# PROJECT SUMMARY

| | |
|---|---|
| Batch details | PGPDSE CHN Jan23 (Group – 2) |
| Team members | 1. Praveen Arokia Raj A<br>2. G.Raghav<br>3. Sathish J<br>4. Harisri Karthikeyan<br>5. Balachandar P<br>6. Mohana Devan R |
| Domain of Project | Tourism Industry |
| Proposed project title | Hotel Reservation Cancellation Prediction |
| Group Number | Group 2 |
| Team Leader | Balachandar P |
| Mentor Name | Mr Jayveer Nanda |

Date: 30-04-2023

Signature of the Mentor

Jayveer Nanda

Signature of the Team Leader

# TABLE OF CONTENTS

# PROBLEM STATEMENT

"Develop a classification model to predict the booking status (canceled or not canceled) based on the provided booking information. The model should use features such as the number of adults and children, length of stay, meal plan, parking requirements, room type, lead time, and previous booking and cancellation history to accurately classify whether a booking will be canceled or not. The objective is to develop a reliable model that can help hotel management to identify potential cancellations in advance and take appropriate measures to reduce the impact on their business.


The model's performance should be evaluated based on metrics such as accuracy, precision, recall, and F1 score. The model should be able to provide actionable insights to the hotel management team for better decision-making and business planning."

# OVERVIEW

## Business problem statement (GOALS)

### 1. Business Problem Understanding:

The hotel booking reservation cancellation prediction can help businesses identify the most important factors that affect hotel reservations and cancellations. It can address key questions such as the most important factors that influence guests' booking decisions and common patterns in cancellations. These insights can be used to optimize booking processes, reduce cancellations, and increase revenue. By analysing the data, businesses can make data-driven decisions to improve their operations and profitability, including marketing strategies, pricing policies, inventory management, and customer service initiatives.


### 2. Business Objective

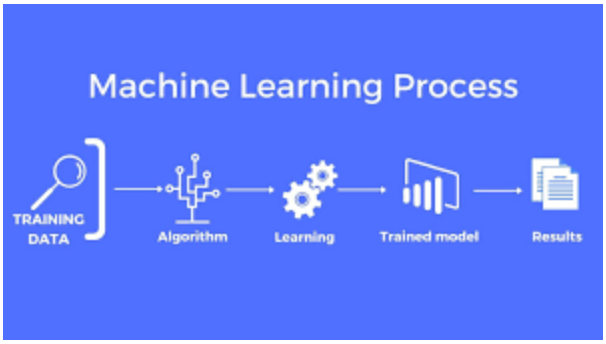
The business objectives for the Hotel Reservations Classification Dataset include maximizing revenue, improving customer satisfaction, reducing cancellations, and streamlining operations. By analyzing the data, hotels can identify trends and patterns that can inform pricing and revenue

management strategies, address negative guest experiences, predict and prevent cancellations, and streamline operations to enhance efficiency and profitability. The ultimate goal is to improve the overall guest experience while optimizing hotel operations for maximum success.

## 3. Approach

a. Data preprocessing: Before any analysis can be done, the dataset needs to be preprocessed. This involves handling missing values, outliers, and categorical variables, as well as standardizing or normalizing numerical variables
b. Descriptive Analysis: This involves analyzing the distribution and summary statistics of the variables in the dataset. This can help identify any issues with the data, such as outliers or data skewness.
c. Inferential Analysis: This involves making inferences about the population based on the sample data. This can help answer questions like, "What is the average number of nights stayed by guests in the hotel?" or "What percentage of bookings is cancelled?"
d. Feature Engineering: This involves creating new features or variables that may improve the performance of predictive models. For example, creating a variable that combines the number of adults and children may be more informative than using those variables separately.
e. Encoding Categorical Variables: This involves converting categorical variables, such as 'market_segment_type' or 'type_of_meal_plan', into numerical variables that can be used in machine learning models. This can be done using methods like one-hot encoding or label encoding.
f. Missing Values and Outliers: This involves handling missing values or outliers in the data. Missing values can be imputed using methods like mean imputation or regression imputation. Outliers can be identified using methods like Z-score analysis or box plots, and can be treated by either removing them or replacing them with more appropriate values.
g. Predictive Modeling: This involves using machine learning algorithms to build predictive models that can forecast future outcomes, such as predicting the likelihood of a booking being cancelled. Different algorithms can be used, such as decision trees, logistic regression, or neural networks, and the performance of the models can be evaluated using metrics like accuracy, precision, recall, and F1-score.
h. Model Evaluation and Selection: This involves evaluating the performance of different predictive models and selecting the best-performing one. This can be done using techniques like k-fold cross-validation or train-test splitting, and evaluating metrics like AUC-ROC or confusion matrix

