INN Hotels Group – Predicting Cancellations of Bookings

Machine Learning(Supervised Learning-Logistic Regression/Decision Trees)

Arpan Dinesh

05/28/2023

Contents

Executive Summary

The INN Hotels group in Portugal faces challenges with high booking cancellations, leading to revenue losses, increased costs, and reduced profit margins.

To address this, a data-driven solution was implemented, utilizing exploratory data analysis, outlier treatment, feature engineering, logistic regression modeling, and decision trees.

The solution focused on predicting cancellations to formulate profitable policies for managing cancellations and refunds. By analyzing patterns and factors contributing to cancellations, the solution provided insights to prioritize bookings, target non-special requests, and leverage online market segments.

These recommendations aimed to optimize operations, minimize revenue losses, and enhance profitability for the INN Hotels group.

Business Problem

The INN Hotels group in Portugal faces a significant challenge with a high number of booking cancellations, resulting in various negative impacts.

These cancellations lead to revenue losses as rooms cannot be resold, increased distribution costs through commissions and advertising, and reduced profit margins due to last-minute price reductions.

Additionally, human resources are required to manage guest arrangements. The emergence of online booking channels has further complicated the issue, as cancellations are no longer limited to traditional booking patterns.

To mitigate these losses and optimize operations, a data-driven solution is needed to predict cancellations, allowing the formulation of profitable policies for managing cancellations and refunds.

Solution Approach

1. Exploratory Data Analysis: Data Overview, Univariate & Bi-Variate Analysis.

2. Data Cleaning: Outlier Check and Treatment, Missing/Duplicate Value Check, Feature Engineering.

3. Logistic Regression: Thresholds, ROC/Precision-Recall Curves, Performance Optimization, Coefficients Interpretation.

4. Decision Tree: Performance Criteria, Pre-Pruning, Post-Pruning, Final Model Selection & Evaluation.

5. Final Model Selection: Compare Logistic Regression Metrics with Decision Tree Metrics.

6. Business Recommendations: Actionable Insights, Profitable Suggestions.

Data Overview

The data contains different attributes of a customer booking. Detailed dictionary:

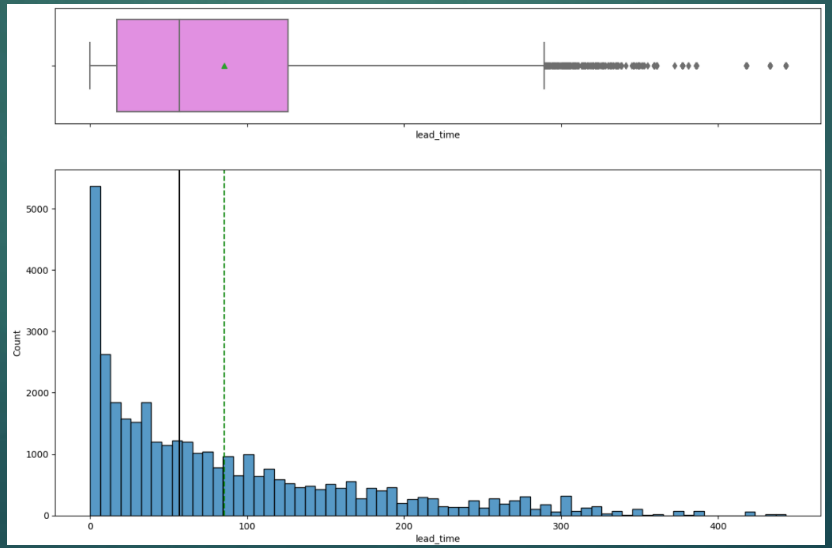
- Booking_ID: the unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
  - Not Selected – No meal plan selected
  - Meal Plan 1 – Breakfast
  - Meal Plan 2 – Half board (breakfast and one other meal)
  - Meal Plan 3 – Full board (breakfast, lunch, and dinner)

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

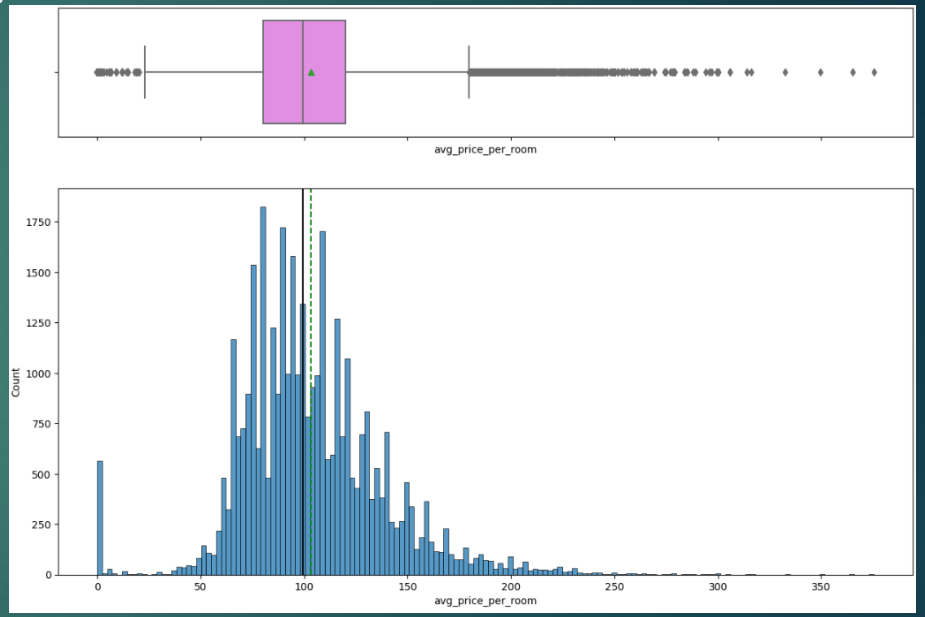
EDA - Univariate(Lead Time)

- The data for lead time is right skewed.
- Customers who took longer than 10 months to arrive after a booking can be classified as outliers.
- The median time for customers to arrive after a booking is made is roughly 2 months.
- Most customers arrived the same day they made a booking.



EDA - Univariate(Average Price/Room)

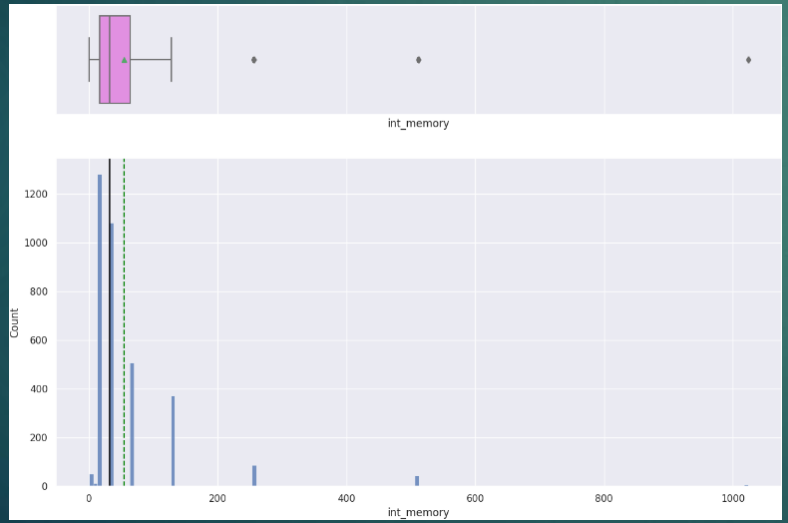
- Prices > $500 are capped to the upper whisker($179.55)
- The median price/room is roughly $100.
- The average price/room is skewed to the right.
- There are many bookings in which customers paid nothing.
- Amongst the booking that had a price of $0, most bookings were complementary while the others were booked online.

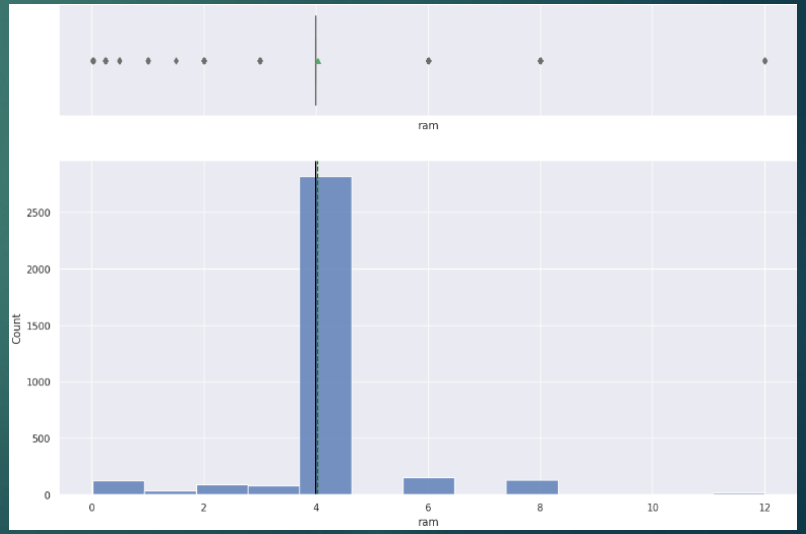


Complementary	354
Online	191

EDA - Univariate(Memory/Ram)

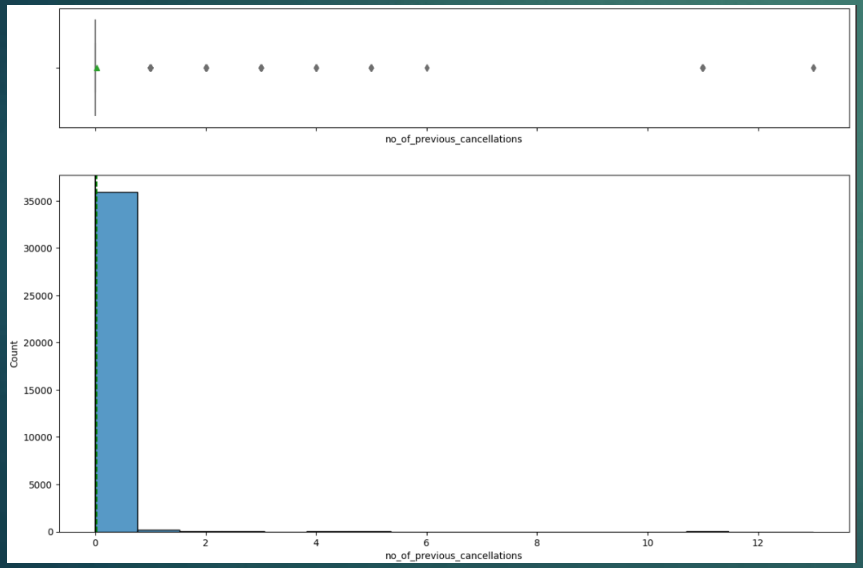
- Most cells have a low internal memory but a few have an incredibly large memory.
- There are 5 fixed values of internal memory a cell can take.
- Most cells have 4GB of RAM.
- A few cells have more than 6 GB of RAM.

