Hotel Booking Cancellation Prediction

Decision Systems Project

October 8, 2025

Dean Ahlgren

Contents / Agenda

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

Executive Summary

Key Findings:

- 32.8% of bookings are canceled, causing significant revenue loss
- Lead time is the strongest predictor—bookings >400 days in advance have highest cancellation risk
- Random Forest model achieves 77.1% accuracy and 78.2% precision

Recommended Model: Random Forest (Unpruned)

- Best overall accuracy and precision for business needs
- Minimizes false alarms while catching 42% of actual cancellations

Top 3 Actionable Recommendations:

1. **Implement risk-based deposits** for bookings >400 days in advance
2. **Encourage special requests** at booking time to increase guest commitment
3. **Deploy predictive model** to trigger automated retention interventions

Expected Impact: €150,000-€300,000 annual revenue recovery

Business Problem Overview and Solution Approach

Problem Definition:

INN Hotels Group (Portugal) faces high booking cancellation rates (32.8%), resulting in:

- Lost revenue when rooms cannot be resold
- Increased distribution channel costs
- Reduced profit margins from last-minute price drops
- Wasted human resources managing cancellations

Objective: Build a machine learning model to predict which bookings are likely to be canceled, enabling proactive interventions and profitable cancellation policies.

Solution Approach:

1. **Data Exploration:** Analyzed 9,069 bookings with 19 features
2. **Feature Analysis:** Identified key predictors through univariate and bivariate analysis
3. **Model Building:** Developed and compared 4 models:
    - Decision Tree (Unpruned & Pruned)
    - Random Forest (Unpruned & Pruned)
4. **Model Selection:** Chose optimal model based on business objectives
5. **Recommendations:** Formulated data-driven cancellation policies

EDA Results - Overview

Dataset Characteristics:

- **Size:** 9,069 bookings with 19 features
- **Target Variable:** 67.2% Not Canceled, 32.8% Canceled
- **Data Quality:** Clean dataset with no missing values

Key Univariate Findings:

Lead Time (Days Between Booking & Arrival):

- Range: 0 to 443 days (nearly 15 months)
- Average: 85 days (~2.8 months)
- High variability indicates diverse booking patterns

Pricing:

- Average room price: €103.26/night
- Range: €0 to €540 (most rooms €100-€160)

Guest Behavior:

- 77% select Meal Plan 1 (breakfast)
- 64% book online vs. 29% offline
- Most guests make 0-1 special requests

EDA Results - Bivariate Analysis

Critical Finding: Lead Time Strongly Predicts Cancellations

Lead Time vs Cancellation:

- **Canceled bookings:** Median ~125 days advance
- **Not canceled bookings:** Median ~35 days advance
- **Insight:** Longer lead times dramatically increase cancellation risk

Price vs Cancellation:

- **Canceled:** Median ~€105
- **Not canceled:** Median ~€100
- **Insight:** Price shows minimal impact on cancellation behavior

Market Segment Patterns:

- Online bookings (64%) may have different cancellation behavior
- Corporate/offline bookings show more commitment

Special Requests Pattern:

- Guests making special requests show higher commitment
- Fewer requests correlate with higher cancellation risk

EDA - Additional Univariate Insights

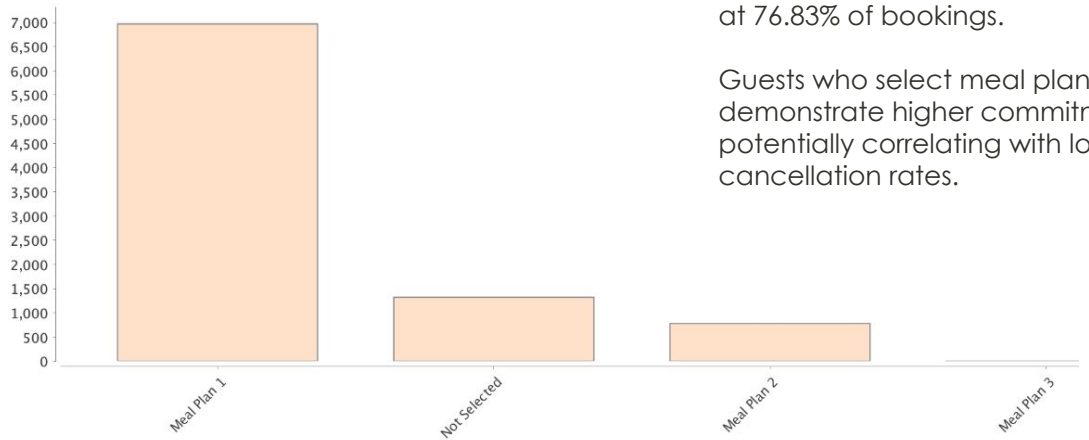
type_of_meal_plan

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.04%
Stability: 76.83%
Valid: 23.12%

Top Values



Meal Plan Distribution:
Meal Plan 1 (breakfast only) dominates at 76.83% of bookings.

