DigiHaul Data Science Lead
Take-home test

Problem statement:

Road haulage is essential for the people and businesses of the UK. Approximately 90% of all goods transported by land in Great Britain are moved directly by road. DigiHaul is a digital transport business, specialising in managing, consolidating and integrating data from both Carriers and Shippers to deliver seamless end-to-end logistics service.

Shippers book shipments on the DigiHaul platform, detailing the scheduled collection and delivery time windows / locations and required vehicle types for carriers to consider. Once a carrier accepts a job and collection is scheduled, DigiHaul's driver app facilitates real-time tracking of shipments through GPS signals, **subject to carriers granting permissions for location logging**.

Dataset:

The dataset include the following files, and the data access link will be provided via email:

1. "GPS_data.csv":

Column Name	Explanation
SHIPMENT_NUMBER	Unique identifier of shipment
LAT	Latitude of the position
LON	Longitude of the position
RECORD_TIMESTAMP	Time stamp[1]

2. "Shipment_bookings.csv"

Column Name	Explanation
SHIPMENT_NUMBER	Unique identifier of shipment
SHIPPER_ID	Shipper ID
CARRIER_ID	Carrer ID
VEHICLE_SIZE	Vehicle type
VEHICLE_BUILD_UP	Vehicle trailer build up
FIRST_COLLECTION_POST_CODE	Collection post code (can be NULL)
LAST_DELIVERY_POST_CODE	Delivery post code (can be NULL)
FIRST_COLLECTION_LATITUDE	Collection latitude
FIRST_COLLECTION_LONGITUDE	Collection longitude
LAST_DELIVERY_LATITUDE	Delivery latitude
LAST_DELIVERY_LONGITUDE	Delivery longitude
FIRST_COLLECTION_SCHEDULE_EARLIEST	Scheduled earliest collection time
FIRST_COLLECTION_SCHEDULE_LATEST	Scheduled latest collection time
LAST_DELIVERY_SCHEDULE_EARLIEST	Scheduled earliest delivery time
LAST_DELIVERY_SCHEDULE_LATEST	Scheduled latest delivery time

[1] "+0100" means British Summer Time

3. “New_bookings.csv” – same format as “Shipment_bookings.csv”, but for different time periods

Tasks:

1. Operational teams rely on KPIs like on-time collection and on-time delivery to gauge carrier performance. What percentage of shipments met the on-time delivery threshold (arriving no later than 30 minutes past the scheduled delivery window) between October 1st and December 31st, 2023? Please outline your assumptions.

2. Predict the likelihood of delay for the list of shipments in “New_bookings.csv” dataset.

3. Outline the technical design for deploying the prediction model through an online endpoint.

4. Develop a presentation summarising the analysis findings and suggested next steps tailored for senior non-technical stakeholders.

   Bonus points if the following could be considered:

   a. Demonstrate best practices in coding.
   b. Utilise additional data sources by making API calls.

Please send the GitHub link containing the presentation and python/R scripts to wenjia.tang@digihaul.com. You may be required to demonstrate the solution live during subsequent interviews.