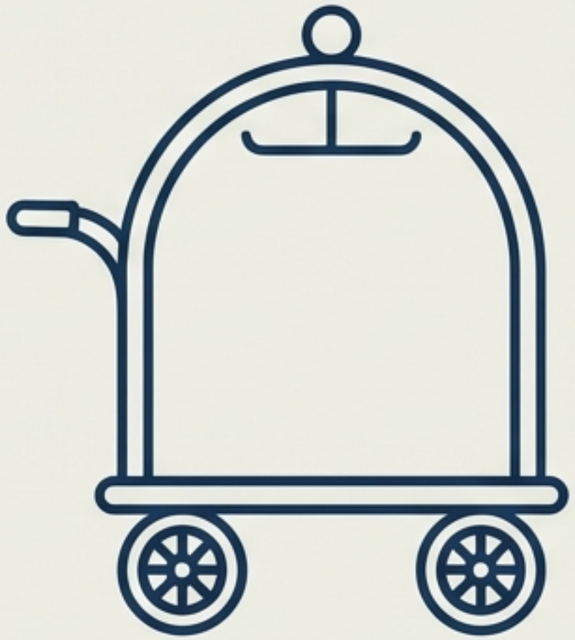
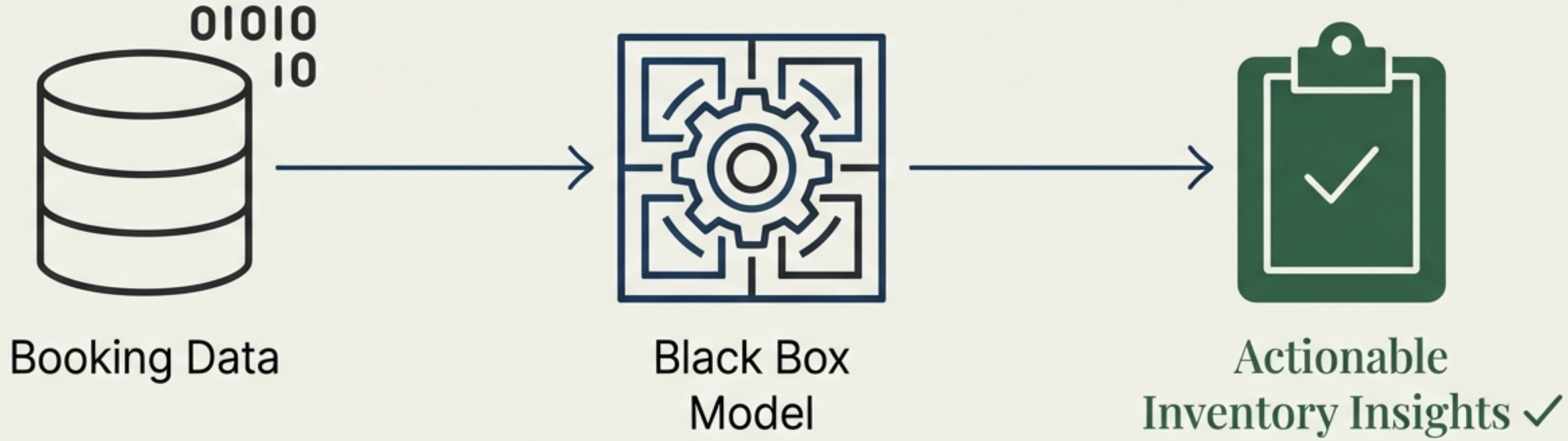
From Data to Decision: Precision in Hotel Booking Prediction

A Comparative Analysis of Machine Learning Models for Inventory Management.



