**Data Mining - Project Report**

**Molly Siebenaler, Regan Day, Mollie Cox, Lindsey Bolton**

**1 Executive Summary**  
The main goals of this project are to build a predictive model and provide new strategic direction for the hotel. The model we built determines the most important factors in whether a guest cancels their hotel room. Based on our analysis, we will recommend strategies and policies to minimize the number of cancellations and the loss of revenue from cancellations. Our model determined that lead time, average price of hotel room, and market segment type are the three most important factors. A higher lead time, more expensive hotel room, and corporate groups make it most likely that the reservation will be cancelled. This report builds on these ideas with the background of the business problem, the goals of the project, a description of the data used, exploratory analysis results, our data mining solution, and our final conclusion.

Our recommendations revolve around the idea of implementing the predictive model on the hotel’s website to determine whether or not a customer is likely to cancel before they make their reservation. Additionally, we suggest that the hotel creates a loyalty program to create more repeat customers since they are less likely to cancel their reservation. Finally, we recommend targeting corporate groups by offering incentives and personalized services or raise cancellation fees to decrease chances of cancellation. These are all data-driven strategies that our analysis found to potentially lower the number of cancellations occurring at this hotel.

**2 Problem Description**

**2.1 Background:**

In our initial analysis, we determined that 33% of rooms have been cancelled at this hotel in years 2017-2018 (Raza, 2023). This significant portion of cancellations was the primary reason our client hired us. They are hoping we can predict whether a customer will cancel their reservation prior to their check-in date to increase their profits.

In this project, we will determine what strategies the hotel can implement to encourage customers to keep their reservation. Initial possibilities for these suggestions could include promotions, cancellation fees, loyalty programs, etc. These strategies will specifically target the specific problem of high cancellations but could also encourage overall customer loyalty to the hotel.

**2.2** **Business Goal and Data Mining Goal:**

In the hotel industry, a high number of room cancellations leads to decreased profits. Therefore, the business goal of this project is to keep hotel rooms fully booked all throughout the year to maximize profits. We will determine and suggest strategies and policies that discourage cancellations to ensure higher profits, while maintaining customer satisfaction.

Similarly, the data mining goal of this project is to use our analysis to narrow down the most important factors for whether someone cancels their booking. Based on what the model and data show, we will connect relevant strategies that hotels could implement to prevent people from cancelling their reservations. This directly answers the business goal previously described by keeping rooms full, thus increasing profits for the hotel.

The prediction target is booking status because this determines whether the customer cancelled their booking or not. We will use the prediction target to build our models, along with various other factors, to determine the most important elements of a reservation for cancellations.

**3 Data Description**

**3.1 Data:**

The data source that was analyzed is the [“Hotel Reservations Dataset”](https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset) from Kaggle. Kaggle is licensed and is a public domain which has public access. There are 18 features and 36,275 rows of data from 2017 and 2018. We discarded booking\_ID, arrival\_year, arrival\_date, and no\_of\_special\_requests because we determined they are irrelevant to our analysis. We are using the 16 most important features in our analysis. Some of these include the number of adults and children included in the reservation, the number of weekend and week nights booked, type of meal plan selected, if they reserved a parking space, the room type they reserved, how many days in advance they booked, the month they arrived, the market segment type, if they’re a repeated guest at this hotel, if they had previous cancellations, the average price of room they reserved, and if they had any special requests. Finally, our target variable is booking status, which indicates if the booking was cancelled or not.

