Predicting Hotel Demand using Machine Learning

Omkar Ratnoji Tawade
Data Analytics
National College of Ireland
line 4: Dublin, Ireland
x1923213@student.ncirl.ie

***Abstract*—This is a project design document on predicting demands of hotel using machine learning techniques. Hotel industry is now slowly reviving after the ongoing pandemic. The need to know the demand of hotels becomes vital to manage the business. The expectation of sanitization and other facilities has gone up. Machine learning algorithm XGBoost will help to know the various reason which affects the demand of the hotels and depending on those factors we will produce a model which will predict the demand of the hotel.**

Keywords—Machine learning, Random Forest, XGBoost

I. INTRODUCTION OF PROJECT DATASET

The dataset is downloaded from the Kaggle. The data was extracted from the hotel's property management system SQL database. Nuno Antonio, Ana de Almeida, Luis Nunes have explained the process of extracting the data from such a system [1]. The dataset contains the data of two types of hotels which are resort hotels and other is city hotels. The dataset contains the booking information from 1st July 2015 to 31st August 2017. The dataset contains 32 variables and 119390 records in total.

hotel
is_canceled
lead_time
arrival_date_year
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
meal
country
market_segment
distribution_channel
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
reserved_room_type
assigned_room_type
booking_changes
deposit_type
agent
company
days_in_waiting_list
customer_type
adr
required_car_parking_spaces
total_of_special_requests
reservation_status
reservation_status_date

Figure 1. List of variables in the hotel demand dataset.

Figure 1 listed all the variables in the dataset. Now we will understand the meaning of each variable.

i. Hotel: It helps to differentiate the records based on resort hotel and city hotel. It will help us in our initial analysis to analyze which hotel is performing better comparatively. It is a categorical variable

ii. is_canceled: It is very easy to understand, when this variable contains the value 1 then it indicates that booking is canceled and when it contains value 0 it suggests that booking is not canceled. This variable will be our dependent variable. It is also a categorical variable

iii. lead_time: The number of days it took to enter the record into PMS from the arrival date. It is an integer variable.

iv. arrival_date_year: It suggests the year of the arrival date. It will help to analyze which hotel is performing well over the years.

v. arrival_date_month: It suggests the month of the arrival date. It will help to analyze which months hotel demands are high and which months the demands are low.

vi. arrival_date_week_number: It suggests the week number of the arrival date. It is an integer variable. We will observe the visualization of this variable and then can analyze if there any seasonality present.

vii. arrival_date_day_of_month: It suggest the arrival day. It will help to analyze whether hotel demands increase on month-end or not.

viii. stays_in weekend_nights: This variable will be very useful to predict the hotel demands. In a normal scenario, hotel demand increases on the weekends, but we will see if this hypothesis is right after observing the visualizations.

ix. stays_in_week_nights: This variable suggests the hotel occupied for the number of nights between Monday to Friday. This will be also an important variable to determine the hotel demands.

x. adults: Number of adults.

xi. children: Number of children. It will help to analyze which hotel is preferred for family.

xii. babies: Number of babies

xiii. meal: It suggests the type of meal booked. There are four categories in this variable which are BB – Bed and Breakfast, HB – Half board (breakfast and other meal usually dinner), FB – Full Board (breakfast lunch and dinner), Undefined/SC – No meal package.

xiv. country: Country of customer who booked the hotel. It will help to know whether hotel demand is influenced by tourists or not.

xv. market_segment: It indicates for which purpose the hotel was booked. There are five categories in this variable which are Online Tourist Agent, Offline Tourist Agent or Tours Operators, Groups, Direct, and Corporate. It will help to know which hotel is used for which particular market segment.

xvi. distribution_channel: It is similar to market_segment and also stores the five categories similar to that of market_segment variable

xvii. is_repeated_guest: It will be one of the important variables in order to predict the demand of the hotels. If guests are repeated, then it suggests that they like the hotel and its facilities and there is a chance that these guests will recommend the hotel to others, and thereby the demand for the hotel will increase. It is a categorical variable.

xviii. previous_cancellations: It indicates the number of cancellations made by an individual before the current booking. If the number of cancellations is significant then the hotel should manage it or increases its resources well.

