Jazmaine Vitta

May 16th, 2023

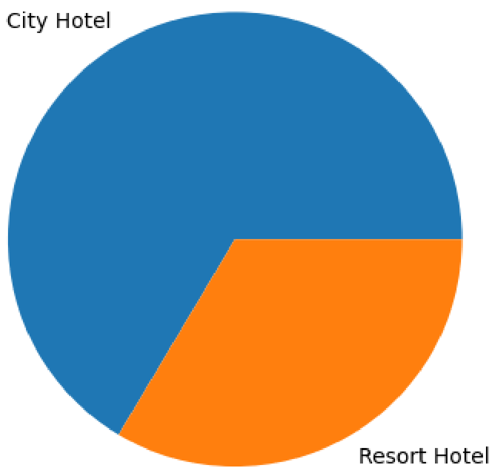
Predicting Hotel Booking Cancellations

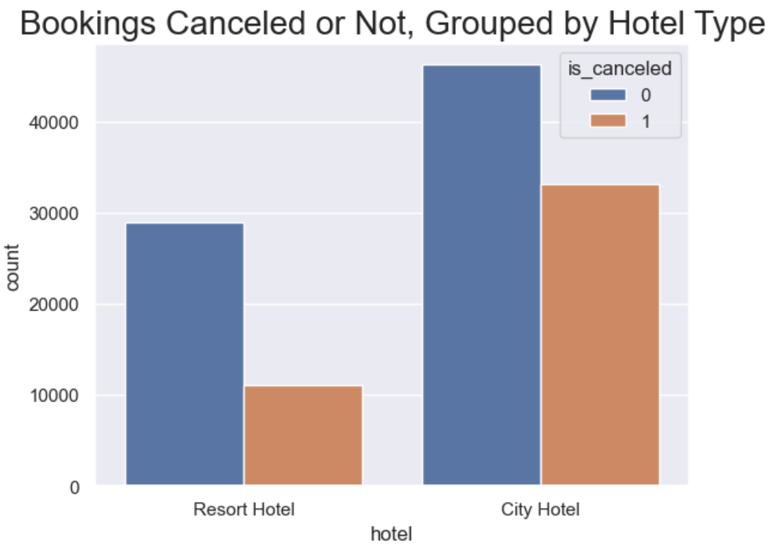
Link to Kaggle Dataset

The inability to anticipate whether or not a customer's hotel booking will be canceled undoubtedly causes major setbacks for the hotel business. As a result of cancellations, hotels can not realistically estimate the total number of guests each night causing them to be inadequately equipped to manage customers, resulting in a loss in revenue and countless other complications. Due to the uncertainty that comes with the possibility of cancellations, hotels are unable to accurately forecast their daily demands. In most cases when a customer does not show up on the arrival date of their stay or cancels close to the date of their booking, those rooms are either resold at a lesser rate than they were originally booked for or even worse not resold at all. These issues not only lose revenue for the business but also cause dissatisfaction with customers. To beat the negative effects of cancellations hotels attempt to overbook rooms which comes with much more issues. Cancellations are a direct cause of loss of revenue in numerous ways.

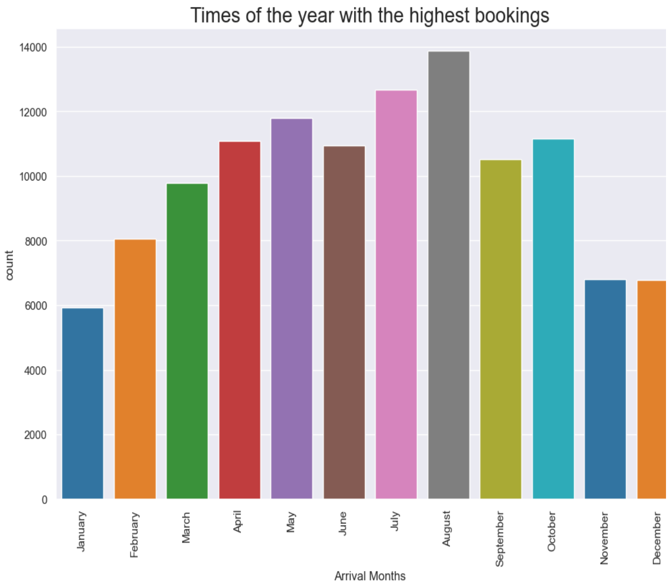
By building a classification model to predict the likelihood of a cancellation, I can assist businesses in overcoming the uncertainty and loss in revenue caused by cancellations. Using a data set from Kaggle.com the data was first cleaned and then trained to find any correlations or causations to the change in booking status. The data set used to build this model contains booking information during the years of 2015 to 2017. The data is made up of a total of 119,389 booking observations with 32 columns, of which 20 are categorical and 12 are numerical. Throughout my process of preparing the data I was able to uncover the answers to many interesting questions. Based on the information the target feature is the 'is_canceled' column.

