

## DSC 630 Milestone 3: Data Selection and Project Proposal

Chitramoy Mukherjee

Predictive analysis on Hotel Booking Cancellation using

hotel\_booking.csv dataset

## **Will I be able to answer the questions I want to answer with the data I have?**

After reviewing columns and rows in the dataset to ensure it contains relevant information such as booking details, guest demographics, room types, reservation status booking dates, guest characteristics, and booking outcomes will be able to create the predictive model to answer the questions we are looking to answer for this project. By carefully evaluating Data Content, Question Suitability, Data Quality and Predictive Power factors we can determine "hotel\_bookings.csv" dataset can support your predictive analysis and help answer your desired questions about hotel bookings. If the dataset lacks certain information or questions require additional data, we may need to explore other sources or adjust your analysis approach accordingly.

## **What visualizations are especially useful for explaining my data?**

Visualizations play a crucial role in exploring and explaining the patterns and insights within hotel booking dataset. Below visualizations you can create to understand and communicate your data effectively:

### **1. Bar Charts:**

- Visualize categorical variables such as hotel type (city or resort), customer type (transient, contract, group), and room type.
- Compare counts or percentages across different categories to identify trends and preferences.

### **2. Histograms:**

- Explore the distribution of numerical variables such as lead time (number of days between booking and arrival), stays on weekend nights, stays on week nights, etc.
- Understand the frequency and spread of values within each variable.

### 3. **Pie Charts:**

- Show the distribution of categorical variables like customer types, market segments, or distribution channels. Highlight the proportion of each category within the dataset.

### 4. **Box Plots:**

- Visualize the distribution of numerical variables such as lead time, booking duration, or room rate, across different categories like hotel type or customer type. Identify outliers, quartiles, and overall distribution characteristics.

### 5. **Time Series Plots:**

- Analyze temporal patterns by plotting variables like booking date, arrival date, or lead time over time. Identify seasonal trends, booking peaks, or patterns in cancellations.

### 6. **Scatter Plots:**

- Explore relationships between numerical variables, such as lead time and booking duration, or between numerical and categorical variables. Identify correlations or patterns in the data.

### 7. **Heatmaps:**

- Visualize correlations between variables in the dataset, especially useful for understanding relationships between numerical variables. Identify areas of high or low correlation, which can inform feature selection for predictive modeling.

## **Do I need to adjust the data and/or driving questions?**

Adjusting the data and refining driving questions may be necessary to ensure the success and relevance of your predictive analysis. Need to perform some basic adjustments in the dataset to Check for missing values, duplicates, and inconsistencies in the dataset. Clean the data by filling missing values, removing duplicates, and resolving inconsistencies to ensure data quality. Create new features from existing ones if they can provide additional predictive power. I can

derive features like total stays (weekend nights + weeknights), booking lead time, or booking season from the available data. Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding. This allows to include categorical variables in predictive models effectively.

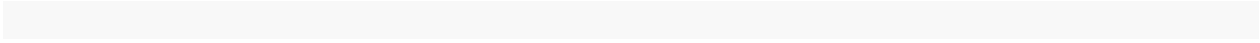
### **Do I need to adjust my model/evaluation choices?**

Will Consider different types of predictive models suitable for the dataset and objectives. Common models for hotel booking prediction include logistic regression, decision trees, random forests, gradient boosting machines (GBM), support vector machines (SVM), and neural networks. Balance between model complexity and interpretability based on our needs and priorities. While complex models like neural networks may offer high predictive performance, simpler models like logistic regression or decision trees might be easier to interpret and explain. Will explore ensemble methods like random forests or gradient boosting to combine predictions from multiple models. Ensemble methods often improve predictive performance and robustness compared to individual models. Will evaluate metrics that align with the goals of the predictive analysis. Common metrics for binary classification tasks in hotel booking prediction include accuracy, precision, recall, F1 score, ROC-AUC, and precision-recall curve. Will Evaluate model performance in the context of business relevance. Consider how well the predictive model aligns with the ultimate business goals and whether the predictions are actionable and useful for decision-making.