xix. previous_bookings_not_canceled: It indicates the number of bookings made before the current booking. This will indicate whether people opt for a hotel more than one time.

xx. reserved_room_type: This is a categorical variable. It has five categories which are basically five alphabets. To maintain anonymity, they didn't record the room_type in the original format. It will help to know which rooms are in great demand.

xxi. assigned_room_types: This is also a categorical variable. It has the same codes which are in reserved_room_type. It would help to analyze if hotel managed their resources well and able to assign rooms which are actually reserved by their customers.

xxii. booking_changes: It is an integer variable. It records the number of changes done in the booking. It will help to analyze if hotels are flexible to allow the changes.

xxiii. deposit_type: It is a categorical variable and consists of three categories which are No Deposit, Non-Refund, and Refund.

xxiv. agent: It contains the id of the travel agency which made the booking. It will help to analyze how many bookings are made with the help of a travel agency.

xxv. company: It contains the id of the companies which helped the customer with their bookings. The hotel's demand can also know by observing the number of agent and company bookings.

xxvi. days_in_waiting_list: It indicates the number of days the booking was on the waiting list before it was confirmed. This is also again one of the most important variable in order to know the hotel demands.

xxvii. customer_type: This is again a categorical variable. Contract, Group, Transient, and Transient-Party are the four categories. When the booking has a contract associated with it then it comes under the contract category. When booking not belongs to part of the group or does not have any contract then it comes under the transient category. When a booking is transient, but it is related to at least other transient bookings then it comes under the transient-party booking. It will help the hotel to utilize its space efficiently by analyzing the statistics of this variable.

xxviii. adr: ADR stands for average daily rates. It is calculated as the sum of all lodging transactions divided by the number of nights. It basically gives an idea about the number of daily bookings and it will be also one of the important independent variable.

xxix. required_car_parking_spaces: It indicates the number of parking spaces required by the customers. By analyzing this variable hotels can manage their parking resources efficiently.

xxx. total_of_special_request: It indicates the number of requests made by the customers.

xxxi. reservation_status: It is a categorical variable. Check Out, Canceled, and No Show are the three categories present in this variable. This will help to analyze the hotel demands in the past.

xxxii. reservation_status_date: It indicates the date at which the last status was set. It is the date on which the customer canceled the booking or did the checkout from the hotel.

II. GOALS OF THE PROJECT

The goal of the project is to predict the demand of hotels based on the previous cancelation data and other independent variables. The hotel industry collects a good amount of data right from the booking, check-in, maintenance of room to the checkout. A lot of knowledge or information can be acquired from this data. This analysis can help the hotel to manage its resources well to meet the demands of its customers. Also, this analysis will help the hotel to know the possibility of cancelation of the reservation. These last-minute cancelation leads to loss for the hotel and also impact the managing resources well. Apart from predicting the cancelation rate, we will analyze few hypotheses. There are two types of hotels in the dataset which are city hotels and resort hotels. Our first hypothesis is that a city hotel is costlier than a resort hotel because a city hotel may have more facilities than a resort hotel. City hotels will be used more for commercial purposes so there is a high chance that they may have a price higher than a resort hotel. Our second hypothesis is that booking a hotel with a travel agent will cost less than booking a hotel directly. Booking agents always have different offers which attract customers to book a hotel via a travel agent. These offers are mainly discount on the price whereas hotels don't provide such discounts as their prices are always fixed. These hypotheses will help the hotel to know what the preferences of their customers were. The hotel can alter its price structure depending upon the output of the analysis of these hypotheses. If the travel agent is providing more bookings, then the hotel can introduce the loyalty program for the customer who is booking directly.

III. ETHICAL CONCERNS

Hotel industries collect a large chunk of data, so there may be a problem like an ethical concern while capturing the data. These may be due to transaction errors and unreported bookings. These bookings are made under the table and usually, these transactions are made through cash. There are also other ethical issues in hotel industries that are very common like stealing of food, beverages, shampoo, towels, and toilet papers. Hotels ask for the identity card from their customer during the booking. So, while using the booking data for analysis purposes hotels must remove all personal data like age, name, phone number, and email id, transaction details. Nowadays, there is a huge risk of cyber-attacks so the company should take care of such data. Hotels must insist their customers pay through card since card readers are a safe option for transactions. They should avoid payment through the phone because there is a high risk of cyber-attacks. When customers book hotels from the travel agent then there is a risk of compromising the data. These travel agents offer a discount to the customers and in return, they access their search data like the content they watched, products they browsed.

