Understanding and Reducing Hotel Booking Cancellations

A Data-Driven Approach

Capstone Project

Model Documentation

Business Analytics and Insights

Maria Alejandra Barrios Meneses

Student Number: 301314502

August, 2024

Content

Executive Summary

0.1. Executive Introduction

The purpose of this document is to present a comprehensive analysis of hotel booking data to understand the factors contributing to cancellations and develop predictive models to address these challenges. By leveraging advanced machine learning techniques and data analysis, the project aims to provide actionable insights that can enhance decision-making processes and improve overall booking stability.

0.2. Executive Objective

The primary objective of this project is to identify key predictors of hotel booking cancellations and to develop robust models that can accurately forecast cancellation risks. By understanding these factors, the study seeks to offer strategic recommendations that will help the business optimize its pricing strategies, refine room management practices, and improve customer experience.

0.3. Executive Model Description

Four machine learning models were developed and evaluated (Decision Tree, Optimized Decision Tree, Random Forest, and Logistic Regression). Each model was assessed based on its accuracy and ROC AUC metrics, which measure the proportion of correctly classified instances and the model's goodness of fit, respectively. The Optimized Decision Tree, fine-tuned using Grid Search, was selected as the best model because combines interpretability with enhanced accuracy.

0.4. Executive Recommendations

Actionable recommendations will be provided after a deep exploration of key drivers that influence the likelihood of hotel cancellation and understand how those behaviors are affecting

the business. Strategic recommendations to enhance booking stability and optimize business operations will include:

- **Customer Communication:** Improve clarity and transparency in bookings terms, engage proactively with early bookers, and provide additional support for booking changes.
- **Personalized Incentives:** Implement reward programs, offer discounts, and provide incentives to encourage customer commitment and minimize cancellations.
- **Marketing Campaigns:** develop target campaigns and promote the benefits of maintaining reservations to reduce cancellations.

Introduction

1.0. Background

In the hospitality industry, managing cancellations is critical for maintaining profitability and operation efficiency. Cancellations not only impact revenue but also create challenges in resource management, operation efficiency, and customer satisfaction. The Hotel Booking Trends report, made by SiteMinder (2023), reveals that hotels experience an average cancellation rate between 20% to 25%. Those cancellations rates significantly impact hotel's revenue; according to the SHR Group (2023) cancellation can lead to revenue losses up to 15% annually and create operational challenges like overbooking, under booking, and inefficient resource allocation.

However, the ability to predict and manage cancellations can greatly enhance a hotel's ability to maintain occupancy rates and optimize their overall operations. Recent advancements in machine learning and data analytics offer new opportunities to address this issue. Hollander (2023) highlights that predictive analytic enables hotel to make data-driven decisions, enhancing

profitability, efficiency, and guest satisfaction by optimizing demand forecasting, pricing strategies, and staffing levels (para. 27).

By leveraging historical booking data and employing sophisticated modeling techniques, business can gain deeper insights into the factors driving cancellations and develop strategies to mitigate their impact.

2.0. Problem Statement

The core problem addressed in this project is the high cancellation rates disrupt inventory management, staffing schedules, and financial forecasting, leading to revenue loss and operational inefficiencies.

This project is crucial to understand and predict cancellation, identifying key factors contributing to cancellations, develop predictive models to forecast cancellation likelihood, and offer actionable recommendations to reduce cancellation rates and enhance booking stability.

3.0. Objectives & Measurement

To tackle the issue of booking cancellations, this project sets several key objectives, supported by comprehensive measurement methods.

Objectives

- Analyze booking data to identify factors significantly influencing cancellation rates.
- Create and evaluate predictive models to forecast booking cancellations with high accuracy and ROC AUC score.
- Develop practical strategies based on the model findings to reduce cancellation rates and enhance operational efficiency.

Measurement

Model Performance Metrics

- **Accuracy:** this metric assesses the proportion of correct classifications; the higher number is better.
- **ROC AUC:** this metric evaluates the model's ability to distinguish between canceled and non-canceled bookings, as well the higher number is better.

Business Metrics

- **Revenue per Available Room (RevPAR):** measures the average revenue earned per available room, assessing the financial impact of cancellation on revenue generation. RevPAR should increase as the predictive model helps reduce booking cancellations.
- **Average Daily Rate (ADR):** analyze the average revenue earned per occupied room, reflecting the effect of cancellations on room pricing strategies and overall revenue. With reduced cancellation rates, hotels can potentially stabilize or even increase ADR.

4.0. Assumptions and Limitations

This section outlines the core assumption that supports the project's methodology and acknowledges potential limitations that could affect the results or applicability of the findings.

Assumptions

- **Data quality and completeness:** the project assumes that the historical data is accurate, complete and representative of the hotel's operation over the analyzed period.
- **Model Generalization:** the predictive model developed assumes that the insights gained can be generalized to other periods, locations, or hotel chains with similar characteristics.
- **Stable Market Conditions:** the project assumes that market conditions, such as demand fluctuations, economic factors, and competitor actions, remain stable during the period of analysis.

Limitations

- **Data Limitations:** the analysis is limited by the availability and granularity of data. The lack of detailed customer profiles or external factors could limit the accuracy of predictions.
- **Modeling Limitations:** the predictive models rely on specific machine learning techniques, which may not capture all possible patterns or interactions.
- **Operations Challenges:** the practical implementation of recommendations may face resistance or challenges due to organizational constraints, technology limitations, or market competition.
- **Dependance on Historical Trends:** the models depend heavily on historical data and may not adapt quickly to new trends or shifts in consumer behavior.

Data Sources

5.0. Data Set Introduction

The dataset was obtained from the data article "Hotel Booking demand datasets" (Nuno, et al., 2019) available at the Journal of Science Direct. The article provided two datasets with hotel booking information capturing both actual arrivals and cancellations; the first dataset (H1) belongs to a resort hotel located in the region of Algarve, Portugal, while the second (H2) belongs to a city hotel located in Lisbon, Portugal. Both datasets maintain an identical structure with 31 variables. H1 dataset contains 40,060 entries, while H2 dataset contains 79,330 entries. Each entry corresponds to a hotel booking with arrival scheduled from July 1, 2015, to August 31, 2017.

6.0. Initial Data Cleansing or Preparation

The two datasets, H1 and H2, were merged since most of the Hotel Chains operate both resort and city hotel. The column *resort_hotel* was created for each dataset to be able to identify

if the record corresponds to a resort hotel or a city hotel (1 – resort hotel, 0 – city hotel). The dataset *df* is the result of the merging process, containing 119,390 entries and 32 variables.

The initial variable names had a *CamelCase* structure and were changed to *snake_case* structure by separating words with underscores, enhancing readability and consistency. Additionally, the variable name “adr” was changed to “average_daily_price”.

After checking the stored variable types, five variables were proceeded to change their data types for the logical ones: *is_canceled* (integer to category), *reservation_status_date* (object to date time), *children* (float to integer), *is_repeated_guest* (integer to category), and *resort_hotel* (integer to category).

The column *arrival_date* was created as the result of joining the columns: *arrival_date_year, arrival_date_month*, and *arrival_date_day_of_month*.

Duplicates entries were checked, and it was found 31,994 duplicated entries, that could occur because bookings with identical attributes may happen on the same day and since the dataset does not contain any unique identifier, like the booking id, it is difficult to classify it they are a true duplicate. However, the model was tested with and without duplicates entries, and it was found that the model performs better with duplicates; therefore, duplicates entries were kept.

After proceeding with the initial preparation, the dataset contains 119,390 entries and 33 columns.

7.0. Data Dictionary

The dataset contains 33 variables of which 16 are numerical, 15 are categorical, and 2 are Date types. A detailed description of the variables is presented is shown in Table 1.

Table 1. Variables Description

Variable	Type	Description
hotel_type	Categorical	Hotel type reserved, indicating: **0** – the booking corresponds to a City Hotel **1** – the booking corresponds to a Resort Hotel
is_canceled	Categorical	Value indicating: **0** - the booking was not canceled **1** - the booking was canceled
reservation_status	Categorical	Reservation last status, assuming one of 3 categories: **Canceled** - booking was canceled by the customer; **Check-Out** - customer has checked in but already departed; **No-Show** - customer did not check-in and did not inform the reason why.
reservation_status_date	Date	Date at which the last status was set.
arrival_date_year	Integer	Year of arrival date.
arrival_date_month	Categorical	Month of arrival date with 12 categories: January to December.
arrival_date_day_of_month	Integer	Day of the month of the arrival date.
arrival_date	Date	Date of the customer arrival.
lead_time	Integer	Number of days between the booking date and the arrival date.
arrival_date_week_number	Integer	Week number of the arrival date
stays_in_weekend_nights	Integer	Number of weekend nights of the booking (Saturday or Sunday)
stays_in_week_nights	Integer	Number of weeknights of the booking (Monday to Friday)
adults	Integer	Number of adults
children	Integer	Number of children
babies	Integer	Number of babies
meal	Categorical	Type of meal booked, assuming one of 5 categories: **BB** - Bed & Breakfast; **HB** - Half Board (breakfast and one other meal); **FB** - Full Board (breakfast, lunch and dinner); **SC** - Sell Catering (no meals are included); **Undefined**
country	Categorical	Country of origin

market_segment	Categorical	Type of customer or booking category, assuming one of 8 categories: **Direct** - booking was made directly by the customer; **Corporate** - booking was made by corporate clients or business travelers; **Online TA** - booking was made through online travel agencies (e.g., Expedia, Booking.com); **Offline TA/TO -** booking was made through traditional non-digital agencies or tour operators; **Complementary** – bookings that might be part of a complementary offer, often used in promotional or special events. **Groups** - booking was made for a group; **Aviation** - booking was made through airline partnerships. **Undefined**
distribution_channel	Categorical	Distribution channel which bookings were processed, assuming one of 5 categories: **Direct** - booking was made directly through the hotel's own channels (e.g., website, phone); **Corporate** - booking was made through corporate agreements or channel; **TA/TO** - booking was made through Travel Agencies or Tour Operators; **GDS** - Global Distribution System. **Undefined**
is_repeated_guest	Categorical	Value indicating: **0** - the booking was not from a repeated guest **1** - the booking was from a repeated guest
previous_cancellations	Integer	Number of previous bookings cancelled by the customer
previous_bookings_not_canceled	Integer	Number of previous bookings not cancelled by the customer
reserved_room_type	Categorical	Code for the type of room assigned, assuming one of 10 categories: A, B, C, D, E, F, G, H, L, P
assigned_room_type	Categorical	Code for the type of room assigned, assuming one of 12 categories: A, B, C, D, E, F, G, H, I, K, L, P

booking_changes	Integer	Number of changes made to the booking from the moment the booking was entered until check-in or cancellation
deposit_type	Categorical	Refers to the deposit made to guarantee the booking, assuming one of three categories: **No Deposit** - no deposit was made; **Non Refund** - a deposit was made with a value equal to the total cost of stay; **Refundable** - a deposit was made with a value under the total cost of stay.
agent	Categorical	ID of the travel agency that made the reservation
company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking
days_in_waiting_list	Integer	Number of days the booking remained in the waiting list before being confirmed.
custome_type	Categorical	Type of booking, assuming one of four categories: **Contract** - refers to bookings linked to an allotment or any other form of contract. **Group** - refers to booking associated with a group. **Transient** - refers to a transient booking that is not associated to other transient booking. **Transient-party** - refers to a transient booking that are associated with at least other transient booking.
average_daily_rate	Float	Average Daily Rate per night
required_car_parking_spaces	Integer	Number of car parking spaces required by the customer.
total_of_special_requests	Integer	Number of special requests made by the customer.

