DESIGN ANALYSIS APPROACH FOR HOTEL BOOKING CANCELLATION PREDICTION

by

Ayush Kumar (1901430130014)

Chittranjan Kumar (1901430130016)

Kanik Goel (1901430130027)

Under the Supervision of

Ms Yashi Bhardwaj

(Assistant Professor)

Department of Information Technology

IMS Engineering College affiliated to A.K.T.U, Lucknow

NH-09, Adhyatmik Nagar, Near Dasna

May, 2023

VISION OF THE INSTITUTE

To make IMSEC an Institution of Excellence for empowering students through technical education coupled with incorporating values and developing engineering acumen for innovations and leadership skills for the betterment of society.

MISSION OF THE INSTITUTE

1. To promote academic excellence by continuous learning in core and emerging Engineering areas using innovative teaching and learning methodologies.
2. To inculcate values and ethics among the learners.
3. To promote industry interactions and produce young entrepreneurs.
4. To create a conducive learning and research environment for life-long learning to develop the students as technology leaders and entrepreneurs for addressing societal needs.

VISION OF THE DEPARTMENT

To be a department of excellence by imparting state-of-the-art technical education and preparing globally competent professionals to contribute innovatively to the real-time requirements of industry and society.

MISSION OF THE DEPARTMENT

M1. To provide strong fundamental and technical skills through effective teaching-learning practices and hands-on experience with the latest tools and technologies.

M2. To encourage students to become industry-ready professionals by possessing multidisciplinary skills, leadership abilities, and research-oriented understanding.

M3. To impart entrepreneurship skills, and develop a sense of respect for social values and professional ethics among the upcoming IT professionals.

PROGRAM OUTCOMES (POs)

Engineering Graduates will be able to:

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for **sustainable development.**

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and teamwork: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12: Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM EDUCATIONAL OUTCOMES (PEOs)

The Graduates will be able to:

PEO1: develop strong competency to formulate, analyze, and solve problems of the IT industry using the necessary mathematical, scientific, and engineering fundamentals.

PEO2: apply the technical knowledge and competency for a successful career in the software industry and progressively hold more responsible positions.

PEO3: demonstrate ethical behavior as technical professionals and a sense of responsibility towards the impact of technology on society.

PEO4: demonstrate critical thinking, professional communication, teamwork, and entrepreneurial skills necessary for high productivity towards nation-building with a commitment of pursuing lifelong learning.

PROGRAM SPECIFIC OUTCOME (PSOs)

Upon completion of this program, the student will be able to:

PSO1: apply knowledge and skills required for software development, database administration, and entrepreneurship in emerging fields like artificial intelligence, data analytics, networking, and cloud computing.

PSO2: apply programming languages, tools and techniques to demonstrate the acquired technical skills for seeking solutions to the problems of various interdisciplinary challenges.

CO-PO-PSO MAPPING FOR ACADEMIC SESSION 2022-23

Course Name: Hotel Booking Cancellation Predication Model

AKTU Course Code: KIT851

Semester / Year: VIII/ 4th

NBA Code: C412

Subject Coordinator: Ms. Yashi Bhardwaj

Course Outcomes:

CO. No.	DESCRIPTION	COGNITIVE LEVEL (BLOOMS TAXONOMY)
C412.1	Analyse and understand the real-life problem and apply their knowledge to get programming solution.	K4 , K5
C412.2	Engage in the creative design process through the integration and application of diverse technical knowledge and expertise to meet customer needs and address social issues.	K4 , K5
C412.3	Use the various tools and techniques, coding practices for developing real life solution to the problem.	K5 , K6
C412.4	Find out the errors in software solutions and establishing the process to design maintainable software applications.	K4 , K5
C412.5	Write the report about what they are doing in project and learning the team working skills.	K5, K6

CO-PO-PSO Mapping:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO 10	PO 11	PO 12	PSO1	PSO2
C412.1	3	3	3	3	3	1	1	2	2	3	3	3	3	3
C412.2	3	3	3	3	3	1	1	2	2	3	3	3	3	3
C412.3	3	3	3	3	3	1	1	2	2	3	3	3	3	3
C412.4	3	3	3	3	3	1	1	2	2	3	3	3	3	3
C412.5	3	3	3	3	3	1	1	3	3	3	3	3	3	3
Avg.	**3**	**3**	**3**	**3**	**3**	**1**	**1**	**2.2**	**2.2**	**3**	**3**	**3**	**3**	**3**

DECLARATION

I hereby declare that the work, which is being presented in the Project, entitled **"Design an analysis of Hotel Booking Cancellation Predication"** in partial fulfilment for the award of degree of **"Bachelor of Technology"** in **Information Technology**, and **submitted to the Department of Information Technology**, IMS Engineering College, Ghaziabad, affiliated to Dr. A.P.J Abdul Kalam Technical University, Uttar Pradesh, Lucknow is a record of my own investigations carried under the guidance of **Ms. Yashi Bhardwaj, Assistant Professor**, IMS Engineering College, Ghaziabad.

I have not submitted the matter presented in this Project anywhere for the award of any other Degree.

Signature with Date:

Name: Ayush Kumar

Roll No: 1901430130014

Signature with Date:

Name: Chittranjan Kumar

Roll No: 1901430130016

Signature with Date:

Name: Kanik Goel

Roll No: 1901430130027

Signature with Date:

CERTIFICATE

I hereby certify that the work which is being presented in the project report entitled **"Design an analysis of Hotel Booking Cancellation Predication"** by **"Ayush Kumar, Chittranjan Kumar, Kanik Goel"** in partial fulfillment of requirements for theaward of degree of B.Tech. (IT) submitted in the Department of IT at **"IMS Engineering College"** under A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, LUCKNOW is an authentic record of my own work carried out under the supervision of **Ms Yashi Bhardwaj.**

Prof. Yashi Bhardwaj
Project Supervisor
Assistant Professor, IT
IMSEC, Ghaziabad

Counter sign by:

Dr. Pushpendra Singh
HOD, IT
IMSEC, Ghaziabad

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to ***(*Ms Yashi Bhardwaj, *Assistant Professor)*** Department of Information Technology, IMSEC, Ghaziabad, India for her generous guidance, help and useful suggestions.

I express my sincere gratitude to Mr. Updesh Kumar Jaiswal, Project Coordinator in Department of Information Technology, IMSEC, Ghaziabad, for his stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I am extremely thankful to Prof. Vikram Bali, Director, IMSEC, Ghaziabad, for providing me infrastructural facilities to work in, without which this work would not have been possible.

Signature with date:
Name: Ayush Kumar
Roll No: 1901430130014
Signature with date:
Name: Chittranjan Kumar
Roll No: 1901430130016
Signature with date:
Name: Kanik Goel
Roll No: 1901430130027
Signature with date:

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

a. ML: Machine Learning
b. AI: Artificial Intelligence
c. DL: Deep Learning
d. ANN: Artificial Neural Network
e. CNN: Convolutional Neural Network
f. RNN: Recurrent Neural Network
g. RF: Random Forest
h. KNN: K-Nearest Neighbors
i. TP: True Positive
j. FP: False Positive
k. TN: True Negative
l. FN: False Negative

DESIGN AN ANALYSIS APPROACH OF HOTEL BOOKING CANCELLATION PREDICTION

by

Ayush Kumar (1901430130014)

Chittranjan Kumar (1901430130016)

Kanik Goel (1901430130027)

Under the Supervision of

Ms Yashi Bhardwaj

(Assistant Professor)

Department of Information Technology

IMS Engineering College affiliated to A.K.T.U, Lucknow

NH-09, Adhyatmik Nagar, Near Dasna,

District: Ghaziabad, UP

May, 2023

ABSTRACT

Hotel managers find it beneficial to predict hotel booking cancellations as it enables them to enhance room inventory management, pricing strategies, and customer satisfaction by proactively addressing potential problems. In this study, we present a machine learning-centered method for forecasting hotel booking cancellations.

Initially, we will gather and pre-process the necessary data for the analysis. This process entails acquiring a dataset containing information about previous hotel bookings, encompassing customer details, booking specifics, and cancellation status. Subsequently, the data will undergo cleaning and preparation for modeling, involving tasks such as handling missing values, transforming categorical variables into numerical representation, and scaling numerical variables.

Following that, we will proceed with the selection and implementation of suitable machine learning algorithms to construct predictive models. We will assess and compare the performance of various algorithms, such as logistic regression, decision trees, and random forests, utilizing a range of evaluation metrics like accuracy, precision, and recall. Additionally, we will explore the potential benefits of ensemble methods, which combine predictions from multiple models to enhance overall performance.

Once we identify the most effective model, we will evaluate its performance on a separate test set to determine its ability to generalize to new data. Furthermore, we will conduct a feature importance analysis to identify the key factors influencing hotel booking cancellations.

In addition to developing predictive models, we will conduct a comprehensive analysis of the factors contributing to hotel booking cancellations. This analysis will involve examining the relationships between various features such as booking lead time, length of stay, and customer demographics, and their impact on the likelihood of cancellation. These insights will assist in devising strategies to mitigate cancellation rates, such as targeted marketing campaigns or incentives for early bookings.

Overall, our proposed approach will equip hotel managers with a robust tool for predicting hotel booking cancellations and a deeper understanding of the factors influencing them. This, in turn, will enable data-driven decision-making to optimize room inventory, pricing strategies, enhance customer satisfaction, and ultimately boost revenue.

