HARO HOTEL

Haro Hotel Booking Cancellation

Haikal Mahfuzh Riza

February 2, 2026

LinkedIn Profile

Content Outline

HARO HOTEL

01
Executive Summary

Haro Hotel, Project Background, Analytical Approach

02
Analysis and Insights

Exploratory Data Analysis (EDA), Prediction Model, Dashboard Overview

03
Takeaway

Business Impact, Recommendation,

Final Business Summary

04
Appendix

Executive Summary

Problem	Insight	Action	Implication
Booking **cancellation rate increased** from 15% (2017) to **37%** (2018).	Cancellations are driven by **low-commitment bookings.**	Making **a tree-based classification prediction model** for upcoming reservations.	By increasing commitment only where risk is high, **Haro Hotel** can **reduce cancellations to 20%** without weakening its digital booking experience.
• Putting **~€4.02M** annual revenue at **risk**.	• Key drivers are early bookings, online, higher-priced stays, fewer special requests, and longer planned stays.	• Prediction model can **detect ~77%** of high-risk cancellations. • Implement targeted pre-arrival **engagement** and selective commitment.	• Target high-risk bookings using engagement and selective restrictions to recover up to **35% of at risk revenue.**

Haro Hotel

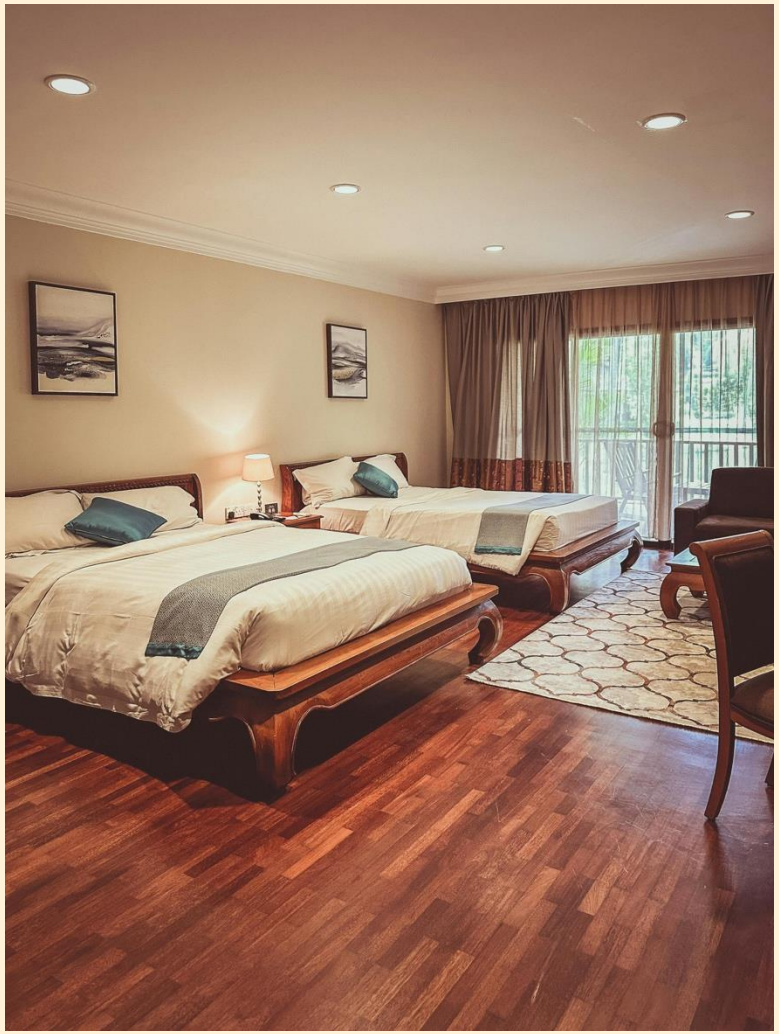
Haro Hotel is a leading futuristic hotel resort using Japanese style, focusing on **digitalization** of their room reservation process.

Digitalization booking drives growth but structurally **increases cancellation** risk through flexibility.

Data source: [Hotel Reservations Classification Dataset]

*) This analysis is based on a publicly available dataset and does not reflect real-world business insights. **Haro Hotel** is a fictional entity, and the results presented here are purely for educational purposes.

Project Background



Objective:
Reduce hotel **booking cancellations** from **37%** to **20%** in **6** months.

The project scopes are:

- Generating insights on which **key factors** driving cancellations.
- **Predicting** whether a hotel booking will be cancelled or not.
- Strategizing on how to **reduce** number of **cancellations.**

The analysis will **not** include logistics efficiency or supplier-side operations.

Analytical Approach

Diagnostic analysis on what factors affecting the cancellation in the past using python library statsmodel and **predictive analysis** on future reservations whether it will be cancelled or not, using classification Machine Learning (ML) library in **python**.






Business Understan-ding → Data Preparation → Data Cleaning → Exploratory Data Analysis → Data Validation & Prediction Model → Data Visualiza-tion

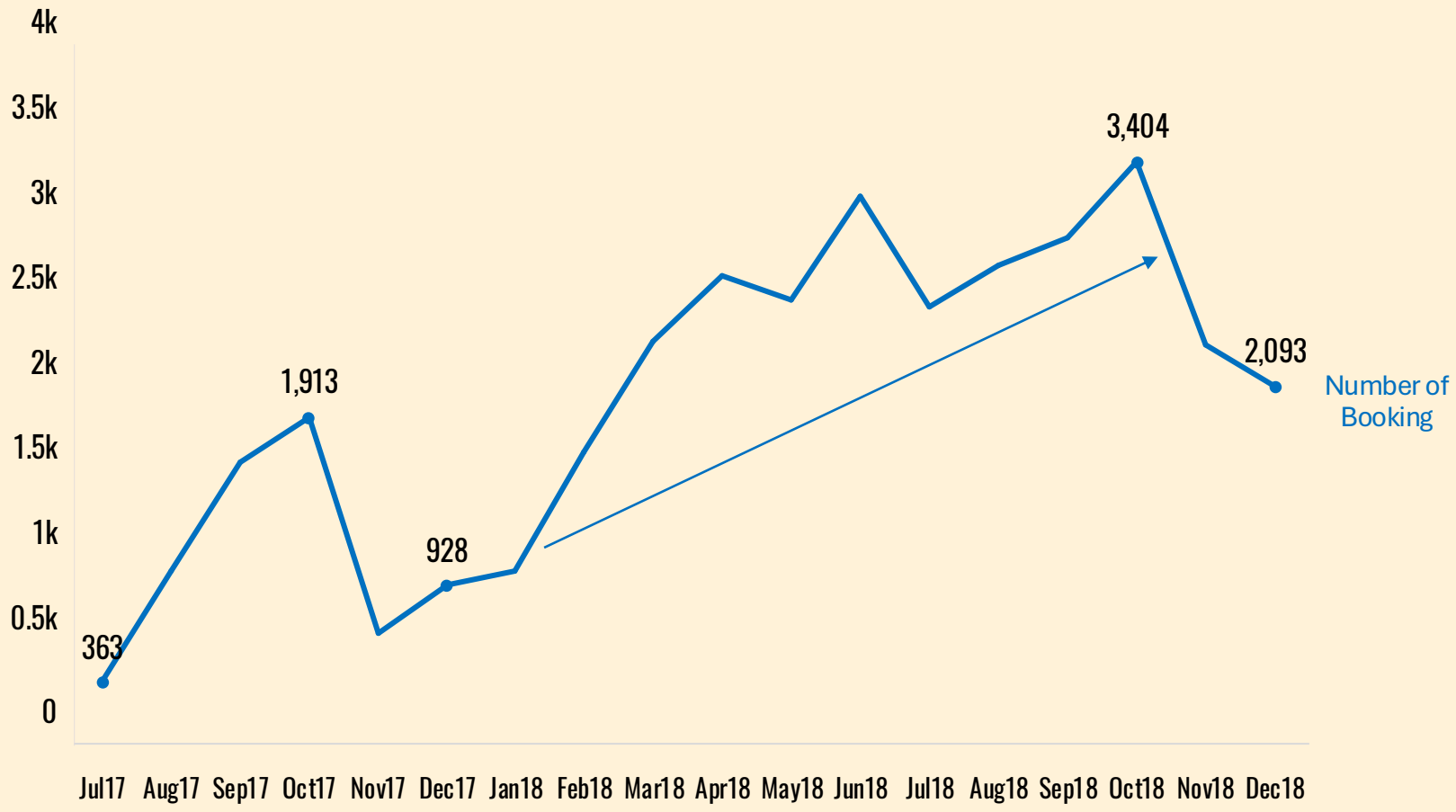
Analysis and Insights

[Python Google Collab Link]

