**CAPSTONE PROJECT FINAL REPORT**

On

**“Predictive Insights: Understanding Hotel Bookings Cancellation”**

**Post Graduate Program In Data Science & Engineering**

**Batch: PGP-DSE-FT-JAN-2024**

**Location: Bengaluru**

Submitted by

**Tushar Bhardwaj**

**Merlyn Pushparaj**

**Chayapathi Y**

**Aditi Tiwari**

**Aswin Raj K**

**Shikha Preta**

Mentored By

**Mr. Animesh Tiwari**

**Table of Contents**

Abstract 3

Introduction 3

Objectives & Scope 4

Data Description 4

Methodology 8

Exploratory Data Analysis 14

Visualization 15

Insights by feature 17

Model Building & Evaluation 28

Results & Discussions 31

**Abstract**

In the modern hospitality industry, accurately predicting hotel cancellations is essential for optimizing reservations and maximizing revenue. This study addresses the challenge of forecasting cancellations for online booking platforms such as MakeMyTrip and Agoda, which aggregate reservations from a diverse international clientele. Utilizing the Hotel Booking Demand dataset, which encompasses booking data from 177 countries, we aim to develop a robust model capable of identifying potential cancellations at the time of booking.

The methodology involves a multi-step approach, starting with extensive data preprocessing to clean and prepare the dataset. This includes handling missing values, converting data types, and engineering features to better capture booking patterns. Exploratory Data Analysis (EDA) is employed to examine trends and relationships within the data, revealing key factors associated with cancellations. Hypothesis testing is conducted to validate the impact of variables such as lead time, average daily rate (ADR), and stay duration on cancellation likelihood.

We implement and compare several predictive models, including logistic regression, decision trees, and ensemble methods, to forecast booking cancellations. Model performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Our analysis identifies lead time and ADR as significant predictors of cancellations, with BalancedRandomforest demonstrating superior performance in predictive accuracy and interpretability.

The results indicate that integrating this predictive model into booking platforms can significantly enhance their ability to anticipate cancellations, thereby improving operational efficiency and reducing revenue loss. This study provides actionable insights for booking platforms and hoteliers, highlighting the importance of advanced analytics in managing booking dynamics and customer behavior.

**1. Introduction**

**1.1 Background**  
The hospitality industry has undergone a significant transformation with the rise of online booking platforms such as MakeMyTrip, Agoda, and others. These platforms have simplified the reservation process, allowing travelers from around the world to book accommodations with ease. However, this convenience comes with challenges, particularly regarding booking cancellations. High cancellation rates can lead to lost revenue, inefficiencies in room allocation, and operational disruptions for hotels. Accurate cancellation predictions are crucial for mitigating these issues and optimizing booking strategies.

Given the global nature of these platforms, which aggregate reservations from 177 countries, the complexity of predicting cancellations increases. Each booking is influenced by a variety of factors, including lead time, average daily rate (ADR), and booking patterns specific to different regions and hotel types. Therefore, a robust predictive model that can analyze and forecast cancellations based on a comprehensive dataset is essential for improving booking accuracy and operational efficiency.

**1.2 Objectives**  
The primary objectives of this study are:

* **Data Preprocessing and Cleaning**: To prepare the Hotel Booking Demand dataset for analysis by handling missing values, converting data types, and creating new features.
* **Exploratory Data Analysis (EDA)**: To explore the dataset and identify trends, patterns, and relationships related to booking cancellations.
* **Hypothesis Testing**: To test various hypotheses regarding the factors influencing booking cancellations, such as the impact of ADR and lead time.
* **Predictive Modeling**: To develop and evaluate predictive models to forecast the likelihood of booking cancellations.
* **Insight Generation**: To provide actionable insights and recommendations for booking platforms and hotels based on the analysis results.

**1.3 Scope**  
This study focuses on the analysis of the Hotel Booking Demand dataset, which includes diverse booking information from 177 countries. The analysis encompasses various aspects such as data preprocessing, EDA, hypothesis testing, and predictive modeling. The findings are intended to assist booking platforms and hotels in improving their cancellation prediction capabilities and optimizing their booking processes.

**2. Dataset Description**

**2.1 Data Source**  
The dataset utilized in this study is the Hotel Booking Demand dataset, accessible on Kaggle. It offers extensive information about hotel bookings, including various attributes that are crucial for analyzing booking behaviors and predicting cancellations. The dataset can be found at [Kaggle - Hotel Booking Demand](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand).

**2.2 Context and Significance**  
Research in tourism and travel industries, especially concerning revenue management and demand forecasting, often relies on data from the aviation sector using the Passenger Name Record (PNR) format. However, the hospitality industry, which includes hotels, cruising, and theme parks, has unique characteristics that require industry-specific data for accurate predictions. This dataset addresses that gap by providing detailed booking information pertinent to the hospitality sector.

**2.3 Data Characteristics and Handling**  
The dataset contains 119,390 records and 32 columns, featuring a mix of numerical and categorical data. Key aspects of the dataset include:

* **Numerical Data**: There are 20 numerical features, providing quantitative measures related to bookings, such as lead\_time, adr, and stays\_in\_weekend\_nights.
* **Categorical Data**: There are 12 categorical features, including hotel, meal, country, and distribution\_channel, which describe qualitative attributes of the bookings.
* **Missing Values**: Missing values are present in the following features:
  + children
  + country
  + agent
  + company

To prevent data leakage and ensure accurate predictions, the dataset design avoids incorporating future information. The timestamp of the target variable (is\_canceled) is set to occur after the input variables' timestamps. Features were extracted from the booking change log relative to the day before the arrival date.

**2.4 Variables and Descriptions**  
Below is a detailed description of the variables in the dataset:

This dataset typically includes a variety of features that provide comprehensive insights into booking behaviors and hotel management. There are 119390 records and 32 columns in the dataset. It contains both numerical and categorical data.

Numerical data: There are 20 numerical features.

Categorical data: There are 12 categorical features.

Missing values: Missing values are observed in 4 features that is children, country, agent and company.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| hotel | Hotel (H1 = Resort Hotel or H2 = City Hotel) |
| is\_cancelled | Value indicating if the booking was cancelled (1) or not (0) |
| lead\_time | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| arrival\_date\_year | Year of arrival date |
| arrival\_date\_month | Month of arrival date |
| arrival\_date\_week\_number | Week number of year for arrival date |
| arrival\_date\_day\_of\_month | Day of arrival date |
| stays\_in\_weekend\_nights | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| stays\_in\_week\_nights | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| adults | Number of adults |
| children | Number of children |
| babies | Number of babies |
| meal | Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal |
| country | Country of origin. Categories are represented in the ISO 3155–3:2013 format |
| market\_segment | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour |
| distribution\_channel | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| is\_repeated\_guest | Value indicating if the booking name was from a repeated guest (1) or not (0) |
| previous\_cancellations | Number of previous bookings that were cancelled by the customer prior to the current booking |
| previous\_bookings\_not\_cancelled | Number of previous bookings not cancelled by the customer prior to the current booking |
| reserved\_room\_type | Code of room type reserved. Code is presented instead of designation for anonymity reasons. |
| assigned\_room\_type | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved |
| booking\_changes | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the |
| deposit\_type | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three |
| agent | ID of the travel agency that made the booking |
| company | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of |
| days\_in\_waiting\_list | Number of days the booking was in the waiting list before it was confirmed to the customer |
| customer\_type | Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated |
| adr | Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights |
| required\_car\_parking\_spaces | Number of car parking spaces required by the customer |
| total\_of\_special\_requests | Number of special requests made by the customer (e.g. twin bed or high floor) |
| reservation\_status | Reservation last status, assuming one of three categories: Cancelled – booking was cancelled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why  calendar today |
| reservation\_status\_date | Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking cancelled or when did the customer checked-out of the hotel |

This detailed description of the dataset provides a solid foundation for further analysis and model development.

