Hotel Booking Cancellation Prediction

Abstract

Booking cancellations have substantial impact on hotel management systems. It affects the services, decrease revenue, and limits the production of accurate booking forecasts. To overcome the effect of cancellations, some hotels use strict cancellations policies and define overbooking tactics. However, these regulations have a negative impact on number of bookings and on hotel reputation which in turn leads to revenue loss. The goal of this project is to minimize the losses by using the power of data science and machine learning algorithms to build a classification model to predict which bookings are likely to cancel. This is helpful for revenue management, supplies purchases and pricing decisions as well.

Design

I used several machine learning algorithms such as logistic regression, k-nearest neighbors, decision tree, and random forest classifiers, before settling on random forest as the model with best performance.

Data

I used hotel booking demand dataset which contains more than 120,000 observations and 32 features. The data are retrieved from two different hotels: Resort hotel and City hotel. A few feature highlights include lead time, number of previous cancellations, and deposit type.

Algorithms

Data Preparation

- Handle missing values
- Encode categorical data
- Fix the data types
- Handling features
- Correlation

Models

- Logistic Regression
- Decision Tree Model
- Random Forest Classifier

Model Evaluation

The dataset was split into 80% for training and 20% for testing. The Classification Report of random forest model results are:

- Accuracy 0.885
- F1 0.87 macro
- precision 0.88 macro
- recall 0.87 macro

Tools

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