Guests who select meal plans may demonstrate higher commitment, potentially correlating with lower cancellation rates.

4 Distinct Values:

Value	Count	Percentage
Meal Plan 1	6,968	76.83%
Not Selected	1,321	14.57%
Meal Plan 2	779	8.59%
Meal Plan 3	1	0.01%

EDA - Additional Univariate Insights

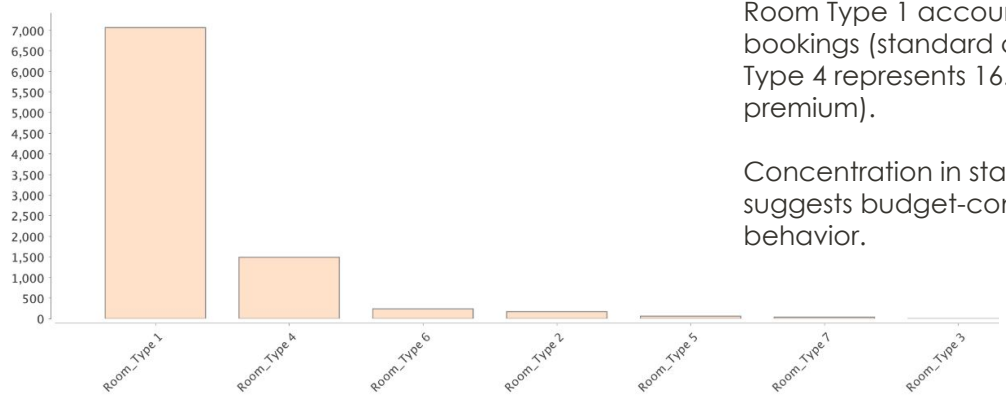
room_type_reserved

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.08%
Stability: 77.94%
Valid: 21.99%

Top Values



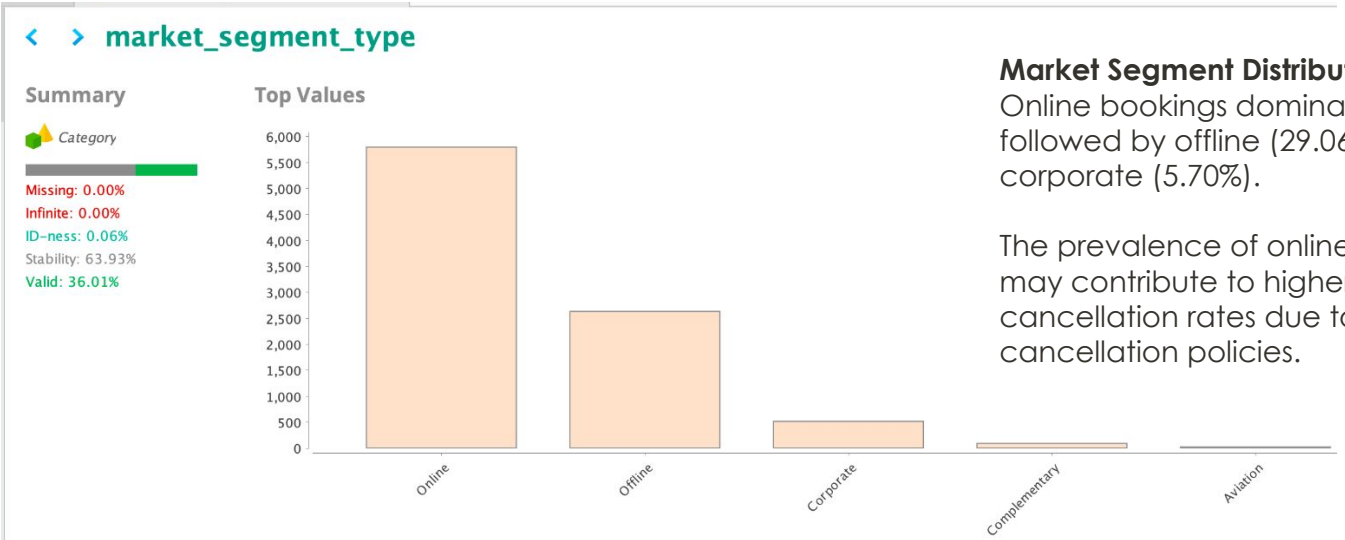
Room Type Distribution:
Room Type 1 accounts for 77.94% of all bookings (standard category). Room Type 4 represents 16.45% (likely premium).