Data Exploration

8.0. Data Exploration Techniques

8.1. Descriptive Statistics

The descriptive statistics provide a summary of the central tendency, dispersion, shape of the distribution of the dataset's features. This section presents the key statistics for both numerical and categorical variables.

Numerical Variables

- ***lead_time:*** the average lead time is 104.01 days with a standard deviation of 106.86 days, indicating significant variability in how far in advance guests book their stays.
- ***stays_in_weekend_nights*** **and** ***stays_in_week_nights:*** most of the bookings have a relatively short stay time, with a mean of 0.93 nights for weekends and 2.5 nights for weekdays. There are some long-term bookings with a maximum of 19 nights for weekends and 50 nights for weekdays.
- **Number of guests** ***(adults, children, babies):*** the average booking includes approximately 1.86 adults, 0.10 children, and 0.01 babies, indicating that most of the bookings are for adults. There is an extreme outlier for *adult*, 55 adults, which could represent a group booking. There are 180 bookings that had zero for the variables *adults, children* and *babies,* meaning that there were not any guests associated to an actual booking. The records where the number of adults is 0 and there is at least one child or baby are 223, representing a unique subset of the dataset that should be handle appropriately for accurate modeling.
- **previous_cancellations and** ***previous_bookings_not_canceled:*** they present low average values, 0.09 for previous cancellation and 0.14 for previous bookings not cancelled suggest that previous bookings and cancellations are rare. Both variables might be crucial in predicting cancellations, as customers with a history of cancellations are more likely to cancel again.
- ***booking_changes*** **and** ***days_in_waiting_list:*** the mean of 0.22 for booking changes indicates that most of the bookings are not modified; however, there is a maximum value

of 21. Days on the waiting list also show a large variance, with a mean of 2.32 days and a maximum of 391 days.

- ***average_daily_rate:*** the mean average daily rate is $101.83 with a standard deviation of $55.01, indicating variability in pricing. There are two extreme values of -$6.38 and $5,400 suggest data errors or exceptional data cases that need to be analyzed.
- ***required_car_parking _spaces*** **and** ***total_of_special_requests:*** these variables present low means, 0.06 and 0.57 respectively, indicating that most bookings do not require car parking or special requests.

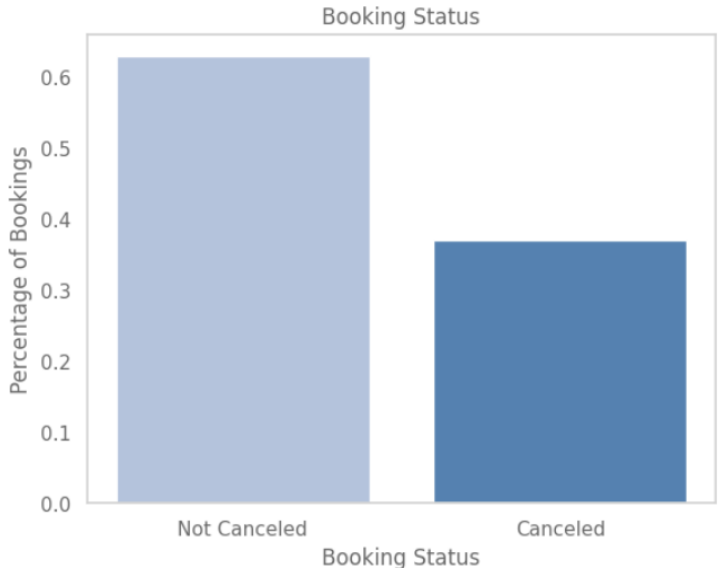
Categorical Variables

- ***hotel_type***: most of the reservation were booked for the City Hotel, accounting 66.44% for City Hotel and 33.55% for Resort Hotel.
- ***is_repeated_guest***: most of the bookings were not made from a repeated guest (97%).
- ***meal***: BB (Bed & Breakfast) meal type is predominant, accounting for 77% of bookings. There is a category Undefined that represents 1% of bookings that should be treated as SC (Self Catering) since any meal were included in the booking.
- ***country:*** most of the guests come from Portugal (40.87%), followed by the United Kingdom (10.2%), and France (8.78%). There are 177 unique values for a country.
- ***market_segment:*** bookings with Online TA (Travel Agents) segment is the largest, comprising 47.3% of bookings (e.g., Booking.com, Expedia). There are just 2 records that have ‘Undefined’ as market segment.
- ***distribution_channel:*** bookings with TA/TO (Travel Agents/Travel Operators) channel is the largest channel, with 82%. The channel ‘Undefined’ has 5 records that should be handled.

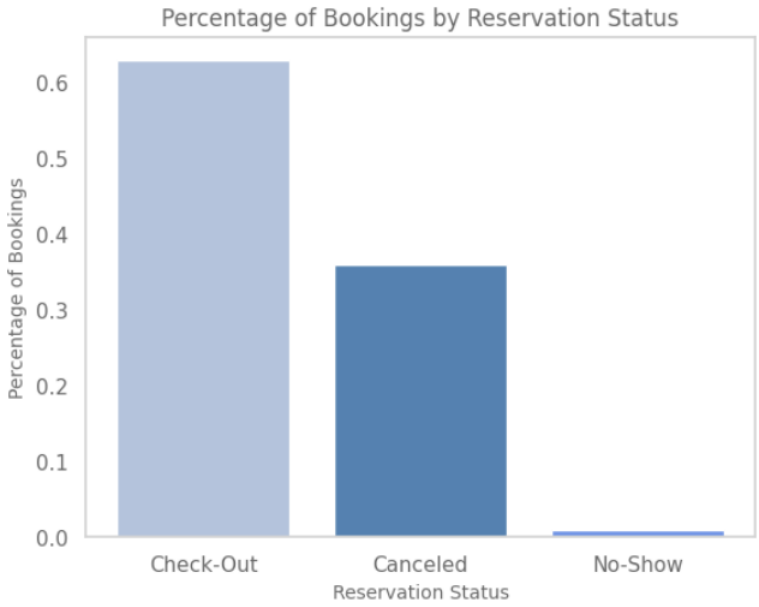
- ***reserved_room_type* and *assigned_room_type:*** the most common reserved room type is 'A' (72.02%) and the same applies to the assigned room type (62%).
- ***deposit_type:*** most bookings are 'No Deposit' with a 87.64% of reservations, following for 'Non Refund' (12.21%) and 'Refundable' (0.14%).
- ***agent* and *company:*** there are 334 unique agent IDs where the ID '9' accounting for 26.78% of bookings; meanwhile there are 353 unique company IDs where 94.3% of bookings have 'NULL'. According to Nuno et. al (2019) the 'NULL' category for the variables *agent* and *company* indicates "not applicable" instead of a missing value, meaning that the booking was not made through an agent or company (para. 13).
- ***customer_type:*** most customers are 'Transient' (75%) followed by 'Transient-Party' (21%). 'Contract' and 'Group' bookings are less common comprising 3.4% and 0.48% respectively.
- ***reservation_status:*** this is a secondary variable that confirms that the proportion between canceled and not canceled bookings, where 63% are 'Check-Out' bookings and 37% are 'Canceled' bookings.

8.2. Visualization

The following figure illustrates the distribution of canceled and not canceled bookings, with 63% of bookings marked as 'Not Canceled' (0) and 37% as 'Canceled' (1). This distribution highlights that a significant proportion of bookings are completed as planned, while a substantial minority are subject to cancellation.

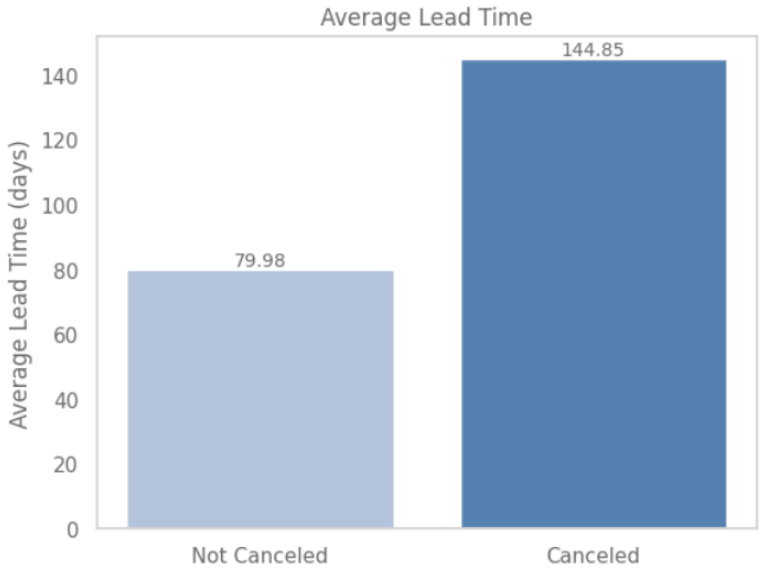


The *reservation_status* variable provides additional information for the Canceled bookings detailing which bookings were canceled prior to arrival date (36%) and which bookings did not check in and did not provide any reason for cancellation (1%). Given that *is_canceled* effectively captures whether a booking was completed or not, the *reservation_status* can be considered redundant and will be dropped for future modeling.