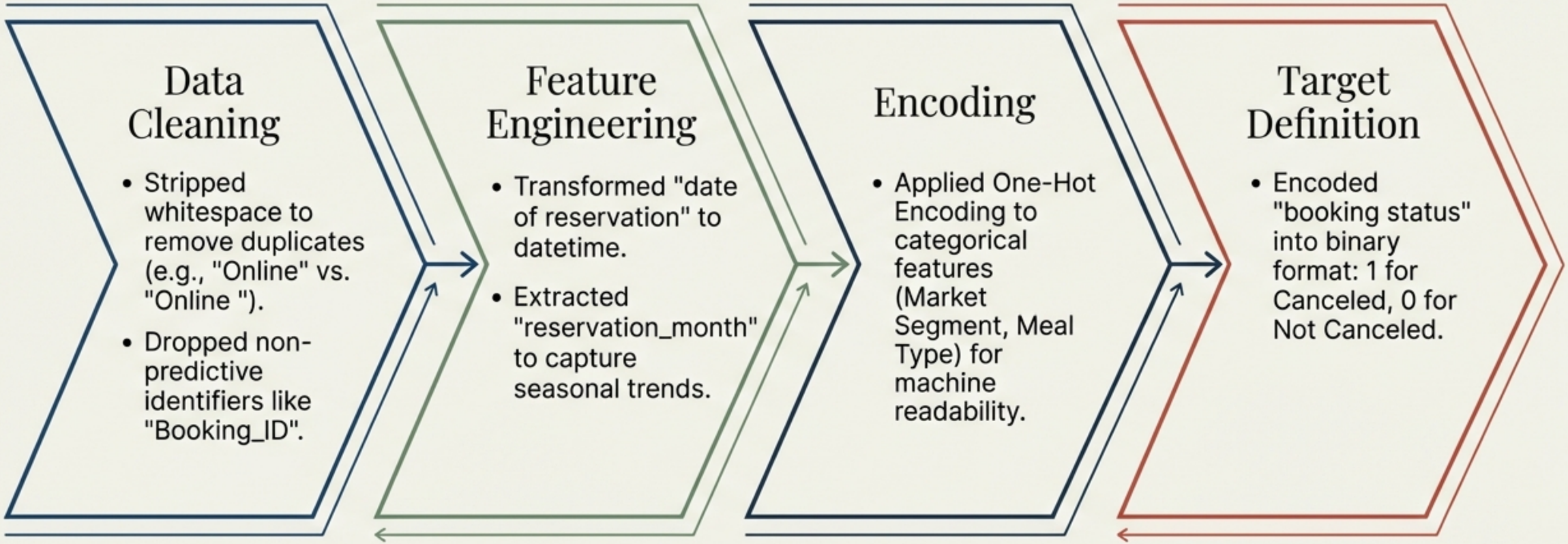
Hotels face a persistent binary problem: Will a guest arrive, or will they cancel? This deck outlines a data-driven approach to solving this inventory puzzle using advanced classification analysis.

The Objective: Solving the Inventory Puzzle



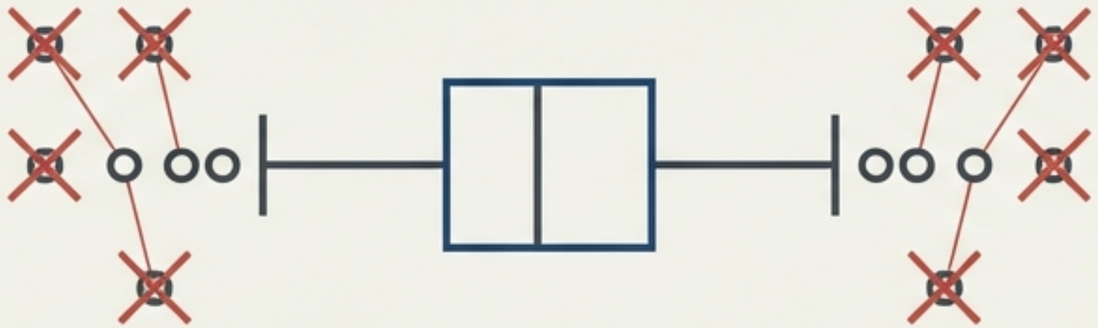
The goal is to predict the ‘Canceled’ (1) vs. ‘Not Canceled’ (0) status of bookings. By accurately forecasting these states, we enable the hotel to anticipate revenue gaps and manage overbooking strategies with confidence.

