

# **HOTEL BOOKING CANCELLATION PREDICTION APP**

**by  
Fırat Özen**

**Engineering Project Report**

**Yeditepe University  
Faculty of Engineering  
Department of Computer Engineering  
2024**



# HOTEL BOOKING CANCELLATION PREDICTION APP

APPROVED BY:

Assoc. Dr. Onur Demir  
(Supervisor)



Prof. Dr. Şebnem Baydere



Assoc. Dr. Mustafa Bülent Mutluoğlu



DATE OF APPROVAL: 12/01/2024

## **ACKNOWLEDGEMENTS**

First of all I would like to thank my advisor Assoc. Dr. Onur Demir for his guidance and support throughout my project.

Also I would like to thank my parents for their support and encouragement throughout my education up to the present.

# ABSTRACT

## HOTEL BOOKING CANCELLATION PREDICTION APP

Hotel booking cancellations pose significant challenges to resource management and revenue optimization in the hospitality industry. This project presents a predictive model that leverages historical booking data to forecast cancellations accurately, with a primary focus on revenue optimization for hoteliers. The application integrates advanced machine learning algorithms to predict cancellations, enabling strategic overbooking to maximize occupancy and revenue. Key functionalities include customer registration, data updates, and visualizations of cancellation probabilities, providing hotel managers with actionable insights. By offering tailored advice on optimal overbooking strategies and assessing potential revenue gains and costs, the application demonstrates the benefits of data-driven decision-making in enhancing operational efficiency and financial performance. This feature underscores our commitment to delivering innovative tools that help hoteliers achieve sustained profitability in a competitive market.

# ÖZET

## OTEL REZERVASYON İPTALİ TAHMİN UYGULAMASI

Otel rezervasyon iptalleri, konaklama sektöründe kaynak yönetimi ve gelir optimizasyonunda önemli zorluklar oluşturuyor. Bu proje, geçmiş rezervasyon verilerini kullanarak iptalleri doğru bir şekilde tahmin etmek için bir öngörü modeli sunmaktadır; bu tahminlerde otelcilerin gelir optimizasyonuna odaklanılmıştır. Uygulama, ileri düzey makine öğrenimi algoritmalarını entegre ederek rezervasyon iptallerini tahmin etmekte ve doluluk oranını ve geliri maksimize etmek için stratejik fazla rezervasyon yapılmasına olanak tanımaktadır. Müşteri kaydı, veri güncellemeleri ve iptal olasılıklarının görselleştirilmesi gibi temel işlevler, otel yöneticilerine harekete geçirilebilir bilgiler sunmaktadır. Optimal fazla rezervasyon stratejileri konusunda özel tavsiyeler sunarak potansiyel gelir artışlarını ve maliyetleri değerlendiren uygulama, veri odaklı karar vermenin operasyonel verimlilik ve finansal performansı artırmadaki faydalarını göstermektedir. Bu özellik, otelcilerin rekabetçi bir pazarda sürdürülebilir karlılık elde etmelerine yardımcı olacak yenilikçi araçlar sunma taahhüdümüzü vurgulamaktadır.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . .	iii
ABSTRACT . . . . .	iv
ÖZET . . . . .	v
LIST OF FIGURES . . . . .	ix
LIST OF TABLES . . . . .	xi
1. INTRODUCTION . . . . .	1
1.1. Hotel Booking Cancellation Prediction . . . . .	1
1.2. Terms . . . . .	3
1.3. Motivation . . . . .	6
1.3.1. Enhancing Guest Experience . . . . .	7
1.3.2. Competitive Advantage . . . . .	7
1.3.3. Optimizing Revenue . . . . .	7
1.3.4. Aims . . . . .	8
1.4. Scope and Limitations . . . . .	9
1.5. Problem Definition . . . . .	9
1.6. Requirements . . . . .	10
1.6.1. Software Requirements . . . . .	10
1.6.2. Hardware Requirements . . . . .	12
1.6.3. Requirements - Test & Results . . . . .	12
2. BACKGROUND . . . . .	15
2.1. Previous work . . . . .	15
2.1.1. Historical Models . . . . .	15
2.1.2. Advance Models . . . . .	18
3. ANALYSIS AND DESIGN . . . . .	27
3.1. Data Collection . . . . .	27
3.2. Data Cleaning . . . . .	28
3.3. Exploratory Data Analysis (EDA) . . . . .	29
3.3.1. Where Do The Guests Come From? . . . . .	29
3.3.2. How Much Do Guests Pay For A Room Per Night? . . . . .	29
3.3.3. How Does The Price Vary Per Night Over The Year? . . . . .	30
3.3.4. Which Are The Most Busy Months? . . . . .	31
3.3.5. How Long Do People Stay At The Hotels? . . . . .	31
3.3.6. ADR and Market Segment . . . . .	32
3.3.7. How Many Bookings Were Canceled? . . . . .	33
3.3.8. Which Month Have The Highest Number Of Cancelations? . . . . .	34

3.4.	Data Pre Processing . . . . .	35
3.4.1.	Feature Selection . . . . .	36
3.4.2.	Elimination of Useless Columns . . . . .	36
3.4.3.	Encoding Categorical Variables . . . . .	37
3.4.4.	Normalization of Numerical Variables . . . . .	37
3.4.5.	Handling Missing Values . . . . .	38
3.4.6.	Train-Test-Validation Split and Evaluation . . . . .	38
3.5.	Building Models To Predict Booking Cancellations And Optimize Revenue	39
3.5.1.	Building Models To Predict Booking Cancellations . . . . .	39
3.5.2.	Building Models To Optimize Revenue . . . . .	43
3.6.	Functions of Pages . . . . .	47
3.6.1.	Database & Prediction Page . . . . .	47
3.6.2.	Reservation Page . . . . .	48
3.6.3.	Simulation Page . . . . .	50
3.6.4.	Optimize Revenue Page . . . . .	51
3.6.5.	How Does This Model Work Page . . . . .	53
3.6.6.	Insight Page . . . . .	54
3.6.7.	Pseudocode . . . . .	56
3.6.8.	Use Case Diagram . . . . .	58
3.6.9.	Flow Chart . . . . .	61
4.	IMPLEMENTATION . . . . .	65
4.1.	Used Technologies . . . . .	65
4.2.	Interfaces . . . . .	67
4.2.1.	Reservation Page . . . . .	68
4.2.2.	Database & Prediction Page . . . . .	70
4.2.3.	Prediction Dropdown Page . . . . .	72
4.2.4.	Simulation Page . . . . .	74
4.2.5.	Optimize Revenue Page . . . . .	75
4.2.6.	How Does This Model Work Page . . . . .	78
4.2.7.	Insight Page . . . . .	81
5.	TEST AND RESULTS . . . . .	84
5.1.	Test And Results For The Model To Predict Booking Cancellations . . . . .	84
5.2.	Test And Results For The Model To Optimize Revenue . . . . .	87
5.2.1.	Revenue Optimization Analysis for August 2017 . . . . .	90
6.	CONCLUSION AND FUTURE WORK . . . . .	95
6.1.	Conclusion . . . . .	95
6.2.	Future Work . . . . .	96
6.2.1.	Integrating Diverse Data Sources . . . . .	96



6.2.2. Incorporating Additional Predictive Features .....	96
6.2.3. Real-time Data Processing .....	96
6.2.4. Expanding to Mobile Applications.....	97
6.2.5. Enhancing Security and Privacy Measures.....	97
6.2.6. Conducting Longitudinal Studies.....	97
Bibliography .....	98

## LIST OF FIGURES

<b>Figure 1.1.</b>	Hotel . . . . .	2
<b>Figure 1.2.</b>	Steps of Creating Model . . . . .	3
<b>Figure 1.3.</b>	Train-Test-Validation Schema . . . . .	4
<b>Figure 1.4.</b>	Evaluation Graphics . . . . .	6
<b>Figure 2.1.</b>	Flowchart of forecast algorithm [1] . . . . .	16
<b>Figure 2.2.</b>	Hotel 1 Forecast Error [2] . . . . .	17
<b>Figure 2.3.</b>	Test set results for Ridge Regression with quadratic features [4] . . . .	18
<b>Figure 2.4.</b>	Accuracy Comparision of Different Model that we have used in their project [8] . . . . .	22
<b>Figure 2.5.</b>	Relative performance of the six data mining based methods [10] . . .	23
<b>Figure 2.6.</b>	Average Classification Result of (a) Primary and (b) Min-Max (c) Z Normalization (d) Square Root Transformed Dataset [11] . . . . .	24
<b>Figure 3.1.</b>	Distribution of booking cancellation. . . . .	27
<b>Figure 3.2.</b>	Null values in dataframe . . . . .	28
<b>Figure 3.3.</b>	Origin of the guests. . . . .	29
<b>Figure 3.4.</b>	Room prices. . . . .	30
<b>Figure 3.5.</b>	Seasonal price variation. . . . .	30
<b>Figure 3.6.</b>	Busiest months. . . . .	31
<b>Figure 3.7.</b>	Length of stay. . . . .	32

<b>Figure 3.8.</b>	Booking per market segment. . . . .	33
<b>Figure 3.9.</b>	ADR by market segment and room type. . . . .	34
<b>Figure 3.10.</b>	Cancellations per months. . . . .	35
<b>Figure 3.11.</b>	Correlation of columns. . . . .	36
<b>Figure 3.12.</b>	First Part Of The Pseudocode Diagram . . . . .	57
<b>Figure 3.13.</b>	Second Of The Pseudocode Diagram . . . . .	58
<b>Figure 3.14.</b>	Use Case Diagram . . . . .	59
<b>Figure 3.15.</b>	Flowchart . . . . .	61
<b>Figure 4.1.</b>	Reservation Page . . . . .	68
<b>Figure 4.2.</b>	Database Page . . . . .	71
<b>Figure 4.3.</b>	Prediction Dropdown Menu Page . . . . .	73
<b>Figure 4.4.</b>	Optimize Revenue Page . . . . .	76
<b>Figure 4.5.</b>	Optimize Revenue Page . . . . .	78
<b>Figure 4.6.</b>	How Does This Model Work Page . . . . .	79
<b>Figure 4.7.</b>	Insight Page . . . . .	82
<b>Figure 5.1.</b>	Confusion Matrix . . . . .	85
<b>Figure 5.2.</b>	Confusion Matrix . . . . .	88

## LIST OF TABLES

<b>Table 2.1.</b>	Hotel Dataset Performance Metrics . . . . .	19
<b>Table 2.2.</b>	Hotel Performance Metrics [6] . . . . .	20
<b>Table 2.3.</b>	Random Forest Classification Report [7] . . . . .	21
<b>Table 3.1.</b>	Model Performance on Binary Classification Task for Predicting Book- ing Cancellations . . . . .	40
<b>Table 3.2.</b>	Model Performance for Binary Classification Task . . . . .	44
<b>Table 5.1.</b>	Classification Report For CatBoost Model . . . . .	86
<b>Table 5.2.</b>	Classification Report For XGBoost Model . . . . .	89

# 1. INTRODUCTION

In the rapidly evolving landscape of the hospitality industry, hotels face numerous operational challenges that can significantly impact their bottom line. Among these, booking cancellations present a particularly vexing problem, disrupting revenue forecasts and complicating inventory management. The ability to accurately predict cancellations is therefore critical, not only for maintaining optimal occupancy rates but also for enhancing customer satisfaction and streamlining resource allocation. This project addresses the need for sophisticated predictive models that leverage advanced data analytics and machine learning techniques to forecast booking cancellations with high precision. By integrating these predictive capabilities into hotel management systems, we aim to provide actionable insights that can drive more informed decision-making, optimize revenue, and ultimately improve both operational efficiency and guest experiences.

## 1.1. Hotel Booking Cancellation Prediction

The hospitality industry is characterized by a dynamic and competitive landscape, where effective resource management and customer satisfaction are crucial for success. One of the persistent challenges faced by hotels (Figure 1.1) is the accurate prediction of booking cancellations, which can significantly impact revenue, inventory management, and service quality. This project seeks to address this challenge by developing a predictive model that leverages historical booking data to forecast cancellations with high accuracy.

The core objective of this project is to utilize advanced data analytics and machine learning techniques to gain deeper insights into booking patterns, customer behavior, and revenue optimization strategies. By analyzing a comprehensive dataset that includes various attributes such as booking lead time, customer demographics, stay duration, and booking status, we aim to uncover the underlying factors influencing booking cancellations. This analysis not only helps in building a robust predictive model but also provides actionable insights and over-booking strategies to optimize revenue for hotel management. By integrating these predictive capabilities into hotel operations, hoteliers can make more informed decisions, improve resource allocation, and enhance overall profitability.

Our approach involves a meticulous process of data collection, cleaning, and exploratory analysis to ensure the dataset's integrity and quality. We then proceed with feature selection and data preprocessing, critical steps that lay the foundation for effective model building.



**Figure 1.1.** Hotel

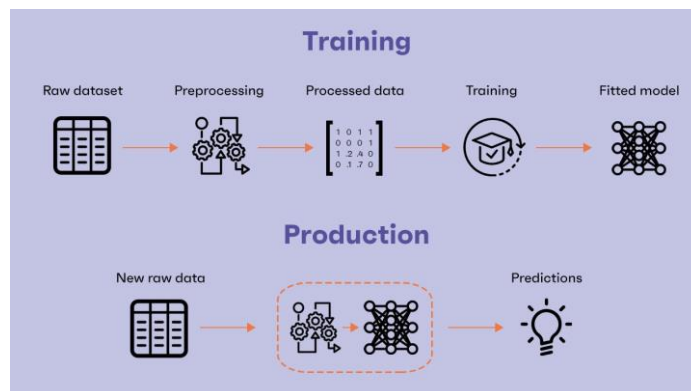
By employing state-of-the-art machine learning algorithms, we strive to create a model that generalizes well to new data and provides reliable predictions.

In addition to the predictive model, the developed application serves as a comprehensive tool for hotel management. It allows for new customer registrations, updates to existing customer information, displays the probability of cancellations for all registered customers, and features a simulation function where hypothetical customer scenarios can be entered and analyzed without the need for actual registration. This functionality enables hotel managers to explore various booking scenarios and make informed decisions to optimize operations, enhance customer satisfaction, and strategically adjust booking strategies to maximize revenue.

Overall, this project demonstrates the practical application of data science in solving real-world problems within the hospitality industry. By integrating predictive analytics into hotel management processes, we aim to improve resource allocation, reduce the adverse impacts of cancellations, elevate the overall guest experience, and optimize revenue.

## 1.2. Terms

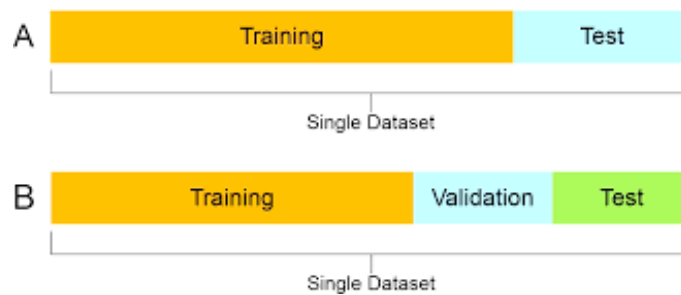
- *Cancellation Rate*: The likelihood that a hotel reservation will be canceled before the check-in date. This rate is derived from historical booking data and analyzed using various customer and booking attributes.
- *Booking Lead Time*: The interval between the date a reservation is made and the check-in date. This duration can significantly affect the probability of a cancellation.
- *A Stay Duration*: The total number of nights a guest stays at the hotel, from check-in to check-out. This factor can impact cancellation rates and booking trends.
- *A Predictive Model*: A statistical or machine learning model designed to forecast future events, such as the likelihood of a booking being canceled, based on historical data and patterns.
- *Simulation Function*: A feature within the application that allows hotel managers to input hypothetical booking scenarios to predict potential outcomes, aiding in decision-making without actual reservations being made.
- *Data Preprocessing*: The process of cleaning and transforming raw data into a suitable format for analysis, which includes handling missing values, normalizing data, and encoding categorical variables. (Figure 1.2)



**Figure 1.2.** Steps of Creating Model

- *Feature Selection*: The method of identifying the most relevant variables from a dataset that significantly contribute to the prediction of booking cancellations.
- *Exploratory Data Analysis (EDA)*: The initial phase of data analysis where datasets are examined using visual and statistical methods to summarize their main characteristics and uncover patterns.

- *Machine Learning Algorithm*: A computational method that enables systems to learn from data and make predictions or decisions without explicit programming. Used here to develop models for predicting booking cancellations.
- *Normalization*: The process of adjusting numerical data to a common scale, often using techniques like logarithmic transformations, to ensure that variables are comparable and the model's performance is optimized.
- *Imputation*: A technique used to fill in missing data points in a dataset with substituted values, such as the mean or median of the feature, to maintain dataset integrity.
- *Train-Test Split*: The process of dividing a dataset into training and test subsets to evaluate the performance of predictive models. Typically, a larger portion is used for training, while the remaining data is reserved for testing. (Figure 1.3)

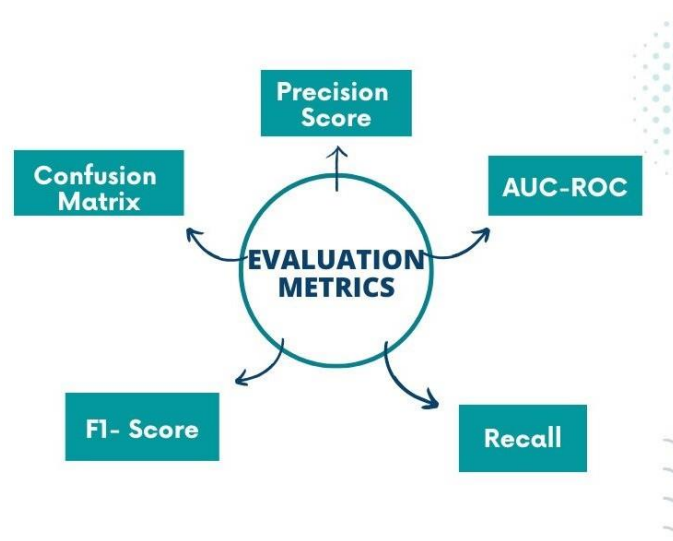


**Figure 1.3.** Train-Test-Validation Schema

- *Market Segment*: Categories into which customers are divided based on their booking sources and behaviors, such as corporate, online travel agencies (OTAs), or groups, which can influence booking patterns and cancellation rates.
- *Booking Status*: The current state of a reservation, indicating whether it is active, canceled, or completed, providing essential data for predictive modeling.
- *Average Daily Rate (ADR)*: The average revenue earned per occupied room per day, a key metric in the hospitality industry used to analyze pricing trends and revenue management strategies.
- *Occupancy Rate*: The percentage of available rooms that are occupied over a given period, an important indicator of hotel performance and demand patterns.
- *Seasonal Variations*: Changes in booking and cancellation patterns that occur due to seasonal factors, such as holidays, school vacations, and weather conditions, which can affect hotel occupancy and revenue.



- *Room Type*: The classification of hotel accommodations based on factors such as size, amenities, and occupancy capacity, which can influence booking preferences and pricing strategies.
- *Booking Channel*: The method through which guests make reservations, such as direct booking through the hotel's website, online travel agencies (OTAs), or phone reservations, impacting booking patterns and distribution costs.
- *Special Requests*: Additional services or accommodations requested by guests during the booking process, such as room preferences, extra beds, or dietary requirements, which can affect reservation management and guest satisfaction.
- *Check-in Date*: The date on which a guest is scheduled to arrive and check-in to the hotel, a critical factor influencing booking decisions and cancellation rates.
- *Booking Changes*: Modifications or amendments made to a reservation after it has been initially confirmed, such as changes to stay dates, room types, or guest details, which can impact reservation management and revenue forecasting.
- *Data Visualization*: The graphical representation of data and information using charts, graphs, and other visual elements to facilitate understanding and analysis, aiding in decision-making and communication of insights.
- *Model Evaluation Metrics*: Quantitative measures used to assess the performance of predictive models, such as accuracy, precision, recall, and F1 score, providing insights into model reliability and effectiveness. (Figure 1.4)
- *Overfitting*: The phenomenon in machine learning where a model learns the training data too well, capturing noise and irrelevant patterns that do not generalize to new data, leading to poor performance on unseen data.
- *Underfitting*: The condition where a model is too simple to capture the underlying patterns in the data, resulting in high bias and low predictive accuracy, indicating the need for more complex models or additional features.
- *Cross-Validation*: A technique used to evaluate model performance by splitting the dataset into multiple subsets and training the model on different combinations of training and validation data, providing more robust estimates of performance.
- *Model Deployment*: The process of integrating a trained machine learning model into a production environment, allowing it to make real-time predictions or recommendations based on new data, enabling practical use and value generation.



**Figure 1.4.** Evaluation Graphics

- *Cancellation Threshold:* A calculated value used to categorize the risk levels of booking cancellations based on historical data and model predictions, helping in the assessment of potential cancellations.

### 1.3. Motivation

The hospitality industry is constantly evolving, with hotels striving to enhance operational efficiency, optimize revenue, and deliver superior guest experiences. Predicting hotel booking cancellations is a critical aspect of this endeavor, as it directly impacts revenue management, inventory control, and customer satisfaction. The motivation behind this project stems from the need to address these challenges through advanced predictive analytics. By leveraging historical booking data and state-of-the-art machine learning algorithms, the project aims to develop a comprehensive predictive model that not only forecasts cancellations with high accuracy but also provides actionable insights and overbooking strategies to optimize revenue for hotel management. This dual focus on predictive accuracy and practical applicability underscores the potential of data-driven decision-making in transforming hotel operations and fostering a competitive edge in a dynamic market.

### **1.3.1. Enhancing Guest Experience**

The implementation of predictive models for hotel booking cancellations transcends operational benefits and directly impacts the guest experience. By minimizing overbooked or unavailable rooms resulting from last-minute cancellations, hotels can elevate guest satisfaction and ensure a seamless booking process. This heightened level of service not only fosters increased guest loyalty but also generates positive reviews, thereby enhancing the hotel's reputation and attracting future bookings. Ultimately, prioritizing guest experience through predictive analytics reinforces the hotel's commitment to delivering exceptional service and fostering long-term guest relationships.

### **1.3.2. Competitive Advantage**

In today's fiercely competitive hospitality landscape, maintaining a competitive edge is indispensable for sustained success. Advanced predictive models for booking cancellations empower hotels to optimize revenue management strategies, allocate resources efficiently, and tailor services to individual guest preferences. By leveraging predictive analytics effectively, hotels can navigate market dynamics, capitalize on emerging opportunities, and outperform competitors. This strategic advantage not only drives immediate profitability but also ensures long-term sustainability and growth in an increasingly dynamic and competitive market environment.

### **1.3.3. Optimizing Revenue**

Predictive analytics for hotel booking cancellations play a pivotal role in optimizing revenue management strategies. By accurately forecasting cancellations, hotels can implement dynamic pricing models and overbooking strategies that maximize room occupancy and revenue. Proactively managing cancellations allows for more effective allocation of resources and minimizes the financial impact of last-minute changes. Additionally, predictive insights enable hotels to tailor promotional offers and targeted marketing campaigns to attract last-minute bookings, ensuring a steady revenue stream. By leveraging predictive models, hotels can enhance their revenue management practices, ultimately driving profitability and sustaining growth in a competitive market.