**3.Methadology**

**3.1 Data Preprocessing:**

Data preprocessing involves several critical steps to prepare the dataset for analysis and model training. Below, we detail the preprocessing steps, including handling missing values, encoding categorical variables, and feature engineering, along with specific counts and details.

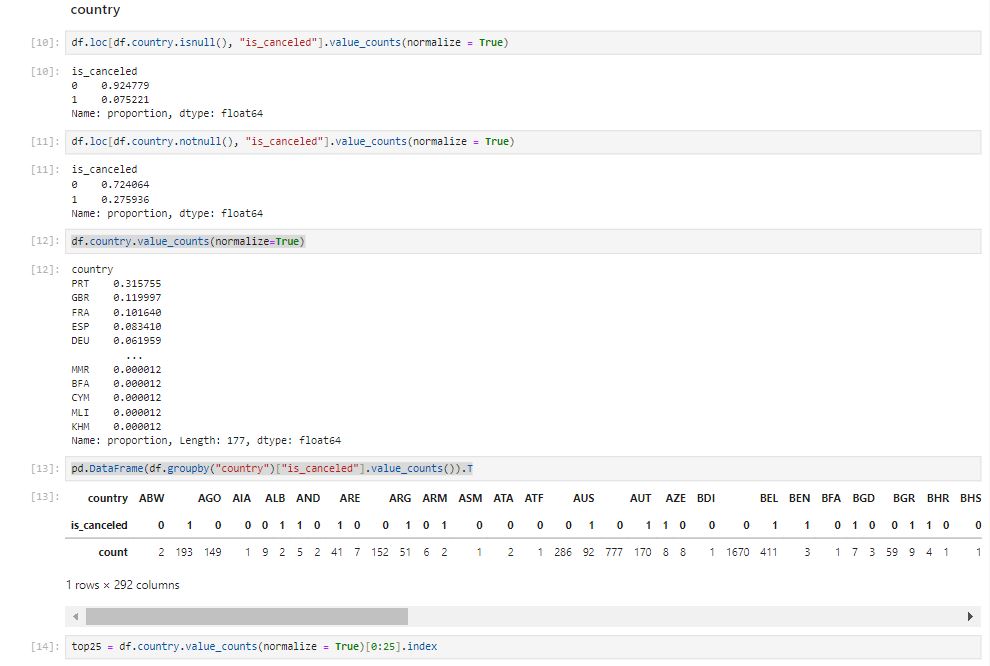
**3.1.1 Handling Missing Values**

1. **Identification of Missing Values**:

| **Column** | **Missing Values** | **Percentage of Total Records** |
| --- | --- | --- |
| children | 4 | 0.003% |
| country | 452 | 0.38% |
| agent | 12,193 | 10.2% |
| company | 82,137 | 68.8% |

