Prepared by Ken Venturi, May 13, 2023

Booking Cancellation Analysis

Project Assessment: Decision Systems

This image produced on MidJourney

Contents / Agenda

Executive Summary

Cancellations are directly correlated with ***lead time as the primary influencer***. As the reservation arrival date appears, less and less people as a percentage of bookings, cancel. The most direct path to addressing this is through the following:

1. **Cancellation Policies, Refunds, Credits and Amenities**. A "no cancellation within a certain date of arrival for cash" policy may need to be implemented. I any case, constricting the policies to favor earlier cancellation and punish last minute cancellations is warranted, already common practice, and so INN Hotel Group policies might need adjusting.

2. **Reward Earlybird Bookings** and offer benefits to further prevent them from cancelling as the arrival date approaches. Being proactive with discretionary assets, consistently reward the behavior or not cancelling with repeat customers and earlybird customers through premium status programs.

Average Price Per Room has a looser but still important correlation with cancellations. One might presume that the more the room costs the more likely a cancellation. The data says is positiviely correlated and so when price goes up, so does the booking cancellation rate.

It is also evident that the **timing of travel** (actual date of the year or season) have a direct impact on cancellation. That would be consistent with holiday (seasonal) travel and also implies most travelers are traveling for vacation and have plans they set up in advance or booked last minute, both of which are not likely to be cancelled. The intermediate travel planner would likely be the largest cancellation category.

Finally**, Length of stay matters**. It appears that the longer the stay, the less cancellation. This might also be an indicator of vacation travelers and requires additional definition of the stays. Weeks of the day would be a strong indicator.

Overall, there is a definitive move forward set of options for reducing cancellations. Additional data, research and analysis will help to execute a better plan once cancellation are predicted, but the predictor developed herein provides a strong directional guidance for which to focus.

Project Outline: Practical Implementation that turns analysis into action

- **Run all booking data through model** both retro actively and going forward to further flesh out model accuracy and precision and prediction value. Going forward, a dynamic data connection between the predictor and the booking system will allow it to offer "Cancellation Probability" set of data to a hotel bookings manager or the front desk and allow management the most accurate data for its predictor.
- **Modify returns policies to favor earlier cancellations**. Statically drive the reduction of cancellation refunds. Offer future stays in leu of cash at some point in the process when cancellation for cash would no longer be the policy.
- **Prepare staffing to mirror the predictions of cancellations to reduce costs at the operating level**. Leverage standby employment to cut costs, adjust hotel HAVAC and systems to cut cost, and allow the local operators to manage the expenses of the building according to projected visitation.
- **Implement a modest *Overbooking Policy***. Statistically gamble on the cancellation rate in a manner that starts with a positive balance of room requests over rooms available. As cancellations come in, move toward the goal of rooms available matching room requests. If in the end the rooms are not sufficient, initiate the referral program at the critical date.
- **Implement a modest *Overbooking Referral Program***. INN Hotel Group is large and often has many "sister" locations in any given market. With this business model, it is better to overbook and refer to other related hotels than it is to overshoot booking predictions and have surprise cancellations. Possibly need to establish a strong referral bonus program to make guests appreciate the transfer to a sister hotel (free night?).
- **Further analysis** is recommended to further validate findings and to address anomalies and interesting events happening in data. There are underlying themes occurring in the phenomena that can not yet be explained in the data. See "Recommended Further Analysis" in the appendix for specifics.

Problem Statement

Cancelled Bookings has become a major limiting factor for the INN Hotel Group as customer policies and travel trends have changed. This costs the company in asset utilization, revenues, operating efficiencies and in turn profits and return on capital. All of which are regularly used financial metrics in public stock valuations and therefor creating even great equity value loss.

Business Objective

Develop better predictive capabilities to allow management to anticipate cancellations, predict trends, identify which bookings are likely to cancel. This ultimately allows INN to adjust pricing and promotions to maximize revenues, trim costs last minute and optimize the company's policies with respect to cancellations and cash refunds. All of these influencers will allow the company to recover significant value in both cash flows and equity value.

Context

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.
2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
4. Human resources to make arrangements for the guests.

Project Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions.

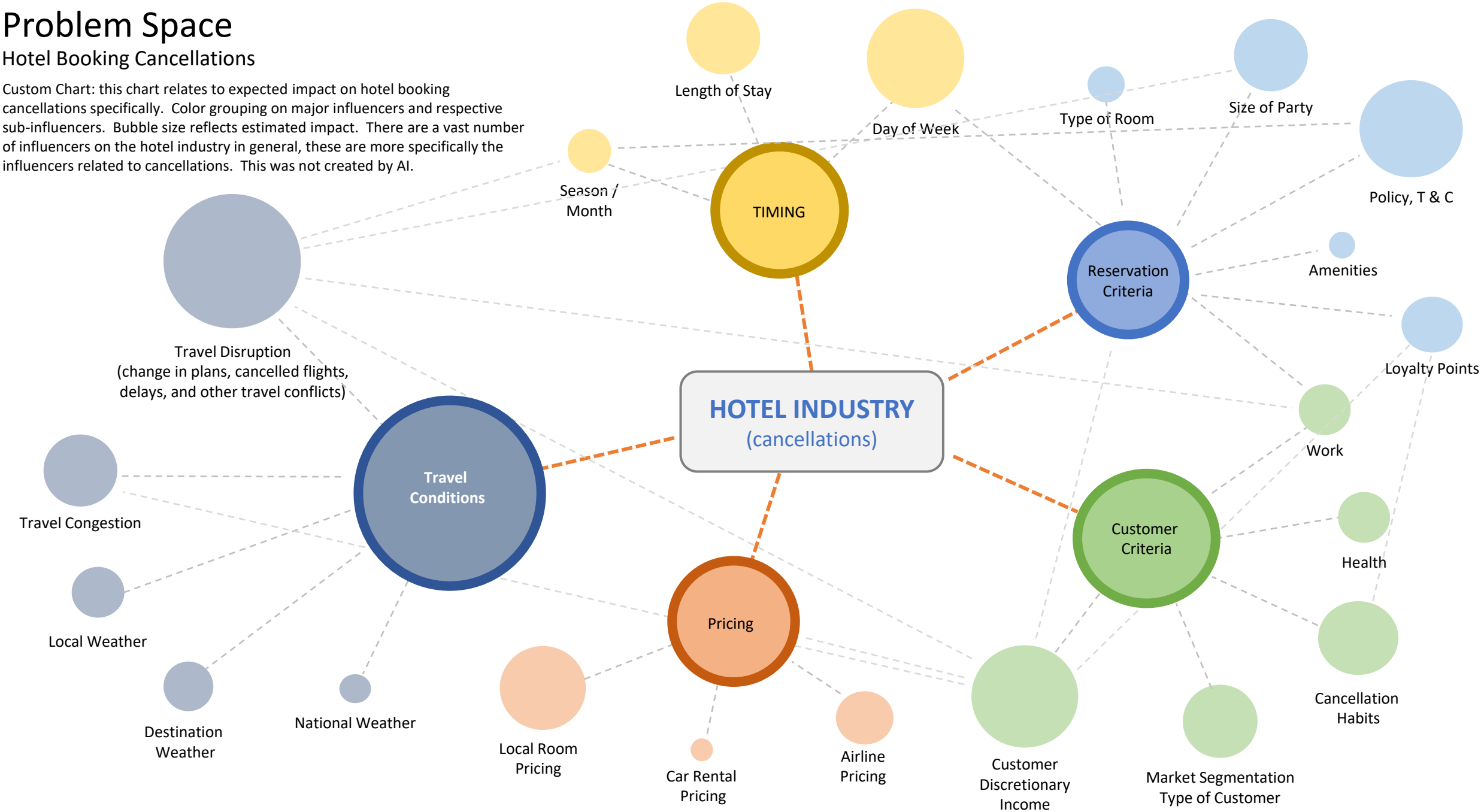
The Project will analyze the data provided to:

1. find which factors have a high influence on booking cancellations,
2. build a predictive model that can predict which booking is going to be canceled in advance, and
3. help in formulating profitable policies for cancellations and refunds.

Problem Space

Hotel Booking Cancellations

Custom Chart: this chart relates to expected impact on hotel booking cancellations specifically. Color grouping on major influencers and respective sub-influencers. Bubble size reflects estimated impact. There are a vast number of influencers on the hotel industry in general, these are more specifically the influencers related to cancellations. This was not created by AI.



Data Overview

The data provided is of various reservations (booking_id) made with the hotel in which not only the detailed room and reservation information was captured but details on the customer including attributes like repeat customer, number previous cancels, number of special requests are also available for analysis.

There are 18 attributes (removing booking_id) in the data set. For a definition of data terms see Appendix.

Overall, the data is in fairly good structure with 9,069 booking records.

Data Attributes

no_of_adults
no_of_children
no_of_weekend_nights
no_of_week_nights
type_of_meal_plan
required_car_parking_space
room_type_reserved
lead_time
arrival_year
arrival_month
arrival_date
market_segment_type
repeated_guest
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
avg_price_per_room
no_of_special_requests
booking_status

