**Project Report**

**1.**

**Retargeting customers who initially didn't buy a package makes business sense for several reasons:**

1. Engagement Reminder: Customers who showed interest but didn't make a purchase may have been distracted, busy, or not yet ready to commit. By retargeting them, you're reminding them of their initial interest and giving them another opportunity to engage with your product or service.

2. Increase Conversion Rates: Studies have shown that retargeting can significantly increase conversion rates. This is because it targets an audience that has already shown some level of interest in your offering.

3. Cost-Effectiveness: It's generally more cost-effective to retarget existing leads than to acquire new ones. Acquiring new customers often requires more resources in terms of marketing spend, time, and effort.

4. Brand Recall and Recognition: Even if a potential customer didn't make a purchase initially, retargeting helps reinforce your brand in their memory. This increased exposure can lead to higher brand recall and recognition.

**2.**

A screenshot of a computer program

Description automatically generated

As we can see from the above screenshot the execution of test/control group seems well in case of “abandoned.csv” as Test and Control are roughly similar in number. But in “reservation.csv” it is opposite, as Control is overnumbered by Test group by a lot, hence it is not properly balanced.

**3.**

A table of numbers with letters

Description automatically generated

…continued

I couldn’t take a long screenshot but the above code lists number of control and test groups for all the given distinct states in the dataset.

A table of numbers with letters and numbers

Description automatically generated

…continued

I couldn’t take a long screenshot but the above code lists number of control and test groups for all the given distinct states in the dataset.

**4.**

Potential data keys to match customers would be Email, Incoming\_Phone and Contact\_Phone. On testing I observed that other column entities had a huge difference in distinct numbers when comparing in both datasets but only these 3 came out to be consistent. Apart from that we can also consider below points,

1. Email:

- Email addresses are typically unique to each individual. They serve as a direct and specific identifier for a person. In most cases, people have distinct email addresses, and they rarely change. This makes it a reliable key for matching values across datasets.

- Consideration: It's important to ensure that the dataset does not contain duplicate email addresses.

2. Incoming\_Phone:

- Phone numbers, especially incoming phone numbers, are typically unique to each individual. They serve as direct contact points and are rarely shared or reused. This makes them a good candidate for a key.

- Consideration: It's important to validate that the dataset does not contain duplicate or incorrect phone numbers.

3. Contact\_Phone:

- Similar to incoming phone numbers, contact phone numbers are usually unique to each individual. They are crucial contact points and are expected to be accurate and distinct for each person.

- Consideration: As with incoming phone numbers, we have to ensure that the dataset does not contain duplicate or incorrect contact phone numbers.

When selecting a key, it's crucial to consider the uniqueness and reliability of the information.

For the rest of the columns,

1. Caller\_ID:

- Both the datasets have different Caller\_IDs with none in common.

2. First\_Name, Last\_Name:

- First name and Last name of any two persons could be similar but even in that case their emails would not be the same. Either the username or Email provider would be different.

3. Street, City, Address and Zipcode:

- All of them could be the same for a lot of people, hence they cannot be considered for comparison.

**5.**

A screenshot of a computer

Description automatically generated

These lines of code are performing matching operations between columns of ab and rs and updating specific columns in the `ab` dataset to indicate whether a match was found or not. The matching criteria are based on the values in email addresses, incoming phone numbers, and contact phone numbers. The result is a set of binary indicators (0 or 1) indicating whether a match was found for each condition.

Now, to identify customers in treatment/control group who purchased/didn’t purchase, firstly we need to find the matching records. We can do so by using the below condition,

ab$match\_email | ab$match\_incoming | ab$match\_contact | ab$match\_incoming\_contact | ab$match\_contact\_incoming

**Basic Logic used:**

If any of the conditions (in between the ORs) is true for a given row, the entire condition will evaluate to `TRUE`. Otherwise, it will evaluate to `FALSE`.

In summary, this condition is used to identify rows where at least one of the specified matching conditions is met.

After executing the corresponding code in R Script we’ll have a new column called “pur” in abandoned.csv which would tell us who purchased/didn’t purchase the package.

**Output:**



By using the below code with given conditions, we can very easily know that which persons from Treatment/Control Group purchased/didn’t purchase the package.

A close-up of a code

Description automatically generated

We can get the data in the form of numbers by cross-tabulating purchase and Test\_Control from abandoned dataset.

A screenshot of a test

Description automatically generated

From here we can get the numbers for all the 4 possibilities.

