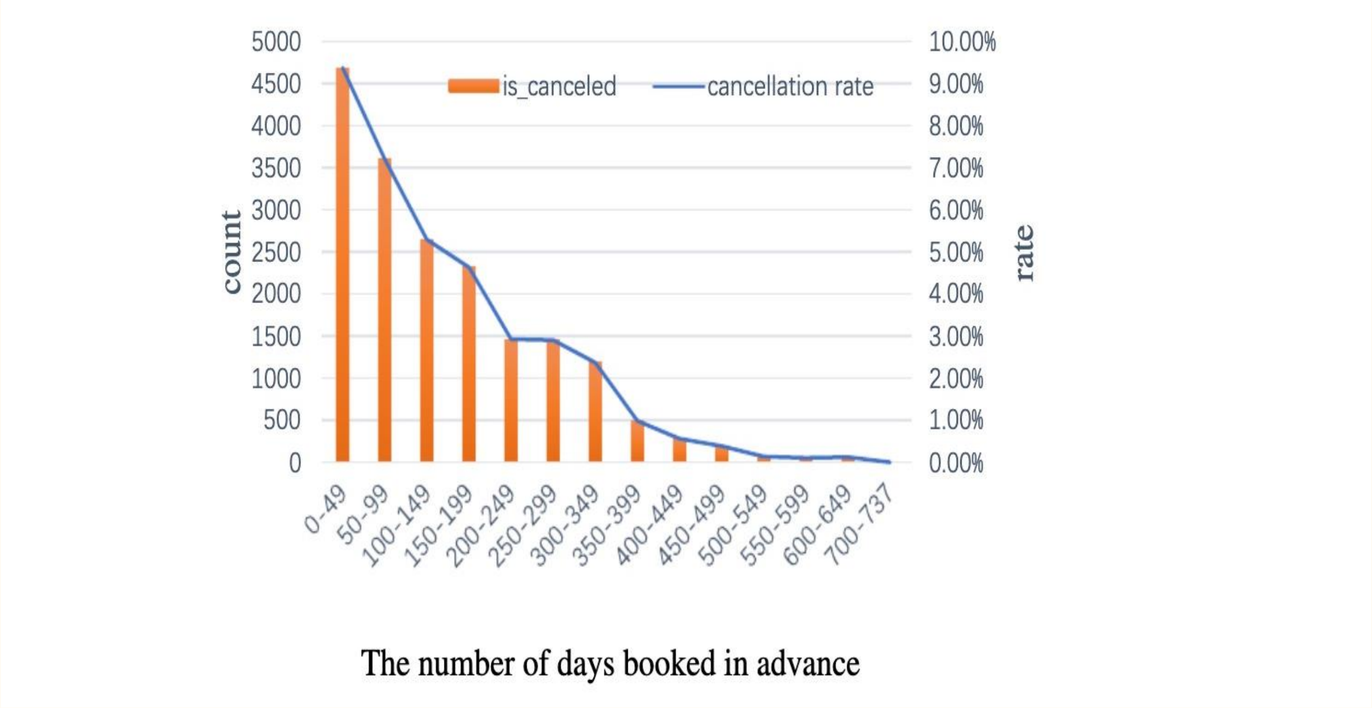
Major Reasons of Affecting Hotel Booking Cancellations:

Analysis of Hotel Booking Cancellation Predictors

“ Travel Accommodation Sector

INCREASE after Pandemic ”



“ What could be

the **major reasons** of affecting

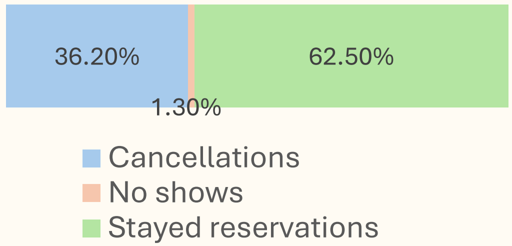
Hotel Booking Cancellations? ”