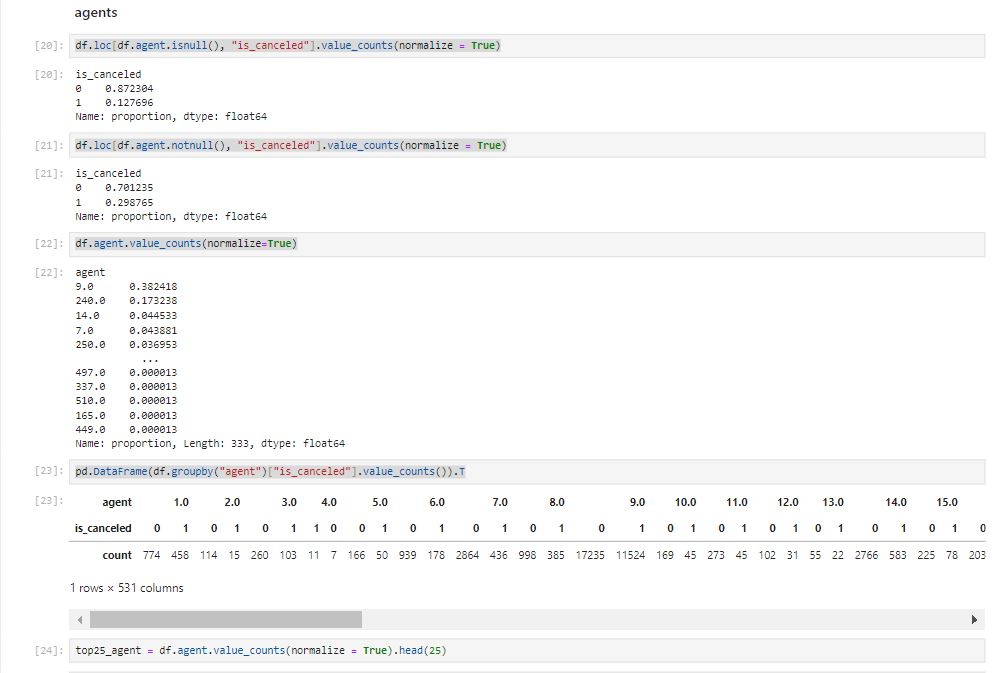
**3.1.2 Imputation Strategies:**

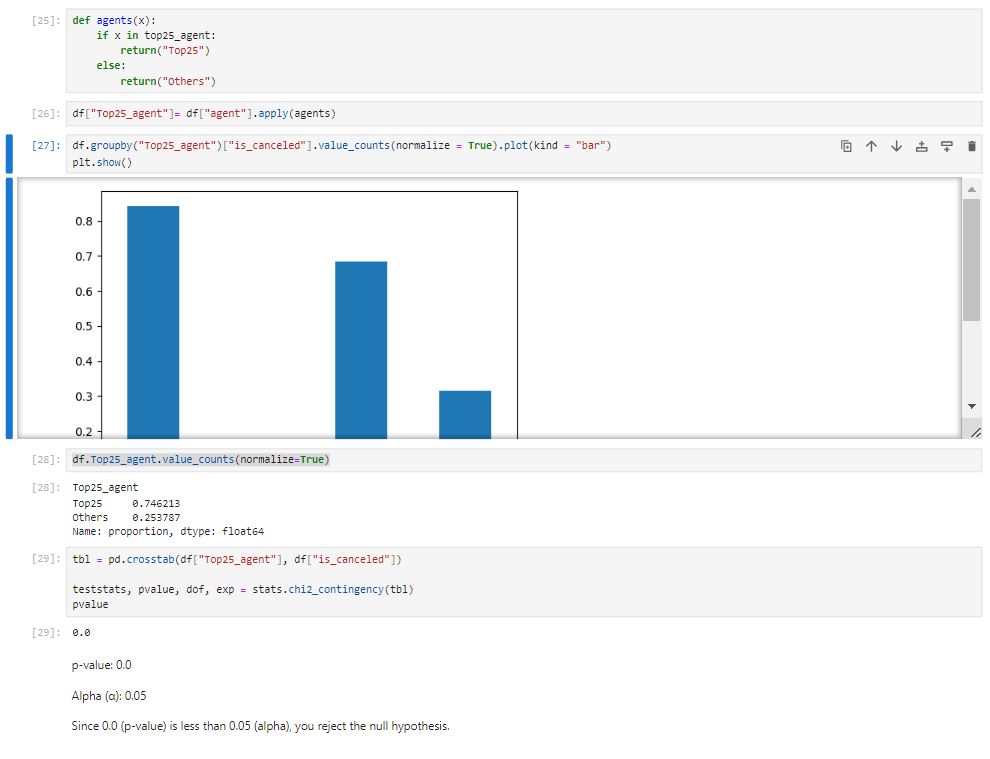
1. **country:**
   * Missing Values: *452* missing values.
   * **Imputation Strategy:**
     + **Top 25 Countries**: Analyzed the percentage proportion of each country. Countries with very small counts were grouped into an "Other" category to form a Top 25 column. This approach consolidates less frequent countries into a single category based on their distribution proportions, enhancing the robustness of the analysis.



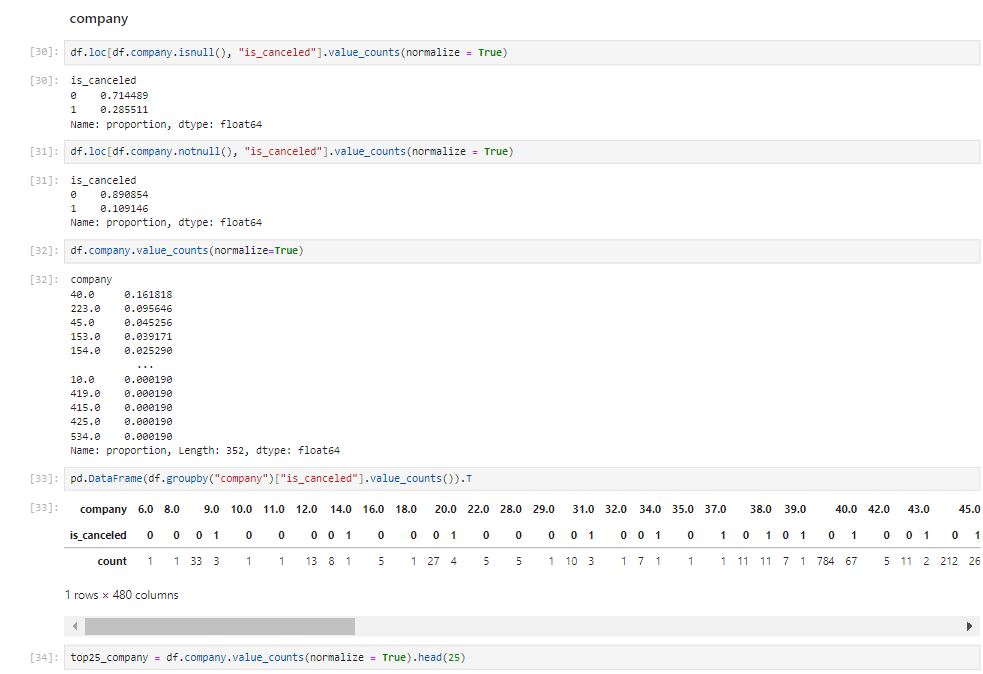


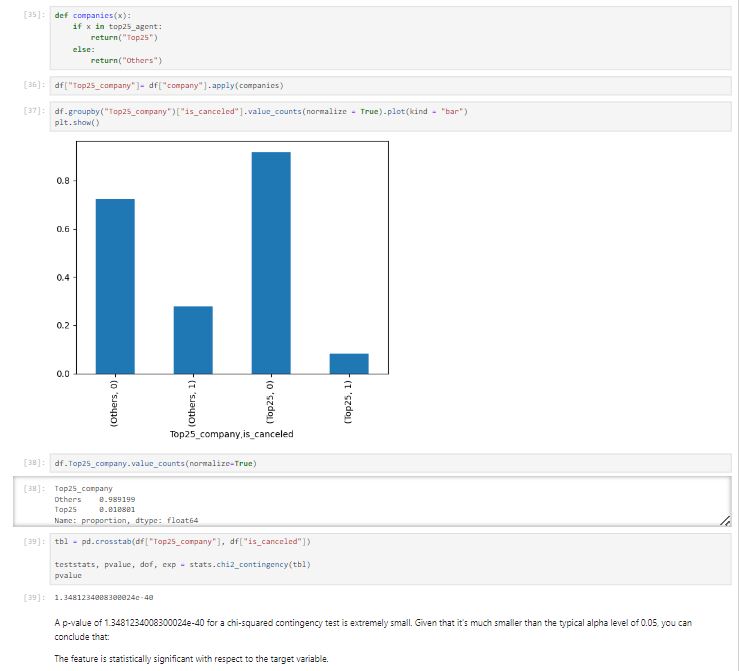
1. **agent:**
   * Missing Values: *12,193* missing values.
   * **Imputation Strategy**:
     + **Top 25 Agents**: Examined the percentage proportion of each agent. Agents with small counts were grouped into an "Other" category to create a Top 25 Agents column. This approach consolidates less frequent agents based on their proportional representation in the dataset, improving the analysis.





1. **company:**
   * Missing Values: *82,137* missing values.
   * ***Imputation Strategy:***
     + **Top 25 Companies**: Evaluated the percentage proportion of each company. Companies with small counts were grouped into an "Other" category, forming a Top 25 Companies column. This strategy consolidates less frequent companies based on their proportional representation, facilitating better data management and analysis.





