DSO 545 Final Project Report

Hotel Reservation Cancellations

Jagriti Sharma, Matteo Mennini,
Mihir Sabnis, Muskan Aggarwal, Shyamal Oza

Motivation

During our trip to Yosemite National Park during fall break, where the golden hues of autumn surrounded us in a breathtaking display, our casual chit-chat with the hotel staff at a nearby beachfront resort took a surprising turn. We were joking around about their impressive earnings when the hotel manager dropped a bombshell – the resort and, apparently, the whole hotel industry have been facing major financial losses. The culprit? A growing trend of cancellations by customers. This revelation, amid our laughter and seaside hangout, hit us hard. It became the spark for our academic curiosity, pushing us into the complex world of hotel cancellations. What began as a light-hearted conversation morphed into a determined mission to understand the ins and outs of this financial challenge. This incident motivated us to explore the intricacies, envisioning not only turning around one hotel's fortune but also making a dent in the broader landscape of the entire hotel industry.

The Business Question

The dataset of a beach-view resort and spa that we have chosen shows that 33% of customers have canceled their reservations. Which are the variables that influence the likelihood of hotel booking cancellations? Can we predict the likelihood of cancellations?

The problem of cancellations:

1. **Revenue Loss**: When customers cancel close to their date of arrival, the hotel faces a significant challenge in reselling the room, resulting in direct revenue loss. The unpredictable nature of last-minute cancellations complicates revenue

forecasting and financial planning. Beyond room charges, cancellations may lead to the loss of potential revenue from additional services such as dining, spa, or event bookings. This underscores the broader financial impact beyond room sales alone.

2. **Operational Disruptions:** Last-minute cancellations disrupt hotel operations by affecting staff scheduling. Staffing levels are planned based on anticipated occupancy, and sudden changes can lead to understaffing or overstaffing issues. This results in increased labor costs and operational inefficiencies. Operational disruptions extend to resource management, impacting the allocation of resources such as cleaning services, maintenance, and concierge assistance. Efficient resource allocation becomes challenging when faced with sudden changes in occupancy levels.
3. **Inventory management:** Last-minute cancellations can cause problems in buying and maintaining perishable inventory. This is especially true when cancellations happen during holiday seasons when special arrangements are made. Other than a loss of revenue, when considering the industry as a whole this can also have an environmental impact.

Methodology Followed:

Step 1: Data Cleaning and Preparation

Data Cleaning for Reliability: ensure accuracy by addressing missing values, null values, data types, and data inconsistencies.

Initial Data Patterns Understanding: conduct exploratory data analysis to grasp key patterns and trends that influence cancellations to aid the creation of visualizations.

Step 2: Generating Visualizations

Segment Analysis & Correlations: Using visual tools like matplotlib, seaborn, and pyplot to identify market segments driving cancellations and uncover relations among variables.

Dynamic Visuals for Insights: Employing interactive visuals to depict evolving cancellation trends and explore data relationships effectively.

Step 3: Predictive Modeling

Forecasting for Resource Planning: Developing a predictive model to foresee cancellations and assist in strategic resource allocation.

Actionable Insights for Management: Providing actionable predictions to aid in decision-making for staffing, inventory, and operational adjustments.

Dataset Information

How it was collected:

The real-data dataset was collected from a hotel's reservation system from 2017 to 2018 and sourced from Kaggle. The specific details of the hotel, including its name, are not disclosed by the author for privacy reasons.

What it contains:

The dataset includes information about hotel reservations and cancellations.

Dependent variable:

'booking_status' : The status of the reservation (canceled or not canceled).

