A/B Testing: A Data-Driven Approach to Optimize Marketing Campaigns

Akhmad Taufik Ismail

April 5, 2023

Introduction and Background

Marketing campaigns play a pivotal role in promoting products and establishing a connection with potential customers. Nonetheless, the success of these campaigns is not uniform across the board, as it can be impacted by several factors, including message type, content, design, and audience preferences. Hence, it is imperative for marketing firms to employ data-driven techniques to test and optimize their campaigns to achieve their desired objectives.

The process of A/B testing, which is also commonly referred to as split testing or bucket testing, is a viable means of achieving this objective. Essentially, A/B testing is a method utilized to assess the effectiveness of various elements of a marketing campaign, including ads, landing pages, and other associated components. In order to carry out an A/B test, a specific aspect of the campaign is modified, and both versions are executed simultaneously, with performance data being collected. By analyzing the test results, it is possible to identify the superior version and subsequently implement any necessary changes [4].

In the present circumstance, the marketing approach of the company is restricted to Public Service Announcements (PSAs), which are informational messages intended to enlighten the public about a social concern. The organization's novel approach is to utilize ads, which are persuasive messages designed to influence the public into purchasing a product or service. Nevertheless, the company is uncertain about the efficacy of ads in enhancing user conversions. Consequently, the company has been requested to conduct A/B testing to compare the impact of PSA and advertisements.

The challenge lies in determining the optimal marketing campaign by selecting the most suitable message type (PSA or ads) tailored to the target audience. The objective is to boost the count of users who convert.

This experiment will employ A/B testing methods, which are a technique for comparing two versions of a marketing element and determining which performs better. A/B testing can assist marketers in determining what resonates best with their target audiences and improving campaign effectiveness [1].

This report will summarize the results of an A/B test conducted by a marketing firm to compare two different message types: public service announcements (PSA) and advertisements. We will go over the test's methodology, data analysis, and conclusions.

Setting Up Problem

Experiments Goal

In order to increase revenue, the company seeks to determine the most effective marketing approach by conducting an A/B test to comprehensively evaluate the impact of two distinct types of marketing messages on user behavior. Presently, the company utilizes Public Service Announcements (PSA) to educate users about its products and services, but it wants to explore the potential of ads as an alternative method to attract and retain customers. The A/B test will involve randomly assigning users to either a PSA or an ad

group, and then measuring their conversion rates. In this context, conversion rate pertains to the percentage of users who take a desired action, such as subscribing to a newsletter, purchasing a product, or downloading an application. The experiment aims to test the hypothesis that ads will produce a higher conversion rate than PSA, which will provide valuable insights into the optimal marketing strategy for the company to increase revenue.

Metrics

A driver metric can be defined in this experiment as the key metric that the company seeks to improve or increase in order to achieve its primary goal of increasing revenue. In this case, the conversion rate could be the driver metric because it is directly related to revenue generation. The company's goal is to increase conversion rates by implementing the best marketing strategy, which will be determined through A/B testing of PSA and ads.

A guardrail metric, on the other hand, is one that the company will monitor to ensure that the marketing strategy has no unintended negative consequences for other important aspects of the business. A guardrail metric for this experiment could be the bounce rate, which is the percentage of users who leave the website without taking any action. If the company notices a significant increase in bounce rate after implementing the new marketing strategy, this could indicate that the ads are not resonating with the target audience, and the company may need to adjust its approach to avoid any negative impact on user behavior.

Variants

1. **Control Group**: The control group will be exposed to PSA, which is the current marketing strategy used by the company. The control group will be used as a baseline to compare the performance of the test group.
2. **Treatment Group**: The treatment group will be exposed to ads, which is the new marketing strategy that the company is considering. The test group will be used to determine the effectiveness of the new marketing strategy.

Hypothesis

- H_0 : The conversion rate of the control group (PSA) is equal to or lower than the conversion rate of the treatment group (ads).
- H_1 : The conversion rate of the treatment group (ads) is higher than the conversion rate of the control group (PSA).

