STOR 320 Final Project

Walker Burgin, Tara Ghorpadkar, David Snider, Sid Vanam

7/27/2021

Vila Vita Resort in Algarve, Portugal. Provided by The Leading Hotels of the World.

Introduction

Our group used data describing hotel bookings from two hotels in Portugal to answer questions relevant to the travel industry. Our first goal was to predict room rates with confidence. Such a tool would be valuable for travel agencies, hotels, and customers. Although hotels can provide rates only for a finite number of months, a model that predicts future rates may enable travel agencies to book vacations further into the future. Using this model, travel agencies would better serve as a middleman by providing more utility to customers and hotels. Customers could have more booking options with greater price certainty, while hotels could increase ease of transaction. Hotels could also understand seasonal and cyclical consumer demand patterns earlier and thus prepare their budget accordingly.

Our second goal was to predict whether the customer would cancel their booking. In our dataset, about one-third (0.37) of bookings ended up getting canceled. Hotels that can accurately predict customer cancellation can proactively modify operations. Hotels could double book rooms that have high cancelation to try and still make a profit. Alternatively, predicting canceled books can help customers. For example, when asking if a customer wishes to join a waitlist, the hotel can inform them of the likelihood of cancellation or the probability of getting that room.

Data

The dataset contains 119,319 bookings for two hotels located in Portugal over about two years. There are 32 variables included in the data set; our research questions concerned a select few. **Our first question uses `"arrival_date"` to predict the variable `"avg_adr"`.** Our group combined the variables `"arrival_date_year"`, `"arrival_date_month"`, and `"arrival_date_day_of_month"` into the variable `"arrival_date"`, to record the client's date of arrival in elapsed days since July 1, 2015 (the earliest arrival date). For modeling, we converted `"arrival_date"` into a Date type, and used a function to obtain days elapsed from 07/01/2015. We created the variable `"avg_adr"` using the variable `"adr"`. `"adr"` is the client's total transaction costs divided by their staying days. In creating `"avg_adr"`, we calculated the mean ADR value for one day for each hotel. After establishing these variables, one can easily observe the increasing oscillating relationship between `"avg_adr"` and `"arrival_date"`, which motivated our first question.

Table 1: Example Data for Q1

arrival_date	avg_adr
0	96.56
1	58.67
2	74.53
3	63.78

Our second question predicts the variable `"is_canceled"`, which has a value of 1 if the booking was canceled, and 0 otherwise. Analysis revealed that certain variables had a statistically significant relationship with is_canceled. The following are the variables we analyzed. `"hotel"` describes whether the observation was at the city hotel or the resort hotel. `"is_repeated_guest"` is a binary variable, valued at 1 if the guest has booked previously. `"lead_time"` describes the number of days between entrance into the booking system and the customer's arrival. `"previous_bookings_not_canceled"` describes the number of the client's previous bookings that were not canceled, while `"previous_cancellations"` describes those that were canceled.

Table 2: Example Data for Q2

is_canceled	hotel	is_repeated_guest	adr	lead_time	previous_bookings_not_canceled	previous_cancellations
0	Resort Hotel	0	98.00	14	0	0
0	Resort Hotel	0	98.00	14	0	0
0	Resort Hotel	0	107.00	0	0	0

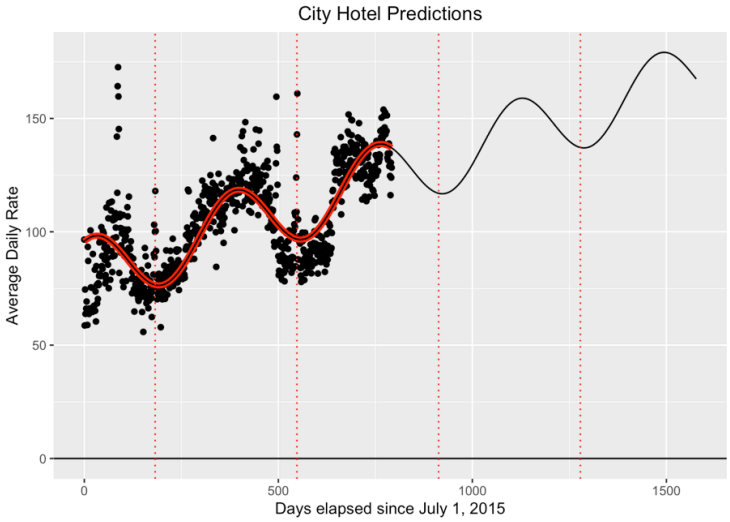
0	Resort Hotel	0	103.00	9	0	0
1	Resort Hotel	0	82.00	85	0	0

