MELBOURNE BUSINESS SCHOOL

Marketing Analytics 2024

Group 7

Syndicate Assignment

Ruihan Deng

Chenxi Jia

Qingyang Kong

Ying Xu

Erin Zeng

Xiaoyang He

1. Data Set

1.1 Data Validation

To ensure accurate statistical inference and model robustness, the first step is to check the data validation.

First of all, by using the R function any(is.na(data)) and any(data < 0), it was observed that the data set has no missing value and all the value is non-negative. It is not reasonable to have a negative impression.

1.2 Balance Check

A balance check on the outcome variable was performed.

	Control Group	Test Group
Purchase	1290	11434
Not Purchase	1366	11213

The population of the test group is much higher than the control group.

T-test		
mean in group 0	mean in group 1	P-value
7.929217	7.869078	0.8987

However, after conducting a t-test, it is found that average impressions for both test group and control group are similar, and they are not significantly different from each other.

After checking distribution/ frequency of impressions across test groups, the total impressions of the websites are not balanced at all. However, since the sample is random given by the background information, we assume that there is no Selection Bias or Omitted Variable Bias.

test	total_imp_1	total_imp_2	total_imp_3	total_imp_4	total_imp_5	total_imp_6
0	1529	8370	2161	3959	91	4950
1	22026	78363	237	36260	1148	40177
test	proportion_imp_1	proportion_imp_2	proportion_imp_3	proportion_imp_4	proportion_imp_5	proportion_imp_6
0	0.07260209	0.3974359	0.102611586	0.1879867	0.004320988	0.2350427
1	0.12359506	0.4397203	0.001329884	0.2034667	0.006441802	0.2254462

1.3 Check correlations

The correlations of imp_1 to imp_6 have been checked by test/control group. It is found that they are not strongly correlated. The output is in the appendix.

1.4 customers behavior

It was found that some customers visit more than one website, the maximum is 6 websites. If the customer visits more than one website, it will be hard to justify which website affects the decision to purchase.

Min	Median	Mean	Max
1	1	1.419	6

2. Answers to the questions

1) Is online advertising effective for Star Digital?

1.1) Mean Treatment Effect.

The average purchase rate of the control group is 48.57% and test group is 50.49%. So, the average purchase rate of the test group is slightly higher.

1.2) Regression Models

A naive ATE is followed using Linear Regression and Logistic Regression.

	OLS model	Logit Model
AIC	36731.07	35077

The AIC of the Logistic Regression model is smaller, which is 35077. Based on the Logistic Regression model, the estimated coefficient for the test group is 0.07676 with a p-value of 0.0614. This result was statistically significant with a confidence interval of 90%, which suggests that there is a slight increase in purchase rate due to online advertising. However, it is not strong enough to be deemed conclusive evidence of effectiveness at 5% level. The odds ratio is approximately 1.0798, meaning that the odds of making a purchase are about 7.98% higher in the test group compared to the control group. It indicates that the advertising intervention may lead to an increased probability of purchase for users in the test group but the evidence is not strong enough.

1.3) Machine Learning Model

To further evaluate the effectiveness of online advertising for Star Digital, we used an XGBoost model to predict the likelihood of user purchase based on advertising exposure across six websites and whether the user was part of the test group exposed to the ads. The results showed that the test parameter had a positive impact on possibility of purchase, as indicated by the positive average marginal effect (AME) value of 0.0003165357. Furthermore, the feature importance analysis revealed that the Gain value for the test variable was 0.01185603, indicating that its effect on improving the model's performance was small. Additionally, the Cover and Frequency values for test were

0.03864357 and 0.03525377, respectively, suggesting that this variable was not widely used in decision-making splits within the model and did not have a significant impact overall. The results from the XGBoost model indicate that online advertising for Star Digital had a very limited effect on driving purchases.

1.4) Conclusion

Based on analysis above, it can be concluded that online advertising increases the odds of the customer purchasing the Star digital subscription package by 7.98%, but the evidence is not strong enough at 95% confidence interval. It may be due to online advertising on some of the websites not being effective compared to others. In other words, the possible reason is online advertising on some of the websites is effective but not on all websites.