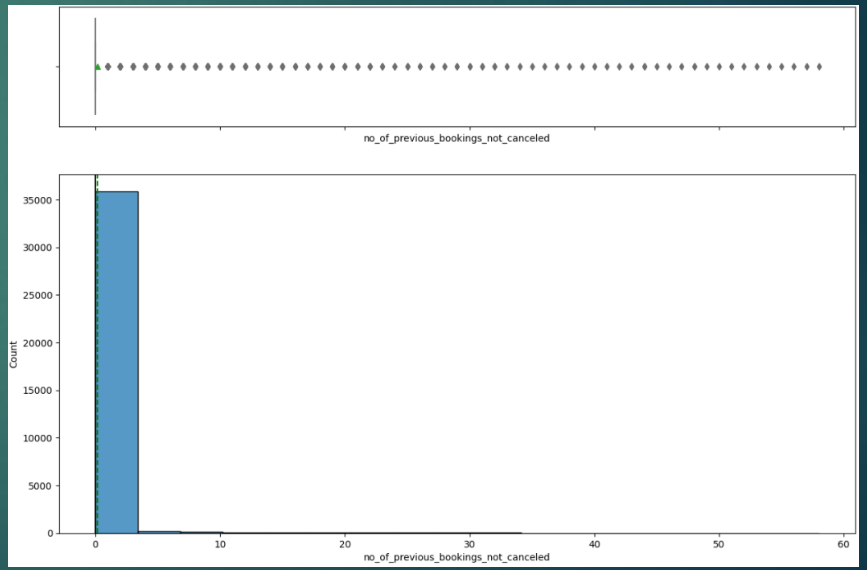




EDA - Univariate(Booking Status)

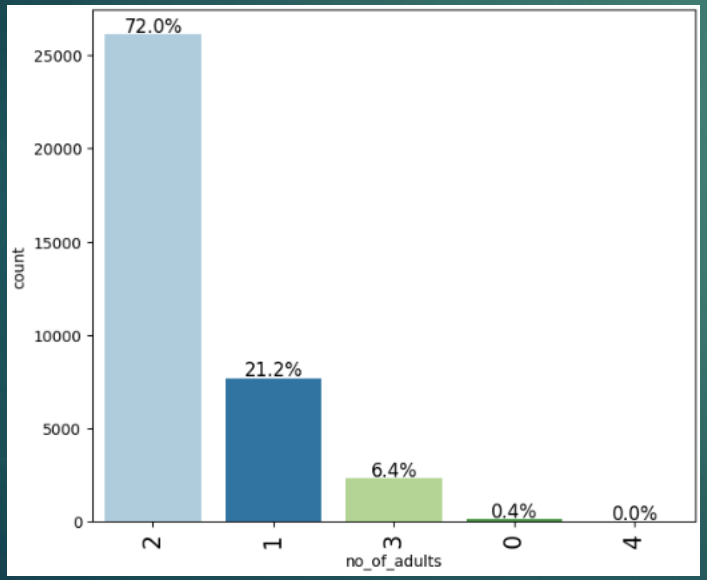
- As expected, very few customers have made previous cancellations.
- Not many previous bookings have not been cancelled.



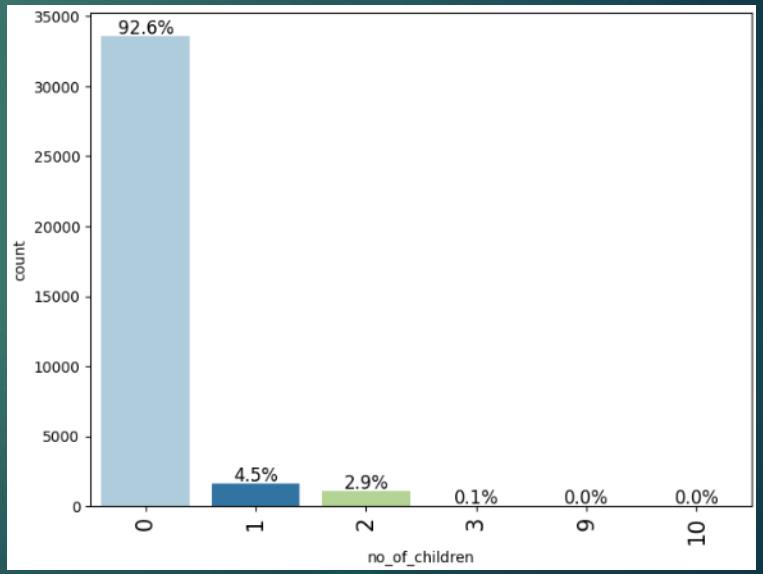


EDA - Univariate(Number of Adults/Children)

- Most bookings had 2 adults.
- Seems like most bookings were made by couples.

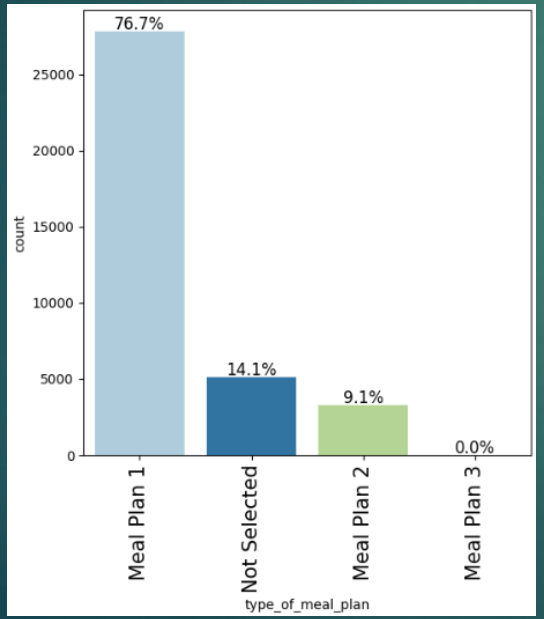


- Bookings with kids constitute less than 6% of all bookings made.
- 1 or 2 adults make up more than 90% of all bookings.

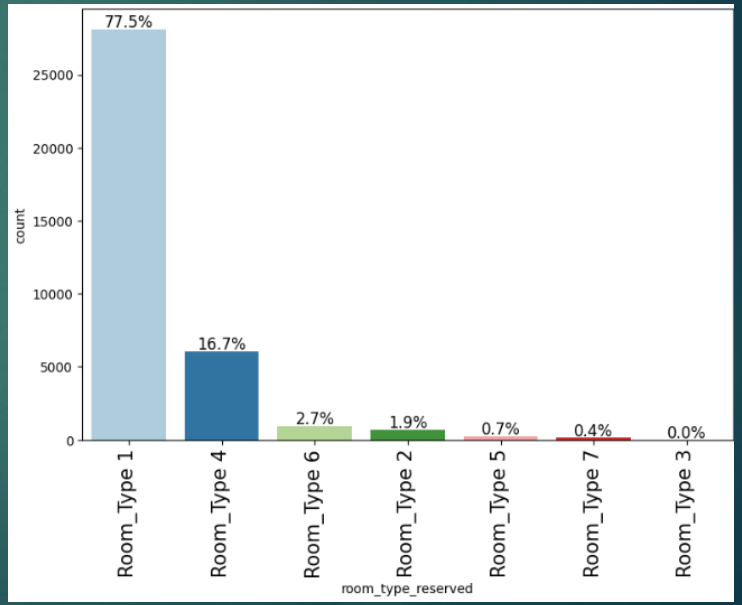


EDA - Univariate(Room Type/ Meal Plan)

- Most bookings chose Meal Plan 1.
- 14% of customers did not choose a meal plan.

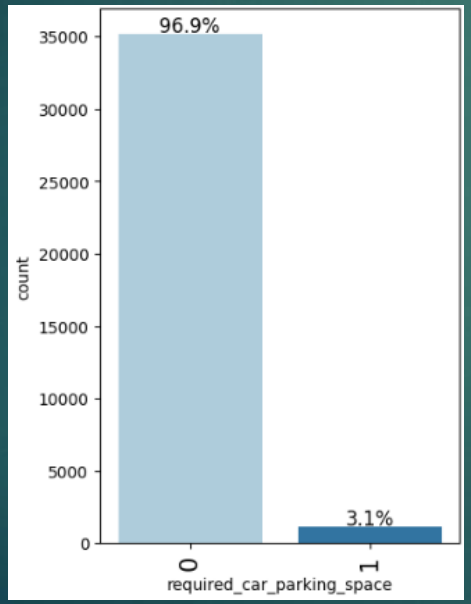


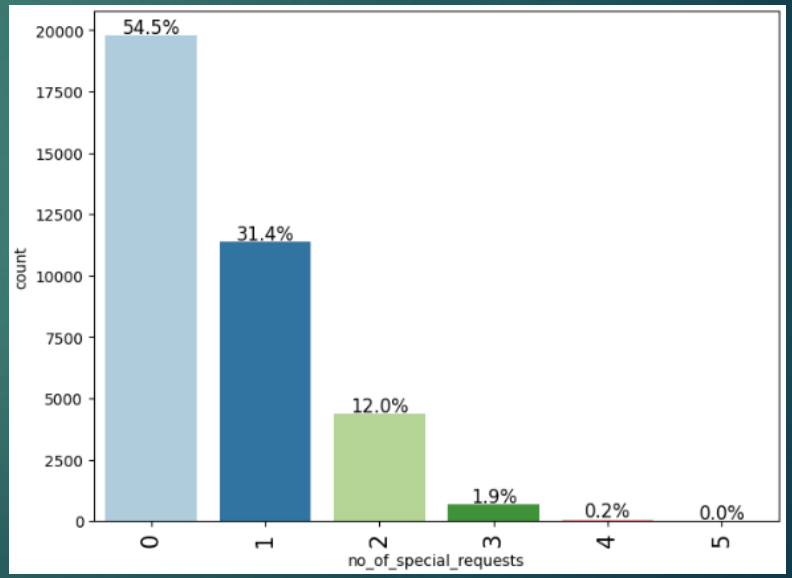
- 77% of bookings are in room type 1.
- 95% of bookings are of room types 1, 4 or 6.