i. Business Recommendations: Based on the insights gained from the data analysis and modelling, business recommendations can be made to improve hotel operations and profitability. For example, if the analysis shows that cancellations are more likely for bookings with a longer lead time, the hotel may decide to offer discounts to encourage earlier bookings or implement a more flexible cancellation policy.




There are different possible approaches to analyze the Hotel Reservations Classification Dataset, including Exploratory Data Analysis to understand the relationships between variables, Feature Engineering to create new features for predictive modeling, Predictive Modeling to use algorithms like decision trees to predict booking cancellations or other outcomes, Model Evaluation and Selection to determine the best-performing model, and Business Recommendations to make data-driven decisions that improve hotel operations and profitability

## 4. Conclusions

The dataset contains potential insights that can be used to improve hotel operations. Seasonal demand patterns can be identified by analyzing arrival year, month, and date variables. Guest preferences can be identified by analyzing variables such as room type reserved, meal plans, and special requests. High-risk guests can be identified using variables related to previous cancellations and bookings. Revenue management opportunities can be found by analyzing the average price per room. Finally, the lead time variable can be used to improve operational efficiency by identifying booking trends and streamlining the check-in process.

With this information, hotels can make data-driven decisions to enhance operational efficiency and customer satisfaction. For instance, hotels can use seasonal demand patterns to inform staffing levels and pricing strategies. They can personalize the guest experience by offering amenities that align with guest preferences. Hotels can use insights on high-risk guests to implement proactive measures to mitigate cancellations. Analyzing average room prices can inform revenue management strategies and aid in determining room rates. By utilizing the lead time variable, hotels can predict and prepare for future booking trends and improve the check-in process.

Overall, the Hotel Reservations Classification Dataset provides a wealth of insights that can inform business strategies for hotels to enhance customer satisfaction, reduce cancellations, and maximize revenue.

# TOPIC SURVEY IN BRIEF

## 1. Problem understanding

This dataset contains a wealth of information about hotel reservations, including details about guests, booking information, and the status of the reservation. The problem understanding for this dataset is to identify patterns and trends in guest behavior and preferences that can help hotels optimize their business processes and improve the overall guest experience. By analyzing variables such as lead time, room type reserved, and market segment type, hotels can identify opportunities to personalize the guest experience, improve operational efficiency, and optimize pricing strategies.

Some specific objectives that can be addressed using this dataset include identifying factors that influence the likelihood of a booking being canceled, identifying the most popular room types and meal plans, and identifying the most common reasons for guest requests. Additionally, the dataset can be used to identify patterns in guest behavior, such as the frequency of weekend versus weeknight bookings and the average number of nights guests stay at the hotel. These insights can help hotels to better understand their guests and tailor their services to meet their needs.

Overall, the problem understanding for this dataset is to gain insights into the booking patterns and preferences of hotel guests, with the ultimate goal of improving business performance, guest satisfaction, and revenue growth.

## 2. Current solution to the problem

The current solutions to the problem of hotel booking cancellations involve implementing stricter cancellation policies or requiring deposits for bookings, as well as analyzing data to identify patterns and contributing factors to cancellations. However, these solutions may not be effective in the long term, and hotels need a more data-driven approach to predict and prevent cancellations. By relying solely on stricter policies or deposits, hotels risk losing potential customers who may opt for more flexible options. On the other hand, data analysis can help hotels understand the underlying reasons for cancellations and take corrective actions accordingly. For example, if the analysis shows that customers frequently cancel due to dissatisfaction with room amenities or customer service, the hotel can make necessary improvements to address these issues. Moreover, a more accurate and data-driven approach, such as predictive modeling using machine learning algorithms, can help hotels forecast cancellations with greater accuracy and identify the key drivers of cancellations, thus enabling them to take proactive measures to prevent them.

