INN Hotels Project Business Presentation

Prepared by Dheeraj Mishra

As of January 15, 2022

Contents

- Business Problem Overview
- Data Overview
- Exploratory Data Analysis (EDA)
- Model Evaluation Criterion
  - What's important?
  - How to reduce losses?
- Model Performance Summary
  - Logistic Regression
  - Decision Tree
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- INN Hotels Group is facing problems with the high number of booking cancellations and would like data-driven solutions to predict which booking is going to be canceled in advance.

- A significant number of hotel bookings are called off due to cancellations or no-shows. Flexible cancellation options are beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with.

- The cancellation of bookings impact a hotel on various fronts:
  - Loss of resources (revenue) when the hotel cannot resell the room.
  - Additional costs by increasing commissions or advertising to help sell these rooms.
  - Lowering prices last minute, resulting in reducing the profit margin.
  - Human resources to make arrangements for the guests.

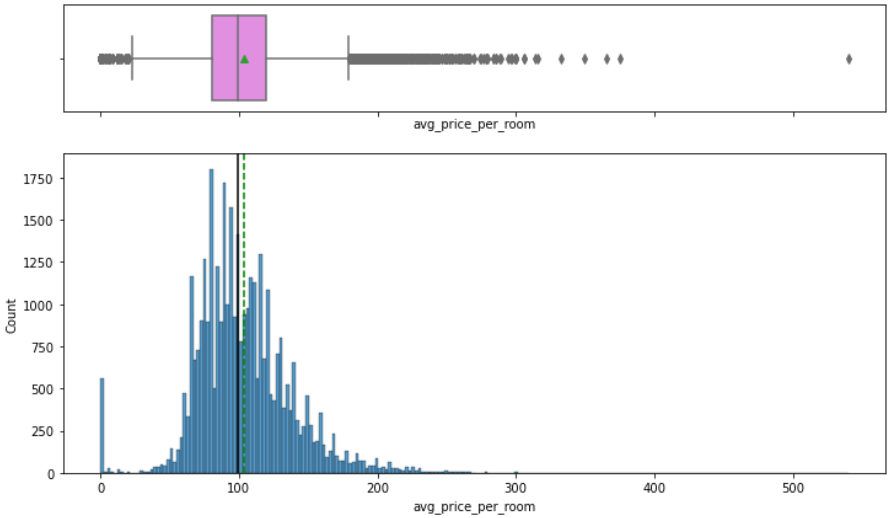
- The task at hand is to analyze the data provided and find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds..

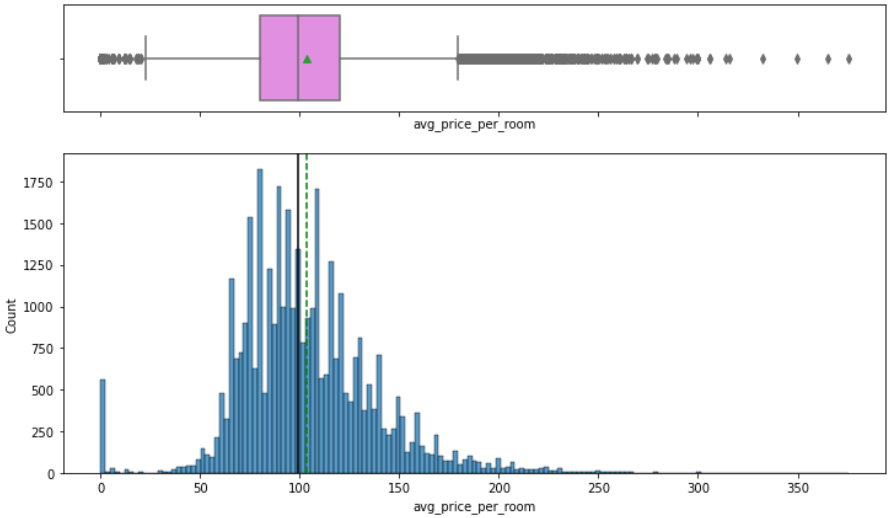
Data Overview

- The data contains information about 36,275 bookings and their characteristics.
- The characteristics include lead time, arrival date-month,year, average room price, meal plans, special requests, weekend vs week nights stay, repeat guests, # of adults and kids, and more.
- We will build predictive data model to predict which booking is going to be canceled in advance.
- No Missing values
- “0” values
  - Column avg_price_per_room has “0” values. We will be imputing it by replacing the “0” with column median value by market_segment_type
- Outliers
  - There are outliers in avg_per_price_per_room column. We will assign the outliers the value of upper whisker where price is greater than equal to 500
  - Outliers are present in the no_of_children columnn. We will fix this by replacing 9, and 10 children with 3
- Categorical features such as type_of_meal_plan, room_type_reserved, and market_segment_type will be encoded as integer using the dummy one-hot value

Exploratory Data Analysis

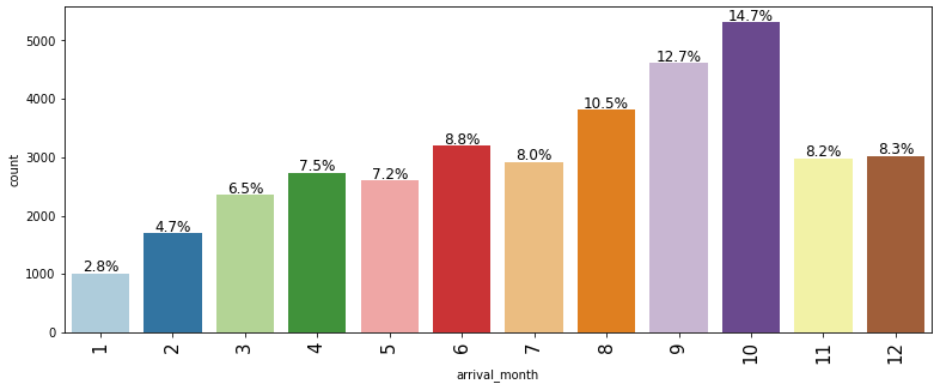
- The avg_price_per_room is heavily skewed.
- Data preparation will be applied to both to reduce the extreme skewness.