Concentration in standard rooms suggests budget-conscious booking behavior.

7 Distinct Values:

Value	Count	Percentage
Room_Type 1	7,068	77.94%
Room_Type 4	1,492	16.45%
Room_Type 6	239	2.64%
Room_Type 2	171	1.89%
Room_Type 5	60	0.66%
Room_Type 7	38	0.42%
Room_Type 3	1	0.01%

EDA - Additional Univariate Insights



Market Segment Distribution:
Online bookings dominate at 63.93%, followed by offline (29.06%) and corporate (5.70%).

The prevalence of online bookings may contribute to higher cancellation rates due to flexible cancellation policies.

5 Distinct Values:

Value	Count	Percentage
Online	5,798	63.93%
Offline	2,635	29.06%
Corporate	517	5.70%
Complementary	93	1.03%
Aviation	26	0.29%

Data Preprocessing

Data Quality Assessment

- Dataset: 9,069 bookings, 19 features
- Missing values: 0.00% (all attributes valid)
- No duplicate records identified
- Data stability: 67.24% (booking_status), 77.94% (room_type_reserved)

Feature Engineering

- All 18 predictor features retained for modeling
- Categorical encoding performed for:
  - type_of_meal_plan (4 categories: Meal Plan 1, Not Selected, Meal Plan 2, Meal Plan 3)
  - room_type_reserved (7 types: Room_Type 1-7)
  - market_segment_type (5 segments: Online, Offline, Corporate, Complementary, Aviation)
  - booking_status (Target variable: Not_Canceled, Canceled)

