**Hotel Booking Cancellation**

## Abstract

The goal of this project was to use classification models to predict with higher accuracy whether a booking will be cancelled. Using supervised machine learning techniques. I worked with data provided by *Kaggle* ( <https://www.kaggle.com>).

## Data

The dataset contains 119,391 observations with 32 features about booking information for a city hotel and a resort hotel, such as Arrival Time, when the booking was made, length of stay, the number of adults, children, and/or babies etc. There are 32 columns 12 of them are Categorical and 20 are Numerical.

## Algorithms

Feature Engineering

* Dropping columns that are not useful.
* Creating Numerical and Categorical data frames.
* Handling non-numeric data.
* Dealing with Missing Data and Correlation Matrix.
* Reviewing correlations between features.
* Splitting data into training and test set.

**Models**

1. Logistic Regression
2. Random Forest

Model Evaluation and Selection

The entire training dataset was split into 80%-20%. This will helps use the 80% of our data to train a model, and leave 20% of unseen data to test our trained model.

Accuracy Score of Logistic Regression is: 0.808153678382686

Precision score:

0.8458283606379516

Recall Score:

0.5889996977938954

f1\_score:

0.6944283614661738

Random Forest:

Accuracy Score of Random Forest is: 0.9547017867628554

Precision score:

0.9850509437113747

Recall Score:

0.8911302508310668

f1\_score:

0.935739785799286

## Tools

In this project will use the following libraries: pandas, numpy, matplotlib.pyplot, seaborn and sklearn.