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

2.1) Regression Model

First, the regression of purchase on test*log(tot_impressions) were done by Linear Regression and Logistic Regression. The AIC of Logistic Regression (logged impression) is the smallest, which is 33493.

	OLS model	OLS model (logged impression)	Logit Model	Logit Model (logged impression)
AIC	36145.36	35162.28	34198	33493

Based on the Logistic Regression model, the coefficient of test:log(tot_impressions) is 0.073458. It is suggested increasing of 1 unit of frequency of Impression in the test group, the possibility of purchasing will increase 7.62% compared to control group. However, the difference in impact between the test group and the control group is a little bit weak, with p=0.05876(by coefficient test, close to significance).

After that, Linear Regression and Logistic Regression were done to find the best fit of the interaction of test and each impression. The model with the smallest AIC, which is 31390, is Logistic Regression Model (logged impression)

	OLS model	OLS model (logged impression)	Logit Model	Logit Model (logged impression)
AIC	35730.96	33637.63	35731	31390

The output is in appendix 3. Based on the result of the selected Logistic Regression model, the P-values of all the interactions are not significant. The interaction between test

and imp_1/ test and imp_5 are negative and others are positive. It suggests that increasing the frequency on certain websites may have a negative effect on the purchase.

2.2) Marginal Effect

The output is in appendix 4. The marginal effects analysis showed that advertising frequency at specific sites had different impacts on the probability of purchase. For instance, test:log(imp_3 + 1) had a significant positive effect of 0.1176, suggesting that exposure to ads on site imp_3 had a more positive impact on users in the test group compared to those in the control group. But test:log(imp_5 + 1) had a significant negative effect of -0.19, suggesting that exposure to ads on site imp_5 had a more negative impact on users in the test group compared to those in the control group.

2.3) Conclusion

Increasing the frequency of overall advertising increases the probability of purchase. But increasing the frequency of advertising on websites 1 and 5 will decrease the probability.

3) Which sites should Star Digital advertise on?

To determine which sites Star Digital should allocate its advertising budget to, a comprehensive analysis involving regression modeling, cost-effectiveness analysis, and return on investment (ROI) calculations was applied.

3.1) Regression modeling

Firstly, a Logistic Regression Model is applied to analyze the impact of advertising, where the advertising impressions from Sites 1-5 are combined as a variable due to the advertiser cannot control which website it can advertise on as these websites are part of a single ad network, and Site 6 as a separate variable.

The output is in appendix 5, The coefficient of Sites 1-5 in the test group (test: sum1to5) was 0.01462 with a p-value of 0.0411274 at the confidence interval of 95%, which shows a high level of significance. This indicates that increasing advertising impressions on Sites 1-5 significantly enhances the likelihood of purchase. Additionally, the variable 'test: imp_6' shows a lower positive coefficient of 0.01348 with a higher p-value of 0.0574255, which is close to significant, but the effect is not as strong as that for Sites 1-5.

3.2) Cost-effectiveness analysis

Secondly, the cost-effectiveness of advertising these two choices is estimated based on the cost per thousand impressions, which shows that Sites 1-5 were valued at 25 dollars, while Site 6 was 20 dollars. A log-transformed regression model is established to examine the relationship between advertising costs and purchase likelihood. The coefficient of 'test: log (cost_1to5 + 1)' was 0.74871 with a P-value of 0.05254, which was close to significant. Although the 'test: log (cost_6 + 1)' had a slightly higher coefficient of 0.74966, the p-value of 0.19567, which is not significant enough.

This result reveals the increase in advertising expenditure on Site 1-5 has a slightly lower improvement compared to expenditure on Site 6. However, the P-value of test:log(cost_6 + 1) is not significant. Considering difference in coefficient is negligible, Variables with lower P-values are generally more reliable predictors, as their effects are less likely to be due to random noise. Star Digital should devote a larger portion of its advertising budget to Sites 1-5.

