**Introduction**

Vungle is one of the key leading players in in-app video advertising, founded in 2011. Vungle had mobile apps embedded with video ads that encouraged users to install more apps. In this process, there were four major stakeholders: the user using the mobile device, the publisher of the app, the sponsor of the video ad, and the platform that matched the ads.

It has grown into one of the most influential platforms in mobile advertising, reaching over 100 million users every month. The company derives most of its revenue from ad conversions, specifically through app installations from these video ads. The first year for Vungle was too expensive and did not bring in the revenue they had projected. The turning point in Vungle’s history came with the receipt of 120,000 dollars as seed funding in 2012.

This case study will help evaluate, with the help of machine learning, a newly developed ad-serving algorithm created by Kritzer and Guerin. The primary objective was to see whether the new algorithm could outperform the current Vungle algorithm in serving ads better to align with past user engagement data in an effort to maximize ad conversions and increase the Effective Revenue per 1000 impressions (eRPM).

**Problem Statement**

Vungle’s main business revolves around offering a platform that allows developers to seamlessly place video ads within their mobile apps. The company's revenue is driven by the performance of these ads, particularly by metrics such as impressions, click-through rates, and conversions. Its primary revenue model, known as cost-per-install (CPI), ensures that advertisers are charged only when users download an app after viewing an ad. This approach directly ties Vungle’s success to the effectiveness of the ad in prompting user installations.

The challenge for Vungle is increasing user engagement and more conversions within its ad-serving platform. Even small improvements in click-through rates or conversion rates have the potential to yield significant revenue gains for the company. An opportunity may exist with the development of a new machine-learning-driven ad-serving algorithm and thus calls for a carefully structured A/B test to evaluate its effectiveness.

**Method**

A/B Testing Framework: The critical concepts in hypothesis testing are as follows:

* One is called the null hypothesis, and the other the alternative or research hypothesis.
* The testing begins with the assumption that the null hypothesis is true.
* The goal is to determine whether there is enough evidence to infer that the alternative hypothesis is true.
* There are two possible decisions: Conclude that there is enough evidence to support the alternative hypothesis. Conclude that there is not enough evidence to support the alternative hypothesis.

Metrics Used:

* eRPM
* Conversion Rate

**eRPM:**

Null Hypothesis: H0: eRPMA = eRPMB

Alternative Hypothesis: H1: eRPMA ≠ eRPMB

**Conversion Rate:**

Null Hypothesis: H0: ConversionRateA = ConversionRateB

Alternative Hypothesis: H1: ConversionRateA ≠ ConversionRateB

**Analysis**

When analyzing Vungle's algorithm performance, eRPM (effective Revenue Per Mille) serves as our key decision metric. The data spans 30 days, tracking daily eRPM values for both Algorithm A and Algorithm B.

**Initial Analysis**

A simple comparison reveals that Algorithm B generates an average eRPM that is $0.131 higher than Algorithm A:

Algorithm A mean eRPM: $3.347

Algorithm B mean eRPM: $3.459

However, this basic arithmetic comparison is insufficient for making a statistically sound business decision. The difference could be due to random chance rather than actual performance superiority. This limitation leads us to employ more rigorous statistical methods.

**Need for Statistical Testing**

To properly evaluate if Algorithm B's higher eRPM represents a genuine improvement, we must use statistical testing methods that can:

* Account for daily variations in performance
* Determine if the difference is statistically significant
* Provide confidence levels for our conclusions

This is why we turn to the t-test methodology, which can help us determine whether the observed difference in eRPM is statistically meaningful or simply due to random variation in the data.

**T-test:**

T-Test considers the two samples as independent, which in this case are algorithms A and B. The t-test looks at the average values (means) of two groups and considers how spread out the numbers are within each group. Think of this as comparing two separate groups that have nothing to do with each other. Like comparing heights of students from two different schools. The test helps determine if there's a real difference between these independent groups or if any observed difference is just due to chance.

Our hypotheses are:

H₀: eRPM\_A = eRPM\_B

H₁: eRPM\_A ≠ eRPM\_B

The null hypothesis states that there is no difference between the mean eRPM of Algorithm B and Algorithm A while the alternative hypothesis states the inverse that there is difference between mean eRPM of Algorithm B and A. We set our significance level (α) at 0.05, which represents a 95% confidence level - a standard threshold in statistical testing that balances the risks of false positives and false negatives.

When conducting the independent t-test, we obtained a p-value of 0.137. Since this p-value exceeded our significance level of 0.05, we failed to reject the null hypothesis. This means we couldn't conclude with statistical confidence that there was a meaningful difference between the algorithms' performance. However, this result proved inconclusive for several reasons.

First, the independent t-test assumed the samples were independent of each other, which wasn't true in our case since both algorithms ran simultaneously.

Second, the test didn't account for daily fluctuations that affected both algorithms' performance simultaneously, such as changes in user traffic or behavior patterns.

These limitations led us to consider a more appropriate testing method - the paired t-test - which would better account for the temporal relationship between our observations.

