INN Hotels

Supervised Learning Classification

02/13/2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Booking Trends:

- Guests booking cheaper rooms, with shorter lead times, requiring parking, being repeat guests, with more special requests, from Corporate/Offline segments less likely to cancel.
- Guests booking expensive rooms, with longer lead times, through Online segment more likely to cancel.

Policy Suggestions:

- Consider separate cancellation/refund policies for business vs. personal travel based on model insights.
- Introduce rewards for business travelers to encourage bookings and reduce cancellations.
- Use models to prioritize room availability for repeat and business guests, especially when overbooked.
- Utilize model predictions to manage room allocations efficiently, reallocating rooms from likely cancellations to secure bookings.

Recommendations:

- Do not solely rely on models; complement with management's industry experience for capacity management.
- Provide cost estimates for true/false positives/negatives to refine model optimization towards highest expected profits

Business Problem Overview and Solution Approach

Problem Definition

- **Context:** High booking cancellation rates cause significant revenue loss and operational challenges for hotels. The ease of online cancellations has intensified this issue.

Objective

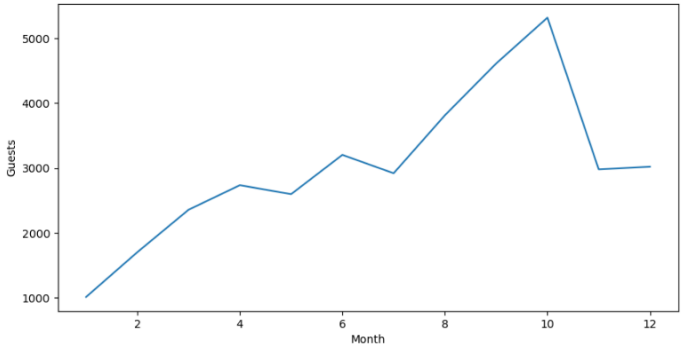
- **Goal:** Develop a Machine Learning model to predict booking cancellations, enabling INN Hotels Group to manage and mitigate cancellation impacts effectively.

Solution Approach

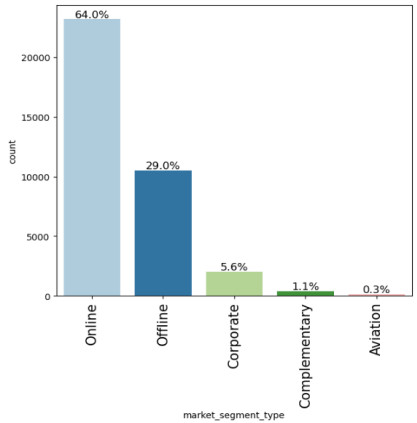
- **Data Preparation:** Clean and preprocess data, encode categorical variables.
- **Feature Engineering:** Identify key factors influencing cancellations.
- **Model Building:** Train models like Logistic Regression, Random Forest, etc., on the training set.
- **Evaluation:** Assess model performance using metrics like accuracy and F1-score on the test set.
- **Implementation:** Deploy the model, recommend policies to reduce cancellations and their impact.

EDA Results

- Busiest month in the hotel
  - October with a count of 5,317 arrivals.



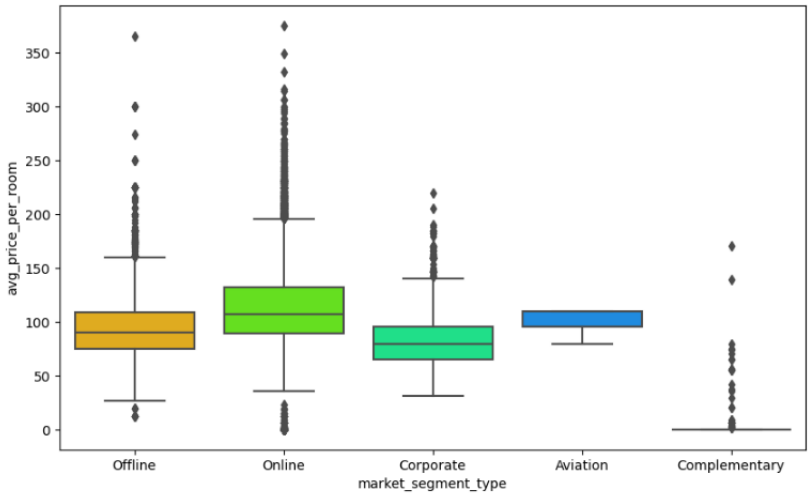
- Most common market segment
  - 23,214 bookings are made online (64%)



