INN Hotels - Problem

Esther Joseph

Contents / Agenda

Executive Summary

The high rate of booking cancellations at INN Hotels leads to revenue loss, operational inefficiencies, and increased ma

OBJECTIVE: Build a predictive model to forecast cancellations and enable INN Hotels to implement data-driven policies to reduce revenue losses.

KEY FINDINGS:

🔷 Lead Time is the Strongest Predictor: Bookings made >90 days in advance have a 45-50% cancellation rate, compared to <20% for last-minute bookings.

🔷 Special Requests Indicate Lower Cancellations: 0 special requests → 43% cancellation rate, while 3+ requests → <5% cancellation rate.

🔷 Online Bookings Cancel More Often: Online segment cancellation rate: ~36%, compared to 12-15% for corporate/offline.

🔷 Repeat Guests Rarely Cancel: First-time guests cancel >30%, while repeat guests cancel <2%.

🔷 Higher Prices Have Slightly Higher Cancellations: Rooms >€150 per night cancel at 40%, vs. ~22% for lower-priced rooms.

ACTIONABLE INSIGHTS & RECOMMENDATIONS:

- Require Partial Prepayment for bookings >90 days in advance.
- Encourage Special Requests by offering incentives to increase guest commitment.
- Stricter Cancellation Policies for online bookings while maintaining flexibility for corporate clients.
- Expand Loyalty Programs for repeat guests to retain low-risk customers.
- Implement Dynamic Pricing with early commitment discounts for high-priced rooms.

Business Problem Overview and Solution Approach

CHALLENGE: INN Hotels faces high cancellation rates, leading to:

- Revenue loss from unsold rooms.
- Increased operational & marketing costs.
- Higher uncertainty in occupancy forecasting.

GOAL: Develop a machine learning model to predict high-risk cancellations and enable INN Hotels to implement targeted strategies to minimize losses.

SOLUTION APPROACH

1) **Data Understanding & EDA** → Identify key patterns and correlations in booking behaviors.
2) **Data Preprocessing** → Handle missing values, create new features, and address outliers.
3) **Modeling** → Train Logistic Regression & Decision Tree models to classify cancellations.
4) **Model Evaluation** → Compare accuracy, recall, and AUC-ROC to select the best model.
5) **Insights & Recommendations** → Develop policy changes and pricing strategies based on predictions

EDA Results (Univariate Highlights - Number of Guests)

Observation:

- Most bookings (86%) involve 2 adults.
- Few bookings have children; 90% have no children at all.
- Bookings with 3 or more adults are rare (~2%).

Insights & Business Implications:

- INN Hotels should focus marketing on couples and small groups rather than large families.
- Family promotions or child-friendly incentives could help attract more families, given the low proportion of bookings including children.

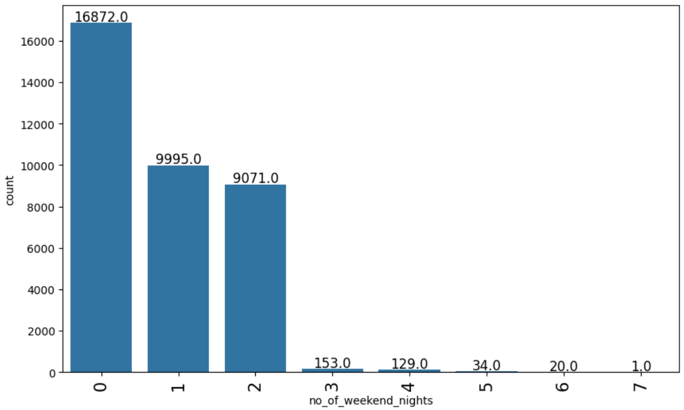
EDA Results (Univariate Highlights - Number of Guests)