### **Are my original expectations still reasonable?**

Our initial expectation is to build a predictive model using the hotel\_booking.csv data to determine whether a hotel booking would be canceled, which is crucial for hotels as

cancellations affect revenue and operational planning. After reviewing the dataset and attributes and volume of data available to build the model, expectation seems to be reasonable. Will be able to answer the questions from the available data and attributes. If I need some reference data along with this data to answer my questions, I will source it during my EDA and model development part.



## DSC 630 Milestone 2: Data Selection and Project Proposal

Chitramoy Mukherjee

### Predictive analysis on Hotel Booking Cancellation using

hotel\_booking.csv dataset

Predictive analysis in hotel booking cancellation involves using historical data and statistical techniques to forecast the likelihood of a guest canceling their reservation before their scheduled arrival date. In this project, we aim to build a predictive model to determine whether a hotel booking would be canceled, which is crucial for hotels as cancellations affect revenue and operational planning. Dataset contains many features related to booking, such as lead time, deposit type, and special requests, which adds to the model's complexity. This prediction model will help the business to determine the probability of cancellation of a booking and create a backup plan to overcome any loss. From a customer perspective we can analyze what will be the best time to book a hotel and get the best pricing for it.

***Dataset information and steps:***

Hotel\_booking.csv dataset contains 2015-2017 timeframe data for City hotel and Resort hotel booking. Below are the variables and its description from hotel\_booking.csv dataset.

Index	Variable	Description
1	hotel	Type of hotel (Resort Hotel, City Hotel)
2	is_canceled	Reservation cancellation status (0 = not canceled, 1 = canceled)
3	lead_time	Number of days between booking and arrival
4	arrival_date_year	Year of arrival
5	arrival_date_month	Month of arrival
6	arrival_date_week_number	Week number of the year for arrival
7	arrival_date_day_of_month	Day of the month of arrival
8	stays_in_weekend_nights	Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9	stays_in_week_nights	Number of weeknights the guest stayed or booked
10	adults	Number of adults
11	children	Number of children
12	babies	Number of babies
13	meal	Type of meal booked (BB, FB, HB, SC, Undefined)
14	country	Country of origin of the guest
15	market_segment	Market segment designation
16	distribution_channel	Booking distribution channel
17	is_repeated_guest	If the guest is a repeat customer (0 = not repeated, 1 = repeated)
18	previous_cancellations	Number of previous bookings that were canceled by the customer
19	previous_bookings_not_canceled	Number of previous bookings that were not canceled by the customer
20	reserved_room_type	Type of reserved room
21	assigned_room_type	Type of assigned room
22	booking_changes	Number of changes made to the booking
23	deposit_type	Type of deposit made (No Deposit, Refundable, Non Refund)
24	agent	ID of the travel agent responsible for the booking
25	company	ID of the company responsible for the booking
26	days_in_waiting_list	Number of days the booking was in the waiting list
27	customer_type	Type of customer (Transient, Contract, Transient-Party, Group)
28	adr	Average Daily Rate
29	required_car_parking_spaces	Number of car parking spaces required
30	total_of_special_requests	Number of special requests made
31	reservation_status	Last reservation status (Check-Out, Canceled, No-Show)
32	reservation_status_date	Date of the last reservation status

33	name	Guest's name
34	email	Guest's email address
35	phone-number	Guest's phone number
36	credit_card	Last four digits of the guest's credit card

Below are the key steps involved for project execution,

**Historical Data:** Historical booking data serves as the foundation for predictive analysis. It includes information on past reservations, cancellations, guest demographics, booking channels, and other relevant variables. Data needs to be cleaned up and handled and loaded for further processing.

**Feature Selection:** Identifying relevant features or variables that influence cancellation behavior is crucial. These may include lead time, booking channel, seasonality, room type, price, guest demographics, and external factors such as events or holidays.

