Analysis of Hotel Bookings & Cancellation Prediction

Data Intensive Computing 587 PHASE 3
Manikanta Kalyan Gokavarapu - mgokavar - 50465129
Rakesh Kumar Gavara - rgavara - 50483851

1. Problem Statement

The aim of this project is to detect patterns and tendencies in customer actions that can be utilized to improve revenue management, enhance resource allocation, optimize customer service, and shape marketing and promotional strategies. Through an analysis of historical booking and cancellation data, the **goal is to develop precise forecasts about future cancellations** and so that users can make necessary operational adjustments to meet customer requirements and maximize profitability.

2. Code

2.1 Fully documented Code:

We have attached fully documented code for phase 1, 2 and 3 under mgokavar_rgavara_phase_3.zip -> src -> (phase1_code, phase2_code and phase3_code).

2.2 Working instructions to demo/use our finished product:

Step - 1:

Extract the zip file "mgokavar_rgavara_phase_3.zip" and navigate to the code folder which will be here *'src -> phase3_code'* . Open phase3_code folder in a python IDE (Ex: Pycharm) and should contain the files as shown in the figure 1 below. Install all the required packages highlighted in the below figure 1 in your local IDE.

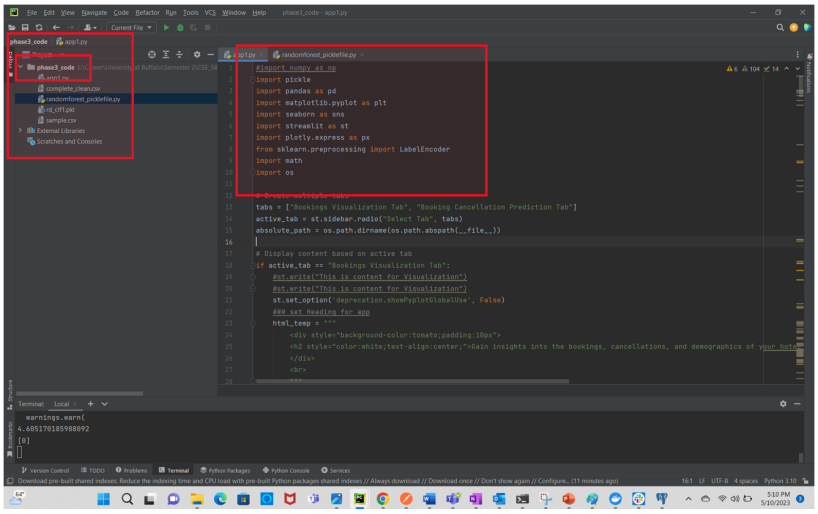

Figure [1]

Step - 2

Click on the ‘app1.py’ file on the left and click on the terminal icon as shown in Figure 2.

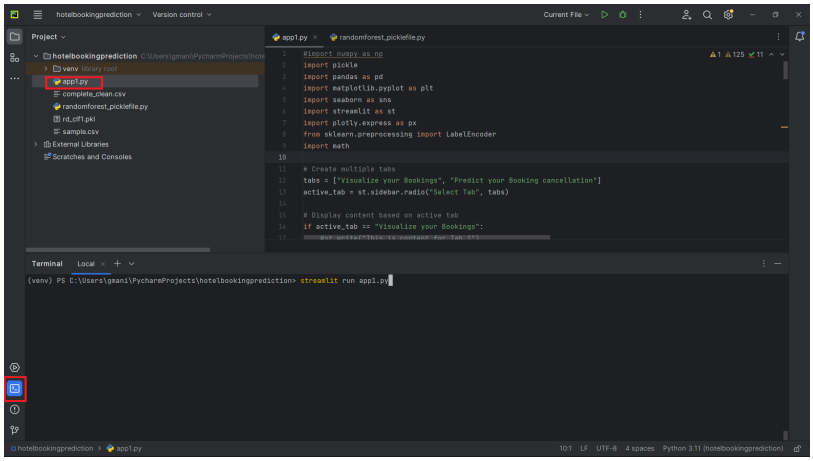

Figure [2]

Step - 3

Once the terminal is open type the command **“streamlit run app1.py”** and press enter as shown in Figure 3.

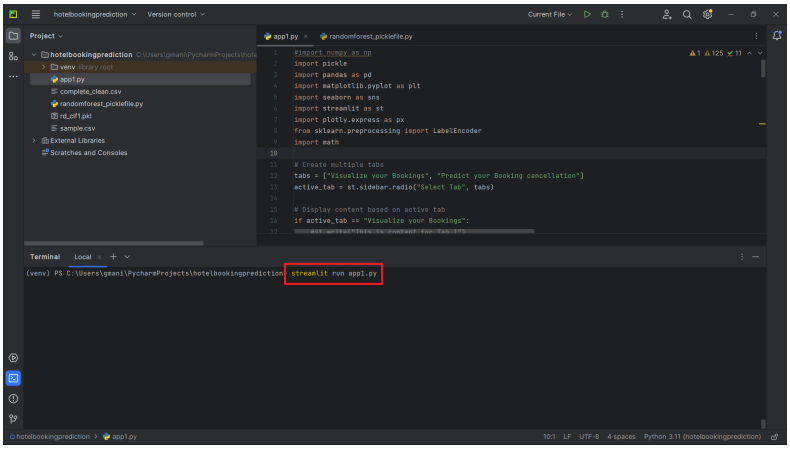

Figure [3]