LITERATURE REVIEW | Background Information

Summary of the booking.com survey of booking cancellations

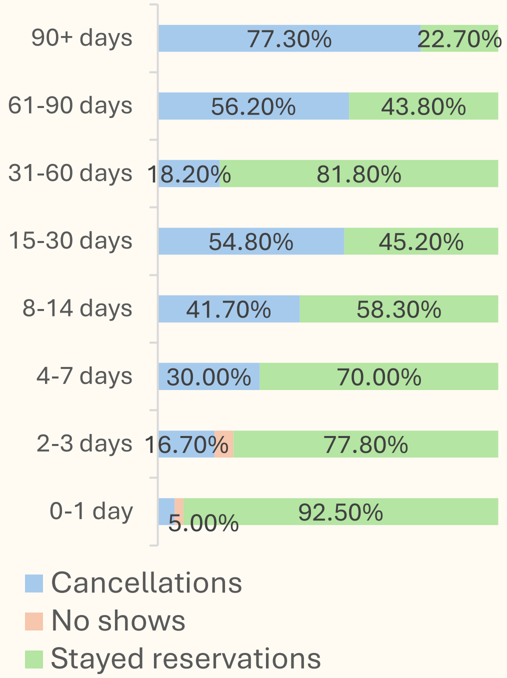
https://partner.booking.com/en-us/solutions/cancellations-characteristics-report

Overall Reservations



According to booking.com survey, 36.2% of accommodation bookings are cancelled.

Reservation status per Bookwindow



Growth Market

Travel accommodation market expected to reach $1,974.30 billion, globally, by 2031 at 11.3% (Global Opportunity Analysis and Industry Forecast, 2023-2032, 2024).

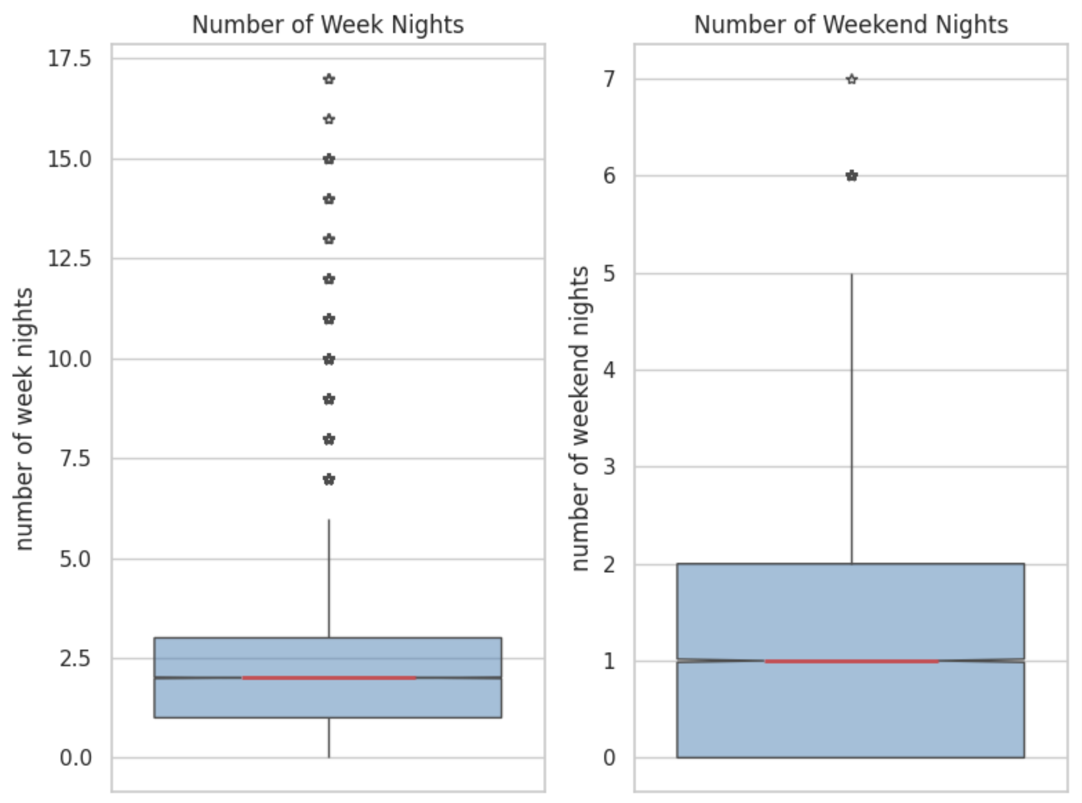
DATASET | Major reasons of affecting hotel booking cancellation

