INNS Hotel Project

Greater Learning Data Science

01 February 2025

By: Ryan J. Stambaugh

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Customer Behavior Insights

📌 **Insight:**

1. **Lead Time Strongly Impacts Cancellations**
   - The model shows that **longer lead times significantly increase cancellation rates**.
   - Customers booking far in advance are more likely to cancel their reservations.

✅ **Recommendations:**

1. **Reduce Cancellation Risks for Long Lead-Time Bookings:**
   - Implement **non-refundable pricing options** for long lead-time bookings.
   - Introduce **prepaid discounts** to incentivize commitment.
   - Send **personalized reminders** and exclusive discounts closer to the check-in date to retain bookings.

Executive Summary

Booking Preferences & Cancellation Trends

📌 **Insight:**

1. **Required Car Parking Space Reduces Cancellations**
   - Guests who **request a car parking space** are **less likely to cancel** their bookings.
   - This suggests that these customers may be **more committed to their travel plans**.

✅ **Recommendations:**

1. **Increase Parking Space Reservations:**
   - Offer **discounted parking packages** to encourage reservations.
   - Highlight the **scarcity of parking availability** (e.g., "Limited parking spots available!").
   - Implement an **upsell strategy** where guests can secure a spot for a small fee.

Executive Summary

Market Segment Impact on Booking Status

📌 **Insight:**

1. **Corporate and Offline Bookings Have Higher Cancellation Rates**
   - **Corporate and Offline bookings are more likely to be canceled** compared to Online bookings.
   - This indicates a possible lack of engagement or flexibility in these segments.

✅ **Recommendations:**

1. **For Corporate Bookings:**
   - Implement **contract-based discounts** or loyalty rewards for repeat business travelers.
   - Introduce **flexible modification options** instead of outright cancellations.
1. **For Offline Bookings:**
   - Encourage **online reservations with exclusive web-only discounts**.
   - Improve communication (e.g., follow-up calls, confirmation emails) to increase commitment.
   - Require a **deposit for offline reservations** to reduce cancellations.

Executive Summary

Threshold Adjustments for Better Cancellation Predictions

📌 **Insight:**

1. **Adjusting the Classification Threshold Improves Recall**
   - The default **threshold (0.5) balances precision and recall** but increasing it to **0.37** captures **more cancellations at the cost of precision**.

✅ **Recommendations:**

1. **Choose a Threshold Based on Business Priorities:**
   - If **reducing lost revenue** from cancellations is the top priority → Use **Threshold = 0.37** (higher recall).
   - If **minimizing false cancellations (incorrectly predicting cancellations)** is more critical → keep **Threshold = 0.5** (balanced performance).

Executive Summary

Revenue Protection Strategies

📌 **Insight:**

1. Non-Refundable vs. Refundable Booking Structures Can Reduce Cancellations

- The model indicates that **some segments are more prone to cancellation** than others.
- A **dynamic pricing strategy** can help reduce risk.

✅ **Recommendations:**

1. Introduce Tiered Booking Options:

- **Fully Refundable (higher price)** for travelers needing flexibility.
- **Partially Refundable (moderate price)** for bookings that allow changes.
- **Non-Refundable (discounted price)** to incentivize committed bookings.

2. Use Dynamic Pricing Based on Cancellation Risk:

- Increase prices for **high-risk segments** (e.g., long lead time, offline bookings).
- Offer early-bird discounts to **lock in customers earlier**.

Business Problem Overview and Solution Approach

📌 **Problem:**

1. **A significant number of hotel bookings are called off due to cancellations or no-shows**
   - The model indicates that **some segments are more prone to cancellation** than others.
   - Cancellations cause a loss of revenue
   - Additional costs to fill rooms by lowering prices last minute to increase bookings

Executive Summary

Business Strategy Recommendations

Strategy	Action Plan
Reduce long lead-time cancellations	Offer prepaid rates, discounts near check-in, and automated reminders
Optimize corporate and offline bookings	Require deposits, offer flexible rescheduling, and improve online conversion
Encourage parking reservations	Provide discounts for early parking bookings and upsell parking as an add-on
Threshold optimization	Adjust classification thresholds based on business needs (recall vs. precision trade-off)
Dynamic pricing strategy	Adjust pricing based on book risk, and non-refundable vs. refundable tiers

Executive Summary

Next Steps

1.Test pricing strategies with A/B testing for refundable vs. non-refundable bookings.

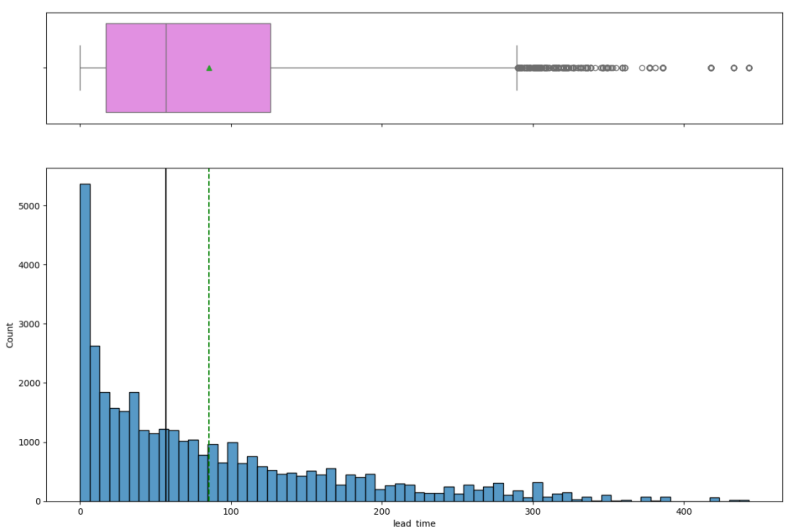
2. Monitor cancellation patterns over time to refine the model further.