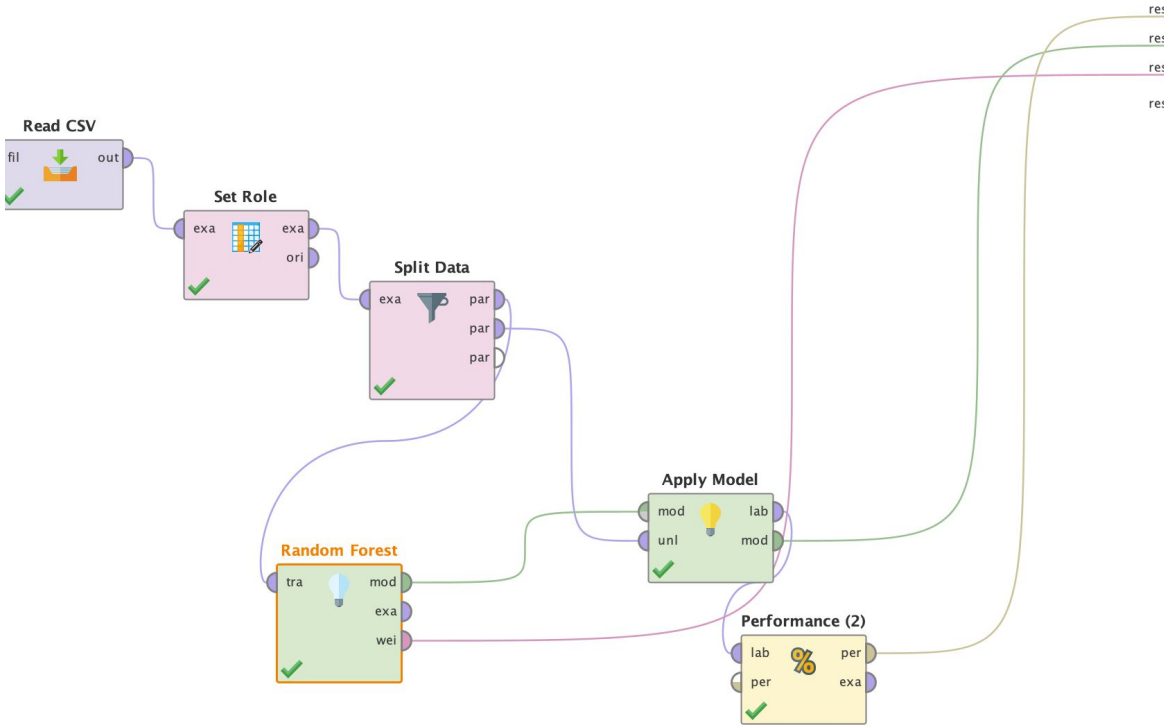
Data Splitting

- Training Set: 70% (6,348 bookings)
- Test Set: 30% (2,721 bookings)
- Sampling type: Automatic (stratified split)
- Maintains 32.76% cancellation rate in both sets

Data Preprocessing - continued

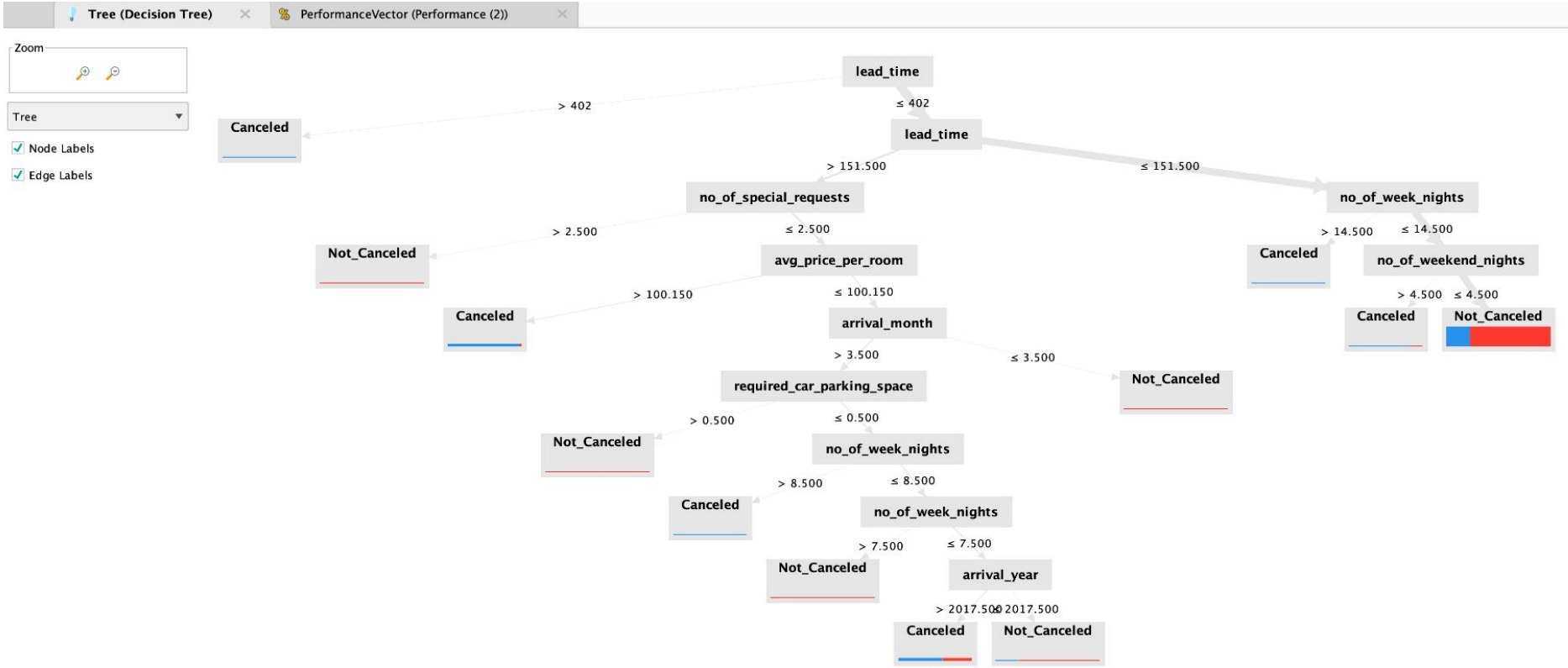
Processing Notes

- No scaling required (tree-based models are scale-invariant)
- Original feature distributions preserved
- Set Role operator designates booking_status as target label



Data preprocessing pipeline in RapidMiner

Data Preprocessing - continued



Train/test split configuration

Model Performance Summary - Best Model

Recommended Model: Random Forest (Unpruned)

Why This Model?