There are 14 predictors (5 **qualitative** and 9 **quantitative**).

Type of meal (binary: 'selected' or 'non selected')w	Number of adults (numeric: number of adults included in the booking)
Car parking space (binary: '0' – car parking is not included '1' – car parking is included)	Number of childern (numeric: number of children included in the booking)
Room type (nominal: types of rooms)	Number of weekend nights (numeric: number of weekend night included in the booking)
Market segment type (binary: 'online' or 'offline')	Number of week nights (numeric: number of week nights included in the booking)
Repeated (binary: '0' – booking is not repeated '1' – booking is repeated)	Lead time (numeric: numbers of days between booking date and arrival date)
Special request (numerical: reason for choosing the school)	Average price (numeric: average booking price)
P-not-C (numeric: number of previous booking not cancelled by customers)	P-C (numeric: number of previous booking cancelled by customers)

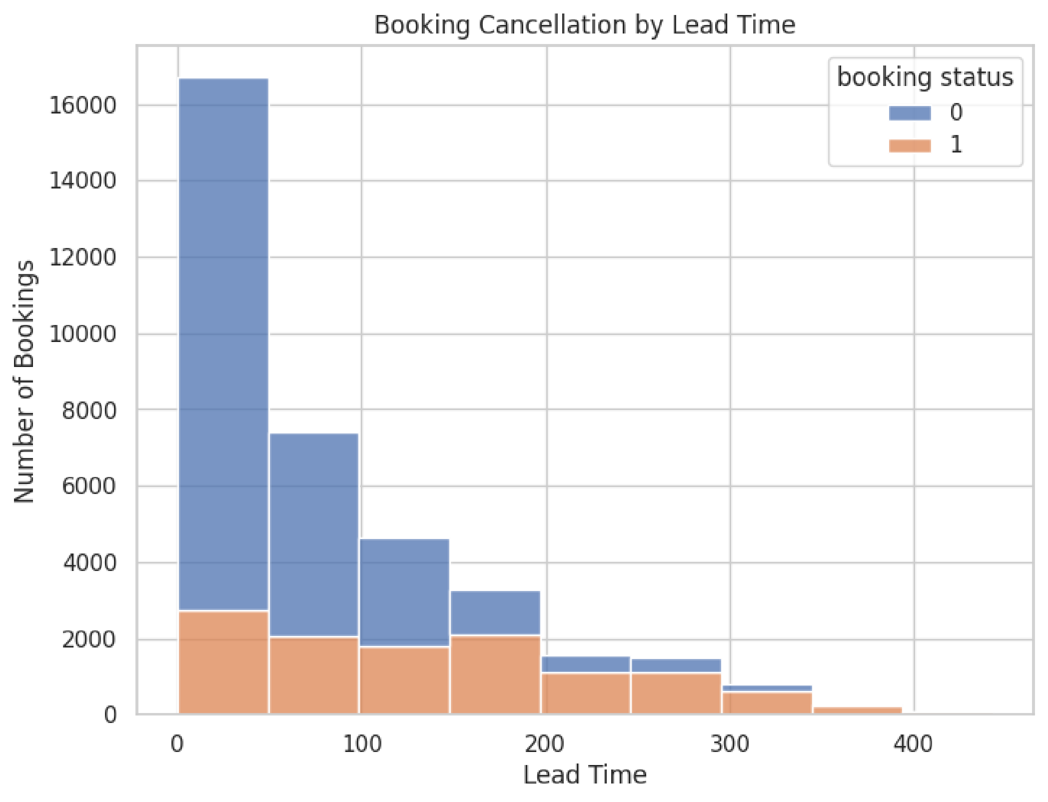
EXPLORATORY DATA ANALYSIS #1| Weekday-Weekend Boxplots & Booking cancellation histogram

Booking days of Weekday - Weekend



: There are more **variations** in weekday booking.

