**Machine Learning: Hotel Reservation Cancellation Prediction**

Exploring Predictive Models for Hotel Booking Cancellation Anticipation

Sai Mohan Anjesh

Aditya Dasari

# Abstract:

In today's world, hotel bookings have become an integral part of travel planning. This project utilizes advanced machine learning techniques to predict the likelihood of a hotel reservation being honored. By analyzing a vast dataset that includes diverse booking attributes and customer history, we have developed predictive models that can anticipate booking outcomes. These models have the ability to identify if customers are likely to cancel their reservations, allowing hotel managers to make informed decisions. Through training models such as Regularized Logistic Regression, Support Vector Machine, and Bagging and Boosting Models on over 36,000 cases, our team has generated two possible outcomes: a customer either cancels their reservation or honors it. Our models have proven valuable in providing the industry with insights to improve hotel bookings and customer satisfaction.

# Motivation:

A significant number of hotel reservations are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. The news article on USA today was the genesis of this project. We read that there had been bogus reservations and cancellations in the recent times in the lodgin industry and thought it would be better to solve this problem for the companies hence helping hotel mangaemnt to properly mangaem their revenue and sales. [Travelers’ rights: When reservations aren’t honored (usatoday.com)](https://www.usatoday.com/story/travel/columnist/mcgee/2017/11/08/when-reservations-arent-honored/841685001/)

# Problem Definition and Goals:

The primary objective is to address the pressing challenge of managing reservation cancellations and no-show incidents within the hospitality industry. Through accurate prediction, this approach aims to enhance revenue management strategies and operational efficiency, empowering hotels to optimize resource allocation, refine pricing strategies, and implement targeted interventions to minimize revenue loss and streamline operations effectively. The goal is to use binary classification methods to predict if a hotel reservation is gonna get cancelled. For better context, we are more interested in understanding people who would end up cancelling a booked room.

## Data Dictionary:

Booking\_ID: unique identifier of each booking

no\_of\_adults: Number of adults

no\_of\_children: Number of Children

no\_of\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no\_of\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

type\_of\_meal\_plan: Type of meal plan booked by the customer:

required\_car\_parking\_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room\_type\_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.

lead\_time: Number of days between the date of booking and the arrival date

arrival\_year: Year of arrival date

arrival\_month: Month of arrival date

arrival\_date: Date of the month

market\_segment\_type: Market segment designation.

repeated\_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no\_of\_previous\_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no\_of\_previous\_bookings\_not\_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg\_price\_per\_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no\_of\_special\_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking\_status: Flag indicating if the booking was canceled or not.

Prior to training machine learning models, it is important to clean the data which involves discarding irrelevant features and preprocessing the data. The exact steps for data exploration and preprocessing are described in the section titled "Data Exploration and Preprocessing". After the dataset has been cleaned and prepared, different classifiers are trained and fine-tuned. Finally, the performance of the models is compared using accuracy and AUC as the performance measures.

# Related Work:

"Predicting Hotel Booking Cancellations using Machine Learning Techniques" by João Lopes, Pedro Almeida, and Orlando Belo

[Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/8260781/authors#authors)

There are many papers in this related field of hospitality demand & revenue management. Interestingly papers talk about the methods and this topic is still a long-term challenge to solve.

This paper (Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation, Authors : Yiying Chen1, Chuhan Ding2, Hanjie Ye3, Yuchen Zhou4 - [Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation | Atlantis Press (atlantis-press.com)](https://www.atlantis-press.com/proceedings/icfied-22/125971599) ) provides three possible substitutes for the neural network including logistic regression, k-Nearest Neighbor (k-NN), and CatBoost. The dataset used in this paper was adapted from Kaggle, a set of booking data from two types of hotels (resort hotel and city hotel) in Portugal, and the corresponding customers’ information. We select some key variables as the predictor to train and test the prediction models based on three machine learning algorithms. After preprocessing the raw data, i.e., standardizing, dealing with missing data, recoding some variables, and scaling, researchers conducted the prediction and compare each model through three metrics (confusion matrix, accuracy score, and F1-score) [1].

