Predictive Model for Hotel Booking Cancellations

Introduction

Since the globalization, the hospitality industry has flourished over the recent decades. Nowadays, multiple options for accommodation are available to the tourists around the globe. And as an outcome, hotel booking cancellation has become the biggest concern for hotel industry. Cancellations for the booking leads to revenue losses, wastage of inventory, reduction in hotel's online reputation. The dataset used here is about a city hotel and a resort hotel in Portugal. Dataset is a made up of variables like arrival and departure date, number of adults, children and babies, meal, country, deposit type, agent, company, etc. The main goal of the project is to build a predictive model to predict the hotel booking cancellations that shall be carried out using various tree-based algorithms like Random Forest, Decision Tree, Extreme Gradient Boosting, Extra Tree Classifier. Later, the model having highest accuracy shall be picked.

Literature Review

N. Antonio, A. De Almeida, and L. Nunes (2018) published a manuscript "Hotel booking demand datasets". Two real datasets were released regarding hotel demand in the article, one of the hotels was Resort Hotel (H1) and the other was City Hotel (H2). Both the hotels had similar structure with 32 variables. For the current project, I shall be merging both the datasets and conduct data analysis.

N. Antonio, A. De Almeida, and L. Nunes (2019) published a research paper "An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations" to forecast bookings cancellation likelihood. Authors made two important research contributions based on continuously learning automated machine learning system, firstly evolution of training method and weighting mechanism and secondly a measure called Minimum Frequency used to determine precision of predictions over time. As a result, the systems helped drawing finer decisions along with better estimation of demand. The paper lists few scopes for future studies and limitations, of which was the imbalanced dataset and I shall try to eradicate this issue.

"Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry," research article was published by M. S. Satu, K. Ahammed and M. Z. Abedin (2020) to inspect the efficacy of various machine learning methods in hotel booking cancellation process. Amongst all the methods applied information gain feature selection methods showed the best result. Taking inspiration from the paper, I shall compare the results of the information gain feature with the selected decision tree-based models.

Y. Azhar, G. A. Mahesa, and M. C. (2021) Mustaqim published a manuscript labelled “Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm”. In order to derive optimum collection of parameters for prediction as well as truncate the losses faced due to hotel booking cancellations, authors decided to employ hyperparameter optimization to random forest algorithm to obtain the best performing model. I plan to employ hyperparameter optimization/tuning prior to fitting decision trees in order to spike up model performance.

Dataset

The dataset is selected from an article published under the title of “Hotel booking datasets” by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data accounts of demand in two hotels in Portugal, the first one is Resort Hotel (H1) and the second is a City Hotel (H2). Considering this is a real dataset all personally identifying information like names of resort, hotel and guests are concealed in order to maintain privacy. The original data was cleaned by Thomas Mock and Antoine Bichat for TidyTuesday on February 11, 2020. The clean data consists of 119390 rows and 32 variables where, 40,060 and 79,330 observations belonged to Resort Hotel (H1) and City Hotel (H2) respectively. The apprehend hotel bookings are expecting guests between July 1, 2015 and August 31, 2017.

Variables	Description	Data Type	Statistical Data Type	Null Count
hotel	Hotel (H1 = Resort Hotel or H2 = City Hotel)	object	Nominal	0
Is_cancelled	Value indicating if the booking was canceled (1) or not (0)	int64	Nominal	0
lead_time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	int64	Discrete	0
arrival_date_year	Year of arrival date	int64	Ordinal	0
arrival_date_month	Month of arrival date	object	Ordinal	0
arrival_date_week_number	Week number of year for arrival date	int64	Ordinal	0
arrival_date_day_of_month	Day of arrival date	int64	Ordinal	0
stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	int64	Discrete	0

stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	int64	Discrete	0
adults	Number of adults	int64	Discrete	0
children	Number of children	float64	Continuous	4
babies	Number of babies	int64	Discrete	0
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)	object	Nominal	0
country	Country of origin. Categories are represented in the ISO 3155–3:2013 format	object	Nominal	488
market_segment	Market segment designation: Direct, Corporate, Online TA, Offline TA/TO. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"	object	Nominal	0
Distribution_channel	Booking distribution channel: Direct, Corporate, TA/TO. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"	object	Nominal	0
Is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0)	int64	Nominal	0
previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking	int64	Discrete	0
previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking	int64	Discrete	0
reserved_room_type	Code of room type reserved. Code is presented instead of designation for anonymity reasons	object	Ordinal	0
assigned_room_type	Code for the type of room assigned to the booking.	object	Ordinal	0

	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	int64	Discrete	0
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.	object	Nominal	0
agent	ID of the travel agency that made the booking	float	Continuous	16340
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	float	Continuous	112593
days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer	int64	Discrete	0
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	object	Nominal	0

adr	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights	float64	Continuous	0
required_car_parking_spaces	Number of car parking spaces required by the customer	int64	Discrete	0
total_of_special_requests	Number of special requests made by the customer (e.g., twin bed or high floor)	int64	Discrete	0
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	object	Nominal	0
reservation_status_date	Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel	object	Ordinal	0