*Data Dictionary*

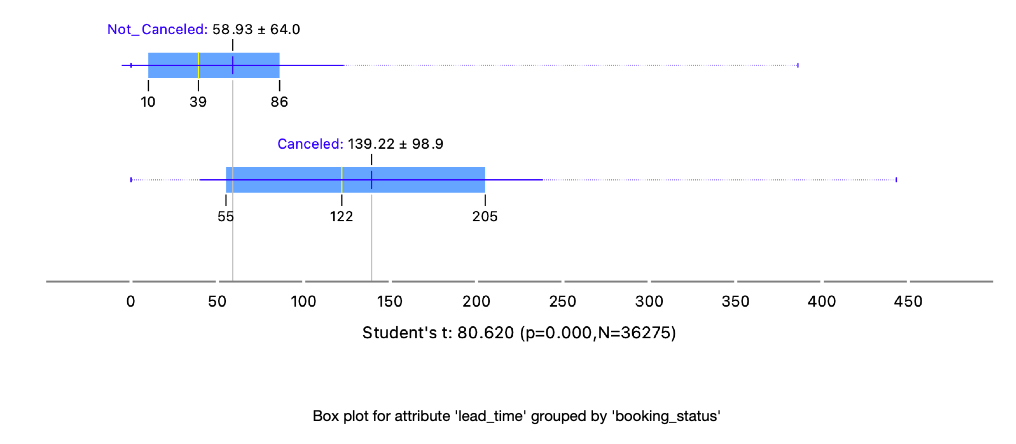
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| no\_of\_adults | Numeric | Number of adults staying in hotel room booking |
| no\_of\_children | Numeric | Number of children staying in hotel room booking |
| no\_of\_weekend\_nights | Numeric | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| no\_of\_week\_nights | Numeric | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| type\_of\_meal\_plan | Character | Type of meal plan booked by the customer (Meal Plan 1 – Breakfast, Meal Plan 2 – Breakfast and one other meal, Meal Plan 3 – Breakfast, Lunch, and Dinner, 0 – None Selected) |
| required\_car\_parking\_space | Numeric | Does the customer require a car parking space? (0 is no, 1 is yes) |
| room\_type\_reserved | Character | Type of room reserved by the customer (Types 1-7) |
| lead\_time | Numeric | Number of days between the date of booking and the arrival date |
| arrival\_month | Numeric | Month of the arrival date |
| market\_segment\_type | Character | Market segment designation (Aviation,  Complementary, Offline, Online, Corporate) |
| repeated\_guest | Numeric | Is this customer a repeated guest? (0 is no, 1 is yes) |
| no\_of\_previous\_cancellations | Numeric | Number of previous bookings that were cancelled by the customer prior to the current booking |
| no\_of\_previous\_bookings\_not\_canceled | Numeric | Number of previous bookings not cancelled by the customer prior to the current booking |
| avg\_price\_per\_room | Numeric | Average price per day of the reservation (dynamic) |
| no\_of\_special\_requests | Numeric | Total number of special requests made by the customer |
| **booking\_status** | **Character** | **Flag indicating if the booking was cancelled or not** |

**That target variable in this dataset is booking\_status.**

**3.2 Exploratory Analysis:**

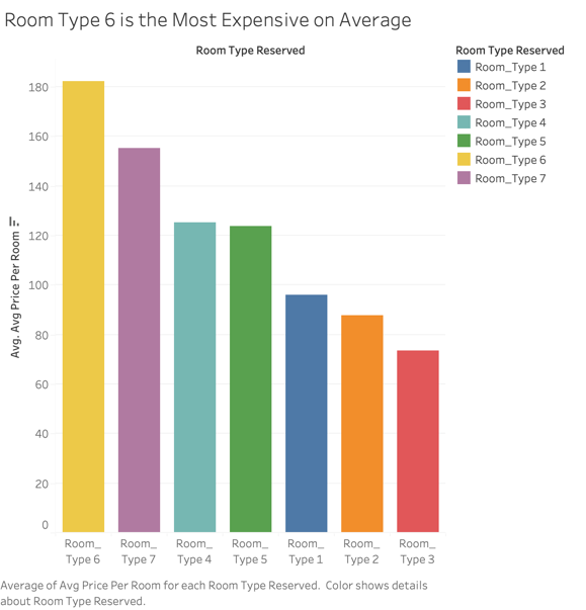
***Figure 3.21***

**Hotel reservations booked closer to a guest’s arrival date are less likely to be canceled.**



*The average lead time for reservations that are not canceled (58.93 days) is less than the lead time for reservations that are canceled (139.22 days).*