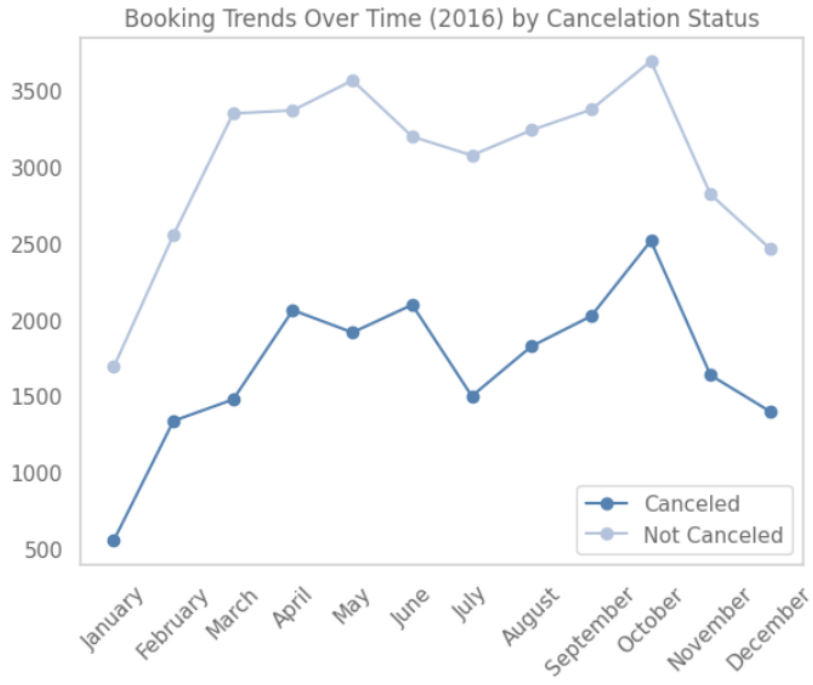


The average lead time for bookings that are not canceled is approximately 79.98 days; in contrast, the average lead time for bookings that are canceled is higher, approximately 144.85

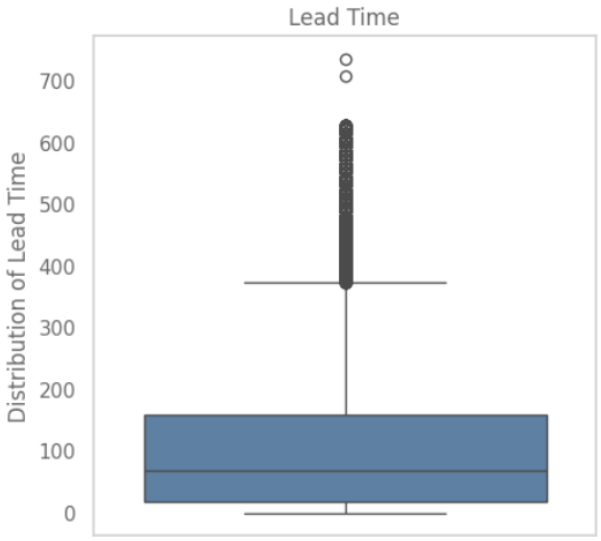
days. This suggests that reservations made further in advance are more likely to be canceled. Conversely, last-minute or closer-to-date bookings are made with more certainty and lower likelihood of being canceled.



The next graph shows the booking trend over time sorting just the bookings for year 2016 since is the only year with complete information for all the months. Bookings typically peak from July to September, where September is the month with the highest number of bookings. Additionally, there is a noticeable decline in winter months especially December and January.

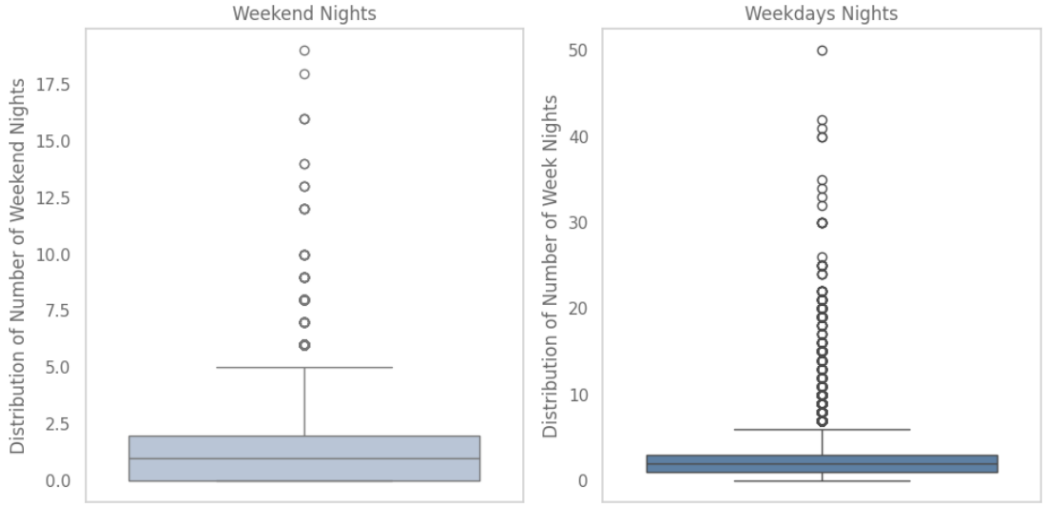


The median lead time is 69 days, representing the interval between the booking date and the arrival date. However, outliers extend beyond this median (2.52%), with the maximum lead time reaching up to 737 days.

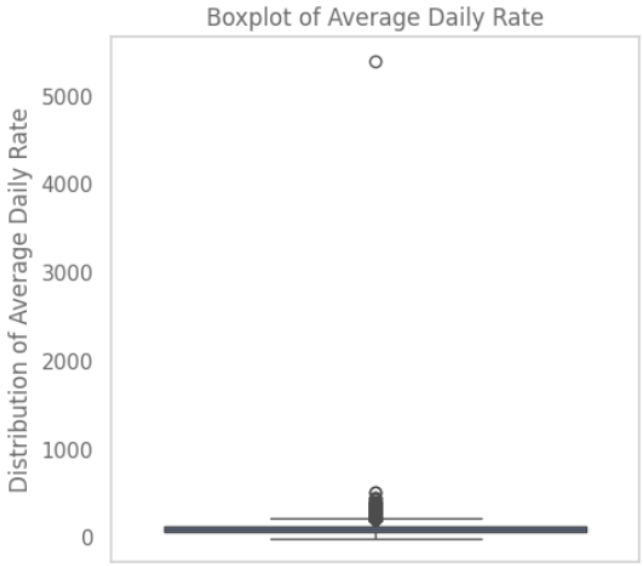


Weekend nights stay shows a median of 1 night, with a concentration of 0 and 2 nights, suggesting preference for shorter stays. The distribution extends up to a maximum of 19 nights, with a few proportion of outliers (0.22%). On the other hand, weeknight stay has a higher

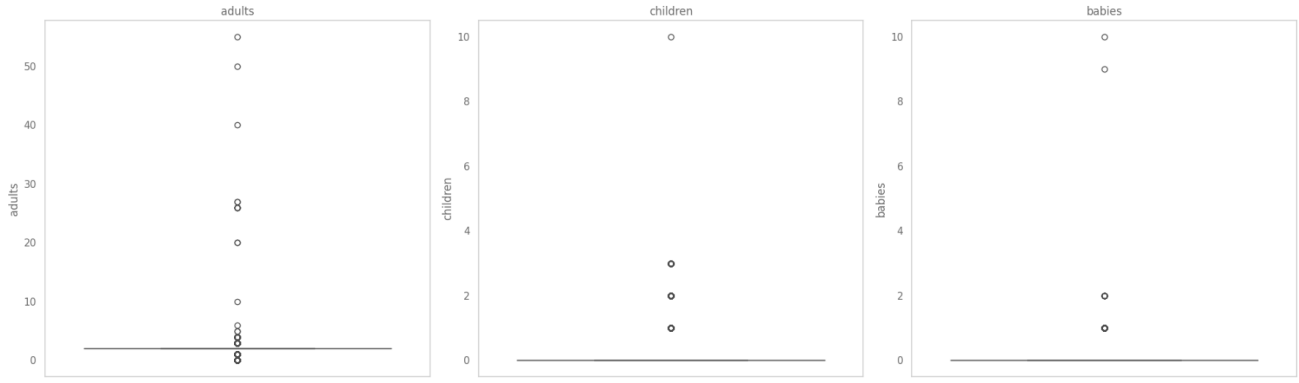
median stay of 2 nights with a range from 1 to 3 nights, with 2.81% outliers extending between 6 to 50 nights.



The boxplot analysis of the *average_daily_rate* highlights significant diversity in pricing. The median is $94.5, with most rates falling between $69.3 and $126. The minimum record rate is $-6.38, which suggests the presence of data entry errors or special cases, while the maximum rate reaches $5,400. Outliers are notably presence, with 3.18% of the data points exceeding the upper bound of $211.

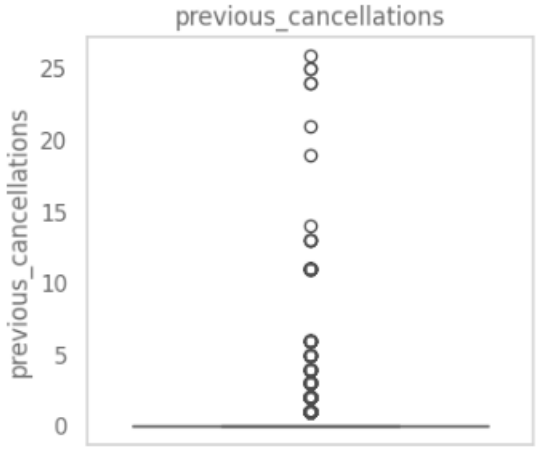


The boxplot analysis of guest demographics reveals the number of adults per booking consistently shows a median of 2, whereas for children and babies the median is 0, indicating that most bookings do not include children or babies. Despite this, there are occasional bookings for up to 55 adults, 10 children, or 10 babies.

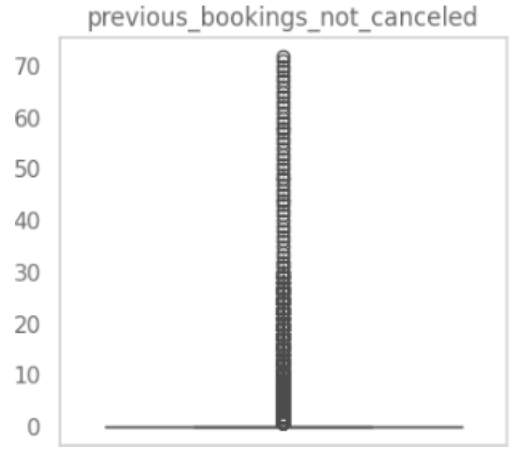


The statistics for these five variables (*previous_cacellations, previous_bookings_not_cancelled*, *booking_changes*, *days_in_waiting_list,* and *required_car_parking)* follow a similar pattern, where the data is heavily skewed towards zero, with a median of 0 and both the first and third quartiles at 0. Those variables present a small proportion of data points consider outliers:

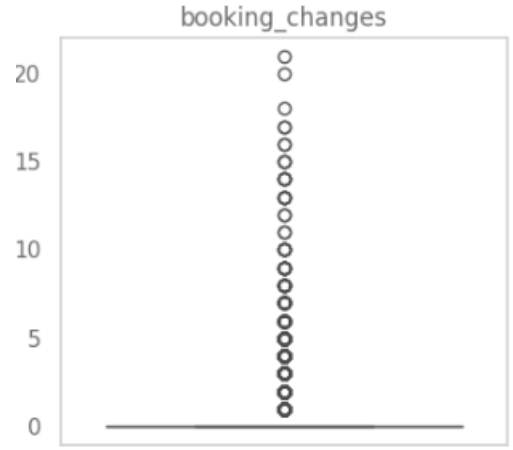
- For ***previous_cancellation***, the maximum value reaches 26, with a small proportion of guests with a history of cancellations (5.43%).



