PHASE 3

PROJECT TITLE: HOTEL BOOKING PREDICTION

Name: Sahithi Bakaram **UBID:** sahithib

Name: Spoorthy Reddy Avancha **UBID:** savancha

1. INTRODUCTION:

Nowadays, hotels are occupying a significant place in travelers' thoughts as they go from one place to another. Hotels genuinely feel like you're getting a break from all the daily routine and helps you get a fresh perspective. People typically reserve hotels for a variety of occasions, from the necessity to luxury, such as when they require a co-working area for business meetings, when they wish for a staycation/vacation, or when they must leave work early for an event and need to change.

The hotels, on the other hand, need to know when the busy times are, who tends to cancel bookings, if parking spaces are in high demand, and how many people will be staying. We want to provide a summary of the features based on the dataset in this project.

2. PROBLEM STATEMENT:

As travel demand increases, it becomes harder to get a quality hotel room. Using "Hotel Booking Demand database", we use machine learning to create an approach for hotel booking by utilizing data and booking attitudes and behaviors. The travelers are increasingly impacted by the daily trend of hotel cancellations during border times. Using "Hotels booking Demand" analysis we can know the Hotel Cancellation Predictions, thereby users may choose whether to book a hotel in advance or not. Further, it will also address some of the following questions:

Which hotel is the most popular?

What elements affect to hotel cancellations?

What types of rooms are most popular?

From the Hotel perspective, knowing that the booking has a higher probability of getting cancelled in advance will help them to handle it so that they can minimize the loss and manage their revenue. Also, having insights about the booking might help them to improve the facilities they can provide so that they can enhance the user experience. For example,

1. Knowing about the frequency of the booking type i.e., if people are visiting for a business, you can enhance their stay experience by having a meeting room
2. If the guests include kids, you can have medical facilities ready incase of an emergency. You can also have a kids play area and a meal plan designed for them.

The dataset can be used for drawing multiple insights to increase hotel's revenue.

By comprehending the business need, the data was examined, cleaned and performed out exploratory data analysis and plotted different types of relationships.

The datasets that are currently available were gathered with the intention of creating classification schemes for the likelihood of cancellation of hotel reservations. However, these datasets use extend beyond this cancellation prediction issue because of the properties of the variables of the datasets.

These datasets can be important for research in revenue management, machine learning, or data mining, as well as in other sectors, due to the lack of real business data for scientific and educational reasons.

3. DATA SOURCE:

- The "Hotels Booking Demand" data is taken from Kaggle.com. The data was first published in the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.
- Portugal provides most of the hotel data.

4. DATASET DETAILS:

The Hotel Booking Demand Dataset consists of hotel booking data from two different hotels located in Portugal. Hotel H1 is located at the resort region of Algarve and Hotel H2 is at the city of Lisbon. The Hotel names include Resort Hotel (H1) and City Hotel (H2). The dataset has the observations of 3 years.

The dataset contains 32 columns and 119390 rows in total. Resort Hotel and City Hotel comprises of 40060 and 79330 observations respectively.

5. DATA CLEANING / PROCESSING:

The below mentioned steps are done as part of the Data Preprocessing:

i. Deleting the duplicate values if any.
ii. Handling the missing values.
iii. Deleting the unwanted rows and columns.
iv. Converting all columns to proper data types and precision
v. Identifying the Outliers
vi. Shape of the Distribution
vii. Measures of Center: Mean, Median
viii. Modality
Modality gives the most frequent value in the dataset. We have calculated the mode values for each feature.
ix. Quantiles
We split the date into four parts using quantiles i.e., 25%, 50%, 75%, max.
25% refers to the first quartile, 50% refers to the median, 75% refers to the third quartile.
x. Removing the outliers
By calculating Inter Quartile Range IQR, Q1, Q3, we remove the outliers.

6. Exploratory Data Analysis (EDA):

After Exploratory Data Analysis, we found out the correlations between is_cancelled and other independent variables. This correlations helps us to make improvements in the Hotel

Management System by improving techniques, thereby cancellation of bookings can be minimized.

7. Algorithms

After data cleaning, we considered the features that are important and even combined some columns. Now the cleaned dataset has 27 columns and 86639.