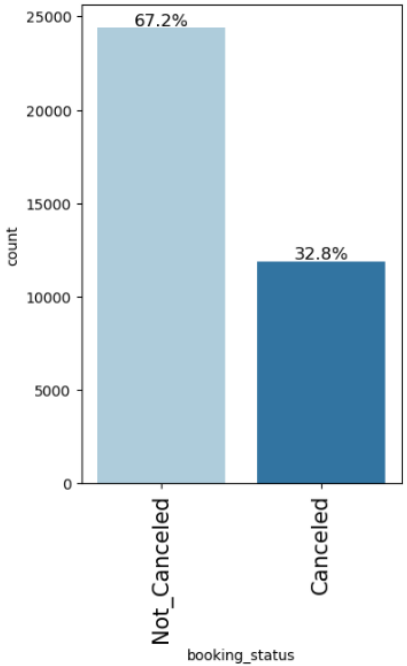
Link to Appendix slide on data background check

EDA Results

- Difference in room prices in different market segments
  - Online bookings contain the highest avg. price per room



- Percentage of bookings cancelled

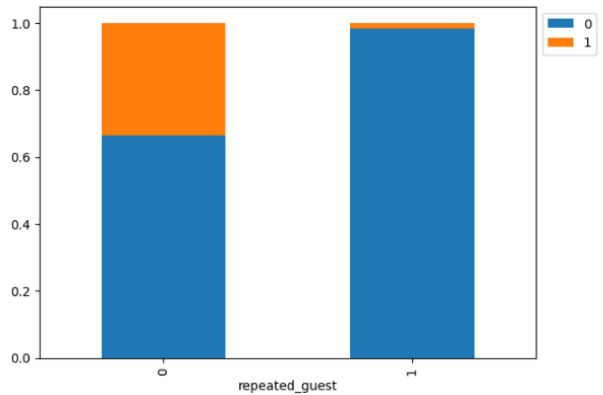


Link to Appendix slide on data background check

EDA Results

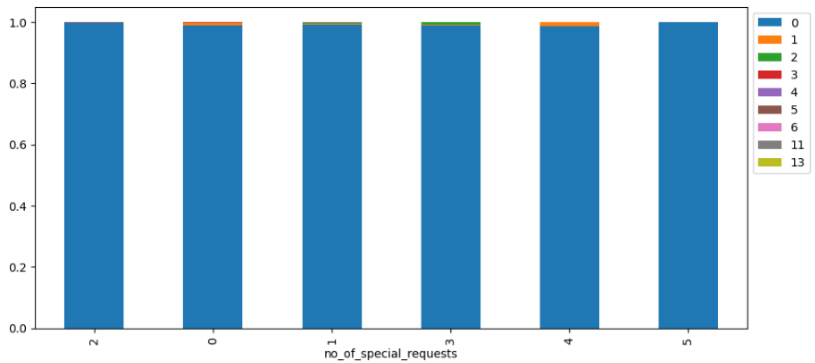
- Percentage of repeating guests cancel
  - 16 of 930 total repeating guests cancel (2%)

booking_status	0	1	All
repeated_guest			
All	24390	11885	36275
0	23476	11869	35345
1	914	16	930



- How special requests affect booking cancellation
  - As the number of special requests increase, booking cancellation decreases.
  - 0 requests has 19,777 cancellations
  - 5 requests has only 8 cancellations