The hotel revenue grew to **€5.67M** in 2018 and the number of booking keeps increasing throughout the year.

Total Revenue, Number of Booking

€5.67M
Total Revenue Net (2018)



HARO HOTEL

The hotel revenue grew to **€5.67M** in 2018 and the number of booking keeps increasing throughout the year. However, the cancellation rate was also rising by **21.9%,** putting revenue loss around **€4.02M**.

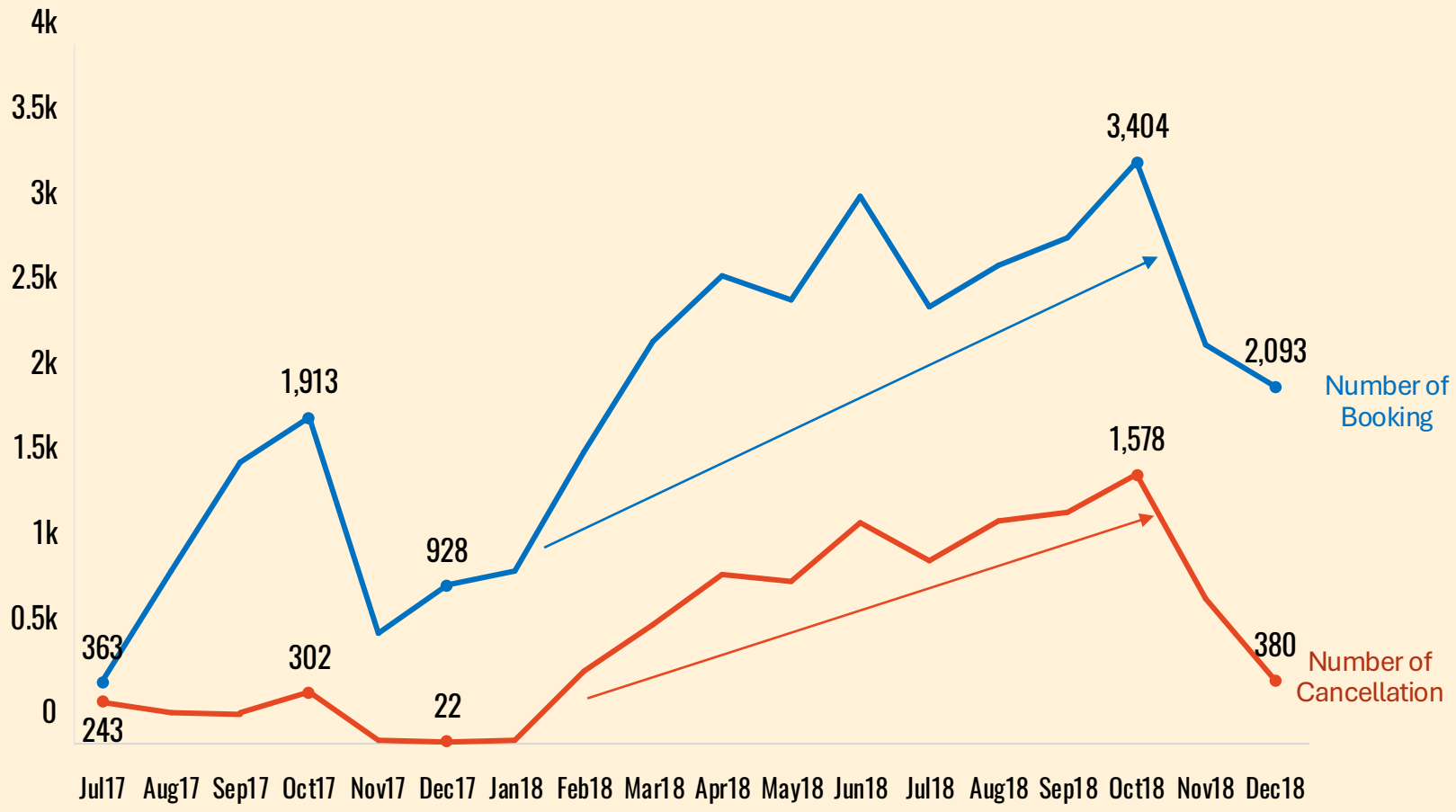
Total Revenue, Number of Booking, Revenue Loss, Number of Cancellation, Cancellation Rate

€5.67M
Total Revenue Net (2018)

€4.02M
Revenue Loss from Cancellation (2018) *

36.7%
Yearly Cancellation Rate (2018)

▲21.9%
Cancellation Rate Diff. (2018 vs. 2017)



HARO HOTEL

*) With assumption all of cancellations have not been replaced by another booking.

Online bookings show significantly higher cancellation rates due to lower switching costs, while the other segment types are more likely not to cancel their bookings.

Market Segment Type

	Total	Cancelled	Not Cancelled	Diff.
Online	64%	71%	60%	+11%
Offline	29%	27%	30%	-3%
Corporate	6%	2%	7%	-5%
Complementary	1%	0%	2%	-2%
Aviation	0%	0%	0%	0%

Accept H1: Bookings through online platform are more likely to be cancelled.

