Reducing Hotel Room Cancellation Rates

Ponmurugu Thanga Tharun, Risha Sunil Shetty, Tristan Amadeus Surya

Department of Computer Science, Nanyang Technological University

This document is the lab project report for the SC1015 Data Science and Artificial Intelligence course at Nanyang Technological University. The project aims to quantify the qualitative description of different liquids using machine learning techniques and use it as an effective tool for lab experiments from images.

1. Introduction

A significant number of hotel bookings are often called off due to cancellations and no-shows. Reasons for cancellations may include change of plans, scheduling conflicts, etc. Cancellations are often made easier by the option to do so free of charge or at a cancellation fee. Although this may be beneficial for consumers, it is a revenue-diminishing factor for hotels to deal with, where losses are particularly high on last-minute cancellations. New platforms hosting online booking channels have dramatically changed customers' booking possibilities and behavior adding to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking patterns and guest characteristics. This pattern of cancellations of bookings impacts 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.

Our project aims to use a model Machine Learning solution that can help in predicting which booking is likely to be canceled.

2. Dataset

INN Hotels Group - Portugal has gathered data to produce the ”INNHotelsGroup” dataset [1]. The dataset contains 36275 entries and 19 data points with no duplicate values or missing values.

3. Features

1. Data Dictionary
2. Booking_ID: Unique identifier for each booking
3. No_of_adults: Number of Adults
4. No_of_children: Number of Children
5. No_of_weekend_nights: Number of weekend nights (Saturday/Sunday) the guest stayed/booked to stay at the hotel
6. No_of_week_nights: Number of weekday nights (Monday to Friday) the guest stayed/booked to stay at the hotel
7. Type_of_meal_plan: Type of meal plan booked by the customer:
    a. Not Selected – No meal plan selected
    b. Meal Plan 1 – Breakfast
    c. Meal Plan 2 – Half board (breakfast and one other meal)
    d. Meal Plan 3 – Full board (breakfast, lunch, and dinner)
8. Required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
9. Room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
10. Lead_time: Number of days between the date of booking and the arrival date
11. Arrival_year: Year of arrival date
12. Arrival_month: Month of arrival date
13. Arrival_date: Date of the month
14. Market_segment_type: Market segment designation.
15. Repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
16. No_of_previous_cancellations: Number of previous bookings that were canceled by the customer before the current booking
17. No_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer before the current booking
18. Avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (euros)
19. No_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
20. Booking_status: Flag indicating if the booking was canceled/not canceled.

	count	mean	std	min	25%	50%	75%	max
No. of Adults	36275	1.84	0.52	0	2	2	2	4
No. of Children	36275	0.11	0.40	0	0	0	0	10
No. of Weekend Nights	36275	0.81	0.87	0	0	1	2	7
No. of Week Nights	36275	2.20	1.41	0	1	2	3	17
Required Parking Car Space	36275	0.03	0.17	0	0	0	0	1
Lead Time	36275	85.23	85.93	0	17	57	126	443
Arrival Year	36275	2017.82	0.38	2017	2018	2018	2018	2018
Arrival Month	36275	7.42	3.07	1	5	8	10	12
Arrival Date	36275	15.60	8.74	1	8	16	23	31
Repeated Guest	36275	0.03	0.16	0	0	0	0	1
No. of Previous Cancellations	36275	0.02	0.37	0	0	0	0	13
No. of Previous Bookings not Canceled	36275	0.15	1.75	0	0	0	0	58
Average Price per Room	36275	103.42	35.09	0	80.3	99.45	120	540
Number of Special Requests	36275	0.62	0.79	0	0	0	1	5

4. Algorithms

We realized that the essence of the project was to understand the intricacies of different machine learning algorithms and to learn which algorithm gives good results for which use case. With this philosophy, we use exploratory data analysis to find the best predictors of booking cancellations. Afterward, we used data modeling, such as logistic regression, support vector machine, decision tree, and random forest, to predict the cancelation probability of the booking.

We observed that libraries like *Scikit Learn* allowed us to tweak different aspects of an algorithm, but not to the extent of our implementation. We chose to try out numerous algorithms using the *Scikit Learn* library [2] in **Python**, ordered from simplest to most complicated:

1. Logistic Regression: This is a simple and efficient algorithm for binary classification problems. It works well when the features are linearly separable. It also provides probabilities for the outcomes, which can be useful for understanding the model.

2. Support Vector Machines: SVMs are effective in high-dimensional spaces and best suited for problems with complex boundaries.

3. Binary Tree: Binary decision trees are particularly useful when you have categorical variables that divide the data into two distinct groups.

4. Decision Trees: Decision trees can handle both numerical and categorical data. They are easy to understand and interpret.

5. Random Forests: A Random Forest is an ensemble of decision trees. It is more robust and less prone to overfitting.

After trying out all the models, we will determine their accuracy using the tuned model. Then, we will visualize the models. Using those model visualizations, we will find the variable that corresponds well to the cancellation. We will then visualize the feature importance graph for presentation purposes.

5. Analysis and Results

Before we implement the algorithms, we must first analyze our dataset.

1. Lead Time

- The histogram in **(Fig 1.2)** reveals a right skewed-distribution, whereby the majority of the lead times are under 126.
  - It is common for individuals to plan vacations approximately 3 to 6 months in advance.
- The boxplot in **(Fig 1.2)** highlights the presence of numerous outliers.
  - These outliers may be attributed to factors such as special deals or promotions offered by hotels, prompting individuals to book well in advance.

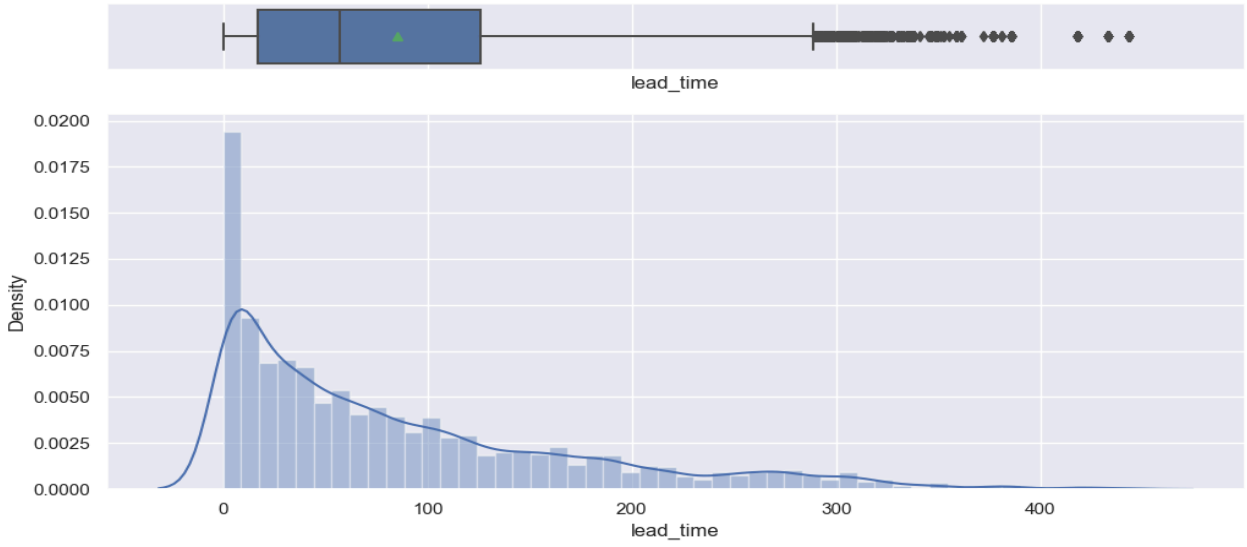


Fig 1.2 : Histogram & Boxplot to display distribution of Lead Time

2. Average Price of a Room

- The mean and median for the average price of a room are close to 100.
- The majority of the prices range from around 80.30 to 120.00.
- The average price of a room **(Fig 1.3)** ranged from 0 to almost 375 dollars.
  - This variation is expected, considering the diverse range of room types available in hotels, spanning from standard rooms to suites.
- The dataset included rooms that were priced at 0 dollars. **(Fig 1.3)**
  - We found that those were rooms provided as complimentary services by the hotel. Perhaps as part of a promotional deal or loyalty programs for members.