Step - 4

Once you run the command, the UI will open on the local host ‘**http://localhost:8501/**’ as shown below and by default the 'Bookings visualization’ tab will be displayed. As shown in below Figure 4**.**

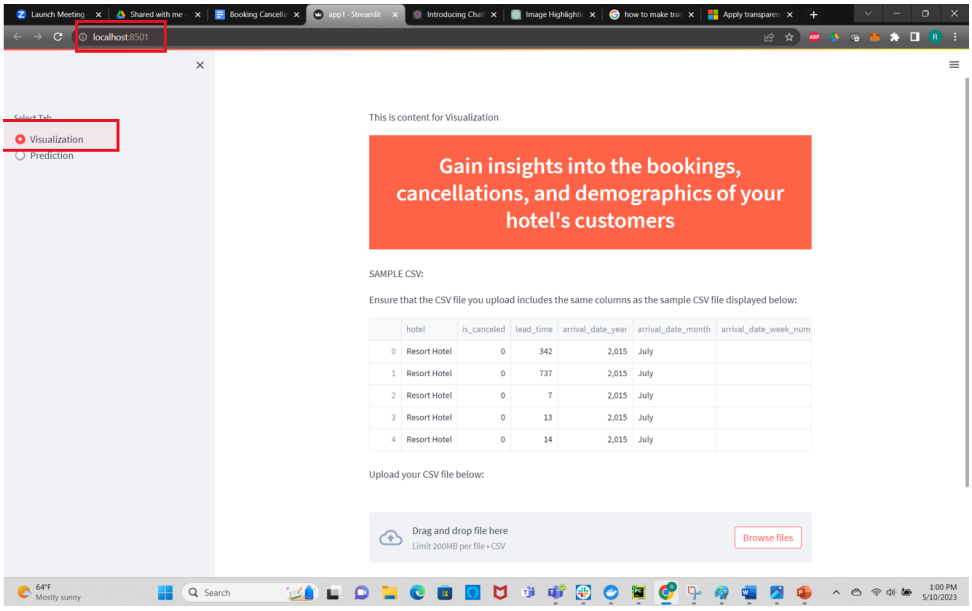

Figure [4]

Step - 5

To your right you can view the booking Visualization window where users can upload their sample dataset. Click on Browse Files button and upload the data file as shown in the below Figure 5.

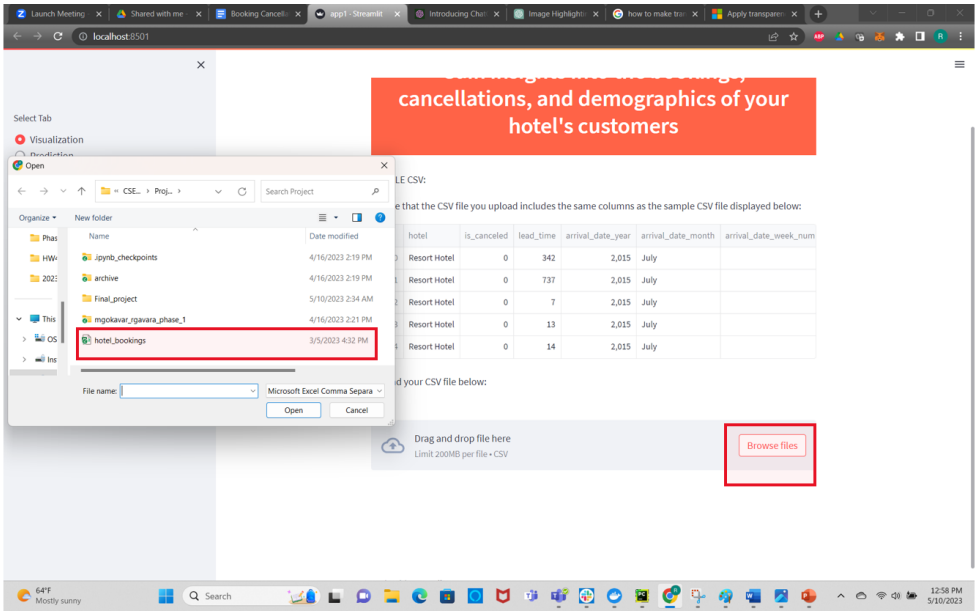

Figure [5]

Step - 6

After file upload users should be able to view the first 5 records of the input dataset as shown in the below Figure 6.