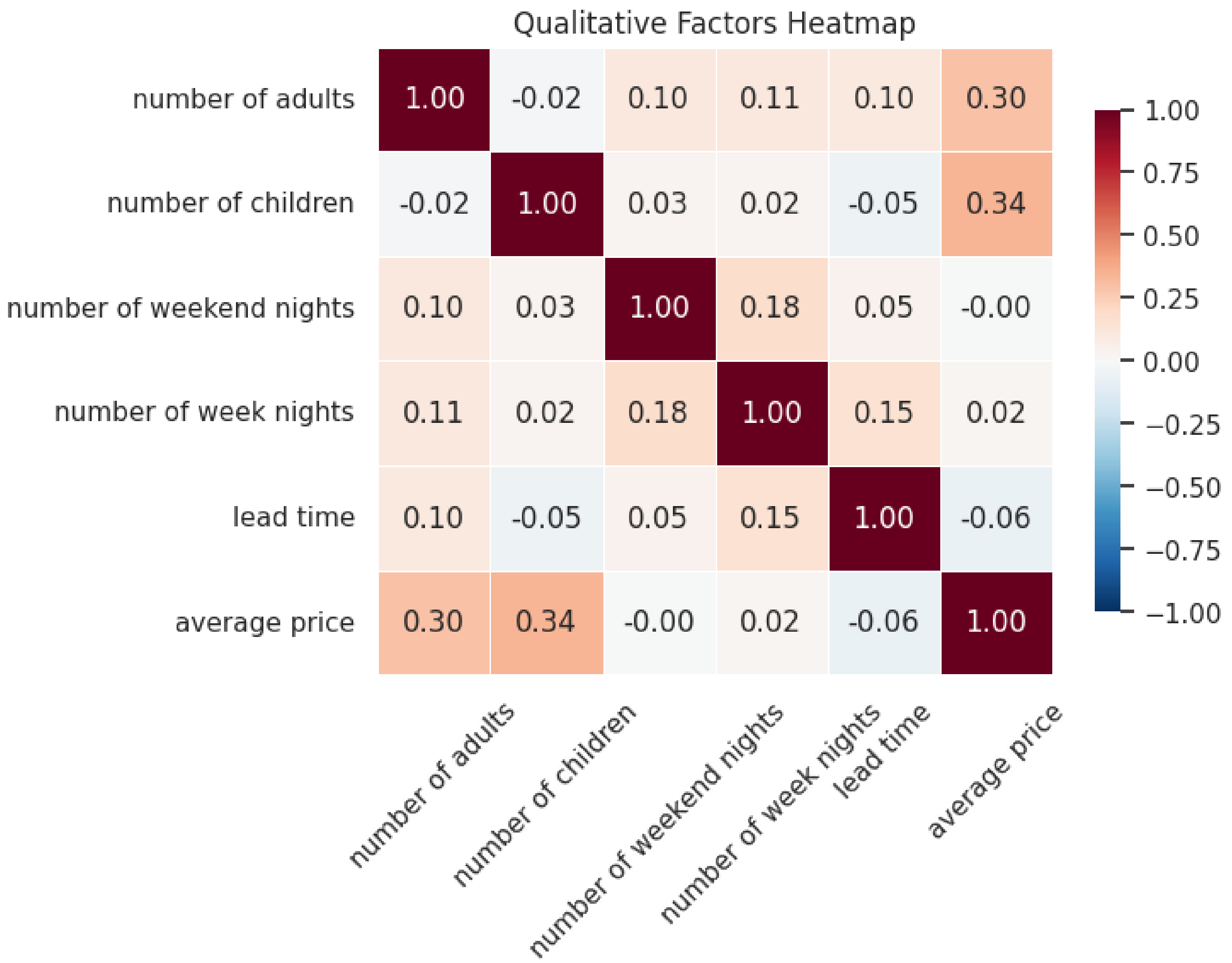
Booking cancellation by Lead Time



: It seems that hotel booking cancellations **increase** with **short-term** lead times