In simpler terms, the null hypothesis ($H0$) for this experiment is that there is no difference in conversion rates between the control group (PSA) and the treatment group (ads). The alternative hypothesis ($H1$) is that the conversion rate of the treatment group (ads) is higher than the conversion rate of the control group (PSA).

Design Experiments

Randomization Unit

In order to guarantee that the experiment generates precise and significant results, it is imperative to perform the A/B test at the user level. This implies that each individual user will be randomly allocated to either the control group, which will be exposed to the PSA message, or the treatment group, which will receive the ad message. By employing this randomization process, it ensures that there is no partiality in the allocation of users to the distinct groups, and any observed differences in the conversion rates between the two groups can be credited to the message type, and not to any other extraneous factors.

Target of Randomization Unit

In order to achieve the objective of determining the impact of message type on user behavior, targeting the randomization unit at the individual user level is the most appropriate approach for this experiment. This is because user behavior is influenced by their individual preferences and interests. By randomly assigning each individual user to either the control group or the treatment group, the experiment can be conducted in a way that is consistent with the company's goal of identifying the optimal marketing strategy.

Sample Size

The sample size for the experiment must be calculated to ensure that the results are statistically significant. The following formula can be used to calculate sample size [5]:

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

- n: Sample size
- σ: Standard deviation
- $z_{1-\alpha/2}$: Critical value for $\alpha/2$
- $z_{1-\beta}$: Critical value for β
- δ: Minimum detectable effect

Because the standard deviation of the conversion rate is unknown, it can be estimated using the Bernoulli distribution approach, which is given by:

$$\sigma = \sqrt{\hat{p}(1-\hat{p})}$$

- $\hat{p}$: Estimated conversion rate / baseline of conversion rate.

Therefore, to calculate the sample size for the experiment, an estimate of the baseline conversion rate is required, which can be obtained from historical data or industry benchmarks. Once the baseline conversion rate is estimated, the minimum detectable effect and the desired levels of statistical significance (i.e., α and β) can be determined. Using these values, the sample size can be calculated using the formula provided above. This ensures that the experiment is conducted with a sufficient sample size to yield statistically significant results.

Significant Level (α)

In order to ensure that the results of the experiment are reliable and meaningful, it is essential to set a significance level that is appropriate for the study. The significance level, denoted by α, is the probability of obtaining a result that is as extreme or more extreme than the one observed in the experiment, assuming the null hypothesis is true. In this experiment, a significance level of 5% has been chosen, which means that there is a 5% chance of obtaining a result as extreme or more extreme than the one observed, even if the null hypothesis is true.

The choice of a significance level of 5% is common in A/B testing because it strikes a balance between detecting meaningful differences between the control and treatment groups, while minimizing the likelihood of false positives. This value is widely used in industry and has been shown to be effective in a variety of experimental settings. To ensure that the results of the experiment are statistically significant, the p-value, which represents the probability of obtaining a result as extreme or more extreme than the observed result, must be less than or equal to α.

Power Level ($1 - \beta$)

The industry standard for power is 80%, which is explained as the probability of detecting a real difference between variants. One reason for choosing 80% power is that it balances the risks of type I and type II errors [2]. However, the appropriateness of 80% power may vary based on the context and limitations of the experiment.

Another reason for selecting 80% power is that it provides a reasonable level of confidence and precision for most AB testing applications. Kohavi et al. explain that if an experiment is repeated 100 times with different random samples, 80 of them will detect an effect of the same or larger magnitude as the true effect [6]. This suggests that researchers can gain meaningful insights from AB testing without requiring overly large or expensive samples. However, the appropriateness of 80% power may vary based on the context, goals, and limitations of the statistical methods employed in the experiment.

