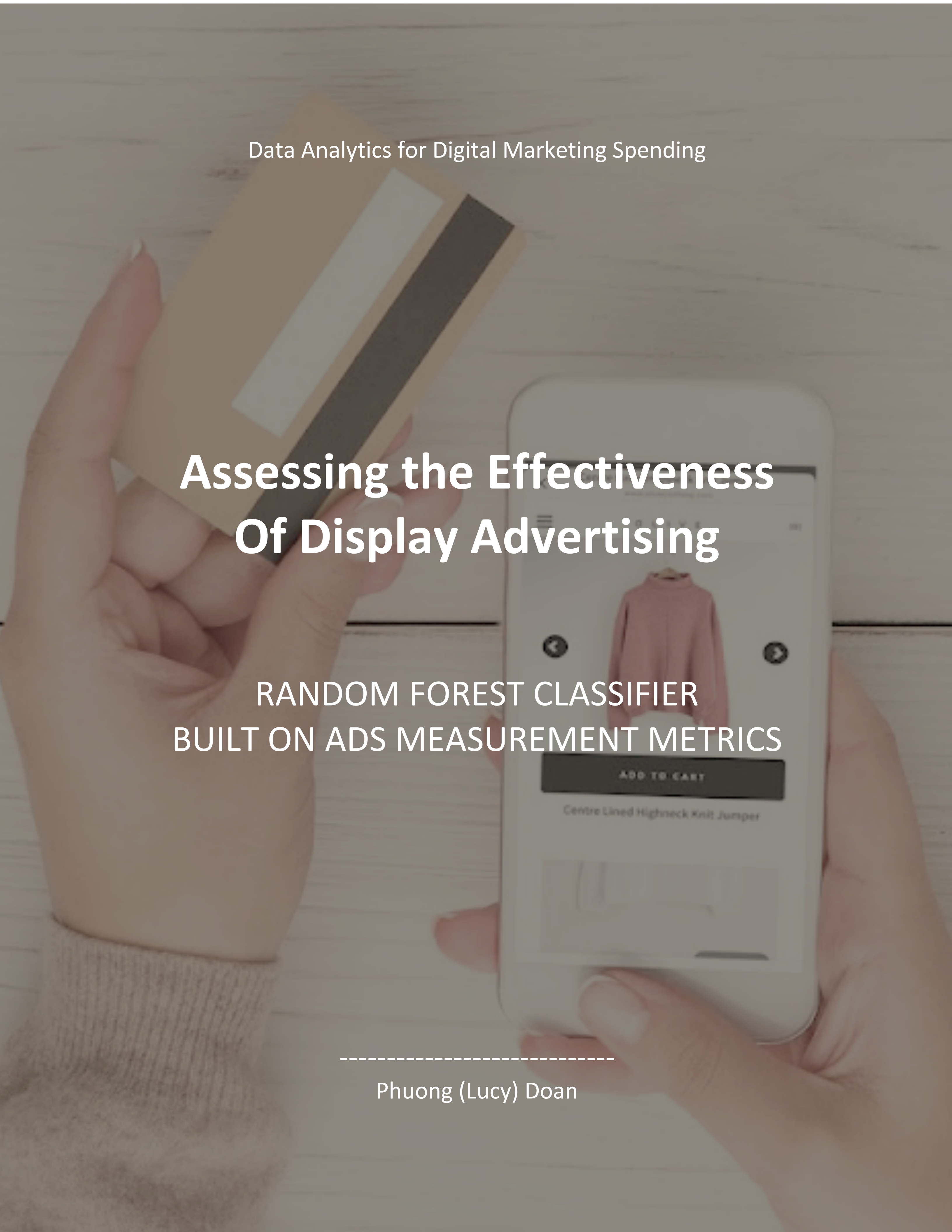


Data Analytics for Digital Marketing Spending

Assessing the Effectiveness Of Display Advertising

RANDOM FOREST CLASSIFIER BUILT ON ADS MEASUREMENT METRICS

Phuong (Lucy) Doan

Table of Contents

I. Problem Description

1.1. Business Objective

Star Digital is a multichannel video service provider with annual advertising spends over US$100 million. Since 2012/2013, the company has gradually increased the share of online advertising share, especially in banner ads. This advertising channel can not only help Star Digital to reach out to a large number of customers who spend a significant part of their time online, but also enable the company to better measure the effectiveness of their advertising campaigns. Therefore, the questions imposed to Star Digital marketers are whether online advertising creates a significant impact on sales, if yes, then where to display the ads, and how much of budget to be allocated for each channel.

