QMB: 6304  
Midterm Project: The Advertising Campaign Analysis

September 27, 2023

**Background**: A prominent travel agency has conducted an advertising campaign. Two datasets - ’Aban- doned.csv’ (ABD) and ’Reservation.csv’ (RES) - provide insights into customer interactions and outcomes from this campaign.  
**Objective**: Determine the statistical success of a retargeting campaign by matching and analyzing the datasets.

**Introduction**

’Abandoned.csv’ contains data about customers who engaged but didn’t purchase a vacation package. Notice the potential missing data and duplicates. These customers were divided into test and control groups for a retargeting campaign. ’Reservation.csv’ documents customers who eventually purchased vacation packages.

**Task**: Establish whether the retargeting campaign was statistically effective.

**1 Business Justification**

1. Explain why retargeting customers who initially didn’t buy a package makes business sense.

* Retargeting customers makes business sense as it helps increase conversion rates, is cost-effective, provides valuable data insights, and maintains a competitive advantage.

1. Analyze the test/control division. Does it seem well-executed?

* Test group size – 4266 and Control Group size – 4176. The size of the test and control groups are not balanced.
* The assignment to test and control groups may not be randomized.
* In our analysis, both the imbalance in group sizes and the suggestion that the assignment may not be randomized could be concerning. An ideal test/control division should be balanced and randomized to ensure that any observed effects can be attributed to the treatment (retargeting campaign) rather than differences in the groups. If this is a real-world scenario, we may want to investigate further why the division is imbalanced and whether randomization procedures were followed correctly. An imbalance could introduce bias, and non-randomized assignment may affect the validity of our results. Adjustments or further investigation may be needed to ensure the reliability of the analysis.

1. Compute summary statistics for the test variable, segmenting by available State data.

* See R file

**2 Data Alignment**

1. From your examination of both files, propose potential data keys to match customers.

* From my Examination of both Files, the potential data keys are Email, Incoming\_Phone and Contact\_Phine

1. Detail your procedure to identify customers in:

* Treatment group who purchased.

**345**

* Treatment group who didn’t purchase.

**3921**

* Control group who purchased.

**93**

* Control group who didn’t purchase.

**4083**

* A blue background with white text

  Description automatically generated

1. Are there unmatchable records? If yes, provide examples and exclude them from the analysis.

* Yes there are duplicate records in the dataset based on Email, Incoming\_Phone and Contact\_Phone which is cleaned and both clean and duplicated data is represented in the R file.

1. Provide a cross-tabulation of outcomes for treatment and control groups.

* A screenshot of a computer

  Description automatically generated
* This is the cross tabulation of data after cleaning duplicated rows

1. Replicate the cross-tabulation for five randomly chosen states, detailing your selections.

* A screenshot of a computer screen

  Description automatically generated
* Refer to R file for the code.

**3 Data Refinement**

1. Generate a cleaned dataset with columns: Customer ID — Test Group — Outcome — State Available — Email Available. Each row should correspond to a matched customer from the datasets. *(Ensure you attach this cleaned dataset upon submission.)*

* A screenshot of a computer

  Description automatically generated
* Refer to R file for code snippet.

**4 Statistical Assessment**

1. Execute a linear regression for the formula: Outcome = α + β \* Test Group + error. Share the results.

* A screenshot of a computer program

  Description automatically generated
* The linear regression results demonstrate that there is a statistically significant difference in outcomes between the control and test groups, making it a valid approach for assessing the effectiveness of the treatment.

1. Justify that this regression is statistically comparable to an ANOVA/t-test.

* A computer screen with white text

  Description automatically generated
* The ANOVA result displays the sum of squares, degrees of freedom, and residuals, which are typical components of an ANOVA output. Therefore, the regression we conducted is indeed statistically comparable to an ANOVA, and it's used to assess the same hypothesis – whether there are significant differences between groups.

1. Debate the appropriateness of the regression model in making causal claims about the retargeting campaign’s efficacy.

* The regression model used in this analysis can help identify associations and correlations between the Test\_Group and Outcome variables but is not suitable for making causal claims about the retargeting campaign's efficacy. Causality requires a more rigorous experimental design, including random assignment and control of potential confounding variables, which this observational regression lacks. While it can inform about relationships, attributing causality to the retargeting campaign would require a different study design.