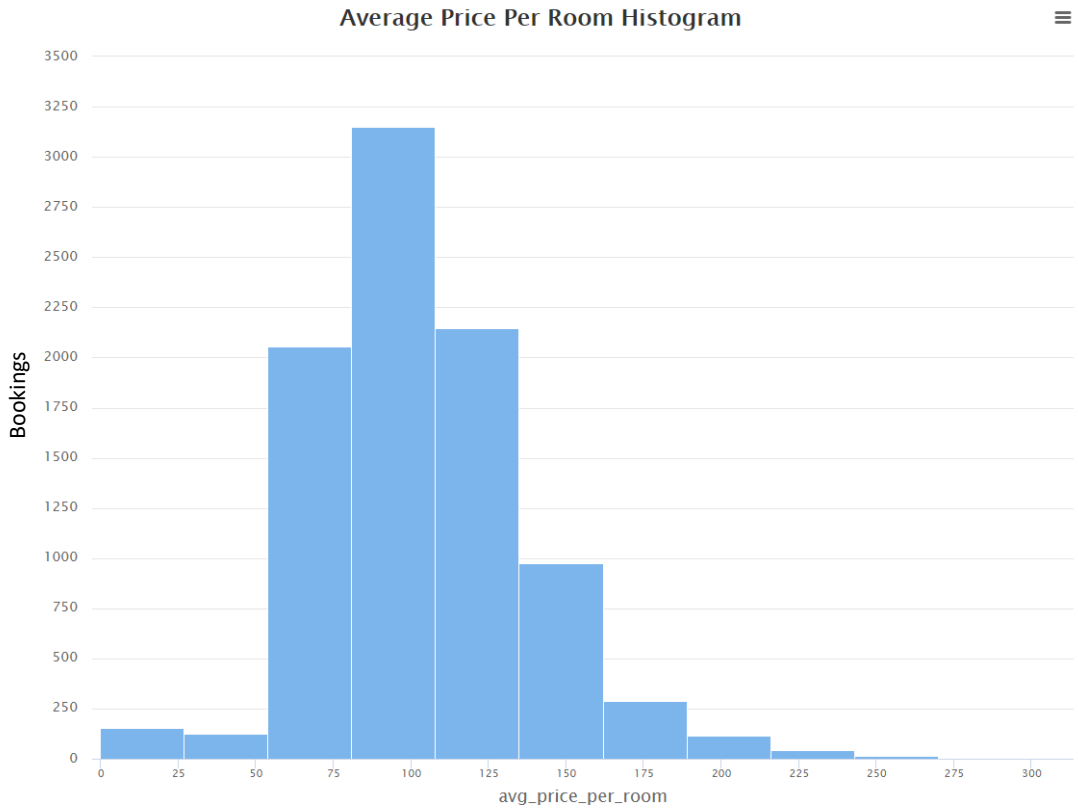
EDA RESULTS

Hotel Booking Cancellation Prediction

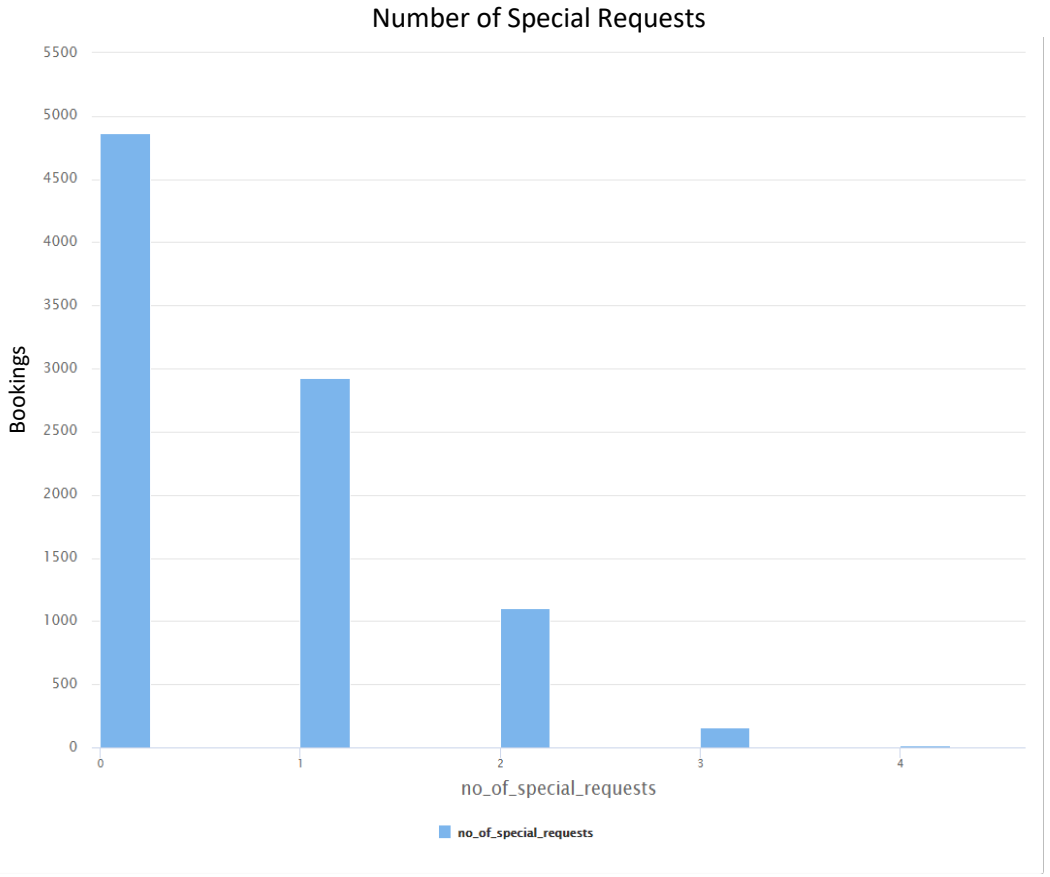
- Out of 9,069 there were 2,971 cancellations
- 32.7% of Bookings Cancelled
- Average people per stay is 1.94 people (1.84 adults and .1 children)
- Average number of nights per stay is 3 (2.2 week nights and .8 weekend nights)
- Average Lead time is 85 days from booking to arrival date
- The vast majority of bookings came from Online
- Average Ticket Price is $103.26 with a maximum price of $540 and plenty of rooms comped at $0.
- Most rooms sold between $54 and $162
- Most people ordered Meal Plan #1 or nothing and most did not have special requests
- Room Types 1 and 4 were most popular with 1 having over 75%
- As a percentage of total bookings, cancellations decrease as lead times decrease
- Results from 2017 and 2018 vary with respect to cancellations

EDA RESULTS
Hotel Booking Cancellation Prediction

The bulk of bookings fall between $54 and $161 per night room rate. Clearly in 2017 and 2018 these hotels are in the mid to lower market for pricing overall and would likely attract budget minded consumers.

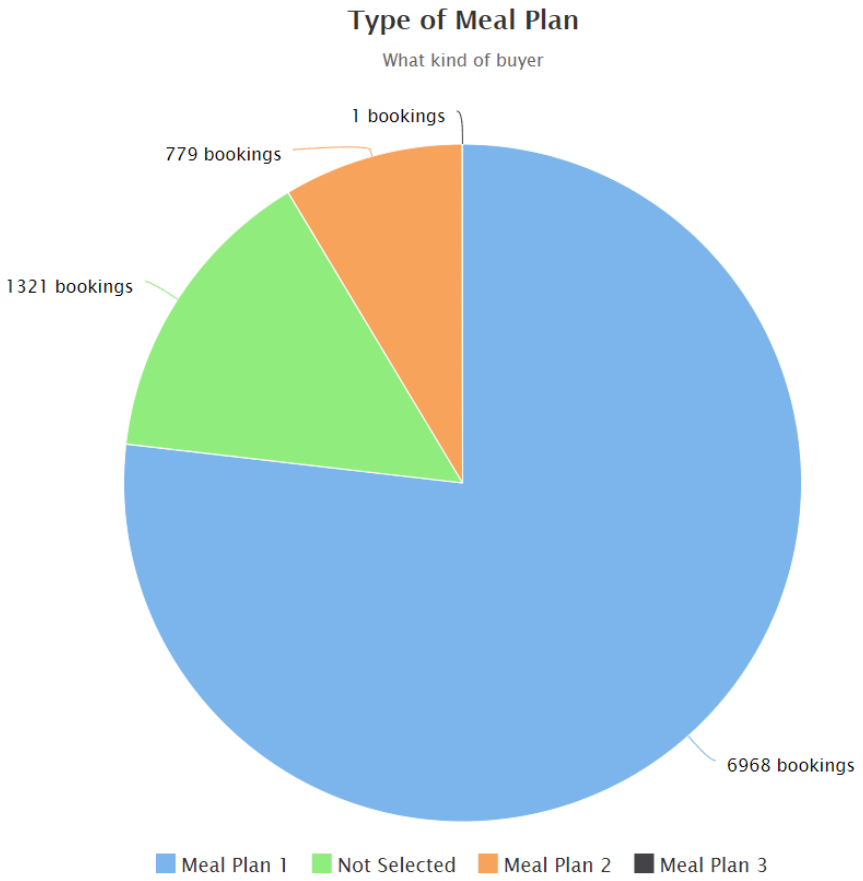


Just over half of bookings have no special requests. Just 1 request per booking are more then half of the rest.

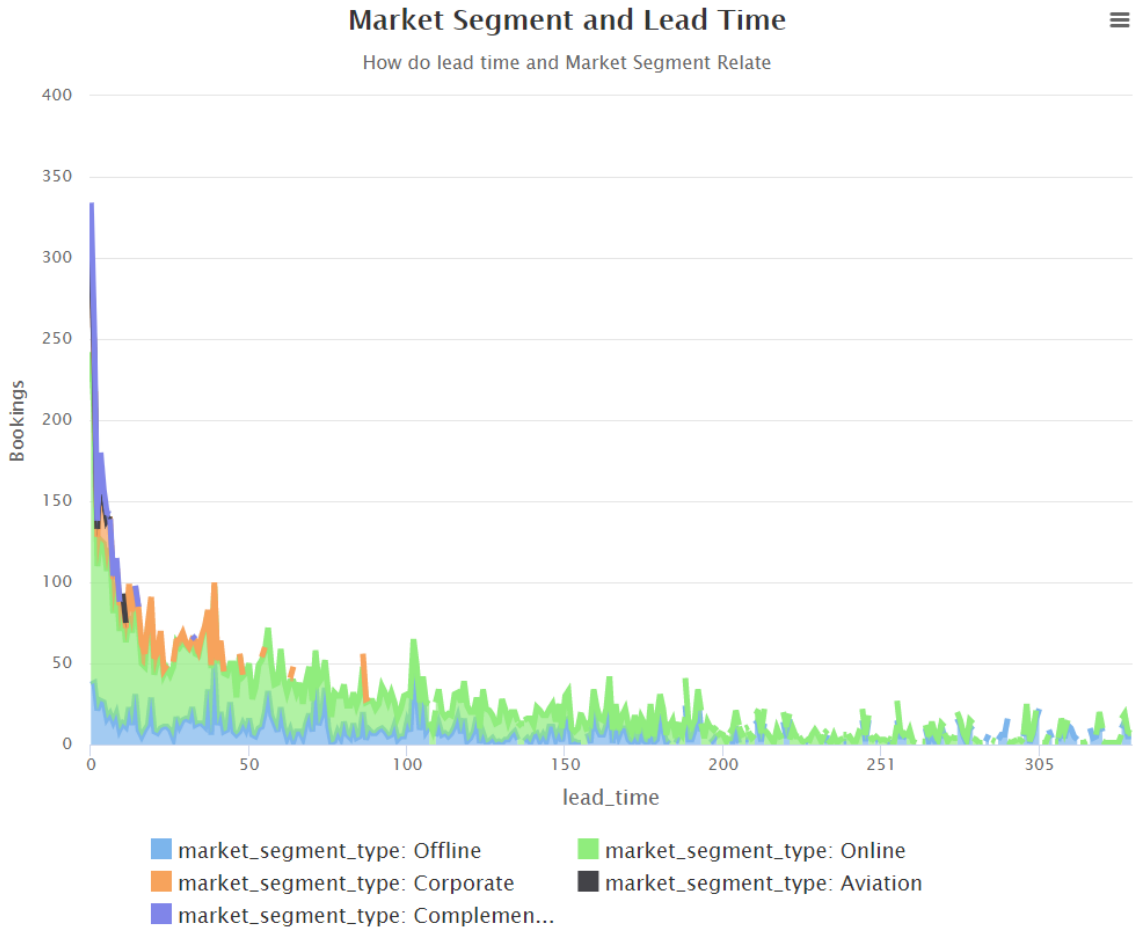


EDA RESULTS
Hotel Booking Cancellation Prediction

Most guests ordered the Type 1 Meal plan (breakfast only) or did not order a meal plan at all. These are the lower costs options and again reinforcing the notion of the budget minded constomers.