Independent variables:

- 'no_of_adults': number of adults
- 'no_of_children': number of children
- 'lead_time: number of days between the date of booking and the arrival date
- ‘avg_price_per_room': average price per day of the reservation, prices of the rooms are dynamic
- 'type_of_meal_plan': type of meal plan booked by the customer.
- 'required_car_parking_space': whether the customer requires a car parking space
- 'room_type_reserved': type of room reserved by the customer
- ’no_of_weekend_night’: number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- ‘ no_of_week_nights’: number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- ‘arrival_year’: year of arrival date
- ‘arrival_month’: month of arrival date
- ‘arrival_date’: day of arrival date
- ‘market_segment_type’: customer’s market segment designation
- ‘repeated_guest’: whether the customer has stayed at the hotel in the past
- ‘ no_of_previous_bookings_not_canceled’: number of previous bookings not canceled by the customer prior to the current booking

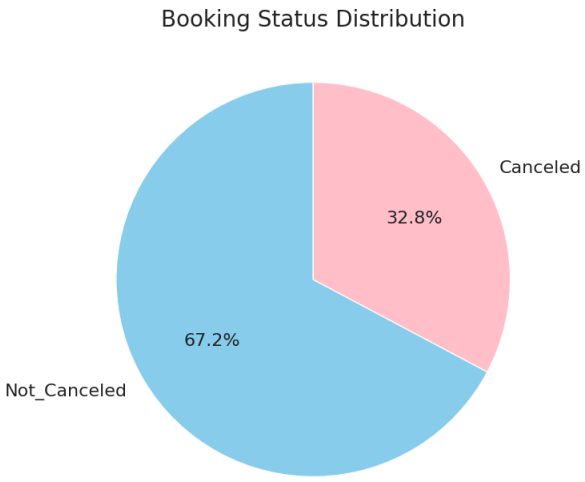
- ‘no_of_previous_cancellations’: number of previous bookings that were canceled by the customer prior to the current booking.
- ‘no_of_special_requests’: total number of special requests made by the customer

Data Preparation:

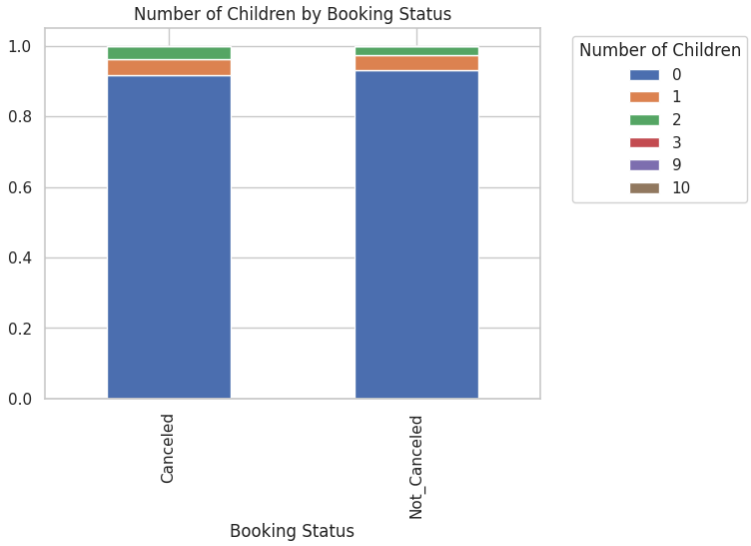
Our dataset exhibited completeness with no missing or null values. However, duplicate entries were identified and subsequently removed in the Excel sheet. Additionally, we addressed date inconsistencies by adjusting dates that fell outside the valid range to the last date of the respective month. To enhance the dataset, we introduced new calculated columns, including "Arrival DateTime" (concatenation of 'arrival_date,' 'arrival_month,' and 'arrival_year'), "Number of Nights" (sum of 'no_of_week_nights' and 'no_of_weekend_nights'), and "Booking Cost" (product of 'number_of_nights' and 'avg_price_per_room'). These enhancements contribute to a more robust and insightful dataset for subsequent analyses.

Exploratory Data Analysis (EDA)

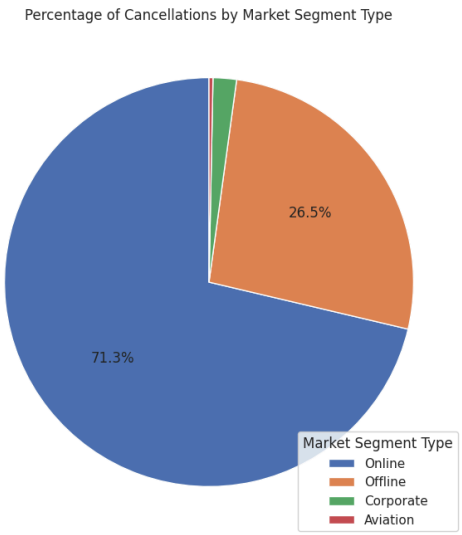
We employed a pie chart to visually analyze the distribution of booking outcomes—canceled versus not canceled. Our findings underscore a notable cancellation rate of 32.8%, signifying a substantial portion of lost bookings for the hotel.



