Hotel Booking Cancelation Prediction

Abhishek Puranik
Sanil Patel
Aditi Chauhan
Prachi Mahapatra

Executive Summary

This report comprises of a predictive model deployed on the selected dataset of the hotel, to find the probability of cancellation of their reservations. It contains related visualizations, sorting of the customers and is intended to help the marketing team of the hotel in setting their advertising strategies to their potential customers.

Business Problem Statement

In todays fast paced world, many hotel and resort businesses face the challenge of forecasting the possibility of cancelation of reservations by their customers. This issue affects their inventory, supply chain and meal plan budget.
In this project, we are trying to predict the likelihood of cancellation of reservation of the resort and city hotel using our prediction model. This will help the hotel management to find their target customers, advertise to them smartly and to increase their overall profit.

Dataset Description

The data set has been taken from Kaggle.com.
Dataset link: https://www.kaggle.com/jessemostipak/hotel-booking-demand

The Data set contains the booking information of two category a city hotel and a resort hotel which has information such as when the booking was made, length of stay, the number of adults, children, and/or babies, among other things. We have **32 columns and 119311 rows**.

Variables	Description
hotel	Hotel (H1 = Resort Hotel or H2 = City Hotel)
is_canceled	Value indicating if the booking was canceled (1) or not (0)
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
arrival_date_year	Year of arrival date
arrival_date_month	Month of arrival date
arrival_date_week_number	Week number of year for arrival date
arrival_date_day_of_month	Day of arrival date
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
adults	Number of adults
children	Number of children
babies	Number of babies
meal	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	Country of origin. Categories are represented in the ISO 3155–3:2013 format
market_segment	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
distribution_channel	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0)
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking

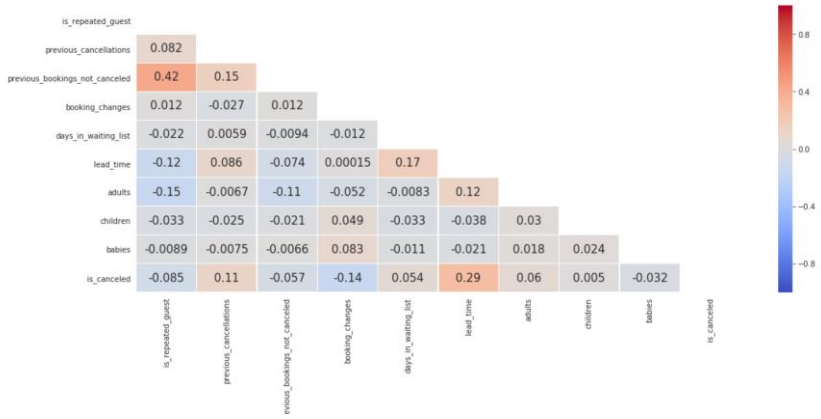
previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons
assigned_room_type	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
Agent	ID of the travel agency that made the booking
Company	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Type of booking, assuming one of four categories:
	Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
Adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces	Number of car parking spaces required by the customer
total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
reservation_status_date	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel
How many observations in the dataset?	119311
How many Binary variable?	Binary: 2, Nominal: 11
How many continuous variables?	19
What is the target variable?	is_cancelled
If binary or nominal: what percentage of the variables belong to each class?	Binary and nominal: 40.62%
If continuous: what is the mean value of target variable?	Target variable is binary
Before doing any further processing, what would your prediction of target variable be?	As our target variable is of binary nature, the result will have only two possible outcome, i.e. 0: not cancelled and 1: cancelled.
After preprocessing no of variables selected	10
Total no of Dummy Variables	69

Dashboard Link:
Other visualizations related to our dataset
https://public.tableau.com/views/Hotel_Booking_Dashboard/Dashboard1?:display_count=y&publish=yes&:origin=viz_share_link