To measure the causal effect of display advertising on sales conversion, one needs to measure what users would have done without seeing the campaign ads. Therefore, the company designed a control experiment, in which participants were randomly assigned to either a test group or a control group, where they would see Star Digital ads or a charity ads respectively. With this experimental design, it ran the campaign with advertisements shown on 6 websites and delivered 170 million impressions to about 45 million users during 2 months in 2012. The major objective of the advertising campaign was to generate subscription to a package offering of Star Digital.

1.2. Data Mining Objective

The data for the consumers who were part of this experiment, both in the experimental and control groups, was tracked. The study sample was chosen so that 50% of the users purchased the subscription and the remaining 50% did not. The advertising prices are $20 and 25% per thousand impressions for site 6 and each of the remaining 5 sites respectively. The raw data contains 25,303 samples, and the detailed description is as below.

Variable Name	Description	Type
purchase	Whether the consumer eventually purchased at Star Digital or not (1 = purchased; 0 = not purchased)	Integer
test	Whether the consumer belonged to test group (1) or control group (0)	Integer
imp_1	The number of ads impression for either Star Digital or charity that the consumer saw at website #1	Integer
imp_2	The number of ads impression for either Star Digital or charity that the consumer saw at website #2	Integer
imp_3	The number of ads impression for either Star Digital or charity that the consumer saw at website #3	Integer
imp_4	The number of ads impression for either Star Digital or charity that the consumer saw at website #4	Integer
imp_5	The number of ads impression for either Star Digital or charity that the consumer saw at website #5	Integer
imp_6	The number of ads impression for either Star Digital or charity that the consumer saw at website #6	Integer

II. Initial Data Preparation

2.1. Metrics and modeling identification

- **Metrics Identification**

Since the major objective of the campaign was sales of Star Digital's subscription, the conversion in this study is the users who chose to purchase the package. Besides, the reach is the number of unique users who have seen the ads.

- **Modeling Identification**

In order to measure the impact of impressions on each website on the purchasing decision of customers, we attempt to solve a binary classification problem using random forest classifier. This is a supervised machine learning algorithm used to predict the probability of 2 output classes, purchasing or not purchasing the company's subscription service.

2.2. Data Quality Checking

All variables are integer and have no missing value.

```
#Check data types
DigitalData.dtypes
```

```
id          int64
purchase    int64
test        int64
imp_1       int64
imp_2       int64
imp_3       int64
imp_4       int64
imp_5       int64
imp_6       int64
dtype: object
```

```
#Check if there is any missing data
DigitalData.isnull().sum()
```

```
id          0
purchase    0
test        0
imp_1       0
imp_2       0
imp_3       0
imp_4       0
imp_5       0
imp_6       0
dtype: int64
```

Regarding two binary variables, "purchase" and "test", the minimum, maximum, and quantile values are either 0 or 1, thus showing no sign of data errors.

	id	purchase	test	imp_1	imp_2	imp_3	imp_4	imp_5	imp_6
count	2.530300e+04	25303.000000	25303.000000	25303.000000	25303.000000	25303.000000	25303.000000	25303.000000	25303.000000
mean	7.089534e+05	0.502865	0.895032	0.930917	3.427775	0.094771	1.589495	0.048967	1.783464
std	4.084545e+05	0.500002	0.306518	5.629510	13.755455	1.505434	6.683091	0.570752	7.010298
min	2.700000e+01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.538805e+05	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	7.083440e+05	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
75%	1.062738e+06	1.000000	1.000000	0.000000	2.000000	0.000000	0.000000	0.000000	2.000000
max	1.413367e+06	1.000000	1.000000	296.000000	373.000000	148.000000	225.000000	51.000000	404.000000

