Project Presentation

Hotel Cancelations Predictions
No Code AI & Machine Learning

10/04/2025

Amit Mangotra

Contents / Agenda

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

Executive Summary

- **Key Factors influencing cancellations**
  - ❑ **Lead Time, Avg Room Price, Arrival Date, Arrival Month, Length of Stay and No of Special Requests** are the key features influencing cancellations. Higher the lead time, higher the avg room price and longer the stay duration, higher are the chances of cancellations. Winter months have lower cancellation rates while guests with more special requests have a lower cancellation rate. Other factors that influence cancelations are *market segment, no of adults & room type*. *Repeat Guests* are not an important factor currently, because of lower number of repeat guests, but hotel should develop loyalty programs as repeat guests have significantly lower cancelation rates.
- **Recommendations for reducing cancelations**
  - ❑ Open reservations only after a certain lead-time threshold e.g. 200 days or request a non-refundable deposit upfront for lead time over 200 days. Similarly, request a non-refundable deposit for high average room rate reservations and longer duration stays. This deposit amount can be a percentage of total booking amount. The average room rate and lead time thresholds can be set by looking at EDA.
  - ❑ Loyalty programs can have two targets: customer profiles with low cancelations rates and the rest. The low cancelation segments include single guests, corporate and aviation guests, guests with special requests and meal plan requests and guests arriving in winter months.
  - ❑ Develop business segment, which include corporates and aviation, as they have low cancellation rates.

Executive Summary

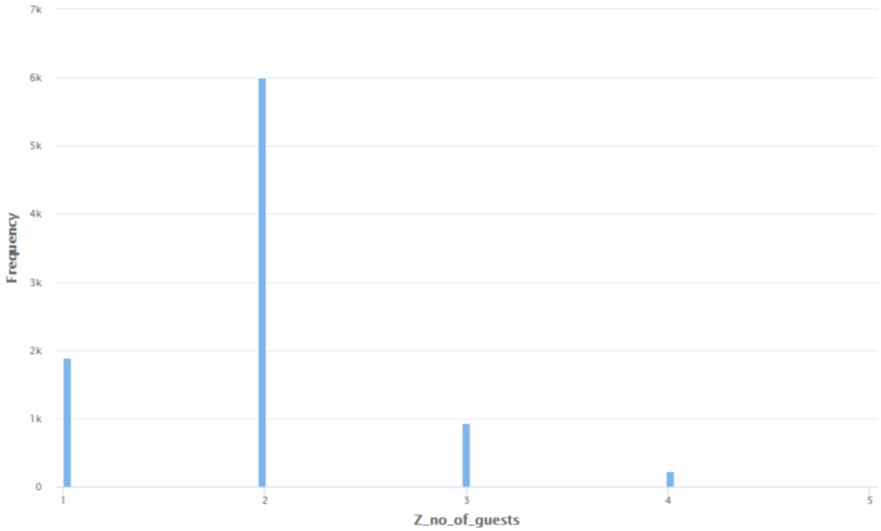
- **Recommendations based on predicted cancelations**
  - ❑ Create a two-tier refund amount based on cancellation prediction. For predicted cancellations, the refund amount would be lower. This would require building a real time system that predicts cancellations while the customer is booking, be it online or offline.
  - ❑ Allow overbooking for room types with predicted cancellations. The overbooking limit can be a conservative number so that no customers are turned away . The high booking-high cancellation months of August, September and October can generate important revenues using overbookings.
  - ❑ By allocating daily wage or seasonal staff based on cancelation predictions, these variable costs can be reduced leading to higher profits.
- **Fine tune data for developing future models**
  - ❑ If *online* and *offline market segments* refer to retail segment only and do not include *corporate* and *aviation* segments, then these can be renamed. Further, by adding *cancellation lead time*, one can identify cancellations that happen at short notice and put in place suitable measures. Finally, there can be a rule setting that ensures that *avg room price* is zero only for complimentary segment. Currently it is zero also for some bookings from online *market segment*
  - ❑ The impact of *arrival data* has needs to be probed. Including a *dayname* feature would help with this analysis.

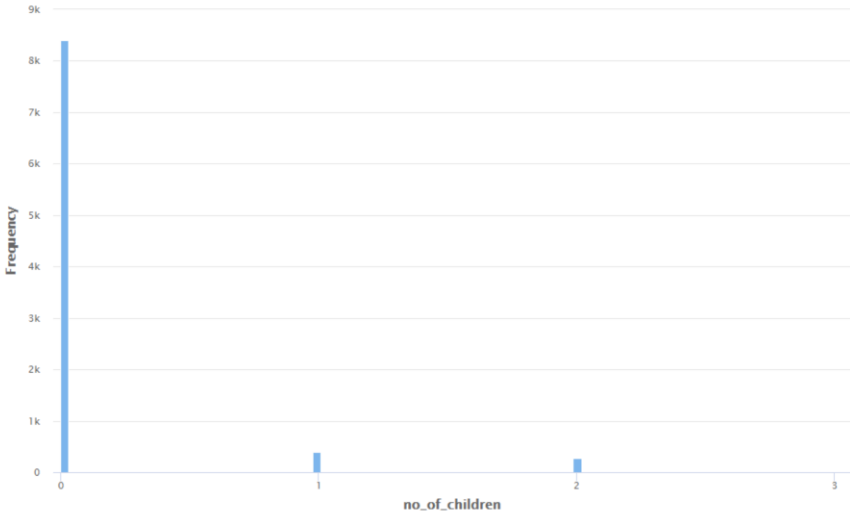
Business Problem Overview and Solution Approach

- The Hotel would like to predict cancelations and build profitable policies for cancellations and refunds
- Solution approach / methodology
  - ❑ EDA
  - ❑ Data Preprocessing
  - ❑ Model Building using Decision Trees, Random Forests
  - ❑ Model Performance Evaluation
  - ❑ Insights and Recommendations

EDA Results – Univariate Analysis

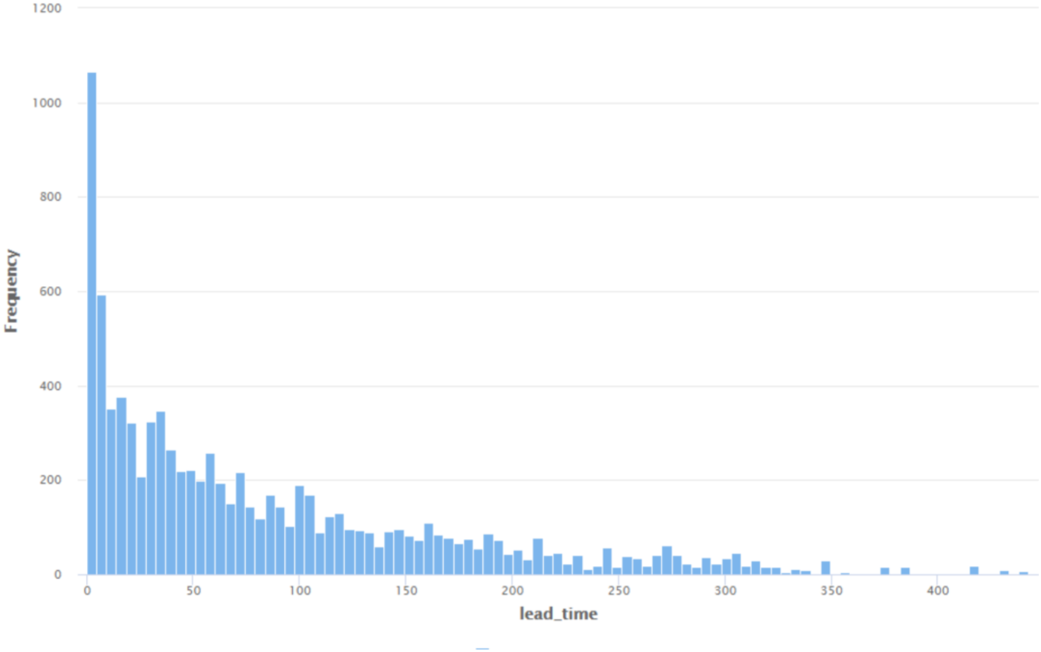
- **Type of Guests:** Mostly Adults guests





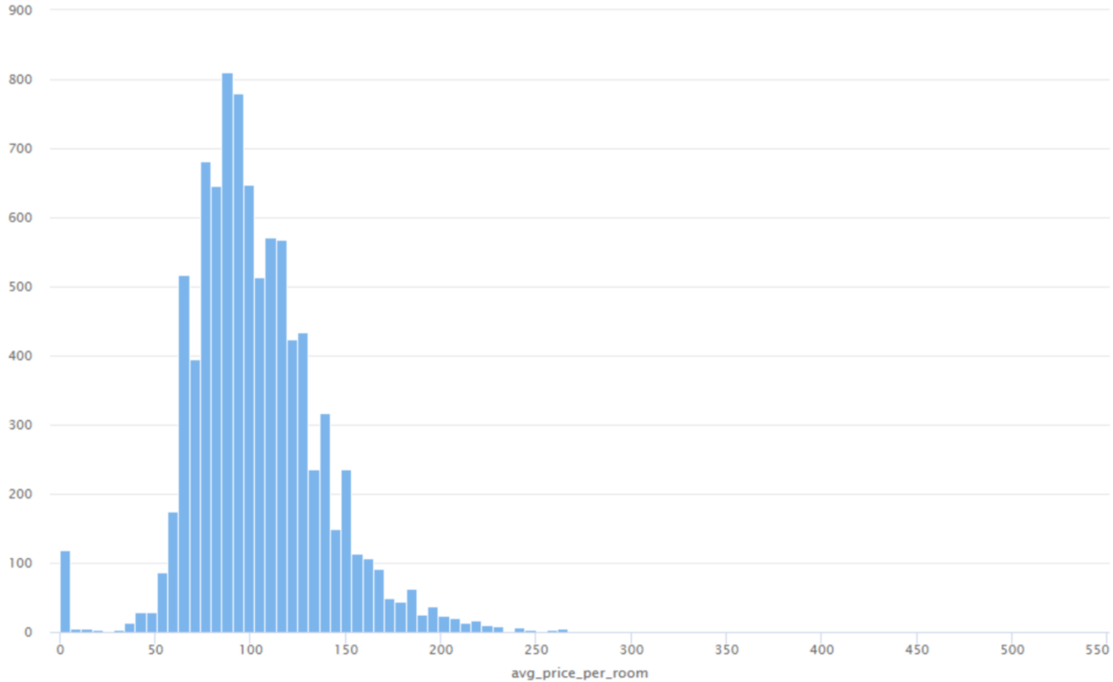
EDA Results – Univariate Analysis

- **Lead Time:** ~11% reservations have a lead time <= 4 days; 25% reservations happen within a lead time of 15 days. But bookings are happening even 200 days in advance. We will later see if there is a cancelation trend for such high lead time bookings



