“Are They Going to Cancel?” Hotel Analysis

By: Damani Walker

Overview

When planning a trip, we all dread the moment when one of our friends makes a last minute cancellation.

What about from the perspective of the hotels?

Can they predict when a person or party is going to cancel based on different variables?

Data Preparation

The dataset is a csv gathered from kaggle.com, which is a data science community, that provide different datasets.

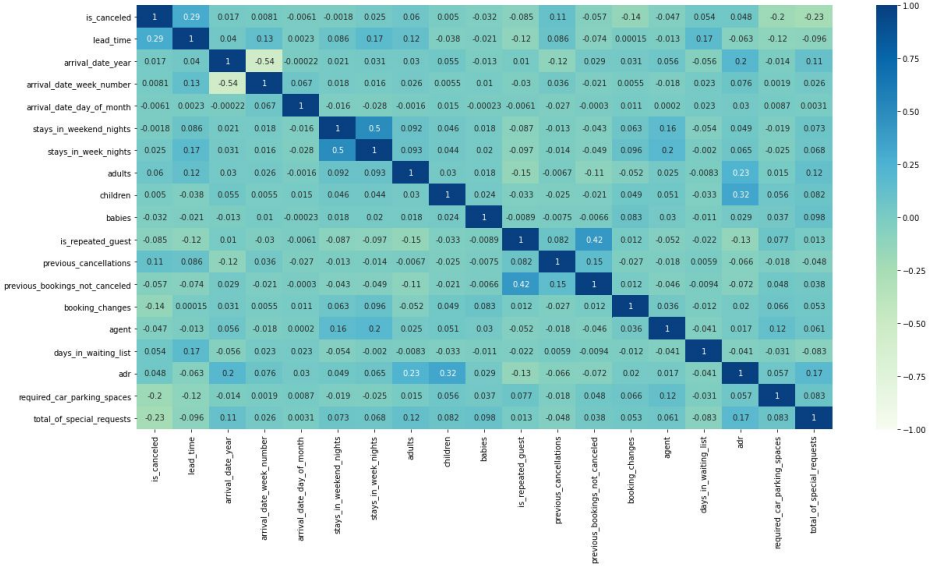
Data needed to be cleaned of nulls and duplicates and provided over 100,000 rows and 32 columns of features to utilize (one of my issues when working with machine learning).

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	agent	company	days_in_waiting_list
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit	NaN	NaN	0
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit	NaN	NaN	0
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit	304.0	NaN	0
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit	240.0	NaN	0

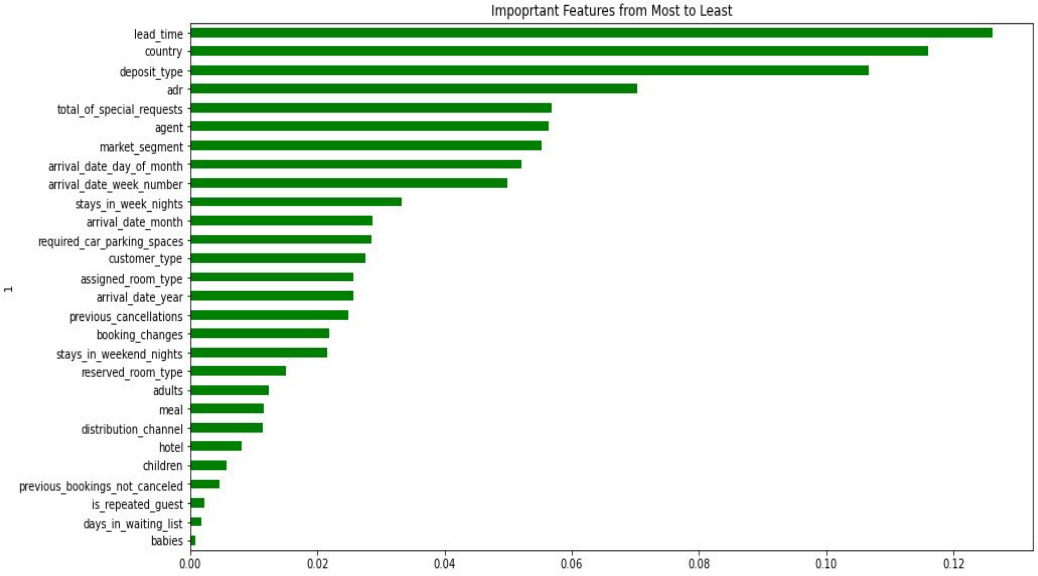
Initial Analysis

The variable I chose as my target was the “is_canceled” (first column).

