Homework 5 - Attribution Exercise

Drew Ficken | Ido Goren | Pranav Himatsingka

3/09/2019

Problem Overview

. CustomerID: Unique customer identifier

. Branding Display (M): Online displays to generate brand awareness, very lightly targeted

. Paid Brand Search (M): paid search impressions for search terms associated with the advertiser's brand

. Paid Generic Search (M): paid search impressions for search terms associated with the travel or specific travel destinations

. Targeted Display (M): Online displays targeted to travel-prone audiences

. Retargeted Display (M): Online displays targeted to customers who recently visited the advertiser's website

. Seasonality: an indicator to distinguish between hi/low season

. TV (M): Adstocked TV GRPs

. Purchase: A 0/1 indicator of whether the customer made a booking or not

. Customer Propensity Score: a propensity score derived from a separate model based on demographic and historical behavior attributes indicating a measure of likelihood to book with the advertiser

. transaction_datetime: time stamp of transaction or reference date

. transaction_date: date of transaction or reference date

Question 1 - Build a Model

Build a model that predicts the probability of conversion as a function of all other customer metrics. [Note: We suggest formulating the problem as a logistic regression model. We have tested the approach on the dataset. Other model structures might be appropriate as well.]

```
out = glm(data = samp.dat, purchase ~ branding_display + tv +
    retargeting + targeted_display + paid_brand_search + paid_generic_search +
    base_propensity + factor(month), family = binomial)

# knitr::include_graphics('regression out.JPG')
```

This was given to us in the template. As you can see from the logit regression summary, all of the covariates are significant (while some of the months aren't necessarily statistically significant, if we want to include the significant month effect in the regression we have to leave the factor(month) and therefore all months in the model). The non-month covariates are all extremely significant.

See Regression # 1 in dashboard for regression summary.

Question 2a - Calculate Attribution

For each of the converting observations, calculate the expected change in conversion probability Δ P(M) associated with each of the marketing activities (M) and attribute the fraction Δ P(M)/Total predicted

conversion probability to each of the marketing activities (M). Δ P(M) is a measure of the expected number of incremental bookings driven by marketing treatment M.

All of the marketing variables (aka branding display, paid brand search, paid generic search, targeted display, retargeted display, and tv) have already undergone covariate transforms to account for the time decay and saturation. For this problem, we'll have to get the Base, which is measured by turning off all touch-points to get the contribution delta. This delta is renormalized to assess the percentage contribution of the k-th touch-point.

Therefore, base is Pr(X=0) = 1/(1 + exp(intercept))

Then, to get the contribution delta, we find Pr(Y=1 | X_mit) - base.

For a multinomial logit regression, we know that the probability of conversion is:

P = $1/(1+\exp(-\alpha - B_1 * X_1 - \ldots B_n * X_n))$, so in order to find the lift or delta from a single marketing activity we subtract the coefficient that we arrived at for that variable from the negated alpha in the denominator.

See Delta Table #1 for Marketing Activity Deltas and Weights.

```
coeff = out$coefficients

base = 1/(1 + exp(-coeff[1]))

branding_display_delta = 1/(1 + exp(-coeff[1] - coeff[2])) -
    base

tv_delta = 1/(1 + exp(-coeff[1] - coeff[3])) - base

targeted_display_delta = 1/(1 + exp(-coeff[1] - coeff[4])) -
    base

retargeted_display_delta = 1/(1 + exp(-coeff[1] - coeff[5])) -
    base

paid_brand_search_delta = 1/(1 + exp(-coeff[1] - coeff[6])) -
    base

paid_generic_search_delta = 1/(1 + exp(-coeff[1] - coeff[7])) -
    base


### put this on dashboard
delta_df <- rbind(base, branding_display_delta, tv_delta, targeted_display_delta,
    retargeted_display_delta, paid_brand_search_delta, paid_generic_search_delta)


total_weight = base + branding_display_delta + tv_delta + targeted_display_delta +
    retargeted_display_delta + paid_brand_search_delta + paid_generic_search_delta


base_weight = base/total_weight

branding_display_weight = branding_display_delta/total_weight

tv_weight = tv_delta/total_weight
```

```
targeted_display_weight = targeted_display_delta/total_weight

retargeted_display_weight = retargeted_display_delta/total_weight

paid_brand_search_weight = paid_brand_search_delta/total_weight

paid_generic_search_weight = paid_generic_search_delta/total_weight

weight_df <- rbind(base_weight, branding_display_weight, tv_weight,
    targeted_display_weight, retargeted_display_weight, paid_brand_search_weight,
    paid_generic_search_weight)

show_df <- cbind.data.frame(delta_df, weight_df)
colnames(show_df) <- c("Delta", "Weight")
rownames(show_df) <- c("Intercept", "Branding Display", "TV",
    "Targeted Display", "Retargeted Display", "Paid Brand Search",
    "Paid Generic Search")

# kable(show_df)
```

Question 2b

Calculate the total attributed conversion by marketing activity.

