A PROJECT ON

"HOTEL BOOKING CANCELLATION PREDICTION"

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYTICS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Submitted By:

Anand Dattatray Shinde (80274)

Sumit Vasant Shinde (80278)

Mr. Nitin Kudale
Centre Coordinator

Mrs. Manisha Hingne
Course Coordinator

CERTIFICATE

This is to certify that the project work under the title 'Hotel Booking Cancellation Prediction' is done by Anand Shinde & Sumit Shinde in partial fulfillment of the requirement for award of Diploma in Big Data Analytics Course.

Mr. Aniket Panval
Project Guide

Mrs. Manisha Hingne
Course Coordinator

Date:
Wednesday, February 21, 2024

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT, Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of SunbeamInstitute of Information Technology, Pune for their support.

Anand Shinde
DBDA Sept 2023 Batch,SIIT
Pune

Sumit Shinde
DBDA Sept 2023 Batch,
SIIT Pune

TABLE OF CONTENTS

1. Introduction

1.1 Abstract:

The "**Hotel Booking Cancellation Prediction** " project aims to provide an accurate forecast of hotel booking cancellations. Cancellations can have a significant impact on revenue, which affects decisions in the hotel industry. To address this issue, the project uses a machine learning-based cancellation model that combines data science tools with human judgment and interpretation to predict cancellations.

The project demonstrates how predictive analysis can contribute to synthesizing and predicting booking cancellations. It also aims to give users relaxation by providing full prediction and analysis to help them make informed decisions about applying to a particular hotel.

The project uses various algorithms such as Logistic, KNN, Random Forest, Decision Tree, etc. to classify data and evaluation matrices to separate categorical data into specific types. By entering certain fields, users can obtain cancellation predictions at the desired level. This helps prevent poor room management by hotels and improves the customer experience.

1.2 Why this problem needs to be Solved?

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behaviour. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impacts a hotel on various fronts:

- Loss of resources (revenue) when the hotel cannot resell the room.
- Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
- Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
- Human resources to make arrangements for the guests.

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled.

1.3 Dataset Information

The data contains the different attributes of customers' booking details. There are 19 columns i.e. features and 53252 rows. The detailed data dictionary is given below.

- **Booking_ID:** unique identifier of each booking
- **no_of_adults:** Number of adults
- **no_of_children:** Number of Children
- **no_of_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- **no_of_week_nights:** Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- **required_car_parking_space:** Does the customer require a car parking space? (0 - No, 1- Yes)
- **room_type_reserved:** Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
- **type_of_meal_plan:** Type of meal plan booked by the customer:
    - Not Selected – No meal plan selected
    - Meal Plan 1 – Breakfast
    - Meal Plan 2 – Half board (breakfast and one other meal)
    - Meal Plan 3 – Full board (breakfast, lunch, and dinner)

- **lead_time:** Number of days between the date of booking and the arrival date
- **arrival_year:** Year of arrival date
- **arrival_month:** Month of arrival date
- **arrival_date:** Date of the month
- **market_segment_type:** Market segment designation.
- **repeated_guest:** Is the customer a repeated guest? (0 - No, 1- Yes)
- **no_of_previous_cancellations:** Number of previous bookings that were cancelled by the customer prior to the current booking
- **no_of_previous_bookings_not_canceled:** Number of previous bookings not cancelled by the customer prior to the current booking
- **avg_price_per_room:** Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- **no_of_special_requests:** Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- **booking_status:** Flag indicating if the booking was cancelled or not.

Chapter 2
Basic Concepts/ Literature Review

Tools - The following tools are required for developing this project.

Jupyter notebook - JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

This project was created using the following languages and libraries. An environment with the correct versions of the following libraries will allow re-production and improvement on this project.

Python version: 3 - Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation via the off-side rule.

Matplotlib version: 3.0.3 - Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.

Seaborn version: 0.9.0 - Python Seaborn library is a widely popular data visualization library that is commonly used for data science and machine learning tasks. You build it on top of the matplotlib data visualization library and can perform exploratory analysis. You can create interactive plots to answer questions about your data.

