**Data Analysis on**

**Hotel Booking System**

**By Team 3,**

**Altamash Quereshi**

**Apoorv Akhouri**

**Darpan Shah**

**Ramya Ramachandran**

**Vaibhavi Sahane**

**Under the guidance of,**

**Prof. Edward Stohr**

**INDEX**

Introduction…………………………………………………………………. 3

Objective……………………………………………………………………. 4

Modelling Technique……………………………………………………….. 5

Accuracy Measures…………………………………………………………. 7

Data Description……………………………………………………….......... 8

Data Cleaning………………………………………………………………. 10

Exploratory Data Analysis………………………………………………….. 11

Model Inference…………………………………………………………….. 16

Discriminant Analysis………………………………………………………. 17

Conclusion & Reference...………………………………………………….. 19

**INTRODUCTION**

Hotel industry is one of the leading industry in 21st century, in order to keep track of the growing business and keep records of the day to day activities such as room booking, cancellation status and how do these parameters affects the business; our team has come up with the perspective of data analysis and implementing a predictive model based on huge set of data (downloaded from kaggle) based on activity of room booking, cancellation status and leading time with certain cancellation policy decided by our team.

Our prior objectives for the project is to predict the cancellation of customer hotel booking and optimizing the revenue of the hotel. Based on the cancellation policy which is charging customers a penalty price if they are unable to cancel the booking before 72 hrs of the day of check-in. In order to predict the cancellation status of the hotel booking, a random forest classifier has been implemented with the help of selected variables based on VIF (variance inflation matrix) and correlation matrix. Python (Jupyter Notebook) have been implemented to fetch desired outcomes. Whereas, for the revenue analysis we have performed Discriminant Analysis which uses a return matrix and confusion matrix to decide whether a revenue has been obtained or the cost has been incurred. If the prediction turned out to be correct, we can work on saving a particular booking from getting cancelled.

Also, for the better understanding of the output and relation between various variables of the dataset, visualization has been implemented with the help of visualization tool Tableau. Our project consists of a bar graph and line graph for a clearer approach of displaying output in a more desired way.

Finally, we are grateful for the guidance and motivation given by Prof Edward Stohr to move forward with our project idea and implementation.

**OBJECTIVES:**

1. Predict the cancellation of customer hotel booking
2. Optimize the revenue for the hotel

**PROBLEM STATEMENTS:**

1. To prevent the cancellation status of the hotel booking with the help of prediction model and cancellation policy.
2. To calculate the revenue of the hotel based on booking status by discriminant analysis.

**MODELLING TECHNIQUE**

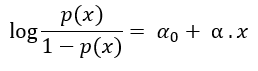
**Logistic Regression:**

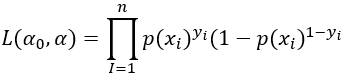
A statistical model typically used to model a binary dependent variable with the help of logistic function. It establishes a relationship between dependent and independent variables. Another name for the logistic function is a sigmoid function and is given by:

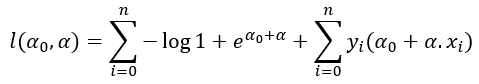
This function assists the logistic regression model to squeeze the values from (-k, k) to (0, 1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

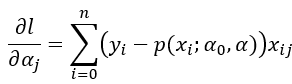
The reason behind this model is that just like Linear Regression, logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task.

**Math behind Logistic Regression:**

So it all start with a linear function p(x) and then using log function with p(x) we are able to bound this function to 0 to 1. So the function will be like

Since Logistic regression predicts probabilities, we can fit it using likelihood. Now the likelihood can be written as:

Further, after putting the value of p(x):

In order to increase the probability of occurring we can use Maximum Likelihood function and differentiating the equation with respect to different parameter and setting it to zero.

= 0

**Assumptions:**

* The dependent variable is categorical. Dichotomous for binary logistic regression and multi-label for multi-class classification
* Attributes and log odds i.e. log (p / 1-p) should be linearly related to the independent variables
* Attributes are independent of each other (low or no multicollinearity)
* In binary logistic regression class of interest is coded with 1 and other class 0.

**Discriminant Analysis:**

It is a statistical technique which allows the researcher to study the difference between two or more group of objects with respect to several variations simultaneously. In the social sciences, there are a wide variety of situations in which this technique may be useful.

Here, due to the extremely large size of the data, we leave the predictions to be done on Python from where we extract the confusion matrix & apply it with the average tariff of booking the hotel for one night.

