Project 4 - INNHotels

BY: LUKE HARRIS

Contents

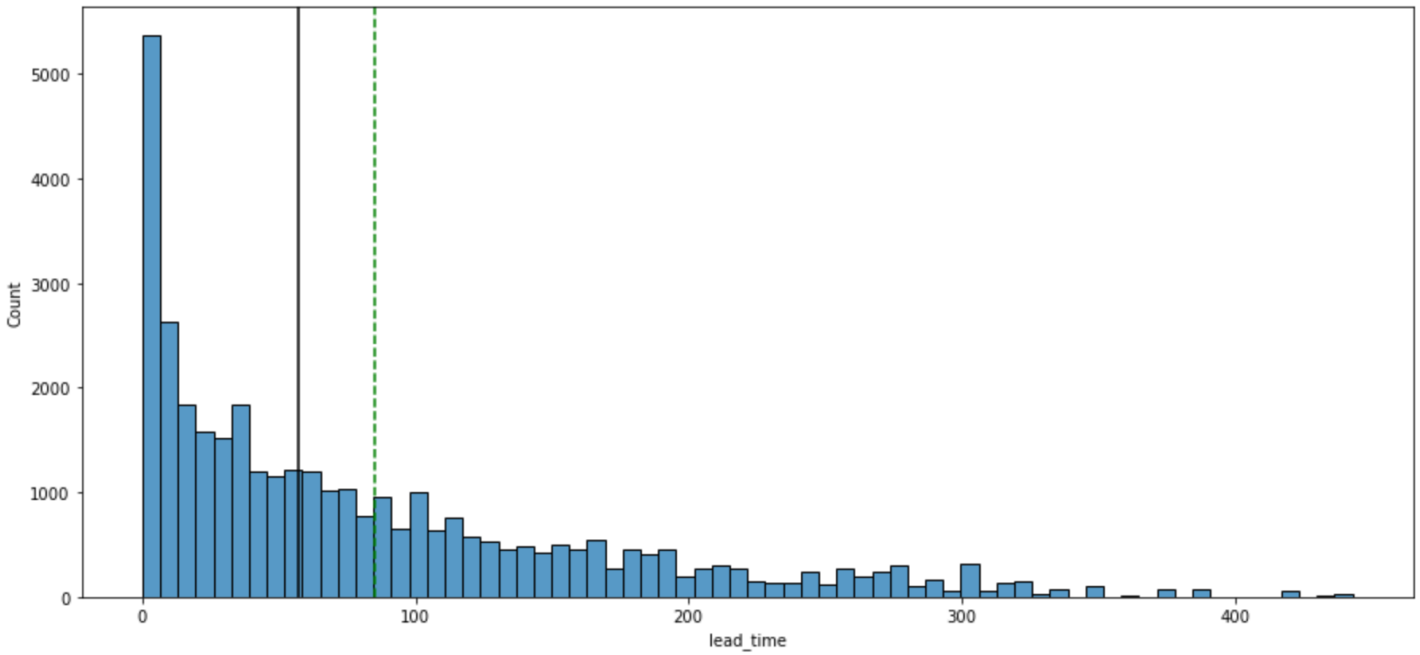
Executive Summary

INN Hotels Group has been having issues with a high number of booking cancellations and has asked us to create a predictive model to predict whether a customer will cancel their booking in advance. Cancellations cause a loss of revenue for the company when the room cannot be resold, sometimes the room price must be lowered last minute and this cuts into the profit margin, also, this requires HR to decide for the guests. This model will help inform INN Hotels to decide a profitable policy for cancelations and bookings.

Exploratory Data Analysis

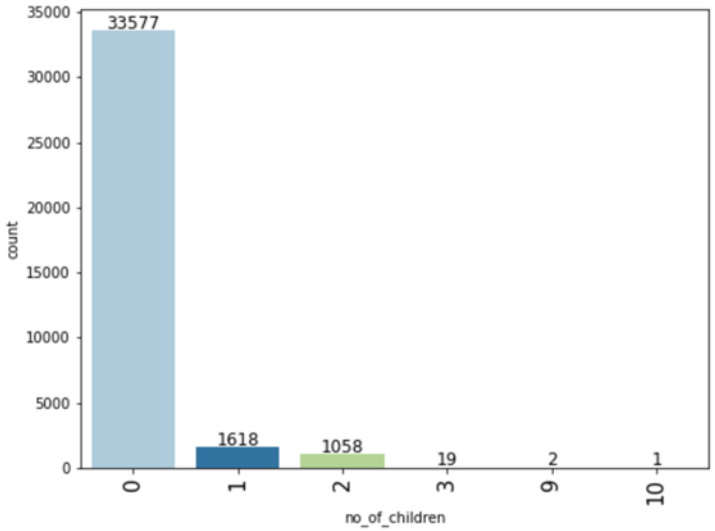
In the next few slides, I will show the analysis of the data we see before creating the model. This will include bivariate analysis as well as single variable analysis. We use these data to create a big picture about what is happening before creating a model to help us understand what is going on within the model.