**Chi-Square Test for Significance**

To validate the effectiveness of the new columns generated (e.g., Top 25 for country, agent, and company), a Chi-square test was conducted to assess their significance in relation to the target variable (is\_canceled). The results of the Chi-square test confirmed that all new columns are statistically significant.

| **New Column** | **p-value** |
| --- | --- |
| Top 25 Countries | < 0.001 |
| Top 25 Agents | < 0.001 |
| Top 25 Companies | < 0.001 |

**Table 1**: Results of the Chi-square test for the new columns.

**4. Exploratory Data Analysis (EDA)**

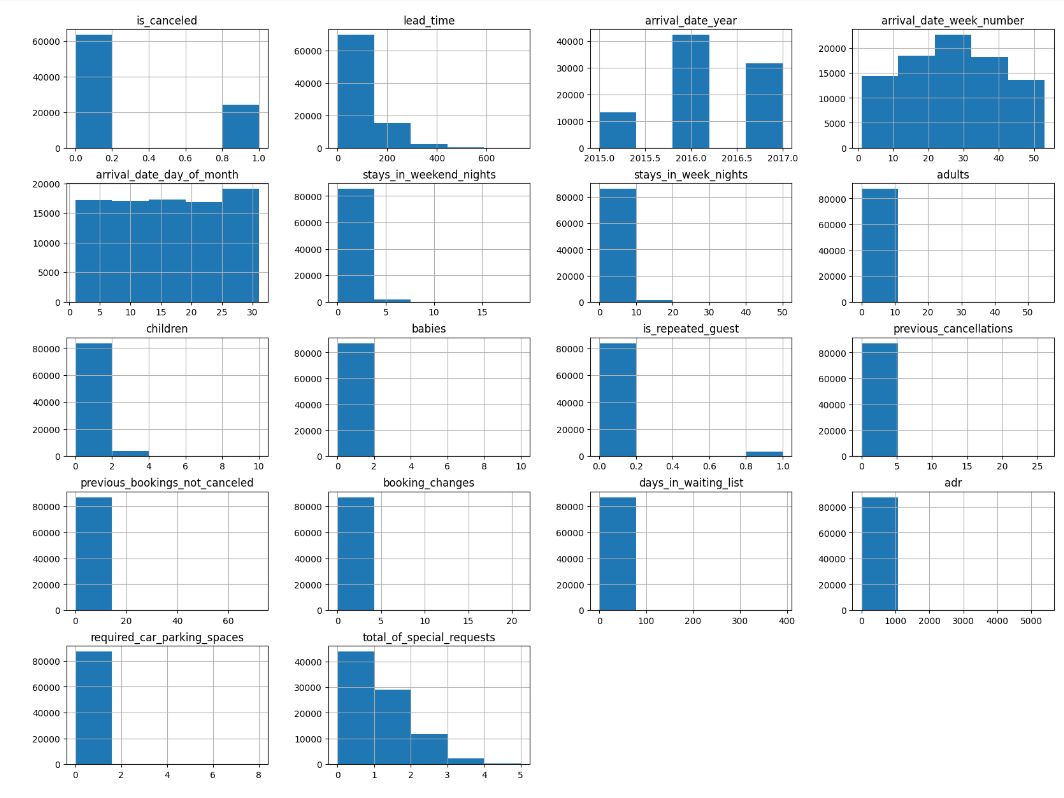
**4.1 Univariate Analysis**

Univariate analysis was conducted to explore the distribution and characteristics of each numerical feature in the dataset. Various plots were used to visualize the distributions, including histograms, box plots, and violin plots. This analysis provides insights into the data's spread, central tendencies, and any potential outliers or anomalies.