3.3) Return on investment (ROI)

Lastly, the ROI for advertising on these two sites is calculated based on the estimated conversion rates and advertising costs. The lifetime value of a purchase was set as 1200 dollars. The ROI of Site 1-5 was 705.754, while the ROI for Site 6 was higher at 813.447. However, although the ROI for Site 6 was slightly higher, the conversion effectiveness of Site 6 in driving purchase was not significant as discussed above, which means the higher ROI may be caused by its lower cost per thousand impressions and random noise rather than the effectiveness.

3.4) Conclusion

Based on the analysis, it is more beneficial for Star Digital to invest its budget on Sites 1-5 since advertising on these sites has shown a significant positive influence on the probability of purchase with high effectiveness in driving conversions. However, since advertising on Site 6 shows an insignificant effect on the user's purchase decisions, the budget allocated to Site 6 could be better utilized elsewhere. Overall, Star Digital should devote a larger portion of its advertising budget to Sites 1-5 to maximize purchase and reassess whether it aligns with the desired marketing objectives.

4) Which one would you recommend buying ads from?

4.1) Saturated LPM Model

Website	Coefficient	P-value
1	-0.014391	0.061
2	0.032773	4.50E-07
3	-0.032731	0.0618
4	0.503706	< 2e-16
5	-0.258138	< 2e-16
6	-0.002677	0.6885

From Saturated LPM Model, it shows that imp_2 and imp_4 both have positive and significant results. so, after comparing the Effect per ad unit of imp_2 and imp_4, it shows imp_4 is much higher than imp_2.

4.2) Regression Model

During our analysis, a variety of statistical models were used to evaluate the impact of individual website ad impressions on consumer purchasing decisions.

	OLS model	OLS model (logged impression)	Logit Model	Logit Model (logged impression)
AIC	35731.19	33632.35	35731	31389

The results show that imp_4 has the largest coefficient and P-value <2e-16.

Website	Coefficient	P-value
1	-0.044152	0.18682
2	0.130453	1.52E-14
3	-0.248166	0.01836
4	1.416081	< 2e-16
5	-1.465621	5.12E-15
6	0.091166	6.97E-06

Although imp_2 and imp_6 also have a significant positive impact on purchases, the effect is not as high as imp_4.

The marginal effect also supports the conclusion above. The imp_4 has the highest positive marginal effect. The output of the marginal effect is in appendix 6.

4.3) Cost effect

The balance between advertising cost and effectiveness is a key consideration when companies make advertising spending decisions. After taking account the cost of advertising into the selected Logistic Regression Model in 4.2, imp_4 has largest coefficient and P-value < 2.2e-16. It indicates that if Star Digital invests the same amount of money, imp_4 can give the highest increasing of probability of purchase.

Website	Coefficient	P-value
1	-0.044152	0.18682
2	0.130453	1.52E-14
3	-0.248166	0.01836
4	1.416081	< 2e-16
5	-1.465621	5.12E-15
6	0.091166	6.97E-06

The marginal effect also supports the conclusion above. The imp_4 has the highest positive marginal effect. The output of the marginal effect is in appendix 7.

4.4) Machine Learning Model

The xgboost and random forest are used.

AME	xgboost	random forest
cost1	-0.03693184	-0.01735366

cost2	0.03806438	0.03580317
cost3	0.01638675	0.01959553
cost4	0.0920528	0.08233464
cost5	-0.22856475	-0.1693412
cost6	0.04816691	0.02211454

Imp_4 has the highest AME in both models, which suggests that imp_4 is the best choice.

4.5) Conclusion

Based on the analysis above, we recommend buying ads from website 4.

5) Recommendations

The short come of currently study is obviously:

1, Lack of Cost-Effectiveness Analysis: The study does not adequately evaluate the cost-effectiveness of advertising, such as return on investment (ROI).

2, Neglect of Non-Linear Effects: The diminishing returns from additional impressions (non-linear effects) are not explicitly analyzed.

3, Attribution Challenges: Despite utilizing a robust test group methodology, the study does not account for multi-channel or cross-channel attribution.