Lead Time

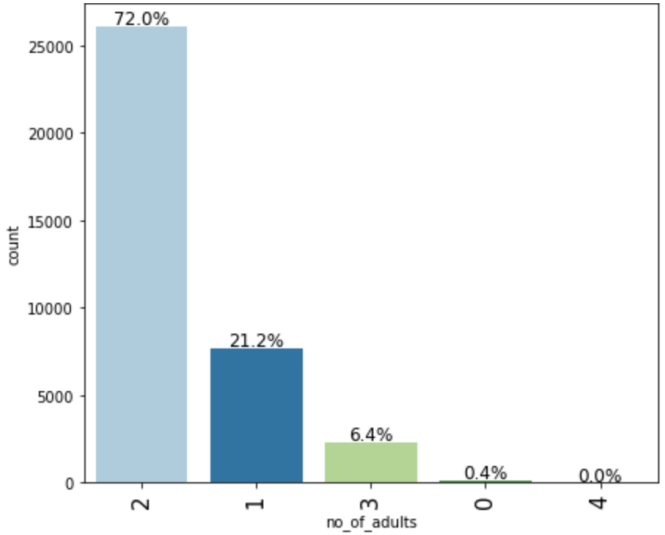


We can see here that lead time is highly right skewed indicating that most people book hotels with little time in advance. The mean lead time someone books a reservation is 85 days, with many people booking the week or even the day of.

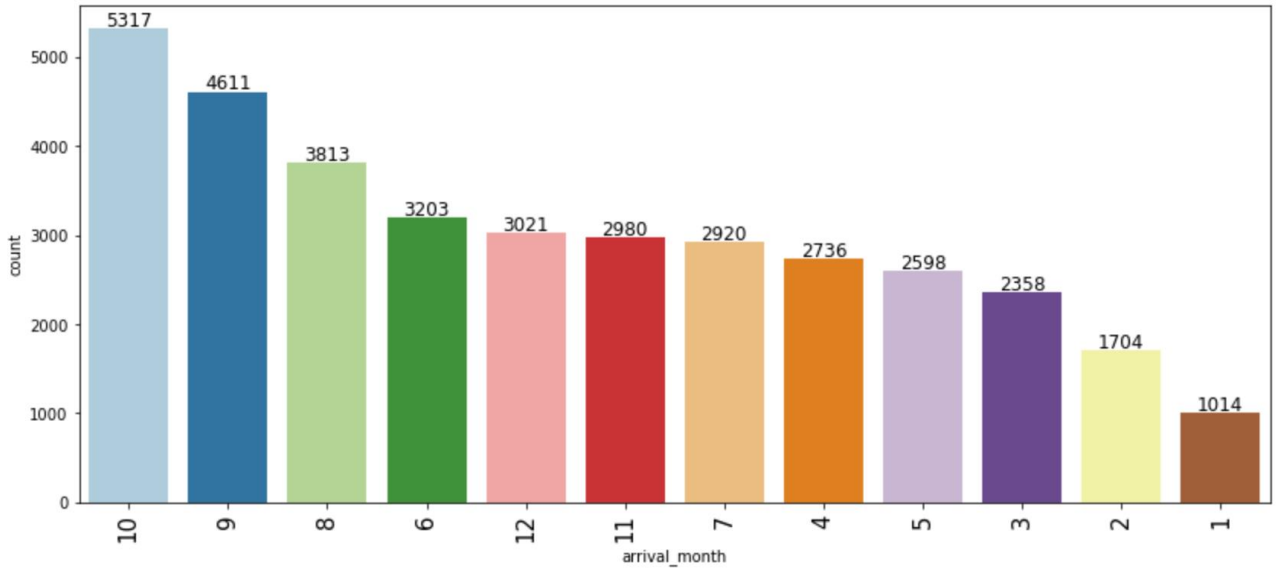
Number of Children and Adults



Here, we can see the number of children and adults there are per room. 72% of the time, it is only adults in the room. They are also the ones who always pay, so accommodating the adults is far more important than giving accommodations for children.

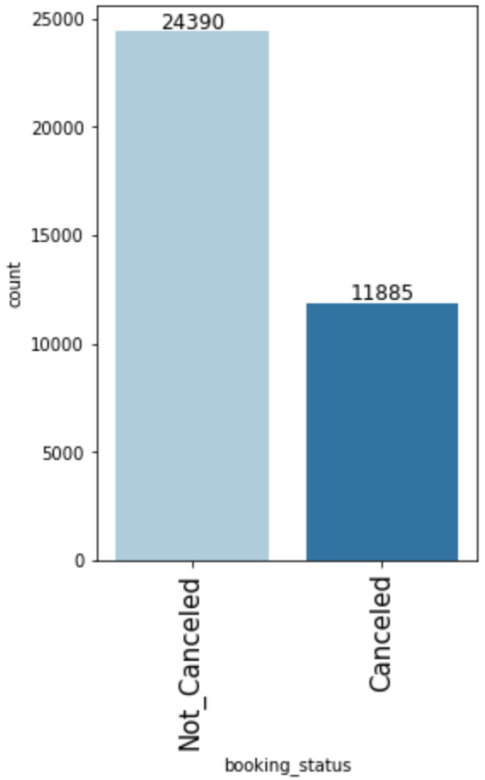


Arrival Month



We see here that most hotel rooms are booked in the summer and fall months. Many people take vacations, especially to areas like Portugal during these months. This could inform us when better cancellation deals can be given. It is more likely to get someone last minute during September or October than it would be in January.