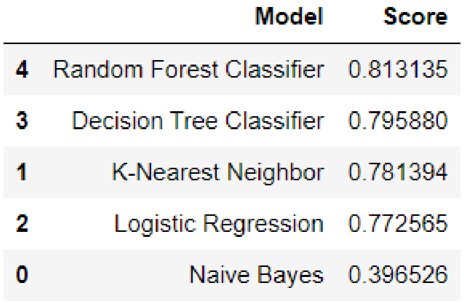
After all the data preprocessing and data cleaning is done, the dataset is separated into response and predictor variables, followed by a partition into two parts, where 80% is used to train the model and the rest 20% is used as a test dataset. We fit and transform the training data using StandardScaler() and then transform the test data as well. Finally, the prediction of whether a room is canceled or not is performed based on other independent variables using machine learning algorithms.

Here classification algorithms such as, Naive Bayes, KNN , Logistic regression, Decision tree, and Random Forest are considered as the dataset comes under the classification with the inference of prediction

Using each algorithm, we fit the model with the training dataset, and then predict cancelation or not for the records in the test dataset.

Model Comparison:

Of all the models, Random Forest Classifier is the best model as the accuracy score for it is higher than any other models considered.

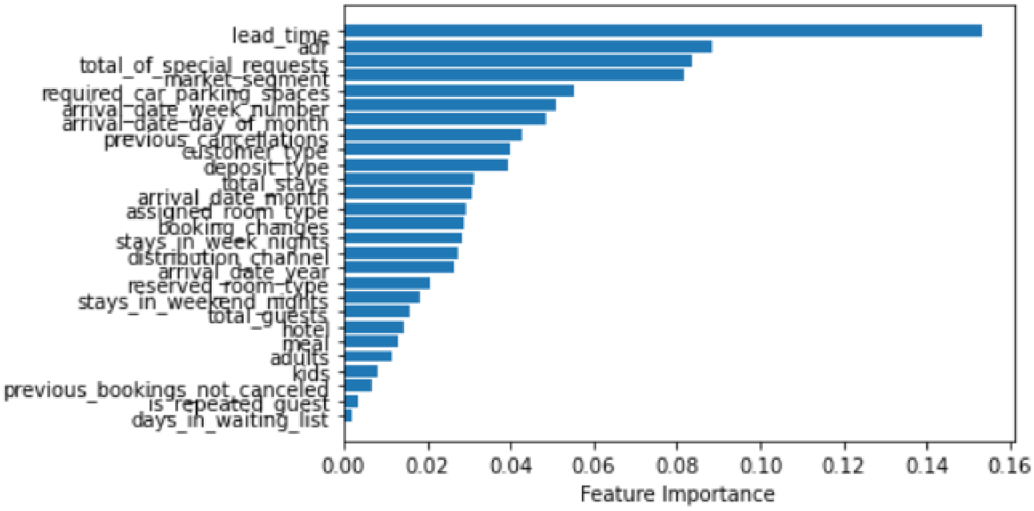


	Model	Score
4	Random Forest Classifier	0.813135
3	Decision Tree Classifier	0.795880
1	K-Nearest Neighbor	0.781394
2	Logistic Regression	0.772565
0	Naive Bayes	0.396526

Accuracy of all models considered

Feature Importance:

The figure below gives the feature importance for the Random Forest model based on the Gini criteria. Feature importance do not indicate if a feature have a negative or positive impact on the outcome variable, it only show how much information each variable contains in order to determine which class an observation belongs to. In Figure Leadtime is by far the most dominant variable, with the rest of the variables.