- For ***previous_booking_not_cancelled***, the maximum value is 72 with 3.03% of the data points being outliers, indicating a minority of guests have a substantial history of non-canceled bookings.

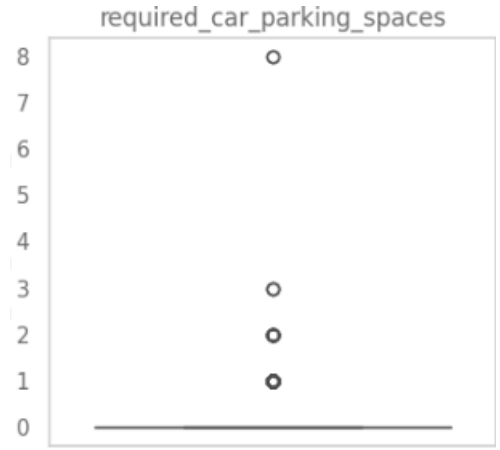


- For ***booking_changes***, the maximum value is 21 changes with 15.14% of data points being outliers, suggesting there are frequent modifications to bookings.

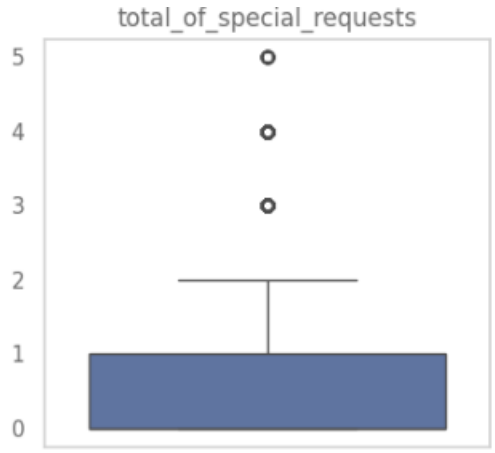


- For ***days_in_waiting_list***, 3.1% of data points classified as outliers with a maximum of 391 days in waiting list, this indicates that while waiting lists are uncommon, when they do occur, they can be extensive.

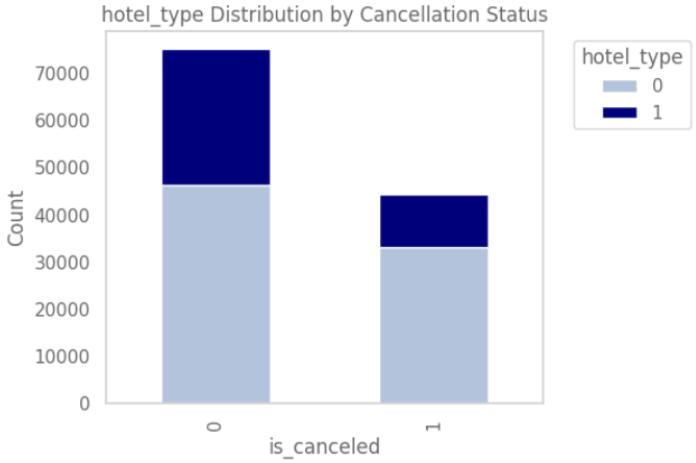
- For ***required_car_parking_spaces***, the maximum value is 8 and 6.21% of data points accounting for 1 or more parking spaces, suggesting occasional bookings with parking space requirements.



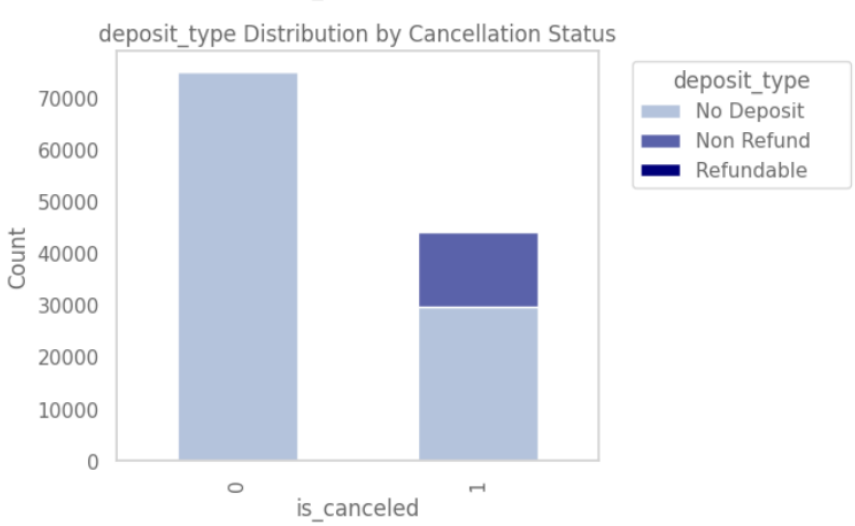
The variable ***total_of_special_requests*** show a slight variation, with a median of 0, Q1 at 0, and Q3 at 1. The upper bound is 2.5, and the maximum value is 5, with 2.41% of data points exceeding the upper bound. This indicates that some of the guests do not make special requests, some do, and a few make multiple requests.



In terms of hotel type, a substantial majority of customers who cancel their bookings have reserved City Hotel, accounting for 74.85% of cancellations, compared to 25.15% for resort hotel. Conversely, among customer who do not cancel, the distribution is more balanced, with 61.5% for City Hotel and 38.5% for Resort Hotel. This suggests that city hotels are more likely to be associated with cancellations.

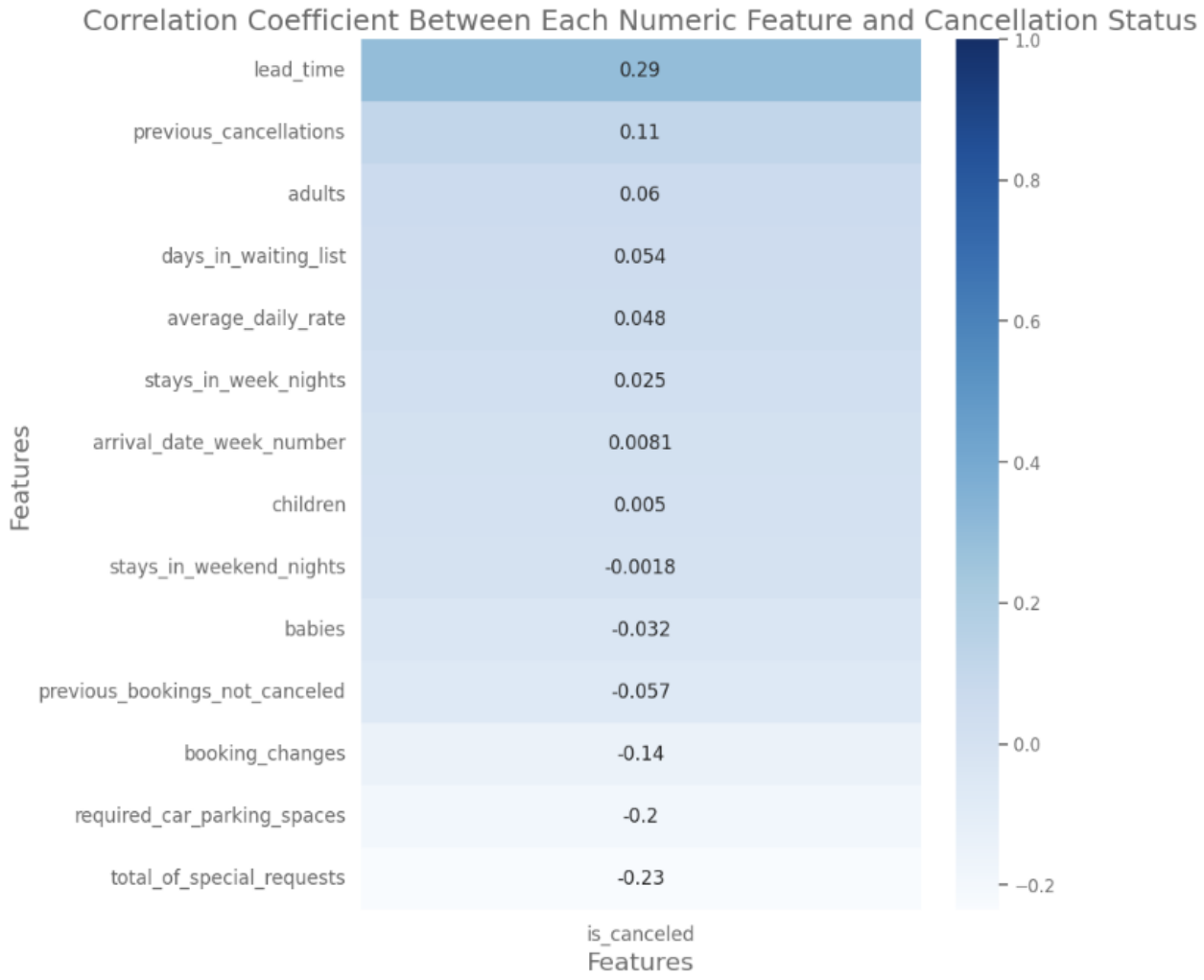


Regarding deposit types, 99% of customers who do not cancel their bookings select the 'No Deposit' option. In contrast, customers who do cancel, 67% also choose 'No Deposit', while 32% opt for 'No Refundable' option. This 32% represents a significant proportion, indicating that a considerable number of customers still cancel their bookings even when their deposit is non-refundable.



8.3. Correlation Analysis

The correlation analysis investigates the relationship between various factors influencing hotel booking cancellations. By identifying significant correlations, this analysis reveals underlying patterns and dependencies among variables such as lead time, number of special requests, requiring of parking spaces, and customer history (particularly prior cancellation) . Understanding these correlations offers insights into the drivers of booking behaviour and highlights areas for potential intervention. The next graph illustrates the correlations between the numerical variables with the target variable “is_canceled”.



lead_time is the most highly correlated feature with whether a booking is canceled, with a correlation of 0.29. this is expected as a longer lead times increases the likelihood of unforeseen circumstances impacting travel plans, thereby contributing to higher cancellation rates.

total_of_special_requests shows the second feature with the strongest correlation (-0.23) with booking cancellation. The likelihood of cancellation decreases as the number of special requests increases. This suggests that fulfilling the special needs may reduce the probability of a customer cancelling their reservation. Related to special requests, the variable **required_car_parking_spaces** is the third feature with the strongest correlation (-0.2),

suggesting that as the number of parking requests increases, the likelihood of cancellation decreases.

The feature **previous_cancellation** presents a correlation of 0.11 with whether a booking is canceled, suggesting that the chance of cancellation increases having a cancellation history.

9.0. Data Cleansing

9.1. Missing values

Handling missing values is crucial for ensuring the accuracy and reliability of data analysis and machine learning models. In this study, the following are the variables with missing values:

- ***country**:* this variable presents 488 missing values which represent a 0.40% of the dataset.
- ***agent* and *company**:* for these variables some entries were labeled as 'NULL' (16,340 for *agent* and 112,593 for *company*). According to Nuno et. al. (2019) those entries should not be considered as missing values, but rather as 'not applicable', indicating bookings made without an agent or company.