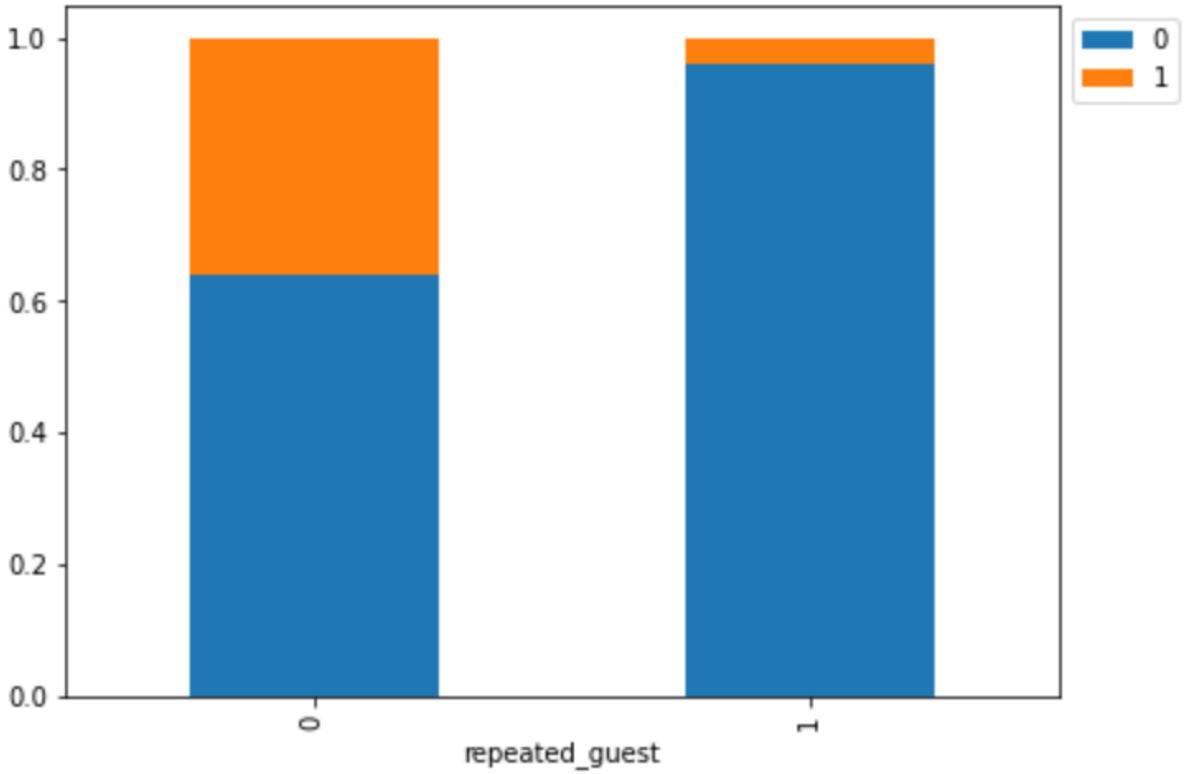
Not Cancelled vs. Cancelled



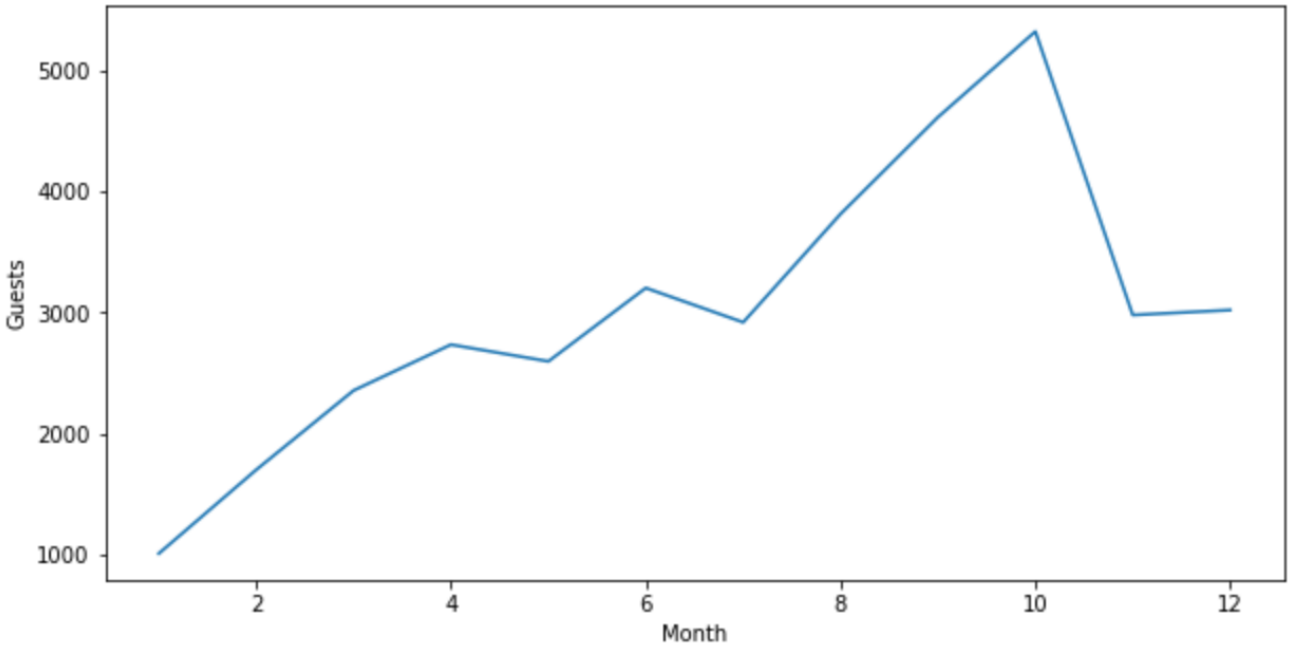
And here we see the number of rooms that were cancelled and not cancelled. As bad as cancellations are, there are still ore than double the number of not cancelled rooms so that's a good sign. Now, we must work to get that number as low as possible.