Chapter-1

INTRODUCTION

The hotel industry thrives on effective management of bookings and occupancy rates. Hotel managers face the challenge of optimizing room inventory and pricing while ensuring customer satisfaction. One critical aspect of this challenge is predicting hotel booking cancellations. Anticipating cancellations can enable hotel managers to proactively address potential issues, adjust pricing strategies, and make better decisions regarding room availability. In recent years, advancements in machine learning and data analytics have opened new opportunities to develop accurate prediction models for hotel booking cancellations.

This project aims to propose and implement a Hotel Booking Cancellation Prediction Model using machine learning techniques. The primary goal is to assist hotel managers in optimizing their operations by accurately forecasting booking cancellations. By doing so, they can effectively manage room inventory, minimize revenue losses, and enhance customer satisfaction.

The proposed model will leverage historical data on hotel bookings, including information about customers, booking details, and whether the bookings were cancelled or not. Through a series of data pre-processing steps, including handling missing values, converting categorical variables into numerical form, and scaling numerical variables, the dataset will be prepared for modelling.

Various machine learning algorithms will be explored and compared to identify the best-performing model. Algorithms such as logistic regression, decision trees, and random forests will be evaluated using standard evaluation metrics like accuracy, precision, and recall. Additionally, ensemble methods, which combine the predictions of multiple models, will be considered to potentially improve the model's performance.

To evaluate the model's generalization ability, a hold-out test set will be utilized. This evaluation will provide insights into how well the model performs on unseen data, determining its practical utility in real- world scenarios. Furthermore, a feature importance analysis will be conducted to identify the key factors influencing hotel booking cancellations. Understanding these factors will enable hotel managers to prioritize specific strategies to reduce cancellation rates, such as targeted marketing campaigns or early booking incentives.

In conclusion, this project aims to develop a robust Hotel Booking Cancellation Prediction Model that will assist hotel managers in optimizing their operations. The model's accurate predictions will facilitate effective decision-making regarding room inventory management, pricing strategies, and customer satisfaction. By leveraging machine learning techniques and analysing the factors that contribute to cancellations, hotel managers can make data-driven decisions to increase revenue and create a more streamlined booking process.

1.2 Principle of the Project

The objective of this project is to employ machine learning techniques to forecast hotel cancellations and analyze the factors that contribute to them. Specifically, we will investigate the following inquiries:

• Q1: Is it possible to predict hotel guest cancellations using machine learning algorithms, considering the available dataset?

• Q2: Which factors have the greatest impact on predicting cancellations?

1.3 Benefits of the Research

The utilization of a Hotel Booking Cancellation Prediction Model powered by machine learning offers numerous benefits to hotel managers and the hospitality industry. Here are some key advantages of implementing such a model:

1. Enhanced Inventory Management: By accurately predicting booking cancellations, hotel managers can optimize their room inventory. This enables them to allocate resources more effectively, ensuring optimal occupancy levels and minimizing revenue losses due to unoccupied rooms.
2. Improved Pricing Strategies: The prediction model allows hotel managers to adjust their pricing strategies based on the likelihood of cancellations. They can dynamically modify prices to attract potential guests while considering the risk of cancellations, maximizing revenue generation.
3. Proactive Issue Resolution: Anticipating cancellations empowers hotel managers to address potential issues in advance. They can reach out to guests with upcoming bookings who are at a higher risk of cancelling, offering personalized incentives or assistance to encourage them to keep their reservations. This proactive approach improves customer satisfaction and loyalty.
4. Resource Optimization: By accurately predicting cancellations, hotels can optimize resource allocation and minimize waste. They can adjust staffing levels, food and beverage supplies, and other operational aspects based on expected occupancy rates, leading to cost savings and efficient resource management.
5. Strategic Decision-Making: The prediction model provides valuable insights into the factors that influence cancellations. By analyzing these factors, hotel managers can make informed decisions and implement targeted strategies to mitigate cancellations. This may involve adjusting marketing campaigns, improving customer service, or implementing policies that reduce cancellation rates.

6. Competitive Advantage: Adopting a data-driven approach to predict hotel booking cancellations gives establishments a competitive edge. It allows them to stay ahead of market trends, adapt to changing customer behavior, and offer superior services tailored to guest preferences.

Overall, the implementation of a Hotel Booking Cancellation Prediction Model using machine learning empowers hotel managers to optimize operations, increase revenue, improve customer satisfaction, and gain a competitive advantage in the dynamic hospitality industry.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

- Mehrotra (2006) noted that precise demand forecasting is a key determinant of revenue management. Talluri (2004) have recognized the importance of revenue management forecasting by confirming that revenue management systems need a quantity forecast, and more precisely, " its performance depends critically on the quality of these forecasts".
- Because of the need for predicted demand, the cancellation of reservations, as in the hospitality industry and other service industries that deal with advanced reservations, do not show the true demand for their services, as there are often a insignificant number of cancellations (Morales et al., 2010).
- The cancellation of bookings is a well-known issue in the revenue management sector related to the service industries, and especially to the hospitality industry. With the growing effect of the internet on the way consumers search and purchase travel services in recent years (Noone & Lee, 2010), research in this topic have been increased, and particularly on the subject of controls used to mitigate the effects of Cancellations on revenue allocation, cancelation policies and overbooking (Ivanov, 2014; Talluri et al., 2004).
- In addition to that, other authors like Morales & Wang (2010) and Ivanov & Zecher (2012) acknowledged the crucial role of demand forecast where forecasting is crucial.
- Actually, Morales (2010) describes that “it is hard to imagine that one can predict whether a booking will be cancelled or not with high accuracy simply by looking at PNR information”. Although, it is suggested in the next chapters that the classification of whether a room reservation will be cancelled is feasible.
- According to Ivanov (2014), the registration of cancellations is an important factor for recognizing data trends and thus creating better forecasts, overbooking and cancellation policies.
- According to A. R. Bajaj and N. N. Jani (2017), the study by Bajaj and Jani focused on hotel booking cancellation prediction using data mining techniques.

They employed various data mining algorithms, including decision trees and Naïve Bayes, to analyze factors such as booking channel, booking date, and customer characteristics in predicting cancellations.

- According to Li and Wang (2018), Li and Wang conducted research on hotel booking cancellation prediction using Support Vector Machines (SVM).

  Their study highlighted the effectiveness of SVM in capturing complex patterns and relationships between variables, such as booking lead time, arrival date, and room type, to predict cancellations accurately.
- According to R. Prathiba, P. J. Ravi, and M. A. Saleem Durai (2019),the study by Prathiba et al. compared various machine learning techniques for hotel booking cancellation prediction. They evaluated the performance of algorithms such as decision trees, Random Forest, and K- Nearest Neighbors, considering factors like booking date, customer demographics, and length of stay.
- According to Fang and Ren (2019), Fang and Ren focused on predicting hotel booking cancellations using neural networks. They explored the application of different neural network architectures, such as Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM), and evaluated their performance using features like booking lead time, customer characteristics, and room type.
- According to A. M. Mahmood and H. A. Ali (2020), Mahmood and Ali conducted a comparative study of machine learning algorithms for hotel booking cancellation prediction. They compared the performance of algorithms, including Decision Trees, Naïve Bayes, and Random Forest, using variables such as booking date, room rate, and customer demographics.
- According to Li and Hu (2020), Li and Hu proposed a machine learning approach for predicting hotel booking cancellations. Their study incorporated features like booking channel, booking lead time, and customer review scores to develop prediction models based on algorithms such as Logistic Regression and Random Forest.
- According to Kim and Yoon (2020), the prediction of hotel booking cancellations using machine learning techniques. They employed algorithms like Logistic Regression, Random Forest, and Gradient Boosting Machines, considering factors such as booking lead time, hotel rating, and cancellation history.
- According to Liu and Cao (2021), they focused on hotel booking cancellation prediction based on machine learning algorithms. Their study examined the performance of algorithms such as Support Vector Machines (SVM), Random Forest, and Gradient Boosting Machines, using variables like booking date, hotel location, and customer characteristics.

2.2 INFERENCE FROM LITERATURE

There are several factors that can influence the likelihood of a hotel booking being cancelled, and machine learning models can be used to predict cancellations based on these factors. Some potential factors that could be considered in a machine learning model for predicting hotel booking cancellations include:

- Booking date: Bookings made closer to the date of arrival are more likely to be cancelled compared to bookings made well in advance.
- Length of stay: Longer stays are generally less likely to be cancelled compared to shorter stays.
- Rate type: Non-refundable rate types are generally less likely to be cancelled compared to refundable rate types.
- Booking source: Bookings made through online travel agencies (OTAs) may be more likely to be cancelled compared to bookings made directly through the hotel's website.
- Customer demographics: Certain demographic groups may be more likely to cancel their bookings. For example, younger travelers may be more likely to cancel due to last-minute changes in plans.
- Destination: Popular tourist destinations may have higher rates of cancellation due to changes in travel plans or unexpected events.
- Seasonality: Cancellation rates may vary seasonally, with higher rates during peak travel periods.
- By considering these and other factors, machine learning models can be trained to predict the likelihood of a hotel booking being cancelled. These predictions can be useful for hotels in managing their room inventory and forecasting demand.

