HOTEL BOOKING ANALYSIS

Sonkar Abhishek , Akshay Pawar ,
Gourav Patil , Ritesh Neulkar , Kevin Varsani
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Analytics in the hotelier world today is important, and nowadays this business cannot be run with some sensible and smart use of data.

Here I demonstrate how to use data to analyze three business important concepts in the fields of revenue manamegment and marketing.

The analysis tries to answer three questions

1. How strong is the seasonality in these hotels?
2. At what time in week the hotels are more busy ?
3. Can we predict a cancellation, just with the information available at the moment this reservation has been made?

The first question relates to detecting seasonality patterns. A basic principle to establish a pricing strategy. The second addresses the time which has more bookings as couples or groups Moreover, this awkward behavior can make revenues having a hard time trying to sell rooms in the nearby dates. Finally, be able to detect whether or not a reservation will end up being canceled, is a powerful weapon for the marketing department, with important potential gains.

Keywords:machine learning,seasonality, cancellation, stay duration

1.Problem Statement

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

This hotel booking dataset can help you explore those questions!

Dataset contains following features:

i. hotel
ii. is_canceled
iii. lead_time
iv. arrival_date_year
v. arrival_date_month
vi. arrival_date_week_number
vii. arrival_date_day_of_month
viii. stays_in_weekend_nights
ix. stays_in_week_nights
x. adults

xi. Children babies
xii. meal
xiii. country
xiv. market_segment
xv. distribution_channel
xvi. is_repeated_guest
xvii. previous_cancellations
xviii. previous_bookings_not_canceled
xix. reserved_room_type
xx. assigned_room_type
xxi. booking_changes
xxii. deposit_type
xxiii. agent
xxiv. company
xxv. days_in_waiting_list
xxvi. customer_type
xxvii. adr
xxviii. required_car_parking_spaces
xxix. total_of_special_requests
xxx. reservation_status
xxxi. reservation_status_date

2. Introduction

This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has from the data.

3.Analysis log

- ❖ Cancelled Booking Analysis
- ❖ Hotelwise Analysis
- ❖ Agent Analysis
- ❖ Reservation Analysis
- ❖ Timewise Analysis
- ❖ Week Stay Analysis

4. Modelling

For seasonality patterns detection, visual descriptive analysis has been carried out, in addition to time series analysis, from which, the seasonal component of the series was extracted, and the strength of the seasonality component was measured. The method can be accessed in Forecasting : Principles And Practice by Rob J Hyndman.

For group differences, descriptive analysis and hypothesis testing have been carried out. For the measurement of the difficulty os selling rooms in the nearby dates of a group's stay, Cohen's d was also computed to measure effect size.

Regarding cancelation predictions. A random forest was perform, splitting the data into a training and test set (70 / 30)

5. Evaluation

For the first two questions, results are evaluated with visualizations and hypothesis test outputs. For the third question, the model was evaluated using accuracy, precision, recall, and F1 score.

As expected, there is a strong seasonality for the resort and city hotel, both in ADR and room nights. Groups have shorter stays in the resort hotel, book with significantly less time in advance, at lower ADRs. In the city hotel, it was found that rooms in dates nearby group stays are sold at an average lower ADR. Regarding the prediction of cancellations, the model obtained an 85 % accuracy, 81 % precision, 77 % recall, and a 79 % f1-score.

Measuring the difficulty of selling rooms nearby group's stays is a very interesting task, which can be expanded much more than just detecting significant results and effect sizes after dividing the data into two groups: we can change the boundaries for what is considered a large group (in the analysis, 30 % of maxim occupation), and how many nearby dates are selected (in the study, up to 6 days away from a group stay were considered as nearby dates) and analyze how varies the difficulty of selling those rooms when we tweak some of these inputs. More complex models can be performed and more interesting results can be obtained. For a future analysis of this dataset I will delve into this question, provided that the limitations of this data set allow me to do so.

As for the prediction of cancellations concerns. it is clear that better results can be achieved in a more exhaustive machine learning process, that includes more models into consideration. Besides, this data is somewhat limited (only two years). A wider time window and more features, which sure will be at the hands of every hotelier in the business, better results could be obtained.