**Model Development:** Predictive models, such as logistic regression, decision trees, random forests, or neural networks, are trained on historical data to predict the likelihood of cancellation for future bookings. These models learn from patterns and relationships in the data to make predictions.

**Model Evaluation:** Models are evaluated using metrics such as accuracy, precision, recall, or area under the ROC curve (AUC) to assess their predictive performance. Cross-validation techniques help validate the model's generalizability.

**Implementation and Deployment:** Once a predictive model is developed and validated, it can be deployed into hotel management systems to generate real-time predictions for upcoming bookings. These predictions inform decision-making processes related to pricing, inventory management, and resource allocation.



### ***Significance:***

Predictive analysis helps hotels optimize pricing strategies by adjusting room rates dynamically based on predicted cancellation probabilities and demand fluctuations. Anticipating cancellations allows hotels to better manage inventory, allocate resources efficiently, and minimize the impact on operations. By proactively managing cancellations, hotels can minimize overbooking situations and ensure a smoother guest experience, ultimately leading to higher guest satisfaction and loyalty. Predictive insights enable hotels to make informed decisions regarding marketing campaigns, promotions, and capacity planning, thereby maximizing revenue potential and competitiveness in the market.

### ***Types of Models:***

For this project, I plan to utilize several machine learning models to predict the chances of cancellation of hotel booking. Will implement and tune classification models including Decision Trees, Random Forest, and XGBoost. Will Emphasize achieving high F1-score for class 1, ensuring comprehensive identification of booking cancellations.

### ***Evaluation of Results:***

Will select evaluation metrics suitable for the binary classification problem of cancellation prediction. Common metrics includes below,

- Accuracy: Measures overall correctness of predictions.
- Precision: Indicates the proportion of predicted cancellations.
- Recall: Measures the proportion of actual cancellations that are correctly predicted.
- F1 Score: Harmonic means of precision and recall, useful for imbalanced datasets.

- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Measures the model's ability to discriminate between positive and negative classes.

### ***Learning Objectives:***

Through this project, I aim to gain a deeper understanding of predictive analytics techniques and their application. Identify trends and factors influencing booking cancellations to anticipate future cancellations more accurately. Gain insights into optimizing revenue management and resource allocation strategies based on predicted cancellation probabilities. Understand guest behavior and preferences regarding cancellations to enhance guest satisfaction and improve overall hotel operations.

### ***Risks and Ethical Implications:***

One of the major risks for utilizing guest data for predictive analysis raises privacy risks, necessitating strict adherence to data protection regulations and ensuring secure handling of sensitive information to prevent unauthorized access or misuse. There's a risk of algorithmic bias leading to unfair treatment of certain groups, such as discriminating against guests from specific demographics or regions. Ethical considerations demand the identification and mitigation of biases to ensure fair and equitable predictions. Lack of transparency in the predictive modeling process can erode trust and lead to unintended consequences. Ensuring transparency, accountability, and clear communication about the model's limitations and implications is essential to maintain ethical standards and stakeholder trust.

### ***Contingency Plan:***

If the original project plan does not work out as expected, I will reassess the data sources and modeling techniques to identify alternative approaches. This may involve exploring additional

datasets, adjusting preprocessing steps, or experimenting with different machine learning algorithms. Regular communication with project advisors and peers will help in troubleshooting and adapting to any challenges encountered during the project. If I must change the topic if the results are not expected, I will work on the breast cancer prediction modeling. I have checked the dataset on that which I can use for this project.

### ***Additional Considerations:***

In addition to developing predictive models, I plan to explore feature importance to understand which variables contribute most to readmission risk. Furthermore, I will visualize model predictions and insights to facilitate interpretation and decision-making by healthcare professionals. Continuous refinement of the predictive models based on real-world feedback and new data will be essential for improving their accuracy and applicability in clinical settings.

### ***References:***

Dataset hotel\_booking.csv is sourced from <https://www.kaggle.com>.