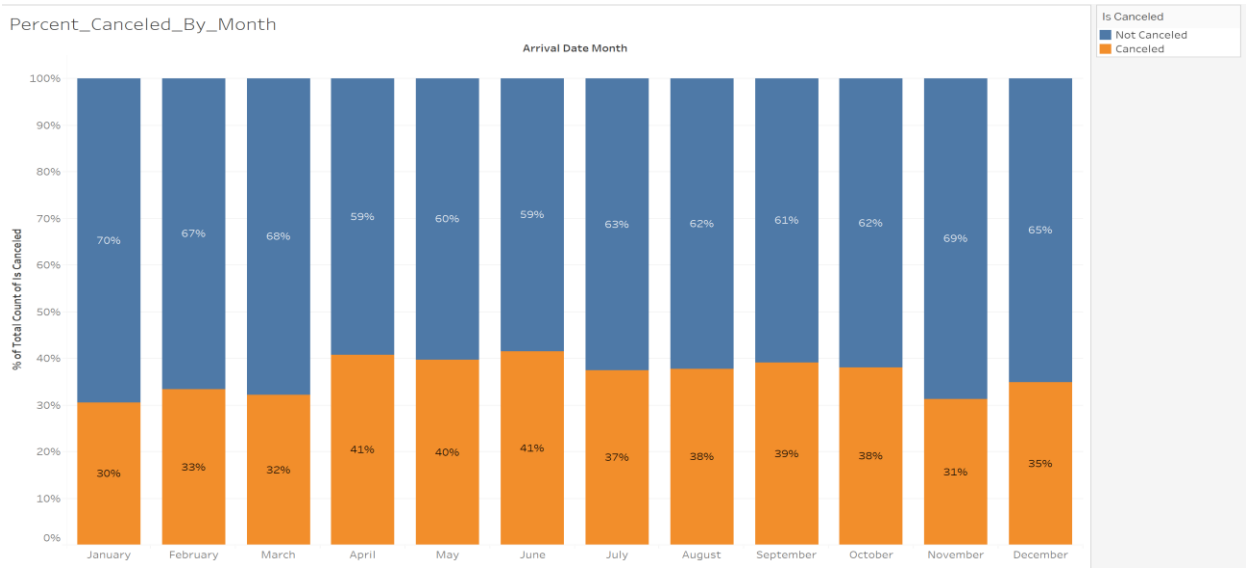
Data Visualization: Exploratory Data Analysis

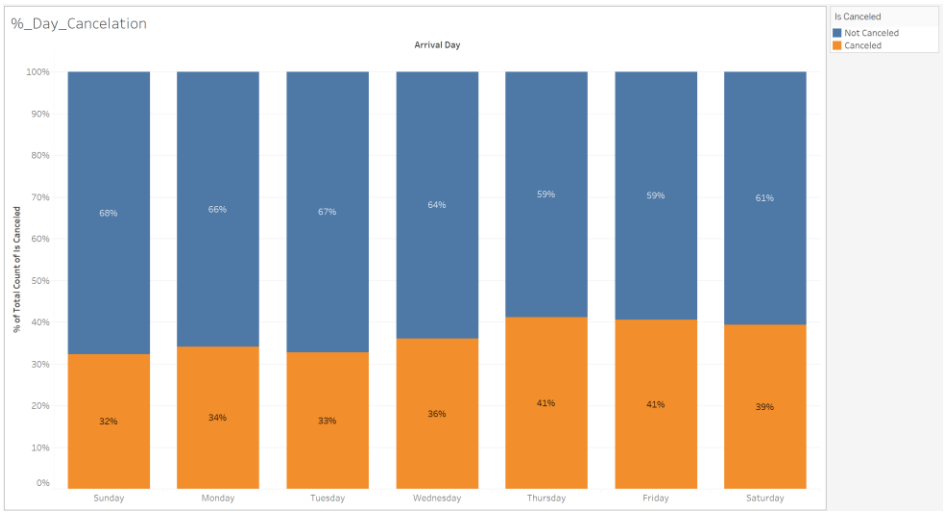
We will visualize our data using Tableau and Python. Through these visualizations, we can interpret patterns and outliers in the data.