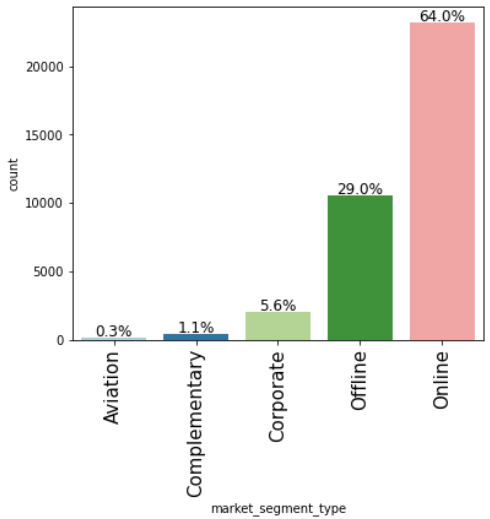


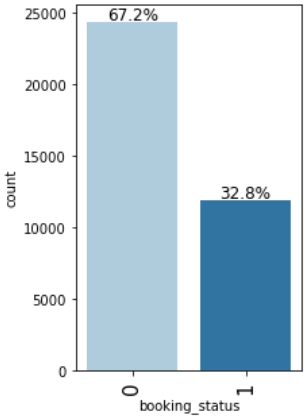


Exploratory Data Analysis

- October is has the most number of reservations in the data, followed by September and August.
- Online market segment dominates reservation with 64% of the market.
- More than ~32% of bookings get cancelled

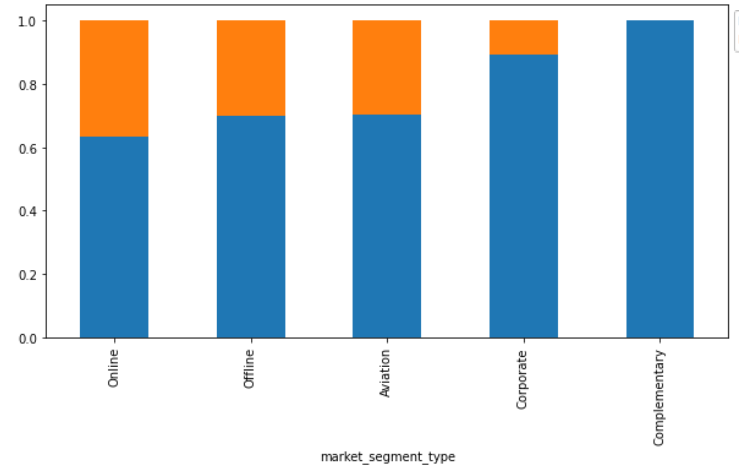


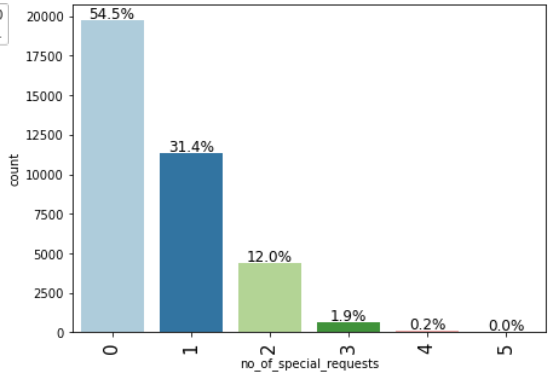


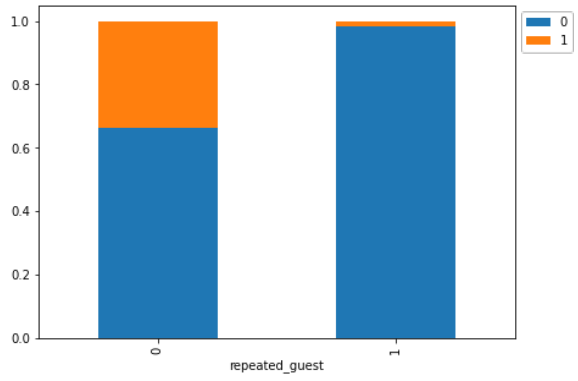


Exploratory Data Analysis

- Booking done through Online market-segment have highest proportion of cancellation
- 31% bookings have at least one special request
- Repeat guests are loyal to brand and show significant less cancellation