The Foundation: Engineering the Pipeline



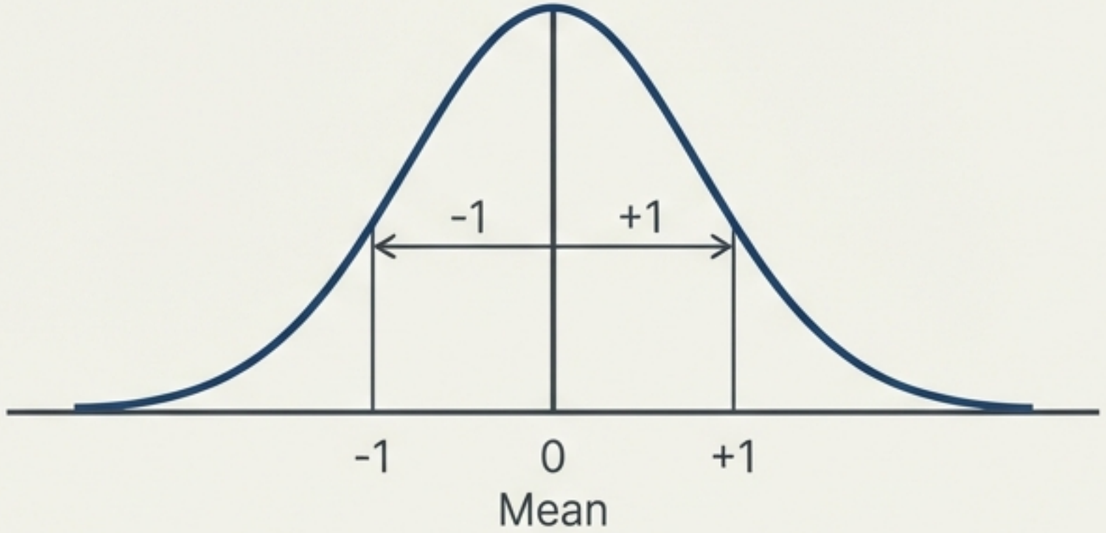
Quality Control: Outliers and Scaling

Outlier Removal



Used the Interquartile Range (IQR) method to detect and remove extreme values in ‘lead time’ and ‘average price’. Rationale: Ensures extreme values do not skew the model’s learning process.

Standard Scaling



Applied StandardScaler to normalize features to Mean=0 and Variance=1. Rationale: Critical for algorithms like Logistic Regression to converge efficiently.

The Contenders: Three Approaches to Classification

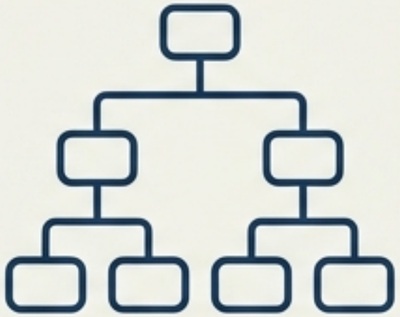
Logistic Regression

The Baseline

A statistical model assuming a linear relationship between input features and the log-odds of the outcome.

Decision Tree

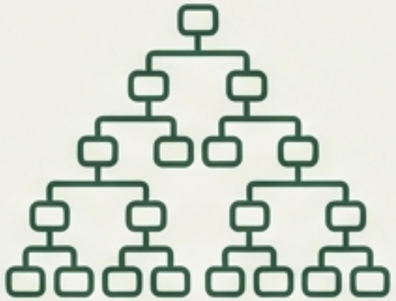
The Non-Linear



A non-parametric model that splits data into subsets. Captures non-linear relationships but is prone to overfitting.