Repeated Guests vs. Booking Status

On the left we see the guests who have never been with us before, and on the right, we see repeated guests. The orange 1's indicate cancellation and the blue 0's indicate non cancellations. As we see here, there is a much less likely chance that someone who repeatedly uses us will cancel. Therefore, it is appropriate to give more friendly cancellation policies towards guests who repeatedly use our service.



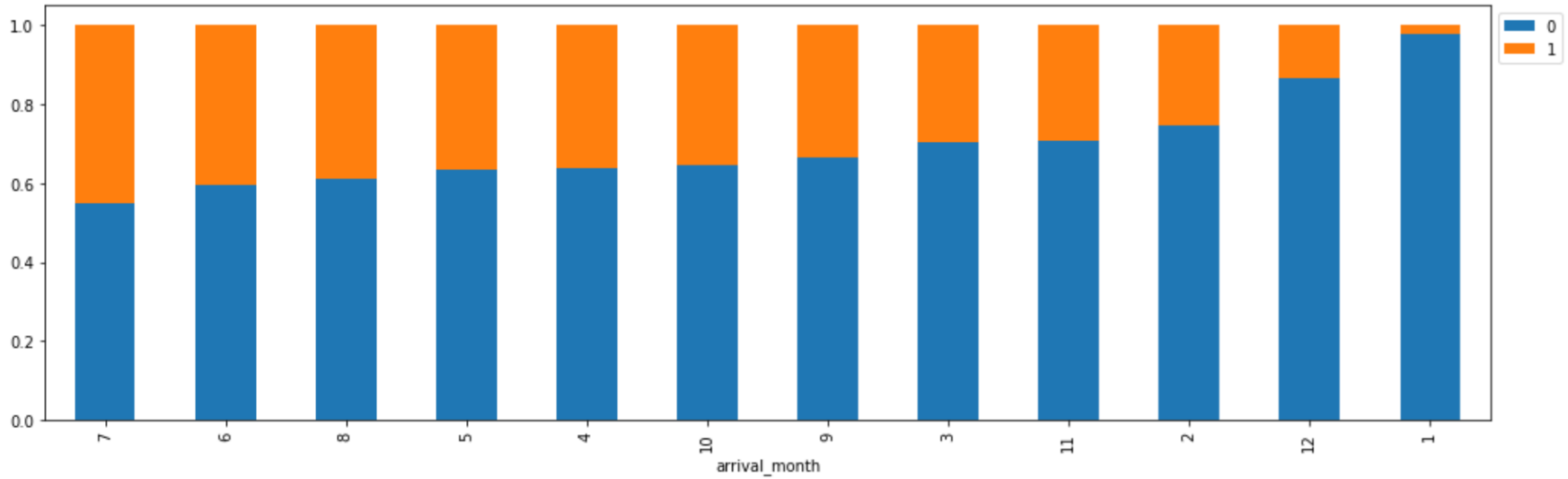
Guests Per Month



For this graph, I cleaned the data to only show non-cancelled guests. We see a large spike during the summer months as indicated in previous figures. With demand this high, we could maintain the highest profit margin by not lowering the prices so much for last minute cancellations. We could also give the most lenient cancellation policies. This could be a win-win for both customers and INN Hotels.

Arrival Month vs. Cancellation

As I suspected, the most cancellations happen during the months in which demand is the highest. I hypothesize that people find better deals online and cancel their previous hotel for a better deal. We could use this to undercut competitors while demand is high and when demand is low people don't cancel as much so we don't need to give as much of an incentive to not cancel.