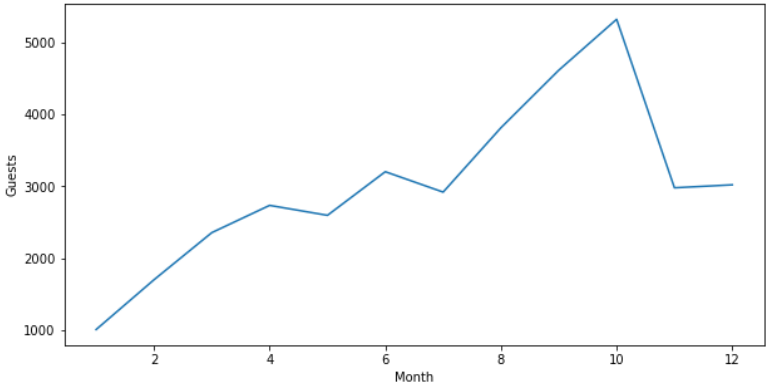


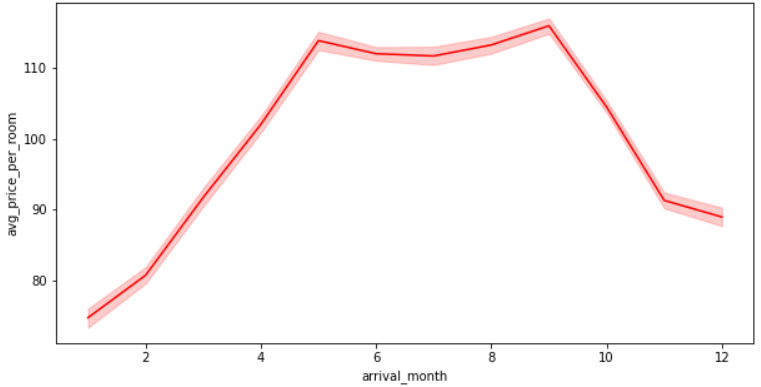


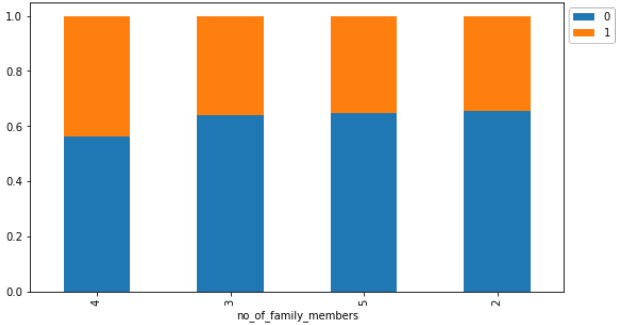


Exploratory Data Analysis

- Bookings are lowest in January, then gradually rise until October and decline sharply afterwards
- Average Room Price is during the peak months (May to September) is hovers slightly above 110
- Cancellation increases when group size (no of people in the booking) increases

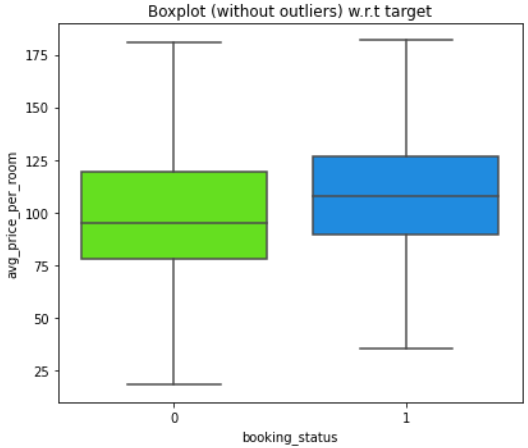


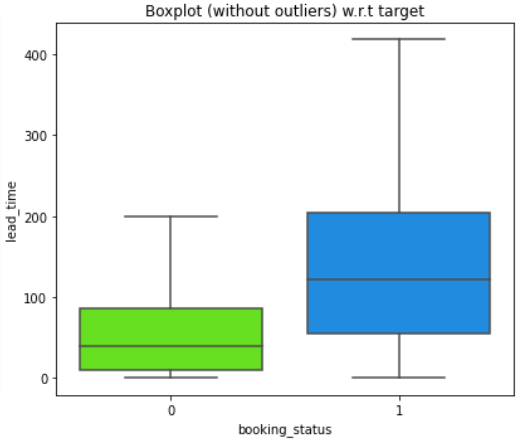


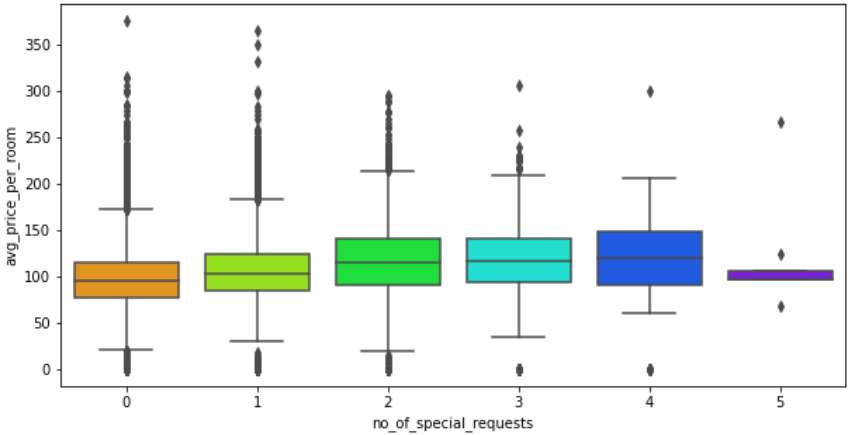


Exploratory Data Analysis

- Pricier the room and/or greater the lead time, more chance of the cancellation.
- Bookings that were cancelled have an average room price over 100 and avg lead time of 120 days.
- Bookings that were not cancelled have an average room price under 100 and avg lead time of 40 days.
- Average cost of the room goes up as number of special requests increase