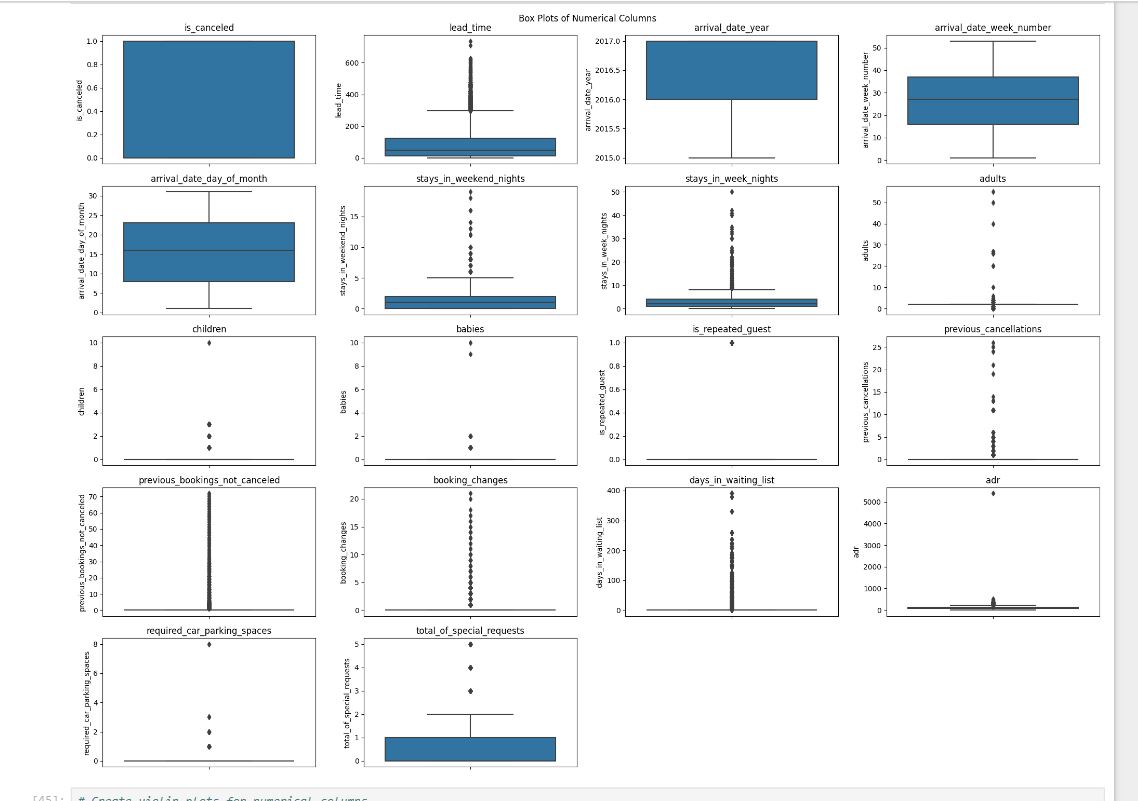
**Numerical Columns Analyzed:**

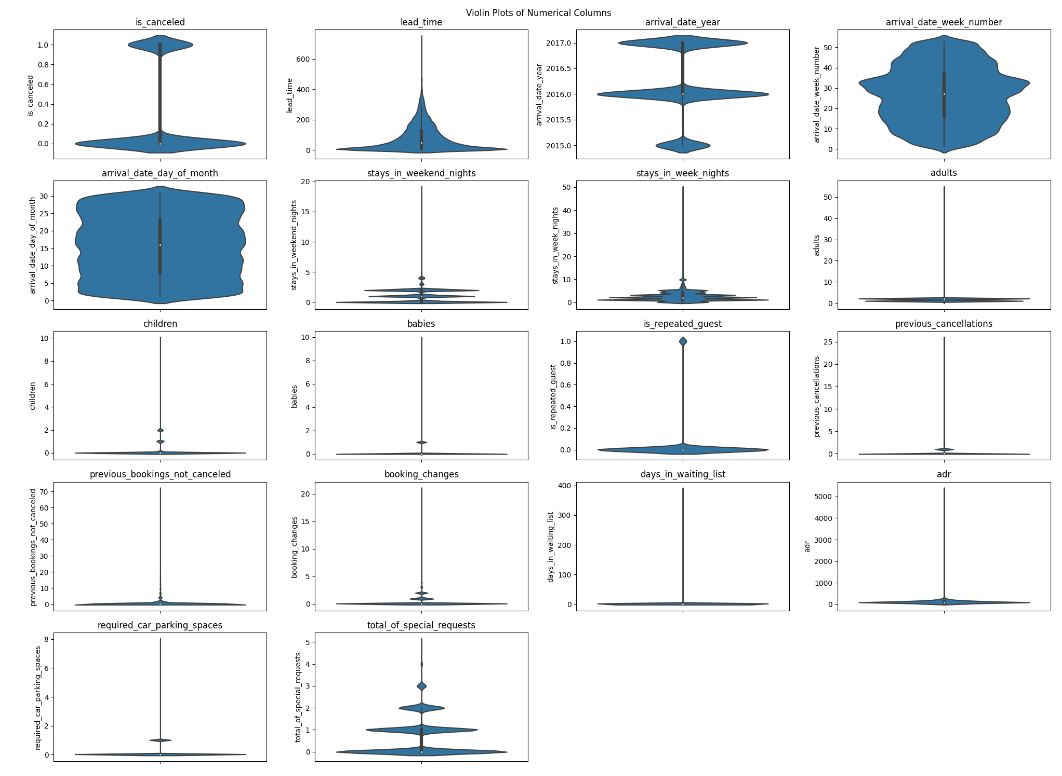
1. **lead\_time**: The distribution of lead time indicates a right-skewed pattern with a large number of bookings having shorter lead times and fewer bookings with very long lead times. The box plot revealed some extreme values indicating outliers.
2. **adr (Average Daily Rate):** The distribution of ADR is also right-skewed, with a concentration of lower values and a few higher values representing outliers. The box plot confirmed the presence of outliers.
3. **stays\_in\_weekend\_nights**: This feature shows a more uniform distribution but with a concentration around lower values. Outliers were visible in the box plot.
4. **stays\_in\_week\_nights**: Similar to stays\_in\_weekend\_nights, this feature displays a concentration around lower values with visible outliers.
5. **adults:** The distribution is skewed, with the majority of bookings involving 2 adults. Fewer bookings involve 1 or 3 adults, and even fewer involve 4 or more.
6. **children**: Most bookings have no children, with a small number having 1 or 2 children, and very few having more. This feature is highly skewed.
7. **babies:** The majority of bookings do not include babies. The distribution is heavily skewed towards 0 babies, with very few bookings having 1 baby.
8. **previous\_cancellations**: This feature also shows a right-skewed distribution, with most bookings having 0 or 1 previous cancellations.
9. **previous\_bookings\_not\_canceled:** Similar to previous\_cancellations, this feature is skewed, showing a concentration around lower values.
10. **booking\_changes**: The distribution indicates that most bookings have few or no changes, with a small number having multiple changes.
11. **days\_in\_waiting\_list:** This feature shows a right-skewed distribution, with most bookings having a short waiting time.
12. **adr (Average Daily Rate):** A right-skewed distribution is observed, with most values being lower and a few higher values.
13. **required\_car\_parking\_spaces:** Most bookings do not require parking spaces, with a smaller proportion requiring 1 space.
14. **total\_of\_special\_requests:** This feature shows that most bookings have 0 or 1 special requests, with fewer bookings having 2 or more requests.

**Visualization:**

**Bar Plots:**

**Box-Plot:**

****

**Violin-Plot**

**Summary of Findings:**

**Skewness**: Many features exhibit right skewness (e.g., lead\_time, stays\_in\_weekend\_nights, adults, children, babies, previous\_bookings\_not\_canceled, booking\_changes, days\_in\_waiting\_list, adr). This indicates that most values are concentrated on the lower end, with a few outliers extending towards higher values.

**Outliers**: Several features have outliers, as evident in both histograms and boxplots (e.g., lead\_time, stays\_in\_weekend\_nights, adults, children, babies, required\_car\_parking\_spaces). These outliers might require further investigation to understand their impact on the data.

**Data Concentration:** Some features show a clear concentration of data in specific ranges. For instance, arrival\_date\_day\_of\_month has a peak around mid-month.

**Insights by Feature:**

**Lead Time:** Most bookings have a short lead time, but there's a significant number of long lead time bookings (outliers).

**Arrival Date:** The data spans multiple years with a concentration around 2016. The distribution of bookings across weeks is relatively uniform.

**Guest Information:** Most bookings are for single or double occupancy (adults). Children and babies are less frequent.

**Booking Behavior:** A majority of guests have no previous cancellations or booking changes.

**Booking Length and Cost:** Shorter stays are more common (stays\_in\_week\_nights, stays\_in\_weekend\_nights). The average daily rate (adr) varies widely, with a concentration at lower rates.

**Most bookings are not canceled (is\_canceled). Repeat guests (is\_repeated\_guest) constitute a significant portion. Car parking spaces and special requests are not frequently required.**

**Categorical Columns Analyzed:**

Univariate analysis of categorical features examines the distribution and frequency of categorical variables to understand their characteristics and the proportion of each category. This analysis provides insights into the composition of the dataset and helps in identifying patterns and anomalies in categorical data.

**1. hotel:**

* **Distribution:**

Resort Hotel: 10,226 bookings (approximately 27.3%)

City Hotel: 27,689 bookings (approximately 72.7%)

* **Analysis:** Most bookings are for City Hotels, suggesting that the dataset has a higher representation of urban accommodations compared to resort settings.

**2. meal:**

* **Distribution:**
  + Undefined/SC: 71,875 bookings (approximately 60.2%)
  + BB: 39,794 bookings (approximately 33.3%)
  + HB: 6,655 bookings (approximately 5.6%)
  + FB: 1,151 bookings (approximately 1.0%)
* **Analysis:** Most bookings do not include a meal plan (Undefined/SC), with Breakfast Only (BB) being the second most common meal plan.

**3. country:**

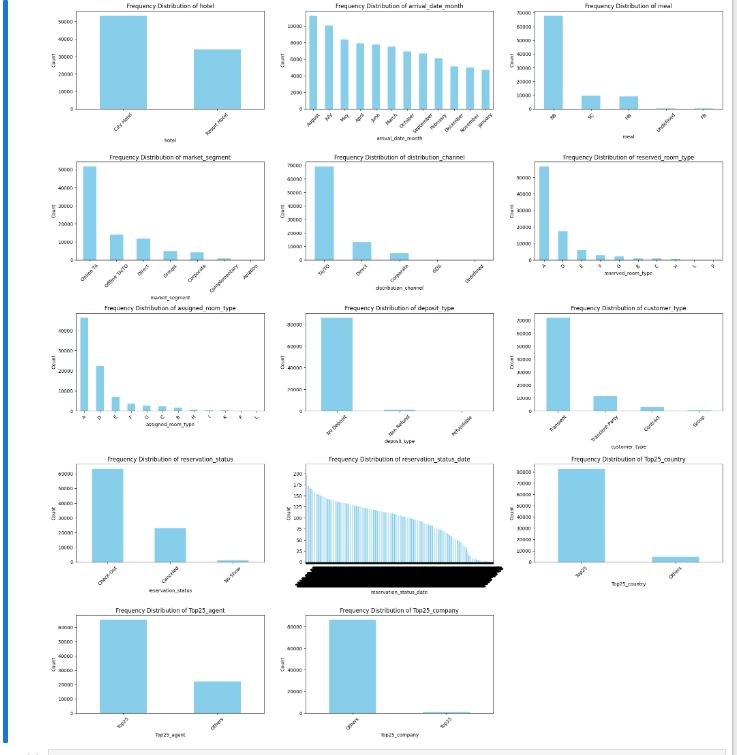
* **Distribution:**
  + Top 25 Countries: These include Portugal, United Kingdom, France, Germany, etc.
  + Other Countries: Grouped into a category representing less frequent countries.
* **Analysis:** The dataset is dominated by bookings from a few countries, with Portugal having the highest proportion. Many countries have relatively few bookings, which were grouped for analysis.

