Shopper Hiring Problem

Analyzing A/B Test Results on Shopper Hiring Funnel

Jason Dexter (5/20/2019)

Contents

PROJECT + ASSIGNMENT OVERVIEW

Problem Statment: Low conversion rates of new hires b/c of drop out during hiring funnel.

Potential Solution: Initiate applicant background check earlier in hiring funnel (on day one).

Objective: Analyze A/B Test results and assess the viability of posed solution for improving conversion rates.

Ultimately, we want to see if initiating the background check sooner:

1) Increases the liklihood of applicants starting as shoppers.
2) Gets the shoppers to start more quickly.

Questions Driving Analysis - need to be answered/delivered via slide-deck (decision-making-audience):

1) What can we conclude at this point from the A/B test?
2) How confident should we be in this conclusion?
3) Is this change cost-effective?
4) How should we think about the cost-effectiveness or return on investment of this change?
    - Consider alternate costs: $50 or $100 instead of $30 (be as specific as possible)
5) What other observations and recomendations do you have for us, based on this data?
    - E.g., what else did you find that seems relevant, or what else would you want to test if we ran an additional experiment?

Project Phase 1.0: Initial Exploratory Data Analysis (EDA)

Objective w/EDA: Gain understanding of event-level data so I can use insights to aggregate/summarize data to the applicant-level. The applicant-level data will be what I use to analyze the A/B test results.

```
# Set global options for code chunks
knitr::opts_chunk$set(
    echo    = TRUE,
    message = FALSE,
    warning = FALSE)

# Tell RMarkdown to recognize the root directory of my Rproj file
knitr::opts_knit$set(root.dir = rprojroot::find_rstudio_root_file())
```

```
# Load libraries + source plotting function
library(tidyverse)
library(lubridate)
library(tidyquant)
library(stringr)
library(sigr)
source("00_Scripts/plot_ggpairs.R")

# Read in raw data
applicant_raw_tbl <- read_csv("00_Data/applicant_data.csv")
```

1.1 VIEW DATA + ASSESS DATA TYPES + ASSESS MISSING DATA

```
# View data
applicant_raw_tbl %>% head(4)
```

```
## # A tibble: 4 x 6
##   applicant_id channel           group     city    event           event_date
##          <dbl> <chr>             <chr>     <chr>   <chr>           <chr>
## 1        10001 web-search-engi~ control   Asgard  application_d~ 10/1/18
## 2        10002 social-media      control   Midgard application_d~ 10/1/18
## 3        10003 web-search-engi~ treatment Midgard application_d~ 10/1/18
## 4        10004 social-media      treatment Asgard  application_d~ 10/1/18
```

```
# Assess missing data (NA values): could be other ways of data missing (this is a good 1st look)
applicant_raw_tbl %>%

    # Iterate across columns and calculate % missing
    map_df(~ sum(is.na(.)) / length(.)) %>%
    knitr::kable(caption = "No NA values present in dataset")
```

Table 1: No NA values present in dataset

applicant_id	channel	group	city	event	event_date
0	0	0	0	0	0

1.2 INITIAL DATA CLEANING

```r
# Clean raw data where needed
applicant_tbl <- applicant_raw_tbl %>%
    # Parse dates in event_date column
    mutate(event_date = mdy(event_date))
```

1.3 INSPECT CATEGORICAL DATA

- Here I did not include the output b/c I just did a quick look at distinct groups per categorical variable.
- This included channel, group, city, and event type.
- I'm getting a sense of what categories exists, their values, and what I might use later to **explain any variation discovered.**

1.4 INSPECT DISTINCT APPLICANTS AND GROUP SAMPLE SIZE(S)

- Looks like the study was intentionally setup as a 2/3 control & 1/3 treatment (setup/study design)

Table 2: % Distinct Applicants by Test Group

group	n	pct_in_group
control	14501	66.8%
treatment	7197	33.2%

1.5 HOW LONG WAS THE A/B TEST RUN? SAME FOR BOTH GROUPS?

- Looks like ~41 days and the date ranges are the same for both groups.
- Upon initial inspection I suspect the A/B test was specifically for Oct, 2018.

Table 3: Min and Max event dates by group.

group	min(event_date)	max(event_date)
control	2018-10-01	2018-11-11
treatment	2018-10-01	2018-11-11

Was this an OCT Test & extra 11 days allow time for conversion?

My thoughts here are that we don't want to include applicants that never had a chance to successfully convert to a hired shopper.

