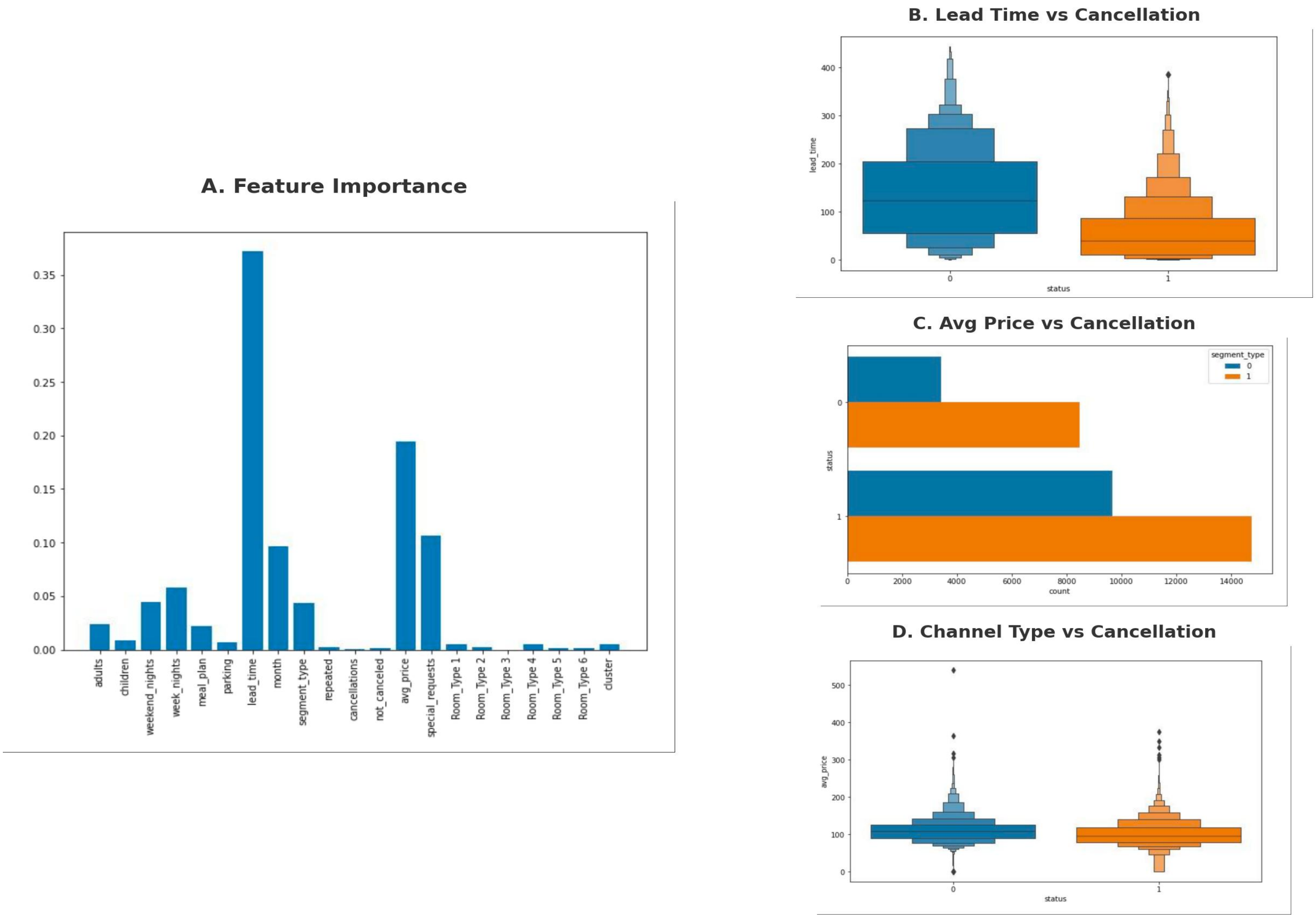
Predicting Hotel Reservation Cancellations

Link to GitHub: https://github.com/Rosen23/Hotel-Reservation-Cancellation-Project

Main Figure



B. Lead Time vs Cancellation

A. Feature Importance

C. Avg Price vs Cancellation

D. Channel Type vs Cancellation

Legend

Main Panel – Feature Importance

- Bars: Represent individual features included in the Random Forest model.
- Bar height: Indicates the relative importance (contribution) of each feature to the prediction.
- X-axis: Feature names.
- Y-axis: Feature importance score (normalized).
- Color: Uniform color; does not represent categories, only the magnitude of importance.