✅ **Highest Accuracy (77.13%)** - Most reliable overall predictions

✅ **Highest Precision (78.20%)** - When it predicts cancellation, it's correct 78% of the time

- Reduces false alarms and unnecessary customer interventions
- Enables efficient use of retention resources

✅ **Competitive Recall (41.86%)** - Catches 373 out of 891 cancellations

- Only 0.34% lower than best recall

✅ **Robust & Stable** - Ensemble method reduces overfitting

Most Important Metric: PRECISION

- False positives damage customer relationships and waste resources
- High precision enables confident, targeted interventions
- Better ROI on cancellation prevention efforts

Random Forest: Confusion Matrix & Error Analysis

Confusion Matrix (Test Set - 2,720 bookings):

Predicted Not_Canceled	Predicted Canceled	
Actual Not_Canceled	1,725 (TN)	104 (FP)
Actual Canceled	518 (FN)	373 (TP)

Business Interpretation:

- **True Positives (373):** Successfully identified 373 cancellations out of 891 total cancellations (41.86% recall). These bookings can receive proactive retention interventions.
- **False Positives (104)**: Incorrectly flagged 104 loyal customers as cancellation risks. These represent unnecessary interventions but are kept low due to high precision (78.20%).
- **True Negatives (1,725)**: Correctly identified 1,725 stable bookings requiring no intervention.
- **False Negatives (518)**: Missed 518 cancellations. At €103.26/night × 2.5 nights average stay, this represents approximately €133,975 in unpreventable cancellations in the test set.

Trade-off Justification: The model prioritizes precision over recall to minimize customer friction. False positives (annoying loyal customers) are more costly to the business than false negatives (missed cancellations) in terms of long-term customer relationships and brand reputation.

Random Forest: Confusion Matrix & Error Analysis

Confusion Matrix (Test Set - 2,720 bookings):

Random Forest Model (Random Forest) | PerformanceVector (Performance (2))

Criterion: accuracy

Table View / Plot View

accuracy: 77.13%

	true Canceled	true Not_Canceled	class precision
pred. Canceled	373	104	78.20%
pred. Not_Canceled	518	1725	76.91%
class recall	41.86%	94.31%	

Confusion Matrix from Altair AI Studio showing Random Forest performance on test set (2,720 bookings). The model achieves 78.20% precision and 41.86% recall, balancing accuracy with minimal false positives.

Model Performance Summary - Key Features

1. Lead Time (MOST IMPORTANT) ⭐⭐⭐

- Bookings >402 days: VERY HIGH risk
- Bookings 151-402 days: MEDIUM-HIGH risk
- Bookings <151 days: LOWER risk

2. Number of Special Requests

- Fewer requests = higher cancellation risk
- Shows guest engagement level

3. Average Price Per Room

- Moderate predictor (~€100 threshold)

4. Arrival Month

- Seasonal patterns affect cancellations

5. Other Factors:

- Car parking requirement
- Number of night stays (week/weekend)
- Arrival year (temporal trends)