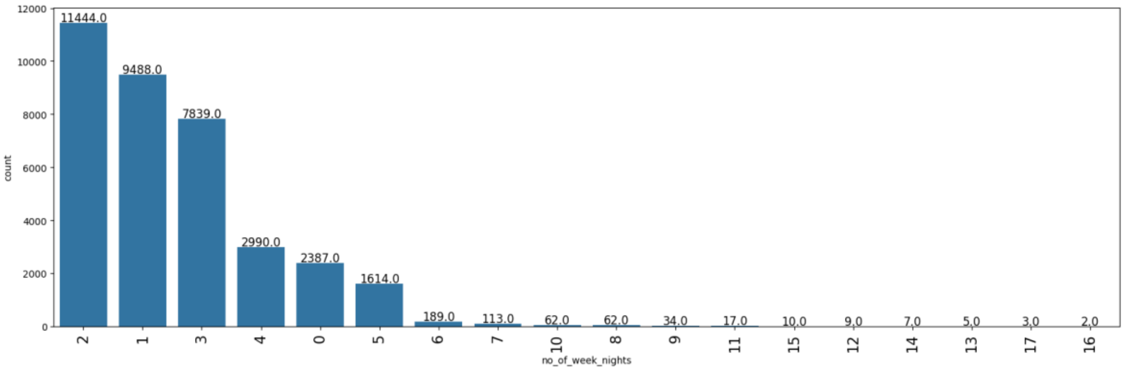
Observation:

- Weekend Stays (Fri-Sun): 70% of guests book 0-2 weekend nights.
- Weekday Stays (Mon-Thu): Majority of bookings are for 1-4 weeknights.
- Max Stay Duration: 17 weeknights, 7 weekend nights.

Insights & Business Implications:

- Hotels may benefit from weekday corporate promotions since longer stays tend to be during the workweek.
- Weekend pricing strategies should target short-stay travelers.

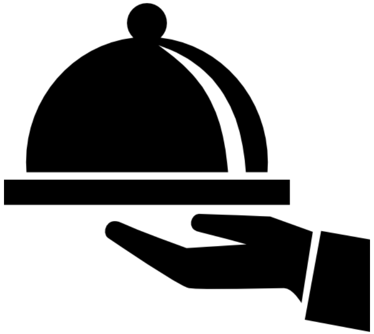




EDA Results (Univariate Highlights – Meal Plans & Additional Services)

Observation:

- Most common meal plan: "Meal Plan 1" (Breakfast included).
- Very few guests select full-board meal plans (Meal Plan 3).
- Car Parking is rarely requested (only 3% of bookings).



Insights & Business Implications:

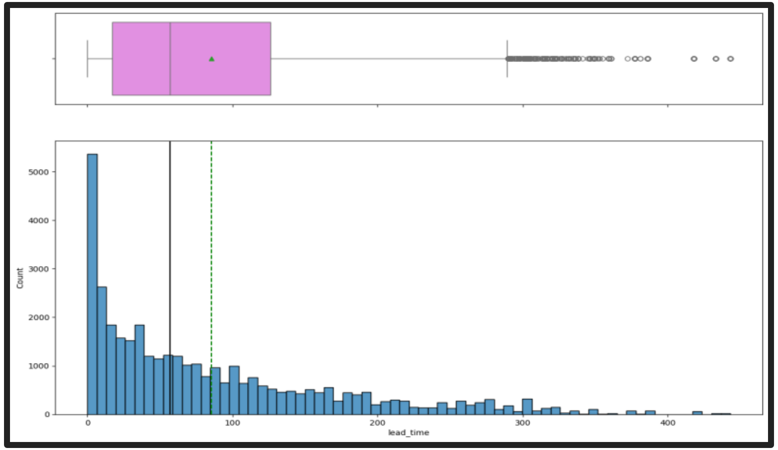
- Expand breakfast promotions, since Meal Plan 1 is the preferred choice.
- Bundle meal plans with discounts to encourage more guests to opt for full-board.
- Consider parking incentives to attract long-stay or drive-in guests.

