Prepared by Sama Osama

WEEK 2 OF CELLULA INTERNSHIP

Task'2

Booking Cancellation Prediction

Hotel Reservation Dataset Analysis &
Modeling

Dataset Overview

- 36,285 rows × 17 columns
- Covers reservation behavior, pricing, and booking outcomes
- Target variable: booking_status (Canceled vs. Not_Canceled)

	Booking_ID	number of adults	number of children	number of weekend nights	number of week nights	type of meal	car parking space	room type	lead time	market segment type	repeated	P-C	P-not-C	average price	special requests	date of reservation	booking status
0	INN00001	1	1	2	5	Meal Plan 1	0	Room_Type 1	224	Offline	0	0	0	88.00	0	10/2/2015	Not_Canceled
1	INN00002	1	0	1	3	Not Selected	0	Room_Type 1	5	Online	0	0	0	106.68	1	11/6/2018	Not_Canceled
2	INN00003	2	1	1	3	Meal Plan 1	0	Room_Type 1	1	Online	0	0	0	50.00	0	2/28/2018	Canceled
3	INN00004	1	0	0	2	Meal Plan 1	0	Room_Type 1	211	Online	0	0	0	100.00	1	5/20/2017	Canceled
4	INN00005	1	0	1	2	Not Selected	0	Room_Type 1	48	Online	0	0	0	77.00	0	4/11/2018	Canceled
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36280	INN36282	2	0	0	2	Meal Plan 2	0	Room_Type 1	346	Online	0	0	0	115.00	1	9/13/2018	Canceled
36281	INN36283	2	0	1	3	Meal Plan 1	0	Room_Type 1	34	Online	0	0	0	107.55	1	10/15/2017	Not_Canceled
36282	INN36284	2	0	1	3	Meal Plan 1	0	Room_Type 4	83	Online	0	0	0	105.61	1	12/26/2018	Not_Canceled
36283	INN36285	3	0	0	4	Meal Plan 1	0	Room_Type 1	121	Offline	0	0	0	96.90	1	7/6/2018	Not_Canceled
36284	INN36286	2	0	0	5	Meal Plan 1	0	Room_Type 4	44	Online	0	0	0	133.44	3	10/18/2018	Not_Canceled

36285 rows × 17 columns

Data Preprocessing Steps

1- Standardized column names (lowercase, removed spaces)

3- Check for Null values and duplicates

5- Dropped invalid date entries (NAT)

2- Fixed mislabeled columns (e.g., p-c → preservation_canceled)