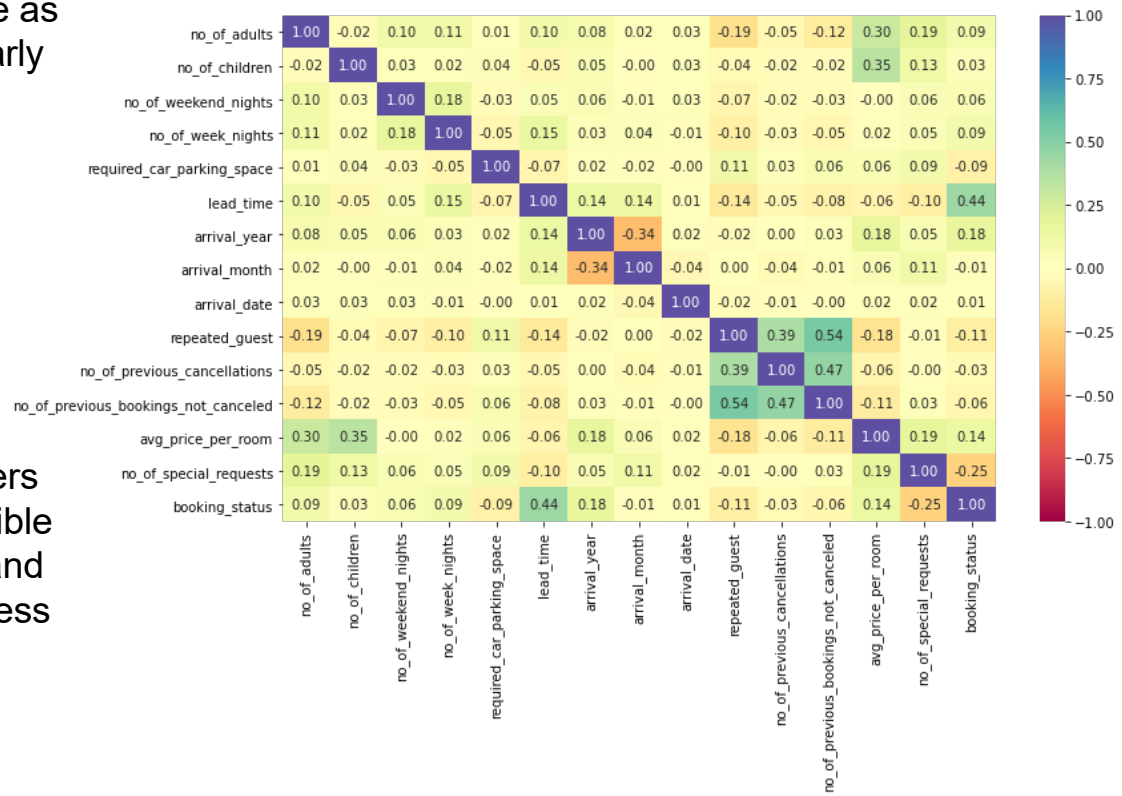




Exploratory Data Analysis

- The lead time is highly correlated with the booking status. This makes sense as the people tend to lock the price by early booking while their rest of travel plans and other arrangement may not have finalized, leading to last minute cancellation

- The number of special requests is negatively correlated with booking status. This makes sense as customers with more special needs are less flexible with their options to find other deals and hence to stick with original booking (less cancellation)



Model Evaluation Criterion

- **Model can make wrong predictions as:**
  - **False Negative**: Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
  - **False Positive**: Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.
- **Which case is more important?**
  - Both the cases are important as:
  - If we predict 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.
  - If we predict 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. This might damage the brand equity.
- **How to reduce the losses?**
  - Hotel would want `F1 Score` to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.
  - F1 score is defined as the harmonic mean between precision and recall

Model Performance Summary

- We want to predict which booking is going to be canceled in advance, based on the characteristics provided to us.
- We will build two models using two different algorithm approaches:
  1. Logistic Regression
  2. Decision Tree
- We have used available data to for building a robust Linear Regression model and Decision Tree, both using the train data and check the performance on test data to understand the predictive power of our models.
- For Logistic Regression model, we will remove multicollinearity from the data to get reliable coefficients and p-values.
- For Decision Tree, we will perform Pre and Post Pruning and find best tree with highest F1 score
- The most significant predictors of the booking cancellation are:
  - Lead Time
  - Avg Price Per Room
  - Number of Special Request
  - Market Segment - Online
  - Repeat Customer

Model Performance Summary – Logistic Regression

Print summary of the model

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Sat, 15 Jan 2022	Pseudo R-squ.:	0.3282
Time:	21:25:25	Log-Likelihood:	-10811.
converged:	True	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-916.4562	120.466	-7.608	0.000	-1152.565	-680.347
no_of_adults	0.1090	0.037	2.921	0.003	0.036	0.182
no_of_children	0.1533	0.062	2.473	0.013	0.032	0.275
no_of_weekend_nights	0.1085	0.020	5.493	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.395	0.001	0.018	0.066
required_car_parking_space	-1.5945	0.138	-11.563	0.000	-1.865	-1.324
lead_time	0.0157	0.000	59.210	0.000	0.015	0.016
arrival_year	0.4527	0.060	7.583	0.000	0.336	0.570
arrival_month	-0.0425	0.006	-6.589	0.000	-0.055	-0.030
repeated_guest	-2.7370	0.557	-4.917	0.000	-3.828	-1.646
no_of_previous_cancellations	0.2288	0.077	2.982	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.306	0.000	0.018	0.021
no_of_special_requests	-1.4697	0.030	-48.883	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1644	0.067	2.472	0.013	0.034	0.295
type_of_meal_plan_Not Selected	0.2857	0.053	5.401	0.000	0.182	0.389
room_type_reserved_Room_Type 2	-0.3553	0.131	-2.711	0.007	-0.612	-0.098
room_type_reserved_Room_Type 4	-0.2826	0.053	-5.325	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7359	0.208	-3.533	0.000	-1.144	-0.328
room_type_reserved_Room_Type 6	-0.9672	0.151	-6.397	0.000	-1.264	-0.671
room_type_reserved_Room_Type 7	-1.4329	0.293	-4.888	0.000	-2.008	-0.858
market_segment_type_Corporate	-0.7915	0.103	-7.694	0.000	-0.993	-0.590
market_segment_type_Offline	-1.7851	0.052	-34.358	0.000	-1.887	-1.683