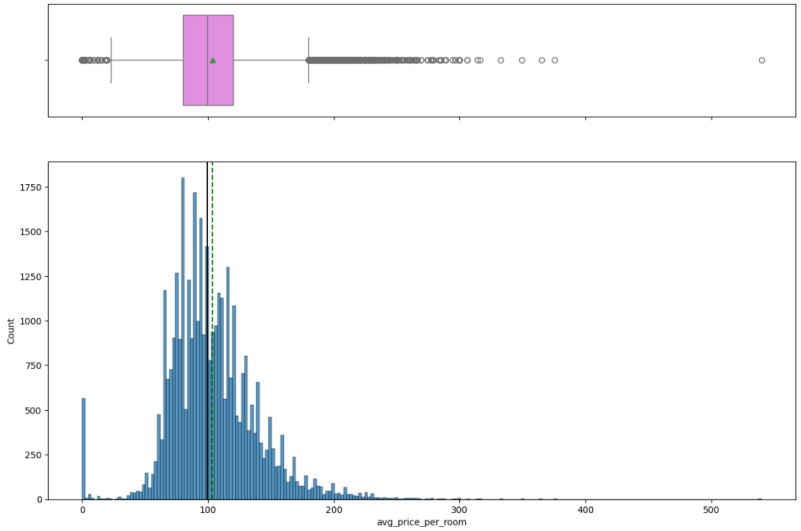
3. Enhance customer engagement (SMS/email reminders, pre-arrival perks) to secure bookings.

Implementing these steps can help **increase revenue, improve customer retention, and reduce unnecessary cancellations**.

EDA Results

Observations for LEAD Time and Average Price Per Room:

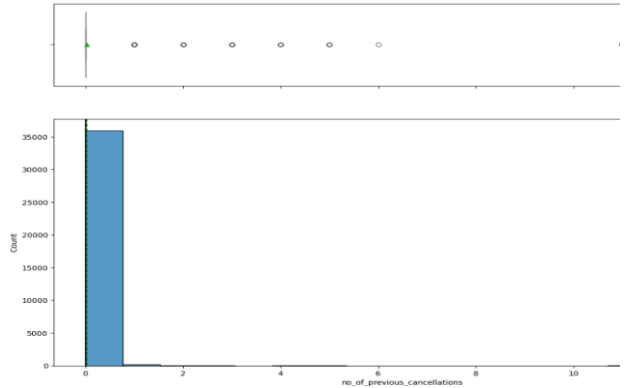




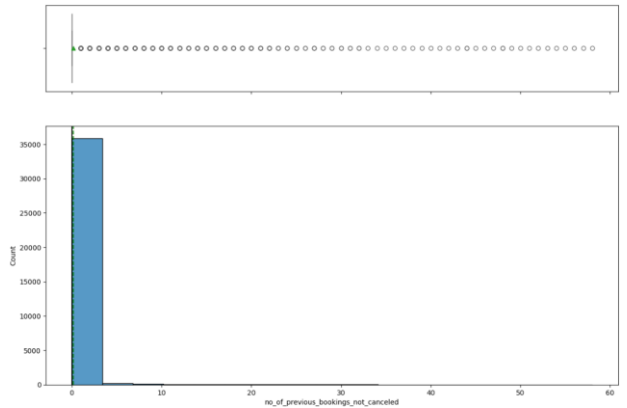
EDA Results

Great Learning
POWER AHEAD

Number of previous cancelations

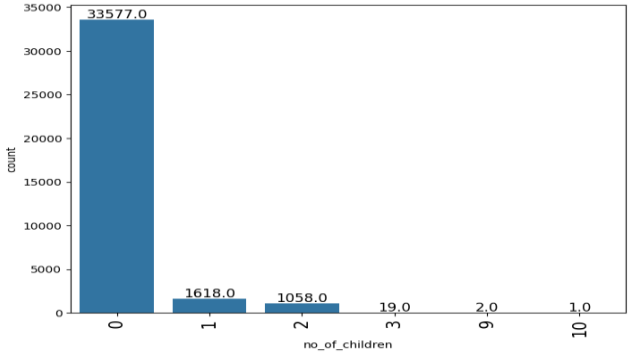


Number of previous bookings not canceled

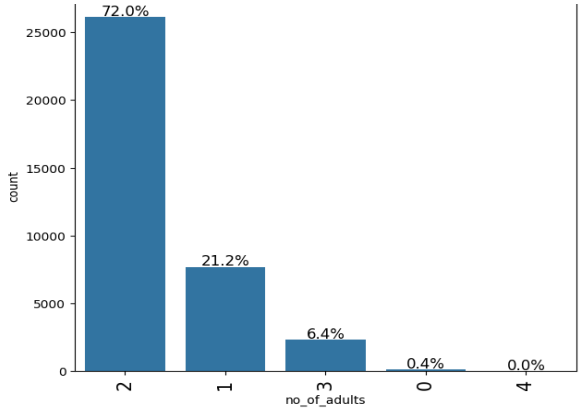


EDA Results

Number of Children

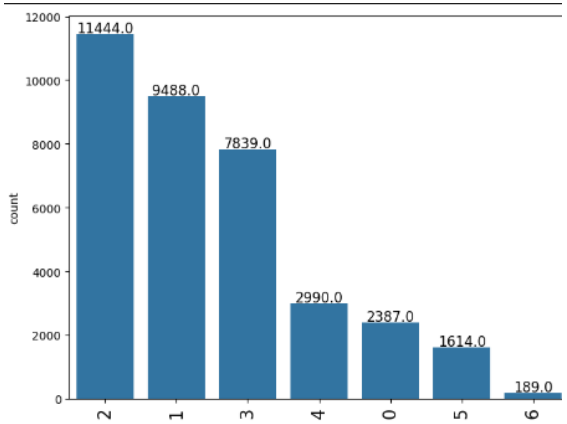


Number of Adults

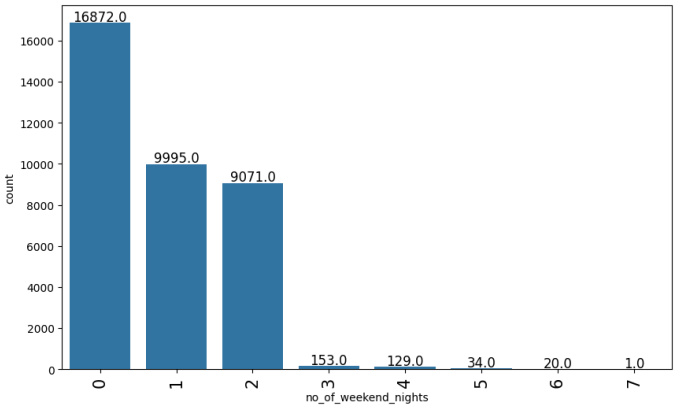


EDA Results

Number of Weeknights

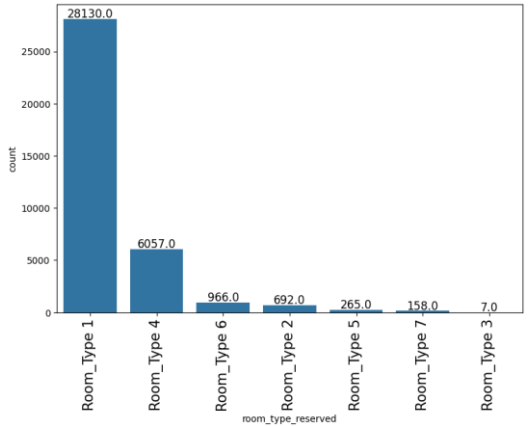


Number of Weekend Nights

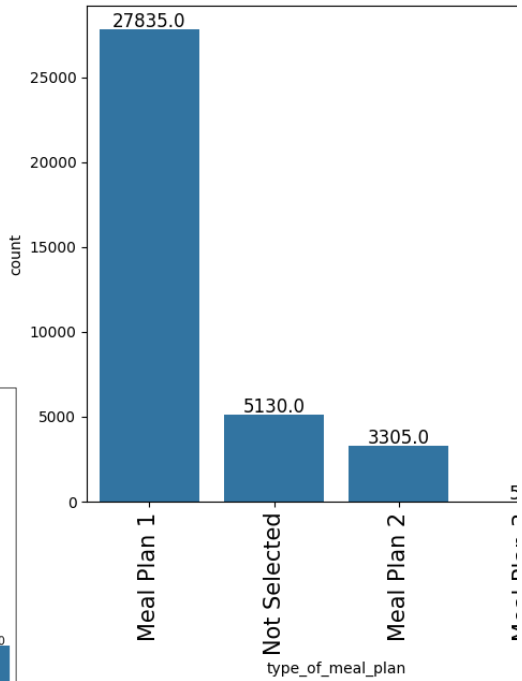


EDA Results

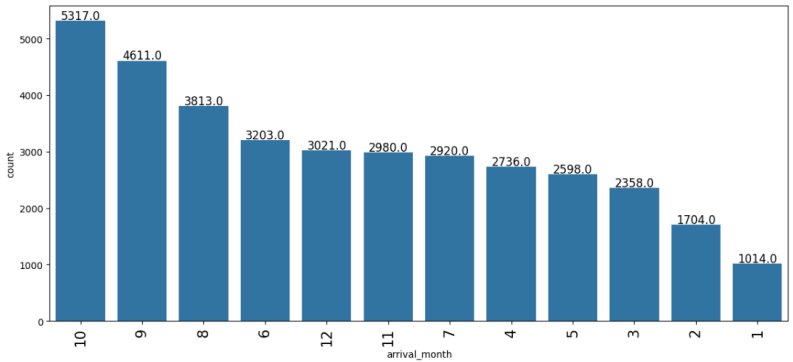
Room Type Reserved



Type of Meal Plan