**6.**

**-----------------------------For unmatched records-----------------------------**

**In the first version I submitted this part was not there as “unmatched records” was a bit ambiguous. Then, after further analysis on dataset I added this in second version.**

Yes, we do have unmatchable records. We can get the records and number of records by using the code below.

A screen shot of a computer

Description automatically generated

Code explanation:

mutate() function from the “dplyr” package is used for creating or modifying variables (columns) in a dataset. Here we are creating a new one for unmatched records.

Basically, this code is creating a new column “unm” in the dataset “ab”. The values in “unm” will be “1” if none of the matching conditions are met, and “0” if at least one of the matching conditions is met, which also complies with our idea of the unmatched columns. This column is used to indicate whether a particular record is unmatched based on the specified matching criteria.

Below snippet shows the number of unmatched columns,



Finally, we can remove the unmatched columns by executing the following code,

#Removing unmatched columns

ab = ab[ab$unm==0,]

This will put all the matched columns (i.e. where ab$unm==0), back in dataframe “ab”.

**----------------------------------­­­­--------------------------------------------------**

**­-----------------------------For removing duplicate records-----------------------------**

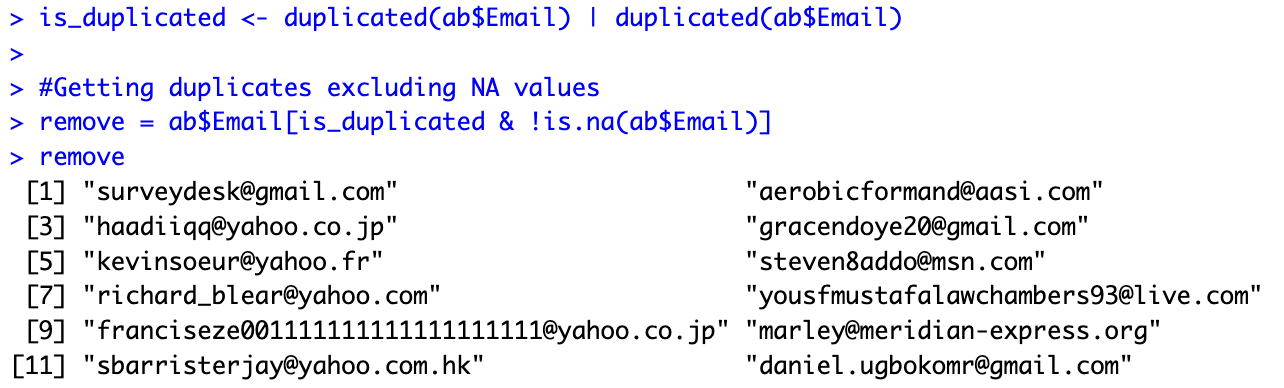
**After the last touchpoint (on Oct 24th 3:00 PM), I have analyzed dataset and added this section to remove duplicate records, as discussed and clarified in the touchpoint.**

Yes, there are unmatchable records, i.e., there are duplicate entries in the abandoned dataset. Basically, some values in the potential keys that we have chosen are duplicate. Let’s see what values are repeating in all the three keys, namely Email, Incoming\_Phone and Contact\_Phone.

**Examples:**

In the code below we have excluded NA values as there are a lot of them in the dataset, if included we’ll get a huge number of matches across the whole dataset.

Email:



Above screenshot shows the duplicate Email addresses present in Email column in abandoned dataset.

Incoming\_Phone:

A screenshot of a computer

Description automatically generated

Above screenshot shows the duplicate Incoming phones present in Incoming\_Phone column in abandoned dataset.

Contact\_Phone:

A screenshot of a computer

Description automatically generated

Above screenshot shows the duplicate Contact phones present in Contact\_Phone column in abandoned dataset.

**Let’s Verify with some of the duplicates:**

Below screenshots are from a Jupyter notebook (code in Python), as I was not able to find a better way to represent this in R.