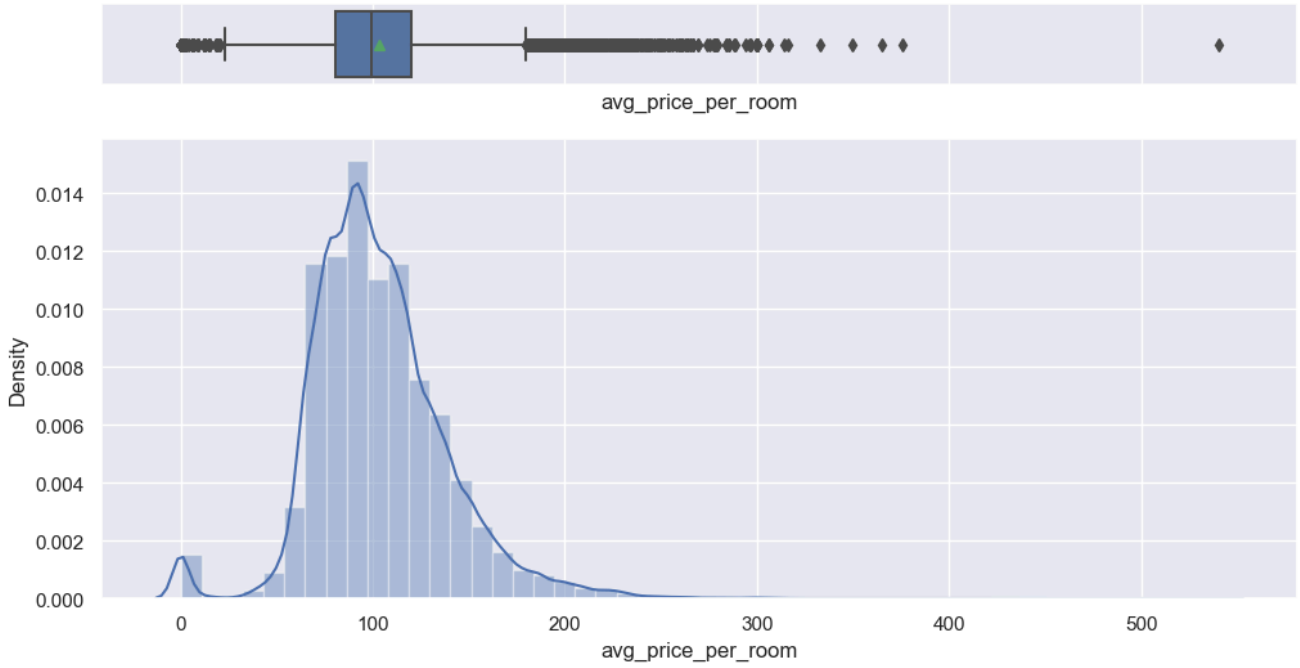


Fig 1.3 : Histogram & Boxplot for the average price of a room

3. Number of Children

- The count plot revealed that in 93% **(Fig 1.4)** of the cases, customers were not traveling with children.
  - This is expected as most travelers prefer not to travel with children due to multiple factors, which include, the desire to relax as well as wanting to reduce expenses.
- The data contains values where the number of children is 9 or 10.
  - Given the unlikeliness of having 9 or 10 children, we replaced these data values with 3 children.

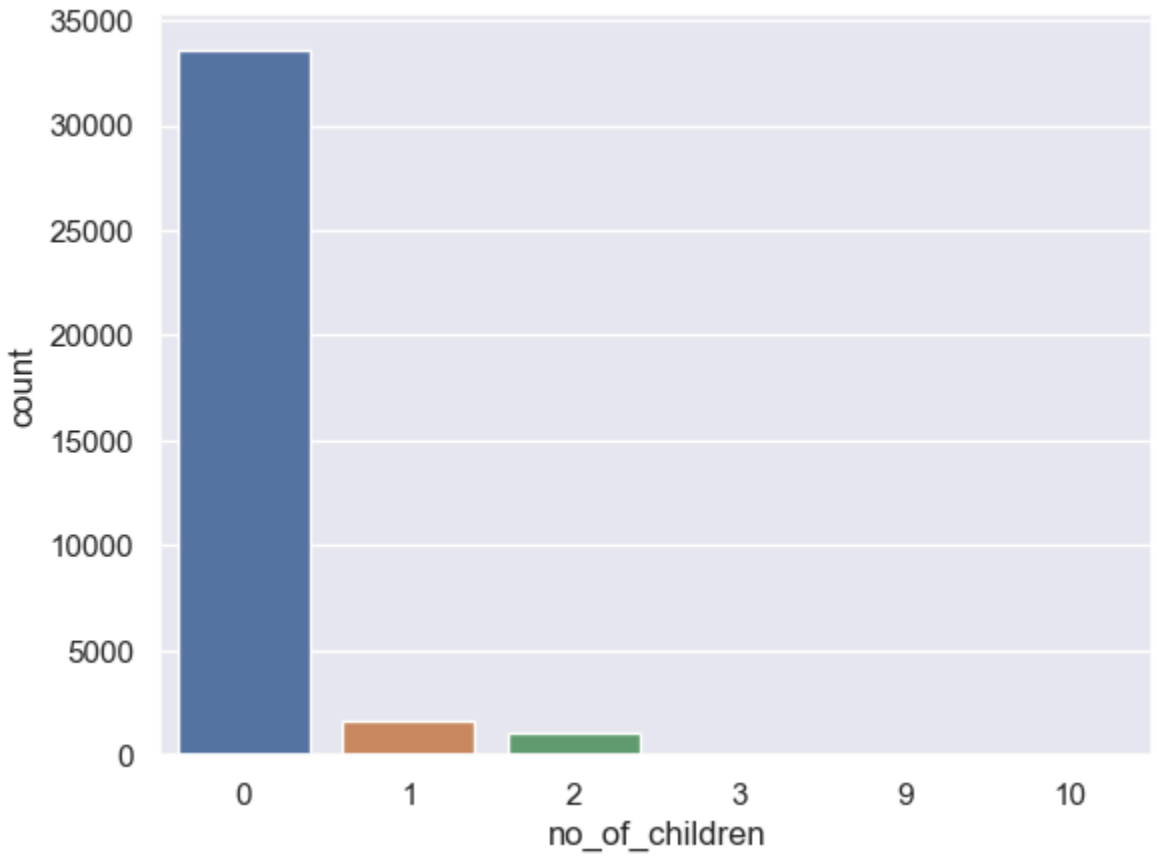


Fig 1.4 : Count plot for the number of children

4. Arrival Month

- The count plot in **(Fig 1.5)** reveals that for this particular dataset, October is the busiest month for hotel arrivals, followed by September and August.
  - This was expected as it was found that October is the best time to travel as the weather all around the world is generally favorable.
- Over 35% of all bookings were made for one of these three months.
- 14.7% of the bookings were made for an October arrival.

