CS 663 - Machine Learning Spring, 2023

Lab 01 – Hotel Reservation Cancellations

The ability to book hotels online and the ease at which different options are presented before to users has dramatically changed booking possibilities and customers' behaviour. Many customers chose to cancel reservations at the very last moment and impacts hotels' business significantly. This is often made easier by the option to do so free of charge or preferably at low cost, which is beneficial to hotel guests but is a less desirable and possibly revenue-diminishing factor for hotels to deal with.

In this lab, your task is to understand a dataset and derive answers for questions that could be helpful for determining import features for understanding customers who are likely to cancel booking in the future. You are expected to use techniques in pandas to manipulate the dataset, visualize it and reach a conclusion. (Note: You have the freedom to use any package for plotting but recommended packages are matplotlib and seaborn.)

Data

The dataset is available on **Canvas**:

Name: **Hotel_Reservations.csv**

Process

Your implementation must:

- Browse to and start a repo from there. (In the future, there may be "starter code" for you.) Choose your name from the list and accept the assignment. If you do not see your name, contact the instructor / TA.
- Use a single Jupyter notebook to answer the questions. **Make sure your submission has executed cells and output**.
- Read the data into a data frame, which you may assume is in the same directory as the notebook.
- Inspect the dataset to determine the total count of non-numerical columns and numerical columns along with identifying the total null values for each columns.
- Determine the percentage of cancelled vs. not cancelled bookings.
- Combine the no_of_weekend_nights and no_of_week_nights and call the new column called 'total_days'. (This is often a process involved in feature engineering where you combine different columns to obtain a single feature.)
- Calculate the percentage of bookings made for different number of days and plot a bar graph in increasing order of number of days.
- Create a categorical variable called 'day' with 4 different classes, namely: weekends, weekdays, both, none… depending upon whether the customer has booked exclusively for weekends, exclusively for weekdays, etc.
- Create 2 separate plots based on 'no_of_weekend_nights' and 'no_of_week_nights' in regards to Cancelled vs Not Cancelled bookings and derive conclusions based on these plots. (Hint: explore the seaborn package.) Explain any conclusions you reach in markdown cells.
- Create a box plot to understand how avg_price_per_room is distributed as per the new 'day' column generated above and derive a conclusion in regards to how it impacts the booking. Explain any conclusions you reach in markdown cells.
- Create a bar plot to understand how 'no_of_special_requests' impacts the booking made by the customer. Explain any conclusions you reach in markdown cells.
- Generate a new feature called 'no_of_people' that is the same of the number of children and adults and generate a plot in the descending order of the count of number of people.
- Lastly, plot the 'repeated_guests' with respect to booking_status and convey the meaning based upon the plot.

Code Quality Instructions

- Assignments will be checked for plagiarism. Any work determined to be plagiarised will receive no credit and may be referred to Academic Integrity as a violation.
- Create a proper markdown cells for each steps in the ‘Process’ and neatly code the solutions for the same.
- While this assignment requires that you generate a number of plots in matplotlib or seaborn, the purpose of the assignment is to derive conclusions about the data from the plots. As such, your notebook should show the conclusions clearly, and these conclusions must be derived from the required work.

Grading

Grades for this assignment will be determined by the grader as follows:

- 100% = Code functions, is well-documented and clearly shows visual representations as per implementation requirements mentioned above.
- 75% = Code functions but is not well-documented or does not clearly show visual representations as per implementation requirements mentioned above.
- 50% = Code functions but is not well-documented -AND- does not clearly show visual representations as per implementation requirements mentioned above.
- 0% = No submission / code does not function

Submission

Submit the link to your GitHub repository on Canvas.