Hotel Reservation Cancellation Risk Model

Capstone 2 Final Report

By Cameron Hicks

Hotel reservation cancellations can have a significant impact both on top line revenue and bottom line profit, especially cancellations closer to the check in date leaving a small window for new bookings to fill the room. In this project we explore a data set of reservations for one resort hotel and one city hotel in Portugal to create a Machine Learning model that can assess the probability of a reservation canceling. Hotel leaders can utilize this information to determine revenue and operational strategies that will produce the best possible business results.

1 - Data

For this project we used two data sources.

The **Hotel Booking Demand** (provided by kaggle) provides the reservation/booking data for one resort hotel and one city hotel in Portugal. This data set contains over 119k reservation records, and includes information for booking cancellations, booking lead time, dates of stay, and other data that we were able to explore if it has relevance to the cancellations.

The Hotel Booking Demand dataset only includes information available to the actual reservations. One of the largest variables that could impact hotel cancellations is weather conditions. To ensure we fully explored this variable we used the **Open Meteo Weather API** to fill in weather conditions for each reservation.

- Hotel Booking Demand csv - provided by kaggle - https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data
- Open Meteo Weather API - https://open-meteo.com/

2 - Approach

a) **Data Wrangling and Preprocessing**
To build a functional predictive hotel reservation cancellation model I first needed to clean, combine and prepare the data to be analyzed. I explored missing and null values and discovered that there were missing values for the Children, Country, Agent, and Company columns. I determined that if there was not a

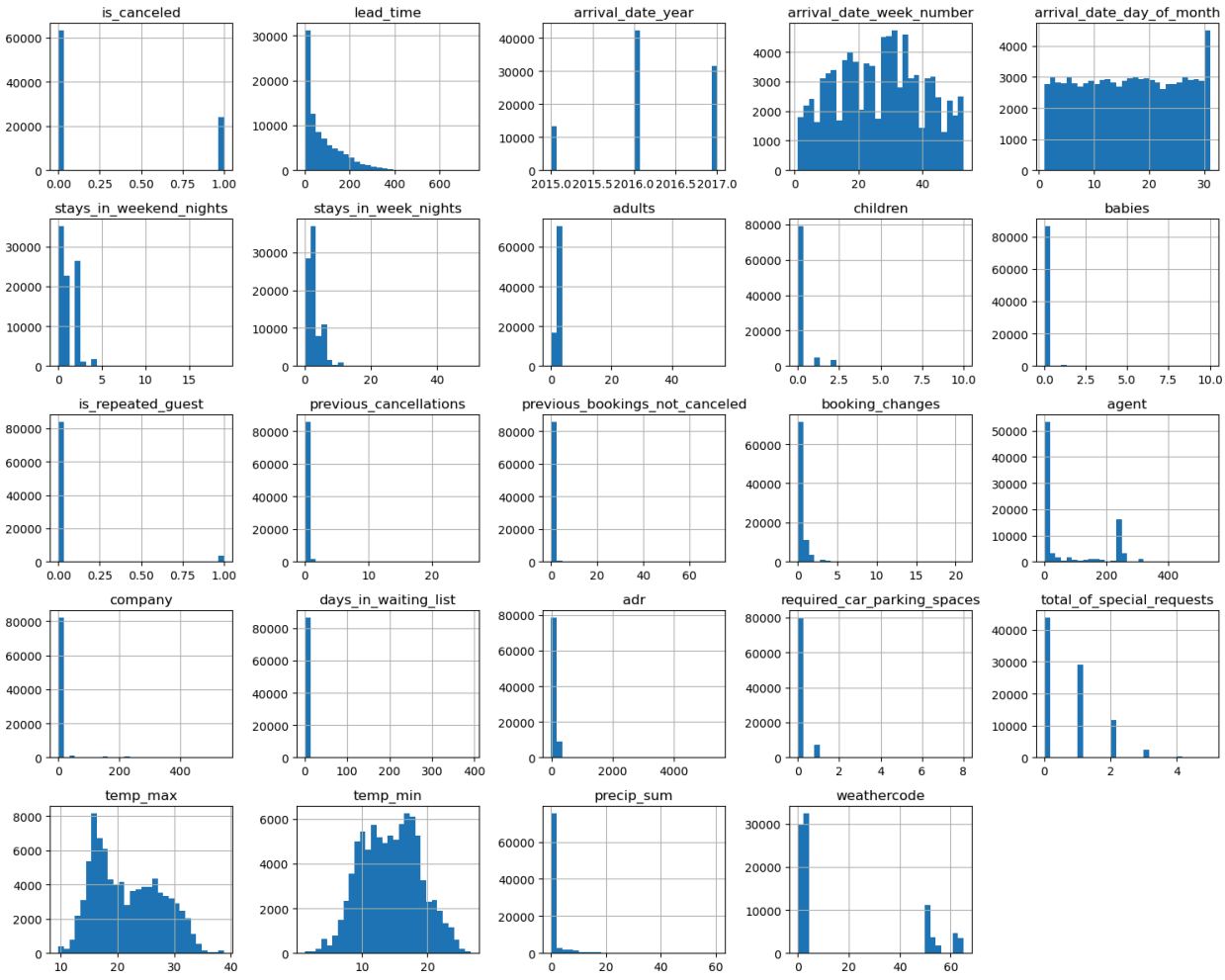
value for Children then the travelers did not have children, so the null values were replaced with ‘0’. I also determined that reservations with null values for Agent and Company indicated that those travelers did not use an agent and were not traveling with their company, so those were also filled with 0. The missing values for Country were filled with the string ‘Unknown’ to preserve the data. Once the missing values were properly handled, I continued to prepare the data by dropping duplicate reservation records and correcting the data types.

Then I proceed to collect and merge weather data from the Weather API. My target weather data was the forecasted weather for the arrival date as of the cancellation date to find if there was a correlation between cancellations and the weather forecasts. To collect this I had to gather all of the unique arrival dates from the reservation data to increase efficiency while pulling from the Weather API. I also needed to pull the unique cancellation dates to pull the weather forecasts. Using those two sets of unique dates, I was able to create a data frame with the weather data I needed and combine it with the reservation data so each reservation record now included data for the temperature high and low, precipitation summary, and a description of the weather conditions.

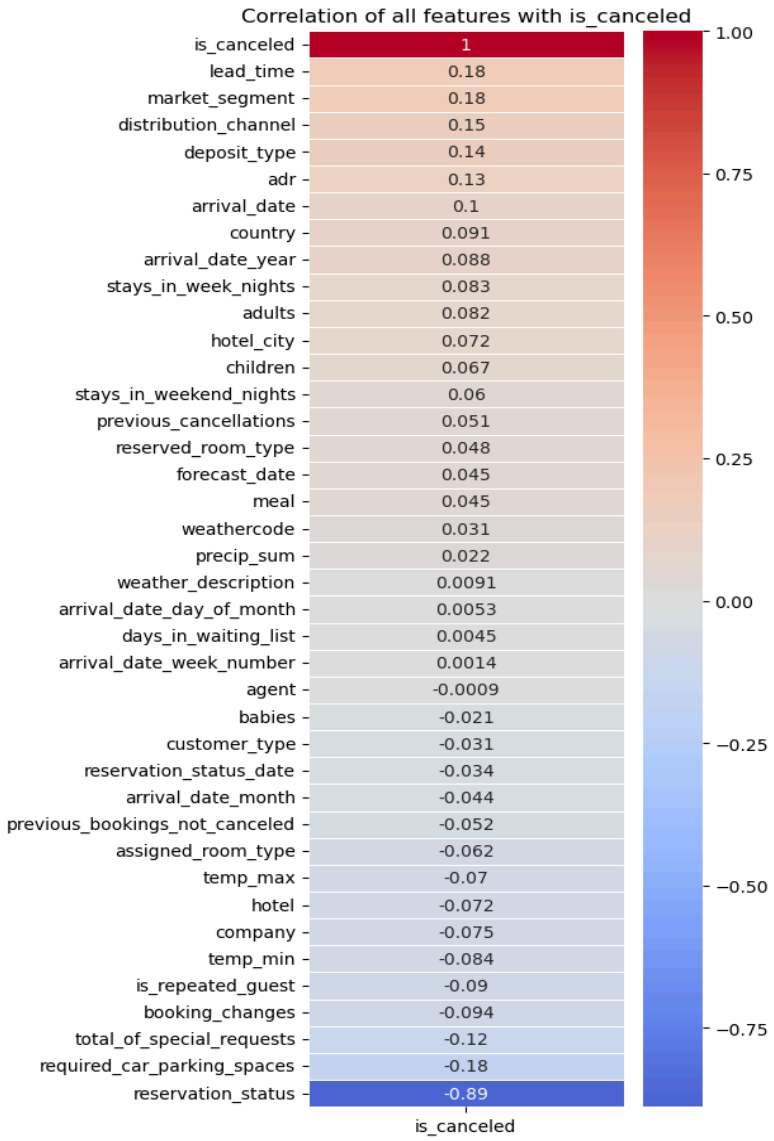
From my combined data frame, I then proceeded to the preprocessing step. There I created dummy features for the Categorial features in the data frame, set and standard scaler, and performed a train test split on the data.

b) Exploratory Data Analysis

In the exploratory data analysis section I began to explore the features of the data set to see what may have an impact on our target feature - cancellations. I found that the lead time was heavily right skewed indicating that most reservations were made within a year of their arrival date, and the closer the arrival date the higher number of reservations were made. Also, the temp max and the temp min both showed a majority of the temps being between 10 - 20 degrees celsius, indicating there is not a large range of temperature variance. The temp max did show clusters of results of higher temperatures, indicating that there is a hot season for the hotels.