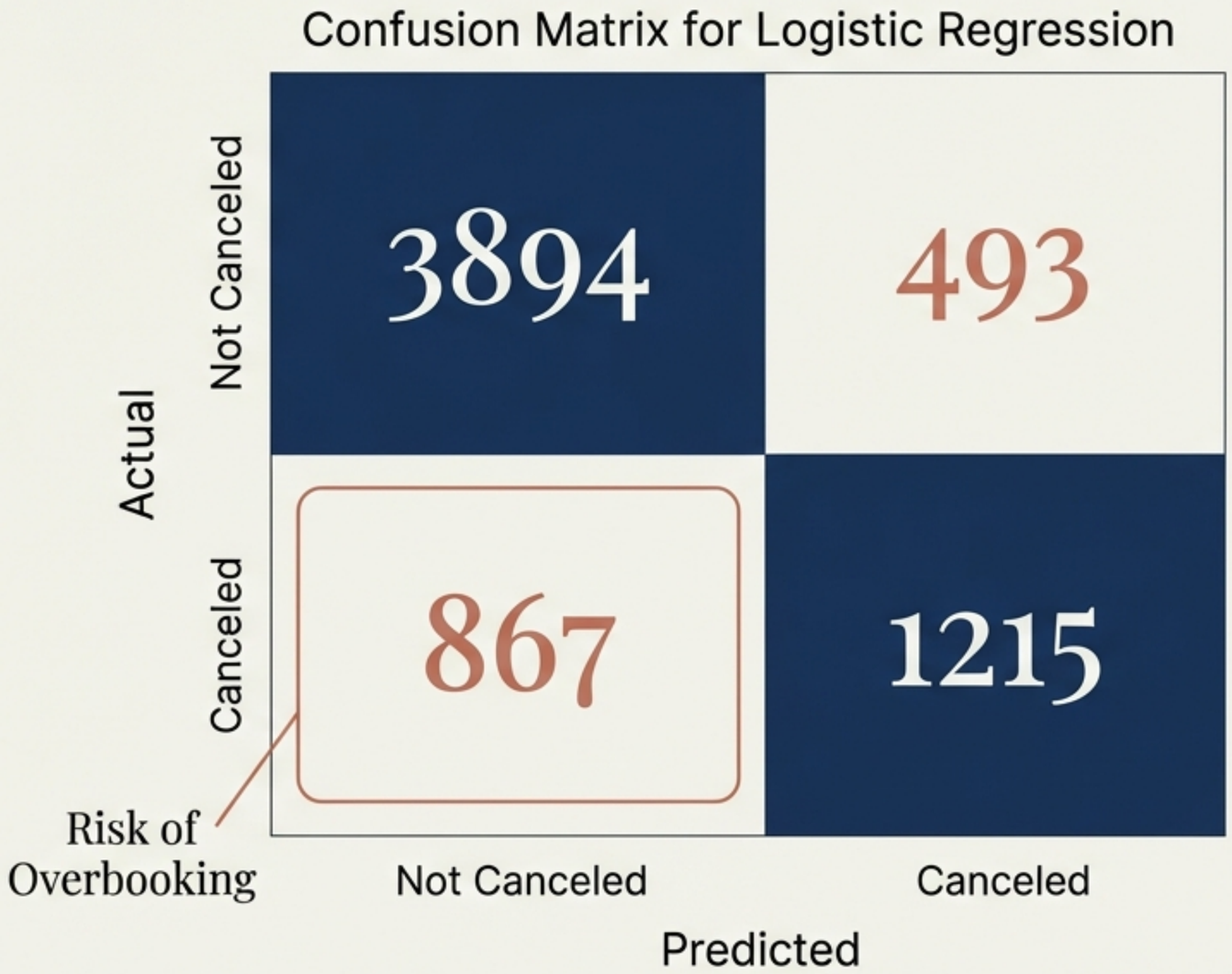
Random Forest

The Ensemble



Constructs a multitude of decision trees at training time and outputs the mode (vote) of the classes. Reduces overfitting.