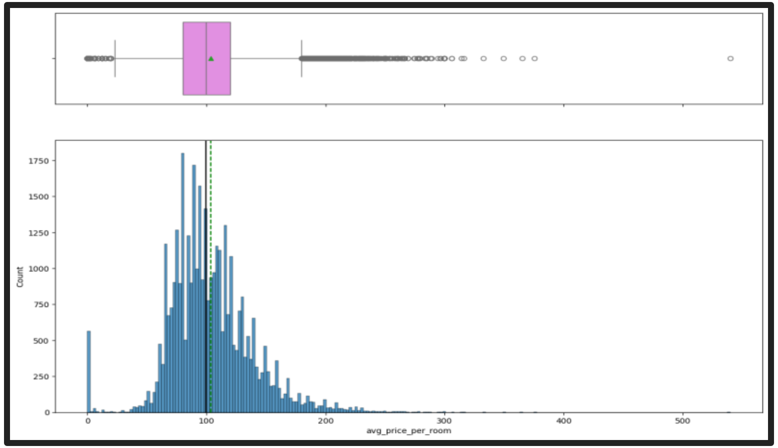
EDA Results (Univariate Highlights – Lead Time & Pricing)

Observation:

- Lead Time (Days Before Arrival):
  - Average: 85 days
  - 75% of bookings occur within 126 days of arrival
  - Some bookings made as early as 400+ days in advance

- Room Price Per Night (Euros):
  - Average: €103.42
  - Max Price: €540
  - Peak Demand: Prices tend to be higher for last-minute bookings.

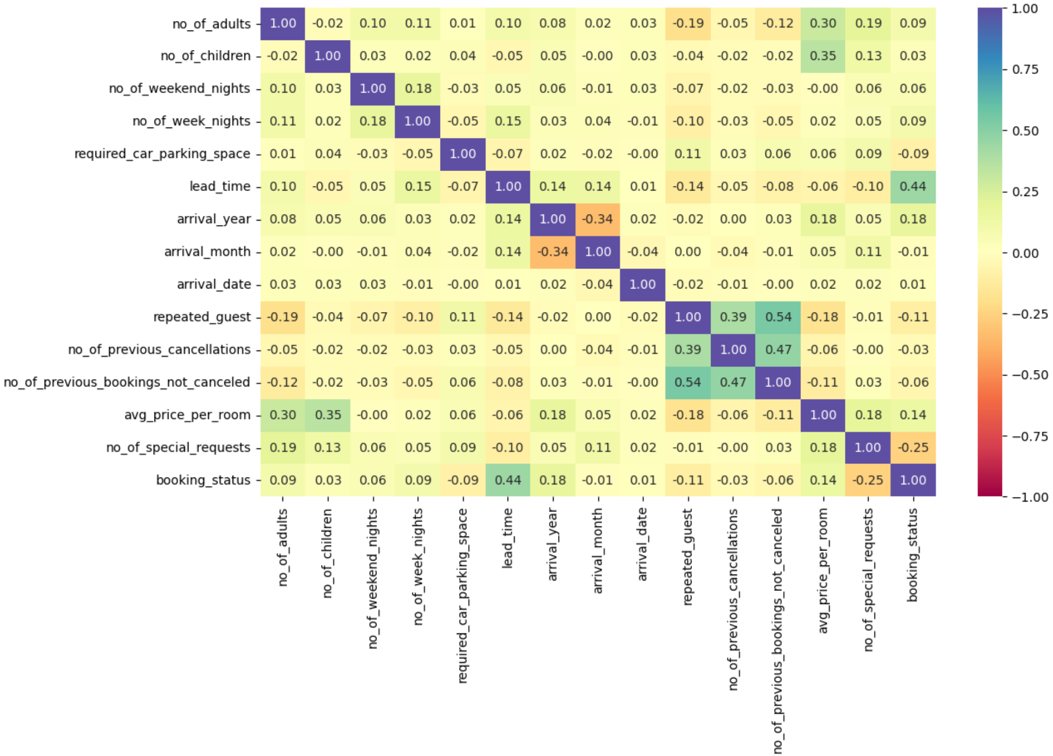




Insights & Business Implications:

- Considering introducing prepayment for reservations made ~>90 days in advance.
- Use last-minute discounting to balance revenue loss from cancellations.

