Predicting Hotel Booking Cancellations Using Machine – Learning Algorithm

Murali S

Date: 23-09-2022

Abstract

Booking cancellations have a substantial impact in demand management decisions in the hospitality industry. Cancellations limit the production of accurate forecasts, a critical tool in terms of revenue management performance. To circumvent the problems caused by booking cancellations, hotels implement rigid cancellation policies and overbooking strategies, which can also have a negative influence on revenue and reputation. Using data sets from hotels and addressing booking cancellation prediction as a classification problem in the scope of data science.

Main objective of this paper is possible to build model for predicting booking cancellations by using the Machine Learning algorithm like logistic regression. We deal this problem as classification problem. From this result, it allow hotel managers to accurately predict net demand and build better forecasts, improve cancellation policies.

1.0 Introduction

Revenue management is defined as "the application of information systems and pricing strategies to allocate the right capacity to the right customer at the right price at the right time". Since hotels have a fixed inventory and sell a perishable "product", as a way to make the right room available to the right guest, at the right time, hotels accept bookings in advance. Bookings represent a contract between a customer and the hotel (Talluri & Van Ryzin, 2004). This contract gives the customer the right to use the service in the future at a settled price, usually with an option to cancel the contract prior to the service provision. Although advanced bookings are considered the leading predictor of a hotel's forecast performance (Smith, Parsa, Bujisic, & van der Rest, 2015), this option to cancel the service puts the risk on the hotel, as the N. António, A. Almeida, L. Nunes, Tourism & Management Studies, 13(2), 2017, 25-39 26 hotel has to guarantee rooms to customers who honor their bookings but, at the same time, has to bear with the opportunity cost of vacant capacity when a customer cancels a booking or does not show up (Talluri & Van Ryzin, 2004).

1.1 Objective

This paper aims to demonstrate how data science can be applied in the context of hotel revenue management to predict bookings cancellations. Moreover, show that booking cancellations do not necessarily mean uncertainty in forecasting room occupation and forecasting revenue. This is achievable by:

1. Identifying which features in hotel PMS's databases contribute to predict a booking cancellation probability.

2. Building a model to classify bookings with high cancellation probability and using this information to forecast cancellations by date.

3. Understanding if one prediction model fits all hotels or if a specific model has to be built for each hotel.

1.2 Problem Statement:

1.2.1 What is the Task?

To detect the whether new booking will conform or going to cancel

1.2.2 Why does the problem need to solve?

Suppose if the booking is going to cancel, the hotel manager will allocate the room to other customer's. In weekend or festival times there will be huge demand on hotel. If the room is cancel by the customer there is no profit nor loss. To make Profit, the hotel cancellation detection is more useful.

1.2.3 How we solve this problem?

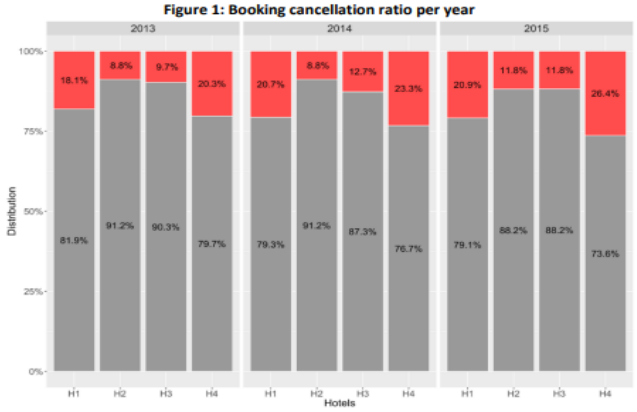
Using the machine learning algorithm in supervised learning. We use Logistic regression to solve this problem.

2.0 Business Need Assessment

As presented in Figure 1, in the four studied hotels, booking cancellations have been increasing since 2013, with thw exception of H3 when, from 2014 to 2015, the hotel imposed a more rigid cancellation policy. In 2015, booking cancellation rates ranged from 11.8% to 26.4%, which are in harmony with what was observed by Morales & Wang (2010).

Acting on bookings marked as having a high probably of being canceled can go from offering hotel services (e.g., spa treatments, free dinner or airport transfer) to discounts in certain services or entrances to local amusement parks. These actions could mitigate booking cancellations and therefore reduce the hotel's risk. These actions generate costs for the hotel, but by reducing the need to overbook, or at least, by enabling a better overbooking policy, the costs related to cancellations occurred respectively, 25, 54, 33, and 55 days after bookings were made. Surprisingly, it is not during months of high demand (the high-season months highlighted in Figure 2) that lead time and cancellation time are higher. In fact, these times are higher when there are special events in the region than during high season.