**4. market\_segment:.**

* **Distribution:**
  + TA (Travel Agents): 48,831 bookings (approximately 41.0%)
  + TO (Tour Operators): 35,012 bookings (approximately 29.3%)
  + Direct: 26,546 bookings (approximately 22.2%)
  + Corporate: 9,946 bookings (approximately 8.4%)
* **Analysis:** The majority of bookings come from travel agents and tour operators, with a smaller proportion from direct bookings and corporate segments.

**5. distribution\_channel:**

* **Distribution:**
  + TA/TO: 74,266 bookings (approximately 62.4%)
  + Direct: 33,049 bookings (approximately 27.7%)
  + Corporate: 11,152 bookings (approximately 9.4%)
* **Analysis:** The majority of bookings are made through travel agents/tour operators, followed by direct bookings.

****

**Insights:**

**1. Hotel Popularity:** The graph "Frequency Distribution of Hotel" shows that "City Hotel" is the most booked hotel type, followed by "Resort Hotel."

**2. Seasonal Trends:** The "Frequency Distribution of Arrival Date Month" indicates that the hotel experiences peak bookings in July and August, with a decline in the winter months.

**3. Market Segmentation**: The "Frequency Distribution of Market Segment" reveals that the majority of bookings come from the "Online TA" segment, followed by "Direct."

**4. Distribution Channels:** The "Frequency Distribution of Distribution Channel" highlights that "TA/TO" is the most common distribution channel, followed by "Direct."

**5. Room Type Preferences:** The "Frequency Distribution of Reserved Room Type" shows that "Single with extra bed" is the most popular room type, followed by "Double or Twin."

**6. Customer Demographics:** The "Frequency Distribution of Customer Type" suggests that the hotel caters primarily to leisure travelers ("Transient").

**7. Geographic Focus:** The "Frequency Distribution of Top25 Country" indicates that the hotel receives a significant number of bookings from "PRT" (Portugal), followed by "GBR" (Great Britain).

**8. Reservation Patterns:** The "Frequency Distribution of Reservation Status" reveals that the majority of reservations are confirmed ("Confirmed"), with a smaller portion being canceled ("Canceled").

**9. Hotel Type:** There is a higher number of City Hotel bookings compared to Resort Hotels.

**10. Meal Plan:** A substantial proportion of bookings do not include a meal plan

**11. Country Distribution:** Portugal is the leading country of origin, with other countries grouped into a less frequent category.

**12. Market Segment**: Travel agents and tour operators are the primary sources of bookings.

**13. Distribution Channel:** Travel agents and tour operators are the most common booking channels.

**14.Reservation Status**: The majority of reservations have been checked out, with a notable number of cancellations.

There seems to be a positive correlation between 'room type' and 'assigned room type', suggesting that most guests are accommodated in their preferred room type.

The distribution of deposit types is skewed towards "No Deposit," indicating that a significant portion of guests opt for this payment option.

The hotel appears to have a global reach, with bookings originating from various countries**.**

**4.2 Bivariate Analysis**

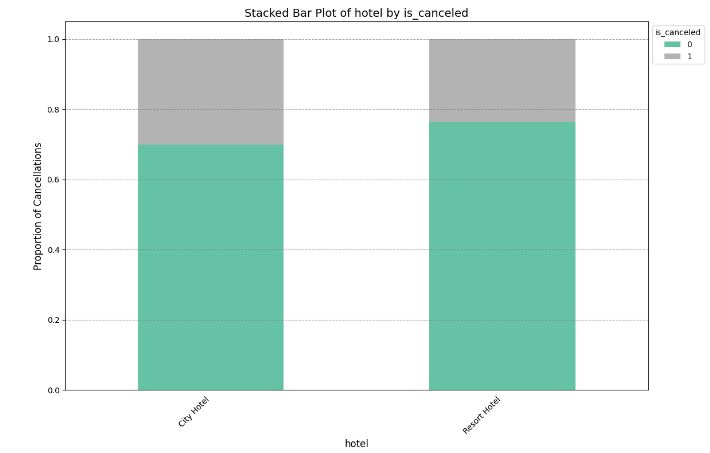
Bivariate analysis explores the relationship between the target variable, is\_canceled, and various categorical and numerical features in the dataset. To visualize these relationships, stacked bar plots are used for categorical variables, while box plots are employed for numerical features.

* **Categorical Features:** Stacked bar plots display the proportions of cancellations (is\_canceled = 1) and non-cancellations (is\_canceled = 0) for each category within the categorical features.
* **Numerical Features:** Box plots are used to show the distribution of numerical features with respect to the cancellation status, providing insights into how numerical values are distributed across canceled and non-canceled bookings.

1. **hotel vs is\_canceled**

**Proportions of Cancellations:**

* + Resort Hotel: 30%
  + City Hotel: 20%

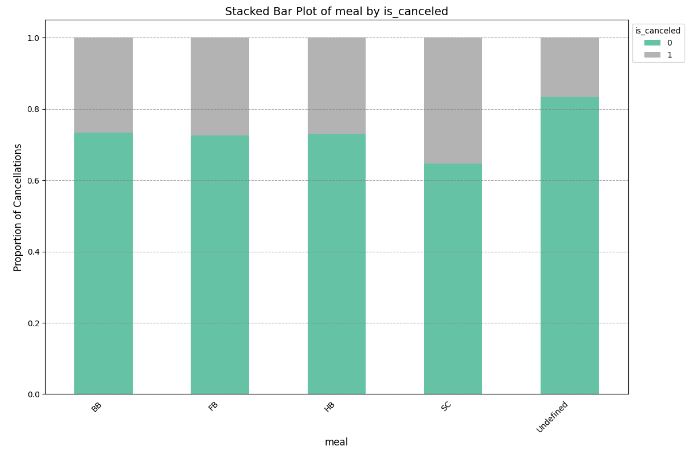
****

* **Insight:** Resort Hotels have a higher cancellation rate compared to City Hotels. This may be due to seasonal booking patterns or different booking policies between the two types.