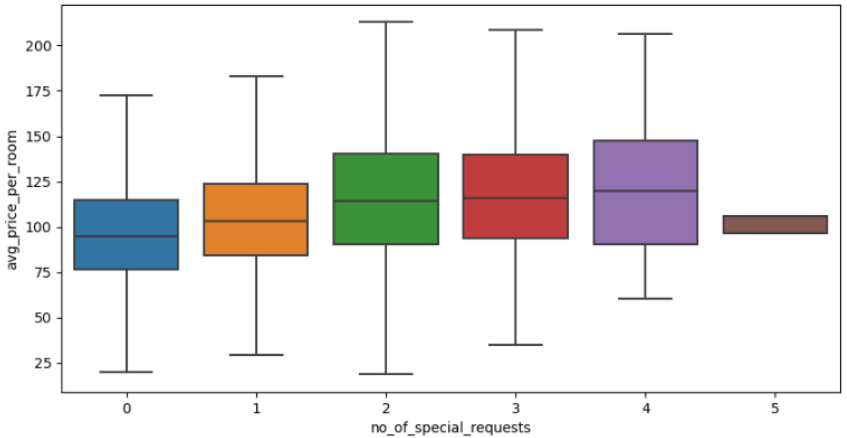
no_of_previous_cancellations	0	1	2	3	4	5	6	11	13	All
no_of_special_requests										
2	4347	4	1	2	1	7	1	1	0	4364
All	35937	198	46	43	10	11	1	25	4	36275
0	19553	153	25	28	3	3	0	8	4	19777
1	11284	39	14	13	6	1	0	16	0	11373
3	668	1	6	0	0	0	0	0	0	675
4	77	1	0	0	0	0	0	0	0	78
5	8	0	0	0	0	0	0	0	0	8



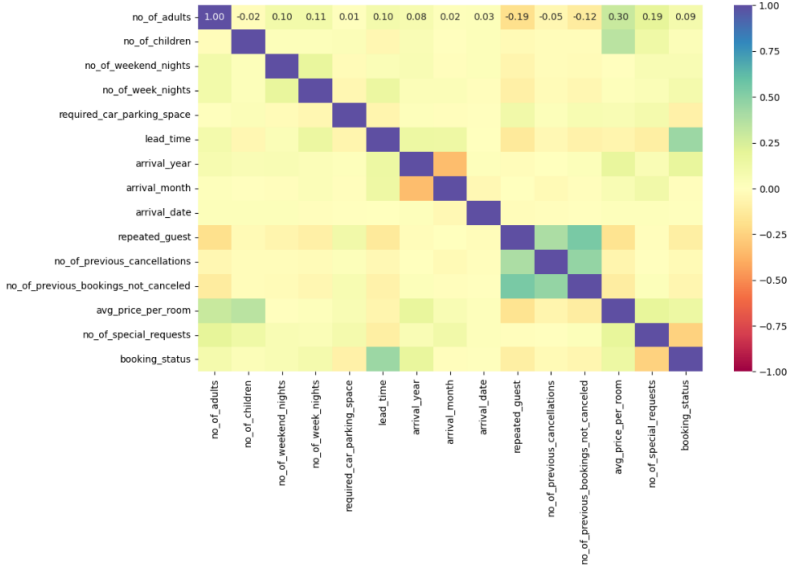
Link to Appendix slide on data background check

EDA Results

- Special request impact on room price
  - The average price per room seems to increase as the number of special requests increase



