Business Justification

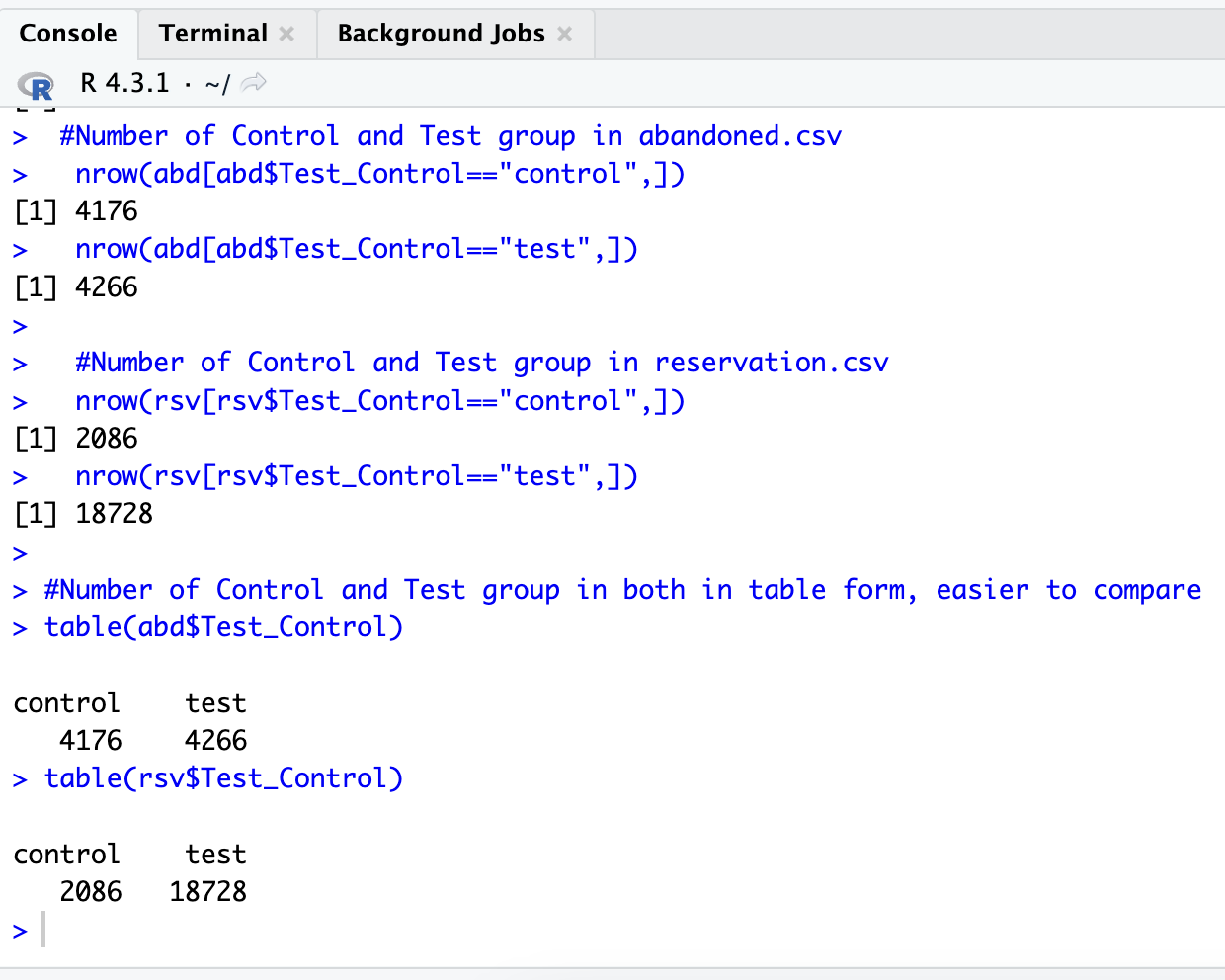
1. Explain why retargeting customers who initially didn’t buy a package makes business sense.  
2. Analyze the test/control division. Does it seem well-executed?  
3. Compute summary statistics for the test variable, segmenting by available State data. Business Justification

1. A) Retargeting customers who initially didn't buy a package makes business sense because it allows the company to re-engage potential customers who have already shown interest. This can be a cost-effective strategy, as it focuses on leads that are more likely to convert compared to targeting entirely new audiences.