Figure 1: Booking cancellation ratio per year



In Fig 2: the booking cancellation patterns are different from each of these four hotels

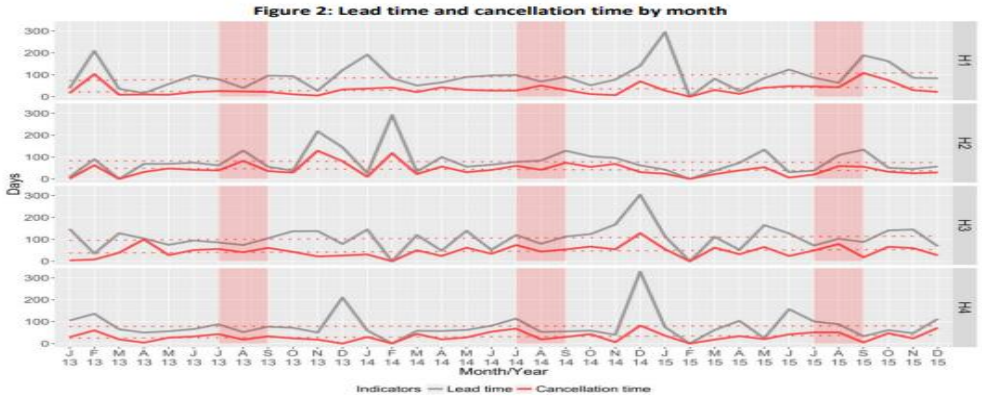
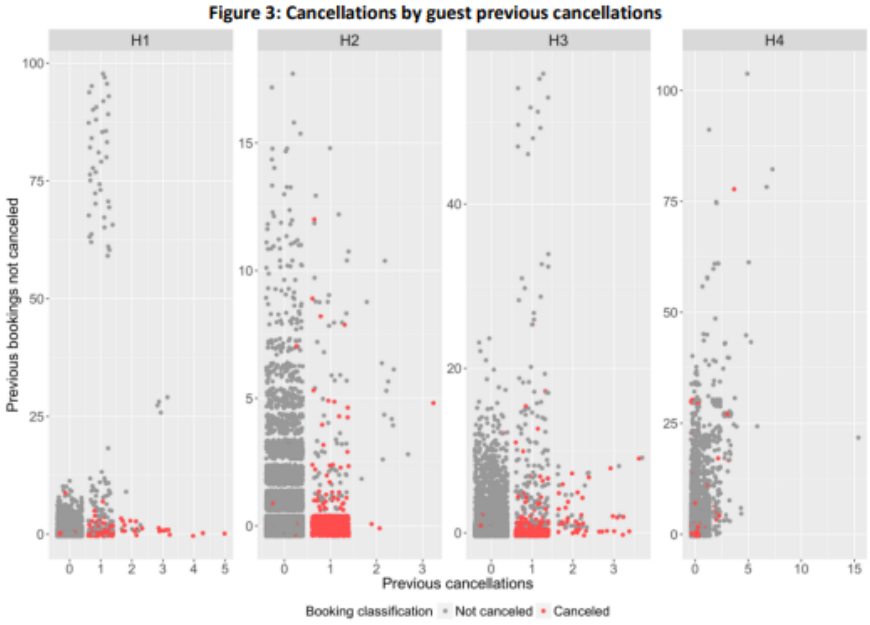
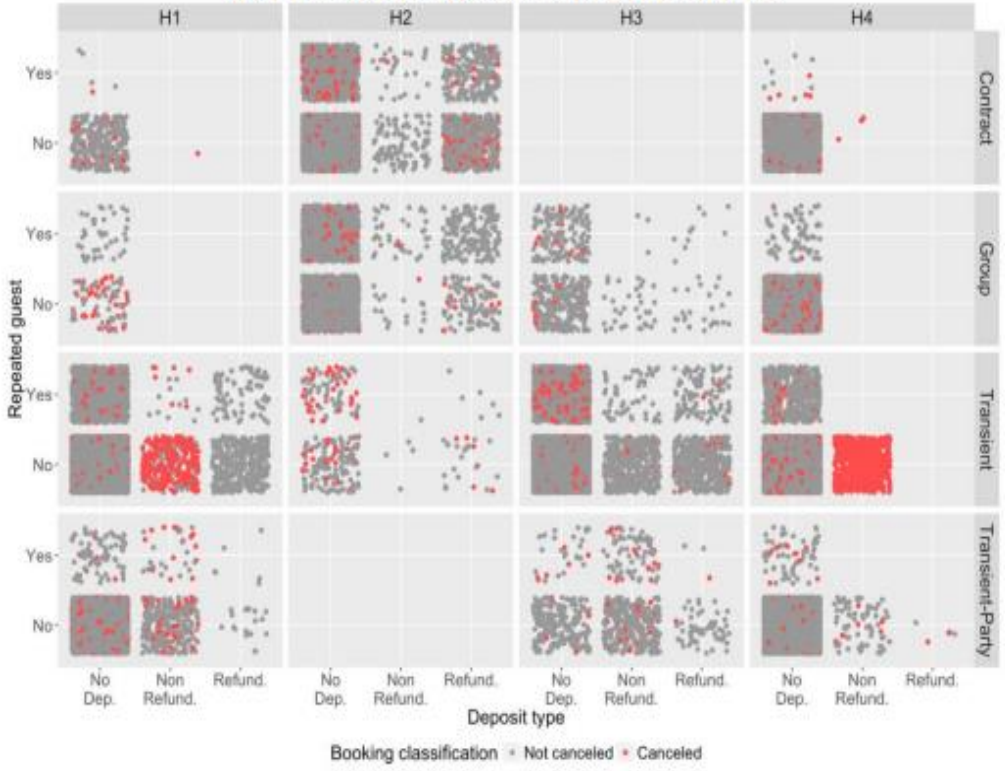