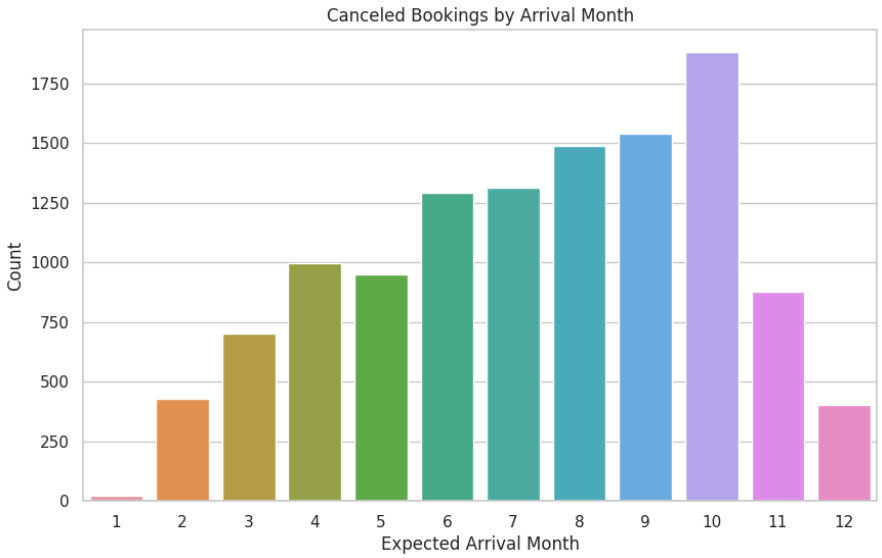
Our objective was to delve into the customer demographics associated with hotel bookings. To achieve this, we employed a stacked bar plot, comparing the number of adults and children against the booking status (Canceled and Not_Canceled). Our analysis revealed a prevailing trend where bookings predominantly featured two adults, and notably, nearly 90% of the bookings did not include children. This insight sheds light on the predominant composition of bookings in terms of adult occupancy and the prevalence of child-free reservations.



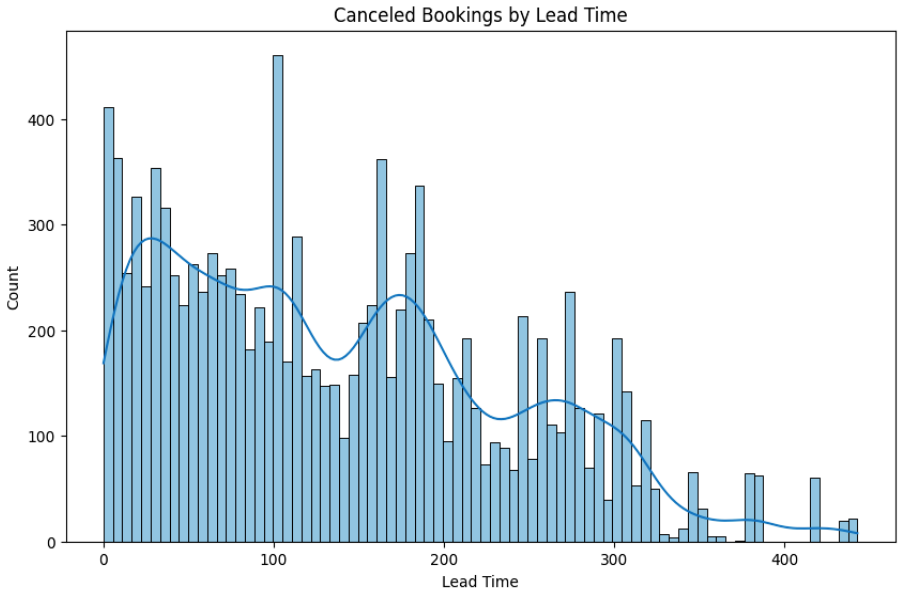
Our investigation extended to identifying the primary market segment contributing to a significant number of cancellations. Our analysis uncovered a noteworthy trend, with approximately 71% of cancellations attributed to the online market, while the offline market accounted for only 26.5% of cancellations. This finding underscores the disproportionate impact of cancellations from the online market segment, highlighting a potential area for targeted interventions or strategies to mitigate cancellations in this specific market domain.