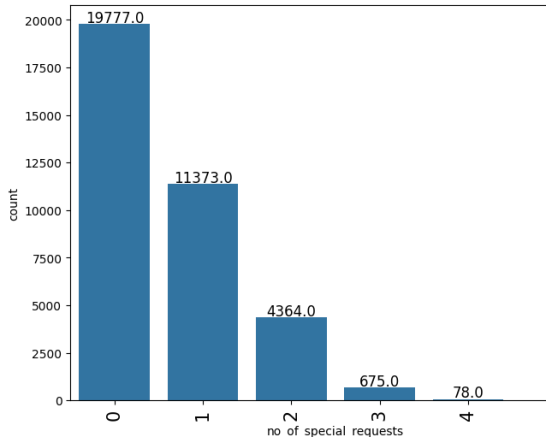
Arrival Month

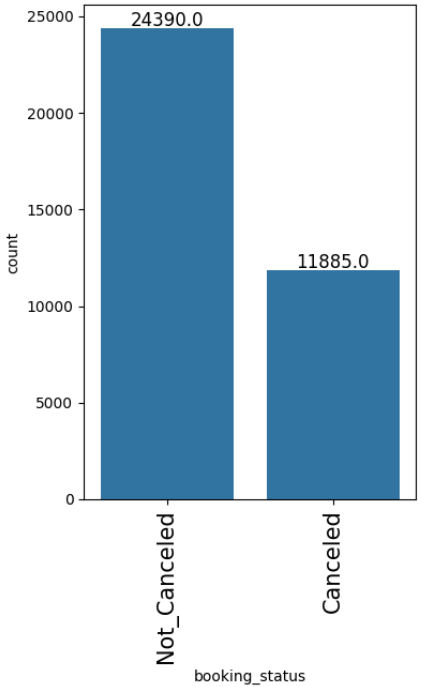


EDA Results

Number of Special Requests

Booking Status





EDA Results

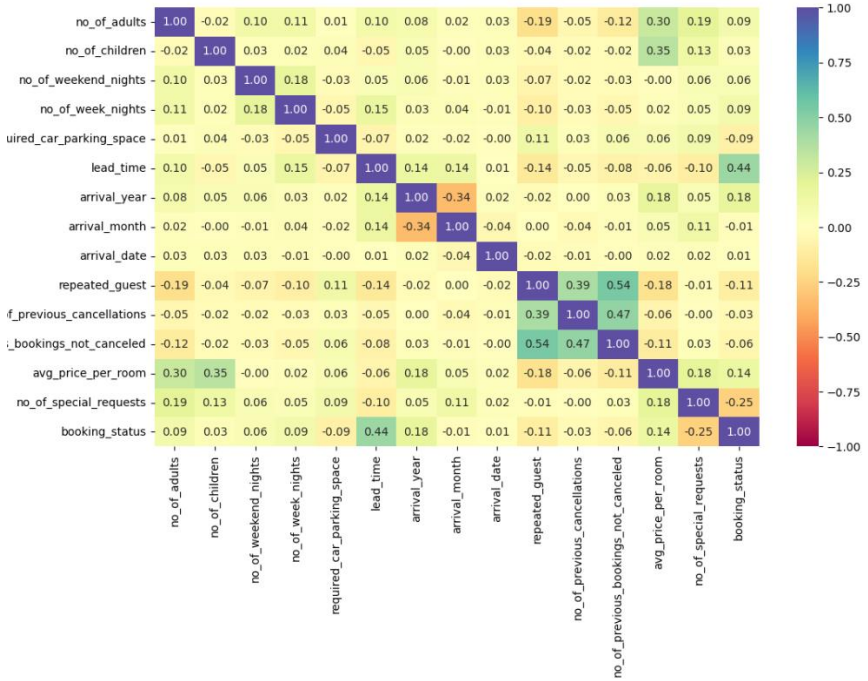
Heatmap of the Correlation Matrix

- Display of correlation coefficients

- Each cell represents the correlation value between two features

- -1 to 1 range with +1 being perfect positive correlation and -1 being perfect negative correlation. 0 being no correlation.

- Red = close to +1 and Blue = close to -1

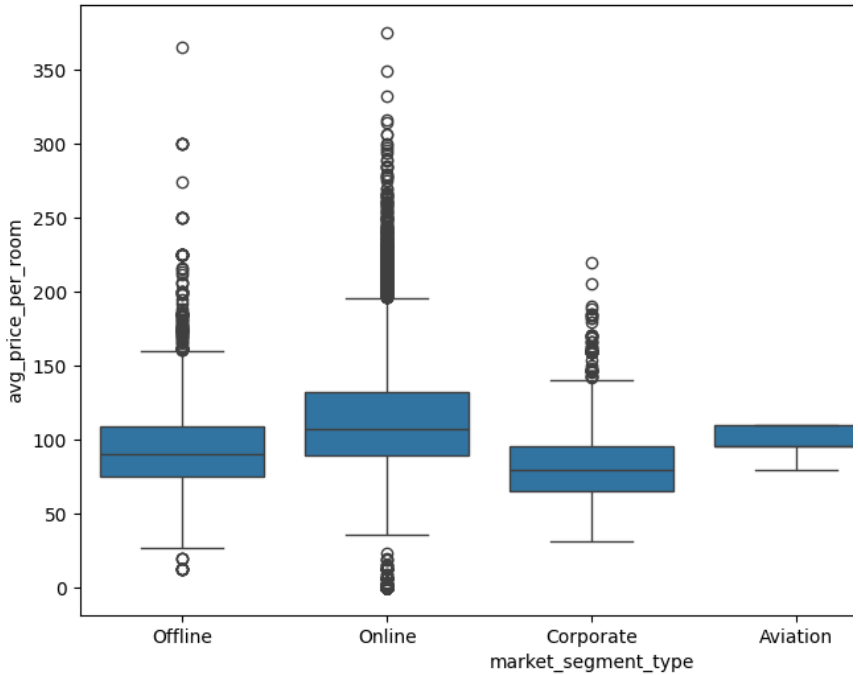


EDA Results

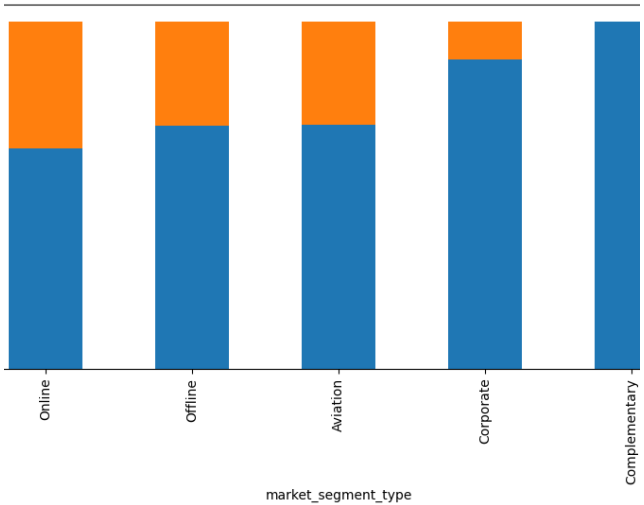
Market Segment Type

Vs.

