Hotel Cancellation Prediction

Introduction

(a) Background

The hotel owner hired a new manager for his two hotels in the UK. One hotel is located in the city of London called the City Hotel, and the other one is located in a suburban area called Resort Hotel.

On the first day of the manager stepping into the two hotels, he found that the room occupation rate is under his expectation. He, then, checked the management system and found that many orders were being canceled.

The manager is not happy about this situation. He wishes to learn more about the business of the two hotels and figure out why the customers frequently cancel their reservations. The manager pulled out all the reservation data of the two hotels from the management system and sent it to the data department. Judy, the Data Scientist, received this case.

(b) Goal

The goal of this case is to figure out the reason why customers cancel their orders and reduce the cancellation rate. By figuring out the factors that may cause a customer to cancel the order, the owner could improve the service to prevent the cancellation. If we can figure out what type of customers has a higher possibility to cancel their reservation, the owner could push some policies on such customers to reduce the cancelation. Or if we can precisely predict if a customer will cancel the reservation, we could have some oversell policy.

Datasets

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020.

I downloaded this dataset from the Kaggle Hotel booking demand. This data set contains 119,390 observations and 32 columns of booking information for a city hotel and a resort hotel. It includes 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, among other things. The time period of this dataset is from July 2015 to August 2017.

Reference:https://www.kaggle.com/jessemostipak/hotel-booking-demand

Data Cleaning and Wrangling

(a) Duplicate rows

There are 31994 duplicate rows. As I explore these duplicate rows, I found these rows may not be useless information. It could be some identical customers or a group of customers making very similar orders. It is not wise to just drop all the duplicated rows. In this case, I will keep all these rows.

(b) Remove Incorrect rows

There are about 180 reservations without any adults or children. A reservation must make for at least an adult or a child, so I removed the 180 rows.

There is also one row with a negative average daily rate. This could be an error, and it is removed.

There are 1810 rows that have a 0 average daily rate. By exploring these rows, most of these reservations are labeled check-out, meaning these reservations have been completed. The 0 average daily rate could due to some promotions or events, so I decided not to remove these rows.

(c) Impute Missing Values

There are 4 missing values in the children column, 478 missing values in the country columns, 16280 in the agent columns, and 112441 in the company columns. Fill the missing values in the Children columns with 0. Impute the missing values in the Country column with “No Country” meaning that the customer is not willing to disclose where he/she come from. Fill the missing values in the Agent and Company columns with “No agent” and “No company” meaning that the order is not made through an agent or company.

(d) Add New Columns

Transform the text month column into a numerical month column. Transform the arrival date year, arrival date month, and arrival date day of the month into a DateTime column “arrival_datetime” for future analysis. I also transformed the reservation status date column into a DateTime column.

(e) Outliers

There is a significant outlier in the average daily rate where the rate is as high as $5400. This reservation is for 2 adults and for the most common room type. This order is also been canceled, this could mean that this order is canceled because of the price error. This row is been removed after analysis.

There are also lots of outliers in the days_in_waiting_list and lead_time (Number of days that elapsed between the entering date of the booking into the PMS and the arrival date). However, by seeing the bar plots of them are left-skewed, long waiting days are unusual, but could be valid and useful information. By considering that long waiting time or making the order too long ago could be a trigger to cancellation, this information should be kept.

By the end of the data cleaning process, the dataset has 119208 observations and 35 columns. Compared to the original dataset, I removed 182 observations and add 3 new columns.