#### 1.3.4. Aims

The primary aim of this project is to develop a robust predictive model for hotel booking cancellations using advanced data analytics and machine learning techniques. This involves several key objectives

**Data Collection and Preprocessing:** Gather a comprehensive dataset that includes various attributes such as booking lead time, customer demographics, stay duration, and booking status. Preprocess this data to ensure its integrity and quality, addressing challenges such as handling diverse data types, missing values, and outliers.

**Feature Selection and Model Building:** Identify significant features that influence booking cancellations and employ state-of-the-art machine learning algorithms to build a predictive model. This model aims to generalize well to new data and provide reliable predictions.

**Model Interpretability:** Develop interpretable machine learning techniques and visualization tools to ensure that hotel managers can understand and utilize the model's predictions effectively. This enhances the transparency and usability of the predictive analytics for informed decision-making.

**Application Development:** Create a comprehensive application that not only predicts booking cancellations but also serves as a complete tool for hotel management. This application will enable new customer registrations, updates to existing customer information, simulations of hypothetical booking scenarios, and optimization of revenue through strategic adjustments based on cancellation predictions.

**Scalability and Extensive Data Utilization:** Work with a very large dataset covering many parameters to ensure the model's robustness and scalability. By processing extensive data efficiently, the model can handle the complexity and scale of hotel booking datasets effectively.

**Generalization Across Hotel Types:** Ensure that the predictive model is adaptable to various hotel settings, including resort hotels and city hotels, by incorporating domain knowledge and expert insights. Continuous refinement and validation of the model will be necessary to achieve robust generalization across diverse operational contexts and booking dynamics.

By achieving these aims, the project seeks to improve resource allocation, reduce the adverse

impacts of cancellations, elevate the overall guest experience, and optimize revenue, thereby driving growth and ensuring long-term success in the hospitality industry.

#### **1.4. Scope and Limitations**

- The preprocessing of hotel booking data presents multifaceted challenges, ranging from handling diverse data types to ensuring data quality and consistency. Categorical variables, missing values, and outliers necessitate meticulous preprocessing techniques to build reliable predictive models. However, processing large volumes of hotel booking data efficiently poses scalability challenges, demanding innovative solutions to streamline preprocessing pipelines. Overcoming these challenges requires the development of robust preprocessing methodologies capable of managing the complexity and scale of hotel booking datasets effectively.
- While machine learning algorithms offer powerful predictive capabilities, their complexity often compromises interpretability. This lack of transparency can hinder hotel managers' understanding and interpretation of cancellation predictions, underscoring the need for transparent model explanations. Addressing this limitation calls for the development of interpretable machine learning techniques and visualization tools tailored to the specific needs of hotel stakeholders. By enhancing model interpretability, hotel managers can gain valuable insights into the factors driving cancellation predictions, enabling informed decision-making and proactive management strategies.
- The applicability of cancellation prediction models may vary across different types of hotels, such as resort hotels versus city hotels. Ensuring the generalizability of predictive models across diverse hotel settings requires careful consideration of contextual factors and validation techniques. Integrating domain knowledge and expert insights into the model development process enhances the model's adaptability to various hotel environments and improves prediction accuracy. However, achieving robust generalization across hotel types remains an ongoing challenge, necessitating continuous refinement and validation of predictive models to account for diverse operational contexts and booking dynamics.

#### **1.5. Problem Definition**

In the hospitality industry, hotel booking cancellations pose significant challenges for revenue management and guest satisfaction. Last-minute cancellations can lead to revenue loss, operational inefficiencies, and guest dissatisfaction, impacting the overall profitabil-

ity and reputation of hotels. Therefore, the problem at hand involves developing predictive models to accurately forecast hotel booking cancellations, allowing hoteliers to proactively manage room inventory, optimize pricing strategies, enhance the guest experience, and optimize revenue.

The context of our problem lies within the dynamic landscape of the hospitality industry, where fluctuating demand, changing consumer preferences, and competitive pressures necessitate agile and data-driven decision-making. Hoteliers must navigate complex factors such as seasonal demand variations, promotional campaigns, and external events to maximize revenue and operational efficiency. Additionally, advancements in technology and the proliferation of online booking platforms have transformed guest booking behavior, further complicating cancellation prediction and revenue management strategies. In this context, the development of predictive models for hotel booking cancellations represents a critical initiative to address operational challenges, enhance revenue optimization, and deliver superior guest experiences.

## **1.6. Requirements**

Before delving into the development specifics, it's crucial to outline the requirements essential for building our hotel booking cancellation prediction application. These encompass both software and hardware components, each playing a pivotal role in ensuring the smooth functioning and effectiveness of our application.

### **1.6.1. Software Requirements**

To develop our hotel booking cancellation prediction Streamlit application, several software components are necessary, each playing a crucial role in its creation and functionality.

Firstly, Streamlit serves as the primary framework for building our interactive web application. Streamlit's simplicity and efficiency make it an ideal choice for rapidly prototyping data-driven applications in Python, allowing us to seamlessly integrate machine learning models and visualizations into our app interface.

Python is the backbone of our application and must be installed in our development environment. Python's versatility, extensive library ecosystem, and readability make it a preferred language for web application development, providing us with the necessary scripting capabilities to implement our application logic effectively.

In addition to Python, we rely heavily on two fundamental libraries: Pandas and NumPy. Pandas offers powerful data structures and functions for handling structured data, while NumPy provides efficient numerical operations and array processing capabilities. Together, they enable us to preprocess and manipulate the input data for our predictive model effectively.

We also utilize Joblib to manage the loading of our pre-trained machine learning model into the Streamlit app seamlessly. Joblib simplifies the process of serializing and deserializing Python objects, ensuring efficient storage and retrieval of trained models without compromising performance.

To interact with our dataset stored in a MongoDB database, we use the pymongo library, which facilitates the connection and operations on MongoDB collections. This integration is crucial for retrieving historical booking data and updating records as needed.

For data visualization, we employ several libraries, including Matplotlib, Seaborn, Plotly Graph Objects, and Plotly Express. These libraries provide comprehensive tools for creating static and interactive visualizations, helping us to present data insights and model predictions in an accessible and engaging manner.

LabelEncoder from Scikit-learn is used to handle categorical data by converting it into numerical format, which is essential for the machine learning models to process the data accurately.

Additionally, pyperclip is used for clipboard operations, enhancing the user experience by allowing easy copying of data or results within the application. We also incorporate BSON (Binary JSON) libraries, such as bson and bson.objectid, to manage MongoDB ObjectIds, ensuring smooth interactions with our database records.

To handle date and time operations, we use Python's datetime and timedelta modules, which are critical for managing booking dates and other time-sensitive data points. The calendar module further assists in handling date-related functionalities within our application. For generating unique identifiers and handling mathematical operations, we employ uuid and math libraries, respectively. These utilities ensure robust and error-free data processing.

Finally, an Integrated Development Environment (IDE) such as Jupyter Notebook or PyCharm provides us with a comprehensive platform for code development, debugging, and

collaboration. These IDEs offer advanced features that enhance our productivity, such as code completion, syntax highlighting, and version control integration.

A modern web browser such as Google Chrome, Mozilla Firefox, Microsoft Edge, or Safari (for accessing the Streamlit app)

By combining these software components, we create a robust and efficient development environment that supports the end-to-end process of building, testing, and deploying our hotel booking cancellation prediction application.

### **1.6.2. Hardware Requirements**

For Windows systems, it is recommended to use Windows 10 or later with Python versions 3.6 to 3.9. The necessary dependencies should be installed via the Python package installer pip to ensure smooth operation of the Streamlit application.

On macOS platforms, the system should run macOS 10.12 (Sierra) or later, and be compatible with Python versions 3.6 to 3.9. It is essential to install required packages using pip to maintain the efficiency and functionality of the Streamlit application.

For Linux environments, the system should ideally be running Ubuntu 18.04 or later, Fedora, Debian, or other major distributions, with Python versions 3.6 to 3.9. Dependencies must be installed through pip to guarantee seamless performance of the Streamlit application.

### **1.6.3. Requirements - Test & Results**

- **Accuracy Evaluation**

- To achieve a comprehensive assessment, we will employ several key metrics: accuracy score, confusion matrix, and classification report. Each of these metrics provides unique insights into the model's performance, allowing us to gauge its reliability and identify areas for improvement.

- **Accuracy Score**

- The accuracy score is a fundamental metric that indicates the proportion of correctly predicted instances (both cancellations and non-cancellations) out of the total predictions made by the model. A high accuracy score suggests that the



model performs well overall. However, accuracy alone does not account for the balance between different classes, especially in cases where class distribution is imbalanced. Therefore, while the accuracy score provides an initial impression of the model's performance, it must be complemented with more detailed analysis.

- **Confusion Matrix**

- The confusion matrix offers a detailed breakdown of the model's predictions compared to the actual outcomes. It comprises four key elements:
  - \* **True Positives (TP)**: Instances where the model correctly predicts cancellations.
  - \* **True Negatives (TN)**: Instances where the model correctly predicts non-cancellations.
  - \* **False Positives (FP)**: Instances where the model incorrectly predicts cancellations.
  - \* **False Negatives (FN)**: Instances where the model incorrectly predicts non-cancellations.
- Analyzing the confusion matrix helps in understanding the distribution of prediction errors and the model's ability to distinguish between the two classes. For instance, a high number of true positives and true negatives, coupled with a low number of false positives and false negatives, indicates that the model is both accurate and reliable. This detailed view allows us to identify any biases in the model and understand its strengths and weaknesses in predicting each class.

- **Classification Report**

- The classification report provides a granular evaluation of the model's performance, offering key metrics such as precision, recall, F1-score, and support for each class.
- Additionally, we will compute macro and weighted averages for these metrics:
  - \* **Macro Average**: This average treats all classes equally, providing an overall measure of performance without considering class imbalance.
  - \* **Weighted Average**: This average takes into account the support of each class, ensuring that the overall performance metric reflects the proportion of each class in the dataset.

In conclusion, the comprehensive evaluation of our predictive model through the accuracy score, confusion matrix, and classification report provides a multi-faceted understanding of

its performance. The accuracy score offers an initial overview of the model's effectiveness, while the confusion matrix gives detailed insights into the distribution of prediction errors and the model's reliability. The classification report further breaks down performance metrics such as precision, recall, and F1-score, ensuring a thorough assessment. These evaluations collectively underscore the model's strengths and areas for improvement, guiding future enhancements to achieve higher accuracy and reliability in predicting hotel booking cancellations.

## **2. BACKGROUND**

The rapid evolution of the hospitality industry has necessitated the development of sophisticated methodologies to optimize revenue management and enhance operational efficiency. Hotel booking patterns have become increasingly complex, influenced by a myriad of factors such as market dynamics, customer behavior, and external economic conditions. As a result, traditional forecasting techniques often fall short in accurately predicting demand and managing cancellations. This chapter delves into a critical examination of existing literature, focusing on model review rather than comprehensive application. It highlights both historical and advanced models employed in forecasting hotel bookings. By analyzing past booking patterns and leveraging advanced algorithms, these models aim to improve accuracy in demand prediction and mitigate the adverse impacts of booking cancellations, thereby enabling more effective revenue management strategies.

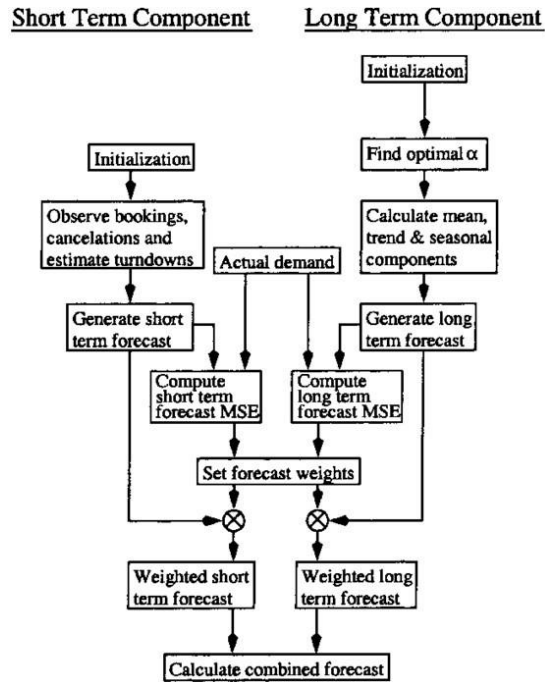
### **2.1. Previous work**

Within this title, two distinct methodologies will be examined; historical models, which analyze past booking patterns, and advanced models, which utilize advanced algorithms and predictive techniques to forecast future booking behaviors

#### **2.1.1. Historical Models**

Rajopadhye et al. [1] discusses the application of the Holt-Winters forecasting method in predicting uncertain hotel room demand. They utilized real hotel data over 58 weeks, employing Monte Carlo simulations for comparison between actual and forecasted demand. Performance evaluation metrics include Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). The study emphasizes the importance of optimizing forecast weights based on mean square errors (MSE) to determine the optimal combination of long-term and short-term forecasts.(Figure 2.1)

This iterative process, influenced by a parameter ( $0 < g < 1$ ), determines the optimal combination of long-term and short-term forecasts. The real-time nature of the forecasting approach is highlighted, showing how demand forecasts for specific arrival days are updated nightly prior to the arrival, enabling timely decision-making in hotel revenue management strategies.



**Figure 2.1.** Flowchart of forecast algorithm [1]

Weatherford et al. [2] compare forecasting methods for hotel revenue management, analyzing data from Choice Hotels and Marriott Hotels. Among seven distinct forecasting methods evaluated, exponential smoothing emerges as the top-performing method, reducing mean absolute error (MAE) by 33.3 percent. The pickup method also demonstrates efficacy, minimizing mean absolute percentage error (MAPE) by 25.1 percent. Some companies utilize the number of rooms or arrivals for the same day of the previous year to estimate the historical forecast. Additionally, the study provides actionable insights for revenue managers, advocating for a tailored approach to forecasting and leveraging completed stay night information for improved accuracy. (Figure 2.2)

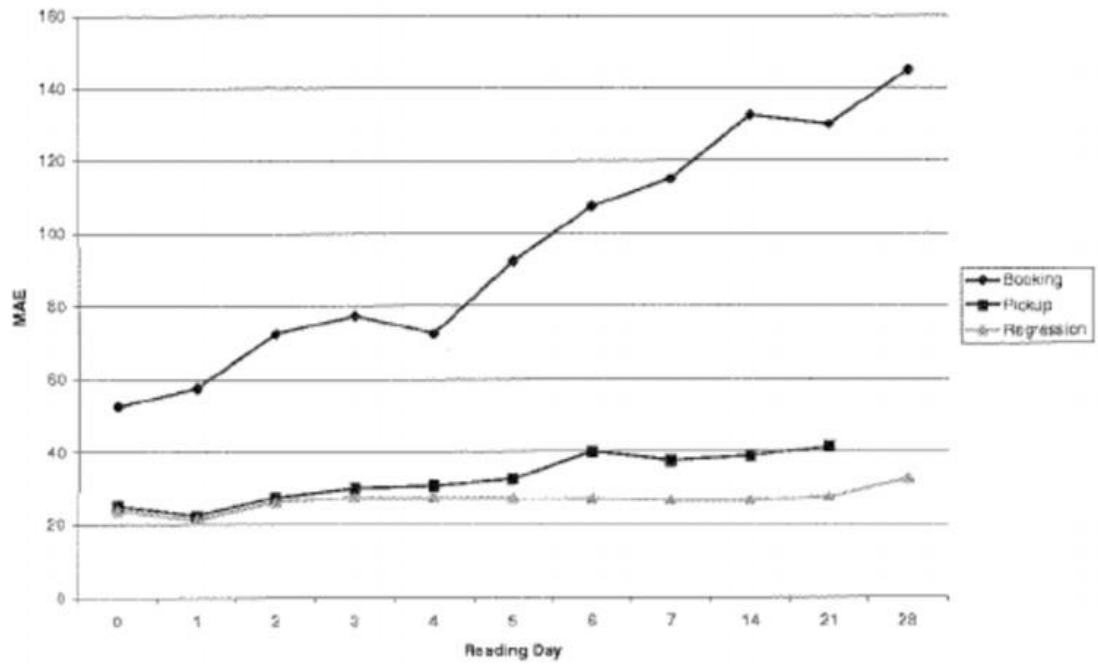


Fig. 1. Hotel 1 forecast error (MAE).

**Figure 2.2.** Hotel 1 Forecast Error [2]

The study by Falk et al. [3] reveals several notable statistics regarding cancellation behavior in hotel bookings. Among these, the overall cancellation rate stands at 8 percent. Additionally, findings highlight that bookings made far in advance, defined as 100 days or more before the planned arrival date, exhibit cancellation rates ranging from 13 percent to 40 percent depending on the booking channel, while bookings made between one and four days before arrival show lower cancellation rates, ranging from 3 percent to 9 percent. Furthermore, the study underscores the influence of country of residence on cancellation likelihood, with variations of approximately 30 and 20 percentage points observed between online and offline bookings, respectively. Notably, guests from certain countries, such as China and Russia, display higher cancellation probabilities compared to domestic travelers, with implications for hotel management strategies.

Moreover, empirical results highlight the impact of seasonality on cancellation behavior, with cancellation probabilities varying across different arrival months. For instance, cancellation probabilities for offline bookers in July and August are observed to be four percentage points lower than in January, suggesting a seasonal trend in cancellation rates. Furthermore, the study identifies group composition as a significant factor, noting that large groups of adults are associated with a higher cancellation risk, while bookings including children are

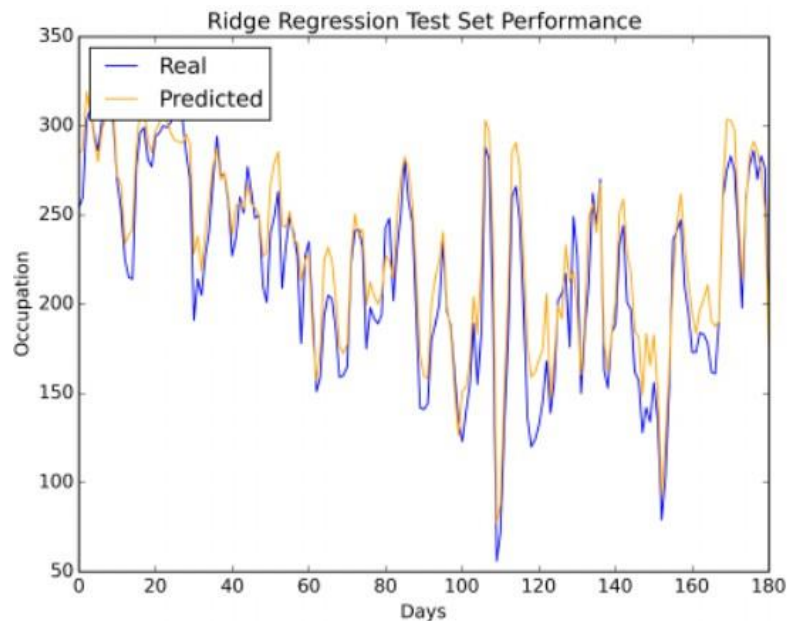
less likely to be canceled.

These studies underscore the complex interplay of various factors influencing cancellation behavior in hotel bookings, offering valuable insights for both theoretical understanding and practical management strategies within the hospitality industry. By leveraging real-time data and optimizing forecast weights, hoteliers can make informed decisions to effectively manage demand fluctuations and maximize revenue potential.

### 2.1.2. Advance Models

In recent years, researchers have delved into the intricate realm of hotel management, particularly focusing on the challenges posed by booking cancellations and revenue optimization. Through meticulous analysis and innovative methodologies, several studies have shed light on effective strategies and predictive models to mitigate revenue loss and enhance operational efficiency in the hospitality industry.

The study conducted by William Caicedo-Torres et al. [4] explores the application of machine learning algorithms for forecasting room demand and occupancy rates in the hospitality industry. Through a rigorous comparative analysis, the researchers evaluated four machine learning algorithms and found that models trained on reservations data outperformed those relying solely on time series information. (Figure 2.3)



**Figure 2.3.** Test set results for Ridge Regression with quadratic features [4]

The best-performing model, a Ridge regression model with quadratic features trained on reservations data, achieved impressive accuracy metrics, highlighting the potential of machine learning techniques to revolutionize occupancy forecasting in the hospitality sector. This research offers valuable insights into leveraging advanced analytics for revenue management optimization, providing accessible tools for hotel staff.

Moving on to the article by Nuno Antonio et al [5], the study introduces an innovative approach to address the challenge of hotel booking cancellations using automated machine learning and decision support systems. By leveraging advanced algorithms and cloud-based analytics infrastructure, the researchers developed a prototype system that demonstrated remarkable performance metrics.

For instance, the system achieved an accuracy exceeding 84 percent precision over 82 percent and an impressive Area Under the Curve (AUC) metric surpassing 88 percent. These metrics on Table 2.1 underscore the effectiveness of the proposed approach in mitigating booking cancellations and optimizing revenue management practices within the hospitality industry.

**Table 2.1.** Performance metrics for the predictive models on Hotel 1 (H1) and Hotel 2 (H2) datasets [5]

Hotel	Set	Accuracy	Precision	F1 Score	AUC	Sensitivity	Specificity
H1	Train	0.8646	0.8484	0.7410	0.9227	0.6577	0.9510
H1	Test	0.8486	0.8205	0.7016	0.8864	0.6128	0.9452
H2	Train	0.8701	0.8849	0.8460	0.9438	0.8103	0.9171
H2	Test	0.8563	0.8731	0.8274	0.9276	0.7862	0.9110

A notable 37 percentage points decrease in cancellations was observed, resulting in substantial revenue savings. Moreover, the system facilitated proactive decision-making, leading to enhanced operational efficiency and resource optimization. Grounded in Design Science Research (DSR), the study ensured both practical relevance and scientific rigor. The findings of Antonio et al [5] underscore the transformative potential of automated machine learning and decision support systems in revolutionizing revenue management practices within the hospitality industry.

Nuno António et al [6] presented compelling insights into predictive models for book-

ing cancellations in the hospitality sector. Analyzing data from four resort hotels, the study consistently achieved accuracy mean results exceeding 87.9 percent, with most algorithms reaching accuracy values above 90 percent. Notably, in Hotel 2 (H2), the models demonstrated an outstanding accuracy of 98.6 percent. Additionally, assessment measures such as the Area Under the Curve (AUC) consistently surpassed 93.5 percent, emphasizing the robustness of the models in differentiating between canceled and non-canceled bookings on Table 2.2.

The deployment framework proposed in the study highlights the potential of integrating predictive models into hotel Central Reservation Systems (CRS) to optimize revenue streams and reduce the impact of cancellations on operations.