Side Panel B – Lead Time

- X-axis: Booking status (0 = canceled, 1 = not canceled).
- Y-axis: Lead time in days (number of days booked in advance).
- Layered boxes: Represent multiple quantile ranges in the distribution.

・Thick central box: Middle 50% of the data (IQR).
・Narrow upper/lower boxes: Higher and lower quantile ranges.
・Color: Single color representing the distribution of each group.

Side Panel C – Room Price
・X-axis: Booking status (0 = canceled, 1 = not canceled).
・Y-axis: Average room price per night.
・Layered boxes: Show the distribution of values across quantiles.
・Median line: Indicates the central tendency of the price distribution.
・Color: Same color scheme used for consistency; not representing categories.

Side Panel D – Channel Type
・Bars (horizontal bars): Represent the number of bookings in each booking status category.
・X-axis: Count of bookings (frequency).
・Y-axis: Booking status (0 = canceled, 1 = not canceled).
・Colors: Represent booking channel type (0 = offline, 1 = online).
・Bar length: Indicates how many bookings belong to each combination of status × channel type.

Key Findings

Customer Behavior Insights
・Lead time is strongly associated with higher cancellation probability.
・Higher average room prices correlate with greater likelihood of cancellation, suggesting price-sensitive behaviors.
・Customers with more special requests are less likely to cancel.
・Online bookings have significantly higher cancellation rates compared to offline bookings.
・Seasonal trends show that cancellation rates vary across peak travel months.

Cluster Insights (K-Means)
・K-Means clustering identifies four distinct customer groups differing in children count, booking channels, room type preferences, and average price levels.
・One cluster represents family-oriented travelers (more children, larger room types, higher prices).
・Another cluster represents mid – high price customers (room type 5 preference, stable lower cancellation rate).

Data & Method

Data Source: Dataset collected via kaggle “Hotel Reservations Dataset” dataset.
Total records: 36275
Variables include:

・Guest details (adults, children)
・Booking structure (room type, meal plan)
・Behavioral variables (lead_time, previous cancellations)
・Pricing (avg_price_per_room)
・Temporal variables (arrival_month)
・Target variable: booking_status (Canceled / Not Canceled)

Preprocessing
・Removed irrelevant fields (e.g., ID)
・Converted meal_plan to ordered numerical values (0 – 3)
・Converted status and segment_type to binary variables
・Standardized features for K-Means clustering

Method 1: K-Means Clustering
・Conducted on standardized numerical features
・Elbow method was inconclusive → Silhouette Score used
・Optimal number of clusters: k = 4
・Cluster labels appended as additional features

Method 2: Random Forest Classification
・Train-test split: 75% / 25%
・Hyperparameter tuning with GridSearchCV
・Best model achieved with n_estimators ≈ 40
・Test accuracy: 0.8967

Top predictive features:
1.lead_time
2.avg_price
3.special_requests
4.arrival_month
5.segment_type

Significance Statement

This figure is significant because it provides a comprehensive view of the primary determinants of hotel booking cancellations by combining feature importance with granular behavioral distributions. Understanding how variables such as lead time, price sensitivity, and booking channels relate to cancellation outcomes is essential for optimizing hotel operations. The visualization supports data-driven decision-making by identifying high-impact predictors, helping hotels refine pricing policies, allocate inventory more efficiently, and reduce revenue volatility caused by last-minute cancellations.
High cancellation rates create uncertainty in capacity planning and reduce expected revenue.