Our inquiry extended to discerning the seasonality trends regarding cancellations over two years. Through the visualization of a bar chart, we identified a pattern wherein the number of cancellations escalated progressively, reaching a peak in October—the onset of the holiday season. Conversely, the month of January exhibited the lowest number of cancellations. This temporal analysis provides valuable insights into the fluctuating patterns of cancellations throughout the year, enabling strategic considerations for managing booking cancellations effectively.

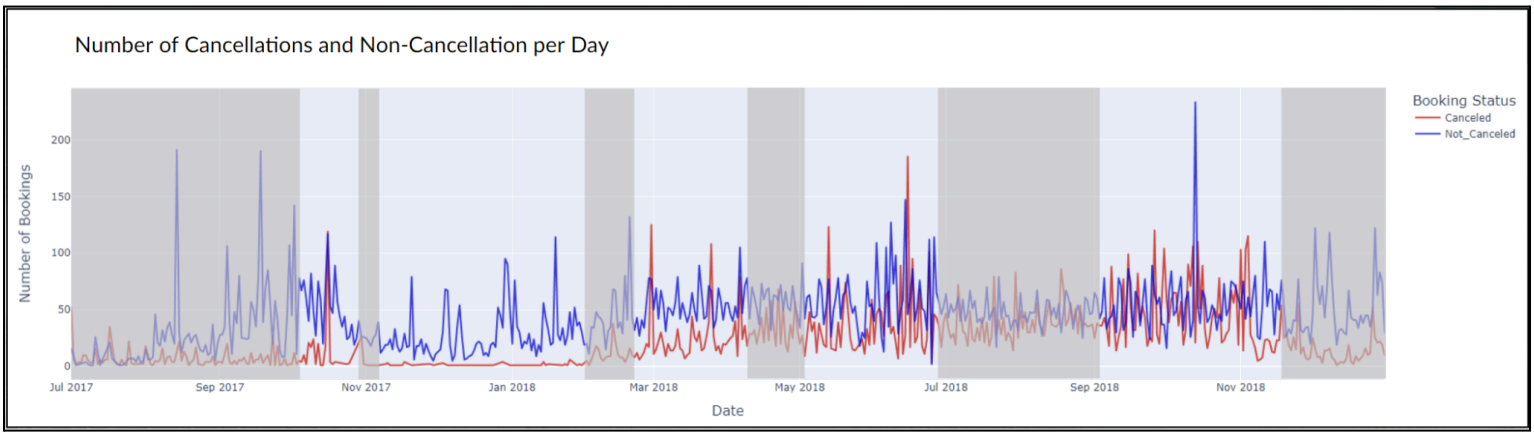


To better investigate booking cancellations we wanted to understand how long these customers have had to make their cancellations. We created a bar chart with a trend line that shows the relationship between lead time and the number of cancellations made. The lead time is the time between reservation and the arrival date of the booking. We can see that as the lead time increases (customers book more in advance), the number of total cancellation has decreased. This could also be because there are fewer bookings made with a longer lead time. We can see some anomalous spikes around 100 days, 150 days and 180 days, which could be due to better prices or other factors not captured by the dataset.