Standard Deviation (σ)

The Bernoulli distribution is commonly utilized for modeling binary outcomes that involve only two possible outcomes. In the context of the Bernoulli distribution, the standard deviation can be calculated as follows:

$$\sigma = \sqrt{\hat{p}(1 - \hat{p})}$$

This statistical distribution is particularly useful for measuring the variation of binary data, which is limited to two values. The selection of the Bernoulli distribution for determining standard deviation values in proportion data is therefore a suitable approach for capturing the variability of binary data.

In an AB testing experiment, the Bernoulli distribution can be used to model the binary outcome of interest, such as the number of conversions in response to two different versions of a marketing campaign (i.e., PSA and Ads). The standard deviation value of the Bernoulli distribution can be determined from the observed data collected during the experiment.

Differences Between Control and Treatment Groups (δ)

The minimum detectable effect, denoted by δ, is the smallest difference between the control and treatment groups that the experiment can detect with a certain level of statistical significance. In this experiment, the minimum detectable effect is set to 0.01, which means that the experiment can detect a difference of 1% in the conversion rate between the control and treatment groups.

Experiment Duration

To ensure that the results of the experiment are reliable and accurate, the duration of the experiment has been set to two weeks. The reason for this duration is because the experiment is conducted on a website or app, and it is observed that the experiences seasonality in the number of visitors. Therefore, to avoid any bias in the results due to variations in the number of visitors, the experiment is being conducted during the peak season. In this way, the experiment can be conducted under conditions that are more representative of the typical user experience, and the results obtained can be relied upon to make data-driven decisions [8]. The two-week duration provides sufficient time to collect data from a large number of users in both the control and treatment groups, while minimizing any potential external factors that could influence the results.

Analyzing and Interpreting the Data

AA Test

It is crucial to perform an A/A test prior to an A/B test to establish a firm baseline for the experiment. The purpose of conducting an A/A test is to identify any disparities in the data or flaws in the testing setup, such

as a bug in the tracking code or a sample population bias. This test can also help validate the statistical methods used and the reliability of the findings [7].

If an A/A test reveals a statistically significant difference between the control and treatment groups, it suggests that there may be something incorrect with the testing setup or that the sample population is biased. Consequently, incorrect conclusions and ineffective optimization may result from the A/B test. Conducting an A/A test can guarantee that the testing setup is functioning correctly and that the sample population is impartial, which can enhance the precision and dependability of the A/B test results.

Performing an A/A test before conducting an A/B test is essential because it establishes a strong foundation for the experiment, recognizes any anomalies in the data or testing setup, and improves the accuracy and consistency of the A/B test results.

Minimum sample size is a crucial factor in experimental design as it helps to ensure that the experiment has adequate power to detect the effect size of interest. Initially, the minimum sample size for an AB test was calculated to be 1664 users per group. However, after conducting an AA test, the control 1 and control 2 groups were found to be statistically different, indicating a bias in the experimental design.

To address this issue, the experiment was redesigned based on significant level, power, standard deviation, and minimum detectable effect, resulting in a new minimum sample size of 3334. The AA test was then repeated, and the results showed that there was no longer any bias in the experimental design.

Eliminating bias is an essential step to ensure the accuracy and reliability of the experiment, which can then proceed to the AB test. The AB test allows for a comparison between the control group and the treatment group to determine which version is more effective in achieving the desired outcome. By first conducting the AA test to establish a baseline and identify any potential bias in the experimental design, and then addressing any issues before conducting the AB test, the experiment can be conducted with a higher level of confidence in the accuracy and reliability of the results.

AB Test

Data Quality

The implementation of robust data quality processes is of paramount importance to guarantee the credibility and precision of the data collected during the experiment. The data quality process involves meticulous procedures for acquiring, archiving, and scrutinizing data to confirm that it is devoid of errors and discrepancies. This is particularly significant in AB testing as it fosters the assurance of reliable and meaningful outcomes of the experiment.