EDA - Univariate(Car Parking Space/# of Special Requests)

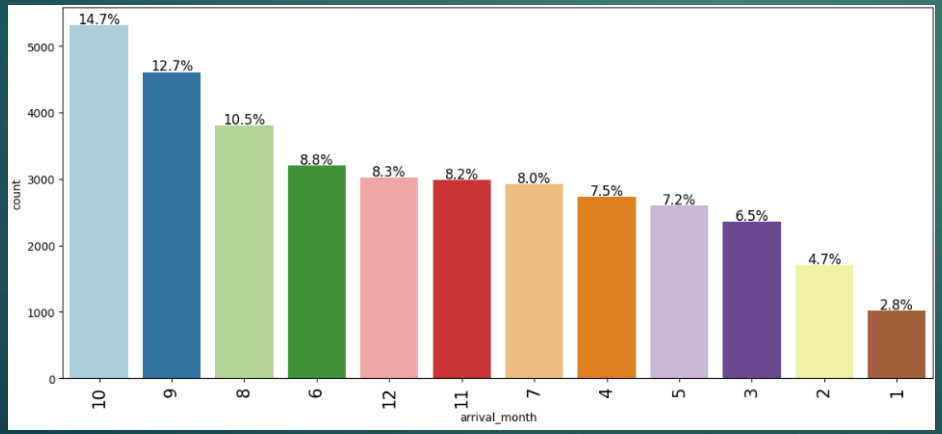
- Most bookings did not require a carking space.
- More than 50% of bookings did not have a special request.
- 2% of bookings had 3 special requests.
- 31% of bookings had 1 special request.

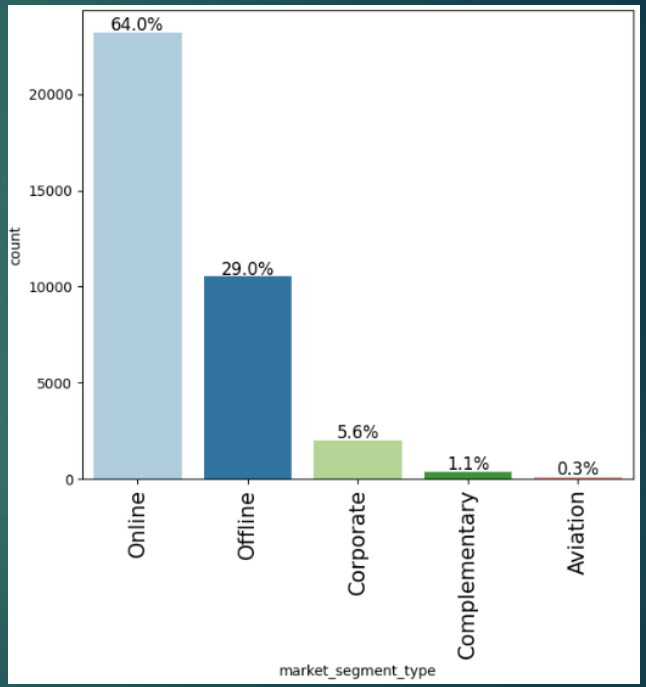




EDA - Univariate(Arrival Month/ Market Segment Type)

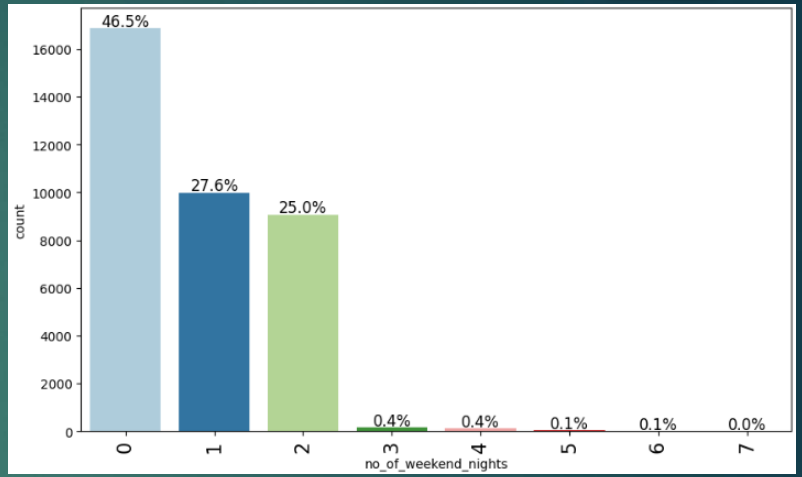
- More than 36% of bookings are made between August to October.
- 93% of bookings are in online or offline market segments.
- The start of the year sees the least number of bookings.
- Only 0.3% of bookings are from the aviation market segment.

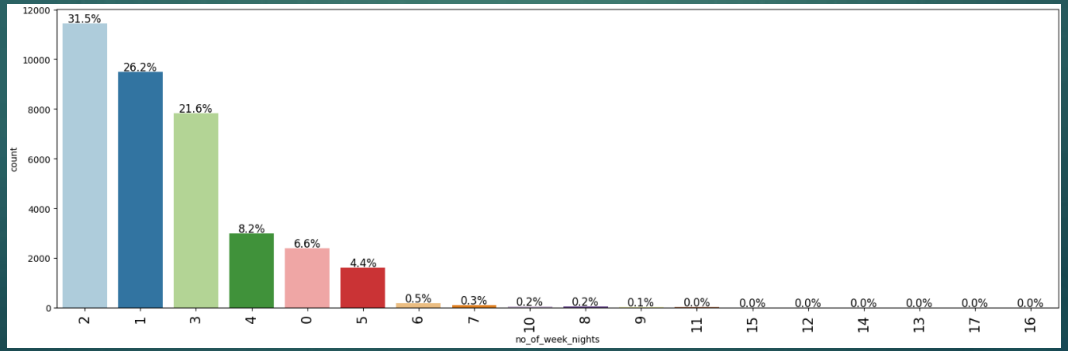




EDA - Univariate(Number of Weekday/Weekend Nights)

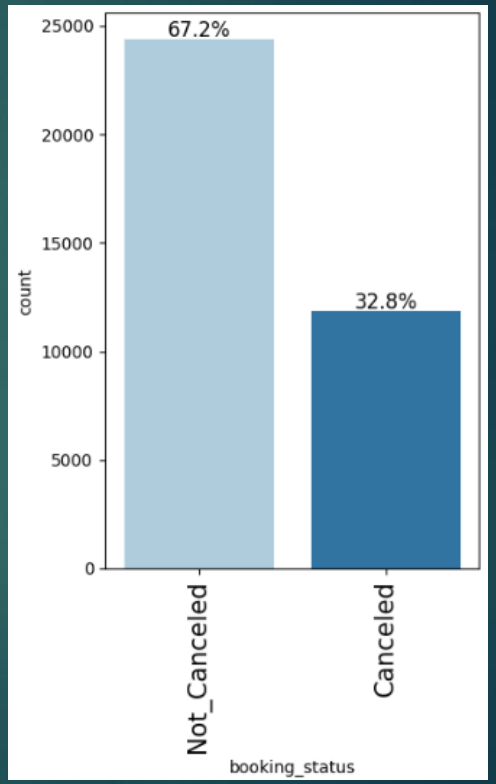
- Surprisingly, 46.5% of bookings were made on weekdays only.
- 56% of bookings are made for 1 or 2 week nights.
- Only 6.6% of bookings are made for no weekday.
- Few bookings are made for more than 2 weekend nights possible suggesting bookings for couple of weeks.





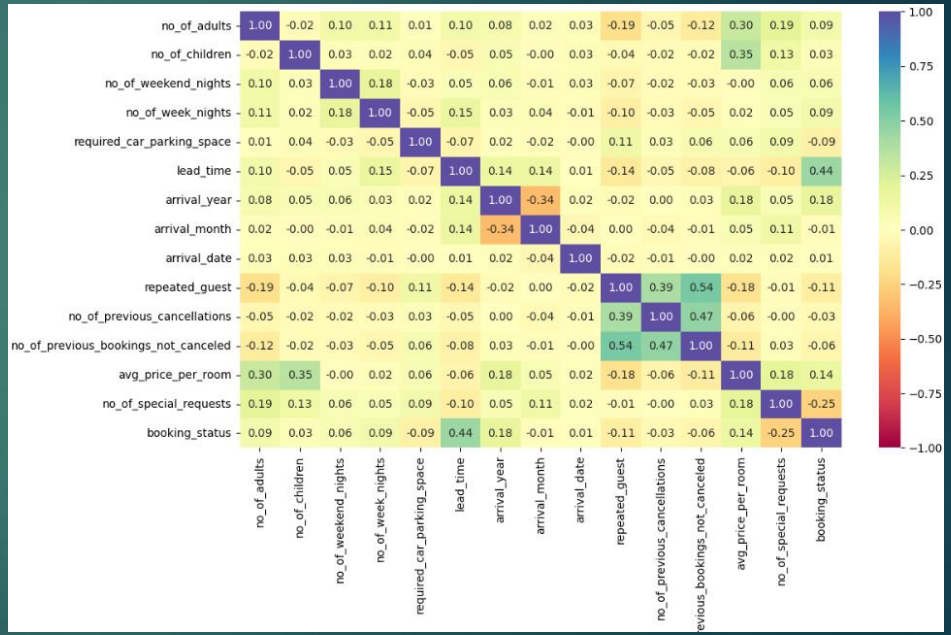
EDA - Univariate(Booking Status)

- The response variable in question has 67% of not canceled bookings.
- 33% of bookings got cancelled.
- Canceled bookings is encoded a 1, while not canceled bookings are encoded a 0.
- Although 67% of bookings were not cancelled, 33% of canceled bookings could cost INN Hotels a lot and would need to be accurately predicted.



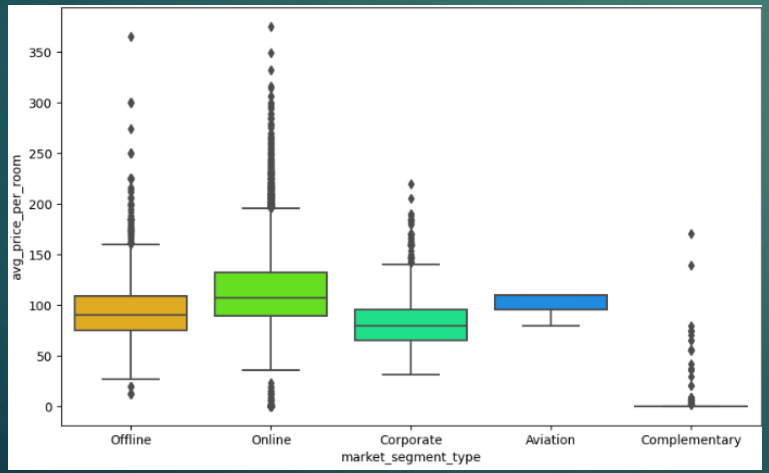
EDA – Bivariate(Correlation)

- None of the variables have a very strong positive or negative correlation with each other.
- The strongest positive correlation is between repeated guest and no of previous bookings not cancelled.
- The strongest negative correlation is between arrival year and month which is irrelevant to our analysis.
- Interestingly, the average price per room is not strongly correlated to any other variables.