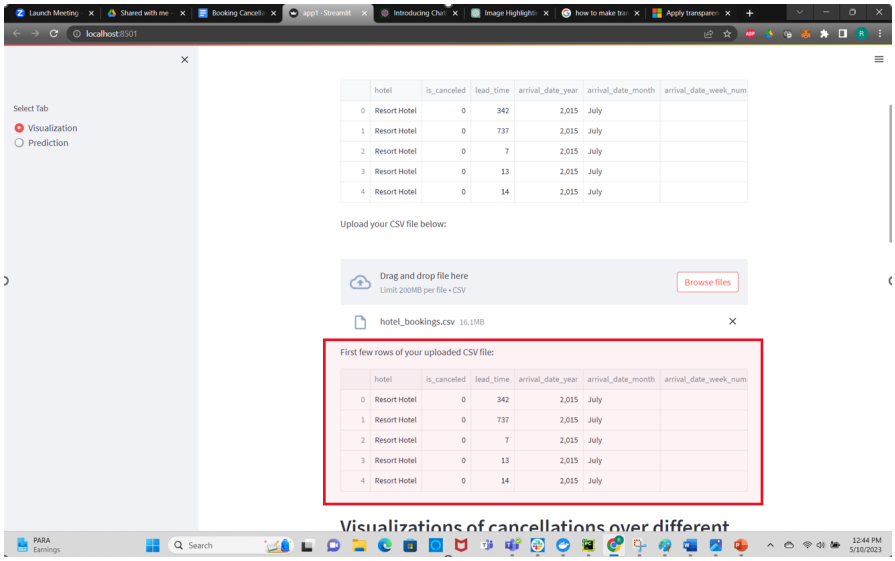

Figure [6]

Step - 7

If a user wants to view the Bar plot analysis(Cancellations count vs input parameters) he/she can select any input parameter from the drop down as shown in the below Figure 7.

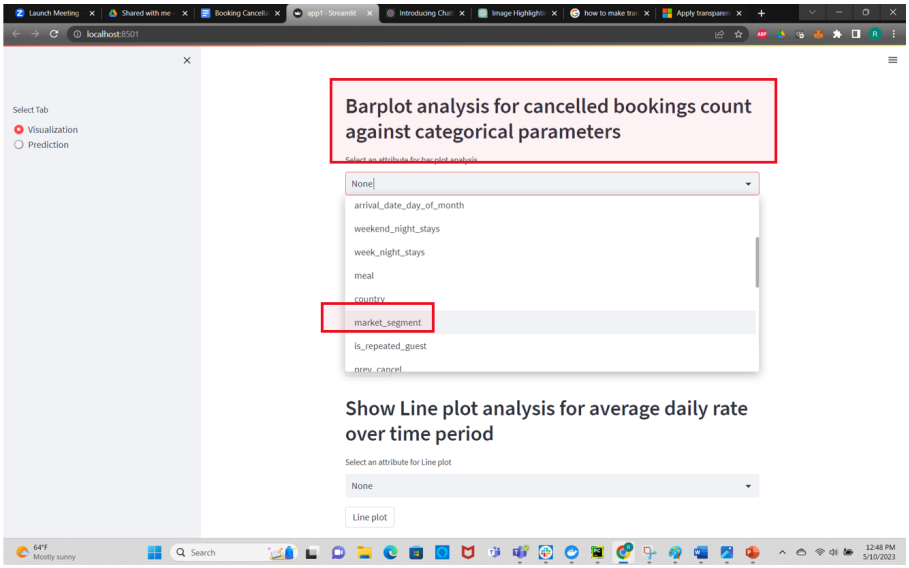

Figure [7]

Step - 8

Once the user clicks on the Bar plot a visualization will be displayed as shown in the below Figure 8. Similarly we have other visualization techniques such as Pie plot, Regression plot and so on.

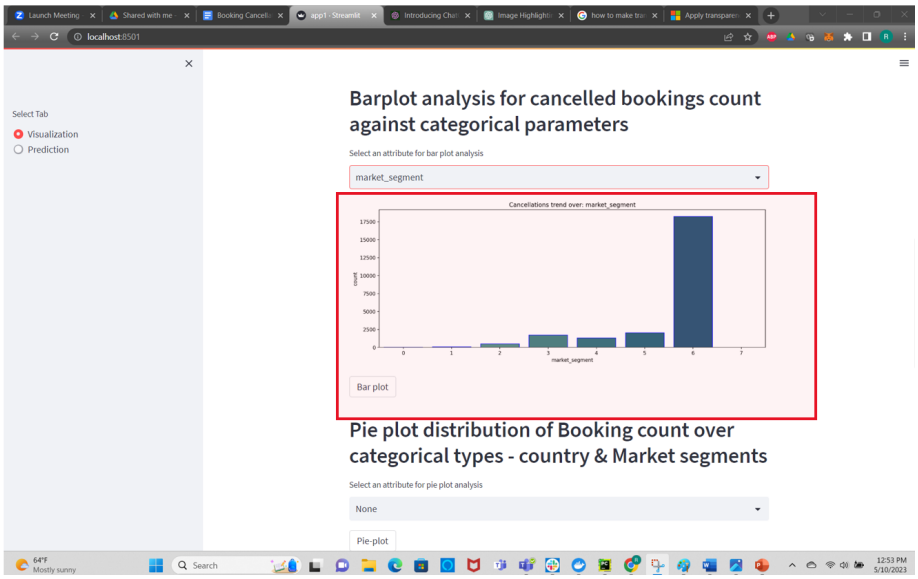

Figure [8]

Step - 9

Click on the Booking Cancellation Prediction tab on the left then the Booking Cancellation Prediction tab will open as shown in below Figure 9.

