Predicting Hotel Booking Cancellation

Business Objective

A renowned hotel chain has in the recent past been confronted with the problem of increased cancellations. The management has observed that there are many instances where guests are booking a hotel room and are then making a cancellation. It has been decided by the business that they would want to implement a new solution which can proactively alert the staff about a booking at risk of cancellation.

Problem Statement

You have been assigned the task of building an SVM model which can predict the customers who are at risk of cancelling their booking. You need to build a model that can predict the is cancelled column.

Tools: Python

Data Description

The dataset provided for this activity consists of 119390 rows and 32 columns. You can find the dataset here. There are many categorical variables in the data whose category mappings are saved in Json files over here.

Below is a brief description of each column in the dataset:

- hotel: categorical variable, information on category of hotel
- is_cancelled: target variable to be predicted.
- lead_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
- arrival_date_year: Year of arrival date
- arrival_date_month: Month of arrival date
- arrival_date_week_number: Week number of year for arrival date
- arrival_date_day_of_month: Day of arrival date
- stays_in_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
- stays_in_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- adults: Number of adults
- children: Number of children
- babies: Number of babies
- meal: Type of meal booked.
- country: Country of origin
- market_segment: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.

- distribution_channel: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- is_repeated_guest: Value indicating if the booking name was from a repeated guest (1) or not (0)
- previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.
- previous_bookings_not_canceled: Number of previous bookings not cancelled by the customer prior to the current booking.
- reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- assigned_room_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due
- booking_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS.
- deposit_type: Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No
- agent: ID of the travel agency that made the booking.
- company: ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for
- days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type: Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of
- adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- required_car_parking_spaces: Number of car parking spaces required by the customer
- total_of_special_requests: Number of special requests made by the customer (e.g. twin bed or high floor)
- reservation_status: Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out
- reservation_status_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to

Model Building

Task - 1

- Do an exploratory data analysis and find out the list of candidate predictors. You can use bivariate plots that measure percentage of canceled bookings across different categories in case of a categorical variable and across deciles in case of continuous variables.
- You will need to use the Json files (details in the data section) to map the values of some of the categorical variables listed in the dataset.

Task - 2

- You need to experiment with different kernels (linear, RBF, polynomial etc) in the SVM model. In your experiments provide details of:
    - The kernel used and the classifier performance on either a single validation set or over folds of data where you have done a cross validation.
    - The training times as well as inference times for each hyperparameter you've chosen.
- Justify the model that you finally selected.
- Save your final model by using joblib or pickle.

Task - 3

- Create a code pipeline to read the saved model and do predictions using that. This should ideally be a separate python file with the logic to load and do inference coded in either a python function or python class.
- Provide your understanding of the next steps that the client/ end-user needs to follow to deploy your model at their end. Think about the below lines:
    - Any technical/infrastructure requirements that the client needs to meet?
    - What files do you need to provide them?
    - What kind of data cleaning and preprocessing would the client need to do before using the model?
    - How will the client use your model on new data?
    - How will the client know that the model is performing well on new data points?
    -

Model Validation (Task - 4)

- Use a k-fold validation strategy. You should track either all or a combination of following metrics:

i. AUC
ii. Confusion Matrix
iii. Accuracy
iv. F1 score

Deliverables

- A well-designed deck outlining the conclusions and the analysis (.ppt)
- A well-structured code pushed on GitHub (Write an informative README, well-structured code/notebooks)

- *Optional*: A blog post on medium/personal blog/blogger/LinkedIn