The Baseline: Logistic Regression Performance



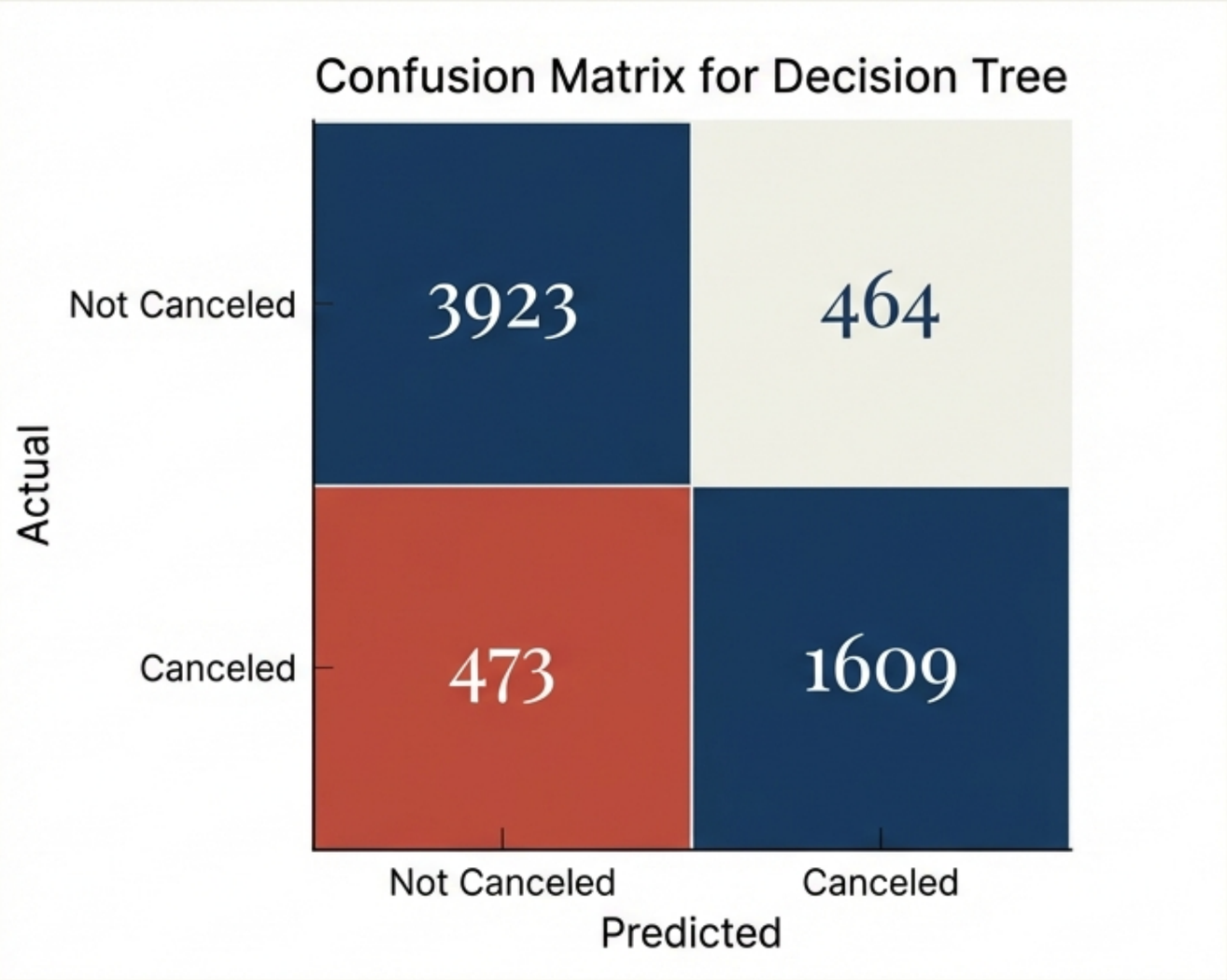
Accuracy: 78.98%

Precision: 71.14%

F1-Score: 64.12%

The matrix reveals a high number of False Negatives (predicting "Not Canceled" when the guest actually canceled). This is the riskiest model for a hotel; it fails to catch many actual cancellations.

