Predict Hotel Booking Cancellation

A. Business Understanding

Hotels continually face the significant challenge of guests canceling their reservations before arrival. These cancellations lead to revenue loss, poor utilization, and operational inefficiency which is problematic given the low profit margins in the industry[1]. The significance of this problem is quantifiable through exploratory data analysis on our hotel demand booking dataset because of the 119,390 total bookings analyzed, 37% (44,224) were canceled before the reservation arrival date. The monetary estimate of these 44,224 bookings was estimated at 16.7 million euros through a calculation of the total cost of reservations the product of the average daily rate and night booked) canceled before the arrival date.

The data mining solution of predictive analysis would offer the best solution to address this business problem by predicting the likelihood that a consumer will cancel their reservation before their arrival date. The results of the predictive analysis will provide hotels with better insights allowing for better occupancy and revenue maximization.

B. Data Understanding

A dataset of hotel bookings from various countries between 2015 – 2017 from the open-source resource platform Kaggle was used to address the business problem. Before preparing and modeling the data, strengths, limitations, and biases were evaluated and visualizations were created exploring important aspects of the data. Strengths of the data included important predictive variables such as previous stays or cancellations, type of hotel, the month of the stay, and the average daily rate. Limitations of the data included the locations of the hotel. The hotels within the dataset were primarily located in Europe, with 40.6% located in Portugal, however, there were also hotels from other countries and continents. As a result of this, the variability in the location of these hotels may limit some of the predictivity power of the models. The largest bias associated with the data was due to seasonality/events. The presence of seasonal or one-time events occurring in the hotel locations could contribute to the unusual depiction of trends during a specific time period, which could result in seasonal, selection, or cultural bias based on the demographic population attending these events. The presence of this bias will attempt to be controlled for in our prediction models.

A series of data was created visually to better understand cancellations and the segments that were correlated with them. The first visualization analyzes the amount of hotel bookings and cancellation rates by hotel type. From Figure 1, there are more bookings in city hotels when compared to resort hotels, but the percentage of booking cancellations in city hotels is more than in resort hotels. We can assume that customers of city hotels are usually for business trips and tend to cancel their bookings more frequently in comparison to customers of resort hotels who usually have planned in advance for traveling purposes and have lower probabilities of canceling.

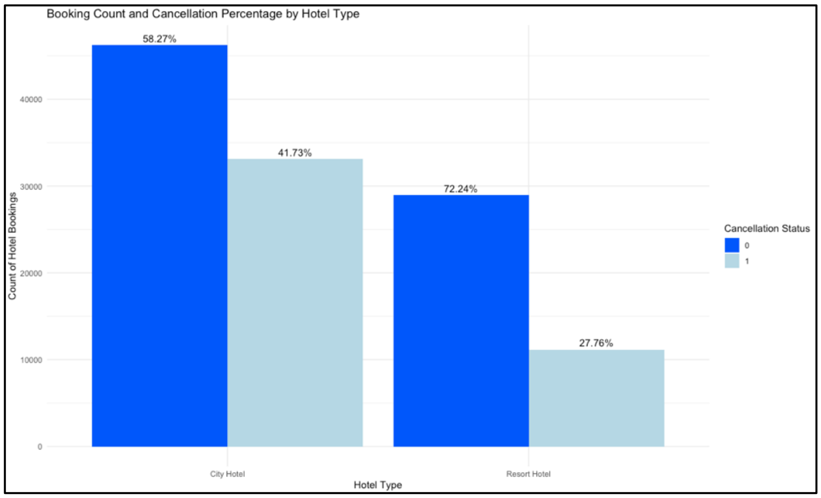

Figure 1

The second visualization evaluates the booking count and cancellation percentage by repeated guest status. From Figure 2, repeated guests are less likely to cancel the bookings which may be because the customers are loyal, or they might be frequent visitors for work purposes.