1. Correlation matrix: is_canceled is related to lead time



2. Month and day has the highest number of cancellations?



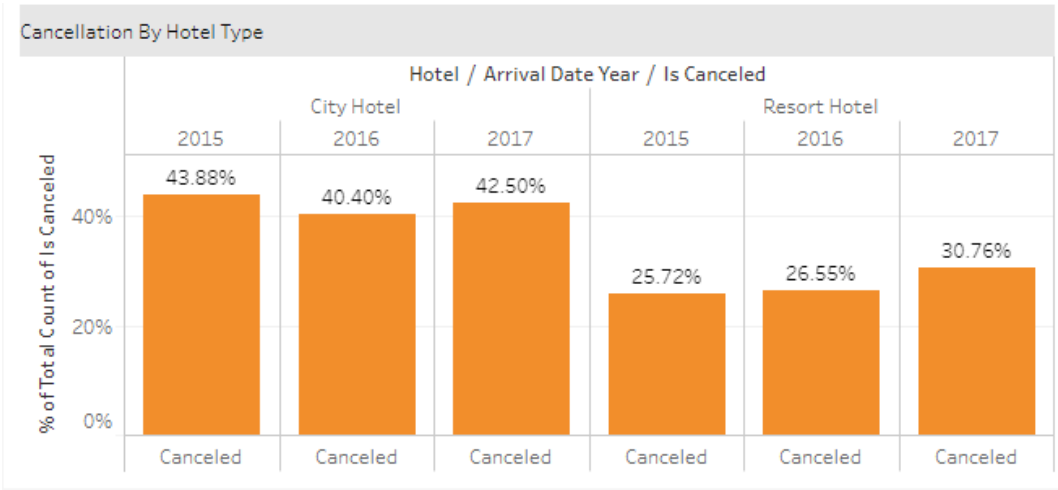


April and June had maximum cancelation almost close to 41%.
In terms of days Thursday and Friday had more cancellation than other days of the week.

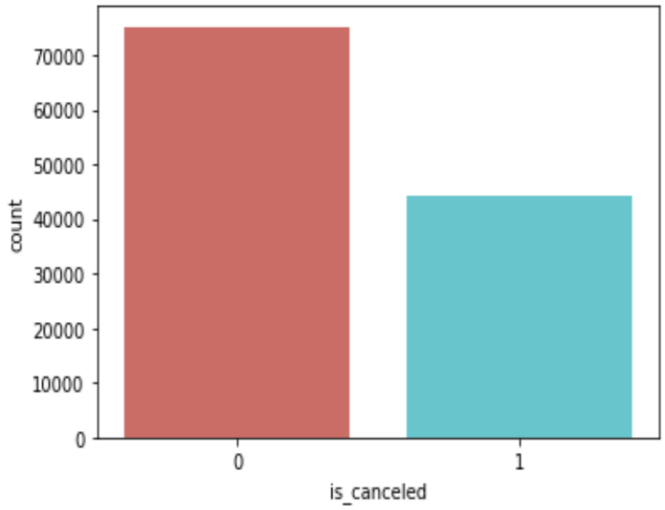
3. Hotel Type has maximum Cancellation and which shows how much-repeated guest has cancelled.

From the above graph we can clearly see that City Hotel had maximum cancellation.
Also, the not repeated guest has done more cancelation than the other indicating that both the hotel and resort is a place where people would like to spend their time.

4. Percentage of cancellation of city hotel and resort hotel in the year 2015, 2016 and 2017.



DATA PROCESSING



The Graph Represents the no of Cancelled and not cancelled Customers booking in past three years 2014,2015 and 2016.

Step 1:

There were total 31 independent variables and intuitively we have dropped 5 variables because these variables will not have much influence on the prediction of the **cancelation status**. Also, Company variable has 90% null values and was not contributing much to the prediction.

'agent', 'company', 'reservation_status', 'reservation_status_date','country' are the variables we removed. Hence, we are left with only 27 variables for our further prediction.

There are 9 categorical variables in our data. We have used LabelEncoder() and OneHotEncoder() to normalize the labels of such variables.

Step 2:

We divide the data into three parts i.e. training:70%, test:20% and validation:10%.
Following is the result:

Training: 83570
Test: 23878
Validation: 11938

Step 3:

Following are the two models which we are using for our predictive analysis:

- Logistic Regression
- Classification Tree

Logistic Regression:

Recursive Feature Elimination (RFE) which uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

Following are the top 10 features that we have selected.

	Feature	Ranking
60	assigned_room_type_11	1
65	customer_type_1	1
16	required_car_parking_spaces	1
63	deposit_type_2	1
61	deposit_type_0	1
43	market_segment_6	1
62	deposit_type_1	1
57	assigned_room_type_8	1
10	previous_cancellations	1
58	assigned_room_type_9	1

We implemented the Logit function to check for p-value. We have dropped 2 variables ['assigned_room_type_11' and 'required_car_parking_spaces'] which have their p value more than 0.05.