➢ METHODS

A. Predicting Hotel Bookings Cancellation with a Machine Learning Classification

Booking cancellation is a very common thing in today's world, which can cause severe losses to the business owners. This paper describes how AI is used to identify which booking can be cancelled and prevent some los ses. The machine learning model should be evaluated in the real time environment for accuracy. Prediction m odel of hotel booking cancellation no doubt the issue that can be resolved in the context of Design Science Re search (DSR), as it need to develop an artifact, here in this particular case, a form of Revenue Management Sy stem (RMS), fulfilling the two requirements of DSR.

B. Rising rate of Cancellation in the Hotel Industry

The growing trend of Hotels Industry is beneficial for Hotels but there are some problems too such as Rising Rate of C ancellation. The user cancels the booking of the hotel after seeing the reviews given by the people who booked the hot el already.

In Some case hotel owners treat the customer in bad way, this also affect the reputation as well as cancellations. The g rowing trend of Hotels Industry is beneficial for Hotels but there are some problems too such as Rising Rate of Cance llation. The user cancels the booking of the hotel after seeing the reviews given by the people who booked the hotel al ready. Now a days, people expecting a better accommodation at the Hotel site, if people found any lag in accommodat ion then they give poor rating of that Hotel. So, if we looking at percentage of cancellation then we found that the per centage of cancellation is increasing day by day. From a survey Cancellation rate rose from under 33% in 2014 to 40 % in 2018. Also, during the COVID-19 pandemic this percentage gets increased because peoples book their Hotels in very earlier time and after changing situation day by day during pandemic and State Government implementing Lockd own in particular State, people sudden changes their plans and cancel the Hotel Bookings.

C. Aspect based Sentiment Oriented Summarization of Hotel Reviews

The reviews and the feedbacks of the customer play an important role in the image as well as the revenue system of th e hotel. But most of the travelers don't read all reviews. The system analyzes the reviews and feedback by the custome rs. The feedbacks of the customer are gathered from the hotel's website and the stored as classes.

As per the study, the model analyzes the overlooked information by the customers and takes some essential steps. Fin ally, after processing all the data collected an emotional analysis is done. The hotels thereby can take the required step s to improve their service.

D. Application Of machine Learning in Hotel Industry A Critical Review

The growth in IT industry also affects the Hotel Industry. However, this change is quite slow. Many researchers are fo cused on testing and applying new artificial intelligence technology and learning equipment in the hotel industry. The study offers a brief knowledge about the use of Machine Learning and its combined technology in the hotel and touris m industry. Machine learning is quite trending these days.

CHAPTER 3
METHODOLGY

3.1 Processing Steps

There are four processing steps. Following approach is set for our approach.

a) Collection of Data

As you are aware, machines initially acquire knowledge from the data provided to them. It is crucial to gather dependable data to enable your machine learning model to identify accurate patterns. The accuracy of your model depends on the quality of the data you input. If you have erroneous or outdated data, the outcomes or predictions will be incorrect and irrelevant.

It is essential to ensure that you utilize data from a trustworthy source as it directly impacts the results of your model. High-quality data is pertinent, with minimal missing or duplicated values, and encompasses a comprehensive representation of various subcategories or classes.

To compare the accuracy of different machine learning models, a sample dataset is employed, which is available on KAGGLE. The dataset comprises 32 attributes and 119,390 records.

b) Cleaning Data

i) The accuracy of a machine learning model greatly depends on the cleanliness of the data, meaning it is free from any unwanted noise. Data cleaning refers to the process of eliminating corrupted or erroneous records from a dataset.

ii) Examples of bad data include:

a) Duplicated data, which can be addressed by utilizing the "drop duplicates ()" method to remove duplicates.

b) Incorrect data, which can be dealt with by deleting the respective column, especially in the

case of large datasets.

c) Data in the wrong format, which needs to be converted or transformed into the correct format.

d) Empty cells, where values can be filled in by using measures such as mean, mode, or median.

c) Choosing a Model

1. Logistic Regression:

- Logistic Regression is a supervised learning algorithm used for binary classification problems. It is based on the concept of linear regression but uses a logistic function to map the input features to a probability of belonging to a specific class.
- Logistic Regression assumes a linear relationship between the input features and the log-odds of the output class. It computes the probability of the output class using the logistic function, which squashes the linear regression output into the range [0, 1]. It then makes predictions by assigning the class with the highest probability.

2. Decision Tree:

- Decision Tree is a supervised learning algorithm that can be used for both classification and regression tasks. It builds a flowchart-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or class label.
- Decision Trees are constructed by recursively partitioning the data based on the feature that best splits the data according to a certain criterion, such as maximizing information gain or minimizing Gini impurity. They are easy to understand and interpret and can handle both numerical and categorical distances.

3. Random Forest

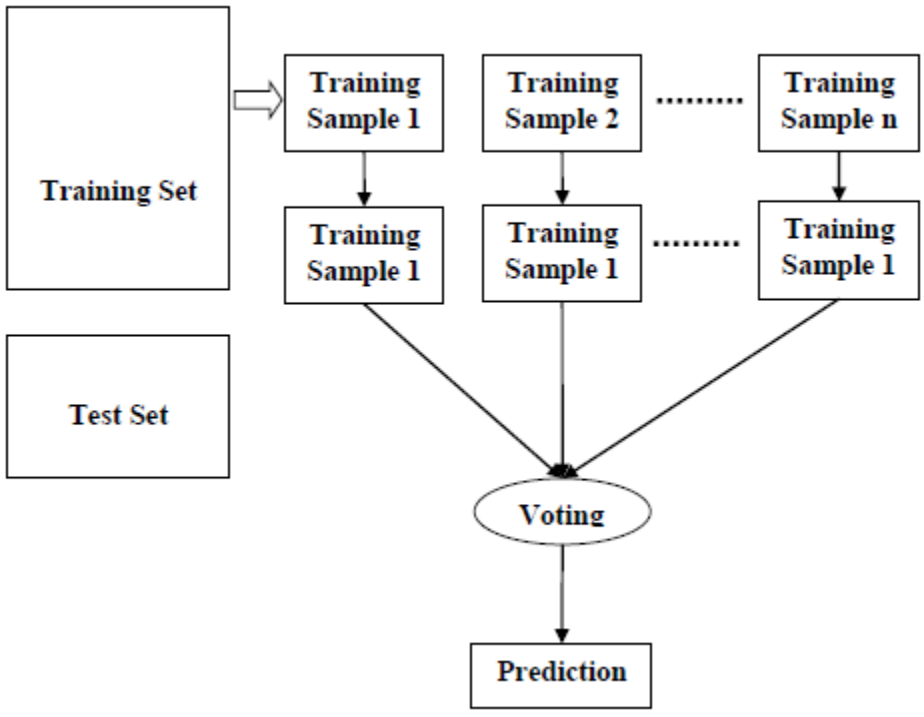
Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

- **Step 1** − First, start with the selection of random samples from a given dataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result.

The following diagram will illustrate its working −



4. KNN (K-Nearest Neighbors):

- KNN is a supervised learning algorithm used for both classification and regression tasks. It works based on the principle that objects with similar characteristics tend to belong to the same class or have similar output values.
- KNN determines the class or value of an unseen instance by comparing it to the k nearest neighbors in the training data. The value of k, a user-defined parameter, determines the number of neighbors considered. In classification, the class of the majority of the nearest neighbors is assigned to the unseen instance, while in regression, the average or weighted average of the neighbors' values is used.

5. Naive Bayes:

- Naive Bayes is a supervised learning algorithm based on Bayes' theorem and the assumption of feature independence. It is commonly used for text classification and spam filtering tasks.
- Naive Bayes calculates the probability of a given class label based on the presence of certain features. It assumes that the features are conditionally independent, meaning that the presence of one feature does not affect the presence of another feature. Despite this "naive" assumption, Naive Bayes can be surprisingly effective in practice and is computationally efficient. It uses the Bayes' theorem to update the probability estimates based on the observed features and makes predictions by selecting the class label with the highest probability.

3.2 Data Understanding

Table 3.1. Variables extracted from each booking from Booking database.