4, Limited Metrics: The analysis primarily focuses on purchase outcomes and fails to incorporate broader metrics like brand awareness, website visits, or other intermediate indicators.

Based on the short come found above, we recommend as below:

1, Enhanced Experimental Design with Expanded Test Groups: customers in certain test groups will only be able to view the Star Digital Ad in specific channel and if they visit other channel, they will see charity ad instead. To separate the test group, the experiment design expands to test budget splits, non-linear frequency effects, multi-channel or cross-channel attribution. Furthermore, we can test various ad formats on the same website to identify which combinations drive the highest purchase.

2, Mitigate Diminishing Returns: Set a maximum viewing frequency for the Star Digital Ad to minimize diminishing effects. After the threshold is reached, display alternative ads (e.g., charity ads) on the same channel.

3, Segmentation Analysis: Incorporate customer characteristics such as age, location, income, and preferences to perform segmentation analysis. Evaluate ad performance across different customer segments to refine targeting strategies and improve effectiveness.

4, Track User Engagement Metrics: Monitor engagement indicators like time spent on-site, pages visited, and interaction rates to better understand the relationship between ads and purchase behavior.

5, Incorporate ROI as a Key Performance Metric: Use ROI to evaluate the cost-effectiveness of advertising on each channel, enabling data-driven budget optimization.

6. Extend the Tracking Period for Long-Term Impact: Assess the long-term effects of ads on customer retention and lifetime value. For example, tracking subscription retention rates over a year could provide deeper insights into site-specific and strategy-specific ad performance.

Appendix

1.Correlations for test group:

	imp_1	imp_2	imp_3	imp_4	imp_5	imp_6
imp_1	1.0000000	0.24562579	0.23678970	0.3726926	0.13616629	0.12599409
imp_2	0.2456258	1.00000000	0.07055948	0.2543954	0.05819335	0.05770546
imp_3	0.2367897	0.07055948	1.00000000	0.1554085	0.06069751	0.05068536
imp_4	0.3726926	0.25439543	0.15540845	1.0000000	0.07281820	0.10656838
imp_5	0.1361663	0.05819335	0.06069751	0.0728182	1.00000000	0.03712775
imp_6	0.1259941	0.05770546	0.05068536	0.1065684	0.03712775	1.00000000

2.Correlations for control group:

	imp_1	imp_2	imp_3	imp_4	imp_5	imp_6
imp_1	1.00000000	0.09924886	0.17573893	0.17629996	0.026150299	0.055955714
imp_2	0.09924886	1.00000000	0.26148930	0.27486053	0.147008485	0.011255882
imp_3	0.17573893	0.26148930	1.00000000	0.33586002	0.123245916	0.033769236
imp_4	0.17629996	0.27486053	0.33586002	1.00000000	0.037466217	0.054168887
imp_5	0.02615030	0.14700849	0.12324592	0.03746622	1.000000000	0.001464441
imp_6	0.05595571	0.01125588	0.03376924	0.05416889	0.001464441	1.000000000

3.Output of glm(purchase ~ test*(log(imp_1+1) + log(imp_2+1) + log(imp_3+1) + log(imp_4+1) + log(imp_5+1) + log(imp_6+1)),family = 'binomial', data = data)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.3395427	0.0693934	-4.8930	9.931e-07	***
test	-0.0704772	0.0737589	-0.9555	0.339321	
log(imp_1 + 1)	0.0990780	0.1125456	0.8803	0.378677	
log(imp_2 + 1)	0.1066825	0.0537407	1.9851	0.047129	*
log(imp_3 + 1)	-0.3400677	0.1078079	-3.1544	0.001608	**
log(imp_4 + 1)	1.3511469	0.1369683	9.8647	< 2.2e-16	***
log(imp_5 + 1)	-0.6806732	0.5733939	-1.1871	0.235190	
log(imp_6 + 1)	-0.0039971	0.0616170	-0.0649	0.948277	
test:log(imp_1 + 1)	-0.1628628	0.1179021	-1.3813	0.167175	
test:log(imp_2 + 1)	0.0274837	0.0566466	0.4852	0.627550	
test:log(imp_3 + 1)	0.5379720	0.3251850	1.6544	0.098055	.
test:log(imp_4 + 1)	0.0777763	0.1453324	0.5352	0.592538	
test:log(imp_5 + 1)	-0.8692192	0.6070475	-1.4319	0.152178	
test:log(imp_6 + 1)	0.1062684	0.0652605	1.6284	0.103446	