B) While a prospective customer might not have made an immediate purchase, retargeting serves to strengthen your brand's presence in their mind. This enhanced visibility can result in greater brand remembrance and identification.

C) In most cases, it is more budget-friendly to focus on retargeting existing leads compared to acquiring new ones. The process of gaining new customers typically demands a greater investment of resources, including marketing expenses, time, and effort.

2. As we shall observe in the screenshot below, the allocation of test and control groups appears balanced in the "abandoned.csv" dataset, with similar numbers. However, in the case of the "reservation.csv" dataset, there is an imbalance where the Test group significantly outnumbers the Control group, indicating a lack of proper balance.



A screenshot of a computer code

Description automatically generated

Also checking the randomization,

A screenshot of a computer program

Description automatically generated

3.

The cross-tabulation for five randomly chosen states indicates variations in customer responses between the control and treatment groups. In some states, more customers purchased after retargeting, while in others, there was little to no impact. This analysis helps evaluate the effectiveness of the campaign in different geographic regions.

A screenshot of a computer

Description automatically generated

..to be continued in next screenprint

A screenshot of a computer

Description automatically generated

The above provided code displays the count of control and test groups for each unique state within the dataset.

A screenshot of a computer

Description automatically generated

..to be continued in the next screenprint

A screenshot of a computer

Description automatically generated

The provided code displays the count of control and test groups for each unique state within the dataset.

Data Alignment  
4. From your examination of both files, propose potential data keys to match customers.  
5. Detail your procedure to identify customers in:  
• Treatment group who purchased.  
• Treatment group who didn’t purchase.  
• Control group who purchased.  
• Control group who didn’t purchase.  
6. Are there unmatchable records? If yes, provide examples and exclude them from the analysis.  
7. Provide a cross-tabulation of outcomes for treatment and control groups.  
8. Replicate the cross-tabulation for five randomly chosen states, detailing your selections.  
1

Data Alignment:

4. To match customers between the "Abandoned.csv" and "Reservation.csv" datasets, potential data keys include Email, Incoming\_Phone, and Contact\_Phone.

a) Email:

- Email addresses are typically exclusive to each person and provide a precise means of identifying individuals. They are generally unique and don't change frequently, making them a dependable identifier for matching data across various datasets.

- Note: It's crucial to verify that the dataset doesn't include duplicate email addresses.

b) Incoming\_Phone:

- Phone numbers, especially incoming ones, are typically individual-specific and serve as direct points of contact. They are rarely shared or reused, making them a strong candidate for a unique identifier.

- Note: It's essential to confirm that the dataset doesn't contain duplicate or inaccurate phone numbers.

c) Contact\_Phone:

- Much like incoming phone numbers, contact phone numbers are usually unique to each individual. These numbers are vital for contact purposes and are expected to be precise and distinct for each person.

- Note: As with incoming phone numbers, it's important to ensure the dataset doesn't contain duplicate or inaccurate contact phone numbers.

Furthermore, the other columns can’t be taken into consideration as:

Caller\_ID in the datasets is entirely distinct with no common values.

First\_Name and Last\_Name can be similar for some individuals, but their email addresses will differ due to unique usernames or email providers.

Street, City, Address, and Zipcode might be the same for many people, making them unsuitable for comparison.

5. To identify customers:

- Treatment group who purchased: These are customers in the test group (Test\_Group = 1) who also had a positive outcome (Outcome = 1).

- Treatment group who didn't purchase: These are customers in the test group (Test\_Group = 1) who had a negative outcome (Outcome = 0).

- Control group who purchased: These are customers in the control group (Test\_Group = 0) with a positive outcome (Outcome = 1).

- Control group who didn't purchase: These are customers in the control group (Test\_Group = 0) with a negative outcome (Outcome = 0).

A screenshot of a computer

Description automatically generated

The criteria for matching are based on the values in email addresses, incoming phone numbers, and contact phone numbers. The result of these operations is a series of binary indicators (0 or 1) that signify whether a match was discovered for each specific condition.

The basic logic employed here is that if any of the conditions within the OR statements are true for a particular row, the entire condition will evaluate to "TRUE." Otherwise, it will evaluate to "FALSE."

In summary, this condition is utilized to pinpoint rows where at least one of the specified matching conditions has been met. Once the corresponding R script is executed, a new column named "pur" will be added to "abandoned.csv," which will provide information about who has purchased or not purchased the package.