This means we need a cutoff date where we don't allow any more applicantes into the analysis. For example, if it takes roughly 11 days for applicants to *complete their first batch* (success), then we need to allow that much time to pass.

- I'm now going to assess the time between application to becoming a succeful hire: *"complete 1st batch"*
- I will use this info to create a cutoff where anyone who applies after date X will not be in the analysis.

1.6 TIDY + TRANSFORM DATA TO STUDY TIME-TO-CONVERSION

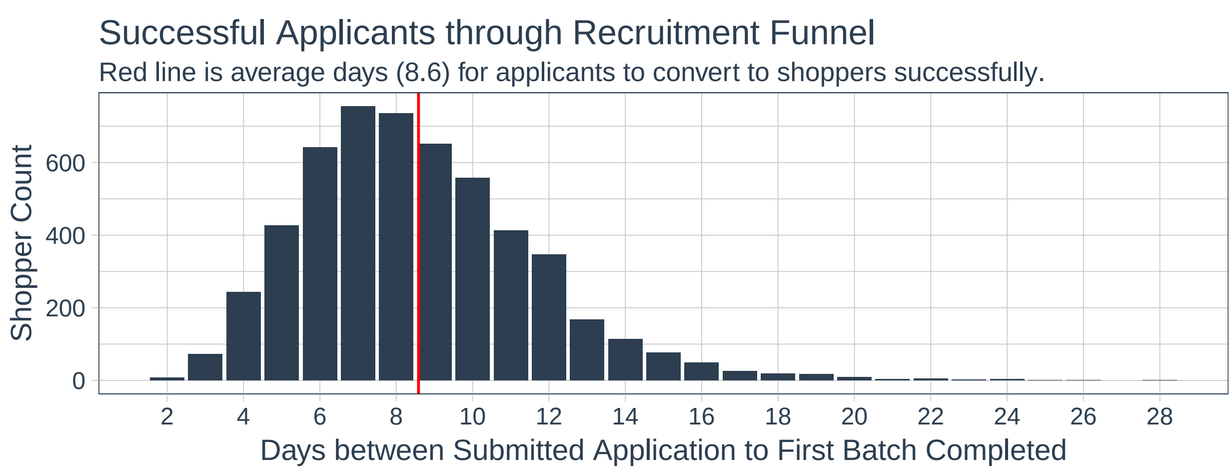
```r
# Aggregate data to assess time between application and successful hire
time_to_conversion_tbl <- applicant_tbl %>%

    # Select columns and filter for event types
    select(applicant_id, group, contains("event")) %>%
    filter(event %in% c("application_date", "first_batch_completed_date")) %>%

    # Pivot and spread event and event_date across columns
    spread(key = event, value = event_date) %>%

    # Filter to get only applicants who converted by dropping rows w/NA values
    filter(!is.na(first_batch_completed_date)) %>%

    # Calculate time between application date and hire date
    mutate(days_to_conversion = first_batch_completed_date - application_date)

# View time to conversion table by pulling 5 sample rows
time_to_conversion_tbl %>% sample_n(5) %>%
    knitr::kable(caption = "Days to conversion by successful applicants.")
```

Table 4: Days to conversion by successful applicants.

applicant_id	group	application_date	first_batch_completed_date	days_to_conversion
12944	treatment	2018-10-08	2018-10-11	3 days
11745	treatment	2018-10-05	2018-10-14	9 days
10761	control	2018-10-02	2018-10-14	12 days
19927	control	2018-10-23	2018-11-04	12 days
26718	treatment	2018-11-03	2018-11-09	6 days

1.7 HOW LONG DOES IT TAKE TO MOVE THROUGH HIRING FUNNEL?

What does the distribution of time-to-conversion in days look like?



1.8 IS IT APPROPRIATE TO USE 10/31/2018 AS A CUTOFF?

The data is nicely distributed and so let's take a look closer at how it's distributed.

- This will inform our cutoff for which applicants go into the analysis of the A/B test.
- The 11 days in NOV might be enough to allow most applicants who will convert, to convert.

High-level summary (table 5) shows that the ~11 days is above the 75th percentile and will capture the majority of 'successes.' Meaning that this should give plenty of time for MOST conversions to have been completed.