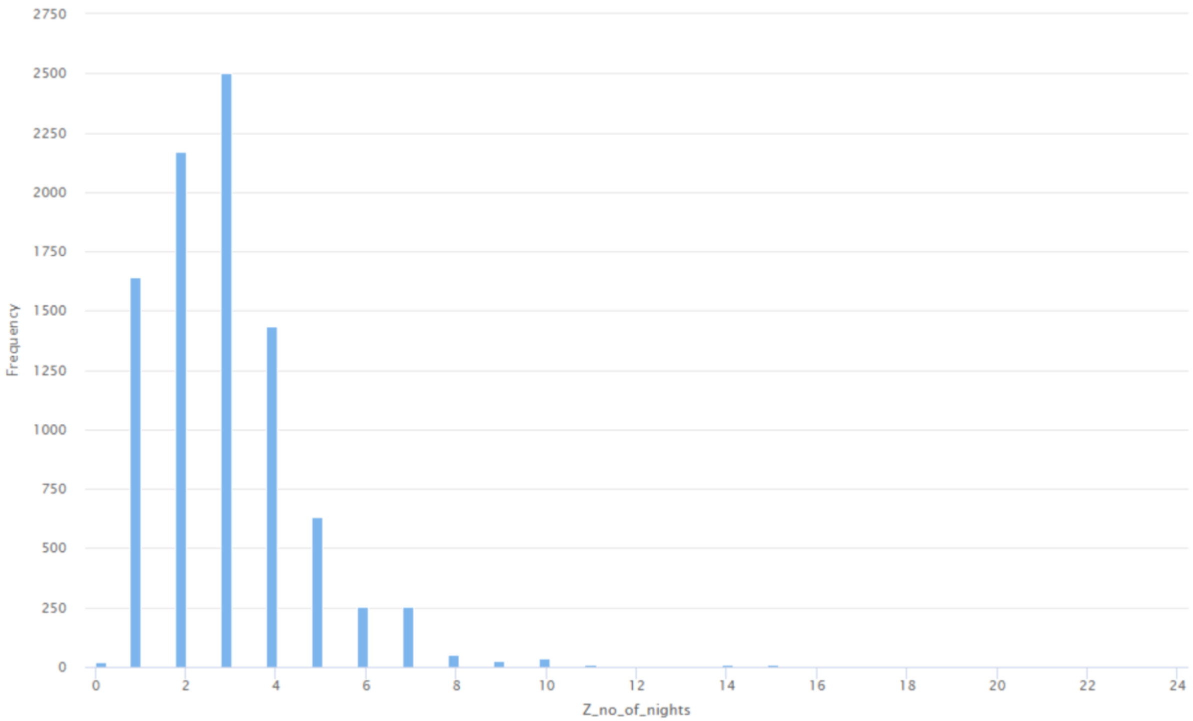
EDA Results – Univariate Analysis

- **Avg Room Price:** It is normally distributed with a mean of around 100 and right skewed, indicating a good revenue management practice. However, one must check if higher priced rooms have higher cancelations.



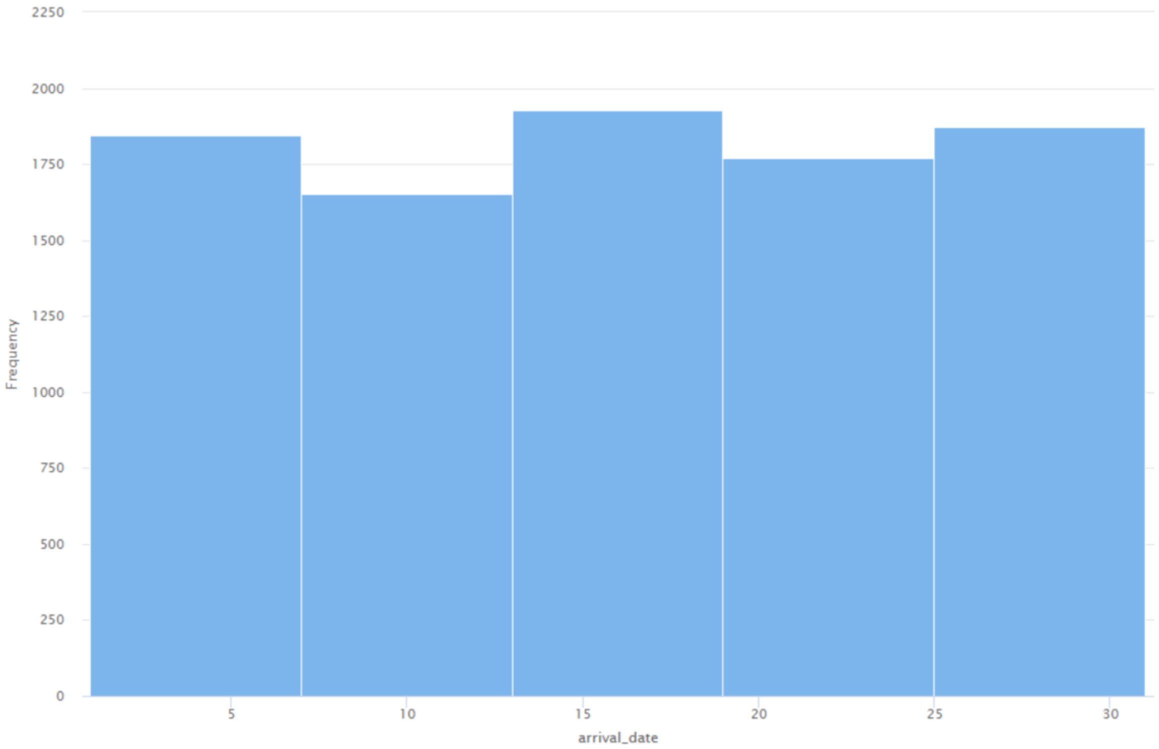
EDA Results – Univariate Analysis

- **Room Nights:** Most guest spend 1-4 nights but there are cases of 10+ nights



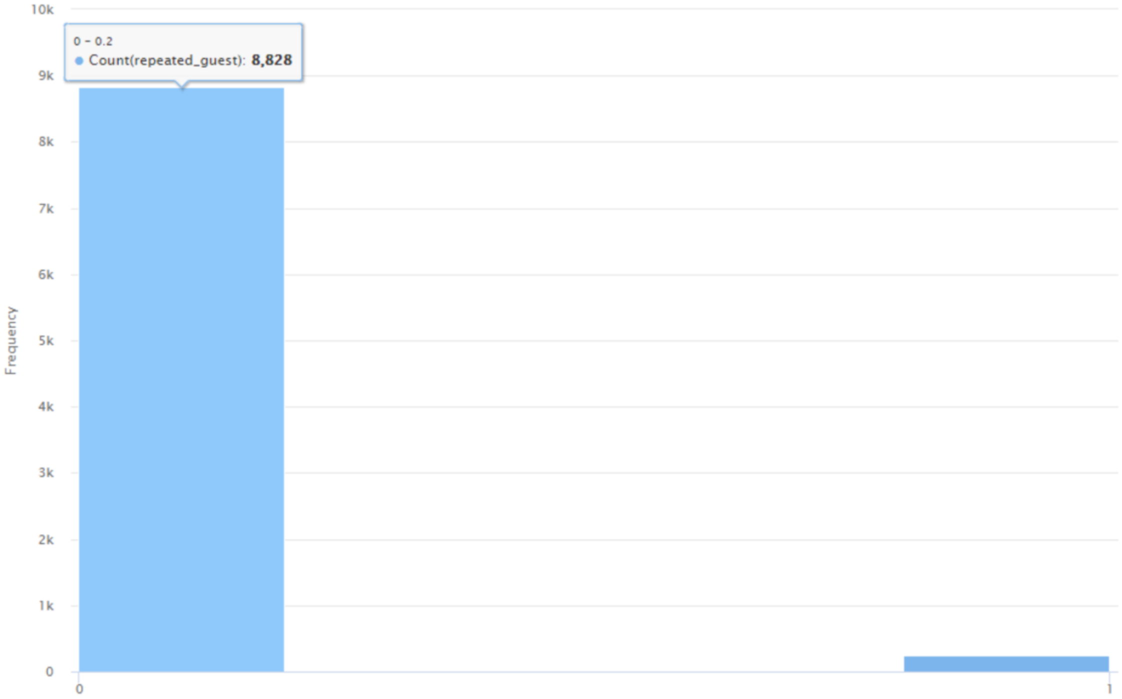
EDA Results – Univariate Analysis

- **Arrival Date:** more bookings in alternate weeks during a month



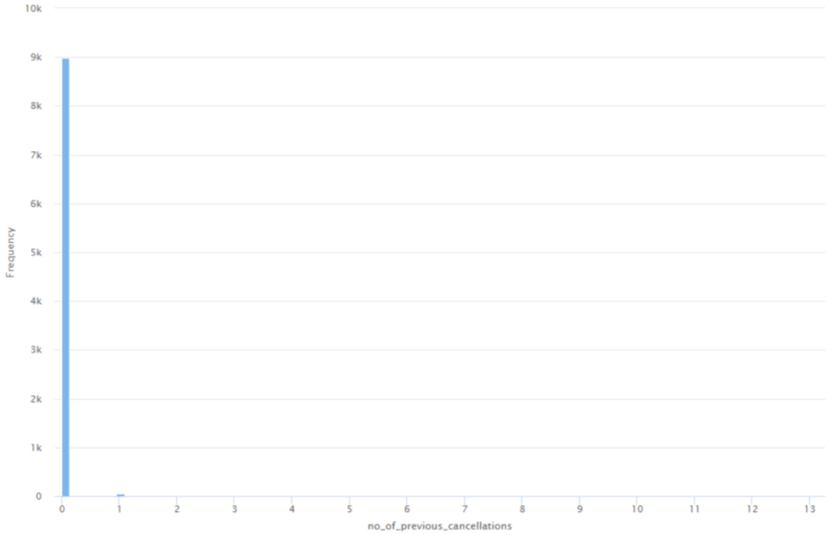
EDA Results – Univariate Analysis

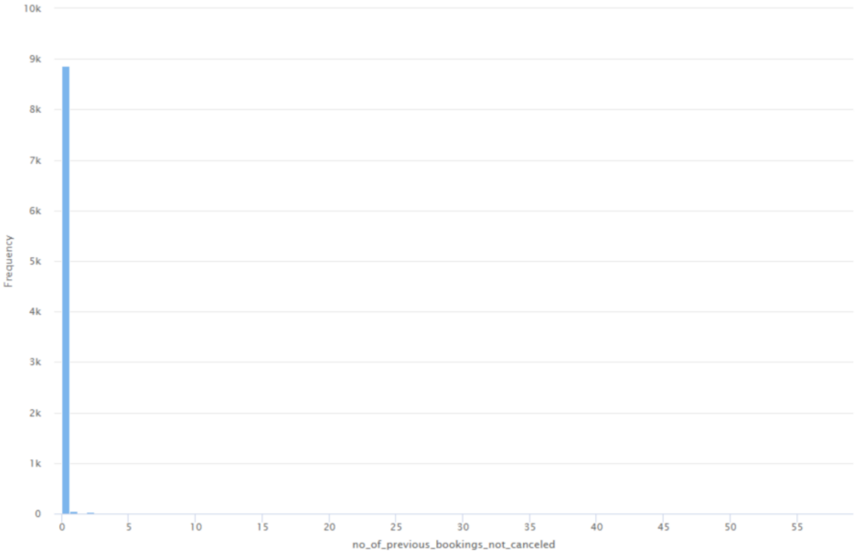
- **Repeat Guests:** Very low number of repeat guests. The hotel is recommended to develop a loyalty program.