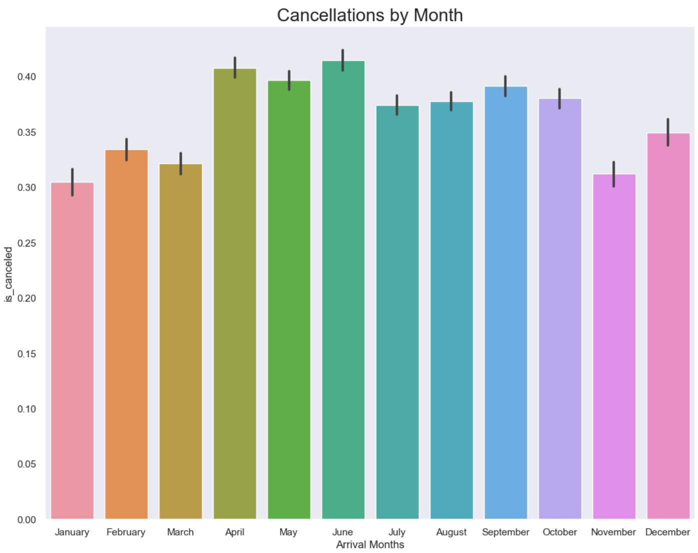
The data contains more than double the amount of city hotels compared to the amount of resort hotels. Between the two, City hotels experience almost as many cancellations as they do bookings, which is a big contrast compared to resort hotels who receive around a third of their bookings canceled.



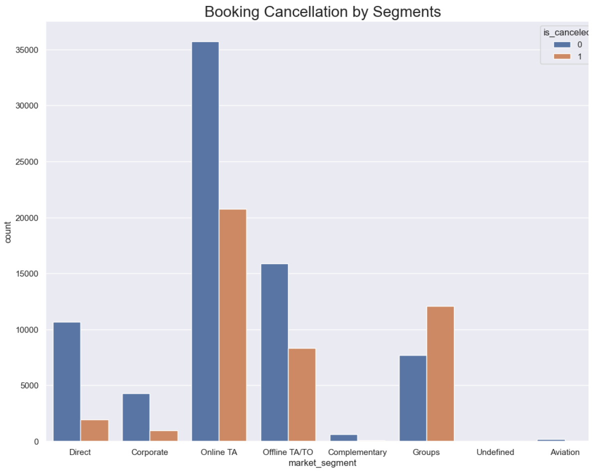


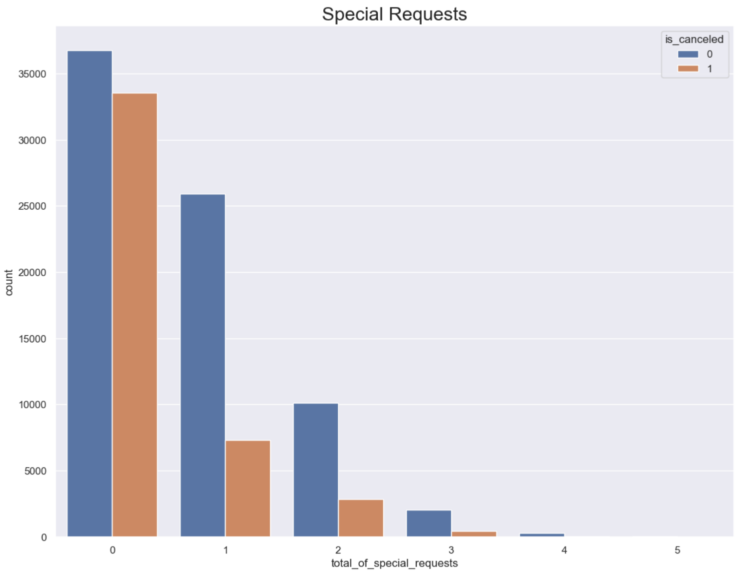
Although the summer months are the most popular time of the year for bookings, the spring months seem to receive the most cancellations.