Figure 2. Ethical issue in Hotel Industry

The hotel should implement different practices in order to solve these ethical issues. Transaction data must be safeguarded like a hotel can have a firewall installed to avoid cyber-attacks. Hotels should use point-to-point encryption technologies to avoid malware attacks on secure networks. Human errors are mostly responsible for the data breach. Mangers of the hotel should be given appropriate knowledge of data protection. Appropriate antivirus must be installed in the hotel employee laptops. Data is usually breached from the employee's laptop. If they open a suspicious email, then it will lead to a breach of data. Phishing attacks are well known in the field of data breaching. Workshops must be carried out to bring awareness about such attacks among the employees. Ethical concerns like stealing food, beverages, towel, and toilet paper from an employee can be minimized by installing cameras everywhere possible. The hotel can do background research of their employee before hiring them. Hotels can have both paper and computer records, and this can be audited regularly to avoid the booking which is made under the table. The hotel can use cloud services to store its data. Cloud services are reliable and efficient for storing data. There are various functionalities in the cloud services which can make the hotel management life easy. Data will be more secured on the cloud because of encryption services used in the cloud and also dynamic password management will help for data protection. The hotel should have good relations with their employees as it will make an employee think for the company and they will do their work more honestly.

IV. THE BUSINESS VALUE OF THE PROJECT

The expectation of customers has risen over the last few years in today's world in terms of offers and service hotels offer them. Also, the process of hotel booking has now become fast such that customers can book a hotel with one click. There is a large amount of data collected by hotel industries so there is a need for analysis to grow the business. The hotel should need to look into this data quickly and then deliver understanding in terms of offer and services. In this project, we are predicting whether booking will be canceled or not this will help the hotel to optimize the price according to the demand. Predicting the correct price is important for the business. The hotel industry is growing fast so there is a large competition. Hotels should have the idea of correct pricing to grow their business. Pricing depends on various factors like the room size, special requirements, holiday season, location. If this price is not predicted clearly then it can impact the business of the hotel. To grow a business, hotels should look or give offers to their repeated guest. In our dataset we have the data of repeated guests, mostly these repeated guests will not cancel their bookings and hotels can arrange additional services for these customers. We also have the data of customers who had previous cancelations. The hotels can also target these customers and can give special offers or services so that they can book their hotel regularly. We also have the data of booking month and year. We can use this data to analyze in which month bookings are canceled most. The hotel can plan to increase discounts or offer in such months. Hotels must be very flexible in order to provide the services. We will see if the cancelation of bookings depends upon the number of special requirements served by the hotel. Even the changes in booking should implement easily and quickly. We will see if the cancelation rate is impacted by an increase or decrease in the number of booking changes executed by the hotel. In our dataset there are two types of hotels which are city hotels and resort hotels, we will analyze and then try to predict which hotel is favored by the children. This will help the hotel business to increase the services for children's activities. International travelers who fly regularly usually book the same hotel if they like a particular hotel so the hotel should analyze the sale of these international customers. Hotels can provide the flexibility of booking for international travelers. We can analyze if the number of waiting days influences the canceling rate or not. Hotels can confirm the booking of international travelers as soon as possible to avoid the cancelation of booking. It is very important for the hotel to predict the cancelation rate in order to utilize these waiting bookings to increase sales. This will help hotel business to get more recommendation from foreign also. The hotel must under their customer genre. In our dataset there are two types of hotel, city hotel will mainly use by the corporates to arrange their meetings and resort hotel is mainly used by the family or travelers because it is an ideal place to relax. Ideally, resort hotels are mainly occupied on the weekend so to increase the number of bookings or increase the sales resort hotel can arrange the services such that this type of hotel can be used by the corporate people to arrange the meeting. Food is also an important factor in terms of hotel demand. The hotel can analyze the relation between meals and the cancelation rate. If the relation is negative, then

the hotel can increase the quality of food and can correct the pricing of the food.

V. Preliminary Visualizations

We will now analyze our dataset using visuals and will try to obtain information from those visualizations.