Sklearn version: 0.20.3 - Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

Numpy 1.24.3 - NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Pandas 2.0.1 - pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

Data Science Concepts Used:

Machine Learning (ML) - Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Hypothesis Testing - Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data.
Hypothesis Testing is basically an assumption that we make about the population parameter.
Time Series- Time Series Analysis in Python considers data collected over time might have some structure; hence it analyzes Time Series data to extract its valuable characteristics.

Data Visualization - Data visualization is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data. Using data visualization, we can get a visual summary of our data. With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data.

Data Cleaning - Data cleaning is the process of identifying and fixing incorrect data. It can be in incorrect format, duplicates, corrupt, inaccurate, incomplete, or irrelevant. Various fixes can be made to the data values representing incorrectness in the data.

Data Exploration - Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

Feature Engineering- Feature engineering is a machine learning technique that leverages data to create new variables that aren’t in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

3. Project Planning

Requirements:

- How Cancellations are Affected by the Market Segment Of Booking?
- How is cancellation affected by the lead time of a booking?
- What is the rate of cancellation of booking outside Portugal and booking that is made in Portugal?
- What are the other factors that are affecting the booking cancellation of hotels?
- Which machine learning algorithm has the highest accuracy while predicting the cancellations of hotel bookings?

Steps To Be Followed:

- **Data Cleaning:**

Data cleaning means removal or fixing of bad data in our data set. Bad data could be: Empty cells; Duplicates; Data in wrong format; Wrong data.

- **Exploratory Data Analysis:**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

- **Feature Selection for machine learning process:**

Feature selection is the process of decreasing the input variable to our model by using only relevant data and getting rid of noise in the data. It is the process by which relevant features are automatically chosen for our machine learning model based on the type of problem we are trying to solve.

- **Model Building:**

Model building in machine learning is the process of creating a mathematical representation by generalizing and learning from training data. Then, the built machine learning model is applied to new data to make predictions and obtain results.
There are 3 types of machine learning models:

- Supervised Learning.
- Unsupervised Learning.
- Reinforcement Learning.

System Design

3.2.1 Design Constraint Jupyter Notebook:

The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers simple, streamlined, document-centric experience. Project Jupyter is a project to develop open-source software, open standards, and services for interactive computing across multiple programming languages.