In this experiment, the data quality has been exemplary, as no invalid data, missing values, or duplicated data were detected. Thus, with the successful completion of this initial stage of verification or sanity check, the subsequent stage will now commence.

Sample Ratio Mismatch

Sample Ratio Mismatch (SRM) is a common issue that can arise during A/B testing. This happens when the allocation of users between the test groups is significantly different from the expected allocation proportions, also known as the sample ratio [3]. Detecting SRM is important since it can affect the reliability of A/B test results. A simple chi-squared test can be used to identify SRM, but it only serves as a diagnosis. The real challenge in dealing with SRM is figuring out where along the often-lengthy process the samples became skewed [9].

To detect SRM, the first step is to define the null and alternative hypotheses (H_0 and H_1). The null hypothesis assumes that no SRM has been detected, while the alternative hypothesis posits that SRM has been detected. Once the hypotheses have been established, the next step is to calculate the chi-square statistic. The chi-square statistic compares the observed and expected frequencies and determines whether the difference is statistically significant.

The chi-square statistic is calculated as follows:

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

The third step is to define decision rules. In statistical test decisions, the comparison of the chi-square statistic with the critical value or the comparison of the p-value with alpha can be used. If the calculated chi-square value exceeds the critical value, the null hypothesis is rejected, and it is concluded that SRM has been detected. Similarly, if the p-value is less than alpha, the null hypothesis is rejected, and SRM is detected.

In the current example, the alpha value is set to 0.01, the p-value is 1, the chi-square test value is 0, and the critical value is 6.63. Based on these values, the null hypothesis is not rejected, indicating that there is no evidence of SRM. However, it is essential to note that detecting SRM is only the first step in the process.

Result

The workflow for conducting AB testing consists of several steps. Firstly, the problem to be solved is defined and the experiment is designed. The experiment is then run for a specified duration, which in this case was two weeks. Once the data is collected, it is loaded and a sanity check is performed to ensure there is no invalid, missing, or duplicate data. The experimental data is then divided into two groups: the control group and the treatment group.

Next, the minimum sample size is calculated based on the criteria specified in the experimental design. In this case, the minimum number of samples required for each group was determined to be 3334, resulting in a total sample size of 6668 users. The next step is to sample users for each group based on the minimum number of samples required.

The evaluation metric that has been determined in the experimental design is then calculated, which in this case was the conversion rate. The conversion rate for the control group was found to be 1.8% while for the treatment group it was 2.7%. The lift over baseline was calculated by subtracting the conversion rate in the control group from the conversion rate in the treatment group, resulting in a lift value of 0.95. While the lift is numerically significant, it needs to be verified statistically.

To do this, the null hypothesis and alternative hypothesis are defined. The null hypothesis is that the conversion rate in the treatment group is equal to the conversion rate in the control group ($H_0 : CVR_{treatment} = CVR_{control}$), while the alternative hypothesis is that the conversion rate in the treatment group is greater than the conversion rate in the control group ($H_1 : CVR_{treatment} > CVR_{control}$). The decision rule is then determined based on the z-statistic and p-value. If the z-statistic is greater than the z-critical value, then the null hypothesis is rejected. Similarly, if the p-value is less than alpha, the null hypothesis is rejected.

With alpha set to 0.05, the p-value was calculated to be 0.0043, the z-statistic was found to be 2.62, and the z-critical was 1.96. Since the z-statistic is greater than the z-critical value and the p-value is less than alpha, the null hypothesis is rejected, indicating that the lift in the conversion rate for the treatment group is statistically significant.

However, the calculated lift value of 0.95 and the associated p-value and z-statistic suggested that the treatment group had a significantly higher conversion rate than the control group. The confidence interval now provides a range of plausible values for the difference in conversion rates, which can help to quantify the uncertainty of the estimate.

The lower limit of the confidence interval is 0.0024, which means that we can be 95% confident that the true difference in conversion rates between the treatment and control groups is at least 0.0024 or higher. The upper limit of the confidence interval is 0.168, which means that we can be 95% confident that the true difference in conversion rates between the treatment and control groups is no more than 0.168 or lower.

