Oct 29 2023

GloBox A/B testing report by Tsedal

Summary

GloBox is an online e-commerce company, primarily known amongst its customer base for boutique fashion items and high-end décor products. However, the food and drink offerings have grown tremendously in the last few months, our company wants to bring awareness to this product category to increase revenue. Therefore, a sample is taken and we conducted this A/B testing experiment.

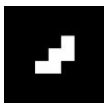
The objective of this A/B testing experiment was to assess the impact of a redesigned landing page with food and drink banner compared to the existing landing page. Two test groups were created for this experiment. The first group represents the control group, which consists of the existing landing page design. The second group represents the treatment group, featuring a food and drink banner on the landing page. Both variations were randomly assigned to website visitors.

To measure the effectiveness of the variations, two test metrics were used: conversion rate and average amount spent per user. A statistic for A/B testing used is an inferential statistic. It is the key to establishing whether the new food and drink banner is leading to changes in the test metrics.

This report contains the methodology, the results, and the recommendation. The appendices contain raw data, SQL query, statistical analysis spreadsheet links, and visualization links.

Methodology

The experiment was conducted from January 25, 2023, to February 06, 2023. A sample size of 48,943 participants was taken. The sample size for control group has 24,343 participants, while treatment group has 24,600 participants.



This A/B testing experiment utilized a randomized controlled trial design to evaluate the impact of two variations on average amount spent and conversion rates. The methodology of the A/B testing is as follow:

Step 1: Extracting data

Knowing that the conversion rate is the number of successful conversions (users who purchased) divided by the total number of users and average amount spent per user is sum of spent amount divided by total number of users. to join or combine data from multiple tables. SQL (Structured Query Language) was used to retrieve the relevant data by joining data from multiple tables. It was utilized to clean the data. This includes removing duplicates using DISTINCT, handling missing values with COALESCE function. Other key words and aggregate functions like COUNT, SUM, AVG, and GROUP BY was employed to calculate metrics and generate summary statistics for the A/B testing variations.

Step2: Calculating statistical analysis

This step is important to construct and calculate the important parameter of a hypothesis test and confidence interval using google spreadsheets.

- ❖ Hypothesis testing

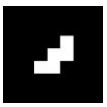
The p-value represents the probability of observing the data, assuming the null hypothesis is true. If the p-value is less than the chosen significance level, reject the null hypothesis and accept the alternative hypothesis.

For conversion rate

Since, a significance level of 0.05 (or 5%), a normal distribution and pooled proportion for the standard error is used. Two sample two tailed z-test with pooled proportion is the appropriate statistical test.

- ✓ H0: There is no significant difference in conversion rates between test groups.
  : i.e P1-P2=P0
- ✓ H1: There is significant difference in conversion rates between test groups.
  : i.e P1-P2<>P0

For average amount spent per user



Since, a significance level of 0.05 (or 5%), a t-distribution and unequal variance is used. Two sample two tailed t test with unequal variance is the appropriate statistical test.

- ✓ H0: There is no significant difference in average amount spent per user between test groups. i.e μ1-μ2=μ0
- ✓ H1: There is significant difference in average amount spent per user between test groups. i.e μ1-μ2<>μ0

❖ Confidence interval

Once the significant difference between the variants is determined, the next step is to calculate the confidence interval to estimate the magnitude of the difference.

For conversion rate

Having, a confidence interval of 0.95, a normal distribution and unpooled proportion for the standard error, two sample two tailed z interval with unpooled proportions is the appropriate method.

For average amount spent per user

Since confidence interval of 0.95, a t distribution and unpooled variance for the standard error is used. Two sample two tailed t interval with unpooled variance is the appropriate method.

Step 3: More analysis

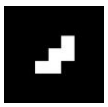
It is calculated online using Statsig sample size calculator for conversions and means.

Check for Novelty effects

The experiment was conducted over a one-week period. User might behave differently when the treatment is new. If we are going to notice a novelty effect, that means the effectiveness of the banner is short-lived.

Power analysis

Conducting power analysis ensures that the A/B test has sufficient statistical power to detect meaningful differences in conversion rates, allowing us to make accurate conclusions about the impact of the redesigned checkout process.



Practical Significance

Considering practical significance alongside statistical significance is crucial for determining whether the observed difference in conversion rates is practically meaningful and has practical implications for the website's performance and business goals.

Step 4: Data Visualization

This is helpful to present the data found from SQL extraction and spreadsheet analysis to generate meaningful insights from the collected data. Tableau features utilized, such as creating calculated fields, and creating parameter utilizing interactivity for exploration like filters, dual axis, reference band is implemented to enable users to analyse the data from different perspectives or segments. Charts like bar, line, histogram are also used to visualize the relationship of test metrics per group, gender, device, and country.

