case_7.3

May 29, 2020

1 How can I experiment with online travel deals to attract more customers?

[2]:

```python
import pandas as pd
import seaborn as sns; sns.set()
sns.set(rc={'figure.figsize':(15,11)})
import matplotlib.pyplot as plt
from IPython.display import IFrame
from math import floor
```

1.1 Case Introduction

Business Context. Your company TravelSmart is a growing company and is competing with a few big players in the market. As part of a marketing promotion to increase the user base, you have decided to give an extra 5% discount for users who book both flights and hotels at TravelSmart.com, rather than flight tickets alone. You are interested in learning how this marketing deal affects a variety of other metrics such as page views, users, sessions, revenue, etc. as compared to the status quo.

Business Problem. The company would like you to answer the following question: **"Should we be offering this 5% bundle discount promotion if our goal is to increase our purchasing customer count?"**

Analytical Context. The scenario above can be viewed as an **A/B test**. A/B testing is a way of determining whether a potential change ("Version B", as opposed to the existing "Version A") improves key business metrics, like sales or conversions.

The case is structured as follows: you will (1) define and state a precise hypothesis; (2) deploy the A/B test; (3) analyze the results of the test; and finally (4) systematically decide the next steps of whether you should go and implement a change or stay with the current version.

1.2 Designing the experiment: developing a concise hypothesis

Spaghetti is delicious, but even more delicious when the pasta is cooked just right. Many chefs use the "spaghetti test" – throw spaghetti at the wall and if it sticks, it's cooked. This is a great example of poor hypothesis design because it doesn't give any indication of when or why the spaghetti should be considered done.

Like cooking spaghetti, before proceeding with any experiment or test (in this case, an A/B test), we need to define a very precise hypothesis so that we have a context on which to evaluate the results of our experiment or test.

1.2.1 Exercise 1:

Suppose we wanted to move forward with the proposed 5% deal mentioned above. Which of the following hypotheses are you testing with your experiment?

(a) The 5% discount for users booking both flights and hotels will ultimately increase revenue.

(b) The 5% discount will bring in more customers.

(c) The 5% discount will incentivize users who book flights to also book hotels.

(d) The 5% discount allows for TravelSmart to increase base prices of hotels without affecting sales numbers, so the net revenue balances out but users are more satisfied by purchasing with a discount.

Answer.

(a) Incorrect. This is known as the ”spaghetti hypothesis”. There are too many uncontrollable variables in this statement and it doesn’t really tell us when our hypothesis is actually true. For example, if TravelSmart’s revenue changed at all, it could have been due to many different reasons (time of year, events, etc.). Generally, these hypotheses also have misleading implications. In this case, if revenues increased, we don’t know how the 5% discount actually affected the change in revenues. Would increasing the discount continuously increase revenues? In general, there is a lack of context to this hypothesis.

(B) Incorrect.

(C) Correct.

(d) Incorrect. Again, there are too many uncontrollable variables in this statement. We have no idea how price changes will be received by the user base.

1.3 Identifying control variables

Now that we have a concise hypothesis, we should think about what might go wrong and how we can prevent/mitigate potential issues. One major aspect to consider is **control variables**. Control variables are quantifiable elements that remain unchanged throughout an experiment. Identifying your control variables is very important for mitigating noisy data because slight variations in variables which you expect to remain constant throughout your experiment can greatly affect the outcomes being measured, in ways that have nothing to do with what you are specifically interested in testing.

For example, say Taylor goes on a diet and increases vegetable consumption for the next few weeks. Taylor wants to learn whether this will lead to weight loss. Taylor regularly takes a walk with his dog every morning. Over the next few weeks, Taylor feels that he is lacking the energy to get up in the morning for his daily walk and often does not go for a walk. Taylor ends up gaining a few pounds.

This is an example of a poorly designed experiment. The few extra pounds could have been due to a number of things, including lack of exercise. Thus it is not possible to learn if vegetable consumption contributed to weight loss from this experiment. If Taylor identified daily exercise to be a control variable, this experiment would have been a lot more successful.

1.3.1 Exercise 2:

Let's identify the potential control variables we need to consider in our experiment. Which of the following variables do you think are significant enough to control for? Select all that apply.

(a) Implementing user experience and visual improvements on the website that were already previously scheduled to deploy.

(b) A brand new, unique route that no other competitor offers.

(c) Performance of your sales team.

(d) Remodeling of the current engineering infrastructure.

Answer.

(a) Not really. From the context of this change, it looks like these UI/UX changes have already gone through A/B testing themselves or were going to get rolled out anyway. Therefore, by default they would already be incorporated into our test all around and we would not need to further account for it.

(b) Correct! This is definitely a variable we can control. If the route is significant enough that it is likely to materially impact the number of purchasing users, then it could confound our results. Let's assume that our manager deems that this route won't be taken frequently enough, so that we can move forward with testing this marketing plan.

(c) Perhaps. Although this directly affects the purchasing users count, we should not force the sales team to maintain a steady level of contribution as it would negatively impact the company (the company's goal should be to grow as fast as possible!). However, we should take this variable into consideration and perhaps implement measures to quantify its effects.