The concept being, if we correctly predict that a customer will not cancel the booking, we can count that as revenue, whereas if we wrongly predict that the customer will not cancel the booking & he/she does, we count that a cost incurred by us. This is a cost incurred because the hotel would be working on customer engagement & retention by sending out promotions/deals to a customer that is going to cancel his/her booking.

On the flip side, if we make the right predictions, we can work on saving that booking from potential cancellation, or in terms of those bookings which are predicted to not be cancelled, the hotel can work on increasing their revenue through customer retention or engagement.

Another way of looking at this is that this helps us by putting a value on how beneficial our model is in terms of revenue corresponding directly with the accuracy of the model.

**ACCURACY MEASURE**

**Mean Squared Error**:

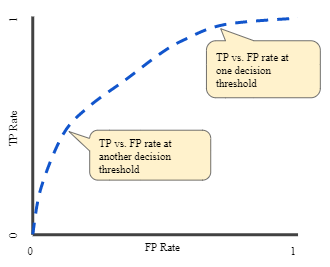
The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.

**Confusion Matrix**:

It is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. This consist of four different parts:

* True Positive (TP) - These are cases in which we predicted yes, and that’s actually yes.
* True Negative (TN) - We predicted no, and it’s actually no.
* False Positive (FP) - We predicted yes, but they actually was no.
* False Negative (FN) - We predicted no, but actually it is yes.

**ROC Curve and AUC:**

An ROC Curve (receiver operating characteristic curve) is a graph showing the performance of a classification model with the help of True Positive Rate and False Positive Rate at different classification thresholds. AUC stands for Area under the ROC curve measures the entire two-dimensional area underneath the entire ROC curve.

**DATA DESCRIPTION**

The datasets available were collected aiming at the development of prediction models to classify hotel bookings likelihood to be canceled. The dataset we had was for the period of three years ​

* 2015 (July to December) ​
* 2016 (January to December) ​
* 2017 (January to August) ​

Initially we had 32 columns in the dataset, out of 32 columns we decided to select 17 columns to make the predictive modelling, optimization and visualization.

* **is\_canceled** - Value indicating if the booking was canceled (1) or not (0)​
* **previous\_cancellations** - Number of previous bookings that were cancelled by the customer prior to the current booking​
* **distribution\_channel**- Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”​
* **stays\_in\_weekend\_nights :** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* **stays\_in\_week\_nights:** Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* **adr:** Average Daily Rate as defined by
* **lead\_time** - Number of days that elapsed between the entering date of the booking into the system and the arrival date​
* **is\_repeated\_guest** - Value indicating if the booking name was from a repeated guest (1) or not (0)
* **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
* **days\_in\_waiting\_list:** Number of days the booking was in the waiting list before it was confirmed to the customer
* **customer\_type:**
* 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
* **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.
* **arrival\_date\_year** – Year of arrival date
* **arrival\_date\_month** - Month of arrival date with 12 categories: “January” to “December”
* **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

**DATA CLEANING**

For cleaning the dataset following steps were followed: ​

* Identified if there were any null values or NA​,
* There were 488 null values identified under the country ​variable. Renamed all the null value to “no country”
* 16,340 null values under agent column, renamed all the null values to “no agent”
* ​1,12,593 null values under company column, rename all the null values to “no company

|  |  |
| --- | --- |
| **Variable name** | **Conversion into Float** |
| Family | If family has no children, then = 0  If family has 1 or more then = 1 |
| distribution\_channel\_new | TA/TO = 1  Corporate = 2  Direct = 3 |
| stays\_in\_weekend\_nights | If family does not stay at night, then = 0  If family stays 1 or more then = 1 |
| adr\_bin | If adr is less than or equal to 50, then= 1  If adr is between 50 to 100 then = 2  If adr is between 100 to 150 then = 3  If adr is greater than 150 then = 4 |
| lead\_time\_bin | If lead time bin is less than or equal to 30, then= 1  If lead time bin is between 30 to 60 then = 2  If lead time bin is between 60 to 90 then = 3  If lead time bin is greater than 90 then = 4 |
| previous\_bookings\_not\_canceled | If previous booking not greater than 1 = 0  If previous booking greater than equal to 1= 1 |
| wrong\_room | If reserved type is equal to assigned type = 0  If reserved type is not equal to assigned type = 1 |
| waiting\_list | If days\_in\_waiting\_list is equal to 0 = 0  If days\_in\_waiting\_list is not equal to 0 = 1 |
| customer\_type\_new | Transient= 1  Contract= 2  Transient-Party= 3  Group = 4 |
| no\_deposit | If deposit\_type is equal to ‘non- refundable’& ‘refundable’= 0  If deposit\_type is equal to ‘No Deposit’= 1 |