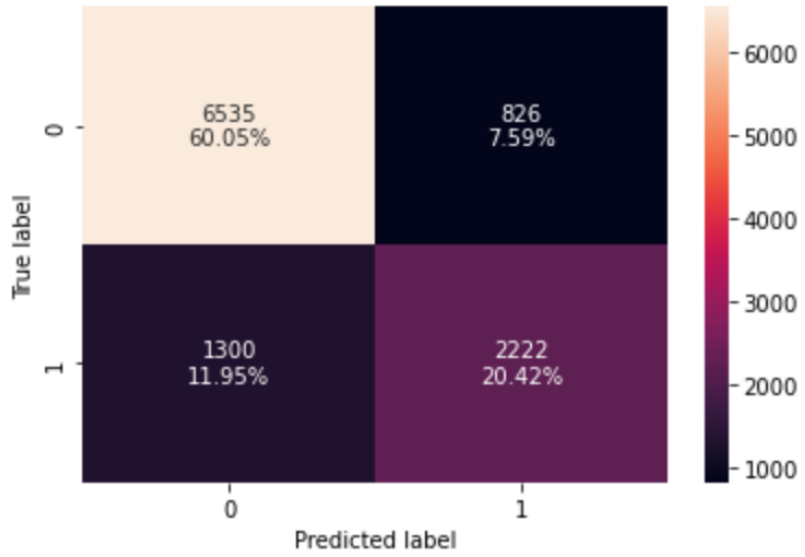
Logistic Regression Model

In the next slide I will discuss the training and testing data and the performance of the model. I am pleased with our model because there are very little type I or type II errors in the performance. For accuracy, recall, and precision each hovers around 80%, 70%, and 70% respectively. Those numbers are high for prediction purposes. However, with pruning methods those numbers should be able to increase greatly. I believe once we have pruned the data, this model will do well for predicting the number of cancellations you will receive in the future.