4- Converted date of reservation to datetime

Feature Engineering

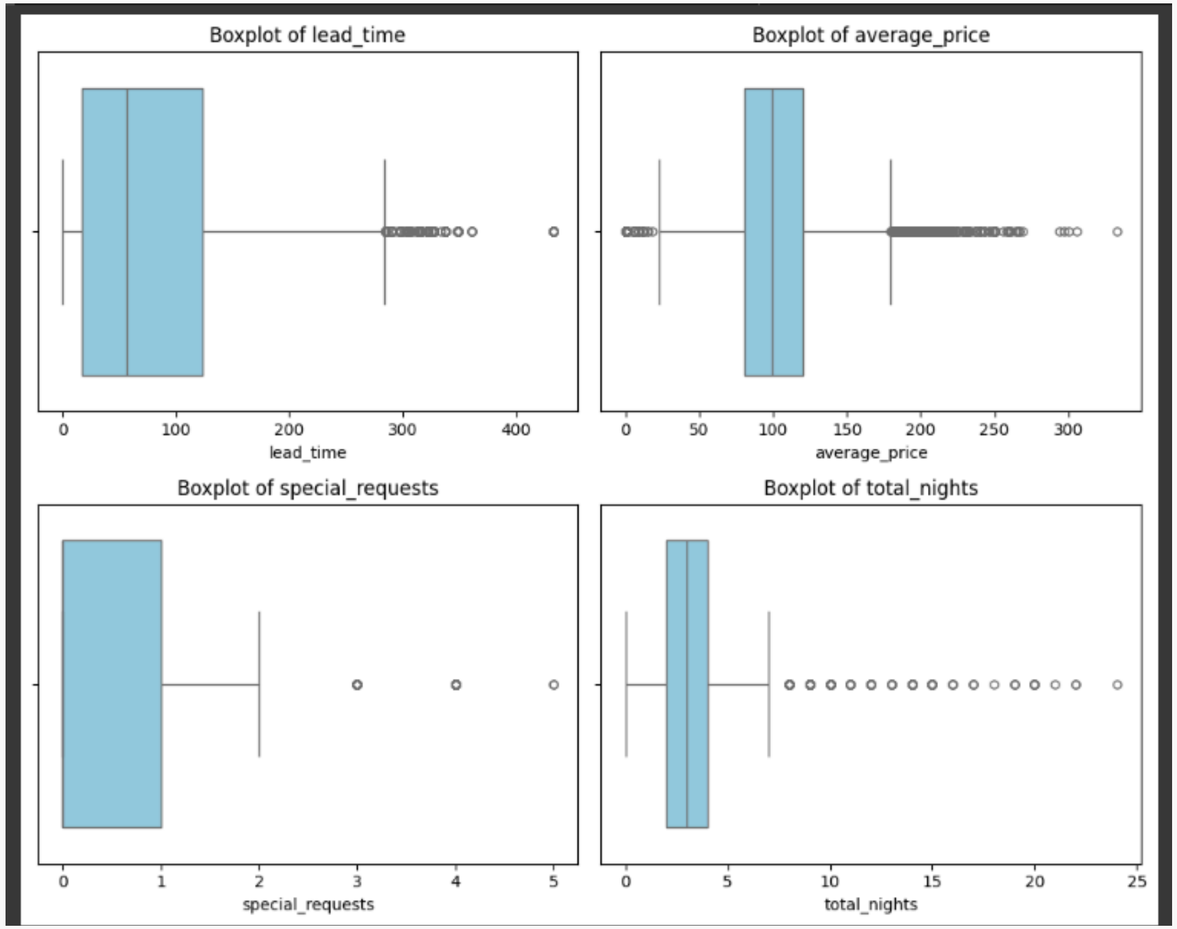
- Extracted: reservation_year and reservation_month from date_of_reservation
- total_nights = number_of_weekend_nights + number_of_week_nights
- Dropped irrelevant fields like booking_id

	booking_id	number_of_adults	number_of_children	number_of_weekend_nights	number_of_week_nights	type_of_meal	car_parking_space	room_type	lead_time	market_segment_type	Visited_Before	Preservation_Canceled	Preservation_not_Canceled	average_price	specia
0	INN00001	1	1	2	5	Meal Plan 1	0	Room_Type 1	224	Offline	0	0	0	88.00	
1	INN00002	1	0	1	3	Not Selected	0	Room_Type 1	5	Online	0	0	0	106.68	
4	INN00005	1	0	1	2	Not Selected	0	Room_Type 1	48	Online	0	0	0	77.00	
8	INN00009	1	1	0	4	Meal Plan 1	0	Room_Type 1	121	Offline	0	0	0	96.90	
10	INN00011	1	0	1	0	Not Selected	0	Room_Type 1	0	Online	0	0	0	85.03	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1	148	Online	0	0	0	98.39	
36275	INN36276	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	Offline	0	0	0	65.00	
36276	INN36277	2	0	2	3	Not Selected	0	Room_Type 1	5	Online	0	0	0	106.68	
36279	INN36281	2	0	1	1	Not Selected	0	Room_Type 1	48	Online	0	0	0	94.50	
36283	INN36285	3	0	0	4	Meal Plan 1	0	Room_Type 1	121	Offline	0	0	0	96.90	

14149 rows × 19 columns

Outlier Detection and Handling using Z-score

- Applied Z-Score method on:
- lead_time, average_price, special_requests, total_nights
- Removed records with Z > 3