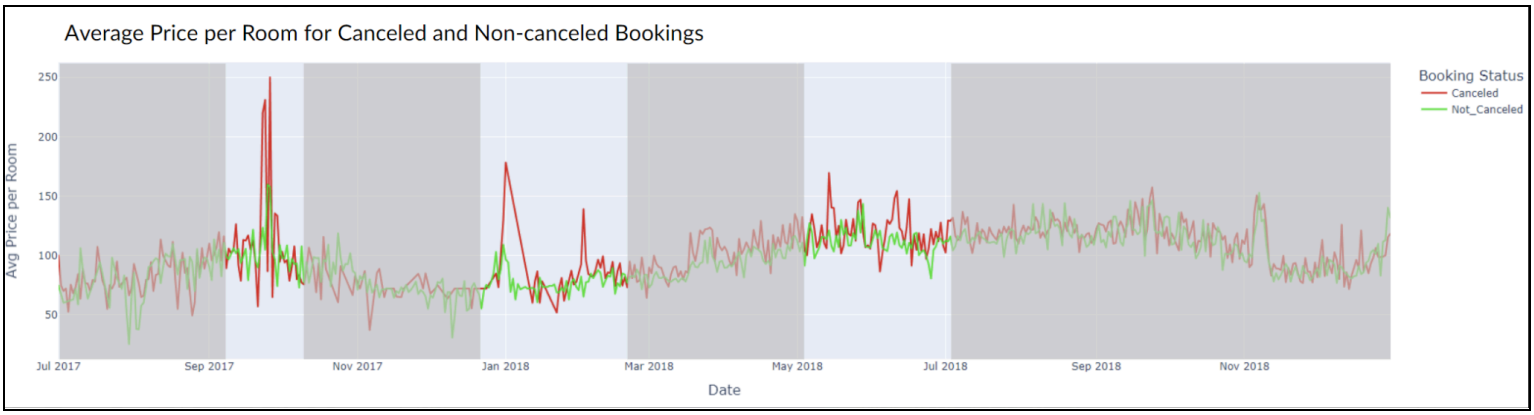


We also wanted to understand how involved a customer is with their booking. From our personal experiences we had an assumption that the more invested you are in your booking, making special requests, the less likely you are to cancel that booking. The stacked bar chart below shows the relationship between the number of booking canceled or not canceled and the number of special requests made. It is in line with our assumption that, as the number of special requests increases, the number of bookings decreases.

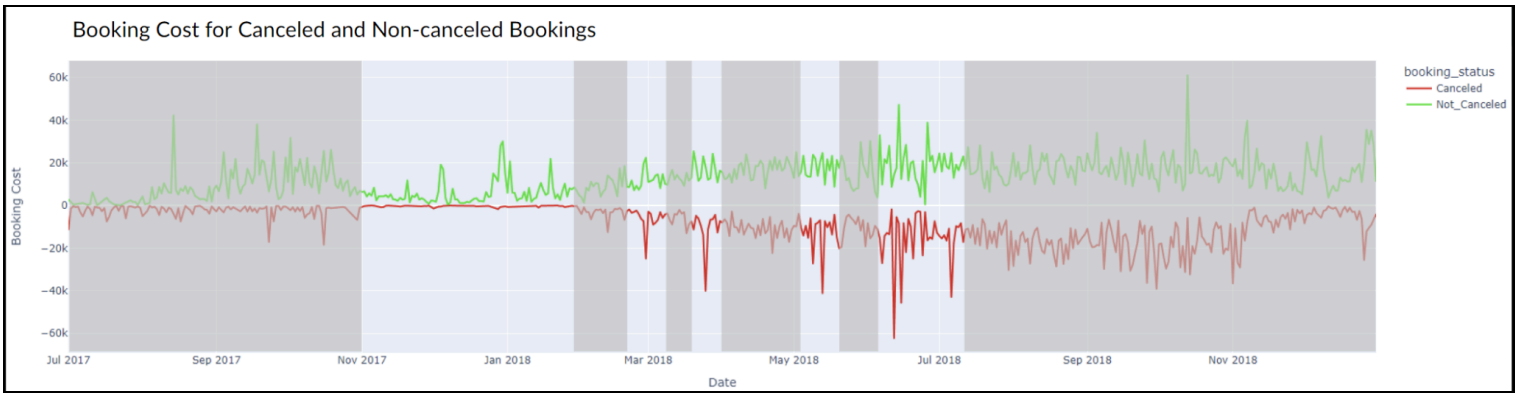
Since this is a time series dataset, we can investigate any potential seasonal patterns and anomalies by plotting bookings over time. In the line graph below we can see the total number of daily canceled and non-canceled booking over time. We can see an anomalous trend from November 2017 - February 2018 where there are little to no booking cancellations made. This could be due to the hotel testing not allowing cancellations for booking made for that period. We can see a high volatility of cancellations from May 2018 to July 2018 which is during a summer which is usually a busy vacation season. Summer also has the largest spike in total cancellations. We can also see high volatility from September 2018 - November 2018 which could be customer planning and finding better deals for the December holiday season.