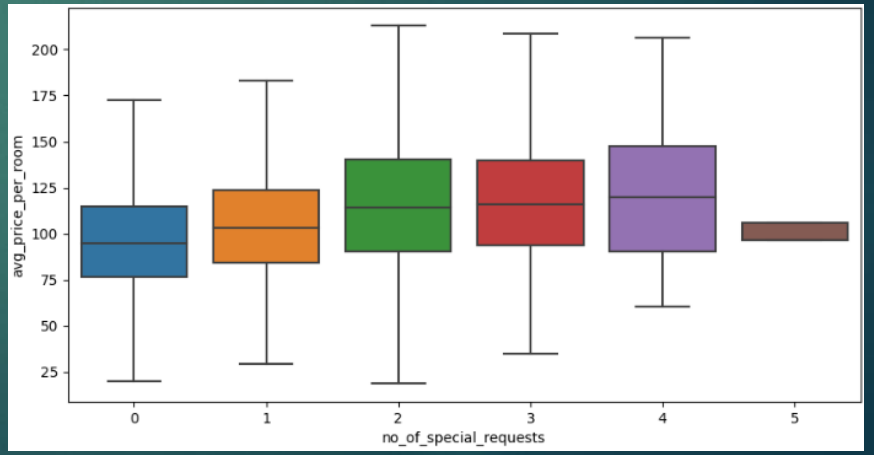


EDA – Bivariate(Price across Market Segments/Special Requests)

- Online markets seem to have the more expensive room prices in general.
- Corporate segments have lower prices possibly due to offers.
- Offline markets are more expensive that corporates but less expensive than online bookings.

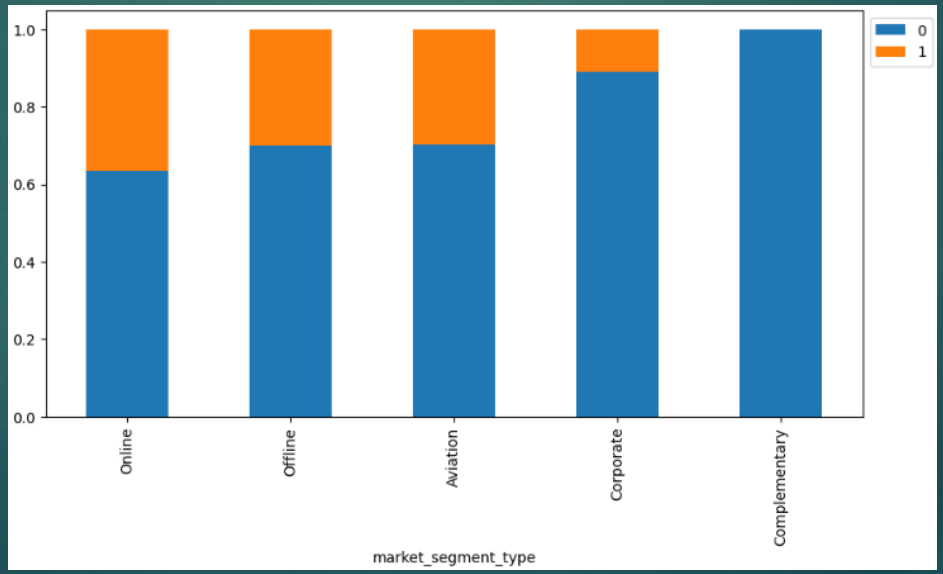


- Expectedly, the lesser the special requests the cheaper the price for a room.
- 2 special requests have the largest variation in prices.
- More than 4 special requests will have at least $75 in price for a room.



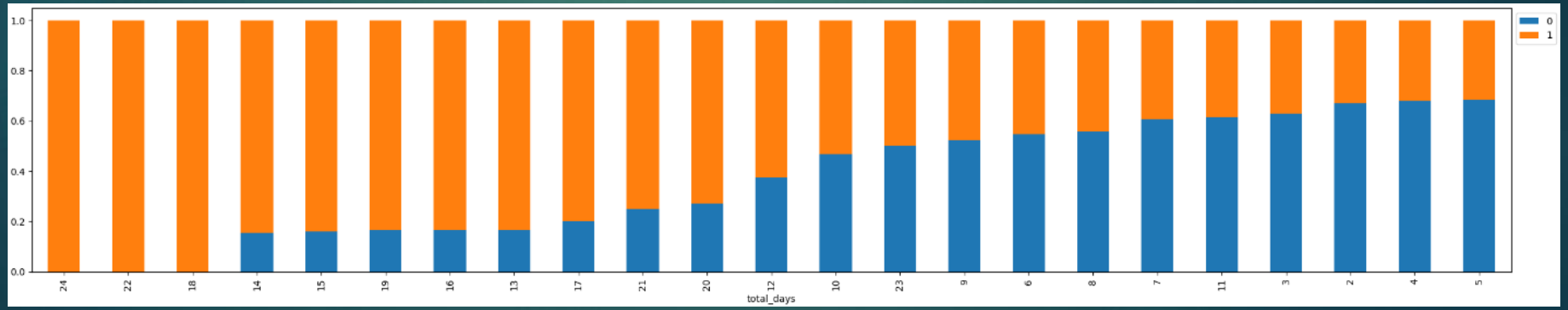
EDA - Bivariate(Booking Status across Markets)

- Expectedly online bookings see the most cancellations.
- Offline and Aviation markets have roughly a 30% cancellation rate.
- Corporate bookings have less than a 20% cancellation rate.
- Not surprisingly complementary bookings have 0 cancellations.



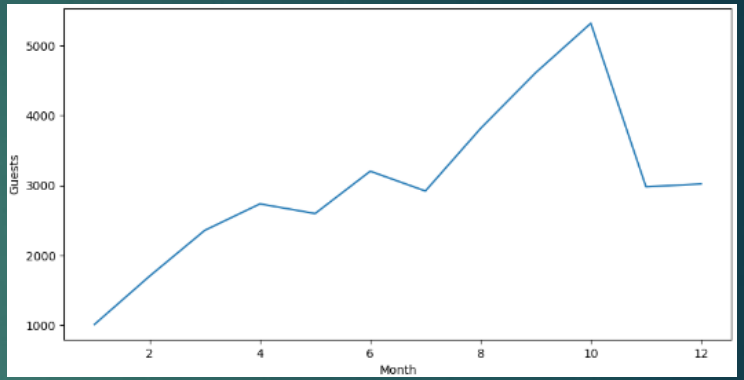
EDA - Bivariate(Booking Status across total days booked)

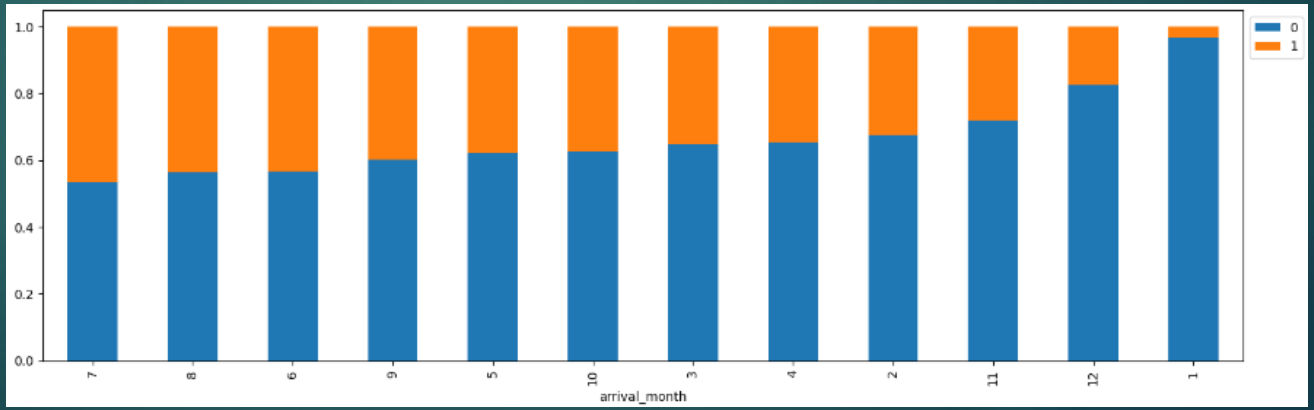
- All bookings for 18+ days had cancellations.
- Expectedly, the chances of customers cancelling their bookings increases with the number of days booked for stay.
- More than half of bookings booked for more than 7 days of stay have been cancelled.
- Bookings for less than 5 days saw the least number of cancellations.



EDA - Bivariate(Booking Status across Months)

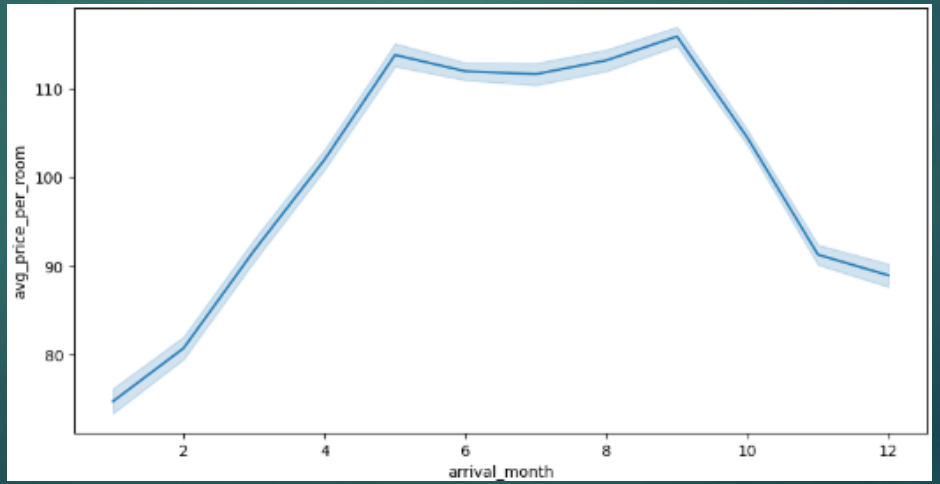
- January had the least cancellations but also the least bookings.
- Of the 3 busiest months(Aug-Oct), October had the least fraction of cancellations.
- Almost 50% of bookings in July got cancelled.