A lot of complimentary rooms are offered last minute (very few lead time days) and online buyers book more last minute.

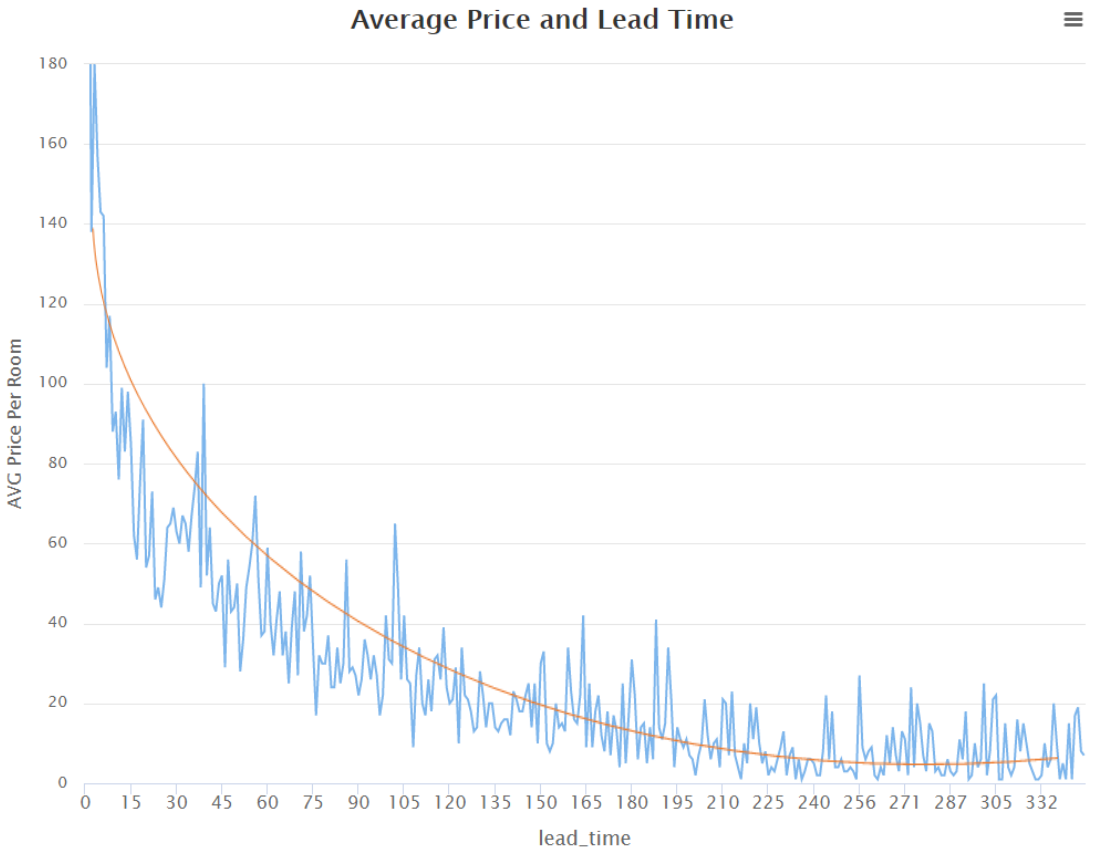


EDA RESULTS

Hotel Booking Cancellation Prediction

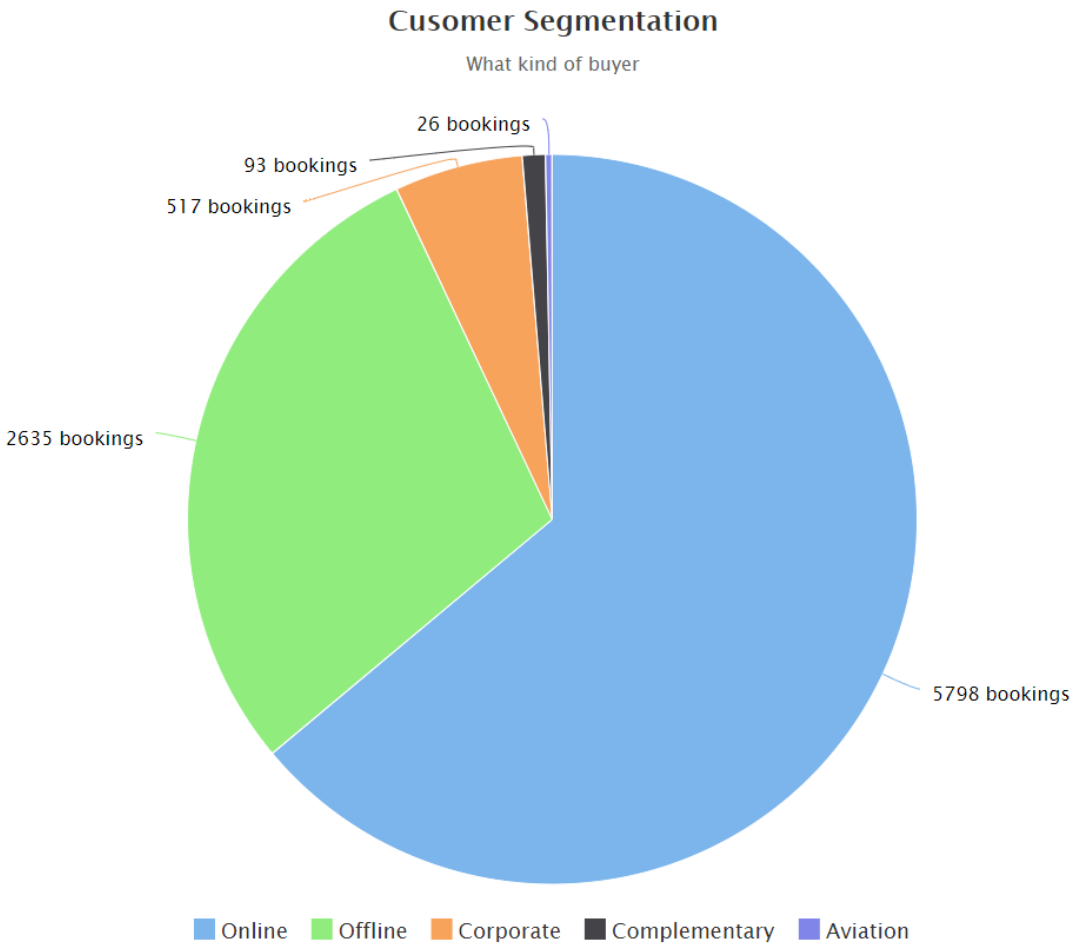
While “not cancelled” bookings are more than double “cancelled” the overall cancellation rate is high. There appears to be a significant opportunity to impact this rate.

Very clear to see that there is a correlation between price charged per room per night and lead time. As the lead time reduces, prices increase.

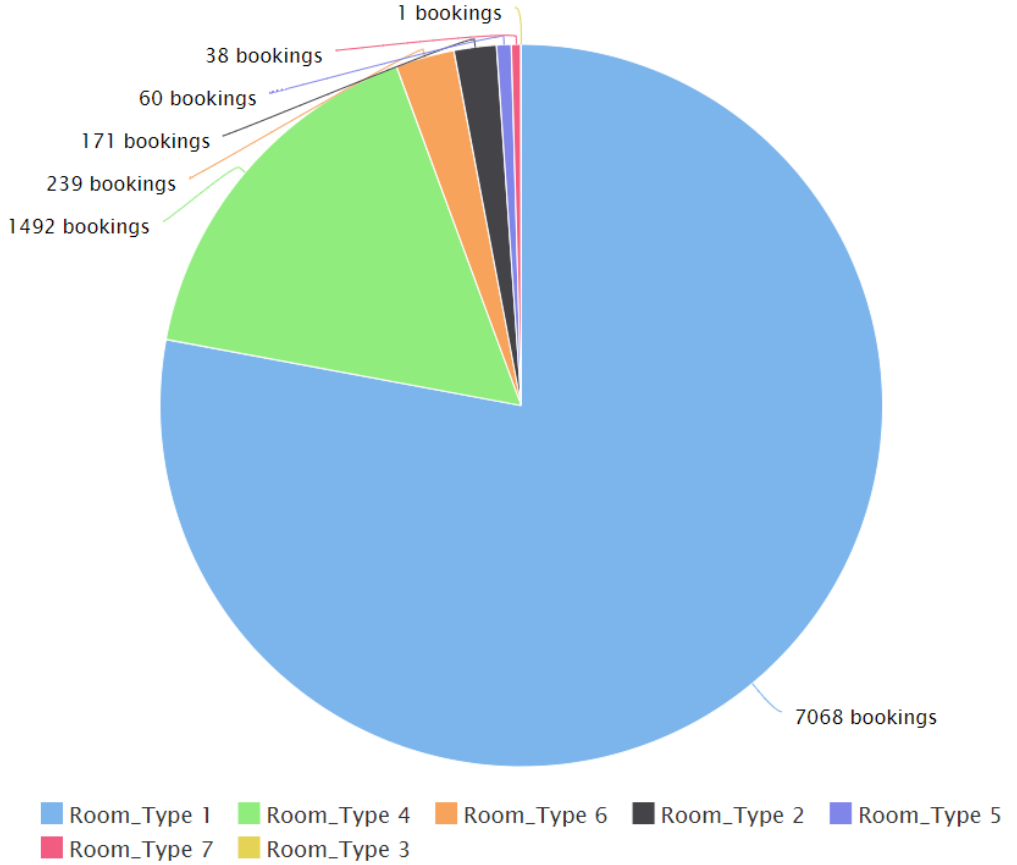


EDA RESULTS
Hotel Booking Cancellation Prediction

Online reservations are by far the largest category of buyer with Offline following as a distant second and corporate behind that.