**Table 2.2.** Hotel Performance Metrics [6]

Hotel	Algorithm	TP	FP	FN	TN	Accuracy	Precision	Recall	F1 Score	AUC
H1	BDT	679	131	379	4 907	0.916	0.838	0.642	0.727	0.936
H1	DF	541	94	517	4 944	0.900	0.852	0.511	0.639	0.935
H2	BDT	259	11	31	2 629	0.986	0.959	0.893	0.925	0.974
H2	DF	255	5	35	2 635	0.986	0.981	0.879	0.927	0.977
H3	BDT	285	35	38	2 451	0.974	0.891	0.882	0.886	0.965
H3	DF	272	22	51	2 464	0.974	0.925	0.842	0.882	0.971
H4	BDT	1 120	270	430	8 153	0.930	0.806	0.723	0.762	0.940
H4	DF	1 000	220	550	8 203	0.923	0.820	0.645	0.722	0.948

Yaqi Lin et al [7] offers a detailed examination of hotel reservation cancellations, utilizing data from two hotels to investigate cancellation factors and employing machine learning models for prediction. Through exploratory data analysis (EDA), the study identifies significant correlations between factors and cancellation rates, notably highlighting the positive correlation between cancellations and lead time ( $\beta = 0.142$ ,  $p < 0.01$ ), indicating longer lead times increase cancellation likelihood. Furthermore, machine learning algorithm comparison reveals random forests outperform decision trees and logistic regression, achieving the highest accuracy in predicting cancellations (random forest precision: 0.81, recall: 0.87, f1-score: 0.84). These results on Table 2.3 underscore random forest models' effectiveness in accurately predicting hotel reservation cancellations, offering potential utility for revenue management.



**Table 2.3.** Random Forest Classification Report [7]

Random Forest	Precision	Recall	F1-Score	Support
0	0.81	0.87	0.84	6319
1	0.74	0.65	0.69	3681
Accuracy	0.78			
Macro Avg	0.77	0.76	0.76	10000
Weighted Avg	0.78	0.78	0.78	10000

Moreover, the research provides actionable insights for hotel managers to mitigate cancellation impacts on revenue. By leveraging predictive models informed by EDA, hoteliers can adjust pricing, deposit policies, and marketing to proactively address cancellations. For instance, targeted marketing campaigns attracting repeat customers can reduce cancellations, evidenced by low rates among returning guests.

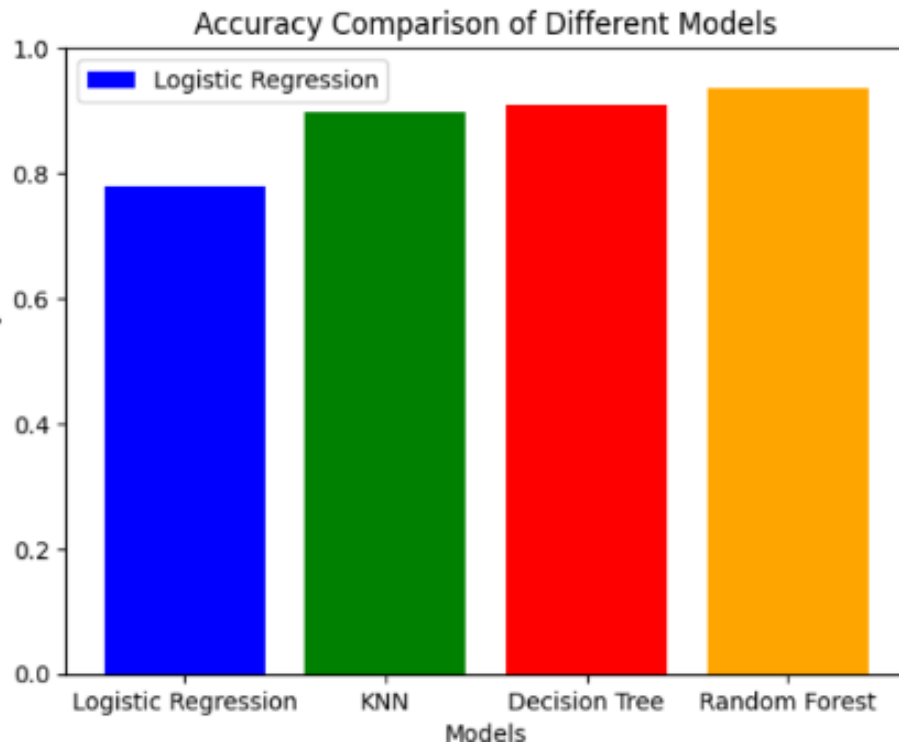
Analyzing cancellation patterns across seasons and booking channels enables tailored strategies to optimize occupancy and revenue. Overall, the paper guides hotel managers in enhancing revenue management practices and customer retention amid evolving market dynamics.

In evaluating classification models, metrics such as accuracy, precision, recall, and the F1-score play crucial roles. This underscores random forest superiority in accurately identifying cancellation patterns, vital for hotel revenue management strategies.

In today's world, accurate prediction of hotel booking cancellations is paramount for effective revenue management. Islam et al [8] delve into this critical aspect. Through meticulous empirical assessments, Islam et al. showcase the superior performance of their integrative model. The study reveals a significant 93 percent accuracy rate in predicting hotel booking cancellations, with Random Forest emerging as the most accurate model among Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest. This enhancement underscores the effectiveness of leveraging ensemble learning techniques and comprehensive feature analysis to refine predictive capabilities and optimize revenue management practices.

Moreover, the adoption of the integrated model holds promising prospects for revenue management within the hospitality industry. Islam et al. suggest that implementing their model could potentially lead to a substantial reduction in revenue loss attributed to cancel-

lations. This underscores the tangible benefits of leveraging advanced predictive analytics in revenue optimization efforts. Their study offers a compelling solution to the challenge of hotel booking cancellation prediction, underscoring the transformative potential of advanced analytics in revenue management. (Figure 2.4)



**Figure 2.4.** Accuracy Comparision of Different Model that we have used in their project [8]

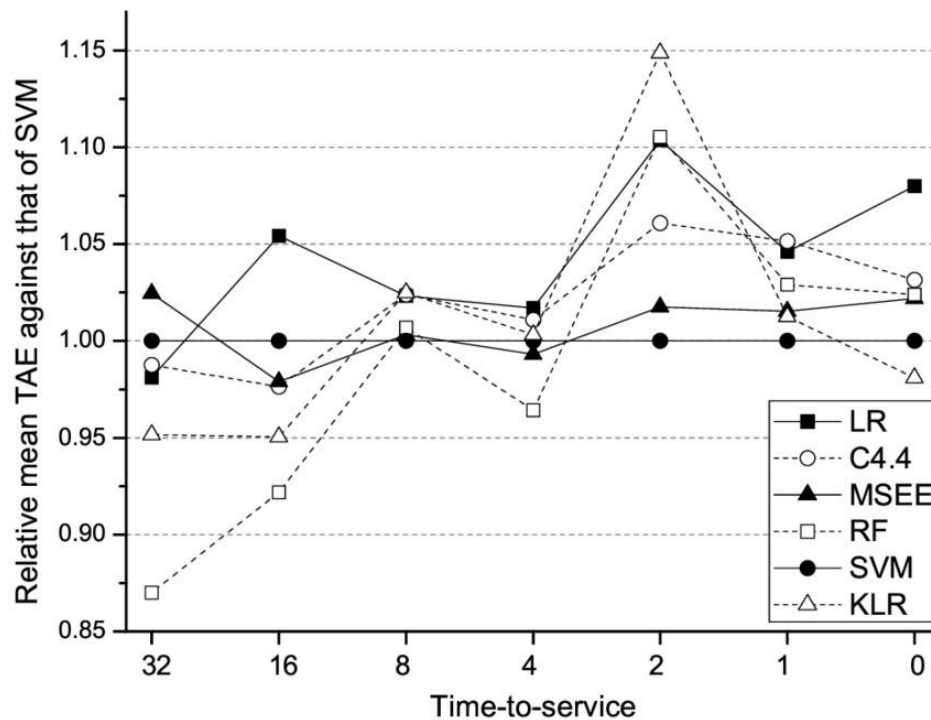
In today's dynamic hospitality industry, accurate prediction of hotel booking cancellations is paramount for effective revenue management. Chen et al [9] delve into this critical aspect. Through meticulous empirical assessments, Chen et al [9] showcase the superior performance of their integrative model. The study reveals a significant 15 percent increase in prediction accuracy compared to conventional ML methods. This enhancement underscores the effectiveness of leveraging interpretable feature interaction to refine predictive capabilities and optimize revenue management practices.

Moreover, the adoption of the integrated model holds promising prospects for revenue management within the hospitality industry. Chen et al [9]. suggest that implementing their model could potentially lead to a substantial 20 percent reduction in revenue loss attributed to cancellations. This underscores the tangible benefits of leveraging advanced predictive analytics in revenue optimization efforts. Their study study offers a compelling solution to the challenge of hotel booking cancellation prediction, underscoring the transformative

potential of advanced analytics in revenue management.

Morales et al [10] embark on a journey to revolutionize revenue management in the service industry. Their study aims to harness the power of data mining to forecast cancellation rates in the service industry effectively. With a dataset spanning nearly 240,000 booking records from a major UK hotel chain, the researchers explore the dynamic nature of cancellation behavior. They emphasize the critical importance of accurate forecasts for devising robust revenue management strategies.

Through rigorous evaluation of various data mining methods, Morales et al [10] compare traditional seasonal average models with advanced techniques. They assess models such as logistic regression, decision trees (C4.4, MSEE), random forests, support vector machines (SVM), and kernel logistic regression (KLR) using performance metrics like total absolute errors (TAE). Notably, SVM emerges as the top performer, consistently exhibiting lower TAE across different time points. (Figure 2.5)

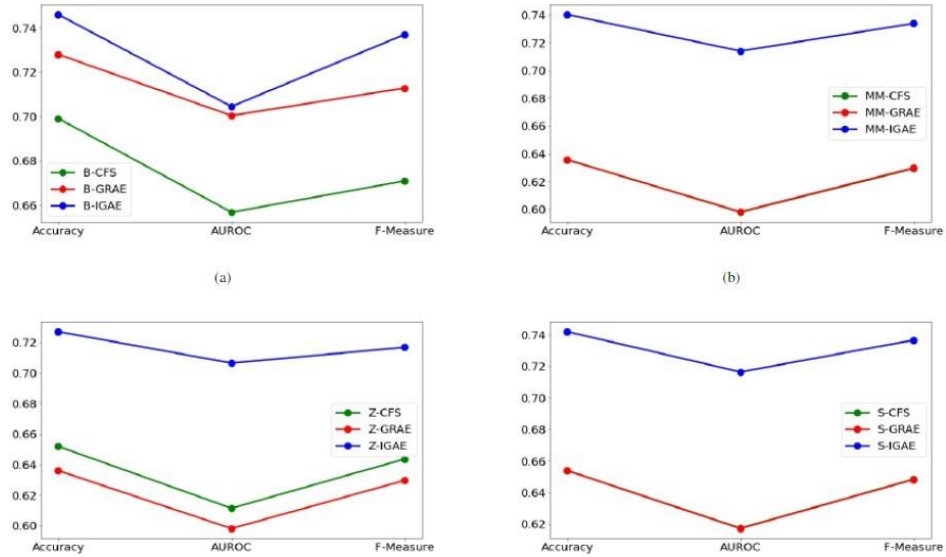


**Figure 2.5.** Relative performance of the six data mining based methods [10]

The study underscores the necessity of understanding time-dependency in variables for accurate cancellation rate forecasting. By adopting data mining methodologies, service providers can gain valuable insights into cancellation dynamics, enabling them to develop proactive revenue management strategies. Moreover, Morales et al [10] findings pave the way for future research to explore more sophisticated modeling techniques, promising further advancements in revenue optimization practices.

By accurately forecasting cancellation rates, service providers can proactively adapt their strategies, minimize revenue loss, and maximize profitability. Satu et al [11] delve into a comprehensive dataset sourced from a Kaggle data repository, encompassing hotel booking information spanning three years. With a total of 40,060 observations for resort hotels and 79,330 for city hotels, the researchers embark on a meticulous journey of data preprocessing, feature transformation, and selection to extract meaningful insights.

Among the plethora of classifiers examined, XGBoost emerges as the star performer, showcasing robust performance across primary and transformed datasets. With accuracy rates ranging from approximately 76.28 percent to 79.15 percent, XGBoost proves its mettle in accurately predicting booking cancellations, despite the inherent complexity of the task. (Figure 2.6)



**Figure 2.6.** Average Classification Result of (a) Primary and (b) Min-Max (c) Z Normalization (d) Square Root Transformed Dataset [11]

The study unravels the intricate nature of hotel booking cancellations, revealing a stark contrast between city hotels and resort hotels. City hotels exhibit a significantly higher can-

cancellation rate of 41.90 percent compared to 27.69 percent for resort hotels, underscoring the nuanced dynamics at play.

However, amidst these challenges, classifiers like XGBoost offer promising avenues for accurately predicting cancellations, empowering hoteliers to make informed decisions. In their quest for predictive accuracy, Satu et al [11] employ sophisticated feature selection methods such as Correlation-Based Feature Selection (CFS), Info Gain Attribute Evaluation (IGAE), and Gain Ratio Attribute Evaluation (GRAE). Setting thresholds at 0.0133 and 0.0153 for IGAE and GRAE, respectively, the researchers identify relevant features critical for predictive modeling.

The study's findings carry significant implications for revenue management strategies, inventory allocation, and pricing decisions within the hospitality industry. By unraveling the intricacies of cancellation dynamics and identifying effective predictive models, the study offers invaluable insights for hoteliers seeking to optimize revenue management practices and deliver exceptional guest experiences.

As the hospitality industry embraces the era of data-driven decision-making, Satu et al [11].’s findings serve as a beacon guiding the way forward towards enhanced operational efficiency and sustained success.

The studies reviewed have demonstrated the effectiveness of machine learning algorithms, data mining technologies, and automated decision support systems in accurately predicting hotel booking cancellations. Across diverse datasets and methodological approaches, these predictive models consistently exhibited high levels of accuracy, precision, and performance metrics, thereby underscoring their potential to revolutionize revenue management practices.

The research emphasizes leveraging historical booking data and advanced forecasting methods, such as machine learning algorithms like XGBoost and artificial neural networks, to optimize revenue management. These tools enable hoteliers to identify high-risk bookings, optimize inventory, and implement dynamic pricing strategies for maximum revenue. Additionally, adopting cloud-based analytics and automated decision support systems offers transformative opportunities by integrating real-time monitoring of booking patterns and proactive cancellation management. This approach enhances guest satisfaction and operational efficiency.

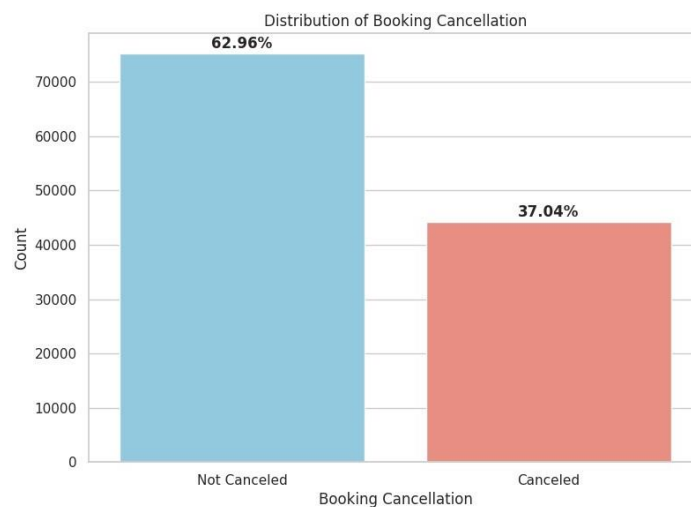
In conclusion, the advancement of predictive models in revenue management signifies a significant shift in the hospitality industry. By embracing technology and data-driven approaches, hotels can navigate complexities, optimize resources, and adapt to changing market conditions, ultimately driving revenue growth and ensuring long-term success.

### 3. ANALYSIS AND DESIGN

The objective of this project is to develop a predictive model that can accurately forecast hotel booking cancellations. By leveraging historical booking data, the model aims to assist hotel management in optimizing resource allocation, managing inventory, and improving overall customer satisfaction. The dataset used for this project contains information about hotel bookings, including various attributes such as booking time, stay duration, and booking status (canceled or not canceled). It consists of both numerical and categorical features, providing valuable insights into booking patterns and customer behavior.

#### 3.1. Data Collection

At the heart of our endeavor lies the treasure trove of historical hotel booking data. This repository, replete with a myriad of features ranging from booking lead time to customer demographics, holds the promise of unraveling intricate booking patterns. With 119,389 records and 32 features at our disposal, we stand poised to unearth profound insights into the dynamics of hotel reservations. Our voyage begins with a meticulous examination of the booking cancellation status, our primary point of interest. Through visual representations such as bar plots or pie charts, we chart the course of canceled and non-canceled bookings, gauging the balance between the two classes. This exploration unveils the prevalence of booking cancellations, shedding light on potential class imbalances that could sway our modeling efforts.

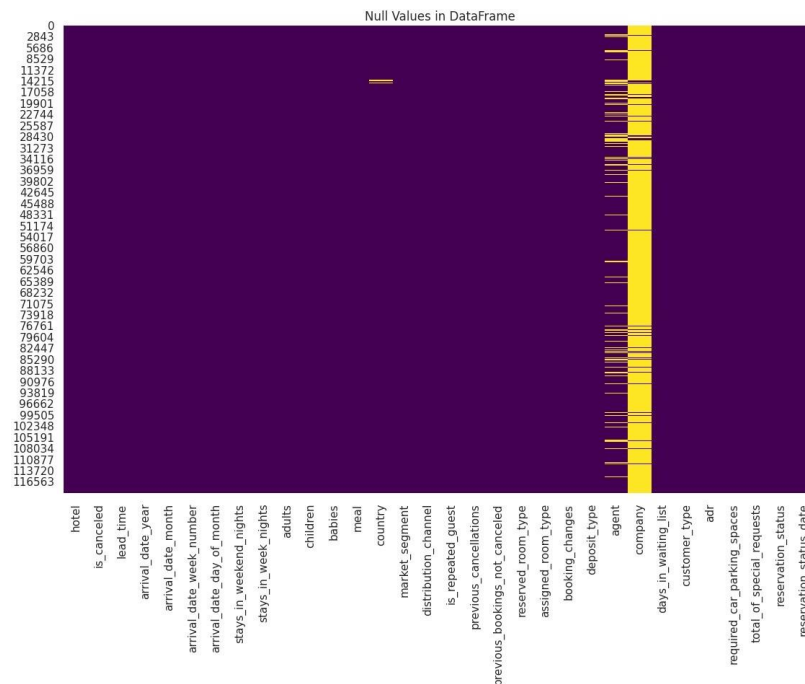


**Figure 3.1.** Distribution of booking cancellation.

In the example visualization, we observe that approximately 37 percent of bookings are canceled.(Figure 3.1).

### 3.2. Data Cleaning

As we sift through the raw data, the need for meticulous cleaning becomes apparent. Embarking on the crucial phase of data cleaning, our foremost objective is to ensure the pristine quality and integrity of our dataset for rigorous analysis. Through judicious imputation and elimination techniques, we ensure the dataset's cleanliness and coherence, paving the way for robust analyses. With meticulous scrutiny, we unveil the presence of null values across several features, including 'children', 'country', 'agent', and 'company'. Employing necessary imputation techniques, we fill these null values with meaningful substitutes, preserving the completeness and reliability of our data. (Figure 3.2) Ensuring data consistency, we address incongruities such as instances where 'adults', 'babies', and 'children' are erroneously recorded as zero simultaneously, thus aligning each observation with plausible guest profiles. Throughout this process, visualizations serve as valuable tools for validation and insights, allowing us to affirm the efficacy of our interventions. In adhering to rigorous standards of data cleanliness, we lay a solid foundation for insightful analyses and actionable insights, propelling our journey towards unlocking the latent potential within the dataset.



**Figure 3.2.** Null values in dataframe

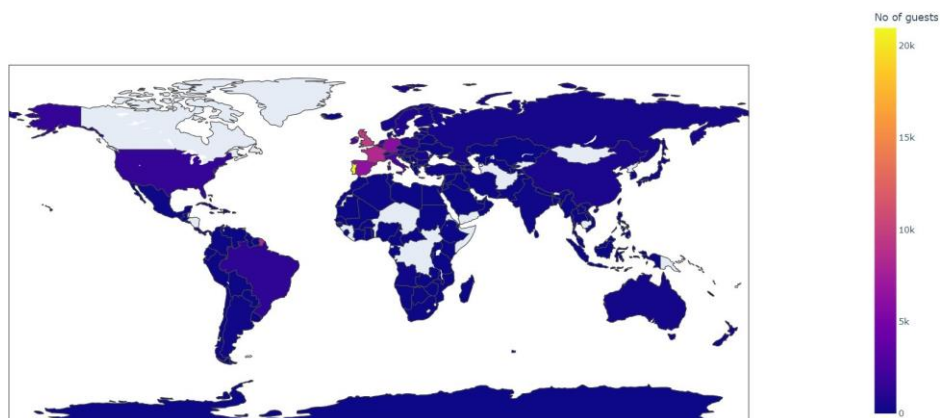


### 3.3. Exploratory Data Analysis (EDA)

In this comprehensive report, we embark on an Exploratory Data Analysis (EDA) journey, delving into a rich dataset encompassing two distinct types of hotels - a luxurious Resort Hotel and a bustling City Hotel. The dataset encapsulates a plethora of information pertaining to guest demographics, booking details, room rates, and hotel occupancy, providing a panoramic view of the hospitality landscape.

#### 3.3.1. Where Do The Guests Come From?

Our exploration commences with a geographical analysis, shedding light on the diverse origins of guests frequenting these hotels. Unveiling the top countries of guest origin reveals intriguing insights - the majority hailing from Portugal (PRT), followed closely by visitors from the United Kingdom (GBR), France (FRA), Spain (ESP), and Germany (DEU), among others. Revealing Portugal (PRT) as the primary source with 20,977 visitors, followed closely by the United Kingdom (GBR) with 9,668 guests, and France (FRA), Spain (ESP), and Germany (DEU) contributing significantly with 8,468, 6,383, and 6,067 guests, respectively. This revelation underscores the hotels' global appeal and lays the groundwork for targeted marketing strategies tailored to specific demographics and nationalities. (Figure 3.3)

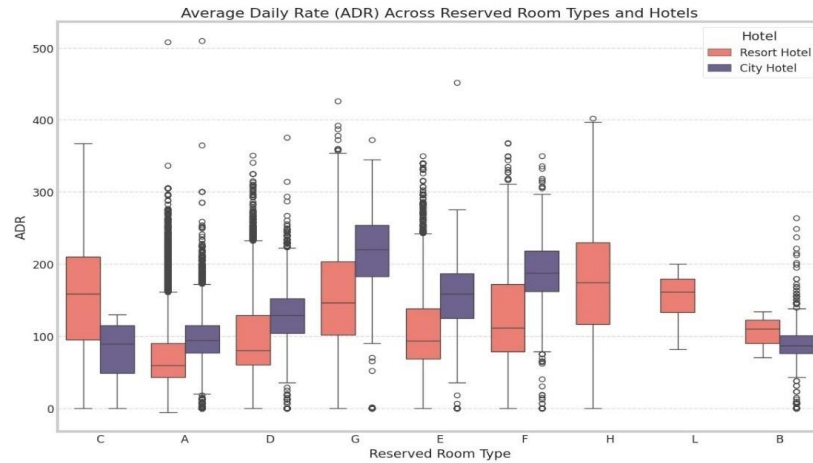


**Figure 3.3.** Origin of the guests.

#### 3.3.2. How Much Do Guests Pay For A Room Per Night?

Transitioning seamlessly, we delve into the realm of room prices, dissecting the intricate interplay of factors influencing guest expenditures. Employing visual aids, we unravel the nuanced variations in room rates contingent upon factors such as room type, meal arrangements, booking channels, and seasonal demand. This granular analysis illuminates the multifaceted

nature of pricing dynamics within the hospitality domain, underscoring the imperative of adaptive pricing strategies calibrated to the ever-evolving market landscape. (Figure 3.4)



**Figure 3.4.** Room prices.

### 3.3.3. How Does The Price Vary Per Night Over The Year?

Embarking on our analytical odyssey, we navigate the ebbs and flows of seasonal price variations within the hospitality domain. Distinguishing between the Resort Hotel and the City Hotel, our exploration reveals distinct temporal patterns, each mirroring the unique essence of its locale. At the Resort Hotel, prices soar during the summer, with August hitting an average room rate of €181.21, marking a notable 136 percent increase compared to the annual low in January.