- Correlation Matrix

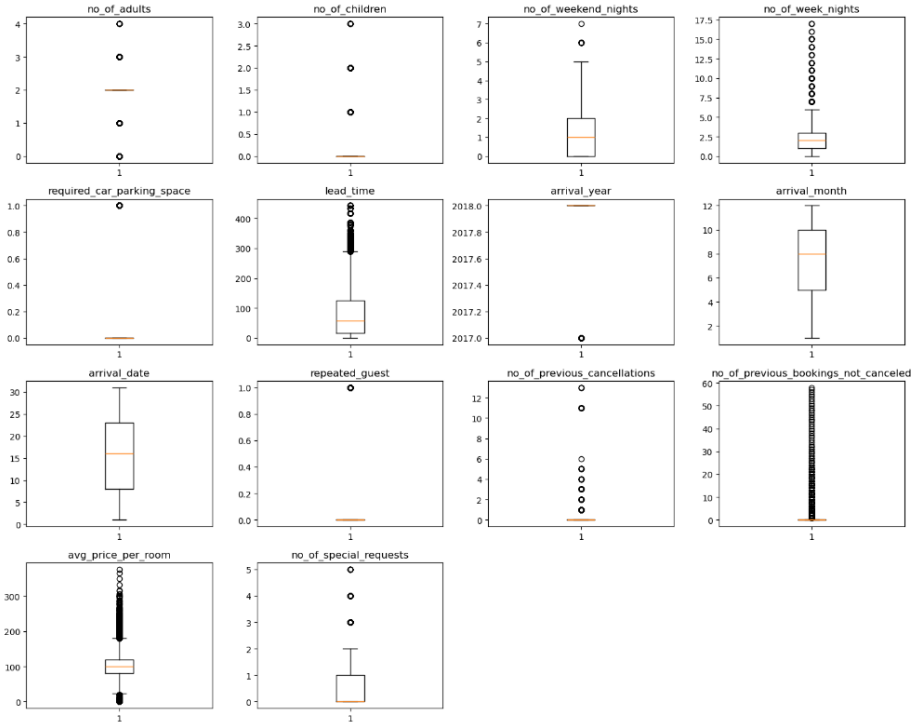


Link to Appendix slide on data background check

Data Preprocessing

- Duplicate value check
  - No duplicate values
- Missing value treatment
  - No null values
- Outlier check (treatment if needed)
  - Created boxplots to determine outliers
- Feature engineering
- Data preparation for modeling
  - Encode categorical features using dummies
  - Train and test split with 70:30 ratio

Outlier Check



***Note**: You can use more than one slide if needed*

Model Performance Summary

Final Logistic Regression Model

- 70/30 train, test split
- Removed Multicollinearity
- Tested multiple thresholds
- Used **booking_status** as the dependent variable.

	Logistic Regression sklearn	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80	0.80	0.80
Recall	0.62	0.73	0.70
Precision	0.74	0.68	0.70
F1	0.68	0.70	0.70

Test set performance comparison:

	Logistic Regression sklearn	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.81	0.80	0.80
Recall	0.63	0.73	0.70
Precision	0.73	0.68	0.69
F1	0.68	0.70	0.70

Model Performance Summary

Final Decision Tree Model

- lead_time and avg_price_per_room were the most important features
- Used CCP as the pruning method
- Potential overfitting

Train	**Sklearn**	**Pre-Prune**	**Post-Prune**
F1- Score	0.990	0.959	0.954
Train	**Sklearn**	**Pre-Prune**	**Post-Prune**
F1- Score	0.797	0.809	0.804

APPENDIX

Data Background and Contents

Objective: Address the challenge of high cancellation rates in hotel bookings, particularly within the INN Hotels Group in Portugal. The goal is to predict booking cancellations using Machine Learning to mitigate revenue loss and operational challenges.

Impact of Cancellations:

- Revenue Loss: Inability to resell canceled rooms.
- Increased Costs: Higher distribution channel fees and marketing expenses to resell rooms.
- Profit Margin Reduction: Necessity to lower prices last minute to resell rooms.
- Resource Allocation: Human resources spent on rearranging guest accommodations.

Data Description: The dataset encompasses various attributes related to customer booking details:

- Booking_ID: Unique identifier for each booking.
- no_of_adults: Number of adults booked.
- no_of_children: Number of children booked.
- no_of_weekend_nights: Stays on weekend nights (Saturday or Sunday).
- no_of_week_nights: Stays on weeknights (Monday to Friday).
- type_of_meal_plan: Meal plan selection (None, Breakfast, Half board, Full board).
- required_car_parking_space: Car parking requirement (Yes or No).
- room_type_reserved: Ciphered room type reserved.
- lead_time: Days from booking to arrival.
- arrival_year/month/date: Arrival date details.
- market_segment_type: Customer market segment.
- repeated_guest: Indicates if the guest has booked before (Yes or No).
- no_of_previous_cancellations/bookings_not_canceled: Historical booking cancellation data.
- avg_price_per_room: Dynamic average price per room.
- no_of_special_requests: Number of special requests made by the guest.
- booking_status: Indicates if the booking was canceled.