9.2. Balance dataset

The dataset exhibited an imbalance class distribution, with 37% of the entries classified as 'Canceled' (1) and 63% classified as 'Not-Canceled' (0). This imbalance can present challenges for machine learning models, as the model may become biased toward the majority class, potentially leading to overfitting and reduced predictive performance for the minority class.

To address this issue, several techniques can be employed to balance the dataset. Resampling methods such as oversampling the minority class or undersampling the majority

class can help to achieve a more equitable distribution of classes. Balancing the dataset is crucial for ensuring that the model learns to accurately predict both classes and generalizes well to new, unseen data, thereby improving the overall performance of the model.

9.3. Outliers

Outliers can significantly affect the result of the booking cancellation analysis, so it is important to detect and handle them appropriately. After addressing missing values and balancing the dataset, a subsequent examination of outliers was conducted revealing changes in the statistics for the variables of interest. All the numerical variables presented data points considered outliers; the next table summarize the proportion of outliers for each feature, in next section it will explaining how the outliers were handled:

Variable Name	Median	Upper Bound	Max value	Outliers
lead_time	78	393	709	2.28%
stays_in_weekend_nights	1	5	19	0.25%
stays_in_week_nights	2	6	42	2.78%
average_daily_rate	95	210	5400	3.27%
adults	2	2	55	5.2%
children	0	0	10	7.17%
babies	0	0	9	0.72%
previous_cancellations	0	0	26	7.06%
previous_bookings_not_canceled	0	0	72	2.46%
bookings_changes	0	0	21	13.41%
days_in_waiting_list	0	0	391	3.56%
required_car_parking_spaces	0	0	8	4.94%
total_of_special_requests	0	2.5	5	2.18%

10.0. Summary

Missing values were identified in the variable ***country,*** with 0.40% of the dataset not having this feature, the approach to handling these missing values is detailed in *section 11.1.* The dataset also exhibited class imbalance, with 37% of entries marked as ‘Canceled’ and 63% as ‘Not-Canceled’, *section 11.2. Resampling* covers the resampling technique used to ensure a

balanced dataset. Lastly, outliers were also identified and will be explained in *section 11.3.* how they were managed to reduce skewness.

Data Preparation and Feature Engineering

11.0. Data Preparation Needs

11.1. Handling missing values

The only variable that presents missing values is *country*, 488 missing values. Those missing values were imputed for the most frequently value, in that case they were imputed for 'Portugal'.

11.2. Resampling

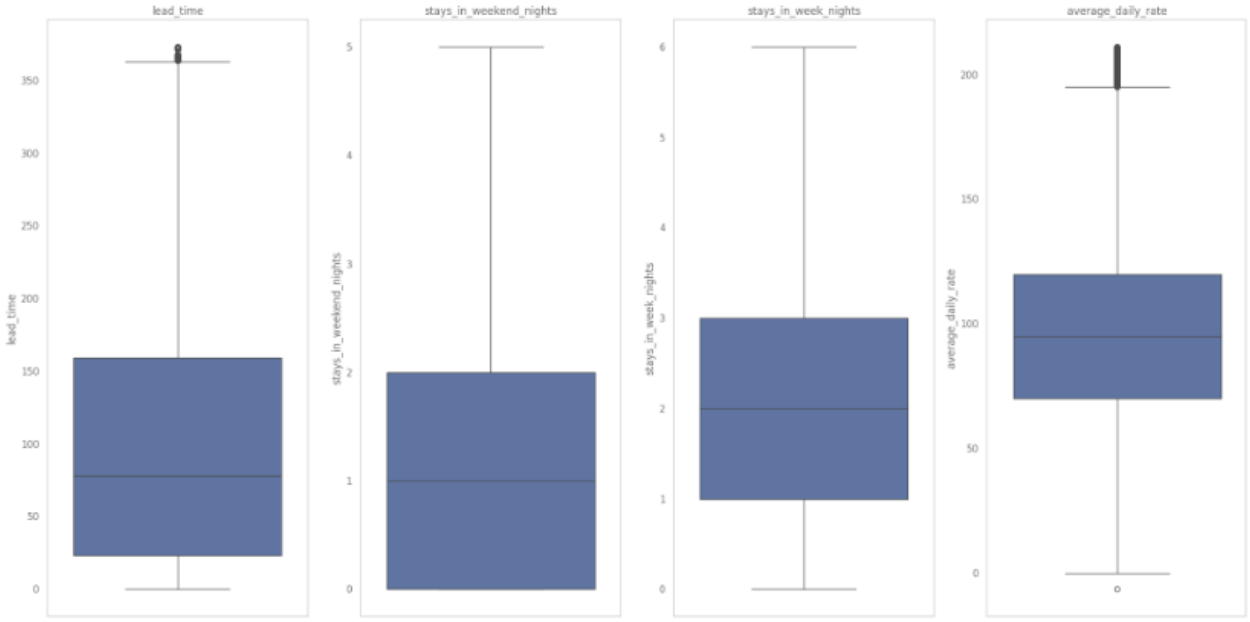
The Downsampling Technique was employed to address the challenge of working with an imbalanced dataset, 37% of entries were classified as 1 'Canceled' and 63% were classified as 0 'Not-Canceled' bookings. Dowsampling involves reducing the number of instances in the majority class to match the number of instances in the minority class. In this case, the majority class, 'Not-Canceled', was downsampled to 44,224 entries that align to the number of 'Canceled' entries.

By applying downsampling, the dataset was balanced to ensure both classes were equally represented, resulting in a total of 88,448 entries. While downsampling can lead to loss a potentially valuable data, it proved to be a straightforward and effective method in that case. Despite the reduction in the number of records, the dataset retained a sufficient volume of data to support robust model performance.

11.3. Managing Outliers

To address the outliers in the dataset, the variables ***lead_time, stays_in_weekend_nights, stay_in_week_nights***, and ***average_daily_rate*** were treated by replacing outliers values with the median of each respective variable. New columns, labeled **<variable>_outlier_flag**, were created to indicate whether a value was original (0) or modified (1).

Following this treatment, outliers remain in ***lead_time*** and ***average_daily_rate***. given their minimal proportion, 0.37% and 1.61% respectively, these outliers were retained as they are.



For the other variables exhibiting outliers, the decision was made to convert them from numerical to binary format. This approach was chosen to facilitate a more straightforward analysis, and it will be explained in section 12. Feature Engineering. With the new variables **<variable>_outlier_flag**, the dataset resulted in 88,448 entries and 37 variables.

11.4. Skew - Transformations

Skewness is a measure of the asymmetry of a data distribution. Skewness value close to zero indicates a relatively symmetric distribution and values greater than +1 or less than -1

signify strong skewness. For those four variables ***lead_time, stays_in_weekend_nights, stay_in_week_nights***, and ***average_daily_rate,*** the skewness was checked, and it was found:

lead_time	0.941946
stays_in_weekend_nights	0.725955
stays_in_week_nights	0.592845
average_daily_rate	0.398724

The variable ***lead_time*** exhibited the highest skewness at 0.94, indicating a moderate right skew. Similarly, ***stays_in_weekend_nights*** and ***stays_in_week_nights*** showed moderate right skewness with values of 0.73 and 0.59, respectively. The ***average_daily_rate*** variable had the lowest skewness at 0.40, suggesting a less pronounced skew. Since they do not have a strong skewness (more than 1), they did not go through a transformation process such as log base 2.

11.5. Censored records

In the context of number of guests, it is observed that 0.31% of the records in the dataset have a value of zero for ***adults*** column. Given that this percentage constitutes a minimal fraction of the overall dataset, the decision has been made to exclude these records rather than attempting imputation. After proceeding with the mention, the dataset resulted in 88,165 entries and 37 variables.

11.6. Excluded columns

The exclusion of these columns streamlines the dataset and focuses the analysis on relevant aspects, thereby enhancing the clarity and relevance of the data for subsequent modeling and insights:

- **reservation_status:** was excluded since it is a redundant variable including the same categories of the target variable.
- ***arrival_date_year, arrival_date_mont, arrival_date_day_of_month, arrival_date_week_number***, and ***arrival_date***: were excluded because are not part of the objective of the model; however, it was created a higher-level feature ***season*** that will be explained in the section 12.1 New variables.

12.0. Feature Engineering

12.1. New variables

- ***length_of_stay:*** this variable has been introduced to the dataset to accurately represent the duration of each booking in days. The column is derived by aggregation the number of nights stays over both weekends (***stays_in_weekend_nights*** column) and weekdays (***stay_in_week_nights*** column). It was identified that 235 records had a zero value for ***length_of_stay***. This anomaly is attributed to customers who, despite staying for less than one full night, are charged for at least one night according to hotel policy. To maintain consistency and reflect the nature of these bookings, these zero values have been adjusted to one night.
- ***Season:*** this column has been introduced to categorize the bookings according to the season of the year, based on the ***arrival_date_month*** column. This modification allows for a more significant analysis of seasonal trends in booking patterns.
- ***dependants:*** has been introduced to denote whether a booking includes any children or babies, thereby consolidating the information previously captured by the separate ***children*** and ***babies*** columns. Consequently, the original ***children*** and ***babies*** columns have been removed from the dataset.

- ***weekend_nights:*** set to 1 if the booking includes one or more weekend nights, and 0 otherwise. Derived from the original variable ***stays_in_weekend_nights***.
- ***prior_cancellations:*** set to 1 of the customer has had experienced one or more previous cancellations, and 0 if no prior cancellations have occurred. Derived from the original variable ***previous_cancellations***.
- ***previous_sucessful_bookings.*** Set to 1 if the customer has had at least one successful booking in the past, and 0 if not. Derived from the original variable ***previous_bookings_not_canceled.***
- ***Bookings_modifications:*** set to 1 if the booking has undergone one or more modifications, and 0 if no changes were made. Derived from the original variable ***booking_changes.***
- ***waiting_list:*** set to 1 if the booking was on the waiting list oner or more days, and 0 otherwise. Derived from the original variable ***days_in_waiting_list***.
- ***requires_parkings:*** set to 1 if the customer requires one or more parking spaces, and 0 if no parking is needed. Derived from the original variable ***required_car_parking_spaces***.
- ***booked_via_agent:*** set to 1 if the booking was made through an agent, and 0 it was made directly by the customer. Derived from the original column ***agent.***
- ***booked_via_company:*** set to 1 if the booking was made through a company, and 0 otherwise. Derived from the original variable ***company.***
- ***room_type_modified:*** has been introduced to capture discrepancies between the reserved and assigned room; where 1 indicates if there was a difference between the room types reserved and assigned, and 0 if they matched. Derived from the original columns ***reserved_room_type*** and ***assigned_room_type***.

- ***region:*** has been introduced to categorize bookings into broader geographic areas: Northern Europe, Southern Europe, Eastern Europe, Western Europe, and Other Region. This adjustment simplifies the dataset by reducing the number of categorical variables from 177 countries to 5 regions, thereby improving the efficiency of analysis. Derived from the column ***country***.