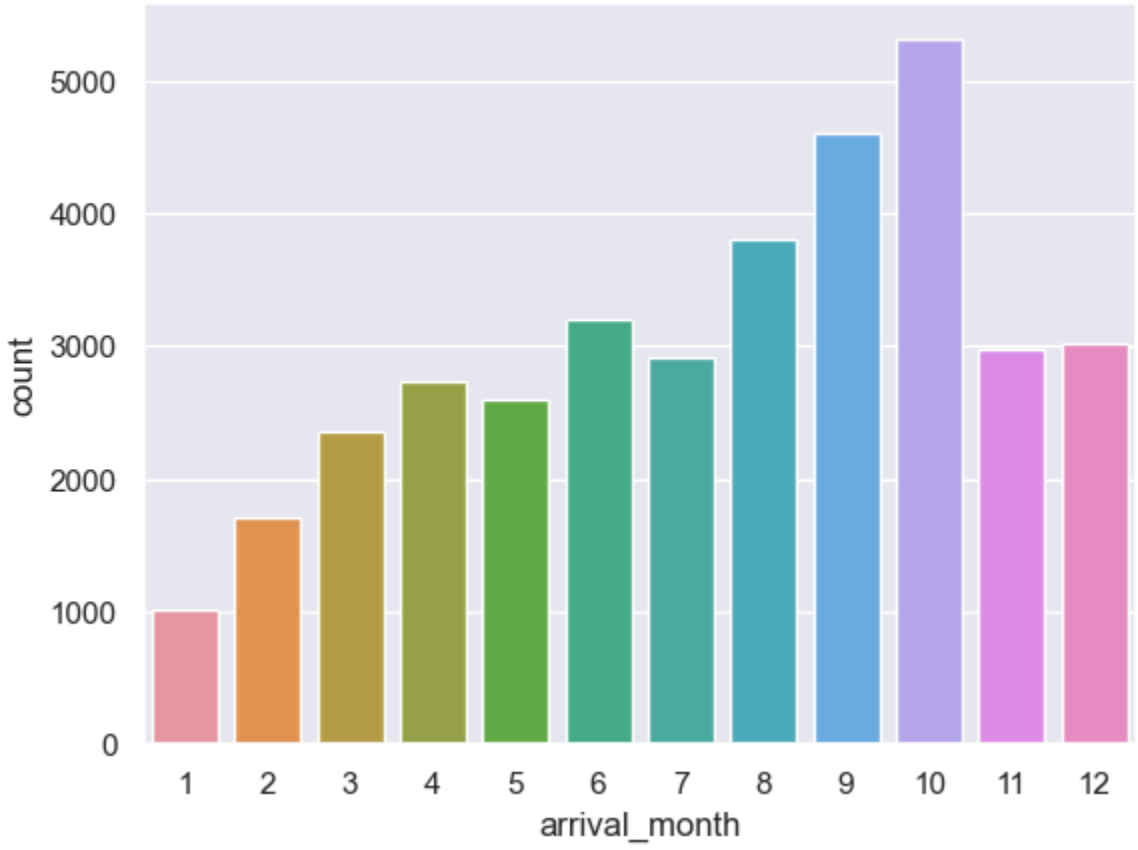


Fig 1.5 : Count plot for the Arrival Month

5. Booking Status

- The count plot **(Fig 1.6)** reveals that 32.8% of the bookings were canceled by the customers.
  - These cancellations may have been due to plan changes, human error, or external factors such as online booking platforms.

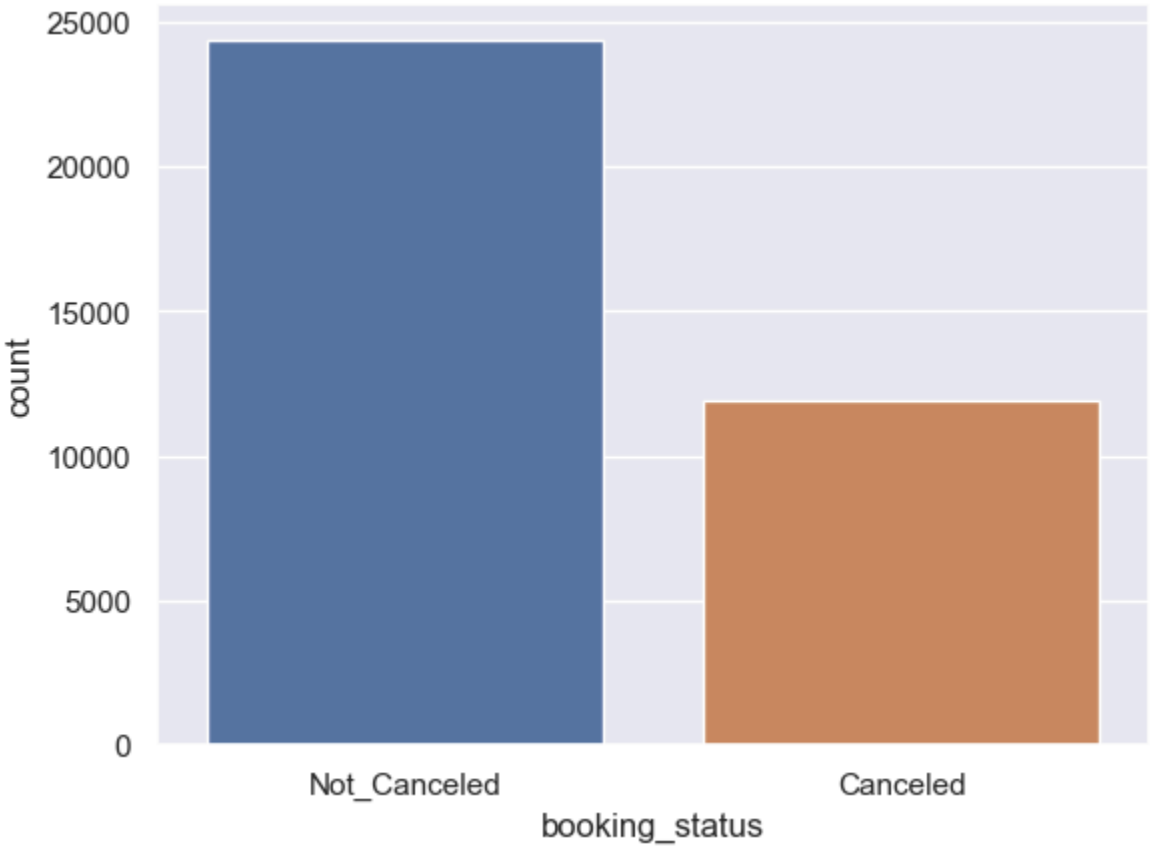


Fig 1.6 : Count plot for booking status

6. Bivariate Analysis (Numeric Features)

1. The number of not canceled previous bookings and repeated guests have the highest positive correlation of 0.54 **(Fig 1.7)**.