Table 5: High-level summary of distribution stats

min	Q1	median	mean	Q3	max	IQR
2 days	6 days	8 days	8.6 days	10 days	28 days	4

Table 6: Summary of distribution stats by experimental group(s)

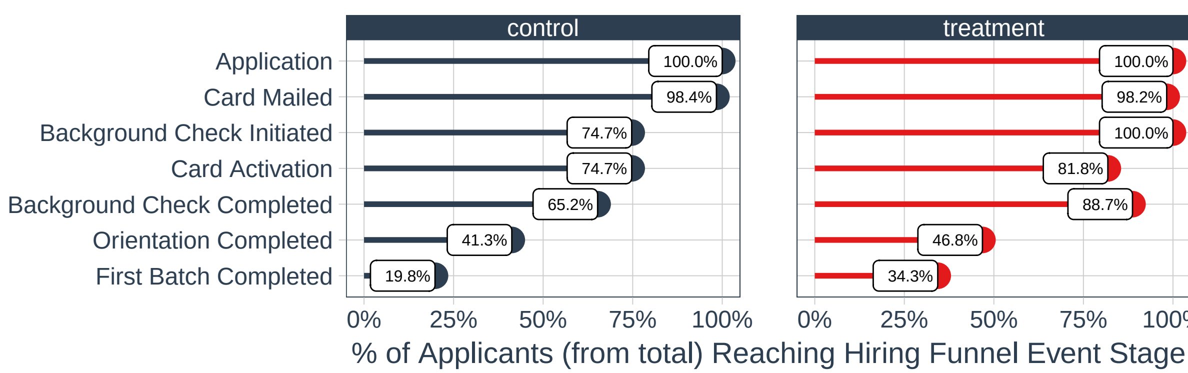
group	min	Q1	median	mean	Q3	max	IQR
control	4 days	8 days	10 days	10.0 days	12 days	28 days	4
treatment	2 days	5 days	7 days	6.9 days	8 days	26 days	3

Key-Takeaways

1) Based on findings, I will use 10/31/2018 as the cutoff date and assume OCT A/B Test.
2) Any applicants who applied after that will be dropped from the analysis.
3) This is our **first indication of differences between treatments** (Table 6).
    - Initial inspection suggests treatment group applicants are converting quicker (10-days vs. 7-days).

1.9 SHOPPER HIRING FUNNEL: CONTROL VS. TREATMENT

Preliminary Results in plot: Data includes NOV applicants. Just wrapping my mind around funnel.



Initial inspection indicates large differences in conversion rates between Control vs. Treatment.

- See differences between groups for '1st Batch Completed' (34.3% vs. 19.8%). **PRELIMINARY**

Project Phase 2.0: Use EDA Insights to Wrangle Data for Analysis

The objective here is to aggregate event-level data to applicant-level.

NOTE: I'm often taking time to do sanity checks on my work at each stage.

2.1 PREP EVENT DATA FOR TIME-BASED CALCULATIONS - JOINED IN 2.2

```
# Pivot data to get 'application_date" and "1st_batch_date" as seperate features
app_date_batch_date_for_joins_tbl <- applicant_tbl %>%

    # Select columns for pivot
    select(applicant_id, event, event_date) %>%

    # Pivot and spread events across columns with date completed as values
    spread(key = event, value = event_date) %>%

    # Select columns needed for calculating days to conversion: 1st_batch_date - app_date = days
    select(applicant_id, application_date, first_batch_completed_date)
```

2.2 WRANGLE DATA INTO THE LEARNING DATA SET FOR ANALYSIS

```
# Construct learning data with target feature: coverted (success/failure)
learning_data_tbl <- applicant_tbl %>%

    # Drop event date. We will add back with joins
    select(-event_date) %>%

    # Setup temp column. For engineering binary features related to event completion
    mutate(yes_no = "Yes") %>%

    # Pivot & spread events across columns to create binary features (fill NA w/"No")
    mutate(event = str_replace(event, pattern = "_date", "")) %>% # remove "_date" for event
    spread(key = event, value = yes_no, fill = "No") %>% # sets event as "yes" or "no"

    # Join data for calculating days to conversion (inner join is fine b/c both have ALL applicants)
    inner_join(app_date_batch_date_for_joins_tbl, by = "applicant_id") %>%

    # Calculate days to conversion for those who successfully completed 1st batch
    mutate(days_to_conversion = (first_batch_completed_date - application_date)/ddays()) %>%

    # Setup Target feature: Success/Failure
    mutate(converted = case_when(
        first_batch_completed == "Yes" ~ "Success",
        TRUE ~ "Failure"
    )) %>%

    # Filter out applicants who applied in November
    filter(application_date <= "2018-10-31")

#learning_data_tbl %>% filter(group == "treatment") %>% count(background_check_completed)
```

2.3 COLUMNS AND ENGINEERED FEATURES IN LEARNING DATA SET

Let's take a quick glimpse of what data we now have at the applicant-level.