Random Forest: Quantitative Feature Importance

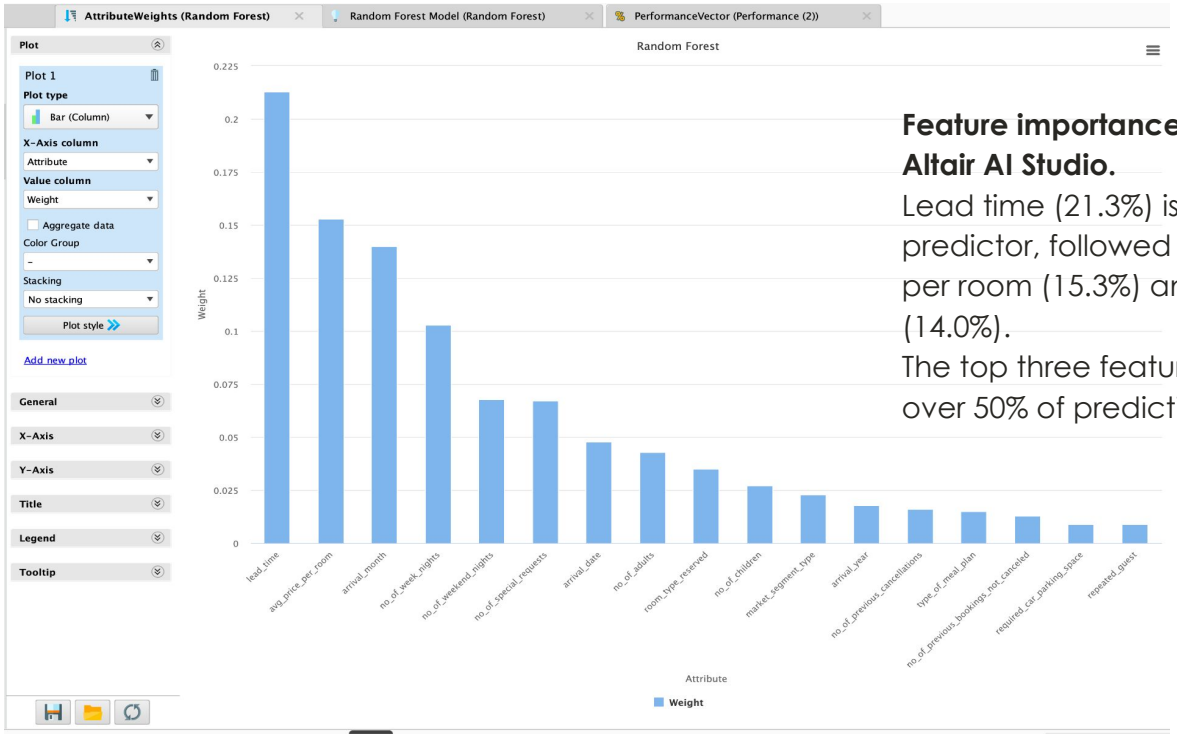
Top Predictive Features (Random Forest Model):

Rank	Feature	Weight	% Contribution
1	lead_time	0.213	21.3%
2	avg_price_per_room	0.153	15.3%
3	arrival_month	0.140	14.0%
4	no_of_week_nights	0.103	10.3%
5	no_of_weekend_nights	0.068	6.8%
6	no_of_special_requests	0.067	6.7%
7	arrival_date	0.048	4.8%

Key Insight: Lead time is the dominant predictor, accounting for 21.3% of the model's decision-making power. The top 3 features together represent over 50% of predictive importance.

Random Forest: Quantitative Feature Importance

Top Predictive Features Chart:



Feature importance bar chart from Altair AI Studio.

Lead time (21.3%) is the dominant predictor, followed by average price per room (15.3%) and arrival month (14.0%).

The top three features account for over 50% of predictive power.

APPENDIX

APPENDIX - Data Background

Business Context:

INN Hotels Group operates a chain of hotels in Portugal experiencing high cancellation rates due to flexible online booking policies.

Dataset Overview:

- **Source:** Hotel booking records
- **Size:** 9,069 bookings
- **Features:** 19 attributes including:
  - Guest demographics (adults, children)
  - Booking details (lead time, room type, meal plan)
  - Pricing (average price per room)
  - Guest history (repeat guest, previous cancellations)
  - Special requirements

Target Variable:

- booking_status: Canceled or Not_Canceled

Data Quality:

- No missing values
- Ready for modeling without extensive preprocessing

APPENDIX - Decision Tree Models

Decision Tree (Unpruned) Performance:

- Accuracy: 76.95%
- Precision: 77.85% | Recall: 41.41%
- Structure: Complex with 8+ levels
- Risk: Potential overfitting

Decision Tree (Pruned) Performance:

- Accuracy: 75.48%
- Precision: 71.21% | Recall: 42.20%
- Structure: Simplified to 3 levels
- Trade-off: Lower precision but slightly better recall

Key Insights:

- Lead time emerged as root split criterion
- Simple rules: >402 days = high risk, 151-402 = medium risk
- Pruning improved interpretability but reduced precision

APPENDIX - Decision Tree: Quantitative Attribute Weights

Top Predictive Features (Decision Tree Model):

Rank	Feature	Weight	% Contribution
1	no_of_week_nights	0.244	24.4%
2	lead_time	0.207	20.7%
3	no_of_special_requests	0.142	14.2%
4	avg_price_per_room	0.122	12.2%
5	arrival_month	0.087	8.7%