* We did not drop any column due to null values
* We renamed few variables and converted them from string to float for a better predictive modeling

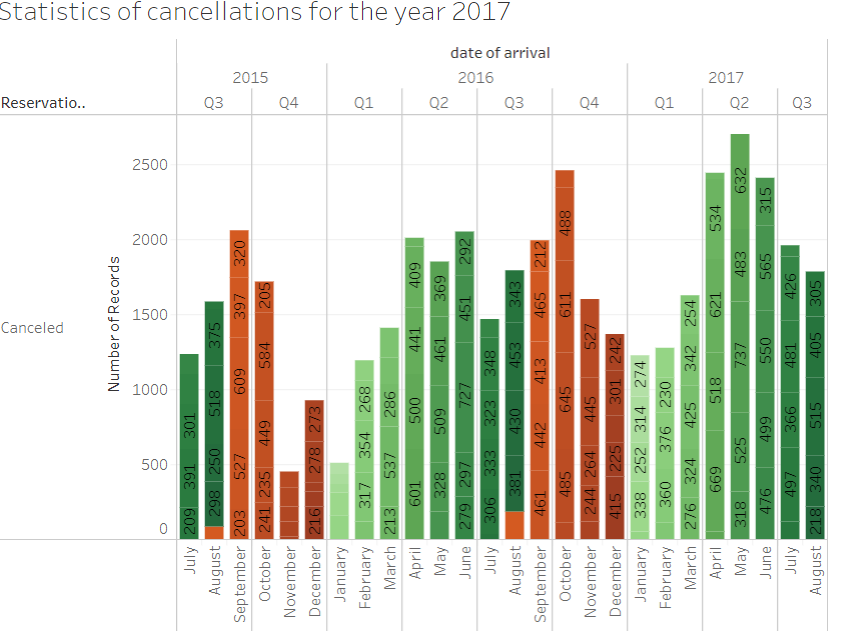
**EXPLORATORY DATA ANALYSIS**

To get to know our data better, we did some findings out of customers in an attempt to know the cancellation process and find out what reasons that change a customer’s mind and hence they cancel their booking.

1. Cancellation for different years.

We tried to plot the cancellations that happened over the year 2015-2017 according to the month and weekly and Quarterly.

**Variables used –**

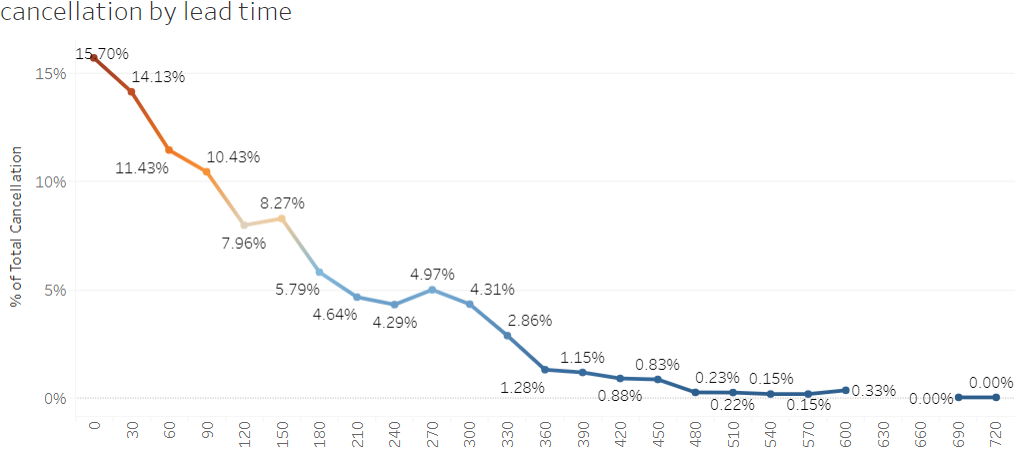
* Reservation status (i.e. is cancelled)
* Month of the date of arrival
* Years (Divided into Quartiles)

The visualization depicts that the cancellation rate increased in the second Quartile of 2017 which was an unusual rise in the cancellation compared to 2015 and 2016, where the most cancellation happened in the 3rd and 4th quarter. The month of December in each year had the least cancellation which is due to the holiday season.

1. Cancellation by lead time

This graph was an important one as if was necessary to figure out how early a customer cancels their booking once done. Early the customer cancels more refund they get. Hence the Company needs to know how much about should be refunded in order to make a profit or engage a customer. Higher the amount charged for cancellation lesser chances will be there that the customer cancels.