The original columns from which these binary variables were derived have been removed to streamline the dataset and enhance its focus on these consolidated indicators.

12.2. Binary indicators

To enhance the dataset's analytical value, flag columns have been created for ***adults*** and ***total_of_special_requests***:

- ***adults_flag***: set to 1 if the booking includes two or more adults, and 0 otherwise. This threshold was determined based on the distribution of the ***adults*** column, where the median and quartiles indicate the most bookings involve two adults.
- ***Total_of_special_request_flag***: set to 1 if the booking includes two or more special requests, and 0 otherwise. This threshold was chosen based on the boxplot statistics for ***total_of_special_requests,*** where the interquartile range and the upper bound suggest that values above two special requests are less common but noteworthy.

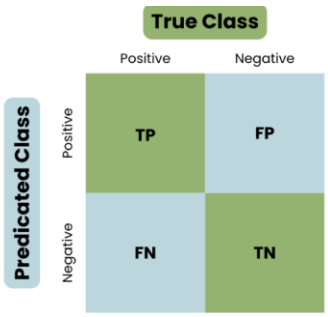
Model Exploration

13.0. Modeling Approach

Four models were developed for predicting the variable ***is_canceled***: a decision tree, an optimized decision tree utilizing grid search, a random forest, and logistic regression. To evaluate these models, the dataset was split into 60% training and 40% validation subsets. The

effectiveness of the model was assessed using accuracy and ROC AUC score. Accuracy measures the proportion of correctly classified instances out of the total number of instances, with higher values indicating better performance. On the other hand, ROC AUS assesses the model's ability to discriminate between classes, with a higher value reflecting a better fit and greater predictive power.

Additionally, confusion matrix was displayed for each model to identify the number of correctly and incorrectly instances classified. Ahmed (2023) indicates that confusion matrix is a tool that provides a tabular representation of the various prediction outcomes and results from a classification problem. The confusion matrix displays 4 important values: true positive, true negative, false positive, and false negative.



True Positive (TP): number of instances where the model correctly predicted as positive class.

True Negative (TN): number of instances where the model correctly predicted as negative class.

False Positive (FP): number of instances where the model incorrectly predicted as positive class.

False Negative (FN): number of instances where the model incorrectly predicted as negative class.

14.0. Model Technique #1: Decision Tree

The Decision Tree model is a supervised learning algorithm used for both classification and regression tasks. It works by recursively partitioning the dataset into subsets based on feature

values, aiming to create a tree-like structure of decisions that leads to a prediction for the target variable (skikit-learn, 2024).

In classification problems, such as predicting booking cancellations, the decision tree splits the data at each node based on the feature that best separates the classes, resulting in a tree that classifies instances into different categories.

For this analysis, the confusion matrix shows that the model correctly predicted 14,333 non-cancellations and 14,596 cancellations. However, there were 3,251 false positives (non-cancellations predicted as cancellations) and 3,086 false negatives (cancellations predicted as non-cancellations).

	Predicted Not Canceled	Predicted Canceled
Actual Not Canceled	14,333	3,251
Actual Canceled	3,086	14,596

The model achieved a good accuracy on the validation set of 82.03%, reflecting its performance on unseen data and suggesting good generalization to new instances. The ROC AUC score was 0.8226 indicating that the model has a string ability to discriminate between classes, it means there is an 82.26% chance that the model will be able to distinguish between positive class and negative class.

14.1. Optimized Decision Tree

The Optimized Decision Tree model was developed using Grid Search to identify the best hyperparameters, leading to enhanced performance. the optimal set of hyperparameters were:

- **Maximum depth:** 25
- **Minimum Impurity:** 0.00005

- **Minimum Samples per Split:** 0.001

These parameters were selected to improve the model's ability to generalize and prevent overfitting by controlling the tree's complexity and ensuring sufficient data at each split.

The confusion matrix reveals that the optimized model correctly predicted 14,768 non-cancellations and 14,575 cancellations. It made 2,816 false positive predictions (non-cancellations classified as cancellations) and 3,107 false negatives (cancellations classified as non-cancellations).

	Predicted Not Canceled	Predicted Canceled
Actual Not Canceled	14,768	2,816
Actual Canceled	3,107	14,575

The model achieved an accuracy of 83.20% on the validation set, indicating a strong performance in predicting booking cancellations on unseen data. The ROC AUC score of 0.9189 demonstrates excellent discriminative power, reflecting the model's ability to effectively distinguish between cancellation and non-cancellation instances.

15.0. Model Technique #2: Random Forest

According to Whitfield (2024), the Random Forest model is an ensemble learning method that constructs multiple decision trees (para. 2). This approach enhances prediction accuracy and robustness by aggregating the results from a diverse set of decision trees, which collectively mitigate the risk of overfitting and capture a broader range of patterns in the data.

For this analysis, the Random Forest model achieved an accuracy of 85.58% on the validation set, indicating a strong performance in classifying bookings accurately. The model correctly predicted 15,111 non-cancellations and 15,070 cancellations. It made 2,473 false

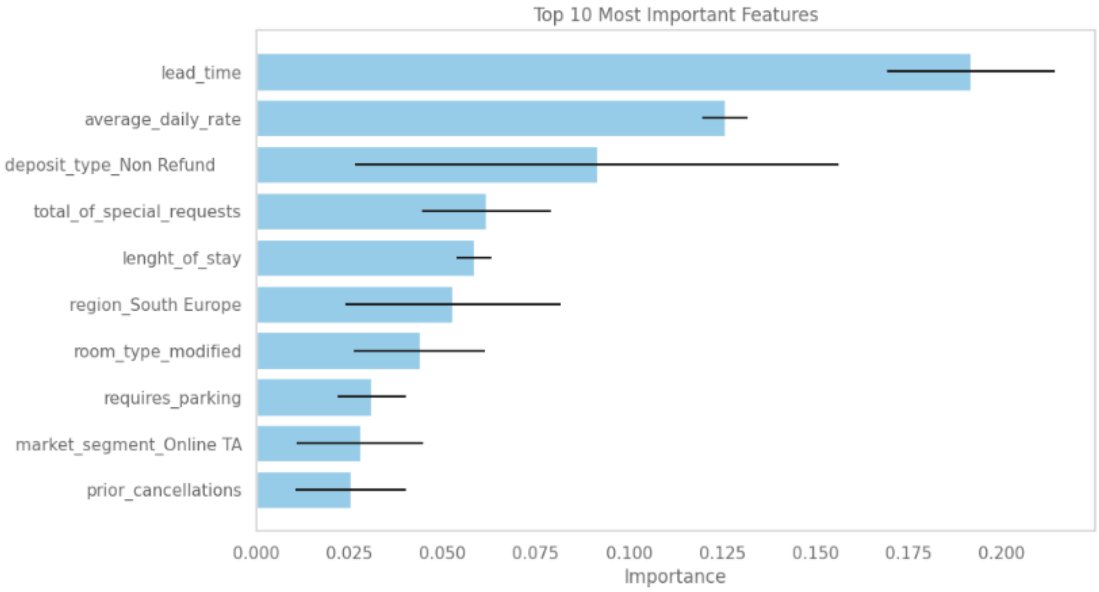
positive predictions (non-cancellations classified as cancellations) and 2.612 false negatives (cancellations classified as non-cancellations).

	Predicted Not Canceled	Predicted Canceled
Actual Not Canceled	15,111	2,473
Actual Canceled	2,612	15,070

The model achieved a ROC AUC score of 0.9389 demonstrating excellent discriminative ability, effectively distinguishing between bookings that will be canceled and those that will not.

A notable advantage of the Random Forest algorithm is its straightforward ability to assess the relative importance of each feature in making predictions (Whitfield, 2024, para. 17). However, the feature importance does not indicate whether a higher or lower value of a feature increases or decreases the likelihood of cancellation.

The following chart shows the 10 most important variables that influence the likelihood of hotel booking cancellation.



The feature importance reveals that ***lead_time*** and ***average_daily_price*** are the most influential features in predicting bookings cancellations, followed by ***deposit_type_Non Refund*** and ***total_of_special_requests***.

16.0. Model Technique #3: Logistic Regression

The Logistic Regression model, according to Kanade (2022), is a supervised machine learning algorithm that is used for binary classification tasks, where the goal is to predict the probability of an outcome based on one or more predictor variables (para. 2). It estimates the probability of a binary response based on the logistic function, which outputs values between 0 and 1. In this analysis, Logistic Regression was used to predict the likelihood of booking cancellations.

The confusion matrix shows that the model correctly predicted 14,456 non-cancellations and 13,717 cancellations. It incorrectly predicted 3,128 non-cancellations as cancellations and 3,965 cancellations as non-cancellations. The model achieved an accuracy of 79.89% on the validation set, demonstrating its effectiveness in correctly classifying bookings.

	Predicted Not Canceled	Predicted Canceled
Actual Not Canceled	14,456	3,128
Actual Canceled	3,965	13,717

The ROC AUC score of 0.8919 indicates that the model has strong discriminatory power distinguishing between cancellations and non-cancellations.

The Logistic Regression provides the coefficients and odds ratios, those values help to understand the influence of each feature on the likelihood of booking cancellations. A positive coefficient indicates that as the predictor increases, the odds of the outcome occurring increase. Meanwhile, odds ratios represent the multiplicative change in the odds of the outcome for a

once-unit increase in the predictor variable. The next table summarizing the coefficients and odds ratios for 10 features used in the model:

	coef	odds
deposit_type_Non Refund	1.729239	5.636364
prior_cancellations	1.021151	2.776388
region_South Europe	0.617020	1.853397
lead_time	0.554866	1.741707
customer_type_Transient	0.447682	1.564681
market_segment_Online TA	0.381181	1.464012
average_daily_rate	0.257049	1.293109
customer_type_Transient-Party	0.183098	1.200932
lead_time_outlier_flag	0.179367	1.196460
lenght_of_stay	0.152470	1.164708

For the variable ***deposit_type_Non Refund,*** having a non-refundable deposit increases the odds of cancellation by approximately 5.63 times compared to other deposit types. In the case of ***prior_cancellations*** variable, a booking with cancellation history raises the odds of a new cancellation by about 177.64%. For customers from South Region have approximately 85.34% higher odds of cancellation compared to other regions.

17.0. Model Comparison

To evaluate the performance of various predictive models for predicting booking cancellations, four approaches were assessed: Decision Tree, Optimized Decision Tree, Random Forest, and Logistic Regression. The metrics used to determine the best model were ROC AUC and accuracy. The next table summarizes the metrics for each model, where showcase the accuracy for training and validation dataset, and the ROC AUC score.

Model	Training Accuracy	Validation Accuracy	ROC AUC score
Decision Tree	0.9952	0.8203	0.2826
Optimized Decision Tree	**0.8674**	**0.8320**	**0.9189**

Random Forest	**0.9952**	**0.8552**	**0.9387**
Logistic Regression	0.8050	0.7989	0.8917
Backward Regression	0.8051	0.8000	0.8917

The Random Forest model reached the higher accurate and ROC AUC score, 85.52% and 0.9387 respectively. Suggesting that it delivers superior performance and robustness; however, its high complexity and computational demands may be a drawback. The extensive number of tress and interactions can lead to longer training times and increased resource usage, potentially complicating deployment and interpretation.