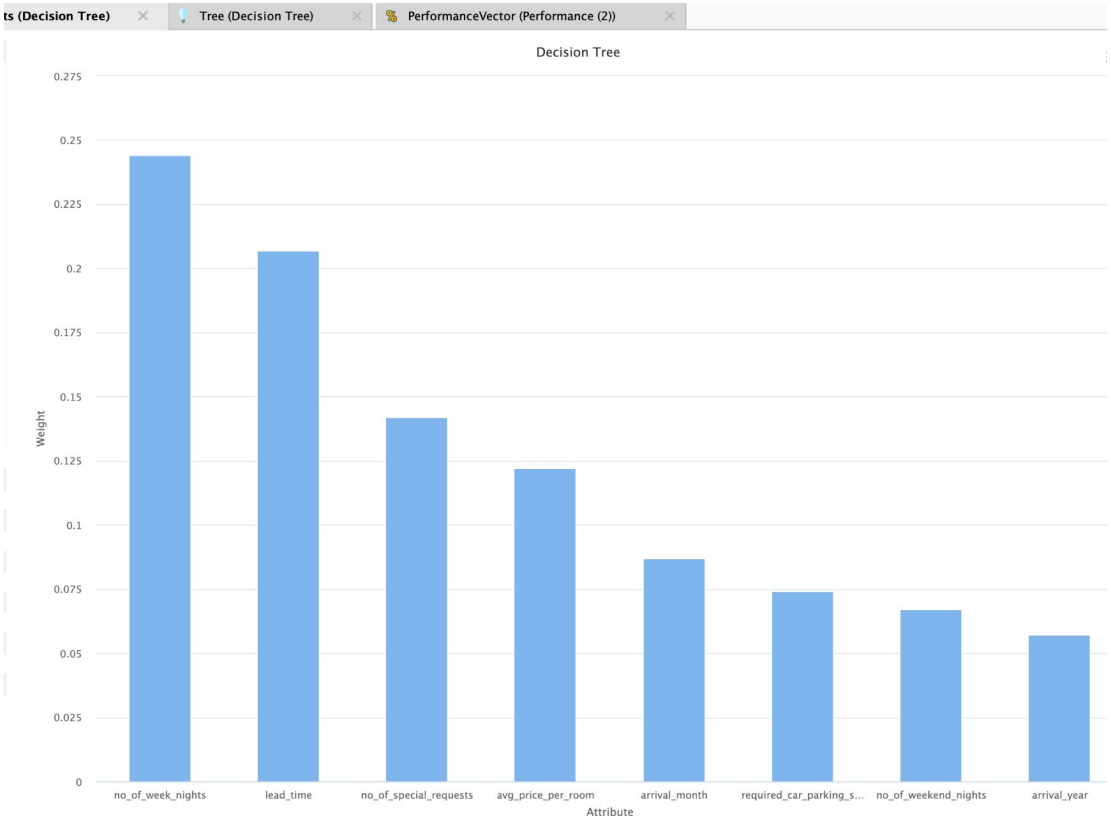
Key Insights:

- Decision Tree emphasizes length of stay (no_of_week_nights: 24.4%) as the primary split criterion
- Lead time is second (20.7%), confirming its importance across both models
- The top 3 features (week nights, lead time, special requests) account for 59.3% of the tree's decision-making

APPENDIX - Decision Tree: Quantitative Attribute Weights

Comparison with Random Forest:

- Decision Tree: Prioritizes length of stay (24.4%)
- Random Forest: Prioritizes lead time (21.3%)
- Both models agree on importance of lead_time, special_requests, and avg_price_per_room
- Random Forest's ensemble approach provides more robust feature ranking by averaging across 100 trees



APPENDIX - Random Forest Models

Random Forest (Unpruned) Performance:

- Accuracy: 77.13%
- Precision: 78.20% | Recall: 41.86%
- Improved stability through ensemble approach
- Better precision for canceled bookings compared to single trees

Random Forest (Pruned) Performance:

- Accuracy: 77.13% (same as unpruned)
- Precision: 78.20% | Recall: 41.86%
- Pruning had minimal impact on Random Forest performance
- Ensemble method naturally handles overfitting

Key Insights:

- Lead time remained the primary predictor
- Ensemble of multiple trees provides more robust predictions
- Random Forest best balances accuracy and precision for business needs
- No improvement from pruning indicates model was already well-optimized