Name	Type	Description
Name	Type	Description
ADR	Numeric	Average daily rate
Adults	Number	Number of adults
AgeAtBookingDate	Number	Age in years of the booking holder at the time of booking
Agent	Categorical	ID of agent (if booked through an agent)
ArrivalDateDayOfMonth	Numeric	Day of month of arrival date (1 to 31)
ArrivalDateDayOfWeek	Categorical	Day of week of arrival date (Monday to Sunday)
ArrivalDateMonth	Categorical	Month of arrival date
ArrivalDateWeekNumber	Numeric	Number of week in the year (1 to 52)
AssignedRoomType	Categorical	Room type assigned to booking
Babies	Numeric	Number of babies
BookingChanges	Numeric	Heuristic created by summing the number of booking changes (amendments) prior to arrival that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meal, ADR, or reserved room type)
BookingDateDayOfWeek	Categorical	Day of week of booking date (Monday to Sunday)
CanceledTime	Numeric	Number of days prior to arrival that booking was canceled; when booking was not canceled it had the value of –1
Children	Numeric	Number of children
Company	Categorical	ID of company (if an account was associated with it)
Country	Categorical	Country ISO identification of the main booking holder
CustomerType	Categorical	Type of customer (group, contract, transient, or transient-party); this last category is a heuristic built when the booking is transient but is fully or partially paid in conjunction with other bookings (e.g., small groups such as families who require more than one room)
DaysInWaitingList	Numeric	Number of days the booking was in a waiting list prior to confirmed availability and to being confirmed as a booking
DepositType	Categorical	Because no specific field in the database existed with the type of deposit, based on how hotels operate, a heuristic was developed to define deposit type (nonrefundable, refundable, no deposit): payment made in full before the arrival date was considered a “nonrefundable” deposit, partial payment before arrival was considered a “refundable” deposit, otherwise it was considered as “no deposit”
DistributionChannel	Categorical	Name of the distribution channel used to make the booking
IsCanceled	Categorical	Outcome variable; binary value indicating if the booking was canceled (0: no; 1: yes)
IsRepeatedGuest	Categorical	Binary value indicating if the booking holder, at the time of booking, was a repeat guest at the hotel (0: no; 1: yes); created by comparing the time of booking with the guest history creation record
IsVIP	Categorical	Binary value indicating if the guest should be considered a Very Important Person (0: no; 1: yes)
LeadTime	Numeric	Number of days prior to arrival that the booking was placed in the hotel
LenghtOfStay	Numeric	Number of nights the guest stayed at the hotel
MarketSegment	Categorical	Market segmentation to which the booking was assigned
Meal	Categorical	ID of meal the guest requested

PreviousCancellations	Numeric	Number of previous bookings to this booking the guest had that were canceled
PreviousStays	Numerical	Number of nights the guest had stayed at the hotel prior to the current booking
RequiredCarParkingSpaces	Numeric	Number of car parking spaces the guest required
ReservedRoomTypes	Categorical	Room type requested by the guest
RoomsQuantity	Numeric	Number of rooms booked
StaysInWeekendNights	Numeric	From the total length of stay, how many nights were in weekends (Saturday and Sunday)
StaysInWeekNights	Numeric	From the total length of stay, how many nights were in weekdays (Monday through Friday)
TotalOfSpecialRequests	Numeric	Number of special requests made (e.g., fruit basket, sea view, etc.)
WasInWaitingList	Categorical	Binary value indicating if the guest was in a waiting list prior to confirmed availability and to being confirmed as an effective booking (0: no; 1: yes)
PreviousBookingsNotCanceled	Numeric	Number of previous bookings to this booking the guest had that were not canceled

3.3 Data Preparation

In this section, the final data sets for the hotel booking cancellation prediction model development were prepared, taking into account the insights gained during the data exploration and quality verification stages. The data preparation process involved the removal of certain observations (rows) and variables (columns) based on previous considerations. Additionally, the "mutual information feature selection filter" was used to validate the data selection process.

Data Selection and Removal

The original data sets were carefully reviewed, and observations and variables were removed based on the findings from the data exploration and quality verification steps. The goal was to retain only the most relevant and informative data for the prediction model development.

To assist in the process of data selection, the "mutual information feature selection filter" was applied. This filter provided additional information about the relevance of each variable and helped make informed decisions regarding the inclusion or exclusion of specific features.

Based on the results obtained from the mutual information feature selection filter and considering the computational requirements of the model, the following variables were identified as less significant and were removed from all data sets:

1. "BookingDateDayOfWeek": This variable, representing the day of the week when the booking was made, was found to have limited impact on the prediction of hotel booking cancellations.

2. "ArrivalDateDayOfWeek": Similarly, this variable, indicating the day of the week for the arrival date, was determined to have minimal influence on the prediction model.

Derived Feature Creation

By introducing this derived feature, we aimed to capture the cancellation behavior in a more meaningful way, consolidating the information from the original columns into a single representative feature. This derived feature was deemed to provide better predictive power for the hotel booking cancellation model.

The data preparation phase successfully resulted in the creation of the final data sets that will be utilized for the development of the hotel booking cancellation prediction model. These refined data sets, after undergoing the necessary modifications and feature engineering, are now ready for the subsequent stages of the project.

CHAPTER – 4

PROBLEM STATEMENT AND SOLUTION APPROACH

4.1 PROBLEM STATEMENT

For hotel management, hotel management is complex because hotels must deal with the customer directly and they have their own demand. A hotel cannot find the optimal solution to customers' demands.

For example, a customer’s special, mismatched information, or has time issues. Therefore, a hotel must try to cause the least problems by predicting the potential for various kinds of problems to solve those problems.

Hotel booking cancellation affects the management of the hotel in many aspects such as affects the management of room turnover, loss of revenue from unused rooms, etc. From these problems, it can be observed that if there is a method to predict cancellations. This will lessen these problems. And it also gives the hotel better management and more income.

4.2 SOLUTION APPROACH

- The hotel needs to know the variables that cause the booking cancellation. The purpose of this project is to make the hotel aware of the factors that cause the booking cancellation. To make the hotel improve and fix the problem areas to satisfy the needs of hotel's customers and reduce the cancellation of the booking.

- If the hotel can predict that the guest wishes to cancel the reservation. The hotel will then be able to manage the unoccupied room and if the hotel can predict it. It will help the hotel to cost less and increase revenue for the hotel as well.

CHAPTER 5
DESIGN and ANALYSIS

Technical Requirements:

4.1 Software

OPERATING SYSTEM	WINDOWS 10
PROGRAMMING LANGUAGE	PYTHON
DATABASE	MYSQL

4.2 Hardware

PROCESSOR	3.4GZ INTEL
HDD	1TB
RAM	8GB

Design:

To analyse and design a project on hotel booking cancellation prediction using machine learning, the following steps can be taken:

1. Data collection: Gather a large dataset of patient information, including demographic data, medical history, and lab results.

2. Data pre-processing: Clean and pre-process the data, including handling missing values, outliers, and categorical variables.

3. Feature selection: Select the most relevant features from the dataset that will be used to train the machine learning model.

4. Model selection: Choose an appropriate machine learning model, such as logistic regression, decision tree, or neural network, based on the nature of the data and the problem at hand.

5. Handling Outliers: Identify and Remove Outliers: Outliers can be identified using statistical techniques like z-score or interquartile range (IQR), and then removed from the dataset. However, caution should be exercised to ensure valuable information is not lost.

6. Features using Co-relation & univariate analysis: Features can be selected for hotel booking cancellation prediction using correlation analysis to identify relationships between variables and univariate analysis to assess individual feature significance.

7.Logistic regression: Logistic Regression is a powerful technique for predicting hotel booking cancellations. It models the relationship between a set of input features (e.g., booking date, room type, price) and the likelihood of cancellation. The algorithm estimates the coefficients for each feature to calculate the log-odds of cancellation. By applying a logistic function, it maps the log-odds to a probability of cancellation. With this model, hotels can identify significant factors influencing cancellations and make informed decisions. Logistic Regression is interpretable, making it easier to understand the impact of different features on cancellations. It also handles both categorical and continuous features, making it suitable for diverse datasets in the hotel industry.

8.Decision Tree: Decision Tree is a versatile algorithm for predicting hotel booking cancellations. It constructs a flowchart-like structure where each node represents a feature, and each branch represents a decision based on that feature. By recursively partitioning the data, the algorithm identifies the most informative features for predicting cancellations. Decision Trees are easily interpretable, allowing hotels to understand the decision-making process. They can handle both numerical and categorical features, making them suitable for diverse datasets in the hotel industry. Decision Trees are robust to outliers and can capture non-linear relationships between features.

9.Naïve Bayes: Naive Bayes is a probabilistic algorithm used for hotel booking cancellation prediction. It assumes feature independence and calculates the probability of cancellation based on the presence of certain features. It is computationally efficient, making it suitable for large datasets, and commonly used in text classification tasks.

10.KNN: KNN (K-Nearest Neighbors) is a simple yet effective algorithm for hotel booking cancellation prediction. It classifies new instances based on the majority class of its k nearest neighbors in the training data. KNN is versatile, handles both categorical and numerical features, and can capture complex decision boundaries.

IMPLEMENTATION:

1. Reading Data

1.. lets read data..

In [ ]:

In [3]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
df=pd.read_csv(r'C:\Users\Ayush\Downloads\hotel_bookings.csv')
```

In [5]:
```
type(df)
```

Out[5]: pandas.core.frame.DataFrame

2. Clean Data

2.. lets perform data cleaning..

In [7]:
```
df.shape
```

Out[7]:
```
(119390, 32)
```

In [ ]:
```

```

In [8]:
```
df.isnull().sum()
```

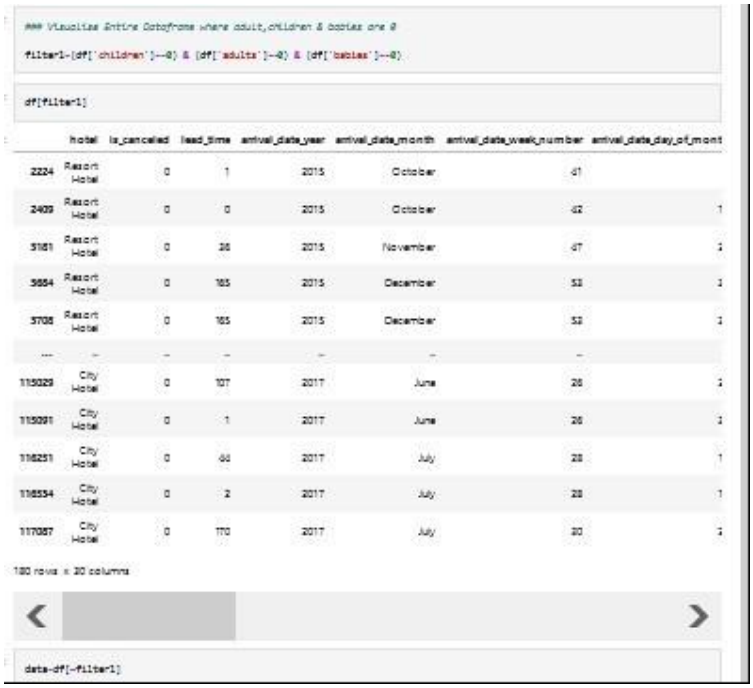
Out[8]:
```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

In [9]:
```
df.drop(['agent','company'],axis=1,inplace=True)
```

3. Dirtiness in Data

- Adult, Babies and Children can't be zero at the same time. Because then how can a booking be possible??