- The Target feature is 'converted' denoting shopper hiring funnel completion: 'Success' or 'Failure'

This is a great data set for us to answer the assigned questions.

- It's also setup nicely for further investigation if we want to do further analysis later to understand the system better e.g., what other factors are driving conversion of applicants to shoppers.

```r
# Transpose data to view glimpse of all features
learning_data_tbl %>% glimpse
```

```
## Observations: 14,982
## Variables: 15
## $ applicant_id               <dbl> 10001, 10002, 10003, 10004, 10005, ...
## $ channel                    <chr> "web-search-engine", "social-media"...
## $ group                      <chr> "control", "control", "treatment", ...
## $ city                       <chr> "Asgard", "Midgard", "Midgard", "As...
## $ application                <chr> "Yes", "Yes", "Yes", "Yes", "Yes", ...
## $ background_check_completed <chr> "No", "Yes", "Yes", "Yes", "Yes", "...
## $ background_check_initiated <chr> "No", "Yes", "Yes", "Yes", "Yes", "...
## $ card_activation            <chr> "No", "Yes", "Yes", "Yes", "Yes", "...
## $ card_mailed                <chr> "Yes", "Yes", "Yes", "Yes", "Yes", ...
## $ first_batch_completed      <chr> "No", "Yes", "No", "Yes", "Yes", "N...
## $ orientation_completed      <chr> "Yes", "No", "Yes", "No", "Yes", "N...
## $ application_date           <date> 2018-10-01, 2018-10-01, 2018-10-01...
## $ first_batch_completed_date <date> NA, 2018-10-20, NA, 2018-10-06, 20...
## $ days_to_conversion         <dbl> NA, 19, NA, 5, 7, NA, 13, NA, 8, 9,...
## $ converted                  <chr> "Failure", "Success", "Failure", "S...
```

Project Phase 3.0: Business Understanding + Business Insights

This phase is to quickly derive: A baseline conversion rate from the control. And then, compare the baseline from control to our treatment conversion rate.

3.1 ASSESS BASELINE CONVERSION RATE

Table 7: Conversion outcomes for control group

converted	applicants	rate_of_outcome
Failure	7346	0.73
Success	2678	0.27

Baseline Conversion Rate: 0.27

Let's see how the treatment did against the baseline (control group)

3.2 COMPARE CONTROL (BASELINE) AGAINST TREATMENT

Let's look at the Control group to get a sense of the baseline rate.

Table 8: Conversion Rates by Group

group	converted	applicants	conversion_rate
control	Success	2678	0.27
treatment	Success	2115	0.43

Key Takeaway: Conversion rate by treatment (0.43) saw a 60% increase agains control (0.27)

3.3 QUICK LOOK TO SEE IF CATEGORY CHANNEL WAS SAMPLED EQUALLY

Table 9: Proportions sampled by group, channel

group	channel	n	pct_channle_by_group
control	job-search-site	1765	0.18
control	shopper-referral-bonus	1332	0.13
control	social-media	2998	0.30
control	web-search-engine	3929	0.39
treatment	job-search-site	860	0.17
treatment	shopper-referral-bonus	659	0.13
treatment	social-media	1429	0.29
treatment	web-search-engine	2010	0.41

Overall, this looks like they were equally sampled. This will build confidence in results.

3.4 QUICK LOOK AT CONVERSION RATES BY CHANNEL

This is a quick look at how conversion rates vary by channel, and by experiment group.

NOTE: This is preliminary.

My concern here is that other factors could influence our conversion rate.

Table 10: Conversion Rates by Control Group

Group	Channel	ConversionRate
control	shopper-referral-bonus	0.34
control	social-media	0.32
control	web-search-engine	0.25
control	job-search-site	0.16

Table 11: Conversion Rates by Treatment Group

Group	Channel	ConversionRate
treatment	shopper-referral-bonus	0.50
treatment	web-search-engine	0.45
treatment	social-media	0.39
treatment	job-search-site	0.38

Key Takeaway: Definitely variation in conversion rates by Channel, Group.