EXPLORATORY DATA ANALYSIS #2 | Correlation Plot

Between qualitative factors, we didn't see meaningful correlation



	number of adults	number of children	number of weekend nights	number of week nights	lead time	average price
number of adults	1.00	-0.02	0.10	0.11	0.10	0.30
number of children	-0.02	1.00	0.03	0.02	-0.05	0.34
number of weekend nights	0.10	0.03	1.00	0.18	0.05	-0.00
number of week nights	0.11	0.02	0.18	1.00	0.15	0.02
lead time	0.10	-0.05	0.05	0.15	1.00	-0.06
average price	0.30	0.34	-0.00	0.02	-0.06	1.00

HYPOTHESIS

NULL HYPOTHESIS (H0)

There is no significant relationship between the lead time/average price/meal type for cancelled and non- cancelled bookings

ALTERNATIVE HYPOTHESIS (H1)

There is a significant relationship between the lead time/average price/meal type for cancelled and non-cancelled bookings

METHODOLOGY | Pre-Processing and Model Building

Key objective: Which model is the best model to explain and predict Hotel booking cancellation

STEP 1: Create Training and Test sets

STEP 2: One-Hot Encoding Categorical Variables

STEP 3: Check for N/As, scale and center

STEP 4: Train regression models

- Model 1 – Linear Regression – Recursive Feature Elimination Regression
- Model 2 – K-Nearest Neighbors (KNN)
- Model 3 – Decision Tree
- Model 4 – Bagged Tree
- Model 5 – Random Forest

STEP 5: Interpret results

STEP 6: Test final model

LINEAR REGRESSION MODEL | Baseline Model: Linear Regression with RFE

Booking Status = + 0.0713 (type of meal 1)
- 0.2041 (type of meal 2)
+ 0.1694 (type of meal 3)
+ 0.1072 (car parking space)
- 0.1529 (room type 2)
- 0.0823 (room type 3)
- 0.0725 (room type 4)
- 0.0779 (room type 5)
+ 0.2218 (room type 6)
- 0.0748 (room type 7)

We use RFE (Recursive Feature Elimination) selection method for the linear regression model.
However, the (Mean Squared Error) was **0.19**, and the R-squared score was **0.14**, indicating a need for a better model to predict the booking status

Confusion Matrix |

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

 # Precision is a metric that tells us about quality of positive predictions. $\left(\frac{TP}{(TP+FP)} = \frac{\#True\ Positives}{\#All\ Predictied\ Positives}\right)$

 # Recall is a metric that tells us about how well the model identifies true positives. $\left(\frac{TP}{(TP+FN)} = \frac{\#True\ Positives}{\#All\ Actual\ Positives}\right)$

 # Accuracy is a metric that measures the proportion of correct predictions.
$\left(\frac{TP+TN}{(TP+FP+FN+TN)} = \frac{\#True\ Positives+Ture\ Negatives}{\#All\ obseravtions}\right)$

 # F1 Score is a metric that balances precision and recall. $\left(2\ x\left(\frac{Precision\ *\ Recall}{Precision + Recall}\right)\right)$

K-NEAREST NEIGHBORS MODEL |

Booking status	Precision	Recall	F1-Score
0 (Not Cancelled)	0.80	0.94	0.87
1 (Cancelled)	0.82	0.53	0.64
Overall	0.80		