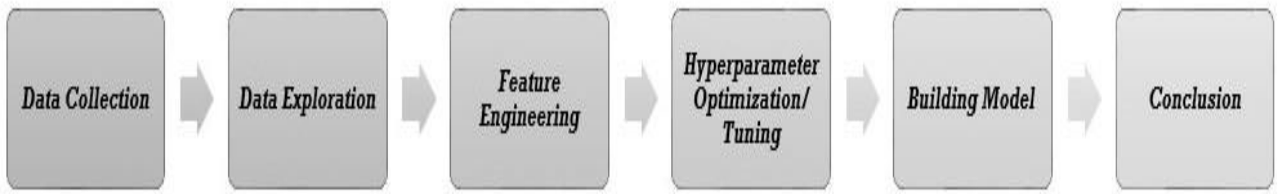
Initial observations about the dataset:

- Higher number of bookings for City Hotel than Resort Hotel
- Out of 32 columns, 20 were numerical and 12 were categorical
- Missing observations found in 4 columns namely: children, country, agent and company
- Reservation_status_date has been given incorrect data-type, instead it should be ‘datetime64’
- Also, children, agent and company are listed as float but in reality, they should be changed to ‘integer’

Descriptive Statistics

Variables	Mean	SD	Median	Min	Max
Is_cancelled	0.37	0.48	0	0	1
lead_time	104.01	106.86	69	0	737
arrival_date_year	2016	0.7	2016	2015	2017
arrival_date_month					
arrival_date_week_ number	27.16	13.61	28	1	53
arrival_date_day_of_month	15.79	8.79	16	1	31
stays_in_weekend_ nights	0.93	0.99	1	0	19
stays_in_week_nights	2.5	1.9	2	1	50
adults	1.86	0.58	2	2	55
children	0.10	0.39	0	0	10
babies	0.1	0	0	0	10
Is_repeated_guest	0.17	0	0	0	1
previous_cancellations	0.08	0.84	0	0	26
previous_bookings_not_canceled	0.14	1.49	0	0	72
booking_changes	0.22	0.65	0	0	21
deposit_type					
agent	86.69	1	229	9	535
company	131.66	6	270	62	543
days_in_waiting_list	2.32	17.59	0	0	391
adr	101.8	50.5	94.6	69.29	5400
required_car_parking_ spaces	0.06	0.25	0	0	8
total_of_special_ requests	0.57	0.79	0	0	5

Approach



Step 1: Data Collection

The original dataset is available as two different tables for H1 and H2. I shall compile them together in the first step both the tables have similar 32 variables.

Step 2: Data Exploration

Second step shall comprise of detailed study of the dataset. Identification of datatypes, number of null values, calculation of descriptive statistics, finding outliers, graphical visualization of the dataset and few other steps of initial data analysis shall be conducted in step 2

Step 3: Hyperparameter Optimization/Tuning and PCA

Hyperparameter Optimization/Tuning is the process to explore and select the set of optimal hyperparameters for an optimal learning algorithm automatically because such model produces the best model output. For this purpose, Grid Search Algorithm shall be used.

At this point I plan to use PCA on the dataset in order to run the various models. I shall compare the results in conclusion.

Step 4: Build Model

Prior to splitting the dataset into test and train datasets a variety of tree-based algorithms shall be employed for model building as mentioned earlier. I am planning to use 4 different models in order to compare the accuracy and recall Random Forest, Decision Tree, Extreme Gradient and Boosting Extra Trees Classifier.

The study feature of the study has imbalanced classification. Hence, SMOTE (oversampling) technique shall be applied.

Step 5: Conclusion

Looking at all the various models which were ran on 4 models it can be concluded that the SMOTE data under hyperparameter tuning condition is the most efficient model for prediction of hotel booking cancellations.

References:

1. Antonio, N., de Almeida, A. and Nunes, L. (2019). Hotel booking demand datasets. Data in Brief. 22, 41-49. https://doi.org/10.1016/j.dib.2018.11.126
2. Antonio, N., de Almeida, A. and Nunes, L. (2019). An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations, Data Science Journal, 18(1), 32. DOI: http://doi.org/10.5334/dsj-2019-032
3. Satu M., Ahammed K. and Abedin M. (2020). Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry, 23rd International Conference on Computer and Information Technology (ICCIT),1-6 https://doi.org/10.1109/ICCIT51783.2020.9392648.
4. Azhar, Y., Mahesa, G. and Mustaqim, M. (2021). Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm. Jurnal Teknologi dan Sistem Komputer, 9, 15-21. http://doi.org/10.14710/jtsiskom.2020.13790