IHG & KSU Data Science Project

Spring 2023 DS7900

Everest: Matt Terry, Ebere Ogburu, Faheem Jamaloodeen, Faruk Muritala, Prativa Basnet 04/30/2023

Contents

Business Problem & Impact

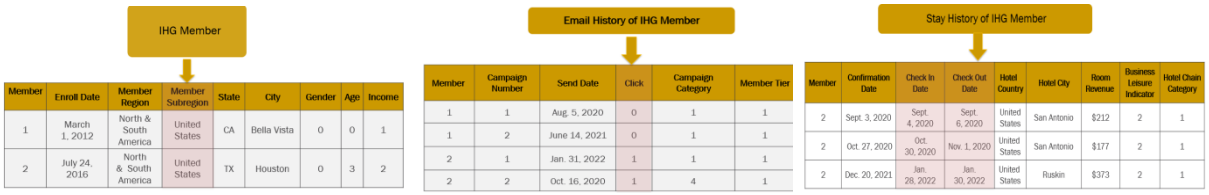
Team Everest had the opportunity to collaborate with our client, Intercontinental Hotels & Resorts, to help IHG solve the challenges they face in their complex database and systems. The client uses traditional email marketing to promote loyalty offers, point promotions, and specific target ad campaigns. In addition, they use commercial email marketing to increase patron awareness and brand awareness by directing a recipient to book or enroll in an offer. They measure the effectiveness of these campaigns and promos by the click through rate also known as CTR which is the ratio of the number of unique patrons who clicked on the link in an email to the total number of unique patrons who were sent the email campaign. While the click through rate is a decent metric to measure the success of a campaign, IHG needs help understanding which patrons click on email campaigns. IHG has tasked Team Everest to determine which variables are the most important in predicting a member's probability of clicking an email campaign and if we can determine the probability of a unique member clicking an email campaign. By creating a model that accurately predicts whether a member will click or not and what variables determine the probability of a member clicking, IHG can use this to evaluate the future of their email campaigns and target their patron demographics. Our goal is to leave IHG with a new mechanism to understand their data in diverse ways and open the doors to new opportunities with however they choose to proceed with our results.

The biggest challenge throughout this project was finding which coding applications we needed to use to run our code and create the best model possible. We leveraged SAS Studio, R Studio, and SAS enterprise miner to run our analyses and analytical methods. We made a last-minute decision to switch our joined dataset and run our code in R Studio to allow our model to achieve greater results. One other challenge not mentioned in our initial discovery was determining which analytical methods and analyses were appropriate to preparing our model, running it, and evaluating the performance of so. Our team discussed the pros and cons of each method and considered the business setting for our decisions.

Project Objectives & Expected Results

Our team's objectives were to create a model that could accurately determine the probability of a member clicking an email and find out what variables were predictive in determining the members probability of opening an email and clicking on an offer link given the datasets IHG provided the team. Our project timeline and weekly milestones can be seen in Appendix A1. IHG can expect us to deliver a model that can accurately predict whether a member will click on an email and generate results of the most predictive variables that determine clicks.

Data Summary & Target Variable.



IHG Member

Member	Enroll Date	Member Region	Member Subregion	State	City	Gender	Age	Income
1	March 1, 2012	North & South America	United States	CA	Bella Vista	0	0	1
2	July 24, 2016	North & South America	United States	TX	Houston	0	3	2

Email History of IHG Member

Member	Campaign Number	Send Date	Click	Campaign Category	Member Tier
1	1	Aug. 5, 2020	0	1	1
1	2	June 14, 2021	0	1	1
2	1	Jan. 31, 2022	1	1	1
2	2	Oct. 16, 2020	1	4	1

Stay History of IHG Member

Member	Confirmation Date	Check In Date	Check Out Date	Hotel Country	Hotel City	Room Revenue	Business Leisure Indicator	Hotel Chain Category
2	Sept. 3, 2020	Sept. 4, 2020	Sept. 6, 2020	United States	San Antonio	$212	2	1
2	Oct. 27, 2020	Oct. 30, 2020	Nov. 1, 2020	United States	San Antonio	$177	2	1
2	Dec. 20, 2021	Jan. 28, 2022	Jan. 30, 2022	United States	Ruskin	$373	2	1

Members (ID) – IHG could not provide the actual member names in the dataset for confidentiality reasons. However, IHG has provided one million members id and a column is associated with a unique numeric identifier for each IHG member, and some demographic information about the member whereas well provided such as Member Enrollment Date, Member Enrollment Channel, Member Region, Member SubRegion, Member Country, Member State, Member Gender - taken from loyalty data, Member Year Age as of send date - taken from loyalty data, Member Income as of send date - taken from loyalty zip -> census median income (by zip) with a Unique Member ID.