Average Price per room

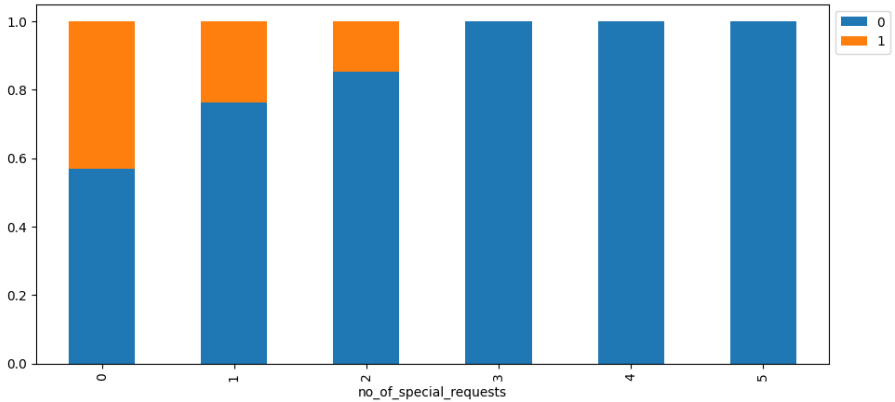


EDA Results



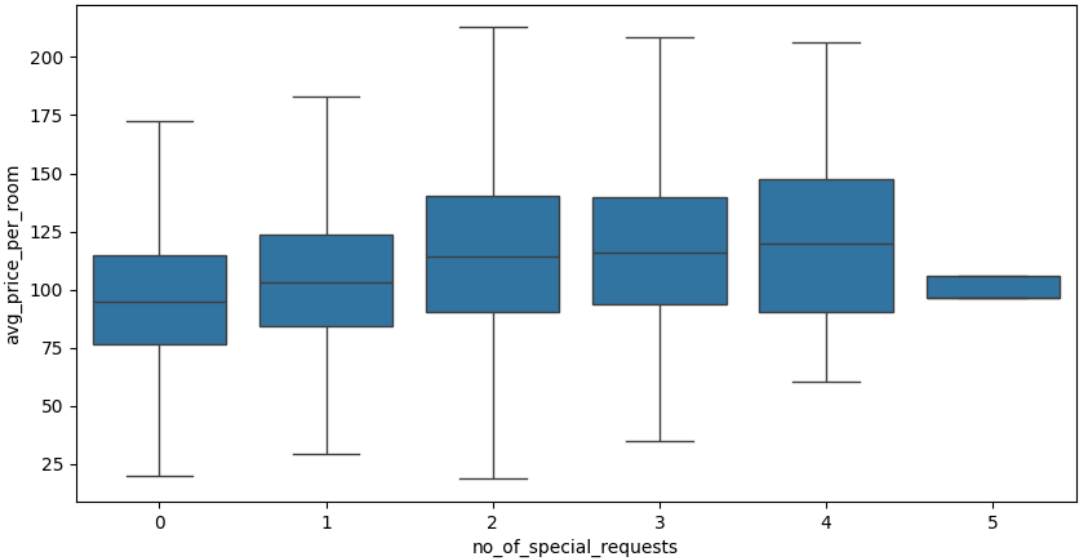
Market segment type

Number of special requests



EDA Results

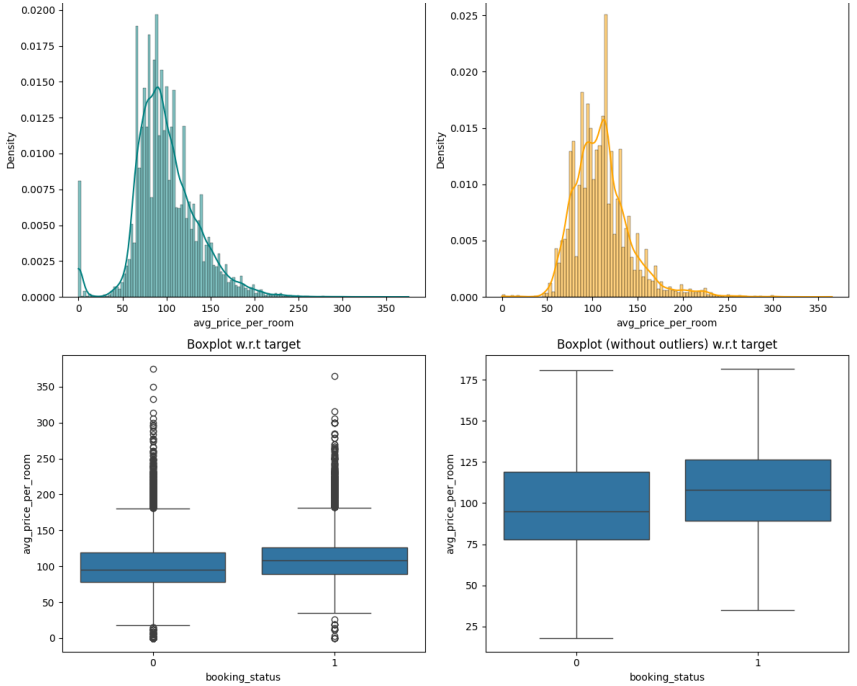
Number of special requests versus the average price per room.



EDA Results

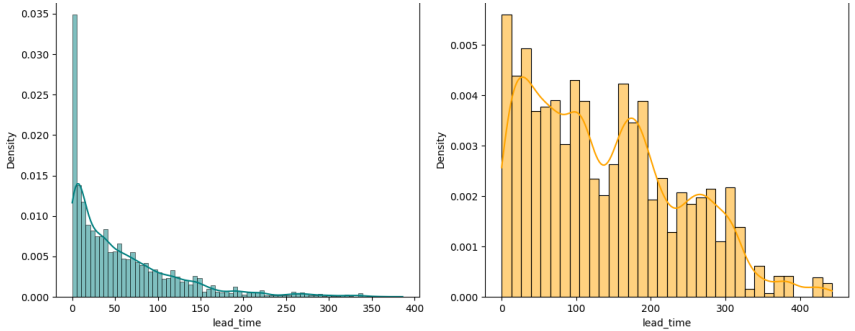
Density graphs of average price per room

Booking Status Boxplot with and without outliers.