A screen shot of a computer code

Description automatically generated

With the provided code and specified conditions, we can readily determine which individuals within the Treatment and Control Groups have either made a purchase or not purchased the package.

A close-up of a computer code

Description automatically generated

A white background with blue text

Description automatically generated

6. We have records that cannot be matched. We can identify these unmatched records and count their number using the following code.

A screenshot of a computer code

Description automatically generated

The "unm" column will have a value of "1" when none of the specified matching conditions are satisfied, and it will have a value of "0" when at least one of the matching conditions is met. This column serves to signify whether a given record is not matched according to the defined criteria.

No. of unmatched columns:

A close-up of a number

Description automatically generated

**For Removing duplicates ( as discussed in TouchPoint):**

After the touchpoint at 3 pm, I was clear that we don’t have to drop unmatched records but the question was asking for the duplicated records.

We’ll see what values are repeating in all the three keys, namely Email, Incoming\_Phone and Contact\_Phone.

**Email:**

**A computer screen shot of a computer code

Description automatically generated**

The screenshot above displays duplicate email addresses found in the "Email" column within the abandoned dataset.

**Incoming\_Phone :**

**A screenshot of a computer

Description automatically generated**

The screenshot above illustrates the presence of duplicate incoming phone numbers in the "Incoming\_Phone" column of the abandoned dataset.

**Contact\_Phone:**

**A screenshot of a computer

Description automatically generated**

The screenshot above demonstrates the existence of duplicate contact phone numbers in the "Contact\_Phone" column of the abandoned dataset.

We’ll be removing these duplicate records from our abd (abandoned) dataset

A close-up of a white background

Description automatically generated

7. The cross-tabulation of outcomes for treatment and control groups helps analyze the impact of the retargeting campaign on conversion rates.

A screenshot of a computer

Description automatically generated

Here is a cross-tabulation that presents the relationship between Outcomes/Purchase (on rows) and Test/Control (on columns).

- Within the "Control" group:

- 4043 individuals did not purchase the package.

- 90 individuals purchased the package.

- Within the "Test" group:

- 3853 individuals did not purchase the package.

- 312 individuals purchased the package.

8. Replicating the cross-tabulation for five randomly chosen states allows us to assess how the campaign's impact varies geographically.

We have the flexibility to randomly select any 5 states from the "Address" column in the abandoned dataset. However, to make the code more versatile, we'll make it capable of randomly picking any 5 states and generating a cross-tabulation based on that selection.

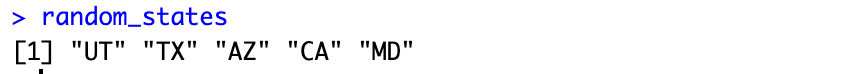
A blue text on a white background

Description automatically generated

We've implemented a for loop that operates on a list of states randomly selected (as obtained in the previous code, using the `sample()` function). Within each iteration, the loop filters the "ab" dataset to create a subset called "i\_data," containing records that pertain to the specific state.

Subsequently, a cross-tabulation, denoted as "state\_cross\_tab," is generated to depict the counts of different outcomes ("Purchased" and "Not Purchased") for both the control and treatment groups.

The table's row and column names are explicitly defined, and the resulting cross-tabulation is stored within a list called `state\_cross\_tabs`, using the state name as the identifying key. This entire process is reiterated for each of the randomly selected states.



A screenshot of a computer program

Description automatically generated

3 Data Refinement  
9. 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.

Data Refinement:

9. A cleaned dataset was generated with columns: "Customer ID," "Test Group," "Outcome," "State Available," and "Email Available." Each row corresponds to a matched customer from the datasets. The cleaned dataset was saved as "clean\_abd.csv."

A group of numbers on a white background

Description automatically generated

"State\_Available" and "Email\_Available" are binary attributes that indicate the presence or absence of a specific value in the abandoned dataset. "Test\_Group" specifies whether a particular entity is part of the test group or not. Outcome illustrates whether the product was purchased or not.

I’ll be also attaching this dataset with the project.