2. Booking status and lead time have the next highest positive correlation of 0.47 **(Fig 1.7)**.

Meanwhile,

1. The average price of a room with the number of adults/number of children has a weak positive correlation of 0.30 and 0.35 respectively **(Fig 1.7)**.

2. The number of special requests has a weak negative correlation of -0.25 with the booking status **(Fig 1.7)**.

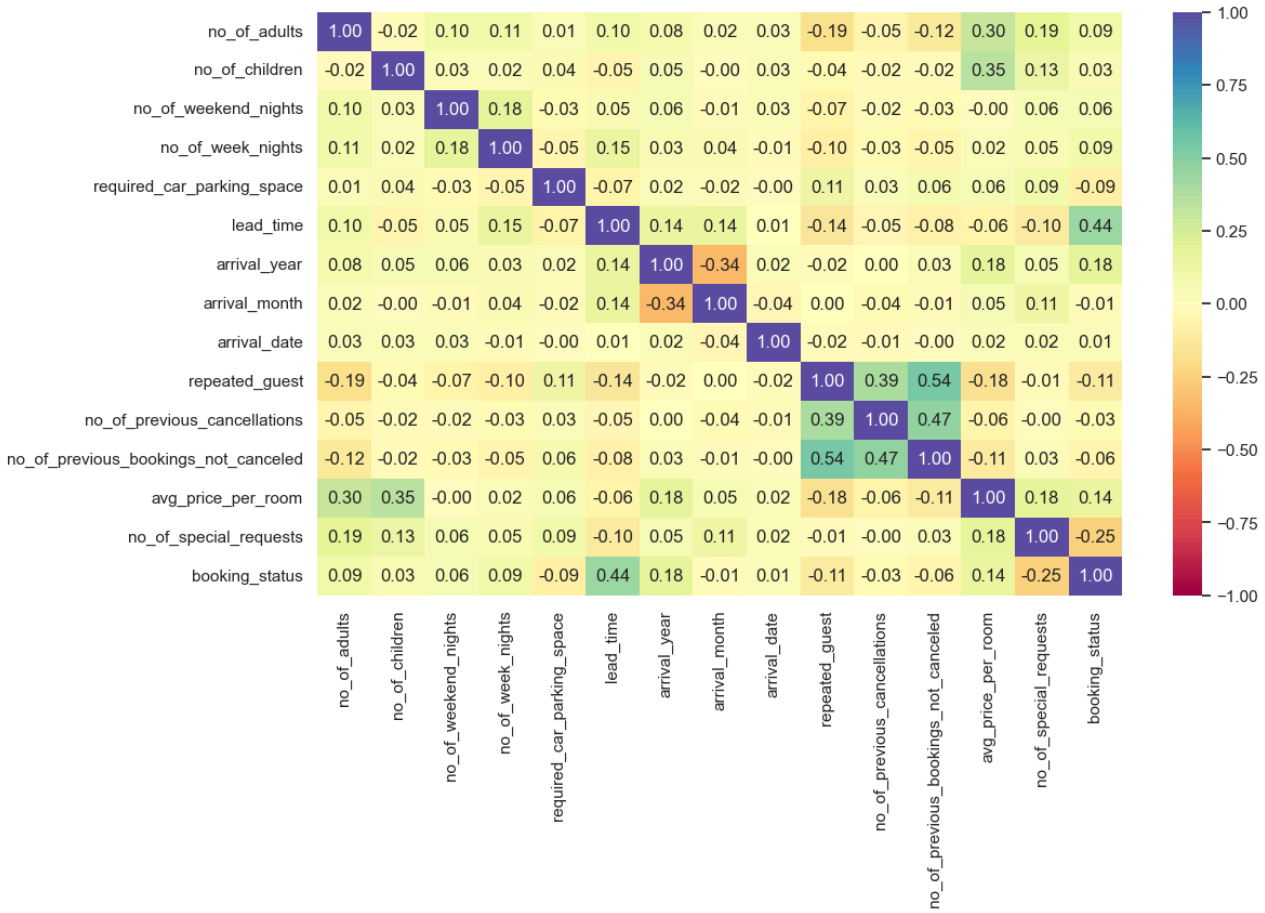


Fig 1.7 Correlation Matrix for the numeric features of the dataset

7. Market Segment Type & Average Price Per Room

- Based on the boxplot **(Fig 1.8)** we observed that online rooms have higher price variations, while offline and corporate room prices are almost similar.
  - Online travel agents propose various prices, and such variations are beyond our control.
- The complementary market segment gets the rooms at cheaper prices, as explained previously.