Email History – Email data of member Cohort which contains 192 million email history sent as dated from year 2020 to 2021. Email information includes email response indicators in addition to dynamics member information which are Unique Campaign Code, Date Campaign/Promotion Sent, Indicator if Member clicked, Campaign Category, Member Tier as of send date, Member Lifecycle code as of send date - aggregated based on historical stay data, Indicator if Member unsubscribed with its Unique Member ID.

Stay History – Stay information as of member Cohort dataset from 2020-2021 which contains five million stay history reports from IHG hotels. Stay information includes only features related to a member's stay activity. The stay history is associated with many chains categories variable better understood by IHG given as CHN_2 - CHN_CAT_1, CHN_3 - CHN_CAT_2, CHN_4 - CHN_CAT_1, CHN_5 - CHN_CAT_1, CHN_6 - CHN_CAT_3, CHN_7 - CHN_CAT_2, CHN_8 - CHN_CAT_1, CHN_9 - CHN_CAT_3, CHN_10 - CHN_CAT_0, CHN_11 - CHN_CAT_0, CHN_12 - CHN_CAT_2, CHN_13 - CHN_CAT_1, CHN_14 - CHN_CAT_2, CHN_15 - CHN_CAT_3, CHN_16 - CHN_CAT_1, CHN_17 - CHN_CAT_0, CHN_18 - CHN_CAT_3, CHN_19 - CHN_CAT_0, CHN_20 - CHN_CAT_3, CHN_21 - CHN_CAT_0, and CHN_22 - CHN_CAT_0. All these variables were tested with appropriate statistical tools and the significance variables were considered in our model.

Target Variable (Clicks) – these contain only two outcomes, Click and No Click associated with 1's and 0's in the dataset respectively. The 0's indicated a member did not click on the promo link sent by IHG and 1's indicates a member clicked on the promo link. From the dataset

provided, out of the 192 million campaigns sent out by IHG only 2 million unique clicks were recorded.

Expected Approach

Our expected approach was to create a binary logistic regression model to predict whether a member clicks or does not click because it would allow us to model the probabilities of a binary outcome based on the predictive variables that we create. Using binary logistic regression will allow our team to estimate and capture the effects of each variable going into the model which gets us closer to determining which variables are significant in determining clicks. For the scope of this project and presenting to business folks' interpretability and transparency was our top priority.

We chose Binary Logistic Regression over decision trees and neural networks. Binary logistic regression is highly explainable, interpretable, and simpler to understand than those other models out there. The only pitfall using the logistic regression in this project was that it would not be able to capture nonlinear relationships in the data. There were lots of variables in the data and it was complex; decision trees and neural networks would have been able to capture more relationships between the predictive variables and the click. A neural network would have been able to excel in large and complex datasets as well. Overall, we did create and test a decision trees model and agreed that given our team skills, abilities, and the time frame we were working with as well as the use case Binary logistic regression was our intended approach.

Data Preparation

Data preparation is a critical component of this data project analysis. It involves the process of cleaning, transforming, and organizing data so that it can be used effectively for analysis. The goal of data preparation is to ensure that the data is framed to suit the approach of this analysis and is accurately relevant to the analysis being performed. This process involves a variety of tasks, including data cleaning, data integration, variable creation, and data transformation.

Data cleaning involves the process of identifying and correcting errors or inconsistencies in the data. This includes filling in missing data such as blanks of guest quantity was imputed with the number of guests of 2, also for the scenario where it recorded a member had a stay in a room however the number of guests was recorded to be zero (0), this also corrected with the calculated mean value (2) of guest quantity.