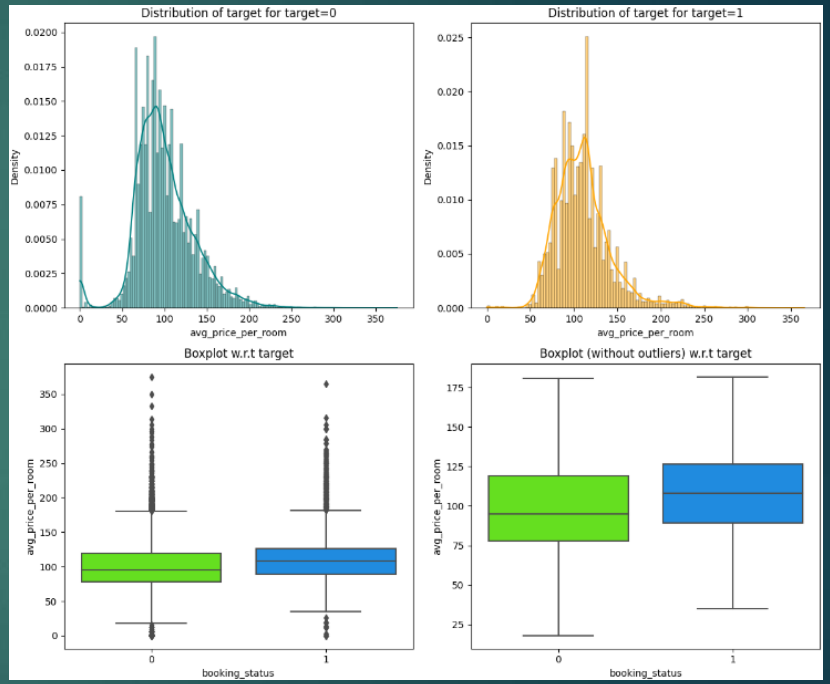
EDA - Bivariate(Room Price across Months)

- Prices are the highest during summers.
- Prices of rooms tend to increase till the end of summer and then go back down as winter approaches.
- January-March sees the cheapest prices of rooms.
- Prices of rooms stays relatively the same from May to September.



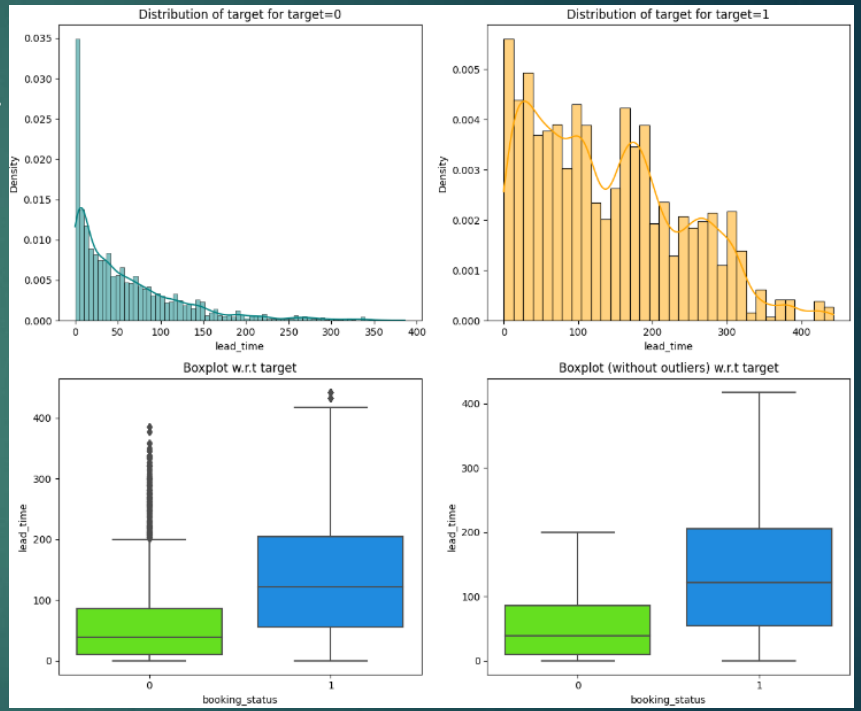
EDA - Bivariate(Booking Status on different Prices)

- Expectedly , cancelled bookings seem to have a higher price for a room than non cancelled booking. Offline and Aviation markets have roughly a 30% cancellation rate.
- Bookings with no cancellations have a larger variation in prices.
- Expectedly, bookings with no cancellations have the cheapest prices.
- A price of roughly $125 saw the most cancellations.



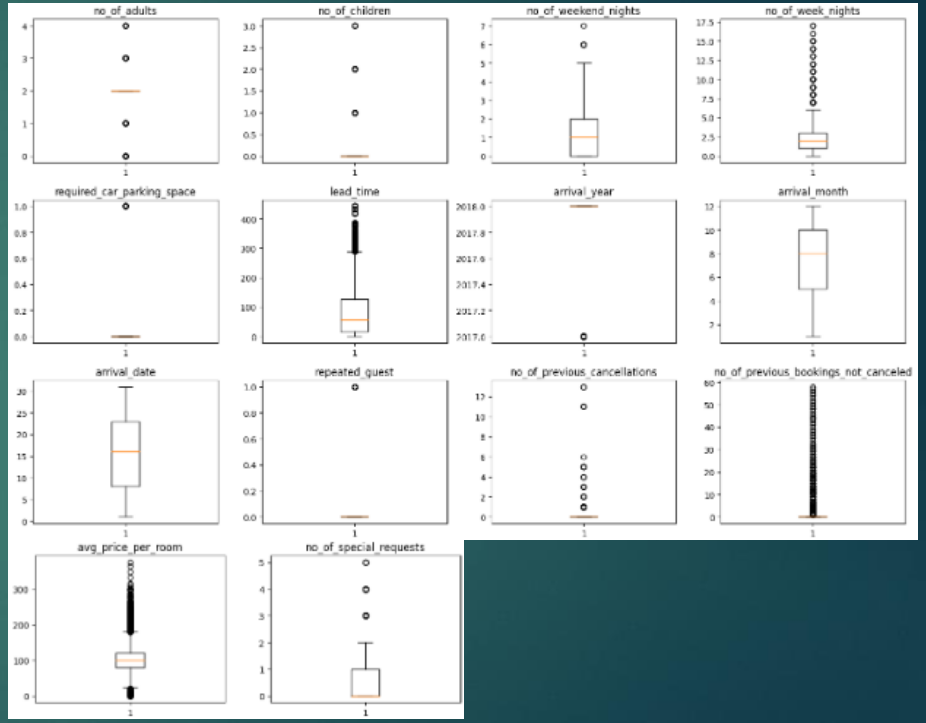
EDA - Bivariate(Booking Status on different Lead Times)

- Expectedly , people who booked rooms well in advance had a higher tendency to cancel bookings. Bookings with no cancellations have a larger variation in prices.
- Having said that, there is a large variation in lead time for people who cancelled their bookings.
- People who did not cancel their booking arrived at the hotel generally 3 months after booking it.
- It would be fair to assume that a person's chances of cancelling a booking increases if their lead time is more than 4 months.



Data Pre-Processing – Outlier Check/Treatment

- Outliers can be noticed in many variables.
- Price of a room above $500 is capped to the upper whisker($180) since a price above $500/room seems unreasonable and will skew results.
- Outliers in number of weekend nights and week nights will be capped to the upper whisker since it is rare to have bookings in hotel rooms for more than a week.
- The rest of the variables with outliers will be left untreated as they all seem to be real data.



Data Pre-Processing

- There are no missing values in the data set. A price of $0 for rooms potentially due to offers or benefits.
- There are no duplicate records in the dataset.
- No feature engineering implemented into logit regression or decision tree modelling. Total Days is computed only for EAD purposes.
- All categorical variables have been converted to dummy variables with binary units(0 or 1).
- The booking status(Cancelled/Not Cancelled) is the dependent(response) variable, and is encoded as 1 for canceled and 0 for not canceled.
- Data is split into train and test set with a 70:30 ratio. The classes in each set are very similar(67% not cancelled & 33% cancelled).

MODEL EVALUATION CRITERION

Types of Possible Errors:

1. Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking. (False Negative)

2. Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking. (False Positive)

Both cases are very important in this case:

1. If the model predicts that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.

2. If the model predicts that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled.

Therefore, the Hotel would want the F1 Score to be maximized: greater the F1 score, higher the chances of minimizing False Negatives and False Positives.

LOGISTIC REGRESSION – Initial Model

- The initial regression model is created(Figure top right) with a default threshold of 0.5
- Some categorical variables were dropped after dummies were created.
- There are some variables with a p-value > 0.05 which can be dropped.
- The model produces an F1 score of 0.68, can predict 63% of all true cancellations, and has an accuracy of 80%(Figure bottom right).
- Important to note that model performance could be improved by removing variables with high VIF and P-values, and by changing thresholds.(Following slides).

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Wed, 31 May 2023	Pseudo R-squ.:	0.3289
Time:	21:44:57	Log-Likelihood:	-10799.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-924.8595	120.900	-7.650	0.000	-1161.819	-687.900
no_of_adults	0.1136	0.038	3.015	0.003	0.040	0.187
no_of_children	0.1599	0.062	2.574	0.010	0.038	0.282
no_of_weekend_nights	0.1135	0.020	5.774	0.000	0.075	0.152
no_of_week_nights	0.0161	0.014	1.188	0.235	-0.010	0.043
required_car_parking_space	-1.5984	0.138	-11.595	0.000	-1.869	-1.328
lead_time	0.0158	0.000	58.915	0.000	0.015	0.016
arrival_year	0.4571	0.060	7.630	0.000	0.340	0.575
arrival_month	-0.0416	0.006	-6.422	0.000	-0.054	-0.029
arrival_date	0.0005	0.002	0.234	0.815	-0.003	0.004
repeated_guest	-2.3623	0.618	-3.825	0.000	-3.573	-1.152
no_of_previous_cancellations	0.2657	0.086	3.095	0.002	0.097	0.434
no_of_previous_bookings_not_canceled	-0.1731	0.153	-1.131	0.258	-0.473	0.127
avg_price_per_room	0.0187	0.001	25.249	0.000	0.017	0.020
no_of_special_requests	-1.4673	0.030	-48.758	0.000	-1.526	-1.408
type_of_meal_plan_Meal Plan 2	0.1709	0.067	2.564	0.010	0.040	0.302
type_of_meal_plan_Meal Plan 3	19.7883	1.35e+04	0.001	0.999	-2.64e+04	2.64e+04
type_of_meal_plan_Not Selected	0.2713	0.053	5.114	0.000	0.167	0.375
room_type_reserved_Room_Type 2	-0.3627	0.131	-2.763	0.006	-0.620	-0.105
room_type_reserved_Room_Type 3	-0.0151	1.313	-0.011	0.991	-2.589	2.559
room_type_reserved_Room_Type 4	-0.2732	0.053	-5.134	0.000	-0.377	-0.169
room_type_reserved_Room_Type 5	-0.7111	0.209	-3.398	0.001	-1.121	-0.301
room_type_reserved_Room_Type 6	-0.9425	0.151	-6.227	0.000	-1.239	-0.646
room_type_reserved_Room_Type 7	-1.3787	0.294	-4.696	0.000	-1.954	-0.803
market_segment_type_Complementary	-48.0777	6.88e+06	-6.99e-06	1.000	-1.35e+07	1.35e+07
market_segment_type_Corporate	-1.2226	0.264	-4.624	0.000	-1.741	-0.704
market_segment_type_Offline	-2.2229	0.253	-8.788	0.000	-2.719	-1.727
market_segment_type_Online	-0.4217	0.250	-1.689	0.091	-0.911	0.068