3.3.2 System Architecture OR Block Diagram

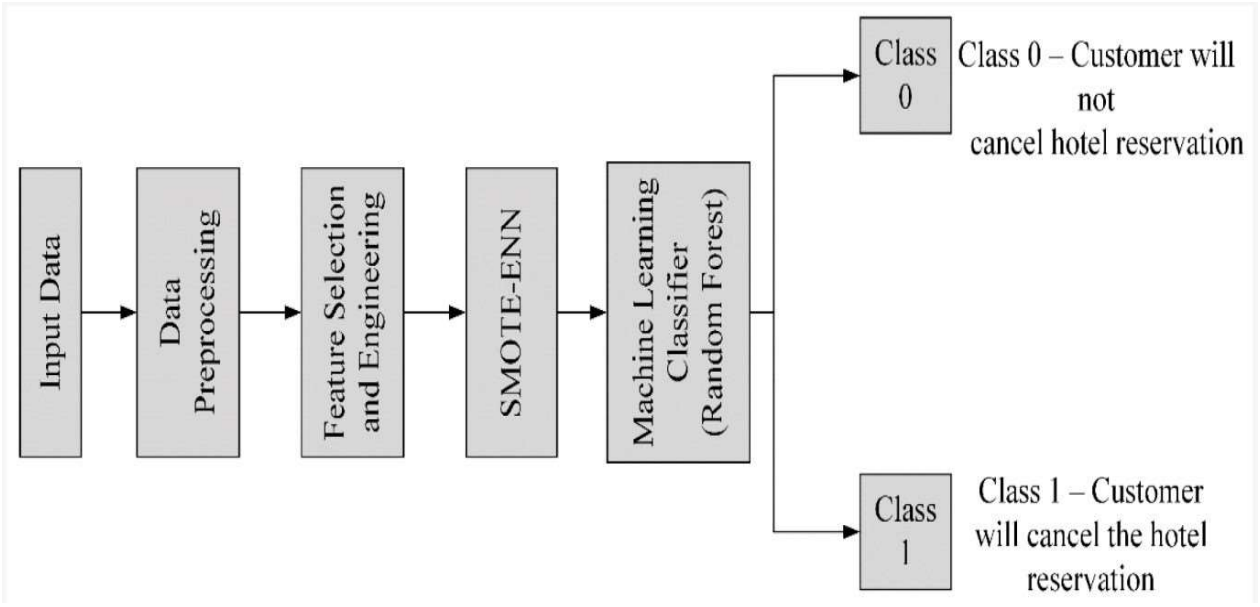
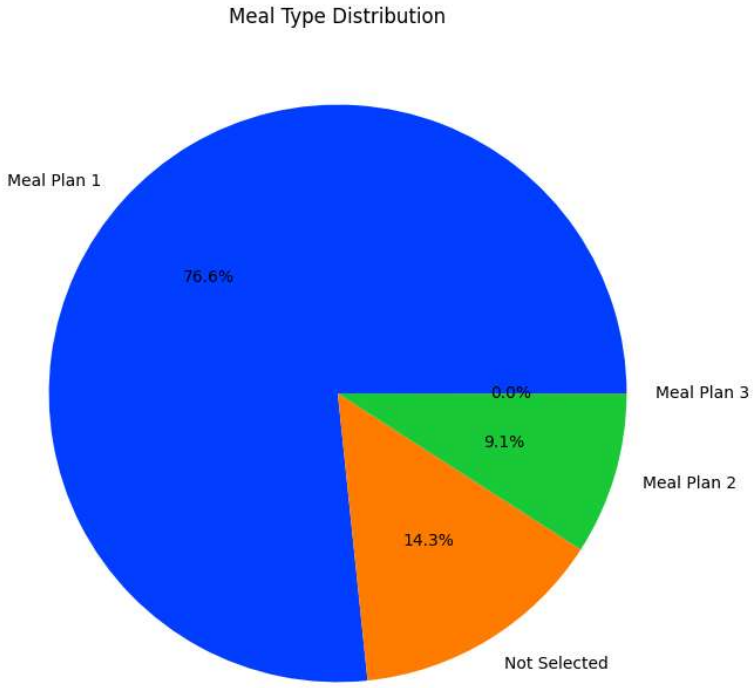


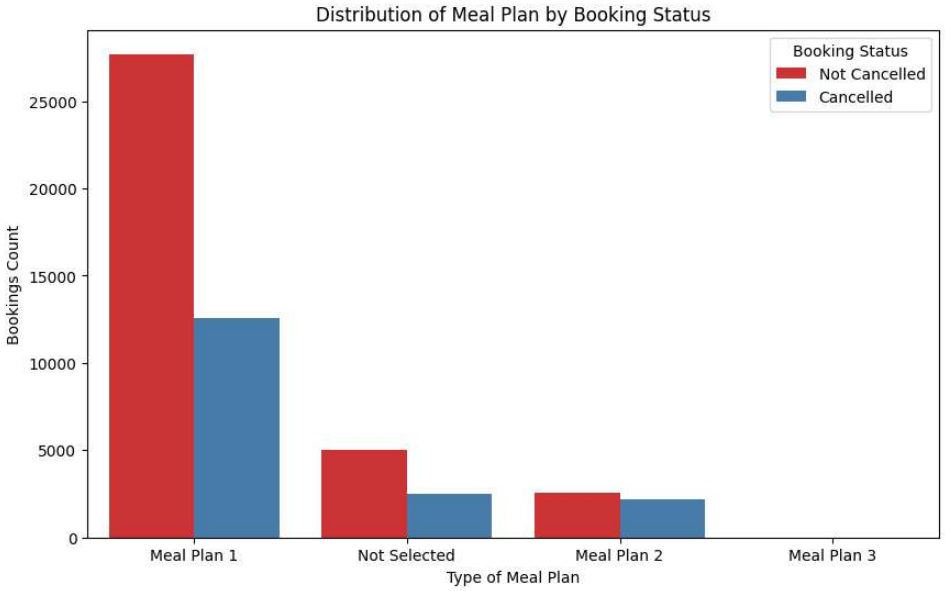
Fig 1: Block Diagram Representation of Conceptual Methodology

4. Exploratory Data Analysis:

The below diagram shows meal plan distribution opted by customers. Meal Plan 1 is most opted meal plan.