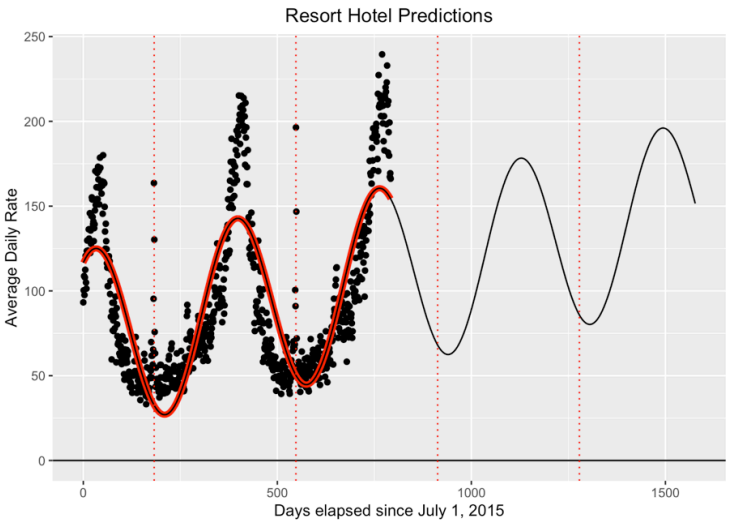
Results

Question 1: Can we predict future room rates with confidence?

To answer our first question, which predicts the average daily rate for both the city and the resort hotels for dates not in the dataset, we first built a model on the data that we had in the data set. We noticed that the data trended in an oscillation - the average daily rate per day increased during summer months and decreased during winter months. These peaks and troughs resembled a sinusoidal graph, so we decided to model the data sinusoidally. The data also had a linear component to it as well. From 2015 to 2017, the peaks and troughs also gradually increased, indicating that the hotels were getting slightly more expensive per year, and we combined this as part of our sinusoidal function to create the most accurate model. To represent the sinusoidal nature in a line of best fit for this graph, we first converted the days elapsed into sin and cos values, and then used a fitting function on the data to find the coefficients for the trend line used to model the data in the data set. Once we found the coefficients, we were able to predict the prices for future dates not in the data set. To do this, we created a specific model function, which used the coefficients determined from the previous fit to determine the line of best fit for future dates. We then binded two years' worth of rows to the original data set - all with adr values set to NA. Using the model function described earlier, we were able to predict average daily rates for the next two years by arrival date. As the trend line shows, the next two years follow the sine curve and also show a linear trend upward in price for both the resort and city hotel, though this upward trend is more pronounced in the city hotel.

The city hotel prediction model indicates that average daily rate per day is higher during summer months than winter months, as the difference in price between the peak price (usually in July) and the lowest price (usually in January) within a year is around 60 - 70 units. However, this difference is much starker in resort hotels. The resort hotels follow a similar pattern, where the average daily rate is higher in summer months than in winter months, but the range between the highest and lowest values within the year is 100 - 125 units. The data is unclear on whether price is measured in Euros (currency of Portugal), or US Dollars, so for analysis of the data, "units" would be a more appropriate term for the measurement of money. This larger difference in resort hotel data is also accompanied by a steeper decline, so that indicates that the price remains high for a shorter amount of time than the city hotels. This also contributes to the fact that city hotel data is more rounded and looks like a series of hills, while the resort data looks sharper and looks more like a series of mountains. This could be due to the fact that resort hotels specifically target customers looking to vacation which usually occurs during summer months. However, city hotels could have vacationers, but could also have customers traveling for work-related or other reasons. Either way, the prediction model helps customers to see when prices will be the highest and to look out for periods where the average daily rate will decrease, and it helps hotel management determine when their revenues from hotel rates will be the highest or lowest and help prepare for when they shift from one specific rate pattern to the next.