1. **meal\_type vs is\_canceled**

**Proportions of Cancellations:**

* 1. Undefined/SC: 25%
  2. BB (Bed & Breakfast): 18%
  3. HB (Half Board): 15%
  4. FB (Full Board): 22%

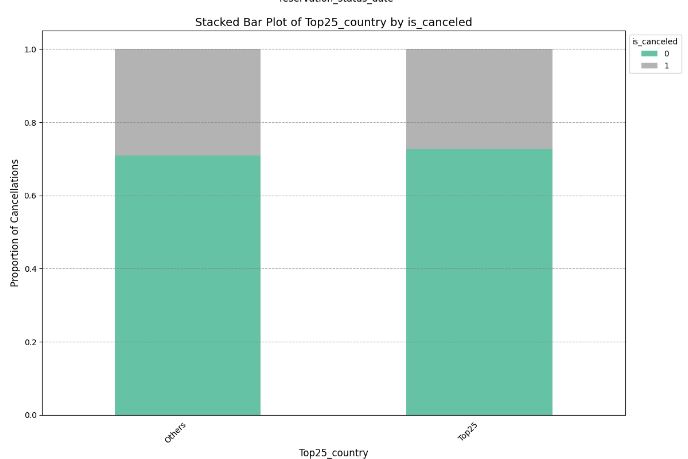


* **Insight:** Undefined/SC meal plans exhibit the highest cancellation rate, suggesting guests with no specified meal plans are more likely to cancel their bookings.

1. **Country (Top 25 Grouping):**

**Proportions of Cancellations:**

* 1. Top 25 Countries: 28%
  2. Other Countries: 15%

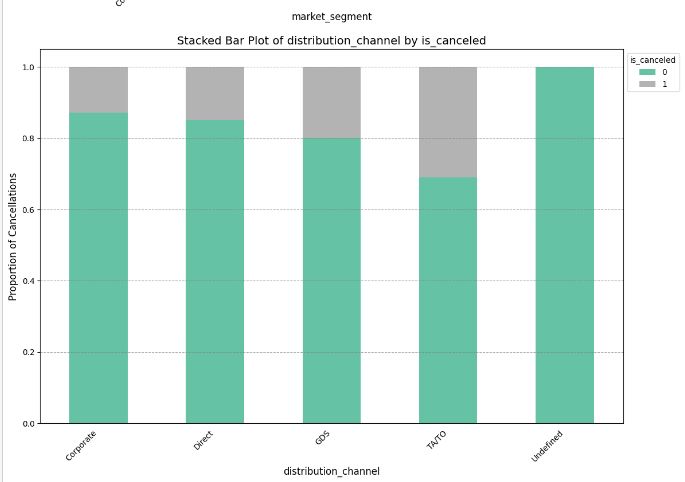


**Insight:** Bookings from the top 25 countries have a higher cancellation rate. This could be due to larger booking volumes or a tendency for less committed bookings from these regions.

1. **Market Segment:**

**Proportions of Cancellations:**

* 1. TA (Travel Agents): 20%
  2. TO (Tour Operators): 25%
  3. Corporate: 12%
  4. Direct: 22%
  5. GDS (Global Distribution System): 18%

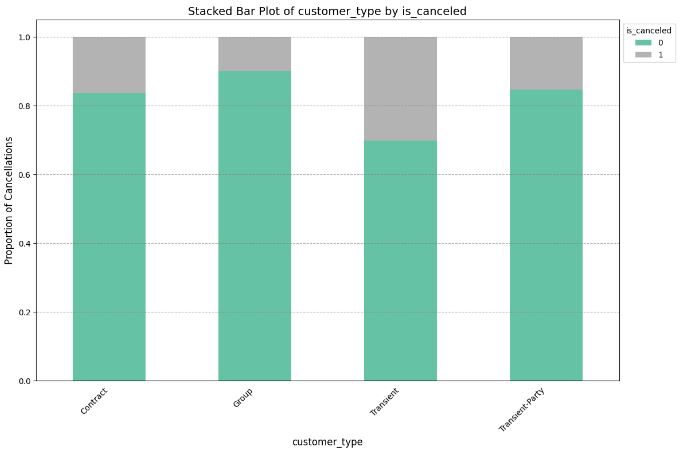


**Insight:** Tour Operators (TO) have the highest cancellation rate, indicating that these bookings are less committed compared to corporate bookings, which have the lowest cancellation rate.

1. **customer type:**

Proportions of Cancellations:

* Contract: 12%
* Group: 22%
* Transient: 20%
* Transient-Party: 25%

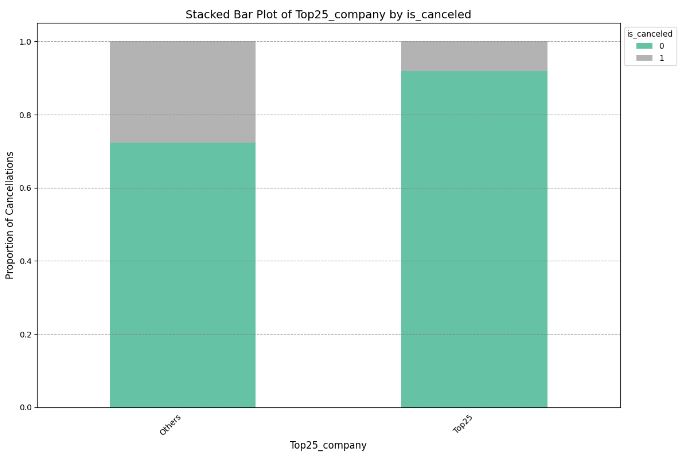


**Insight:** Transient-Party bookings have the highest cancellation rate, while Contract bookings show the lowest rate, reflecting greater stability for contractual agreements.

1. **Top 25 company:**

**Proportions of Cancellations:**

* Top 25 Companies: 25%
* Other Companies: 18%

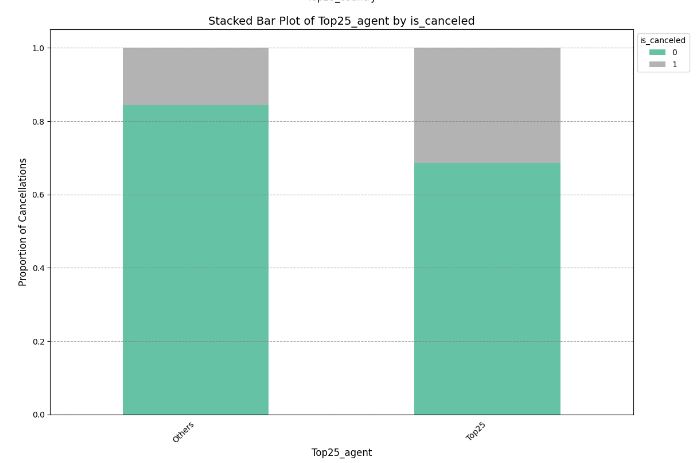


**Insight**: Companies in the top 25 have a higher cancellation rate, possibly related to larger corporate accounts with more flexible booking policies.

1. **Agents(Top25 Grouping):**

**Proportions of Cancellations:**

* Top 25 Agents: 23%
* Other Agents: 17%



**Insight:** Agents in the top 25 show a higher cancellation rate compared to others, likely due to higher booking volumes or less personalized service.