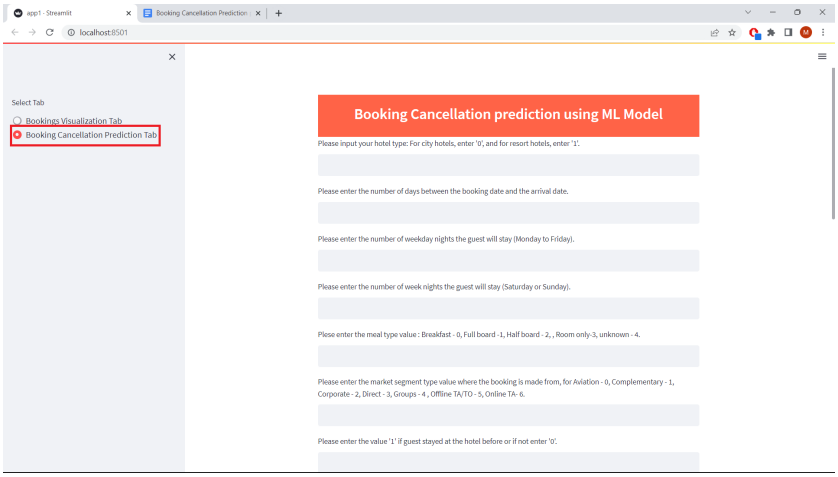

Figure [9]

Step - 10

Once the “Booking Cancellation Prediction Tab” opens up, users can enter their booking record based on the instructions shown in the GUI and At the end of the page they can click on ‘predict’ button to get the result as shown in below figure 10.

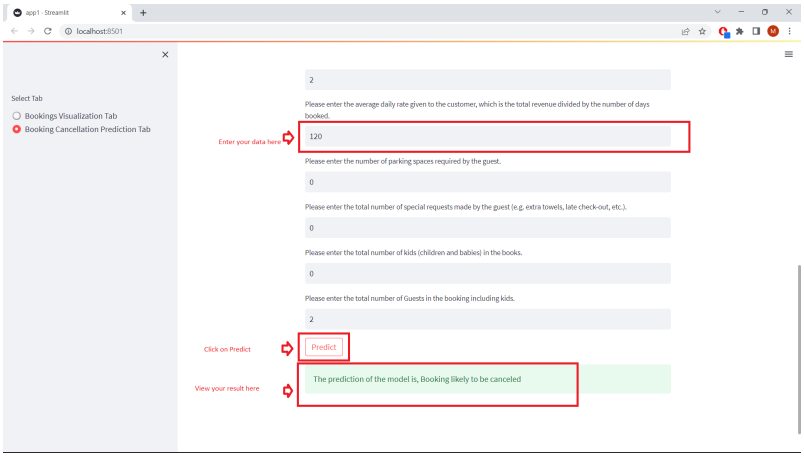

Figure [10]

2.3 Models From Phase 2

2.3.1 Models Analysis and Tuning Details :

We have implemented five models in Phase 2 they are K-Nearest Neighbors, Multinomial Naive Bayes, Random Forest, Logistic Regression, Support Vector Machine and for each model we have done needed tuning and found the best accuracies possible for our dataset as shown in below Figure [11].

Model Name	Accuracy (%)	Tuning used
K-Nearest Neighbors	77.88	Used Optimal K value to get the best results possible.
Multinomial Naive Bayes	75.10	Used hyper parameter tuning, set alpha = 1.0 which is used to avoid zero probabilities(laplace smoothing) and fit_prior = true so the model will automatically estimate

		the prior probabilities of each class from the training data.
Random Forest	80.85	Hyper parameter tuning using n_estimators value which is set 112 and min_sampe_split is set to 4. for best results.
Logistic Regression	77.72	Hyper parameter tuning using max_iter = '3000', solver = 'sag' and penalty = 'l2'
Support Vector Machine	80.71	Hyper parameter tuning in SVM is done using the 'gridSearchCV' method. SVM performance was best at hyperparameters: kernel = 'rbf', C = 1, gamma = 0.1

Figure[11]