***Figure 3.22***



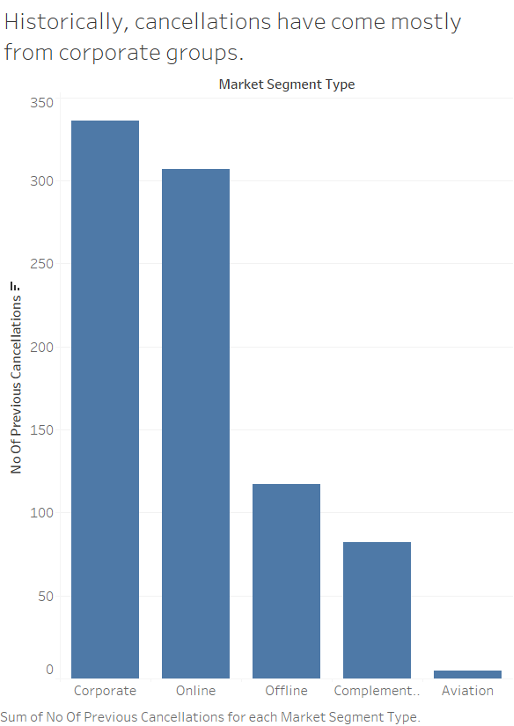
***Figure 3.23***

**While Room Type 1 has the most bookings, it also has the highest cancellation frequency.**

A screenshot of a graph

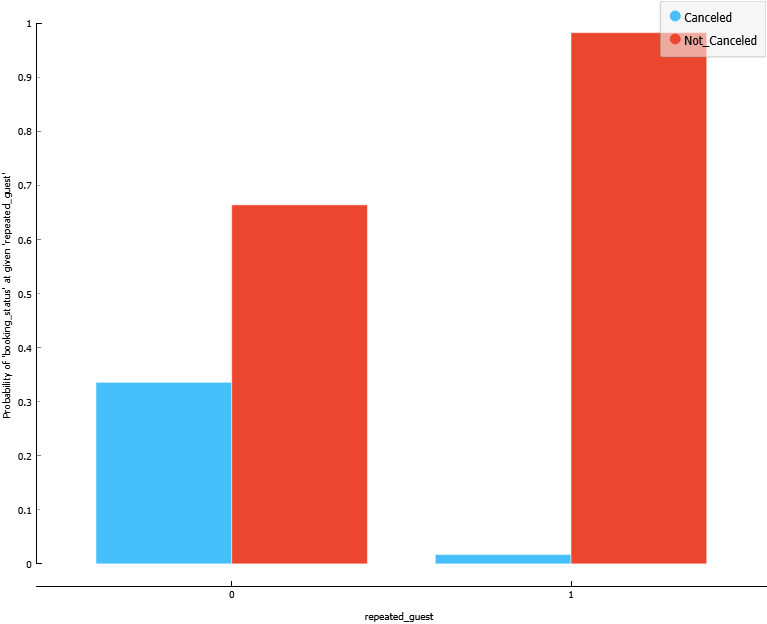
Description automatically generated

***Figure 3.24***



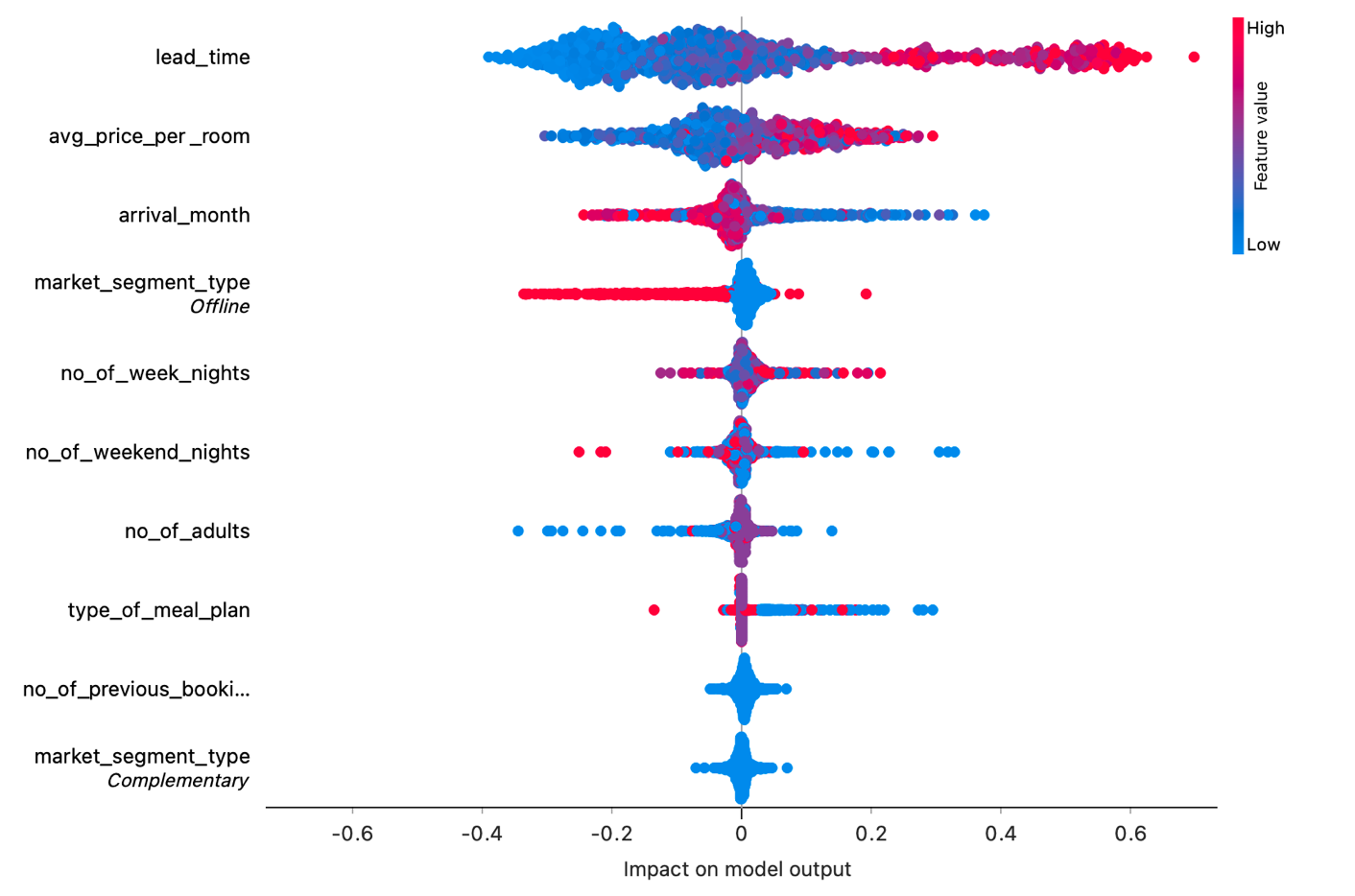
***Figure 3.25***

**Repeat guests are much less likely to cancel their hotel reservations.**



***Figure 3.26***

**Lead time, price, and arrival month have the largest impact on whether a guest will cancel their room.**



**3.3 Data Preprocessing:**

The data was preprocessed in Orange using the pre-process widget, to continuize discrete variables. Additionally, the type\_of\_meal\_plan feature was changed from categorical to numeric in Excel manually. The data was converted from “Meal Plan x” to “x”.