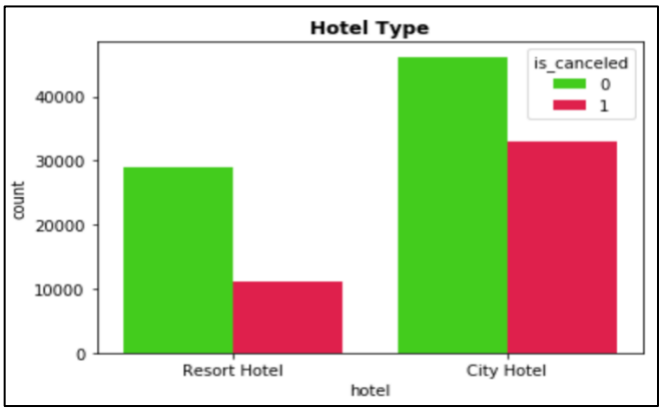

Figure 3. Frequency of Hotels.

Figure 3 suggests that the cancelation rate of the city hotels is more than the resort hotel. Since resort hotels are booked for the group so the cancelation rate is low in the resort hotel. Also, the price of the city hotels might be very low compared to the resort hotel, so the cancelation rate is high for the city hotel.

Figure 4. Frequency of dependent variable.

Figure 4 shows the frequency distribution of our dependent variable. Almost 34% of people who had booked the hotel have canceled their bookings. This figure is high and so there is a need to build the model to predict the cancelation rate of the booking in order to avoid loss and increase the business.

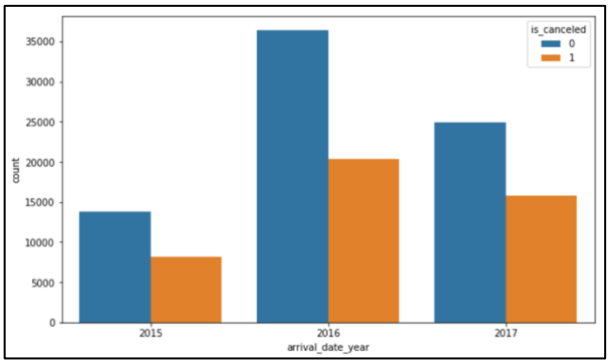

Figure 5. Plot between is_canceled and arrival_date_year.

Figure 5 suggests that the cancelation rate of the hotels is almost the same over the years. Almost 30-35% of bookings are canceled every year.

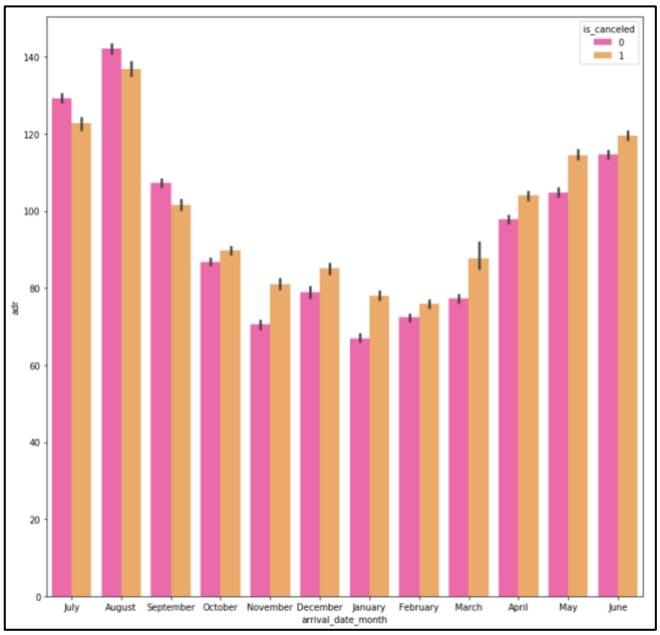

Figure 6. Plot between is_canceled and arrival_date_month

Figure 6 suggests that the number of bookings is more from May to August. These numbers are high due to the holiday season. Also, the cancelation rate is quite low during these months. In other months the cancelations are more comparatively. The hotel can reduce the price in such months to increase the sales in these months. Our model will help the hotel to predict the cancelation rate per month. This will help the hotel to manage its resources efficiently.