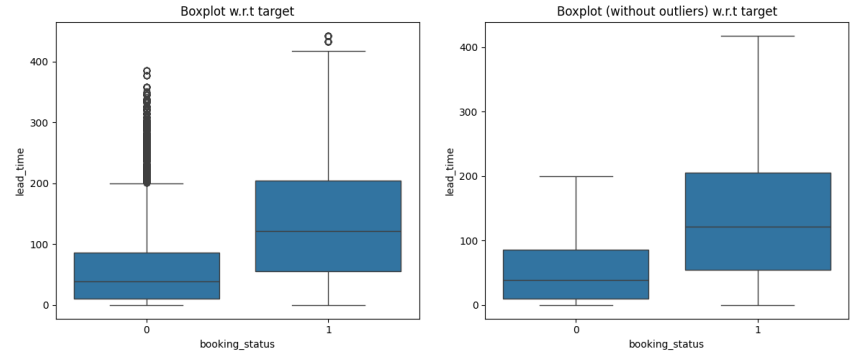


EDA Results

Lead time versus density

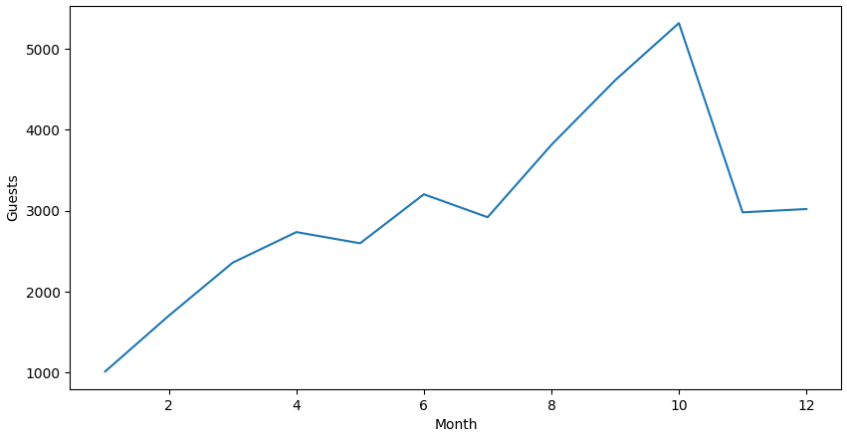


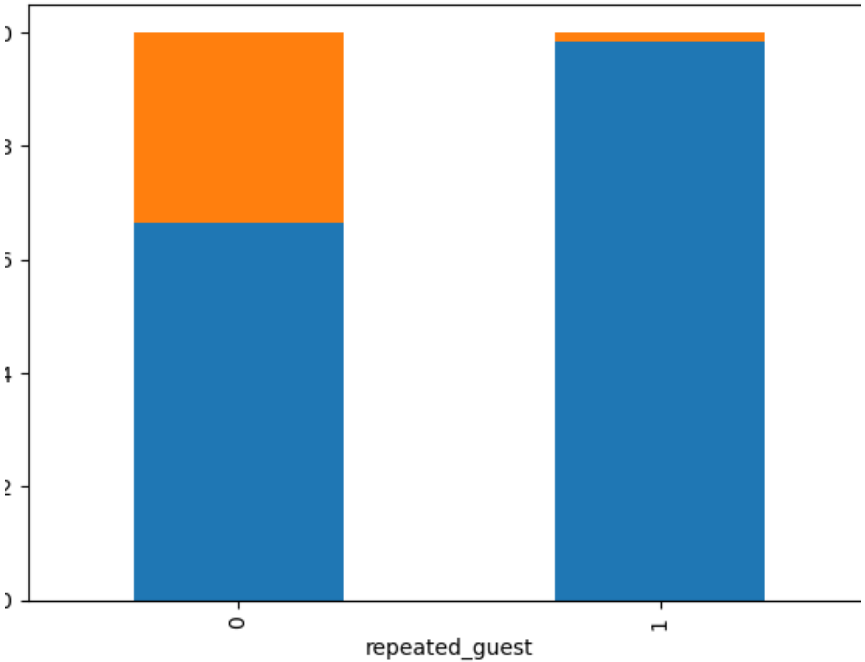
Booking status versus lead time on Boxplot



EDA Results

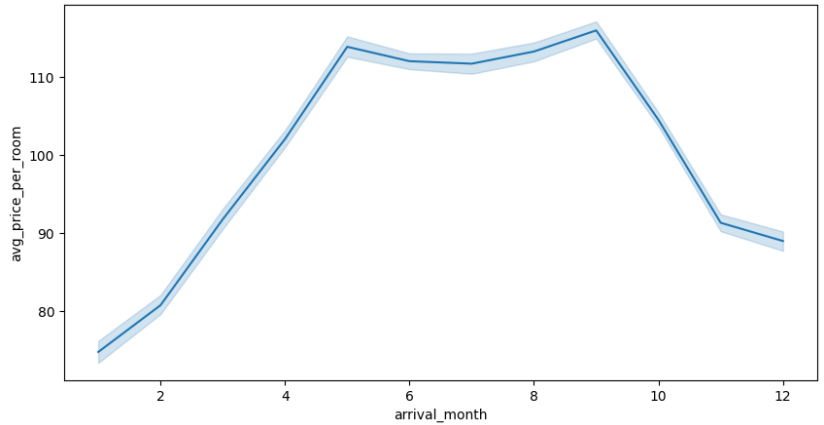
Repeated Guests and guests during different months.

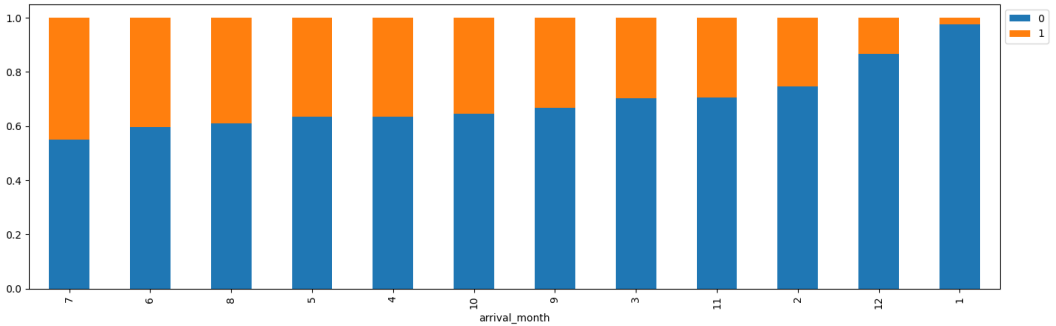




EDA Results

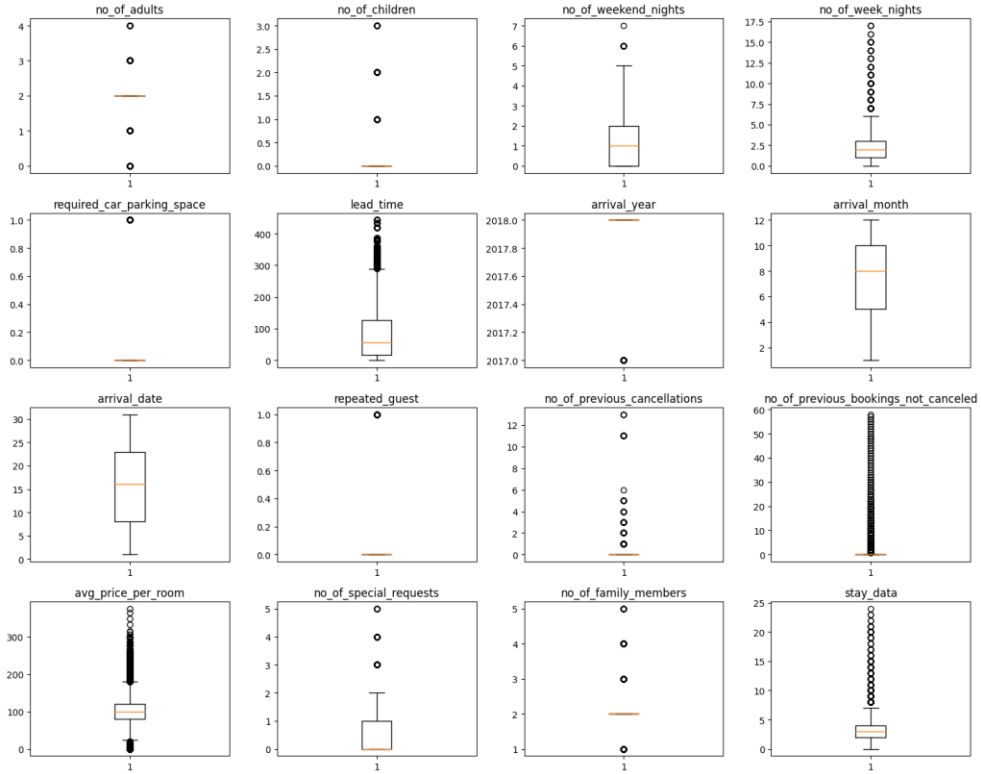
Arrival month versus average price per room.