*) Hypothesis testing using Chi-Square with 95% Confidence Interval

HARO HOTEL

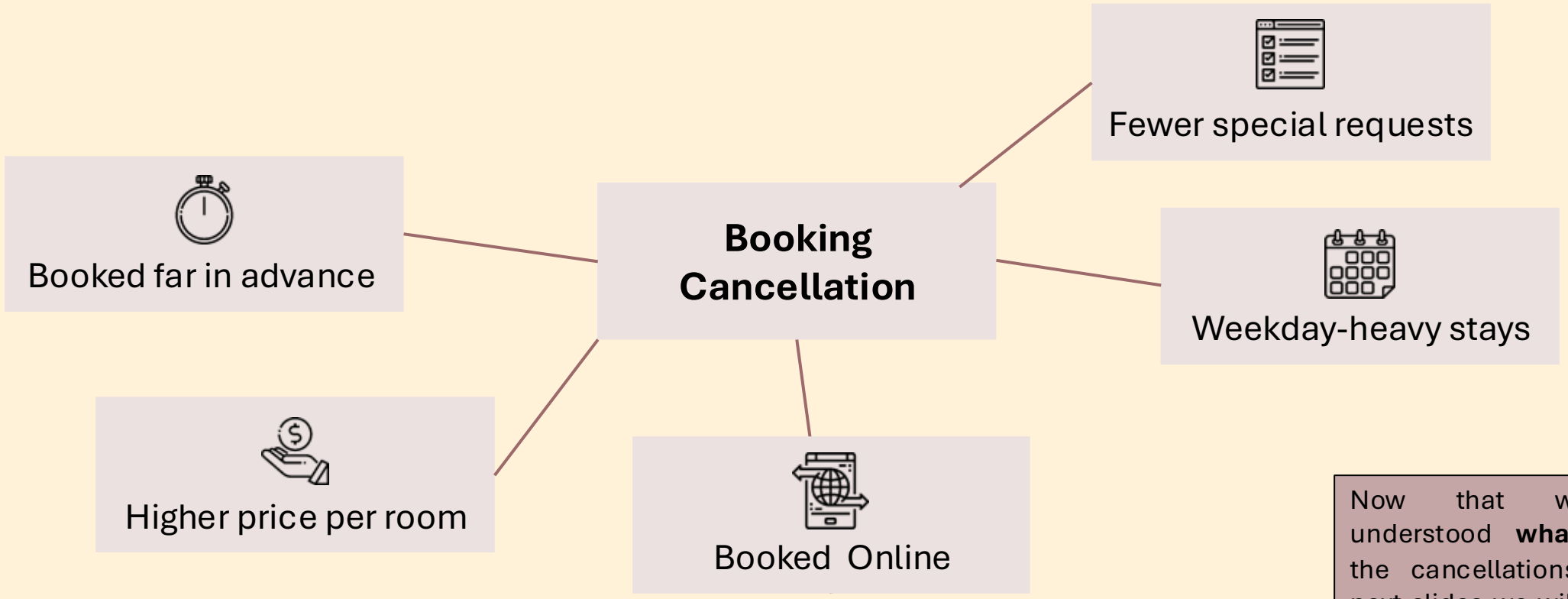
Cancellations consistently increase with indicators of flexibility and low commitment –longer lead time, higher price per room, number of week nights, and fewer number of special request.

	Total	Cancelled	Not Cancelled	Diff.	
Lead Time (time between booking and arrival time)	85.23	139.22	58.93	+80.29	Accept H1: Bookings made far in advance increase the likelihood of cancellation.
Avg. Price per Room	103.42	110.59	99.93	+10.66	Accept H1: Bookings with higher price average room are associated with a higher probability of booking cancellation.
No. Special Request	0.62	0.33	0.76	-0.43	Accept H1: Bookings with more special requests are less likely to be cancelled compared to those without special requests.
No. Week Nights	2.20	2.39	2.11	+0.28	Accept H1: Bookings scheduled with more weekday nights is more likely to be cancelled.

HARO HOTEL

*) Hypothesis testing using T-Test with 95% Confidence Interval

Key drivers of **Haro Hotel** room booking cancellation



Now that we have understood **what caused** the cancellations, on the next slides we will focus on **how to tackle** the issue of this surging cancellation using **prediction model** classification.

HARO HOTEL

Based on the historic key drivers, we generated a tree-based Classification prediction model, which can predict cancellation in the future. The model performs well in enabling early prioritization of high-risk bookings with sufficient accuracy to act.

Prediction Model: **Tree-based Classification prediction** model that balances recall and precision, enabling interpretability and clear rules for operational use.

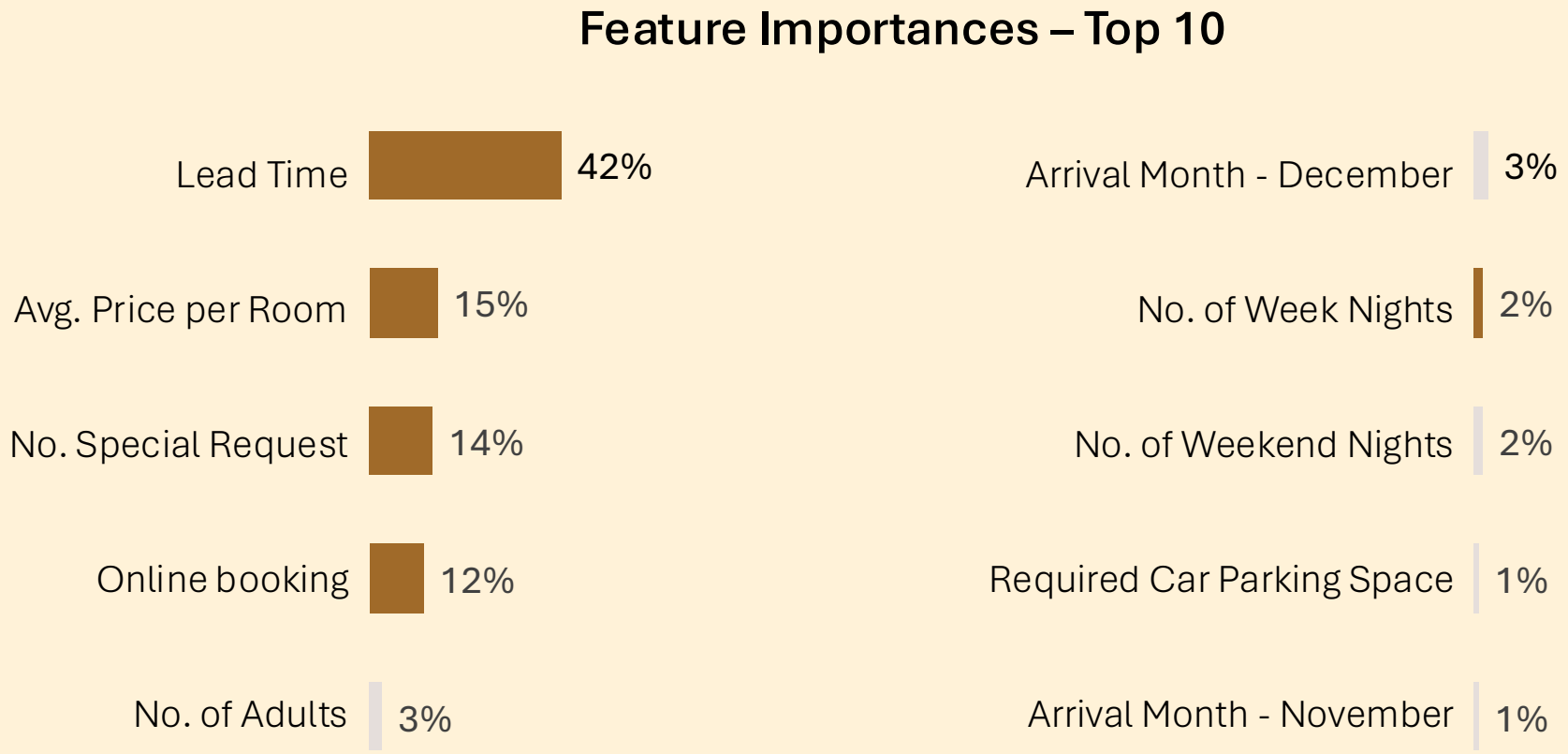
Model Performance			Benchmark
Model Accuracy:	88% →	Reliable differentiation between cancelled vs kept bookings	80-85%
PR-AUC Score:	91% →	Strong signal quality with low false alarms	55-60%
Recall:	77% →	Captures most cancellation cases early enough to intervene	60-70%

This model is sufficiently accurate to prioritize high-risk bookings and trigger targeted retention actions before cancellation occurs. This enables proactive intervention on ~3 out of 4 future cancellations.

HARO HOTEL

Booking timing, price, customer commitment, and online booking explain most cancellation risk – make up to 83% of all features. Feature importance from the prediction model are aligned with from key drivers (from the hypothesis testing).