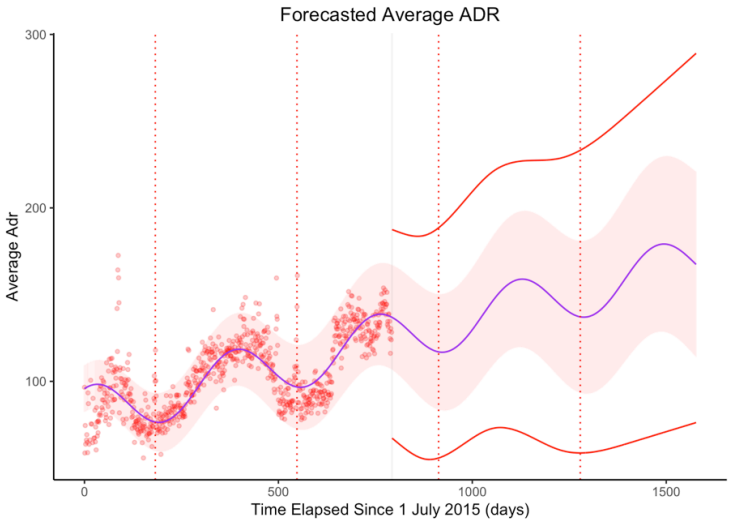


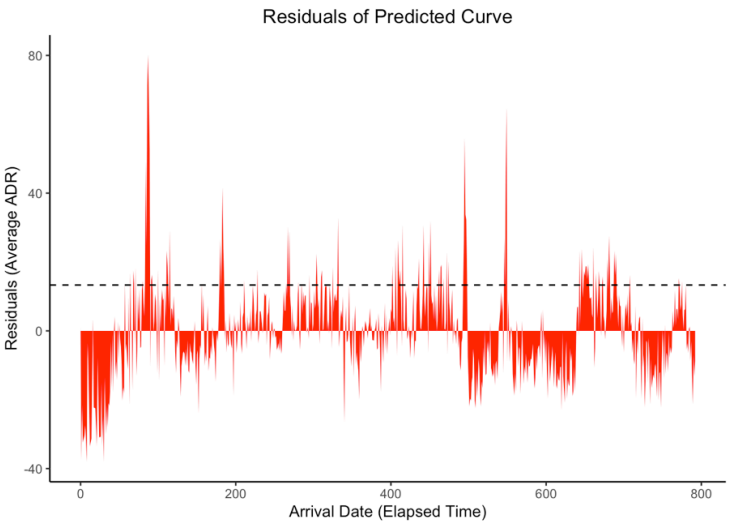
The mathematical equation for city hotel data:

$$14.66\cos(\frac{2\pi x}{365.25}) + 5.60\sin(\frac{2\pi x}{365.25}) + 0.0553x + 81.10$$

The mathematical equation for resort hotel data:

$$46.47\cos(\frac{2\pi x}{365.25}) + 26.28\sin(\frac{2\pi x}{365.25}) + 0.0486x + 70.13$$





Our prediction model is displayed in “Forecasted Average ADR”, with confidence intervals defined by the RMSE of 13.31. Our second chart, “Residuals of Predicted Curve”, identifies which segments of the prediction model are less accurate than others at mapping the dataset.

Question 2: Which model best predicts cancellation?

How do significant variables correlate with cancellation?

To determine which variables we should use in our models, we started by identifying the variables that had a significant relationship with cancellation. Pictured are graphs showing the significant variables' relationship with cancellation, along with a table of P values. The P values for the first 4 variables were calculated using a T-test. The P values for the last 2 were calculated using a two proportion Z-test.