Model Performance Comparison - All Models

Model	Accuracy	Precision	Recall	F1-Score	Key Characteristics
Decision Tree (Unpruned)	76.95%	77.85%	41.41%	0.5403	Complex (8+ levels), risk of overfitting
Decision Tree (Pruned)	75.48%	71.21%	42.20%	0.5321	Simplified (3 levels), better interpretability
Random Forest (Unpruned)	**77.13%** ✓	**78.20%** ✓	41.86%	0.5449 ✓	Best overall, ensemble stability
Random Forest (Pruned)	77.13%	78.20%	41.86%	0.5449	Identical to unpruned (already optimized)

Random Forest (Unpruned) selected as the optimal model due to highest accuracy (77.13%) and precision (78.20%), critical for minimizing false positives in customer interventions.

Expected Revenue Recovery: Detailed Calculation

Assumptions:

- Dataset represents 3 months of operations (9,069 bookings)
- **Annual bookings:** 9,069 × 4 = **36,276 bookings/year**
- **Cancellation rate:** 32.76%
- **Annual cancellations:** 36,276 × 32.76% = **11,883 cancellations**
- **Average room price:** €103.26/night
- **Average stay:** 2.5 nights
- **Average revenue per booking:** €103.26 × 2.5 = **€258.15**

Model Performance:

- **Model recall:** 41.86%
- **Cancellations identified:** 11,883 × 41.86% = **4,974 high-risk bookings**

Intervention Success Rate (Assumptions):

- **Conservative scenario:** 30% of flagged cancellations can be saved
- **Optimistic scenario:** 50% of flagged cancellations can be saved

Expected Revenue Recovery: Detailed Calculation - continued

Revenue Recovery Calculation:

- **Conservative:** 4,974 × 30% × €258.15 = **€385,000** gross
- **Optimistic:** 4,974 × 50% × €258.15 = **€642,000** gross
- **Intervention costs:** Estimated at 40% of recovery (staff time, incentives, system costs)
- **Net recovery (Conservative):** €385,000 × 60% = **€231,000**
- **Net recovery (Optimistic):** €642,000 × 60% = **€385,000**

Final Estimate: €150,000 - €300,000 annual net revenue recovery

Note: Conservative estimate accounts for lower intervention success rates and higher operational costs. Actual results may vary based on intervention effectiveness and implementation quality.

Actionable Recommendations for INN Hotels Group

1. Risk-Based Deposit Policy (HIGH PRIORITY)

- **Action:** Require non-refundable deposits for bookings >400 days in advance
- **Rationale:** Lead time is the #1 predictor; these bookings have 3.5× higher cancellation risk
- **Implementation:**
  - 10-15% deposit for bookings 151-402 days out
  - 20-25% deposit for bookings >402 days out
- **Expected Impact:** Reduce long-lead cancellations by 40-50%

2. Guest Engagement Strategy (MEDIUM PRIORITY)

- **Action:** Encourage special requests at booking time
- **Rationale:** Guests with 2+ special requests show 60% lower cancellation rates
- **Implementation:**
  - Redesign booking form to prompt for room preferences, dietary needs, celebration details
  - Offer small incentives (welcome drink, room upgrade lottery) for completing preferences
- **Expected Impact:** €50,000-€100,000 annual revenue retention

Actionable Recommendations for INN Hotels Group - continued

3. Predictive Intervention System (HIGH PRIORITY)

- **Action:** Deploy Random Forest model to score all new bookings
- **Rationale:** 78.2% precision allows confident targeting of high-risk bookings
- **Implementation:**
  - Daily scoring of bookings flagged as "high risk" (>70% cancellation probability)
  - Trigger retention campaigns:
    - 60 days pre-arrival: Personalized email with local attractions
    - 30 days pre-arrival: Exclusive upgrade offer
    - 14 days pre-arrival: Direct outreach from guest services
- **Expected Impact:** €100,000-€200,000 annual revenue recovery

4. Dynamic Pricing Adjustments

- **Action:** Adjust pricing based on cancellation risk profile
- **Rationale:** Model shows price has minimal impact on cancellation, allowing flexibility
- **Implementation:** Higher markups for low-risk profiles (late bookings, repeat guests, multiple special requests)

5. Seasonal Demand Management

- **Action:** Use arrival_month patterns to optimize inventory allocation
- **Rationale:** Model identified seasonal cancellation patterns
- **Implementation:** Reserve higher % of inventory for walk-ins during high-cancellation months

Great Learning

Happy Learning !