Analysis of the Hotel Demands and Bookings

Abstract

The goal of this project is to predict future hotel booking cancellation. Predicting future booking cancellation can help hotels in increasing operational efficiency, optimizing their marketing strategies, and maximising revenue. I utilised public hotel booking dataset provided by Kaggle and performed an explanatory data analysis that revealed interesting patterns and trends. Following that, several classification models were trained and evaluated using 5-fold cross validation. Based on the results obtained, random forest is the best classifier for booking cancellation prediction with an f-macro of 0.89.

Design

This project originates from the Kaggle competition "Hotel booking demand". The data presents a binary class status of bookings for hotels in different countries. Building a machine learning model that classify booking statuses accurately can help hotels plan for cancellation and refund policies, staffing schedules as well as targeting customers with offers and discounts. It is also important to understand key booking cancellation factors and how those factors relate to booking cancellation.

Data

The dataset consists of 119,390 observations with 32 features. The individual sample/unit of analysis in this project is a single booking made by a hotel customer. There are 32 features related to the booking, including booking date, lead time, number of adults, children, babes, deposit type and previous cancellations. The heatmap of pairwise correlation of all columns in the datafram reveals statistically significant correlations between the target variable and reservation_ status, lead_time, country, deposit_type. By checking reservation_ status values, it appears that it is highly correlated with the target and should be eliminated from further analysis. Moreover, an explanatory data analysis was undertaken to inform baseline models and feature engineering.

Algorithms

Feature Engineering & Selection

1. Converting features such as [total_of_special_requests, required_car_parking_spaces, booking_changes] to binary variables to highlight strong signals.
2. Mapping reserved_room_type and assigned_room_type features to a new binary feature "Got Required Room"
3. Converting categorical features to binary dummy variables
4. Scaling numerical features

Models

Naïve base, Logistic regression, k-nearest neighbors, Support vector machine and random forest classifiers were trained, fine tuned and evaluated using kfold cross-validation. The best overall model was Random Forest followed by SVM.

Model Evaluation and Selection

The dataset was split using stratified splitting into 80% for training and 20% for testing. All the models were fine tuned using either grid search or random search with 10-fold cross validation. The performance metrics for the models with the best hyperparameter were then calculated using 10-fold cross validation. Finally, the performance of the best model on the test set is reported here. Since the data is unbalanced F-macro is used as the mean metric together with macro precision and macro recall. Table 1 shows the final random forest 10-fold CV scores.

Metric	Cross Validation	Holdout
F1	0.88	**0.89**
Precision	0.89	0.89
Recall	0.81	0.88

Tools

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

Communication

Please check the notebook https://github.com/safa212/Metis-Four-Week-Data-Science-Bootcamp-/blob/main/Project%231/HotelBookingPrediction-FP-SafaAlsafari.ipynb