2.3.2 Selected Model for phase 3 and its Hyper Parameter tuning details.

- We have selected Random Forest Algorithm for our final data product because out of the 5 algorithms implemented Random Forest achieved the highest accuracy of 81%. This is because the algorithm's result is achieved by majority voting of multiple minor models (decision trees) results. So, the effect of noisy data in the algorithm is very less. This resulted in getting good accuracy for the dataset used.
- The second reason for selecting this algorithm is because if we see the classification report in figure 12 we could notice the precision, recall, f1-score and support of both the classes i.e 0 - booking not canceled, 1- booking canceled. For the class 0 the model is performing well as precision, recall, f1-score are close to 1 and has high support with good number of instances for the class. For class 1 also the model performs well even when there are fewer instances.

```
Accuracy Score of Random Forest is : 0.8085376162299239
Confusion Matrix :
[[17210  1645]
 [ 3338  3833]]
Classification Report :
              precision    recall  f1-score   support

           0       0.84      0.91      0.87     18855
           1       0.70      0.53      0.61      7171

    accuracy                           0.81     26026
   macro avg       0.77      0.72      0.74     26026
weighted avg       0.80      0.81      0.80     26026
```

Figure [12]

- To tune the random forest model further we removed some redundant input features like arrival_year, adults, children, babies, reserved_room_type and assigned_room_type.
- Hyper parameter tuning in random forest is achieved by finding optimal n_estimators value to get the best results. Here n_estimators is the number of trees in the forest. In figure 13 we can see that at 112 trees we are getting the best accuracy for our algorithm and in the similar fashion tried different values of min sample splits which means samples required to split an internal node and found the optimal value which is 4 for my dataset. By using this hyper parameter tuning and removing redundant features, I have achieved a good accuracy of 81% for the model.

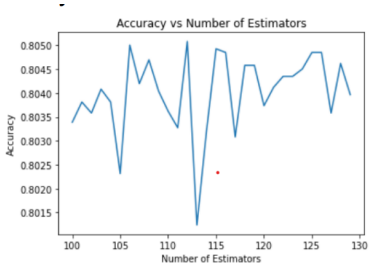

Figure [13]

2.4 Recommendations related to your problem statement based on your analysis

2.4.1 What can users Learn from the product ?

The product contains two tabs**.**

1. **Bookings Visualization tab: (users - hotel chains)**

   a. Visualization using Bar plot - With this visualization Hotel owners will be able to view the cancellations count value against various categorical values such as market_segment, country and so on in the form of bar graphs as below.

For example: With the below bar plot Hotel owners can identify which market segment got the highest ,least cancellation count.

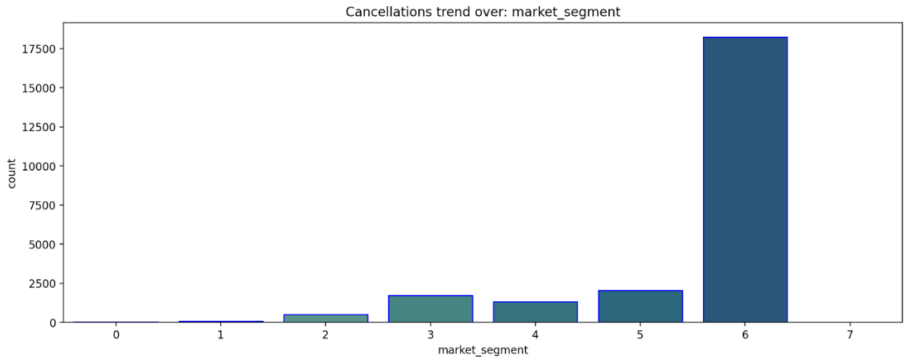

Figure [14]

b.Visualization using pie chart - With this visualization Hotel owners will be able to view a quantified percentage of bookings against two categorical parameters such as country and customer type.

For example : With the below pie chart Hotel owners can visualize the bookings distribution in terms of percentage among different countries such as France, Italy, USA and so on.

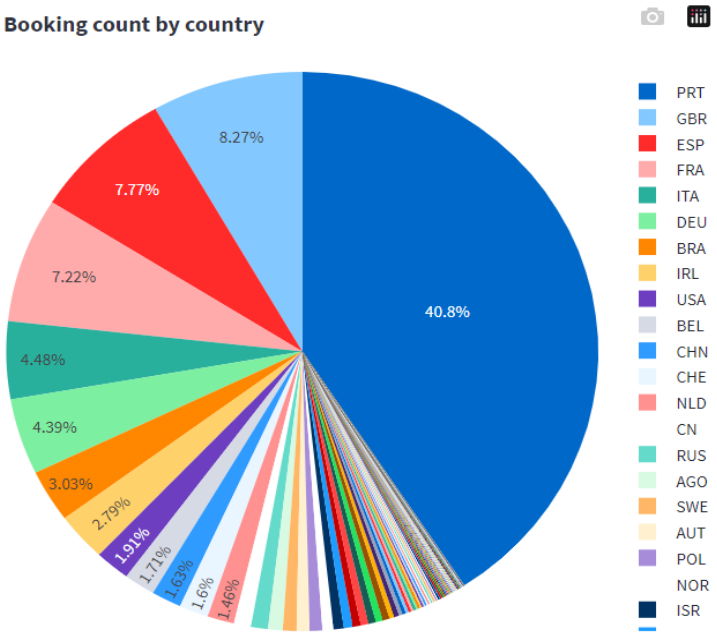

Figure [15]

c.Visualization using Line plot - With this visualization Hotel owners will be able to view the trend of average daily rate of hotel rentals against time period.

