Executive Summary on

Hotel Booking Cancellation Prediction

Overview

This repository contains a project focused on predicting hotel booking cancellations. The dataset includes a variety of features related to hotel bookings, such as lead time, deposit type, and special requests. The primary objective is to develop a predictive model that accurately determines whether a hotel booking would be canceled. This prediction is essential for hotels as cancellations can significantly impact revenue and operational planning.

Problem

In this project, we aim to build a predictive model to determine whether a hotel booking would be canceled, which is crucial for hotels as cancellations affect revenue and operational planning. The dataset contains a high number of features related to booking, such as lead time, deposit type, and special requests, which adds to the complexity of the model. The challenge lies in the data preprocessing steps, which include feature selection and engineering, handling missing values, and noise in the data. Additionally, we are going to train different models, evaluate their performance using the right metrics, and interpret the model by analyzing the most important features in the context of hotel booking cancellations.

Objectives

The objectives of the project are as follows:

- Explore the Dataset: Investigate the dataset's basic information, including summary statistics for numerical and categorical variables.
- Preprocessing Steps:
    - Select and engineer features to enhance predictive performance.
    - Handle missing values effectively.
    - Address noise in the data to improve model robustness.
    - Encode categorical variables for use in machine learning models.
- Model Building:

o Implement and fine-tune classification models, including Decision Trees, Random Forest, and logistic Regression.

o Emphasize achieving high F1-score for class 1 predictions to comprehensively identify booking cancellations.

• Evaluate and Compare Model Performance: Utilize accuracy, precision, recall, F1-score, and AUC to gauge models' effectiveness.

• Analyze Feature Importance: Understand which features have the most significant impact on model predictions and interpret their relevance in the context of hotel booking cancellations.

Dataset Description

The dataset comprises various metrics related to hotel bookings. The features of the dataset are described in the table below:

1. hotel :(H1 = Resort Hotel or H2 = City Hotel).
2. is_canceled Value: showing if the booking had been cancelled (1) or not (0).
3. lead_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. arrival_date_year: Year of arrival date.
5. arrival_date_month: The months in which guests are coming.
6. arrival_date_week_number: Week number of year for arrival date.
7. arrival_date_day_of_month: Which day of the months guest is arriving.
8. stays_in_weekend_nights: Number of weekend stay at night (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. stays_in_week_nights: Number of weekdays stay at night (Monday to Friday) in the hotel.
10. adults: Number of adults.
11. children: Number of children.
12. babies: Number of babies.
13. meal: Type of meal booked.
14. country: Country of origin.
15. market_segment: Through which channel hotels were booked.
16. distribution_channel: Booking distribution channel.
17. is_repeated_guest: The values indicating if the booking name was from a repeated guest (1) or not (0).

18. previous_cancellations: Show if the repeated guest has cancelled the booking before.
19. previous_bookings_not_canceled: Show if the repeated guest has not cancelled the booking before.
20. reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
21. assigned_room_type: Code for the type of room assigned to the booking. Code is presented instead of designation for anonymity reasons.
22. booking_changes: How many times did booking changes happen.
23. deposit_type: Indication on if the customer deposited something to confirm the booking.
24. agent: If the booking happens through agents or not.
25. company: If the booking happens through companies, the company ID that made the booking or responsible for paying the booking.
26. days_in_waiting_list: Number of days the booking was on the waiting list before the confirmation to the customer.
27. customer_type: Booking type like Transient – Transient-Party – Contract – Group.
28. adr: Average Daily Rates that described via way of means of dividing the sum of all accommodations transactions using entire numbers of staying nights.
29. required_car_parking_spaces: How many parking areas are necessary for the customers.
30. total_of_special_requests: Total unique requests from consumers.
31. reservation_status: The last status of reservation, assuming one of three categories: Canceled – booking was cancelled by the customer; Check-Out
32. reservation_status_date: The last status date.

File Descriptions

- 📁 Hotel_Booking.csv: CSV file containing the hotel booking dataset.

Conclusion

It is concluded that the Random forest classification model performs the best for this type of task. It has highest accuracy score and ROC/AUC score (capability of model to distinguish between the classes)