The second-best model, with higher accuracy and ROC AUC score, is the Optimized Decision Tree. By fine-tuning hyperparameters through grid search, the optimized decision tree demonstrates an improvement in validation accuracy to 83.20% and achieves a robust ROC AUC of 0.9189. This model strikes a balance between performance and complexity, it provides a good comprise between accuracy, interpretability, and computational efficiency.

Model Recommendation

18.0 Model Selection

The model selected for predicting whether a booking will be canceled or not is the Optimized Decision Tree, with an accuracy of 83.20% and a ROC AUC score of 0.9189.

Although the Random Forest model demonstrated slightly superior performance with a higher ROC ACU and validation accuracy, 0.9887 and 85.52% respectively, the Optimized Decision Tree was selected for deployment.

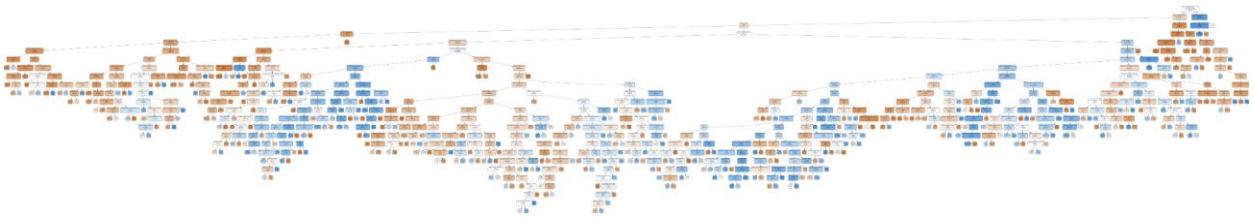
The Decision Tree model provides several advantages that align with practical needs. Firstly, it offers high interpretability due to its straightforward representation of decision rules, which can be easily understood and communicated to stakeholders.

Secondly, the transparency in decision-making allows for the identification of the most significant features affecting cancellation likelihood, enhancing the model's usability in strategic planning and decision support.

Additionally, the Optimized Decision Tree has a simpler structure compared to Random Forest, which involves an ensemble of multiple trees and can be computationally intensive. The reduced complexity of the Decision Tree translates to lower maintenance requirements, making it easier to update and adapt over time.

While the Random Forest's complexity allows for robust performance, the Optimized Decision Tree's balance of performance and simplicity makes more practical choice for operational use. Thus, the Optimized Decision Tree emerges as a strong candidate offering a favorable balance between accuracy, interpretability, and operational efficiency.

The following image is the Optimized Decision Tree with maximum depth of 25, in section 20. Model Sensitivity to Key Drivers we are going to explore the most important nodes.



19.0 Model Theory

Decision Trees are a type of non-parametric supervised learning algorithm used for classification and regressions tasks (IBM, 2024). They work by recursively splitting the dataset into subsets based on feature values, by creating a tree-like structure where each node represents a decision rule and each leaf node represents a predicted outcome. A decision tree follows the following structure:

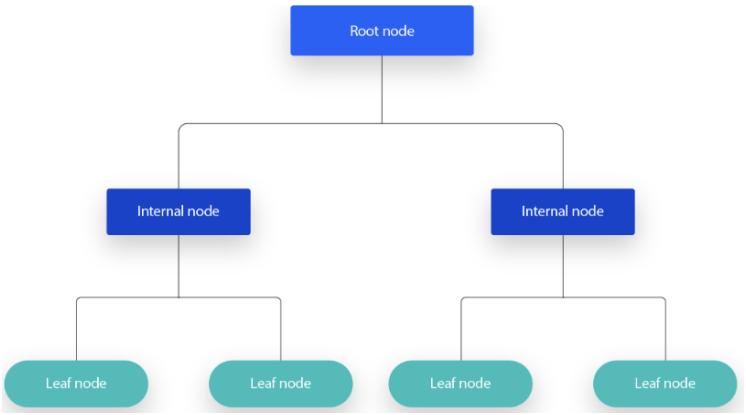


Figure 1. IBM Decision Tree Diagram. https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

The diagram is the visual representation of a decision tree, where it starts with the rood node that represents the most important variable. The outgoing branches are the internal nodes (or decision nodes) that represent the points where the dataset is split based on a specific condition. The leaf nodes represent all the possible outcomes, in this case represents if the booking will be canceled or not based on certain conditions.

Grid Search is a hyperparameter optimization technique used to enhance the performance of the model (Shah, R. 2024. Para. 3). It involves systematically exploring a predefined set of hyperparameters, such as ‘max_depth’, ‘min_samples_split’ and ‘min_impurity_decrease’, to find the combination that yields the best model performance. The hyperparameters used for this analysis were:

- **max_depth:** controls the maximum number of levels or layers in the decision tree. This hyperparameter was set to 25 means the tree can grow up to 25 levels deep.
- **min_impurity_decrease:** this parameter determines the minimum reduction in impurity required to split a node in the tree. Impurity measures how mixed the class labels are within a node, the default criterion for impurity is Gini. Singh (2024) explains that Gini impurity evaluates the probability of misclassification when a random data point is

assigned a class label according to the class distribution within a specific node (para. 2), where 0 indicates a perfect pure node and 0.5 indicates the maximum impurity. The value using Grid Search was set 0.00005, this value is small enough to allow for detailed splits but ensures that only significant reductions in impurity lead to additional nodes.

- **min_samples_split:** specifies the minimum number of samples a node must have before it can be split into further nodes. In this case a small threshold was set, 0.001, for example in a dataset with 50,000 samples, 0.001 would correspond to 50 samples.

Grid Search helps in identifying the optimal settings that balance model complexity and performance. This approach aims to improve the model's ability to generalize to new data, preventing overfitting and enhancing accuracy.

19.1 Model Assumptions and Limitations

The decision tree model, particularly in its optimized form, provides a robust approach to predictive modeling but comes with certain assumptions and limitations that must be considered.

- **Homogeneity of data:** decision tree presumes that data within each lead node is relatively homogeneous, with the assumption that the splitting criteria will create nodes that are distinct in terms of the target variable.
- **Overfitting:** decision trees are prone to overfitting, especially with a high depth. This means they can become too complex, capturing noise in the training data rather than underlying patterns.
- **Bias towards features with more levels:** decision trees can be biased towards features with more levels of categories.
- **Counter-intuitive variables:** occasionally, variables may exhibit unexpected effects, which can be counter-intuitive and suggest issues with data quality or feature

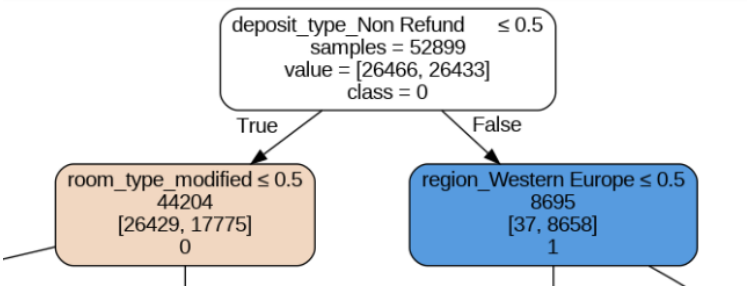
engineering. This requires further investigation to ensure the model's validity and accuracy.

- **Accuracy of the data:** the model assumes that the data is accurate and is a valid representation of the population.

20.0 Model Sensitivity to Key Drivers

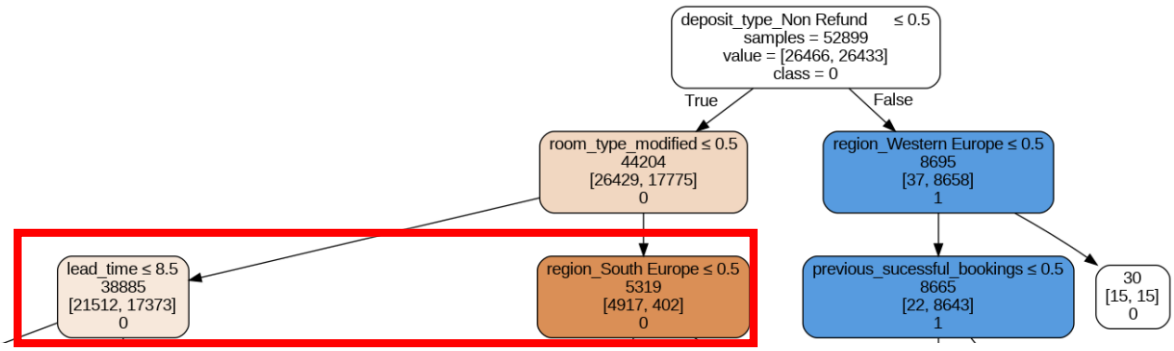
The sensitivity of the Decision Tree model to key drivers is a critical aspect of its performance. The model structure allows to focus on the most influential variables in predicting booking cancellations. By prioritizing these key drivers, the Decision Tree provides valuable insights into how changes in specific features impact the likelihood of cancellations.

Non-Refundable Deposit



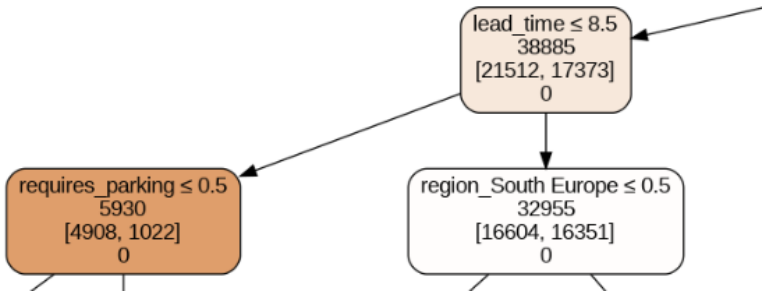
The primary factor that influences cancellation rates is bookings with **non-refundable deposits** showing a strong split in cancellation behavior. Surprisingly, non-refundable bookings have a higher chance of cancellation (99.6%). Despite that non-refundable bookings typically come with stricter terms and conditions, including the inability to cancel or modify the reservation without incurring significant penalties, the data suggests that customers are still more likely to cancel these bookings. This counter-intuitive finding suggests that customers commit to Non-Refundable rates with the intention of securing a lower price but later change their plans due to unforeseen circumstances.

Difference in room reserved vs. room assigned



The data suggests that when customers receive the exact room they reserved, they are more likely to cancel. Exact room assignments experience a significantly higher cancellation rate, 44.6% (17,373 records out of 38,885), compared to those with different room assignments, 7.55% (402 out of 5,319 records). This might imply that customers who get exactly what they requested are more likely to have their plans change, perhaps due to the certainty provided by receiving their preferred room. Alternatively, it could reflect a higher level of satisfaction with the booking experience when customers receive a different room, possibly due to better room quality or additional amenities.