Exploratory Data Analysis and Initial Findings

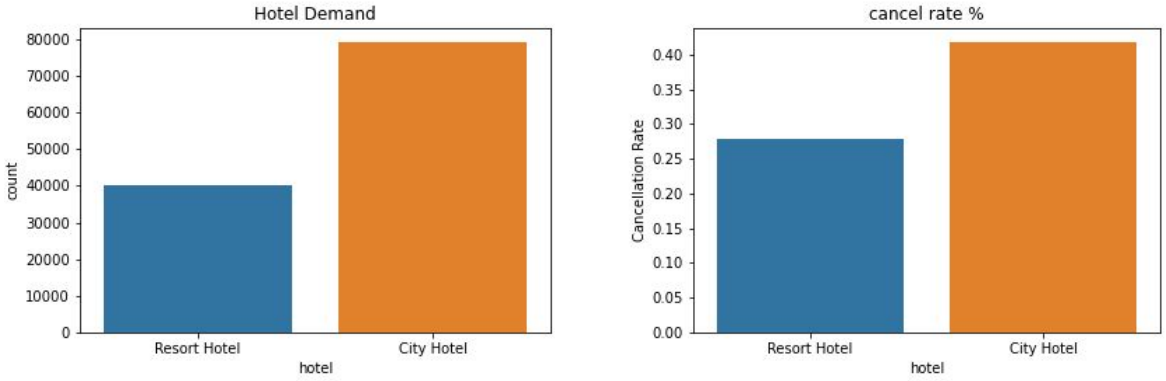
(a) Revenue and Losses Analysis

First, I explored the total revenue and loss due to the cancellation of both of the hotels.

- The overall revenue from July 2015 to August 2017 is $7,513,799 and total loss due to cancellation is $4,636,542 (61.7% of the total revenue)
- The revenue from July 2015 to July 2016 (one year period) is $3,109,679 and the loss due to cancellation is $1,752,480 (56.4% of the total revenue)
- The revenue from August 2016 to August 2017 (one year period) is $4,404,119 and the loss due to cancellation is $2,884,062 (65.5% of the total revenue)

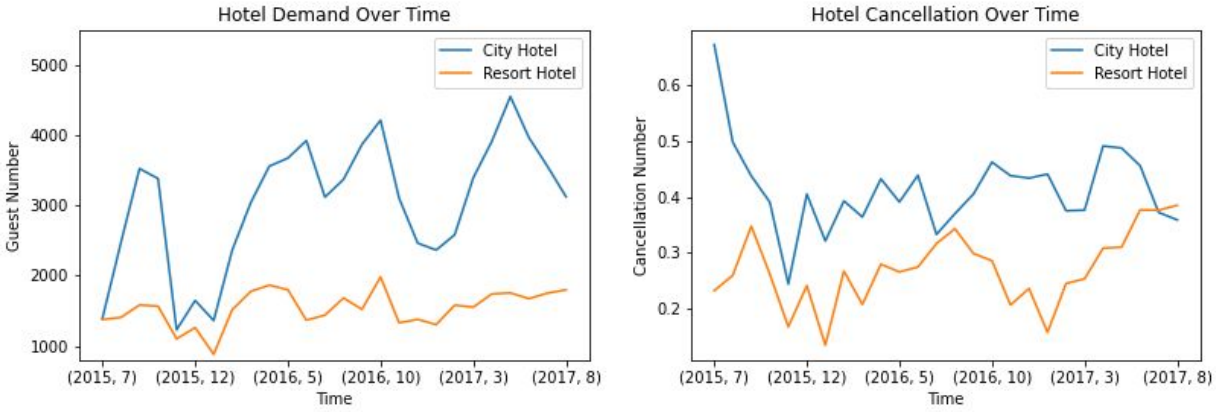
The loss due to cancellation is always more than half of the revenue. If we can find some ways to turn some of the cancellation losses into revenue, the hotel owner would be very happy.

From the hotel demand and cancellation analysis (2 figures shown below), we can tell that the City Hotel has more reservations (demand) than the Resort Hotel. However, the cancellation rate of the City Hotel is also higher than that of the Resort Hotel.



The following two figures show the hotel demand and the hotel cancellation rate over time (from July 2015 to August 2017). From the Hotel demand graph on the left, we are glad to see that there is an upward trend in hotel demand! We can also see that the demand for the City Hotel in winter is lower, and it bounced back in the spring. The fluctuation of the Resort Hotel is smaller, but we can still see the demand is less in winter.