This question is asking us to look at the total number of observations in the data set that we are working with, and determine how many individual conversions are due to the individual marketing activities. The intercept has a large weight of 0.4986, so for comparison sake we know that there are a decent number of people that would convert with zero marketing activities. It's also important to note the high coefficient value of the base propensity (odds ratio 7.25), which isn't considered to be a marketing activity but is a covariate that has a large effect on conversion rate. We'll discuss this in problem 3.

One key distinction separating this scenario from some of our previous assignments is that these marketing activities aren't represented as a factor model of 1's and 0's, showing if they did occur or didn't. Instead, they've been transformed, so the deltas don't represent the overall percentage of conversions derived from the marketing activity. Rather, we take the sum of these covariate transforms for each marketing activity and multiply it by the delta to see how many real conversions can be attributed to that marketing activity. For the intercept, we can actually just multiply the delta by the total number of observations to find how many people would have converted anyways.

The average transform column was included to show why some of the marketing activities that have high deltas don't actually result in a high number of conversions.

See Table #1 in dashboard for Marketing Activity Conversions.

```
conv_ct = nrow(samp.dat[samp.dat$purchase == 1])

intercept_conversions = nrow(samp.dat) * base

branding_display_conversions = sum(samp.dat$branding_display) *
    branding_display_delta

tv_conversions = sum(samp.dat$tv) * tv_delta

targeted_display_conversions = sum(samp.dat$targeted_display) *
    targeted_display_delta
```

```
retargeted_display_conversions = sum(samp.dat$retargeting) *
    retargeted_display_delta

paid_brand_search_conversions = sum(samp.dat$paid_brand_search) *
    paid_brand_search_delta

paid_generic_search_conversions = sum(samp.dat$paid_generic_search) *
    paid_generic_search_delta

conv_df <- rbind(intercept_conversions, branding_display_conversions,
    tv_conversions, targeted_display_conversions, retargeted_display_conversions,
    paid_brand_search_conversions, paid_generic_search_conversions)

mean_df <- rbind("NA", mean(samp.dat$branding_display), mean(samp.dat$tv),
    mean(samp.dat$targeted_display), mean(samp.dat$retargeting),
    mean(samp.dat$paid_brand_search), mean(samp.dat$paid_generic_search))

show_df1 <- cbind.data.frame(delta_df, weight_df, mean_df, conv_df)
colnames(show_df1) <- c("Delta", "Weight", "Average Transform",
    "Conversions")
rownames(show_df1) <- c("Intercept", "Branding Display", "TV",
    "Targeted Display", "Retargeted Display", "Paid Brand Search",
    "Paid Generic Search")

# kable(show_df1)
```

Question 3a

Repeat steps 1 and 2 but omit the customer propensity score. [Note: A (bad) model might suggest that a marketing treatment reduces the probability of conversion. In that case, attributions could be negative.]

Compare attribution results

See Regression #2 in Dashboard for new model's regression summary that no longer includes propensity score.

See Table #2 in Dashboard for Marketing Activities Deltas, Weights, and expected conversions.

Question 3b

Why do we see the directional changes we see?

As you can see from the differences between Table #1 and Table #2 in the dashboard, when we remove the customer propensity score we run into a targeting bias issue, as mentioned in the "AttributionDoneRight" slides. The consumers that have a high propensity score are already inclined to buy our product which we measured with past behavior. By not including this variable, we add much more weight to the intercept (from delta of .055 to .143), as well as the targeted display, retargeted display, paid brand and generic searches (.004 to .079, .005 to .107, .018 to .135, and .017 to .26, respectively). On the reverse side, tv and branding display actually decrease both in delta (.006 to .003 and .003 to -.005, respectively) and in conversions (340 to 200 and 79 to -117, respectively), going from the most effective to the least effective marketing activities in terms of conversions. One of the assumptions when running a regression is that the population of customers that we're running the regression on is that these customers are selected randomly, which is not the case in

this scenario. That's why this customer propensity score needs to be included. If this score wasn't significant and had a odds ratio very close to 1.00 we wouldn't necessarily have to include it, but as we saw in the first regression the odds-ratio of this propensity score is actually 7.25.

Question 3c

What would be the business consequences of not accounting for exogenous differences in consumer affinity for the product?