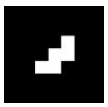
Results

The results from the spreadsheet of the A/B testing experiment are presented below:

1. Conversion Rate:

Control group achieved a conversion rate of 3.92%, while treatment group achieved a higher conversion rate of 4.63%. Based on hypothesis testing a two-sample z-test was performed to compare the conversion rate of control and treatment group, resulting in a p-value of 0.000111. The observed difference of 0.71% was statistically significant since 0.000111 < 0.05, indicating that we REJECT the null hypothesis i.e., the banner experience led to a significant improvement in the conversion rate compared to the existing design. This demonstrates the effectiveness of the food banner version in driving more users to complete the purchase. The confidence interval for the difference in proportion between control and treatment group ranged from 0.0035 to 0.0107, suggesting that the population conversion rate will likely to be within this interval with 95% confidence.

With Baseline parameter of 3.92%, practical significance of 5% equal number of sample size and statistical power 80%. We need 153,900 sample size each of total sample size 307,800.



2. Average amount spent per user:

Variation A had an average amount spent per user3.37$, whereas Variation B had 3.39$. Based on hypothesis testing the observed difference of 0.02$ is statistically insignificant(P>0.05). Based on hypothesis testing a two-sample t-test was performed to compare the average amount spent per user of control and treatment group, resulting in a p-value of 0.94. The observed difference of 0.02$ was not statistically significant since P-value<0.05, indicating that we FAIL TO REJECT the null hypothesis i.e., the banner experience doesn't lead to a significant improvement in the average amount spent per user compared to the existing design. The 95% confidence interval for the difference in proportion between control and treatment group ranged from -0.4387 to 0.4714.

With Baseline parameter of 3.37$, pooled standard deviation of 25.8, practical significance of 5% equal number of sample size and statistical power 0.8. We need 153,900 sample size each of total sample size 362,342.

Recommendation

Although the increment in the conversion rate is a promising result, we didn't see enough improvement in the other success metrics to be confident in releasing the landing page with banner experience as the new default design for the user. Because the banner takes up high-value real estate on the main page. We need more revenue than conversion rate. We need to CONTINUE ITERATING. Since we saw only one of the metrics is statistically significant. Since both of the metrics show a statistically significance less than practical significance 5%. This makes our test practically insignificant. In this case, our test is not powerful enough to actually detect that difference. So, we need to analyse the minimum detectable effect sample size. The sample size causes the risk of sampling error and reduces the statistical power of the results. So, we have to re-run the test and collect more data since this system is automated and doesn't take much cost. The limitation here will be it takes time. We should consider launching if one or both metrics show a statistical increase that is above the practical significance.

Appendices

Appendix A: User level aggregate table

This appendix contains the extracted data from the raw data using SQL editor called beekeeper. The Null values are handled. It contains user ID, country, gender, device, test group, conversion, spent per user. It can be found in the spreadsheet link below:

https://docs.google.com/spreadsheets/d/1w_gA0Elin0bOVc2pkRgR9wqMlr_I3pysSzysF3n_qN4/edit#gid=2099685271

Appendix B: SQL code to extract the User level aggregate table and for novelty test

```sql
SELECT uu.id,
       COALESCE(uu.country,'unknown') AS country,
       COALESCE(uu.gender,'unknown') AS gender,
       COALESCE(aa.device,'unknown') AS device,
       COALESCE(gg.group,'unknown') AS test_group,
       CASE WHEN SUM(aa.spent)>0 THEN '1'
       ELSE '0'
       END AS conversions,
       COALESCE(SUM(aa.spent),0) AS spent_per_user
FROM users uu
LEFT JOIN activity aa
ON uu.id=aa.uid
LEFT JOIN groups gg
ON uu.id=gg.uid
GROUP BY uu.id,aa.device,gg.group
```

id	country	gender	device	test_group	conversions	spent_per_user
1000000	CAN	M	unknown	B	0	0
1000001	BRA	M	unknown	A	0	0

Appendix C: Visualizations

This appendix contains visualizations to complement the main report. It is done with a Business Intelligence tool called Tableau. It includes bar charts showing the test metrics per test group, gender, device, country. For the further analysis line chart are used to check novelty effect. These visualizations enhance the clarity and understanding of the experiment's results. The links are provided below:

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Pertestgroup?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Spentdistribution?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Avg_Perdevice?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Conv_perdevice?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Avg_Pergender?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Avg_Pergender_1?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Percountry?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Conv_C_I?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Avg_C_I?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/GloBox_16982716796330/Dashboard1?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/Noveltyeffect_16983541154600/Sheet1?publish=yes

https://public.tableau.com/app/profile/sedal.admasu/viz/Noveltyeffect_16983541154600/Sheet2?publish=yes

Appendix D: Statistical Analysis

This appendix provides the spreadsheet statistical analysis conducted. It includes information on the hypothesis tests used, assumptions made, p-values, and confidence intervals. The appendix also contains additional statistical outputs, such as effect sizes or statistical power calculations, that were considered during the analysis. It can be found in the conversion rate and average amount spent per user sheet

https://docs.google.com/spreadsheets/d/1w_gA0Elin0bOVc2pkRgR9wqMlr_l3pysSzysF3n_qN4/edit#gid=2099685271