For example: With the below line plot analysis Hotel owners can visualize the average daily rate over time(months, days or other time period metrics) and identify steep and stable rate of change of rates over time.

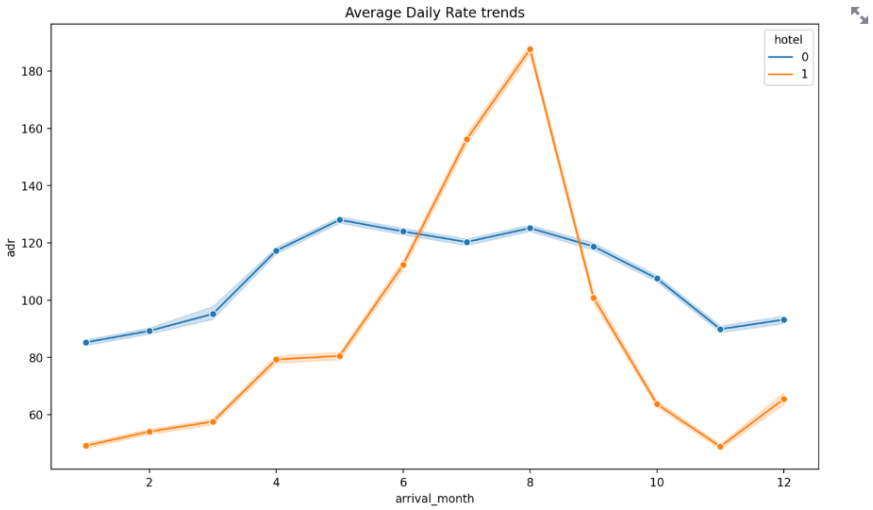

Figure [16]

d.Visualization using Violin plot - With this visualization Hotel owners will be able to view the trend of cancellations over meal types. This visualization is extremely helpful when contrasting the distributions of various groups or classifications because it makes it simple to identify variations in central tendency, variation, and skewness and violin plot also helps in revealing patterns and trends in data.

For example - With the below Violin plot analysis Hotel owners can easily identify with meal plan 0 the cancellations are higher for hotel zero.

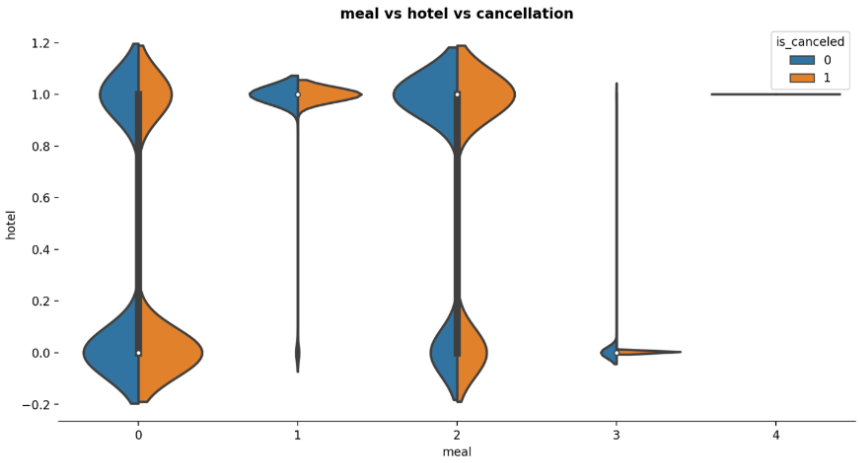

Figure [17]

e.Visualization using Cat plot - this is similar to line plot but with one big distinction. With this visualization we can visualize categorical data whereas in line plot we can visualize continuous data.

For example - With the below Cat plot analysis Hotel owners can observe that the customers who belong to the 'Transient' category do more cancellations compared to others.