III. Analytical Questions

3.1. Is online advertising effective for Star Digital?

Conversion Rate	Test Group	Control Group	Overall
Sample Size	22647	2656	25303
# Conversions	11434	1290	12724
Conversion Rate	50.49%	48.57%	50.29%

Site	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
# Conversions	2647	6014	394	4228	152	6378
Control	*6.14%*	*22.89%*	*9.53%*	*14.98%*	*0.41%*	*23.46%*
Test	*10.97%*	*23.87%*	*0.62%*	*16.91%*	*0.62%*	*25.41%*

Overall, the conversion rate of the test group was slightly higher than that of control group while individually, only except for site #1, sites #2 through #5 had higher conversion rate in test group than in control group. However, the performance gap is minimal for all sites #2 through #5.

Based on conversion rate for the entire sample, display advertising did not prove to perform significantly better than otherwise. However, the fact that conversion rate of control group outnumbered the rate of test group on site #1 may suggest an overall improved conversion rate if reducing or removing ads displayed on this site.

On a side note, all the calculation details can be found on Excel spreadsheet associated with this study report.

3.2.The frequency effect of advertising on purchase

Site	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
# impressions	*23555*	*86733*	*2398*	*40219*	*1239*	*45127*
# reach	*4753*	*11502*	*719*	*4654*	*503*	*13777*
Frequency	4.96	7.54	3.34	8.64	2.46	3.28
Probability of purchase	10.46%	23.77%	1.56%	16.71%	0.60%	25.21%

The ads frequency is calculated by dividing the number of impressions by the number of reaches. It is clear that site #4 had the highest ads frequency but resulted in a decent purchasing probability, whereas site #6 had among the lowest ads frequency but ranked first in terms of probability of purchase. This implies that increasing the frequency of ads displaying does not necessarily lead to higher purchasing of customers.

3.3.Which sites to advertise on

- **Impact of impressions on purchase**

In order to measure the impact of impression on purchase at each site, a random forest classifier model is developed, with the dependent variable being the purchase and the independent variables being the impression count on 6 websites.

After splitting the dataset into training and testing sets, we initialize the Random Forest Classifier, then let the model learn from the training set, and make the predictions.

```
RF_class = RandomForestClassifier(n_estimators=50)
RF_class = RF_class.fit(X_train, y_train)
y_pred_RF = RF_class.predict(X_test)
```

The mode performance measurements, precision, recall and f1-score, all fall within the range 0.6 – 0.8, which validate the quality of the model.

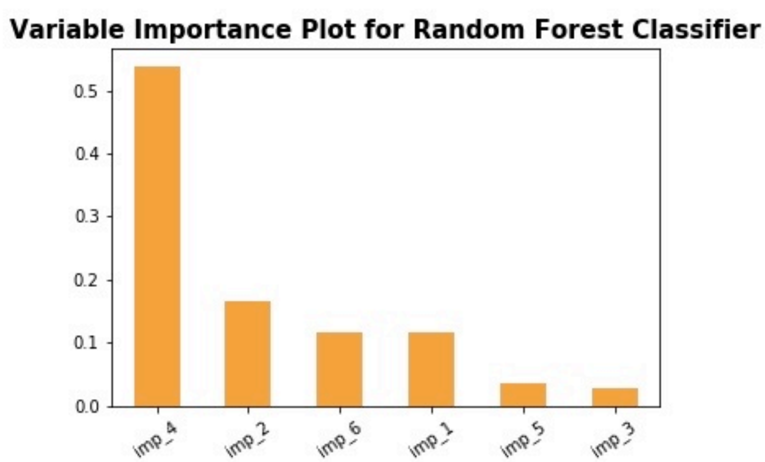
precision	0.706072
recall	0.654954
f1-score	0.631000

The below plot shows that the impressions on site #4 has the most influence on customer's purchasing behavior, followed by sites #2 and 6#

```
#Variable importances
RFclass_VarImportance = pd.Series(RF_class.feature_importances_, index = X.columns)

#Visualize with bar chart
RFclass_VarImportance.nlargest(6).plot(kind="bar", rot=35, color="orange")
plt.title("Variable Importance Plot for Random Forest Classifier", size=15, weight="bold")
```

```
Text(0.5, 1.0, 'Variable Importance Plot for Random Forest Classifier')
```