(d) Not really. Although this is a net benefit to the company, it doesn't affect our performance indicators for the time period during which this experiment is performed.

Control variables are extremely hard to get an exhaustive list of, but that's okay. In the real world, there are a lot of variables that affect everyone's behavior and are simply unavoidable. Choice (c) above is a perfect example of this. This does not mean such variables should be neglected, but they don't have to be strictly controlled; they merely need to be accounted for to a reasonable extent.

1.4 Defining KPIs

Once we have identified the relevant hypothesis and variables, we are nearly ready to deploy our experiment. The last thing to consider is choosing what variable we should be measuring so that evaluating the outcome of this experiment is very clear. These variables are called **Key Performance Indicators (KPIs)**. There are a lot of different metrics that we can consider, but all of them can be classified as either micro or macro conversions.

Macro conversions are the primary conversions for a particular service. In this case, macro conversions refer to the customer-accreting metrics of the website. Micro conversions relates to the smaller components that add up to macro conversions. The table below are a few examples of macro and micro conversions for TravelSmart:

Macro Conversions	Micro Conversions
Total Number of Ticket Sales	Views on a Specific Route
Membership Purchases	Time Spent on Landing Page
Sales and Leads Conversions	Number of Referral Accounts

Let's do the following experiment to better understand macro/micro conversions. Suppose your ran your deal for 60 days, and on Day 60 and we arrived at some data given in the following plot:

The following table summarizes the above plot. Let's analyze what's going on in this fictitious scenario.

[3]:

```python
def color_negative_red(value):
    if '%' not in str(value):
        return ''
    value = float(value.rstrip('%'))
    if value < 0:
        color = '#d65f5f'
    elif value > 0:
        color = '#5fba7d'
    else:
        color = 'black'
    return 'background-color: %s' % color

days = 60
df_experiment = pd.read_csv('micro_macro_dataset.csv')
pre_experiment = df_experiment.loc[1:days]
post_experiment = df_experiment.loc[days+1:2*days]
pre_means = pre_experiment.mean().apply(floor)
post_means = post_experiment.mean().apply(floor)
percentage_changes = ((post_means - pre_means)/pre_means*100).map("{:.2f}%".
 ↪format)

means = pd.DataFrame({
    "Pre-Experiment Mean": pre_means,
    "Post-Experiment Mean": post_means,
    "Percentage Changes" : percentage_changes
}).transpose()
means.style.applymap(color_negative_red)
```

[3]:

```
<pandas.io.formats.style.Styler at 0x111056d68>
```

Although we see that there are more visits to these specific pages, the number of tickets purchased and hotels booked did not change significantly. This reflects poorly on our experiment because it seems like there was no change to ticket sales, but we actually lost money by offering lower ticket and hotel prices. In this case, we say that the macro conversions were unaffected by this experiment and we deem the experiment unsuccessful.

Our major takeaway is that whenever we define an experiment, it is important to define and track macro conversions as our Key Performance Indicators. Micro conversions are also important, but the primary focus should be on macro conversions.

For our experiment, the most sensible macro conversion to use as a KPI is the number of distinct users who purchased a ticket or booked a hotel, because this is directly correlated to our user base count, which is exactly what we are trying to expand.

1.5 Treatment and control groups

All right, we are now ready to deploy our experiment. As implied in the name of this test, we have to split our user base into two groups, groups A and B. That way, we can apply the experiment on group B and see how it compares with the group with no changes, group A. We refer to group B as the treatment group (the ”treatment” here refers to the deal being offered) and group A as the control group.

1.5.1 Exercise 3:

Imagine that you have 100,000 users. How would you plan to deploy the experiment?

(a) You will start out by exposing 2% (2000) of users to the experiment. If the experiment sees no noticeable impact, ramp up the experiment to 50% (50000) users.

(b) You will start out by exposing 2% (2000) users to the experiment. If the experiment sees no noticeable impact, ramp up the experiment to 7% (7000) users, and later on to more and more users.

(c) You will expose 10% (10000) users to the experiment. This is a balanced size of the treatment group; it is just large enough to see the results of the experiment and exposes a small enough population in case the experiment is unsuccessful.

(d) You will give the deal to 50% (50000) users with the intention to compare against the control group. This increases the information gain and increases efficiency.

Answer.

(a) Incorrect. While starting out any experiment, it is most important to mitigate risk first on a small population before ramping up to a large population. Imagine our product contains a bug, causes degradations or is not very good – by exposing all of our users to it, we can cause severe long-term damage to our business. This answer gets that part right; however, after making sure things run smoothly on 2% of the users, it is good practice to ramp up in phases, before exposing half of your user base.

(b) Correct! This answer understands that you should ramp up your experiment in phases.

(c) Incorrect. Imagine our product contains a bug, causes degradations or is not very good – by exposing a significant fraction of our users to it can cause severe long-term damage to our business. That is why we want to perform a phased release, or an A/B test to make sure that only a subset of our users sees the first version of the product.