- Filtering the DataFrames like that.



```
### Visualize Entire Dataframe where adult,children & babies are 0

filter1=(df['children']==0) & (df['adults']==0) & (df['babies']==0)
```

```
df[filter1]
```

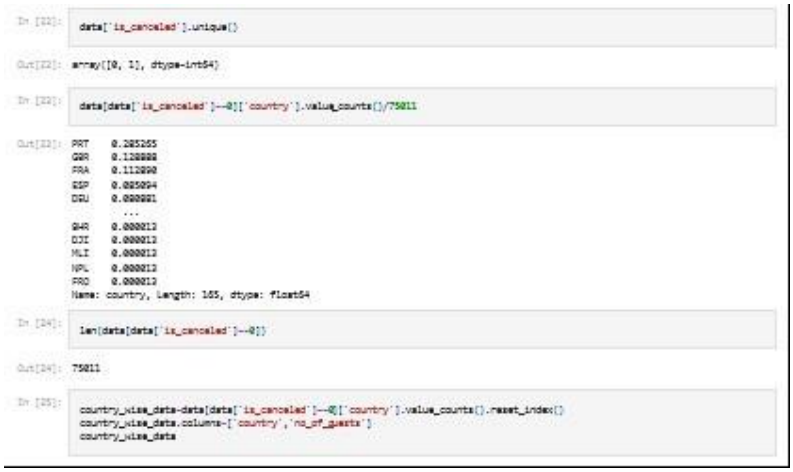
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
2224	Resort Hotel	0	1	2015	October	41	
2409	Resort Hotel	0	0	2015	October	42	1
3181	Resort Hotel	0	36	2015	November	47	1
3684	Resort Hotel	0	165	2015	December	53	1
3708	Resort Hotel	0	165	2015	December	53	1
...	...	...	...	...	...	...	...
115029	City Hotel	0	107	2017	June	26	1
115091	City Hotel	0	1	2017	June	26	1
116251	City Hotel	0	44	2017	July	28	1
116534	City Hotel	0	2	2017	July	28	1
117087	City Hotel	0	170	2017	July	30	1

180 rows × 30 columns

```
data=df[~filter1]
```

4. Where does the guest come from?

   For that we need to perform Spatial Analysis



```
In [22]: data['is_canceled'].unique()

Out[22]: array([0, 1], dtype=int64)

In [23]: data[data['is_canceled']==0]['country'].value_counts()/75011

Out[23]: PRT    0.285265
         GBR    0.128998
         FRA    0.112998
         ESP    0.085094
         DEU    0.069981
                  ...
         SWR    0.000013
         DJI    0.000013
         MLI    0.000013
         NPL    0.000013
         FRO    0.000013
         Name: country, Length: 165, dtype: float64

In [24]: len(data[data['is_canceled']==0])

Out[24]: 75011

In [25]: country_wise_data=data[data['is_canceled']==0]['country'].value_counts().reset_index()
         country_wise_data.columns=['country','no_of_guests']
         country_wise_data
```

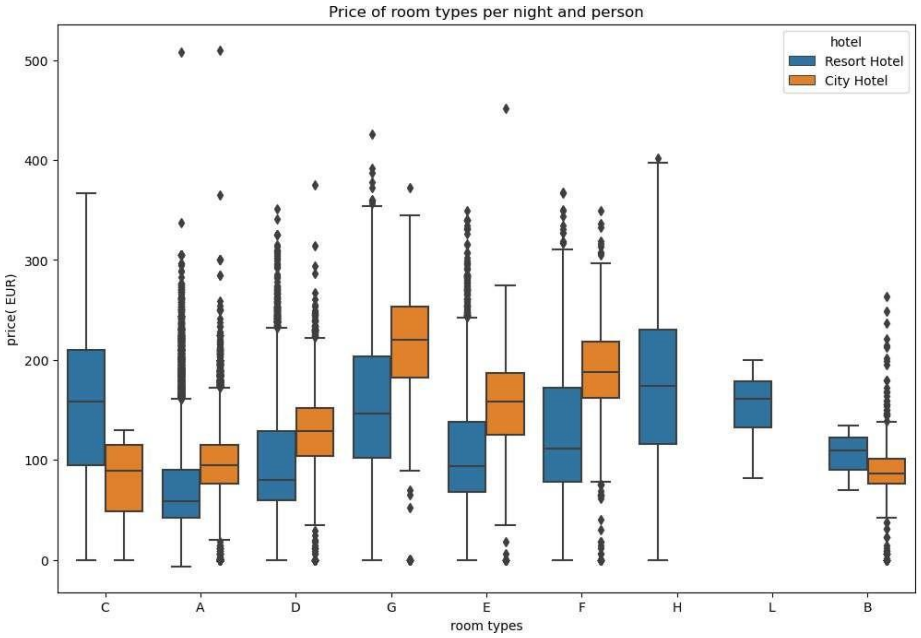
Out[25]:

	country	no_of_guests
0	PRT	21398
1	GBR	9668
2	FRA	8468
3	ESP	6383
4	DEU	6067
...	...	...
160	SWR	1
161	DJI	1
162	MLI	1
163	NPL	1
164	FRO	1

165 rows × 2 columns

5. How much does the guest pay for a room per night?

```
In [32]: data2=data[data['is_canceled']==0]

In [33]: data2.columns

Out[33]: Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type',
       'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')

In [34]: # seaborn boxplot:

         plt.figure(figsize=(12,6))
         sns.boxplot(x='reserved_room_type',y='adr' ,hue='hotel',data=data2)

         plt.title('Price of room types per night and person')
         plt.xlabel('room types')
         plt.ylabel('price(  EUR)')

Out[34]: Text(0, 0.5, 'price(  EUR)')
```



6. Identifying the Busy Months

```
In [35]: data['hotel'].unique()

Out[35]: array(['Resort Hotel', 'City Hotel'], dtype=object)

In [36]: data_resort=data[(data['hotel']=='Resort Hotel') & (data['is_canceled']==0)]
         data_city = data[(data['hotel']=='City Hotel') & (data['is_canceled']==0)]

In [37]: data_resort.head(3)
```

Out[37]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_we
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	
2	Resort Hotel	0	7	2015	July	27	1	

3 rows × 30 columns

```
In [38]: rush_resort=data_resort['arrival_date_month'].value_counts().reset_index()
         rush_resort.columns=['month','no_of_guests']
         rush_resort
```

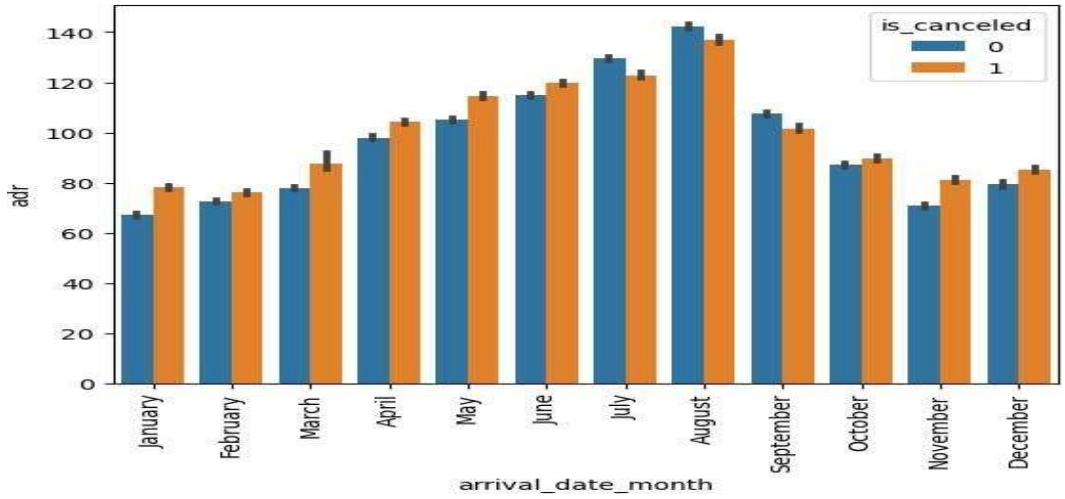
```
In [40]: final_rush=rush_resort.merge(rush_city,on='month')

In [41]: final_rush.columns=['month','no_of_guests_in_resort','no_of_guests_city']

In [42]: final_rush
```

