Springboard Data Science Capstone Project #2:

Predicting the Likelihood of Hotel Cancellations

Kelly Pham

November 29, 2022

Problem Statement:

In today's day and age, travel is everywhere – regardless if it is for business or leisure. With the increase of travel comes an increase for lodging, and there is a lot of competition for hotels to appeal to travelers. Due to the competition, guests are seeking lodging that provides amenities for reasonable prices. The lodging industry has been hit hard by the Covid-19 pandemic, so it is beneficial for hotels to increase bookings, and decrease cancellations, since hotels do lose money through cancellations. By predicting cancellations, hotels can decide whether to 'overbook' available rooms to make up for the cancellation. Moreover, we will identify what features are more or less likely to contribute to cancellations, so we can focus advertising more towards groups that are less likely to cancel.

Dataset:

The dataset was acquired through kaggle.com (https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand), and its size is 119390 rows by 32 columns. The original source of the data is the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. This data set contains booking information for a city hotel in Lisbon and a resort hotel in Algarve, both in Portugal; some information included are when the booking was made, the length of stay, the number of people, the deposit type, and the nationality of the guest. All personally identifying information has been removed from the data. The authors of this article obtained the data through the hotels' Property Management System (PMS) SQL databases, and wrangled them. The PMS assured no missing data exists in its database tables. However, in some categorical variables like Agent or Company, "NULL" is presented as one of the categories. This should not be considered a missing value, but rather as "not applicable". For example, if a booking "Agent" is defined as "NULL" it means that the booking did not came from a travel agent.

Data Wrangling:

The dataset was already wrangled, but there were still a few issues to deal with.

The following columns have missing values: *company*, *agent*, *country*, and *children*.

- The '*company*' column was about 94% null. The only solution was to drop the column.
- The '*agent*' column was stored as floats, and it is assumed that NULL values implied the booking did not come from a travel agent; we encoded this with the float 999.
- We replaced null values from the '*country*' column with code 'XXX' to denote unknown country.
- We assumed the missing values for the '*children*' column was because those guests did not have children, so they did not fill in that section. We filled in those values with 0.

We dropped the following rows that made no sense:

- 3 rows with 1-2 adults and 9-10 children, a highly unlikely scenario.
- A couple rows with ADR < 0.

Since the goal is to predict cancellations, the target variable is '*is_canceled*'.
A lot of the data is heavily right-skewed – so we explored the columns '*stays_in_weekend_nights*', '*stays_in_week_nights*', '*adults*', '*children*', '*previous_cancellations*', and '*adr*'. Further analyses of these columns suggest that the longer stays and larger groups are likely big events like weddings.

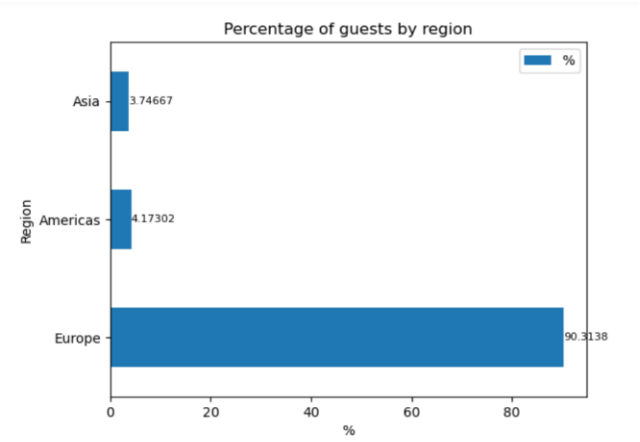
Exploratory Data Analysis (EDA):

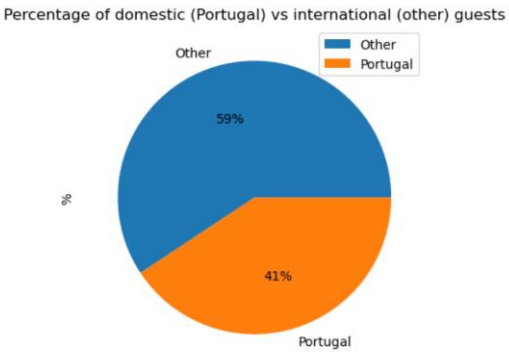
There are 178 unique values for the '*country*' column, so we bin them based on the continents of each country. We create a new column called *'region'* with values Europe, Americas, Asia, Oceania, Africa, and Unknown. To do so, we import the country codes dataset (https://www.kaggle.com/datasets/andradaolteanu/iso-country-codes-global), joined it to the cleaned dataset via a left join on the ISO-3 code of both tables, and extracted the continents to the *'region'* column. Additionally, it would be useful to consider domestic (Portuguese) vs international travelers rather than look at every single country.

New features, such as '*arrival_season'*, were created as well.

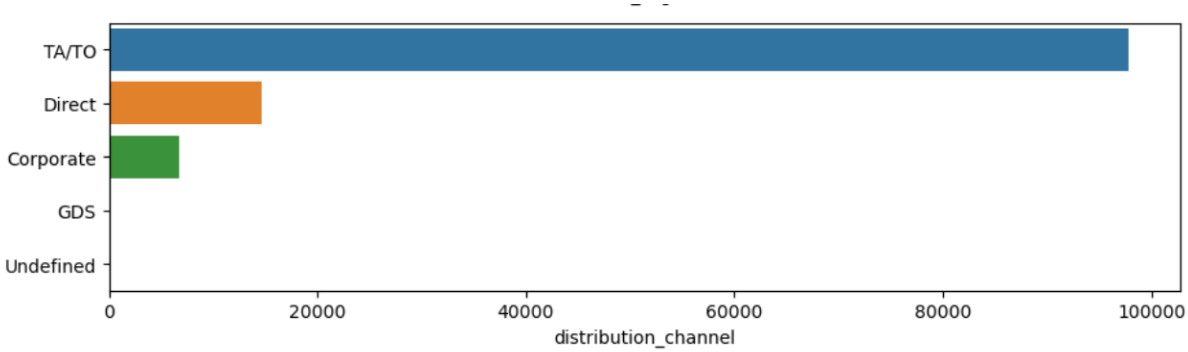
Below are some findings from the EDA:

- Over 90% of travelers are from Europe.
- About 44% of travelers were domestic.



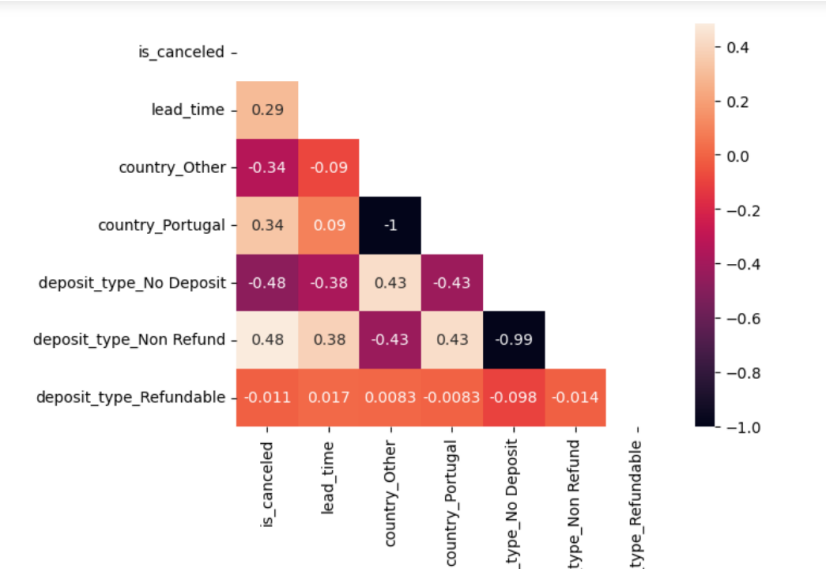


- The majority of travelers booked through Travel Agents/ Tour Operators.

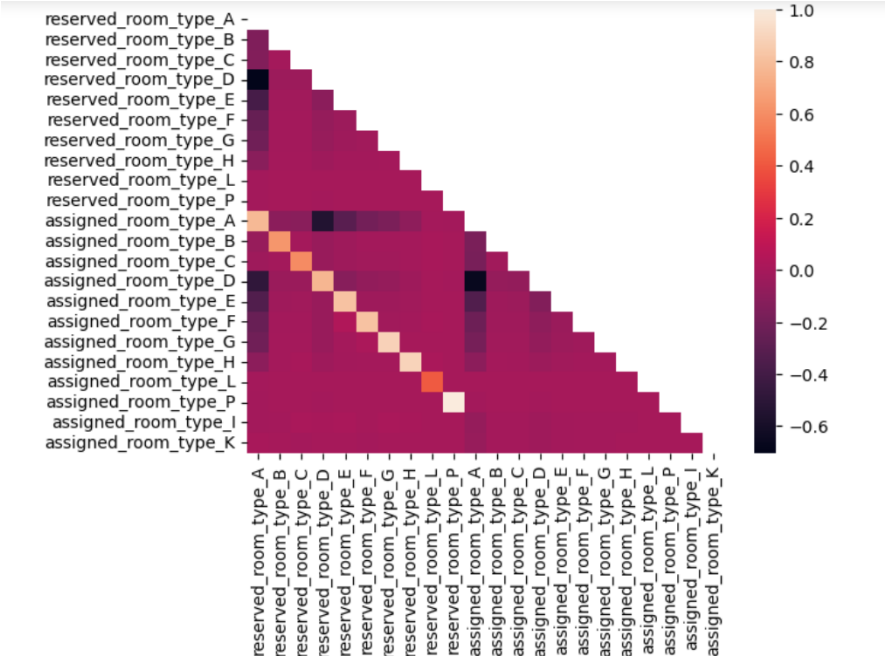


- Most common market segment is through an online travel agent.
- The median cancellation time for the city hotel is 55 days before the arrival date, and 48 days for the resort hotel.
- The median length of stay for both the city and resort hotels is 3 nights.

- There is a positive correlation between nonrefundable deposit and being from Portugal.
- There is also a positive correlation between Portugal and cancellations.



- There is a strong positive correlation between the room types reserved and the room types assigned, which makes sense.



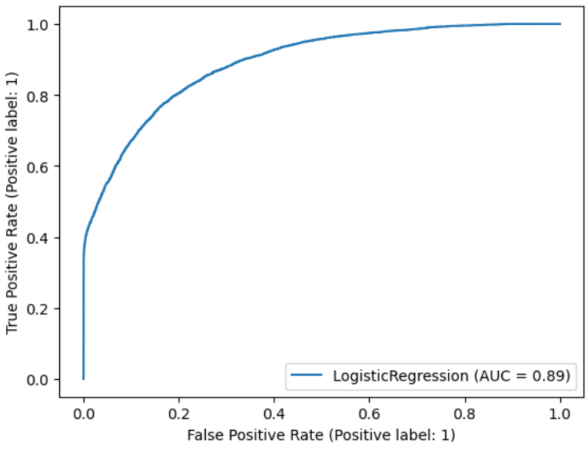
- About 37% of the bookings are cancellations (ie, our dataset is unbalanced).

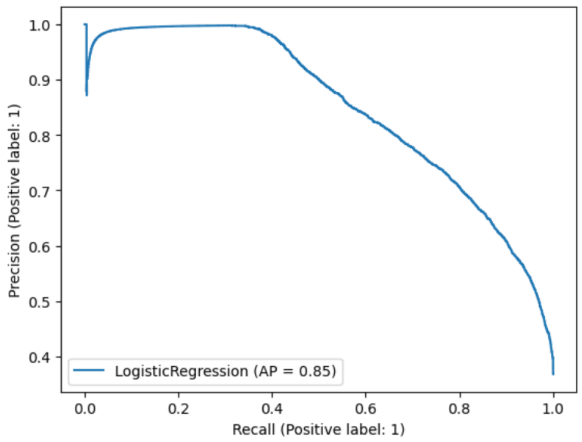
Preprocessing and Model Selection:

First, we drop the datetime columns because they will not be used in models – we are not doing time series analysis. Then, we one-hot-encode the categorical variables. We chose this method instead of label encoding because the categorical variables in this dataset are nominal and not ordinal. Finally, we split the data into training and test sets.

Baseline Model:
For the baseline model, we predict the most common occurrence; that is, we predict all bookings are not cancelled. This model had an accuracy score of 62.9% on the training data and 63.4% on the test data. This baseline model has an ROC score of 50%.

Logistic Regression:
Next, we try logistic regression while keeping all features. This model has an accuracy score of 81% on the both the test and train sets. However, since our data is unbalanced, we also look at the ROC and precision-recall curves which are shown below:





We see this model has an ROC score of 0.89 and Precision-Recall score of 0.85, which is pretty good. Hence, we will not tune the hyperparameters, and we will try the random forest model next.

Random Forest:
This time we tune the hyperparameters via CV randomized search over a 5-fold cross-validation. We try the following hyperparameter values for the random forest model:

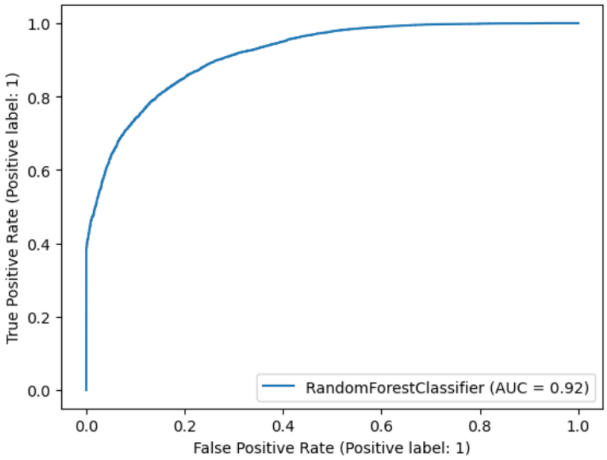
- *n_estimators*: 50, 100, 150, 200, 300
- *min_samples_leaf*: range(1,10)
- *max_features*: sqrt, log2, none
- *max_depth*: 3, 5, 9, None

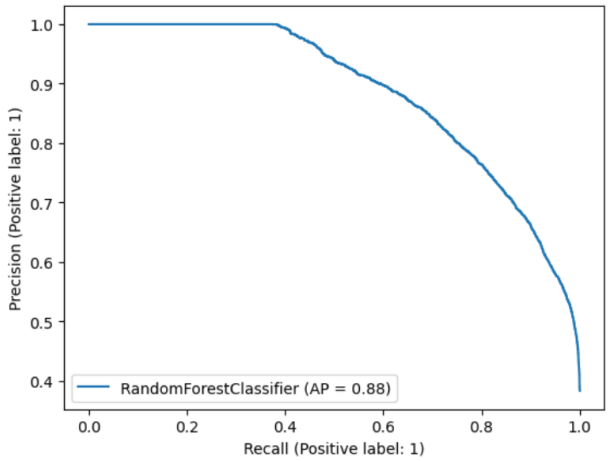
- *criterion*: gini, entropy

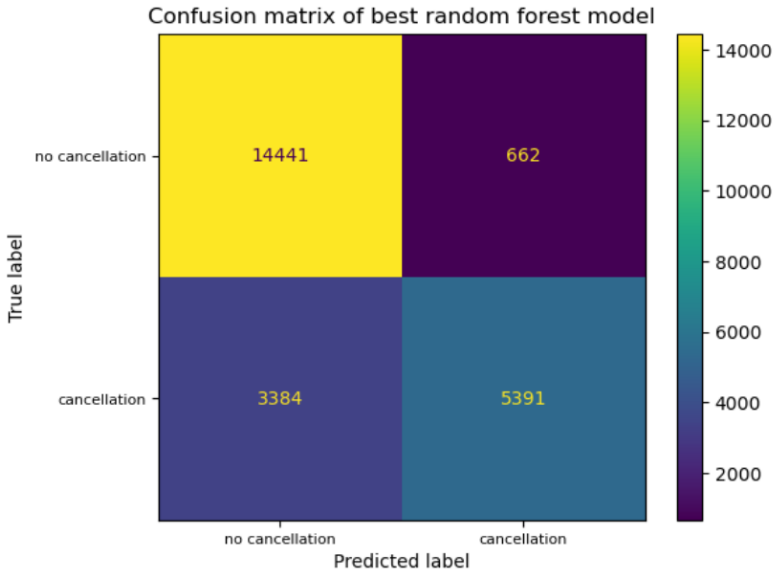
Cross-validation results suggest the best random forest model is the one with the following hyperparameters:

{*n_estimators*: 300, *min_samples_leaf*: 1, *max_features*: sqrt, *max_depth*: 9, *criterion*: gini}

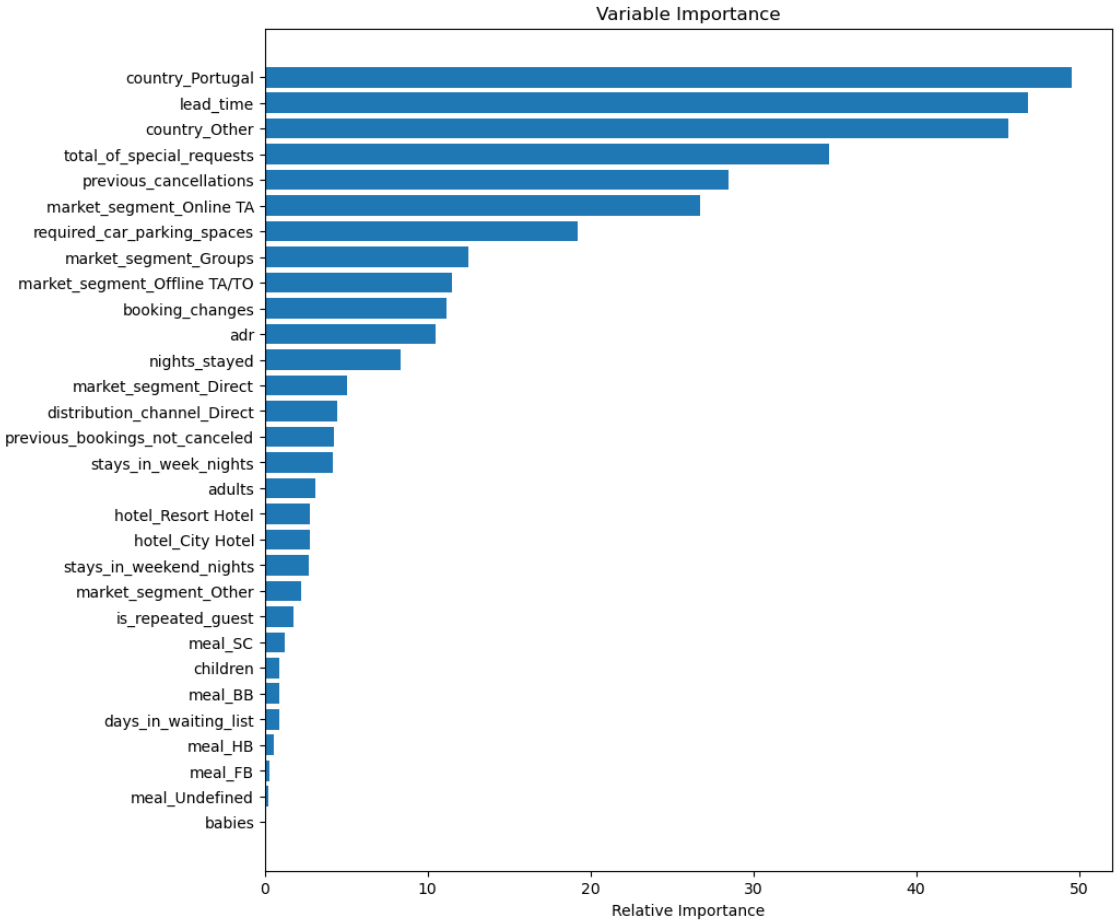
This 'best' random forest model has an accuracy score of 83% on the test set, and ROC score of 0.92, Precision-Recall score of 0.88, which is slightly better than the logistic regression model.







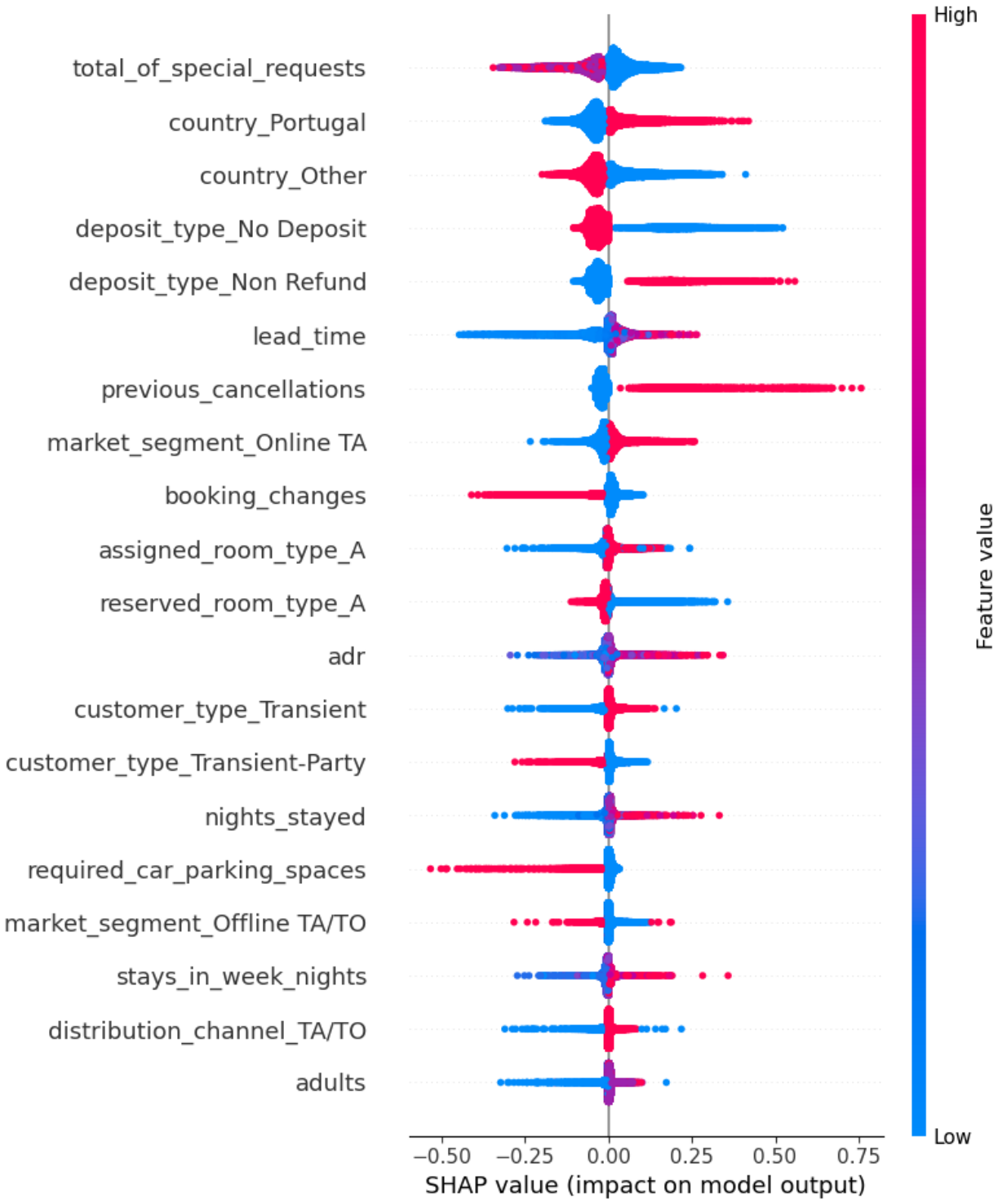
Model Selection:
We select the random forest model due to its higher ROC and PR score. Using the builtin *feature_importances_* method from Scikit Learn random forest, we see the variables '*country*' and '*lead_time*' were used the most.