EDA Results – Univariate Analysis

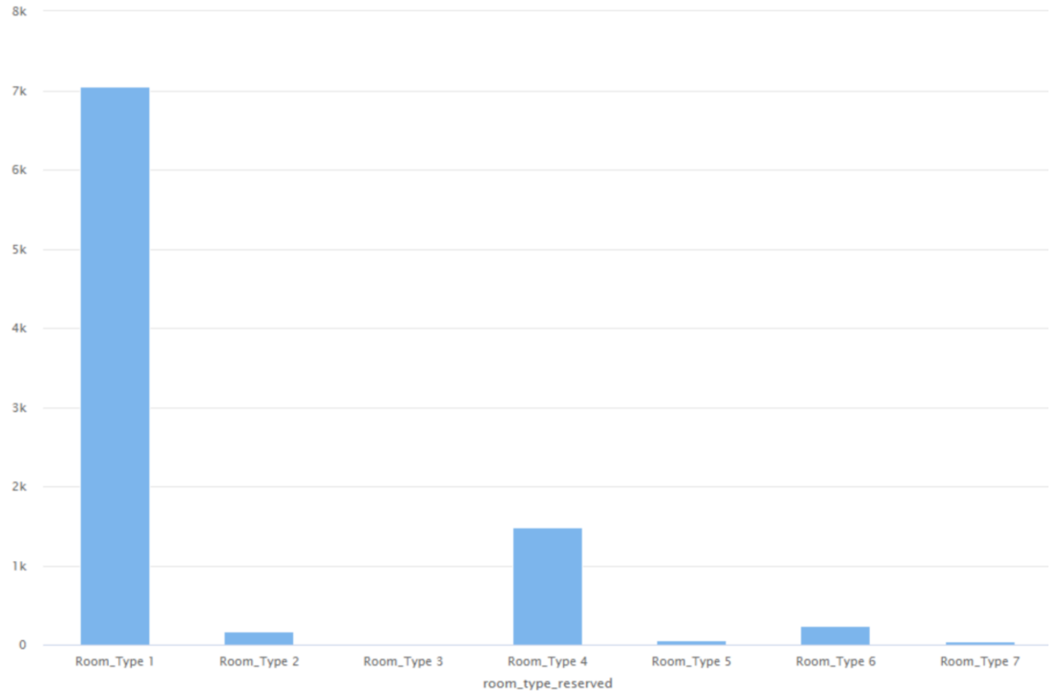
- **Cancellation History:** Since there are mostly first-time customers, of course they have no cancelation history. However, if there are 33% cancellations, there is a serious problem for the hotel. It is not attracting repeat customers and also have high cancelation rates.





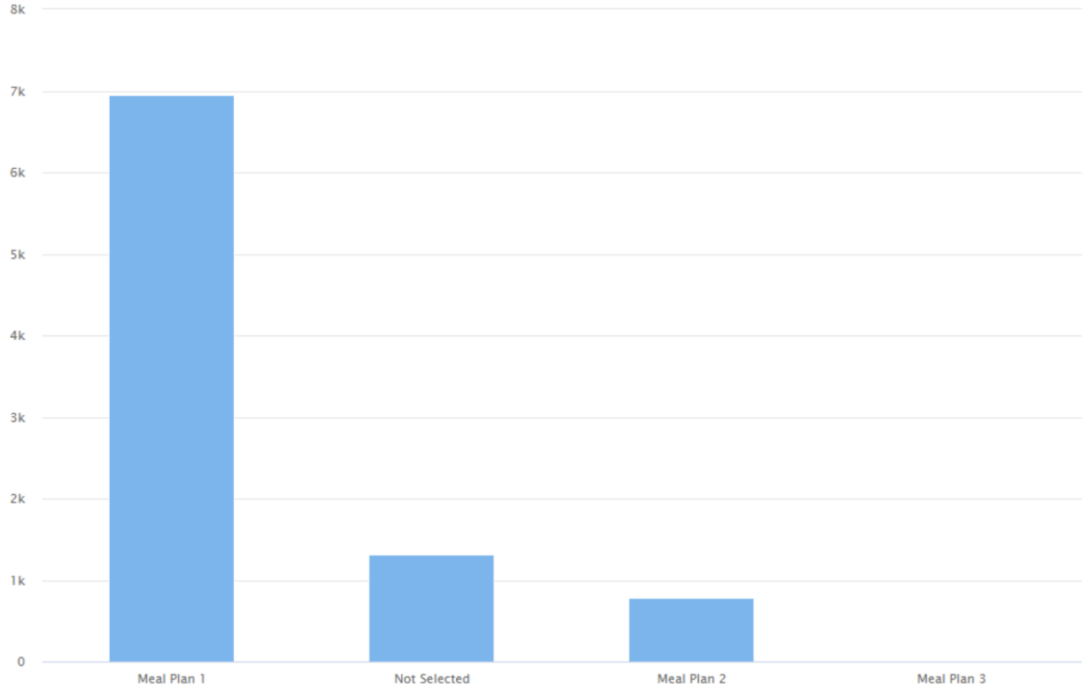
EDA Results – Univariate Analysis

- **Room Type:** Most reservations are for Room Type 1 while Room Type 4 has the next highest reservations. If there is overcapacity in these other room types, can the hotel convert these into room types 1 and 4 ?



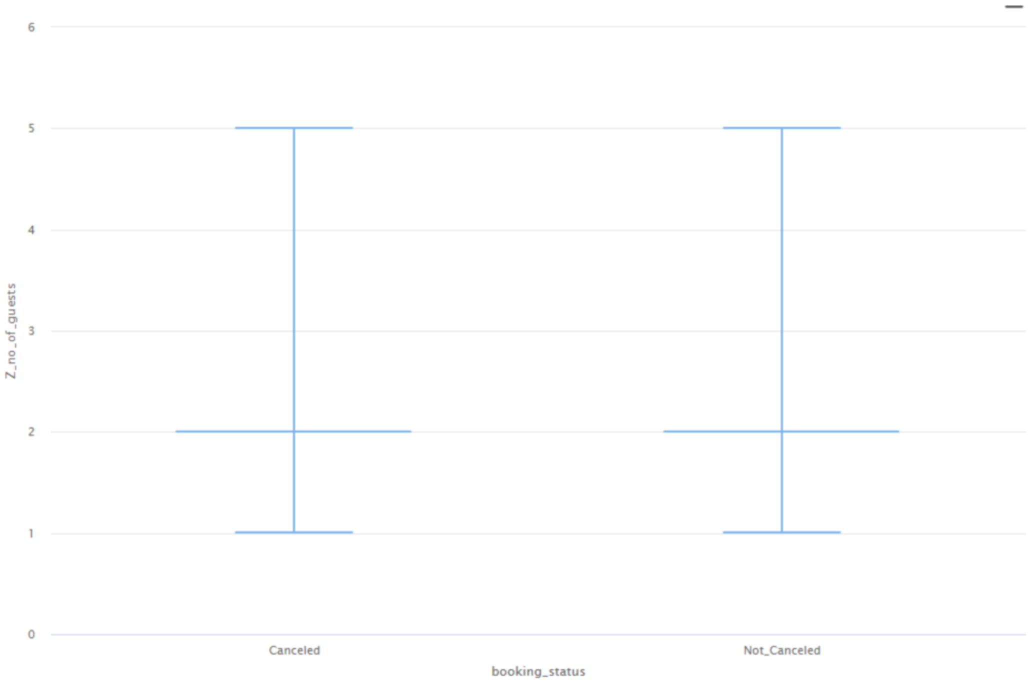
EDA Results – Univariate Analysis

- **Meal Plan:** Most reservations opt for plan 1 i.e. only including breakfast



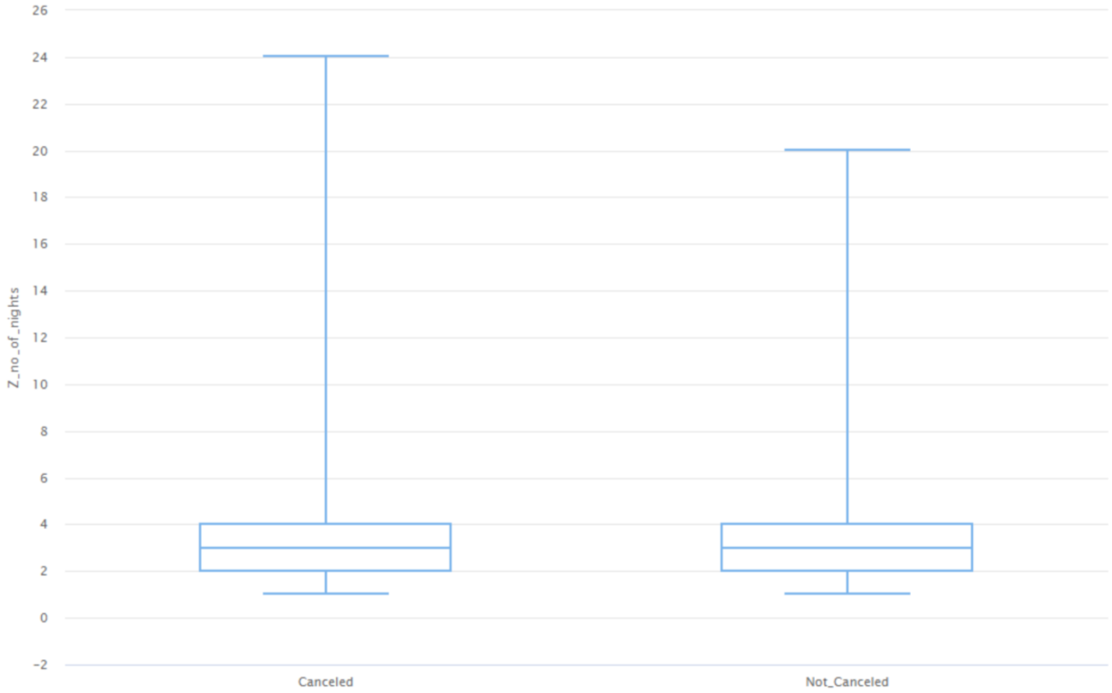
EDA Results – Bivariate Analysis

- **No of Guests:** Cancelations do not depend on number of guests



EDA Results – Bivariate Analysis

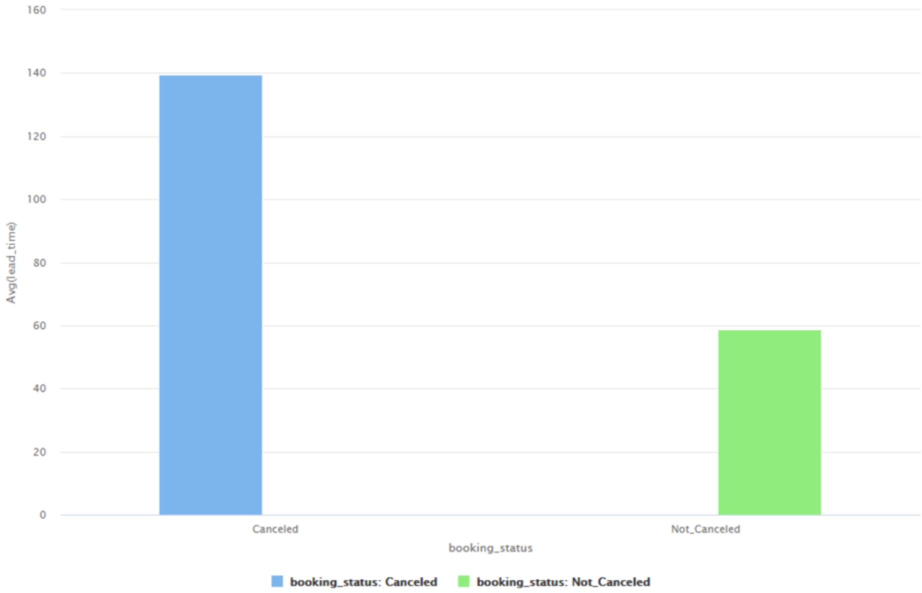
- **No of Nights:** 25% of cancelations are for bookings between 4-24 days.

