International Journal of Research in Engineering, Science and Management
Volume 6, Issue 5, May 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

Indoor Data Prediction for Hotel Using Python

Kartikey Singh[1*], Garima Pandey[2], Anjali Yadav[3]
[1,2,3]Student, Department of Computer Science and Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow, India

***Abstract*: This project includes a record of actual hotel reservations for a city and a resort hotel. It includes information on reservations, cancellations, and guest information. The project's primary goal is to comprehend and display data from the perspectives of hotels and customers. We will perform exploratory data analysis with python to get insight from the data and then predict bookings.**

***Keywords*: Hotel booking, resort hotels, city hotels.**

1. Problem Statement

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 been removed from the data.

Now, we have to:

"Explore and analyze the data to discover important factors that govern the bookings and predict bookings."

2. Introduction

In this project, we will try to figure out the following questions:

1. How many bookings were cancelled?
2. What is the booking ratio between Resort Hotel and City Hotel?
3. Which is the busiest month for hotels?
4. From which country most guests come?
5. How long people stay in the hotel?
6. Which was the most booked accommodation type (Single, Couple, Family)?
7. How many bookings will happen in the next 6 months?

3. Analysis to be Done

- ANALYSIS #1: Checking the status of reservation and the repeated guest of city hotel and resort hotel
- ANALYSIS #2: Country name from where most of the guests are coming
- ANALYSIS #3: Month wise bookings
- ANALYSIS #4: Bookings by various sources
- ANALYSIS #5: Monthly cancellations
- ANALYSIS #6: Cancellations by type of customer
- ANALYSIS #7: Duration of stay of people
- ANALYSIS # 8: Comparison of Duration of Stay in both types of hotel
- ANALYSIS #9: Most booked accommodations type
- ANALYSIS #10: Booking ratio of type of hotel
- PREDICTION: Prediction of hotel bookings for the next 6 months.

4. Access to Dataset

We have stored our Dataset in Google Drive in form of .CSV File, in order to access the data from drive we have to follow the below given steps:

- Mounting the Notebook from Drive
- Providing the Path to Read the .CSV File

Codes to Access the Data:

Mounting the Notebook from Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

Providing the Path to Read the CSV File

```
path = '/content/drive/MyDrive/AlmaBetter/Python for Data Science/Capstone Project/Copy of Hotel Bookings.csv'

df = pd.read_csv(path)
```

5. Importing the Packages

In order to perform the analysis, we will import some necessary packages of Python.

Codes for Importing the Packages

```
!pip install pycountry # Using !pip, To import a library that's not in Colaboratory by default
import pandas as pd # IMPORITNG PANDAS
import seaborn as sns # IMPORTING SEABORN
import matplotlib #IMPORTING MATPLOTLIB
import pycountry as pc #IMPORTING PYCOUNTRY
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import matplotlib.ticker as mtick
%matplotlib inline
```

Codes for Additional Predefined Setting

```
sns.set_style('whitegrid')
matplotlib.rcParams['font.size'] = 14
```

*Corresponding author: kartikeysingh0802@gmail.com

```
matplotlib.rcParams['figure.figsize'] = (12, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

6. Data Preprocessing

1) Checking for missing or null values

Code to check Null Value

```
df.isnull()
```

hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	4
babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593

We can see that some of the columns in given dataset is having Null Values, so we have to remove the Null Values.

Codes to Remove Null Values

```
df['country'].fillna(0,inplace=True)
df['agent'].fillna(0,inplace=True)
df['company'].fillna(0,inplace=True)
```

2) Separating and creating different datasets of both the hotel types. (i.e., Resort and City Hotel)

Codes to Fill Null Values

```
df['country'].fillna(0,inplace=True)
df['agent'].fillna(0,inplace=True)
df['company'].fillna(0,inplace=True)
```

7. Analysis

A. Checking the Status of Reservation and the Repeated Guest of City Hotel and Resort Hotel

Result by Analysis: It helps the company to get the detail about the Check-out and Cancel Reservation Status

Code for Analysis # 1