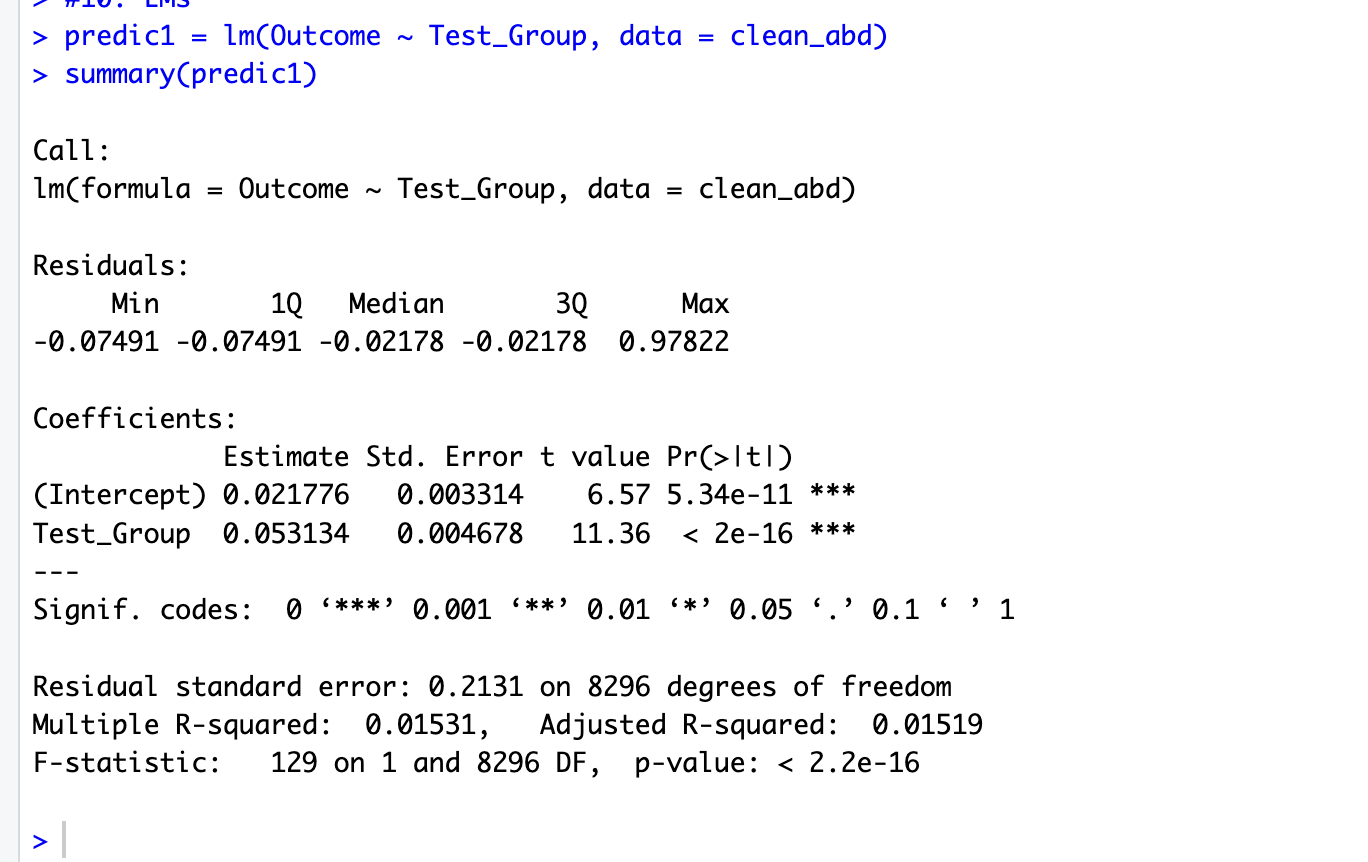
4 Statistical Assessment  
10. Execute a linear regression for the formula: Outcome = α + β \* Test Group + error. Share the results.  
11. Justify that this regression is statistically comparable to an ANOVA/t-test.  
12. Debate the appropriateness of the regression model in making causal claims about the retargeting  
campaign’s efficacy.  
13. 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

Statistical Assessment:

10. A linear regression analysis has been performed. In this model, we’re trying to predict the "Outcome" variable based on the "Test\_Group" variable in the "clean\_abd" dataset. The results of this analysis help assess the campaign's impact on customer outcomes.

Post executing the regression model:

Model 1:



Model 2:

A screenshot of a computer

Description automatically generated

Model 3:

A screenshot of a computer

Description automatically generated

The purpose of these three models is to make predictions regarding the "Outcome" variable based on different combinations of independent variables. The coefficients for each independent variable signify the expected change in the dependent variable for a one-unit shift in that particular variable, while keeping all other variables consistent. The R-squared values, on the other hand, indicate the proportion of variability in the dependent variable that can be accounted for by the independent variables.