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.477469
         Iterations: 35
```

```
C:\Users\ADITI\Anaconda3\lib\site-packages\statsmodels\base\model.py:512: ConvergenceWarning: Maximum Likelihood optimizatio
n failed to converge. Check mle_retvals
  "Check mle_retvals", ConvergenceWarning)
```

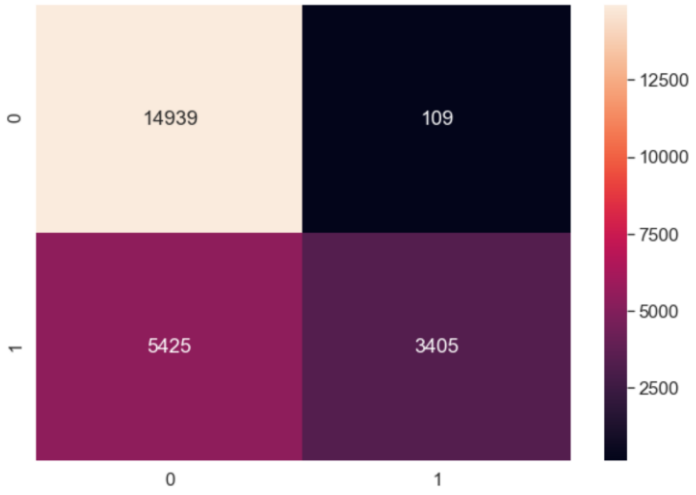
```
                           Logit Regression Results
==============================================================================
Dep. Variable:           is_canceled   No. Observations:                83570
Model:                         Logit   Df Residuals:                    83560
Method:                          MLE   Df Model:                            9
Date:                Sat, 25 Apr 2020   Pseudo R-squ.:                 0.2758
Time:                        21:54:21   Log-Likelihood:                -39902.
converged:                      False   LL-Null:                       -55100.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
assigned_room_type_11          16.1860    472.524      0.034      0.973    -909.944     942.316
deposit_type_2                 -1.1963      0.236     -5.071      0.000      -1.659      -0.734
deposit_type_1                  4.7833      0.123     38.988      0.000       4.543       5.024
market_segment_6                1.0492      0.018     57.229      0.000       1.013       1.085
assigned_room_type_9           -2.4053      0.389     -6.191      0.000      -3.167      -1.644
deposit_type_0                 -1.4925      0.015    -99.691      0.000      -1.522      -1.463
customer_type_1                -1.2315      0.177     -6.950      0.000      -1.579      -0.884
previous_cancellations          1.7489      0.050     35.016      0.000       1.651       1.847
assigned_room_type_8           -3.3406      0.583     -5.733      0.000      -4.483      -2.199
required_car_parking_spaces   -33.2806   1398.515     -0.024      0.981   -2774.319    2707.758