### 3. Proposed solution to the problem

The proposed solution for the hotel booking cancellation prediction problem is to build a predictive model using machine learning algorithms, which can be trained on historical data to predict the likelihood of a booking being canceled and identify the factors contributing to cancellations. This solution involves data preparation, feature selection and engineering, model selection and training, model evaluation and optimization, and deployment and monitoring. The model's performance is then evaluated, optimized, and deployed in a production environment to predict booking cancellations. The use of predictive modeling can provide hotels with a more accurate and effective methods for predicting cancellations, which can help personalize the guest experience and improve customer satisfaction.

# CRITICAL ASSESSMENT OF TOPIC SURVEY

### *1. Find the key area, gaps identified in the topic survey where the project can add value to the customers and business*

The project can add value to customers and the business by personalizing the guest experience, optimizing revenue management, analyzing seasonal demand, and improving operational efficiency. Potential gaps that could be addressed include predicting booking cancellations, identifying high-risk guests, and optimizing marketing strategies. By leveraging machine learning algorithms to analyze data such as meal plans, room types, pricing, and booking trends, the project can improve customer satisfaction, reduce revenue loss, and improve operational efficiency.

### *2. What key gaps are you trying to solve?*


a. Hotels could identify guests who are at a higher risk of canceling their bookings by analyzing variables such as previous cancellation history and special requests.

b. Optimization of marketing strategies can be achieved by analyzing variables such as market segment type and booking patterns.

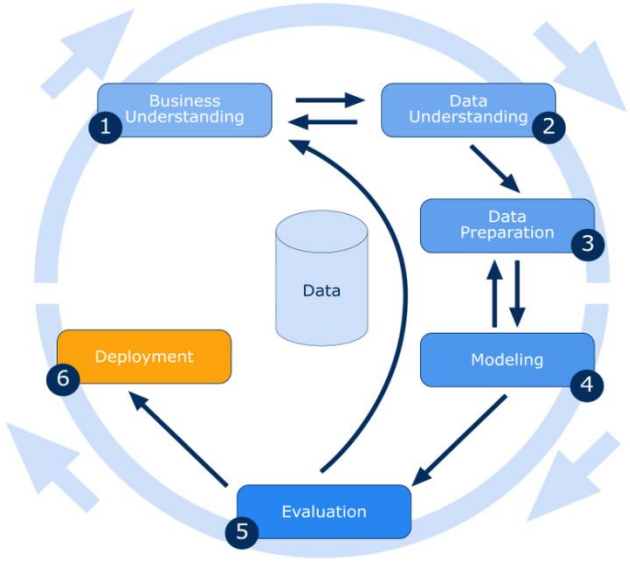
c. Addressing these gaps can improve operational efficiency, reduce revenue loss, and improve customer satisfaction.

d. By analyzing variables such as room types, meal plans, and special requests, hotels can tailor their offerings to better match guest preferences and improve the

guest experience.

e. Seasonal demand patterns can be identified by analyzing variables such as arrival year, month, and date, allowing hotels to adjust pricing and staffing levels accordingly.
f. The lead time variable can be used to streamline the check-in process and improve operational efficiency by identifying booking trends and staffing needs
g. Revenue management opportunities can be found by analyzing the average price per room and adjusting pricing strategies based on demand and availability.
h. h. In addition to predicting booking cancellations, machine learning algorithms can also be used to predict room demand and occupancy rates, enabling hotels to optimize inventory and pricing in real time.
i. Analyzing customer feedback and sentiment can provide valuable insights into areas for improvement and help hotels prioritize investments in areas such as customer service and amenities.
j. By leveraging data to personalize the guest experience, hotels can build stronger relationships with their customers and foster loyalty, leading to increased revenue and repeat bookings.