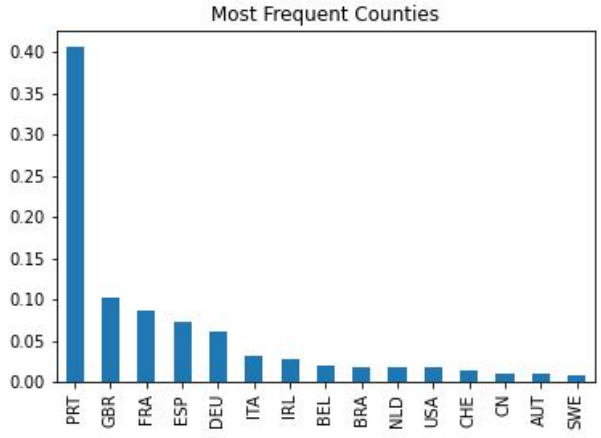
The hotel cancellation rate graph on the right tells that the City Hotel has a significant decrease in cancellation rate in 2015, then the cancellation rate fluctuates around 0.4. The Resort Hotel's cancellation rate fluctuates around 0.25. Recently, the cancellation rate of the Resort Hotel is increasing and starting to exceed the cancellation rate of the City Hotel.

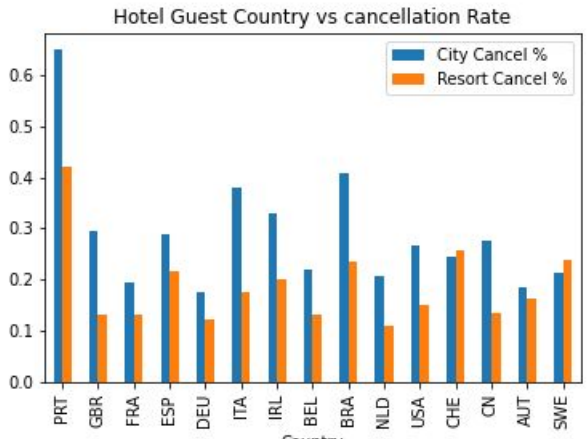


(b) Customer Countries Analysis

From the Most Frequent Countries figures on the left below, we see that most guests are from Europe Countries. Portugal's guests are the majority. However, Portugal guests have a relatively high cancellation rate (as shown on the Hotel Guest Country vs cancellation Rate figure), especially to the City

Hotel (almost 70%). From the country analysis, we can see that country could be a useful factor to fit into a model.



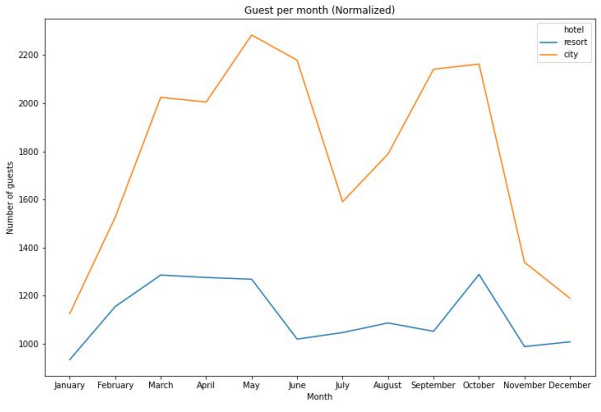


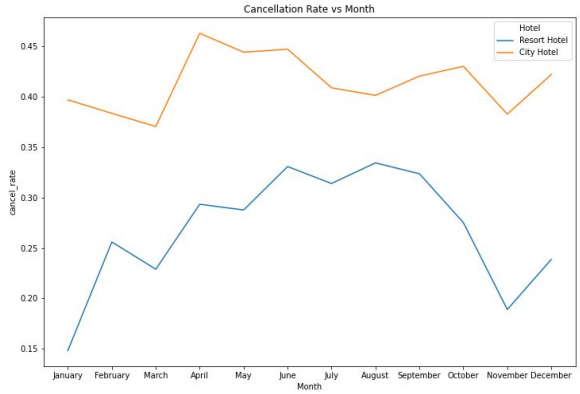
(c) Seasonality Analysis

The following two figures show the normalized trend of the number of guests per month and Price Per Night Per Person over months. Since the data is from July 2015 to August 2017, we need to normalize the data so we can have a clear look at the trend over months.

From the "Guest Per Month (Normalized)" figure, we can see that both the City Hotel and the Resort Hotel have a similar shape in customer numbers. They all have more guests during the Spring and Autumn. The number of guests is the lowest in winter. And surprisingly, both hotels have a drop in customer numbers in summer. Overall, we can see a clear seasonality demand to both of the hotels.