**Paired T-test:**

Think of this as comparing "before and after" measurements of the same thing. In Vungle's case, we're comparing Algorithm A and B's performance on the same days, which is like measuring twins who experience the same conditions.

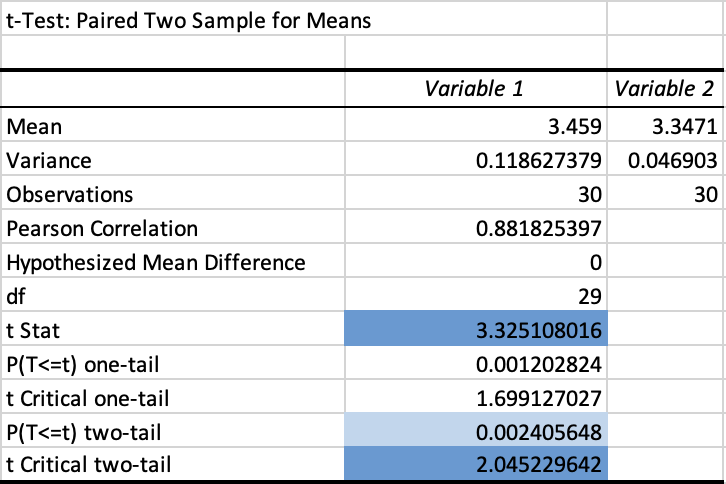
Each day provides a pair of measurements (A and B). Both algorithms run under identical daily conditions. If Monday is busy for Algorithm A, it's also busy for Algorithm B. The test looks at the daily differences between A and B.

For example, when Algorithm A earns $3.327 and Algorithm B earns $2.953 on June 1st, the paired test considers this difference while accounting for the fact that they occurred under the same conditions. This makes it a more reliable test for determining if Algorithm B's average improvement of $0.131 per thousand impressions is statistically significant.The paired t-test is particularly powerful here because it recognizes that eRPM measurements from the same day are naturally connected, making it the ideal choice for Vungle's A/B testing scenario.

In the paired t-test analysis of Vungle's algorithms, we formulated our hypotheses differently to reflect the paired nature of the data. The null hypothesis (H₀) stated that the mean difference in eRPM between Algorithms B and A was zero (H₀: eRPM\_B - eRPM\_A = 0), while the alternative hypothesis (H₁) proposed that there was a non-zero difference between them (H₁: eRPM\_B - eRPM\_A ≠ 0). We maintained our significance level (α) at 0.05, representing our 95% confidence threshold.

The paired t-test yielded a p-value of 0.002406, substantially lower than our significance level of 0.05. This result led us to reject the null hypothesis, providing strong statistical evidence of a meaningful difference between the algorithms' performance.

The test revealed that Algorithm B generated an average of $0.131 more revenue per thousand impressions than Algorithm A, with a high correlation coefficient of 0.882 between the paired observations. This result proved more reliable than the independent t-test because it properly accounted for the natural daily variations affecting both algorithms, such as weekend traffic spikes, holiday patterns, and daily user behavior cycles. The paired approach effectively controlled for these temporal factors by comparing the algorithms' performance on the same days, thereby reducing the noise in our analysis and providing a clearer picture of the true performance difference between the algorithms.



The statistical significance of Algorithm B's superior performance is further supported by comparing the t-statistic against the critical value. In our paired t-test analysis, we obtained a t-statistic of 3.325, which is notably larger than the critical value of 2.045 (two-tailed test at α = 0.05). This comparison provides strong evidence for rejecting the null hypothesis, as our test statistic falls well outside the acceptance region. The fact that our t-statistic (3.325) exceeds the critical value (2.045) by such a margin indicates that the observed difference in eRPM between Algorithms A and B is not only statistically significant but also demonstrates a robust effect size. This substantial difference between the t-statistic and critical value, combined with our low p-value of 0.002406, provides compelling statistical evidence that Algorithm B's improved performance of $0.131 per thousand impressions represents a genuine improvement rather than random variation.

**Z-test:**

We have conducted a different statistical testing, Z-test with a different metric to make further analysis. Z – test is conducted on proportions, in our case, the conversion rate of Algorithm A and the conversion rate of Algorithm B, taking into account the larger sample size. Below are the total impressions delivered by Algorithm A and Algorithm B. Both Algorithms A and B have around 236M and 15M impressions respectively. Though the number varies significantly, this cannot be used to draw conclusions.

|  |  |
| --- | --- |
| Total Impressions A | 236,459,402 |
| Total Impressions B | 15,825,376 |

The difference between the conversion rates of the two algorithms is not huge, but for a system with impressions in millions, any tiny change could have a significant impact.

|  |  |
| --- | --- |
| Conversion Rate A | 0.004024649% |
| Conversion Rate B | 0.003540579% |

Below are the hypothesis set for the analysis. Null hypothesis states that there is no significant difference between the two algorithms. The Alternate hypothesis states that the new algorithm has the potential to do better and outperform the existing algorithm.