# METHODOLOGY TO BE FOLLOWED




## 1. Business Understanding:

Hotel booking cancellation is a common and the most loss causing problem that the tourism industry is facing. The factors that influence cancellation and predicting the probability of cancellation will be crucial for the industry on the whole to prevent loss and enhance operational efficiency

## 2. Data Understanding:

1. Booking_ID - unique identifier of each booking
2. no_of_adults - Number of adults
3. no_of_children - Number of Children
4. no_of_weekend_nights - Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

5. no_of_week_nights - Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
6. type_of_meal_plan - Type of meal plan booked by the customer
7. required_car_parking_space - Does the customer require a car parking space?
8. room_type_reserved - Type of room reserved by the customer.
9. lead_time - Number of days between the date of booking and the arrival date
10. arrival_year - Year of arrival date
11. arrival_month - Month of arrival date
12. arrival_date - Date of the month
13. market_segment_type - Market segment designation
14. repeated_guest - Is the customer a repeated guest?
15. no_of_previous_cancellations - Number of previous bookings that were canceled by the customer prior to the current booking
16. no_of_previous_bookings_not_canceled - Number of previous bookings not canceled by the customer prior to the current booking
17. avg_price_per_room - Average price per day of the reservation; prices of the rooms are dynamic
18. no_of_special_requests - Total number of special requests made by the customer
19. booking_status - Flag indicating if the booking was canceled or not

```
In [5]: 1 hotel_reservations.dtypes

Out[5]: Booking_ID                            object
        no_of_adults                           int64
        no_of_children                         int64
        no_of_weekend_nights                   int64
        no_of_week_nights                      int64
        type_of_meal_plan                     object
        required_car_parking_space             int64
        room_type_reserved                    object
        lead_time                              int64
        arrival_year                           int64
        arrival_month                          int64
        arrival_date                           int64
        market_segment_type                   object
        repeated_guest                         int64
        no_of_previous_cancellations           int64
        no_of_previous_bookings_not_canceled   int64
        avg_price_per_room                   float64
        no_of_special_requests                 int64
        booking_status                        object
        dtype: object

In [6]: 1 hotel_reservations.shape

Out[6]: (36275, 19)
```

## 3. Data Preparation:

a. Missing Value Treatment
b. Outlier Treatment
c. Encoding Categorical Variables
d. Scaling and Transformation of numeric data if necessary
e. Smote if required to improve the scope

i. **SMOTE** (Synthetic Minority Over-sampling Technique) is a popular data augmentation technique used to address the problem of class imbalance in a dataset. In this technique, synthetic samples are generated for the minority class by interpolating existing samples. The algorithm randomly selects a sample from the minority class and then identifies its k nearest neighbors. The synthetic samples are then generated by randomly selecting one of the k neighbors and creating a new sample that lies on the line joining the selected neighbor and the original minority sample.


ii. In the given dataset, SMOTE could be used to address the data imbalance in the booking status column. By generating synthetic samples for the minority class (e.g., cancelled bookings), the dataset can be rebalanced to ensure that the machine learning models are not biased towards the majority class (e.g., successful bookings). This can help improve the accuracy of the models and ensure that the predictions are more representative of the underlying data distribution.

## 4. Modeling:

a. Select appropriate machine learning algorithms like, XGB, Random Forest, KNN based on the problem statement and the data characteristics.
b. Train the models on the training dataset and tune their hyper parameters to optimize performance.

## 5. Evaluation:

a. Evaluate the models on the testing dataset using appropriate metrics and compare their performance to select the best model
b. Validate the model's performance using cross-validation techniques.

# NOTES FOR PROJECT TEAM

| | |
|---|---|
| **Original owner of data** | Attribution 4.0 International (CC BY 4.0) |
| **Data set information** | https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset |
| **Any past relevant articles using the dataset** | https://www.sciencedirect.com/science/article/pii/S2352340918315191<br>https://www.researchgate.net/publication/325980818_Understanding_of_online_hotel_booking_process_A_multiple_method_approach |
| **Reference** | https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset |
| **Link to web page** | www.kaggle.com |