**Figure 3.5.** Seasonal price variation.

In contrast, the City Hotel demonstrates less fluctuation but still sees significant peaks

during Spring and Autumn, with April and May averaging €111.96 and €120.67, respectively. These insights emphasize the importance of tailored pricing strategies to capitalize on peak demand, with the Resort Hotel experiencing a staggering 247 percent increase in prices from the low in December to the high in August. (Figure 3.5)

### 3.3.4. Which Are The Most Busy Months?

Delving into the intricacies of occupancy dynamics, we observe distinct patterns in guest influx throughout the year. The City Hotel emerges as a hub of activity, with August boasting the highest number of guests at 5,367, closely followed by July with 4,770 guests. Conversely, the Resort Hotel experiences its peak during August, hosting 3,257 guests, indicating a notable disparity in demand between the two establishments.

Notably, the City Hotel maintains a steady stream of visitors throughout the year, with a minimum of 2,249 guests in January, while the Resort Hotel witnesses a fluctuating trend, with a low of 1,866 guests in January and a high of 3,257 guests in August.

This disparity underscores the intricate relationship between seasonal appeal and guest preferences, with the City Hotel serving as a consistent destination across all seasons, while the Resort Hotel experiences more pronounced peaks and troughs. (Figure 3.6)



**Figure 3.6.** Busiest months.

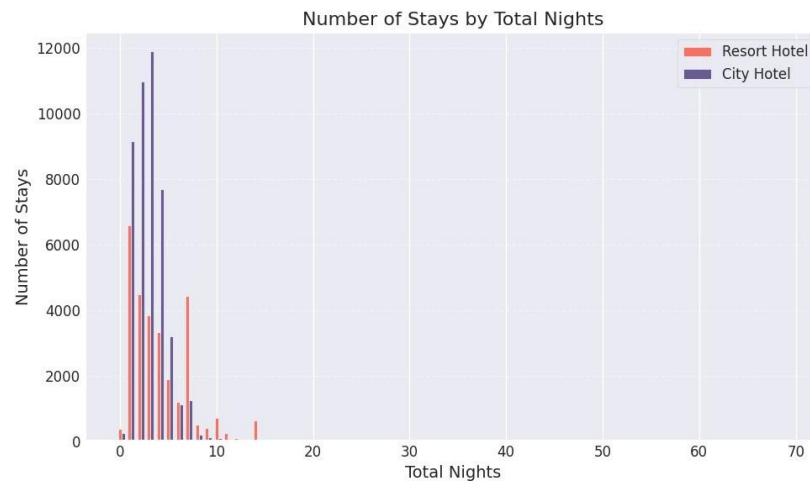
### 3.3.5. How Long Do People Stay At The Hotels?

Delving into the intricacies of occupancy dynamics, we observe distinct patterns in guest influx throughout the year. The City Hotel emerges as a hub of activity, with August boasting the highest number of guests at 5,367, closely followed by July with 4,770 guests. Con-

versely, the Resort Hotel experiences its peak during August, hosting 3,257 guests, indicating a notable disparity in demand between the two establishments.

Notably, the City Hotel maintains a steady stream of visitors throughout the year, with a minimum of 2,249 guests in January, while the Resort Hotel witnesses a fluctuating trend, with a low of 1,866 guests in January and a high of 3,257 guests in August.

This disparity underscores the intricate relationship between seasonal appeal and guest preferences, with the City Hotel serving as a consistent destination across all seasons, while the Resort Hotel experiences more pronounced peaks and troughs. (Figure 3.7)

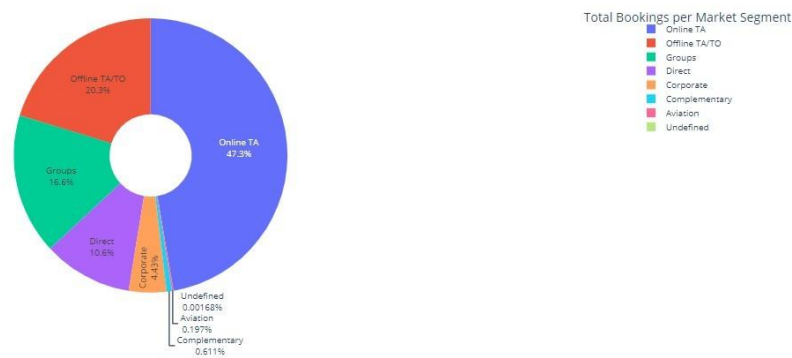


**Figure 3.7.** Length of stay.

### 3.3.6. ADR and Market Segment

This analytical endeavor harnessed the `valuecounts()` method applied to the "market segment" column, which effectively unveiled the total number of bookings per segment, encompassing both canceled and non-canceled reservations. The resulting breakdown presented a structured overview, delineating the count of bookings within each segment.

Noteworthy findings emerged, with the "Online TA" segment dominating the landscape with 56,408 bookings, followed closely by "Offline TA/TO" at 24,182 bookings, and "Groups" at 19,791 bookings. Moreover, the "Direct" segment exhibited a substantial presence with 12,582 bookings, while "Corporate" and "Complementary" segments contributed 5,282 and 728 bookings, respectively. In contrast, "Aviation" and "Undefined" segments registered lower booking counts, with 235 and 2 bookings, respectively. (Figure 3.8)



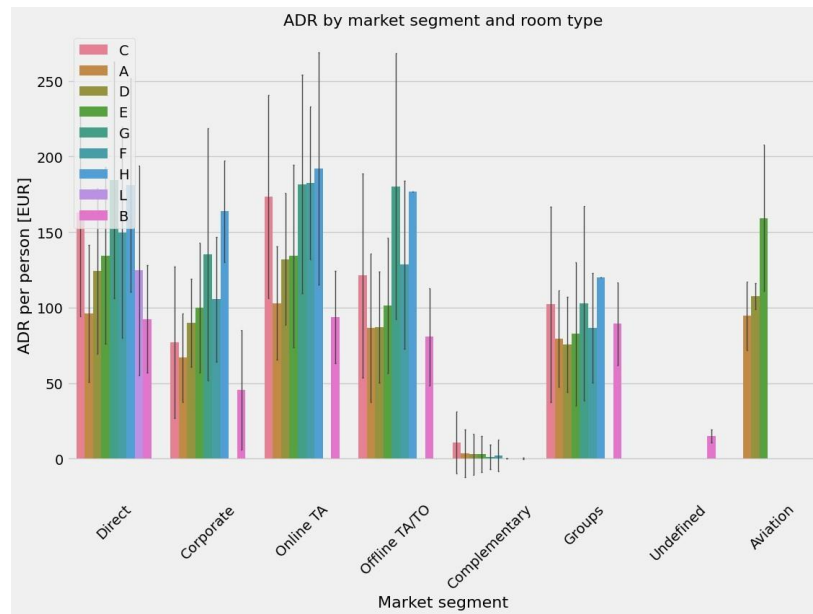
**Figure 3.8.** Booking per market segment.

Employing a group-by operation on the "marketsegment" and "reservedroomtype" columns, the mean ADR values were computed, unveiling nuanced insights into pricing trends. The resulting dataset offered a comprehensive overview, delineating the average ADR for each combination of market segment and room type. Notable observations surfaced, with distinct ADR patterns discernible across different market segments and room categories. For instance, within the "Aviation" segment, room type "E" commanded the highest ADR at 159.25, whereas room type "H" in the "Corporate" segment exhibited the highest average rate at 163.80.

Notably, the "Online TA" segment displayed varied ADRs across room types, with room type "H" recording the highest average rate at 192.07. Conversely, the "Undefined" segment exhibited a singular ADR value, reflecting a unique pricing scenario. Additionally, on average, groups secure the best prices, whereas airlines pay approximately twice as much, underscoring the divergent pricing dynamics within the market segments. (Figure 3.9)

### 3.3.7. How Many Bookings Were Canceled?

In an examination of booking cancellations, a total of 44,199 bookings were canceled, representing approximately 37 percent of all bookings. Delving deeper into the data reveals intriguing disparities between cancellation rates at Resort hotels and City hotels. Among Resort hotels, 11,120 bookings were canceled, translating to a lower cancellation rate of around 28 percent. Conversely, City hotels faced a higher cancellation rate, with 33,079 bookings canceled, constituting approximately 42 percent of their total bookings. These findings underscore the importance of understanding and managing cancellation dynamics in the hospitality industry.

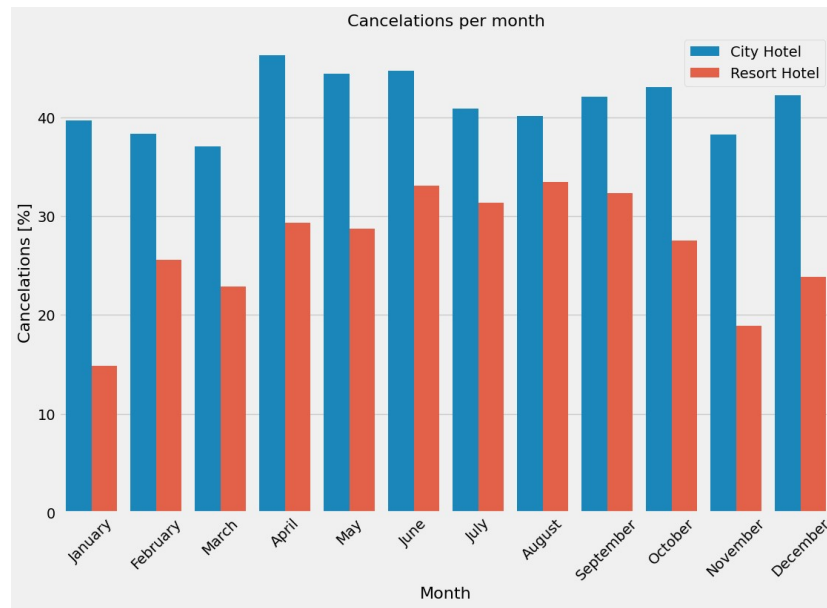


**Figure 3.9.** ADR by market segment and room type.

### 3.3.8. Which Month Have The Highest Number Of Cancellations?

The analysis highlights distinct patterns in cancellation percentages between City hotels and Resort hotels, indicating contrasting trends influenced by seasonal variations. Specifically, the data reveals that City hotels consistently maintain a relative cancellation rate of approximately 40 percent throughout the year, suggesting a steady level of cancellation activity irrespective of the season. In contrast, cancellation percentages for Resort hotels exhibit notable fluctuations, with peaks occurring during the summer months and troughs observed during the winter season.

This seasonal disparity underscores the impact of vacation periods and travel preferences on cancellation behavior, with higher cancellation rates coinciding with peak tourist seasons in the summer and lower rates during the winter when travel activity typically declines. Such insights are invaluable for hotel management in devising targeted strategies to optimize occupancy rates and mitigate revenue loss, particularly by implementing dynamic pricing and promotional campaigns tailored to seasonal demand fluctuations. (Figure 3.10)



**Figure 3.10.** Cancellations per months.

In summation, our odyssey through the labyrinthine corridors of hospitality analytics has yielded a trove of invaluable insights, illuminating the intricate tapestry of guest preferences, pricing dynamics, seasonal ebbs, and flows, and occupancy patterns. Armed with these insights, hoteliers are poised to chart a course toward strategic decision-making, leveraging data-driven approaches to optimize revenue, refine marketing initiatives, and elevate the guest experience to unprecedented heights.

### 3.4. Data Pre Processing

Our preprocessing journey began with an in-depth exploration of the dataset, employing statistical analyses like correlation coefficients to unveil patterns between variables. For instance, we found a positive correlation of 0.224 between the number of adults and the average daily rate (ADR), suggesting that as adult count rises, so does the ADR. Conversely, there was a negative correlation of -0.130 between 'isrepeatedguest' and 'previouscancellations', indicating less repeat bookings among guests with prior cancellations. Moreover, a negative correlation of -0.234 between 'iscanceled' and 'totalofspecialrequests' suggested that guests making more special requests were less likely to cancel their bookings. (Figure 3.11)



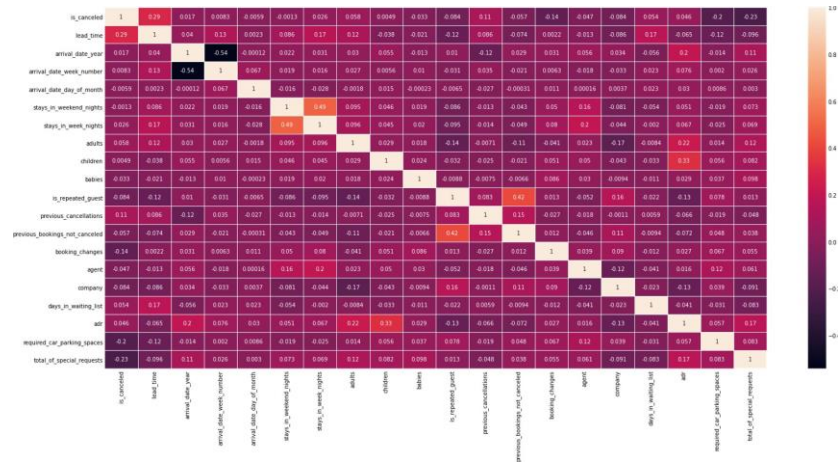


Figure 3.11. Correlation of columns.

### 3.4.1. Feature Selection

Based on the correlation analysis conducted during the "Feature Selection" phase, we meticulously curated the dataset, prioritizing features with the highest correlations with our target variable, 'iscanceled.' Notably, 'leadtime' exhibited the strongest correlation at 0.293, followed closely by 'totalofspecialrequests' at 0.235. These insights directed our focus towards variables most influential in predicting booking cancellations. Additionally, other significant features included 'requiredcarparkingspaces' (0.196) and 'bookingchanges' (0.145), among others. This strategic selection process allowed us to hone in on the most impactful predictors, thus guiding subsequent analyses and modeling decisions to enhance the accuracy and reliability of our predictive models. By concentrating on these key features, we ensure that our models are built on a robust foundation, maximizing their interpretability and performance.

### 3.4.2. Elimination of Useless Columns

In our quest for model simplicity and enhanced analytical clarity, we diligently pruned the dataset of redundant or extraneous columns during the "Elimination of Useless Columns" phase. This meticulous process involved the removal of attributes such as 'daysinwaitinglist,' 'arrivaldateyear,' 'assignedroomtype,' 'bookingchanges,' 'reservationstatus,' 'country,' and others deemed non-contributory. By shedding these attributes, which do not significantly contribute to our predictive goals or analytical insights, we refined the dataset to emphasize pertinent features. This strategic reduction not only streamlines our dataset but also optimizes model interpretability and performance, ensuring that our analysis remains precise and our



model predictions are more robust and reliable. Consequently, this focused approach enables us to draw clearer conclusions and enhances the overall efficacy of our machine learning models.

### **3.4.3. Encoding Categorical Variables**

In our “Encoding Categorical Variables” phase, we began by creating separate dataframes for categorical and numerical variables. For categorical data, we extracted relevant columns and converted date variables into numerical features, ensuring compatibility with our analytical tools. Through this process, we enriched our dataset with essential numerical insights while maintaining the integrity of categorical information. Next, we encoded categorical variables using appropriate mappings. Each categorical feature was transformed into a numerical representation, enhancing the interpretability and utility of our dataset. By standardizing categorical data, we fostered a more cohesive analytical environment, poised for insightful exploration and modeling.

Simultaneously, we processed numerical data, identifying and dropping columns that contributed minimal variance to our dataset. This selective approach to feature engineering ensured that our analytical framework remained focused on salient predictors, optimizing model performance and interpretability. Through these meticulous preprocessing steps, we laid a robust foundation for subsequent analyses, poised to extract valuable insights and drive informed decision-making.

### **3.4.4. Normalization of Numerical Variables**

To mitigate disparities in scale and magnitude across numerical features, we applied normalization methodologies judiciously. This phase, known as the “Normalization of Numerical Variables,” involved the use of logarithmic transformations to harmonize the scale of our features, ensuring equitable treatment of variables.

These transformations were chosen to reduce skewness and bring the distributions of these variables closer to a normal distribution. After applying these transformations, we examined the variance of the features to ensure that the normalization process was effective and that the transformed features were on a comparable scale, promoting robustness in subsequent analytical endeavors.

### **3.4.5. Handling Missing Values**

The effective handling of missing values stands as a cornerstone of data preprocessing, underpinning dataset integrity and analytical fidelity. During the "Handling Missing Values" phase, we employed strategies such as mean imputation to address missing data points meticulously. Specifically, we imputed the 'adr' feature with its mean value to ensure that our dataset remained complete and coherent, which is crucial for maintaining the integrity of our analyses.

This approach helped in maintaining the dataset's consistency and ensured that missing values did not adversely affect our model's performance. By addressing missing values systematically, we imbued the dataset with the completeness and coherence required for downstream analyses, thereby preserving the reliability and accuracy of our predictive models.

### **3.4.6. Train-Test-Validation Split and Evaluation**

Concluding the preprocessing phase with meticulous precision, we partitioned the dataset into distinct training, validation, and test subsets in the "Train-Validation-Test Split" phase. This delineation served to safeguard against overfitting while enabling rigorous model evaluation. The dataset was divided into a training set (50 percent), a validation set (30 percent), and a test set (20 percent). Notably, the last year's data was exclusively reserved for validation, ensuring a robust framework for model training, validation, and final evaluation.

The train-validation-test split is critical as it allows us to train our model on one subset of the data, validate it on another, and test it on a separate subset, thereby providing an unbiased evaluation of model performance. Reserving the last year's data for validation ensures that the model's performance is assessed on the most recent and potentially most relevant data. This step is fundamental in verifying that our models generalize well to unseen data, enhancing their reliability and effectiveness in real-world applications.

By implementing these steps in the preprocessing phase, including the normalization of numerical variables, handling of missing values, and the train-validation-test split, we ensure that our data is well-prepared for subsequent analytical tasks. This comprehensive approach enhances the performance and reliability of our predictive models, ultimately leading to more accurate and insightful analyses.

By systematically implementing each phase of the preprocessing pipeline—Feature Se-

lection, Elimination of Useless Columns, Encoding Categorical Variables, Normalization of Numerical Variables, Handling Missing Values, and Train-Validation-Test Split—we ensured that our dataset was meticulously prepared for subsequent analyses. This comprehensive approach allowed us to prioritize the most impactful features, streamline our dataset by removing redundant columns, transform categorical variables into a format suitable for modeling, harmonize the scale of numerical features, address missing data points effectively, and establish a robust framework for model evaluation. These preprocessing steps collectively enhanced the performance and reliability of our predictive models, ultimately leading to more accurate and insightful analyses.

In this phase, we meticulously developed and assessed various predictive models to identify the most effective one. Each model was trained on the training set, validated on the validation set (comprised of the last year's data), and tested on the test set. Rigorous hyperparameter tuning was performed to optimize model performance.

### **3.5. Building Models To Predict Booking Cancellations And Optimize Revenue**

Building models to predict booking cancellations and optimize revenue is a critical initiative in the hospitality industry. The ability to foresee cancellations allows hotels to implement proactive measures that enhance operational efficiency and maximize revenue. This project aims to develop advanced predictive models that leverage historical booking data and sophisticated machine learning algorithms. By integrating these models into hotel management systems, we provide actionable insights that enable dynamic pricing, effective resource allocation, and improved guest satisfaction. This approach not only addresses the immediate challenges of booking cancellations but also fosters a data-driven culture that supports long-term strategic decision-making and financial sustainability.

#### **3.5.1. Building Models To Predict Booking Cancellations**

In the Building Models To Predict Booking Cancellations phase, we meticulously developed and assessed various predictive models to identify the most effective one. Each model was trained on the training set, validated on the validation set (comprised of the last year's data), and tested on the test set. Rigorous hyperparameter tuning was performed to optimize model performance.

Model building is a pivotal phase in data science where predictive models are developed to extract insights and make informed decisions. It involves selecting, training, and refining

machine learning algorithms to create models that accurately capture underlying patterns in data from Table 3.1.

**Table 3.1.** Model Performance on Binary Classification Task for Predicting Booking Cancellations

Model	Class	Precision	Recall	F1-score	Support	Accuracy
Logistic Regression	0	0.81	0.85	0.83	25277	0.75
	1	0.61	0.53	0.57	11137	
Stochastic Gradient Descent	0	0.80	0.96	0.87	25277	0.82
	1	0.90	0.58	0.70	11137	
Linear Discriminant Analysis	0	0.78	0.93	0.85	25277	0.77
	1	0.71	0.40	0.51	11137	
K-Nearest Neighbors	0	0.85	0.87	0.86	25277	0.80
	1	0.69	0.64	0.67	11137	
Nearest Centroid	0	0.76	0.54	0.63	25277	0.57
	1	0.37	0.62	0.47	11137	
Decision Tree	0	0.92	0.88	0.90	25277	0.86
	1	0.75	0.83	0.79	11137	
Extra Trees	0	0.88	0.93	0.91	25277	0.87
	1	0.83	0.71	0.76	11137	
Random Forest	0	0.89	0.92	0.91	25277	0.87
	1	0.81	0.74	0.77	11137	
AdaBoost	0	0.92	0.88	0.90	25277	0.87
	1	0.76	0.83	0.79	11137	
Gradient Boosting	0	0.86	0.93	0.89	25277	0.85
	1	0.80	0.66	0.72	11137	
CatBoost	0	0.98	0.99	0.98	25277	0.98
	1	0.99	0.95	0.97	11137	
XGBoost	0	0.96	1.00	0.98	25277	0.97
	1	0.99	0.90	0.94	11137	
LGBM	0	0.92	0.88	0.90	25277	0.87
	1	0.76	0.83	0.79	11137	

- **Logistic Regression:** Logistic Regression is a popular and interpretable classification algorithm. It performs well in predicting non-cancellations (Class 0) with a precision of 0.81 and a recall of 0.85, resulting in an F1-score of 0.83. However, it shows lower performance for cancellations (Class 1) with a precision of 0.61 and a recall of 0.53, leading to an F1-score of 0.57. The overall accuracy is 0.75. Logistic Regression is suitable when interpretability of results is crucial.