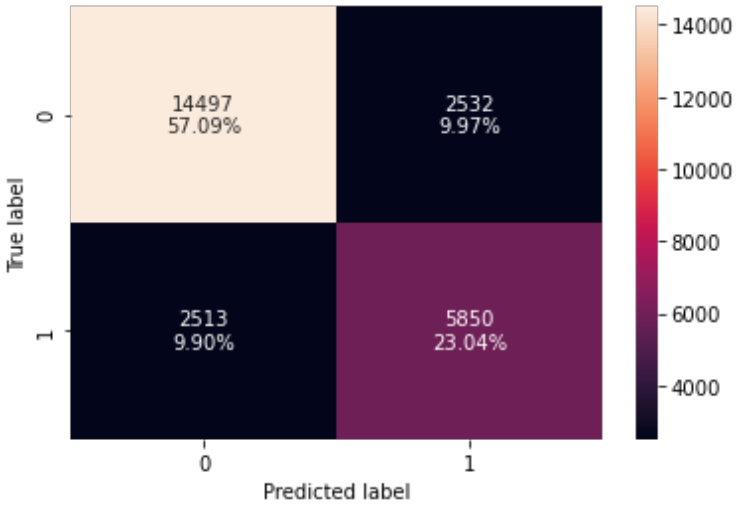
Coefficients and % change in odds of the model

	const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year
Odds	0.00000	1.11518	1.16569	1.11458	1.04254	0.20301	1.01583	1.57259
Change_odd%	-100.00000	11.51781	16.56850	11.45776	4.25381	-79.69879	1.58299	57.25888

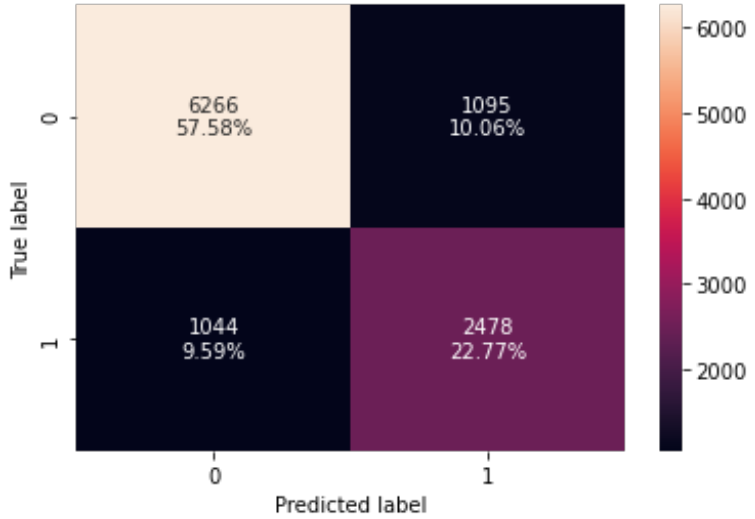
Model Performance Summary – Logistic Regression

Comparing performance using Confusion Matrix

Training-Data Performance

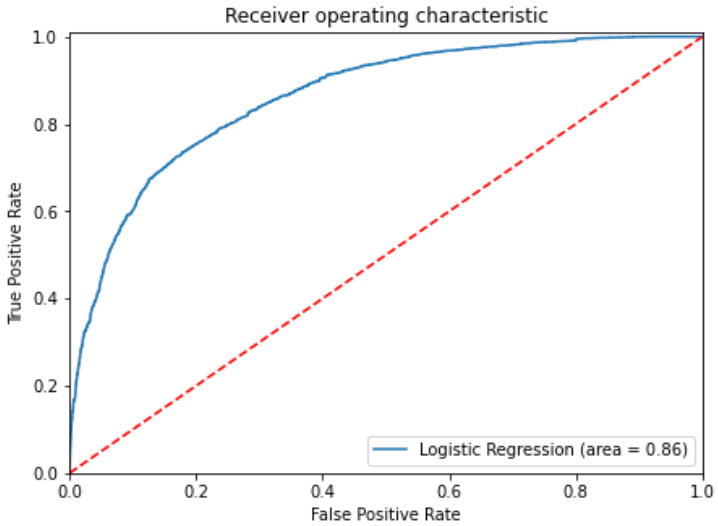


Testing-Data Performance



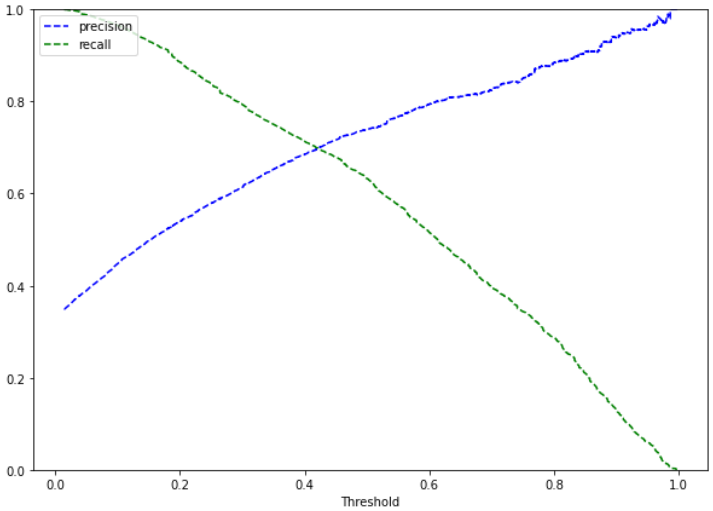
Model Performance Summary – Logistic Regression

Using
AUC-ROC curve
Optimal Threshold = 0.37



- The optimal cut off is at 0.37 where tpr is high and fpr is low.

Using
Precision-Recall curve
Optimal Threshold = 0.42



- At the threshold of 0.42, we get balanced recall and precision.