Prediction Model

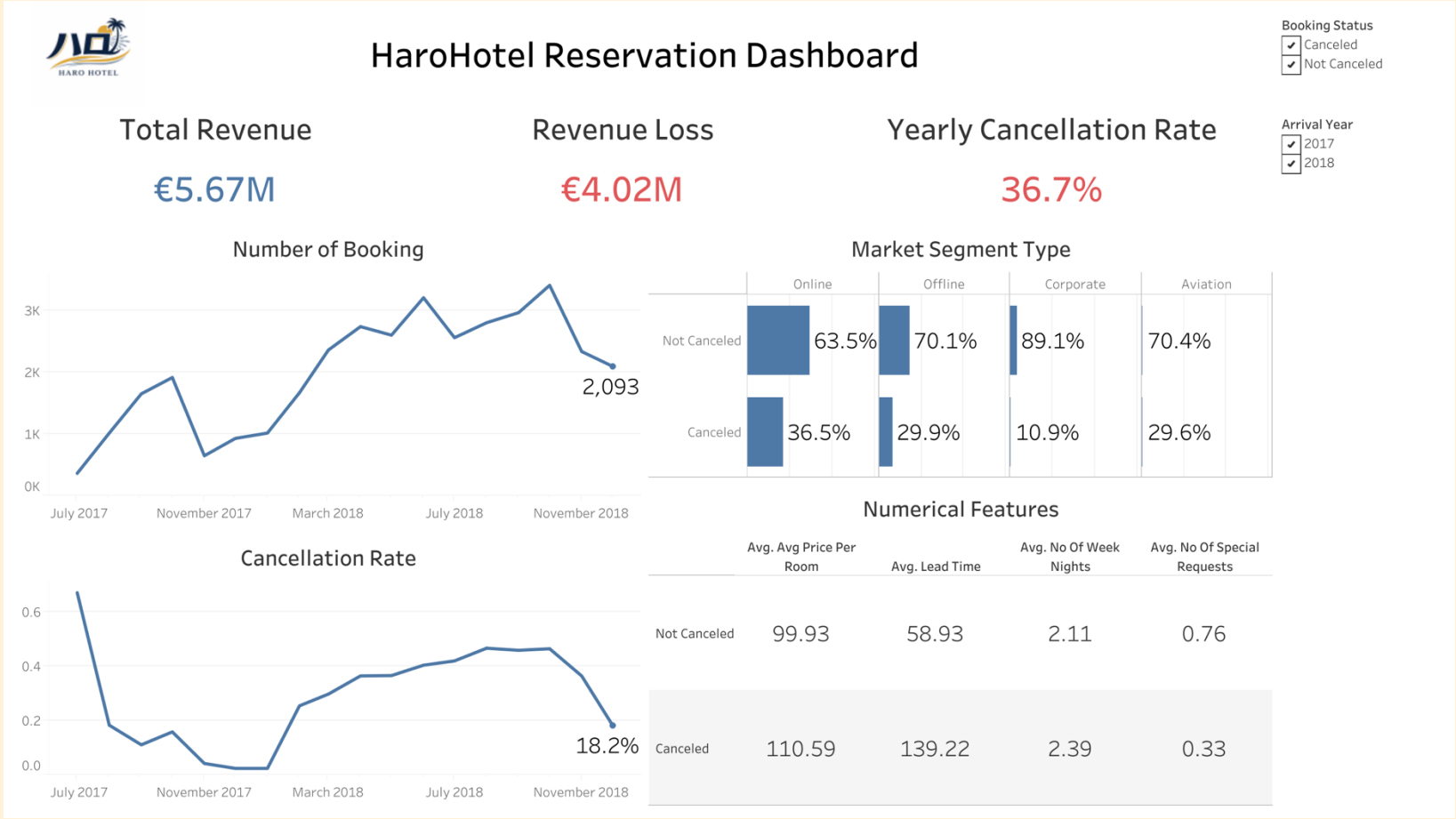


HARO HOTEL

Lead time alone explains ~40% of cancellation risk, followed by price and engagement signals — confirming where intervention should focus.

Operational dashboard visualization to continuously track cancellation bookings based on the features that affecting cancellation rate.

Dashboard Visualization using Tableau



[Tableau Public Link]

Summary

Business Impact

Active efforts on the controllable features

Detect **77%** of the cancellation

- If time between booking and arrival **> 60 days** → Implement **milestone-based confirmation** (e.g., reminder at 30/14/7 days) and apply **partial deposits**
- **Premium** room booking → Non-refundable discounts or tiered cancellation policies
- No special request → Activate **engagement campaign**, ask for room preference, meal upgrades, add on, etc.
- Booking via **online** booking → Apply **stricter** cancellation windows

Probable scenarios with respected of confidence level

Retain the detected cancellation → prevent revenue loss

Retention of Detected Cancellation	Cancellation Rate*	Impact of Yearly Revenue*
< 30%	28%-36%	Up to 17%
30-60%	21%-27%	18%-35%
> 60%	20% or lower	> 35%

*) Calculation on Appendix Slide 39

HARO HOTEL

Recommendation

Observation	Business Impact	Isolation	Prioritization	Recommendation
Longer lead time between booking date and arrival date significantly increase the likelihood of cancellation.	Long-lead bookings account for a large portion of cancellations due to higher planning **uncertainty**.	Lead time more than 60 days before arrival. 	**High** – long lead time can be managed with engagement which will enrich customer journey	Implement milestone-based **confirmation** (e.g., reminder at 30/14/7 days), partial **deposits**.
Higher average room prices are associated with increased cancellation probability.	Each cancellation results in a **larger revenue loss** per booking.	Avg. price per room above premium price threshold. 	**High** – higher price related with higher potential revenue loss	Introduce **non-refundable discounts**, or tiered cancellation policies for premium-priced rooms.
Bookings with few or **no special requests** are more likely to be cancelled.	Indicating **commitment** issue, but **scalable** to address.	No special requests submitted when booking. 	**Medium** – customer demand leads to customer loyalty	Trigger pre-arrival **engagement** campaigns (room preference, meal upgrades, add-ons) to increase commitment and reduce cancellation likelihood.
Online bookings exhibit a significantly higher cancellation rate, especially when combined with long lead times.	Represents the **largest share** of total cancellations and directly impacts revenue.	Online channel 	**Medium** – high impact, high volume	Apply stricter cancellation windows while offering flexible rescheduling options.

HARO HOTEL

Final Business Summary

Booking cancellations are a commitment management issue rather than a demand problem.
A relatively small set of booking profiles—typically online, booked far in advance, higher-priced, and lacking personalization—accounts for a disproportionate share of cancellations and revenue volatility. These patterns reflect low switching costs and low emotional commitment, not service or operational shortcomings.

This risk can be addressed through targeted, data-led intervention.
By identifying high-risk bookings early, the business can selectively increase commitment through tiered cancellation rules, partial deposits, milestone-based reconfirmation, and pre-arrival engagement, while preserving flexibility for low-risk bookings. This approach avoids blunt policy changes that could undermine digital growth.

Scenario analysis indicates material upside.
Retaining even 30–60% of predicted cancellations could reduce overall cancellation rates to approximately 21–27% and recover an estimated 18–35% of at-risk revenue, positioning cancellation management as a controllable strategic lever rather than a reactive cost.

HARO HOTEL

Thank you.

For any enquiries, please contact:

- Haikal Mahfuzh Riza
- haikalriza@gmail.com / +62812 9383 7040

Special thanks to:

- Kak Ibnu – Team 3 Amsterdam Team Leader
- Kak Lintang – Amsterdam Section Manager
- Team 3 teammates
- RevoU Amsterdam section members

for their support throughout this project.