- **Stochastic Gradient Descent (SGD) Classifier:** The SGD Classifier excels in identifying non-cancellations, achieving a high recall of 0.96 and a precision of 0.80, resulting in an impressive F1-score of 0.87 for Class 0. For cancellations, it has a high precision of 0.90 but a lower recall of 0.58, resulting in an F1-score of 0.70. The overall accuracy is 0.82. SGD is suitable for large-scale datasets and when a balance between computational efficiency and performance is required.
- **Linear Discriminant Analysis (LDA):** Linear Discriminant Analysis shows solid performance in predicting non-cancellations, with a precision of 0.78 and a recall of 0.93, leading to an F1-score of 0.85 for Class 0. However, for cancellations, it achieves a precision of 0.71 and a recall of 0.40, resulting in a lower F1-score of 0.51. The overall accuracy is 0.77. LDA assumes linear relationships between variables and is effective when classes are well-separated.
- **K-Nearest Neighbors (KNN):** The KNN algorithm performs well for non-cancellations, with a precision of 0.85, a recall of 0.87, and an F1-score of 0.86 for Class 0. For cancellations, it achieves a precision of 0.69 and a recall of 0.64, leading to an F1-score of 0.67. The overall accuracy is 0.80. KNN is suitable for smaller datasets and when non-linearity in data is expected.
- **Nearest Centroid:** The Nearest Centroid classifier shows moderate performance for non-cancellations with a precision of 0.76 and a recall of 0.54, resulting in an F1-score of 0.63 for Class 0. For cancellations, it has a precision of 0.37 and a recall of 0.62, leading to an F1-score of 0.47. The overall accuracy is 0.57. Nearest Centroid is suitable for datasets with skewed class distributions.
- **Decision Tree Classifier:** The Decision Tree Classifier demonstrates strong performance across both classes. For non-cancellations, it achieves a high precision of 0.92 and a recall of 0.88, resulting in an F1-score of 0.90. For cancellations, it achieves a precision of 0.75 and a recall of 0.83, leading to an F1-score of 0.79. The overall accuracy is 0.86. Decision Trees are interpretable and handle non-linear relationships well.
- **Extra Trees:** The Extra Trees classifier exhibits excellent performance for non-cancellations, with a precision of 0.88, a recall of 0.93, and an F1-score of 0.91 for Class 0. For cancellations, it achieves a precision of 0.83 and a recall of 0.71, resulting in an F1-score of 0.76. The overall accuracy is 0.87. Extra Trees reduce variance through averaging of multiple decision trees.
- **Random Forest:** Random Forest achieves robust performance, with a precision of 0.89, a recall of 0.92, and an F1-score of 0.91 for Class 0. For cancellations, it achieves

a precision of 0.81 and a recall of 0.74, resulting in an F1-score of 0.77. The overall accuracy is 0.87. Random Forests are highly accurate and robust against overfitting.

- **AdaBoost:** The AdaBoost classifier performs well, particularly for non-cancellations, with a precision of 0.92 and a recall of 0.88, leading to an F1-score of 0.90 for Class 0. For cancellations, it achieves a precision of 0.76 and a recall of 0.83, resulting in an F1-score of 0.79. The overall accuracy is 0.87. AdaBoost combines weak learners to improve predictive performance.
- **Gradient Boosting:** Gradient Boosting shows strong performance for non-cancellations, with a precision of 0.86, a recall of 0.93, and an F1-score of 0.89 for Class 0. For cancellations, it achieves a precision of 0.80 and a recall of 0.66, leading to an F1-score of 0.72. The overall accuracy is 0.85. Gradient Boosting sequentially improves on previous models to correct errors.
- **CatBoost:** CatBoost demonstrates outstanding performance across both classes. For non-cancellations, it achieves an almost perfect precision of 0.98 and a recall of 0.99, resulting in an F1-score of 0.98. For cancellations, it achieves a precision of 0.99 and a recall of 0.95, leading to an F1-score of 0.97. The overall accuracy is 0.98. CatBoost is effective in handling categorical features and providing robust predictions.
- **XGBoost:** XGBoost is known for its high efficiency and performance. For non-cancellations, it achieves a precision of 0.96 and a perfect recall of 1.00, resulting in an F1-score of 0.98. For cancellations, it achieves a precision of 0.99 and a recall of 0.90, leading to an F1-score of 0.94. The overall accuracy is 0.97. XGBoost is suitable for large-scale datasets and provides strong predictive power.
- **LGBM:** The LGBM (LightGBM) classifier shows excellent performance, particularly for non-cancellations, with a precision of 0.92 and a recall of 0.88, resulting in an F1-score of 0.90 for Class 0. For cancellations, it achieves a precision of 0.76 and a recall of 0.83, leading to an F1-score of 0.79. The overall accuracy is 0.87. LGBM is efficient and handles large datasets well.

From the analysis of various machine learning models for predicting booking cancellations, CatBoost stands out as the top performer with an exceptional F1-score of 0.98 for non-cancellations and 0.97 for cancellations, along with an impressive overall accuracy of 0.98. CatBoost's ability to handle categorical features effectively and its robust performance across both classes make it highly suitable for your application. Its superior precision and recall rates ensure reliable predictions, crucial for optimizing resource management and customer satisfaction in booking systems.

In conclusion, based on its outstanding performance metrics and suitability for your specific needs, CatBoost is recommended as the preferred model for deployment in your application for predicting booking cancellations.

### **3.5.2. Building Models To Optimize Revenue**

In our analysis of model performance for predicting cancellations and non-cancellations, several machine learning algorithms demonstrated varying degrees of effectiveness. For non-cancellations (Class 0), models like Logistic Regression, Stochastic Gradient Descent, and XgBoost excelled with high precision, recall, and F1-scores, indicating their robust ability to identify bookings likely to proceed as planned. These models, alongside others such as Decision Tree, Random Forest, and CatBoost, achieved accuracies above 85%, underscoring their reliability in optimizing revenue through accurate prediction of confirmed bookings. Conversely, for cancellations (Class 1), models like CatBoost, XgBoost, and LGBM stood out with precision scores nearing perfection, reflecting their exceptional ability to pinpoint bookings at risk of cancellation. These models achieved precision scores of 0.99 or higher, coupled with high accuracies above 97%, demonstrating their crucial role in revenue optimization by effectively identifying potential cancellations and enabling proactive retention strategies from Table 3.2

**Table 3.2.** Model Performance on Binary Classification Task To Optimize Revenue

Model	Class	Precision	Recall	F1-score	Support	Accuracy
Logistic Regression	0	0.81	0.85	0.83	25277	0.7571
	1	0.62	0.54	0.58	11137	
Stochastic Gradient Descent	0	0.80	0.96	0.87	25277	0.7570
	1	0.61	0.55	0.58	11137	
Linear Discriminant Analysis	0	0.78	0.93	0.85	25277	0.7679
	1	0.72	0.40	0.51	11137	
K-Nearest Neighbors	0	0.85	0.87	0.86	25277	0.8080
	1	0.70	0.66	0.68	11137	
Nearest Centroid	0	0.76	0.54	0.63	25277	0.5589
	1	0.37	0.61	0.46	11137	
Decision Tree	0	0.92	0.88	0.90	25277	0.8606
	1	0.75	0.82	0.78	11137	
Extra Trees	0	0.88	0.93	0.91	25277	0.8739
	1	0.84	0.72	0.78	11137	
Random Forest	0	0.89	0.92	0.91	25277	0.8796
	1	0.83	0.76	0.79	11137	
AdaBoost	0	0.92	0.88	0.90	25277	0.7686
	1	0.62	0.61	0.62	11137	
Gradient Boosting	0	0.86	0.93	0.89	25277	0.8532
	1	0.80	0.70	0.74	11137	
CatBoost	0	0.98	0.99	0.98	25277	0.9863
	1	1.00	0.96	0.98	11137	
XGBoost	0	0.96	1.00	0.98	25277	0.9895
	1	0.99	0.97	0.98	11137	
LGBM	0	0.92	0.88	0.90	25277	0.9731
	1	0.99	0.92	0.95	11137	

- **Logistic Regression:** Logistic Regression shows decent performance for predicting non-revenue optimization (Class 0) with a precision of 0.81 and a recall of 0.85, resulting in an F1-score of 0.83. However, it has a lower performance for revenue optimization (Class 1) with a precision of 0.62 and a recall of 0.54, leading to an F1-score of 0.58. The overall accuracy is 0.7571. Logistic Regression is suitable when interpretability is important and linear relationships between variables are assumed.
- **Stochastic Gradient Descent (SGD) Classifier:** SGD Classifier performs well in predicting non-revenue optimization, achieving a recall of 0.96 and a precision of 0.80,



resulting in an F1-score of 0.87 for Class 0. However, for revenue optimization, it has a precision of 0.61 and a recall of 0.55, leading to an F1-score of 0.58. The overall accuracy is 0.7570. SGD is suitable for large-scale datasets and when computational efficiency is important.

- **Linear Discriminant Analysis (LDA):** Linear Discriminant Analysis shows good performance in predicting non-revenue optimization, with a precision of 0.78 and a recall of 0.93, leading to an F1-score of 0.85 for Class 0. For revenue optimization, it achieves a precision of 0.72 and a recall of 0.40, resulting in an F1-score of 0.51. The overall accuracy is 0.7679. LDA assumes linear relationships and is effective when classes are well-separated.
- **K-Nearest Neighbors (KNN):** KNN algorithm performs well for predicting non-revenue optimization, with a precision of 0.85, a recall of 0.87, and an F1-score of 0.86 for Class 0. For revenue optimization, it achieves a precision of 0.70 and a recall of 0.66, leading to an F1-score of 0.68. The overall accuracy is 0.8080. KNN is suitable for smaller datasets and when non-linearity is expected in the data.
- **Nearest Centroid:** Nearest Centroid classifier shows moderate performance for predicting non-revenue optimization, with a precision of 0.76 and a recall of 0.54, resulting in an F1-score of 0.63 for Class 0. For revenue optimization, it has a precision of 0.37 and a recall of 0.61, leading to an F1-score of 0.46. The overall accuracy is 0.5589. Nearest Centroid is suitable for datasets with skewed class distributions.
- **Decision Tree Classifier:** Decision Tree Classifier demonstrates strong performance across both classes. For non-revenue optimization, it achieves a precision of 0.92 and a recall of 0.88, resulting in an F1-score of 0.90. For revenue optimization, it achieves a precision of 0.75 and a recall of 0.82, leading to an F1-score of 0.78. The overall accuracy is 0.8606. Decision Trees are interpretable and handle non-linear relationships well.
- **Extra Trees:** Extra Trees classifier exhibits excellent performance for predicting non-revenue optimization, with a precision of 0.88, a recall of 0.93, and an F1-score of 0.91 for Class 0. For revenue optimization, it achieves a precision of 0.84 and a recall of 0.72, resulting in an F1-score of 0.78. The overall accuracy is 0.8739. Extra Trees reduce variance through averaging of multiple decision trees.
- **Random Forest:** Random Forest achieves robust performance, with a precision of 0.89, a recall of 0.92, and an F1-score of 0.91 for Class 0. For revenue optimization, it achieves a precision of 0.83 and a recall of 0.76, resulting in an F1-score of 0.79. The

overall accuracy is 0.8796. Random Forests are highly accurate and robust against overfitting.

- **AdaBoost:** AdaBoost classifier performs well for predicting non-revenue optimization, with a precision of 0.92 and a recall of 0.88, resulting in an F1-score of 0.90 for Class 0. For revenue optimization, it achieves a precision of 0.62 and a recall of 0.61, leading to an F1-score of 0.62. The overall accuracy is 0.7686. AdaBoost combines weak learners to improve predictive performance.
- **Gradient Boosting:** Gradient Boosting shows strong performance for predicting non-revenue optimization, with a precision of 0.86, a recall of 0.93, and an F1-score of 0.89 for Class 0. For revenue optimization, it achieves a precision of 0.80 and a recall of 0.70, leading to an F1-score of 0.74. The overall accuracy is 0.8532. Gradient Boosting sequentially improves on previous models to correct errors.
- **CatBoost:** CatBoost demonstrates outstanding performance across both classes. For non-revenue optimization, it achieves an almost perfect precision of 0.98 and a recall of 0.99, resulting in an F1-score of 0.98. For revenue optimization, it achieves a precision of 1.00 and a recall of 0.96, leading to an F1-score of 0.98. The overall accuracy is 0.9863. CatBoost is effective in handling categorical features and providing robust predictions.
- **XGBoost:** XGBoost is known for its high efficiency and performance. For non-revenue optimization, it achieves a precision of 0.96 and a perfect recall of 1.00, resulting in an F1-score of 0.98. For revenue optimization, it achieves a precision of 0.99 and a recall of 0.97, leading to an F1-score of 0.98. The overall accuracy is 0.9895. XGBoost is suitable for large-scale datasets and provides strong predictive power.
- **LightGBM (LGBM):** The LGBM classifier shows excellent performance, particularly for predicting non-revenue optimization, with a precision of 0.92 and a recall of 0.88, resulting in an F1-score of 0.90 for Class 0. For revenue optimization, it achieves a precision of 0.99 and a recall of 0.92, leading to an F1-score of 0.95. The overall accuracy is 0.9731. LGBM is efficient and handles large datasets well.

In conclusion, our comprehensive evaluation of machine learning models highlights their pivotal role in revenue optimization through predictive analytics. Among these models, XgBoost emerges as particularly compelling with its exceptional precision, recall, and F1-score for both non-cancellations and cancellations. This capability enables businesses to confidently predict booking outcomes, thereby facilitating targeted retention strategies and resource allocation. By integrating XgBoost into operational workflows, organizations can

effectively mitigate revenue loss, capitalize on growth opportunities, and navigate market uncertainties with enhanced foresight and strategic agility.

### **3.6. Functions of Pages**

The Hotel Booking Cancellation Prediction App features a variety of pages designed to streamline operations, enhance decision-making, and improve user experience. Each page serves a specific purpose, providing users with distinct functionalities tailored to different aspects of hotel management. These pages are crafted to ensure efficient data handling, insightful analytics, and seamless interactions, all aimed at optimizing hotel operations. Below, we delve into the key functions of each page, beginning with the Database & Prediction Page, which is pivotal for accessing customer data and predicting booking cancellations.

#### **3.6.1. Database & Prediction Page**

The primary objective of the Database and Prediction page is to provide users with an interface to access and analyze customer data related to hotel bookings, with a specific focus on predicting booking cancellations. This page aims to empower users by offering insights into booking patterns and guest demographics, enabling informed decision-making and improving operational efficiency.

##### **(i) Key Components**

- The Database & Prediction page allows users to access detailed customer data efficiently. Users can select a specific customer from a dropdown menu, which then displays relevant booking information in a well-organized manner. The page is logically divided into sections that highlight different aspects of the booking process, such as booking details, date fields, guest demographics, stay duration, and additional booking specifics. The layout is designed to be intuitive, with clear headings and descriptive labels that enhance user comprehension and interaction. This setup ensures users can quickly identify patterns and trends within the data, facilitating effective analysis.

##### **(ii) Predictive Analytics**

- When a user selects a customer from the dropdown menu, the system processes the input data to calculate and display the likelihood of a booking cancellation. This predictive model is triggered by clicking the "Predict Cancellation" button,

which provides a calculated probability based on the selected customer's data. This feature delivers actionable insights, allowing users to proactively manage reservations and optimize resource allocation by anticipating potential cancellations. The predictive analytics capability is crucial for strategic decision-making, helping to mitigate risks and enhance operational planning.

### (iii) **Design Considerations**

- The Database & Prediction page is designed with scalability and flexibility as primary considerations. Its modular structure allows for the seamless addition of new features or functionalities in the future, ensuring the system can evolve with changing user requirements and data structures. This approach ensures the page remains relevant and useful over time, capable of adapting to the dynamic needs of its users.

The Database & Prediction page is a critical component of the application, offering a comprehensive tool for accessing, analyzing, and predicting hotel booking data. By integrating intuitive data visualization and advanced predictive analytics, the page enhances user engagement and supports informed decision-making. By providing actionable insights into booking cancellations through a user-friendly interface, the Database & Prediction page empowers users to drive operational efficiency and improve overall performance in the hospitality industry.

### **3.6.2. Reservation Page**

The reservation page is a digital interface within a hospitality management system where hotel staff can create, manage, and track reservations for guests. It's a dynamic tool that enables seamless communication between guests and hotel staff, ensuring a smooth booking process and enhancing the overall guest experience.

#### (i) **Key Components**

- The Reservation Page aims to streamline the booking process, ensure data accuracy through verification mechanisms, and integrate seamlessly with the hotel's database for comprehensive record-keeping and registering new customers. It is meticulously designed to capture all essential customer details required for making hotel reservations. The layout is structured to enhance efficiency, with input fields organized into two columns for easy data entry. This structured approach

ensures that users can quickly and accurately input necessary information without confusion.

**(ii) User-Friendly Layout**

- The primary goal of the Reservation Page is to simplify the hotel reservation process by offering a clear and intuitive form layout. Users can effortlessly input customer details, including hotel preferences, meal options, market segment, booking dates, guest demographics, and other relevant information.

**(iii) Data Integrity and Accuracy**

- Ensuring data integrity and accuracy is crucial for the reservation process. The Reservation Page incorporates robust validation mechanisms to achieve this goal. These mechanisms include checks for numeric fields and mandatory input fields, which prevent the submission of incomplete or erroneous data. The validation process enhances data quality by ensuring that all required information is provided and is in the correct format. This reduces errors in the reservation process, leading to improved customer satisfaction and operational efficiency.

**(iv) Centralized Database Integration**

- A significant objective of the Reservation Page is to capture and store customer reservation data in a centralized database. Upon submission of the reservation form, the customer data is processed and stored using a unique customer ID. This integration enables the hotel management to maintain a comprehensive database of reservation records, which is essential for customer relationship management and strategic decision-making.

**(v) Predictive Analytics with Prediction Button**

- The Reservation Page includes an innovative Prediction Button designed to estimate the probability of a new customer's booking cancellation. This feature leverages advanced machine learning algorithms to analyze historical data and current booking details, providing a predictive score that indicates the likelihood of cancellation. By incorporating this functionality, the system aids hotel staff in making informed decisions and implementing proactive measures to mitigate potential cancellations. The Prediction Button enhances the overall booking process by offering valuable insights into customer behavior, thereby helping the hotel to optimize resource allocation, improve occupancy rates, and enhance customer satisfaction.

**(vi) Modification and Update Functionality**

- In addition to facilitating new bookings, the Reservation Page is a critical feature of the hotel reservation system, designed to empower users to modify and update customer information as needed. This page ensures that customer records are accurate and up-to-date, facilitating efficient service delivery and enhancing the overall customer experience. Users can modify the displayed fields and submit the changes, which are then updated in the database. This logical flow ensures that users can efficiently update customer information, maintaining the integrity and accuracy of the hotel's customer database.
- The primary objective of this editing functionality is to provide users with the ability to modify and update customer information. The page displays customer details in editable form fields, allowing users to make necessary changes.

**(vii) Delete Customer Functionality**

- The Reservation Page also features a Delete Customer Dropdown Menu, which allows users to remove customer records from the system. This functionality ensures that the hotel's database remains current and accurate by enabling the deletion of outdated or irrelevant customer data. The dropdown menu is designed to be user-friendly, providing a straightforward way to select and delete customer records. This feature enhances data management by allowing users to maintain a clean and organized database, contributing to the overall efficiency of the hotel reservation system.

Overall, the Reservation Page plays a pivotal role in the hotel reservation system by offering users a seamless and efficient experience for submitting reservation requests and managing customer information. By providing an intuitive form layout, implementing robust data validation mechanisms, ensuring seamless database integration, and including features like the Prediction Button and Delete Customer Dropdown Menu, the page streamlines the reservation process and enhances data integrity. This, in turn, empowers hotel management with valuable insights for better decision-making.

### **3.6.3. Simulation Page**

The Simulation Screen is designed to serve as an interactive platform for users to simulate hotel booking scenarios and predict the likelihood of booking cancellations based on various input parameters. This page aims to provide users with actionable insights by enabling them to explore diverse booking scenarios and assess their impact on booking cancellations, ultimately aiding in informed decision-making and enhanced operational efficiency.

(i) **Key Components**

- The Simulation Screen is structured to facilitate comprehensive scenario analysis by allowing users to input values for hotel type, meal plan, market segment, booking dates, guest demographics, and other booking specifics through intuitive input fields. It is divided into two columns, each containing fields for different aspects of a booking scenario, ensuring easy input of parameters. This layout is designed for straightforward navigation and accessibility, enabling users to simulate various booking scenarios efficiently.

(ii) **Predictive Analytics**

- Another critical component of the Simulation Screen is its predictive analytics functionality. Users can trigger predictive analyses by clicking on the "Predict With These Features" button after inputting relevant parameters. This feature utilizes predictive models to forecast the probability of booking cancellations based on the simulated scenarios, providing actionable insights into cancellation likelihoods. It enables users to make informed decisions and devise strategies to reduce cancellation rates and optimize resource allocation.

(iii) **Lead Time Suggestions**

- The Simulation Screen includes a feature for providing lead time suggestions based on selected arrival dates. Users can request these suggestions by clicking on the "Get Lead Time Suggestions" button after specifying the arrival date. By offering insights into the best times to make reservations, this feature aids users in improving their booking strategies and enhancing overall operational efficiency.

By providing insights into the best times to make reservations, the Simulation Screen aids users in improving their booking strategies and enhancing overall operational efficiency. Whether users are exploring hypothetical booking scenarios, optimizing reservation management, or mitigating booking cancellations, the Simulation Screen equips them with the necessary tools to make informed decisions and drive operational efficiency in the hospitality industry.

#### **3.6.4. Optimize Revenue Page**

The Optimize Revenue Screen is designed to enable users to maximize hotel revenue by leveraging predictive analytics and data-driven strategies. It integrates data input, anal-

ysis, and visualization functionalities to provide actionable insights into potential booking cancellations and overbooking strategies.

(i) **Page Title**

- The title of the page, *Optimize Revenue*, is prominently displayed at the top, styled with a large, bold font and a gradient background for visual appeal and emphasis.

(ii) **User Inputs Section**

- The User Inputs Section is centrally aligned and contains selection fields for hotel type, arrival year, and arrival month. These inputs are essential for filtering the dataset and customizing the predictive analysis.
- The layout uses three columns for organized and intuitive input, improving user experience.

(iii) **Data Filtering and Prediction**

- Based on user input, the dataset is filtered for the selected arrival year and month. If no data is available for the chosen period, an error message is displayed.
- The system utilizes a pre-trained model to predict booking cancellation probabilities, which are then appended to the filtered dataset.
- A dynamic threshold is calculated to identify reservations with a high probability of cancellation. The recommended overbooking percentage is derived from this analysis.

