Report on user engagement A/B testing CTA(Bridebook)

By: Hamza Jibran Butt
Ipynb can be viewed at:
https://colab.research.google.com/drive/1DHrMuNEKHQgrd2g2qTmcyZb-X-szb8GH?usp=sharing
(A downloaded .py version of this script is available as attachment for code reference)

Design of Experiment and formulating a Hypothesis:

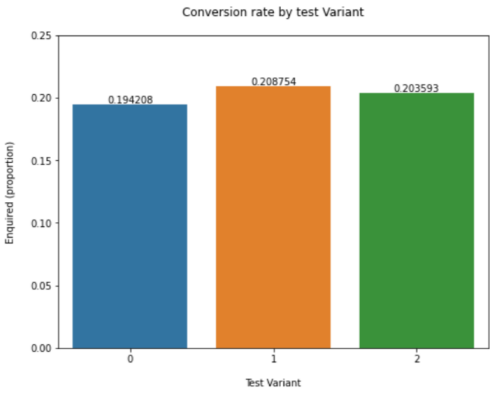
The objective of the A/B test is to try and increase user engagement with the end goal of increasing the number of registered users that make enquiries to venues. To this end we have a control webpage and two variant webpages.

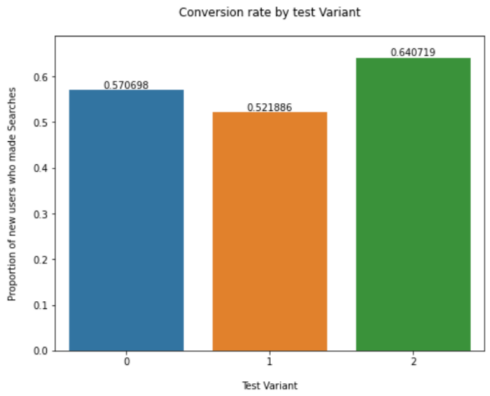
For this, the approach taken is to compare each hypothesis with the control variant. If both are better than the control then both variants will be compared for the best option as we have 2 variants of the same control we want to compare.

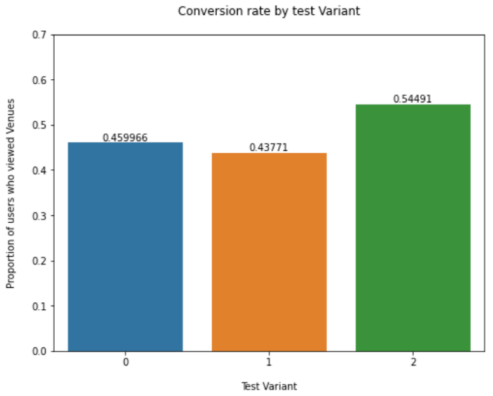
To compare conversion rates for engagement.Our main feature in the dataset was 'venEnq' i.e. How many enquiries to venues a registered user made. However other metrics for user engagement also shed some interesting light. These are the numbers of 'Venue Searches' and 'Venues Viewed' in the first week.

The outcome of the test?

Initial conversion analysis was done on samples for the control and test variants. Conversions for whether a user enquired, Searched or viewed venues in the first week. The results can be observed in the following graphs:







From an initial analysis we see that in the control, 19.4% of the users from the control variant enquired about a venue. Both variants 1 and 2 scored better in terms of users enquiring a venue; 20.8% and 20.3% based on this sample.
Important user engagement metrics such as Venue 'Searches' and 'Views' see an increasing conversion rate for Variant 2; 64.0% and 54.4% as opposed to control-test that had conversion rates of 57.0% and 45.9% for Searches and views.

This looks like the variants are performing better but are they significantly different from the control test for us to adopt. For this a z-test to calculate the p-value and Confidence intervals at 95% are assumed as standard practice, hence alpha = 0.05. We will conduct an A/B test between the control and each variant.

Z-test for venue enquiry between Control and Variant1:

- P-value = 0.608
- CI for control group = [0.162,0.226]
- CI for Variant 1 = [0.162, 0.255]

Z-test for venue enquiry between Control and Variant2:

- P-value = 0.788
- CI for control group = [0.162,0.226]
- CI for Variant 2 = [0.143, 0.265]

Both these tests suggest that there is not enough of a significant difference for us to consider either of the variants to perform better than control. From the Confidence Intervals we learn that Variant 2 may have the best or worst performance whereas the CI for Variant1 compared to control suggest that Variant 1 would not perform worse than control but not significantly better either.

Business Recommendations to suggest based on this?

My recommendation would be to continue the A/B test to get more data points to make sure that there is a significant difference in outcome or not in different variants. The results do indicate an improvement in user enquiries to venues. I believe that with more data it could become statistically apparent that Variant 2 is significantly better than the control.

Additional questions based on more time/data

In this test I considered the conversion of a registered user based on whether they made an inquiry to a venue. Although we were unable to determine a significant difference between the test variants and control. Initial conversion calculations were done for whether users viewed venues or made searches which show promising improvement.

Given more time I would have calculated a correlation matrix to observe how these features relate with users making enquiry and based on that may consider doing a z-test and checking for a significant difference in co-relating features.

Any other comments or improvements.

According to the dataset we accumulated a total of 1051 records in a period of two weeks. 587, 297 and 167 users for Control, Variant1 and Variant2 respectively. I believe the test does not have sufficient records for the variants. Ideally we would like to have the same ratio of users for each test variant. To that end I would recommend that this test would ideally extend to 1 month.

Apart from that the dataset was very clean and thorough and sufficient features were present in the dataset.