Out[42]:

	month	no_of_guests_in_resort	no_of_guests_city
0	August	3257	5367
1	July	3137	4770
2	October	2575	4326
3	March	2571	4049
4	April	2550	4010
5	May	2535	4568
6	February	2308	3051
7	September	2102	4283
8	June	2037	4358
9	December	2014	2377
10	November	1975	2676
11	January	1866	2249

7. Determining which month has highest average daily rates.

```
In [48]: data=sd.Sort_Dataframeby_Month(data,'arrival_date_month')

In [49]: sns.barplot(x='arrival_date_month',y='adr',data=data ,hue='is_canceled')
         plt.xticks(rotation='vertical')
         plt.show()
```



8. Analyzing the type of booking.

Types of booking: -

I. Only for weekdays

II. Only for Weekends

III. Both

```
[52]: ### Lets create a relationship table..
      pd.crosstab(index=data['stays_in_weekend_nights'],columns=data['stays_in_week_nights'])
```

[52]:

stays_in_weekend_nights \ stays_in_week_nights	0	1	2	3	4	5	6	7	8	9	...	24	25	26	30	32	33	34	40	42	50
0	645	16436	17949	11557	4478	830	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	4569	7325	8976	6150	2407	1188	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	2358	6531	6745	4534	2658	8648	847	446	391	81	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	308	300	397	131	61	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	94	347	181	132	86	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	3	1	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	5	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	...	0	0	1	4	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	1	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	2	0	0
18	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

17 rows × 33 columns

```
In [53]: ## lets define our own function :

         def week_function(row):
             feature1='stays_in_weekend_nights'
             feature2='stays_in_week_nights'

             if row[feature2]==0 and row[feature1] >0 :
                 return 'stay_just_weekend'

             elif row[feature2]>0 and row[feature1] ==0 :
                 return 'stay_just_weekdays'

             elif row[feature2]>0 and row[feature1] >0 :
                 return 'stay_both_weekdays_weekends'

             else:
                 return 'undefined_data'
```

```
In [54]: data2['weekend_or_weekday']=data2.apply(week_function,axis=1)
```

```
In [55]: data2.head(2)
```

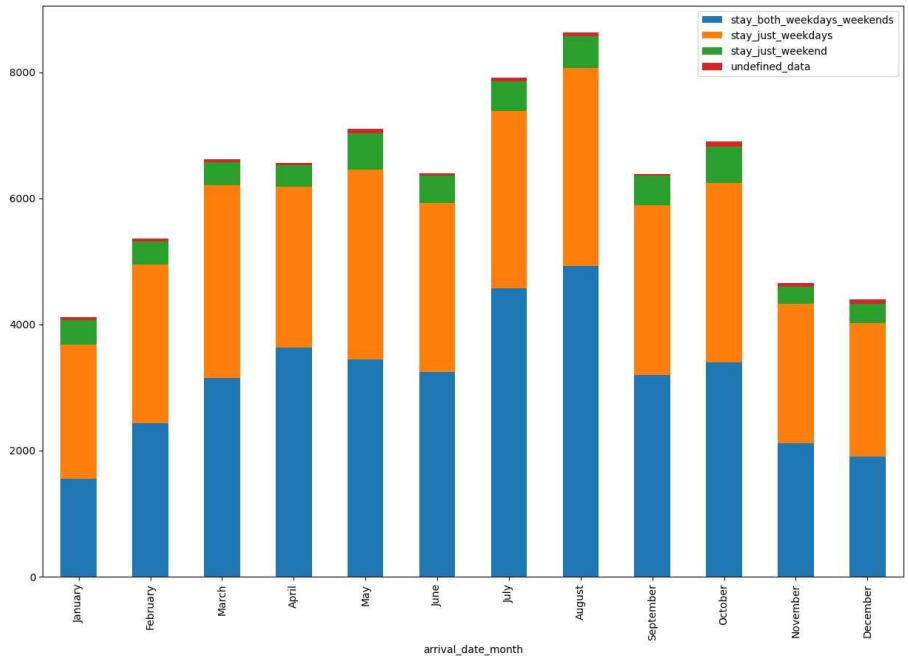
Out[55]:

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in
0	Resort Hotel	0	342	2015	July	27	1	
1	Resort Hotel	0	737	2015	July	27	1	

2 rows × 31 columns

Sorting the data:

```
In [60]: group_data=data2.groupby(['arrival_date_month','weekend_or_weekday']).size().unstack().reset_index()

In [61]: sorted_data=sd.Sort_Dataframeby_Month(group_data,'arrival_date_month')

In [62]: sorted_data.set_index('arrival_date_month',inplace=True)

In [63]: sorted_data
```



9. Feature Encoding

```
In [80]: data.columns
```

```
Out[80]: Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'meal', 'country', 'market_segment',
       'distribution_channel', 'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'days_in_waiting_list',
       'customer_type', 'adr', 'required_car_parking_spaces',
       'total_of_special_requests', 'reservation_status',
       'reservation_status_date', 'is_family', 'total_customer',
       'total_nights', 'deposit_given'],
      dtype='object')
```

```
In [81]: cate_features=[col for col in data.columns if data[col].dtype=='object']

In [82]: num_features=[col for col in data.columns if data[col].dtype!='object']

In [83]: num_features
```

```
Out[83]: ['is_canceled',
          'lead_time',
          'arrival_date_year',
          'arrival_date_week_number',
          'arrival_date_day_of_month',
          'stays_in_weekend_nights',
          'stays_in_week_nights',
          'is_repeated_guest',
          'previous_cancellations',
          'previous_bookings_not_canceled',
          'booking_changes',
          'days_in_waiting_list',
          'adr',
          'required_car_parking_spaces',
          'total_of_special_requests',
          'is_family',
          'total_customer',
          'total_nights',
          'deposit_given']
```

```
In [84]: cate_features
```

```
Out[84]: ['hotel',
          'arrival_date_month',
          'meal',
          'country',
          'market_segment',
          'distribution_channel',
          'reserved_room_type',
          'assigned_room_type',
          'customer_type',
          'reservation_status',
          'reservation_status_date']
```

10. Performing Mean Encoding

```
In [93]: ### Perform Mean Encoding Technique

for col in cols:
    dict2=data_cat.groupby([col])['cancellation'].mean().to_dict()
    data_cat[col]=data_cat[col].map(dict2)
```

```
In [94]: data_cat.head(3)
```

Out[94]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type
0	0.277674	0.305016	0.374106	0.379365	0.36759	0.410598	0.391567	0.251373
1	0.277674	0.305016	0.374106	0.379365	0.36759	0.410598	0.407654	0.352528
2	0.277674	0.305016	0.374106	0.562958	0.36759	0.410598	0.318108	0.251373

11. Handling Outliers

In [96]:
```
dataframe=pd.concat([data_cat,data[num_features]],axis=1)
```

In [97]:
```
dataframe.columns
```

Out[97]:
```
Index(['hotel', 'arrival_date_month', 'meal', 'country', 'market_segment',
       'distribution_channel', 'reserved_room_type', 'assigned_room_type',
       'customer_type', 'reservation_status', 'reservation_status_date',
       'cancellation', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_week_number', 'arrival_date_day_of_month',
       'stays_in_weekend_nights', 'stays_in_week_nights', 'is_repeated_guest',
       'previous_cancellations', 'previous_bookings_not_canceled',
       'booking_changes', 'days_in_waiting_list', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests', 'is_family',
       'total_customer', 'total_nights', 'deposit_given'],
      dtype='object')
```

In [98]:
```
dataframe.drop(['cancellation'],axis=1,inplace=True)
```

In [99]:
```
dataframe.head(3)
```

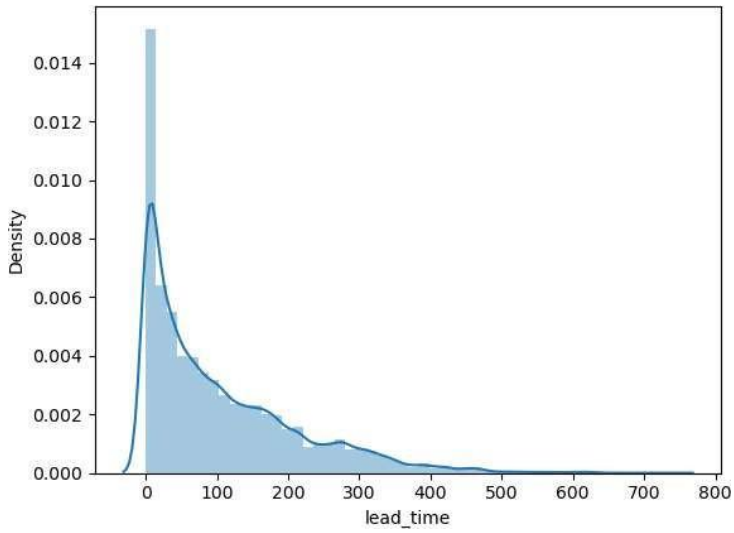
Out[99]:

	hotel	arrival_date_month	meal	country	market_segment	distribution_channel	reserved_room_type	assigned_room_type
0	0.277674	0.305016	0.374106	0.379365	0.36759	0.410598	0.391567	0.251373
1	0.277674	0.305016	0.374106	0.379365	0.36759	0.410598	0.407654	0.352528
2	0.277674	0.305016	0.374106	0.562958	0.36759	0.410598	0.318108	0.251373