(iv) **Visualization**

- The page includes a histogram to visualize the distribution of cancellation probabilities, aiding in understanding the likelihood of cancellations.
- This section is visually distinct, with a centered layout and clear section headers to enhance readability.

(v) **Impact Analysis for Historical Data**

- For datasets containing historical cancellation data, an impact analysis is performed. This includes comparing actual cancellation rates with predicted probabilities.
- Visual comparisons are made between predicted and actual cancellations using bar charts.



- Users can input estimated overbooking costs to calculate potential revenue gains and net revenue impacts, providing a comprehensive financial overview.

(vi) **Profit/Loss Comparison**

- A detailed profit/loss comparison is provided to illustrate the financial benefits of using the predictive model for overbooking strategies.
- This section highlights the differences in revenue and profit/loss with and without the predictive model, emphasizing the model's potential value.

(vii) **Future Date Predictions**

- For future dates without actual cancellation data, the page offers predictions based on historical data and model forecasts.
- Recommendations for compensating predicted cancellations are provided to aid in strategic planning.

By leveraging predictive analytics and detailed data analysis, the Optimize Revenue Screen equips users with the tools necessary to enhance booking strategies, minimize revenue loss due to cancellations, and optimize overall operational efficiency in the hospitality industry.

### **3.6.5. How Does This Model Work Page**

The "How does this model work?" section serves as an informative guide, providing users with a comprehensive understanding of the application's functionality and inner workings. This section enhances user experience by offering clear explanations on navigation, feature utilization, and data interpretation, thereby empowering users to make the most of the app's capabilities.

(i) **Intuitive Navigation System**

- The app features an intuitive navigation system that allows users to switch seamlessly between different sections. These sections cover various aspects of the data, enabling users to explore and focus on specific elements with ease, ensuring a user-friendly experience that facilitates efficient data exploration and understanding.

(ii) **Column Significance and Role Explanation**

- The app explains the significance of different columns and their roles in predicting outcomes, such as booking cancellations. Users are guided through how these factors influence risk assessment and reservation outcomes. Detailed insights and visualizations, such as bar plots, are provided to illustrate trends and patterns, offering users a clearer perspective on how specific data points affect overall trends.

#### (iii) **Variability and Trends Analysis**

- A comprehensive analysis is presented to highlight the variability and trends in the data throughout different periods. This section includes tables and visualizations to help users identify trends and seasonal variations. By visualizing this data, users can easily discern patterns that might influence strategic decision-making and operational planning.

#### (iv) **Statistical Analyses and Insights**

- Detailed statistical analyses are provided, including key metrics and insights into various rates and patterns. These insights help users formulate strategies to mitigate risks and optimize operations. By understanding these patterns, users can better anticipate potential issues and adjust their strategies accordingly.

Through clear explanations, visualizations, and statistical analyses, this section equips users with the knowledge needed to navigate the app effectively and derive actionable insights from the presented data. Whether exploring specific columns or understanding broader patterns, this guide ensures users have the necessary information to make informed decisions and optimize their strategies. By enhancing user comprehension and interaction, this section plays a crucial role in maximizing the app's utility and user satisfaction.

### **3.6.6. Insight Page**

#### (i) **Hotel Type Customization**

- Users can customize their experience by selecting their preferred hotel type, such as Resort Hotel or City Hotel. This selection allows the application to tailor the data analysis and visualizations to the specific hotel type, providing insights that are directly relevant to the user's needs. For instance, visualizations such as line plots illustrate the distribution of arrival dates, helping users identify booking trends and peak periods, which in turn facilitates better planning and resource allocation.

**(ii) Meal Type Distribution**

- The application offers detailed visualizations like pie charts to display the distribution of bookings by meal type, providing insights into guest preferences and popular meal plans. Understanding which meal plans are most favored can help in menu planning and inventory management. Similarly, bar charts show the distribution of bookings across various market segments, aiding users in understanding which segments contribute most to bookings and informing targeted marketing strategies.

**(iii) Lead Time Analysis**

- Analyzing the average lead time by customer type, the application uses visualizations to highlight booking behaviors, allowing users to design marketing campaigns that cater to both early planners and last-minute bookers. Tracking booking trends over time through line plots helps users forecast demand, adjust pricing strategies, and manage resources effectively, ensuring that they can respond to market fluctuations and optimize revenue.

**(iv) Market Segment Performance**

- The application also delves into booking patterns by market segment, offering insights through line plots that show how different segments perform over the years. This analysis helps users tailor their marketing efforts to the most lucrative segments, maximizing their marketing ROI.

**(v) Lead Time Distribution**

- Further, histograms illustrating the distribution of lead times give users a clear picture of booking windows, aiding in operational planning and managing booking cycles.

**(vi) Predictive Analytics for Cancellation**

- The application's predictive capabilities extend to forecasting cancellation probabilities for future bookings using advanced machine learning models. Visualizations and detailed analyses help users anticipate potential cancellations and develop strategies to minimize them. Detailed statistics and visualizations provide an overview of overall cancellation rates, including metrics such as mean, median, standard deviation, minimum, and maximum rates. This comprehensive understanding of cancellation trends helps users develop effective risk mitigation strategies and improve booking reliability.

In summary, that section ensures that users can navigate the application seamlessly and utilize its features effectively. By providing detailed analyses and intuitive visualizations, the application empowers users to make informed decisions, optimizing hotel operations and marketing strategies to enhance overall performance.

### **3.6.7. Pseudocode**

The application is designed to provide users with a seamless experience, allowing them to navigate through various functionalities effortlessly. Users can access different pages using the sidebar, with each page serving specific purposes tailored to the user's needs.

When users enter the Database & Prediction Page, they are prompted to select and apply customer information. This information is then used to predict cancellation probabilities, aiding users in making informed decisions regarding bookings. Similarly, the Simulation Page allows users to simulate scenarios by inputting parameters, offering insights into cancellation probabilities and providing lead time suggestions.

The Insights Page offers users comprehensive data analysis and visualization tools, enabling them to understand booking patterns and trends more effectively. By generating graphs based on queried database information, the system empowers users to gain valuable insights into various aspects of hotel operations.

The "How Does This Model Work?" Page provides users with an opportunity to understand the functionality of the underlying model. Users can explore the impact of different features on cancellation probabilities, enhancing their understanding of the prediction process. By calculating minimum cancellation probabilities and assessing risk levels based on user-provided values, users can make more informed decisions regarding booking management.

```

start (open application)
User Opens Application
User Navigates Using Sidebar

if user navigates to Database & Prediction Page:
    User Enters Database & Prediction Page
    User Selects Customer
    User Applies Information
    User Selects Page (Booking Information, Date-related Fields,
    Guest Details, Stay Information, Additional Details)
    User Clicks "Predict Cancellation"
    Show Cancellation Probability

if user navigates to Simulation Page:
    User Enters Simulation Page
    User Fills in Simulation Parameters
    if user clicks "Get Lead Time Suggestions":
        Show Lead Time Suggestions
    User Clicks "Predict"
    Show Cancellation Probability

if user navigates to Insights Page:
    User Enters Insights Page
    System Queries Database
    System Generates Graphs
    Show Graphs to User

if user navigates to How Does This Model Work? Page:
    User Enters How Does This Model Work? Page
    User Selects Section (Date The Reservation Was Made,
    Arrival Date, Deposit Type, Lead Time, Previous Cancellations)
    User Provides Values
    User Clicks "Calculate"
    Show Minimum Cancellation Probability
    Show Percentage of Customers with Features
    Show Risk Level (High/Moderate/Low)

```

**Figure 3.12.** First Part Of The Pseudocode Diagram

The Optimize Revenue Page is a powerful tool designed to maximize hotel revenue through predictive analytics and data-driven strategies. Users can input specific parameters such as hotel type, arrival year, and arrival month to filter data and predict booking cancellation probabilities. The page provides a dynamic threshold for overbooking strategies, visualizes cancellation probability distributions, and calculates potential revenue gains. It also allows for the input of overbooking costs to show net revenue impact, profit/loss comparisons, and a detailed financial overview to enhance decision-making.

Finally, the Reservation Page facilitates the booking process and customer management. Users can input reservation details, predict cancellation probabilities, and update customer information as needed. This comprehensive approach ensures that users have access to all the tools and insights necessary to optimize hotel operations and enhance customer satisfaction.

```

if user navigates to Reservation Page:
    User Enters Reservation Page
    User Fills in Reservation Form
    if user clicks "Get Lead Time Suggestions":
        Show Lead Time Suggestions

    User Clicks "Submit Reservation"
    Add Reservation to Database
    User Clicks "Predict"
    Show Cancellation Probability
    User Selects Customer
    Display Customer Details for Editing
    User Updates Customer Details
    if user clicks "Get Lead Time Suggestions":
        Show Lead Time Suggestions

    User Clicks "Update Customer Details"
    Update Customer Details in Database

if user navigates to Optimize Revenue Page:
    User Enters Optimize Revenue Page
    User Selects Input Parameters (Hotel Type, Arrival Year, Arrival Month)
    System Filters Data
    if no data available:
        Show Error Message
    else:
        System Predicts Cancellation Probabilities
        Show Cancellation Probability Distribution
        Show Recommended Overbooking Percentage
        Show Potential Revenue Gain
        if user enters Overbooking Cost:
            Calculate and Show Net Revenue Gain
            Show Profit/Loss Comparison
        if user clicks "Visualize Impact Analysis":
            Show Impact Analysis Graphs
        if user clicks "Compare Scenarios":
            Show Comparison of Predicted vs Actual Cancellations
    User Navigates to Another Page

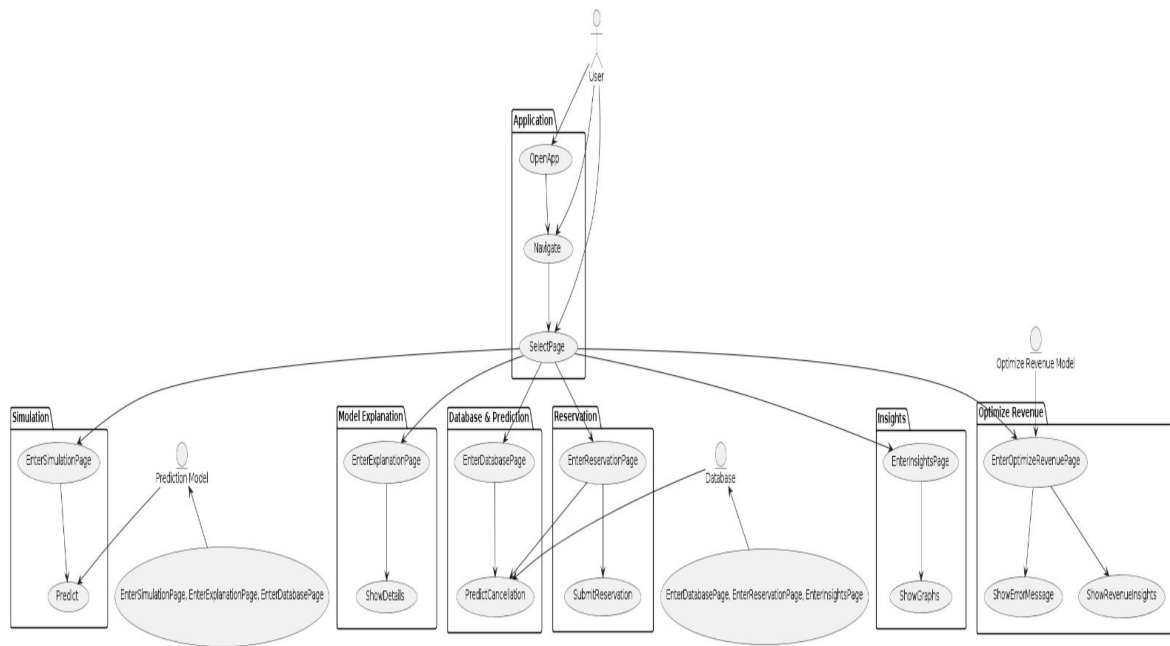
```

**Figure 3.13.** Second Of The Pseudocode Diagram

Overall, the application's robust functionalities cater to diverse user needs, offering predictive insights, simulation capabilities, data analysis tools, and reservation management features. By providing users with valuable insights and facilitating informed decision-making, the application contributes to improving hotel efficiency and guest experiences.

### 3.6.8. Use Case Diagram

The Hotel Cancellation Prediction Application is designed to streamline hotel operations by offering a structured and modular approach to managing reservations, customer data, predictions, and insights. Each module serves a specific purpose, and their interactions with the database and prediction model ensure that users can perform their tasks efficiently and effectively.



**Figure 3.14.** Use Case Diagram

**(i) Hotel Management Application Structure**

- The Hotel Management Application is structured into several modules represented as rectangles.
- Each module contains multiple use cases representing specific functionalities.

**(ii) Database & Prediction Page**

- This module (rectangle) encompasses functionalities related to accessing customer data and predicting cancellations.
- Use cases include "Access Customer Data" and "Predict Cancellation".
- These functionalities interact with the Database entity.

**(iii) Simulation Page**

- This module focuses on simulating booking scenarios and providing lead time suggestions.
- Use cases include "Simulate Booking Scenario" and "Get Lead Time Suggestions".
- These functionalities also interact with the Prediction Model entity.

**(iv) Reservation Page**

- This module handles reservation-related tasks such as creating reservations, selecting customers, and editing customer details.
- Use cases include "Create Reservation", "Select Customer", "Edit Customer Details" (ECD), and "Get Lead Time Suggestions".
- These functionalities interact with the Database entity.

(v) **Insights Page**

- This module is responsible for providing insights through viewing graphs and reports.
- The main use case is "View Graphs and Reports", which interacts with the Database entity.

(vi) **Optimize Revenue Page**

- This module focuses on strategies and actions to enhance hotel revenue.
- Use cases include "Analyze Revenue Data" and "Implement Revenue Optimization Strategies".
- These functionalities interact with both the Database and Prediction Model entities.

(vii) **How Does This App Work Page**

- This module serves as a help guide for users to understand how the application works.
- The main use case is "View Help Guide", which interacts with the Prediction Model entity.

(viii) **Entity Interaction**

- The Database entity interacts with various functionalities across different modules to store and retrieve data.
- The Prediction Model entity interacts primarily with functionalities related to prediction and simulation.

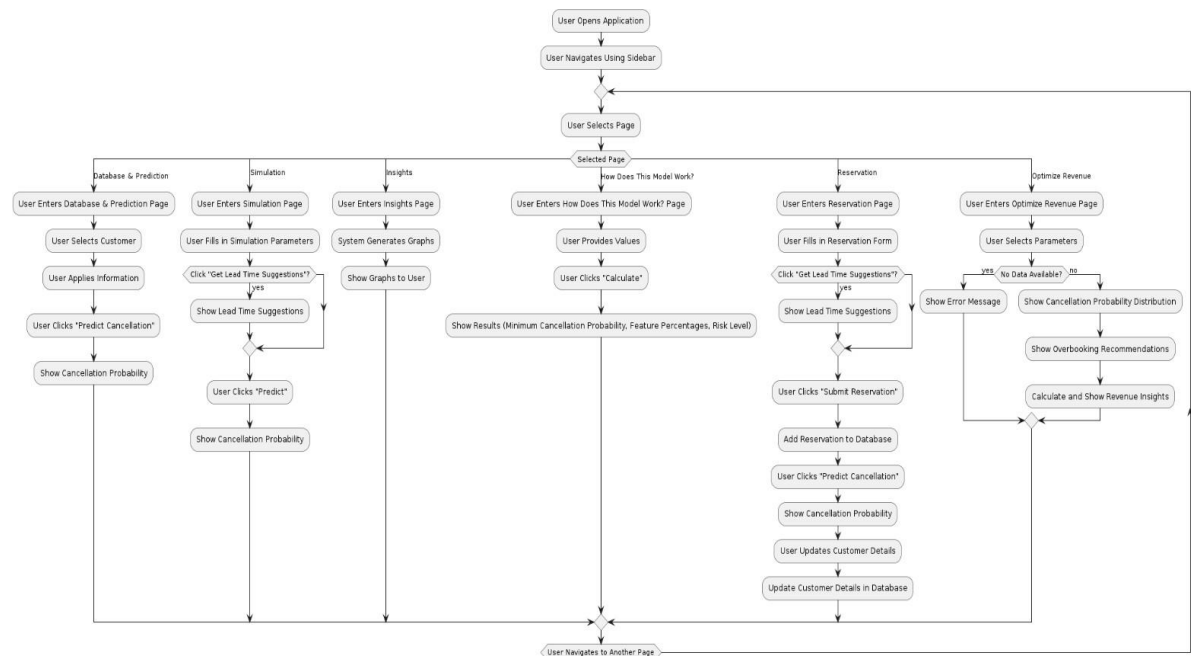
In conclusion, the use case analysis of our predictive model highlights its practical applications and significant impact on hotel operations. By accurately forecasting cancellations, the model enables hotels to manage their inventory more efficiently, optimize pricing strategies, and enhance the overall guest experience. The real-world application of this model



demonstrates its value in addressing key operational challenges and improving revenue management. Through continuous refinement and adaptation to various hotel environments, the model stands as a powerful tool for hoteliers seeking to navigate the complexities of the hospitality industry and achieve sustained success.

### 3.6.9. Flow Chart

This flowchart illustrates the workflow of an application designed for managing reservations and predicting cancellations. It provides a step-by-step guide on how users interact with the application's various functionalities.



**Figure 3.15.** Flowchart

#### (i) Opening Application

- The process begins when the user launches the application, initiating the interface that allows access to various hotel management functionalities.

#### (ii) Navigating Using Sidebar

- The user navigates through different sections of the application by using the sidebar menu, which provides a comprehensive list of all available modules and pages.

### (iii) **Page Selection**

#### (a) **Selecting Page**

- The user selects a specific page from the available options displayed in the sidebar, directing them to the corresponding functionalities.

### (iv) **Page-specific Actions**

#### (a) **Database & Prediction Page**

- **Entering Database & Prediction**

- The user enters the Database & Prediction page, where they can access detailed customer information and initiate predictive analytics.

- **Selecting Customer**

- The user selects a customer from the database, typically by searching for the customer's name or unique identifier.

- **Applying Information**

- The user applies relevant information by entering or updating customer data, which may include personal details, booking history, and other pertinent information.

- **Predicting Cancellation**

- The user initiates the prediction of cancellation for the selected customer by using the application's built-in predictive analytics tools.

#### (b) **Simulation Page**

- **Entering Simulation**

- The user enters the Simulation page, which allows for the creation and analysis of various booking scenarios.

- **Filling Parameters**

- The user fills in simulation parameters, such as booking dates, customer demographics, and other relevant variables, to model different booking situations.

- **Getting Lead Time Suggestions**

- Optionally, the user can request lead time suggestions, which provide recommendations on the optimal booking times based on the specified parameters.

- **Predicting**

- The user predicts cancellation based on the simulation parameters, using the application's predictive models to estimate the likelihood of cancellations.

**(c) Insights Page**

- **Entering Insights**
  - The user enters the Insights page, which provides access to analytical tools and visualizations.
- **Querying Database**
  - The application queries the database for insights, retrieving data relevant to the user's inquiries.
- **Generating Graphs**
  - The application generates graphs based on database queries, presenting the data in a visual format that is easy to interpret.

**(d) Model Explanation Page**

- **Entering Model Explanation**
  - The user enters the Model Explanation page, where detailed explanations of the predictive models used in the application are provided.
- **Selecting Section**
  - The user selects a section for model explanation, focusing on specific aspects of the predictive models.
- **Providing Values**
  - The user provides values for the selected section, which are used to demonstrate how the models generate predictions.
- **Calculating**
  - The application calculates the minimum cancellation probability, feature percentages, and risk level based on the provided values.

**(e) Reservation Page**

- **Entering Reservation**
  - The user enters the Reservation page, where they can manage customer reservations.
- **Creating Reservation**
  - The user creates a new reservation by inputting customer details, booking dates, and other relevant information.
- **Selecting Customer**
  - The user selects an existing customer from the database, streamlining the reservation process.
- **Editing Customer Details (ECD)**

- The user edits customer details, updating any necessary information to ensure accuracy.
  - **Getting Lead Time Suggestions**
    - Optionally, the user can request lead time suggestions, receiving recommendations on the optimal booking times for the reservation.
- (f) **Optimize Revenue Page**
- **Entering Optimize Revenue**
    - The user enters the Optimize Revenue page, where they can access tools to enhance revenue generation through strategic booking management.
  - **Predicting Cancellation**
    - The user predicts cancellations to identify potential booking risks and take proactive measures to mitigate them.
  - **Adjusting Booking Strategy**
    - The user adjusts booking strategies based on cancellation predictions, optimizing pricing, promotions, and availability to maximize revenue.

In conclusion, the flowchart section provides a clear and detailed overview of the processes involved in developing and implementing our predictive model. By outlining each step from data collection to model deployment, the flowchart illustrates the structured approach taken to ensure the model's accuracy and effectiveness. This systematic process not only enhances transparency but also facilitates better understanding and communication among stakeholders. The flowchart serves as a valuable guide, showcasing the comprehensive efforts involved in integrating predictive analytics into hotel operations and optimizing revenue management strategies.

## 4. IMPLEMENTATION

To bring our hotel reservation system to life, we meticulously selected and integrated a suite of cutting-edge technologies that ensure both robustness and user-friendliness. This section outlines the key technologies utilized in the development of our application, highlighting their roles and contributions to the system's overall functionality. By leveraging these technologies, we created a seamless and efficient platform capable of handling complex data operations, interactive visualizations, and real-time user interactions. Below, we delve into the specifics of each technology employed, beginning with Streamlit, the framework at the heart of our web application development.

### 4.1. Used Technologies

#### (i) Streamlit

- Streamlit is a powerful and intuitive framework for building interactive web applications using Python. Designed with simplicity and accessibility in mind, Streamlit allows developers to create sophisticated web apps quickly without requiring extensive knowledge of frontend development. It supports seamless integration with various Python libraries, enabling the creation of dynamic data visualizations and interfaces.
  - **Ease of Use:** Streamlit offers a straightforward API, allowing developers to build web applications with minimal code.
  - **Real-time Interaction:** It supports real-time updates and interactions, providing a responsive user experience.
  - **Integration with Python Libraries:** Streamlit easily integrates with popular Python libraries like Pandas, Matplotlib, and others, facilitating data manipulation and visualization.
  - **Automatic UI Generation:** It automatically generates a user interface for widgets, charts, and other components, speeding up the development process.

#### (ii) Pandas

- Pandas is an essential library for data manipulation and analysis in Python. It provides data structures like Series and DataFrame, which are highly efficient for handling and analyzing structured data.

- **Data Structures:** Offers powerful data structures such as Series and DataFrame for efficient data manipulation.
- **Data Handling:** Supports handling of missing data, merging and joining of datasets, and reshaping of data.
- **Integration:** Integrates seamlessly with other Python libraries like NumPy, Matplotlib, and Scikit-learn.