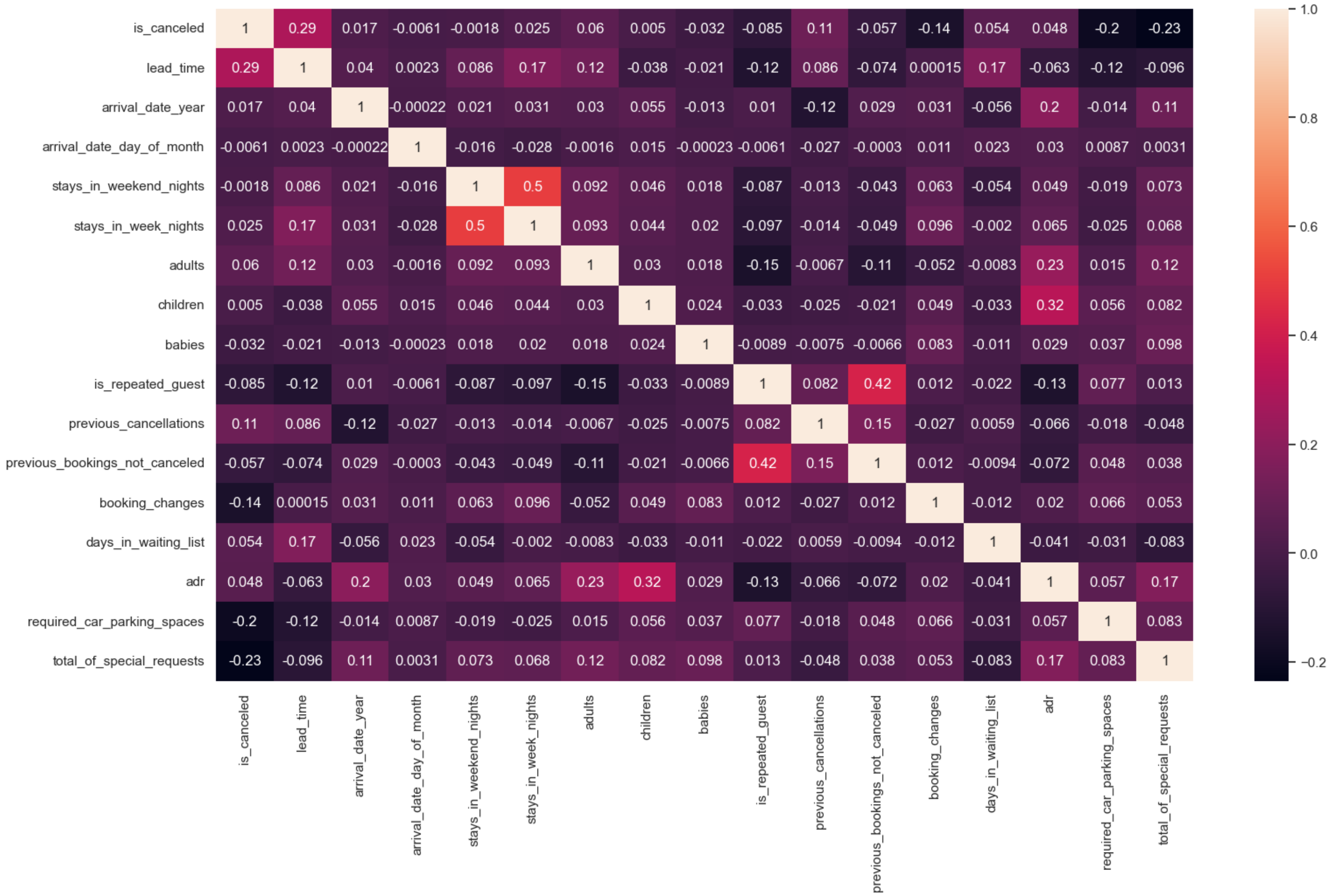


Customers who book directly with the hotel are much less likely to cancel as opposed to customers who book with a travel agent or online via a third party website. Bookings made with at least one special request are drastically less likely to cancel compared to those made with zero requests.