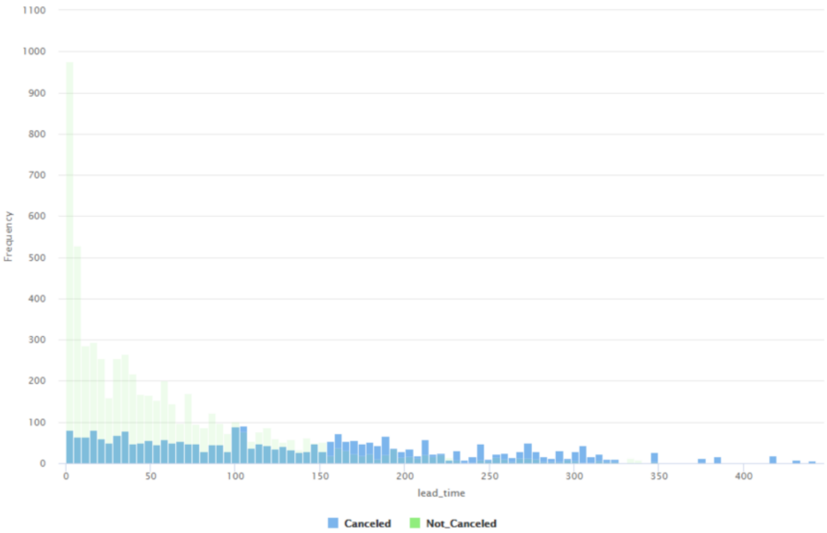


Proprietary Content @Amit Mangotra. Unauthorised use and distribution prohibited

EDA Results – Bivariate Analysis

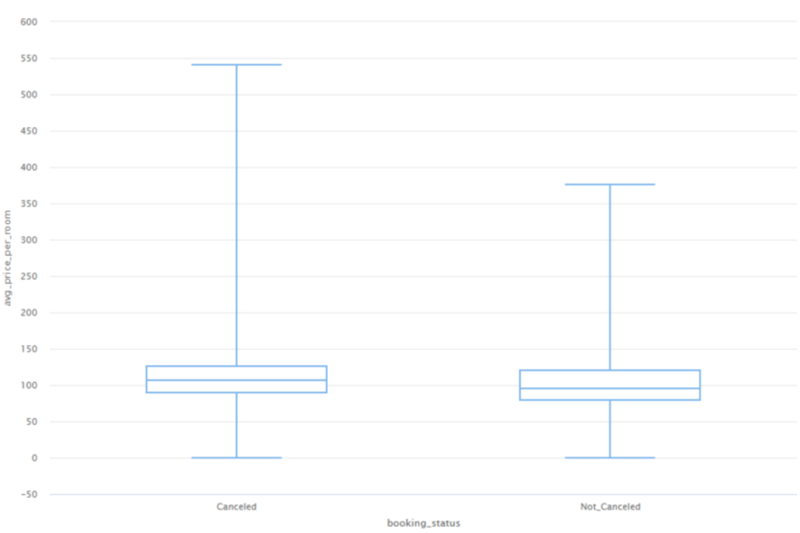
- **Lead Time:** Cancelations have a higher average lead time. Higher the lead time, more chances of cancelations. There are cancelled bookings with lead times of 150-400+ days!

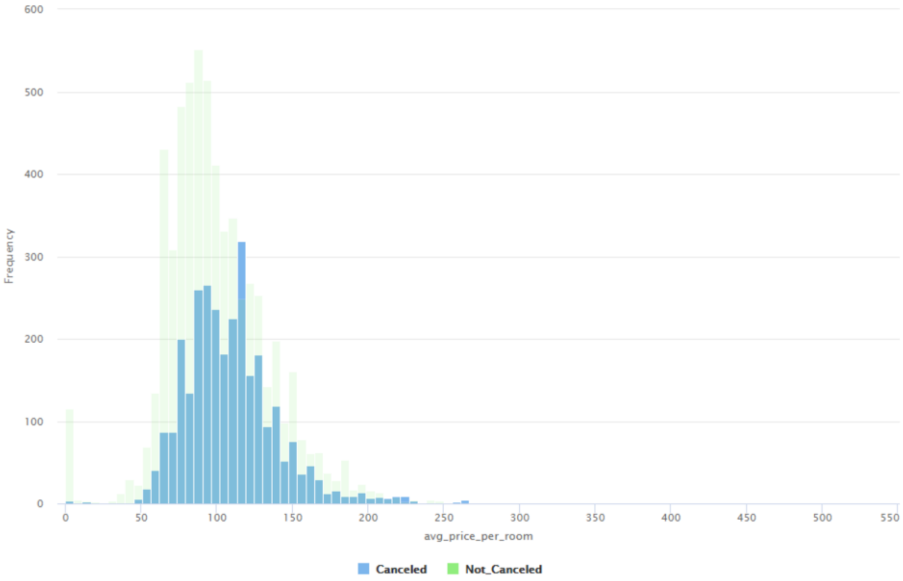




EDA Results – Bivariate Analysis

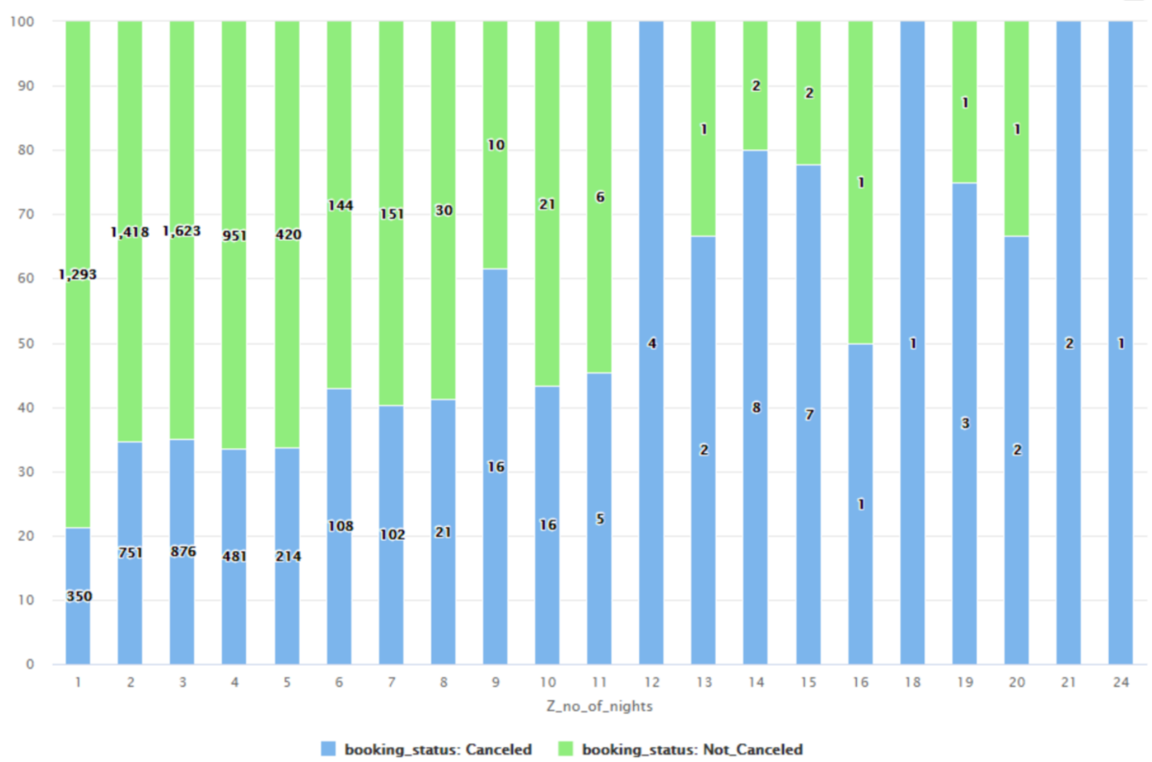
- **Avg Room Price:** Cancelations have a higher median avg room price and cancelation rates increases at higher prices





EDA Results – Bivariate Analysis

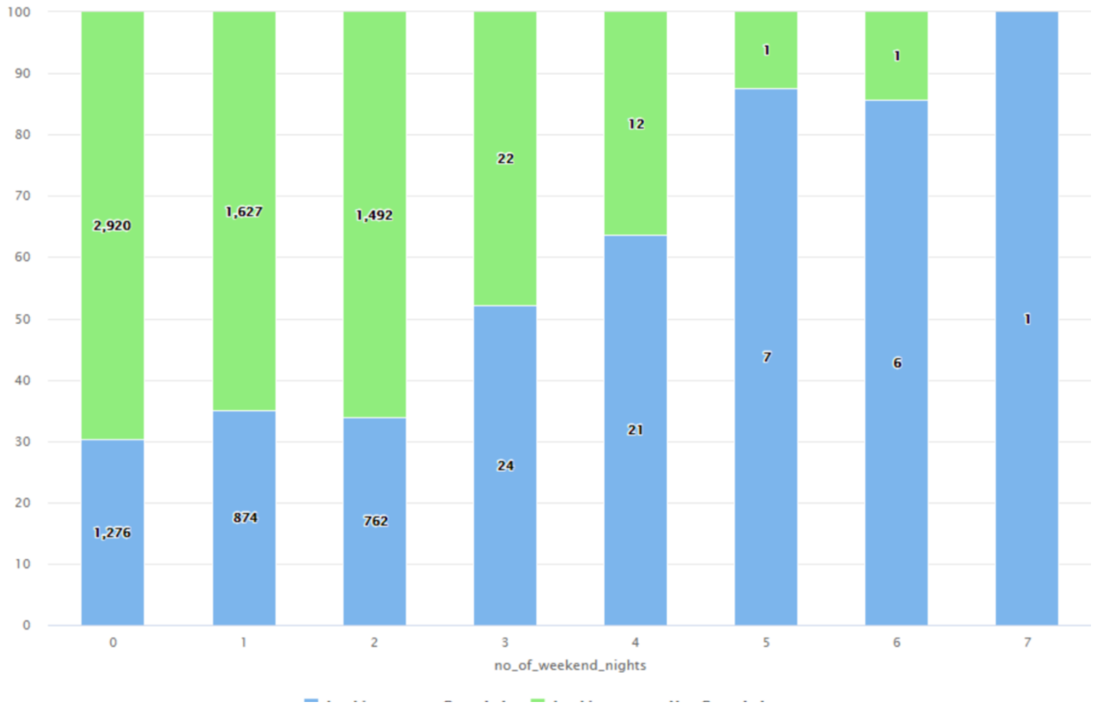
- **No of Nights:** % share of cancelations increases as the duration of stay increases