Now that we have an understanding of the seasonality of cancellations we can see how this might affect a hotel company. First we would need to also understand the trends in the average room price per day. The line graph below allows us to better understand possible customer behavior and price sensitivity. When prices are lower there is a higher volatility in cancellations that could indicate customers canceling as they find better prices. There is a huge spike for 3 separate days in mid-September reaching 150 cancellations. This could be anomalous due to a large party of customers canceling. There is another major spike right before New Years 2017 which could be customers canceling expensive bookings for cheaper last minute booking options. The greatest sustained volatility is in the summer vacation season.



Finally with this information we can take a deeper look into the actual potential revenue loss from these canceling bookings compared to non-canceled bookings. In the line graph below we can see multiple spikes of large potential revenue loss possibly due to large changes in average room price. The greatest loss is around the summer vacation season. The summer season is the costliest to this hotel and would be the toughests season to manage budgets, rebooking, offers, pricing and availability planning.



Predictive Model

We decided to use a machine learning model to predict the likelihood of booking cancellation to better assist the hotel with resource planning. For this specific purpose, we decided to use a decision tree model for two main reasons. First, this model is ideal when working with classification. Our objective is to classify the binary variable “'booking_status” as either “Canceled” or “Not Canceled”. Second, decision trees are a white box, meaning that we are aware of the mathematical reasons why a certain outcome is predicted. This gives increased visibility and can help us motivate the outcome to the hotel’s management, as well as being easier to explain to non technical audiences in general.

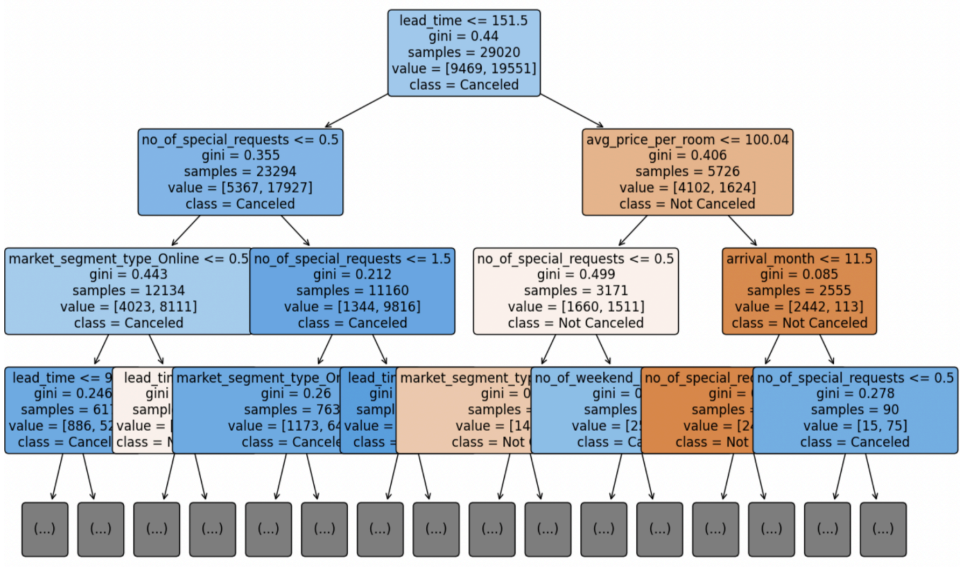
Model Preparation

To prepare the model we first dropped the dependent “'booking_status” variable, as well as other irrelevant variables. We also dropped the "Arrival DateTime" variable due to the library scikit-learn not being able to handle date-type variables. We created dummy