Then wanted to see how the other variables correlated with the target variable.



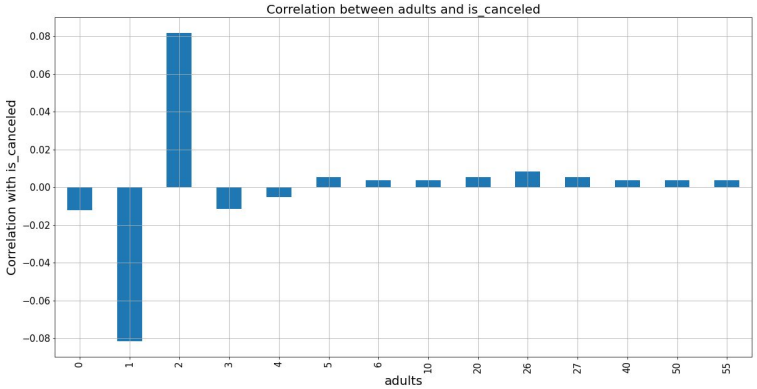
Initial Analysis (cont'd)

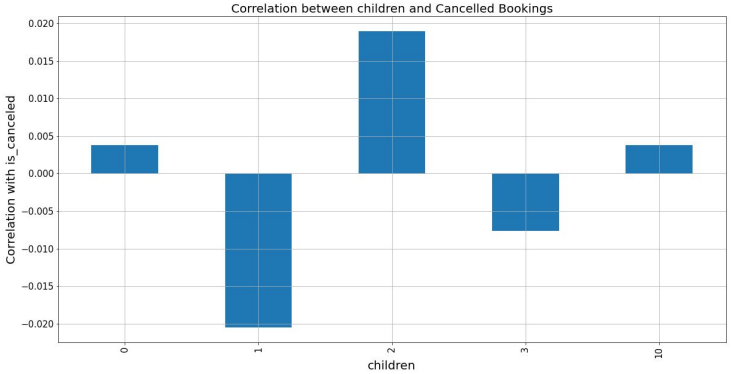


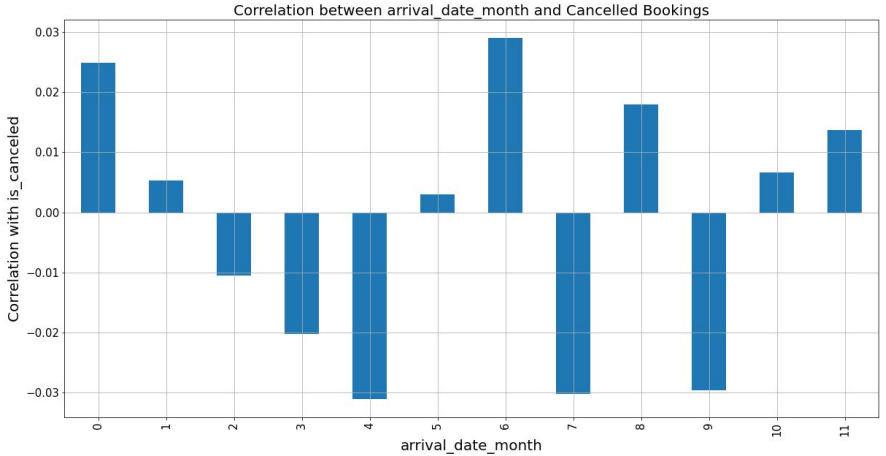
Here we have a graph showing the most important to the least important features:

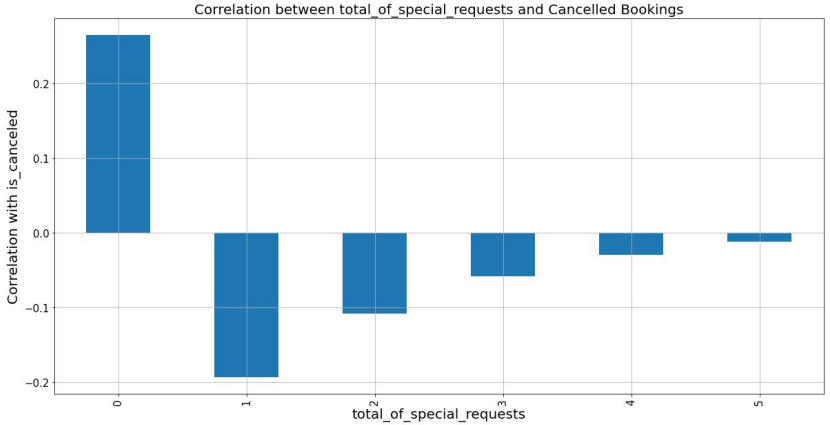
1. **Lead_time** - Number of days between the entering date of the booking and the arrival date
2. **Country** - Country of origin
3. **Deposit_type** - Indication on if the customer made a deposit to guarantee the booking
4. **Avg_daily_rate (ADR)** - Cost per night
5. **Total_special_requests** - Number of special requests made by the customer (e.g. twin bed or high floor)
6. **Agent** - ID of the travel agency that made the booking
7. **Market_segment** - Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
8. **Arrival_day_of_month** - day of the month arriving
9. **Arrival_week_number** - Week number of year for arrival date
10. **Required_car_parking** - Number of car parking spaces required by the customer

Other Notable Analysis









Machine Learning Models

LogisticRegression Model Report

	pre	rec	spe	f1	geo	iba	sup
0	0.79	0.92	0.59	0.85	0.73	0.56	18853
1	0.81	0.59	0.92	0.68	0.73	0.52	10995
avg / total	0.80	0.80	0.71	0.79	0.73	0.54	29848

BalancedRandomForest Classifier Report (Main)

	precision	recall	f1-score	support
0	0.93	0.89	0.91	18853
1	0.83	0.88	0.85	10995
accuracy			0.89	29848
macro avg	0.88	0.89	0.88	29848
weighted avg	0.89	0.89	0.89	29848

Naive Random Oversampling + BRF Report

	pre	rec	spe	f1	geo	iba	sup
0	0.91	0.92	0.84	0.92	0.88	0.79	18853
1	0.87	0.84	0.92	0.85	0.88	0.77	10995
avg / total	0.89	0.89	0.87	0.89	0.88	0.78	29848

EasyEnsembleClassifier Model Report

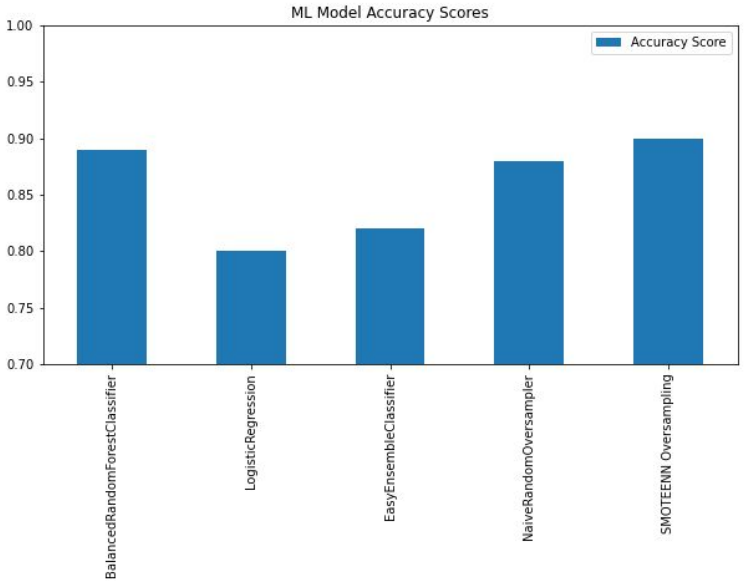
	pre	rec	spe	f1	geo	iba	sup
0	0.88	0.85	0.80	0.86	0.82	0.68	18853
1	0.76	0.80	0.85	0.78	0.82	0.68	10995
avg / total	0.83	0.83	0.82	0.83	0.82	0.68	29848

SMOTEENN + BRF Report

	pre	rec	spe	f1	geo	iba	sup
0	0.94	0.91	0.90	0.92	0.90	0.82	18853
1	0.85	0.90	0.91	0.87	0.90	0.81	10995
avg / total	0.91	0.90	0.90	0.91	0.90	0.82	29848

Conclusion

With that being said, can hotels predict if a booking will be canceled?



Based on the accuracy reports of the different models, it is safe to say that YES, it is possible for hotels to predict if a person will cancel.

Since the accuracy for most of the models are 80% and above, with the SMOTEENN at 90%, I do believe it is possible.

Questions?

Thank You