```
sns.countplot (x= 'reservation_status', data= df, hue= 'hotel').set_title ('Reservation Status')
```

Fig. 1. Reservation status

Fig. 2. Status of repeated guest

B. Country Name from where most of the Guests are coming

Result by Analysis: This Analysis helps company to make custom or attractive packages for the country from where fewer guests are coming

Code for Analysis # 2

```
sns.barplot (y= list(df.country.value_counts().head (12)), x= list(df.country.value_counts().head(12).index))
```

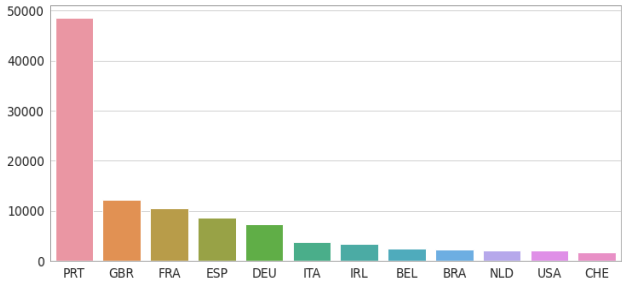

Fig. 3.

C. Month Wise Bookings

Result by Analysis: By this Analysis Company can figure Out which month have less no. of bookings, and provide some offers in that months to increase the sales.

Code for Analysis # 3

```
## Order of months
new_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December']
df_not_canceled = df[df['is_canceled'] == 0]
## Select only City Hotel
sorted_months = df_not_canceled.loc[df.hotel=='City Hotel'
,'arrival_date_month'].value_counts().reindex(new_order)
```

```
x1 = sorted_months.index
y1 = sorted_months/sorted_months.sum()*100

## Select only Resort Hotel
sorted_months = df_not_canceled.loc[df.hotel=='Resort
Hotel'
,'arrival_date_month'].value_counts().reindex(new_order)

x2 = sorted_months.index
y2 = sorted_months/sorted_months.sum()*100

## Draw the line plot

fig, ax = plt.subplots(figsize=(18,6))

ax.set_xlabel('Months')
ax.set_ylabel('Booking (%)')
ax.set_title('Booking Trend (Monthly)')

sns.lineplot(x1, y1.values, label='City Hotel', sort=False)
sns.lineplot(x2, y2.values, label='Resort Hotel', sort=False)

plt.show()
```

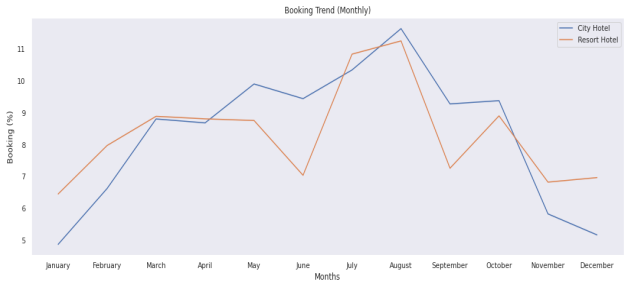

Fig. 4.

D. *Bookings by Various Sources*

Result by Analysis: By this, Company can get the sources of Booking and get the Idea on which platform they have to focus

Code for Analysis # 3

```
## Order of months
new_order = ['January', 'February', 'March', 'April',
'May', 'June', 'July', 'August', 'September', 'October',
'November', 'December']
df_not_canceled = df[df['is_canceled'] == 0]
## Select only City Hotel
sorted_months = df_not_canceled.loc[df.hotel=='City
Hotel'
,'arrival_date_month'].value_counts().reindex(new_order)

x1 = sorted_months.index
y1 = sorted_months/sorted_months.sum()*100

## Select only Resort Hotel
sorted_months = df_not_canceled.loc[df.hotel=='Resort
Hotel'
,'arrival_date_month'].value_counts().reindex(new_order)

x2 = sorted_months.index
y2 = sorted_months/sorted_months.sum()*100

## Draw the line plot

fig, ax = plt.subplots(figsize=(18,6))

ax.set_xlabel('Months')
ax.set_ylabel('Booking (%)')
ax.set_title('Booking Trend (Monthly)')