Lead time

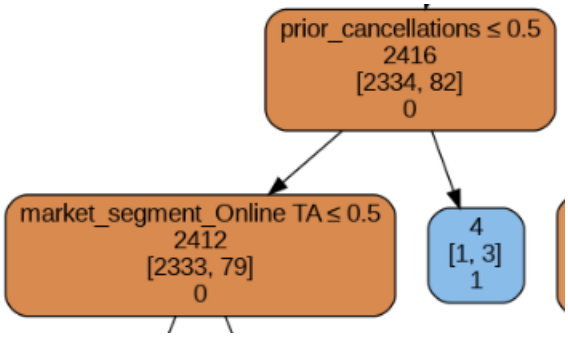


The data suggests that longer lead times present a higher cancellation rate on bookings. In this particular node, bookings with lead times less than 8.5 days have a likelihood of cancellation of 17.23% compared to those with greater than 8.5 days with a 49.61% chance of cancellation. This finding suggests that customers booking with a short lead time are likely making more

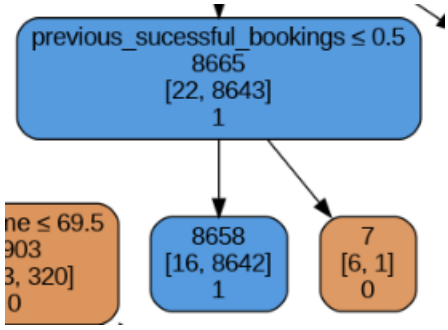
immediate plans, to possibly due to urgent needs or spontaneous decisions. These bookings are less prone to cancellation since the decision to book is closely tied to a specific and imminent need.

On the other hand, customers booking well in advance may experience changes in their plans over time, leading to a higher cancellation rate. This could be due to evolving travel plans, changes in circumstances, or uncertainty about future travel.

Booking history



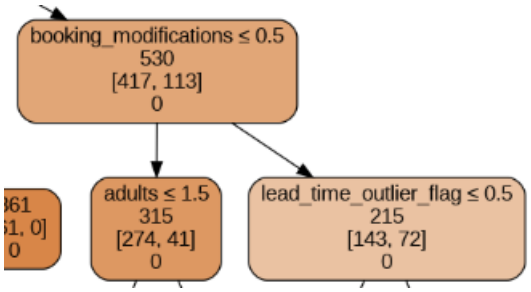
Another key driver behind cancellation is previous cancellation. The model found that customers with a history of cancellations represent a higher risk of future cancellations. The likelihood of cancellation is 75% when customers have previous cancellations; on the contrary, the likelihood of cancellation for customers without any cancellation history is 3.2%.



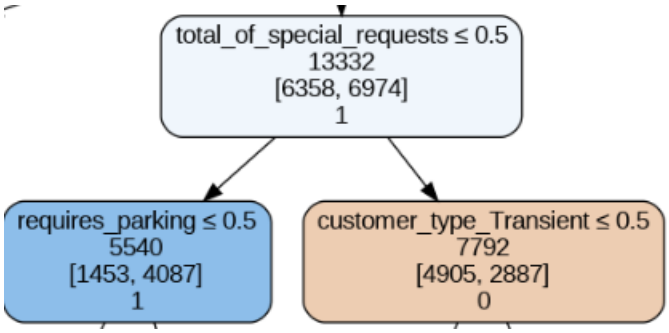
The other key factor in regards with customer's booking history is previous successful bookings. The data indicates that first time guests with any previous successful booking have an exceptionally high chance of cancelling (99.8%), compared to a much lower chance (14.3%) for

guests with prior successful bookings. This highlights a significant difference in behavior between first-time and repeat customers.

Booking modifications



The insights that customers who modify their booking are more likely to cancel (33%) compared to those who do not modify their booking (13%) suggests a strong link between booking changes and higher cancellations risks. Modifying a booking often indicates a change in the customer's plans or circumstances including change of room, additional guests, change on dates, among others. The process of modifying a booking can sometimes be complex or frustrating, which may lead customers to ultimately cancel their reservations, especially if they encounter difficulties or perceive the process as inconvenient.



The model suggests that bookings with special requests have lover chance of cancellations compared to bookings without special requests, suggesting that customers who make specific requests are more likely to follow through with their reservations. In that node, the behavior indicates that there is a 37% chance of cancellation when the reservation has any

special requests, versus 73% when the reservation does not have special requests. The reason behind this behavior is because special requests often reflect a customer's specific needs or preferences, which can increase their investment in the booking.

Conclusion and Recommendations

21.0. Impacts on Business Problem

By examining the characteristics and patterns associated with high cancellation rates, the business can gain deeper understanding of how these factors influence overall performance. This section explores the implications of these drivers, highlighting their impact on revenue, customer satisfaction, and operational efficiency.

Non-Refundable Deposit

Understanding that Non-Refundable bookings are highly prone to cancellations can help to refine pricing strategies. If cancellations are frequent despite the non-refundable nature, this could suggest that customers are driven by price but may face changing circumstances. Adjusting the pricing model to include more flexible options or offering incentives for more committed bookings could be beneficial.

High cancellation rates in Non-Refundable bookings could also highlight a disconnect between customer expectations or communication. By addressing this through enhanced customer service or clearer booking terms, the business can manage customer expectations better and potentially reduce cancellations.

Difference in room reserved vs. room assigned

Knowing that the likelihood of cancellation is higher when the reserved and assigned rooms match, this insight reveals an opportunity to refine inventory and room assignments strategies. This finding might indicate a need for better management of room allocations and an

understanding of customer preferences. This could help in optimizing room assignments to balance customer expectations and operation efficiency.

If customers are more likely to cancel when receive the exact room they booked, improving communication around room assignments and the benefits of flexibility might reduce cancellations. Ensuring that customers understand the value of the assigned room could help mitigate the higher cancellation rates associated with matching room assignments.

Lead time

Higher cancellation rates for longer lead times can lead to significant revenue losses, as early reservations might be less likely to materialize. This impacts overall profitability and operation efficiency resulting in overbooking or underutilization of resources. To mitigate the high cancellation rates for bookings made with long lead times, the business might consider offering incentives or flexible terms that encourage commitment,

Booking history

Understanding the high likelihood among previous offenders allows for proactive engagement. It is important to collect feedback from customers with a history of cancellations to understand the reasons behind their behaviors and address any issues that may be causing dissatisfaction.

The extremely high cancellation rate among first-time guests with a successful booking history suggests a lack of commitment or a possible pattern of bookings with the intention to cancel. This could be due to factors such as uncertainty, lower stakes, or the use of the booking as a placeholder.

Booking modifications

The impact of booking modifications on a business can be substantial, leading to increased uncertainty in revenue and operational challenges. Customers who frequently modify their bookings are more likely to cancel, which disrupts room availability and can result in lost revenue.

Special requests from customers can have a positive impact on business by increasing customer satisfaction and reducing cancellation rates. Accommodating special requests enhances the customer experience, leading to improved loyalty and repeat business.

22.0. Recommended Next Steps

22.1. Model Recommendations

To enhance the predictive accuracy and reliability of the model, it is essential to first investigate the quality of the data. Certain variables in the current dataset exhibit counter-intuitive behavior, which may indicate underlying issues with the data quality or feature engineering. Addressing these anomalies is crucial for ensuring that the model's insights are valid and actionable.

Additionally, exploring customer segmentation through techniques such as K-means or Hierarchical Clustering could provide deeper insights into customer behavior patterns. By segmenting customers based on their booking habits and characteristics, strategies can be tailored more effectively.

22.2. Business Recommendations

Based on the insights gained from analyzing booking cancellations, it is essential to develop targeted strategies to mitigate their impact and enhance overall business performance. The recommended next steps focus on actionable interventions that address identified issues and

capitalize on opportunities for improvement. By implementing these strategies, the business can refine its approach to pricing, room management, booking policies, and customer engagement.

Customer Communication and Service

- **Clarify Booking Terms:** Improve customers service by clearly outlining the terms and conditions of Non-Refundable bookings. Ensuring customers fully understand these terms can help manage their expectations and reduce the likelihood of cancellations.
- **Engage Proactively:** For customers who book far in advance, implement regular updates and personalized offers. This engagement can help maintain customer interest and commitment to their reservations.
- **Support Booking Changes:** Provide additional support and guidance for customers who make changes to their bookings. By managing their expectations and assisting them throughout the modification process, the risk of cancellations can be mitigated.
- **Improve Request Handling:** Enhance the process for handling and fulfilling special requests. Efficiently managing these requests can lead to higher customer satisfaction and more stable bookings, reducing the likelihood of cancellations.

Personalized Incentives

- **Reward Loyalty:** Develop incentive programs or rewards for repeat guests to encourage continued positive booking behavior. Recognizing and valuing loyal customers can foster greater commitment and reduce cancellations.
- **Advance Booking Discounts:** Offer discounts or special rates for customers who book well in advance. This can decrease the likelihood of cancellations with longer lead times.

- **Incentivize Special Requests:** Provide additional incentives for bookings with special requests, such as complimentary upgrades or services. This approach can enhance the customer experience and promote stability in bookings.

Marketing campaigns

- **Targeted Campaigns:** Launch marketing campaigns focused on regions with high cancellations rates. By addressing specific regional issues, targeted campaigns can help reduce cancellations and increase booking stability.
- **Promote Reservation Benefits:** Emphasize the advantages of maintaining the original reservation and highlight flexible rebooking options. Effective promotion of these benefits can encourage customers to honor their bookings and reduce cancellation rates.

References

Ahmed, N. November 2023. *What is a confusion matrix in machine learning? The model evaluation tool explained.* Datacamp. https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning

Antonio, N. Almeida, A. Nunes, L. February 2019. *Hotel booking demand datasets*. Science Direct. https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib5

Hollander, J. May 31, 2023. *Hotel Data Analytics: What You Need to Know About Big Dara in Hospitality.* Hotel Tech Report. https://hoteltechreport.com/news/hotel-data-analytics

IBM. August 2024. *What is a decision tree?* IBM. https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,internal%20nodes%20and%20leaf%20nodes.

Kanade, V. April 8, 2022. *What is Logistic Regression? Equation, Assumptions, Types and Best Practices.* Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Scikit learn. August 2024. *Decision Trees*. Scikit learn. https://scikit-learn.org/stable/modules/tree.html

Shah. R. August 13, 2024. *Tune Hyperparameters with GridSearchCV.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/

SHR Group. June 21, 2023. *We need to talk about cancellations*. SHR Group. https://shrgroup.com/2023/06/21/we-need-to-talk-about-cancellations/

Singh, H. June 4, 2024. *GSplitting Decision Trees with Gini Impurity.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees-gini-impurity/

SiteMinder. February 22, 2023. *Hotel Booking Trends Report*. Hotel Business. https://hotelbusiness.com/siteminder-traveler-confidence-surges-post-pandemic/