(d) Incorrect. While starting out any experiment it is more important to mitigate risk first on a small population, before information gathering. Imagine our product contains a bug, causes degradations or is not very good. By exposing all of our users to it we can cause severe long-term damage to our business. However, after making sure things run smoothly on 2% of the users, it is good practice to ramp up in phases, before exposing half of your user base.

1.5.2 Exercise 4:

Consider 2 options for deploying your experiment:

I. The first 2% of users get the new deal.
II. Randomly pick 2% of the users to get the new deal.

Which of the following statements is true?

(a) Option I is best, as it is the fastest way to learn about whether users like the deal

(b) Option II is better than option I because you are more likely to get a more representative group of users than option I

(c) Randomly picking users makes the whole experiment difficult to analyze because we do not have control over who gets the deal, and thus what one learns from this experiment is uncertain

(d) Option I and II are not really different unless we have more information

Answer.

(a) Incorrect. Suppose you roll out the experiment as U.S. EST 9 am. Then most of the users in the first 2% are going to be from the US, and specifically the East Coast. However, if a significant fraction of the users who are likely to book both the flight and travel are from the West Coast or from Asia, then you will exclude these groups.

(b) Correct! Randomization avoids the issue mentioned in the previous explanation for choice (a) of biased sampling.

(c) Incorrect. Even though randomization makes the outcome of a particular user getting the deal uncertain, it makes sure that the users who get the deal indeed are a good representation of our target user base. Later in this case study, we will see how to account for the randomness in the experiment using statistical methods.

(d) Incorrect. Even in the absence of any further information, randomization (Option II) is a better way to run your experiment.

1.6 Exploring Randomization

Randomly giving 2% of users the deal does not mean that exactly 2% of users at any given time have received the deal. As users come visit your website throughout the day, the actual percentage of users getting the deal fluctuates around 2%, as the following visualization shows:

```
[4]: IFrame("./histogram.html", height=450, width="100%")
```

```
[4]: <IPython.lib.display.IFrame at 0x125e7cba8>
```

1.6.1 Exercise 5:

Suppose you ran the experiment for a day, randomly giving the deal to 2% of the users. Suppose that you have a user base of over one million. However, at the end of the day, you observe that only about 1% of the users actually received the deal.

(a) 1% is well within the variation you would expect because of the randomization, and thus you can proceed with ramping up the experiment.

(b) 1% is definitely much lower than what you would expect, even accounting for the variation. Thus you might suspect a bug in your software and consult your engineering team to check if all of the intended users are getting the deal or not.

(c) Yes, it is outside the range of what you would expect the sample size to be, but you can still proceed with ramping up the experiment because there are still a large number of users who got the deal.

Answer.

(a) Incorrect. As the visualization above shows, 1% is much smaller than the actual number of users who should get the deal even accounting for randomness.

(b) Correct! As the visualization above shows, 1% is much smaller than the actual number of users who should get the deal even accounting for randomness. You can also check for this using various statistical tests. If the fraction of users getting the deal is much different from the expected fraction, it is worth checking with the engineering team to see if there is a bug in your software before ramping up the experiment.

(c) Incorrect. While it is true that you have some information from the users who got the deal, if the fraction of users getting the deal is much different than the expected fraction, there may be other bugs in your deployment system, so you should do another check to make sure there isn't a more serious error going on.

1.7 Ramping up the experiment

Now that you have run your experiment for 2% of the user base for a couple of days, it is the time to ramp up your experiment. Before discussing how and when to ramp up, let's first discuss how to understand the results of an A/B experiment.

Let's suppose you decide to randomly give the deal to 10% of users and have run it for a couple of days. If you are interested in purchasing users, you may calculate the difference in percentage of purchasing users between the two groups. Suppose we have run the experiment and observed the following outcome for the difference in percentage of purchasing users between group A and group B:

95% confidence interval

(0.7%, 1.2%)

p - value for the hypothesis that there is no difference in purchasing users between the two groups

0.03

The 95% confidence interval indicates that it is 95% likely that the percentage of purchashing users in group B is between 0.7% and 1.2% greater than the percentage of purchashing users in group A. In this case, 0% is outside the interval, and we can conclude that the percentage of purchasing users is higher for users who are given the deal. Furthermore, the p - value of 0.03 indicates that this result is significant.

If the experiment is run for longer, the confidence interval will be narrower. This is because the number of users given the deal increases with time. This increase in information leads to more precision in our estimate.

1.8 Conclusions

In this case, we learned about conducting A/B tests online. You learned about how to properly define a hypothesis, control variables, KPIs, and finally set up and analyze the results of the experiment. We learned in this case that the deal did indeed increase the percentage of purchasing users by a statistically significant 1%.

1.9 Takeaways

A/B tests are a powerful tool to conduct rapid experimentation and quickly learn about the preferences of your users. However, any A/B test you run must ensure proper randomization. It ensures that you will not accidentally exclude any user base and derive conclusions based on a biased sample. With an A/B test, one usually starts the experiment on a small percentage of users so as to mitigate risk. After this initial step, the experiment can be ramped up.