sns.lineplot(x1, y1.values, label='City Hotel', sort=False)
sns.lineplot(x2, y2.values, label='Resort Hotel', sort=False)

plt.show()
```

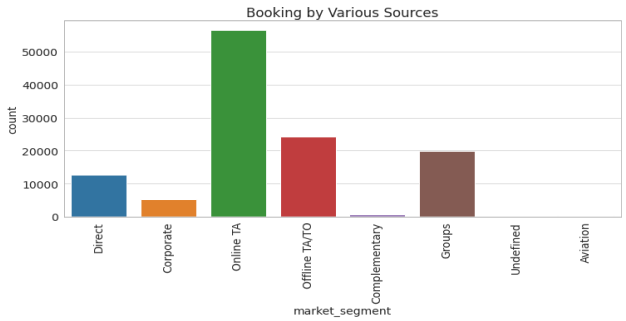

Fig. 5. Booking by various sources

E. *Monthly Cancellations*

Result by Analysis: By this Analysis Company can figure out which month have max. Number of cancellations so that they can have the Idea of Reason for Cancellation like month containing less holidays etc.

Code for Analysis # 5

```
df_copy1= df.copy()
df_copy1['no_of_bookings']= 1
df_copy1.loc[(df_copy1.arrival_date_month== 'July')|
(df_copy1.arrival_date_month==
'August'),'no_of_bookings']/=3
df_copy1.loc[~((df_copy1.arrival_date_month== 'July')|
(df_copy1.arrival_date_month==
'August')),'no_of_bookings']/=2
df_copy1.sample(100)
b= df_copy1[df_copy1.is_canceled == 1]
month_sorted =
['January','February','March','April','May','June','July','
August','September','October','November','December']
sns.set(style="dark")
plt.title("Monthly cancellations")
sns.barplot(x= 'arrival_date_month', y= 'no_of_bookings',
hue= df_copy1['hotel'], estimator= sum, data= df_copy1)
plt.xticks(rotation = 90)
plt.show()
```

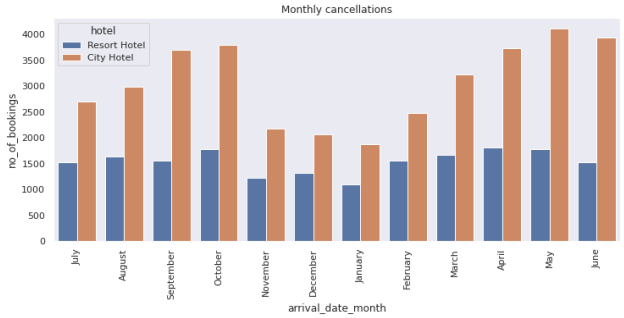

Fig. 6. Monthly cancellations

F. *Cancellations by Type of Customer*

Result by Analysis: By this Insight, Company can easily focus on the type of Customers who cancel their Visit / Stay

Code for Analysis # 6

```
a = df.groupby("customer_type")['is_canceled'].describe()

sns.barplot(x=a.index, y=a["mean"] * 100)
```

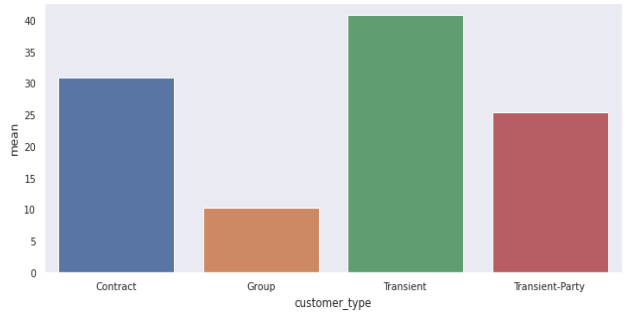

Fig. 7.

G. *Duration of Stay of People*

Result by Analysis: This is just for Information of the Duration of Stay.

Code for Analysis # 7

