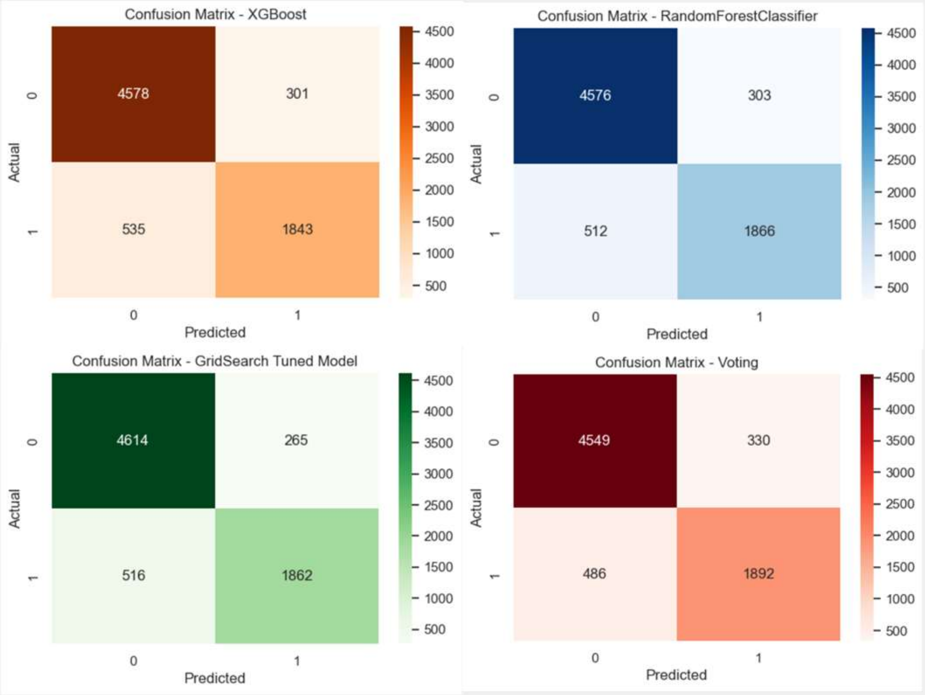
Predicting Hotel Booking Cancellations

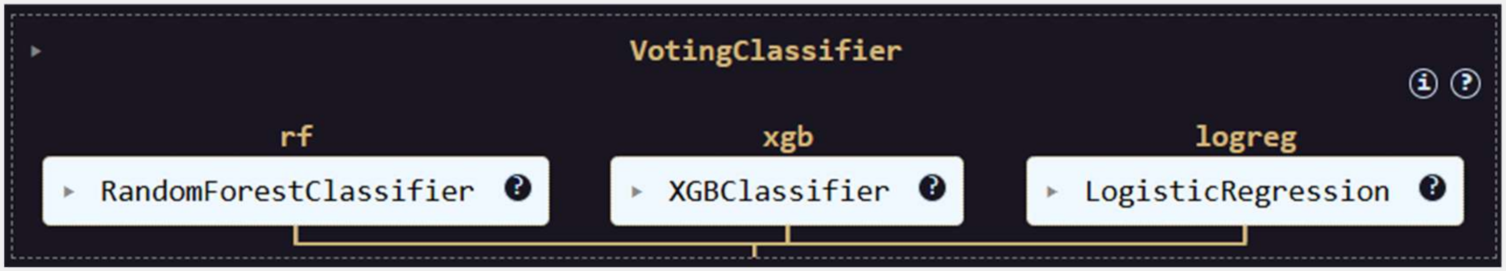
Project Goal:

- Solve a binary classification problem.
- Predict whether a hotel booking will be canceled (1) or not (0).

Approach:

- Conducted thorough preprocessing, outlier handling, and feature engineering.
- **Trained multiple models:** Random Forest, XGBoost, Logistic Regression.
- Combined them using a Voting Classifier for improved accuracy.