1. Integrate State and Email dummies into the regression. Also consider interactions with the treatment group. Compare these results to the previous regression and provide insights.

* The results of the different regression models are as follows:
* A screenshot of a computer program

  Description automatically generated
* **Model 1 (model)**: Outcome regressed on Test\_Group alone, showing a significant positive association between Test\_Group and Outcome (p < 0.01).
* A computer screen shot of a program

  Description automatically generated
* **Model 2 (model\_iteration\_1)**: Outcome regressed on State\_Available and Email\_Available, revealing both predictors to be statistically significant, with positive associations with Outcome (p < 0.01).
* A computer screen with numbers and letters

  Description automatically generated
* **Model 3(model\_iteration\_2)**:: Outcome regressed on Test\_Group, State\_Available, and Email\_Available, showing Test\_Group to be statistically significant with a positive association with Outcome (p < 0.01). State\_Available and Email\_Available remain statistically significant with positive associations.
* A screenshot of a computer

  Description automatically generated
* **Model 4(model\_iteration\_3)**:: Outcome regressed on Test\_Group, State\_Available, Email\_Available, and their interactions. Test\_Group and the interaction term Test\_Group:Email\_Available are statistically significant with positive associations (p < 0.01).
* These results suggest that while the Test\_Group has a positive effect on the Outcome, adding interactions or additional predictors such as State\_Available and Email\_Available does not significantly improve the model's explanatory power.
* Below is a Stargazer summary Table
* A screenshot of a computer screen

  Description automatically generated
* In Conclusion, Based on the statistical analysis, the retargeting campaign appears to be statistically effective to some extent. The linear regression model showed that the "Test\_Group" variable, representing exposure to the campaign, has a statistically significant effect on the outcome variable. However, the effect size is relatively small, suggesting that while the campaign has a statistically significant impact, it may not be the sole driver of customer decisions. Therefore, while there is evidence of effectiveness, it may be necessary to complement the retargeting campaign with other strategies or improvements to maximize its impact on the desired outcomes.

**5 Reflections**

1. Reflect on the project:

* **Would you modify the experiment design if given a chance?**

As a Data Scientist, I would consider modifying the experiment design if the original objectives weren't met, the sample size was inadequate, or variables needed adjustment. I would also assess group composition, experiment duration, and data quality, and consult experts to ensure a more robust design aligned with research goals.

* **Could alternative paths be taken with better-quality data?**

Yes, alternative paths could be taken with better-quality data. High-quality data can lead to more reliable and detailed analyses, enabling the exploration of additional variables, uncovering subtle effects, and enhancing the accuracy of predictions.

* **Are there actionable business implications from this analysis?**

Yes, there are actionable business implications from this analysis. While the specific implications would depend on the goals of the retargeting campaign and the company's objectives, the analysis provides insights into the impact of the campaign on customer outcomes. These insights can inform decisions related to campaign optimization, resource allocation, and customer targeting, ultimately leading to more effective marketing strategies.

1. Self-assessment: Rate your effort (0-100) and anticipated performance. Elaborate if needed, men- tioning any collaborations.

* Self-assessment:

I would rate my effort at 90 for this project. I put in a considerable amount of effort in data cleaning, analysis, and providing detailed explanations. I strived to address your questions comprehensively.

In terms of anticipated performance, I believe the provided analysis is thorough and should meet the project's objectives. However, the ultimate assessment would depend on your feedback and whether the analysis aligns with your expectations and requirements.

Regarding collaborations, this project was completed individually, without direct collaboration with others.

**Note**: This is an individual project. While discussions with peers are encouraged, mention all individuals you collaborated with. Ensure your final submission includes your Rscript, analyses, and the refined dataset.

# Provide an explanation in comments or as a text output

cat("Retargeting customers makes business sense as it helps increase conversion rates, is cost-effective, provides valuable data insights, and maintains a competitive advantage.")