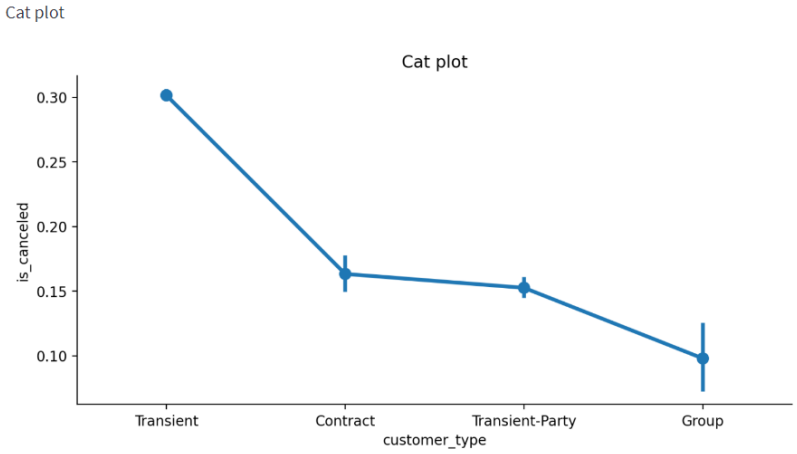

Figure [18]

f.Visualization using Regression plot - Regression plots can help identify patterns in the relationship between two variables, such as whether the relationship is linear or nonlinear. This can be particularly useful for identifying complex relationships that may not be immediately apparent from looking at the raw data.

For example - With the below Regression plot analysis Hotel owners can observe that the customers who have booked for more nights are most likely to cancel their bookings.

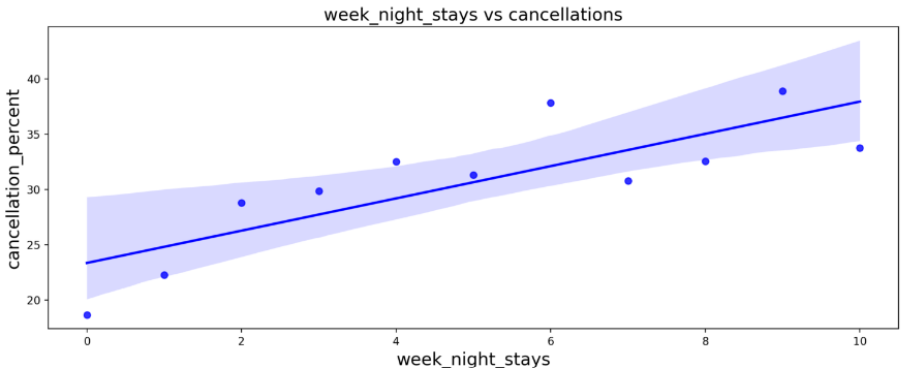

Figure [19]

2. **Booking Cancellation prediction tab:**
    - The Booking Cancellation prediction app can be utilized by users like individual hotel owners to forecast whether their customers will fulfill their booking or not.
    - The Booking Cancellation prediction tab both the results/outputs are shown in below Figure 20 and Figure 21.

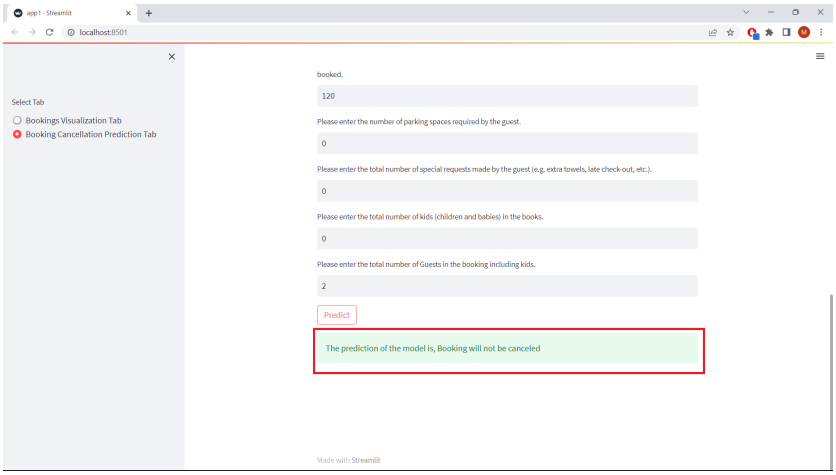

Figure [20].

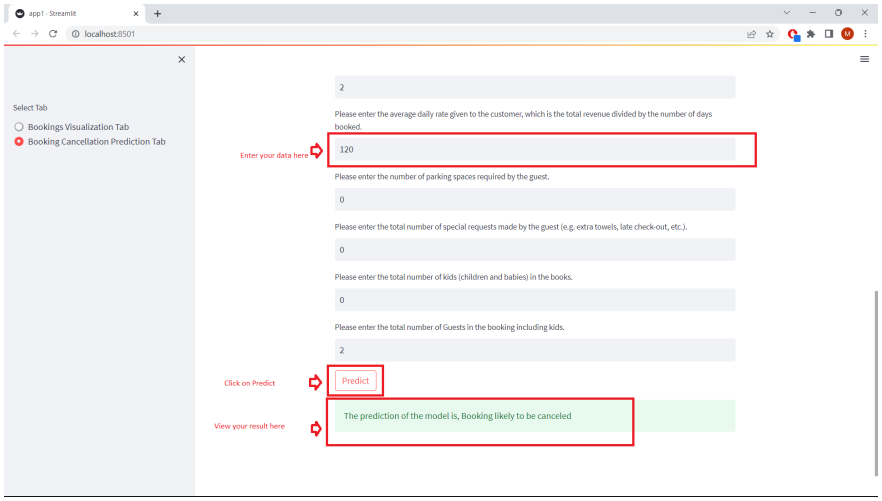

Figure [21].