Model Performance Summary – Logistic Regression

Training-Data Performance

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.8055	0.7929	0.8013
Recall	0.6327	0.7355	0.6995
Precision	0.7391	0.6687	0.6979
F1	0.6817	0.7005	0.6987

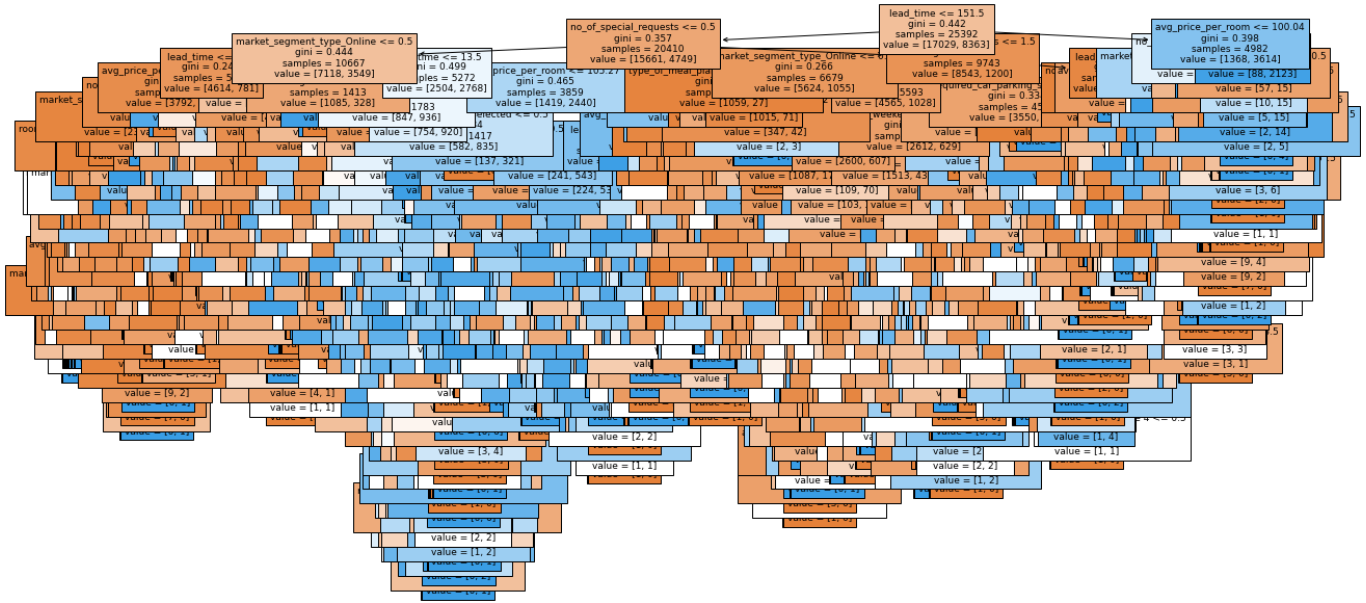
Testing-Data Performance

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.8047	0.7961	0.8035
Recall	0.6309	0.7394	0.7036
Precision	0.7290	0.6668	0.6935
F1	0.6764	0.7012	0.6985

- We have been able to build a predictive model that can be used by the INN-Hotels Group to find the Bookings that could get cancelled with an f1_score of ~0.70 on the training set and formulate policies accordingly.
- All the logistic regression models have given a generalized performance on the training and test set.
- Coefficient of some levels of no_of_previous_cancellations, lead_time, avg_price_per_room, type_of_meal_plan_Not Selected, no_of_children, #of_adults, # of_weekend_nights are positive an increase in these will lead to increase in chances of a Booking cancellation.
- Coefficient of market_segment_type_Offline, required_car_parking_space, no_of_special_requests, and room_type_reserved_Room_Type 7 are negative increase in these will lead to decrease in chances of a Booking cancellation.
- We recommend **best logistic regression** model to be the one with maximum F1 score, which is **Logistic Regression with 0.42 Threshold** value. This model also gives generalized performance on both train and test data.

Model Performance Summary – Decision Tree

Model without any Pruning – is very complex with significant depth and nodes



Tree is overfitting with high Accuracy, Recall, Precision, and F1 Score

Training-Data Performance

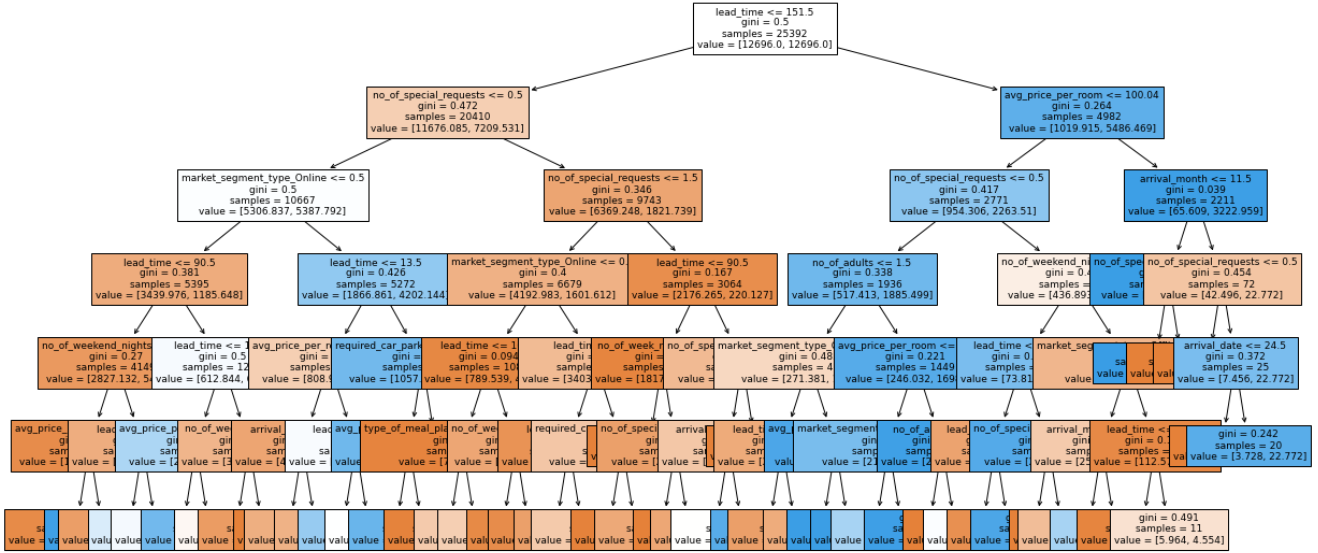
Accuracy	Recall	Precision	F1
0.99421	0.98661	0.99578	0.99117