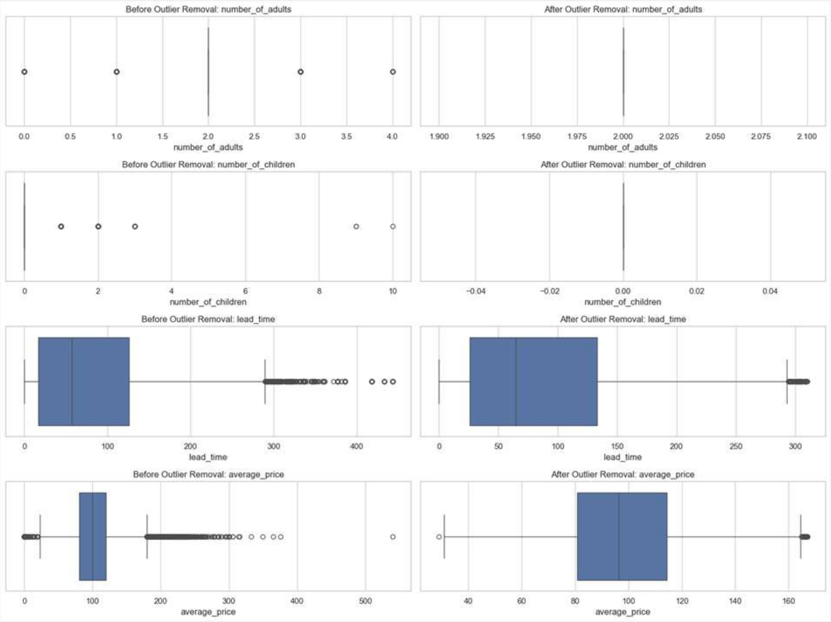


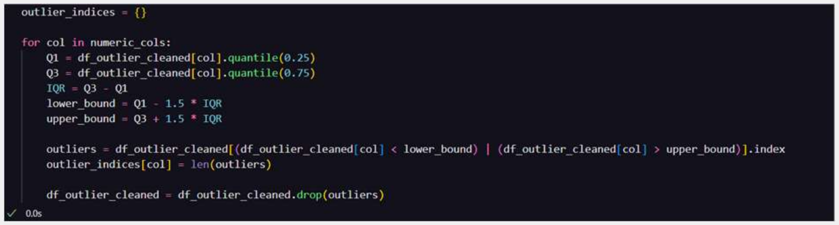


Data Cleaning and Preprocessing

Steps Taken:

- Checked for null values and confirmed the dataset had no missing entries.
- Removed extra whitespaces and standardized column names.
- Converted `date_of_reservation` to datetime format.
- Dropped irrelevant or constant columns.
- Applied one-hot encoding to categorical features using `pd.get_dummies()`.
- Split features and target variable: X and `y = booking_status`.
- Removed outliers using IQR on numeric columns.



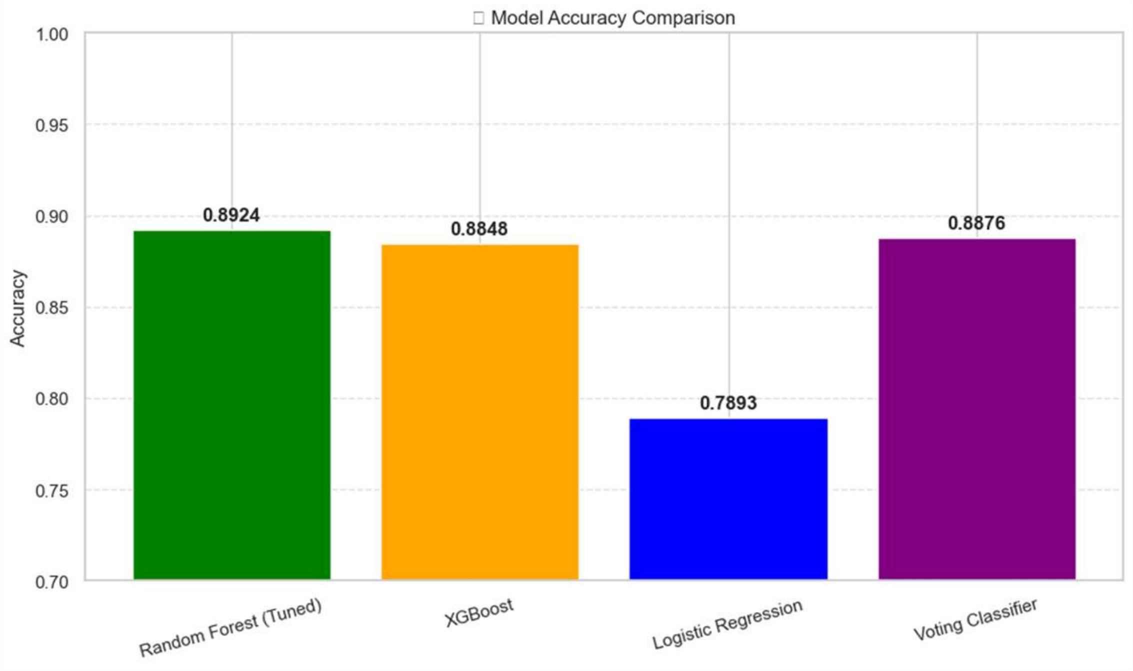


```
outlier_indices = {}

for col in numeric_cols:
    Q1 = df_outlier_cleaned[col].quantile(0.25)
    Q3 = df_outlier_cleaned[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = df_outlier_cleaned[(df_outlier_cleaned[col] < lower_bound) | (df_outlier_cleaned[col] > upper_bound)].index
    outlier_indices[col] = len(outliers)

    df_outlier_cleaned = df_outlier_cleaned.drop(outliers)
```