```
total_nights =
df_not_canceled['stays_in_weekend_nights']+
df_not_canceled['stays_in_week_nights']
x,y = get_count(total_nights, limit=15)

plot(x,y, x_label='Number of Nights', y_label='Booking
Percentage (%)', title='Night Stay Duration (Top 15)',
figsize=(10,5))
```

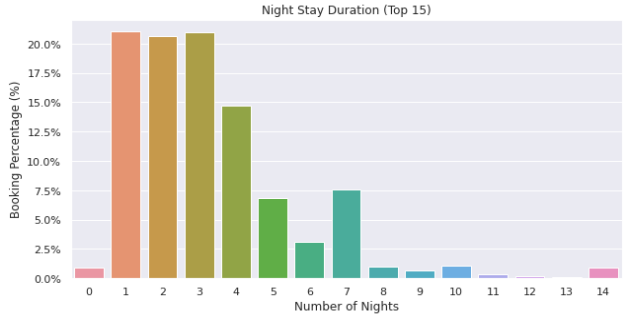

Fig. 8. Night stay duration

H. *Comparison of Duration of Stay in both types of Hotel*

Result by Analysis: This Analysis provide the comparison by which company get to know that in which type of hotel, people like to stay more.

Code for Analysis # 8

```
df_not_canceled.loc[:,'total_nights'] =
df_not_canceled['stays_in_weekend_nights']+
df_not_canceled['stays_in_week_nights']

fig, ax = plt.subplots(figsize=(12,6))
ax.set_xlabel('No of Nights')
ax.set_ylabel('No of Nights')
ax.set_title('Hotel wise night stay duration (Top 10)')
sns.countplot(x='total_nights', hue='hotel',
data=df_not_canceled,
order =
df_not_canceled.total_nights.value_counts().iloc[:10].index
, ax=ax);
```

Fig. 9. Hotel wise night stay duration

I. *Most Booked Accommodation Type*

Result by Analysis: By this Analysis, company get to know, which accommodation type is highly booked, so that company can increase or decrease the types of accommodation in their property accordingly.

Code for Analysis # 9

```
## Select single, couple, multiple adults and family
single = df_not_canceled[(df_not_canceled.adults==1) &
(df_not_canceled.children==0) &
(df_not_canceled.babies==0)]
couple = df_not_canceled[(df_not_canceled.adults==2) &
(df_not_canceled.children==0) &
(df_not_canceled.babies==0)]
family = df_not_canceled[df_not_canceled.adults +
df_not_canceled.children + df_not_canceled.babies > 2]

## Make the list of Category names, and their total
percentage
names = ['Single', 'Couple (No Children)', 'Family /
Friends']
count = [single.shape[0],couple.shape[0], family.shape[0]]
count_percent = [x/df_not_canceled.shape[0]*100 for x in
count]
```

```
## Draw the curve
plot(names,count_percent,  y_label='Booking (%)',
title='Accommodation Type', figsize=(10,7))
```

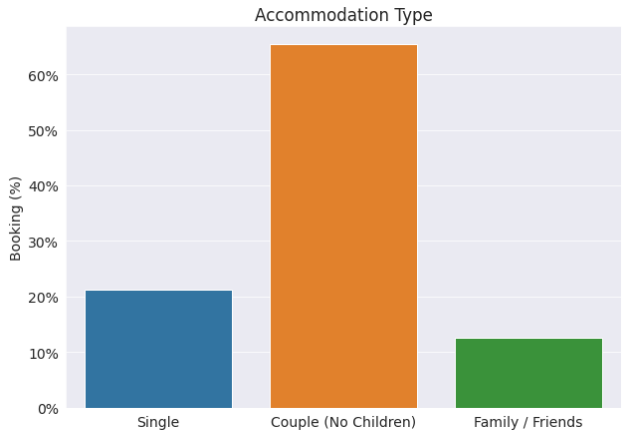

Fig. 10. Accommodation type

J. Booking Ratio of Type of Hotel

Result by Analysis: Finally, by this analysis, company gets the Ratio of booking of both types of Hotels.

Code for Analysis # 10

```
x,y = get_count(df_not_canceled['hotel'])
plot(x,y, x_label='Hotels', y_label='Total Booking (%)', titl
e='Hotel comparison')
```

Fig. 11. Hotel comparison

K. Prediction (Linear Regression) of Hotel Bookings for the next 6 Months

Result by Analysis: Finally, by this Linear Regression prediction, company get the details of hotel bookings for the next 6 months

Code for Prediction