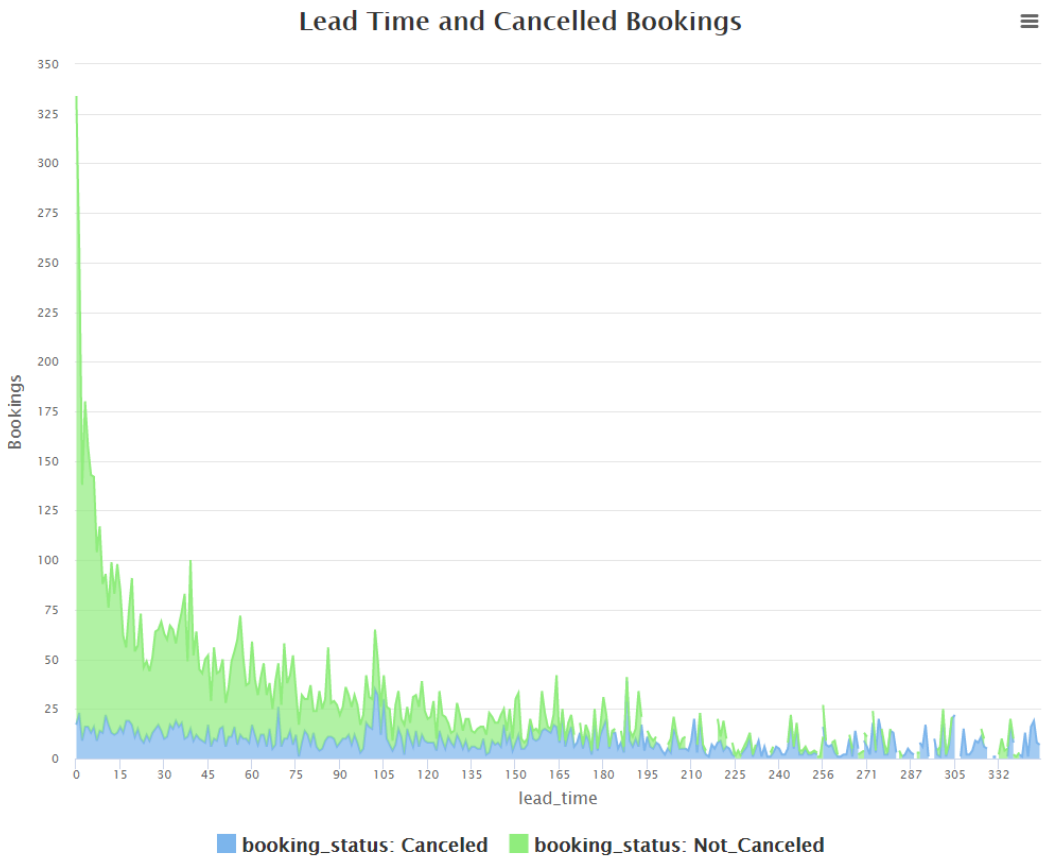


Room Type 1 is by far the majority booked room with Type 4 as a distant second. Since the hotel has decided to keep the features of online rooms confidential and have not disclosed a key for room type, it is difficult to draw conclusions from the data.



EDA RESULTS
Hotel Booking Cancellation Prediction

Cancelled bookings become a significantly smaller percentage of the total as lead times reduce. Essentially those that book last minute cancel much less.



The EDA results appear to convey that the INN customer tends to be budget minded, usually books online, do not require many special requests, and also tend to book last minute.

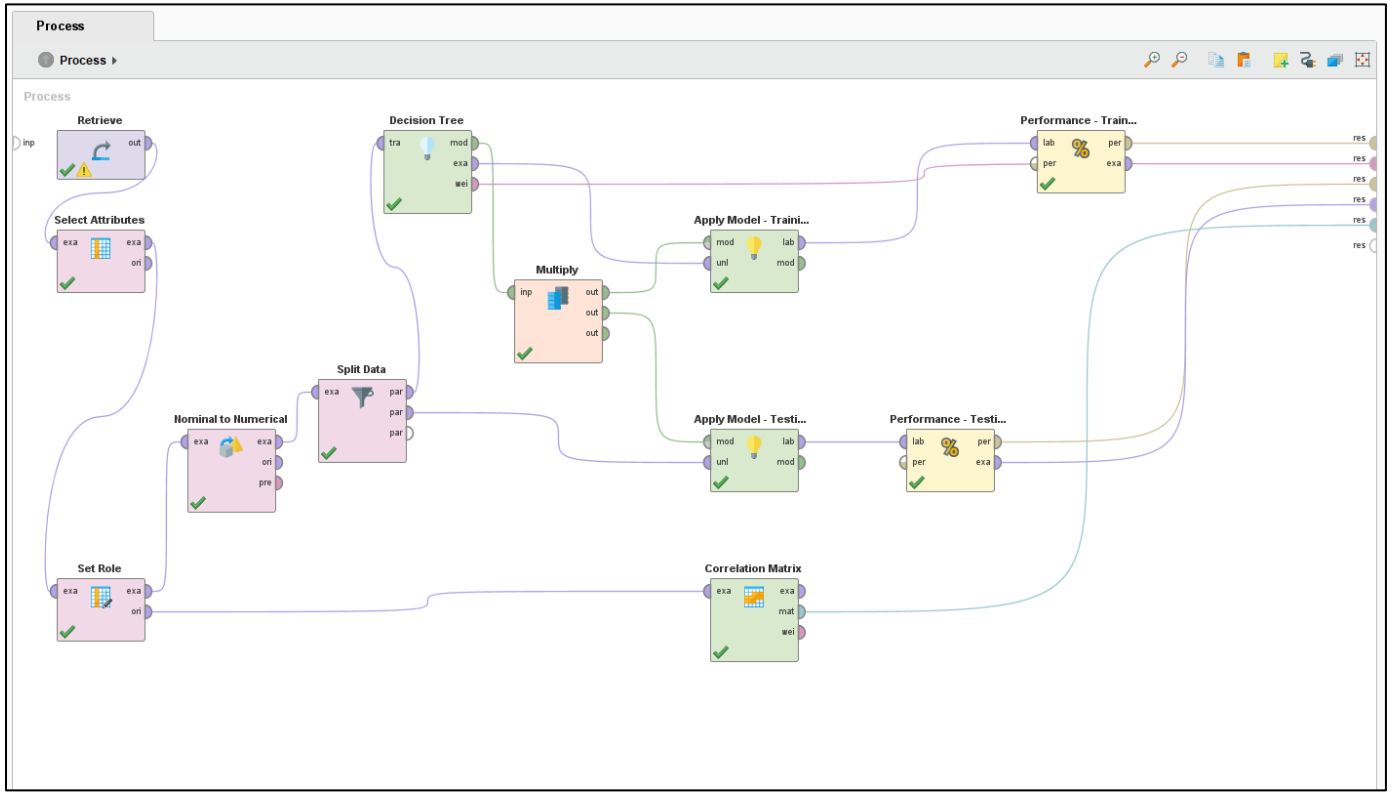
Those that book with less lead time pay higher prices and tend not to cancel their reservations nearly as much as the early bird longer lead time bookings.

Understanding lead times,

Data Pre-processing
Hotel Booking Cancellation Prediction

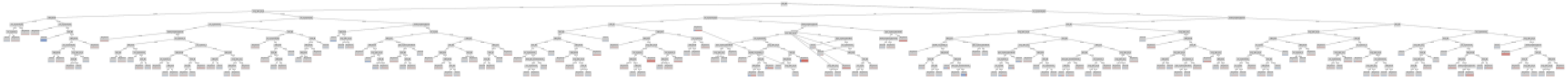
- There are no missing values in the data.
- As a unique identifier, “Booking_ID” column would contribute no value to analysis and is being removed from the data set.
- Dummy encoding has been performed to convert nominal variables into numeric variables for four data values (Market Segment, Room Type, Booking Status, Room Type). This was placed in the process not to impact the correlation table, but to prepare the data for the Split data Operator.
- Dataset has been split into training and testing set in the ratio of 80:20.
- After multiple iterations a Decision Tree depth of 11 was settled on to maximize performance results.

The Initial Decision Tree Process



Model Performance Summary: Decision Tree Pruned

Hotel Booking Cancellation Prediction



The Chosen Model

The ***Decision Tree Pruned*** produced the best performance results based on a myriad of considerations including:

1. simplicity,
2. accuracy,
3. precision and
4. recall performance.

And, the F1 Score corroborates with the highest score as well giving us a fairly strong level of confidence that the pruned decision tree is our best predictor.

Most Important Features

The data was fairly consistent across the models in determining the most important factors with the top results being:

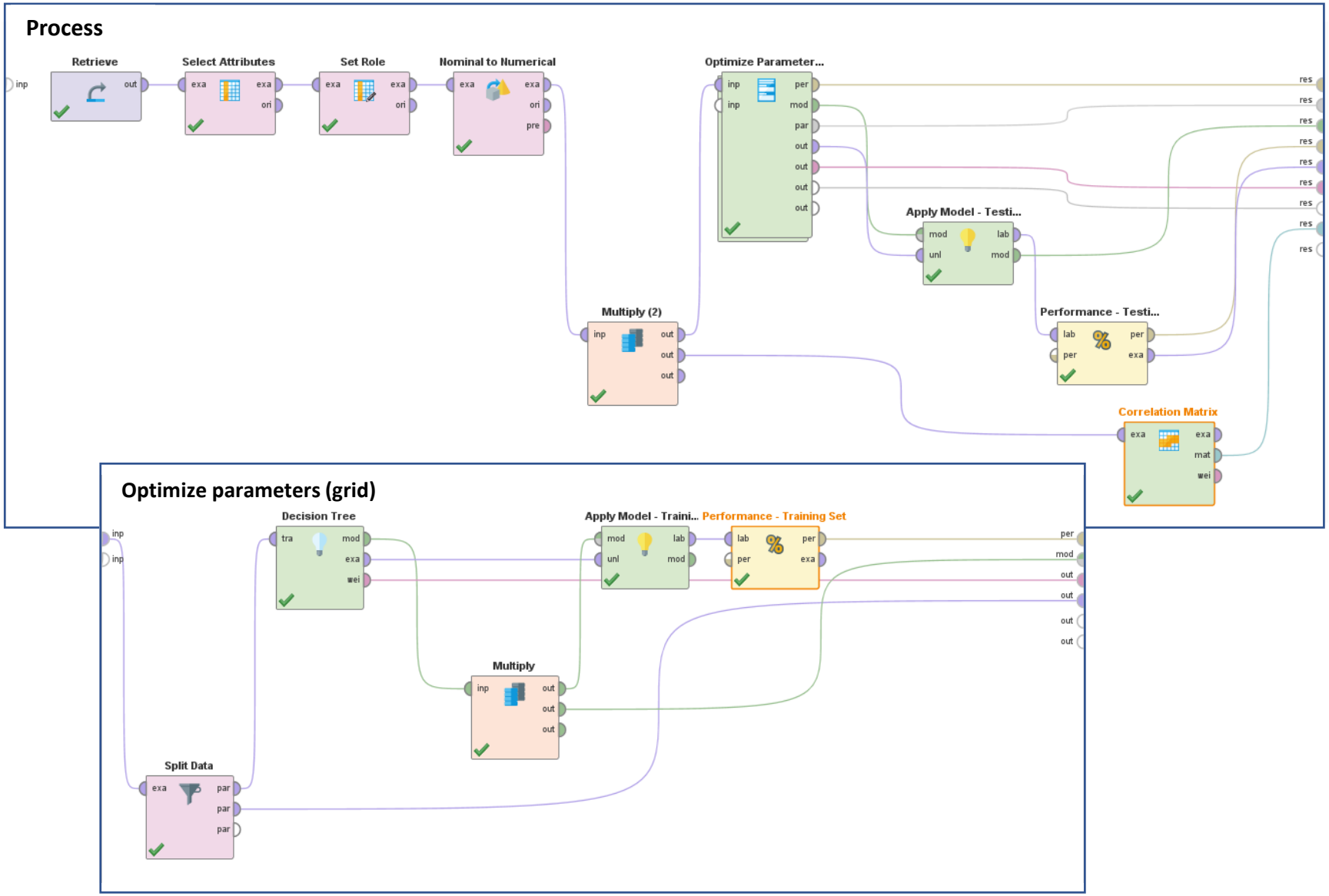
1. Leadtime
2. Average Price
3. Timing (month and date combined)
4. Length of Stay (week nights and weekend nights combined)