Data Preprocessing

Boxplots of all data points.



Model Performance Summary

- Evaluation Criteria
  - Model can make wrong predictions – Predicting customer will not cancel but will cancel their booking or predicting a customer will cancel but will not actually cancel their booking.
- Determining which case is most important
  - Must avoid predicting that a booking will not be canceled, and the booking gets canceled. This is a loss of potential revenue.
  - We must also avoid that a booking will get canceled and the booking doesn't get canceled. In this case, the hotel might not be prepared for the customers and provide unsatisfactory service.
- How to reduce losses?
  - Decision tree classifier with F1 score chosen as the best combination for a parameter.
  - F1 score maximizes the chance of minimizing false negatives and false positives.
  - Model parameters:
    - **Max Depth: 5, Max Leaf Nodes: 50, Min Samples Split: 10, Random State: 1**
  - Two most important features:
    - **Degree of Spondylolisthesis and Pelvic tilt**

Model Performance Summary

- **Final Model: Logistic Regression with the following parameters:**
  - Intercept: -922.8266
  - P-Value < 0.05
    - Required Parking Space (-1.5943)
    - Lead Time (0.0157)
    - Arrival Month (-0.0417)
    - Market Segment Type (Corporate & Offline had highest significance)
  - Important Features
    - **Lead time** – the longer the lead time, the more likely the booking is to be canceled
    - **Required parking spaces** – customers needing a parking space were less likely to cancel
    - **Market segment type** – corporate and offline segments had a significant influence on the booking status

Model Performance Summary

Training Performance Comparison

Metric	Default Threshold	0.37 Threshold	0.42 Threshold
Accuracy	80.55%	79.26%	80.08%
Recall	63.28%	73.53%	69.87%
Precision	73.90%	66.83%	69.71%
F1 Score	68.18%	70.02%	69.79%

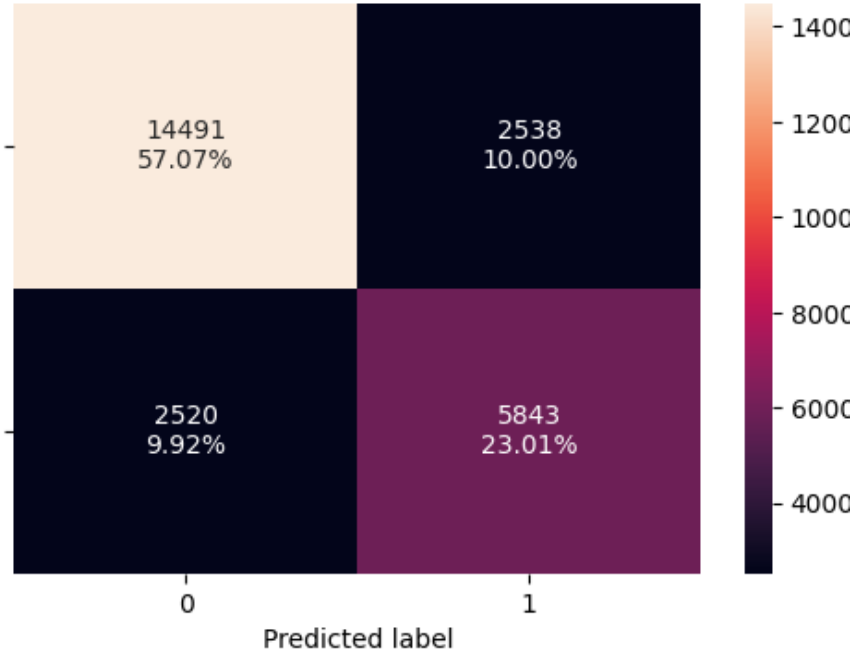
Testing Performance Comparison

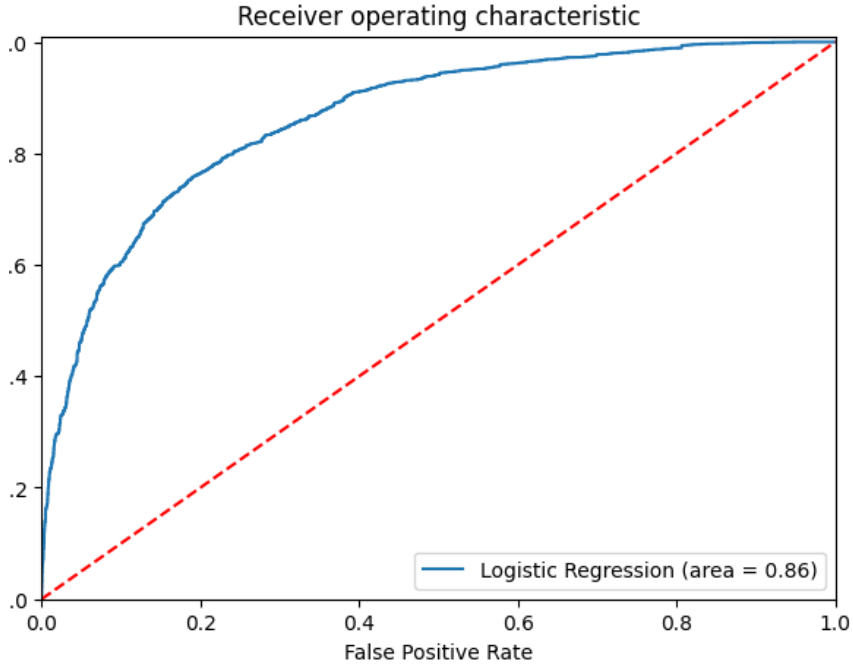
Metric	Default Threshold	0.37 Threshold	0.42 Threshold
Accuracy	80.43%	79.64%	80.34%
Recall	63.15%	73.93%	70.27%
Precision	72.78%	66.75%	69.37%
F1 Score	67.61%	70.16%	69.60%

Key Takeaways

- Increasing the threshold to 0.37 improved the recall but reduced precision.
- The best balance between recall and precision was found at the default threshold (0.5).