Testing-Data Performance

Accuracy	Recall	Precision	F1
0.87136	0.8092	0.79653	0.80282

Model Performance Summary – Decision Tree

Model after Pre Pruning – is less complex with reduced depth and # of nodes



Tree is less overfitting however there is room for improvement for better predictions

Training-Data Performance

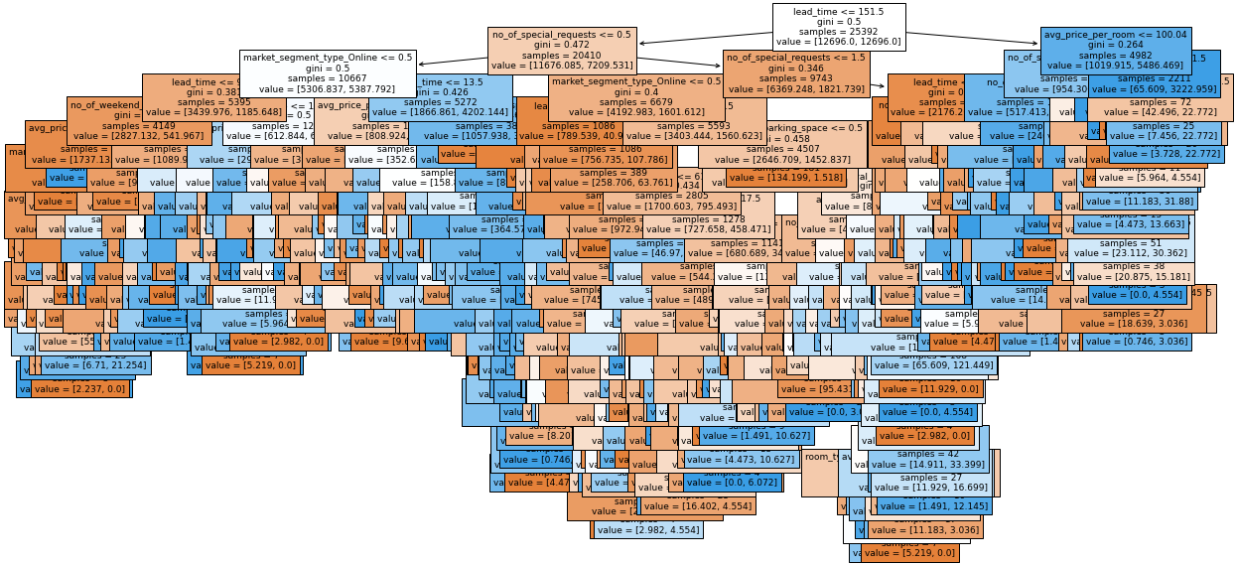
Accuracy	Recall	Precision	F1
0.83097	0.78608	0.72425	0.7539

Testing-Data Performance

Accuracy	Recall	Precision	F1
0.83497	0.78336	0.72758	0.75444

Model Performance Summary – Decision Tree

Model after Post Pruning
little complex compare to Pre-Pruning model but delivers better performance!



Tree maximizes the F1 score and doesn't suffer overfitting!

Training-Data Performance

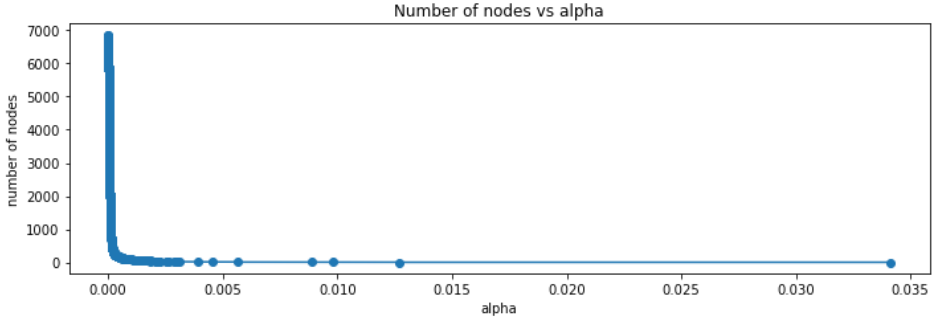
Accuracy	Recall	Precision	F1
0.89989	0.90303	0.81353	0.85594