Data integration involves the process of over-sampling clicks in the training dataset. After carefully considering the pros and cons of oversampling it is essential to employ the oversampling techniques to the clicks for its significance. This technique involves creating additional instances of the minority class of clicks which help improve the performance of predictive models. The class imbalance was observed to be 97.82% and 2.18% for the click rate

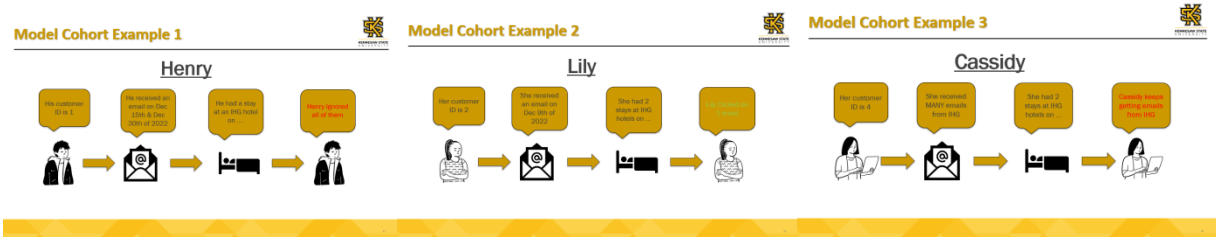
and the No click rate respectively. A 50.57/49.43 percent rate of oversampling was applied during the training phase of our model, and we achieved the balance percentage rate for the click rate and the No click rate.

Variable creation necessitates the creation of derived variables, which are calculated based on existing variables, or the transformation of existing variables given into a new format such as the creation of new features, aggregation of records, converting categorical variables to numerical variables, and binning of numeric variables. The goal of variable creation is to provide additional insights that measure specific characteristics and behaviors and information not available in the original data and ensure the new variables are reliable and valid, and accurately reflect the underlying processing that is more appropriate for analysis. Such as Stay Length, US region, Bin Guest Quantity, and other variables are available in Appendix …….

Data preparation is a time-consuming and complex process, but it is essential for ensuring that the results of data analysis are accurate and reliable. The data quality used in the analysis significantly impacts the accuracy of the results. From the historical dataset received which is dated from January 2020 to January 2022, we set a fixed 3-month window for the model cohort timeline and trace 6 months back from the most recent email sent to a member within the fixed 3-month window and aggregate the stay history of each member.

Taking time to properly prepare the data ensured that the analysis is working with the most relevant and accurate information possible. The effectiveness of this data preparation helps to identify potential issues in the data, which were addressed before the analysis begins. Overall, data preparation is a critical step in this project and was given careful consideration to ensure the best possible results.

Model Cohort Example



The model Cohort helps to identify the different members attribute after aggregation the history, the above first examples show that a member which had a stay in IHG hotel receive one email and fails to click on the promo link, while the second member also receive an email, had two stays in the hotel and decided to click on one of the email promo link, while the other

member received many emails from IHG, and also had stays but decided to not click on any of the promo link sent by IHG.

Variables that were significant

In statistical analysis, a backward selection test was applied to assess the significance of variables in a model by removing the least significant variable. Typically, a significance level of 5% is employed to determine if a variable is statistically significant in this project. If a variable's p-value is greater than 0.05, it is considered not significant at this level for this model. The table above shows variables significant in a model, meaning they do have a statistically significant effect on the click's response. These variables have shown a strong correlation with the outcome variable.

Model Preparation

In machine learning, a widespread practice is to split a given dataset into two subsets: a training set and a testing set. This is implored to evaluate model performance. The training set is used to train a model, and the testing set is used to evaluate its performance on unseen data. The 70/30 partition is a common split ratio used in machine learning for binary classification problems. This means that 70% of the data is used for training, and the remaining 30% is used for testing.

There is no one-size-fits-all answer to why this split ratio is used in binary logistic regression (BLR). However, there are a few reasons why it may be a popular choice:

1. Bias-variance tradeoff: If the training set is too small, the model may not be able to capture the underlying patterns in the data, leading to underfitting. On the other hand, if the training set is too large, the model may overfit and fails to generalize to new data. The 70/30 partition strikes a balance between these two extremes, allowing the model to capture the important patterns in the data while avoiding overfitting.
2. Computational efficiency: The larger the training set, the longer it takes to train the model. The 70/30 partition provides a good balance between having enough data to train the model effectively and keeping the training time reasonable.

3. Statistical significance: A larger testing set provides a more statistically significant evaluation of the model's performance. A 70/30 partition provides a large testing set, which can give a reliable estimate of how the model will perform on new, unseen data.
4. Data availability: In some cases, the size of the dataset may limit the amount of data that can be allocated for testing. In these cases, a 70/30 partition can be a reasonable choice to ensure that there is enough data to train the model while still having a meaningful evaluation of its performance.
5. Data imbalance (addressed through oversampling): In some datasets, the distribution of classes may be imbalanced, meaning that one class is much more prevalent than the other. In these cases, it may be necessary to adjust the partition ratio to ensure that both the training and testing sets contain a representative sample of both classes. A 70/30 partition may still be used, but the ratio may be adjusted to ensure that the minority class is adequately represented in both sets.