 # K=2 is the best number for predicting

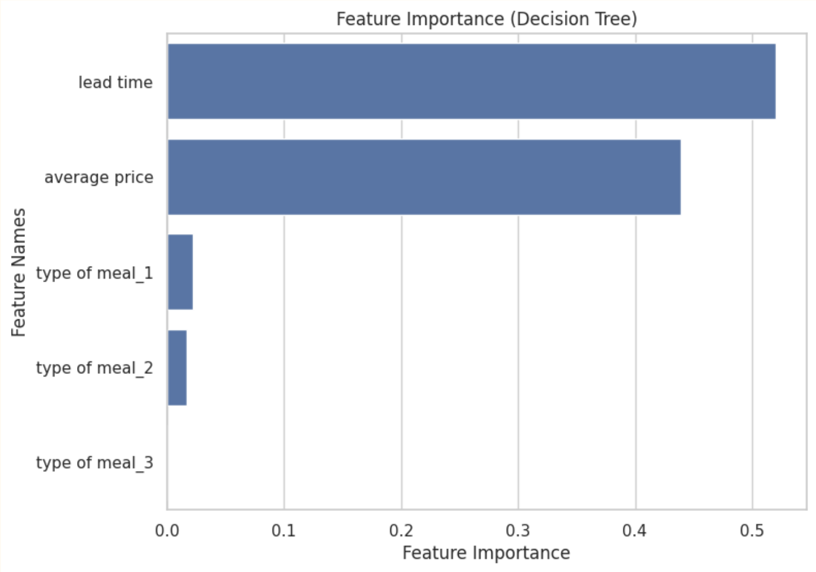
 # Overall accuracy is 80%

 # For class 0 (Not cancelled): High precision(0.80) and very high recall(0.94) indicate the model is very good at identifying class 0. Additionally, hight F1-Score(0.87) confirms good overall performance for class 0

 # For class 1 (Cancelled): Great precision(0.82) but lower recall(0.53) suggest the model is accurate when it predicts class 1, but it missess many actual class 1 instances. F1-Score(0.64) shows the imbalnce between precision and recall

DECISION TREE MODEL |

Booking status	Precision	Recall	F1-Score
0 (Not Cancelled)	0.84	0.87	0.85
1 (Cancelled)	0.71	0.66	0.68
Overall	0.80		



 # Overall accuracy is 80%

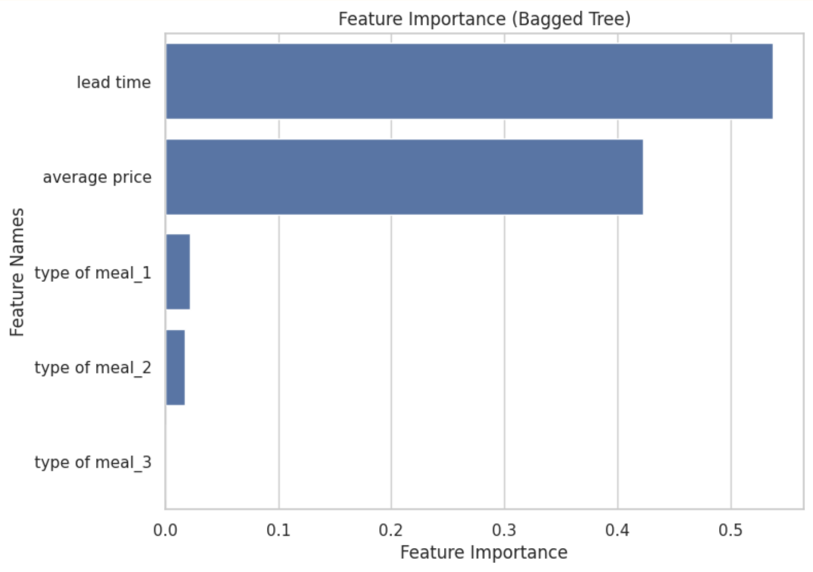
 # For class 0 (Not cancelled): High precision(0.84) and recall(0.87) suggest the model is very good at identifying class 0. F1-Score(0.85) is also high, which shows overall performance is great for class 0