Model Building - Logistic Regression

- Checking Multicollinearity
  - Other than dummy variables, no other predictors show a VIF above 5.

	feature	VIF
0	const	39091160.37
20	market_segment_type_Online	70.87
19	market_segment_type_Offline	63.67
18	market_segment_type_Corporate	16.82
17	market_segment_type_Complementary	4.44
8	repeated_guest	1.78
11	avg_price_per_room	1.75
10	no_of_previous_bookings_not_canceled	1.65
5	arrival_year	1.42
9	no_of_previous_cancellations	1.39
4	lead_time	1.38
1	no_of_adults	1.27
6	arrival_month	1.27
14	type_of_meal_plan_Meal Plan 2	1.26
12	no_of_special_requests	1.24
16	type_of_meal_plan_Not Selected	1.20
2	no_of_children	1.19
13	total_nights	1.09
3	required_car_parking_space	1.04
15	type_of_meal_plan_Meal Plan 3	1.01
7	arrival_date	1.01

Model Building - Logistic Regression

- Results based on coefficients and odds
  - Repeated Guest: Being a repeated guest reduces the cancellation odds by ~94% or 0.06 times, with all else constant.
  - Market Segment Type Offline: Bookings from the Offline segment see an ~83% or 0.17 times reduction in cancellation odds, holding other factors steady.
  - Parking Space Requirement: Needing a parking space cuts cancellation odds by ~80% or 0.20 times.
  - Average Cost Per Night: Each increase in nightly cost raises cancellation odds by ~2% or 1.02 times.
  - Additional Night Stay: Each extra night booked boosts cancellation odds by ~6% or 1.96 times.

	Odds	Change_odds
const	0.00	-100.00
repeated_guest	0.06	-93.90
market_segment_type_Offline	0.17	-82.71
required_car_parking_space	0.20	-79.61
no_of_special_requests	0.23	-76.72
market_segment_type_Corporate	0.44	-55.58
arrival_month	0.96	-3.90
lead_time	1.02	1.57
avg_price_per_room	1.02	1.65
total_nights	1.06	5.65
no_of_adults	1.08	7.63
no_of_previous_cancellations	1.26	26.25
type_of_meal_plan_Meal Plan 2	1.27	27.30
type_of_meal_plan_Not Selected	1.41	41.44
arrival_year	1.60	59.53

Model Building - Logistic Regression

Model performance: lg1

- Coefficient Sign: Negative coefficients reduce cancellation probability with attribute increases; positive coefficients increase it.
- P-value Significance: Variables with p-values under 0.05 are significant at a 5% level, though multicollinearity may impact these values.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25371
Method:	MLE	Df Model:	20
Date:	Mon, 12 Feb 2024	Pseudo R-squ.:	0.3266
Time:	08:07:37	Log-Likelihood:	-10836.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-947.8054	120.138	-7.889	0.000	-1183.271	-712.340
no_of_adults	0.0698	0.036	1.914	0.056	-0.002	0.141
no_of_children	-0.0450	0.046	-0.986	0.324	-0.134	0.044
required_car_parking_space	-1.5856	0.138	-11.494	0.000	-1.856	-1.315
lead_time	0.0156	0.000	59.066	0.000	0.015	0.016
arrival_year	0.4685	0.060	7.871	0.000	0.352	0.585
arrival_month	-0.0391	0.006	-6.085	0.000	-0.052	-0.027
arrival_date	0.0004	0.002	0.206	0.837	-0.003	0.004
repeated_guest	-2.3829	0.617	-3.862	0.000	-3.592	-1.174
no_of_previous_cancellations	0.2687	0.085	3.166	0.002	0.102	0.435
no_of_previous_bookings_not_canceled	-0.1749	0.154	-1.134	0.257	-0.477	0.127
avg_price_per_room	0.0162	0.001	24.363	0.000	0.015	0.018
no_of_special_requests	-1.4556	0.030	-48.576	0.000	-1.514	-1.397
total_nights	0.0534	0.009	5.656	0.000	0.035	0.072
type_of_meal_plan_Meal Plan 2	0.2485	0.066	3.777	0.000	0.120	0.378
type_of_meal_plan_Meal Plan 3	16.7413	2512.467	0.007	0.995	-4907.604	4941.087
type_of_meal_plan_Not Selected	0.3337	0.051	6.527	0.000	0.234	0.434
market_segment_type_Complementary	-28.9673	2751.609	-0.011	0.992	-5422.022	5364.088
market_segment_type_Corporate	-1.1178	0.264	-4.242	0.000	-1.634	-0.601
market_segment_type_Offline	-2.0655	0.252	-8.201	0.000	-2.559	-1.572
market_segment_type_Online	-0.2948	0.249	-1.183	0.237	-0.783	0.194