variables for the categorical variables “type_of_meal_plan”, “room_type_reserved”, and “market_segment_type”. While dropping unnecessary variables and creating dummies should not have an impact on a decision tree model’s performance, we decided to do so because it may make the visualization of the tree’s structure clearer and easier to understand. We also performed some pre-pruning by limiting the tree depth to 13 and setting the minimum sample split size to 30 to minimize the likelihood of overfitting. In a real-world scenario, after cross-validating the model with new data, these and other hyperparameter tuning practices could be changed to evaluate the tradeoff between accuracy and visibility.

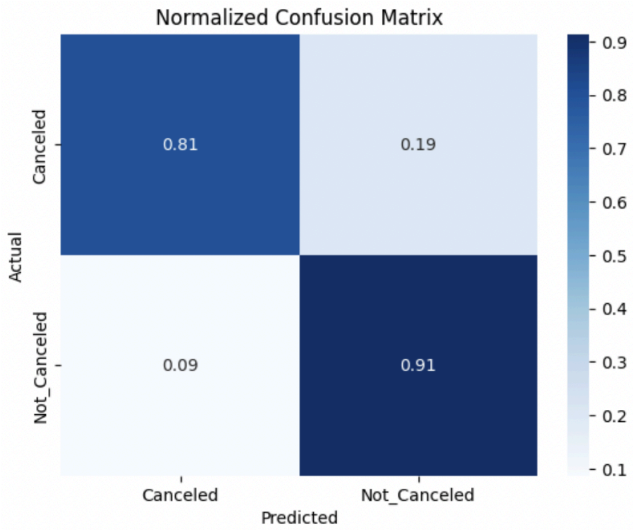
Results

We divided the dataset into a standard 80/20 split for training and testing. Overall, we were able to achieve an 88% accuracy. Moreover, precision, recall, and F1-score are all the same, which indicates that the number of instances in each class is balanced.



Visualization of the first levels of the decision tree.

We now need to understand how this model performs in each possible scenario. Below, we can see how we visualized a confusion matrix using a heatmap.



Confusion matrix.

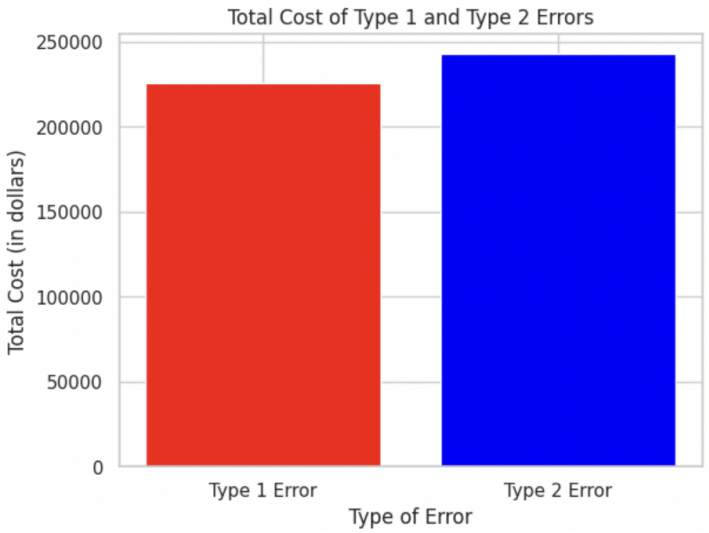
When making a prediction, we have a 19% chance of encountering a type 1 error, or false predicted booking, which leads to lower rate of occupancy and revenue loss, and a 9% chance of encountering a type 2 error, or false canceled booking, which leads to overbooking. We calculated the total cost of each type of error by using the following code. We used two different average prices for confirmed and and canceled bookings due to average room price being a factor in the likelihood of room cancellation.