Decision Tree Pruned

```
lead_time > 145.500
|   avg_price_per_room > 100.150
|   |   arrival_month > 11.500
|   |   |   no_of_special_requests > 0.500
|   |   |   |   no_of_week_nights > 3: Canceled {Canceled=2, Not_Canceled=0}
|   |   |   |   no_of_week_nights ≤ 3: Not_Canceled {Canceled=0, Not_Canceled=2}
|   |   |   no_of_special_requests ≤ 0.500: Not_Canceled {Canceled=0, Not_Canceled=19}
|   |   arrival_month ≤ 11.500
|   |   |   no_of_special_requests > 2.500: Not_Canceled {Canceled=0, Not_Canceled=10}
|   |   |   no_of_special_requests ≤ 2.500
|   |   |   |   lead_time > 150.500: Canceled {Canceled=580, Not_Canceled=0}
|   |   |   |   lead_time ≤ 150.500
|   |   |   |   |   arrival_month > 5.500
|   |   |   |   |   |   no_of_special_requests > 1.500
|   |   |   |   |   |   |   arrival_month > 8: Canceled {Canceled=2, Not_Canceled=0}
|   |   |   |   |   |   |   arrival_month ≤ 8: Not_Canceled {Canceled=0, Not_Canceled=3}
|   |   |   |   |   |   no_of_special_requests ≤ 1.500
|   |   |   |   |   |   |   avg_price_per_room > 126.880
|   |   |   |   |   |   |   |   lead_time > 148.500: Not_Canceled {Canceled=0, Not_Canceled=2}
|   |   |   |   |   |   |   |   lead_time ≤ 148.500: Canceled {Canceled=6, Not_Canceled=1}
|   |   |   |   |   |   |   avg_price_per_room ≤ 126.880: Canceled {Canceled=12, Not_Canceled=0}
|   |   |   |   |   arrival_month ≤ 5.500: Not_Canceled {Canceled=0, Not_Canceled=6}
|   avg_price_per_room ≤ 100.150
|   |   no_of_special_requests > 0.500
|   |   |   no_of_weekend_nights > 0.500
|   |   |   |   market_segment_type = Offline > 0.500: Not_Canceled {Canceled=0, Not_Canceled=54}
|   |   |   |   market_segment_type = Offline ≤ 0.500
|   |   |   |   |   no_of_week_nights > 4.500
|   |   |   |   |   |   arrival_date > 24.500
|   |   |   |   |   |   |   lead_time > 229: Canceled {Canceled=6, Not_Canceled=0}
|   |   |   |   |   |   |   lead_time ≤ 229: Not_Canceled {Canceled=0, Not_Canceled=2}
|   |   |   |   |   |   arrival_date ≤ 24.500
|   |   |   |   |   |   |   no_of_weekend_nights > 3.500: Canceled {Canceled=2, Not_Canceled=0}
|   |   |   |   |   |   |   no_of_weekend_nights ≤ 3.500
|   |   |   |   |   |   |   |   lead_time > 166: Not_Canceled {Canceled=4, Not_Canceled=20}
|   |   |   |   |   |   |   |   lead_time ≤ 166: Canceled {Canceled=3, Not_Canceled=1}
|   |   |   |   |   |   no_of_week_nights ≤ 4.500
|   |   |   |   |   |   |   no_of_week_nights > 1.500
|   |   |   |   |   |   |   |   arrival_month > 10.500
|   |   |   |   |   |   |   |   |   no_of_week_nights > 2.500: Not_Canceled {Canceled=0, Not_Canceled=5}
|   |   |   |   |   |   |   |   |   no_of_week_nights ≤ 2.500
|   |   |   |   |   |   |   |   |   |   arrival_date > 19: Canceled {Canceled=4, Not_Canceled=0}
|   |   |   |   |   |   |   |   |   |   arrival_date ≤ 19: Not_Canceled {Canceled=0, Not_Canceled=2}
|   |   |   |   |   |   |   |   arrival_month ≤ 10.500
|   |   |   |   |   |   |   |   |   lead_time > 233.500
|   |   |   |   |   |   |   |   |   |   lead_time > 244: Not_Canceled {Canceled=1, Not_Canceled=8}
|   |   |   |   |   |   |   |   |   |   lead_time ≤ 244: Canceled {Canceled=3, Not_Canceled=1}
|   |   |   |   |   |   |   |   |   lead_time ≤ 233.500: Not_Canceled {Canceled=2, Not_Canceled=59}
|   |   |   |   |   |   |   no_of_week_nights ≤ 1.500
|   |   |   |   |   |   |   |   arrival_date > 26.500
|   |   |   |   |   |   |   |   |   arrival_date > 30: Not_Canceled {Canceled=0, Not_Canceled=2}
|   |   |   |   |   |   |   |   |   arrival_date ≤ 30: Canceled {Canceled=6, Not_Canceled=0}
|   |   |   |   |   |   |   |   arrival_date ≤ 26.500
|   |   |   |   |   |   |   |   |   avg_price_per_room > 77.850
|   |   |   |   |   |   |   |   |   |   arrival_month > 9.500: Canceled {Canceled=3, Not_Canceled=2}
|   |   |   |   |   |   |   |   |   |   arrival_month ≤ 9.500: Not_Canceled {Canceled=3, Not_Canceled=14}
|   |   |   |   |   |   |   |   |   avg_price_per_room ≤ 77.850: Not_Canceled {Canceled=0, Not_Canceled=7}
|   |   |   |   no_of_weekend_nights ≤ 0.500
|   |   |   |   |   lead_time > 175.500
|   |   |   |   |   |   no_of_special_requests > 2.500: Not_Canceled {Canceled=0, Not_Canceled=3}
|   |   |   |   |   |   no_of_special_requests ≤ 2.500
|   |   |   |   |   |   |   arrival_date > 23.500
|   |   |   |   |   |   |   |   lead_time > 247.500: Not_Canceled {Canceled=0, Not_Canceled=3}
```

Model Performance Summary: Decision Tree Pruned RapidMiner Process

Hotel Booking Cancellation Prediction



Model Performance Summary
Hotel Booking Cancellation Prediction

	Most Important		Highly Important				Highly Important		Highly Important
Model	F1 Score (test data)	Training Set Accuracy	Test Set Accuracy	Training Set Recall (FP - FN)	Training Set Precision (TP - TN)	Test Set Recall (FP - FN)	Test Set Weighted Recall	Test Set Precision (TP - TN)	Test Set Weighted Precision
Decision Tree	0.79	90.14	86.44	83.72 – 93.28	85.85 – 92.16	77.61 – 90.64	84.17	80.31 – 89.27	84.79
Decision Tree Pruned***	0.79	89.46	86.88	81.11 – 93.52	85.92 – 91.04	76.43 – 91.97	84.20	82.25 – 88.91	85.58
Random Forest	0.74	87.00	85.12	68.53 – 96.00	89.31 – 86.23	63.97 – 95.41	79.69	87.16 – 84.47	85.81
Random Forest Pruned	0.27	72.56	72.38	16.24 – 100	100.00 – 71.01	15.66 – 100	57.83	100.00 – 70.89	85.44

The model of choosing is ***Decision Tree Pruned*** as it produced the highest F1 Score, Accuracy, Weighted Recall, and very high Weighted Precision. And a small variance of just 3.42% between Training and Test Accuracy and Recall. The DT Pruned consistently performed better than the more extrapolated random forest models and produced the clearest definition (Elbow) of which attributes bare the most weighted significance.

The Random Forest Pruned model is the clear laggard in producing a strong result for this data set. The remaining three models all performed similarly.

These are all overall good models and in this case the Accuracy, F1 Scores, reduced computational requirement, and the clear definition of key principle components lead to considering the simple and most efficient model of the simple ***Decision Tree Pruned*** .

Appendix

This image produced on MidJourney

Appendix: Dataset Overview
Hotel Booking Cancellation Prediction

- 9,069 Rows (each a unique booking)
- 19 Columns (Booking_ID is a unique key and will not add value to the analysis and therefor there 18 attributes)
- Data is well structured, there are no missing values and formats appear to be consistent

Row No.	Booking_ID	no_of_adults	no_of_childr...	no_of_week...	no_of_week...	type_of_meal_plan	required_ca...	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_seg...	repeated_g...	no_of_previ...	no_of_previ...	avg_price_p...	no_of_speci...	booking_sta...
1	INN23152	1	0	0	2	Meal Plan 1	0	Room_Type 1	188	2018	6	15	Offline	0	0	0	130	0	Canceled
2	INN21915	1	0	0	2	Meal Plan 1	0	Room_Type 1	103	2018	4	19	Offline	0	0	0	115	0	Canceled
3	INN24290	2	0	1	4	Not Selected	0	Room_Type 1	33	2018	4	18	Online	0	0	0	90.540	0	Canceled
4	INN31921	2	0	0	3	Meal Plan 1	0	Room_Type 1	64	2018	11	22	Online	0	0	0	93.600	1	Canceled
5	INN34718	2	0	1	1	Meal Plan 2	0	Room_Type 1	247	2018	6	6	Offline	0	0	0	115	1	Canceled
6	INN31303	2	0	0	3	Meal Plan 1	0	Room_Type 1	304	2018	11	3	Offline	0	0	0	89	0	Canceled
7	INN34963	1	0	3	5	Not Selected	0	Room_Type 1	275	2018	10	9	Online	0	0	0	91.690	0	Canceled
8	INN14729	2	0	2	0	Meal Plan 1	0	Room_Type 1	146	2018	4	24	Offline	0	0	0	95	0	Canceled
9	INN06771	2	2	2	3	Meal Plan 1	0	Room_Type 2	41	2018	9	4	Online	0	0	0	208.930	0	Canceled
10	INN34053	2	0	2	1	Meal Plan 1	0	Room_Type 4	41	2018	9	18	Online	0	0	0	149.400	1	Canceled
11	INN12335	2	0	1	0	Not Selected	0	Room_Type 1	10	2018	3	13	Online	0	0	0	97	0	Canceled
12	INN32422	3	0	2	1	Meal Plan 1	0	Room_Type 4	128	2018	10	29	Online	0	0	0	123.300	1	Canceled
13	INN07271	1	0	0	1	Meal Plan 1	0	Room_Type 1	177	2018	7	30	Online	0	0	0	99.900	0	Canceled
14	INN06950	2	0	2	2	Meal Plan 1	0	Room_Type 1	180	2018	11	27	Online	0	0	0	85	0	Canceled
15	INN06595	2	0	0	2	Meal Plan 2	0	Room_Type 1	346	2018	9	13	Online	0	0	0	115	1	Canceled
16	INN05722	1	0	0	3	Meal Plan 1	0	Room_Type 1	166	2018	11	1	Online	0	0	0	110	0	Canceled
17	INN27728	3	0	0	1	Meal Plan 1	0	Room_Type 1	220	2018	9	29	Offline	0	0	0	105.400	2	Canceled
18	INN20017	1	0	2	1	Meal Plan 1	0	Room_Type 1	163	2018	10	15	Offline	0	0	0	115	0	Canceled
19	INN05909	2	0	2	2	Meal Plan 1	0	Room_Type 1	174	2018	9	2	Online	0	0	0	119.850	1	Canceled
20	INN24458	2	0	2	4	Meal Plan 1	0	Room_Type 1	160	2018	4	27	Offline	0	0	0	90	0	Canceled
21	INN10762	2	0	0	2	Not Selected	0	Room_Type 1	163	2018	12	2	Online	0	0	0	79.200	0	Canceled
22	INN04584	2	0	1	5	Meal Plan 1	0	Room_Type 4	79	2018	4	4	Online	0	0	0	96.620	1	Canceled
23	INN00902	2	1	0	3	Meal Plan 1	0	Room_Type 1	124	2018	5	31	Online	0	0	0	143.100	1	Canceled
24	INN28424	1	0	0	2	Meal Plan 1	0	Room_Type 1	103	2018	4	19	Offline	0	0	0	115	0	Canceled
25	INN09306	2	0	2	1	Meal Plan 1	0	Room_Type 4	202	2018	10	22	Online	0	0	0	109.800	0	Canceled
26	INN01117	2	0	1	2	Meal Plan 2	0	Room_Type 1	418	2018	9	26	Online	0	0	0	107	0	Canceled
27	INN06459	2	0	1	3	Not Selected	0	Room_Type 1	56	2018	12	26	Online	0	0	0	88.400	0	Canceled
28	INN20404	2	0	0	3	Meal Plan 2	0	Room_Type 1	34	2017	9	23	Offline	0	0	0	224.670	0	Canceled
29	INN23504	2	0	0	1	Meal Plan 1	0	Room_Type 1	3	2017	10	8	Corporate	0	0	0	65	0	Canceled
30	INN31410	2	2	2	7	Meal Plan 1	0	Room_Type 6	151	2018	7	1	Online	0	0	0	167.450	2	Canceled
31	INN01666	1	0	0	5	Meal Plan 1	0	Room_Type 1	230	2018	9	6	Online	0	0	0	111	0	Canceled
32	INN11148	2	0	0	1	Meal Plan 1	0	Room_Type 1	443	2018	4	29	Offline	0	0	0	65	0	Canceled
33	INN19257	2	0	1	1	Meal Plan 1	0	Room_Type 1	218	2018	8	13	Online	0	0	0	69.330	0	Canceled
34	INN22869	2	1	0	3	Meal Plan 1	0	Room_Type 1	42	2018	10	20	Online	0	0	0	164.330	1	Canceled
35	INN36190	2	0	0	3	Meal Plan 1	0	Room_Type 1	195	2018	9	28	Online	0	0	0	126.900	0	Canceled
36	INN00969	2	0	0	1	Meal Plan 1	0	Room_Type 1	105	2018	4	6	Online	0	0	0	75	0	Canceled