 # For class 1 (Cancelled): Good precision(0.82) but lower recall(0.66) suggest the model is accurate when it predicts class 1, but it missess many actual class 1 instances. Lower F1-Score(0.68) suggest the imbalnce between precision and recall

 # Variable Importance Plot: Lead time is the most important variable for predicting the booking status, followed by average price

BAGGED TREE MODEL |

Booking status	Precision	Recall	F1-Score
0 (Not Cancelled)	0.83	0.89	0.86
1 (Cancelled)	0.74	0.65	0.69
Overall	0.81		



 # Overall accuracy is 81%

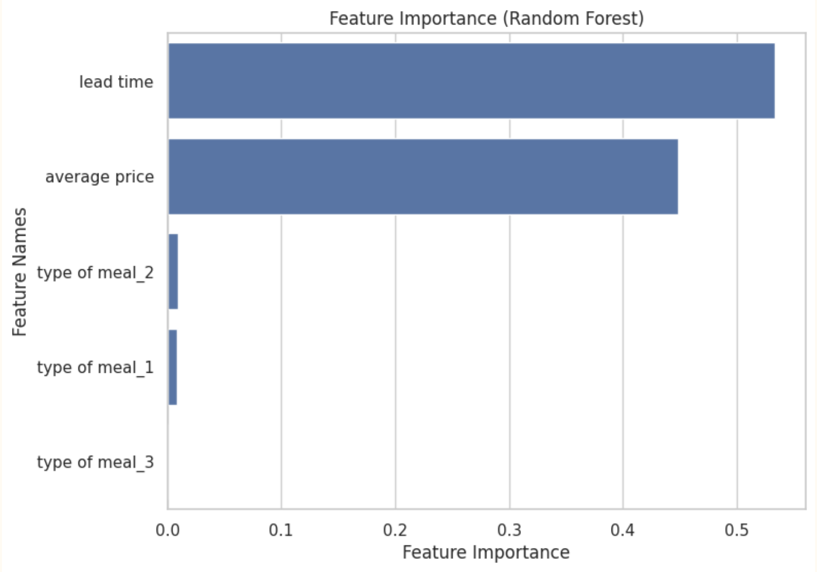
 # For class 0 (Not cancelled): High precision(0.83) and recall(0.89) suggest the model is very good at identifying class 0. F1-Score(0.86) is high, which shows overall performance is great for class 0

 # For class 1 (Cancelled): Good precision(0.74) but recall(0.65) is lower than presicion. It suggest that the model is good at predicting class 1, but it has many missing values in class 1 instances. F1-Score(0.69) also shows the imbalnce between precision and recall

 # Variable Importance Plot: Lead time is still the most important variable for predicting the booking status, followed by average price

RANDOM FOREST MODEL |

Booking status	Precision	Recall	F1-Score
0 (Not Cancelled)	0.84	0.88	0.86
1 (Cancelled)	0.74	0.66	0.69
Overall	0.81		