The cancelation rate also shows seasonality changes. The cancellation rate for the Resort Hotel is the lowest in the winter and increases in spring, summer, and autumn. The cancellation rate for the City Hotel is the highest in the spring the autumn, winter and summer have a relatively low cancelation rate. The seasonality of the cancelation rate is not quite the same as that of the guest number.

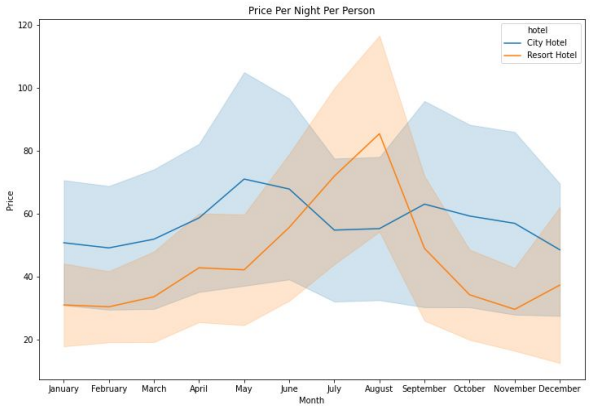


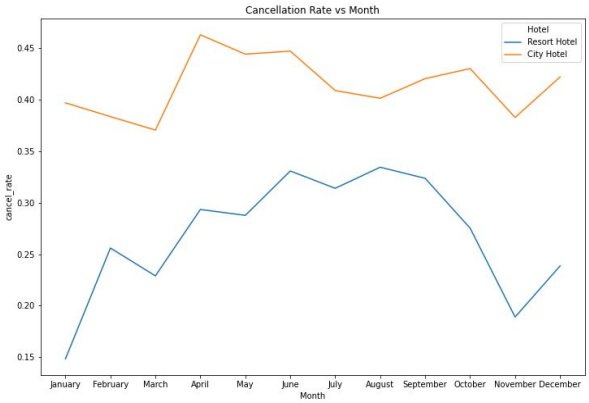


(d) Price Analysis

The following graphs show the trend of hotel price per night per person (on the left) and the cancellation rate over the month (on the right. This graph is the same as the one shown in the seasonality analysis). Both of the graphs are normalized based on months just like what I did in the

seasonality analysis. Compare the "Cancellation Rate vs Month" figure with the “Price Per Night Per Person” figures, we see that the shape of the cancelation rate trend is very similar to the shape of the price trend for both of the hotels. We see that there is an increase in both the cancellation rate and the price from March to September for the Resort Hotel. We also find the trend of the City Hotel for both the cancellation and the price is highly identical. We may consider that the cancellation rate could also be price-driven rather than just seasonality.





Modeling

As our goal of this project is to find out the reason why people cancel their orders and precisely identify customers who will cancel the order, the metric I focused on when building my models was **precision**.

(a) Feature Engineering

Before I build the model, I performed feature engineering. In this feature engineering, I removed 9 columns as they seems not informational for predicting the cancellation rate: 'arrival_date_year', 'arrival_date_day_of_month', 'arrival_date_month', 'reservation_status', 'reservation_status_date', 'reservation_status_date_datetime', 'arrival_datetime', 'company', 'agent'. Most of these features are time-related, time expects month, might not help us predict future cancellations. So I dropped all time-related features here, and only leave the numerical column of arriving month for seasonality analysis. I dropped the company and the agent column because most customers did not make the reservation through an agent or a company.

I also added a feature called family. If the order has both adults and children or babies, then the order is considered to be placed by a family. Then I removed the children and baby columns from the dataset.

As I checked the dataset, I found several columns that have a small amount of “Undefined” values in meal,distribution_channel,market_segment columns. In order to reduce columns numbers as I encode the categorical data, I imputed all the 'Undefined' values with the mode values in the columns.

I used binary encoding for country columns as it has a large amount of different categorical values. Binary encoding is suitable for such a feature. I used OneHotEncoding for the rest of the categorical features.

The dataset used to have 35 columns. After feature engineering, there are 68 columns.

(b) Model Selection

In this project, I tested 3 different machine learning classification models: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier. The testing results are shown in the table below. **Gradient Boosting is the best model in this project**.