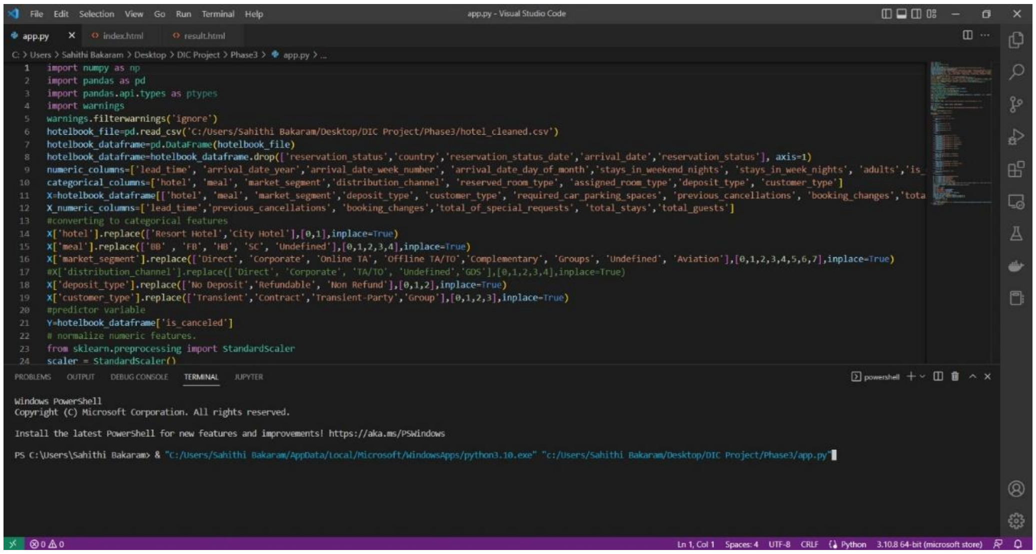
Feature importance through the random forest algorithm

8. Model used in the Data Product

Accuracy score for Random Forest classifier is 81.31% and this is the highest score obtained compared to other models where the score ranged from 39.6%-79.5%.
So, we considered **Random Forest Classifier** in our data product.
We incorporated HTML, CSS for our data product using Flask framework.

9. Working Instructions

app.py-The model that has the highest accuracy score is Random Forest classifier. This model is built, trained, and dumped into the pickle file model.pkl.
Next, we read the model in the pickle file and predict the model using Flask Web Framework to handle the POST requests.



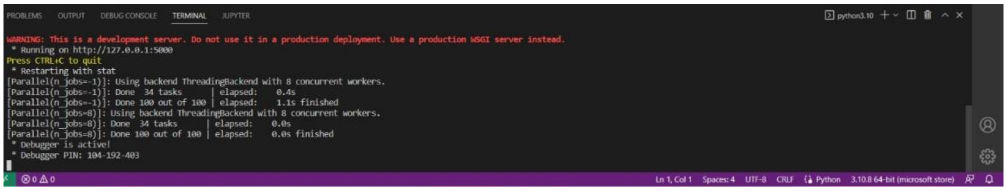
```python
import numpy as np
import pandas as pd
import pandas.api.types as ptypes
import warnings
warnings.filterwarnings('ignore')
hotelbook_file=pd.read_csv('C:/Users/Sahithi Bakaram/Desktop/DIC Project/Phase3/hotel_cleaned.csv')
hotelbook_dataframe=pd.DataFrame(hotelbook_file)
hotelbook_dataframe=hotelbook_dataframe.drop(['reservation_status','country','reservation_status_date','arrival_date','reservation_status'], axis=1)
numeric_columns=['lead_time', 'arrival_date_year','arrival_date_week_number', 'arrival_date_day_of_month','stays_in_weekend_nights', 'stays_in_week_nights', 'adults','is_
categorical_columns=['hotel', 'meal', 'market_segment','distribution_channel', 'reserved_room_type', 'assigned_room_type','deposit_type', 'customer_type']
X=hotelbook_dataframe[['hotel', 'meal', 'market_segment','deposit_type', 'customer_type', 'required_car_parking_spaces', 'previous_cancellations', 'booking_changes','tota
X_numeric_columns=['lead_time','previous_cancellations', 'booking_changes','total_of_special_requests', 'total_stays','total_guests']
#converting to categorical features
X['hotel'].replace(['Resort Hotel','City Hotel'],[0,1],inplace=True)
X['meal'].replace(['BB' , 'FB', 'HB', 'SC', 'Undefined'],[0,1,2,3,4],inplace=True)
X['market_segment'].replace(['Direct', 'Corporate', 'Online TA', 'Offline TA/TO','Complementary', 'Groups', 'Undefined', 'Aviation'],[0,1,2,3,4,5,6,7],inplace=True)
#X['distribution_channel'].replace(['Direct', 'Corporate', 'TA/TO', 'Undefined','GDS'],[0,1,2,3,4],inplace=True)
X['deposit_type'].replace(['No Deposit','Refundable', 'Non Refund'],[0,1,2],inplace=True)
X['customer_type'].replace(['Transient','Contract','Transient-Party','Group'],[0,1,2,3],inplace=True)
#predictor variable
Y=hotelbook_dataframe['is_canceled']
# normalize numeric features.
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Sahithi Bakaram> & "C:/Users/Sahithi Bakaram/AppData/Local/Microsoft/WindowsApps/python3.10.exe" "c:/Users/Sahithi Bakaram/Desktop/DIC Project/Phase3/app.py"
```