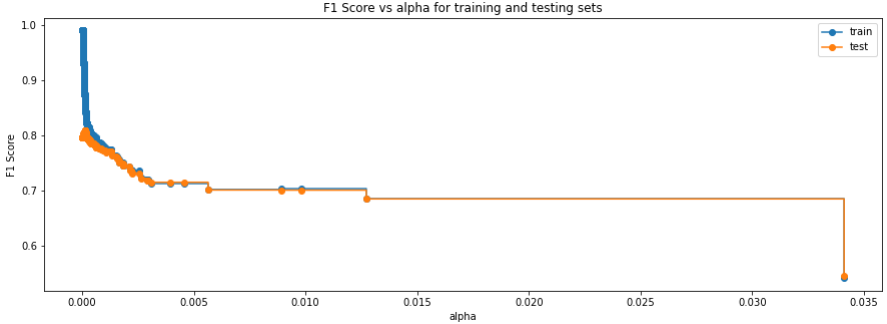
Testing-Data Performance

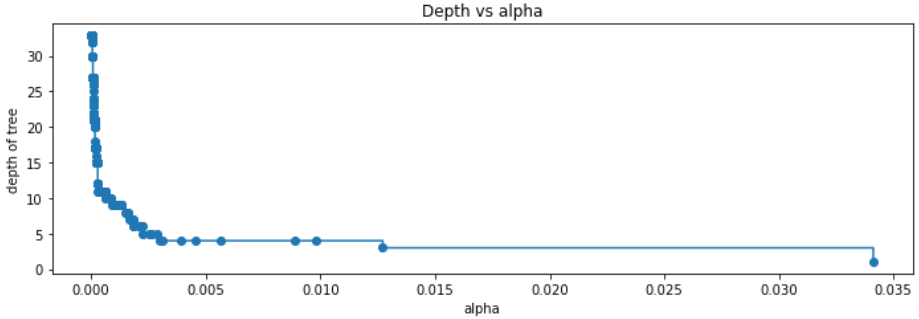
Accuracy	Recall	Precision	F1
0.86888	0.85576	0.76634	0.80858

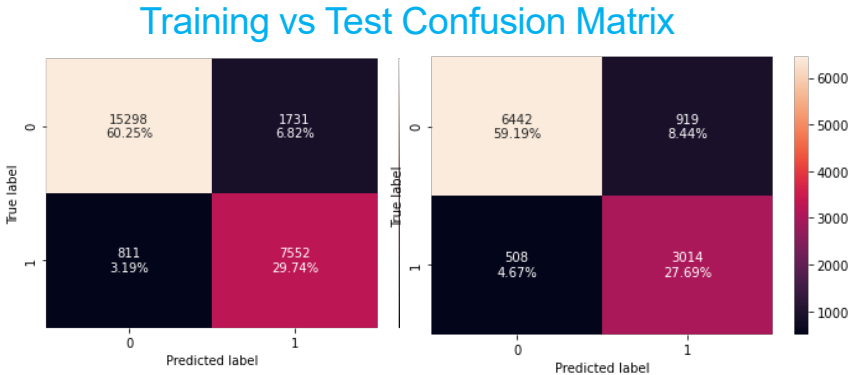
Model Performance Summary – Decision Tree

Using Cost Complexity method for Post Pruning
F1 Score vs alpha for training and testing sets







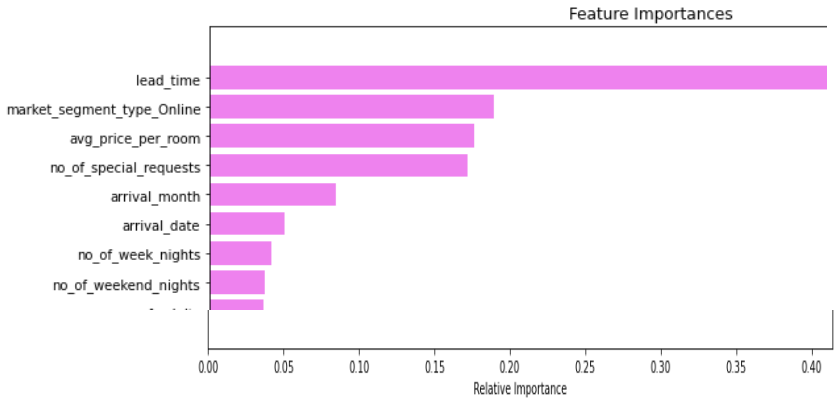


Model Performance Summary – Decision Tree

Comparison of Decision Trees: No-pruning | Pre-Pruning | Post Pruning

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89989
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81353
F1	0.99117	0.7539	0.85594

Important Features of Tree



Post-Pruning Tree delivers the maximum F1 score on train and test and doesn't suffer overfitting!

Training-Data Performance

Accuracy	Recall	Precision	F1
0.89989	0.90303	0.81353	0.85594

Testing-Data Performance

Accuracy	Recall	Precision	F1
0.86888	0.85576	0.76634	0.80858

Business Insights

- August, September and October months are the busiest time for hotel chain, with October taking the biggest share of ~15%
- January is slowest period and booking gradually increases as the year progresses
- Online market segment dominates all booking methods, with 64% share
- ~45% bookings have at least one special request
- ~33% of bookings get cancelled
- Lead Time clearly has strong correlation with booking status
- Number of special requests have a high negative correlation with booking status
- Hotel rates are dynamic and change according to demand and customer demographics.
- Average price per room is highest for the Online market segment
- Corporate segment pays less price per room compare the all other segments, probably due to established discount contracts
- Booking status varies across different market segments.
- 45% guests have at least one special requirements when booking a hotel room
- Booking with special requests have less likelihood of getting cancelled compared to bookings with no special requests
- Bookings with more number of special requests pay higher room price
- positive correlation between booking status and average price per room
- Greater the lead time, bigger the chance of cancellation
- Bookings that were cancelled have an average lead time of 120 days, around 3 months lead
- Bookings that were not-cancelled have an average lead time of 40 days, slightly higher than a month lead time.

Business Recommendations

- Lead time, market_segment_type_Online, avg_price_per_room, and no_of_special_requests (in that order) are the most important variables in determining if the booking will get cancelled
- The lead time has highest impact on booking cancellation. Following up with phone call, text or email with customer during last two weeks prior to arrival date to check if their plan is still in place could give early indication if booking is going to be cancelled.
- For Online market segment, customers may find last minute deal elsewhere and potentially cancel their reservation with INN Hotel. Proactively reaching customers with high-chance of booking-cancellation and by offering price-match could prevent the booking cancellation.
- When processing cancellation, attract customer for hotel-credit along with discount for future stay at the hotel
- There is drop in the hotel reservation during the months of November and December, although these are significant holiday months. Business should look into attracting customers using promotional campaign to offer discounts to past customers to drive the sale slower months.
- Develop an online portal to offer the last minute deals (resulting from cancellations) directly to customers, without paying to third-party distribution channels