Appendix

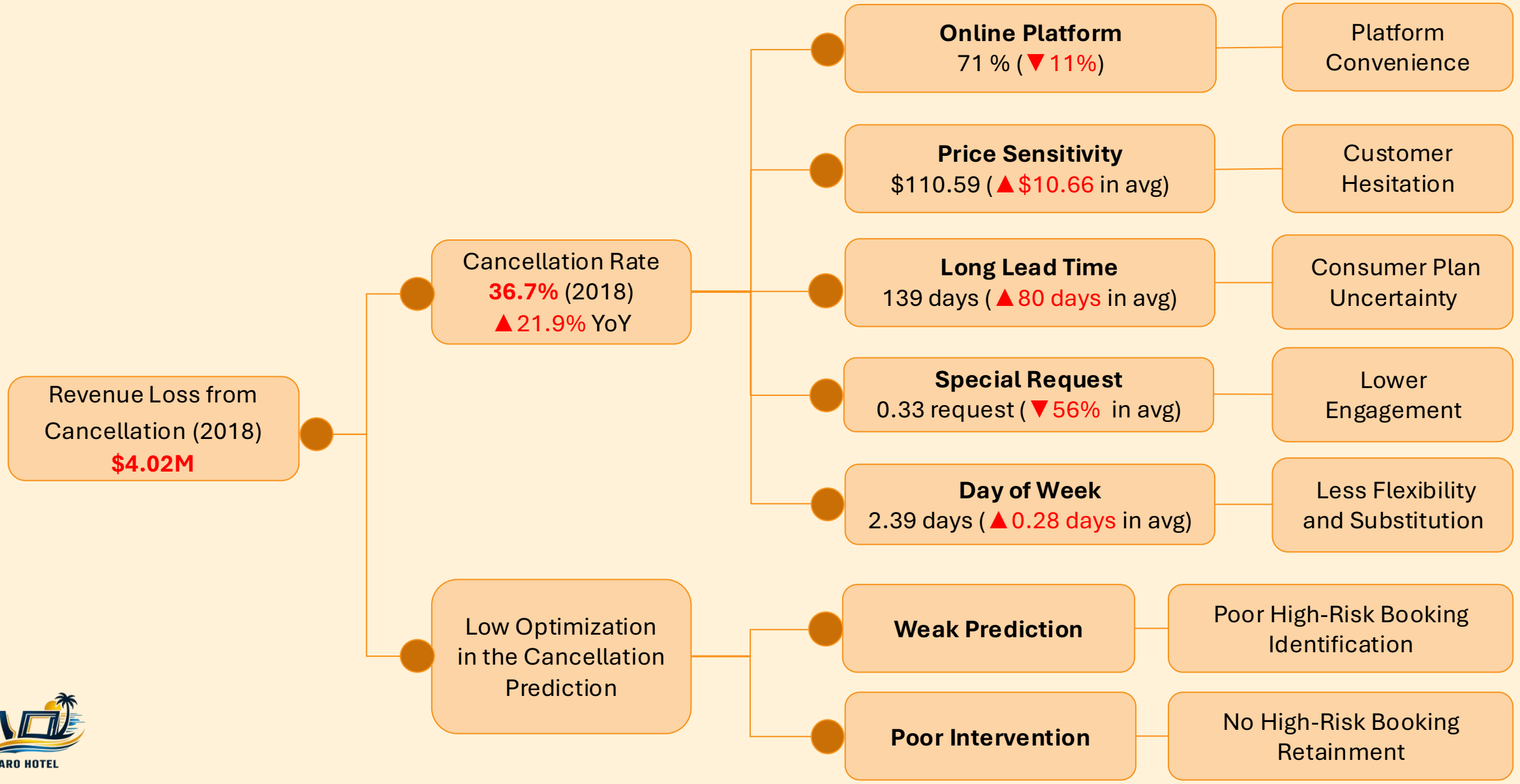
HARO HOTEL

Stakeholders Identification

Decision-maker	Head of Business Strategy
Accountable	Head of Data
Responsible	Data Analyst
Consulted	Marketing Lead, Sales Lead
Informed	Data Team, Marketing Team, Sales Team, Business Strategy Team

HARO HOTEL

Root Cause Analysis



HARO HOTEL

Stakeholders Identification

HARO HOTEL

Issue Branch	Hypothesis	Metrics	Prioritization
Online Platform	**H1:** Customers booking through online platform has higher cancellation rate as the platform giving convenient benefit	Cancellation Rate via Online booking	**Top Priority.** Adapting to business digitalization
Price Sensitivity	**H1:** Higher average room prices are associated with a higher probability of booking cancellation.	Average Daily Rate (average price per room per day)	**Top Priority.** Direct revenue impact
Consumer Plan Uncertainty	**H1:** Longer lead times between booking date and arrival date increase the likelihood of cancellation.	Lead Time (time between booking and arrival time)	**Middle-Top Priority**. Behavioural uncertainty
Low Customer Commitment	**H1:** Bookings with more special requests are less likely to be cancelled compared to those without special requests.	Number of Special Request	**Middle-Low Priority.** Engagement signal
Timing Flexibility (Weekend)	**H1:** Bookings scheduled with more weekday nights is more likely to be cancelled.	Arrival Day	**Low Priority.** External timing Factor

Hypothesis

Prioritized Hypotheses	Metrics	Reasoning
H1: Customers booking through online platform has higher cancellation rate as the platform giving	Cancellation Rate via Online bookings	Business has to adapt to digitalization booking although convenience for customers has downside of raising cancellation
H1: Higher average room prices are associated with a higher probability of booking cancellation.	Average Daily Rate (average price per room per day)	Measures customer price sensitivity and supports pricing optimization to reduce high-risk cancellations
H1: Longer lead times between booking date and arrival date increase the likelihood of cancellation.	Lead Time (time between booking and arrival time)	Identifies high-risk early bookings where proactive retention actions can be applied
H1: Bookings with more special requests are less likely to be cancelled compared to those without special requests.	Number of Special Request	Gives signal for engagement and commitment, enabling prioritization of retention efforts

There is no null in the dataset. Transforming arrival_month column into string data type.

Data Cleaning

arrival_month	count
October	5317
September	4611
August	3813
June	3203
December	3021
November	2980
July	2920
April	2736
May	2598
March	2358
February	1704
January	1014