Model Performance Summary

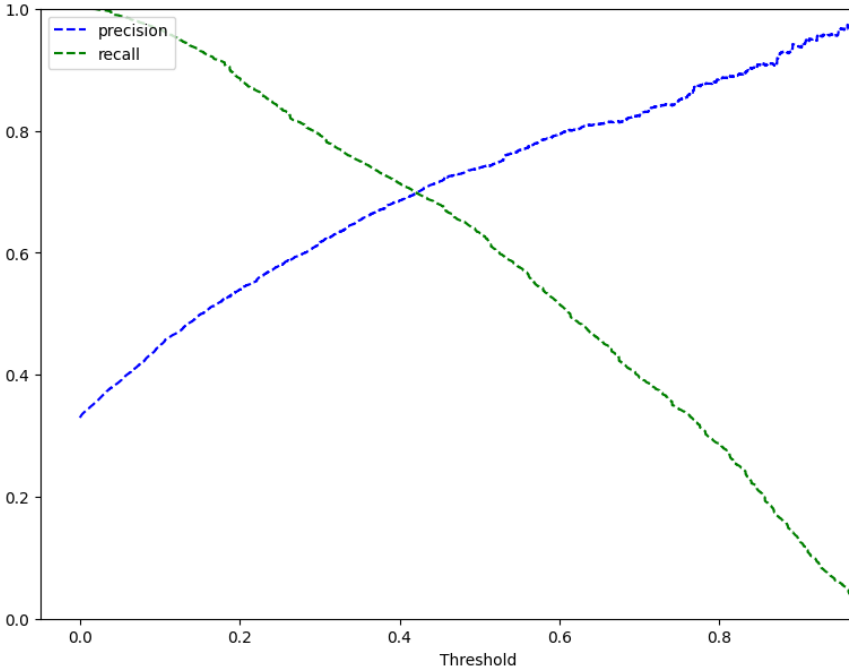




Model Performance Summary

Previous and Current page:

- Confusion matrix shows a moderate # of false positives and false negatives – room for improvement in recall
- ROC-AUC score 0.86, showing strong model performance
- Precision-Recall Curve Analysis showed adjusting the threshold could improve recall with reduced precision.



Model Performance Summary

Conclusion

- Logistic regression model performed with an **80% accuracy** and a **recall of 63%** at the default threshold.
- The best performing model was **Logistic Regression** with the default threshold (0.5).
- If the goal is to increase recall (detecting cancellations more effectively), a threshold adjustment could be made.

APPENDIX

Data Background and Contents

The data for this project was derived from INN Hotels Group of Portugal, which is facing challenges due to high booking cancellations. Those cancellations are primarily caused by schedule changes, conflicts, and last-minute cancellations. The impact of cancellations includes:

- **Loss of revenue due to unsold rooms.**
- **Higher marketing and distribution costs to fill vacancies**
- **Lower profit margins due to last-minute price reductions**
- **Operational burdens on human resources**

The objective of the analysis is to **predict booking cancellations using machine learning** – specifically logistic regression and decision trees.

Model Building - Logistic Regression

Tests conducted to check assumptions of logistic regression:

- **Multicollinearity:** Checked using VIF to ensure that predictor values are not highly correlated
- **Linearity of Log-Odds**: Used Box-Tidwell transformation to confirm numerical predictors have a linear relationship with log-odds of the dependent variable
- **Independence of errors**: Assessed using Durbin-Watson statistic to confirm that residuals are not autocorrelated
- **Absence of outliers**: Used Cook's distance to identify influential points that could distort model performance.
- **Class Balance**: Used class distribution of canceled vs. non-canceled bookings to determine the need for rebalancing techniques like SMOTE.

Model Building - Logistic Regression

Interpretation of Logistic Regression Coefficients and Odds

- **Positive coefficients**: Variables with positive coefficients increase the odds of cancellation.
  - Like higher lead times(time between booking and check-in)
  - Bookings with non-refundable rates have lower odds of cancellation
- **Negative coefficients**: Variables w/neg. coeff. reduce odds of cancellation.
  - Repeat guests and short stays are less likely to cancel
- **Odds Ratio Interpretation**:
  - Ratio of 1.5 means that for every one-unit increase in that predictor, the likelihood of cancellation increases by 50%.

Model Building - Logistic Regression

Model Performance of Logistic Regression

- **Precision & Recall**:
  - Precision measures how many of the predicted cancellations were actual cancellations
  - Recall measures how many of the actual cancellations were correctly identified
- **AUC-ROC Score**:
  - Higher AUC (>0.8) suggests strong predictive ability

Model Performance Evaluation and Improvement - Logistic Regression

Changing Classification Threshold

- **Default threshold is 0.5** – anything above 50% is a cancellation
- **Experimenting with different thresholds improves recall or precision**
  - Lower threshold increases recall but leads to more false positives
  - Higher threshold increases precision but may reduce recall

Model Building - Decision Tree

Model building steps

- **Data preprocessing**
  - Handling missing values, creating categorical variables, splitting into train and test sets
- **Feature Selection**
  - Using Gini Impurity and Information Gain
- **Model Training**
  - Fitting the decision tree classifier to the data set
- **Hyperparameter Tuning**
  - Adjusting max depth, samples split, and samples leaf

Model Performance Evaluation and Improvement - Decision Tree

Decision Tree Model Performance

- **Accuracy score**: % of correctly classified instances
- **Confusion matrix**: True positives, false positives, true negatives, and false negatives
- **Precision & Recall**
- **F1 Score**: Balance between precision and recall
- **ROC-AUC Score**: Evaluated how well the model distinguishes between cancellations and non-cancellations

Impact of Pruning

- **Pre-Pruning**: max depth, min samples split, and min samples leaf
- **Post-Pruning**: ccp alpha (remove weak branches)