A screenshot of a computer

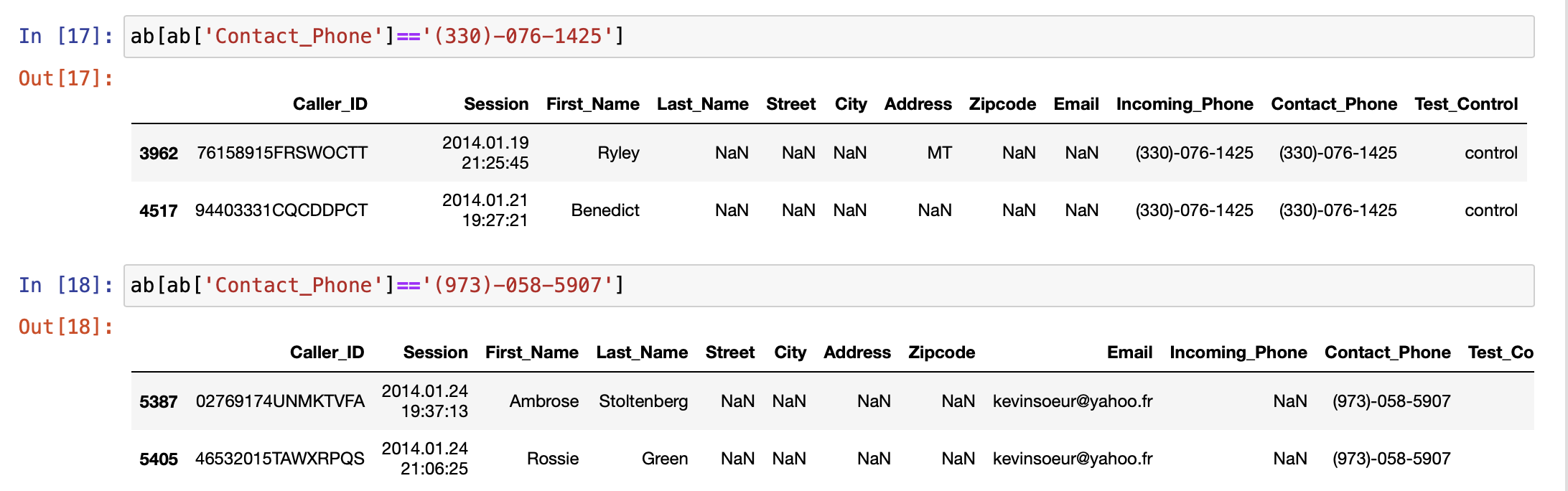
Description automatically generated

Randomly took two emails to check for duplicates, namely [surveydesk@gmail.com](mailto:surveydesk@gmail.com) and [yousfmustafalawchambers93@live.com](mailto:yousfmustafalawchambers93@live.com) from “remove” list for “Email” from previous Examples section. Evidently, both the emails have duplicates which yet again verifies that Emails in “remove ” list are all having duplicates.

A screenshot of a computer

Description automatically generated

Randomly took two Incoming phones to check for duplicates, namely (803)-322-1911 and (863)-602-1297 from “remove” list for “Incoming\_Phone” from previous Examples section. Evidently, both the numbers have duplicates which yet again verifies that Incoming numbers in “remove ” list are all having duplicates.



Randomly took two Contact phones to check for duplicates, namely (330)-076-1425 and (973)-058-5907 from “remove” list for “Contact\_Phone” from previous Examples section. Evidently, both the numbers have duplicates which yet again verifies that Contact numbers in “remove ” list are all having duplicates.

On some further analysis in Python, I also noticed that more than 90% of Incoming\_Phone and Contact\_Phone duplicate values are common. Something which we can also see in above screenshots (apart from the last one).

**----------------------------------­­­­--------------------------------------------------**

**7.**

A close-up of a white background

Description automatically generated

Here is a cross tabulation of Outcomes/Purchase (on rows) and Test/Control (on columns)

Interpreting the numbers:

- In the "Control" group:

- There are 4043 persons who didn’t purchase the package.

- There are 90 persons who purchased the package.

- In the "Test" group:

- There are 3853 persons who didn’t purchase the package.

- There are 312 persons who purchased the package.

**8.**

We can randomly pick any 5 states from the Address column in abandoned dataset. But let’s generalize the code so that it picks any 5 states randomly and generates cross tabulation.

A close-up of a text

Description automatically generated

Code explanation:

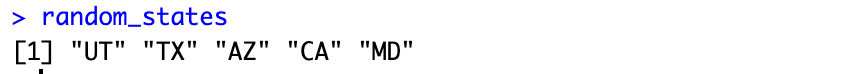
We have a for loop that iterates over a list of randomly selected states (from above code, using sample()). For each state, it filters the dataset “ab” to create a subset “i\_data” containing records specific to that state. Then, it creates a cross tabulation (state\_cross\_tab) that shows the counts of different outcomes ("Purchased" and "Not Purchased") for both the control and treatment groups. The table's row and column names are specified, and the resulting cross tabulation is stored in a list (`state\_cross\_tabs`) with the state name as the key. This process is repeated for each randomly chosen state.

A screenshot of a document

Description automatically generated

Here is a cross tabulation for five randomly selected states namely, UT, TX, AZ, CA and MD.