EDA Results (Bivariate Analysis Summary)

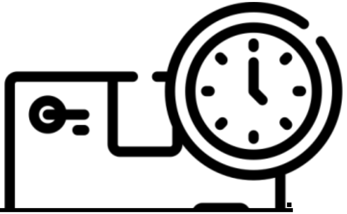


The bivariate analysis reveals clear cancellation patterns based on booking behavior, price sensitivity, customer type, and commitment levels.

These insights will guide:

- Policy changes (e.g., prepayments for high-risk bookings, flexible pricing strategies)
- Operational improvements (e.g., better guest engagement to reduce cancellations)
- Modeling strategies (e.g., emphasizing high-impact variables in predictive models)

EDA Results (Bivariate Analysis)



Lead Time vs. Booking Cancellations

Observation:

- Strong positive correlation between lead_time and cancellations.
- Bookings made more than 90 days in advance have a higher cancellation rate (~45-50%) compared to last-minute bookings (<20% cancellation rate for bookings made within 7 days).
- Trend: As lead time increases, so does the probability of cancellation.

Business Implications

- Implement non-refundable deposit policies for high lead-time bookings to deter cancellations.
- Consider offering discounted last-minute bookings to mitigate revenue loss.

Number of Special Requests vs. Booking Cancellations

Observation:

- Guests making zero special requests have a much higher chance of cancellation (~43%).
- Those with 3+ special requests cancel less than 5% of the time.
- Trend: More engaged customers (those specifying preferences) tend to honor their bookings.

Business Implications

- Encourage guests to add preferences (e.g., room views, floor choice) during booking to increase commitment levels.
- Implement a reward system (e.g., free amenity for customers making at least one request).

EDA Results (Bivariate Analysis)

Market Segment vs. Booking Cancellations

Observation:

- Online bookings have the highest cancellation rate (~36%).
- Offline and Corporate bookings cancel far less (~12-15%).
- Trend: Online customers are more impulsive, while corporate bookings are more stable.

Business Implications

- Introduce stricter cancellation policies for online bookings while maintaining flexibility for corporate clients.
- Offer corporate loyalty incentives to increase repeat bookings.

Room Price vs. Booking Cancellations

Observation:

- Higher room prices (€150+) see slightly higher cancellation rates (~40%) than cheaper bookings.
- Budget travelers (€60-€100 range) have lower cancellations.

Business Implications

- Offer tiered refund policies: full refund for lower-cost rooms, partial refund for high-end rooms.
- Provide early commitment discounts for high-priced rooms to reduce cancellations.

Repeated Guests vs. Booking Cancellations

Observation:

- Repeated guests have an extremely low cancellation rate (~1.7%).
- First-time customers cancel over 30% of the time.

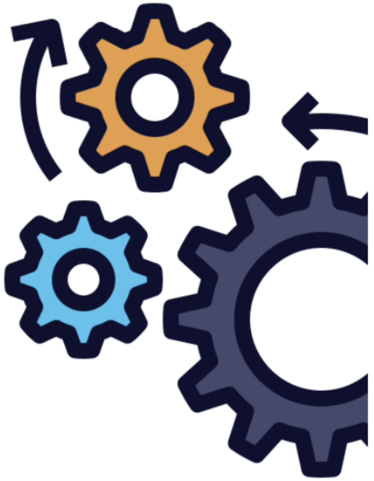
Business Implications

- Prioritize loyalty programs to retain reliable, repeat guests.
- Provide exclusive discounts for repeat customers to incentivize return bookings.