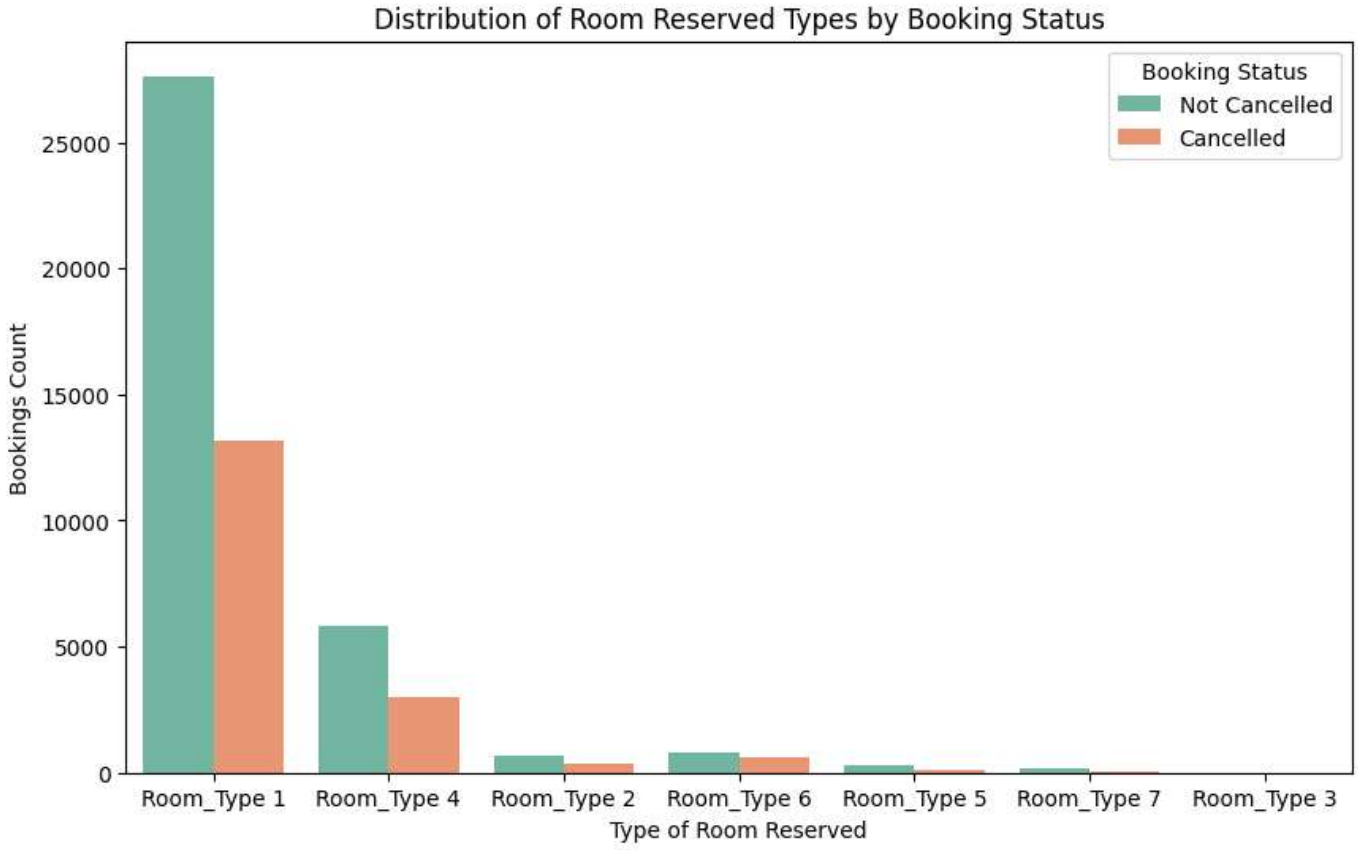
```
type_of_meal_plan
Meal Plan 1     40253
Not Selected     7486
Meal Plan 2      4771
Meal Plan 3         9
Name: count,
dtype: int64
```