Appendix: Data Dictionary
Hotel Booking Cancellation Prediction

Observations of data:

The data provided is of various reservations (booking_id) made with the hotel group in Portugal in which not only the detailed room and reservation information was captured but details on the customer including attributes like repeat customer, number previous cancels, number of special requests are also available for analysis.

There are 18 attributes (removing booking_id) in the data set. For a definition of data terms see Appendix.

Overall, the data is in fairly good structure with 9,069 booking records.

Certain attributes can be grouped to create better directional results. For instance, understanding the summed impact of weekend nights and week nights can give us a better predictor on how much the length of stay in general is impacted.

Several Attributes may required Dummy Processing to convert from polynominal to categorized numeric.

Data 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:
    1. Not Selected – No meal plan selected
    2. Meal Plan 1 – Breakfast
    3. Meal Plan 2 – Half board (breakfast and one other meal)
    4. Meal Plan 3 – Full board (breakfast, lunch, and dinner)
- 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.

Attribute Overview

The data provided is of various reservations (booking_id) made with the hotel in which not only the detailed room and reservation information was captured but details on the customer including attributes like repeat customer, number previous cancels, number of special requests are also available for analysis.

There are 18 attributes (removing booking_id) in the data set. For a definition of data terms see Appendix.

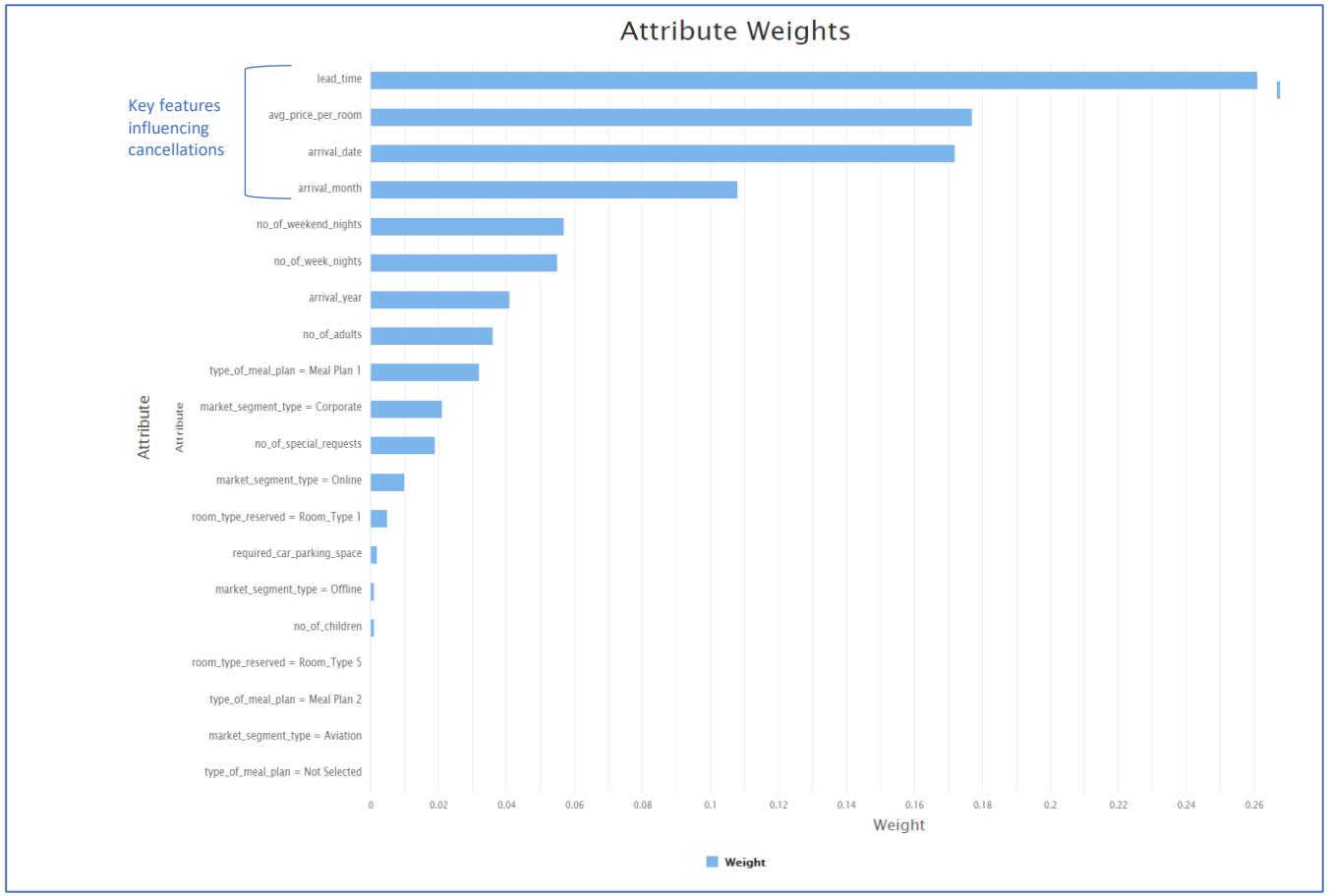
Overall, the data is in fairly good structure with 9,069 booking records.

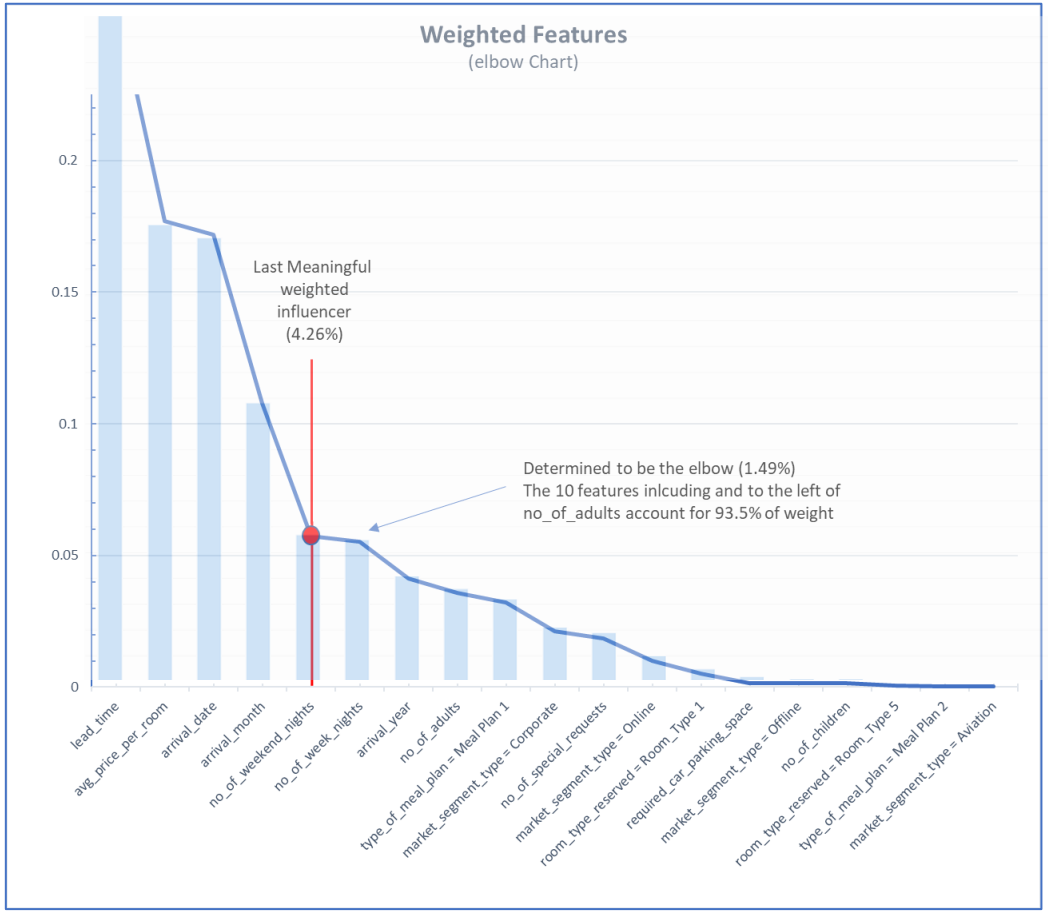
Data Attributes

no_of_adults
no_of_children
no_of_weekend_nights
no_of_week_nights
type_of_meal_plan
required_car_parking_space
room_type_reserved
lead_time
arrival_year
arrival_month
arrival_date
market_segment_type
repeated_guest
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
avg_price_per_room
no_of_special_requests
booking_status

