the correlation between total number of impressions within sites one through five (taken as a whole) and then we also examined a model plotting the effects of sites one through five (individually) against purchasing odds and compared them with site six.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.145722   0.013946  -10.45   <2e-16 ***
sites1to5    0.032438   0.001461   22.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(This output explores the relationship between total number of impressions within sites one through five and purchasing odds)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.175539   0.013954 -12.580   <2e-16 ***
imp_1       -0.004198   0.003762  -1.116   0.2644    
imp_2        0.014695   0.001750   8.397   <2e-16 ***
imp_3       -0.039755   0.015777  -2.520   0.0117 *  
imp_4        0.179368   0.006947  25.818   <2e-16 ***
imp_5       -0.446645   0.048353  -9.237   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(this output explores the relationship between each individual site and purchasing odds)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.022103   0.013433  -1.645   0.0999 .  
imp_6        0.019834   0.002927   6.776 1.24e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(this model shows the relationship of just imp_6 against likelihood of purchase)

From what we can see in the first picture, the combined total of impressions across sites one to five has a positive correlation with purchasing intent/purchasing odds; as can be seen, a one unit increase in the independent variable carries a percent odds increase of ~3.2%. A model with just imp_6 has similar numbers when examined alone or as part of a model with all imp variables (ie ~1.6% in a model with all imp variables, and 2% in a model where it is plotted against purchase alone). Site 6 is not going to be a “waste” of resources (i.e. money spent has a negative effect on purchasing odds) but it is certainly not as high compared with the other model representing sites one through five.

4. Optional Challenge Question -- Which sites should Star Digital advertise on? In particular, should it put its advertising dollars in Site 6 or in Sites 1 through 5?

Star Digital should put its advertising on sites one through five and should ignore site six. If we look at the output of the models:

```
glm(formula = dat$purchase ~ dat$imp_6, family = binomial(), 
    data = dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2187  -1.1764   0.8864   1.1868   1.1868  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.022103   0.013433  -1.645   0.0999 .  
dat$imp_6    0.019834   0.002927   6.776 1.24e-11 ***
---
```

```
> exp(coef(imp6_fit))
(Intercept)   dat$imp_6 
  0.9781393   1.0200318 
> 
```

(the above pictures represent the outputs for the model detailing <u>only</u> site 6)

```
glm(formula = dat$purchase ~ dat$rowsums5, family = binomial(),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0075  -1.1297   0.1269   1.2119   1.2399

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.145722   0.013946  -10.45   <2e-16 ***
dat$rowsums5  0.032438   0.001461   22.21   <2e-16 ***
---
```

```
> exp(coef(imp1to5_fit))
 (Intercept) dat$rowsums5
   0.8643978    1.0329700
```

(these pictures represent the output of the model including combined totals for sites one through 5)

We can see that the exponential value for sites one through five is equal to approximately 1.033 (rounded) which we can interpret as for every one thousand impression increases within these sites we see a corresponding increase in purchasing odds of 3.45%. In contrast, the model containing only imp_6 had an exponential value of approximately 1.02 (rounded) which is interpreted as: for every 1000 impressions in site 6, the odds of purchase increase by approximately 2.003%. As 3.45% is larger than 2%, we can safely state that it is a better use of resources to try and boost impressions in sites one through five rather than focus on site six.