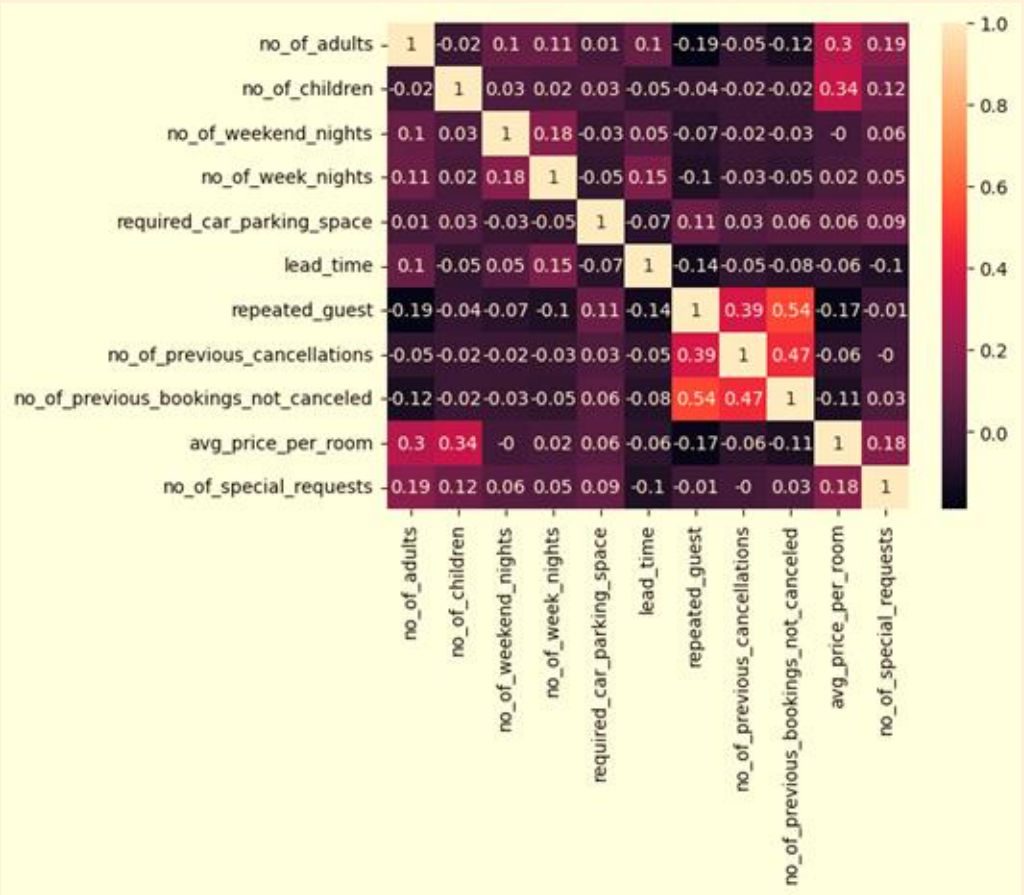
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  object
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(12), object(6)
memory usage: 5.3+ MB
```

HARO HOTEL

Looking at the correlation between the numerical columns, there is a moderate correlation between the previous number of cancellation and repeated guest. Hence, we remove the 'Not Cancelled' feature.

Understand the Dataset & Table Relationships

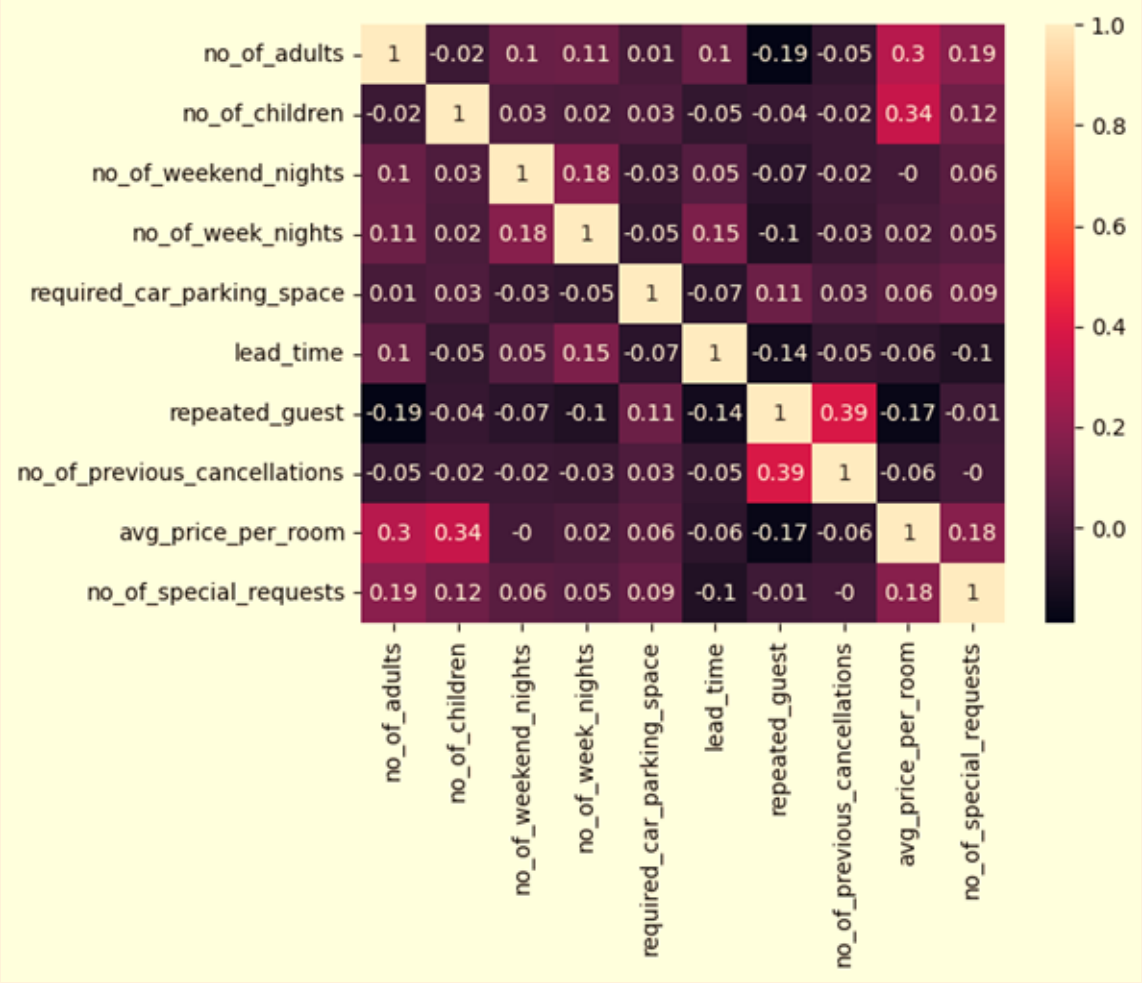
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```



HARO HOTEL

Low multicollinearity between numerical columns after taking out number of previous bookings not cancelled column

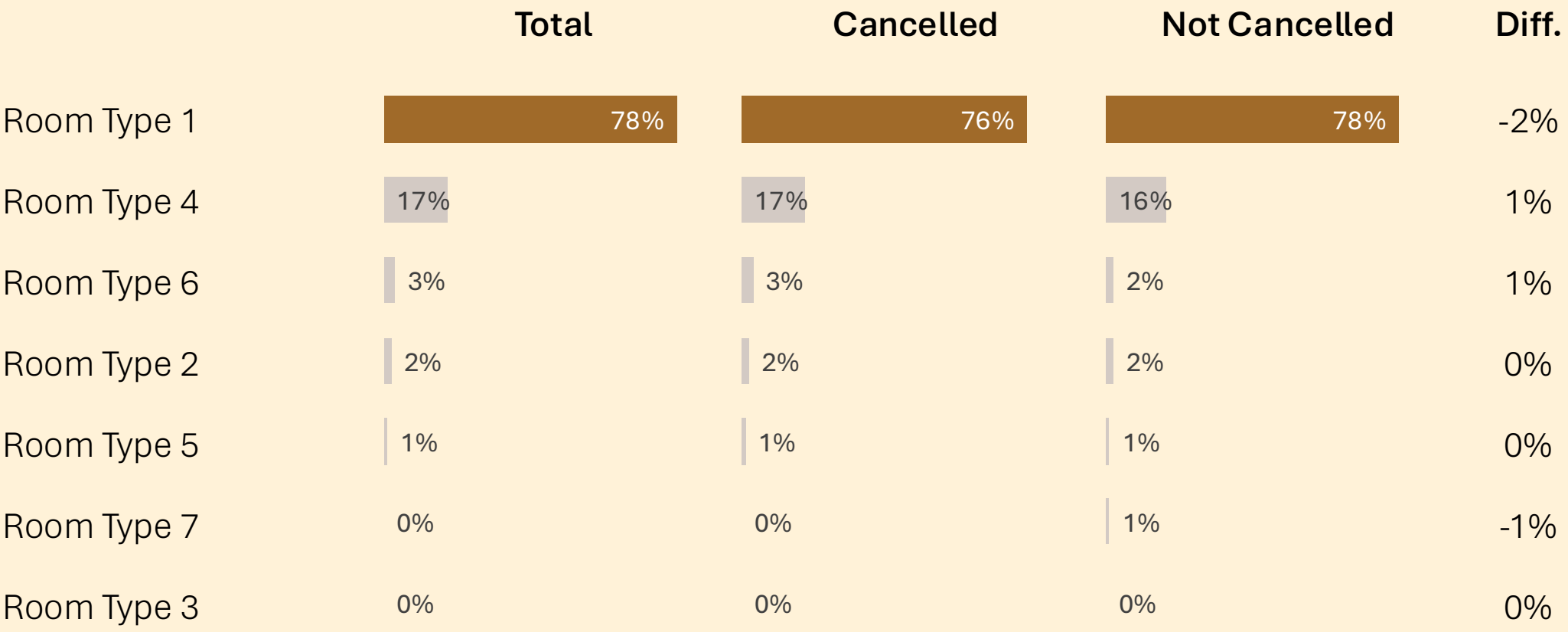
Table Relationship



HARO HOTEL

There were not many differences between the Cancelled bookings and the Non-Cancelled bookings in terms of room type reservation

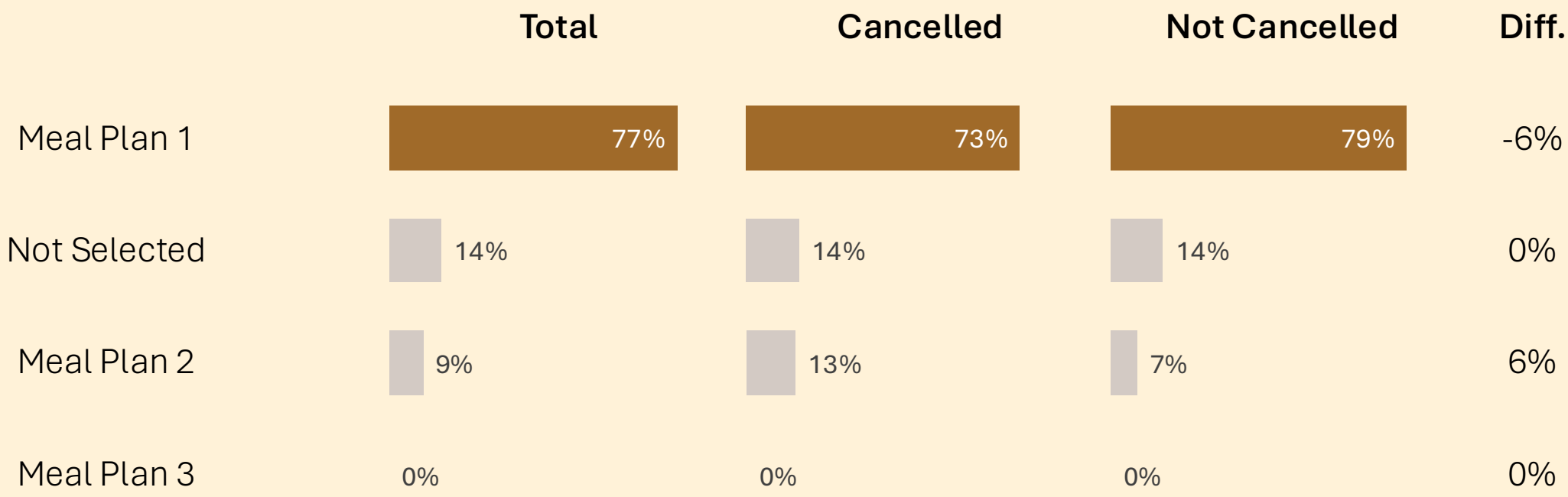
Room Type



HARO HOTEL

In terms of meal plan type, there were also not many differences between the Cancelled bookings and the Non-Cancelled bookings.

Type of Meal Plans



*) Hypothesis testing using Chi-Square test with 95% Confidence Level

Cancellation bookings also statistically proven tend to have higher average lead time, price per room, number of week nights, and fewer number of special request

Numerical Features

	Total	Cancelled	Not Cancelled	Diff.
No. of Weekend nights	0.81	0.89	0.77	0.12
No. of Adults	1.84	1.91	1.81	0.1
No. of Children	0.11	0.12	0.1	0.02
Required Car Parking Space	0.03	0.01	0.04	-0.03
Repeated Guest	0.03	0.00	0.04	-0.04
No of Previous Cancellations	0.02	0.01	0.03	-0.02
No of Previous Bookings Not Cancelled	0.15	0.00	0.23	-0.23

HARO HOTEL

Online Platform

Hypothesis Testing Calculation

Hypothesis	H1: Customers booking through online platform has higher cancellation rate as the platform giving convenient benefit
Statistical Test	Chi-Square, as it is categorical feature