✓ 0.0s

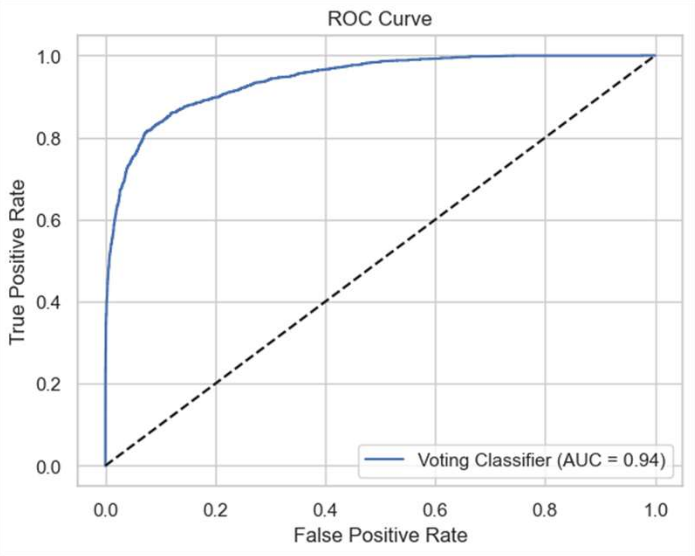
Feature Engineering & Modeling

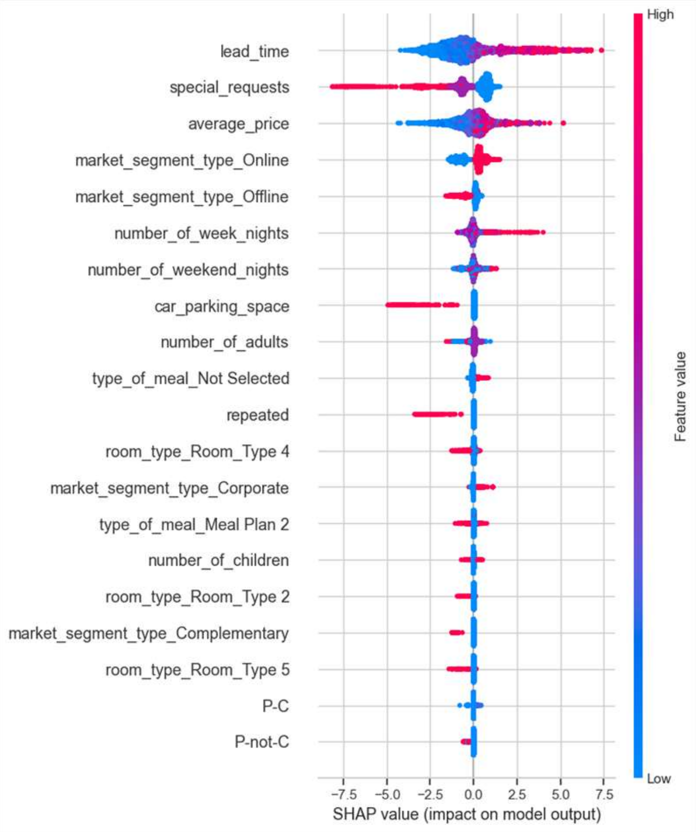
- Removed Booking_ID, reservation date
- Engineered: dummy variables, stratified train-test split
- Trained **Random Forest**, **XGBoost**, **Logistic Regression**
- Combined them with a **Voting Classifier**



Performance Evaluation & SHAP Insights

- Used accuracy, classification report, ROC curve
- ROC AUC for VotingClassifier: ~94%
- Applied **SHAP** to interpret feature impact


ROC Curve
True Positive Rate
False Positive Rate
Voting Classifier (AUC = 0.94)


lead_time
special_requests
average_price
market_segment_type_Online
market_segment_type_Offline
number_of_week_nights
number_of_weekend_nights
car_parking_space
number_of_adults
type_of_meal_Not Selected
repeated
room_type_Room_Type 4
market_segment_type_Corporate
type_of_meal_Meal Plan 2
number_of_children
room_type_Room_Type 2
market_segment_type_Complementary
room_type_Room_Type 5
P-C
P-not-C
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
SHAP value (impact on model output)
High
Feature value
Low