Data Science and Machine Learning Project Proposal

By **Rakhaa Ismail**

1- Introduction:

The average percentage of canceled reservations is currently 24%. Therefore, when you calculate your future revenue from existing reservations, always reduce it by 24% to have an objective estimate. This problem leads to suffering from loss in hotel revenue because of the uncertain booking cancelation of its customers. Moreover, based on my research the prior knowledge of the rate of hotel reservation cancellations highly affects the prices of reservations. So, it will be beneficial if the hotel company has a previous prediction if the customer will cancel the booking or not. Or at least the average of cancelation that may happen. This project will benefit the hotel company. The project will be exploring business questions, which are:

- Which month has the highest number of cancelations?
- Which customer type has the highest cancellation?
- What is the average length of stay at a hotel?

And build machine learning models to help predict whether the customer will cancel the booking or not.

2- Data Description:

The dataset of the project is about hotel booking demand. I chose this dataset because I'm interested in the business field and solving business problems. The dataset was taken from Kaggle website. It contains 119,390 rows and 32 features. The data set represents the reservation records for two hotels in Portugal, but I'm not interested in these specific hotels. The project is about hotels in general. Especially that the features in the dataset are typically reservations information required for every hotel.

The features that I except to work with are: is canceled, lead time, arrival date year, arrival date month, market segment, is repeated guest, previous cancellations, reserved room type, booking changes, deposit type, customer type, adr, required car parking spaces. Below in the table is a detailed description of all features given in the data set.

No.	Features	Description	Data Type
1	Hotel	Dataset collected from two hotels. (H1) refer to hotel called a resort hotel and (H2) is a city hotel.	object
2	is_canceled	Value indicating if the booking was canceled (1) or not (0).	int64
3	Lead_Time	Number of days that elapsed between the entering date of the booking into the system and the arrival date.	int64
4	arrival_date_year	Year of arrival date.	int64
5	arrival_date_month	Month of arrival date with 12 categories: “January” to “December”.	object
6	arrival_date_week_number	Week number of the arrival date.	int64
7	arrival_date_day_of_month	Day of the month of the arrival date.	int64
8	stays_in_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.	int64
9	stays_in_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.	int64
10	adults	Number of adults.	int64
11	children	Number of children.	float64
12	babies	Number of babies.	int64
13	meal	Type of meal booked. Categories are presented as: 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
14	country	Country of origin.	object
15	market_segment	Market segment designation. In categories, ['Direct', 'Corporate', 'Online TA', 'Offline TA/TO', 'Complementary', 'Groups', 'Undefined', 'Aviation'].	object
16	distribution_channel	Booking distribution channel. ['Direct', 'Corporate', 'TA/TO', 'Undefined', 'GDS'].	object
17	is_repeated_guest	Value indicating if the booking name was from a repeated guest (1) or not (0).	int64
18	previous_cancellations	Number of previous bookings that were cancelled by the customer prior to the current booking.	int64

19	previous_bookings_not_canceled	Number of previous bookings not cancelled by the customer prior to the current booking.	int64
20	reserved_room_type	Code of room type reserved.	object
21	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.	object
22	booking_changes	Number of changes/amendments made to the booking from the moment the booking was entered on the system until the moment of check-in or cancellation.	int64
23	deposit_type	No Deposit, 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
24	agent	ID of the travel agency that made the booking.	float64
25	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.	float64
26	days_in_waiting_list	Number of days the booking was in the waiting list before it was confirmed to the customer Indication on if the customer made a deposit to guarantee the booking.	int64
27	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
28	adr	Average Daily Rate - Calculated by dividing the sum of all lodging transactions by the total number of staying nights.	float64
29	required_car_parking_spaces	Number of car parking spaces required by the customer.	int64
30	total_of_special_requests	Number of special requests made by the customer (e.g. twin bed or high floor).	int64
31	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
32	reservation_status_date	Date at which the last status was set.	object

3- Methodology:

- Classification

4- Tools:

- Numpy
- Pandas
- Scikit-learn
- Matplotlib
- Seaborn
- Statsmodels

And other additional tools.