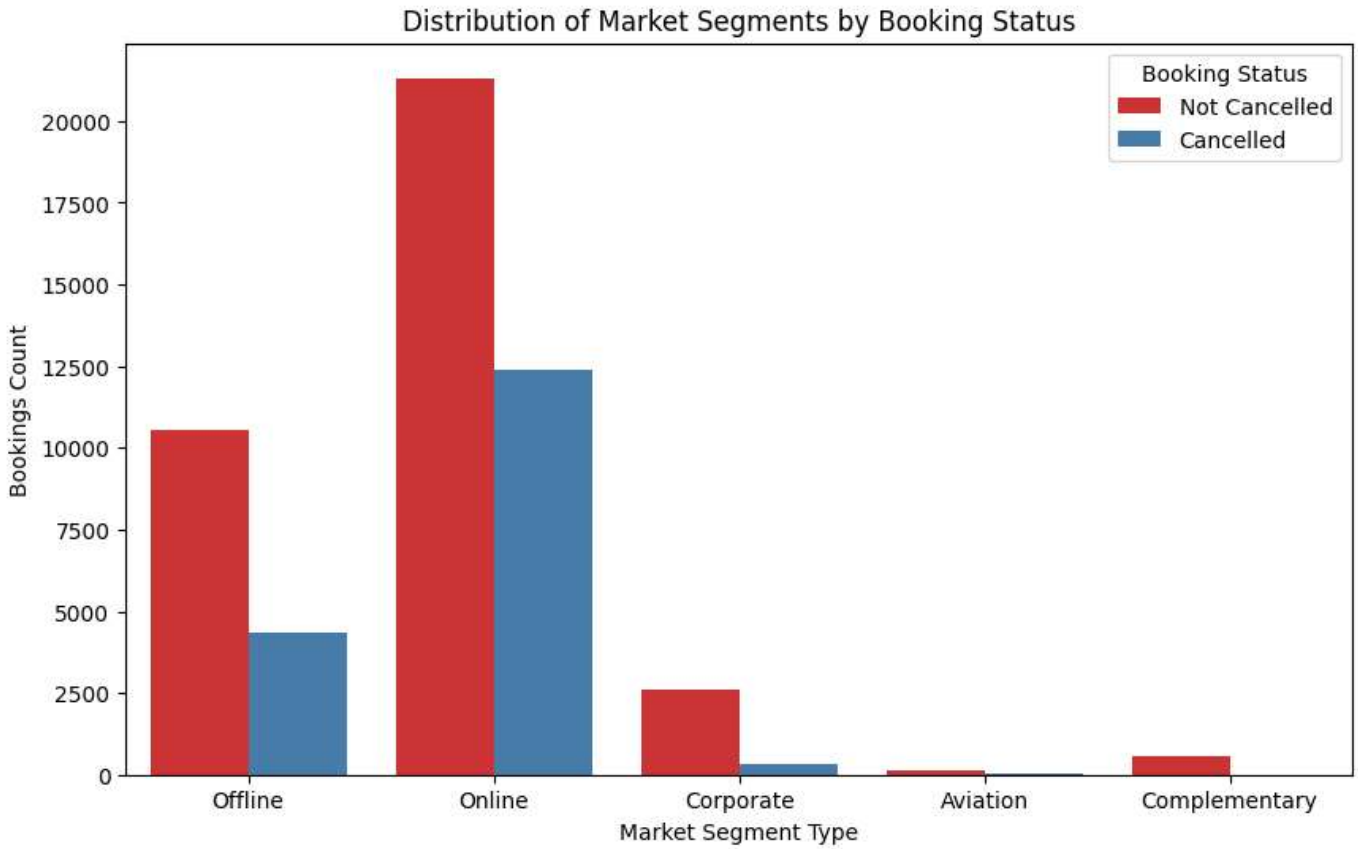
From below chart we can conclude that customers who opt Room_Type 6 are most likely to cancel their booking.

```
room_type_reserved
Room_Type 1    40779
Room_Type 4     8803
Room_Type 6     1404
Room_Type 2      994
Room_Type 5      379
Room_Type 7      230
Room_Type 3        9
Name: count,
dtype: int64
```



```
market_segment_type
Online           33687
Offline          14895
Corporate         2910
Complementary      545
Aviation           188
Name: count, dtype: int64
```