**Variables used** – Lead time

****We created bins for the lead time and calculated the percentage of total cancellation based on lead time.

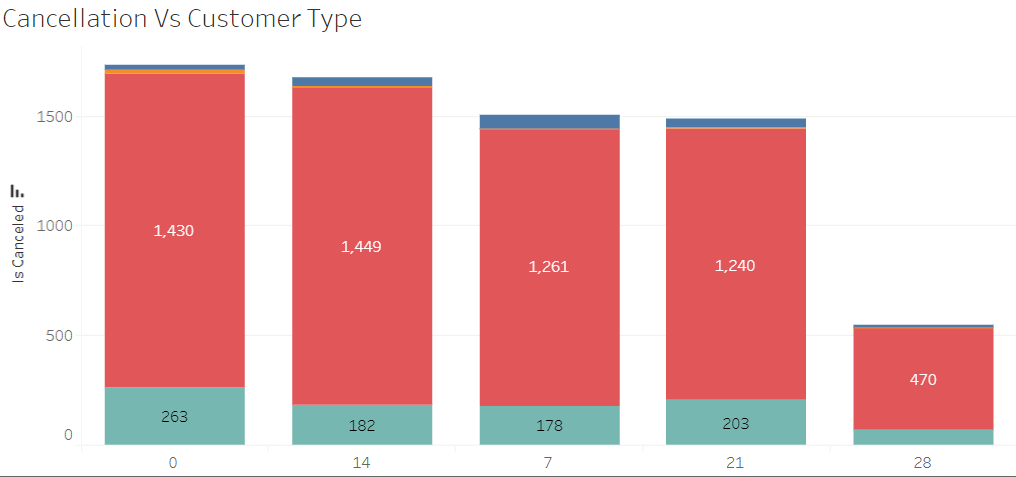
As we can observe from the graph, there is a downward trend in the total % of cancellation. This means when the lead time is longer, fewer people would prefer to cancel the reservation.

1. Cancellation vs Customer type

As stated earlier our data consisted of 4 type of customers that made the bookings. Our objective was to find out which type of customer had the most cancellation.

Variables used –

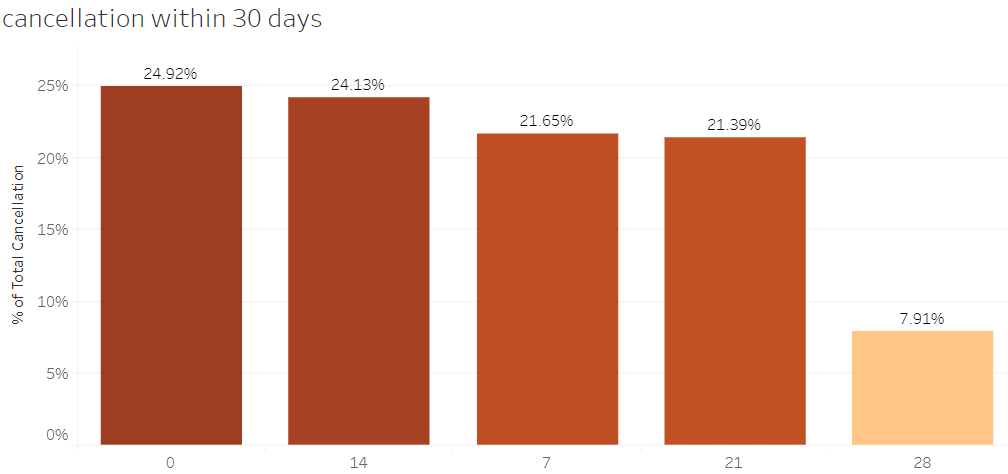
* is cancelled
* Customer type



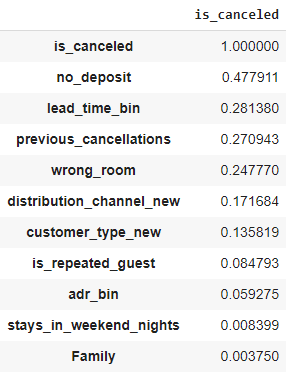
We found that the Transient customers had the most cancellation no matter the number of days that were divided weekly by lead time. The other type of customers had a significantly lower rate of cancellation. The transient party customers had the second highest number of cancellations.

1. Cancellation within 30 days

An important insight was to get about the number of customers that cancelled just a month before their data of arrival. So that to keep the customers engaged the hotels can give them a reminder for their arrival or introduce an new scheme in order to keep the customer engaged.