Relationship between is_canceled and other significant variables

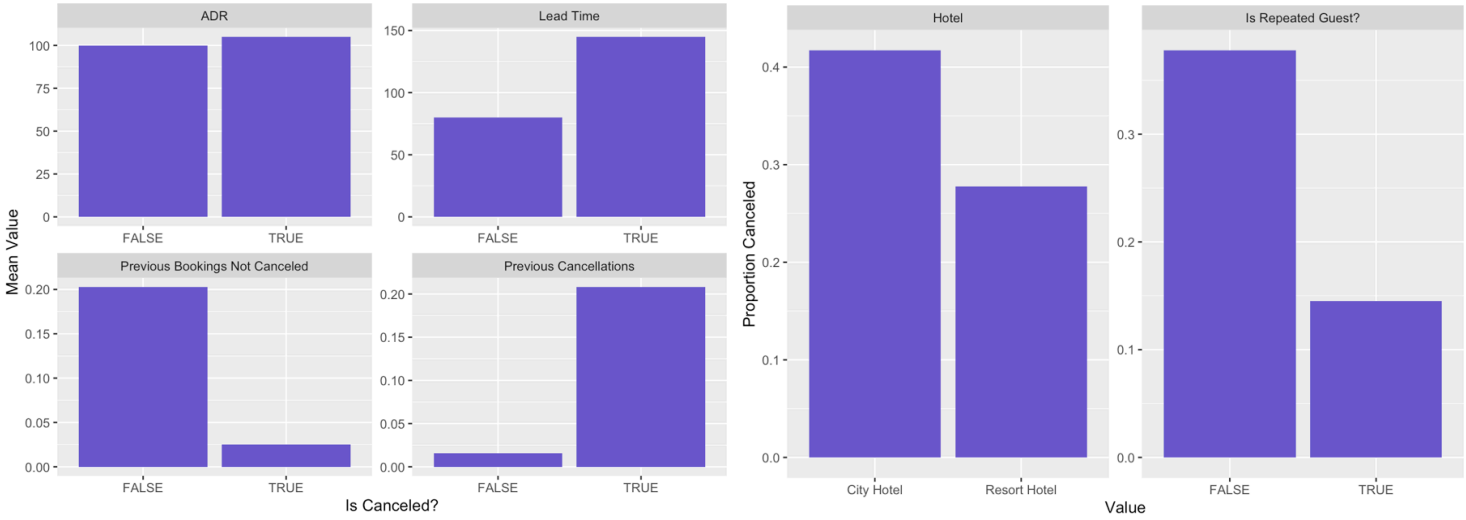


Table 3: P Values from 2-proportion Z Test

Variables	P.values
Average Lead Time	0.00e+00
Average Previous Cancellations	3.44e-196
Average Previous Bookings Not Canceled	5.88e-129
Average ADR	9.76e-59
Hotel	0.00e+00
Is Repeated Customer?	1.78e-188

Model Selection:

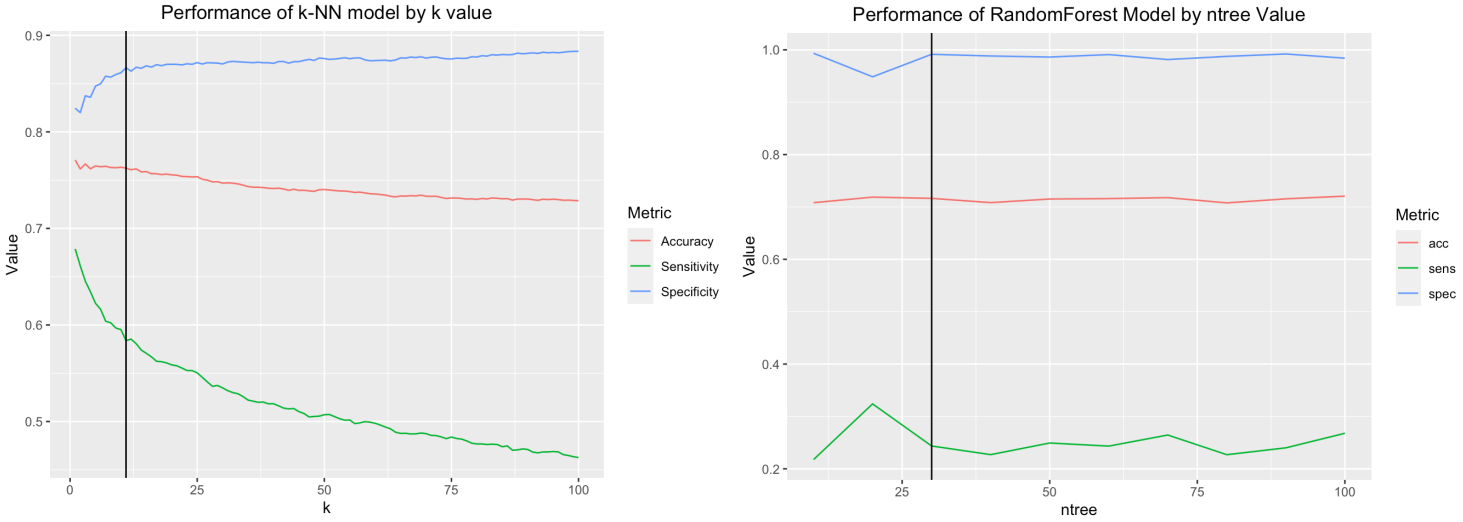
Using the significant variables, we created multiple predictive models: logistic, step-wise logistic, logistic with two-fold interaction, k-NN, and RandomForest. To provide an example of one of our models, Table 4 gives coefficients for the simple logistic model.

Table 4: Coefficients of Simple Logistic Model

Term	Estimate	P.value
(Intercept)	-1.3829	0.000e+00
hotelResort Hotel	-0.4658	3.616e-189
lead_time	0.0049	0.000e+00
previous_cancellations	2.9150	0.000e+00

previous_bookings_not_canceled	-0.6373	1.779e-108
adr	0.0037	4.648e-132
is_repeated_guest	-1.0641	7.943e-34

As for the k-NN and RandomForest models, we needed to determine which k and ntree values we should use. To determine which k value we should use for our k-NN model, we analyzed changes in accuracy, sensitivity, and specificity over time. In the context of our situation, it is more important that hotels correctly classify those who did not cancel, because we want to minimize the hassle from overbooking rooms that weren't actually canceled. However, we still want to maintain accuracy. Thus, we care more about accuracy and specificity than sensitivity. Accordingly, we chose k=11, because it maximizes accuracy×specificity. Below is a visualization of this analysis. As for the RandomForest model, there was little variation of metrics by ntree value. Nevertheless, we chose parameter ntree=30 because it maximized accuracy×specificity.



The logistic models performed the same in terms of accuracy, sensitivity, and specificity, so we consider only the simple logistic model here. We compared simple logistic, k-NN with k=11, and RandomForest with ntree=30. Table 5 gives the metrics of our models' performance.

Model	Sensitivity	Specificity	Accuracy	Spec_Times_Accuracy
1	0.19	0.98	0.69	0.67
2	0.58	0.87	0.76	0.66
3	0.25	0.99	0.72	0.71

Table 5: Model Performance Metrics

Conclusion

Our group chose to analyze two important questions relating to the two hotels sampled in our dataset. One question was: Can we use a model to predict hotel price? And the other was: What is the best model to predict cancellations? To answer the first question, we found that average daily rate has a strong sinusoidal relationship

with arrival date, and so we created a model that predicts future rates based off arrival date with confidence intervals. We answered our question by testing a variety of predictive methods and settling with the one that maximized accuracy times specificity. Its accuracy was 0.72 and its accuracy times specificity metric was 0.71.