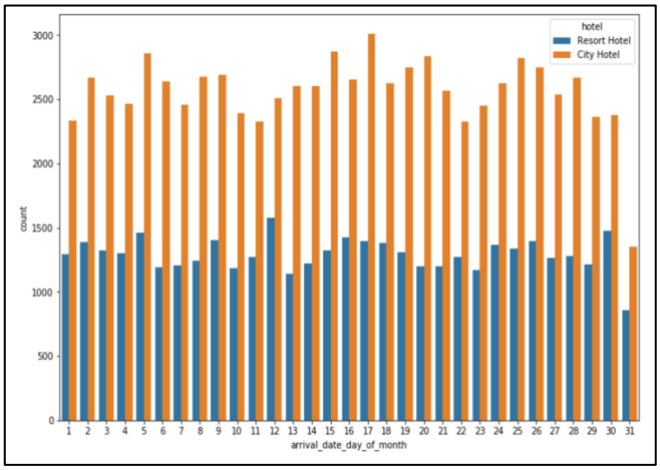

Figure 7. Plot between is_canceled and arrival_date_day_of_month

It is obvious that bookings of city hotel will be more than the resort hotel. The count of bookings of the hotel is more for few intervals, this may be due to the weekends. This shows the seasonality in the bookings. We will see whether it impacts the prediction of the cancelation rate while building the model.

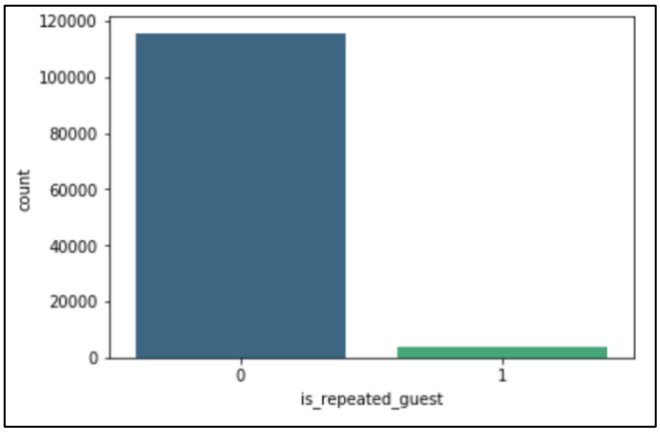

Figure 8. Frequency plot of the repeated guest.

Figure 8 suggests that the number of repeated guests is very low. This figure is quite low, and the hotel must strategize to increase the number of repeated guests because these guests have experienced the facilities of the hotel before, and these guests can give a recommendation to their friends or families for the hotel.

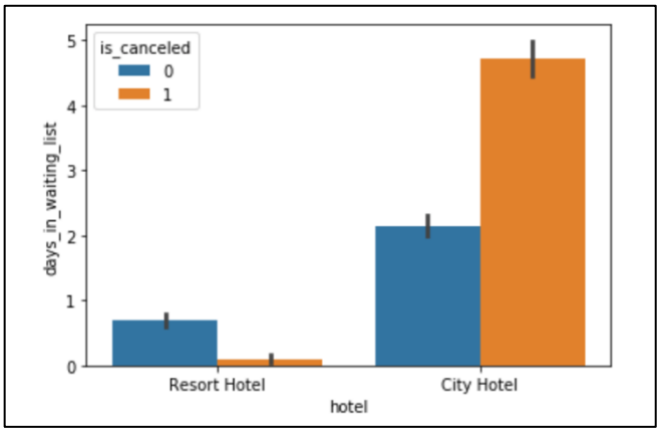

Figure 9. Cancelation rate of the hotels vs Days in waiting list.

Figure 9 is the plot of the hotel's cancelation rate vs day in the waiting list. This can be one of the most important independent variables while building the model. By observing the plot, we get the information that if city hotel booking is on the waiting list for a greater number of days then it is highly possible that booking is to be canceled. If a resort hotel booking is under the waiting list, then it is highly unlikely to be canceled.

