CS 513 B – KDD PROJECT PROPOSAL

Forecasting Cancellation Flags: A Data-driven Approach to Hotel Reservation Cancellation Prediction

Project Group No: 1

Problem Statement:

The high rate of cancellations and no-shows in online hotel reservations has become a challenge for hotels as it impacts their revenue and occupancy rates. While customers benefit from the flexibility of free or low-cost cancellations, hotels have to deal with the revenue-diminishing effect of empty rooms. Hence, there is a need to explore strategies that can help hotels reduce cancellations and no-shows while maintaining customer satisfaction and loyalty. Therefore, the problem at hand is to **predict the cancellation flag of a hotel booking** based on a set of features, in order to assist hotels in managing their resources and revenue more efficiently.

Dataset:

The file includes a variety of attributes related to customers' reservation details, as described below:

- Booking_ID: A unique identifier for each booking
- no_of_adults: The number of adults associated with the booking
- no_of_children: The number of children associated with the booking
- no_of_weekend_nights: The number of weekend nights (Saturday or Sunday) that the guest stayed or booked to stay at the hotel
- no_of_week_nights: The number of week nights (Monday to Friday) that the guest stayed or booked to stay at the hotel
- type_of_meal_plan: The type of meal plan booked by the customer
- required_car_parking_space: A binary indicator (0 or 1) indicating whether the customer requires a car parking space
- room_type_reserved: The type of room reserved by the customer (ciphered or encoded by INN Hotels)
- lead_time: The number of days between the booking date and the arrival date
- arrival_year: The year of the arrival date
- arrival_month: The month of the arrival date
- arrival_date: The date of the month for the arrival date
- market_segment_type: A designation for the market segment associated with the booking
- repeated_guest: A binary indicator (0 or 1) indicating whether the customer is a repeated guest
- no_of_previous_cancellations: The number of previous bookings that were canceled by the customer prior to the current booking

- no_of_previous_bookings_not_canceled: The number of previous bookings that were not canceled by the customer prior to the current booking
- avg_price_per_room: The average price per day of the reservation, taking into account dynamic room pricing (in euros)
- no_of_special_requests: The total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: A flag indicating whether the booking was canceled or not.

Source of Dataset: https://www.kaggle.com/datasets/naveenrenji/hotel-resource-management-dataset

Implementation Strategy and algorithms used:

To conduct this comparison, each group member will be assigned two models to analyze. The four group members will be responsible for comparing the selected models and evaluating the performance of each model. For this analysis, we will identify the strengths and weaknesses of the models and document our findings. We will then use the results of this evaluation to make a recommendation for the best model to implement.

We will use algorithms like :
1. Decision Trees
2. Random Forest
3. Logistic Regression
4. KNN
5. Bagging algorithms like The Random Forest
6. Boosting algorithms like LightGBM, XGBoost, CatBoost, AdaBoost

Algorithms will be utilized in tandem with GridSearchCV and RandomSearchCV to conduct hyper-parameter tuning, allowing for the attainment of optimal results.

Model metrics and Evaluation:

Evaluation of different models used in project.

1. AUC-ROC
2. F!
3. Confusion Matrix
4. Precision

Team Members:

Group 1

1. Naveen Mathews Renji
2. Aatish Kayyath
3. Madhura Shinde
4. Abhishek Kocharekar