Alpha	0.05
P-Value	6.7e-175 (less than Alpha) → accept H1
Interpretation	As the cancellation rate proportion of the online booking is higher, it means online booking has the higher chance to be cancelled.

HARO HOTEL

Lead Time

Hypothesis Testing Calculation

Hypothesis	H1: Longer lead times between booking date and arrival date increase the likelihood of cancellation.
Statistical Test	T-Test, as it is continuous numerical data
Alpha	0.05
P-Value	< 0.001 (less than alpha) → accept H1
Interpretation	As the mean score of lead time for cancelled booking is higher, it means the longer lead time between booking time and arrival time, the more likely it is to be cancelled.

HARO HOTEL

Average Price per Room

Hypothesis Testing Calculation

Hypothesis	H1: Higher average room prices are associated with a higher probability of booking cancellation.
Statistical Test	T-Test, as it is continuous numerical data
Alpha	0.05
P-Value	1.17e-175 (less than alpha→ accept H1
Interpretation	As the mean score of avg. price per room for cancelled booking is higher, it means the higher price they paid, the more likely the booking to be cancelled.

HARO HOTEL

Number of Special Request

Hypothesis Testing Calculation

Hypothesis	H1: Bookings with special requests are less likely to be Cancelled compared to those without special requests.
Statistical Test	Chi-Square, as it is categorical feature
Alpha	0.05
P-Value	< 0.001 (less than alpha→ accept H1
Interpretation	As the cancellation rate proportion of less special request is lower, the lesser the booking has special request, the higher the chance it is to be Cancelled.

HARO HOTEL

No. of Week Nights

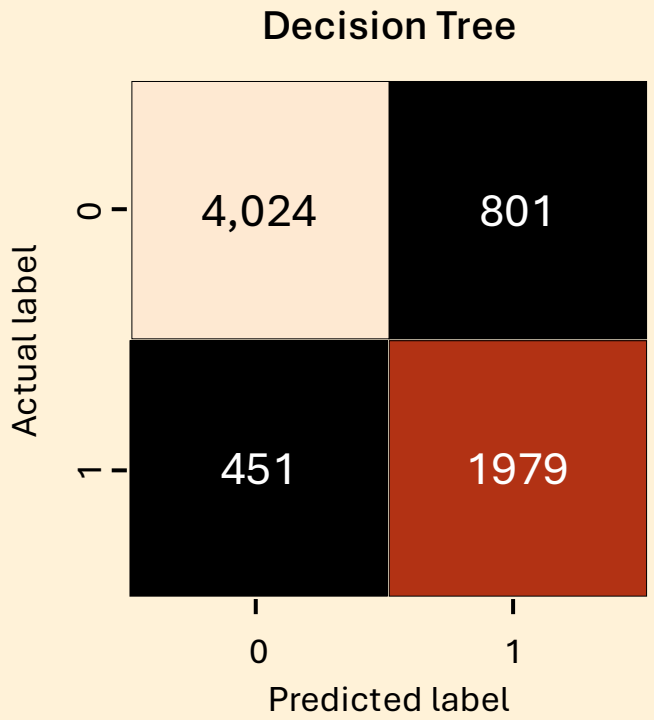
Hypothesis Testing Calculation

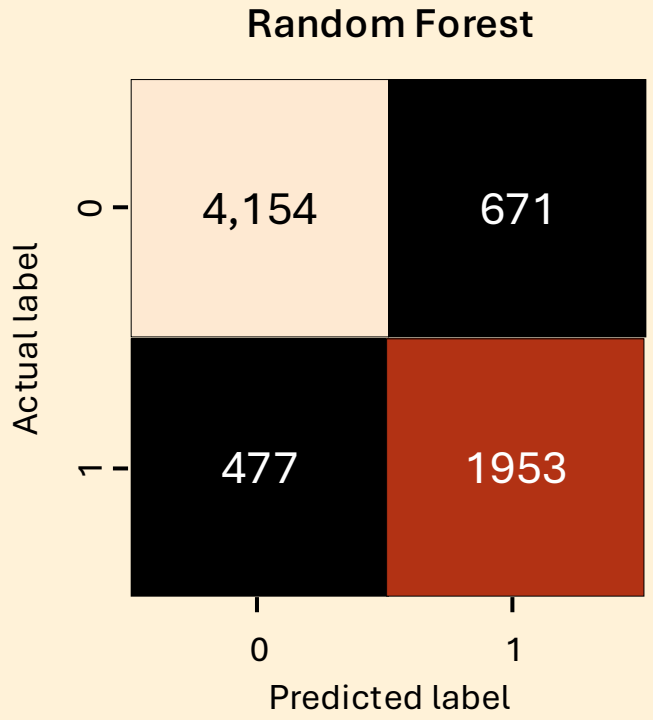
Hypothesis	H1: Higher average room prices are associated with a higher probability of booking cancellation.