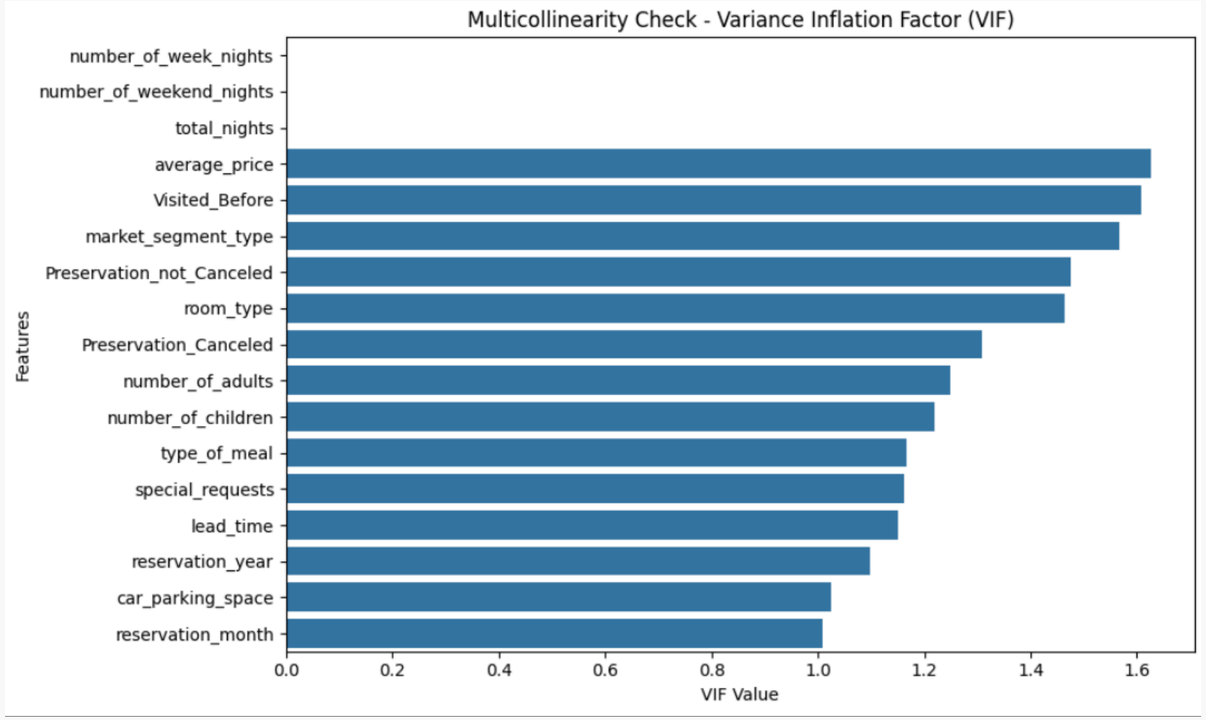


Transformation For The Categorical Data

- Label encoded:
- type_of_meal, room_type, market_segment_type
- Converted booking_status to 0 (Not_Canceled) and 1 (Canceled)
- Applied StandardScaler to numerical features

Multicollinearity Check

- Calculate VIF (Variance Inflation Factor)
- Visualize VIF values as a horizontal bar chart
- Use it to detect multicollinearity among features
- number_of_week_nights, number_of_weekend_nights, and total_nights have VIF = ∞, indicating perfect multicollinearity.
- Most other features have VIF < 2, suggesting low multicollinearity.



Train-Test Split

- 80% training / 20% test
- Scaled and resampled training data
- Final sets:
- X_train_balanced, X_test_scaled, y_train_balanced, y_test

```
Logistic Regression Accuracy: 0.7922

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.89      0.85      1865
           1       0.74      0.61      0.67       965

    accuracy                           0.79      2830
   macro avg       0.78      0.75      0.76      2830
weighted avg       0.79      0.79      0.79      2830
```

Modeling Approaches

Results – Logistic Regression

- Accuracy: e.g., 79.22%
- Precision, Recall, F1-score (from classification report)
- Confirms model's effectiveness in predicting cancellations
- Balanced performance across both classes after SMOTE

Metric	Class 0 (Not Canceled)	Class 1 (Canceled)
Precision	0.81	0.74
Recall	0.89	0.61
F1-Score	0.85	0.67

- **Overall Accuracy:** 79.22%
- **Macro F1 Score:** 76%
- **Weighted F1 Score:** 79%

Thank you