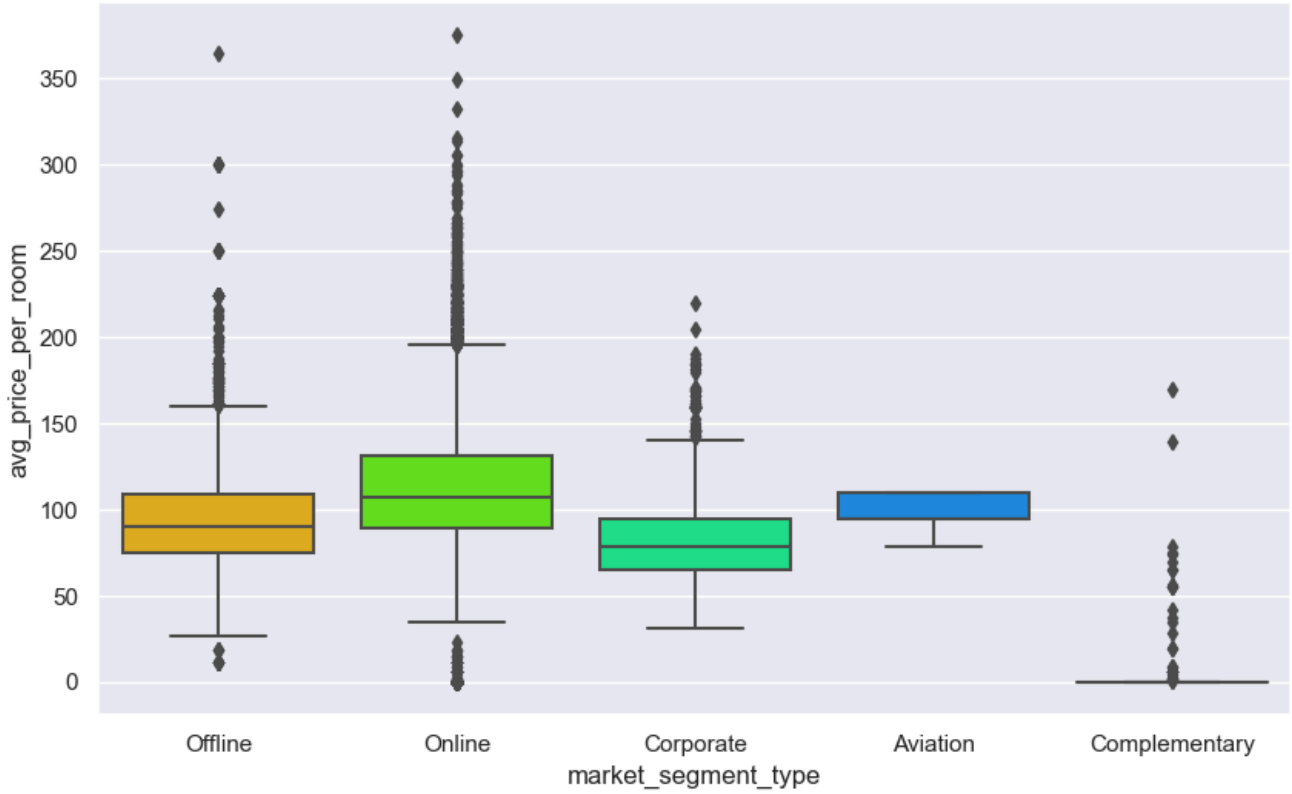


Fig 1.8 : Boxplot (Average Price Per Room against Market Segment Type)

8. Booking Status & Market Segment Type

- The graph in **(Fig 1.9)** suggests that online and aviation markets have the highest cancellations.
  - The highest rate of cancellation was observed in the online market. This is expected, given the ease of cancellation coupled with the ability to compare prices with various booking platforms.
- Offline market is not far back.
- Aviation cancellations may correspond to flight cancellations
  - Flights get canceled from time to time regularly.
  - The hotel should look into developing packages for the aviation market with higher prices but complementary cancellations.
- Complementary bookings are seldom canceled as they are essentially free.
- Corporate has a low cancellation rate compared to other segments
  - Hotels should look into developing packages for corporate customers.

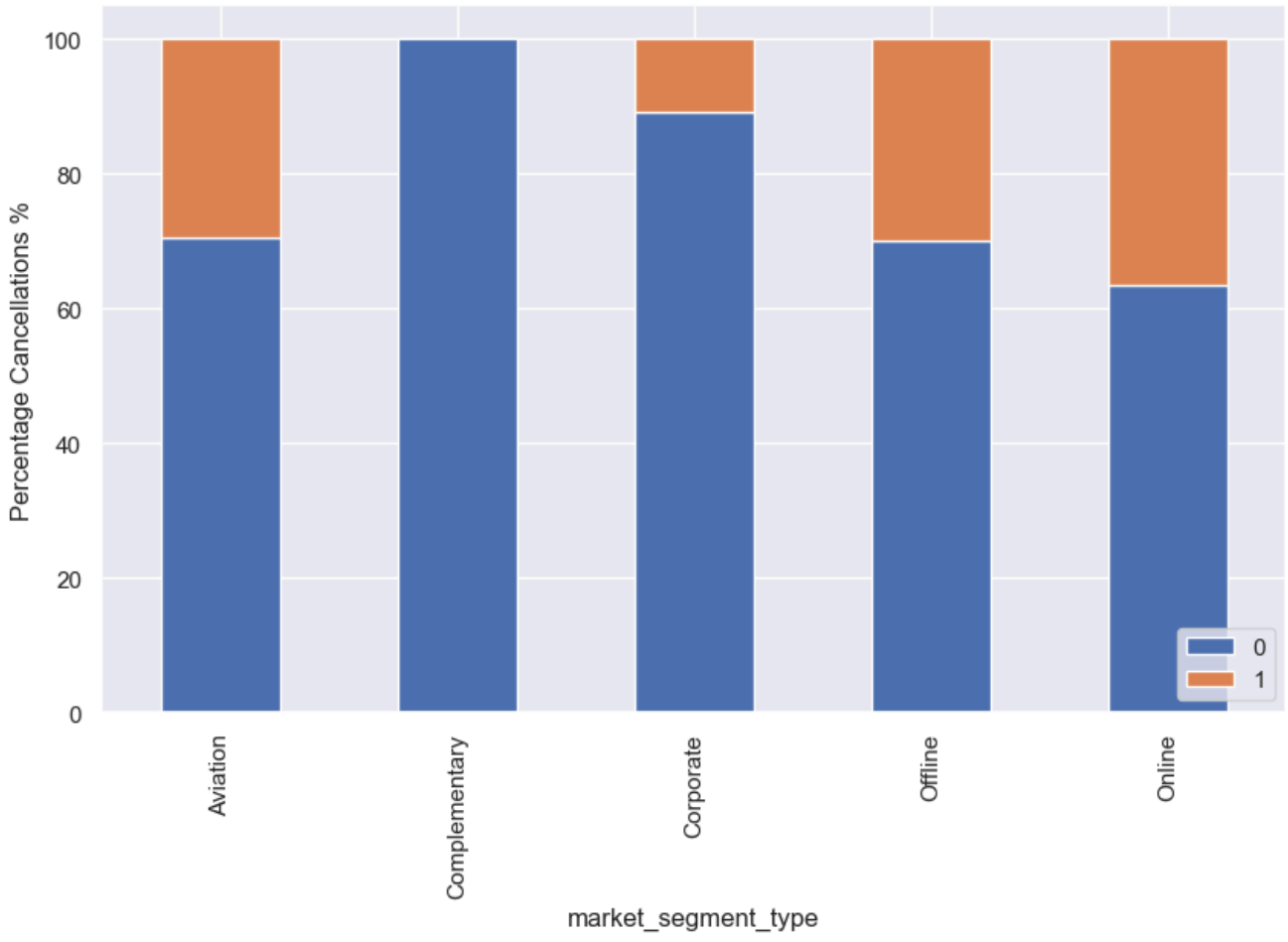


Fig 1.9 : Stacked Bar Plot (Percentage Cancellations against Market Segment Type)

9. Booking Status & Repeated Guests

- The stacked bar plot in **(Fig 2.0)** reveals that repeated guests are less likely to cancel their bookings
  - Most likely due to their loyalty to the brand and their previous positive experience with the brand.
  - There is also a possibility that these repeated guests are seasoned travelers who are able to plan their travel with minimal changes required unlike new travelers.