app.py Execution



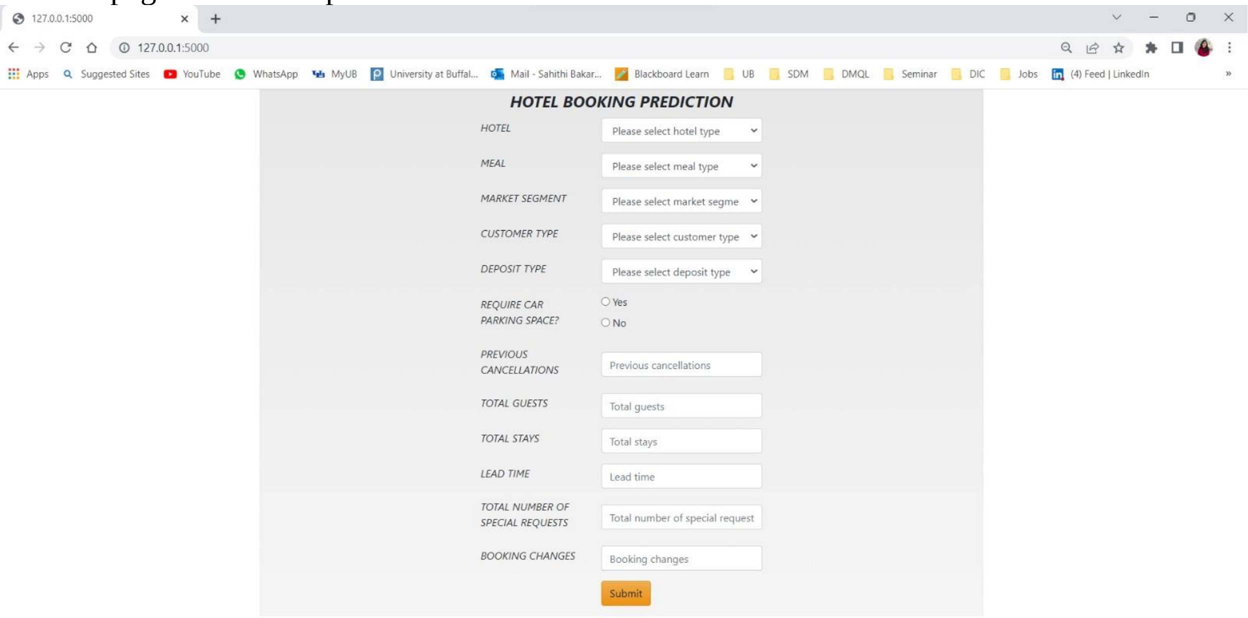
```
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed:    0.4s
[Parallel(n_jobs=-1)]: Done 100 out of 100 | elapsed:    1.1s finished
[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 100 out of 100 | elapsed:    0.0s finished
 * Debugger is active!
 * Debugger PIN: 104-192-403
```

After the execution of app.py,it pops the url of server it is running on

Finally, we run the server using the port 5000.

Using the address http://127.0.0.1:5000, we can access the Hotel Booking Prediction GUI with the redirection from app.py to input.html.

The webpage of our data product is as follows:

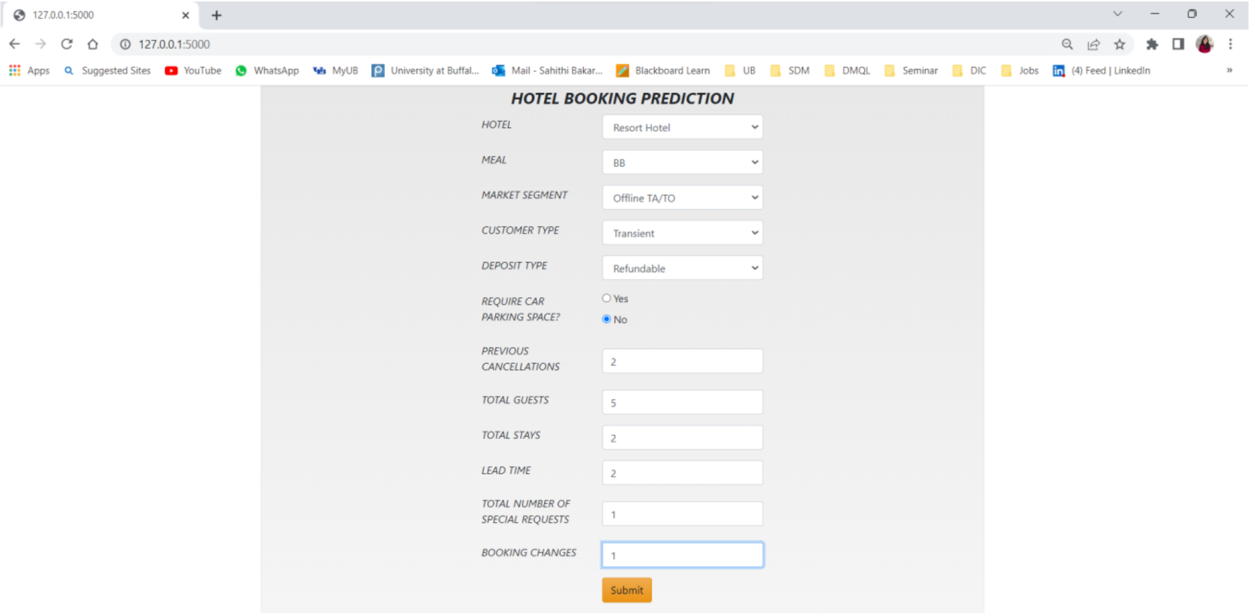


HOTEL BOOKING PREDICTION

HOTEL	Please select hotel type
MEAL	Please select meal type
MARKET SEGMENT	Please select market segme
CUSTOMER TYPE	Please select customer type
DEPOSIT TYPE	Please select deposit type
REQUIRE CAR PARKING SPACE?	○ Yes ○ No
PREVIOUS CANCELLATIONS	Previous cancellations
TOTAL GUESTS	Total guests
TOTAL STAYS	Total stays
LEAD TIME	Lead time
TOTAL NUMBER OF SPECIAL REQUESTS	Total number of special request
BOOKING CHANGES	Booking changes

User can input the details to the GUI for the 12 features below and the get the prediction. Of all these, lead time is the feature that is affecting the most.
Upon entering the input values and submitting it, the input.html is redirected to result.html page.

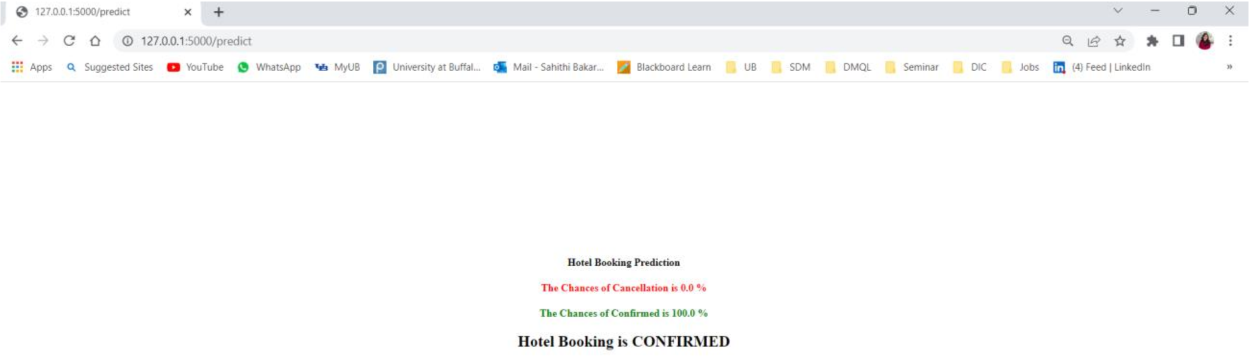
Hotel Booking Prediction as the status, Confirmed:
By giving the below inputs, the Hotel Booking Prediction is in the CONFIRMED status. The chance of probability of Cancellation is lower, typically it is 0%. And the chance of probability of Confirmation is 100%. So, it is more likely to be CONFIRMED.
If we observe, the lead time value is 2.