Overall, these results suggest that the treatment group has a higher conversion rate than the control group, with a difference ranging from at least 0.0024 to no more than 0.168, and the observed lift value of 0.95 falls within this range. Therefore, the results of the AB test are statistically significant and provide evidence to support the adoption of the treatment.

Conclusion and Recommendation

Conclusion

The experiment was designed to test the effectiveness of a new feature on the website of a digital marketing agency. The new feature was designed to increase the conversion rate of the website by providing users with a more personalized experience. The experiment was conducted for a period of two weeks, during which the conversion rate of the control group was found to be 1.8% while the conversion rate of the treatment group was 2.7%. The lift over baseline was calculated to be 0.95, which was statistically significant.

The confidence interval was calculated, and the lower limit of the confidence interval was 0.0024, with an upper limit of 0.168. This means that we can be 95% confident that the true difference between the two groups' conversion rates is between these two limits.

Based on these findings, it is possible to conclude that the ad-based marketing campaign was more effective in increasing conversion rates than the PSA-based campaign. As a result, a business decision could be made to devote more resources to the advertising campaign and adjust the marketing strategy accordingly.

Recommendation

In future research, it is recommended to following:

1. Conduct longer experiments: The current experiment only ran for two weeks. Conducting experiments over a longer period of time can help to identify any potential long-term effects and to observe how the results change over time.
2. Increase the sample size: Although the minimum sample size was met, increasing the sample size can provide more accurate results and can help to reduce the effects of noise in the data.
3. Test different ad types and platforms: Instead of using only one type of ad, testing multiple ad types on different platforms can help to identify which ad type or platform performs the best.
4. Use a different evaluation metric: In this experiment, the conversion rate was used as the evaluation metric. However, other metrics such as click-through rate or engagement rate may provide different insights into the effectiveness of the ads.

By taking these recommendations into account, future experiments can be designed and conducted to provide more accurate and reliable results, which can ultimately lead to better business decisions.

References

[1] Tom Baldwin. *How A/B Testing Can Improve Your Marketing Campaigns*. www.linkedin.com, Mar. 2023. URL: https://www.linkedin.com/pulse/how-ab-testing-can-improve-your-marketing-campaigns-tom-baldwin (visited on 04/02/2023).

[2] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum Associates, 1988. URL: https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.

[3] Emily Healy. *Sample Ratio Mismatch: What Is It and How Does It Happen?* AB Tasty, Nov. 2022. URL: https://www.abtasty.com/blog/sample-ratio-mismatch/ (visited on 04/05/2023).

[4] *How to Use A/B Testing to Maximize Marketing Campaigns*. Lotame, Mar. 2019. URL: https://www.lotame.com/how-to-use-a-b-testing-to-maximize-marketing-campaign-performance/ (visited on 04/02/2023).

[5] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020. DOI: 10.1017/9781108653985.

[6] Ron Kohavi et al. "Online Controlled Experiments at Large Scale". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2013). DOI: 10.1145/2487575.2488217. (Visited on 04/04/2023).

[7] Mani Makkar. *What Is an A/A Test? Why Should You Care? — VWO*. VWO Blog, Dec. 2019. URL: https://vwo.com/blog/aa-test-before-ab-testing/ (visited on 04/04/2023).

[8] Khalid Saleh. *How Long Should You Run an A/B Test for and How to Calculate...* Invespcro, Nov. 2017. URL: https://www.invespcro.com/blog/how-long-should-you-run-an-ab-test-for/ (visited on 04/02/2023).

[9] Andre Ye. *A/B Testing's Worst Enemy: Sample Ratio Mismatch*. DataSeries, June 2020. URL: https://medium.com/dataseries/a-b-testings-worst-enemy-sample-ratio-mismatch-bf95eb0ab1c7 (visited on 04/05/2023).