Figure 3 it is possible to see that most cancellations were made by guests who had already had one previous cancellation and also had less than 5 non canceled bookings. It was also possible to see that customers with a high number of previous bookings at the hotel, rarely canceled. Yet, the maximum number of cancellations per guest or previous bookings are different in each hotel.



In Figure 4 it is possible to verify that the type of customer (contract, group, transient or transient-party) and the deposit type made by them to guarantee the booking (no deposit, non-refundable/paid totally in advance or partially paid) as also different behaviors in terms of cancellations.

Figure 4: Cancellations by customer type, by deposit type



3.0 Target Specification and Characterization

Detail of the Target - Datasets:

1. hotel :(H1 = Resort Hotel or H2 = City Hotel).
2. is_canceled Value: showing if the booking had been cancelled (1) or not (0).
3. lead_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. arrival_date_year: Year of arrival date.
5. arrival_date_month: The months in which guests are coming.
6. arrival_date_week_number: Week number of year for arrival date.
7. arrival_date_day_of_month: Which day of the months guest is arriving.
8. stays_in_weekend_nights: Number of weekend stay at night (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. stays_in_week_nights: Number of weekdays stay at night (Monday to Friday) in the hotel.
10. adults: Number of adults.
11. children: Number of children.
12. babies: Number of babies.
13. meal: Type of meal booked.
14. country: Country of origin.
15. market_segment: Through which channel hotels were booked.
16. distribution_channel: Booking distribution channel.
17. is_repeated_guest: The values indicating if the booking name was from a repeated guest (1) or not (0).

18. previous_cancellations: Show if the repeated guest has cancelled the booking before.
19. previous_bookings_not_canceled: Show if the repeated guest has not cancelled the booking before.
20. reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
21. assigned_room_type: Code for the type of room assigned to the booking. Code is presented instead of designation for anonymity reasons.
22. booking_changes: How many times did booking changes happen.
23. deposit_type: Indication on if the customer deposited something to confirm the booking.
24. agent: If the booking happens through agents or not.
25. company: If the booking happens through companies, the company ID that made the booking or responsible for paying the booking.
26. days_in_waiting_list: Number of days the booking was on the waiting list before the confirmation to the customer.
27. customer_type: Booking type like Transient – Transient-Party – Contract – Group.
28. adr: Average Daily Rates that described via way of means of dividing the sum of all accommodations transactions using entire numbers of staying nights.
29. required_car_parking_spaces: How many parking areas are necessary for the customers.
30. total_of_special_requests: Total unique requests from consumers.
31. reservation_status: The last status of reservation, assuming one of three categories: Canceled – booking was cancelled by the customer; Check-Out
32. reservation_status_date: The last status date.