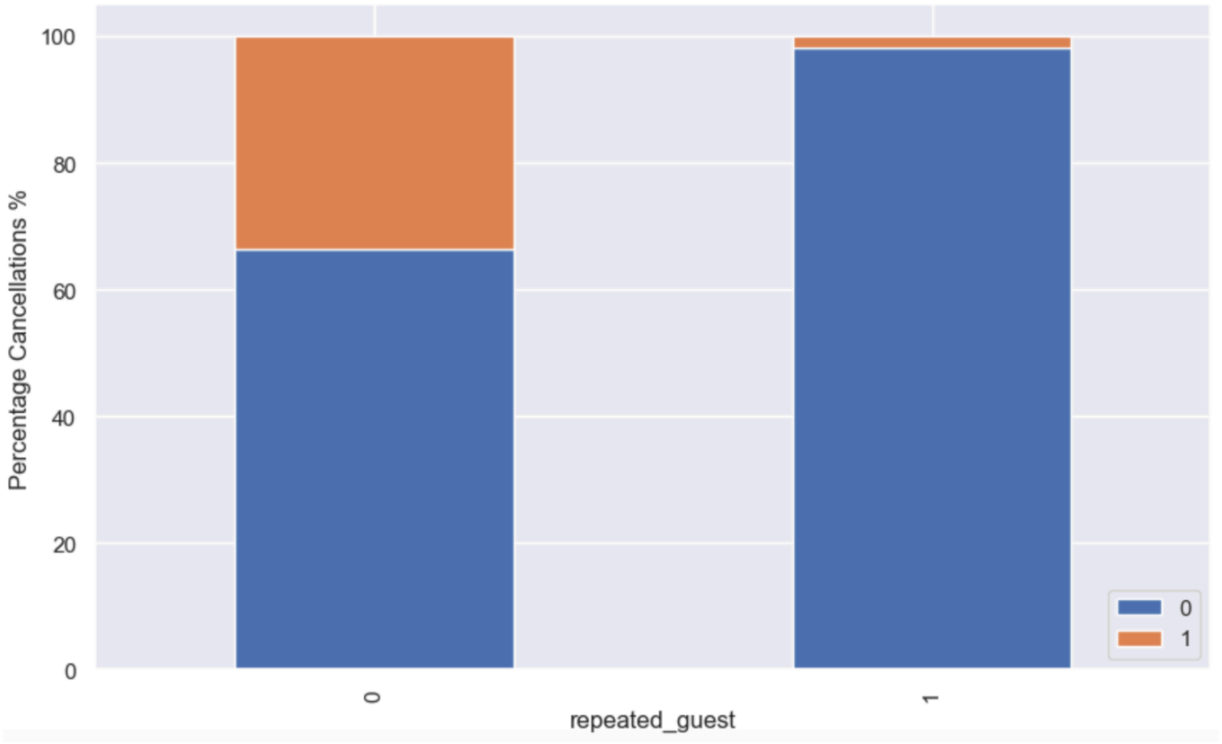


Fig 2.0 : Stacked Bar Plot (Percentage Cancellations against Repeated Guests)

10. Booking Status & Total Days Booked

- The stacked bar plot in **(Fig 2.1)** reveals that as the number of days booked by the customers increases, the chances of cancellation increases as well.
- There was an anomaly for (23 days), however, we did not find any information that suggested to us that it was a valid point to consider hence, we removed the outlier. Which gave us the stacked bar plot shown in **(Fig 2.1)**

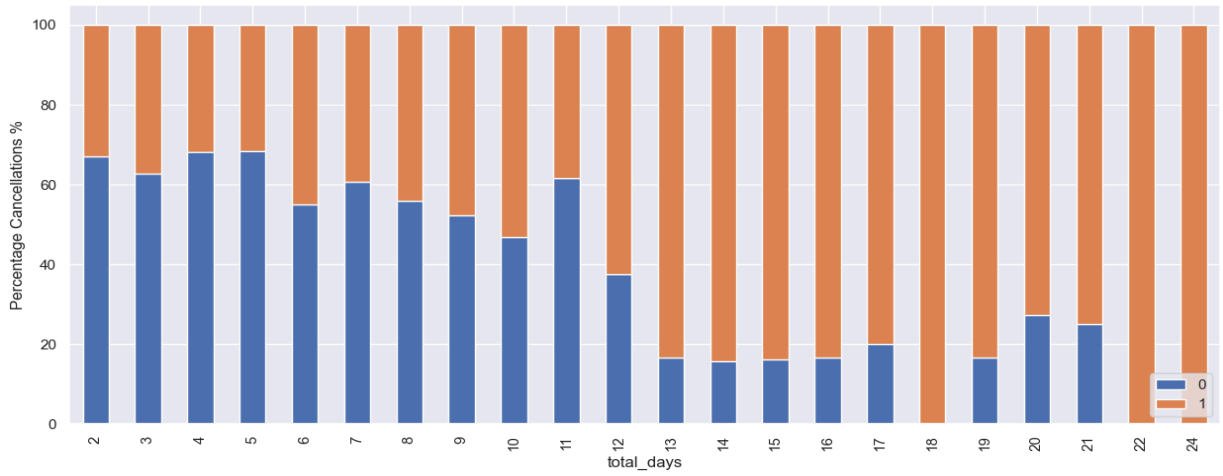


Fig 2.1 : Stacked Bar Plot (Percentage Cancellation against Total Days)

11. Average Price Per Room & Arrival Month

- **Fig 2.2** reveals that the price of rooms is the highest between May to September - (115 Euros per room)
  - ➢ This is expected given that this is the period of school holidays internationally.
  - ➢ Hence causing a rise in demand, which explains the price increase.

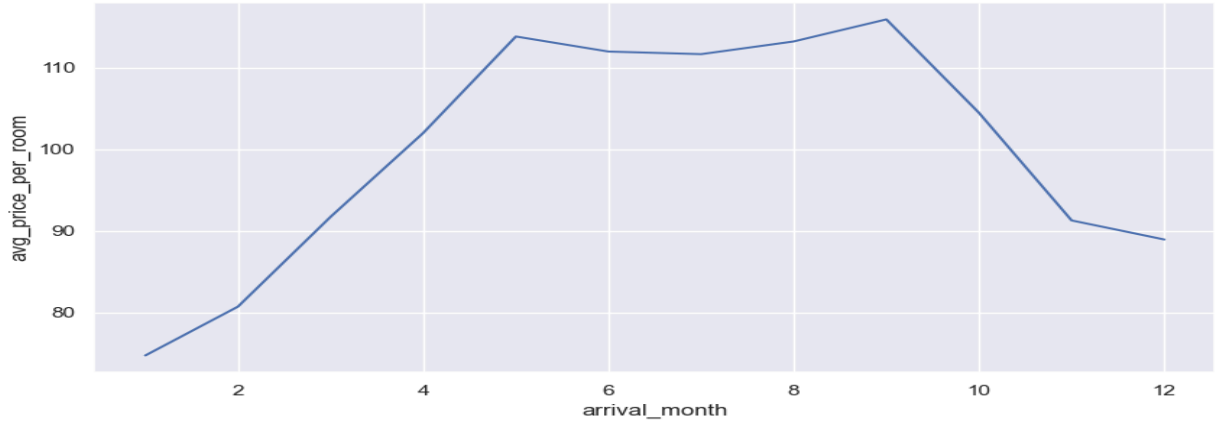


Fig 2.2 : Average Price Per Room against Arrival Month

After conducting our exploratory data analysis, we trained and tested our algorithms on the complete data set to start with. Later we randomly separated the data set into training data and test data so that we had samples from each class. 70% of the data was used as training data and 30% was used as test data. The following figures and tables show the results we observed on implementing algorithms as mentioned in the above section.

Table 1 shows the training and testing accuracy for the different learning methods we implemented. Figure 2 shows the testing accuracy using different techniques on three of the learning methods - LDA, SVM, and Logistic Regression.