Statistical Test	T-Test, as it is continuous numerical data and has normal distribution
Alpha	0.05
P-Value	2.7e-62 (less than alpha→ accept H1
Interpretation	As the mean score of number of weekday nights per room for cancelled booking is higher, bookings with more weekday nights have higher chances to be cancelled.

HARO HOTEL

The Decision Tree and Random Forest model reach 82-84% accuracy in predicting hotel cancellation. However, the PR-AUC Score is still within 64%-66% range. Hence, we apply boosting for better prediction.

Prediction Model Decision Tree and Random Forest



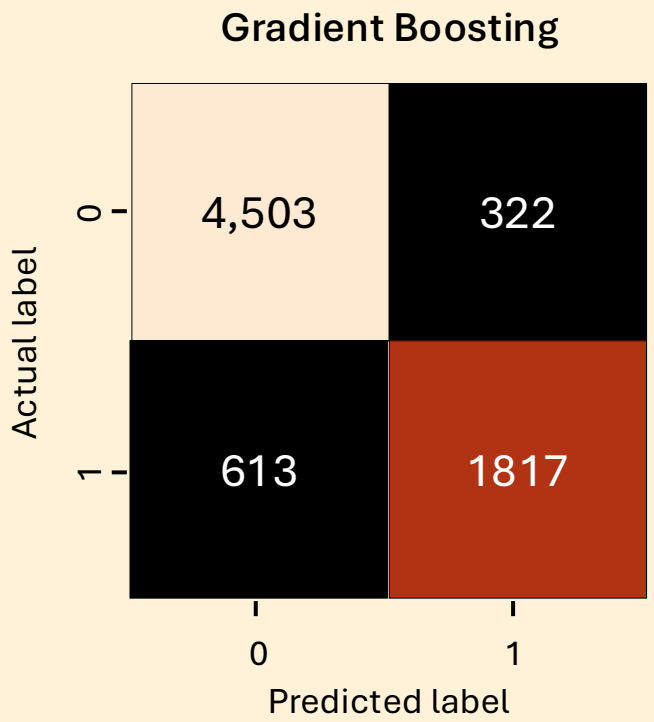


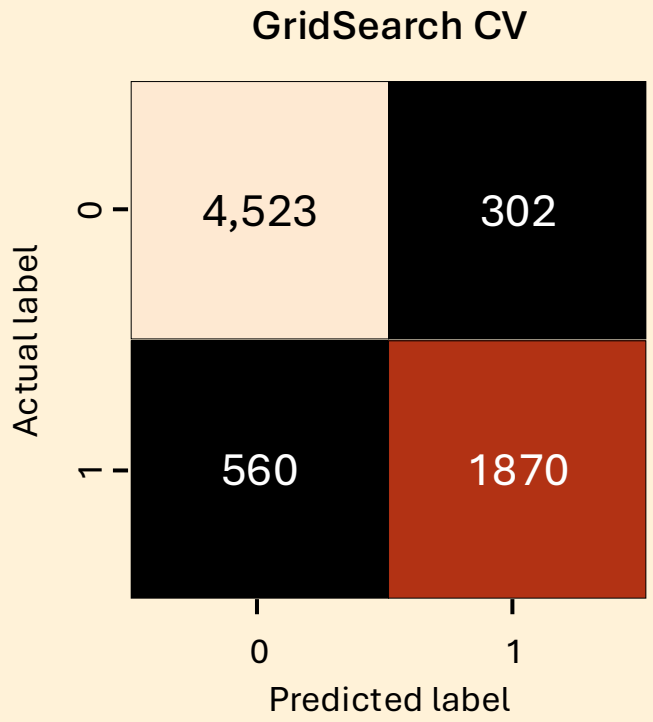
Model Accuracy	82%	84%
PR-AUC Score	64%	66%

HARO HOTEL

Based on the features and tree-based Prediction Model, Gradient Boosting model tuned by GridSearch CV succeed in detecting cancellation by 77% (recall).

Prediction Model Gradient Boosting, Hyperparameter Tuning GridSearch CV





Model Accuracy	87%	88%
PR-AUC Score	72%	91%
Recall	75%	77%

HARO HOTEL

Revenue Saving Scenario Calculation

Revenue 2018: €5,668,855

Prediction Model Recall: 77%

Total Revenue Saving from Recall: 3,300,613

Yearly Cancellation Rate: 36.71%

Retainment from Cancellation	Cancellation Rate Decrease	Final Cancellation Rate	Revenue Saving (€)	% Revenue Saving
5%	1.41%	35%	165,030.65	3%
10%	2.82%	34%	330,061.3	6%
20%	5.65%	31%	660,122.6	12%
30%	8.48%	28%	990,183.9	17%
40%	11.30%	25%	1,320,245.2	23%
50%	14.13%	23%	1,650,306.5	29%
60%	16.960%	20%	1,980,367.8	35%
70%	19.78%	17%	2,310,429.1	41%

HARO HOTEL

*) Calculation file

Classification Performance

Model Performance		Benchmark
Model Accuracy:	How well the model can predict the actual Cancelled and Non-Cancelled bookings	80-85%
PR-AUC Score:	How well the model predict the Cancellation while avoiding false alarm	55-60%
Recall:	How well the model predict the Cancellation out of all the actual Cancelled bookings	60-70%

HARO HOTEL