4.0 External Search

The sources I have used as reference for analysing the need of hotel cancellation and to detect the Cancellation using several resources such as mentioned below:

- https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue
- https://www.redalyc.org/pdf/3887/388751309003.pdf
- https://www.analyticsvidhya.com/blog/2022/03/end-to-end-hotel-booking-cancellation-machine-learning-model/
- https://www.ijraset.com/research-paper/hotel-booking-prediction-using-ml
- https://www.ezeeabsolute.com/blog/hotel-benchmarking/
- https://www.revfine.com/hotel-benchmarking/

4.1 Benchmarking

4.1.1 Hotel Benchmarking

Hotel benchmarking is the process of comparing your property's performance against the competition, adding a layer of context to what success and failure looks like in your

circumstances and environment. The key to unlocking this level of understanding? Historical performance data, both yours and that of the competition.

4.1.2 Why is Benchmarking Good for Hotels?

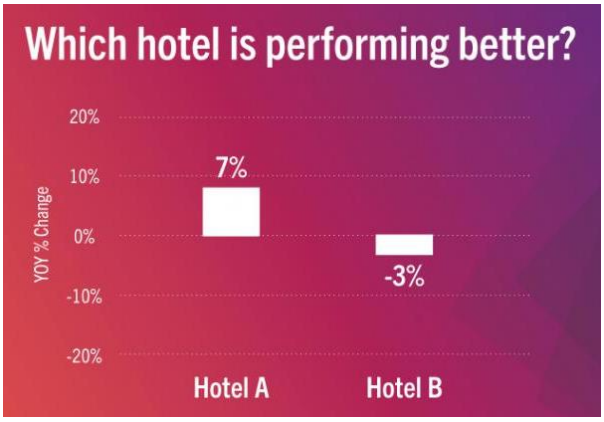
- In the hotel business segment, you need to understand the market and the factors that are involved. It could be the risk factor or even the benefits you are likely to recap.
- This is, however, not possible when you don't follow a certain process and that includes assessment and comparison.
- Benchmarking is a fantastic way to learn about the space you're venturing into or already operating in.
- But from a business perspective, you have to look for the bigger picture. Small profits may seem nice initially, but to compete in an industry that keeps on evolving, you need to learn more in-depth.
- The benefits of benchmarking in the hospitality industry are immense, and big hotel brands are great examples of it. They keep on evaluating their property and whenever it is needed, they make significant changes in the way they operate.

4.1.3 Tips to Get better at Hotel Benchmarking

- Understand the market standard
- Know who you're competing against
- Learn about the relevant metrics to consider
- Narrow down on the data as much as possible
- Compare your findings with past data
- Outline a roadmap for future operations
- Measure your performance

4.1.4 Applying Benchmark to Hotel

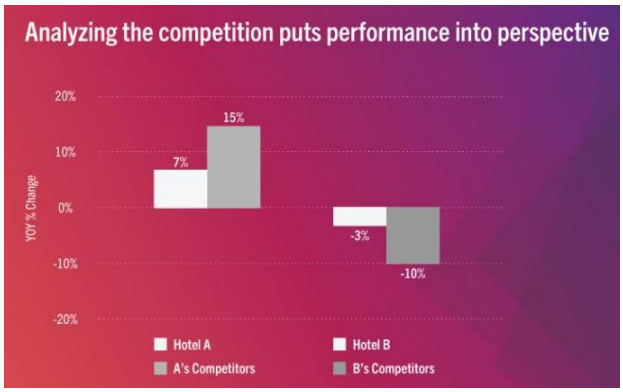
Hotel A VS Hotel B – an example of Benchmarking



At first glance, Hotel A is outperforming hotel B. But not everything is always as it seems.