Table 1: Training Method Accuracy

Method	Training Accuracy	Testing Accuracy
Logistic Regression	79%	79%
Support Vector Machines	80%	80%
RBF Kernel	80%	82%
Decision Trees	89%	89%
Random Forest	99%	99%

Fig 2.3 : Feature Importance Plot (Decision Trees)

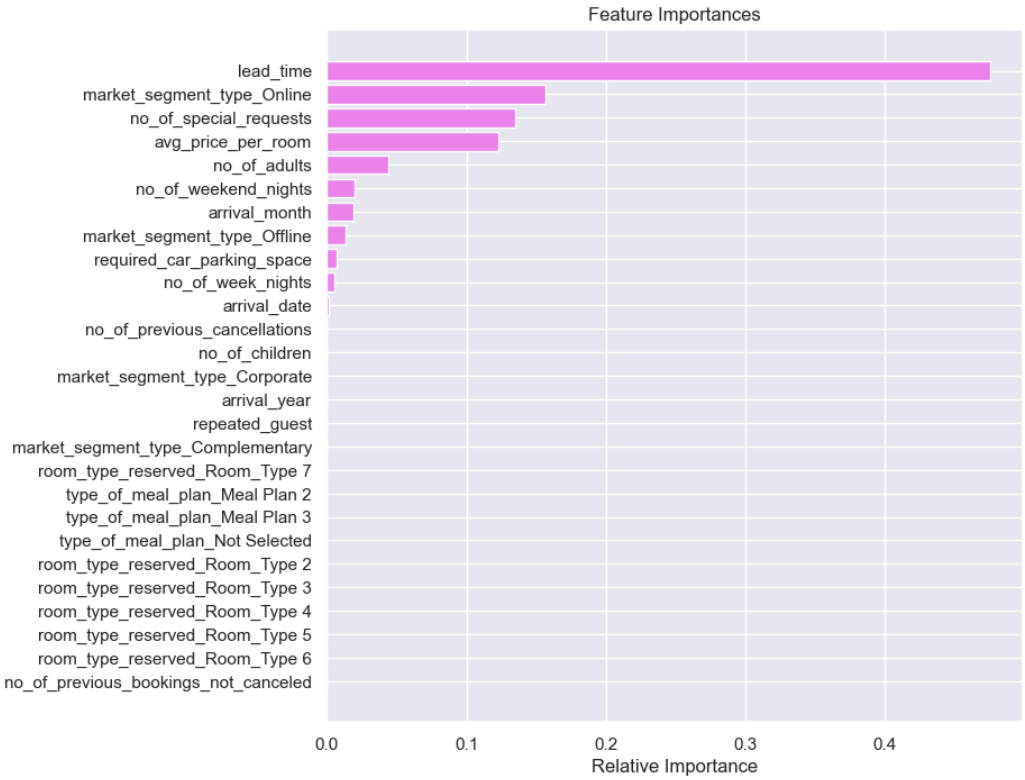
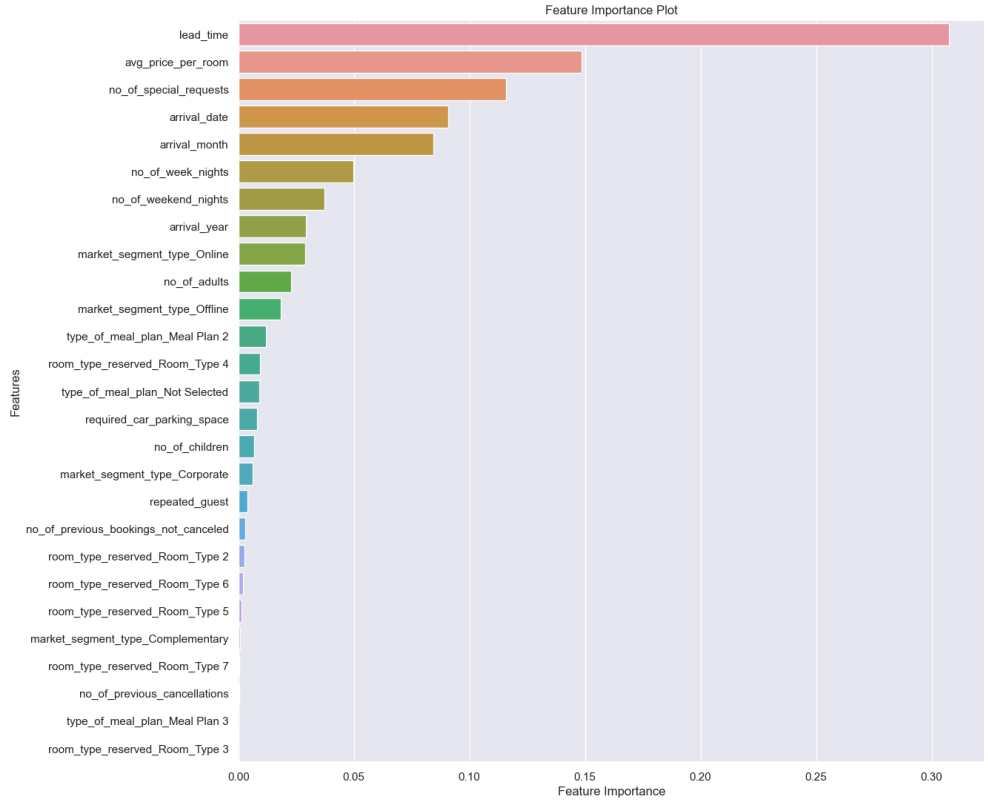


Fig 2.4 : Importance (Random Forest)



Improvements/Advancements

- Previously, when performing univariate analysis on 'Booking Status', we had observed that there was a class imbalance. **(Fig 2.5)**
- Class imbalance may potentially affect the accuracy of our results as there may be a possible bias towards the majority class which is the 'Not canceled' class.
- Hence, to mitigate such a possibility, we performed **Upsampling**. **(Fig 2.6)**
- We then utilized our most accurate model (Random Forest Model) with the upsampled data. **(Fig 2.7)**