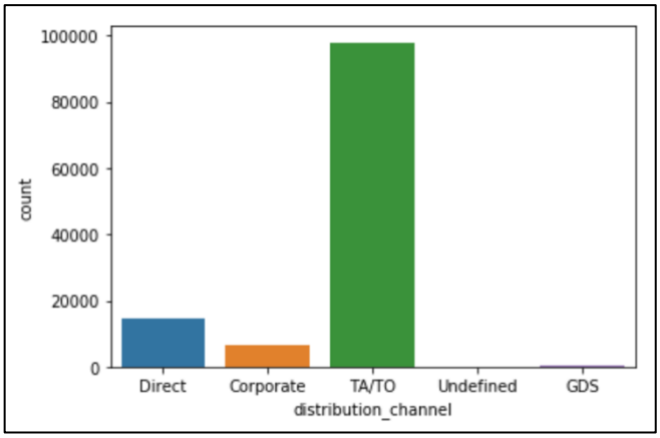

Figure 10. Frequency ploy of distribution_channel

Figure 10 suggests that most people book their hotels through a travel agent (online or offline). The hotel can advertise their hotels both on online platforms and offline to increase the business more.

VI. Applicable Techniques

XGBoost is one of the boosting techniques in machine learning. We will use this algorithm to predict if the booking is going to be canceled or not. Boosting algorithms use the sequential method in which models are produced sequentially. It will generate multiple weak learners and combine their predictions to form one strong rule. XGBoost is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Ensemble learning is the method in which the output of several weak learner models is combined to produce a single model which is efficient and accurate than those weak learner models.

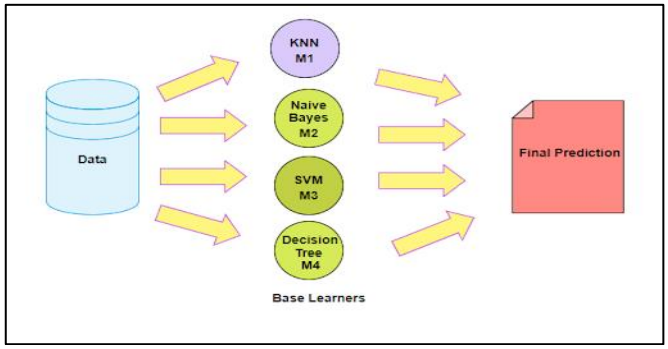

Figure 11. Ensemble Learning

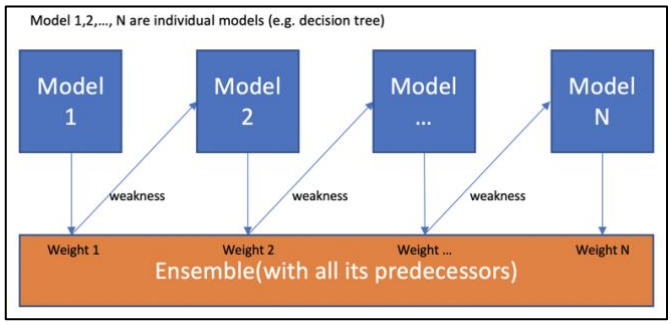

Figure 12. Boosting Algorithm

XGBoost (Extreme Gradient Boost) is an extended version of the gradient boosting algorithm. In gradient boosting models are generated such that the present model is better than the previous model. The overall model improves sequentially at each iteration. The difference in this type of boosting is that weights for misclassified outcomes are not incremented instead of adding weights, gradient boosting try to optimize the loss function of the previous model by adding a new adaptive model that adds the previous models. It basically reduces the error in the prediction of previously generated models. It will be used for classification prediction problems in our project. XGBoost algorithm is known for its speed and accuracy. The speed of this algorithm is fast due to the concept of parallelization. It will make sure that it will use maximum power for computation from your distributed system. XGBoost also uses a cache optimization concept in which xgboost stores it's all intermediate statistics in the cache memory. Since the calculations are stored in the cache

memory, xgboost can easily use it again quickly. In XGBoost there is a regularization parameter, it will help to avoid the model from overfitting. This parameter was not there in the normal gradient boosting algorithm. This regularization parameter helps to increase the performance of the algorithm. Another parameter that helps to increase the performance of the algorithm is the auto pruning of the tree. This will not allow the growth of a tree beyond a certain threshold. This is done in order to main the constant variance in the model.

REFERENCES

[1] Nuno Antonio, Ana de Almeida, Luis Nunes, "Hotel booking demand datasets," vol.22, pp. 41–49, February 2019.