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80553	0.63219	0.73954	0.68167

LOGISTIC REGRESSION – Multicollinearity & P-Value Variables

MULTICOLLINEARITY

- Only the market segment offline and online variables have a VIF value > 5, but they are categorical variables so they will be disregarded.

	feature	VIF
0	const	39496424.64296
1	no_of_adults	1.35324
2	no_of_children	2.09420
3	no_of_weekend_nights	1.05469
4	no_of_week_nights	1.09093
5	required_car_parking_space	1.04011
6	lead_time	1.39829
7	arrival_year	1.43186
8	arrival_month	1.27652
9	arrival_date	1.00675
10	repeated_guest	1.78368
11	no_of_previous_cancellations	1.39571
12	no_of_previous_bookings_not_canceled	1.65208
13	avg_price_per_room	2.07026
14	no_of_special_requests	1.24814
15	type_of_meal_plan_Meal Plan 2	1.27407
16	type_of_meal_plan_Meal Plan 3	1.02526
17	type_of_meal_plan_Not Selected	1.27498
18	room_type_reserved_Room_Type 2	1.10593
19	room_type_reserved_Room_Type 3	1.00330
20	room_type_reserved_Room_Type 4	1.36439
21	room_type_reserved_Room_Type 5	1.02808
22	room_type_reserved_Room_Type 6	2.05601
23	room_type_reserved_Room_Type 7	1.11810
24	market_segment_type_Complementary	4.50036
25	market_segment_type_Corporate	16.91703
26	market_segment_type_Offline	64.07947
27	market_segment_type_Online	71.15263

High P-Value Variables

- 4 variables are dropped due to a p-value > 0.05 and only the select features(p-value > 0.05) are kept.
- A new model(without high p-values) is run and it's performance on the train and test is checked.

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80549	0.63207	0.73951	0.68158

Testing performance:

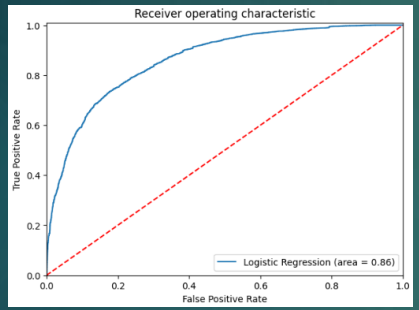
	Accuracy	Recall	Precision	F1
0	0.80557	0.63373	0.72989	0.67842

- Interestingly, the F1 score has not changed much even after dropping insignificant variables.
- The model is generalizing well since the difference between train and test metrics is low, but it's performance could be improved via different thresholds.

LOGISTIC REGRESSION – ROC Curve

TRAINING SET

- The ROC curve on train set gives an AUC of 0.86, with an optimal threshold of 0.37.
- The F1 score on the training set with the new threshold has improved to 0.7.

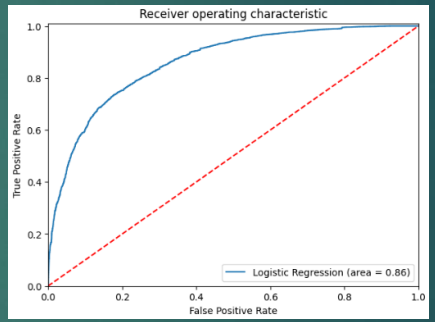


Training performance:

	Accuracy	Recall	Precision	F1
0	0.79379	0.73466	0.67067	0.70121

TEST SET

- The ROC curve on test set also gives an AUC of 0.86, with an optimal threshold of 0.37.
- The F1 score on the test set with the new threshold has improved to 0.7.



Test performance:

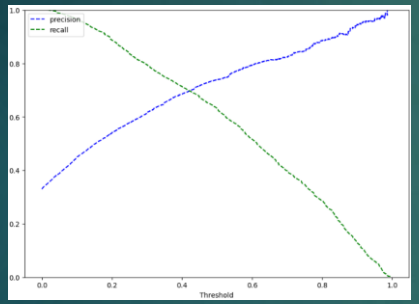
	Accuracy	Recall	Precision	F1
0	0.79564	0.73651	0.66684	0.69995

The F1 Score has improved with a threshold at 0.37 and the model is generalizing very well on the test as well. Further analysis on the F1 score needs to be conducted with a new threshold(Precision/Recall Curve).

LOGISTIC REGRESSION – Precision/Recall Curve

TRAINING SET

- The intersection of the precision-recall curve gives a threshold of 0.42.
- The F1 score on the training set with the new threshold is 0.69.

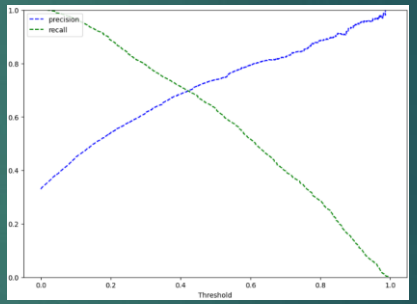


Training performance:

	Accuracy	Recall	Precision	F1
0	0.80021	0.69807	0.69616	0.69712

TEST SET

- The intersection of the precision-recall curve gives a threshold of 0.42.
- The F1 score on the test set with the new threshold has improved to 0.7.



Test performance:

	Accuracy	Recall	Precision	F1
0	0.80373	0.70415	0.69390	0.69899

The F1 Score at a threshold of 0.42 is higher than the threshold at the default threshold of 0.5, but has slightly reduced from the threshold at 0.37. The model is still generalizing very well on the test set too.

LOGISTIC REGRESSION – Model Coefficient Interpretation

- No. of Adults: For every one unit increase in the # of adults, the odds of a cancellation increase by 11.5%
- If people book with a request for a car parking space, the cancellation odds reduce by 79%!
- For every dollar increase in the price of a room, the odds of a cancellation increase by almost 2%.
- For every extra special request a person books, his or her chances of cancelling that booking decrease by 77%!
- If people do not select a meal plan, their chances of cancelling a booking increases by 31%.
- Bookings for room types 5 and above have a much lesser chance of being cancelled as compared to bookings for room types 2 or 4.
- Offline and Corporate market segments reduce the chances of a cancellation by 83% and 55% respectively.

	Odds	Change_odd%
const	0.00000	-100.00000
no_of_adults	1.11547	11.54742
no_of_children	1.16919	16.91886
no_of_weekend_nights	1.12587	12.58669
required_car_parking_space	0.20140	-79.86040
lead_time	1.01593	1.59336
arrival_year	1.57575	57.57528
arrival_month	0.95865	-4.13482
repeated_guest	0.06322	-93.67781
no_of_previous_cancellations	1.25530	25.52994
avg_price_per_room	1.01919	1.91883
no_of_special_requests	0.23049	-76.95084
type_of_meal_plan_Meal Plan 2	1.16856	16.85627
type_of_meal_plan_Not Selected	1.31421	31.42138
room_type_reserved_Room_Type 2	0.69889	-30.11126
room_type_reserved_Room_Type 4	0.76552	-23.44783
room_type_reserved_Room_Type 5	0.48528	-51.47249
room_type_reserved_Room_Type 6	0.38413	-61.58676
room_type_reserved_Room_Type 7	0.24655	-75.34467
market_segment_type_Corporate	0.44665	-55.33512
market_segment_type_Offline	0.16603	-83.39731

LOGISTIC REGRESSION – Model Comparison and Selection

The performance metrics of the model against 3 thresholds: Default(0.5), ROC Curve(0.37) & Precision-Recall Curve(0.42) on the train and test set is given below.

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80549	0.79379	0.80021
Recall	0.63207	0.73466	0.69807
Precision	0.73951	0.67067	0.69616
F1	0.68158	0.70121	0.69712

Testing performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80557	0.79564	0.80373
Recall	0.63373	0.73651	0.70415
Precision	0.72989	0.66684	0.69390
F1	0.67842	0.69995	0.69899

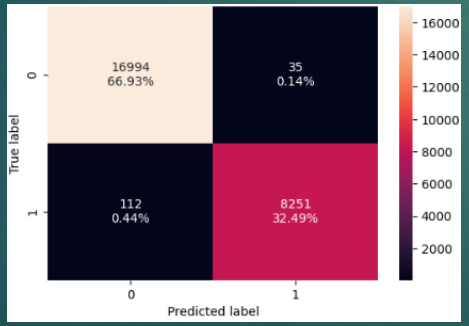
The F1 score is the highest at the 0.37 threshold obtained from the ROC curve, and it generalizes extremely well on unseen data, so that will be the best logistic regression model for INN Hotels to implement for predicting cancellations.

Further analysis on performance improvement will be conducted by developing a decision tree and comparing it's metrics with the logistic regression's.

DECISION TREE– Base Model – Performance Evaluation

TRAIN

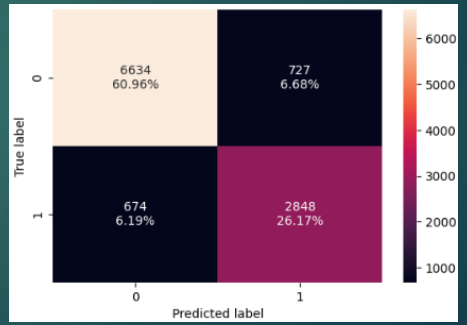
- The decision tree gives excellent metrics on the train set, but is definitely capturing noise.
- The tree is able to predict almost 99% of all cancelled bookings on the training data.
- The difference between the test and training metrics calls for pre-pruning and post-pruning the tree and evaluating the best model.