The Step Up: Decision Tree Classifier



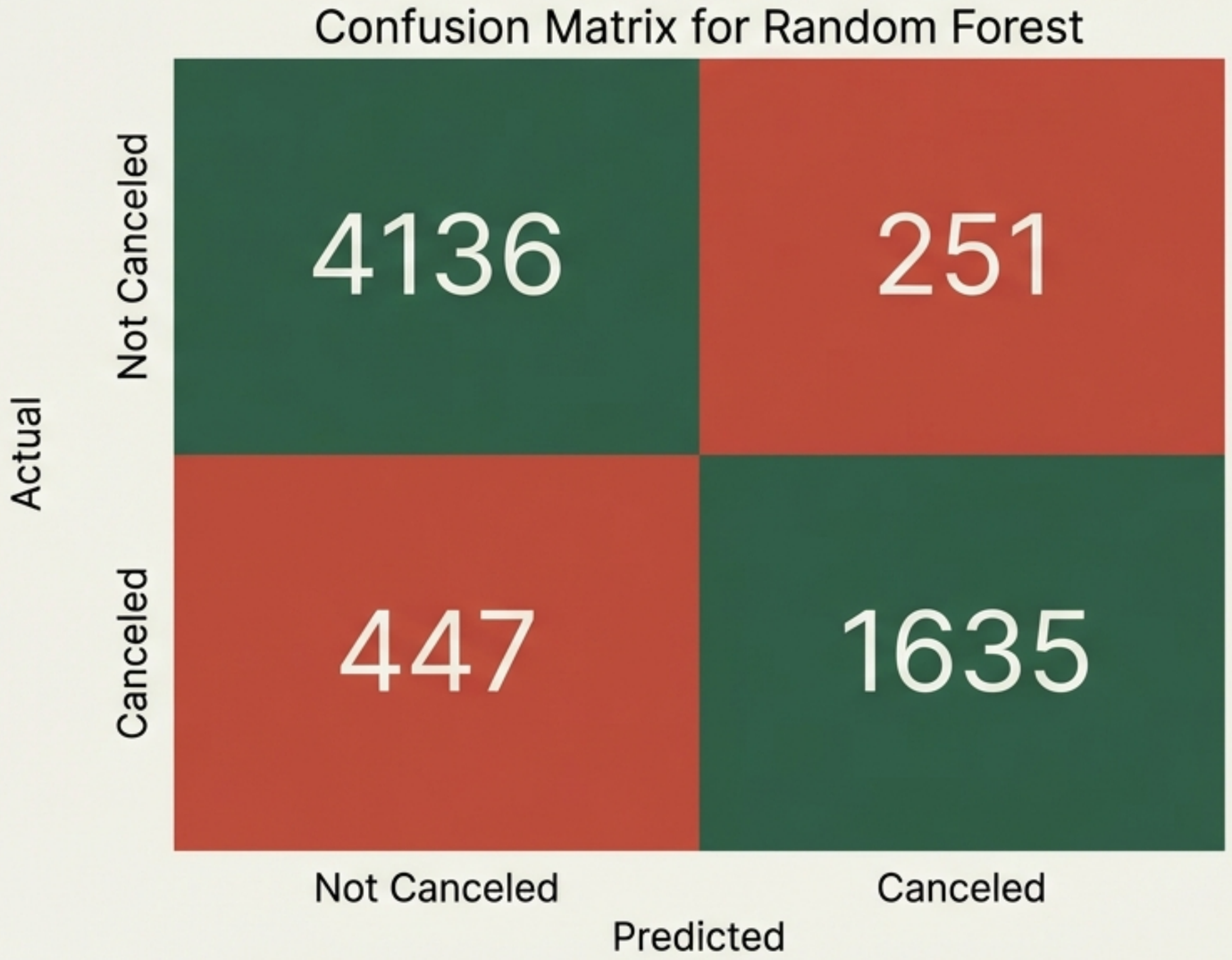
Accuracy: 85.52%

Precision: 77.62%

F1-Score: 77.45%

By handling non-linear relationships, the Decision Tree outperforms the baseline. The False Negative rate drops significantly. However, single decision trees are often unstable and prone to memorizing training data rather than generalizing.