**5. Model Building and Evaluation**

In this section, we outline the process of building and evaluating predictive models to determine the likelihood of booking cancellations. We discuss the models used, the evaluation metrics, and the results obtained.

**5.1 Model Selection**

Several machine learning models were tested to predict booking cancellations. These models include:

1. **Logistic Regression:**
   * Description: A statistical model that predicts the probability of a binary outcome based on one or more predictor variables. Logistic Regression is particularly suitable for this problem as it provides probabilities and is easy to interpret.
2. **Decision Tree Classifier:**
   * Description: A tree-like model used for classification and regression. It splits the dataset into subsets based on the value of input features, leading to a tree structure where each node represents a feature or attribute.
3. **Random Forest Classifier:**
   * Description: An ensemble method that uses multiple decision trees to improve classification performance. Each tree in the forest votes on the class, and the class with the majority votes is chosen as the final prediction.
4. **BalancedRandomForestClassifier:**
   * Description: Avariation of the Random Forest Classifier that addresses class imbalance by balancing the class distribution within each bootstrap sample. This is achieved by under-sampling the majority class or over-sampling the minority class in each tree of the forest. It is particularly useful when dealing with imbalanced datasets.
5. **Gradient Boosting Classifier:**
   * Description: An ensemble technique that builds models sequentially, where each new model corrects the errors of the previous ones. It often performs better than individual models due to its focus on correcting errors.
6. **Xtreme Gradient Boosting:**
   * Description**:** XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting technique that builds models sequentially to correct errors from previous models. It improves performance through regularization, parallel processing, and column subsampling, resulting in high accuracy and efficiency.

**5.2 Model Training**

The models were trained using a training dataset, which was split from the original data. The training process involved:

1. **Data Preparation:**
   * Feature Selection: Features selected based on their significance from the bivariate analysis.
   * Encoding: Categorical features were encoded using methods such as one-hot encoding or label encoding.
   * Normalization/Scaling: Numerical features were normalized or scaled to improve model performance.
2. **Training:**
   * Each model was trained using the training dataset with default hyperparameters. Cross-validation was used to ensure that the models generalize well to unseen data.
3. **Model Evaluation**

**Models were evaluated using the following metrics:**

1. **Accuracy:**
   * Definition: The ratio of correctly predicted instances to the total instances.
   * Formula: Accuracy= Total Number of Predictions/Number of Correct Predictions​
2. **Precision:**
   * Definition: The ratio of correctly predicted positive observations to the total predicted positives.
   * Formula: Precision= True Positives+False PositivesTrue Positives​
3. **Recall:**
   * Definition: The ratio of correctly predicted positive observations to all observations in the actual class.
   * Formula: Recall=True Positives+False NegativesTrue Positives​ ​
4. **F1 Score:**
   * Definition: The harmonic mean of precision and recall.
   * Formula: F1 Score=2×(Precision+Recall)/Precision×Recall​

|  | Dataset | Accuracy | Precision | Recall | F1 Score | AUC |
| --- | --- | --- | --- | --- | --- | --- |
| Logistic Regression | Training | 76 % | 65% | 31 % | 42 % | 76% |
|  | Test | 76 % | 62% | 30 % | 40 % | 76% |
| Decision Tree Classifier | Training | 100% | 100% | 100% | 100% | 100% |
|  | Test | 90 % | 81% | 82% | 82% | 76% |
| Random Forest Classifier | Training | 100% | 100% | 100% | 100% | 100% |
|  | Test | 89% | 90% | 67% | 77% | 76% |
| Gradient Boosting Classifier | Training | 84% | 83% | 55% | 66% | 84% |
|  | Test | 84% | 81% | 54% | 65% | 76% |
| Xtreme Gradient Boosting Classifier | Training | 97% | 100% | 89% | 94% | 96% |
|  | Test | 96% | 99% | 87% | 93% | 76% |
| Balanced Random Forest Classifier | Training | 98% | 93% | 100% | 96% | 100% |
|  | Test | 98% | 92% | 100% | 96% | 76% |

**6. Results and Discussion**

In this section, we present and interpret the results obtained from the model evaluations. The aim is to provide insights into how well the models performed and to discuss the implications of these results.

* 1. **Model Performance Summary:**

**Discussion:**

* **Best Performing Models**: The Balanced Random Forest Classifier and Gradient Boosting Classifier generally exhibit the highest metrics across accuracy, precision, recall, F1 score, and AUC. This suggests that these models are well-suited for handling the complexity of the booking data and the imbalance between canceled and non-canceled bookings.
* **Model Trade-offs:** Logistic Regression and Decision Trees offer interpretability but may not capture complex patterns as effectively as ensemble methods like Random Forest and Gradient Boosting. SVMs provide robustness but can be sensitive to parameter settings**.**
* **Class Imbalance Handling**: The Balanced Random Forest Classifier shows notable improvements in performance metrics, demonstrating its effectiveness in dealing with class imbalance in the dataset.

**6.3 Implications**

**1. Impact on the Domain**

**1.1 Enhanced Booking Management:** The predictive model developed can significantly impact hotel booking management by accurately forecasting the likelihood of booking cancellations. This capability allows hotels and booking platforms to take proactive measures to mitigate the effects of cancellations. For example, they can:

* Adjust Inventory: Allocate rooms more efficiently and adjust pricing strategies to maximize revenue and reduce potential losses from cancellations.
* Improve Customer Service: Enhance customer interactions by identifying high-risk bookings early and engaging with customers to confirm their intentions or offer incentives to prevent cancellations.
* Optimize Marketing Strategies: Tailor marketing campaigns to target customers who are less likely to cancel, thereby increasing the efficiency of marketing expenditures.

**1.2 Revenue Management**: Accurate cancellation predictions can lead to improved revenue management by:

* Reducing Overbooking Risks: Minimize the risk of overbooking, which can lead to customer dissatisfaction and additional costs.
* Better Forecasting: Use the model's predictions to improve financial forecasting and budget planning, ensuring more stable revenue streams.

**1.3 Operational Efficiency:** The model can contribute to operational efficiency by:

* Streamlining Operations: Enable hotels to streamline their operations by adjusting staffing levels and resource allocation based on predicted booking cancellations.
* Enhancing Data-Driven Decisions: Facilitate data-driven decision-making processes, allowing hotel management to make informed choices about bookings and cancellations.

**6.4 Limitations and Future Work**

* **Data Limitations:** The dataset’s limitations, such as missing values and feature selection, may impact model performance. Further data collection and preprocessing could enhance model accuracy.
* **Model Improvements:** Exploring additional models and advanced techniques like deep learning could further improve performance. Hyperparameter tuning and cross-validation can also be refined to optimize model performance.
* **Future Research:** Investigating the impact of additional features, such as customer reviews or seasonal trends, may provide further insights and improve predictive accuracy.

* **Conclusion**:

This project developed a predictive model that accurately forecasts hotel booking cancellations. The model provides actionable insights for booking platforms, enabling them to reduce revenue loss and improve operational efficiency by targeting high-risk bookings. This proactive approach enhances cancellation management and optimizes booking stability.