	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

TEST

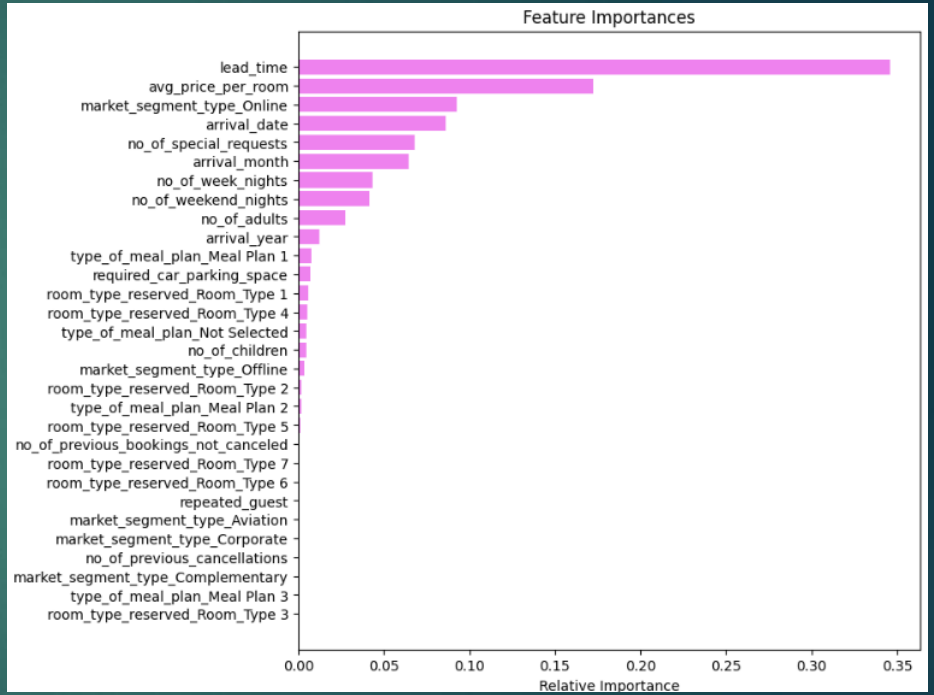
- Expectedly, the tree does not perform as well on the test set as it did on the training set.
- On unseen data, the tree seems to be able to predict 80% of all cancellations.
- The tree still maintains a large F1 score on unseen data.



	Accuracy	Recall	Precision	F1
0	0.87127	0.80863	0.79664	0.80259

DECISION TREE– Base Model – Feature Importance

- From the Base Model(Before Pruning), the most important factor in predicting a booking cancellation seems to be lead time.
- Expectedly, the price of a room is also very useful in predicting a cancellation of a booking.
- Interestingly, number of previous bookings not cancelled and repeated guest are major predictors in booking cancellations.
- It is important to notice how these features change after the tree is pruned(Shown in below slides).

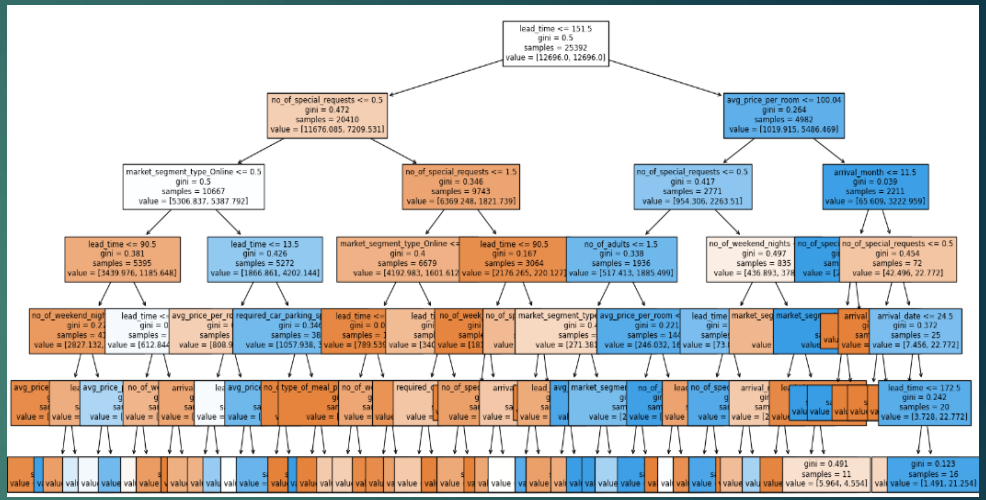


DECISION TREE– Pre- Pruning – Hyper-Parameter Tuning

Pre-Pruning will be conducted to trim the tree before it grows using Hyper-Parameters: Max Depth, Max Leaf Nodes & Min Samples Split.

The grid search function in Python will assess all combinations of these Hyper-Parameters(shown below) with respect to the F1 score(decided earlier), and a separate function will fit the best combination of the Hyper-Parameters(shown bottom right) to the data and trim the decision tree accordingly(shown top right).

The visual of the tree shows a shorter tree with lesser branches.




```
# Grid of parameters to choose from
parameters = {
    "max_depth": np.arange(2, 7, 2),
    "max_leaf_nodes": [50, 75, 150, 250],
    "min_samples_split": [10, 30, 50, 70],
}
```

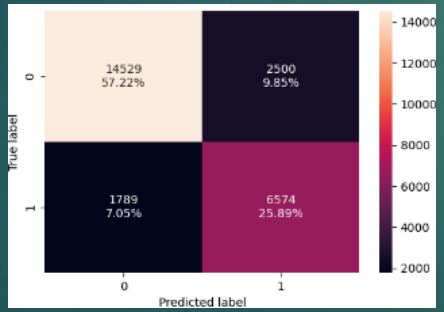
DecisionTreeClassifier

```
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=75,
                       min_samples_split=10, random_state=1)
```

DECISION TREE– Pre-Pruning – Performance Evaluation

TRAIN

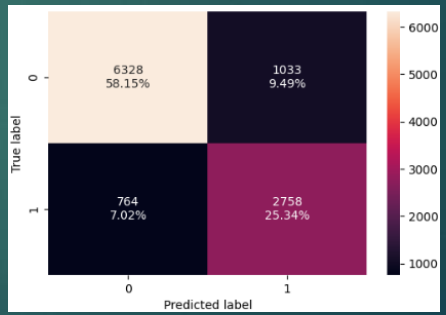
- Expectedly, the pre pruned tree's metrics dive to 0.75 from 0.99(Base Tree) on the F1 score.
- The pruned tree is still doing well on the Recall parameter.
- The similarities between the F1, Precision and Recall scores of the Train and Test sets suggest that this tree does a good job in prediction on unseen data as well.



Accuracy	Recall	Precision	F1
0.83109	0.78608	0.72449	0.75403

TEST

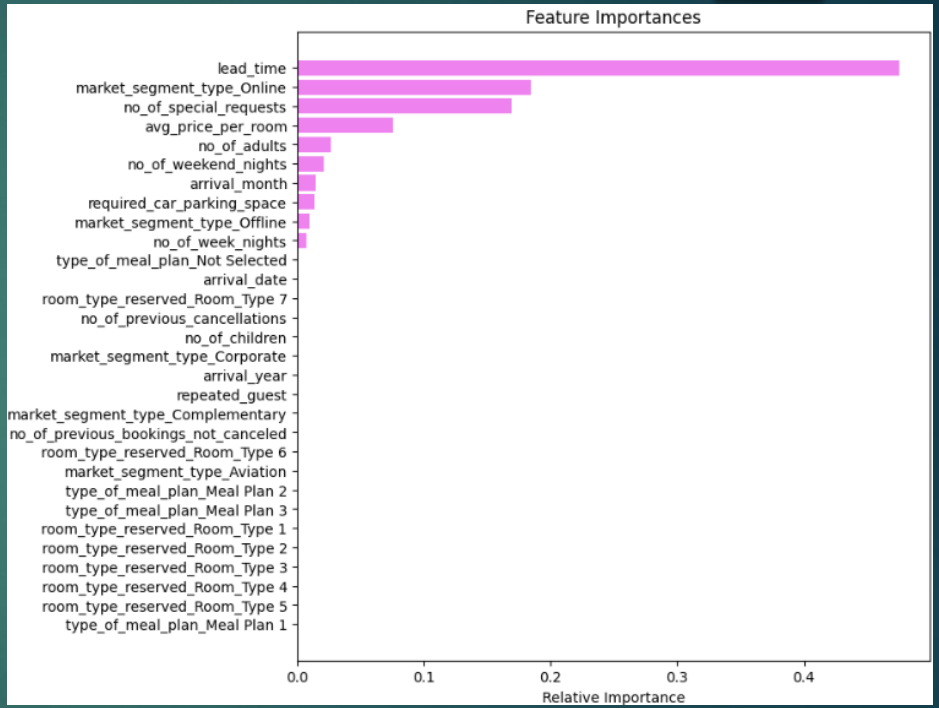
- The pre-pruned tree does better on the test set than the training set with respect to the F1 score.
- The tree is able to predict 78% of cancellations on unseen data which is high.
- It is important to compare this model to the one after post-pruning to finalize the model.



Accuracy	Recall	Precision	F1
0.83488	0.78308	0.72751	0.75427

❖ DECISION TREE– Pre-Pruning – Feature Importance

- Lead time is still the most important factor for predicting cancellations.
- As expected, the pre-pruned tree has trimmed the important features required for accurate predictions.
- The pre-pruned tree retains Market Segment Online, Price of a room, # of special requests and # of days as important features.
- It is important to notice how these features change after the tree is post-pruned(Shown in below slides).

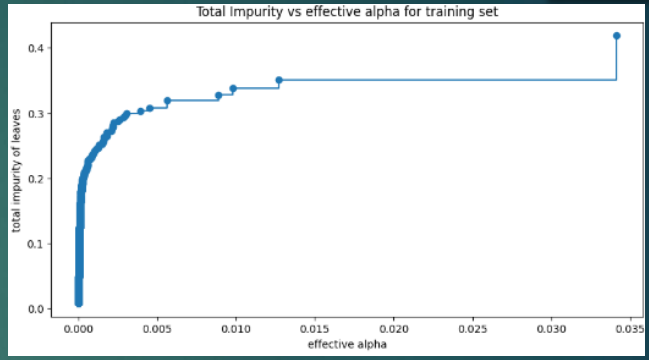


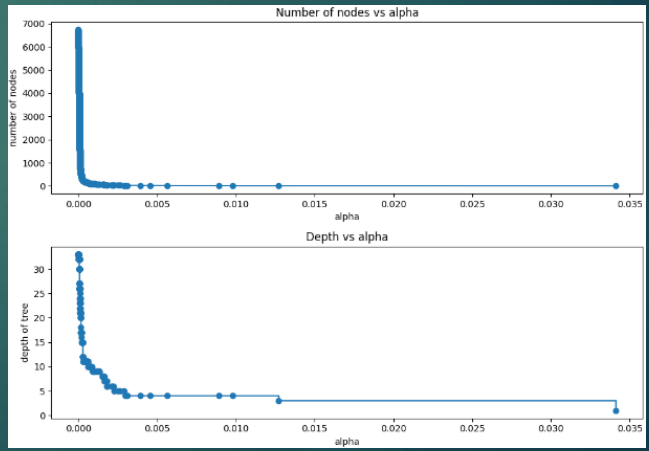
DECISION TREE– Post- Pruning – Cost Complexity Parameter

Post-Pruning will be conducted to trim the tree after it is grown to its fullest using the Post-pruning Parameters: Cost Complexity Parameter(Alpha), Effective Alpha and Cost Complexity Measure.