H0: Conversion rateB = Conversion rateA

H1: Conversion rateB ≠ Conversion rateA

|  |  |
| --- | --- |
| Z Score | -29.55744395 |
| Standard Error | 0.00001637 |

The **pooled conversion rate** plays a crucial role in conducting the z-test by providing a single estimate of the overall conversion rate, assuming the null hypothesis is true (i.e., that there is no difference between the conversion rates of Groups A and B).

Pooled conversion rate =

((Total Impressions A \* Conversion Rate A) + (Total Impressions B \* Conversion Rate B)) / (Total Impressions A + Total Impressions B) = 0.003994

Pvalue for the z-score is extremely small (5.27×10−192, almost zero). Indicating that the observed difference in conversion rates is highly statistically significant, leading to strong evidence to reject the null hypothesis.

**Algorithm A is doing better, considering the Conversion rate metric.**

**Assumptions of T-Test:**

First, the data must be continuous, as the test compares the means of continuous variables, such as eRPM in this case.

Second, the data should be approximately normally distributed; however, for large sample sizes, the Central Limit Theorem justifies the assumption of normality even if the data are not perfectly normal.

Third, the two groups being compared must be independent, meaning that the observations in one group do not influence those in the other.

Lastly, the t-test can accommodate both equal and unequal variances between the groups, depending on the specific test parameters, ensuring flexibility in its application. These assumptions are critical for accurate interpretation of t-test results.

**Assumptions of Z-test:**

**Independence of Samples**: The data for each group (A and B) must come from independent samples, meaning the results of one group should not influence the other.

**Sufficient Sample Size**: The sample sizes must be large enough for the sampling distribution to approximate a normal distribution.

**Random Sampling**: Both samples must be randomly drawn to ensure that they are representative of the respective populations.

**Binary Outcomes**: The variable being tested (e.g., whether a user converts or not) must be dichotomous (binary outcomes like success/failure).

**Null Hypothesis Validity**: The z-test assumes the null hypothesis is true when calculating the pooled proportion and standard error.

**Normality of Sampling Distribution**: The sampling distribution of the difference in proportions must be approximately normal, which holds under the large sample size assumption.

**Results**

**Algorithm B can increase revenue while reducing conversion rates.**

Algorithm B more likely prioritized ads that had lower conversion chances and advertisers willing to pay a higher price for each successful install.

**Implications:**

Algorithm B’s focus on maximizing **short-term revenue** might seem like an obvious win, but the long-term risks it poses could outweigh the benefits:

* **User Experience Decline:** Ads with lower chances of conversion can frustrate users. When users feel ads aren’t relevant or engaging, they’re more likely to disengage with the platform. This erosion in user trust and satisfaction can lead to a decrease in overall impressions over time.
* **Ad Diversity Imbalance:** By prioritizing ads that generate the highest revenue per install, Algorithm B inadvertently creates an ad environment that lacks variety. This not only affects the user experience but also reduces opportunities for advertisers with lower budgets to compete, which could hurt the platform's revenue diversity and resilience.
* **Long-Term Revenue Decline:** While short-term revenue may spike, the reduction in user satisfaction and ad diversity can lead to a gradual drop in impressions. Fewer users interacting with the platform means fewer ads being served, which ultimately reduces both impressions and revenue in the long run.
* **Brand Reputation Risk:** A degraded user experience can damage the brand's reputation among both users and advertisers. Users might seek alternative platforms, and advertisers may hesitate to invest if they perceive the ad ecosystem as unsustainable or skewed.

Prioritizing short-term revenue at the expense of user experience and diversity might seem lucrative initially, but it can create risks that are harder to reverse later. Our recommendations focus on maintaining this balance, ensuring that Vungle’s platform remains engaging and sustainable for users, advertisers, and the business itself.

**Recommendations:**

* **Optimize Revenue-Driven Ads**: Since B prioritizes higher revenue per install, Vungle should balance ad prioritization to maximize revenue while avoiding long-term losses from reduced advertiser impressions and conversions.
* **Experiment with Ad Mix**: Test varying combinations of high-revenue and high-conversion ads to identify an optimal strategy that benefits both advertisers and Vungle.
* **Extend A/B Testing**: Conduct extended tests (e.g., 30-60 days) with diverse geographies and demographics to ensure results are consistent and scalable.
* **Dynamic Ad Prioritization Strategy**: Develop a machine learning model to dynamically prioritize ads based on real-time revenue-per-install and conversion probabilities, ensuring both short-term revenue maximization and long-term advertiser satisfaction.
* **Segmented A/B Testing**: Implement segmented tests across geographies, audience demographics, and ad types to uncover actionable insights for personalized ad delivery strategies that enhance overall ecosystem value.
* **Advertiser Partnership Program**: Create a program that communicates the value of high-revenue installs to advertisers while offering tiered visibility for high-conversion ads, maintaining trust and ensuring mutual growth.