Note: In general individual hotel owners don't have huge amounts of bookings data. So the models will not perform efficiently if we accept custom datasets from users and also our model is already trained using a large amount of input data collected from different hotels in 135 different countries. and the model is tuned to achieve high accuracy. So, we allow the user to give their booking data and we predict whether the booking will be fulfilled or not.

2.4.2 How does it help them solve problems related to your problem statement ?

1. **Uses of Bookings Visualization tab:**

- Smart investment: With the above analysis techniques such as Pie plot analysis the hotel owners will get an overview of some of the trends such as countries with maximum bookings, least cancellations and so on. This information will be very useful for hotel owners who are in the dilemma of expanding their franchise in specified regions.
- Idle pricing: Most of the cancellations happen due to price fluctuations, competition in terms of prices of rival hotels. With our visualization like the one we did line plot analysis Hotel owners can keep the prices less dynamic and more feasible to customers.

- Quality food: Serving good food is one of the primary pillars for achieving success for any hotel. Combination of quality food with variety is very crucial, with our visualization we did using Violin plot hotel owners can identify which meal plan is most liked by customers and accordingly follow the same.
- Know your customers: Most of the hotels fail because they do not understand their customers and their needs. With the visualization such as the one we did for cat plot hotel owners can identify which set of categories of customers do most cancellations and improvise based on the majority requirements.

2. **Uses of Booking Cancellation prediction tab:**

- Reducing Income loss: The application can help hotel owners to minimize revenue loss by predicting which bookings are likely to be canceled. Then, they can take the necessary actions to lessen the likelihood of cancellations or make plans for alternate sources of income.
- Resource Management: Effective resource management can assist hotel operators in determining which reservations are most likely to be canceled. For instance, if a reservation is anticipated to be canceled, the hotel can free up the resources set aside for that reservation.
- Customer service: Better customer service can be provided by hotel owners with the aid of the application. Hotel owners can contact clients and provide incentives to keep their reservations by estimating which bookings are most likely to be canceled. They can also recommend alternate booking choices.
- Operational planning: This application can assist hotel operators in making more effective plans for their operations. For instance, the hotel can direct resources to other areas if it anticipates a cancellation to keep activities running smoothly.

2.4.3 Ideas on how to extend our project further (or) other venues that can be explored related to the problem.

- **Integrate new data sources:** We can look into integrating the Booking Cancellation prediction application with a variety of data sources, such as flight schedules, weather information, or regional events. It may increase the predictability of the prediction model and offer more insightful information to hotel managers and owners by examining the effect of these external factors on cancellations of reservations.

- **Creating recommendation engine:** Starting with the Booking Cancellation prediction app, you can create a recommendation engine that gives hotel owners and managers advice on how to optimize their inventory and pricing strategies in light of anticipated cancellations and no-shows.
- **Usage in other venues:** We can expand the model's scope to make it more useful for other hospitality industry stakeholders like travel agencies, booking platforms, or event planners. The Booking Cancellation prediction app is useful for individual hotel owners, but we can look into ways to expand the model's scope.
- **Extending the Web application:** We can develop a dashboard or automatic alerting tool that provides hotel owners and managers with real-time insights into occupancy, cancellations, and revenue. This can help them make better decisions and respond more quickly to changing market conditions and also the web application can be enhanced and can be integrated with internal hotel portals used by owners.
- **Sentiment analysis:** Use natural language processing techniques to analyze customer reviews and feedback to gain insights into customer sentiment. This analysis can help businesses understand the reasons for cancellations and make improvements to their services.
- **Fraud detection:** Use machine learning algorithms to detect fraudulent bookings and reduce the number of cancellations caused by fraudulent activities.

References:

- Data intensive Computing Lecture Slides
- https://practice.geeksforgeeks.org/courses/data-science-live?utm_source=GfG&utm_medium=gfg_submenu&utm_campaign=DS_Submenu
- https://www.analyticsvidhya.com/
- https://realpython.com/pandas-plot-python/#analyze-categorical-dat
- https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb
- https://www.kaggle.com/code/gautigadu091/categorical-naive-bayes-from-scratch-in-python/notebook
- Devanshi Srivastava, Programs Buzz, Decision Tree: Introduction, 06/04/2021
- Chengyou Chen, Ensemble Learning, Lecture Slides, 10/11/2022.
- https://medium.com/@myselfaman12345/c-and-gamma-in-svm-e6cee48626be
- https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete guide-for-beginners
- https://www.datacamp.com/tutorial/understanding-logistic-regression-python 2.
- https://www.statology.org/plot-logistic-regression-in-python/

- **Data source link:** https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hot els.csv
- Streamlit documentation - https://docs.streamlit.io/