Data Science Home Challenge

Please submit a Jupyter notebook (or similar) with your code, answers and assumptions. Please also submit a version of the notebook saved as a PDF or HTML document.

We do not expect you to spend more than 3 hours on this challenge.

Mail your answers to **ds-home-challenge@prospective.io**. If you have any queries about the tasks send them to the same email address.

NB: When answering a question please write down any assumptions that you make.

Task 1 - Processing and visualising data

On-Time-Performance (OTP) is used to assess how successful a bus company is at delivering their published schedule for a particular route. In this case it will be defined as the percentage of scheduled stops that are reached less than 300 seconds (5 minutes) late. Only the first, last and specified intermediate timing-point (TP) stops should be considered. For the first and intermediate stops the scheduled stop departure time should be used for comparison while the scheduled arrival time should be used for the last stop.

You are given the following data files:

timing_observations.csv

The dataset contains scheduled and observed stop arrivals and departures for route 56 in the outbound direction:

- **report_date**: The date the trip took place.
- **trip_id**: The id of the observed trip.
- **scheduled_start_time**: The time the observed trip was scheduled to start.
- **bus_stop_id**: An id for each stop of the trip.
- **sequence_number**: The sequence the stops are called at.
- **timing_point**: A boolean which specifies whether a particular stop should be considered when calculating OTP.
- **scheduled_arrival**: The scheduled arrival time at each stop.
- **scheduled_departure**: The scheduled departure time from each stop.
- **observed_arrival**: The observed arrival time at each stop.
- **observed_departure**: The observed departure time from each stop.

T1Q1. Calculate and plot the daily OTP for route 56 during March 2022 ie. report_date vs OTP percentage. Note the OTP is expected to vary by day and be systematically different on weekends.

T1Q2. Only including data for weekdays, calculate and plot the hourly OTP for route 56. Use the scheduled_start_time hour to bin the data.

Task 2 - Modelling

You are given the following data files:

data_observations.csv

The dataset contains bus dwell time observations for a random sample of stops. The data dictionary is as following:

- "stop_id": (int) stop identifier
- "date": (str) date of the observation
- "wet_weather_score": (float) a score between 0 and 1 proportional to the amount of rain in the past 24 hours
- "boardings": (int) the number of observed boardings to the bus at the stop
- "alightings": (int) the number of observed alightings from the bus at the stop
- "dwell_time": (float) the time in seconds that the bus spent in the vicinity of the stop; note that the dwell time was recorded even in the instances when the bus did not stop.

data_stops.csv

The dataset contains additional information about the stops. The data dictionary is as following:

- "stop_id": (int) stop identifier
- "is_urban": (flag) indicator variable denoting if the stop is in an urban area

T2Q1. Implement a regression model that predicts the dwell time of a bus at a stop. Assume that only the following predictor variables are be available at prediction time:

- "stop_id": (int) stop identifier
- "wet_weather_score": (float) a score between 0 and 1 proportional to the amount of rain in the past 24 hours
- "boardings": (int) the number of observed boardings to the bus
- "alightings": (int) the number of observed alightings from the bus

Note that you may choose to only use a subset of above variables and you may also want to derive additional ones. Please state your reasoning. Train your model using data from *data_observations.csv*.

T2Q2. Choose appropriate evaluation measures and estimate and reflect on the performance of your model for each of the following:

- on the training data
- for stops with observations recorded in *data_observations.csv*
- for stops without previous observations