Proprietary Content @Amit Mangotra. Unauthorised use and distribution prohibited

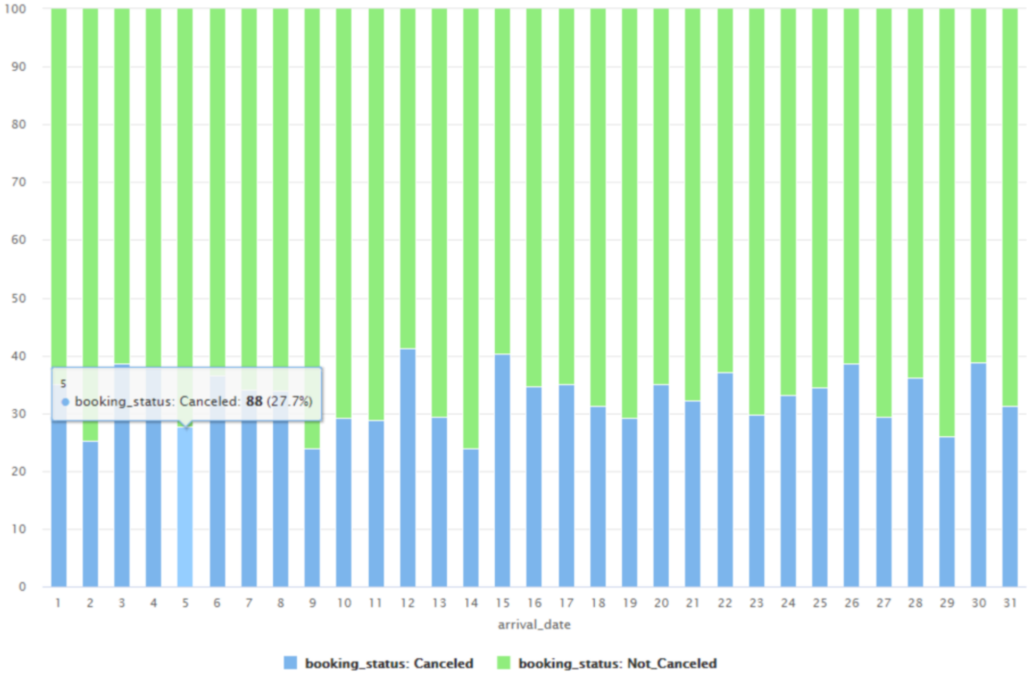
EDA Results – Bivariate Analysis

- **No of weekend nights:** % share of cancelations are low when staying upto 1 weekend and then increase. This is also indicated by increase in % share with increase in total stay duration



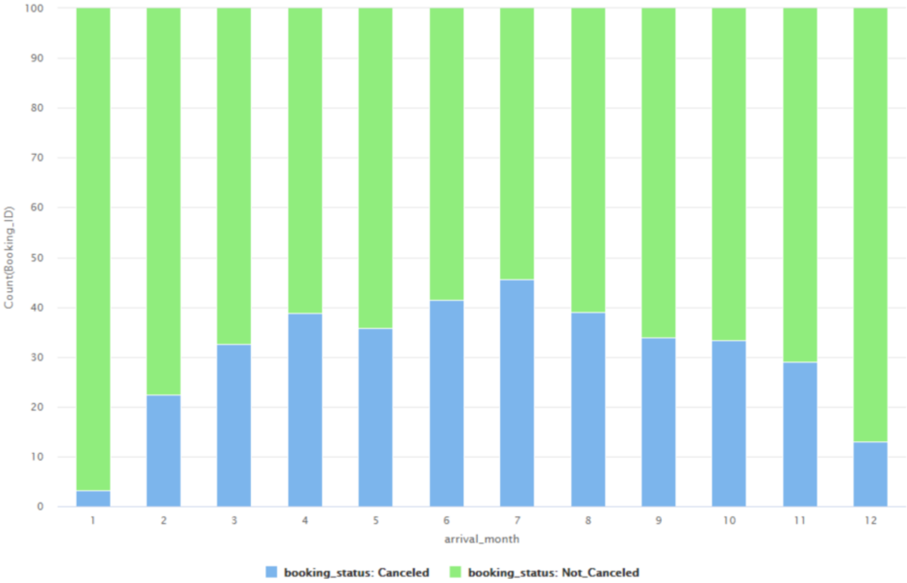
EDA Results – Bivariate Analysis

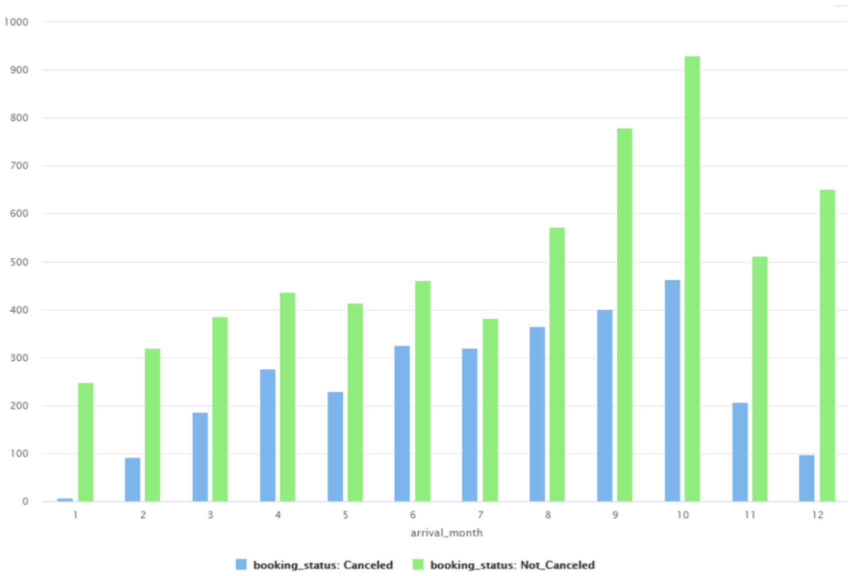
- **Arrival Date:** No apparent distinct pattern of cancelations but some dates do have higher cancelation rates. The model might be able to generate a pattern



EDA Results – Bivariate Analysis

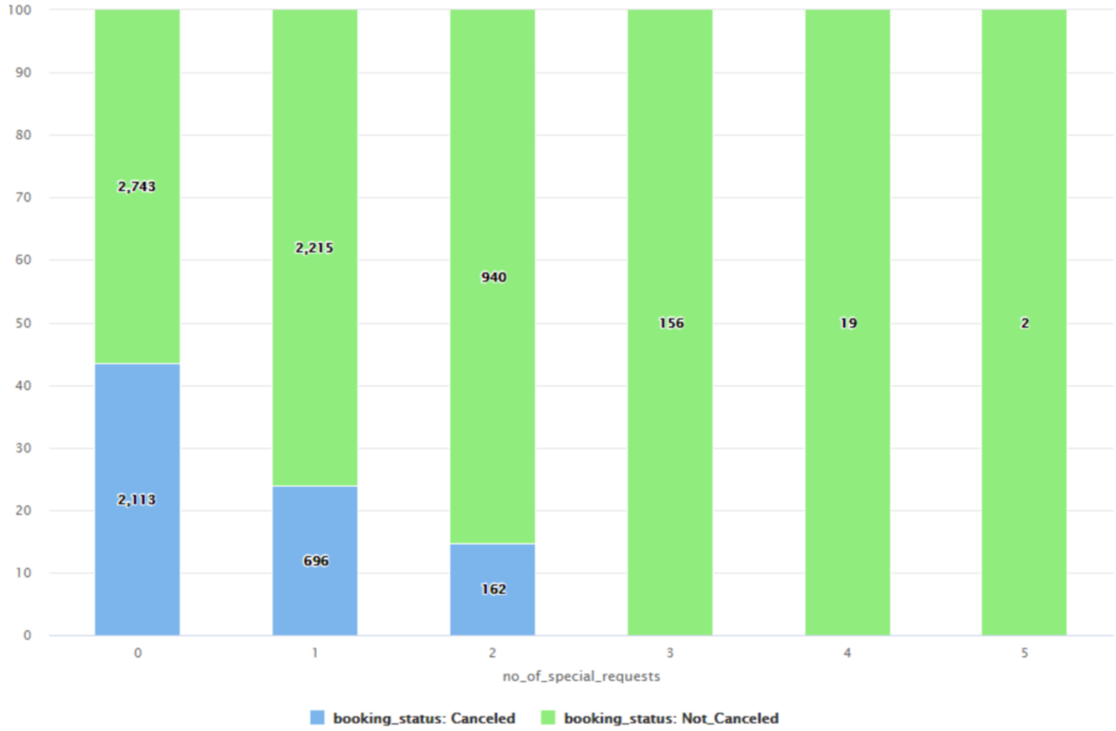
- **Arrival Month:** Around 40% cancelation rates in spring and summer months of April, June and July while lower rates in winter months of Dec and Jan. Sep and Oct have high booking and high cancellation numbers.





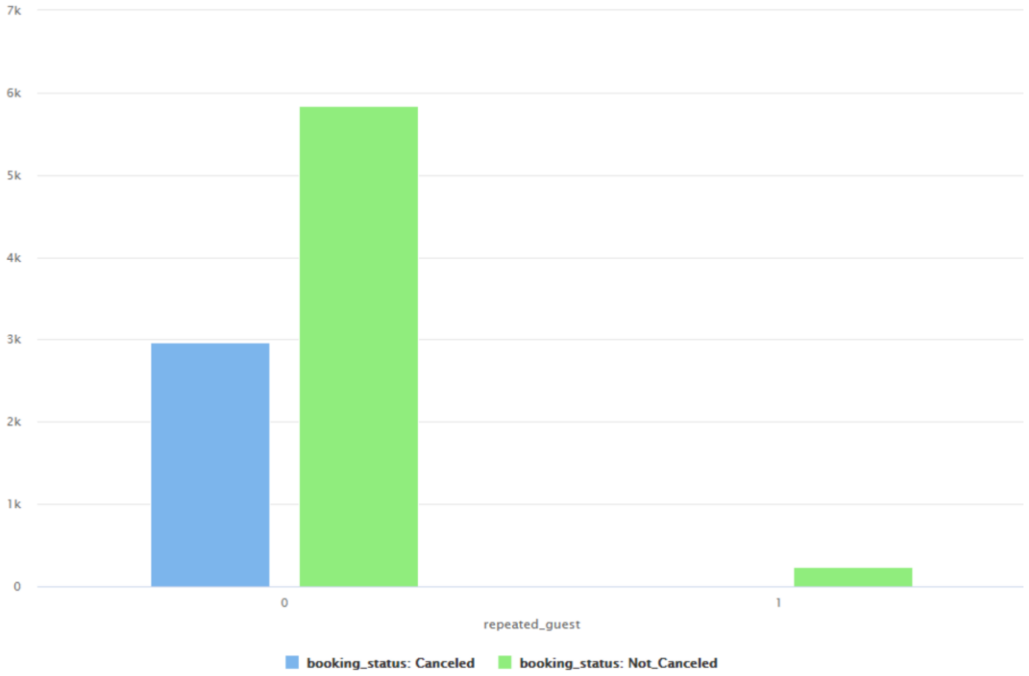
EDA Results – Bivariate Analysis

- **No of special requests:** Guests with special requests have a lower % share of cancelations