The Training and Testing Performance In the Linear Regression Model

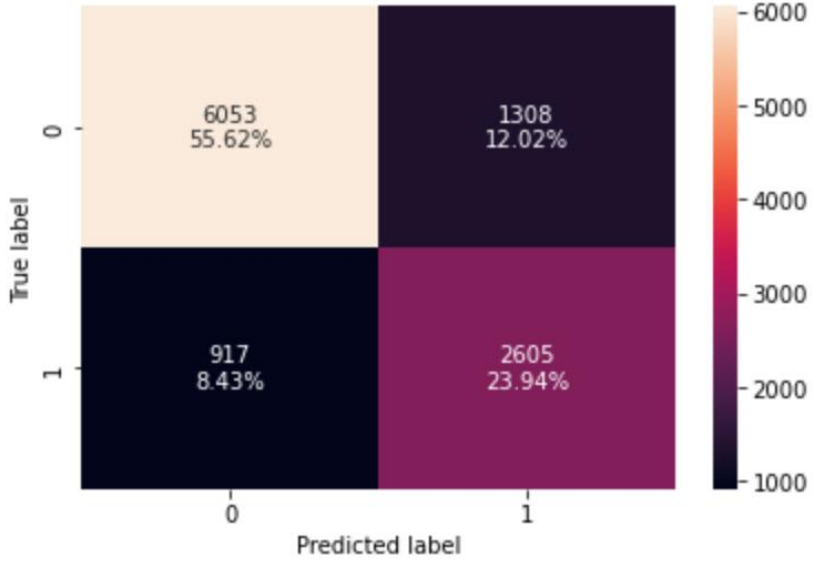
Model on the training performance

	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

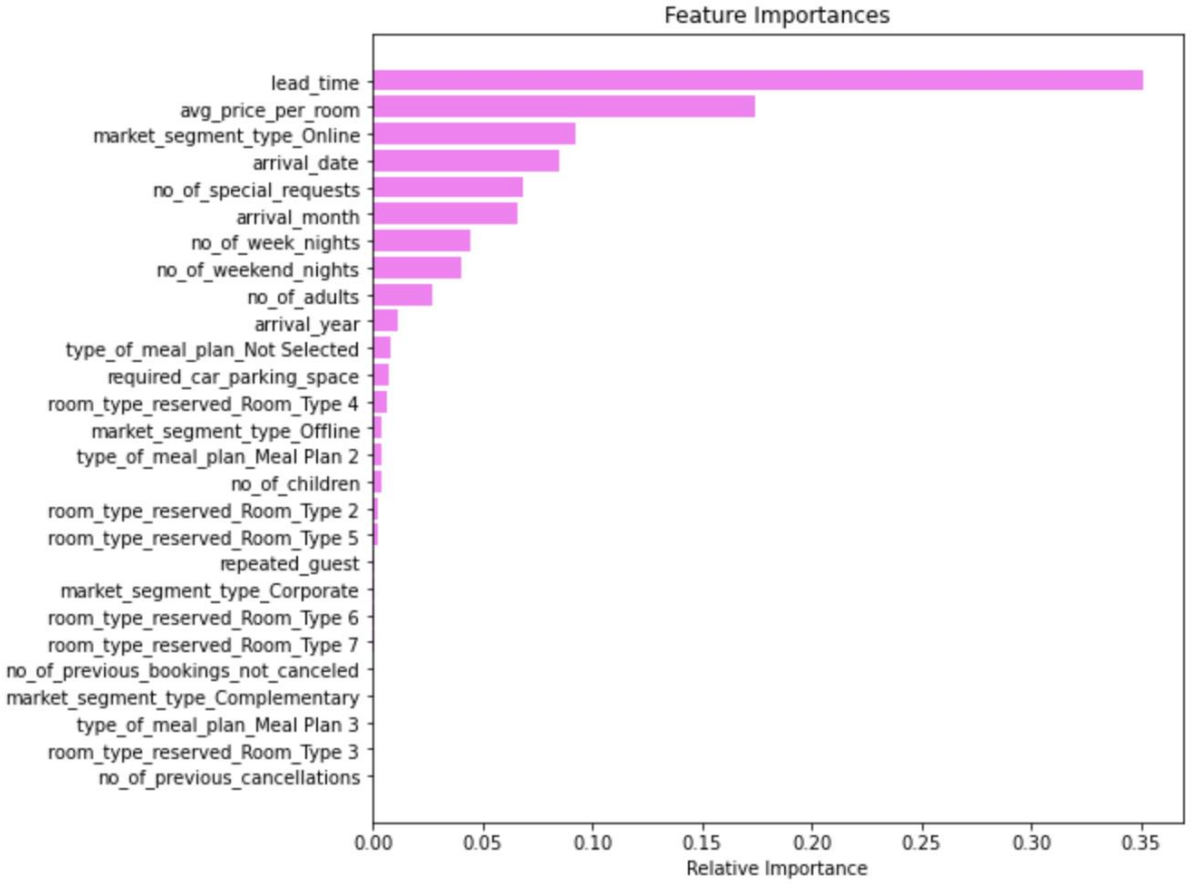


Model on the testing performance

	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852



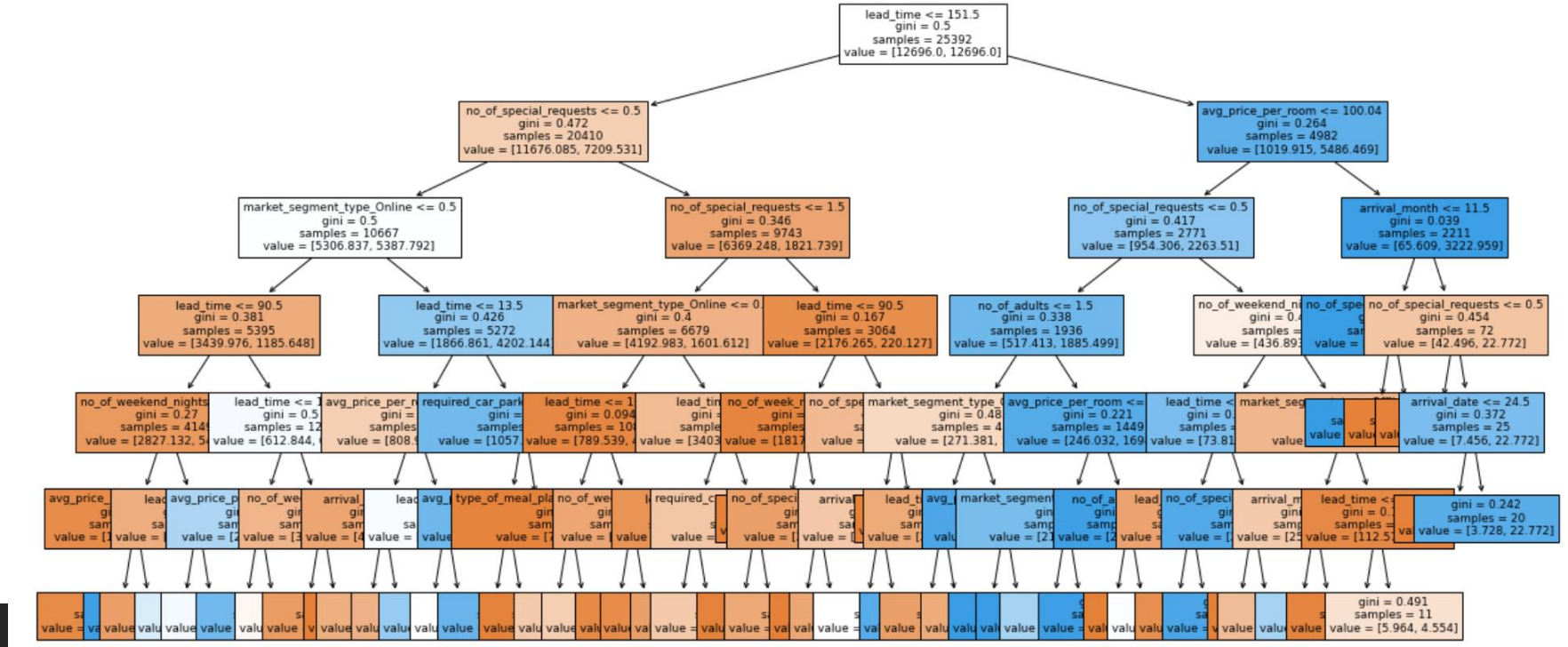
Feature Importance



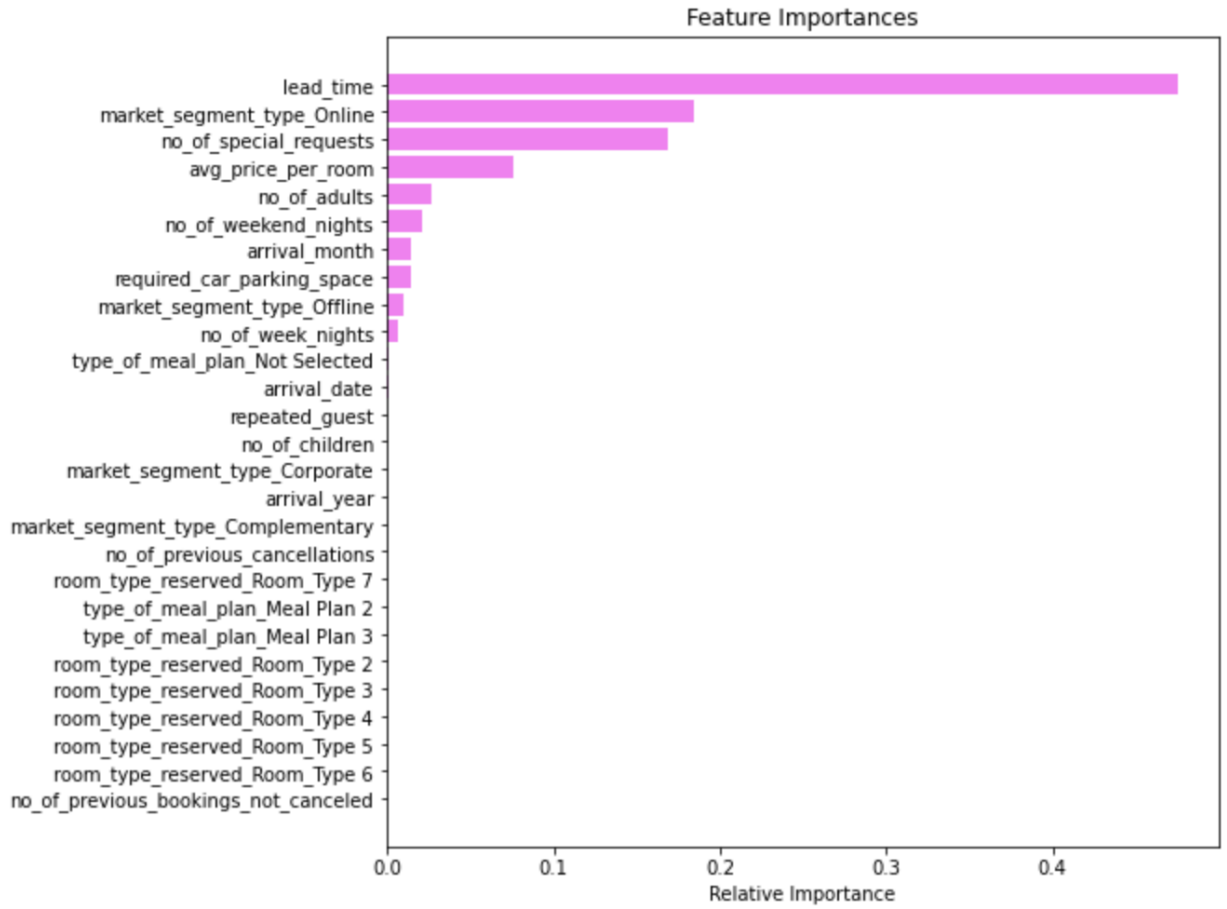
As seen here, the most important variable for our model to determine cancelations is the lead time. Along with other factors, the further out someone chooses to book a room the more of a chance it will be cancelled. The price per room is also important, people who spend more money tend to be more likely to keep their booking. Also, online is very important, these are the people looking for the best deals and will cancel their room if they find a better deal elsewhere.

Decision Tree of the Model

The decision tree was pre pruned to make it more simplistic to read. This also prevents the model from becoming too narrow and not actually giving any real information. We used the ratios given from earlier in the model to determine how much pruning was appropriate.