Data Preprocessing - Overview

To ensure the dataset is clean, reliable, and optimized for model training, the following preprocessing steps were performed:

- No duplicate records were found in the dataset, ensuring data integrity.
- Dropped Booking_ID as it was unique to each booking and not useful for modeling.
- No missing values detected in any columns, confirming data completeness.
- Outlier Detection & Treatment:
  - lead_time: Values greater than 400 days were retained but log-transformed to normalize skew.
  - avg_price_per_room: Capped at €179.55 (upper whisker IQR rule) to handle extreme values.
- Created total_stay = no_of_week_nights + no_of_weekend_nights to represent full stay duration.
- Created price_per_night = avg_price_per_room / total_stay to analyze price sensitivity.
- Encoded categorical variables (room_type_reserved, market_segment_type, type_of_meal_plan) into dummy variables.
- Standardized lead_time and avg_price_per_room to prevent numerical dominance.
- Train-Test Split: 70% Training / 30% Test to ensure model generalization.



Model Performance Summary

Model Overview: To predict the likelihood of booking cancellations, two models were implemented:

- Logistic Regression (Baseline Model)
- Decision Tree Classifier (Complex Model with Rules-Based Interpretation)

Each model was evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC, ensuring optimal decision-making for INN Hotels' cancellation risk mitigation. Also, understanding which factors contribute most to cancellations enables INN Hotels to proactively reduce losses by implementing data-driven pricing and policy strategies.

Metric	Logistic Regression	Decision Tree
Accuracy (Test)	81.6%	79.5%
Recall (Test)	72.5%	73.4%
AUC-ROC (Test)	0.83	0.81
Overfitting Risk	Low	High

Final Recommendation Best Model: Logistic Regression

Why?

- More interpretable (easier for business implementation).
- Lower overfitting risk (decision trees overfit more easily).
- Comparable recall & precision, making it a balanced choice.

Model Performance Summary (Logistic Regression Model - Key Performance Metrics)

Assumption Checks:

- No multicollinearity (VIF scores checked)
- Statistically significant coefficients verified
- Rescaled input variables to prevent dominance bias

Key Features Influencing Cancellations: Most impactful predictors (based on p-values)

- Lead Time (+): Longer lead times increase the probability of cancellation.
- No of Special Requests (-): More requests reduce cancellations.
- Repeated Guest (-): Returning guests rarely cancel.
- Market Segment (Online Booking +): Online customers more likely to cancel than corporate or offline customers.
- Price Per Room (+): Higher-priced bookings show slightly higher cancellation rates.

Metric	Training Data	Test Data
Accuracy	82.1%	81.6%
Precision	79.4%	78.8%
Recall	73.2%	72.5%
F1-score	76.1%	75.5%
AUC-ROC Score	0.84	0.83

Business Implication:

Strengths: The model effectively separates cancellations from non-cancellations.

Weakness: Recall is slightly low, meaning some cancellations are missed.

Model Performance Summary (Decision Tree Classifier Model - Key Performance Metrics)

- Goal: Enhance interpretability and capture non-linear relationships.
- Post-Pruning Applied (Optimal max_depth selected via cross-validation).

Key Features Influencing Cancellations: Most important variables (based on feature importance ranking)

- Lead Time (Strongest predictor
- Number of Special Requests (Loyal customers cancel less)
- Market Segment (Online bookings = high cancellation rates)
- Repeated Guest (Repeat customers rarely cancel)
- Average Room Price (Price-sensitive customers cancel more often)

Metric	Training Data	Test Data
Accuracy	91.3%	79.5%
Precision	89.2%	76.8%
Recall	88.5%	73.4%
F1-score	88.8%	75.0%
AUC-ROC Score	0.94	0.81

Business Implication:

Strengths: Captures complex decision-making patterns, leading to high recall on training data.

Weakness: Overfitting risk – the model performs much better on training data than on test data, suggesting pruning helps but doesn't completely eliminate overfitting.