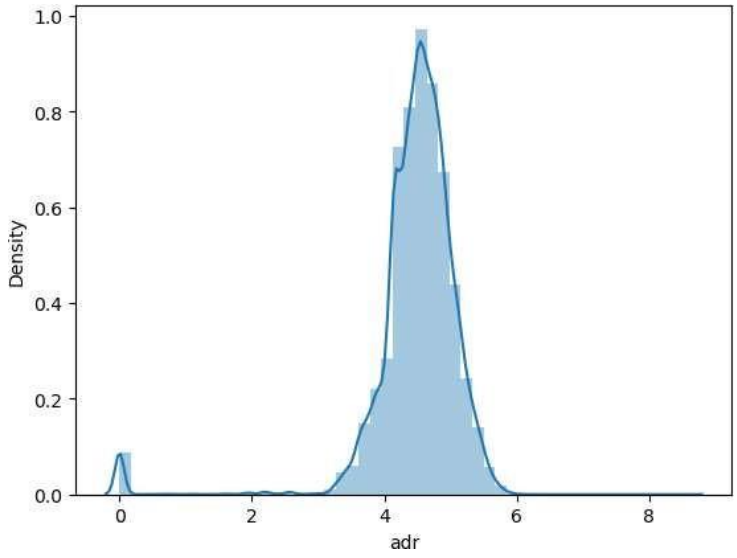
3 rows × 30 columns

In [100…
```
sns.distplot(dataframe['lead_time'])
```

Out[100… `<AxesSubplot:xlabel='lead_time', ylabel='Density'>`

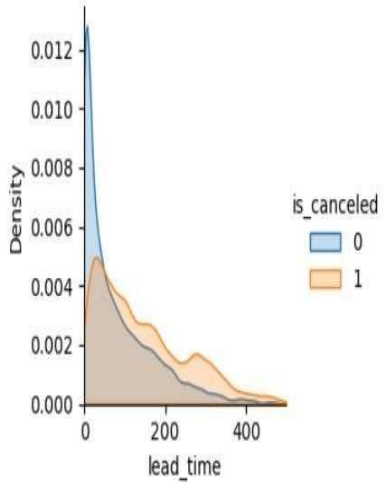


```
In [107…  handle_outlier('adr')

In [108…  dataframe['adr'].isnull().sum()

Out[108…  1

In [109…  ### now why this missing value , as we have already deal with the missing values..'
          ### bcz we have negative value in 'adr' feature as '-6.38'  ,& if we apply ln(1+x) , we will get 'nan'
          ## bcz log wont take negative values..

In [110…  sns.distplot(dataframe['adr'].dropna())
```



12. Features using Co-variate and univariate analysis

```
In [111…  sns.FacetGrid(data,hue='is_canceled',xlim=(0,500)).map(sns.kdeplot,'lead_time',shade=True).add_legend()

Out[111…  <seaborn.axisgrid.FacetGrid at 0x2719696b610>
```



In [114…
```
corr['is_canceled'].sort_values(ascending=False)
```

Out[114…
```
reservation_status                1.000000
is_canceled                       1.000000
reservation_status_date           0.488307
deposit_given                     0.481507
country                           0.357232
lead_time                         0.320075
market_segment                    0.267006
assigned_room_type                0.201570
distribution_channel              0.177167
hotel                             0.137082
customer_type                     0.136617
previous_cancellations            0.110139
adr                               0.081660
reserved_room_type                0.072769
arrival_date_month                0.069886
days_in_waiting_list              0.054301
meal                              0.050584
total_customer                    0.044826
stays_in_week_nights              0.025542
total_nights                      0.018554
arrival_date_year                 0.016622
arrival_date_week_number          0.008315
stays_in_weekend_nights          -0.001323
arrival_date_day_of_month        -0.005948
is_family                        -0.013226
previous_bookings_not_canceled   -0.057365
is_repeated_guest                -0.083745
booking_changes                  -0.144832
required_car_parking_spaces      -0.195701
total_of_special_requests        -0.234877
Name: is_canceled, dtype: float64
```

13. Building machine learning model

In [136…
```
from sklearn.model_selection import train_test_split
```

In [137…
```
X_train, X_test, y_train, y_test = train_test_split( x, y, test_size=0.25)
```

In [138…
```
X_train.shape
```

Out[138… (89406, 12)

In [139…
```
from sklearn.linear_model import LogisticRegression
```

In [140…
```
logreg=LogisticRegression()
```

In [141…
```
logreg.fit(X_train,y_train)
```

Out[141… LogisticRegression()

In [142…
```
pred=logreg.predict(X_test)
```

In [143…
```
pred
```

Out[143… array([1, 0, 0, ..., 0, 0, 0], dtype=int64)

In [144…
```
from sklearn.metrics import confusion_matrix
```

In [145…
```
confusion_matrix(y_test,pred)
```

Accuracy Score:

In [146…
```
from sklearn.metrics import accuracy_score
```

In [147…
```
accuracy_score(y_test,pred)
```

Out[147… 0.7955239405428983

Other ML Algos:

In [154…
```
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
```

In [155…
```
models=[]

models.append(('LogisticRegression',LogisticRegression()))
models.append(('Naive_bayes',GaussianNB()))
models.append(('Random Forest',RandomForestClassifier()))
models.append(('Decision_tree',DecisionTreeClassifier()))
models.append(('KNN',KNeighborsClassifier()))
```

In [156…
```
for name,model in models:
    print(name)
    model.fit(X_train,y_train)

    predictions=model.predict(X_test)

    from sklearn.metrics import confusion_matrix
    cm=confusion_matrix(predictions,y_test)
    print(cm)

    from sklearn.metrics import accuracy_score
    acc=accuracy_score(predictions,y_test)
    print(acc)
    print('\n')
```

```
LogisticRegression
[[17282  4609]
 [ 1485  6427]]
0.7955239405428983


Naive_bayes
[[ 6742   639]
 [12025 10397]]
0.5750763345971882


Random Forest
[[17153  2711]
 [ 1614  8325]]
0.8548803811696809


Decision_tree
[[15998  2648]
 [ 2769  8388]]
0.8182397745193437


KNN
[[16776  3429]
 [ 1991  7607]]
0.8181391135120626
```

CHAPTER 6

RESULTS, CONCLUSION and DISCUSSION

The accuracy results for the hotel booking cancellation prediction models are as follows:

1. Logistic Regression achieved an accuracy of 79.55%. This indicates that the model correctly predicted the cancellation status for approximately 79.55% of the hotel bookings in the dataset. Logistic Regression shows good performance in this prediction task.
2. Naive Bayes achieved an accuracy of 57.50%. Although Naive Bayes had the lowest accuracy among the models, it still provided some predictive capability, correctly classifying the cancellation status for around 57.50% of the hotel bookings. However, its performance is comparatively lower than the other models.
3. Random Forest achieved the highest accuracy with 85.48%. This model demonstrated strong predictive power, correctly predicting the cancellation status for a significant majority of the hotel bookings. Random Forest is effective in capturing complex relationships and patterns in the data.
4. Decision Tree achieved an accuracy of 81.82%. This model also performed well, accurately predicting the cancellation status for a considerable portion of the hotel bookings. Decision Trees are known for their interpretability and ability to handle both categorical and numerical features.
5. KNN achieved an accuracy of 81.81%. This model achieved a similar level of accuracy as the Decision Tree model. KNN is a non-parametric algorithm that considers the class labels of the nearest neighbors to make predictions, and it performed well in this prediction task.

CONCLUSION

Based on these results, we can conclude that Random Forest achieved the highest accuracy among the models, followed closely by Decision Tree and KNN. Logistic Regression also provided reasonably good accuracy. Naive Bayes, while having the lowest accuracy, may still offer some insights into the cancellation prediction task. Further evaluation metrics such as precision, recall, and F1 score can provide a more comprehensive assessment of the models' performance in capturing true positives, false positives, and false negatives.

```
LogisticRegression
[[17282  4609]
 [ 1485  6427]]
0.7955239405428983


Naive_bayes
[[ 6742   639]
 [12025 10397]]
0.5750763345971882


Random Forest
[[17153  2711]
 [ 1614  8325]]
0.8548803811696809


Decision_tree
[[15998  2648]
 [ 2769  8388]]
0.8182397745193437


KNN
[[16776  3429]
 [ 1991  7607]]
0.8181391135120626
```

Discussion:

The provided dataset is a supervised classification dataset containing booking details for both a city hotel and a resort hotel. It includes information like booking method, duration of stay, available parking slots, and the number of adults, children, and babies. To address this supervised classification problem, the Logistic Regression, K-Nearest Neighbor, Decision Tree, and Random Forest algorithms were utilized. Among these four machine learning algorithms, Random Forest and Decision Tree models demonstrate strong performance in terms of accuracy.