	Model	Accuracy	AUC	Tru Positive Rate	Precision
0	Logistic Regression.	0.76	0.70	0.46	0.81
1	Random Forest	0.88	0.86	0.79	0.88
2	Gradient Boosting	0.85	0.95	0.64	0.94

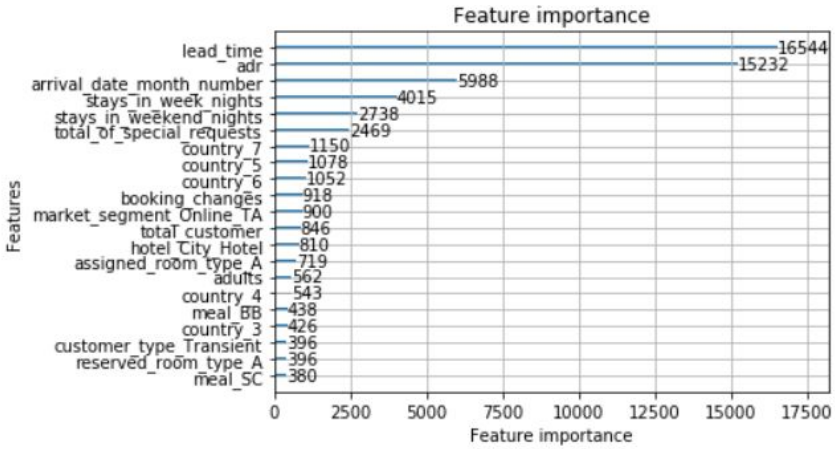
The Logistic Regression model does not perform so well. The logistic regression model only correctly identifies 46% of cancellation. And only 81% of the positive prediction is correct. The True positive rate is very low (0.46), and the precision is not high enough. Compared to the other two models, Logistic Regression is not a good choice for this project.

The Random Forest model. 79% of cancellations are correctly identified and 88% of the positive prediction is correct. Although this model has the highest score on the True Positive rate, the precision rate is still not high enough compared to the Gradient Boosting model. Since precision is the key metric that we should focus on this problem, Random Forest is still not the best model to choose from.

The Gradient model performs the best overall. It has the highest AUC score and precision score. we can see that 64% cancellation is correctly identified. The precision is 94%, meaning that 94% of the positive prediction is correct. This model can help us correctly identify which orders are more likely to be canceled, so we can process certain policies to such orders.

(c) Model Analysis

The following graph shows the top 21 most important features in the Gradient Boosting model. Features are ranked from the most important ones to the least important ones.



Problem 1: The lead_time (number of days that elapsed between the entering date of the booking into the PMS and the arrival date) has the biggest effect on the cancellation. This seems reasonable that, as the customers make their reservations too early, they are more likely to change their plans and cancel their trips.

Suggestion 1: Provide a cancellation penalty if the customers make the reservations too early.

Problem 2: The second most important feature is the average daily rate. As the average daily rate gets more expensive, customers are more likely to cancel their orders. Maybe the customers find out better deals after they made the reservation and decide to switch to other hotels. The hotel owner definitely does not want customers to cancel rooms with high pricing which may cause a big loss to the hotel operations.

Suggestion 2: Send coupons through welcome email after customers make the reservations. The coupons could be used on dining or SPA. Send brochures through emails to the customers once or twice a month to give them understandings of the high-quality service we provide.

Future Works

1. The model gave us information about which factors have more impact on the cancellation rate. However, we still need some more quantitative information in order to provide more detailed advice. For example, we can do analysis and see how much the cancellation rate will increase by adding one unit of lead_time
2. We can do analyses for the City Hotel and the Resort Hotel separately and figure out why the City Hotel has a higher cancellation.
3. According to the features of a customer, such as the country (i.e. where the guest is located), and the order time, we can calculate the probability that a customer will cancel an order. Based on the cancellation probability, calculate the expected value of this customer and give the customer a respective price. For example, we have a Portuguese guest. He ordered a room in the City Hotel in January a year from now. The order is for two adults only. The system estimated the probability of this customer to cancel this order is high, then the system will raise the price for this room for this customer so we can have a higher expected value.