Random\_states variable with five states.



To compute this, I also used distinct keyword in code as we may come across duplicate selections if we select directly from the given column.

**9.**

A screenshot of a computer code

Description automatically generated

After cleaning the dataset, we get something like displayed in above screenshot. For this I used abandoned.csv and selected the usable columns. In new cleaned dataset,

Outcome represents whether the user purchased the package or not. I’ll also attach this dataset with assignment submission.

State\_Available and Email\_Available are binary attributes which tell us whether that value is present in the abandoned dataset or not.

Test\_Group tells whether a given entity is a Test\_Group or not.

**10.**

After running Linear Regression Model here are the results,

**Model 1**

A screenshot of a computer program

Description automatically generated

**Model 2**

A screenshot of a computer program

Description automatically generated

**Model 3**

A screenshot of a computer

Description automatically generated

**Interpretation:**

All the three models are used to predict the “Outcome” variable based on different combinations of independent variables. The coefficient for each independent variable represents how much the dependent variable is expected to change for a one-unit change in that variable, holding other variables constant. The R-squared values indicate the proportion of variance in the dependent variable explained by the independent variables.

Among the three models, Model 3 (pred3) appears to be the best choice because,

1. Adjusted R-squared: Model 3 has the highest adjusted R-squared value (0.02085). The adjusted R-squared considers the number of predictors in the model, providing a more reliable measure of model fit. A higher adjusted R-squared indicates a better fit of the model to the data.

2. Significant Coefficients: All the coefficients in Model 3 are statistically significant (indicated by the asterisks in the summary output) and a p-value<0.05. This suggests that all the predictor variables (Test\_Group, State\_Available, Email\_Available) are contributing meaningfully to the model.

As our data is mostly binary so the R-Squared value would be very low, as expected. But from all the three, Model 3 has the highest R-squared. And after looking at the coefficient p-values in Model 3 it seems all of them are statistically significant (i.e. >0.05) which means adding extra predictors has improved our model.

We can also verify our thoughts and calculations from stargazer output,

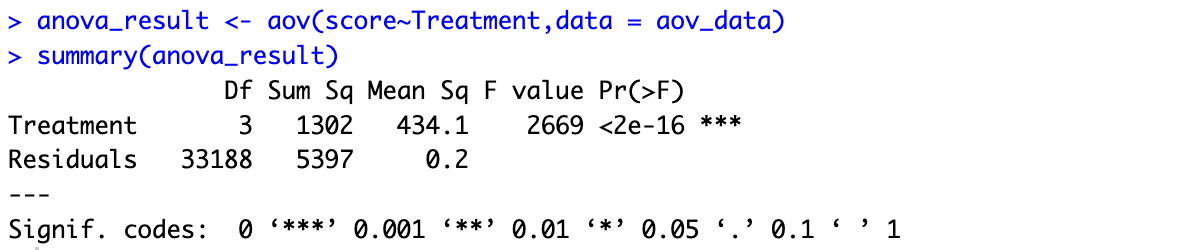
A screenshot of a computer

Description automatically generated

**Conclusion:**

Model 3 is the best.

**11.**



H0: There is no significant difference among group means.

H1: At least one of the group has a significantly different mean.

P-value(2e-16) < Level of significance (0.05)

**Interpretation:**

Based on the output, the ANOVA test shows that there is a highly significant difference between at least one pair of treatment groups. The p-value is extremely small (close to zero), which means we reject the null hypothesis that there are no differences between treatment groups. This result suggests that the treatment variable has a significant effect on the response variable.

**Output from TukeyHSD**

A screenshot of a computer code

Description automatically generated

Tukey's multiple comparisons of means is used to determine if there are significant differences between pairs of groups in a study.

With an extremely low p-value (< 0.05) in ANOVA test, indicating that there are significant differences in means among the treatment groups. This is further confirmed by the Tukey multiple comparisons, which show significant differences in means between all the groups, based on the close to zero P-values (all are <0.05).

A graph with numbers and lines

Description automatically generated

We can see the same trend followed in TukeyHSD plot as well. All the pairs (or mean differences) which do not have zero (or zero line) in their interval have significantly different means. Evidently, all the differ significantly in terms of their mean scores.

For the pairs below (or left) of the zero line, like Outcome-Email\_Available, the mean score for Outcome is significantly lower than Email\_Available. And for the mean differences at the right of the zero line, say pair P1-P2 is there. Then, mean score for P1 would be significantly higher than P2.