Variables used – lead time

The cancellation in the last month did not follow a linear pattern due to uncertain reasons which could be emergencies or a change of mind of the customer. About 25% of the customers cancelled when the lead time was 0 and 8% of the customers cancelled their bookings when the arrival period was 28 days. The 2nd graph in our report showed a linear downward trend but it is highly uncertain to predict about the cancellations in the last month before the arrival.

**Correlation Matrix:**

* From the table to the right we can observe the correlation strength of each variable with the dependent variable (‘is\_canceled).
* Like no deposit has a correlation coefficient of 0.47 with is\_canceled so they are highly correlated.
* A high correlation will help the model to predict the correct value for the dependent variable (‘is\_canceled’).
* Thus we will keep all the variable which display a high correlation coefficient with the dependent variable.

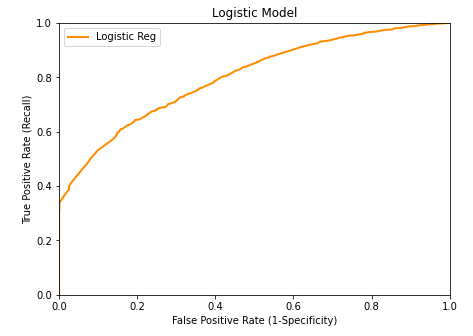
**Variance Inflation Factor:**

A variance inflation factor is basically a tool to help identify the degree of multicollinearity. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.



**MODEL INFERENCE**

* Using the logistic model we have reached to an accuracy of 78% which means we are sure that if the model says the person is going to cancel the booking it is very probable.
* We have a recall rate of 96% which represents we are able to predict the customers who has actually cancelled the booking.
* Precision is 74% that means we able to do a good prediction of actual confirm and cancelled bookings overall.
* We have an AUC of 80% which represents that our model will hold with any classification threshold.



Using this ROC curve we have determined the revenue of the hotel by including the cost involved for true positive and true negative values.

**DISCRIMINANT ANALYSIS**

Some research on what should our hotel tariffs be, lead us to a set of room rates (average across room types) for each month for our Hotel.

|  |  |
| --- | --- |
| January | $ 118.00 |
| February | $ 118.00 |
| March | $ 195.00 |
| April | $ 215.00 |
| May | $ 250.00 |
| June | $ 215.00 |
| July | $ 206.00 |
| August | $ 188.00 |
| September | $ 197.00 |
| October | $ 197.00 |
| November | $ 197.00 |
| December | $ 101.00 |

An average of these rates were used to our Return Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predict Positive | Predict Negative |
| Actual Positive | $ 183.08 | $ - |
| Actual Negative | $ (183.08) | $ - |

Applying it with the confusion matrix we got from our model:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification matrix (Confusion)** | |  |  |
| **(Actual alongside, predicted along top)** | | |  |
|  | Not Cancelled | Cancelled |  |
| Not Cancelled | 18166 | 687 | 18853 |
| Cancelled | 6240 | 4755 | 10995 |

Percent correct classification: 76.79%

|  |  |  |
| --- | --- | --- |
| **Probability of classification of an Individual** | | |
| **(Actual alongside, predicted along top)** | | |
|  | Yes | No |
| Yes | 0.609 | 0.023 |
| No | 0.209 | 0.159 |
|  | 0.818 | 0.182 |

Using the formula below to calculate expected return:

= **$ 2,183,412.08**

**CONCLUSION:**

* Finally, we are able to conclude the extent of our analysis and include the cost of unforeseen cancellations rather than just predicting the cancellation.
* We could see that the hotel is able to make a total profit of 2 Million in a time duration of 28 months.
* But we recommend that this revenue can be increased by using our model and taking the steps below to reduce the cancellations like:
  + In our analysis we found that mostly those people are cancelling the booking for whom lead time is very high which means he has made the booking very early and hence has forgotten about the booking.

**Solution –**

* We can send monthly reminders.
* We can send new deals and promotions etc.

**Note: Ultimately we have to make increase customer engagement with the hotel to reduce cancellations.**

**REFERENCES**

[1] Menard, S., 2010. *Logistic regression: From introductory to advanced concepts and applications*. Sage.

[2] Klecka, W.R., Iversen, G.R. and Klecka, W.R., 1980. *Discriminant analysis* (Vol. 19). Sage.

[3] Hartwig, F. and Dearing, B.E., 1979. *Exploratory data analysis* (No. 16). Sage.

[4] Wang, Q., Wang, J., Zhao, P., Kang, J., Yan, F. and Du, C., 2017. Correlation between the model accuracy and model-based SOC estimation. *Electrochimica Acta*, *228*, pp.146-159.