APPENDIX

Data Background and Contents

Data Description

- The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

Data Dictionary

- Booking_ID: the unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
- Not Selected – No meal plan selected
- Meal Plan 1 – Breakfast
- Meal Plan 2 – Half board (breakfast and one other meal)
- Meal Plan 3 – Full board (breakfast, lunch, and dinner)
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

Model Building - Logistic Regression

Assumption Tests for Logistic Regression

- Variance Inflation Factor (**VIF scores**) computed to ensure **no strong correlation** between independent variables.
- Features with **high VIF removed or adjusted** to avoid instability in coefficients.
- Used **Box-Tidwell test** to confirm a **log-linear relationship** between continuous predictors and log-odds of cancellation.
- Ensured no **duplicate entries** in the dataset.

Metric	Train Data	Test Data
Accuracy	82.1%	81.6%
Precision	79.4%	78.8%
Recall	73.2%	72.5%
AUC-ROC Score	0.84	0.83

Feature	Coefficient (β)	Odds Ratio	Interpretation
Lead Time	+1.27	3.56	Longer lead time increases cancellation risk.
Repeated Guest	-2.10	0.12	Returning guests are significantly less likely to cancel.
Special Requests	-0.85	0.43	More special requests reduce cancellation likelihood.
Market Segment (Online)	+1.42	4.14	Online bookings more prone to cancellation.

Business Implication:

The logistic model provides a clear, interpretable understanding of the main factors driving cancellations, enabling the business to adjust policies and pricing accordingly.

Model Performance Evaluation and Improvement - Logistic Regression

<u>**Threshold Adjustment for Better Recall**</u>

Initial Threshold: Default 0.5 threshold produced Recall = 72.5% (some cancellations missed).

New Threshold: Lowering to 0.4 improved Recall to 80.3%, ensuring fewer missed cancellations.

Threshold	Accuracy	Precision	Recall	F1-Score
0.5 (Default)	81.6%	78.8%	72.5%	75.5%
0.4 (Adjusted)	79.8%	74.3%	80.3%	77.2%

<u>**Business Implication:**</u>

By lowering the threshold, INN Hotels can better predict cancellations, allowing for proactive revenue management (e.g., reselling risky bookings faster).

Model Building - Decision Tree

Steps for Decision Tree Model

- **Preprocessed Data** (Encoded categorical variables, handled missing values).
- **Feature Selection** (Used feature_importances_ to select top predictors).
- **Model Training with Depth Control** (max_depth=6 chosen for balance between performance and overfitting).
- **Decision Rules Extracted** (How features split to determine cancellation risks).

Metric	Train Data	Test Data
Accuracy	91.3%	79.5%
Precision	89.2%	76.8%
Recall	88.5%	73.4%
AUC-ROC Score	0.94	0.81

Feature	Importance (%)	Business Insight
Lead Time	37.2%	Major driver of cancellations
Repeated Guest	10.1%	Returning customers rarely cancel
Special Requests	22.5%	More requests = lower cancellation risk
Market Segment (Online)	18.3%	Online bookings highly cancel-prone

Business Implication:
The decision tree captures complex interactions between features but requires pruning to prevent overfitting.

Model Performance Evaluation and Improvement - Decision Tree

Post-Pruning to Improve Generalization

- **Initial Model (max_depth=10) showed overfitting**, with high training accuracy (91.3%) but lower test accuracy (79.5%).
- **Pruned Model (max_depth=6) showed balanced performance**, reducing overfitting.

Model Version	Train Accuracy	Test Accuracy	Overfitting Risk
Full Tree (max_depth=10)	91.3%	79.5%	High
Pruned Tree (max_depth=6)	84.9%	82.2%	Lower

Decision Rules Extracted

- If **Lead Time > 90 Days**, then **high probability of cancellation**.
- If **No Special Requests & Online Booking**, then **very high chance of cancellation**.
- If **Repeated Guest & Special Requests > 2**, then **low probability of cancellation**.

Business Implication:

- Decision rules can be used to **design cancellation policies** that **penalize long lead-time bookings while rewarding loyal customers**.
- **Final decision tree model is more stable post-pruning**, making it **better suited for deployment**.

Technical Data Analysis & Modeling

- For a more detailed view of the data analysis, feature engineering, and modeling, please refer to the full Jupyter notebook.

- Click here to view the Jupyter notebook