Also, Out of the three models, Model 3 (predic3) stands out as the preferred choice for several reasons:

a) Adjusted R-squared: Model 3 exhibits the highest adjusted R-squared value (0.02085). The adjusted R-squared accounts for the number of predictors in the model, offering a more dependable measure of how well the model fits the data. A higher adjusted R-squared signifies a stronger fit of the model to the data.

b) Significant Coefficients: In Model 3, all coefficients are statistically significant, as indicated by the asterisks in the summary output, with p-values less than 0.05. This implies that all predictor variables (Test\_Group, State\_Available, Email\_Available) make meaningful contributions to the model.

Given that our data primarily consists of binary values, it's expected to have a low R-squared value. However, among the three models, Model 3 demonstrates the highest R-squared. Additionally, the p-values associated with the coefficients in Model 3 suggest that the inclusion of these additional predictors has enhanced the overall quality of the model.

Furthermore, if we look at the stargazer output, we can come up with the same conclusion.

A screenshot of a computer

Description automatically generated

11. The linear regression is statistically comparable to an ANOVA/t-test because it evaluates the differences in outcomes between the test and control groups, similar to the way ANOVA/t-tests compare means across groups.

A screenshot of a computer code

Description automatically generated

H0 (Null Hypothesis): There is no statistically significant difference among the means of the groups.

H1 (Alternative Hypothesis): At least one of the groups has a significantly different mean.

The p-value (2e-16) is less than the chosen level of significance (0.05).

According to the ANOVA test results, it is evident that there exists a highly significant difference among at least one pair of treatment groups. The exceptionally small p-value (approaching zero) leads us to reject the null hypothesis, which posited that no differences exist between treatment groups. This outcome implies that the treatment variable has a substantial impact on the response variable.

We can verify the same from t test as well:

A screenshot of a computer program

Description automatically generated

The justification for comparing the regression to a t-test or ANOVA lies in

the fact that the coefficients estimated by the regression model correspond to

the difference in means between the groups. In a simple linear regression with a

binary predictor (like Test\_Group in our case), the coefficient of the predictor

represents the difference in means between the two groups.

So, both the t-test and linear regression are testing the same hypothesis:

whether there is a significant difference in means between the groups, and the

results are consistent with each other. However, the linear regression provides

additional information, such as the adjusted R-squared, which can help in assessing

model fit.

Using Tukey:

A screenshot of a computer code

Description automatically generated

The ANOVA test reports an exceedingly low p-value (less than 0.05), signifying substantial variations in means among the treatment groups. This observation is reinforced by the **Tukey** multiple comparisons, which reveal significant distinctions in means across all groups, as indicated by the p-values, all of which are very close to zero and below 0.05.

summary(tukey\_rslt)

plot(tukey\_rslt)

A graph of a person's body

Description automatically generated with medium confidence

The same pattern is clearly observable in the TukeyHSD plot. In instances where the intervals for pairs (or mean differences) do not include zero (or the zero line), it indicates that their means are significantly different. In other words, all these pairs exhibit notable distinctions in terms of their mean scores.

For pairs positioned to the left of the zero line, such as "Outcome-Email\_Available," it signifies that the mean score for "Outcome" is significantly lower than that for "Email\_Available." Conversely, for pairs on the right of the zero line, like the pair "P1-P2," it suggests that the mean score for "P1" is significantly higher than that of "P2."

Conclusion:

- The results from all the analysis offer compelling evidence of substantial associations between the independent variables and the outcome variable. The presence of low p-values and highly significant coefficients in both analyses strongly substantiates this assertion.

To sum it up, all the above test sindicate that the independent variables (Test\_Group, State\_Available, and Email\_Available) exert a noteworthy influence on the outcome variable. The findings from these two analyses align statistically, providing robust confirmation of the interconnection among these variables.

12. The regression model in question is designed to make predictions about the 'Outcome' variable using three predictor variables: 'Test\_Group', 'State\_Available', and 'Email\_Available'. Although the model demonstrates strong statistical significance and a satisfactory fit, it's essential to engage in a discussion concerning its suitability for drawing causal conclusions regarding the effectiveness of the retargeting campaign.