From below chart we can see that corporate booking are less likely to get cancelled while online bookings are most vulnerable to cancellation.



5 Implementation

During the project development, the following steps were taken:

5.1 Methodology

Data Loading and Reading: In this step, we loaded the dataset into the environment and read the data using Python's Pandas library.

Data Cleaning and Preprocessing: We removed missing values, duplicates, and outliers from the dataset. We also performed data imputation for missing values using different techniques like mean imputation, mode imputation, and regression imputation. Additionally, we scaled the numerical variables using standard scaling to ensure that all variables have a similar scale. Furthermore, we converted categorical variables into numerical variables using one-hot encoding.

Exploratory Data Analysis: In this step, we analyzed the origin of guests, price paid per night by guests, busiest months for bookings, month with the highest Average Daily Rate (ADR), and whether bookings were made for weekdays, weekends, or both.

Feature Engineering: We created new features like total number of guests, total number of nights booked, and average price per night.

Feature Encoding: We transformed categorical variables into numerical variables using one-hot encoding.

Outlier Detection and Handling: We detected and handled outliers using different techniques like Z-score, Tukey's method, and Isolation Forest.

Feature Selection: We selected important features using correlation and univariate analysis.

Model Building: We built a machine learning model using various algorithms like Logistic Regression, Random Forest, and SVM.

Model Cross-Validation: We performed cross-validation to evaluate the performance of the machine learning model.

Experimenting with Multiple Algorithms: We experimented with multiple algorithms for model building to find the best performing one.

5.2 Testing OR Verification Plan

We tested the machine learning model using various test cases to ensure that the model is performing well on different datasets. We also performed hyperparameter tuning to improve the performance of the model.

Test	Test Case Title	Test Condition	System Behavior	Expected Result
T01	Data Loading Test	The data loading function is called with a valid input file path.	The function should read the input file and load the data into memory.	The data is loaded without errors and is available for further processing.
T02	Outlier Detection Test	The outlier detection function is called with a sample dataset containing known outliers.	The function should identify the outliers and mark them for removal.	The identified outliers should be removed from the dataset and the remaining data should be suitable for further processing.

		PREDICTED	
		Guest Cancel (1)	**Guest Show-up (0)**
ACTUAL		TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
	Guest Cancel (1)	Model correctly predicts that the guest will cancel the booking.	Model predicts that the guest will not cancel the booking while they actually will.
		FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)
	Guest Show-up (0)	Model predicts that the guest will cancel the booking while they actually won't.	Model correctly predicts that the guest will not cancel the booking.

Summary

Type 1 error: False Positive (guest predicted will cancel but actually not)
Consequence: high risk of overbooking
Type 2 error: False Negative (guest predicted won't cancel but actually will)
Consequence: loss revenue due to unutilised rooms