```
x=df[['arrival_date_month' , 'lead_time', 'adults',
'children', 'previous_bookings_not_canceled']]
y=df['is_canceled']

x=pd.get_dummies(x)

x_train, x_test, y_train, y_test=train_test_split(x,y,
test_size=0.2 , random_state=42)

from sklearn.impute import SimpleImputer

# create an instance of SimpleImputer to fill in missing
values with the mean
imputer = SimpleImputer(strategy='mean')

# fit the imputer on x_train and transform both x_train
and x_test
x_train_imputed = imputer.fit_transform(x_train)
x_test_imputed = imputer.transform(x_test)

# train the linear regression model on the imputed data
lr_model = LinearRegression()
lr_model.fit(x_train_imputed, y_train)

# predict the test data and evaluate the model performance
y_pred = lr_model.predict(x_test_imputed)

from sklearn.metrics import r2_score
r2 = r2_score(y_test, y_pred)

months=['June 2023','July 2023', 'August 2023',
'September 2023','October 2023', 'November 2023']
lead_time=30
adults=2
children=0
previous_bookings_not_canceled=1
data={'arrival_date_month':months,
'lead_time':[lead_time]*len(months),
'adults':[adults]*len(months),
'children':[children]*len(months),
'previous_bookings_not_canceled':[previous_bookings_not
_canceled]*len(months)}
x_pred=pd.DataFrame(data)
x_pred=pd.get_dummies(x_pred)

#y_pred=lr_model.predict(x_pred)

y_pred_rounded_truncated = y_pred[:len(months)]
plt.plot(months, y_pred_rounded_truncated )

plt.xlabel('Months')
plt.ylabel('Number of bookings')
plt.title('Predicted bookings for 6 months')
plt.show()
```

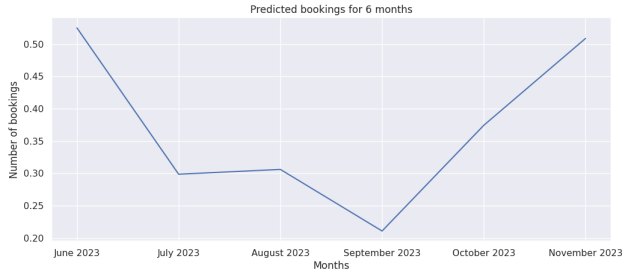

Fig. 12. Predicted bookings for 6 months

8. Conclusion

We used the dataset that contains data about hotel bookings.

We cleaned and preprocessed the data and then we performed the exploratory data analysis to extract information from the data to get the answer like:

- Resort hotel cancelation rate is low (approx. 30%). Most of the time booking not cancelled.
- Most of the population booked the city hotel, also the percentage of repeated guest is high in City hotel.
- Portugal, GBR and FRA are the top countries from where most of the guest comes.
- Most bookings were made from July to August. And the least bookings were made at the start and end of the year.
- Guest uses different channel for making booking, out of which most preferred way were online through TA (Travel Agent).
- Both the Hotels can figure out which month have max. number of cancellations so that they can have the Idea of Reason for Cancellation like month containing less holidays etc.
- Monthly cancellation Analysis can figure out which month have max. number of cancellations so that they can have the Idea of Reason for Cancellation like month containing less holidays etc.
- Both the hotel can easily focus on the type of Customers who cancel their Visit / Stay.
- Both the Hotel Information of the Duration of Stay is available.
- After comparison of duration of stay in both hotel it is easily get to know that in which type of hotel, people like to stay more.
- Most booked accommodation type can increase or decrease the types of accommodation in their property accordingly.
- Booking ratio gives the ratio of booking of both type of hotel.
- Prediction (Linear Regression) provides the company with details of hotel bookings for the next 6 months.

References

[1] https://medium.com/
[2] https://www.geeksforgeeks.org/
[3] https://www.researchgate.ne
[4] https://www.kaggle.com/
[5] https://jovian.ai/