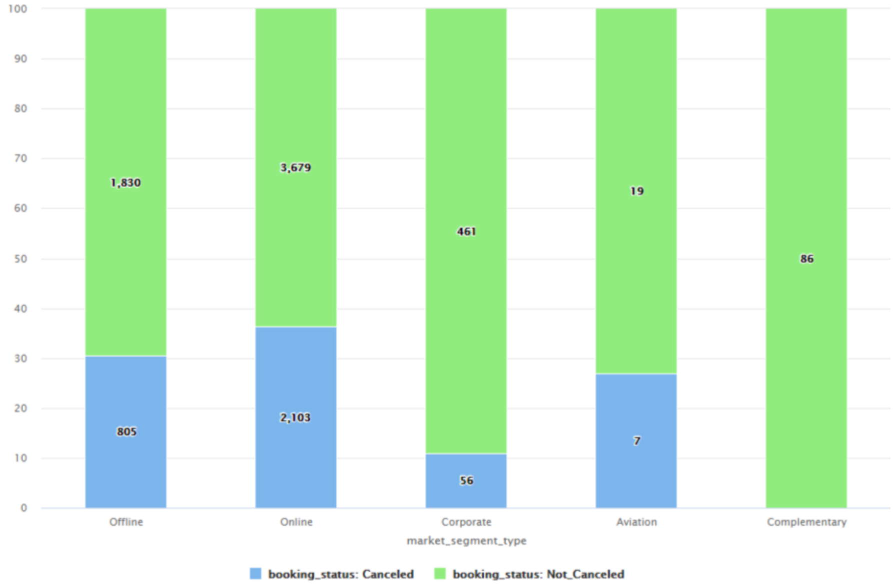
EDA Results – Bivariate Analysis

- **Repeat Guests:** While Repeat Guests have a lower number, they also have a very low cancelation rate, practically close to zero. A loyalty program is recommended to reduce cancelations



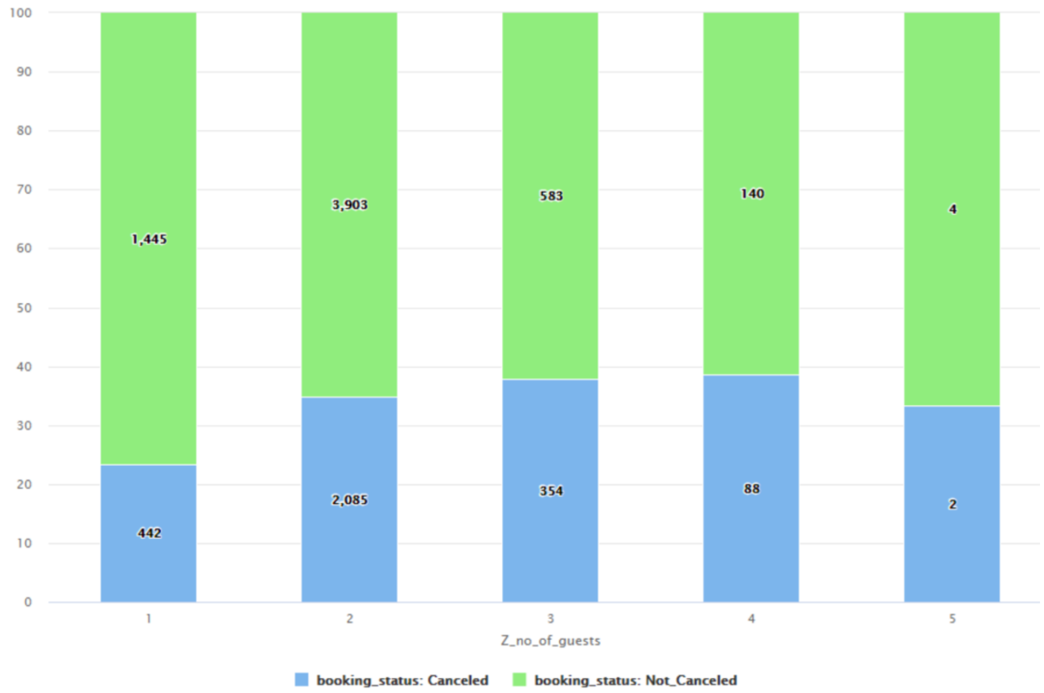
EDA Results – Bivariate Analysis

- **Market Segment Type:** Corporate segment has the lowest cancelation rate while Online has the highest. Complimentary segment do not need to pay and hence never cancel. Info is not provided on whether corporate, aviation and complementary segments have online or offline bookings.



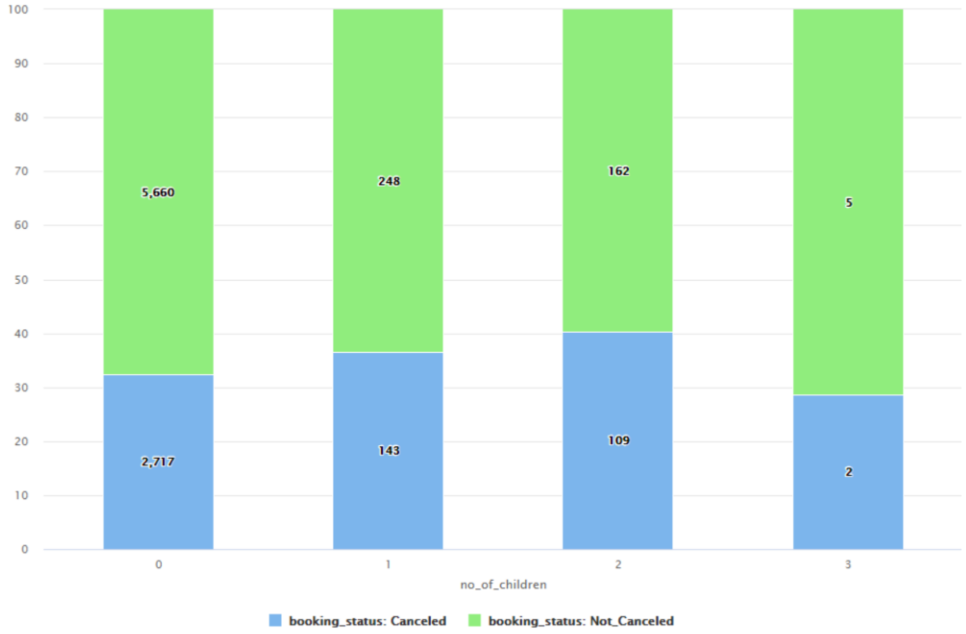
EDA Results – Bivariate Analysis

- **No Of Guests:** Both cancelation numbers and % share of cancelations are high for a 2-guests booking



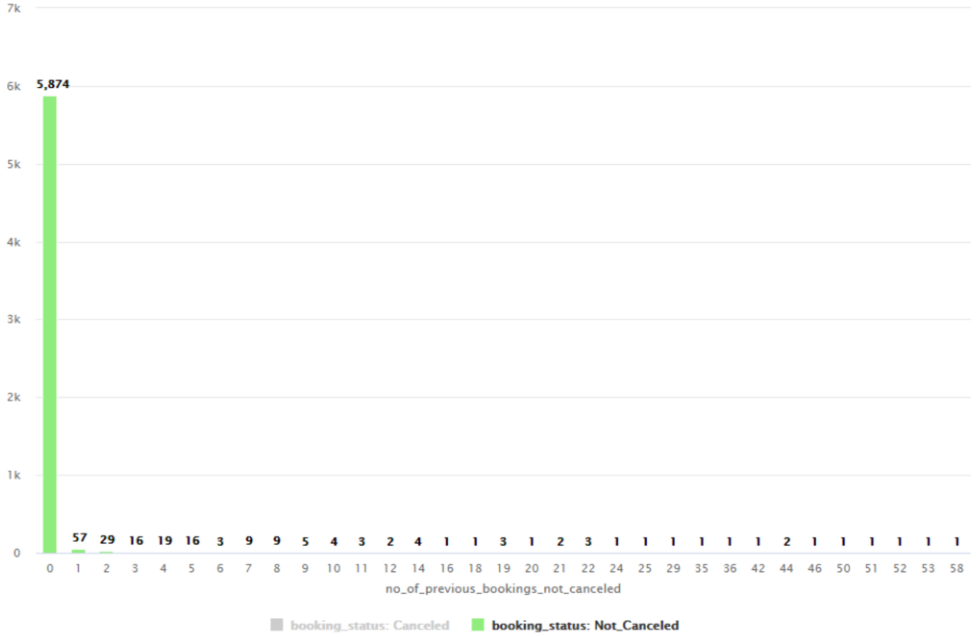
EDA Results – Bivariate Analysis

- **No of Children:** Guests with no children have a lower % share of cancelled bookings but have the highest number of cancelations (excluding the rare cases of bookings with three children)



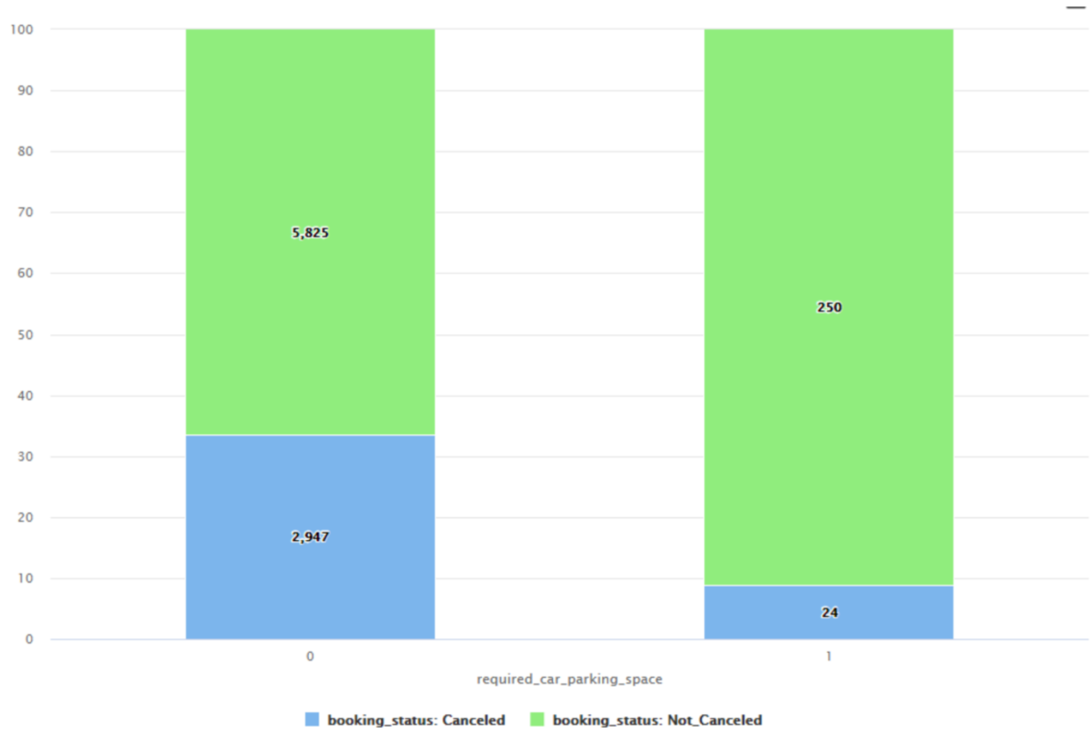
EDA Results – Bivariate Analysis

- **No of previous bookings not cancelled:** Although very few repeat customers, those who have not cancelled before have practically zero cancelation rates, once again reinforcing the need for repeat customers