Metric Analysis

Referring to the *metric evaluation* above, we need to minimise the False Negative Rate (minimise risk of loss revenue). If the model fails to minimise the False Negative Rate (Type 2 error), it means that the guest who is supposed to cancel is predicted by the model as a guest who wouldn't cancel. If this happens, the business owner will experience a loss of customers, which can increase the risk of losing revenue due to unoccupied rooms that could be resold or republished on online travel agencies.
However, we should pay attention to the score of False Positive (Type 1 error) in the prediction results. If it is high, it means that the model incorrectly predicts guests who want to cancel. This will lead to the high risk of overbooking. When the model predicts a guest will cancel the room, we will take action by reselling/republishing the room by online/offline. If the prediction were wrong and someone had booked this occupied room, that would cause overbooking if there was no similar room available.
So, the model we are after is a model that provides accurate predictions in the positive class with a higher recall score to avoid losing revenue and unutilised capacity. However, we need to make sure that the precision score has a good measure to avoid the risk of overbooking. So, we have to balance between precision and recall of 1 positive class. So the main metric of prediction will be **f1-score**, but we will also pay attention to the **recall** score and ensure that it is greater than **precision**. In addition, the purpose of using the f1-score is one method to overcome the imbalance of data in the positive class (Guest cancel) and the negative class (Guest Show-up).

GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries.

It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

6. Results and discussion:

Logistic Regression, Naïve Baye's Classifier, KNN Model, Support Vector Machine, Random Forest, DecisionTree machine algorithms were used to predict whether customer will cancel booking or not. Among the given algorithms Random Forest Machine algorithm was the best performing one.

```
          from sklearn.metrics import confusion_matrix,
accuracy_score, precision_score, recall_score, f1_score

    y_true = y_test
    y_pred = model.predict(x_test)

    # get confusion matrix
    cm = confusion_matrix(y_true, y_pred)
    print(cm)

    # print accuracy
    print(f"accuracy = {accuracy_score(y_true, y_pred) * 100:.2f}%")
    print(f"precision = {precision_score(y_true, y_pred) * 100:.2f}%")
    print(f"recall = {recall_score(y_true, y_pred) * 100:.2f}%")
    print(f"F1 score = {f1_score(y_true, y_pred) * 100:.2f}%")
```

```
[[4799 523]
[ 515 4624]]
accuracy = 90.08%
precision = 89.84%
recall = 89.98%
F1 score = 89.91%
```

7. Conclusion and Future Scope

7.1 Conclusion

Random Forest Has the Best Accuracy Among All Algorithms That We Tried from all the evaluation matrix to predict hotel cancellation classification case, we see that Random Forest has the best accuracy when it comes to predicting hotel cancellation based on certain features (85.2 %). This model enables hotel managers to mitigate revenue loss derived from booking cancellations and to mitigate the risks associated with overbooking (relocation costs, cash or service compensations, and, particularly important today, social reputation costs). This project also allows hotel managers to implement less rigid cancellation policies, without increasing uncertainty. This has the potential to translate into more sales, since less rigid cancellation policies generate more bookings.

7.2 Future Scope

A hotel check-in abandonment forecasting model can provide valuable insights for hotel management to optimize their operations and improve guest experience. Here are some potential **future scopes of such a model:**

1) **Real-time optimization:** The model can be integrated with real- time data feeds to allow hotel staff to make informed decisions on the spot. For example, if the model predicts high check-in abandonment rates during a particular time of day, the staff can allocate more resources to that time to ensure a smooth check-in process.

2) **Personalization:** The model can be customized to take into account the guest's history, preferences, and behavior to provide a more personalized experience. For instance, if the model predicts that a guest is likely to

abandon check-in due to a long queue, the hotel can offer them a personalized check-in service.

3) **Revenue optimization:** The model can help hotels optimize their revenue by predicting the optimal number of rooms to overbook. By considering the historical data and trends, the model can suggest the right amount of overbooking without risking the guest experience.

4) **Customer loyalty:** By reducing check-in abandonment rates, the model can help hotels improve customer loyalty. When guests have a seamless and efficient check-in experience, they are more likely to return and recommend the hotel to others.

5) **Continuous improvement:** The model can help hotels continuously improve their operations by providing insights into the causes of check-in abandonment. By analyzing the data, the hotel can identify areas for improvement and implement changes to reduce abandonment rates over time.

Overall, a hotel check-in abandonment forecasting model has a lot of potential for enhancing guest experience and optimizing hotel operations. As technology advances and more data becomes available, the scope of such a model will only increase.