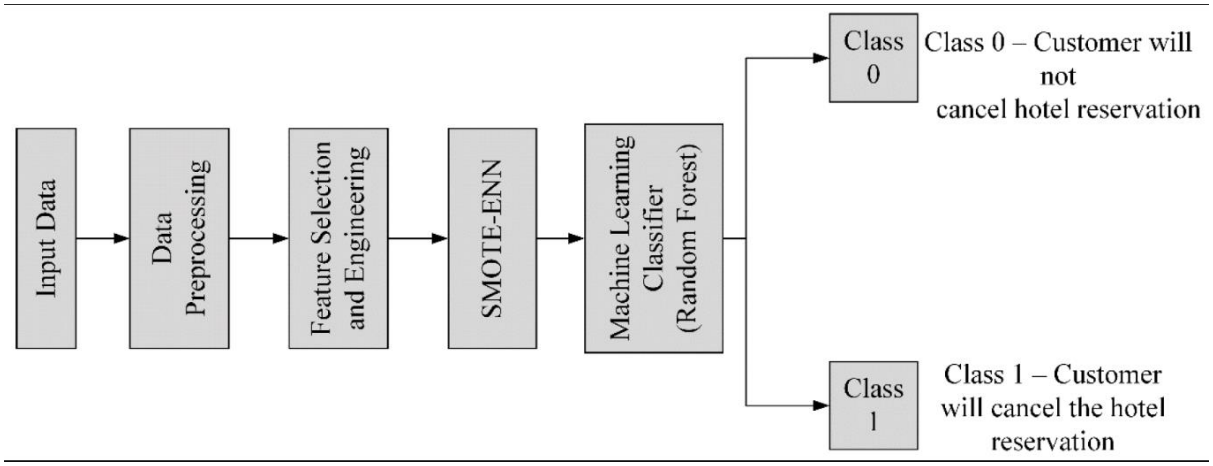
If hotel A and B operate in the same market and have similar pricing, property type and amenities, then hotel A can count that month's performance as win—reporting a 7% increase compared to its competitors 3% decline.

But what if they're operating in two different markets?



Hotel A might have reported 7% growth, but when the competition increased at an average of 15%, benchmarking highlights that this property wasn't able to captialize on the market's growth as effectively as the competition. Back to hotel B, in the context of the competition's 10% decrease, a 3% decline suggests that the property is actually ahead from a performance perspective.Without benchmarking and the context this offers, hotel A and B would most likely be blissfully unaware and unnecessarily concerned, respectively.

4.2 Applicable Patents



- First we collect the hotel data. There are various source to collect the data such Kaggle, govt.in, Google datasets, Github, etc..
- Next in our data there are outliers, missing values, etc. To handle this Data preprocessing steps is required.
- In our Target dataset there are 32 features in that we select the important features, to reduce the dimensionality. It will save time and cost of the model also space.

- After this our model is not smooth due to various factors. For example in our data there are a lot of factors such Festival time booking is High, Summer vacation time booking is High, Rainy season the booking is Quite low. Booking is based on Seasonal change. Here time is one of the factor. To reduce this we smooth our data by using various techniques such Transform the data like Log, Square, etc..
- The our data is perfectly ready we use our machine learning model to detect the Hotel Cancellation

4.3 Applicable Standards

Cancellation Rules - Hotels

- Within 24 Hours from the time of Reservation - No Refund.
- Between 24 Hours to 48 Hours - 70% Refund.
- Before 48 Hours - 90% Refund.
- No Refund for cancellation of Dormitory Booking.

CANCELLATION RULES FOR HILL STATIONS **(1st April to 15th June season period only)**

- Within 24 Hours from the time of Reservation - No Refund.
- Between 24 Hours to 48 Hours - 50% Refund.
- Before 48 Hours - 75% Refund.
- No Refund for cancellation of Dormitory Booking.
- The Same Rule will be applied to Hotel Kanniyakuamri and Courtrallam during season period.

GENERAL RULES HOTELS

- Only one preponement / postponement is permitted before 48 hours subject to availability of accommodation.
- After availing one postponement or preponement no cancellation is permitted.
- Luxury Taxes and Service Tax will be charged as per rules on lodging bills at Hotels.
- The rates are subject to change.
- Online customers can cancel/postponment their Accommodation in TTDC online website using their login ID.
- For online customers, refund for cancellation will be credited to thier card account by online.
- Catering facilities available in all TTDC Group of Hotels.

4.4 Applicable Constraints

- Requires a lot of research to obtain dataset of hotel in order to provide more sophisticated and accurate results.
- Establishing a network for receiving a booking
- Android developer / Mobile application developer
- Data Analysts
- Data Researcher
- Machine Learning Engineer
- Confidential data to be obtained to train the model

4.5 Business Opportunity

Hotel Manager will face these kind of business problems. From this there will be a wide Opportunity for Business Analysts. To overcome this hazardous circumstance, our main objective is to use machine learning, which not only gives faster results but also demonstrates higher accuracy in the Hotel Booking Prediction process. For better performance we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole datasets.