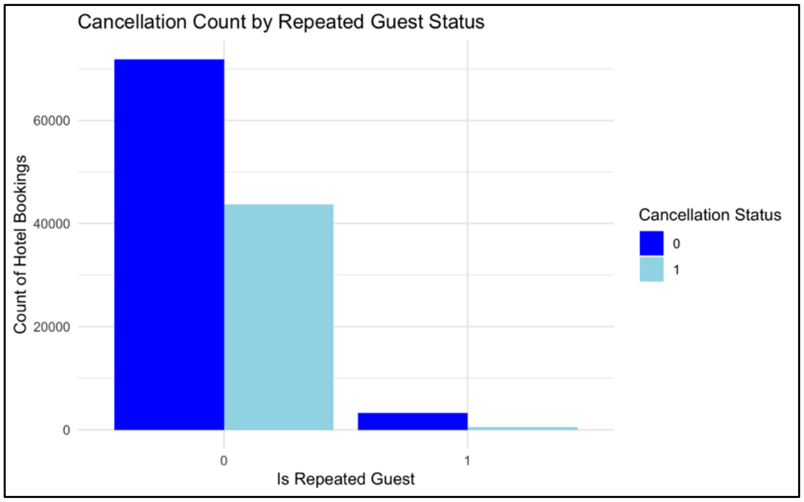

Figure 2

The third visualization shows the booking count for each month by hotel type. From Figure 3, August has the maximum number of bookings for both hotel types, and resort hotels tend to have a peak season whereas city hotels don't have obvious seasonality.

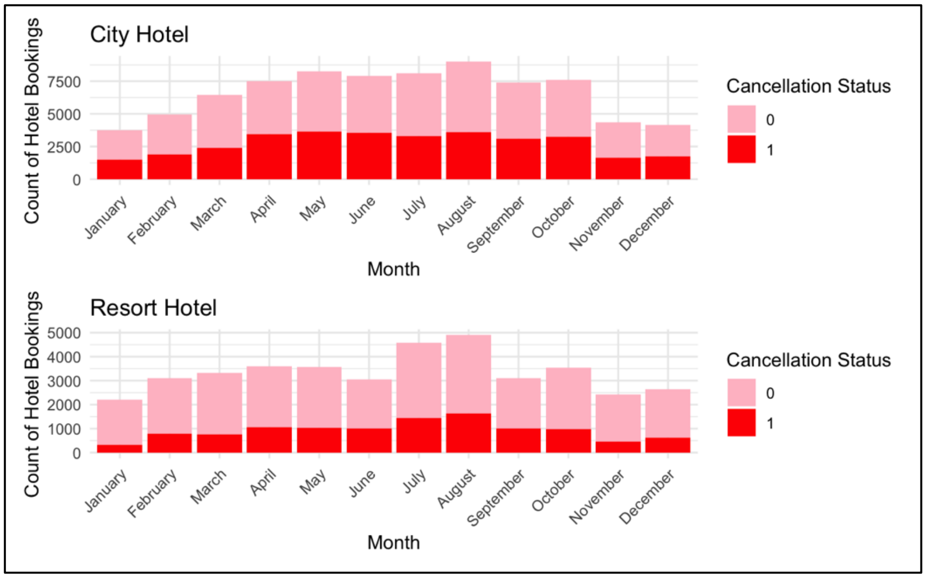

Figure 3

The fourth visualization displays the average daily rate by month. City hotels tend to have higher rates than resort hotels, however, during summer seasons, resort hotels have higher rates. The overall rate is high for both hotel types in vacation seasons.

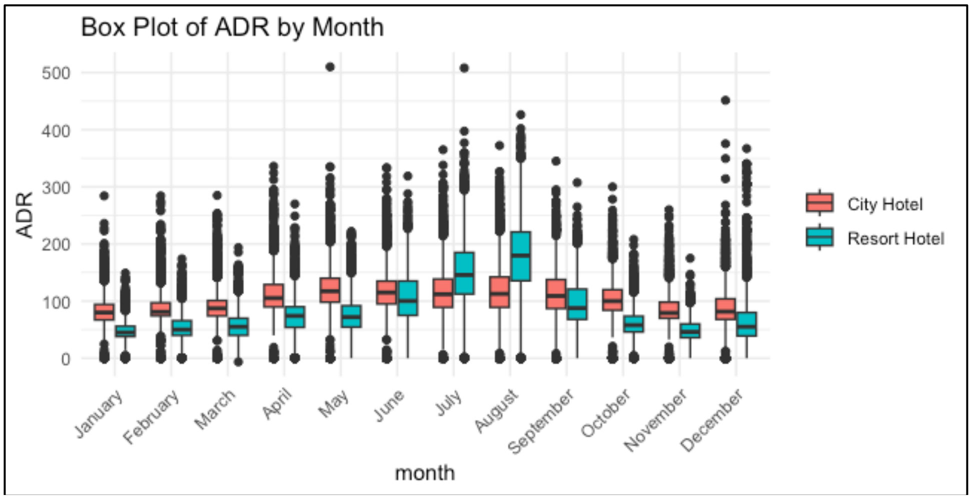

Figure 4

The fifth visualization shows the relationship between the average daily rate and cancellations. Interestingly, the hotel rate didn't significantly affect the cancellation action.