HOTEL BOOKING PREDICTION

HOTEL	Resort Hotel
MEAL	BB
MARKET SEGMENT	Offline TA/TO
CUSTOMER TYPE	Transient
DEPOSIT TYPE	Refundable
REQUIRE CAR PARKING SPACE?	○ Yes ● No
PREVIOUS CANCELLATIONS	2
TOTAL GUESTS	5
TOTAL STAYS	2
LEAD TIME	2
TOTAL NUMBER OF SPECIAL REQUESTS	1
BOOKING CHANGES	1

Submit



Hotel Booking Prediction

The Chances of Cancellation is 0.0 %

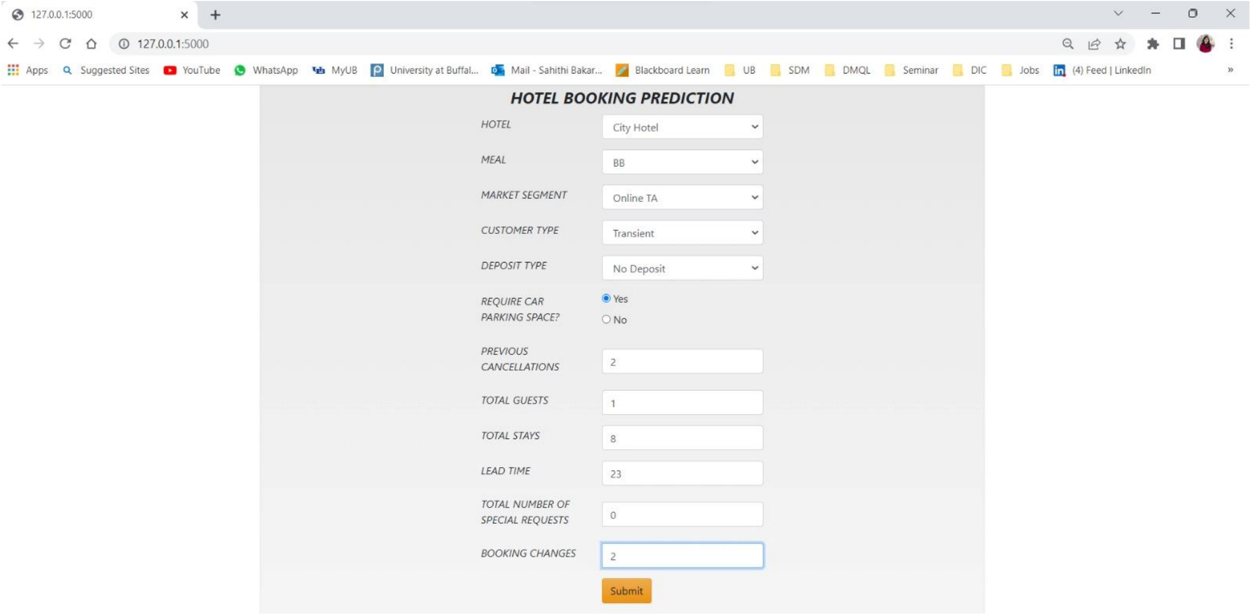
The Chances of Confirmed is 100.0 %

Hotel Booking is CONFIRMED

Hotel Booking Prediction as the status, Cancelled:

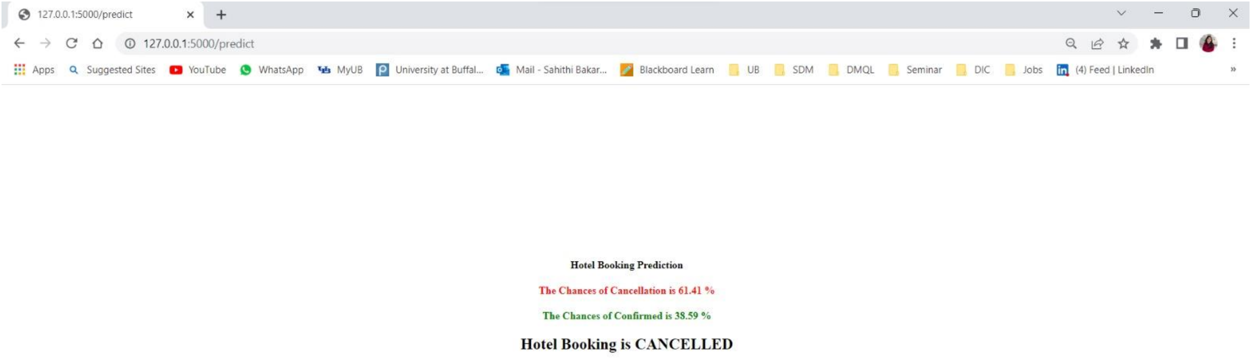
By giving the below inputs, the Hotel Booking Prediction is in the CANCELLED status. The chance of probability of Cancellation is higher which is 61.41%. And the chance of probability of Confirmation is 38.59%. So, it is more likely to be CANCELLED.

If we observe, the lead time value is 23.



HOTEL BOOKING PREDICTION

HOTEL	City Hotel
MEAL	BB
MARKET SEGMENT	Online TA
CUSTOMER TYPE	Transient
DEPOSIT TYPE	No Deposit
REQUIRE CAR PARKING SPACE?	● Yes ○ No
PREVIOUS CANCELLATIONS	2
TOTAL GUESTS	1
TOTAL STAYS	8
LEAD TIME	23
TOTAL NUMBER OF SPECIAL REQUESTS	0
BOOKING CHANGES	2



Hotel Booking Prediction

The Chances of Cancellation is 61.41 %

The Chances of Confirmed is 38.59 %

Hotel Booking is CANCELLED

10. Relationship between different features and the cancellation rate:

We considered top 12 features like lead time, total number of special requests, hotel, meal, market segment, customer type, deposit type, require car space, previous cancellations, total number of guests, total stays, booking changes.

- **Lead time with Cancellation Rate:**

  The likelihood of a booking being canceled is depends with the amount of lead time, the higher the lead time, the higher the cancellation rate, and the shorter the lead time, the less likely the booking will be canceled.

- **Deposit type with Cancellation Rate:**

This dataset includes three different types of deposits: Refundable, Non-Refundable, and No Deposit. No Deposit and Refundable have the lesser chance of cancellations. While, Non-Refundable has more chances of cancellations.

- **Previous Cancellation with Cancellation Rate:**

Booking that was cancelled previously has higher chance of cancelling. While, the booking that was not cancelled has lesser chance of cancelling. This indicates, the booking that was cancelled earlier is more likely to be cancceled again.

- **Parking Space with Cancellation Rate:**

Bookings that require parking space has no cancellations. While, the booking with no parking space are more likely to be cancelled.

- **Booking Changes with Cancellation Rate:**

Customers who made changes to their booking is more likely to be confirmed. Whereas the customers who did not make any booking changes has a probability to be cancelled.

- **Special Requests with Cancellation Rate:**

Customers who made special requests to their booking is more likely to be confirmed. Whereas the customers who did not make any special requests has a probability to be cancelled.

Users can solve the problems related to our problem statement like what elements affect to hotel cancellations? How different features affect the cancellation rate? and many more.

11. Recommendations that our data product gives:

- Restricting the Lead time to be not more than some specified days helps to improve the confirmed status of the booking. Because, booking that are made in advance has higher rate of cancellations.
- From the data analysis, we can observe that the booking made with Refundable and No Deposit booking are mostly cancelling. So, we can set the type of the deposit to be Non-Refundable, thereby, allowing users to be in the confirmed status to stay at the hotel.
- Allowing users not able to book hotels, if they have any previous cancellations will make the users to not cancel the booking ahead for their stay.
- If there are special requests or the meals preferred by the customers, then the hotel management must take care of it. By doing so, the cancellation rate can be minimized.
- Hotel management must incorporate new techniques like making the customers to provide rating and good review on the social networking sites and thereby providing 5%-10% of concession in the price of their booking. This will attract most of the customers and there may be increase in the number of bookings.
- If the online website is interactive, then it will attract the customers. Thereby, the cancellation rate can be minimized.

12. REFERENCES

- https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
- https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
- https://scikit-learn.org/stable/modules/tree.html
- https://towardsdatascience.com/deploy-a-machine-learning-model-using-flask-da580f84e60c