------------------------------------------------------------------------------------------------
```

After the initial analysis of the data using Logistic Regression, we have got the accuracy of **76.79%**.

Confusion Matrix

Threshold Probability	Accuracy
0.05	37.33
0.1	37.73
0.2	63.9
0.4	**76.82**
0.55	75.7
0.65	75.7
0.8	75.13
0.95	75.14

Accuracy of logistic regression model	76.82%
Values predicted correctly	14,939 + 3405= 18,344
Values predicted wrong:	109 + 5,425=5,534



Step 4:

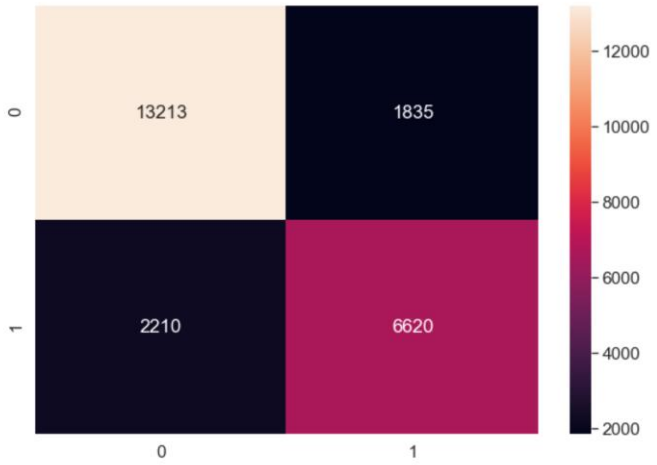
Classification Tree:

We have built decision tree using two criteria i.e. Gini index and Entropy two achieve maximum accuracy.
By implementing the decision tree using **Gini Index**, we found the following results:

Depth	Accuracy
5	79.78
6	80.55
7	80.98
8	81.14
9	81.47
19	82.06
20	82.87
21	**83.02**
25	82.63
120	82.38

Accuracy of decision tree for criterion = 'gini'	83.02%
Values predicted correctly	13213 + 6620= 19823
Values predicted wrongly	2210+1835=4055

Confusion Matrix

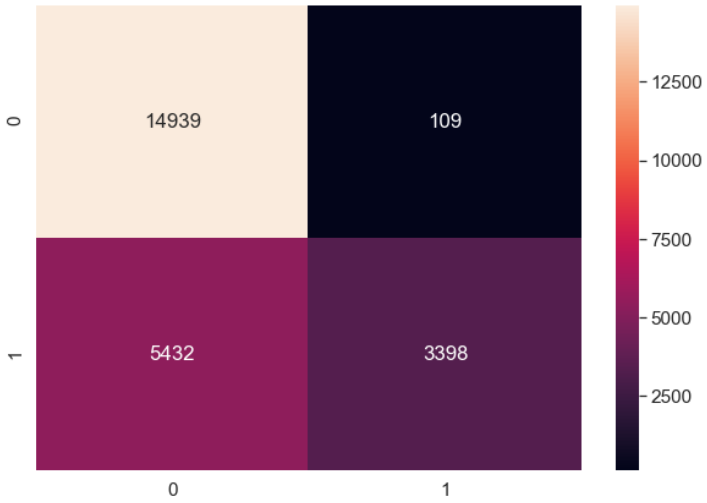


By implementing the decision tree using **entropy**, we found the following results:

Depth	Accuracy
11	81.69
13	81.83
17	82.61
19	82.85
21	82.74
22	**82.86**
23	82.87
24	82.78

Accuracy of decision tree for criterion = 'entropy'	82.87%
Values predicted correctly	13238 + 6552= 19780
Values predicted wrongly	2278+1810=4088

Confusion Matrix



So, we have selected **criteria = gini index** and **depth = 21,** as it gives maximum accuracy for the model.

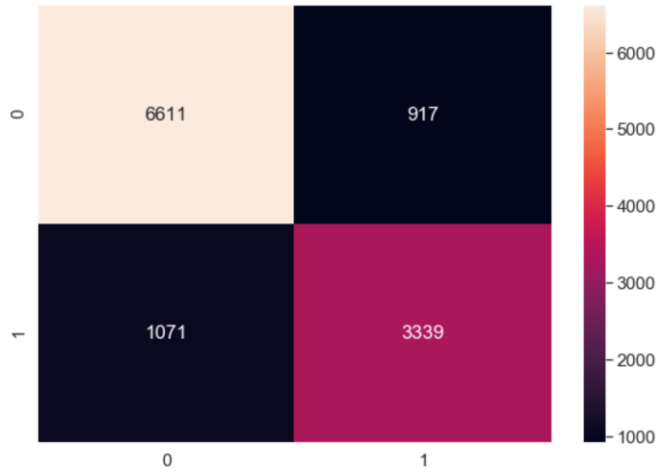
Step 5:

Deploying the selected model on the validation data for 'classification'.

We have selected **Decision Tree** to be deployed on our 10% validation data, as this model has the best accuracy compared to **Logistic Regression**.

Accuracy of decision tree on Test Data (10%)	83.27%
Values predicted correctly	6611 + 3339= 9950
Values predicted wrongly	1071+917=1988