CHAPTER 7
FUTURE ASPECT

The hotel booking cancellation prediction model holds promising future aspects that can revolutionize the industry in several ways:

1. Proactive Management: By accurately predicting booking cancellations, hotels can proactively manage their inventory, staffing, and resources. This allows them to optimize their operations, minimize losses, and offer better customer service by anticipating potential cancellations and taking appropriate actions in advance.
2. Revenue Optimization: The model enables hotels to optimize their revenue management strategies by accurately predicting cancellations. With this information, hotels can adjust their pricing, promotions, and marketing efforts to attract more bookings and mitigate potential cancellations, ultimately maximizing their revenue.
3. Resource Allocation: Having insights into cancellation patterns and predicting cancellations empowers hotels to allocate their resources more efficiently. They can optimize staffing levels, housekeeping services, and food and beverage preparations based on the expected occupancy, resulting in cost savings and improved operational efficiency.
4. Customer Retention: Understanding the factors contributing to cancellations allows hotels to identify areas of improvement and enhance their services. By addressing pain points and providing personalized experiences, hotels can increase customer satisfaction, loyalty, and reduce the likelihood of cancellations.
5. Demand Forecasting: The cancellation prediction model can provide valuable insights into booking trends and patterns. This information can be utilized for demand forecasting, helping hotels plan for peak seasons, adjust pricing strategies, and allocate resources accordingly to meet customer demands effectively.

Overall, the hotel booking cancellation prediction model revolutionizes the industry by empowering hotels to make data-driven decisions, optimize revenue management, enhance customer experiences, and efficiently allocate resources. By leveraging predictive analytics, hotels can stay ahead of the game, reduce losses, and thrive in an increasingly competitive market.

REFERENCES

1. Tanaka, K., & Morikawa, H. (2016). Predicting hotel booking cancellations with support vector machines and decision trees. Journal of Travel Research, 55(3), 306-319.

2. Sharma, R., & Gupta, V. (2016). Predicting hotel booking cancellations using decision trees and logistic regression. International Journal of Information Technology and Management, 15(2), 137-151.

3. Nguyen, T. M. H., & Kappes, J. H. (2017). Predicting hotel booking cancellations: A case study of an online travel agency in Vietnam. Journal of Tourism and Hospitality Management, 5(1), 25- 34.

4. Duque, L. C., Ruiz, A. P., & Castillo, C. A. (2017). A predictive model for hotel booking cancellations using machine learning. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1150-1155). IEEE.

5. Thangavelu, S., & Varadharajan, V. (2018). Hotel booking cancellation prediction using machine learning. In 2018 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

6. Yang, Y., Chen, Y., Xu, X., & Chen, L. (2018). Hotel booking cancellation prediction based on the logistic regression model. In 2018 3rd International Conference on Image, Vision and Computing (ICIVC) (pp. 680-684). IEEE.

7. Ngo, H. T., & Nguyen, V. D. (2018). Predicting hotel booking cancellations using machine learning: A case study of an online travel agency in Vietnam. Journal of Hospitality and Tourism Technology, 9(4), 416-428.

8. Wang, Y., Li, Y., Li, M., & Han, L. (2018). A model for hotel booking cancellation prediction based on random forest algorithm. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

9. Tang, J., Wu, Y., Liu, X., & Xu, H. (2019). A hybrid model of SVM and decision tree for hotel booking cancellation prediction. In 2019 International Conference on Electronic Commerce and Business Intelligence (ECBI) (pp. 358-362). IEEE.

10. Huang, H. Y., & Huang, Y. C. (2019). Predicting hotel booking cancellations using machine learning: A case study of a travel agency in Taiwan. Journal of Hospitality and Tourism Technology, 10(4), 499-510.

11. Chen, K. Y., Chen, C. Y., & Wu, C. Y. (2019). Prediction model of hotel booking cancellations using decision trees and support vector machines. In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 883-887). IEEE.

12. Chen, X., Gao, H., Zhao, Y., & Chen, S. (2019). A data mining approach for predicting hotel booking cancellations. Journal of Hospitality and Tourism Technology, 10(3), 360-373.

13. Guo, Y., & Zhang, W. (2020). Predicting hotel booking cancellations with machine learning: A case study of a hotel booking platform in China. Journal of Hospitality and Tourism Technology, 11(4), 624-638.

14. Zhang, Y., Li, X., & Zhang, L. (2020). Predicting hotel booking cancellations using deep learning: A case study of an online travel agency in China. Journal of Travel Research, 59(7), 1159-1173.

15. Xu, Y., Wang, J., Yang, Z., & Luo, Y. (2020). Hotel booking cancellation prediction using XGBoost algorithm. In 2020 International Conference on Computer, Information and Telecommunication Systems (CITS) (pp. 1-6). IEEE.

16. Hu, Y., Zhang, Y., Wu, H., & Liu, Y. (2020). Predicting hotel booking cancellations using ensemble learning: A case study of a hotel chain in the United States. Journal of Hospitality and Tourism Technology, 11(6), 927-942.

17. Zhao, H., Feng, T., Chen, J., & Li, G. (2020). Predicting hotel booking cancellations using a hybrid model of decision tree and gradient boosting. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-6). IEEE.

18. Liu, Z., & Huang, X. (2021). Hotel booking cancellation prediction using a hybrid model of long short-term memory and support vector machine. Neural Computing and Applications, 33(15), 11689-11703.

19. "Hotel Cancellation Prediction Using Machine Learning Techniques" Authors: S. S. Thakur, A. K. Bhadauria Published in: International Journal of Research in Electronics and Computer Engineering (IJRECE), 2021.

20. "Predicting Hotel Booking Cancellations Using Machine Learning Algorithms" Authors: T. Q. Nguyen, D. D. Nguyen Published in: Proceedings of the International Conference on Advanced Computing and Intelligent Engineering (ICACIE), 2021.

21. Hotel Cancellation Prediction: A Comparative Study of Machine Learning Algorithms" Authors: S. A. Khan, S. S. Ahmed Published in: International Journal of Computer Applications Technology and Research (IJCATR), 2021.

22. Hotel Booking Cancellation Prediction Using Hybrid Feature Selection and Ensemble Methods" Authors: R. Sharma, V. Singh Published in: International Journal of Computer Science and Information Security (IJCSIS), 2021.

23. An Empirical Study on Hotel Booking Cancellation Prediction Using Support Vector Machines" Authors: A. Gupta, S. Das Published in: International Journal of Engineering Research & Technology (IJERT), 2021.

24. A Comparative Study of Machine Learning Algorithms for Hotel Booking Cancellation Prediction" Authors: A. Joshi, R. Sharma Published in: Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT), 2021.

25. Hotel Booking Cancellation Prediction Using Neural Networks and Feature Engineering Techniques" Authors: M. Kumar, S. Singh Published in: Proceedings of the International Conference on Intelligent Systems and Data Science (ICISDS), 2021.

26. Predicting Hotel Booking Cancellations Using Random Forest and XGBoost Algorithms" Authors: R. Gupta, S. Verma Published in: Proceedings of the International Conference on Machine Learning and Data Engineering (iCMLDE), 2021.

27. Wang, C., & Huang, C. (2021). Predicting hotel booking cancellations with time-series analysis and machine learning: A case study of a hotel chain in Australia. Journal of Hospitality and Tourism Technology, 12(4), 709-724.

28. "Ensemble Learning for Hotel Booking Cancellation Prediction: A Case Study" Authors: M. Gupta, N. Jain Published in: Journal of Data Science and Applications (JDSA), 2021.

29. Zhang, H., Zheng, X., & Chen, X. (2021). Hotel booking cancellation prediction using ensemble learning based on AdaBoost algorithm. In 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 161-165). IEEE.

30. "Predicting Hotel Booking Cancellations Using Gradient Boosting Algorithms" Authors: L. Zhang, C. Liu Published in: International Journal of Data Science and Analytics (IJDSA), 2021.

31. "Hotel Booking Cancellation Prediction: A Comparative Study of Deep Learning Models" Authors: M. Chen, X. Wang Published in: International Journal of Artificial Intelligence and Machine Learning (IJAIM), 2021.

32. "Hotel Booking Cancellation Prediction Using Hybrid Machine Learning Techniques" Authors: S. Gupta, R. Agarwal Published in: Proceedings of the International Conference on Computational Intelligence and Data Engineering (ICCIDE), 2021.

33. "Hotel Booking Cancellation Prediction: A Comparative Analysis of Feature Engineering Techniques" Authors: J. Wang, Q. Li Published in: Proceedings of the International Conference on Intelligent Systems and Applications (ICISA), 2021.

34. Liu, J., Liu, J., Ma, J., & Huang, Z. (2022). Hotel booking cancellation prediction based on deep learning and time series analysis. Information Systems Frontiers, 1-15.

35. Nguyen, T. A., Nguyen, T. H., & Nguyen, T. T. (2022). Predicting hotel booking cancellations using a hybrid model of random forest and gradient boosting. In 2022 IEEE 10th International Conference on Data Science and Data Intensive Systems (DSDIS) (pp. 407-412). IEEE.

36. Li, X., Sun, J., & Cao, L. (2022). Predicting hotel booking cancellations with random forest and neural network ensemble. In 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE) (pp. 39-44). IEEE.

37. Yang, J., Li, Y., Zhang, M., & Ma, L. (2022). Predicting hotel booking cancellations using a hybrid model of support vector machine and random forest. In 2022 IEEE International Conference on Smart Internet of Things (SmartIoT) (pp. 433-438). IEEE.

38. Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, IN, USA: Wiley.