These results do not include other potential costs associated with overbooking and mismatched resource planning that are unavailable to us.

```
# Use confirmed room price for Type 1 Error
cost_type_1_error = average_cost_confirmed * percentage_type_1_error
# Use canceled room price for Type 2 Error
cost_type_2_error = average_cost_canceled * percentage_type_2_error
# Total costs
total_cost_type_1_error = cost_type_1_error * canceled_bookings
total_cost_type_2_error = cost_type_2_error * (total_reservations - canceled_bookings)
```

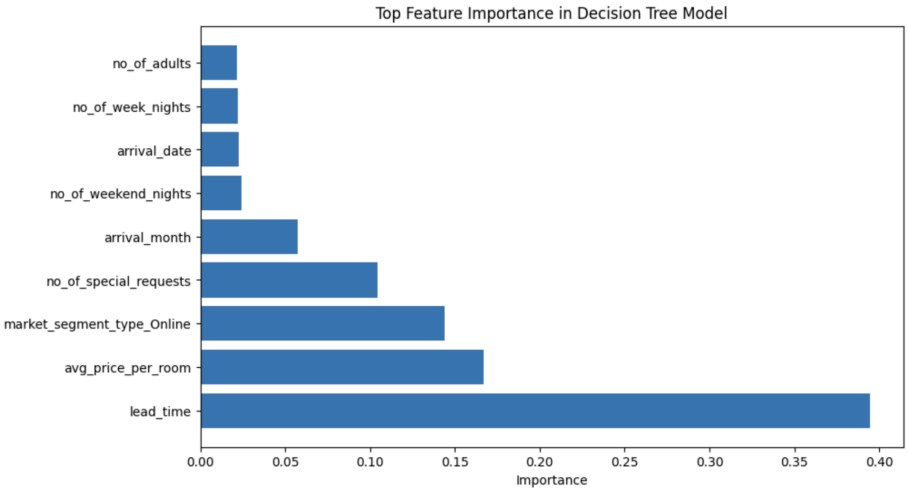
Total cost of Type 1 Error: $225660.12

Total cost of Type 2 Error: $242756.04



Overall, we see that the direct costs associated with type 1 error are slightly lower, and this outcome may be considered more favorable by management.

Below you can see the visualization of the importance of the top features of the model, or which variables had the most impact when predicting booking cancellations.



We can see that lead time is the variable most related to hotel cancellations, and the average room price, market segment, number of special requests, and even aerial month also have a considerable impact.

Conclusion And Final Recommendations

The predictive model can be used for better resource planning and financial forecasting. By predicting the likelihood of cancellation based on other reservation variables the hotel can improve internal processes and inventory management. We can also use estimated costs associated with type 1 and type 2 errors to draw further implications about their associated outcomes. While the direct costs of false confirmed bookings are lower, there could be other costs associated with both lower occupancy and overbooking, which we do not have available based on our current dataset. Some of this data, like inventory costs, could be easily collected, but there could be other costs such as damage to brand reputation. Management should consider these factors when deciding how to implement this model strategically.

Moreover, the hotel can focus on time and price-sensitive offers. Implementing personalized offers can be used as a remarketing and retention tool to drive customer engagement and improve confirmed bookings. By using lead time and seasonal cancellation data along with the likelihood of cancellation we can leverage personalized offers to increase the likelihood of true confirmed bookings and conversions.

Lastly, the hotel can implement a comprehensive reporting system to track the effectiveness of personalized offers and error mitigation strategies. Conducting regular

audits to assess the impact of implemented changes can furthermore lead to the identification of areas for further improvement and revenue maximization.

In conclusion, we were able to take a real-world dataset containing data about hotel cancelations and do extensive data analysis to understand the patterns and trends that lead to cancelation. We created a decision tree model for forecasting and identified the most impactful variables in predicting booking cancelations. From these analyses, we were able to understand the key metrics that can be influenced to reduce booking cancelations for this hotel. While we were able to give targeted recommendations, there are some limitations that we must keep in mind. This data does not have really important metrics such as the actual price of each booking as compared to the average price for the booking. The analysis would be more complete if the dataset also captured the price of canceling the booking and if a premium was paid by the customer to have the option of free cancellation. The data is also for one hotel so there might be some very specific problems that his hotel has that may have influenced the dataset, it is not a representative sample for the overall hotel industry.

Sources

Dataset:

https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset/data