Models (Neural Network and Binary Logistic Regression)

We initially considered implementing a Neural Network model, but we dropped it before any testing was conducted in favor of a model with greater interpretability and transparency.

We can consider the problem of modeling patron click rate on an ad campaign sent from IHG using Binary Logistic Regression; the goal is to model the probability of an event occurring, given a set of predictor variables. In this case, the event of interest is whether a patron clicks on the ad, and the predictor variables may include factors such as the patron's age, gender, subscription length, and other relevant demographic or behavioral information. The logistic regression equation is

$$Y_i = \beta_0 + \beta_{1i} + \cdots + B_k X_{xi} + \in_i$$

where:

- *Y* is the binary response variable (whether the patron clicked on an ad or not)
- *X* is a vector of predictor variables (subscription length, respective season, and hotel chain)
- *β* is a vector of regression coefficients that determine the impact of each predictor variable on the response variable

To estimate the coefficients in the logistic regression equation, we can use maximum likelihood estimation. This involves finding the set of coefficients that maximize the likelihood of the observed data, given the logistic regression model. Once the coefficients have been estimated, we can use the logistic regression equation to predict the probability of a patron clicking on the

ad, based on their demographic and behavioral information. We can also use the model to identify which predictor variables are most strongly associated with the outcome of interest.

It is clear this model has many advantages over the Neural Network such as:

1. Interpretability: Binary Logistic Regression is a simple and interpretable model that allows for easy understanding of the relationship between the input variables and the probability of a patron clicking on an email. This can be useful for identifying important factors that influence patron behavior.
2. Efficiency: Binary Logistic Regression can be computationally efficient and may require less training time compared to more complex methods like Neural Networks or XGBoost.
3. Linear relationship: If the relationship between the input variables and the probability of a patron clicking on an email is roughly linear, Binary Logistic Regression can be a good fit for the data. This is because the model assumes a linear relationship between the input variables and the log odds of the outcome.
4. Avoids overfitting: Binary Logistic Regression is less prone to overfitting compared to some other methods like decision trees or random forests. This can be important when the dataset is small or when the number of input variables is large relative to the number of observations.

Other methods like Neural Networks, XGBoost, and random forests may also be effective at modeling patron click behavior depending on the specific characteristics of the dataset and the problem at hand. These models all behave and perform similarly. They have similar misclassification rates and errors. The important part is the variable creation and justifying why they work the best with our model of choice.

Targeted Model Evaluation and Performance

In this report, we present the performance evaluation of the develop model for the classification task of sentiment analysis on a dataset of members clicking on an email set to 20%, such that if a member probability is above the cutoff value of 20%, the model indicates a click (1) and otherwise. To assess the model's accuracy, we introduce a confusion matrix that summarizes the results of the model's predictions compared to the ground truth labels.

- Confusion Matrix for Validating Data

ACTUAL	PREDICTED	
	Click (1)	No Click (0)
Click (1)	281 (2.09%)	4 (0.03%)
No Click (0)	12987 (96.75%)	152 (1.13%)

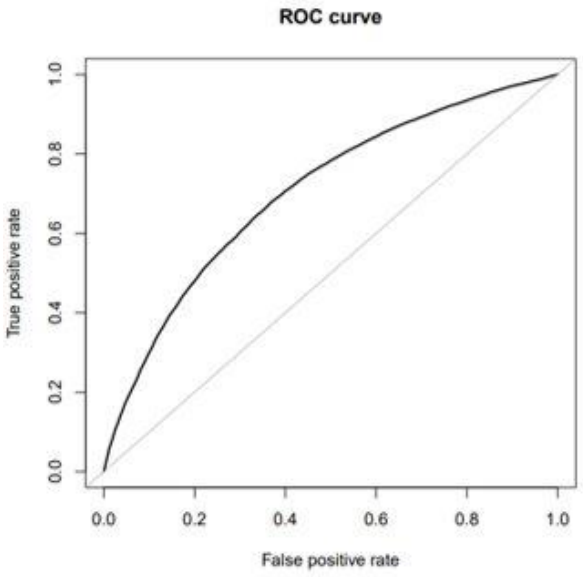
The above confusion matrix for our data validation shows that the model predicted 2.09% of the actual clicks as clicks and 96.75% of the actual no clicks as clicks, also predicted 0.03% of actual Clicks as No Clicks with 1.13% of actual No Click as No clicks. However, this indicates significant room for improvement.