The Champion: Random Forest Classifier



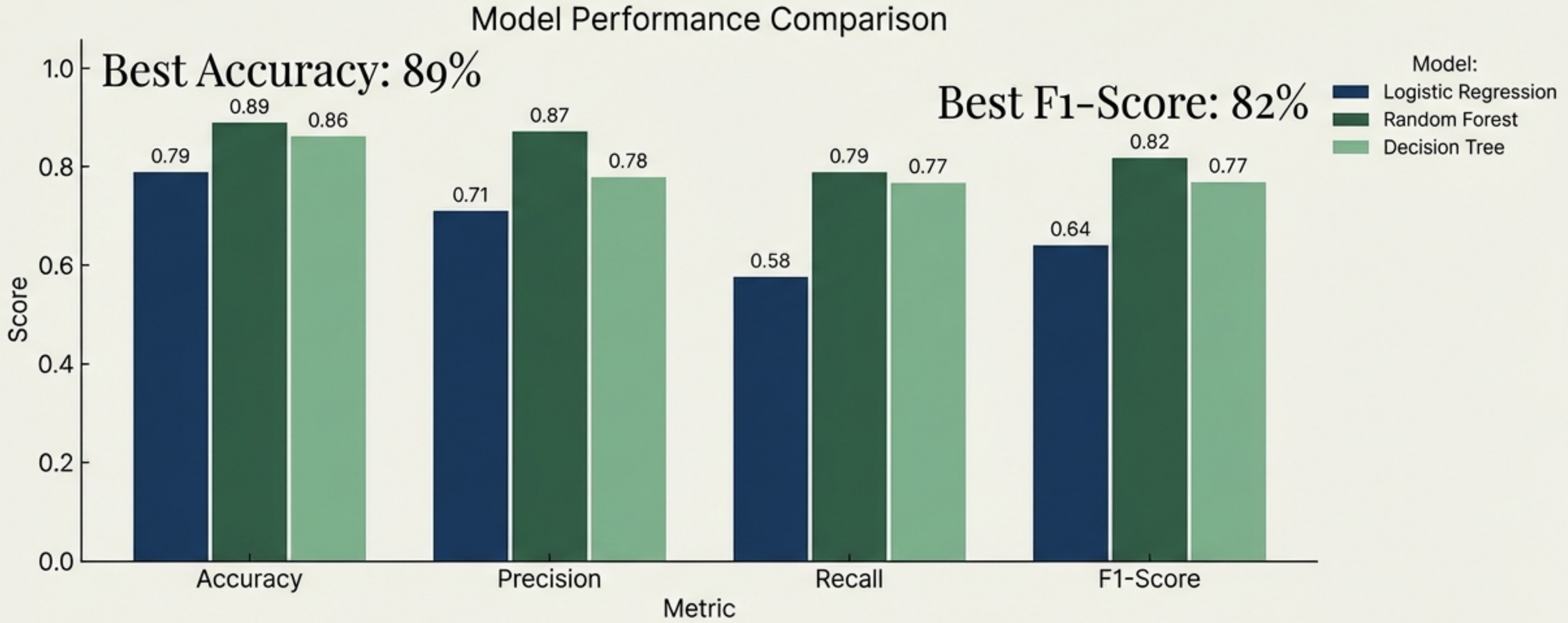
Accuracy: 89.21%

Precision: 86.69%

F1-Score: 82.41%

The ensemble method provides the most robust predictions. Note the 86.69% Precision—when this model predicts a cancellation, it is highly likely to be correct.

Head-to-Head Comparison



Random Forest (Sage Green) consistently exceeds other models in every metric, validating the choice of ensemble methods over linear or single-tree approaches.

Verdict and Strategic Recommendation

Scorecard

The Random Forest Classifier is the verified choice for deployment.

- Achieved highest Accuracy (89.21%).
- Best handling of complex, non-linear data patterns.
- Highest Precision means fewer false alarms.

Business Impact

This model allows the hotel to reliably predict cancellations. Management can now safely adjust overbooking limits and target marketing efforts to fill gaps, minimizing lost revenue from empty rooms.