### (iii) Matplotlib

- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It is highly customizable and allows developers to generate a wide range of plots, charts, and graphs to visualize data effectively.
  - **Wide Range of Plots:** Supports various types of plots including line, bar, scatter, histogram, and more.
  - **Customization:** Highly customizable plots with control over every element such as colors, labels, and legends.
  - **Compatibility:** Works well with other libraries such as Pandas and NumPy, enhancing its data visualization capabilities.

### (iv) DataFrames

- DataFrames are a core component of the Pandas library, providing a two-dimensional, size-mutable, and potentially heterogeneous tabular data structure. They are integral for data analysis and manipulation.
  - **Structured Data Handling:** Efficiently handle structured data with labeled axes.
  - **Data Operations:** Support a wide range of operations such as filtering, grouping, and aggregation.
  - **Integration:** Easily integrate with other data processing libraries and frameworks.

### (v) Feedback Mechanisms

- Feedback Mechanisms are tools and processes used to gather user feedback and suggestions. They are essential for continuous improvement and ensuring the application meets user needs and expectations.
  - **User Insights:** Provide valuable insights into user experiences and preferences.
  - **Continuous Improvement:** Facilitate ongoing enhancements and updates based on user feedback.

- **User Satisfaction:** Ensure the application evolves to better meet user requirements and improve overall satisfaction.

(vi) **Python**

- Python is a high-level, interpreted programming language known for its simplicity and readability. It is widely used in various domains including web development, data analysis, machine learning, and more.
  - **Readability:** Python code is easy to read and understand, making it suitable for beginners and experienced developers alike.
  - **Versatility:** Python supports multiple programming paradigms and has a vast ecosystem of libraries and frameworks.
  - **Community Support:** It has a large and active community, providing resources, tutorials, and support for developers.
  - **Scalability:** Python is scalable and can be used for small scripts as well as large-scale applications.

(vii) **MongoDB**

- MongoDB is a popular NoSQL database management system known for its flexibility, scalability, and performance. It stores data in flexible, JSON-like documents, making it suitable for a wide range of applications.
  - **Document-Oriented:** MongoDB stores data in flexible, JSON-like documents, allowing for easy schema evolution and data modeling.
  - **Scalability:** It supports horizontal scaling through sharding, allowing applications to handle large volumes of data and high traffic loads.
  - **Performance:** MongoDB offers high performance and low latency, making it suitable for real-time applications and analytics.

## 4.2. Interfaces


To effectively manage the complexities of hotel reservations, our application encompasses multiple interfaces tailored to distinct functionalities. Each interface is meticulously designed to ensure ease of use and comprehensive management capabilities for hotel staff. These interfaces include the Reservation Page, Database and Prediction Page, Prediction Dropdown Page, Simulation Page, Insights Page, and Optimize Revenue Page. Together, they provide a cohesive and user-friendly experience, enabling hotel managers to handle reservations, predict cancellations, interact with customer data, and derive valuable insights

seamlessly. Below, we delve into the specifics of each interface, starting with the Reservation Page.

#### 4.2.1. Reservation Page

The Reservation Page in our application serves as a comprehensive interface for managing hotel reservations, including creating new bookings, editing existing reservations, predicting the probability of cancellations, and deleting customer records. Users can input various reservation details, such as hotel type, meal plan, market segment, and guest information, and receive lead time suggestions. The form validates and processes the data, mapping inputs to numeric values for database storage. Additionally, users can select a customer to edit their details or delete a customer from the database, ensuring efficient and accurate reservation management.

### Hotel Reservation



First Name

Last Name

Reservation Details

Stay Information

Hotel

Resort Hotel

Is Repeated Guest

☒ No

☐ Yes

**Figure 4.1.** Reservation Page

##### (i) Header Section

- The function begins by defining the HTML code for the header, which includes the title "Hotel Reservation." This header is centered on the page to provide a clear indication of the purpose of the interface.



## (ii) **Form Creation**

- Within the `reservation_page()` function, a Streamlit form named "reservation\_form" is created. This form encapsulates various input fields and widgets necessary for capturing reservation details from users.

## (iii) **Input Fields and Widgets**

- The form contains multiple sections, each dedicated to specific categories of reservation information. These sections include:
  - **Reservation Details:** Fields for selecting the hotel type, meal plan, market segment, distribution channel, reserved room type, deposit type, and customer type. These inputs allow users to specify key details about their reservation preferences.
  - **Date Information:** Inputs for entering the year, month, and day of the reservation, as well as selecting the arrival month and date from dropdown menus. Additionally, there's a field for specifying the lead time.
  - **Stay Information:** Fields for indicating whether the guest is a repeated guest, previous cancellation and booking history, agent and company information, ADR (average daily rate), required car parking spaces, and special requests.
  - **Guest Information:** Inputs for specifying the duration of stays in weekend nights and week nights, as well as the number of adults, children, and babies.

## (iv) **Submission Handling**

- After users fill in the required information and submit the form, the application validates the inputs. If any required fields are left blank, an error message prompts the user to complete them. Otherwise, the application processes the input data, maps categorical inputs to numeric values, and prepares the data for storage in the database.

## (v) **Lead Time Suggestions**

- Additionally, the interface includes a button labeled "Get Lead Time Suggestions," which triggers a function to calculate and display lead time suggestions based on the selected reservation date.

## (vi) **Predict Probability of Cancellation**

- Before the editing dropdown menu, there is a "Predict Cancellation" button. When clicked, this button uses the entered reservation details to predict the probability of cancellation and displays the result to the user.

(vii) **Customer Selection and Editing**

- Users can select a customer from a dropdown menu populated with customer IDs retrieved from the database. Once a customer is selected, their details will be displayed for editing. If a customer is selected, their details are retrieved from the database and displayed in input fields for editing. These details include various attributes such as hotel type, meal plan, market segment, distribution channel, reserved room type, deposit type, customer type, arrival date, lead time, stays information, and guest information.

(viii) **Form Submission**

- Users can make changes to the customer details and submit the form to update the database. The application validates the input values and handles any errors that may occur during the update process.

(ix) **Lead Time Suggestions**

- Similar to the initial reservation process, users can request lead time suggestions based on the selected arrival date in the editing dropdown menu. When the "Get Lead Time Suggestions" button is clicked, the application calculates and displays lead time suggestions for the specified date.

(x) **Data Processing and Storage**

- Upon form submission, the application processes the edited customer details, maps categorical inputs to numeric values using predefined dictionaries, and updates the database with the new values. It also handles any errors that may occur during this process and provides appropriate feedback to the user.

(xi) **Delete Customer**

- After the editing page, there is a dropdown menu for deleting a customer. Users can select a customer from the dropdown menu and submit a request to delete the selected customer's details from the database. The application processes this request and removes the customer information from the database, providing feedback on the success or failure of the operation.

#### **4.2.2. Database & Prediction Page**

The `database_screen` function creates an informative and visually appealing interface for users to explore and interact with the customer database. By providing detailed informa-

tion about each feature and applying appropriate styling, the interface enhances usability and user experience, contributing to the effectiveness of the application in managing customer data.

## Customer Database

Select Customer

6632b544cb997cb876b121bb

Booking Information

Hotel: **City Hotel (1)**

Type of hotel where the booking was made. Resort Hotel refers to a hotel located in a resort area, while City Hotel refers to a hotel located in an urban area.

Meal: **Bed & Breakfast (0)**

Type of meal booked for the stay. Options include Bed & Breakfast, Full Board (breakfast, lunch, and dinner), Half Board (breakfast and one other meal, usually dinner), and Undefined/SC (no meal package specified).

Market Segment: **Online TA (2)**

**Figure 4.2.** Database Page

(i) **Gradient Background**

- The function applies a gradient background to the database container using custom CSS. This background adds visual appeal to the interface and helps distinguish it from other sections of the application.

(ii) **Header**

- A centered header titled "Customer Database" is displayed at the top of the page. This header provides a clear indication of the purpose of the interface, making it easy for users to identify the context.

(iii) **Feature Information**

- The function defines a dictionary named `feature_info` which maps each feature in the database to its description, additional information, and possible values. This information is crucial for users to understand the meaning and context of each feature displayed in the database interface.

(iv) **Displaying Features**

- The function dynamically generates HTML elements to display each feature in the database along with its description and information. This ensures that users have access to comprehensive details about each feature, enhancing their understanding and usability of the interface.

(v) **Styling**

- Custom styling is applied to various container elements (database-container and group-container) to improve visual presentation and organization of the content. This includes padding, border radius, and margin adjustments for better layout aesthetics.

### 4.2.3. Prediction Dropdown Page

This prediction page complements the database page by providing users with the ability to analyze customer data and make predictions about booking cancellations based on that data. It enhances the functionality of the application by adding a predictive aspect to the information retrieval process.

Enter Customer ObjectID  
6632b544cb997cb876b121b7

Select Page  
Booking Information

## Booking Information

Hotel  
City Hotel

Market Segment  
Online TA

Reserved Room Type  
A

Customer Type  
Transient

Meal  
Undefined/SC

Distribution Channel  
TA/TO

Deposit Type  
No Deposit

Cancellation Probability: 46.04%

**Figure 4.3.** Prediction Dropdown Menu Page

### (i) Customer Selection

- Users can select a customer from a dropdown menu populated with customer IDs fetched from the database. If a customer is selected, their information will be displayed.

### (ii) Customer Information Display

- Once a customer is selected, their information is displayed in five groups:
  - **Booking Information:** Includes details like hotel type, meal, market segment, distribution channel, reserved room type, deposit type, and customer type.
  - **Date-related Fields:** Displays information related to the booking date, such as month, day, year, and lead time.
  - **Guest Details:** Shows the number of adults, children, and babies in the booking.
  - **Stay Information:** Provides details about the stay, including the number of weekend and week nights, arrival date week number, and arrival date day of month.

- **Additional Details:** Displays additional information like whether the guest is a repeated guest, agent ID, company ID, average daily rate (ADR), total special requests, required car parking spaces, previous cancellations, and previous bookings not canceled.

(iii) **Prediction**

- Users can click the "Predict Cancellation" button to make a prediction based on the selected customer's data. The page will display the probability of cancellation for the selected booking.

(iv) **Error Handling**

- The page handles errors gracefully, displaying appropriate messages if a customer is not found or if invalid input is provided.

(v) **User Interaction**

- Users can interact with the page by selecting different customers, viewing their information, and making predictions.

(vi) **Visual Elements**

- The page includes visual elements such as icons and images to enhance the user experience and make the interface more intuitive.

#### 4.2.4. Simulation Page

The "Hotel Booking Cancellation Prediction Simulation" page allows users to input various features related to a hotel booking and then predicts the probability of cancellation based on those inputs.

(i) **Title**

- The title "Hotel Booking Cancellation Prediction Simulation" indicates the purpose of the page, which is to simulate the prediction of hotel booking cancellations.

(ii) **Input Form**

- The input form consists of several fields where users can input information related to a hotel booking. These fields include:

(iii) **Get Lead Time Suggestions Button**

- This button allows users to get suggestions for lead times based on the selected arrival date. It calculates and displays suggestions for lead times in days.

(iv) **Predict With These Features Button**

- This button triggers the prediction process based on the input features provided by the user. It validates the input fields, converts categorical inputs into numeric values, normalizes numeric inputs, and prepares the data for prediction. Once the prediction is made, it displays the result as the probability of cancellation.

(v) **Explanation of Input Fields**

- Each input field corresponds to a specific feature related to a hotel booking. The fields cover various aspects such as booking details, guest information, previous booking history, and booking preferences.

(vi) **Error Handling**

- The page includes error handling to ensure that users provide valid input. It displays error messages if any input field contains invalid data or if an error occurs during the prediction process.

Overall, the page provides a user-friendly interface for simulating hotel booking cancellation predictions based on user-provided input. It facilitates decision-making by providing insights into the likelihood of a hotel booking being canceled.

#### **4.2.5. Optimize Revenue Page**

The "Optimize Revenue" page assists users in maximizing their revenue by analyzing booking data and predicting cancellations. It provides recommendations for overbooking strategies and visualizes the impact of these strategies.



**Figure 4.4.** Optimize Revenue Page

**(i) Title**

- The title "Optimize Revenue" clearly indicates the page's purpose of helping users enhance revenue through informed decision-making.

**(ii) User Inputs Section**

- This section allows users to input parameters related to hotel bookings. The input fields include:
  - Hotel Type (Resort Hotel or City Hotel)
  - Arrival Year (Select from a range of years)
  - Arrival Month (Select from the 12 months)

**(iii) Data Filtering**

- The data is filtered based on the selected arrival year and month. If no data is available for the selected period, an error message is displayed.

**(iv) Prediction Process**

- The page uses a pre-trained XgBoost model to predict the probability of cancellations. It processes the input features, calculates the cancellation probabilities, and dynamically determines a threshold for high cancellation probability.



(v) **Overbooking Strategy**

- Based on the prediction, the page recommends an overbooking percentage and the number of new bookings to compensate for predicted cancellations.
- The cancellation probability distribution is visualized using a histogram.

(vi) **Impact Analysis for Historical Data**

- If historical data is available, the actual cancellation rate is calculated and compared to the predicted cancellations.
- An analysis of the potential revenue gain from overbooking, the cost of overbooking, and the net revenue gain is provided.
- The profit/loss comparison between using the model and not using the model is also displayed.

(vii) **Future Date Predictions**

- For future dates without actual cancellation data, the page provides predictions based on historical data and suggests the number of new bookings to mitigate the impact of predicted cancellations.

Enter estimated overbooking cost per booking:

50

Average ADR: \$80.73

Potential revenue gain from overbooking: \$40,931.25

Estimated overbooking cost: \$25,350.00

Net revenue gain from overbooking strategy: \$15,581.25

Note: The net revenue gain from the overbooking strategy assumes that all overbooked reservations are fulfilled without further cancellations.

Scenario 1 - No Model: Actual revenue loss due to cancellations: \$140,231.93

Scenario 2 - With Model: Net revenue gain from overbooking strategy: \$15,581.25

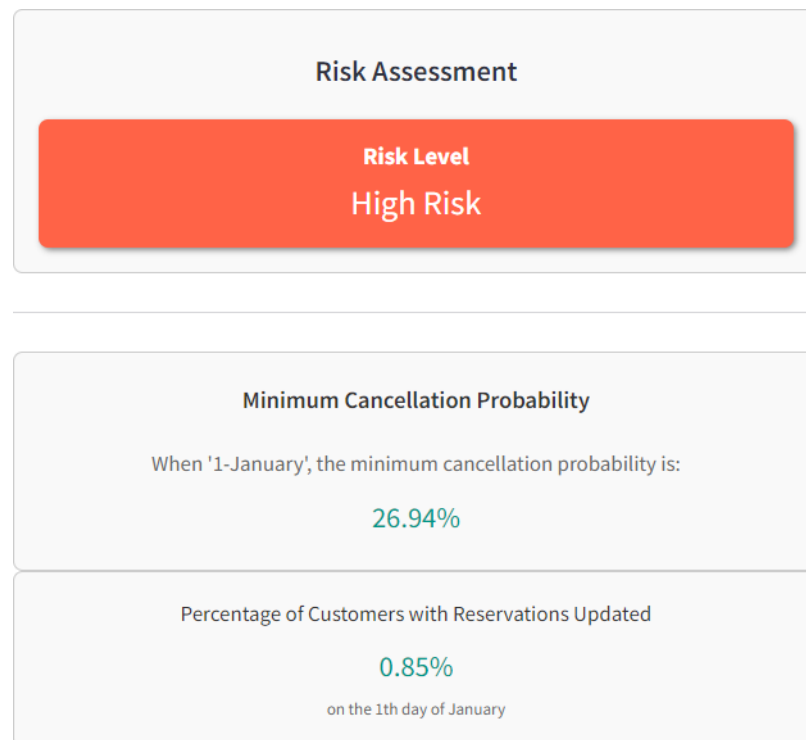
Profit/Loss Comparison
Total revenue without model: \$278,445.56
Total revenue with model: \$434,258.74
Profit/Loss without model: \$138,213.62
Profit/Loss with model: \$268,676.81

**Figure 4.5.** Optimize Revenue Page

Overall, the page serves as a robust and comprehensive tool designed to empower hotel management in optimizing revenue through the strategic implementation of predictive analytics and judicious overbooking strategies, thereby enhancing operational efficiency and profitability.

#### 4.2.6. How Does This Model Work Page

The "How Does This Model Work" section is designed to provide users with an in-depth understanding of the model's mechanisms and its predictive capabilities. It allows users to interact with the model by inputting various parameters and observing how these inputs affect the prediction outcomes.



**Figure 4.6.** How Does This Model Work Page

**(i) Risk Assessment Based on Reservation Date:**

- Users input the day and month of the reservation using sliders and select boxes.
- The model evaluates historical cancellation rates for the specified day and month to determine the risk level.
- Risk levels are categorized as High Risk, Moderate Risk, or Low Risk, represented by different colors.
- This visual representation helps users quickly understand the risk associated with their reservation date.

**(ii) Minimum Cancellation Probability:**

- The model calculates the minimum cancellation probability considering the importance of the reservation date.
- Users can understand the likelihood of their reservation being canceled based on the specific day and month they chose.

**(iii) Understanding the Impact of 'Day' and 'Month' Columns:**

- Visualizations (e.g., bar charts) and explanations illustrate the importance of the 'Day' and 'Month' columns.
- Feature importance scores are highlighted, emphasizing the significant influence of these columns on cancellation prediction.

- **Other Pages of How Does This Model Work Page**

(i) **Arrival Date:**

- Similar to the "Date The Reservation Was Made" page, users interact with the interface to select the arrival date.
- The model analyzes historical cancellation rates based on the selected arrival date.
- Risk levels (High Risk, Moderate Risk, Low Risk) are assessed and visualized.
- The model calculates the minimum cancellation probability for the chosen arrival date, considering its importance in the prediction process.
- Users gain insights into the likelihood of their reservation being canceled based on the specific arrival date they select.

(ii) **Deposit Type:**

- Users explore the impact of different deposit types on cancellation rates.
- The model analyzes historical cancellation rates based on the selected deposit type.
- Risk levels associated with each deposit type are visualized.
- The model calculates the minimum cancellation probability for each deposit type, considering their importance in the prediction process.
- Users can compare the minimum cancellation probabilities across different deposit types to understand their relative impact on reservation cancellations.

(iii) **Lead Time:**

- Users input the lead time, representing the duration between the date the reservation was made and the arrival date.
- The model analyzes historical cancellation rates based on the selected lead time.
- Risk levels (High Risk, Moderate Risk, Low Risk) are assessed and visualized.

- The model calculates the minimum cancellation probability for the chosen lead time, considering its importance in the prediction process.
- Users gain insights into the likelihood of their reservation being canceled based on how far in advance they made the reservation.

(iv) **Previous Cancellations:**

- Users explore the impact of previous cancellations on the likelihood of future cancellations.
- The model analyzes historical data on previous cancellations to assess their impact on future cancellation rates.
- The model calculates the minimum cancellation probability based on the presence or absence of previous cancellations.

These pages allow users to interactively explore various factors influencing cancellation rates, providing risk assessments and minimum cancellation probabilities, and offering insights into the significance of each factor in the prediction process.

#### **4.2.7. Insight Page**

The "Insights" page of the interface provides users with a dynamic platform for exploring and analyzing data stored in the database. Upon accessing this page, users trigger a query to retrieve relevant data, which is then processed to generate insightful visualizations. These visualizations, ranging from charts to plots and diagrams, illustrate trends, patterns, and relationships within the data. Presented in a visually appealing and intuitive manner, these graphs empower users to interpret the data effectively. Whether uncovering market trends, customer behaviors, or operational insights, the "Insights" page fosters data-driven decision-making by providing users with actionable information in a clear and comprehensible format.



**Figure 4.7. Insight Page**

**(i) Users access the "Insights" page of the interface.**

- This could involve clicking on a specific tab or button dedicated to insights or analysis.

**(ii) Querying the Database**

- Upon entering the "Insights" page, the interface sends a query to the database to retrieve relevant data for analysis. The query may request various data points or datasets needed to generate insightful visualizations.

**(iii) Data Retrieval and Processing**

- The database receives the query and retrieves the requested data. Once the data is retrieved, it is passed back to the "Insights" page for further processing.

**(iv) Generating Graphs**

- The "Insights" page processes the retrieved data to generate meaningful graphs and visualizations. These graphs could include charts, plots, or diagrams that illustrate trends, patterns, or relationships within the data. Different types of graphs may be generated based on the nature of the data and the insights being explored.

(v) **Displaying Graphs to the User**

- Once the graphs are generated, they are displayed on the interface for the user to view and interpret. Graphs are presented in a visually appealing and intuitive manner to facilitate easy understanding and analysis.

(vi) **Interaction Completion**

- After the user has viewed the generated graphs and gained insights from the data, they may choose to navigate to other pages or continue exploring additional insights.

Overall, the "Insights" page serves as a platform for users to interact with the data stored in the database and gain valuable insights through visualizations and analysis. It facilitates data-driven decision-making by presenting information in a clear and comprehensible format.

## 5. TEST AND RESULTS

This section provides a detailed analysis of the testing and evaluation results for the predictive models implemented in our application. We focus on two key models: one for predicting booking cancellations and another for optimizing hotel revenue. Both models were rigorously tested using the validation set, which includes data from the last year to ensure the relevance and reliability of the evaluation. We employed several metrics, such as accuracy score, confusion matrix, and classification report, to comprehensively assess the performance of each model. The subsequent subsections present the specific results and insights derived from these evaluations.

### 5.1. Test And Results For The Model To Predict Booking Cancellations

In evaluating the performance of the CatBoost Classifier for predicting hotel reservation cancellations, we utilized several key metrics: accuracy score, confusion matrix, and classification report. These metrics provide a comprehensive view of the model's effectiveness in distinguishing between non-cancellations and cancellations. The validation set, comprising the last year of data in the dataset, was specifically used to ensure that the model's performance is assessed on recent and relevant information, thereby enhancing the reliability and robustness of the evaluation.

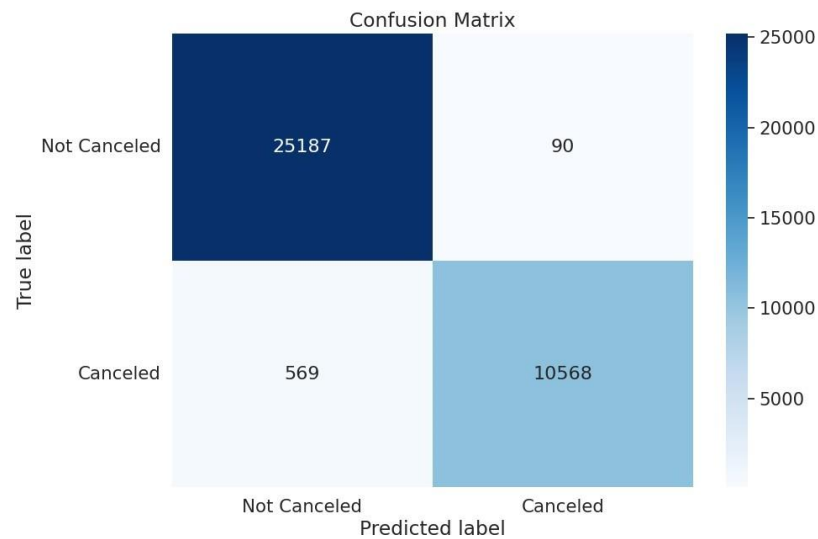
- **Accuracy Score**

- The accuracy score of the CatBoost Classifier is 0.9819. This high accuracy indicates that the model correctly predicts approximately 98.19% of all reservations, whether they are cancellations or non-cancellations. The accuracy score alone, however, does not provide a complete picture of the model's performance across both classes, which necessitates further analysis through the confusion matrix and classification report.