5.0 Concept Generation

This studies if it is possible to predict hotel cancellations with the use of machine learning and which variables that have the greatest significance in the models. To gain a more nuance understanding about the subject the following topics for future research are presented. The time aspect is discarded in this studies but would be interesting to investigate. As mentioned earlier, it is cancellations that take place close to the day of arrival that create problems for hotels. A future study could analyse how well machine learning algorithms can predict cancellations that occur within e.g., 14 days in the future. It indicates that external weather data in form of precipitation and temperature improves the results of the models.

However, as the use historical weather data in realized values, a future study could be based on weather forecasts, since the decision to cancel a reservation can be assumed to be based on the current forecast. Another interesting topic of future research could be to estimate the economic effect on the hotel after implementing a machine-learning prediction system. This future research will probably be in the field of business but can highlight the monetary value of a statistical solution.

Further research could also make use of features from additional data sources, such as weather information, competitive intelligence (prices and social reputation), or currency exchange rates, to improve model performance and measure the influence of these features in booking cancellations.

6.0 Concept Development

Following are the steps for the implementation.

- Exploratory Data Analysis
- Data Pre-Processing
- Model Building
- Model Comparison

Phases Involved in Exploratory Data Analysis

Data Collection:

Collecting data is a starting point of exploratory data analysis. Data collection is the process of finding data from different public sites, or one can buy from private organizations and load data into our system. Kaggle and Github are familiar and known sites that provide free data.

Import Libraries:

After collecting data we have to import the necessary libraries to build machine learning models. Numpy, Pandas, Matploilib, Seaborn are the known libraries used in the machine learning model.

Numpy: NumPy is a Numerical Python Library that helps perform mathematical operations.

Pandas: Panda is an open-source library that helps understand relational or labelled data.

Matplotlib: Matplotlip is a Python visualization library that helps visualize 2D array plots.

Seaborn: Seaborn is a data visualization library built on top of matplotlib.

Data Cleaning:

- Remove missing values, outliers, and unnecessary rows/ columns.
- Check and impute null values.
- Check Imbalanced data.
- Re-indexing and reformatting our data.

Handling Outliers

There are two types of outliers:

- Univariate outliers: Outliers are the data points that are away from the expected range of values. In Univariate analysis single variable is considered to detect outliers.
- Multivariate outliers: These outliers are dependent on the correlation between two variables

Correlation Matrix

A correlation matrix glimpses the correlation between different variables. The correlation between two variables is determined by the correlation coefficient.

Numeric – Categorical Analysis

In numeric-categorical analysis, one variable is a numeric type and the other is a categorical variable. We can use group by function or box plot to perform numeric-categorical analysis.

Categorical-categorical Analysis

The chi-square test determines the association between categorical variables. The Chi-square test calculates based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table.

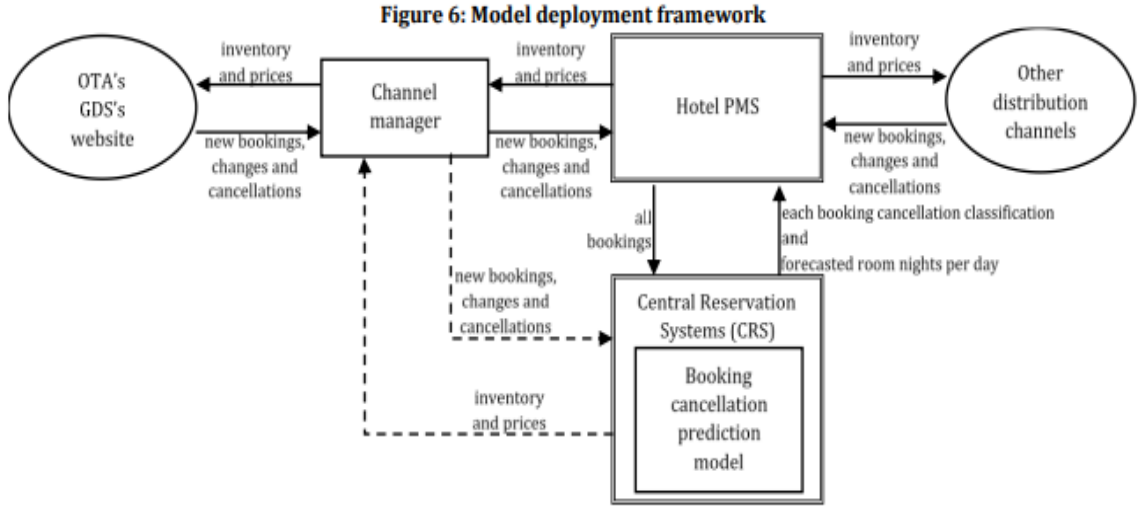
Scatter Plot

A scatter plot represents every data point in the graph format. It shows the relationship of two data means how the values from one column fluctuate according to the corresponding values in another column.