**Justification:**

- Both the regression analysis and the ANOVA test provide strong evidence of significant relationships between the independent variables and the outcome variable. The low p-values and highly significant coefficients in both analyses support this conclusion.

In summary, both the regression analysis and ANOVA test suggest that the independent variables (Test\_Group, State\_Available, and Email\_Available) have a significant impact on the outcome variable. The results of the regression analysis and ANOVA test are statistically comparable, reinforcing the relationship between the variables.

**12.**

The regression model provided attempts to predict the 'Outcome' variable based on three predictor variables: 'Test\_Group', 'State\_Available', and 'Email\_Available'. While the model shows **high statistical significance and good fit**, there are important considerations to debate regarding its appropriateness in making causal claims about the retargeting campaign's efficacy:

1. Statistical Significance:

- All three predictor variables (Test\_Group, State\_Available, and Email\_Available) show high statistical significance, indicating that they are associated with changes in the 'Outcome'.

- However, statistical significance does not imply causation. It only shows that there is an association between the variables. The causal relationship could be in the opposite direction or even bi-directional.

2. Causality vs. Association:

- Establishing causality requires more than just a statistical model. It demands a well-designed experimental setup where other potential factors are controlled.

- Observational studies, like this regression model, can only establish associations, not causality.

3. Unobserved Variables:

- There may be unobserved or uncontrolled variables (confounders) that could be influencing both the predictor variables and the outcome. Without accounting for these, causal claims are unreliable.

In conclusion, while this regression model provides valuable insights into the relationship between the predictor variables and the outcome, it is not sufficient to establish causal claims about the retargeting campaign's efficacy.

**13.**

**Model 1**

A screenshot of a computer

Description automatically generated

**Model 2**

A screenshot of a computer

Description automatically generated

**Model 3**

A screenshot of a computer

Description automatically generated

**Interpretation:**

We can see from the summaries that Model 2 and Model 3 perform very close, both are better than Model 1 in Adj R Squared and P-values (lowest). Lower P-value would result in more statistical significance of a predictor. Between Model 2 and Model 3 I’d prefer Model 2 as both have comparable Adj R-squared and similar levels of coefficient significance (maybe a touch more in Model 2).

More terms (coefficients) in Model 3 can improve model fit but can also increase complexity and make it hard to interpret. So, **Model 2** is the winner here.

A screenshot of a computer

Description automatically generated

A snippet from stargazer so that it’s easy to compare all the values. We can form the same interpretation from here as well.

**Final Comparison:**

So for **final comparison we have Model 3 (pred3) and Model 2 (interaction\_model2)**. The interaction\_model2 has high Adj R-squared, even lower coefficient p-values and same no. of terms (coefficients) which is good for interpretability and causes less complexity. So, **best Linear Model is interaction\_model2**.

**14.**

**Part 1**

The first part that I would like to modify is the Data Collection process. It is impossible to get a perfect dataset but still I’d love to try the same experimentation with a good dataset where it has less NaN values, more columns and preferably more no. of rows as well. The new columns could contain more data about users like National IDs so that it. Gets a bit easy and fool proof in mapping people in different datasets.

The second thing would be to have more data which should also include test\_data for our model. I really want to know how it would perform on such a real world scenario. How we can make it better in those cases?

**Part 2**

As discussed in Part1, we would have endless opportunities with better-quality data. In today’s world data is everything and for a ML model what we put in is what we would get out. So, better quality of data would be a big welcome. If we can have some variables related to personal preference then aside from retargeting, we can also recommend trips. Maybe I am daydreaming but with great data, I believe, we can create a solid recommendation system for Cruise trips.

**Part 3**

Yes, based on this analysis we can work on several business implications like,

Based on the analysis of the retargeting campaign, there are several actionable business implications that can be drawn:

1. Targeted Retargeting Strategies:

- The analysis could reveal specific customer segments that are more responsive to retargeting efforts.

2. Messaging and Content Optimization:

- Insights from the analysis can help in refining messaging and content for retargeting campaigns.

3. Personalization:

- Personalization can be enhanced based on the findings. As discussed, with good quality data we can try building a recommendation system for Cruise trips.

4. Customer Experience Enhancements:

- If there are common pain points or barriers identified through the analysis, the travel agency can work on improving the customer experience to increase conversion rates.

5. Campaign Timing and Frequency:

- The analysis can shed light on the optimal timing and frequency of retargeting efforts.

**15. 92**

Hi Dr. Z,

I have increased the score by 2 points because I believe in this submission I have raised the quality of both code and content, which took me quite some time. I hope you’ll like it ☺

P.S. Added detailed procedure and verification for removing duplicate records.