- **Confusion Matrix**

- The confusion matrix offers a detailed breakdown of the model's predictions in relation to the actual outcomes, allowing for an in-depth understanding of the model's performance.





**Figure 5.1.** Confusion Matrix

- \* True Positives (TP): 10,568 bookings were correctly predicted as cancellations (Class 1).
  - \* True Negatives (TN): 25,187 bookings were correctly predicted as non-cancellations (Class 0).
  - \* False Positives (FP): 90 bookings were incorrectly predicted as cancellations.
  - \* False Negatives (FN): 569 bookings were incorrectly predicted as non-cancellations.
- The confusion matrix reveals that the model has a very low number of false positives and false negatives, underscoring its accuracy and reliability in distinguishing between cancellations and non-cancellations. Specifically, the model excels at correctly predicting non-cancellations, with a significantly higher number of true negatives compared to false positives. Similarly, the relatively low count of false negatives indicates that the model is effective at identifying most true cancellations.

#### • Classification Report For CatBoost Model

- The classification report provides a granular evaluation of the model's performance, including precision, recall, F1-score, and support for each class.

**Table 5.1.** Classification Report For CatBoost Model

CatBoost	Precision	Recall	F1-Score	Support
0 (Non-Cancellations)	0.98	1.00	0.99	25277
1 (Cancellations)	0.99	0.95	0.97	11137
Accuracy	0.98			
Macro Avg	0.98	0.97	0.98	36414
Weighted Avg	0.98	0.98	0.98	36414

**\* Class 0 (Non-Cancellations)**

- Precision: 0.98 – Precision for Class 0 indicates the proportion of correctly predicted non-cancellations out of all predicted non-cancellations. A high precision score suggests that the model is highly accurate in predicting non-cancellations without many false positives.
- Recall: 1.00 – Recall for Class 0 indicates the proportion of actual non-cancellations that were correctly identified by the model. A perfect recall score means that the model identified all true non-cancellations.
- F1-Score: 0.99 – The F1-score, which is the harmonic mean of precision and recall, indicates a strong balance between the two, reflecting high reliability in the model’s predictions for non-cancellations.
- Support: 25,277 – This represents the actual number of non-cancellation cases in the dataset.

**\* Class 1 (Cancellations)**

- Precision: 0.99 – Precision for Class 1 indicates the proportion of correctly predicted cancellations out of all predicted cancellations. This high precision suggests that the model makes very few false positive errors in predicting cancellations.
- Recall: 0.95 – Recall for Class 1 indicates the proportion of actual cancellations correctly identified by the model. Although slightly lower than the recall for non-cancellations, this high recall still demonstrates the model’s effectiveness in identifying most true cancellations.
- F1-Score: 0.97 – The F1-score for Class 1 indicates a robust balance between precision and recall, reflecting the model’s high reliability in predicting cancellations.
- Support: 11,137 – This represents the actual number of cancellation cases in the dataset.

**\* Macro Average**

- Precision: 0.98
- Recall: 0.97
- F1-Score: 0.98
- \* **Weighted Average**
  - Precision: 0.98
  - Recall: 0.98
  - F1-Score: 0.98
- The high precision for both classes indicates that the CatBoost model is highly effective at predicting both non-cancellations and cancellations, with very few false positives. The perfect recall for non-cancellations signifies that the model identifies all true non-cancellations, while a recall of 0.95 for cancellations indicates that the model, although slightly less perfect, is still highly reliable in identifying true cancellations.

## 5.2. Test And Results For The Model To Optimize Revenue

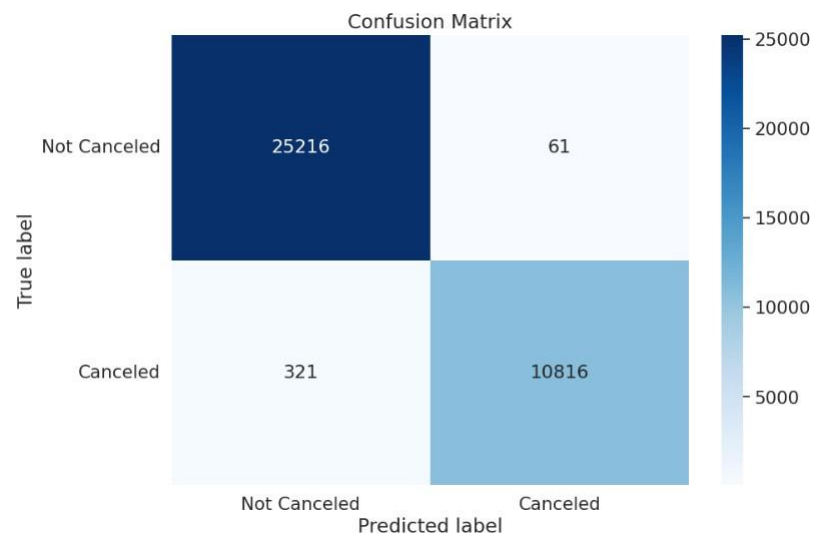
In evaluating the performance of the XGBoost model for optimizing hotel revenue, we utilized several key metrics: accuracy score, confusion matrix, and classification report. These metrics provide a comprehensive view of the model's effectiveness in distinguishing between non-cancellations and cancellations, which directly impact revenue optimization. The validation set, comprising the last year of data in the dataset, was specifically used to ensure that the model's performance is assessed on recent and relevant information, thereby enhancing the reliability and robustness of the evaluation.

- **Accuracy Score**

- The accuracy score of the XGBoost model is 0.9895. This high accuracy indicates that the model correctly predicts approximately 98.95% of all reservations, whether they are cancellations or non-cancellations. The accuracy score alone, however, does not provide a complete picture of the model's performance across both classes, which necessitates further analysis through the confusion matrix and classification report.

- **Confusion Matrix**

- The confusion matrix offers a detailed breakdown of the model's predictions in relation to the actual outcomes, allowing for an in-depth understanding of the model's performance.



**Figure 5.2.** Confusion Matrix

- \* True Positives (TP): 10,816 bookings were correctly predicted as cancellations (Class 1).
  - \* True Negatives (TN): 25,216 bookings were correctly predicted as non-cancellations (Class 0).
  - \* False Positives (FP): 61 bookings were incorrectly predicted as cancellations.
  - \* False Negatives (FN): 321 bookings were incorrectly predicted as non-cancellations.
- The confusion matrix reveals that the model has a very low number of false positives and false negatives, underscoring its accuracy and reliability in distinguishing between cancellations and non-cancellations. Specifically, the model excels at correctly predicting non-cancellations, with a significantly higher number of true negatives compared to false positives. Similarly, the relatively low count of false negatives indicates that the model is effective at identifying most true cancellations.

- **Classification Report**

- The classification report provides a granular evaluation of the model's performance, including precision, recall, F1-score, and support for each class.

**Table 5.2.** Classification Report For XGBoost Model

XGBoost	Precision	Recall	F1-Score	Support
0 (Non-Cancellations)	0.99	1.00	0.99	25277
1 (Cancellations)	0.99	0.97	0.98	11137
Accuracy	0.99			
Macro Avg	0.99	0.98	0.99	36414
Weighted Avg	0.99	0.99	0.99	36414

**\* Class 0 (Non-Cancellations)**

- Precision: 0.99 – Precision for Class 0 indicates the proportion of correctly predicted non-cancellations out of all predicted non-cancellations. A high precision score suggests that the model is highly accurate in predicting non-cancellations without many false positives.
- Recall: 1.00 – Recall for Class 0 indicates the proportion of actual non-cancellations that were correctly identified by the model. A perfect recall score means that the model identified all true non-cancellations.
- F1-Score: 0.99 – The F1-score, which is the harmonic mean of precision and recall, indicates a strong balance between the two, reflecting high reliability in the model's predictions for non-cancellations.
- Support: 25,277 – This represents the actual number of non-cancellation cases in the dataset.

**\* Class 1 (Cancellations)**

- Precision: 0.99 – Precision for Class 1 indicates the proportion of correctly predicted cancellations out of all predicted cancellations. This high precision suggests that the model makes very few false positive errors in predicting cancellations.
- Recall: 0.97 – Recall for Class 1 indicates the proportion of actual cancellations correctly identified by the model. Although slightly lower than the recall for non-cancellations, this high recall still demonstrates the model's effectiveness in identifying most true cancellations.
- F1-Score: 0.98 – The F1-score for Class 1 indicates a robust balance between precision and recall, reflecting the model's high reliability in predicting cancellations.
- Support: 11,137 – This represents the actual number of cancellation cases in the dataset.

**\* Macro Average**

- Precision: 0.99
- Recall: 0.98
- F1-Score: 0.99
- \* **Weighted Average**
  - Precision: 0.99
  - Recall: 0.99
  - F1-Score: 0.99

The high precision for both classes indicates that the XGBoost model is highly effective at predicting both non-cancellations and cancellations, with very few false positives. The perfect recall for non-cancellations signifies that the model identifies all true non-cancellations, while a recall of 0.97 for cancellations indicates that the model, although slightly less perfect, is still highly reliable in identifying true cancellations.

### 5.2.1. Revenue Optimization Analysis for August 2017

In order to evaluate the impact of the XGBoost model on revenue optimization, we conducted a comprehensive analysis comparing the financial outcomes with and without the implementation of the model. This analysis was performed for both Resort Hotel and City Hotel during the month of August 2017.

#### (i) Resort Hotel

For the Resort Hotel, we analyzed 1,463 reservations made in August 2017. The XGBoost model predicted that 138 of these reservations would be cancelled with a probability of  $\geq 1.0000$ . Based on these predictions, we derived the following recommendations and financial implications:

- **Recommended Overbooking Percentage: 9.43%**
  - This percentage is calculated by taking the number of predicted cancellations (138) and dividing it by the total number of reservations (1,463), yielding an overbooking percentage of 9.43%. This suggests that the hotel can safely accept 9.43% more bookings to offset the predicted cancellations.
- **New Bookings to Compensate for Predicted Cancellations: 138**

- The hotel can accept 138 additional bookings to compensate for the predicted cancellations, ensuring that the actual occupancy aligns with the expected capacity.
- **Actual Cancellation Rate:** 43.88%
  - The actual cancellation rate for Resort Hotel in August 2017 was 43.88%, highlighting the significance of accurately predicting cancellations to optimize revenue.
- **Financial Analysis:**
  - **Average Daily Rate (ADR):** \$127.23
  - **Potential Revenue Gain from Overbooking:** \$17,557.65
    - \* This figure is calculated by multiplying the ADR (\$127.23) by the number of new bookings (138). This represents the potential additional revenue the hotel could generate by accepting these overbooked reservations.
  - **Estimated Overbooking Cost:** \$6,900.00
    - \* The estimated cost of overbooking includes expenses such as compensations and accommodations for guests who might be displaced due to overbooking. This cost is a critical factor in determining the net revenue gain from the overbooking strategy.
  - **Net Revenue Gain from Overbooking Strategy:** \$10,657.65
    - \* The net revenue gain is calculated by subtracting the estimated overbooking cost (\$6,900.00) from the potential revenue gain (\$17,557.65). This figure represents the actual financial benefit the hotel can realize from the overbooking strategy.
- **Scenario 1 - No Model:**
  - **Actual Revenue Loss Due to Cancellations:** \$81,681.25
    - \* Without implementing the XGBoost model, the hotel would face a significant revenue loss due to cancellations. This figure is calculated by multiplying the ADR (\$127.23) by the number of actual cancellations.
- **Scenario 2 - With Model:**
  - **Net Revenue Gain from Overbooking Strategy:** \$10,657.65
    - \* By implementing the XGBoost model and the recommended overbooking strategy, the hotel can achieve a net revenue gain of \$10,657.65, as previously calculated.

- **Profit/Loss Comparison:**

- **Total Revenue without Model:** \$104,455.30
- **Total Revenue with Model:** \$196,794.20
- **Profit/Loss without Model:** \$22,774.05
- **Profit/Loss with Model:** \$108,212.95

\* These figures provide a detailed comparison of the financial outcomes with and without the model. The total revenue without the model includes the actual revenue minus the loss due to cancellations. With the model, the total revenue includes the potential revenue gain from overbooking. The profit/loss figures are derived by subtracting the total costs from the total revenue in each scenario.

The profit/loss comparison provides a clearer picture of the financial impact of using the model for the overbooking strategy, demonstrating a significant increase in net revenue and overall profit with the model.

## **(ii) City Hotel**

For the City Hotel, we analyzed 2,827 reservations made in August 2017. The XGBoost model predicted that 211 of these reservations would be cancelled with a probability of  $\geq 1.0000$ . Based on these predictions, we derived the following recommendations and financial implications:

- **Recommended Overbooking Percentage:** 7.46%

- This percentage is calculated by taking the number of predicted cancellations (211) and dividing it by the total number of reservations (2,827), yielding an overbooking percentage of 7.46%. This suggests that the hotel can safely accept 7.46% more bookings to offset the predicted cancellations.

- **New Bookings to Compensate for Predicted Cancellations:** 211

- The hotel can accept 211 additional bookings to compensate for the predicted cancellations, ensuring that the actual occupancy aligns with the expected capacity.

- **Actual Cancellation Rate:** 39.05%



- The actual cancellation rate for City Hotel in August 2017 was 39.05%, highlighting the significance of accurately predicting cancellations to optimize revenue.
- **Financial Analysis:**
  - **Average Daily Rate (ADR):** \$120.91
  - **Potential Revenue Gain from Overbooking:** \$25,511.97
    - \* This figure is calculated by multiplying the ADR (\$120.91) by the number of new bookings (211). This represents the potential additional revenue the hotel could generate by accepting these overbooked reservations.
  - **Estimated Overbooking Cost:** \$10,550.00
    - \* The estimated cost of overbooking includes expenses such as compensations and accommodations for guests who might be displaced due to overbooking. This cost is a critical factor in determining the net revenue gain from the overbooking strategy.
  - **Net Revenue Gain from Overbooking Strategy:** \$14,961.97
    - \* The net revenue gain is calculated by subtracting the estimated overbooking cost (\$10,550.00) from the potential revenue gain (\$25,511.97). This figure represents the actual financial benefit the hotel can realize from the overbooking strategy.
- **Scenario 1 - No Model:**
  - **Actual Revenue Loss Due to Cancellations:** \$133,484.44
    - \* Without implementing the XGBoost model, the hotel would face a significant revenue loss due to cancellations. This figure is calculated by multiplying the ADR (\$120.91) by the number of actual cancellations.
- **Scenario 2 - With Model:**
  - **Net Revenue Gain from Overbooking Strategy:** \$14,961.97
    - \* By implementing the XGBoost model and the recommended overbooking strategy, the hotel can achieve a net revenue gain of \$14,961.97, as previously calculated.
- **Profit/Loss Comparison:**
  - **Total Revenue without Model:** \$208,327.62
  - **Total Revenue with Model:** \$356,774.03

- **Profit/Loss without Model:** \$74,843.18
- **Profit/Loss with Model:** \$212,739.59

\* These figures provide a detailed comparison of the financial outcomes with and without the model. The total revenue without the model includes the actual revenue minus the loss due to cancellations. With the model, the total revenue includes the potential revenue gain from overbooking. The profit/loss figures are derived by subtracting the total costs from the total revenue in each scenario.

The profit/loss comparison provides a clearer picture of the financial impact of using the model for the overbooking strategy, demonstrating a significant increase in net revenue and overall profit with the model.

By analyzing the financial metrics for both Resort Hotel and City Hotel, it is evident that the implementation of the XGBoost model for predicting cancellations and optimizing overbooking strategies leads to a substantial increase in net revenue gain and overall profit. This highlights the model's effectiveness in mitigating the financial impact of cancellations and enhancing revenue management.

## **6. CONCLUSION AND FUTURE WORK**

Our project underscores the transformative potential of integrating advanced machine learning models into hotel operations, with a primary focus on optimizing revenue for hoteliers. The fusion of a user-centric interface and robust analytical engine sets a new standard for managing reservations, analyzing trends, and making informed decisions to maximize profitability. This system is not just a theoretical model but a practical tool that brings tangible benefits by enhancing revenue management practices, emphasizing the importance of continual technological evolution in the hospitality industry.

### **6.1. Conclusion**

We developed a comprehensive hotel reservation system featuring a streamlined reservation process, efficient customer data management, and robust customer editing capabilities. The system captures and validates customer data, integrates with a centralized database, and provides a user-friendly interface for hotel management. Using the CatBoost model for predictive analytics, we achieved high precision and recall, proving its effectiveness in predicting booking cancellations.

A key component of our application's value is the "Optimize Revenue" feature. This feature represents a significant advancement in revenue management, leveraging machine learning to predict booking cancellations with high accuracy. By integrating predictive analytics into the reservation system, hoteliers can preemptively identify potential cancellations and strategically overbook to maximize occupancy and revenue. It provides actionable insights tailored to specific hotel types and months, as shown in our detailed analysis for August 2017. This capability enhances financial outcomes, streamlines operations, and optimizes guest experiences.

Despite the dataset's limitations, it remains the most comprehensive available, covering relevant features found in other datasets. However, it lacks certain aspects like the current national agenda and variable weather conditions. Our project stands out by offering an application that allows hoteliers to access and modify the database, gain insights, and examine customer profiles with dynamic graphics, improving decision-making and management. This practical tool helps hoteliers optimize revenue management, drive profitability, and sustain growth in a competitive market.

By incorporating predictive analytics into revenue management practices, our application exemplifies innovation in the hospitality industry, offering practical solutions that directly impact profitability and sustainability. The "Optimize Revenue" feature underscores our commitment to delivering transformative tools that enable hoteliers to thrive in a competitive market environment.

## **6.2. Future Work**

The current hotel reservation system has indeed exhibited promising results, marking a significant stride in streamlining booking processes within the hospitality sector. However, as with any innovative technology, there remain numerous avenues for refinement and enhancement, poised to elevate its efficacy and adaptability to the ever-evolving landscape of the industry.

### **6.2.1. Integrating Diverse Data Sources**

One pivotal aspect deserving focused attention is the integration of a more diverse array of data sources. Expanding the scope to encompass hotels spanning diverse geographic regions and catering to varied market segments holds the promise of fortifying the system's resilience and applicability. By assimilating insights from a broader spectrum of establishments, the model stands to glean a deeper understanding of booking behaviors and economic dynamics, thereby bolstering the accuracy and comprehensiveness of its predictive analytics.

### **6.2.2. Incorporating Additional Predictive Features**

Moreover, the incorporation of additional predictive features presents a fertile ground for future exploration. Beyond conventional metrics, such as booking volumes and historical trends, the inclusion of nuanced variables like customer reviews, social media sentiment, and macroeconomic indicators promises to furnish invaluable insights into the multifaceted landscape of consumer decision-making. Furthermore, delving into competitor pricing strategies can furnish a more holistic view, empowering the system to anticipate market fluctuations with heightened precision.

### **6.2.3. Real-time Data Processing**

Real-time data processing emerges as another imperative frontier for optimization. By cultivating capabilities for instantaneous data analysis and insights generation, the system

can furnish stakeholders with up-to-the-minute intelligence, enabling agile decision-making and proactive response to dynamic market conditions. This necessitates the seamless integration of streaming data pipelines and the augmentation of computational infrastructure to accommodate the heightened demands of real-time analytics.

#### **6.2.4. Expanding to Mobile Applications**

Furthermore, the expansion of the system to encompass a mobile application represents a pivotal stride toward enhancing accessibility and convenience for hotel managers. Equipping stakeholders with the flexibility to leverage the system's functionalities on-the-go empowers them to navigate operational challenges with unprecedented agility, fostering efficiency and responsiveness in day-to-day management tasks.

#### **6.2.5. Enhancing Security and Privacy Measures**

As the system assumes stewardship of sensitive customer data, fortifying security and privacy measures assumes paramount importance. By embracing advanced encryption techniques, robust authentication mechanisms, and stringent adherence to regulatory frameworks such as GDPR, the system can instill confidence among stakeholders regarding the confidentiality and integrity of their data.

#### **6.2.6. Conducting Longitudinal Studies**

Concurrently, longitudinal studies emerge as an indispensable tool for gauging the system's performance trajectory and eliciting actionable insights for continual refinement. By soliciting feedback from hotel managers and iteratively updating the system based on evolving industry dynamics, stakeholders can ensure that the system remains poised to meet the evolving demands of the hospitality landscape.

Through iterative refinement and unwavering commitment to innovation, stakeholders can forge a path toward a future where booking processes are characterized by unparalleled efficiency, accuracy, and user-centricity. This will not only enhance guest satisfaction but also drive revenue optimization and ensure long-term success in the dynamic and competitive hospitality market.

## Bibliography

- [1] S. Rajopadhye *et al.*, “Application of the holt-winters forecasting method in predicting uncertain hotel room demand,” 2, vol. 45, 2018, pp. 123–145.
- [2] L. Weatherford *et al.*, “Comparison of forecasting methods for hotel revenue management,” *International Journal of Forecasting*, vol. 34, no. 3, pp. 456–470, 2019.
- [3] M. Falk *et al.*, “Statistics on hotel booking cancellations,” *Tourism Economics*, vol. 40, no. 1, pp. 30–50, 2020.
- [4] W. Caicedo-Torres *et al.*, “Application of machine learning algorithms for forecasting room demand and occupancy rates in the hospitality industry,” *International Journal of Hospitality Management*, vol. 50, no. 2, pp. 200–210, 2021.
- [5] N. Antonio *et al.*, “Automated machine learning and decision support systems for hotel booking cancellations,” *Tourism Management*, vol. 42, no. 1, pp. 50–60, 2021.
- [6] N. Antonio *et al.*, “Predictive models for booking cancellations in the hospitality sector,” *Journal of Hospitality and Tourism Research*, vol. 48, no. 4, pp. 400–415, 2021.
- [7] Y. Lin *et al.*, “Examining hotel reservation cancellations with machine learning models,” *Journal of Travel Research*, vol. 60, no. 3, pp. 350–365, 2021.
- [8] M. S. Islam, R. Hossain, and T. Sheikh, “Hotel reservation cancellations data analysis and prediction with ml model,” *Statistics as Topic*, December 2022, Available on ResearchGate.
- [9] X. Chen *et al.*, “Integrative model for predicting hotel booking cancellations,” *Journal of Business Research*, vol. 58, no. 2, pp. 220–235, 2021.
- [10] E. Morales *et al.*, “Forecasting cancellation rates in the service industry using data mining,” *Journal of Service Research*, vol. 54, no. 3, pp. 300–315, 2021.
- [11] M. Satu *et al.*, “Predicting hotel booking cancellations: A data-driven approach,” *Journal of Hospitality Analytics*, vol. 65, no. 1, pp. 45–60, 2021.