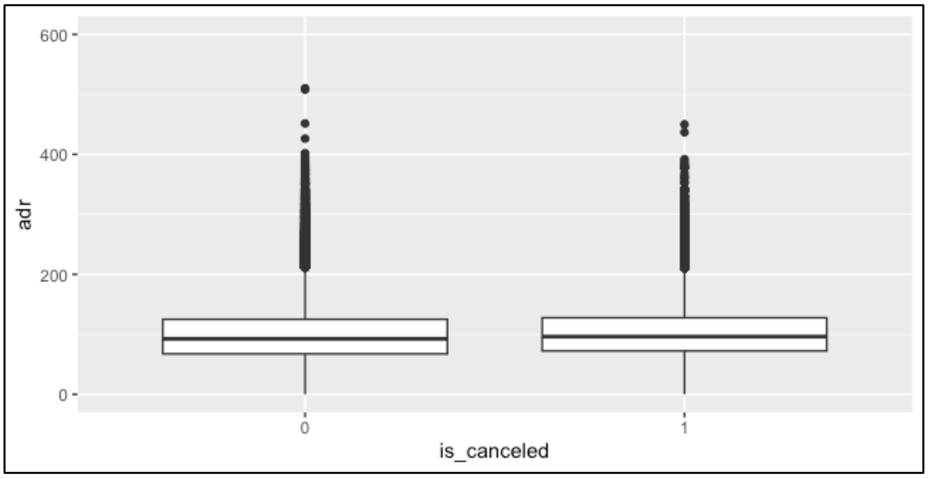

Figure 5

C. Data Preparation

The dataset contains booking information for city hotels and resort hotels, including information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces. There are in total 32 variables and 119,390 observations. The data preparation can be divided into three parts: missing data, NULL values, and variables filtering/transformation. In the first part of missing data, there are 4 N/A values in the children column, and this usually means there are no children, therefore, we replaced N/A with 0. In the second part, NULL values, there were 489 rows in the country column that had NULL values. They were replaced with UNKWN to better interpret the data. Additionally, there were 112,593 rows in the company column that had NULL values. Since the NULL percentage is too high (94%), we decided to drop this column. In the third part, variable filtering/transformation, we dropped columns that were not feasible to prevent data leakage. For example, reservation_status (check-in, no-show, canceled) shouldn’t be shown in the model as hotels wouldn’t know when they predict the cancellation. The Agent column had too many variables that may cause the model to overfit. Therefore, we created a new column called **is_agent** to check whether the customer booked through the agency or not. Lastly, the following variables were dropped: company, reservation_status, reservation_status_date, arrival_date_week_number, stays_in_weekend_nights, agent, arrival_date_year.

D. Modeling and Evaluation

The core task of the analysis is to predict whether a customer will cancel the hotel booking given the observed characteristics. Since the prediction is binary, we will conduct data mining for classification prediction. In this analysis, we will be using different supervised learning models including logistic regression, SVM, classification tree, and XGBoost. We also apply K-fold cross-validation to make use of all data points and lower model performance bias. The performance metric we chose is OOS Accuracy.

The result showed that the XGBoost model has the best OOS performance with an Accuracy of 84.09%, therefore, we will apply the XGBoost model for our prediction. With this model, we expected that the hotel companies could better predict whether a customer is going to cancel the booking or not given the characteristics recorded in the model.

