
# AIAP® Technical Assessment Past Years Series:

# Hotel No-Show Prediction



***AI Singapore***
*innovation 4.0, 3 Research Link #02-04 Singapore 117602 | www.aisingapore.org*
*AI Singapore is a national programme (AISG-RP-2019-050) supported by the National Research Foundation and hosted by the National University of Singapore (Company Registration No: 200604346E)*

# Objectives

Your objective is to predict the No-Show of customers (using the dataset provided) to help a hotel chain to formulate policies to reduce expenses incurred due to No-Shows. In your submission, you are to evaluate at least 3 suitable models for estimating the customers' No-Show.

You are given access to the dataset that contains the customer records from the hotel chain. (a link to retrieve the dataset can be found in the Data section below). With the given dataset, you are to perform the following two tasks:

1. Exploratory Data Analysis
2. Build an End-to-end Machine Learning Pipeline

# Task 1 - Exploratory Data Analysis (EDA)

Using the dataset specified in the **Data** section, conduct an EDA and create an interactive notebook in **Python** that can be used as a presentation to explain the findings of your analysis. It should contain appropriate visualisations and explanations to assist readers in understanding how these elaborations are arrived at as well as their implications.

## Deliverable

1. Notebook in **Python**: an `.ipynb` file named **`eda.ipynb`**.
2. Your EDA should:
    - Outline the steps taken in the EDA process
    - Explain the purpose of each step
    - Explain the conclusions drawn from each step
    - Explain the interpretation of the various statistics generated and how they impact your analysis
    - Generate clear, meaningful, and understandable visualizations that support your findings
    - Organize the notebook so that is it clear and easy to understand

# Task 2: End-to-end Machine Learning Pipeline (MLP)

Design and create a machine learning pipeline in Python scripts (.py files) that will ingest/process the entailed dataset and feed it into the machine learning algorithm(s) of your choice.

The pipeline should be easily configurable to enable easy experimentation of different algorithms and parameters as well as different ways of processing data (e.g. usage of a config file, environment variables, or command line parameters). Within the pipeline, data must be fetched/imported using SQLite, or any similar packages (provided in the `Data` section).

## Deliverables

1. A folder named `**src**` containing Python modules/classes in `.py` format.
2. An executable bash script `**run.sh**` at the base folder of your submission to run the aforementioned modules/classes/scripts. DO NOT install your dependencies in the `run.sh`; this will be taken care of automatically when we assess the assignment if you have created your `requirements.txt` correctly.
3. A `**requirements.txt**` file at the base folder of your submission.
4. A `**README.md**` file that sufficiently explains the pipeline design and its usage. You are required to explain the thought process behind your submitted pipeline in the README. The README is expected to contain the following:
    a. Full name (as in NRIC) and email address (stated in your application form).
    b. Overview of the submitted folder and the folder structure.
    c. Instructions for executing the pipeline and modifying any parameters.
    d. Description of logical steps/flow of the pipeline. If you find it useful, please feel free to include suitable visualization aids (eg, flow charts) within the README.
    e. Overview of key findings from the EDA conducted in Task 1 and the choices made in the pipeline based on these findings, particularly any feature engineering. Please keep the details of the EDA in the `.ipynb`. The information in the `README.md` should be a quick summary of the details from `.ipynb`.
    f. Describe how the features in the dataset are processed (summarized in a table).
    g. Explanation of your choice of models.
    h. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.

## Dataset

The dataset contains the customer records from a hotel chain. Do note that there could be synthetic features in the dataset. Hence, please ensure that you state and verify any assumptions that you make.

You can retrieve the dataset using the following URL:
https://techassessment.blob.core.windows.net/aiap-pys-2/noshow.db

## Instructions for setting up SQLite and querying the database

The dataset can be accessed through the `noshow.db` file. You may find either of the following packages, `SQLite` or `SQLAlchemy`, useful for accessing this database.

You should place the `noshow.db` file in a `data` folder. Your machine learning pipeline should retrieve the dataset using the relative path `data/noshow.db`.

DO NOT submit the `noshow.db` in your final submission.

## List of Attributes

| Attribute | Description |
|---|---|
| booking_id | Unique customer booking ID |
| no_show | If the customer is a No-Show: 0 = Show, 1 = No-Show |
| branch | Hotel branch |
| booking_month | Month the booking was made by the customer |
| arrival_month | Month the customer plan to arrive at the hotel |
| arrival_day | Day date the customer plan to arrive at the hotel |
| checkout_month | Month the customer plan to checkout of the hotel |
| checkout_day | Day date the customer plan to checkout of the hotel |
| country | Nationality of the customer |
| first_time | If it is the first time customer staying in the hotel |
| room | Room type booked by the customer |
| price | Price of the room booked by the customer |
| platform | Platform used to book the room by the customer |
| num_adults | Number of adults staying |
| num_children | Number of children staying |