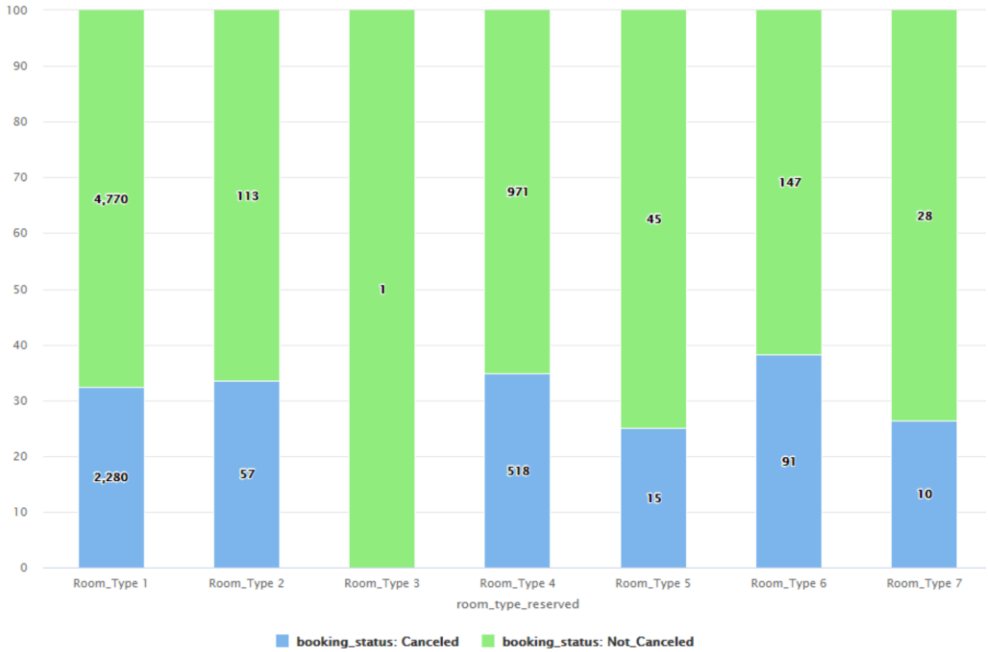
EDA Results – Bivariate Analysis

- **Required Car Parking Space:** Guests requesting car parking space are less likely to cancel



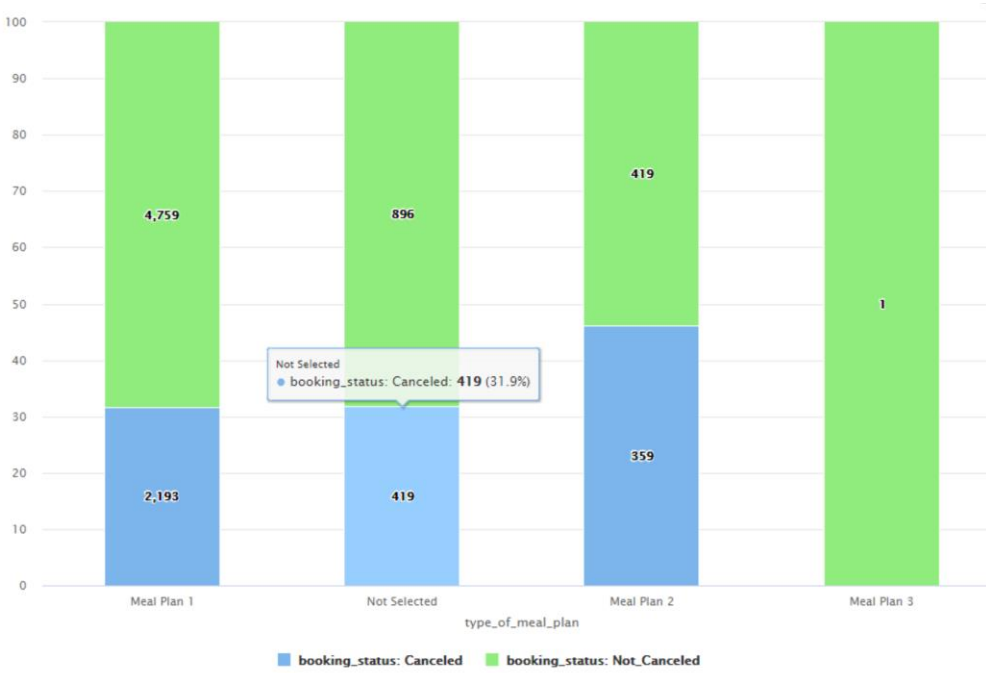
EDA Results – Bivariate Analysis

- **Room Type:** % share of cancelations are the highest for room type 6 and lowest for room type 5



EDA Results – Bivariate Analysis

- **Meal Plan:** Guest with no meal plan selection and plan 1 have a lower % share of cancelations



Data Preprocessing

- ***Data anomalies:***

  Remove 22 bookings with no of nights = 0.

  Remove booking with no of guests = 12 as highly unlikely to have 12 guests in 1 room

  Do nothing with room bookings with avg room price = 0 because these might be free bookings. However, no of nights = 0 is treated as anomaly because hotel must keep track of room nights even for free bookings.

  Remove one bookings with children = 10 as unlikely to have 10 children in 1 room

- ***Outliers:***

  Set lead time upper limit as 340 i.e. approximately 3 standard deviations

  No treatment for outlier number of room nights because this would also involve treating the two constituent attributes: no of weeknights and no of weekend nights

  Retain the high outlier avg prices such as 300+including 540 because high prices are part of revenue management in hotels and because there are both cancelations and no cancelations in 300+ price ranges

- ***Generate attributes:***

  No of guests = no of adults + no of children

  No of nights = no of weekday + no of weekend nights

- ***Remove attributes:***

  Booking ID removed after checking there are no duplicates

Model Performance Summary

- Recall is the most important parameter because hotel wants to minimize false cancelation predictions. Predicting that a booking won't be cancelled, while it is cancelled, will lead to loss of revenues

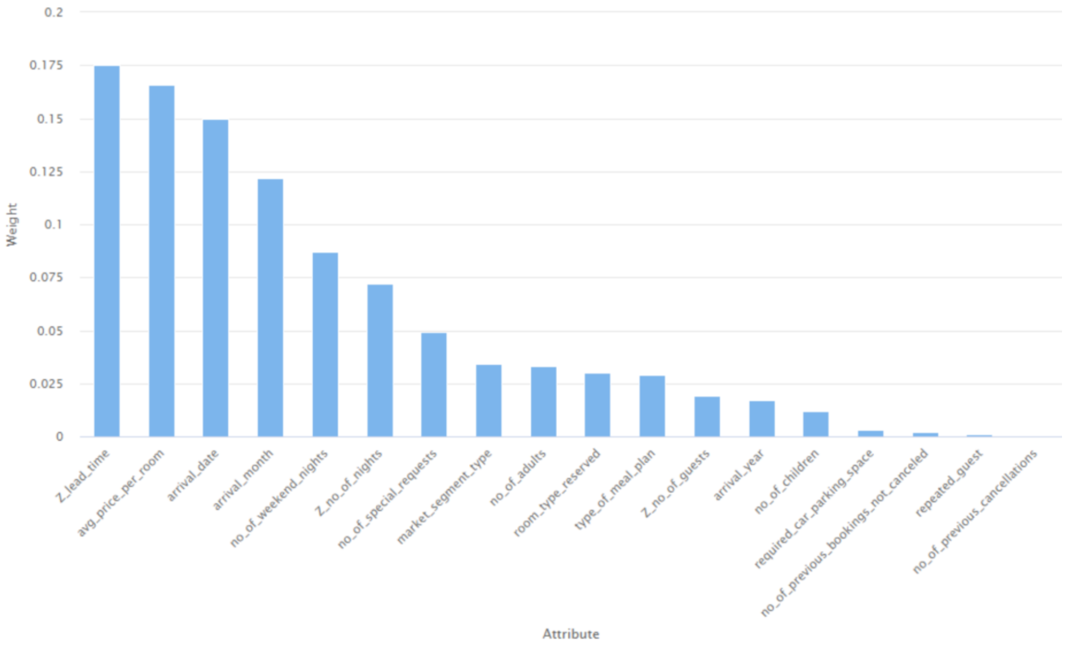
- Even though the Recall is lowest for the Random Forest model, it is the best model among the four models since it is the only model that is not overfitting.

- Model used Pruning, Gini Index, 100 trees and Tree Depth of 10

Model	Recall_Train	Recall_Test
Decision Tree	97.49%	73.84%
Decision Tree Pruned	99.42%	74.17%
Random Forest	98.94%	73.61%
Random Forest Pruned	75.93%	70.18%

Model Performance Summary

- **Feature importance:** *Lead Time, Avg Room Price, Arrival Date, Arrival Month, Length of Stay and No of Special Requests* are the key features influencing cancellations



APPENDIX

Data Background and Contents

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

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:

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)

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

Data Background and Contents

Data Dictionary

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)

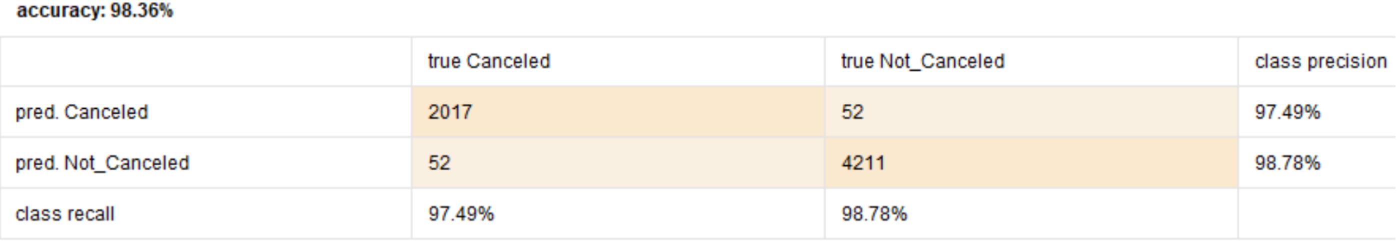
no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking_status: Flag indicating if the booking was canceled or not.