This makes marketers believe that their marketing campaigns are much, much more effective than they really are. This is dangerous because their allocation will be spent heavily on these different touch-points, and not really see the effects that they were expecting given their prior analysis which would be frustrating for all parties involved. Furthermore, besides spending too much overall on these marketing activities, they'd be allocating their resources towards the wrong ones, concentrating heavily on retargeted display rather than branding display and tv. Finally, if they didn't attempt to seek out the right people to target based on their higher propensity scores, they would expect a 14% conversion rate regardless of their history, which is almost 3 times higher than what they'd actually get from a random sample.

Question 4

Create a dashboard that contains the results of your project including estimation results, attribution and any other features you think are useful.

Dashboard - Conversion Results

Below are results that predict the probability of a conversion as a function of all customer metrics, as well as expected change in conversion probability (Delta) and the total attributed conversions.

Regression #1

	purchase		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.06	0.05 – 0.07	**<0.001**
branding display	1.12	1.09 – 1.15	**<0.001**
tv	1.07	1.03 – 1.11	**<0.001**
retargeting	1.07	1.04 – 1.11	**<0.001**
targeted display	1.10	1.05 – 1.16	**<0.001**
paid brand search	1.36	1.23 – 1.51	**<0.001**
paid generic search	1.35	1.22 – 1.50	**<0.001**
base propensity	7.25	6.99 – 7.53	**<0.001**
factor(month)2	0.80	0.62 – 1.04	0.095
factor(month)3	0.70	0.51 – 0.96	**0.027**
factor(month)4	0.96	0.65 – 1.43	0.855
factor(month)5	0.44	0.17 – 1.16	0.096
factor(month)6	0.89	0.76 – 1.04	0.134
factor(month)7	0.90	0.77 – 1.06	0.200
factor(month)8	0.88	0.75 – 1.03	0.106
factor(month)9	0.93	0.78 – 1.09	0.356
factor(month)10	0.93	0.79 – 1.11	0.442
factor(month)11	0.95	0.79 – 1.14	0.555
factor(month)12	0.95	0.79 – 1.16	0.633
Observations	100000		
Cox & Snell's R^2 / Nagelkerke's R^2	0.152 / 0.262		

Regression #2

	purchase		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.17	0.15 – 0.19	**<0.001**
branding display	1.03	1.01 – 1.06	**0.017**
tv	0.96	0.93 – 0.99	**0.015**
retargeting	1.71	1.66 – 1.76	**<0.001**
targeted display	2.00	1.90 – 2.09	**<0.001**
paid brand search	2.31	2.05 – 2.59	**<0.001**
paid generic search	4.04	3.60 – 4.54	**<0.001**
factor(month)2	0.85	0.66 – 1.08	0.177
factor(month)3	0.76	0.57 – 1.03	0.076
factor(month)4	1.13	0.79 – 1.63	0.506
factor(month)5	0.38	0.15 – 0.96	**0.040**
factor(month)6	0.87	0.76 – 1.01	0.072
factor(month)7	0.91	0.79 – 1.05	0.196
factor(month)8	0.90	0.77 – 1.04	0.152
factor(month)9	0.96	0.82 – 1.12	0.573
factor(month)10	0.94	0.80 – 1.10	0.466
factor(month)11	0.98	0.83 – 1.16	0.794
factor(month)12	0.97	0.81 – 1.16	0.713
Observations	100000		
Cox & Snell's R^2 / Nagelkerke's R^2	0.045 / 0.077		

Table #1

	Delta	Weight	Average Transform	Conversions
Intercept	0.0547823	0.4985906	NA	5478.23440
Branding Display	0.0063015	0.0573520	0.540295198215045	340.46808
TV	0.0034856	0.0317235	0.227543357666118	79.31260
Targeted Display	0.0038573	0.0351067	0.058058155744266	22.39496
Retargeted Display	0.0053649	0.0488279	0.11837813144912	63.50913
Paid Brand Search	0.0182073	0.1657101	0.012371822386783	22.52575
Paid Generic Search	0.0178754	0.1626891	0.017630828702422	31.51575

Table #2

	Delta	Weight	Average Transform	Conversions from Activity
Intercept	0.1436300	0.1982256	NA	14362.9982
Branding Display	0.0037158	0.0051283	0.540295198215045	200.7641
TV	-0.0051484	-0.0071053	0.227543357666118	-117.1480
Targeted Display	0.0795740	0.1098211	0.058058155744266	461.9917
Retargeted Display	0.1070925	0.1477998	0.11837813144912	1267.7414
Paid Brand Search	0.1352724	0.1866912	0.012371822386783	167.3566
Paid Generic Search	0.2604419	0.3594394	0.017630828702422	459.1807