Logistic Regression

```
reg_model <- train(is_canceled~., data = hotel.train, method = “glm”, family = “binomial”,
trControl = ctrl)
```

Accuracy	Kappa
0.8143185	0.5873026

SVM

```
svm_cv_model <- train(is_canceled ~ ., data = hotel.train, method = “svmRadial”, trControl =
ctrl, tuneGrid = data.frame(C = 10, sigma = 0.001))
```

Accuracy	Kappa
0.8284401	0.6180982

Classification Tree (Accuracy not available)

```
Tree.model <- train(is_canceled ~ ., data = hotel.train, method = “rpart”, trControl = ctrl)
```

cp	ROC	Sens	Spec
0.01830484	**0.8030059**	0.8762093	0.62882292

XGBoost

```
fit.xgbTree <- train(is_canceled ~ ., data = hotel.train, method=”xgbTree”, metric=”Accuracy”,
trControl=ctrl,tuneGrid=param)
```

Accuracy	Kappa
0.8416617	0.6519294

Please find the summary results located in the table below:

Since the classification tree model is giving us ROC instead of accuracy, we found the XGBoost model which is an ensemble learning method combining the predictions from multiple weak learners (typically decision trees) to create a strong predictive model. It also generated the best OOS accuracy result

Model	OOS Accuracy
Logistic Regression	81.43%
SVM	82.84%
XGBoost	84.09%

Using the XGBoost model, we made predictions of the cancellation rate. Based on the output from R, the accuracy is 84.378%.

Additionally, we applied K-means analysis in order to cluster customers and determine the similarities and characteristics of each group. Based on the clustering result and the insights from EDA we are able to delpoy strategies specifically dealing with the cancellation rate. The result of the K-Means analysis is shown below:

	is_ canceled	lead_time	arrival_date_day _of_month	adults	children	babies	is_repeated_ guest	adr
1	0.348	43.285	16.159	2.020	0.278	0.012	0.011	166.994
2	0.249	31.320	15.562	1.718	0.043	0.007	0.0642	75.273
3	0.451	160.728	16.040	1.911	0.096	0.007	0.004	103.189
4	0.644	326.526	15.488	1.945	0.047	0.003	0.015	85.041

Group 1: Customer groups with more children and babies compared to other groups. This group could be identified as families, and they tend to book the hotel in shorter lead times. The average daily rate is the highest maybe because they booked the family rooms.

Group 2: Customer groups with more repeat guests and book the hotel in shorter lead times. The average daily rate is the lowest. This group could be identified as business customers who tended to book the same hotel that has cooperated with companies. This group also has the lowest cancellation rate, which is aligned with our visualization result.

Groups 3 & 4: Customer groups who book the hotel in longer lead times. Those groups have a higher possibility of canceling the booking. With longer lead times, the customers are more inclined to cancel the booking.

The results of the data mining should be through the Accuracy/ROC. In the comparison of our models, the XGBoost model performed the best and should be used to predict the probability of customers canceling. Expected improvement should be measured by an increase in capacity utilization in these hotels. If hotels can better predict which customers will most likely cancel their reservations then they should be able to book more rooms and increase their capacity utilization. Determining an expected ROI would be difficult to predict because the cancellation rate will most likely remain the same, but hotels should be able to increase their capacity utilization. Because we cannot predict how much the capacity utilization would increase it would be difficult to estimate a precise ROI.

E. Deployment

Based on the analysis above, it seems that whether a customer will cancel the booking depends on hotel type, traveling purposes, booking time, family group, and if the customer is booking the hotel repeatedly. Longer lead time leads to higher probabilities of cancellation, and repeated customers tend to be loyal to the hotel and don’t easily cancel the booking. Most of the factors are related to customers’ traveling habits, purpose, and behavior rather than external factors such as hotel rates. We suggest that the hotels could use our above-suggested XGBoost model for prediction. They could also focus on customers’ booking purposes and booking behavior to determine whether certain customers will cancel the booking. By calculating the percentage of booking rate minus the cancellation, the hotels would be able to release a certain percentage of rooms for overbooking.

One of the biggest risks associated with this plan are the instances where hotels overbook rooms and cancellations are less than predicted. In this situation, the hotel would have more bookings than rooms which could create a serious problem for customers who need a room to stay in. The two ways to mitigate this risk are by setting a conservative threshold for overbooking and partnering with other local hotels. Setting a conservative threshold on the % of rooms that can be overbooked limits the chances of the hotel booking too many rooms each night. Additionally, by partnering with other local hotels, a hotel can send overbooked guests to other local hotels for a lower price than they would have paid at their hotel. The predictive model established in this project will allow hotel companies a better mechanism for maximizing their occupancy rates and profit.

References

1. Verot, Benjamin. Everything You Need to Know About Hotel Cancellations.

   HotelMinder.com. September 19, 2023.

2. Hotel Booking Demand Dataset. Kaggle.com