And yes our project is also based on dataset form Kaggle and hence we have compared our modelling process with another project on Kaggle. [Hotel Reservation Prediction ML|DL - 95% Accuracy (kaggle.com)](https://www.kaggle.com/code/subhranilmondal12/hotel-reservation-prediction-ml-dl-95-accuracy#Cross-Validation-Method-To-Eliminate-Overfitting). In this Kaggle project, the author has trie out various ML models and eventually built a perfect model using Neural Networks with approcimately 95% accuracy [2].

Unlike [1] & [2], this project attempts to perform multiple Machine Learning modelling approaches and eventually find out the right modelling way, which was the Random Forest Model with class weight )with approximately 98%. Also the data processing has been quite different from other 2 aforementioned projects. Most importanty we have used the AUC & Accuracy for estimatignt he performance of the models. Let us deep dive into our techniques in the following sections.

Data Exploration and Processing

The first step in data exploration is understanding the structure of the dataset. We have 19 features and 36275 observations. The given dataset if for the years 2017 and 2018, it has one and half year of data, from July 2017 to December 2018. All the 19 features have a significant value and the first column in the dataset, Booking\_ID, is a unique identifier. This feature can be discarded. There are no features that give demography information. Like any dataset there are both numerical and categorical columns. Numerical columns are "no\_of\_adults, "no\_of\_children", "no\_of\_weekend\_nights", "no\_of\_week\_nights", "required\_car\_parking\_space", "lead\_time", "repeated\_guest" "no\_of\_previous\_cancellations", "no\_of\_previous\_bookings\_not\_canceled", "avg\_price\_per\_room" "no\_of\_special\_requests", "arrival\_date", "arrival\_year", "arrival\_month". There are few categorical columns: type\_of\_meal\_plan", "room\_type\_reserved", "market\_segment\_type", "booking\_status". Assumption: price in the dataset is in $ USD.

Let us deep dive. As, we dove deep into the dataset we found that there were cases of prices to be $0. So we tried inspecting those cases by merging the arrival date, arrival month and arrival year as one date feature and understanding the price trends. Inspecting the price led us to understand the discrpeacny present in the date. Data was entered for the date Feb 29th, 2018, which is a blunder either no the time of data entry or corruption in the data due to data transfer or storage. We have used to a techinique to rectify that problem by analysing the number of obeservations present on day beore Feb 29th (i.e. Feb 28th ) and on the day aftyer Feb 29th (i.e. Mar 1st ). The results are shown below:

|  |  |
| --- | --- |
| Date | total |
| 2018-02-28 | 165 |
| 2018-03-01 | 61 |

Table 1 - Number of observations for the dates - Feb 28 & Mar 01

So to make it fair at the time of modelling we updated the 37 records of Feb 29th date to Mar 1st.

Once the date problem got fixed, we proceeded with understand the price trend and developed a prices chart. For a given single day there are multiple cases so aggregated with for each day at 3 levels – Min , Avg \*& Max price.