Starting from the leaf nodes, the nodes with the highest effective alpha values are pruned or cut off from the tree. Pruning a node involves removing all its descendant nodes and replacing them with a leaf node labeled with the majority class of the remaining samples in that subtree. As the effective alpha increases, more branches are cut from the tree and the total impurity increases.(Figure top right)

As alpha increases, the number of nodes and depth of the tree reduces.(Figures bottom right)



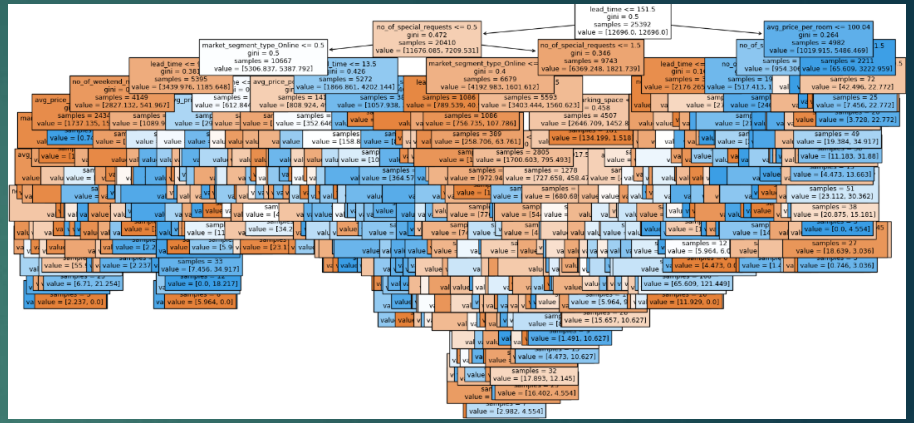


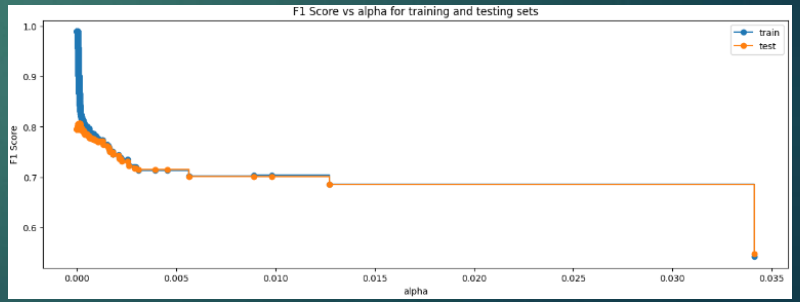
DECISION TREE– Post- Pruning – Final Tree

Cross-validation is used to find the optimal alpha value. The training data is divided into multiple subsets, and each subset is used as a validation set while the remaining data is used for training. The alpha value that yields the best performance on the validation sets (highestF1 score) is selected.

Using the optimal alpha value determined through cross-validation, the decision tree is pruned to the desired level of complexity. (Figure top right).

The F1 score is plotted against alpha for the train and test set. The value of alpha that yields that maximum F1 test score is the optimal alpha the decision tree is trimmed to. max.

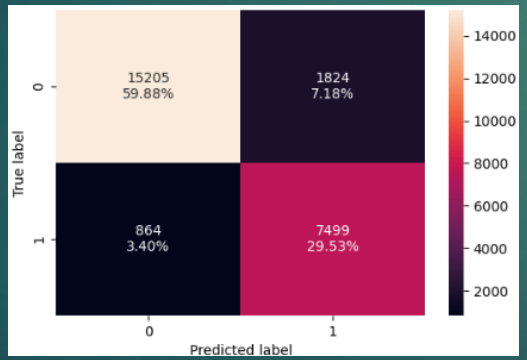




DECISION TREE– Post-Pruning – Performance Evaluation

TRAIN

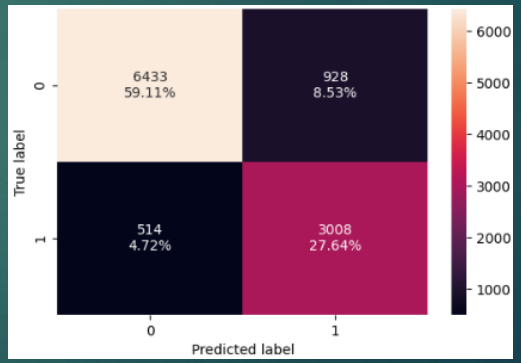
- The post pruned tree's metric jumps from 0.75(pre-pruned) to 0.85 on the F1 score, on the train set.
- All the metrics on the post-pruned tree are larger than the ones on the pre-pruned tree seen previously.
- The similarities of the metrics between train and test data suggests the model can predict accurately on unseen data.



Accuracy	Recall	Precision	F1
0.89414	0.89669	0.80435	0.84802

TEST

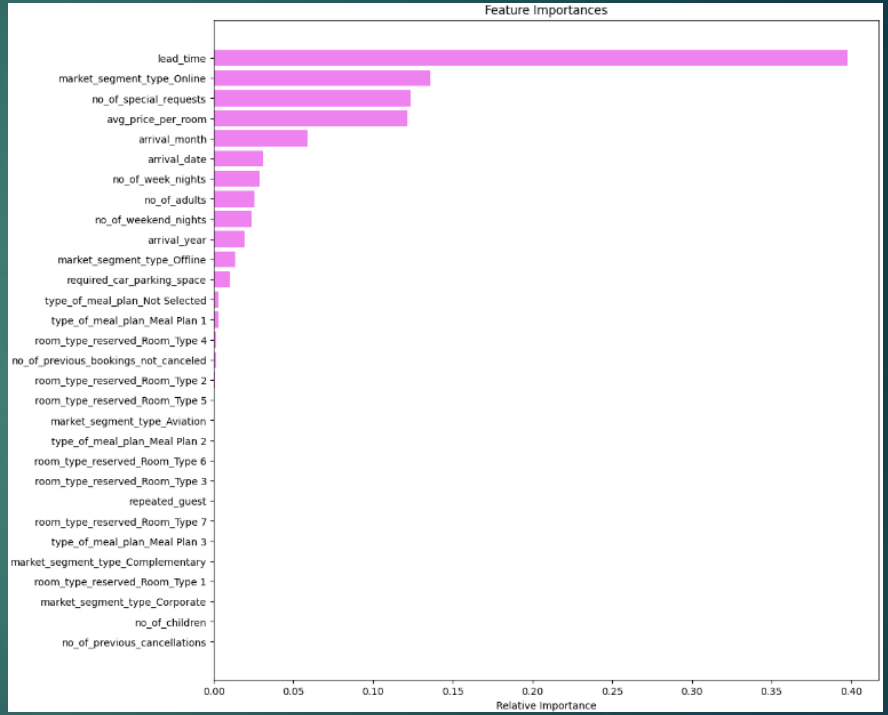
- The post-pruned tree does better than the pre-pruned tree on test data for all metrics.
- The post-pruned tree is able to predict 85% of all true cancellations on unseen data which is 7% higher than what the pre-pruned tree could predict. The Post-Pruned tree seems to be the best choice.



Accuracy	Recall	Precision	F1
0.86750	0.85406	0.76423	0.80665

DECISION TREE– Pre-Pruning – Feature Importance

- Lead time is still the most important factor for predicting cancellations.
- There isn't much change in important features from the pre-pruned tree.
- The post-pruned tree also retains Market Segment Online, Price of a room, # of special requests and # of days as important features.
- Arrival date is a more important feature in the post-pruned tree as compared to the pre-pruned one.



DECISION TREE– Models Comparison

- The F1 score is the metric in concern, and it is maximized in the post-pruning tree.
- The metrics between train and test set are a lot closer in the pre-pruning tree than the post-pruning one.
- A score of 0.806 is a very satisfying F1 score for the model on unseen data.
- The post-pruned tree is able to predict 85% of all true cancellations on unseen data which is 7% higher than what the pre-pruned tree could predict.
- Based on the highest F1 score and performance on both train and test sets, **The Post-Pruned tree is the final decision tree chosen to predict booking cancellations for INN Hotels**.

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83109	0.89414
Recall	0.98661	0.78608	0.89669
Precision	0.99578	0.72449	0.80435
F1	0.99117	0.75403	0.84802

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87182	0.83488	0.86750
Recall	0.80522	0.78308	0.85406
Precision	0.80000	0.72751	0.76423
F1	0.80260	0.75427	0.80665

FINAL MODELS COMPARISON – Logit Regression Model & Decision Tree

Best Decision Tree Metrics(Post-Pruned)	
Accuracy	0.8675
Precision	0.854
Recall	0.764
F1 Score	0.806

Best Logistic Regression Model(0.37 Threshold)	
Accuracy	0.8675
Precision	0.854
Recall	0.764
F1 Score	0.806

- The best logit regression model and the best decision tree's metrics on the test set are compared with each other.
- It is clear that the decision tree's metrics outperform the metric's of the logit regression model on the test set.

INN Hotels must use the Decision Tree for the best prediction capabilities of their future booking cancellations.

BUSINESS RECOMMENDATIONS – Decision Tree

- Prioritize Bookings with Low Lead Time: Focus on bookings with a lead time < 152 days. These bookings have shown a higher likelihood of not being canceled, indicating a higher probability of revenue generation. Allocate resources and attention to these bookings to maximize profitability.

- Target Non-Special Requests: Bookings with few special requests(<= 0.50) have demonstrated a lower likelihood of cancellation. Develop strategies to incentivize customers to opt for standard bookings without additional requests, such as offering discounts or complementary services. This approach can contribute to a more stable revenue stream and reduce the risk of cancellations.

- Leverage Online Market Segments: Focus on bookings from online market segments (<= 0.50). These segments have shown a lower propensity for cancellation compared to offline segments. Invest in targeted marketing efforts and incentives to attract more customers from online channels, thus reducing cancellation costs.

- Monitor Average Price per Room: Bookings with higher average prices (> 196.50) tend to have a higher cancellation rate. Review the pricing strategy and consider implementing dynamic pricing algorithms to optimize room rates and reduce cancellations, particularly for higher-priced bookings.

- Enhance Booking Experience for Longer Stays: Focus on improving the booking experience for guests planning longer stays. Bookings with more weekend nights (> 0.50) and longer lead times (> 68.50) have shown a higher likelihood of cancellations. Provide personalized services, loyalty benefits, and exclusive offers to guests with longer stays to increase their commitment and minimize cancellations.

THANK YOU

Arpan Dinesh