Appendix: Model Building (Decision Tree with Pruning) Weighted Attributes

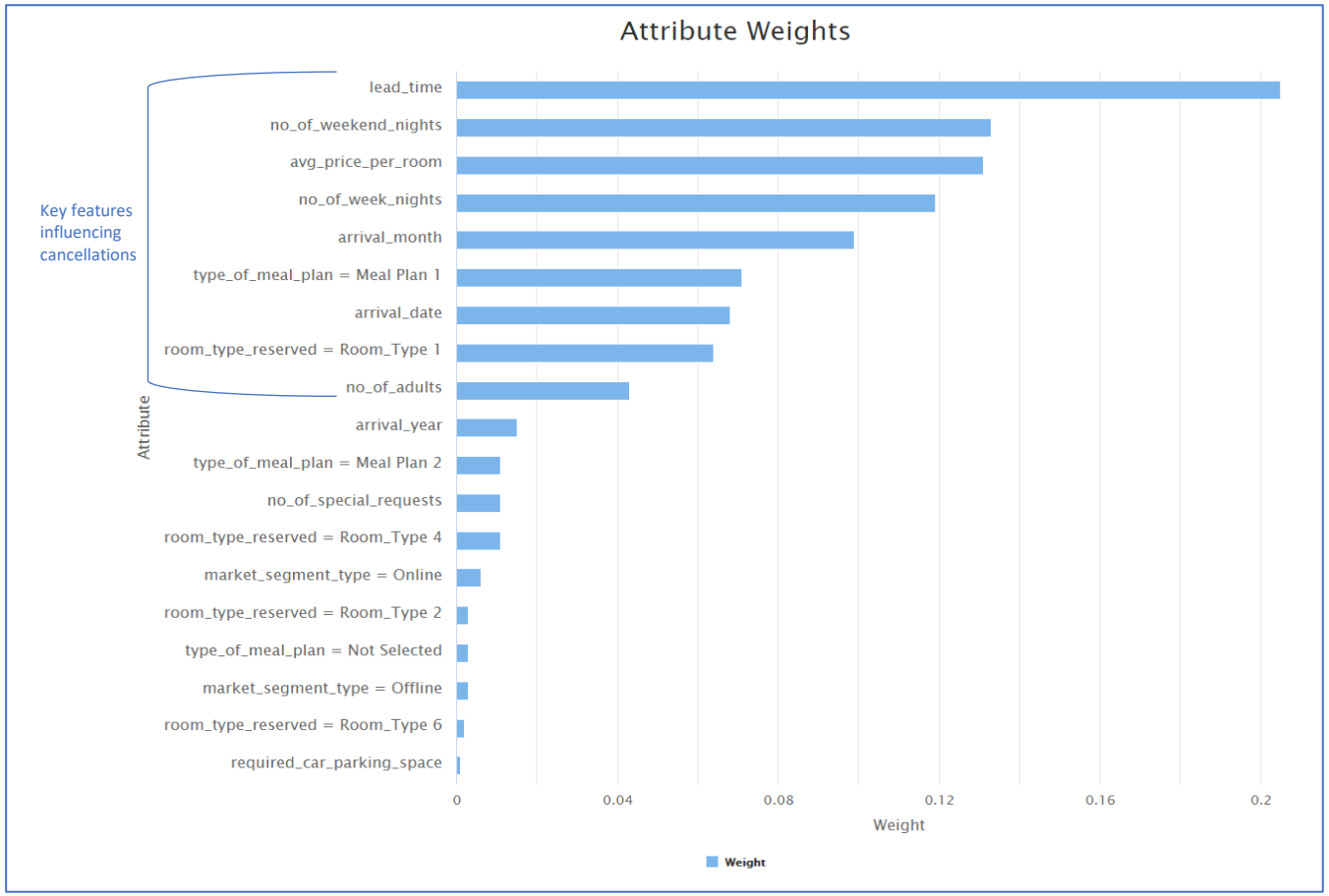
Hotel Booking Cancellation Prediction

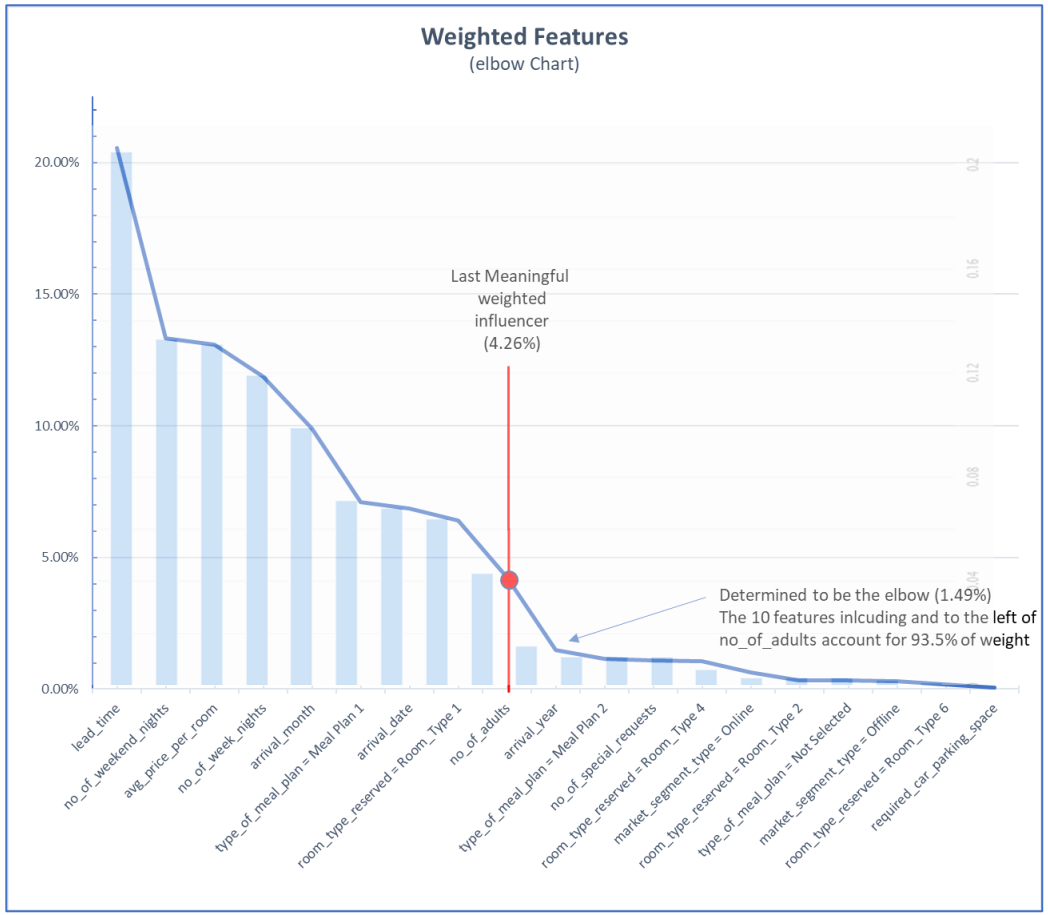




Appendix: Model Building (Decision Tree) Weighted Attributes

Hotel Booking Cancellation Prediction





Appendix: Decision Tree Pruned Correlation Matrix

Hotel Booking Cancellation Prediction

Attributes	type_of...	type_of...	type_of...	type_of...	room_t...	room_t...	room_t...	room_t...	room_t...	room_t...	room_t...	market...	market...	market...	market...	market...	no_of_a...	no_of_c...	no_of_...	no_of_...	require...	lead_ti...	arrival_...	arrival_...	arrival_...	repeate...	no_of_p...	no_of_p...	avg_pri...	no_of_s...
type_of_meal_plan = Meal Plan 1	1	-0.752	-0.558	-0.019	-0.210	0.059	0.182	0.046	0.029	0.028	0.006	0.003	-0.077	0.129	0.029	0.040	-0.025	0.080	0.048	0.080	0.018	-0.034	0.013	-0.019	-0.020	0.074	0.021	0.042	0.002	0.017
type_of_meal_plan = Not Selected	-0.752	1	-0.127	-0.004	0.190	-0.039	-0.167	-0.050	-0.022	-0.022	-0.004	-0.235	0.280	-0.099	-0.022	-0.036	0.012	-0.075	-0.018	-0.063	0.007	-0.130	0.119	-0.016	-0.005	-0.052	-0.010	-0.030	-0.086	0.053
type_of_meal_plan = Meal Plan 2	-0.558	-0.127	1	-0.003	0.079	-0.040	-0.064	-0.006	-0.015	-0.020	-0.003	0.292	-0.236	-0.070	-0.016	-0.019	0.022	-0.026	-0.050	-0.041	-0.036	0.215	-0.170	0.049	0.036	-0.046	-0.019	-0.025	0.107	-0.092
type_of_meal_plan = Meal Plan 3	-0.019	-0.004	-0.003	1	-0.020	-0.001	-0.005	-0.002	-0.001	0.162	-0.000	-0.007	-0.014	-0.003	-0.001	0.103	0.003	-0.003	-0.010	0.006	-0.002	-0.010	0.005	-0.005	0.006	-0.002	-0.001	-0.001	-0.031	0.005
room_type_reserved = Room_Type 1	-0.210	0.190	0.079	-0.020	1	-0.261	-0.834	-0.309	-0.153	-0.122	-0.020	0.229	-0.240	0.079	-0.056	-0.043	-0.258	-0.263	-0.067	-0.100	-0.027	0.086	-0.096	0.011	-0.034	0.036	0.013	0.012	-0.386	-0.146
room_type_reserved = Room_Type 2	0.059	-0.039	-0.040	-0.001	-0.261	1	-0.062	-0.023	-0.011	-0.009	-0.001	-0.071	0.077	-0.031	-0.007	0.026	-0.070	0.161	0.017	0.015	0.018	0.024	-0.025	0.009	0.014	-0.018	-0.009	-0.010	-0.056	0.027
room_type_reserved = Room_Type 4	0.182	-0.167	-0.064	-0.005	-0.834	-0.062	1	-0.073	-0.036	-0.029	-0.005	-0.186	0.207	-0.074	0.071	-0.013	0.290	-0.067	0.067	0.104	-0.007	-0.073	0.107	-0.023	0.025	-0.038	-0.012	-0.018	0.274	0.124
room_type_reserved = Room_Type 6	0.046	-0.050	-0.006	-0.002	-0.309	-0.023	-0.073	1	-0.013	-0.011	-0.002	-0.093	0.105	-0.035	-0.009	0.004	0.052	0.647	0.013	0.004	0.059	-0.047	0.005	0.013	0.014	-0.014	-0.006	-0.009	0.361	0.060
room_type_reserved = Room_Type 5	0.029	-0.022	-0.015	-0.001	-0.153	-0.011	-0.036	-0.013	1	-0.005	-0.001	0.011	-0.052	0.056	-0.004	0.073	-0.020	0.005	-0.006	-0.004	0.009	-0.026	0.029	0.011	0.006	0.020	-0.005	0.034	0.037	-0.017
room_type_reserved = Room_Type 7	0.028	-0.022	-0.020	0.162	-0.122	-0.009	-0.029	-0.011	-0.005	1	-0.001	-0.038	-0.005	-0.001	-0.003	0.197	0.036	0.124	-0.011	0.005	0.018	-0.037	0.005	-0.006	0.001	0.032	0.024	0.023	0.075	0.042
room_type_reserved = Room_Type 3	0.006	-0.004	-0.003	-0.000	-0.020	-0.001	-0.005	-0.002	-0.001	-0.001	1	0.016	-0.014	-0.003	-0.001	-0.001	0.003	-0.003	-0.010	-0.002	-0.002	0.004	0.005	0.009	-0.003	-0.002	-0.001	-0.001	0.008	-0.008
market_segment_type = Offline	0.003	-0.235	0.292	-0.007	0.229	-0.071	-0.186	-0.093	0.011	-0.038	0.016	1	-0.852	-0.157	-0.034	-0.065	-0.078	-0.126	-0.068	-0.014	-0.100	0.283	-0.153	0.029	-0.009	-0.065	-0.021	-0.052	-0.215	-0.345
market_segment_type = Online	-0.077	0.280	-0.236	-0.014	-0.240	0.077	0.207	0.105	-0.052	-0.005	-0.014	-0.852	1	-0.327	-0.071	-0.136	0.240	0.147	0.116	0.086	0.035	-0.155	0.174	-0.020	0.002	-0.179	-0.029	-0.098	0.339	0.386
market_segment_type = Corporate	0.129	-0.099	-0.070	-0.003	0.079	-0.031	-0.074	-0.035	0.056	-0.001	-0.003	-0.157	-0.327	1	-0.013	-0.025	-0.294	-0.059	-0.091	-0.129	0.098	-0.186	-0.043	-0.022	0.013	0.416	0.081	0.266	-0.152	-0.128
market_segment_type = Aviation	0.029	-0.022	-0.016	-0.001	-0.056	-0.007	0.071	-0.009	-0.004	-0.003	-0.001	-0.034	-0.071	-0.013	1	-0.005	-0.087	-0.014	0.011	0.029	0.015	-0.049	0.026	-0.001	0.013	0.030	-0.004	0.002	-0.002	-0.043
market_segment_type = Complementary	0.040	-0.036	-0.019	0.103	-0.043	0.026	-0.013	0.004	0.073	0.197	-0.001	-0.065	-0.136	-0.025	-0.005	1	-0.075	0.011	-0.048	-0.066	0.052	-0.085	-0.055	0.018	-0.007	0.174	0.051	0.088	-0.295	0.033
no_of_adults	-0.025	0.012	0.022	0.003	-0.258	-0.070	0.290	0.052	-0.020	0.036	0.003	-0.078	0.240	-0.294	-0.087	-0.075	1	-0.022	0.118	0.099	-0.011	0.104	0.070	0.014	0.011	-0.193	-0.045	-0.118	0.290	0.191
no_of_children	0.080	-0.075	-0.026	-0.003	-0.263	0.161	-0.067	0.647	0.005	0.124	-0.003	-0.126	0.147	-0.059	-0.014	0.011	-0.022	1	0.019	0.015	0.047	-0.044	0.048	0.001	0.022	-0.036	-0.017	-0.020	0.330	0.116
no_of_weekend_nights	0.048	-0.018	-0.050	-0.010	-0.067	0.017	0.067	0.013	-0.006	-0.011	-0.010	-0.068	0.116	-0.091	0.011	-0.048	0.118	0.019	1	0.197	-0.043	0.052	0.062	-0.029	0.019	-0.060	-0.015	-0.015	-0.005	0.079
no_of_week_nights	0.080	-0.063	-0.041	0.006	-0.100	0.015	0.104	0.004	-0.004	0.005	-0.002	-0.014	0.086	-0.129	0.029	-0.066	0.099	0.015	0.197	1	-0.064	0.160	0.034	0.032	-0.003	-0.090	-0.015	-0.023	0.013	0.052
required_car_parking_space	0.018	0.007	-0.036	-0.002	-0.027	0.018	-0.007	0.059	0.009	0.018	-0.002	-0.100	0.035	0.098	0.015	0.052	-0.011	0.047	-0.043	-0.064	1	-0.072	0.029	-0.015	-0.004	0.115	0.028	0.063	0.053	0.084