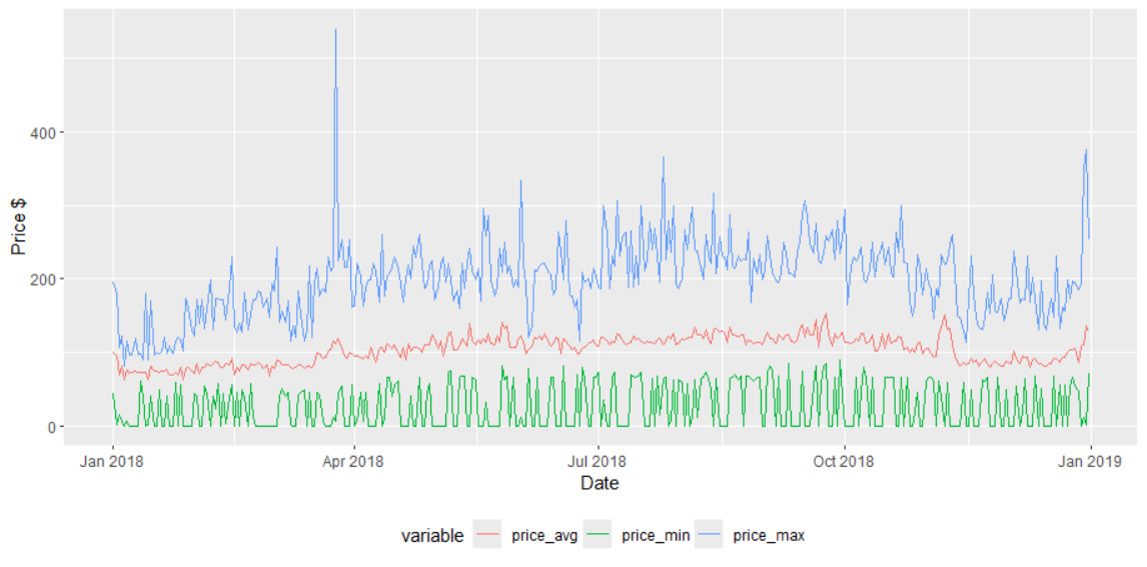


Figure 1 - Price Trend (2018)

From the graph, it is very evident that not all the levels of price is 0 at a given date, which means that the price cases that are 0 (green line evidently shows those cases) are for certain booking options. Then proceeded to understand the market segmentaion of prices 0 cases, and understood that theose are either from Online or Complimentatry Market segment. So our hypothesise to to this either in compliemtnray room option or online offer option one might be able to book wutb $0 price. To strengthen out hunch, we checked the number of cases that were cancelled, we see that only 6 booking IDs of the 546 $0 cases were canceled for rooms. People with free hotel rooms will not tend to cancel in most of the cases because of the human nature.

In data analysis, our next focus on cancalltion of booked rooms. Created a graph that plot sum of cancellation for each day in 2018. We can unsderstadn that during September. October, and November there are many cancellations.

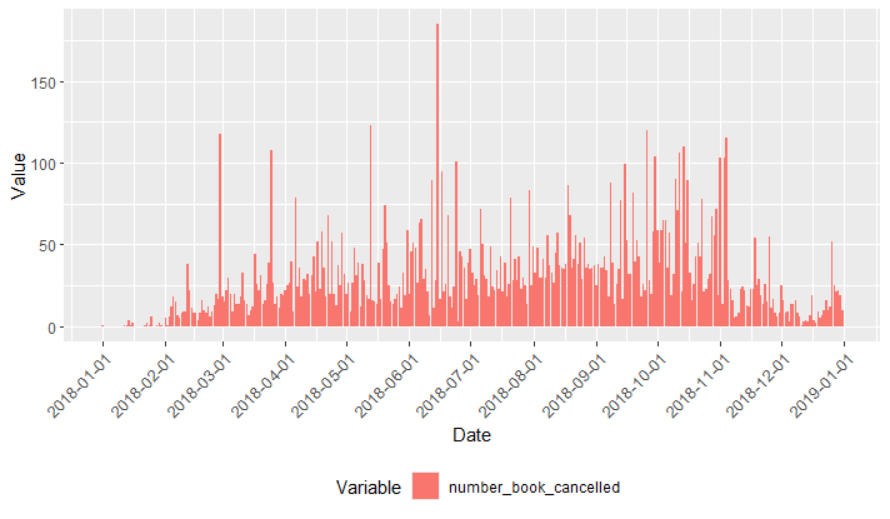


Figure 2 - Number of Cancllations (2018)

Then we focussed on understanding the relationship between various variables.

|  |  |
| --- | --- |
| Figure 3 - jj |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  | |
|  | |
|  |  |
|  |  |

