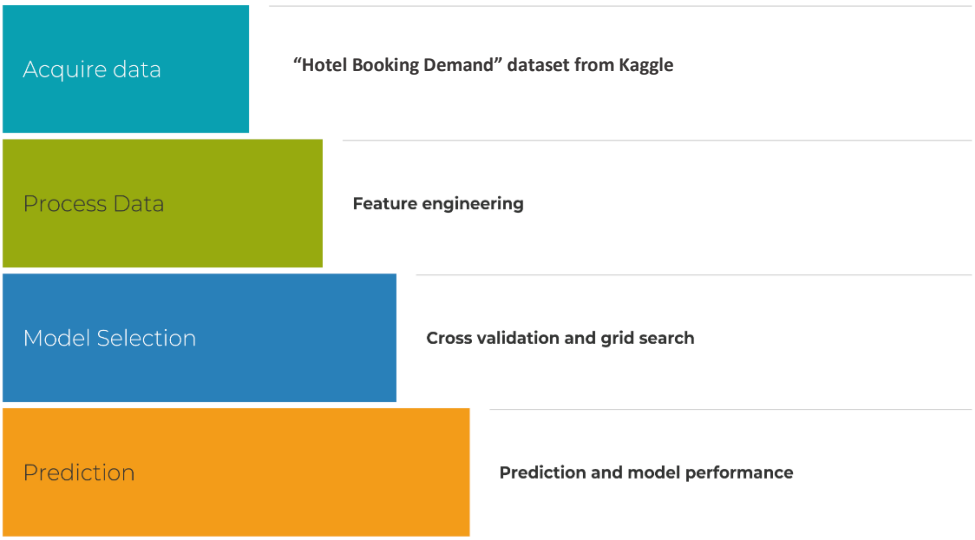
Project Title: **Predicting Hotel Bookings Cancellation**
Machine Learning Classification Model
By: Eman Alwusaybie

Abstract

Hotels get multiple bookings throughout the year and several of these bookings are cancelled. Cancelled bookings mean lost revenue for the business - if the management could predict whether a hotel booking is likely to get cancelled, it would be in their best interest to allocate the room to someone else and earn their business. This project aims to utilize current and prior booking information in order to predict whether a hotel booking is likely to get cancelled.

Design



Data

The dataset used to address this problem is "Hotel Booking Demand" from the Kaggle with 119.000 rows and 31 columns. You can download it from here:
https://www.kaggle.com/jessemostipak/hotel-booking-demand
This data set contains booking information for a city hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

Algorithms

1. Random Forest
2. Regularised Logistic Regression
3. XgBoost

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

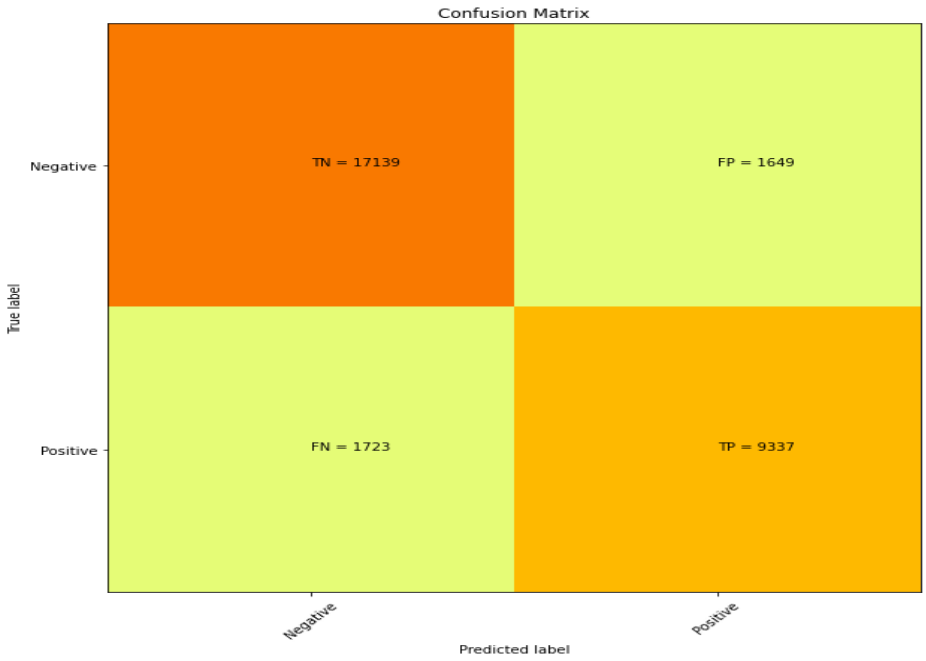
I used Logistic Regression, Random Forest and XgBoost to identify the propensity of a booking getting cancelled. After gridSearch and hyperparameter tunnig, I got the below results:

Logistic Regression: 0.82

Random Forest: 0.84

Xgboost: 0.89

The XgBoost model proved to be a best choice to predict whether a hotel booking would be cancelled given the current and prior booking information about the type of hotel, room, customer, stay, payment status, etc.



	precision	recall	f1-score	support
0	0.91	0.91	0.91	18788
1	0.85	0.84	0.85	11060
accuracy			0.89	29848
macro avg	0.88	0.88	0.88	29848
weighted avg	0.89	0.89	0.89	29848