Fig 2.5 : Booking Status with class imbalance

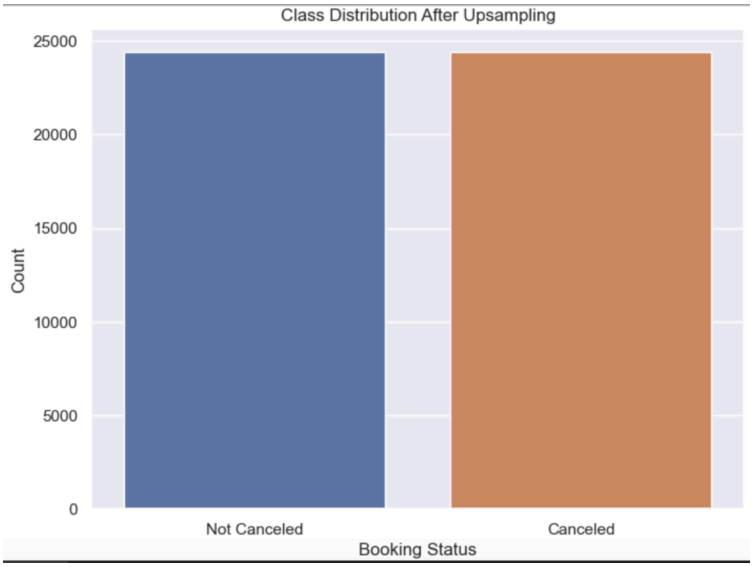


Fig 2.6 : Upsampled Booking Status

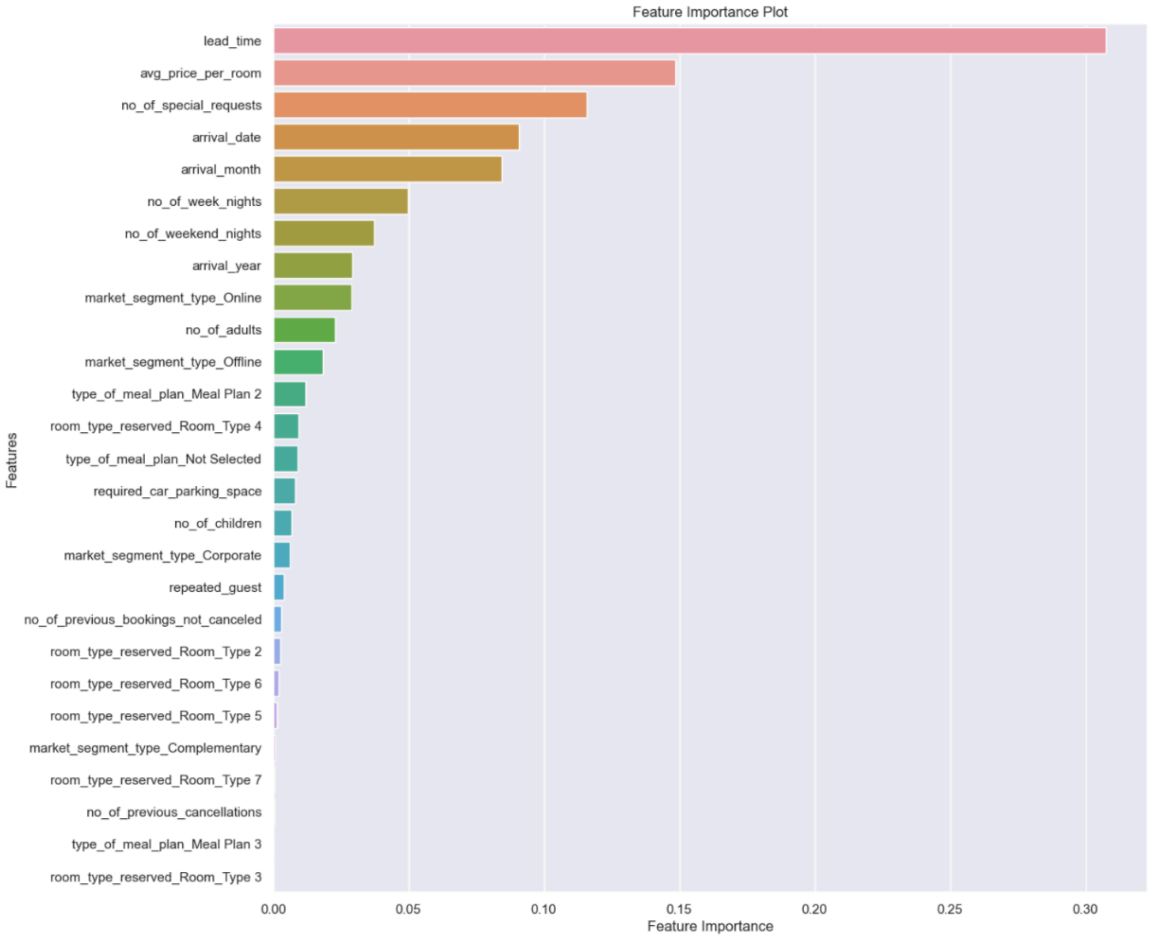


Fig 2.7 : Random Forest Model with the modified data

- As seen in **(Fig 2.7)**, with the upsampled data there is little to no difference in the accuracy of the model, however, by performing upsampling, we mitigated any possibility of bias or ambiguity in the results.

6. Discussion

We have found the Lead Time, Market Segment online, and number of Special requests along with the average price of the room are the most important variables in predicting the guest booking cancelation pattern.

1. Given that lead time has the highest relative importance we recommend that hotels limit how ahead in time guests may book their rooms as bookings with high lead time have a higher rate of cancellation.

2. The results also revealed that bookings with special requests have a low probability of cancellation. Which is understandable as guests are able to make their stay unique through such requests. Therefore, we recommend hotels to offer customisable packages to guests. Which will accommodate special requests as well as provide a personalized stay. This way hotels would be able to retain their guests as each stay would be tailored to the preferences of the guests.

3. The results also revealed that bookings done online have a high probability of cancellation. This is probably due to the fact that guests are now able to compare prices of different hotels with a click of a button. Hence we recommend hotels to provide competitive prices while limiting the number of online bookings hotels accept.

The correlation plot agrees that the lead time most greatly affects hotel cancellations with a 0.44 correlation. The plot also states that the number of special requests and the average price of the room also have some correlation, with a -0.25 and 0.11 correlation, respectively. Therefore, the comparison of feature importance using ML models agrees with the correlation plot.

Table 2: Comparison of feature importance

Feature	Learning Method	Relative Importance
Lead Time	**Decision Trees**	**0.47**
Lead Time	Random Forest	0.31
Average price per room	Random Forest	0.15
Market segment (online)	Decision Trees	0.15
Number of special requests	Decision Trees	0.13
Number of special requests	Random Forest	0.12
Average price per room	Decision Trees	0.12

7. Future Work

1. We implemented the Decision Trees and Random Forests, but our computer took a very long time just to load the dataset. So, we can try to run our project at a higher depth using high-performance computing machines.
2. We can use more variables, such as the ID of users, to determine the probability of booking cancellation
3. We can quantify the probability of having a booking canceled using percentages.
4. We can develop a desktop application that allows hotels to prioritize bookings that are less likely to be canceled

References

[1] https://www.kaggle.com/datasets/pthangatharun/inn-hotels-group/data

[2] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Contributions Overview for Project Team Members - Alphabetical Order

Risha:

Risha played a pivotal role in our mini-project, primarily focusing on Exploratory Data Analysis (EDA). She adeptly handled both univariate and bivariate analyses, providing insightful interpretations of the data. Risha also undertook the task of upsampling models to ensure robust training across all classes, enhancing our understanding of key variables and resulting in valuable recommendations. Her efforts extended to crafting the presentation slides and making critical decisions regarding project approaches, ensuring adherence to deadlines with exceptional quality and efficiency. She also made sure to identify mistakes and rectify them immediately. Her proactive approach included early completion of tasks to gather essential feedback for continuous improvement.

Tharun:

Tharun was instrumental in our mini-project, beginning with extensive dataset exploration to select the most suitable one for our needs. He developed a comprehensive plan for analysis and model building for the INN hotel booking project. Tharun led the construction and refinement of various models including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, evaluating and comparing their efficacy meticulously. His models, incorporating true positive and false positive rates, provided hotels with informed

decision-making tools. Tharun also contributed significantly to project explanations during presentations and took on key coordination responsibilities within the team, facilitating meetings. He also had a proactive approach leading to an early completion of tasks to gather essential feedback for continuous improvement.

Tristan:

Tristan's role was critical in transforming our project notebook results, graphs, and insights into a polished and accessible project report. He synthesized complex findings into a clear narrative, ensuring our report was both informative and reader-friendly.

This revised summary offers a more streamlined and detailed account of each team member's contributions, highlighting their individual roles and achievements within the project.