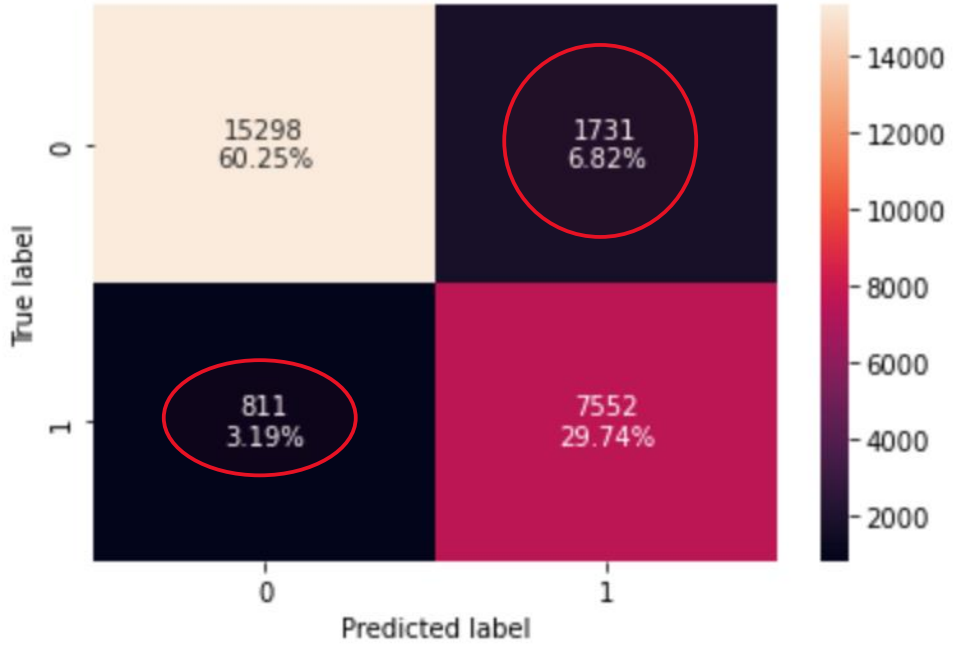


Pre-Pruned Feature Importance



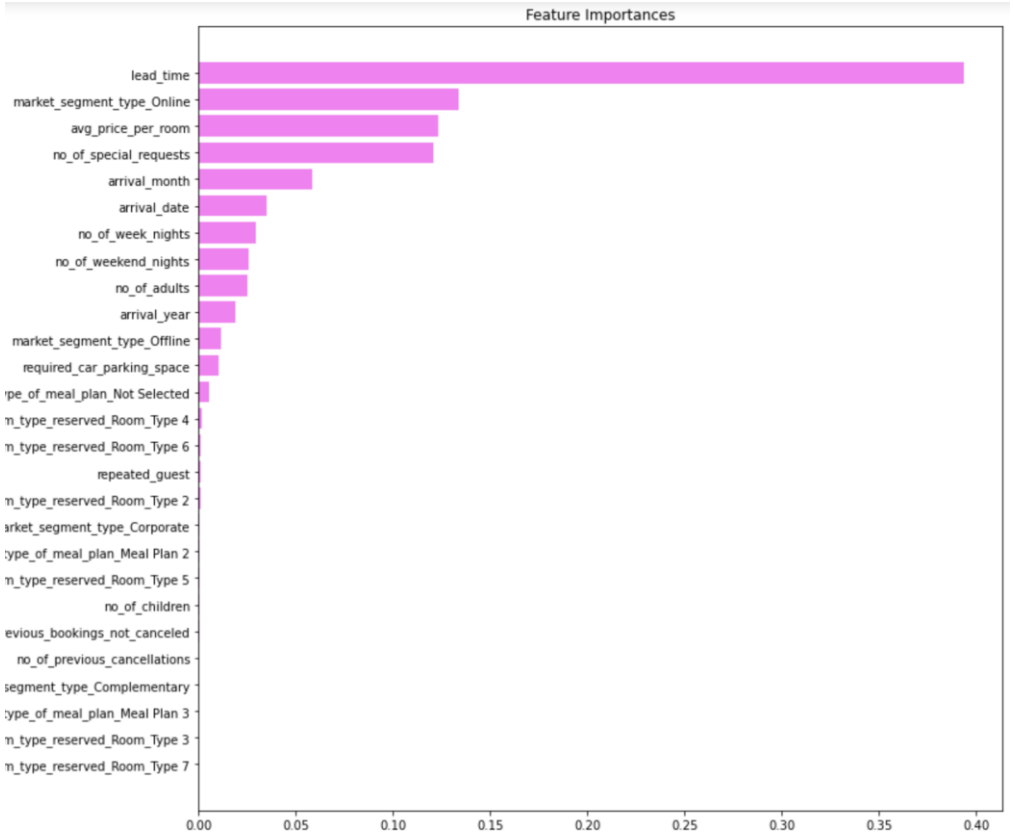
Now that the data is pruned, we see that lead time is still the most important factor for cancellation, but the market segment and special requests has now taken the second and third most important roles in our data.

Post-Pruning



This matrix is after we have post and prepruned the data. As seen here, the type I and type II errors have significantly decreased since the first building of the model. Only 9% of the data are errors in this data.

Post-Pruned Feature Importance



And Finally, we see the feature importance for the post pruning. Lead time and market segment type are still the two most important features. But now they are followed by average price per room in third.

Final Model Performance

	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	0.75390	0.85594

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86888
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76634
F1	0.80309	0.75444	0.80858

Finally we see the how the model performs overall with the pruning and cleaning we have done. The test data is on the left, while the training data is displayed on the right. This model is doing an excellent job indicating the odds of a person cancelling a booking.

Insights and Recommendations

The most important months for hotels are the summer months. Demand is incredibly high which means that if someone were to cancel a booking it is likely that someone could pick up the room by the end of the same day. This means offering lenient cancellation policies during this high demand months would be good for customers and business since there is a good chance you wouldn't need to lower prices too much and cut into the profit margin. The target market should be people who are booking online, with few children, and planning a last-minute vacation. It is highly unlikely someone who books a hotel that same week will end up cancelling. Couples who make decent money, and have booked before, who are looking for a weekend getaway are the largest market for INN Hotels. It is very important to treat repeat customers satisfactory as they are the least likely to cancel a booking. Incorporating a loyalty program and giving deals to these individuals would make them keep coming back for more while also not cancelling trips.