Model Building - Logistic Regression

Model performance: lg2

After removing multicollinearity

- Positive Coefficients: Increases in number of adults, lead time, arrival year comparison (2018 vs. 2017), previous cancellations, average room price, total nights booked, selecting Meal Plan 2, and opting out of a meal plan raise cancellation chances.
- Negative Coefficients: Increases in parking space requirement, arrival month, repeat guest status, special requests, and Corporate or Offline market segments lower cancellation chances.
- P-Values: There are no longer p-values greater than 0.05
- This is the final model

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25377
Method:	MLE	Df Model:	14
Date:	Mon, 12 Feb 2024	Pseudo R-squ.:	0.3254
Time:	08:07:45	Log-Likelihood:	-10855.
converged:	True	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-945.1028	119.540	-7.906	0.000	-1179.397	-710.809
no_of_adults	0.0736	0.036	2.068	0.039	0.004	0.143
required_car_parking_space	-1.5900	0.138	-11.512	0.000	-1.861	-1.319
lead_time	0.0156	0.000	59.576	0.000	0.015	0.016
arrival_year	0.4671	0.059	7.884	0.000	0.351	0.583
arrival_month	-0.0398	0.006	-6.222	0.000	-0.052	-0.027
repeated_guest	-2.7976	0.552	-5.068	0.000	-3.879	-1.716
no_of_previous_cancellations	0.2331	0.076	3.059	0.002	0.084	0.382
avg_price_per_room	0.0164	0.001	26.774	0.000	0.015	0.018
no_of_special_requests	-1.4575	0.030	-48.745	0.000	-1.516	-1.399
total_nights	0.0549	0.009	5.832	0.000	0.036	0.073
type_of_meal_plan_Meal Plan 2	0.2414	0.066	3.680	0.000	0.113	0.370
type_of_meal_plan_Not Selected	0.3467	0.051	6.818	0.000	0.247	0.446
market_segment_type_Corporate	-0.8114	0.102	-7.956	0.000	-1.011	-0.611
market_segment_type_Offline	-1.7548	0.051	-34.686	0.000	-1.854	-1.656

Model Performance Evaluation and Improvement - Logistic Regression

Increasing the threshold on the Train set did not improve the metrics.

- Accuracy remained at 0.8
- Recall decreased
- Precision decreased
- F1-score remained the same

	Logistic Regression sklearn	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80	0.80	0.80
Recall	0.62	0.73	0.70
Precision	0.74	0.68	0.70
F1	0.68	0.70	0.70

Increasing the threshold on Test set also did not improve metrics

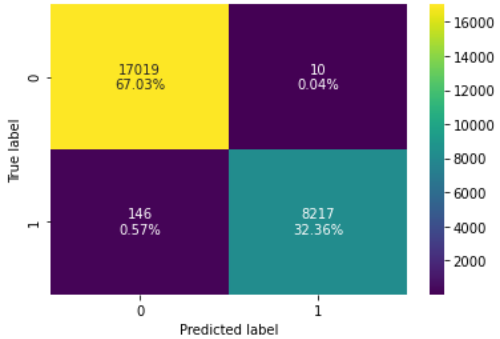
- Accuracy remained at 0.8
- Recall decreased
- Precision increased from 0.68 to 0.69
- F1-score remained the same

	Logistic Regression sklearn	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.81	0.80	0.80
Recall	0.63	0.73	0.70
Precision	0.73	0.68	0.69
F1	0.68	0.70	0.70