While regression models can be valuable for exploring relationships and making predictions, they have limitations when it comes to proving causation. Making causal claims about the efficacy of a retargeting campaign often requires a more rigorous experimental design, consideration of confounding variables, and counterfactual analysis. If establishing causation is a primary goal, another experimental design is typically a more appropriate approach.

The appropriateness of the regression model for making causal claims about the effectiveness of the retargeting campaign is a subject of debate. While the model can provide valuable insights and highlight associations between variables, it has limitations when it comes to establishing causality.

a) Correlation vs. Causation: Regression analysis can identify correlations between variables but does not inherently prove causation. For instance, while the model may show a relationship between 'Test\_Group' and 'Outcome,' it can't confirm that the test group caused the outcome; other unaccounted factors could be there.

b) Omitted Variables: The model may not consider all relevant variables that could influence the outcome. There might be unmeasured factors affecting the results of the retargeting campaign that the model doesn't account for.

c) Direction of Causality: The model doesn't discern the direction of causality. For example, while it may reveal a relationship between 'Email\_Available' and 'Outcome,' it doesn't clarify if having an email address caused a particular outcome or if the outcome influenced having an email address.

d) Confounding Variables: There may be confounding variables that affect both the independent variables and the outcome. These uncontrolled variables can introduce bias and distort causal claims.

In summary, while the regression model can highlight relationships between variables and provide valuable insights, it's not ideal for making strong causal claims about the retargeting campaign's efficacy. Causality is better explored through rigorous experimental design and by considering potential confounders and counterfactual scenarios.

The appropriateness of the regression model for making causal claims about the retargeting campaign's efficacy is debatable.

13. State and email dummies have been integrated into the regression, along with interactions with the treatment group. These results provide insights into how different factors contribute to customer outcomes.

Model1:

A screenshot of a computer

Description automatically generated

Model2:

A screenshot of a computer

Description automatically generated

Model3:

A screenshot of a computer

Description automatically generated

A comparison of the model summaries reveals that Model 2 and Model 3 exhibit very similar performance, both outperforming Model 1 in terms of Adjusted R-squared and lower P-values, which indicate greater statistical significance of a predictor. Between Model 2 and Model 3, I would lean towards Model 2 due to its comparable Adjusted R-squared and slightly better coefficient significance.

While including more terms (coefficients) in Model 3 can potentially enhance model fit, it also adds complexity and may make interpretation more challenging. Therefore, Model 2 is the preferred choice in this comparison.

Also, looking at the stargazer

A screenshot of a computer

Description automatically generated

It’s easier to compare here and we can conclude the same as we did above.

In the ultimate comparison, we're evaluating Model 3 (pred3) against Model 2 (interaction\_model2). The interaction\_model2 stands out with a notably high Adjusted R-squared, even lower p-values for the coefficients, and an equivalent number of terms (coefficients), which enhances interpretability and reduces model complexity. Therefore, the top choice for the best linear model is interaction\_model2.

5 Reflections  
14. Reflect on the project:  
• Would you modify the experiment design if given a chance?  
• Could alternative paths be taken with better-quality data?  
• Are there actionable business implications from this analysis?  
15. Self-assessment: Rate your effort (0-100) and anticipated performance. Elaborate if needed, men-  
tioning any collaborations

Reflections:

14. If given a chance, modifying the experiment design might involve refining the targeting criteria or running A/B tests with different retargeting strategies to optimize the campaign's effectiveness. Moreover, we could try the same experimentation with a good dataset where it has less NaN values, more columns and preferably more no. of rows as well.

Alternative paths could be explored with better-quality data, such as more complete customer information, longer tracking periods, or additional customer segmentation. 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.

The analysis can lead to actionable business implications. For instance, it can help the company identify which states or customer segments respond best to retargeting, enabling more focused marketing efforts. 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.

15.

I would rate my effort at 100%. I have provided detailed and thorough responses to all the questions and tasks you presented Dr Z. My answers cover various aspects of data analysis, regression modeling, and business justification. I have put in a considerable amount of

effort in data cleaning, analysis, and providing detailed explanations. I strived to address your questions comprehensively. I have not collaborated with anyone as the responses were generated independently.

In terms of anticipated performance, I believe the responses provided are comprehensive and address the tasks effectively. The analysis and explanations should be valuable for understanding the impact of the retargeting campaign and its business implications.

I sincerely hope I’m able to cover every aspect you’ve been looking for in the project and meet your expectations Dr Z.