- How well our model performed

Performance Metric	Value
True Positive Rate	98.6%
False Negative Rate	1.41%
True Negative Rate	97.4%
False Positive Rate	2.5%
AUC	70.5%

The confusion matrix for how well the model performed shows that the model correctly classified 98.6% of the positive reviews as positive and 97.4% of the negative reviews as negative, resulting in an overall accuracy of 70.5%. However, the model misclassified 2.5% of positive reviews as negative and 1.41% of negative reviews as positive. By analyzing the confusion matrix, we can identify the model's strengths and weaknesses and use this information to access the model's performance for IHG based on their recommendation.

- ROC



Recommendations

In the context of the problem, there are two ways to maximize the output of a Binary Logistic Regression model.

1. Identify the most influential factors; the coefficients in the logistic regression model help identify which factors have the strongest impact on patron click rate.
2. Predict the click rate of patrons; the trained model may be used to predict the probability of a patron clicking on an ad campaign based on the specific characteristics added to the model. IHG can use these predictions to personalize the experience of each integration using certain marketing campaigns.

This model was advocated because it identifies the most influential factors and predicts the click rate of unique patrons because of the high interpretability associated with each coefficient.

A new proposal that should be explored from our output would be the considerations of trend analysis. By analyzing the results of a logistic regression model over time, IHG can identify trends in patron behavior and adjust their marketing strategies accordingly. For example, one of the variables we considered in our model was the subscription length of each patron.

If the model indicates that patrons are more likely to unsubscribe from the membership list after a certain number of emails, IHG can adjust their marketing campaigns and limit the number of emails sent out to a specific demographic of members to reduce the likelihood of unsubscribing. This analysis could be explored through steps outlined in this table:

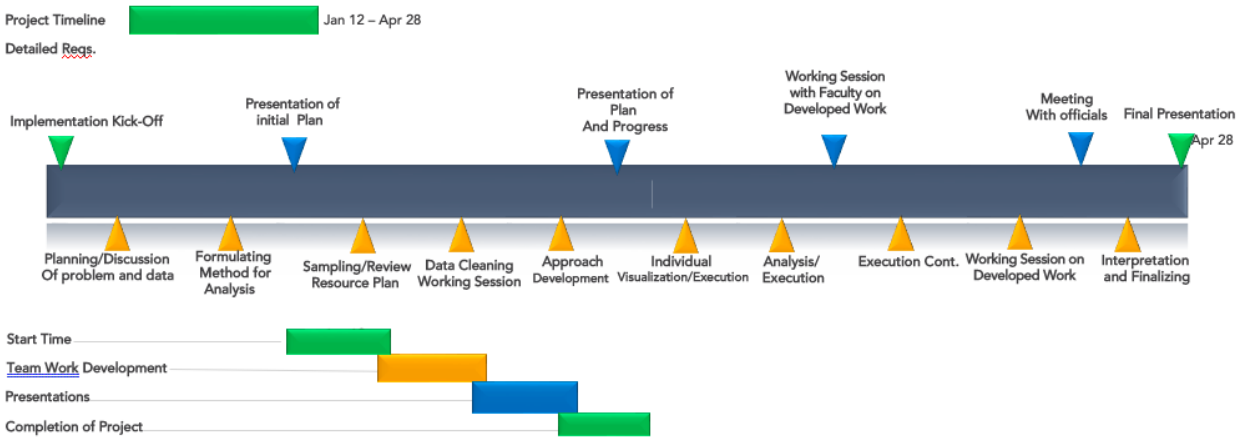
Step	Description	Importance
1	Identify subscriptions	Recognize selecting the relevant unit of analysis
2	Calculate the number of emails received by each patron who has unsubscribed	Provides data for analyzing email frequency and subscription length correlation.
3	Plot a scatter plot of the number of emails received versus the subscription length for patrons who have unsubscribed	Visualizes the relationship between email frequency and subscription length. A scatter plot can reveal patterns or trends between these variables.
	Calculate the correlation coefficient	Quantifies the direction and strength of the correlation between the

4	between the number of emails received and the subscription length for patrons who have unsubscribed	variables. A negative correlation coefficient indicates that as email frequency increases, subscription length decreases.
5	Perform a regression analysis to model the relationship between the number of emails received and the subscription length for patrons who have unsubscribed	Generates a mathematical model for the variables' relationship, which can estimate how much subscription length decreases with each additional email.
6	Repeat the same analysis for patron who have not unsubscribed	Creates a comparison group to examine if the correlation between email frequency and subscription length differs for patrons who have not unsubscribed.