Confusion Matrix



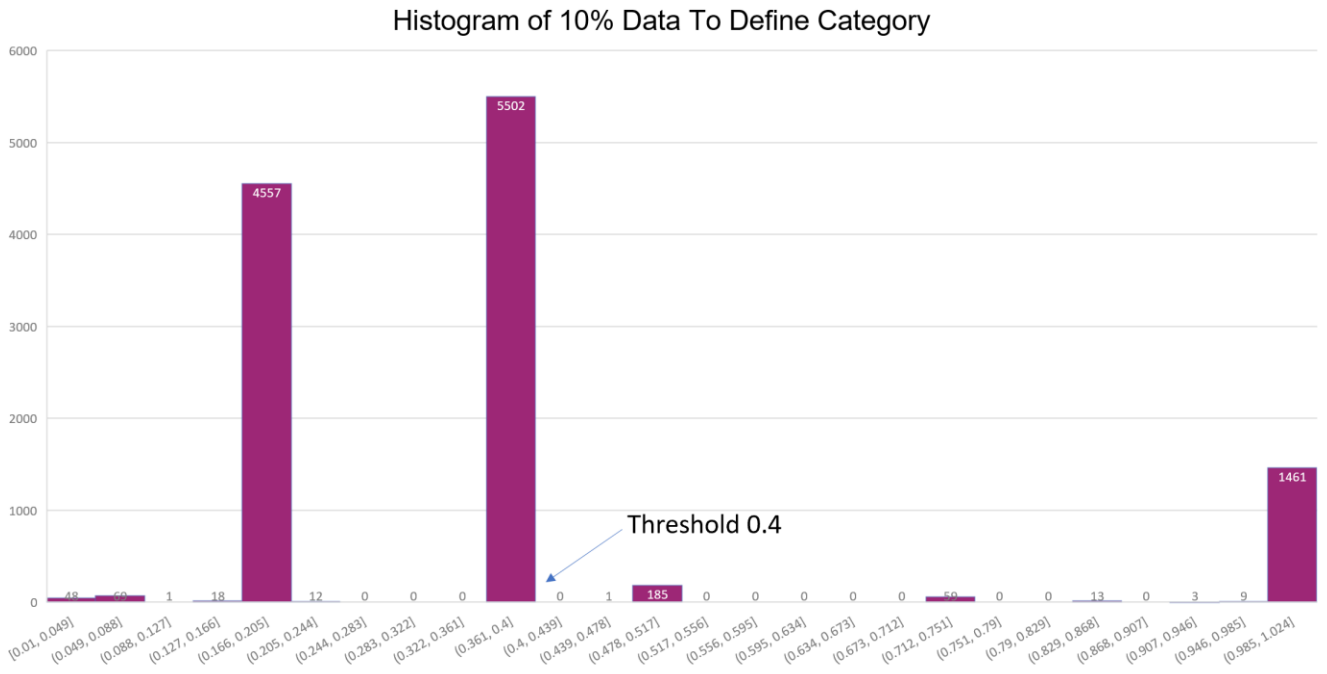
Step 6:

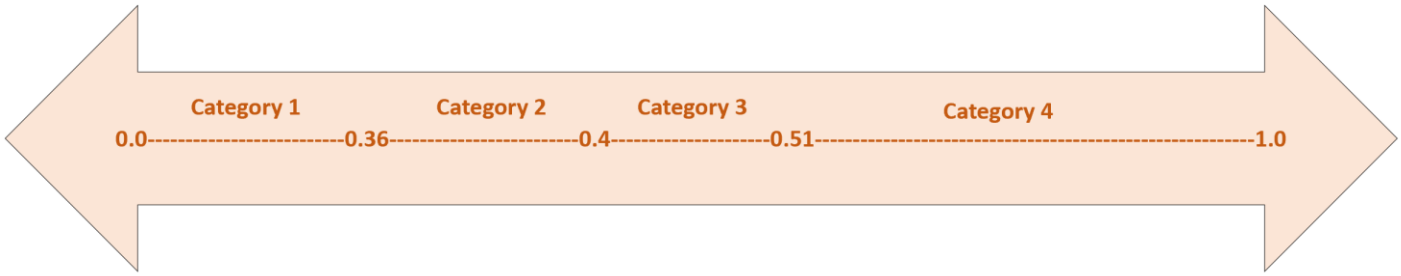
Deploying Logistic Regression model on the validation data for 'category selection'.

To identify our target customers on the basis of probability of cancellation, we'll be using logistic regression over validation data.

This model will be focusing on sorting the customers in different categories based on the probability of cancelling their reservation. The graph below is showcasing the customers in different categories which can be useful for the marketing team of the hotel to decide policies and advertising strategies.

This categorical bifurcation will also help the marketing team to find their target customers in order to provide them with some customize benefits in their reservation plans to prevent their cancellation.





Summary:

Category 2 (0.36-0.4) :
This category contains customers who have slight chance of cancelling the booking ranging around 30%- 40% . With good advertising, there is a healthy chance of such customers not cancelling their booking.

Category 3 (0.4-0.51):
This category consists of potential target customers who are on the verge of cancellation. If targeted advising and customized booking plans are provided to such customers, their cancellation can be prevented. Also, as the count for this category is less than the others, providing customized plans to such customers will not cost much to the pocket.

Category 1 and Category4:
This category consists of the customers who are not suited for the advertising policy of the hotel as they have least chance of cancellation or are on the verge of cancelling the reservation. If the marketing team focuses on such people, it might cost a lot value to the hotel.

Summarizing the above insights, we find our target audience as **Category 2** and **Category 3.** Focusing on the selected reservations, the hotel can gain **maximum profitability**.