Pair Plot

Pair plots compare multiple variables at the same time. Pair plots save spaces and compare various variables at the same time.

7.0 Final Product Prototype with Schematic Diagram



7.1 Deployment

Although the deployment of these models in a production environment was not in the scope of this research, the way they are deployed is critical to their success. For that reason, the elaboration of a framework (see Figure 6) to define how models are deployed is also an important duty of this research.

As represented in Figure 6, the booking cancellation prediction model should not be implemented by itself. In truth, if deployed independently of the hotel other systems, it is unlikely that it would present any valid results in terms of revenue management. Today's speed and complexity imposed on a hotels reservations department is such that advantages of using the model could not be clear if tasks related to the model inputs and outputs had to be done manually.

Some predictor variables vary with time (e.g., "LeadTime") or can assume new values every day, as in the case of changes/amendments to bookings (e.g., "BookingChanges" or "Adults"). Thus, the model should be run every day so that all in-house bookings and results are evaluated on a daily basis.

8.0 Product details

8.1 Data Sources:

- Github
- Kaggle
- Google datasets
- Govt.in

8.2 Software, Frameworks

- Python
- Pycharm

- Jupyter Notebook
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Scikit – Learn

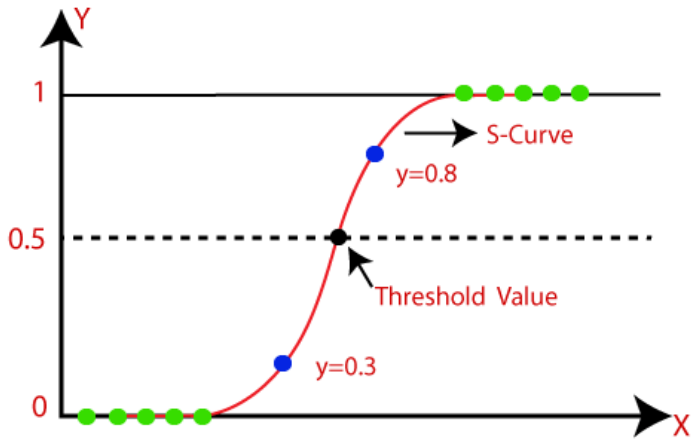
8.3 Algorithms

Logistic Regression

Logistic regression comes under the most popular Supervised Machine Learning algorithms. Logistic regression predicts the categorical dependent variable using a given set of independent variables. The categorical dependent variable should be either Yes or No, 0 or 1, true or false, etc. Logistic regression is much comparable to Linear Regression only implementation is different. Linear Regression solves the Regression problems, and Logistic regression is used for solving the classification problems. Instead of fitting the regression line, In logistic regression, we fit the sigmoid S function, which predicts two maximum values(0,1). The curve from the logistic function indicates the likelihood of something such as whether rain comes or not based on weather conditions.

Logistic Function (Sigmoid Function)

To map the predicted values to probabilities sigmoid function is used. It maps the values between the range of 0 and 1. The threshold value is used in the logistic regression to compute the S Shape. The threshold value defines the probability of either 0 or 1. The value above the threshold tends to be 1, and the values below the threshold tend to 0.



Assumptions for Logistic Regression

- The data must follow a normal distribution.
- The dependent variable must be categorical.
- The independent variable should not have multi-collinearity

Type of Logistic Regression

The Logistic regression is classified into three types based on the categories.

Binomial: If dependent variables hold two categories such as 0 or 1, male or female, pass or fail, it comes into binomial logistic regression.

Multinomial: If a dependent variable holds three or more possible unordered categorical variables, such as cat, dog, lion, it comes into multinomial logistic regression.

Ordinal: If the dependent variable holds three or more possible ordered categorical variables such as low, medium, high, it comes into ordinal logistic regression.