By performing this trend analysis, we can determine if an excess number of emails is linked to patrons unsubscribing prematurely. If there is a negative correlation between the number of emails received and the subscription length, it may indicate that sending too many emails is causing patrons to unsubscribe and should be avoided to increase click rate.

Appendix

Below you can find our project timeline



Data Dictionary For Created Varibles		
Variable	Description	Values
Guest Quantity	Number of guests in a hotel room	1 - ∞
Bin Guest Quantity	Categorization of guest quantity	Single Occupancy (1) Double Occupancy (2) Group (3+)
Season	Categorization of time period based on travel patterns and preferences	Summer, Fall, Spring, Holiday, Normal, High Season, Moderate Season, Low Season
Binned Room Revenue	Categorization of room revenue, Min revenue, Max revenue, Revenue range	High Revenue, Moderate Revenue, Low Revenue
Stay Length	Total aggregation of nights spent in hotel	1 - ∞
Binned Stay Length	Categorization of length of stay	Short-Term Stay (1-3 nights), Medium-Term Stay (4-7 nights), Long-Term Stay (8-14 nights), Extended Term Stay (15+ nights)
US Region	Member State	Northeast, South, Midwest, West
Number of Email	Categorization of number of emails received by member	1-4: 5-10, 11-20, 21-30, 31+
Reward/Points	Indicates if member is enrolled in a rewards/points program	Yes, No

Clicks Aggregation	Total clicks made by member	0 - ∞
Dummy Clicks	Categorization of clicks made	0, 1
Room Revenue	Revenue generated by member's room(s)	0 - ∞
Lead Time	Number of days between email campaign send date and check-in date	Check-in date, Email campaign send date
Booking Window	Number of days between reservation date and check-in date	Check-in date, Reservation date
Member Duration	Number of days since member enrolled	Enrollment date, Email campaign send date
Campaign Frequency	Number of campaigns sent to member within a given time period	Time period, Email campaign send date
AverageRevPerGuest	Average revenue generated per guest	Room revenue, Guest quantity
AverageRevPerNight	Average revenue generated per night	Room revenue, Number of nights
PreLeadTime	Number of days between check-in date and email campaign send date	Check-in date, Email campaign send date

Data Dictionary for Created Variables		
Variable	Description	Values
Guest Quantity	Number of guests in a hotel room	1 - ∞
Bin Guest Quantity	Categorization of guest quantity	Single Occupancy (1) Double Occupancy (2) Group (3+)
Season	Categorization of time period based on travel patterns and preferences	Summer, Fall, Spring, Holiday, Normal, High Season, Moderate Season, Low Season
Binned Room Revenue	Categorization of room revenue, Min revenue, Max revenue, Revenue range	High Revenue, Moderate Revenue, Low Revenue
Stay Length	Total aggregation of nights spent in hotel	1 - ∞
Binned Stay Length	Categorization of length of stay	Short-Term Stay (1-3 nights), Medium-Term Stay (4-7 nights), Long-Term Stay (8-14 nights), Extended Term Stay (15+ nights)
US Region	Member State	Northeast, South, Midwest, West
Number of Email	Categorization of number of emails received by member	1-4: 5-10, 11-20, 21-30, 31+
Reward/Points	Indicates if member is enrolled in a rewards/points program	Yes, No
Clicks Aggregation	Total clicks made by member	0 - ∞

Dummy Clicks	Categorization of clicks made	0, 1
Room Revenue	Revenue generated by member's room(s)	0 - ∞
Lead Time	Number of days between email campaign send date and check-in date	Check-in date, Email campaign send date
Booking Window	Number of days between reservation date and check-in date	Check-in date, Reservation date
Member Duration	Number of days since member enrolled	Enrollment date, Email campaign send date
Campaign Frequency	Number of campaigns sent to member within a given time period	Time period, Email campaign send date
AverageRevPerGuest	Average revenue generated per guest	Room revenue, Guest quantity
AverageRevPerNight	Average revenue generated per night	Room revenue, Number of nights
PreLeadTime	Number of days between check-in date and email campaign send date	Check-in date, Email campaign send date