4.Output of logitmfx(purchase ~ test*(log(imp_1+1) + log(imp_2+1) + log(imp_3+1) + log(imp_4+1) + log(imp_5+1) + log(imp_6+1)), data = data,atmean = FALSE)

```
Call:
logitmfx(formula = purchase ~ test * (log(imp_1 + 1) + log(imp_2 +
    1) + log(imp_3 + 1) + log(imp_4 + 1) + log(imp_5 + 1) + log(imp_6 +
    1)), data = data, atmean = FALSE)

Marginal Effects:
                        dF/dx   Std. Err.       z     P>|z|
test                -0.01543749  0.01533464 -1.0067 0.3140754
log(imp_1 + 1)       0.02166560  0.02264846  0.9566 0.3387675
log(imp_2 + 1)       0.02332849  0.01115117  2.0920 0.0364365 *
log(imp_3 + 1)      -0.07436331  0.02054238 -3.6200 0.0002946 ***
log(imp_4 + 1)       0.29545812  0.02324427 12.7110 < 2.2e-16 ***
log(imp_5 + 1)      -0.14884423  0.07642862 -1.9475 0.0514756 .
log(imp_6 + 1)      -0.00087406  0.01325992 -0.0659 0.9474436
test:log(imp_1 + 1) -0.03561355  0.02365156 -1.5058 0.1321291
test:log(imp_2 + 1)  0.00600993  0.01177201  0.5105 0.6096823
test:log(imp_3 + 1)  0.11763945  0.05777454  2.0362 0.0417321 *
test:log(imp_4 + 1)  0.01700751  0.02433324  0.6989 0.4845886
test:log(imp_5 + 1) -0.19007398  0.08137451 -2.3358 0.0195021 *
test:log(imp_6 + 1)  0.02323793  0.01403557  1.6556 0.0977937 .
```

Output of logitmfx(purchase ~ test*(log(imp_1+1) + log(imp_2+1) + log(imp_3+1) + log(imp_4+1) + log(imp_5+1) + log(imp_6+1)), data = data,atmean = TRUE)

```
Marginal Effects:
                        dF/dx   Std. Err.       z     P>|z|
test                -0.01747343  0.01728593 -1.0108 0.3120896
log(imp_1 + 1)       0.02462413  0.02573908  0.9567 0.3387275
log(imp_2 + 1)       0.02651410  0.01266897  2.0928 0.0363636 *
log(imp_3 + 1)      -0.08451793  0.02332164 -3.6240 0.0002901 ***
log(imp_4 + 1)       0.33580418  0.02605176 12.8899 < 2.2e-16 ***
log(imp_5 + 1)      -0.16916954  0.08683983 -1.9481 0.0514073 .
log(imp_6 + 1)      -0.00099342  0.01507061 -0.0659 0.9474435
test:log(imp_1 + 1) -0.04047673  0.02687572 -1.5061 0.1320490
test:log(imp_2 + 1)  0.00683061  0.01337922  0.5105 0.6096742
test:log(imp_3 + 1)  0.13370362  0.06564016  2.0369 0.0416583 *
test:log(imp_4 + 1)  0.01932996  0.02765494  0.6990 0.4845710
test:log(imp_5 + 1) -0.21602938  0.09245346 -2.3366 0.0194585 *
test:log(imp_6 + 1)  0.02641117  0.01594824  1.6561 0.0977107 .
```