Model Building - Decision Tree / No Pruning

- Recall is the most important parameter because hotel wants to minimize false cancelation predictions
- Base Model using Gini Index and Max Tree Depth of 20
- Model overfits because Train Recall vs Test Recall differs by more than 10 pp (97.49% vs 73.84%)

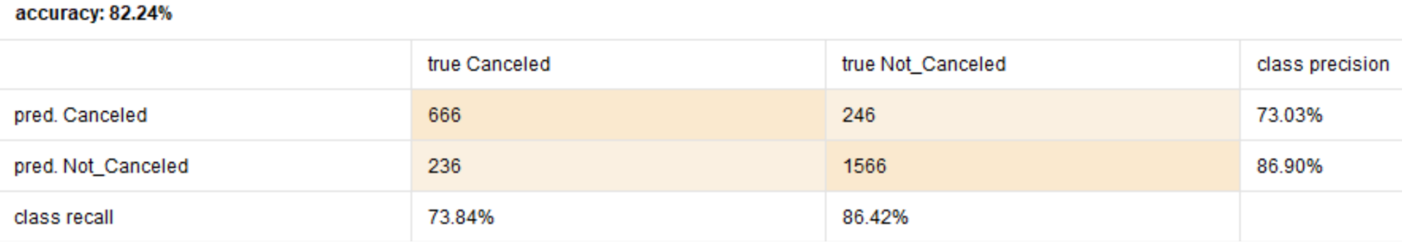
Train



accuracy: 98.36%

	true Canceled	true Not_Canceled	class precision
pred. Canceled	2017	52	97.49%
pred. Not_Canceled	52	4211	98.78%
class recall	97.49%	98.78%	

Test

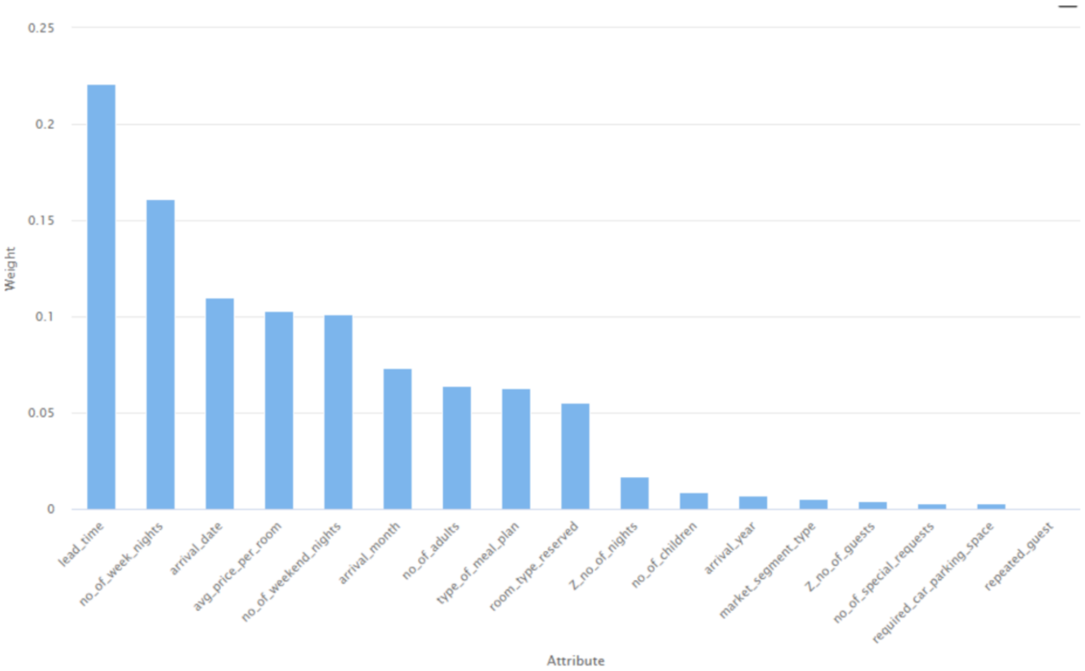


accuracy: 82.24%

	true Canceled	true Not_Canceled	class precision
pred. Canceled	666	246	73.03%
pred. Not_Canceled	236	1566	86.90%
class recall	73.84%	86.42%	

Model Building - Decision Tree / No Pruning

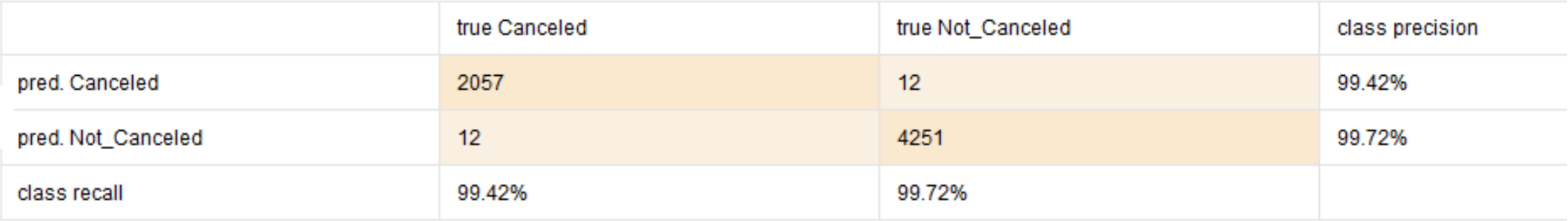
Lead Time, No of Nights, Arrival Date, Avg Price Per Room are the key features influencing cancellations. Other important features are ***Arrival Month, No of Adults, Type of Meal Plan and Room Type Reserved***



Model Building - Decision Tree / Pruned

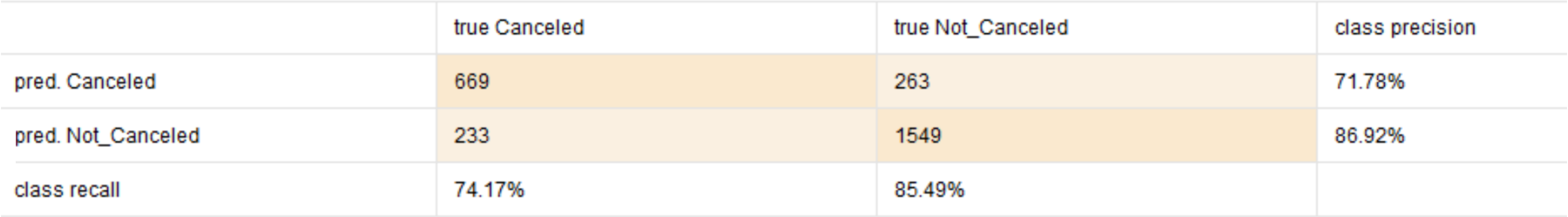
- Pruned Model is slightly better as recall as improved marginally from 73.84% to 74.17%. Model overfits because Train Recall vs Test Recall differs by more than 10 pp

Train



	true Canceled	true Not_Canceled	class precision
pred. Canceled	2057	12	99.42%
pred. Not_Canceled	12	4251	99.72%
class recall	99.42%	99.72%	

Test

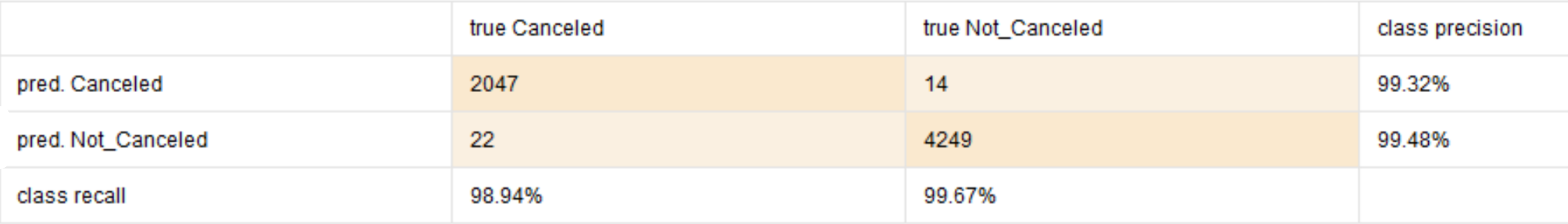


	true Canceled	true Not_Canceled	class precision
pred. Canceled	669	263	71.78%
pred. Not_Canceled	233	1549	86.92%
class recall	74.17%	85.49%	

Model Building – Random Forest/ No Pruning

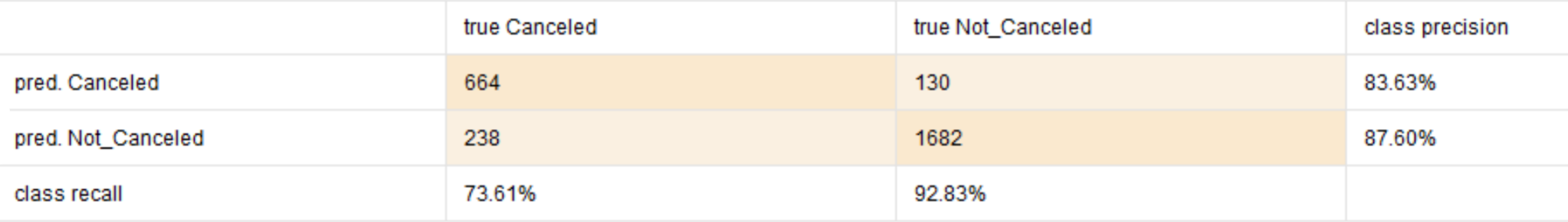
- Model overfits because Train Recall vs Test Recall differs by more than 10 pp

Train



	true Canceled	true Not_Canceled	class precision
pred. Canceled	2047	14	99.32%
pred. Not_Canceled	22	4249	99.48%
class recall	98.94%	99.67%	

Test

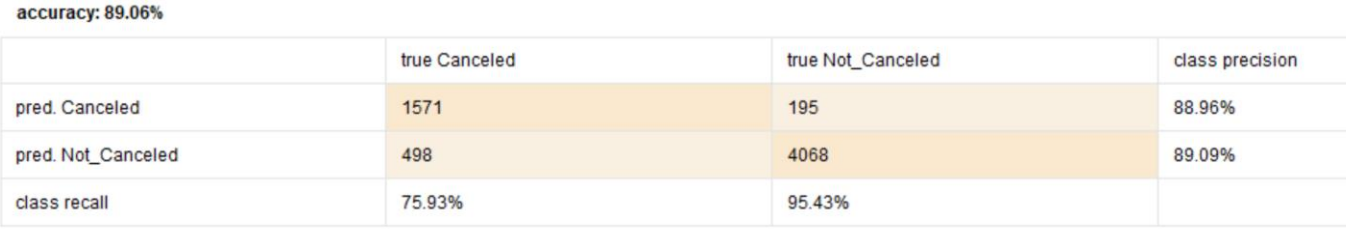


	true Canceled	true Not_Canceled	class precision
pred. Canceled	664	130	83.63%
pred. Not_Canceled	238	1682	87.60%
class recall	73.61%	92.83%	

Model Building - Random Forest/ Pruned

- Model is not overfitting as less than 10pp difference in train and test recall values

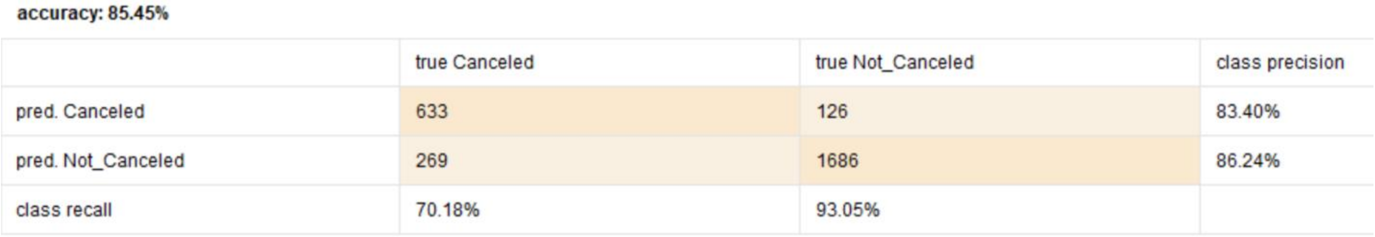
Train



accuracy: 89.06%

	true Canceled	true Not_Canceled	class precision
pred. Canceled	1571	195	88.96%
pred. Not_Canceled	498	4068	89.09%
class recall	75.93%	95.43%	

Test



accuracy: 85.45%

	true Canceled	true Not_Canceled	class precision
pred. Canceled	633	126	83.40%
pred. Not_Canceled	269	1686	86.24%
class recall	70.18%	93.05%	