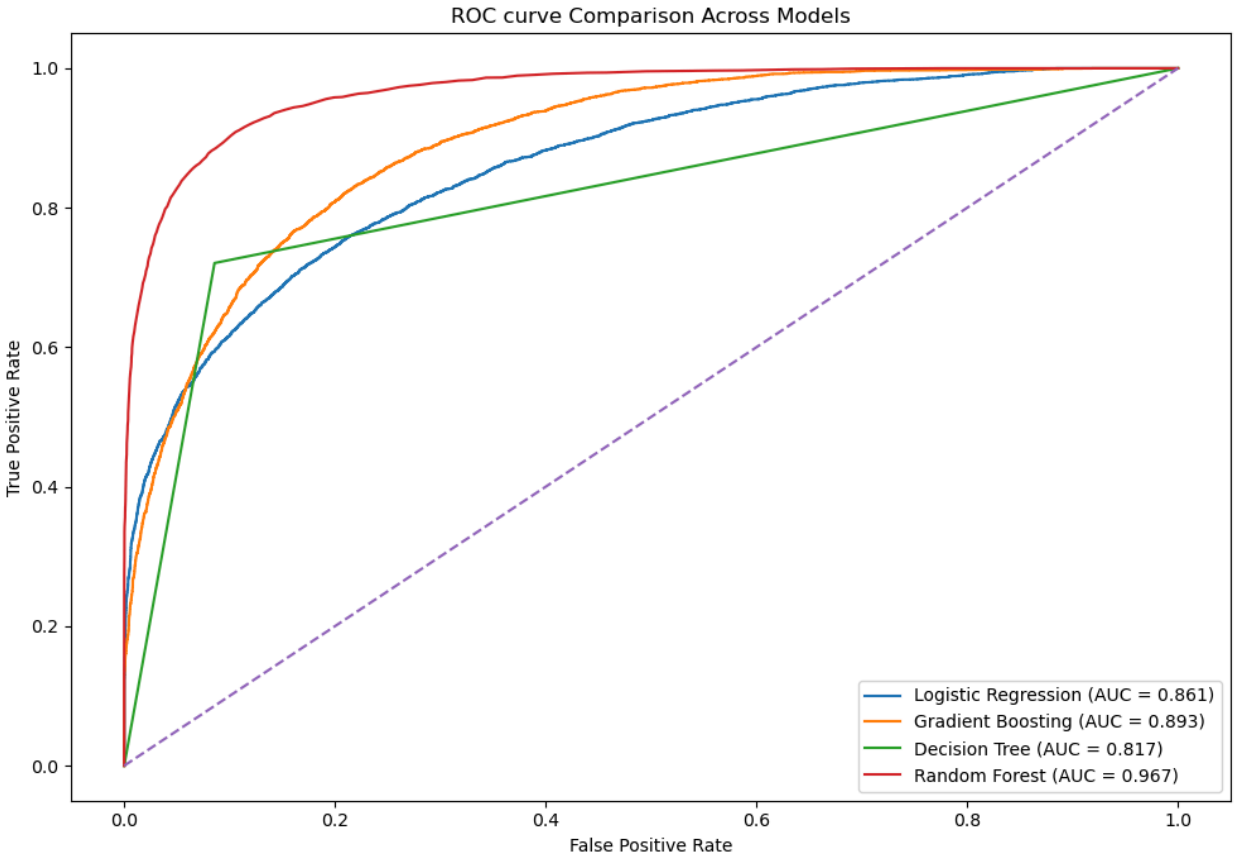
I then began to explore the features and how they correlated to the cancellations and plotted a heatmap to better visualize. I zoomed in on our target feature ‘is_canceled’ to specifically see the correlation with each feature.



c) Modeling

In the modeling section I was able to put it all together and compared several different models to determine that the Random Forest model delivered the best results for this data. In all I compared a Logistic Regression Model, A Gradient Boosting Classifier, a Decision Tree Model, and the Random Forest Model. Comparing the Logistic Regression, Gradient Boosting Classifier, Decision Tree, and Random Forest Models on the metrics of accuracy, precision, F1-score, and ROC-AUC, we can see that the Random Forest model performed the strongest on all metrics. The strong and consistent performance the Random Forest Model showed on the training and test data makes it the clear choice for our Hotel Cancellation Prediction Model.

	Model	Accuracy	Precision	F1	ROC-AUC
0	Logistic Regression	0.829519	0.764877	0.638788	0.861251
1	Gradient Boosting Classifier	0.833295	0.766413	0.651185	0.892918
2	Decision Tree	0.861041	0.761099	0.740353	0.817474
3	Random Forest	0.900172	0.939655	0.789378	0.966561



3. Conclusion and Recommendations

In conclusion, the final model selected was the Random Forest model due to its high performance and reliability. I strongly recommend the hotel revenue management team use this model to apply to each new reservation to understand the probability of

cancellation. The Revenue Management team can use this information to determine how many rooms can be oversold while mitigating the risk of not having enough rooms for all guests. The ultimate goal is to achieve a perfect sell out of rooms even with reservation cancellations.

4. Recommendations for Future Development

While this model achieves the objective of this project, I highly recommend the additional enhancements to further generate business results:

1 - Oversell recommendation model - This model will use the cancellation probabilities generated by this current project to deliver a recommendation of how many rooms can be oversold by each day. This will reduce the time it takes to analyze the probability data, and will generate more accurate oversell recommendations

2 - Cancellation policy recommendations model - This model will use the cancellation probabilities generated by the current model and will deliver recommendations of what cancellation policies to assign to each new reservation based on the cancellation probability. This model will protect the hotel from cancellations by assigning more strict cancellation policies to reservations that have higher cancellation probabilities.