**4. Data Mining Solution**

**4.1 Models:**

Since our business problem looked to predict whether a hotel reservation would be cancelled or not, we trained 4 models that supported classification problems. We used the following models: gradient boosting, random forest, neural network, and a decision tree.

* The gradient boosting model had the following parameters: 108 trees, 0.011 learning rate, depth of 4, and subsets larger than 4.
* The random forest model had the following parameters: 90 trees, 10 attributes considered at each split, depth of 6, and subsets larger than 6.
* The neural network model had the following parameters: 2 hidden layers with 4 neurons in each, and a maximum of 100 iterations.
* The decision tree model had the following parameters: a minimum of 30 instances in the leaves, subsets larger than 20, maximum depth of 200, and must stop when majority reaches 95%.

**4.2 Performance Evaluation:**

The models received the following performance metrics during a 10-fold cross-validation training:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **CA** | **F1** | **Precision** | **Recall** | **MCC** |
| Neural Network | 0.883 | 0.822 | 0.819 | 0.819 | 0.822 | 0.587 |
| Decision Tree | 0.920 | 0.858 | 0.857 | 0.856 | 0.858 | 0.675 |
| Random Forest | 0.906 | 0.852 | 0.847 | 0.852 | 0.852 | 0.656 |
| Gradient Boosting | 0.890 | 0.834 | 0.824 | 0.838 | 0.834 | 0.612 |

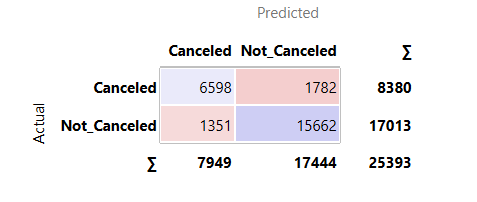
The models received the following performance metrics after performing on our training data set:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **CA** | **F1** | **Precision** | **Recall** | **MCC** |
| Neural Network | 0.885 | 0.825 | 0.821 | 0.822 | 0.825 | 0.593 |
| Decision Tree | 0.943 | 0.877 | 0.876 | 0.875 | 0.877 | 0.718 |
| Random Forest | 0.911 | 0.857 | 0.853 | 0.856 | 0.857 | 0.668 |
| Gradient Boosting | 0.893 | 0.836 | 0.826 | 0.840 | 0.836 | 0.616 |

The models received the following performance metrics after performing on our mock testing set:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **CA** | **F1** | **Precision** | **Recall** | **MCC** |
| Neural Network | 0.878 | 0.819 | 0.815 | 0.815 | 0.819 | 0.573 |
| Decision Tree | 0.916 | 0.851 | 0.850 | 0.850 | 0.851 | 0.655 |
| Random Forest | 0.903 | 0.855 | 0.851 | 0.853 | 0.855 | 0.658 |
| Gradient Boosting | 0.888 | 0.833 | 0.823 | 0.835 | 0.833 | 0.602 |

Based on these metrics, the **decision tree** provides the best model for our data. It has the highest performance metrics overall and does not overfit the data.



The confusion matrix for the decision tree further explains the performance of our model.

**5 Conclusion**

**5.1 Recommendations:**

In summary, we have the following recommendations for our client:

1. We have determined that our decision tree model can accurately predict whether or not a customer will cancel their reservation, based on information they provide prior to officially booking. We recommend they implement this predictive model onto their reservation website, so the model can determine how likely it is that they will eventually cancel their reservation. We will open it up to our client to determine how to best implement this model, based on their ethics and business values. One potential option is raising cancellation fees and rates for specific groups, which would most likely be corporate groups based on our analysis below.
   1. We based this conclusion off our ***Data Mining Solution Models***, which we determined that the decision tree had the best results. Knowing that this decision tree model led to high performance metrics (for testing, training and 10-fold cross validation), this model could predict chance of cancellation before the reservation is made.
2. We advise the implementation of a loyalty and reward program, incorporating a points system. This initiative aims to incentivize repeat guests, fostering a sense of loyalty and encouraging continued bookings.
   1. We based this conclusion off our findings in ***Figure 3.25***, which shows that repeat guests are much less likely to cancel their hotel reservation. The goal of this recommendation is to create even more loyal customers.
3. Prioritizing repeat guests in the booking process, affording them the opportunity to reserve accommodations further in advance. This prioritization is found in the observation that guests with a history of stays at the hotel demonstrate a higher likelihood of commitment, reducing the risk of cancellations.
   1. We based this conclusion of our findings in ***Figure 3.21***, ***Figure 3.25***, and ***Figure 3.26***. Figure 3.27 showed us that lead time had the highest impact on our target variable. The higher the lead time is, the more likely it is that the customer will cancel their reservation. As Figure 3.21 shows, a higher lead time results in a higher likelihood of cancelling. However, Figure 3.26 implies that repeat guests are less likely to cancel their reservations. Offering the perk of reserving further in advance to loyal customers could further increase loyalty to the hotel.
4. To attract and retain new guests, the hotel should consider implementing tailored incentives. This may involve special promotions, discounts, or exclusive perks to make the initial experience exciting and reduce the likelihood of cancellations among first-time visitors. Additionally, a focus on corporate groups presents an avenue for minimizing cancellations. Offering incentives and personalized services for corporate bookings can act as a deterrent to cancellations, as group reservations often entail more stringent planning and coordination.
   1. We based this conclusion off our findings in ***Figure 3.24.*** We noticed that corporate groups are much more likely to cancel their reservation than any other customer segment. Specifically targeting these incentives to corporate accounts could build loyalty, both within the company and its employees.
5. Restructuring what Room Type 1 and Room Type 4 offer guests and how the value proposition is communicated to customers will decrease the frequent number of cancellations coming from these two room types. These two room types are most frequently booked by customers, but also see a greater percentage of cancellations compared to the other types of rooms. Although the cause of this would need to be evaluated in future work, we can determine that there is a problem with these rooms from the consumer’s perspective.
   1. We based this conclusion off our findings ***Figure 3.22*** and ***Figure 3.23***. It’s evident in Figure 3.22 that Room Types 1 and 4 are neither the most expensive nor least expensive room offerings, so it’s likely something other than room price that is impacting the high percentage of cancellations. In Figure 3.23, the frequent pattern of bookings and cancellations for these two rooms is evident. We would need to gather more information from the hotel to determine what these room types mean and how they could impact customer cancellations.

**5.2 Limitations:**

The dataset encountered a challenge due to the absence of explicit descriptions for room types, making it challenging to make out the distinctions among the seven different categories. Specifically, crucial details such as the number of beds, bathroom configurations, and other amenities were not provided. This lack of detailed information hampered the ability to accurately interpret and compare the various room types. A more comprehensive dataset, including explicit specifications for each room type, such as the quantity of beds and bathroom features, would significantly enhance the precision and depth of insights.

Another significant limitation arises from external factors, like the COVID-19 pandemic serving as a prime example. Unforeseen events like global health crises can dramatically alter travel patterns, hotel occupancy rates, and customer behavior, introducing a level of unpredictability. The pandemic's widespread impact on the hospitality industry underscores the need for continuous adaptation and recognition of the dynamic nature of external influences.

**5.3 Future Work**:

Expanding the scope of customer engagement through personalized marketing initiatives could be a future opportunity for this hotel. This would be a way to increase customer loyalty and to increasingly build its brand. Additionally, utilizing customer data from the implemented loyalty program to tailor promotional offers, exclusive deals, and personalized experiences can strengthen the hotel's connection with its customers, driving repeat customers and positive word-of-mouth marketing. Finally, we could learn more about each room type contains and research what causes groups to cancel certain rooms if we continued to work on this project.

**Works Cited**

Raza, A. (2023, January 4). Hotel Reservations Dataset. Kaggle. https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset