Predicting the Likelihood of Hotel Cancellations

Kelly Pham

Springboard, May 2022 Cohort

Problem Statement

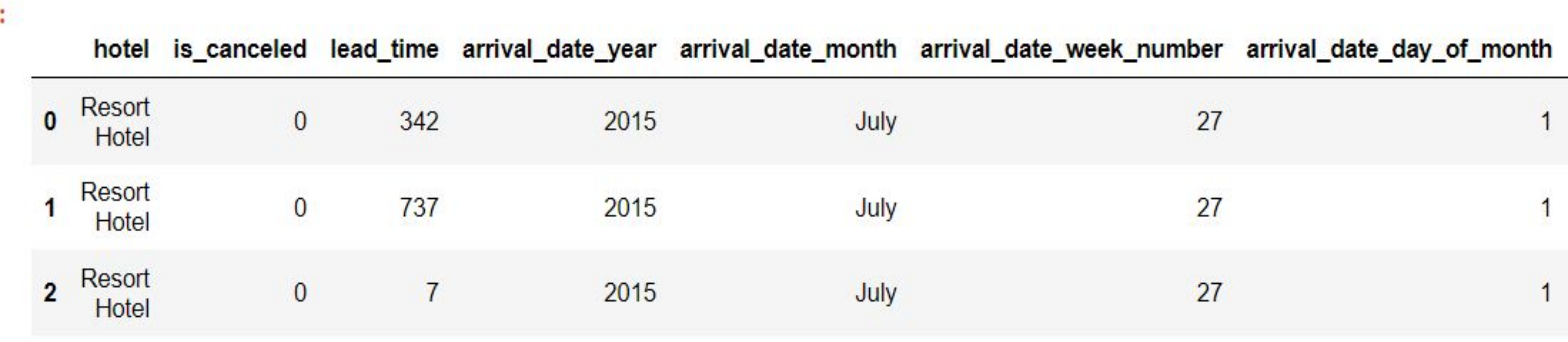
- Travel is everywhere in the modern world, regardless if it’s for business or leisure
- Travellers need lodging
- Average hotel cancellation rate is ~37%
- Cancellations can cost hotels to lose money

- **What factors contribute to hotel cancellation rates?**
- **Can we predict the likelihood of a cancellation, given booking data?**

The Dataset

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

- Size: over 100K rows, 32 columns
- Booking data from 2 hotels in Portugal



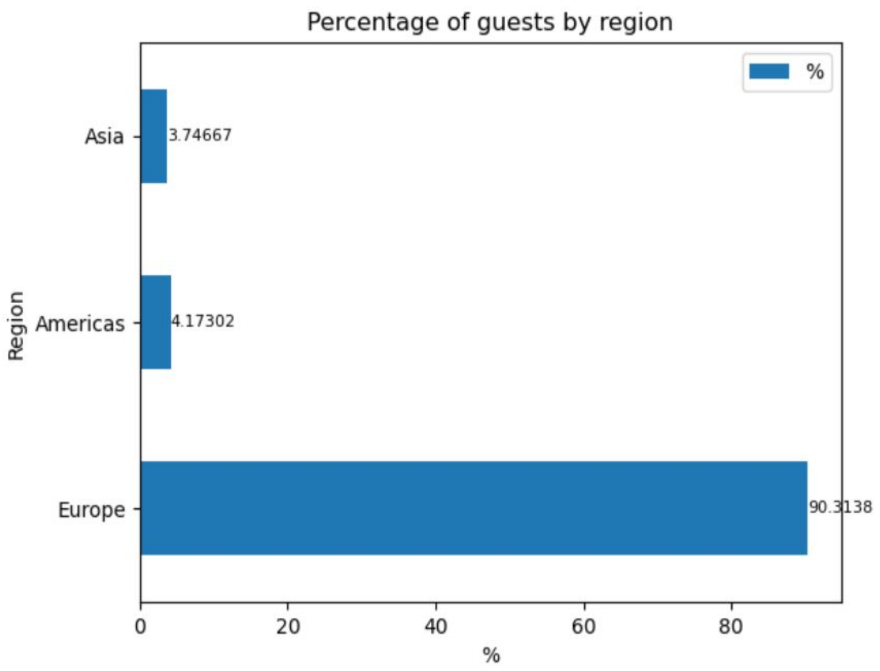
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month
0	Resort Hotel	0	342	2015	July	27	1
1	Resort Hotel	0	737	2015	July	27	1
2	Resort Hotel	0	7	2015	July	27	1

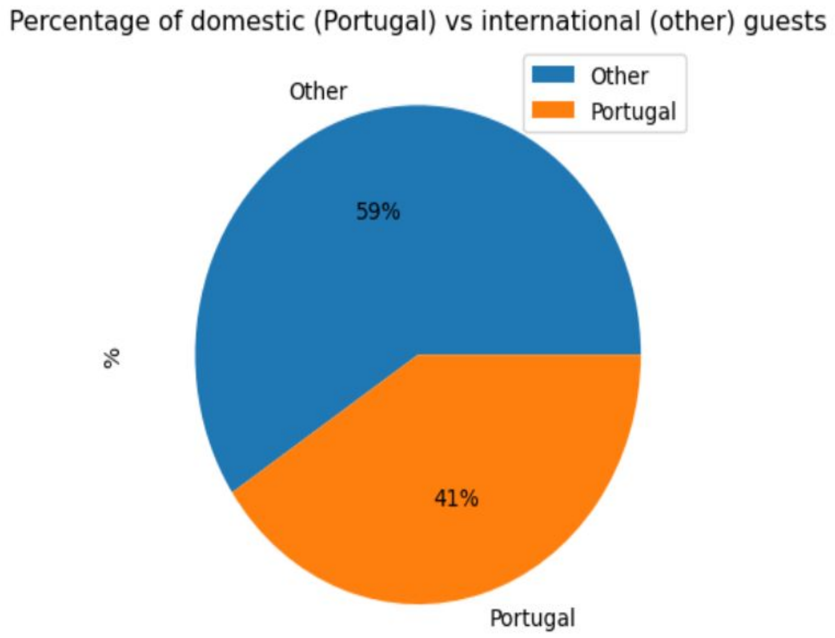
The Dataset - some columns of interest

Column name	Description
hotel	Whether the guest booked a Resort or City hotel
is_canceled	Values 0 (not cancelled) or 1 (cancelled)
lead_time	Number of days between booking and arrival date
country	Nationality of guest, in ISO-3 form
distribution_channel	How a booking was made (e.g: TA, direct, corporate)
deposit_type	Refundable, nonrefundable, no deposit
arrival_date_month	Self explanatory
arrival_date_day_of_month	Self explanatory

Exploratory Data Analysis

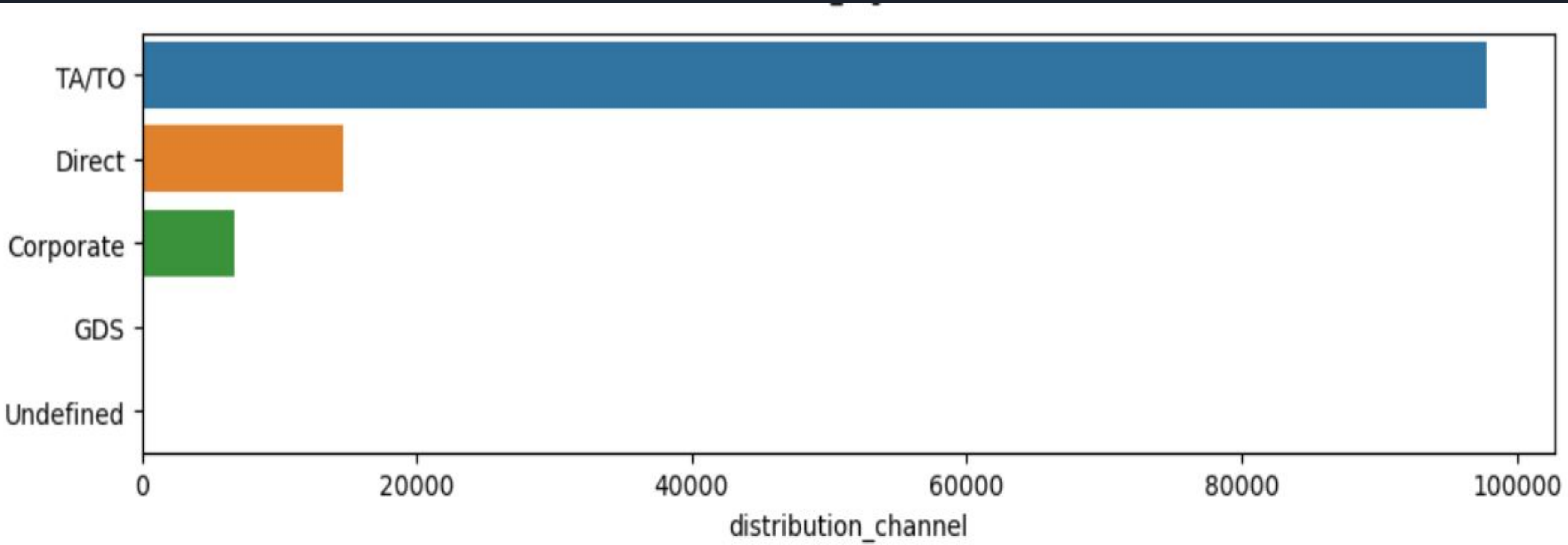
- Guests by region





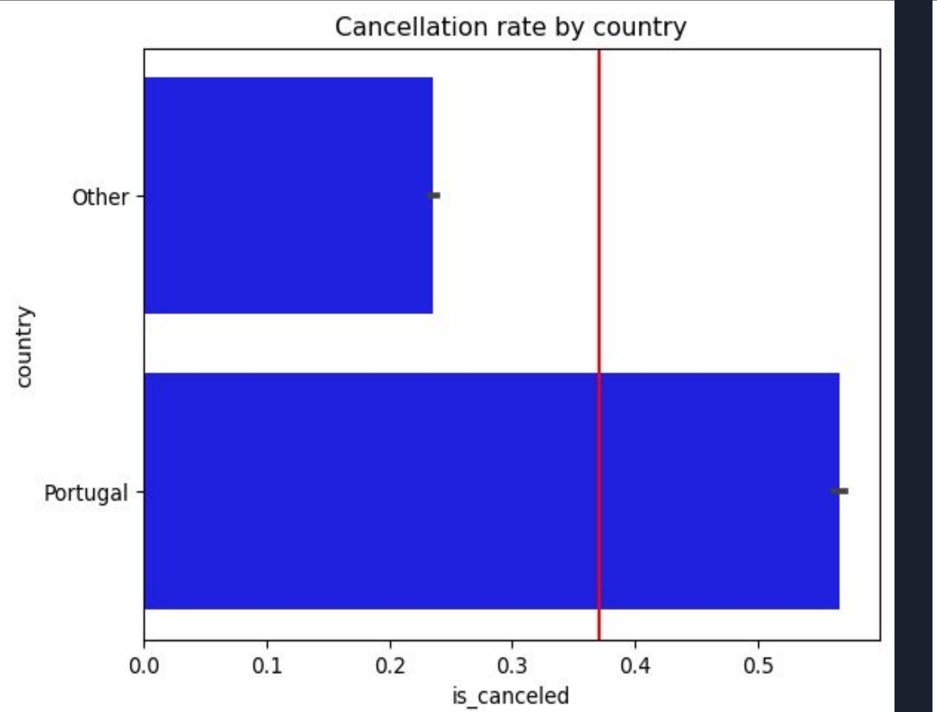
Exploratory Data Analysis

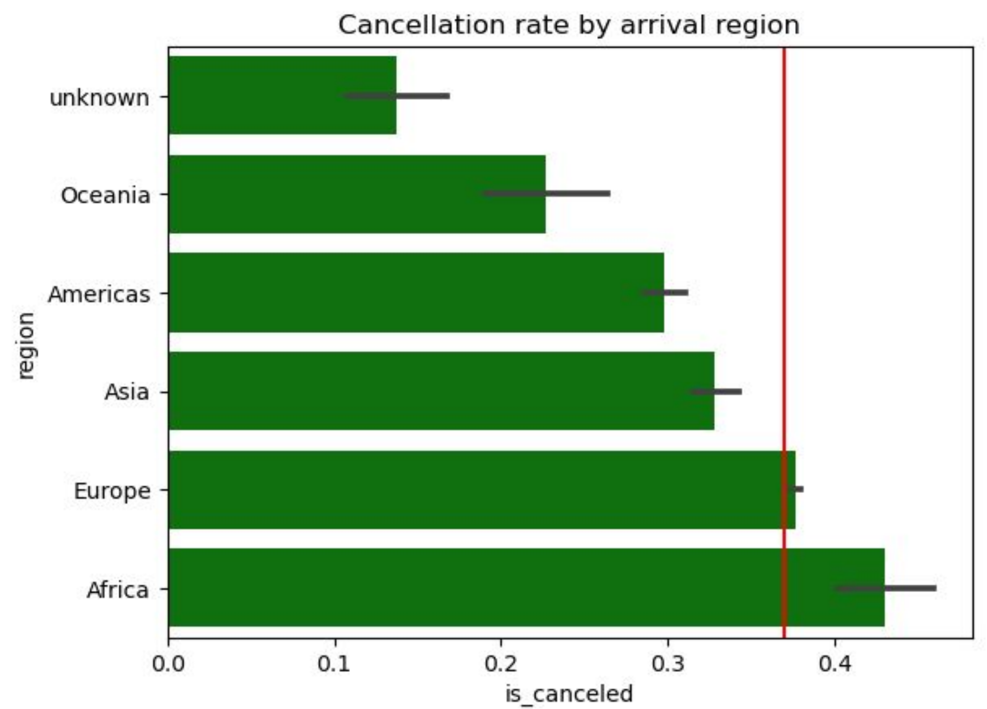
- Most Popular Distribution Channels



Exploratory Data Analysis

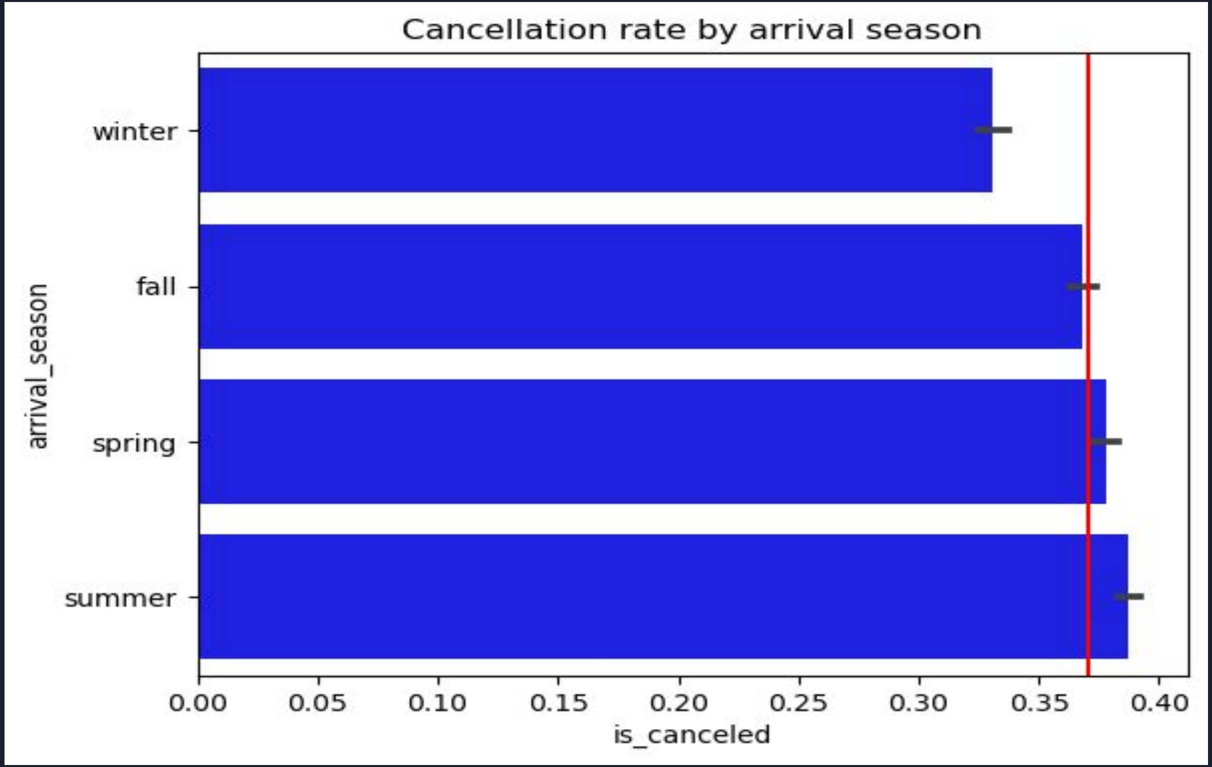
- Average cancellation rate is about 37%, denoted by red line





Exploratory Data Analysis

- Average cancellation rate is about 37%, denoted by red line



Machine Learning Modeling

Type:	Supervised Learning
Binary Classification:	1 for cancelled reservations, 0 for not cancelled
Imbalanced Data:	About 37% of data are labeled with class 1
Tools:	Python’s scikit learn

Algorithms used:

1. Logistic Regression
2. Random Forest

Model Comparison

Logistic Regression:

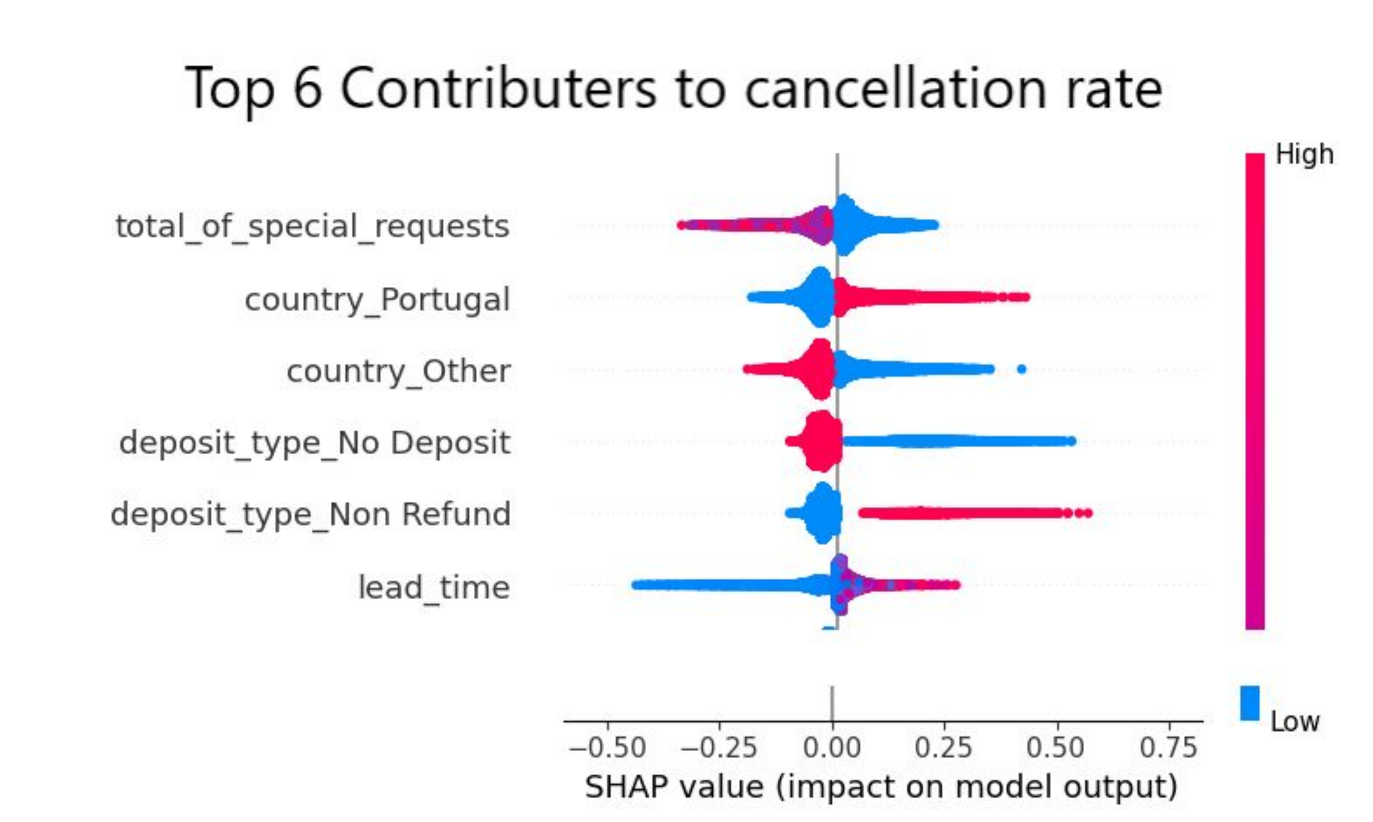
- **Accuracy**: 0.81
- Precision-Recall: 0.85
- ROC: 0.89

Random Forest:

- **Accuracy**: 0.83
- **Precision-Recall:** 0.88
- ROC: 0.92

Random Forest had the better scores on all 3 tests

Random Forest Results



Which features contribute to cancellation rate?

- Being a domestic traveller
  - Domestic travellers know they have more options in their home country
- Nonrefundable deposit type
  - Could be from Rewards program, less financial loss for the guest
- Longer lead time
  - They had more time to ‘shop around’
- Maybe spring and summer seasons
- Maybe being from Africa

Red = from random forest model

Green = from EDA, but needs more exploration

What can hotels do to reduce cancellation?

- Advertise more to international guests
- Offer a ‘local discount’ to retain domestic guests
- Check up on guests with a long lead time, perhaps starting 60 days before their arrival date