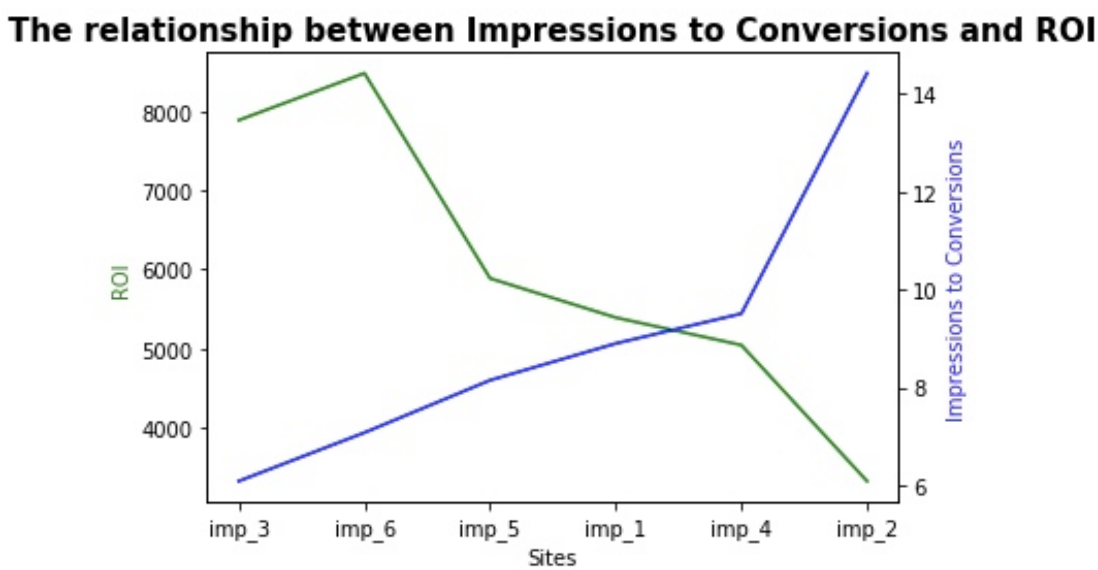


- **Return on Investment on each advertisement site**

Additionally, another indicator used to measure site's performance is ROI, or return on investment, which is equal to total customer lifetime value on each channel divided by total campaign cost of that site.

Site	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
Purchase after impressions	2647	6014	394	4228	152	6378
Customer Lifetime Value ($)	3176400	7216800	472800	5073600	182400	7653600
Campaign cost ($)	588.88	2168.33	59.95	1005.48	30.98	902.54
ROI	5394.01	3328.28	7886.57	5045.97	5888.62	8480.07
Impressions to Conversion	8.90	14.42	6.09	9.51	8.15	7.08

Sites #3 and #6 had the highest ROIs. Interestingly, site #3 had pretty low customer lifetime value (CLV) but low campaign cost, thus producing high ROI. Since the CLV is a function of conversion while campaign cost a function of # impressions, when the advertising price is not different much among different sites, the ROI would depend on the ratio of impressions to conversions. In this case, site #3 with the lowest impressions to conversions ratio resulted in the highest ROI; In contrast, site #2 with the highest impressions to conversions ratio resulted in the lowest ROI.



IV. Conclusions and Recommendations

In general, online advertising did not show a significant advantage on driving conversion rate. However, some sites showed relatively better conversion rate than some others, suggesting that redistributing advertising spends from the low converting sites to high converting sites may result in higher conversion rate overall for the whole campaign.

In a consideration about ads frequency, increasing the frequency of advertising did not necessary lead to higher purchasing of customers. Therefore, the ratio of the number of impressions to the number of unique users seeing the ads is not a factor affecting how likely a customer will buy Star Digital's packages.

Regarding sites' performance, sites #4 and #2 had significantly higher impacts on driving customer's purchase and dominated other sites in terms of impressions to conversions. However, since the ratio of impressions to conversions is inversely proportional to return on investment, the company should advertise on low converting sites instead, such as sites #3 and #5, if it wishes to maximize the value of advertising money.