After that we made the predictive model to predict whether the booking will be cancelled or not

We will:

- Perform the Feature Engineering to make new featuers
- Perform the Data Selection to select only relevant features
- Tranform the Data (Categorial to Numerical)
- Split the data (Train Test Split)
- Model the data (Fit the Data)
- And finally Evaluate our model

File description:

- **requirements.txt** : requeriments file with the modules used in the analysis.
- **Analysis of hotel bookings.ipynb** : jupyter notebook with the analysis.

- **hotel_ bookings.csv** : csv files with the data.
- **Figures** : a folder containing some of the figures produced in the notebook.

The output of the analysis can be obtained directly by running all the cells in the Jupyter Notebook file. There are texts and visualizations explaining the output.

Installations:

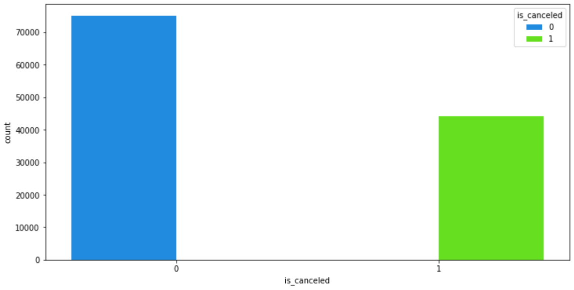
Below is a list of the modules used in the analysis is shown

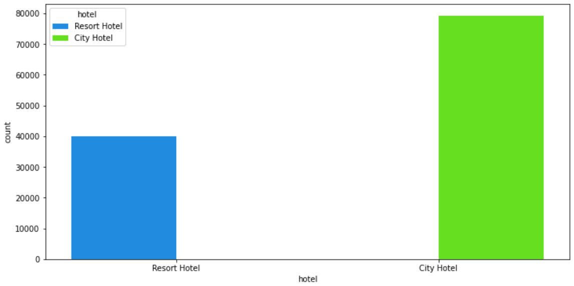
- pandas
- numpy
- matplotlib
- seaborn
- datetime
- dateutil
- statsmodels
- calendar
- scipy
- sklearn
- math

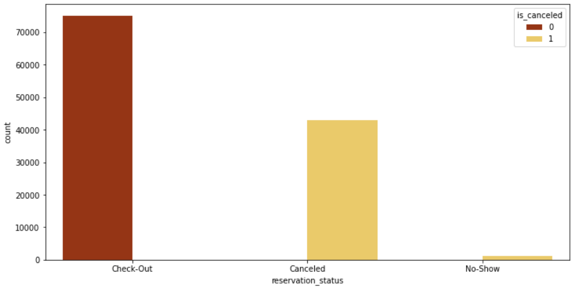
6. Conclusion:

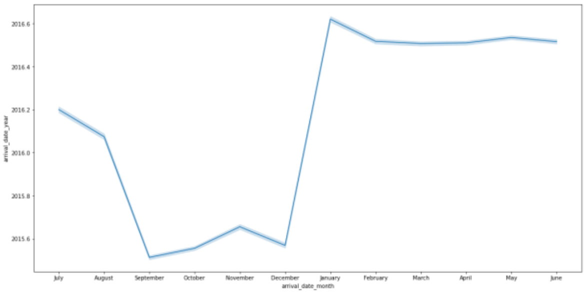
We learned that

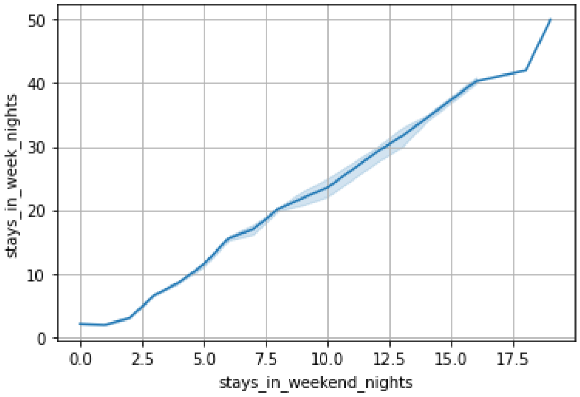
- Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel.
- More than 60% of the population booked the City hotel.
- Stays in week nights is more than weekend nights.
- Most bookings were made from July to August. And the least bookings were made at the start and end of the year.
- Almost 40% is in cancel reservation.











7. Acknowledgements

The data is originally from the article Hotel Booking Demand Datasets, given by Almabetter.

The data was downloaded and cleaned by Sonkar Abhishek , Akshay Pawar, Gourav Patil , Ritesh Neulkar , Kevin Varsani .