lead_time	-0.034	-0.130	0.215	-0.010	0.086	0.024	-0.073	-0.047	-0.026	-0.037	0.004	0.283	-0.155	-0.186	-0.049	-0.085	0.104	-0.044	0.052	0.160	-0.072	1	0.162	0.128	0.001	-0.143	-0.052	-0.077	-0.069	-0.099
arrival_year	0.013	0.119	-0.170	0.005	-0.096	-0.025	0.107	0.005	0.029	0.005	0.005	-0.153	0.174	-0.043	0.026	-0.055	0.070	0.048	0.062	0.034	0.029	0.162	1	-0.355	0.015	-0.032	0.003	0.027	0.180	0.059
arrival_month	-0.019	-0.016	0.049	-0.005	0.011	0.009	-0.023	0.013	0.011	-0.006	0.009	0.029	-0.020	-0.022	-0.001	0.018	0.014	0.001	-0.029	0.032	-0.015	0.128	-0.355	1	-0.032	0.009	-0.044	0.004	0.054	0.103
arrival_date	-0.020	-0.005	0.036	0.006	-0.034	0.014	0.025	0.014	0.006	0.001	-0.003	-0.009	0.002	0.013	0.013	-0.007	0.011	0.022	0.019	-0.003	-0.004	0.001	0.015	-0.032	1	-0.022	-0.007	0.002	0.016	0.016
repeated_guest	0.074	-0.052	-0.046	-0.002	0.036	-0.018	-0.038	-0.014	0.020	0.032	-0.002	-0.065	-0.179	0.416	0.030	0.174	-0.193	-0.036	-0.060	-0.090	0.115	-0.143	-0.032	0.009	-0.022	1	0.397	0.509	-0.162	-0.022
no_of_previous_cancellations	0.021	-0.010	-0.019	-0.001	0.013	-0.009	-0.012	-0.006	-0.005	0.024	-0.001	-0.021	-0.029	0.081	-0.004	0.051	-0.045	-0.017	-0.015	-0.015	0.028	-0.052	0.003	-0.044	-0.007	0.397	1	0.499	-0.059	0.002
no_of_previous_bookings_not_canceled	0.042	-0.030	-0.025	-0.001	0.012	-0.010	-0.018	-0.009	0.034	0.023	-0.001	-0.052	-0.098	0.266	0.002	0.088	-0.118	-0.020	-0.015	-0.023	0.063	-0.077	0.027	0.004	0.002	0.509	0.499	1	-0.097	0.019
avg_price_per_room	0.002	-0.086	0.107	-0.031	-0.386	-0.056	0.274	0.361	0.037	0.075	0.008	-0.215	0.339	-0.152	-0.002	-0.295	0.290	0.330	-0.005	0.013	0.053	-0.069	0.180	0.054	0.016	-0.162	-0.059	-0.097	1	0.183
no_of_special_requests	0.017	0.053	-0.092	0.005	-0.146	0.027	0.124	0.060	-0.017	0.042	-0.008	-0.345	0.386	-0.128	-0.043	0.033	0.191	0.116	0.079	0.052	0.084	-0.099	0.059	0.103	0.016	-0.022	0.002	0.019	0.183	1

Appendix: Decision Tree Pruned Collapsed Binominal Values

Hotel Booking Cancellation Prediction

When considering the individual values in the binominal attributes, the weights are split among them and therefor each individual value option holds just the weight of its count. Considering the grouping as a total could be meaningful as they would give the groupings individual weights against the total summed creating a category weight against the total. That means that ***Length of stay*** (number of week nights + weekend nights of each booking) or ***Family Size*** (number of children + adults per booking) now have larger weights against the total and move up the scale as meaningful.

Attribute	weight
lead_time	0.2613
avg_price_per_room	0.1769
Length of Stay	0.1678
arrival_year	0.0412
Family Size	0.0371
type_of_meal_plan = Meal Plan 1	0.0326
market_segment_type = Corporate	0.0214
no_of_special_requests	0.0185

Appendix: The word of Model Comparison

Hotel Booking Cancellation Prediction

These excel spreads were developed to drive the comparison grid of the various models.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

DT Pruned	Accuracy	Weighted Recall	Weighted Precision
	86.88	84.20	85.58
CONFUSION Matrix	true Canceled	true Not_Canceled	class precision
pred. Canceled	454	98	82.25%
pred. Not_Canceled	140	1122	88.91%
class recall	76.43%	91.97%	

Recall	0.76
Precision	0.82
F1	0.79

FR Pruned	Accuracy	Weighted Recall	Weighted Precision
	72.38	57.83	85.44
CONFUSION Matrix	true Canceled	true Not_Canceled	class precision
pred. Canceled	93	0	100.00%
pred. Not_Canceled	501	1220	70.89%
class recall	15.66%	100.00%	

Recall	0.16
Precision	1.00
F1	0.27

Decision Tree	Accuracy	Weighted Recall	Weighted Precision
	86.44	84.17	84.79
CONFUSION Matrix	true Canceled	true Not_Canceled	class precision
pred. Canceled	461	113	80.31%
pred. Not_Canceled	133	1107	89.27%
class recall	77.61%	90.74%	

Recall	0.78
Precision	0.80
F1	0.79

Random Forest	Accuracy	Weighted Recall	Weighted Precision
	85.12	79.69	85.81
CONFUSION Matrix	true Canceled	true Not_Canceled	class precision
pred. Canceled	380	56	87.16%
pred. Not_Canceled	214	1164	84.47%
class recall	63.97%	95.41%	

Recall	0.64
Precision	0.87
F1	0.74

Appendix: Recommended Future Analysis

Hotel Booking Cancellation Prediction

Future Analysis is warranted for this effort. There appears to be a significant variance between arrival years. This is a good place to drive deeper analysis and try to understand what in the underlying phenomena is causing the variance.

Initial ideas of variable significance did prove out on some level and left more analysis to be conducted on others. Capturing simple information like the following in the customer reservation process would add clarity to the data and offer guidance on policy. Questions like the following would be of immense value.

1. Vacation or business travel
2. Reoccurring Visits to this a city
3. Willing to take a raincheck if overbooking causes a room shortage
4. What other hotel chains do you frequent (explore competition and for sister hotels)

In addition, merging in more of the detailed data the organization might have around the following would be very valuable to add clarity to our principal component analysis. Data including:

1. If booking had a promotion attached to it (transactional register data – GL)
2. If meal choice was made online or at the counter (transactional data – GL)
3. Weekday of Check In and Check Out (actual data from bookings and not derived)
4. Customer Home City and Booking City (from booking information)

Thank you