As shown in the heatmap below, the features most correlated to the target feature ‘is_canceled’ are ‘lead_time’, ‘total_of_special_requests’, ‘booking_changes’ and ‘required_car_parking_spaces’.



	is_canceled	lead_time	arrival_date_year	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests
is_canceled	1	0.29	0.017	-0.0061	-0.0018	0.025	0.06	0.005	-0.032	-0.085	0.11	-0.057	-0.14	0.054	0.048	-0.2	-0.23
lead_time	0.29	1	0.04	0.0023	0.086	0.17	0.12	-0.038	-0.021	-0.12	0.086	-0.074	0.00015	0.17	-0.063	-0.12	-0.096
arrival_date_year	0.017	0.04	1	-0.00022	0.021	0.031	0.03	0.055	-0.013	0.01	-0.12	0.029	0.031	-0.056	0.2	-0.014	0.11
arrival_date_day_of_month	-0.0061	0.0023	-0.00022	1	-0.016	-0.028	-0.0016	0.015	-0.00023	-0.0061	-0.027	-0.0003	0.011	0.023	0.03	0.0087	0.0031
stays_in_weekend_nights	-0.0018	0.086	0.021	-0.016	1	0.5	0.092	0.046	0.018	-0.087	-0.013	-0.043	0.063	-0.054	0.049	-0.019	0.073
stays_in_week_nights	0.025	0.17	0.031	-0.028	0.5	1	0.093	0.044	0.02	-0.097	-0.014	-0.049	0.096	-0.002	0.065	-0.025	0.068
adults	0.06	0.12	0.03	-0.0016	0.092	0.093	1	0.03	0.018	-0.15	-0.0067	-0.11	-0.052	-0.0083	0.23	0.015	0.12
children	0.005	-0.038	0.055	0.015	0.046	0.044	0.03	1	0.024	-0.033	-0.025	-0.021	0.049	-0.033	0.32	0.056	0.082
babies	-0.032	-0.021	-0.013	-0.00023	0.018	0.02	0.018	0.024	1	-0.0089	-0.0075	-0.0066	0.083	-0.011	0.029	0.037	0.098
is_repeated_guest	-0.085	-0.12	0.01	-0.0061	-0.087	-0.097	-0.15	-0.033	-0.0089	1	0.082	0.42	0.012	-0.022	-0.13	0.077	0.013
previous_cancellations	0.11	0.086	-0.12	-0.027	-0.013	-0.014	-0.0067	-0.025	-0.0075	0.082	1	0.15	-0.027	0.0059	-0.066	-0.018	-0.048
previous_bookings_not_canceled	-0.057	-0.074	0.029	-0.0003	-0.043	-0.049	-0.11	-0.021	-0.0066	0.42	0.15	1	0.012	-0.0094	-0.072	0.048	0.038
booking_changes	-0.14	0.00015	0.031	0.011	0.063	0.096	-0.052	0.049	0.083	0.012	-0.027	0.012	1	-0.012	0.02	0.066	0.053
days_in_waiting_list	0.054	0.17	-0.056	0.023	-0.054	-0.002	-0.0083	-0.033	-0.011	-0.022	0.0059	-0.0094	-0.012	1	-0.041	-0.031	-0.083
adr	0.048	-0.063	0.2	0.03	0.049	0.065	0.23	0.32	0.029	-0.13	-0.066	-0.072	0.02	-0.041	1	0.057	0.17
required_car_parking_spaces	-0.2	-0.12	-0.014	0.0087	-0.019	-0.025	0.015	0.056	0.037	0.077	-0.018	0.048	0.066	-0.031	0.057	1	0.083
total_of_special_requests	-0.23	-0.096	0.11	0.0031	0.073	0.068	0.12	0.082	0.098	0.013	-0.048	0.038	0.053	-0.083	0.17	0.083	1

With the knowledge of columns that have the most relation to the target feature I differentiated which features are categorical and which are numerical. This was useful in determining the type of visualization that I later performed. The data in the categorical features were One Hot End Coded to be able to make proper estimates.The data was then split into training data and testing data. Using the now split data a was able to build a Logistic Regression model which predicted booking cancellations with an accuracy of only

81%. With all the data I know the accuracy for predicting could be at least 90%. A Random Forest Model was built and we can now predict whether or not a new booking will be canceled with 93% accuracy. The Random Forest Classifier is no doubt the better model. This model has a much lower cross-validation of 0.926. It also exhibits less variability with a mean absolute error of 0.0700226.

Predicting cancellations is a real problem for the hotel industry! Having a good understanding of this problem and the features that closely relate with cancellations will be very useful to your business. The model I created will no doubt assist in decreasing the possibility of your business losing revenue by being under/over prepared. My Random Forest Model has a high accuracy of 93.09%, to predict hotel booking cancellation. Gathering more observations for resort hotels to help balance the data would also assist in strengthening the predictability of the model. In turn, this recommendation would help the model become even more of an asset to your business.