5.Output of glm(purchase ~ test*(sum1to5 + imp_6), data = data, family = 'binomial')

```
              Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.1665559  0.0470873 -3.5372 0.0004044 ***
test        -0.0060871  0.0506007 -0.1203 0.9042474
sum1to5      0.0194518  0.0064326  3.0239 0.0024950 **
imp_6        0.0039781  0.0057905  0.6870 0.4920822
test:sum1to5 0.0146167  0.0071572  2.0422 0.0411274 *
test:imp_6   0.0134828  0.0070960  1.9001 0.0574255 .
```

6. Output of marginal effect

```
Call:
logitmfx(formula = purchase ~ log(imp_1 + 1) + log(imp_2 + 1) +
    log(imp_3 + 1) + log(imp_4 + 1) + log(imp_5 + 1) + log(imp_6 +
    1), data = obdata, atmean = FALSE)

Marginal Effects:
                  dF/dx  Std. Err.       z     P>|z|
log(imp_1 + 1) -0.0096596  0.0064264  -1.5031  0.132808
log(imp_2 + 1)  0.0285404  0.0036011   7.9255 2.271e-15 ***
log(imp_3 + 1) -0.0542934  0.0173368  -3.1317  0.001738 **
log(imp_4 + 1)  0.3098081  0.0084630  36.6073 < 2.2e-16 ***
log(imp_5 + 1) -0.3206463  0.0264497 -12.1229 < 2.2e-16 ***
log(imp_6 + 1)  0.0199452  0.0043442   4.5912 4.407e-06 ***
```

```
Call:
logitmfx(formula = purchase ~ log(imp_1 + 1) + log(imp_2 + 1) +
    log(imp_3 + 1) + log(imp_4 + 1) + log(imp_5 + 1) + log(imp_6 +
    1), data = obdata, atmean = TRUE)

Marginal Effects:
                  dF/dx  Std. Err.       z     P>|z|
log(imp_1 + 1) -0.0109738  0.0072991  -1.5034  0.132727
log(imp_2 + 1)  0.0324233  0.0040681   7.9700 1.586e-15 ***
log(imp_3 + 1) -0.0616801  0.0196790  -3.1343  0.001723 **
log(imp_4 + 1)  0.3519577  0.0084479  41.6623 < 2.2e-16 ***
log(imp_5 + 1) -0.3642704  0.0297328 -12.2515 < 2.2e-16 ***
log(imp_6 + 1)  0.0226588  0.0049261   4.5997 4.231e-06 ***
```

7. Output of marginal effect

```
Call:
logitmfx(formula = purchase ~ log(cost1 + 1) + log(cost2 + 1) +
    log(cost3 + 1) + log(cost4 + 1) + log(cost5 + 1) + log(cost6 +
    1), data = obdata, atmean = TRUE)

Marginal Effects:
                 dF/dx Std. Err.       z     P>|z|
log(cost1 + 1) -0.087644  0.064708 -1.3545  0.175592
log(cost2 + 1)  0.266581  0.028277  9.4275 < 2.2e-16 ***
log(cost3 + 1) -0.654025  0.204427 -3.1993  0.001378 **
log(cost4 + 1)  2.576511  0.089474 28.7962 < 2.2e-16 ***
log(cost5 + 1) -5.121483  0.523744 -9.7786 < 2.2e-16 ***
log(cost6 + 1)  0.324224  0.051374  6.3110 2.772e-10 ***
```

```
Call:
logitmfx(formula = purchase ~ log(cost1 + 1) + log(cost2 + 1) +
    log(cost3 + 1) + log(cost4 + 1) + log(cost5 + 1) + log(cost6 +
    1), data = obdata, atmean = FALSE)

Marginal Effects:
                 dF/dx Std. Err.       z     P>|z|
log(cost1 + 1) -0.081366  0.060082 -1.3542   0.17566
log(cost2 + 1)  0.247488  0.026443  9.3593 < 2.2e-16 ***
log(cost3 + 1) -0.607183  0.189940 -3.1967   0.00139 **
log(cost4 + 1)  2.391976  0.088782 26.9422 < 2.2e-16 ***
log(cost5 + 1) -4.754672  0.489454 -9.7142 < 2.2e-16 ***
log(cost6 + 1)  0.301002  0.047857  6.2896 3.183e-10 ***
```