Model Building - Decision Tree

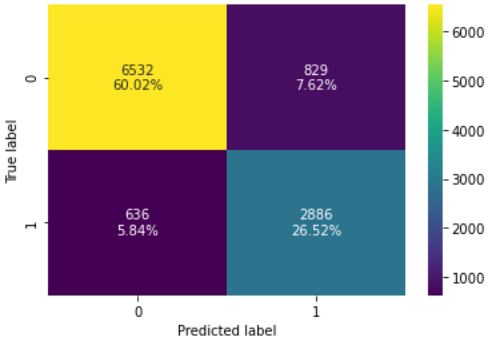
Preprune, Train

- Only 146 misclassified bookings
- Likely overfitted
- F1 Score: 0.99



Preprune, Test

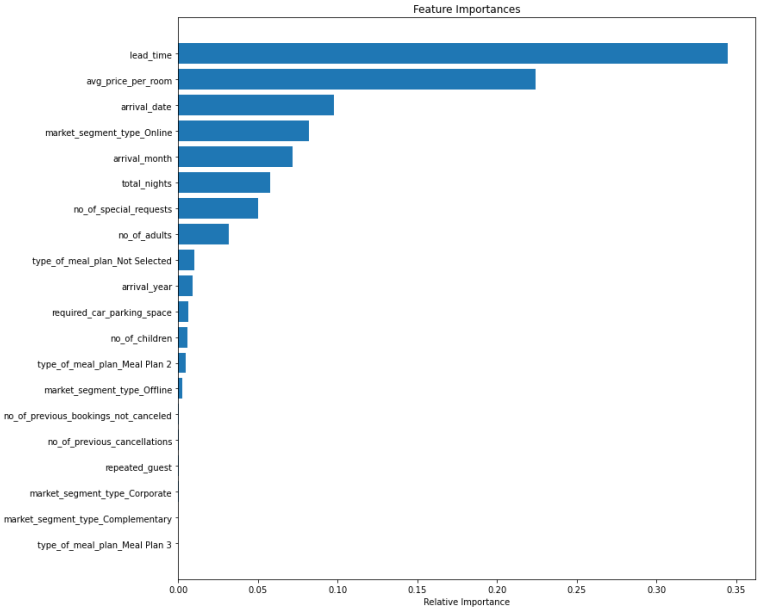
- Large margins from train metrics
- F1 Score: 0.80
- Overfitting confirmed



Model Building - Decision Tree

Check for important features

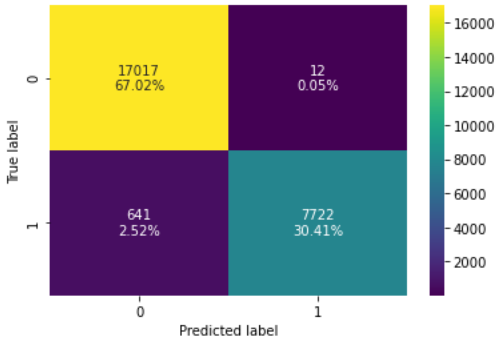
- lead_time and avg_price_per_room are the two most important variables



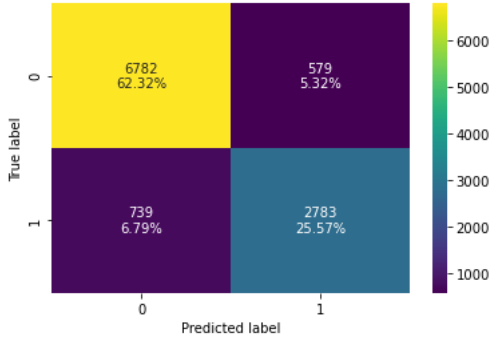
Model Building - Decision Tree

Using GridSearch

Train



Test



- Model performs better after using GridSearch

Model Building - Decision Tree

After CCP (Cost Complexity Pruning)

Training performance comparison:

	Decision-Tree sklearn	Decision-Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
F1	0.99	0.96	0.95

Test performance comparison:

	Decision-Tree sklearn	Decision-Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
F1	0.80	0.81	0.80

- Despite tuning efforts, both pre-pruning and post-pruning slightly reduced overfitting in the decision-tree model. The pre-pruned model still shows a significant performance gap between training and testing, indicating lower generalizability compared to the best logistic regression model.

Great Learning

Happy Learning !