- See job-search-site: 0.16 for control & 0.38 for treatment.

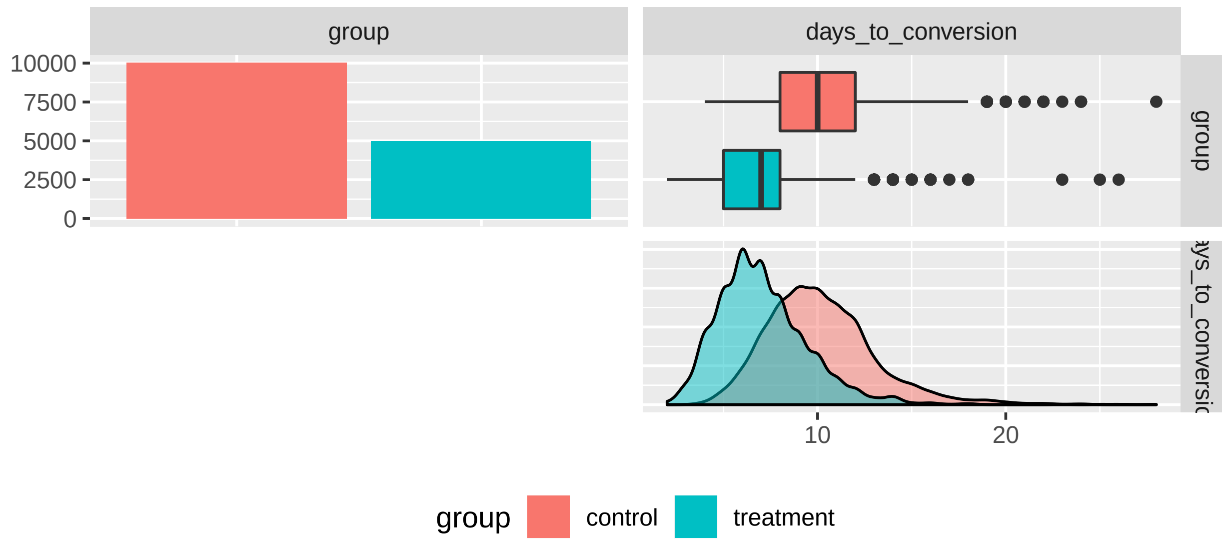
This variation indicates a more thorough investigation would help derive further insights to identify the primary drivers behind applicant conversion rates.

Project Phase 4.0: Data Understanding for Time-To-Conversion

Let's take a look at the distributions for time-to-conversion.

4.1 DOES THE TREATMENT SEE QUICKER START TIMES?

```
learning_data_tbl %>%
    select(group, days_to_conversion, group) %>%
    plot_ggpairs(color = group)
```




```
#learning_data_tbl %>% group_by(group) %>% summarize(mean_days = mean(days_to_conversion, na.rm = T))
```

Key Takeaway: Very Large differences between the two groups.

- We can say with confidence that initiating the background check earlier definitely leads to quicker start times.

Without doing a statistical test, I'd say these two distributions are VERY different and that they'd be significant if looked at closer.

Project Phase 5.0: Analyzing A/B Test Results

5.1 QUICK STATS TO GET SIGNIFICANCE

Everything so far points towards these results being significant (Results related to conversion rates).

I did get the counts below and use this calculator here to determine statistical significance

Site: https://neilpatel.com/ab-testing-calculator/

I also used this site recomended by a friend of mine who is a product analyst

- I used this to get the sample size and look at minimum detectable effect details

Site: https://www.evanmiller.org/ab-testing/sample-size.html

```
# Get counts by group
learning_data_tbl %>% count(group)
```

```
## # A tibble: 2 x 2
##   group         n
##   <chr>     <int>
## 1 control   10024
## 2 treatment  4958
```

```
# Get success and failure by group
learning_data_tbl %>%
    count(group, converted)
```

```
## # A tibble: 4 x 3
##   group     converted     n
##   <chr>     <chr>     <int>
## 1 control   Failure    7346
## 2 control   Success    2678
## 3 treatment Failure    2843
## 4 treatment Success    2115
```

5.1 MY EXPERIENCE WITH A/B TESTING

Profesionally I've not used A/B Testing but am fascinated by scientific experimenation.

- I'd be very interested in building expertise in this area.
- And in more sophisticated methods that complement understanding these systems.

Project Phase 6.0: Craft Plots for Presentation