This biggest advantage of the *feature_importances_* method is a speed of computation - all needed values are computed during the Random Forest training. A drawback is the tendency to prefer numerical features and categorical features with high cardinality.

We also use SHAP analysis to see how much each variable contributes to the random forest model. SHAP analysis is based on game theory and is model-agnostic. The downside is that it takes a lot of time for the machine to compute.

The following graph shows the SHAP values of our random forest model:



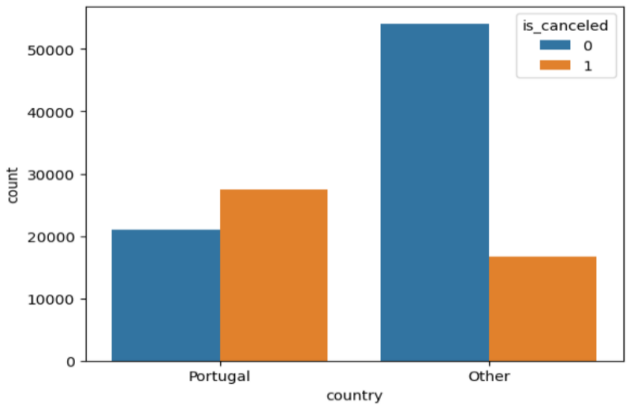
We see the features with the most important impact on the prediction are the number of special requests, country, and deposit type.

Interpretation of the SHAP results:

The less amount of special requests there are, the less likely a guest will cancel. Perhaps this is because more requests means a guest is more committed to showing up.

Nonsuprisingly, local travelers (in our case, from Portugal) contribute to more cancellations than international travellers. Domestic travellers know the local language and hear more about options in their country. Hence, they are more likely to shop around. International travellers are not as familiar with the area, so they might stick to the first hotel they book.

EDA also showed that, for international guests, the difference between those who cancelled and those who did not is drastically different. While for Portuguese travellers, the difference wasn't as extreme. This is shown in the graph below.



Deposit type also contributes to cancellation. Those with no deposit are less likely to cancel, while those with non-refundable deposits are more likely to. Nonrefundable rates are more likely to be much lower than regular BAR rates. It's also possible that some of these rates are paid through points and rewards programs, such as the Expedia Rewards Program. Guests using these programs generally pay very little (or even nothing) out of their own pockets, so it make sense that these rates have higher cancellations; financially, the guests have very little to lose.

Solutions:

Domestic travellers, those with long lead times, and those who booked with nonrefundable deposits are more likely to cancel their reservations. Here are some ways to reduce cancellations:

- Advertise more to international guests, since they are less likely to cancel.
- Offer ‘local discounts’ to domestic guests – this might entice them to stay.
- Check up on guests who booked a long time before their arrival date; send a follow-up email, perhaps starting 60 days before their arrival date.

Future work:

If I had more time, I can try other models such as XGBoost or CatBoost, as well as logistic regression or random forests with highly correlated features dropped. Additionally, I would like to analyze if there are specific reservation dates that have a higher cancellation rate than the average.