 # Overall accuracy is 81%

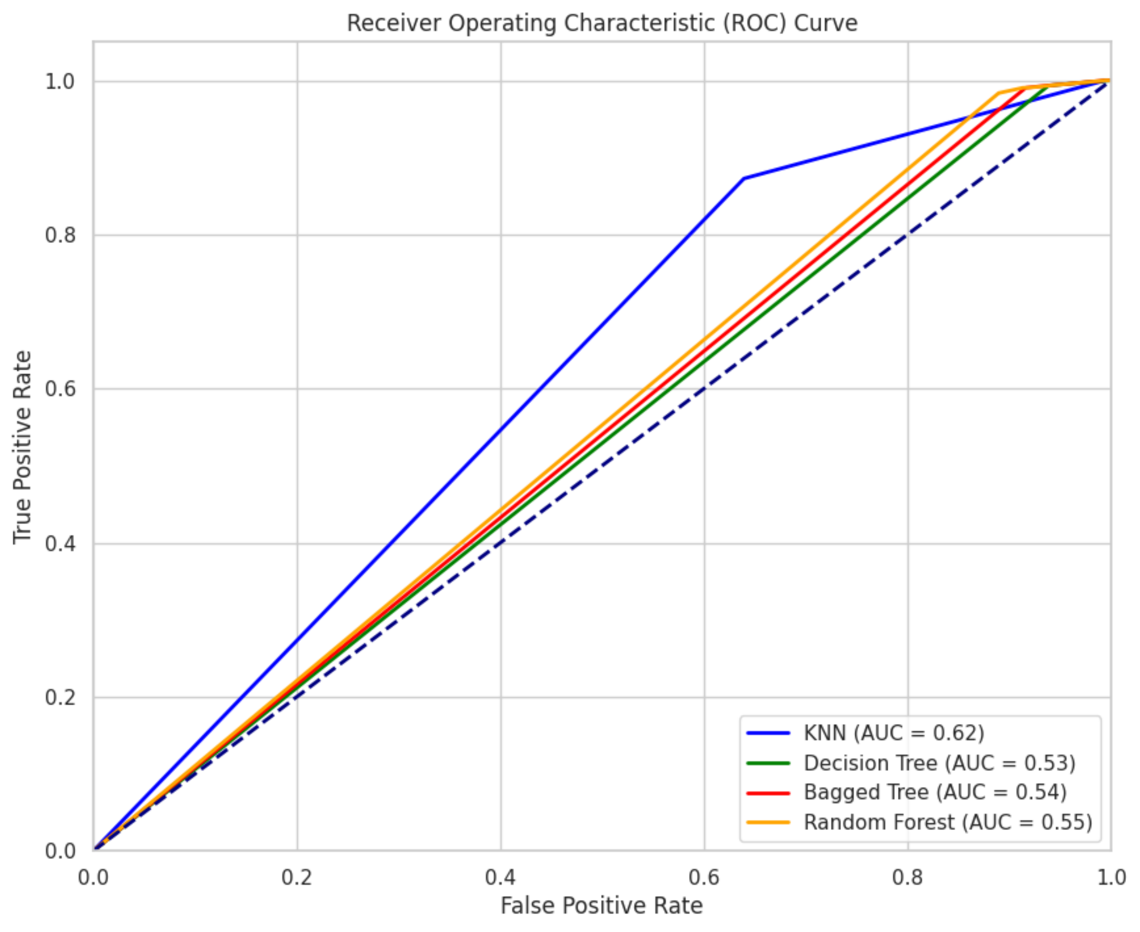
 # For class 0 (Not cancelled): High precision(0.84) and recall(0.88) suggest the Random Forest model is very good at identifying class 0. F1-Score(0.86) shows overall performance is great for class 0

 # For class 1 (Cancelled): Good precision(0.74) but recall(0.66) is lower than presicion. F1-Score(0.69) also shows the imbalnce between precision and recall

 # Variable Importance Plot: Lead time is still the most important variable for predicting the booking status, followed by average price

RESULTS | Models' Performance Comparison

Model Type	Accuracy
KNN	0.7985
Decision Tree	0.8047
Bagged Tree	0.8068
Random Forest	**0.8119**



INTERPRETATION &CONCLUSION | Models' Performance

- **Accuracy** was highest in Random Forest model, followed closely by Bagged Tree, Decision Tree. KNN has the lowest accuracy, but the difference is not substantial (about 1.3% lower than Random Forest).

- **ROC Curve** interestingly shows the KNN has the highest Area Under the Curve (AUC) score, while KNN has the lowest accuracy. It indicates that better overall classification performance across difference thresholds.

- Overall, we interpret that **Random Forest** model performs the best. It demonstrates good performance and balance between precision and recall

  - Highest accuracy (0.8199)

  - Second-highest AUC (0.55)

FURTHER CONSIDERATION | Models' Performance

- The differences in performance are relatively small, which suggest that the problem might be challenging for these models.
- Further tuning or feature engineering could potentially improve performance
- The choice of model might depend on specific requirements (e.g., interpretability, prediction speed, etc.)

THANK YOU

Q&A