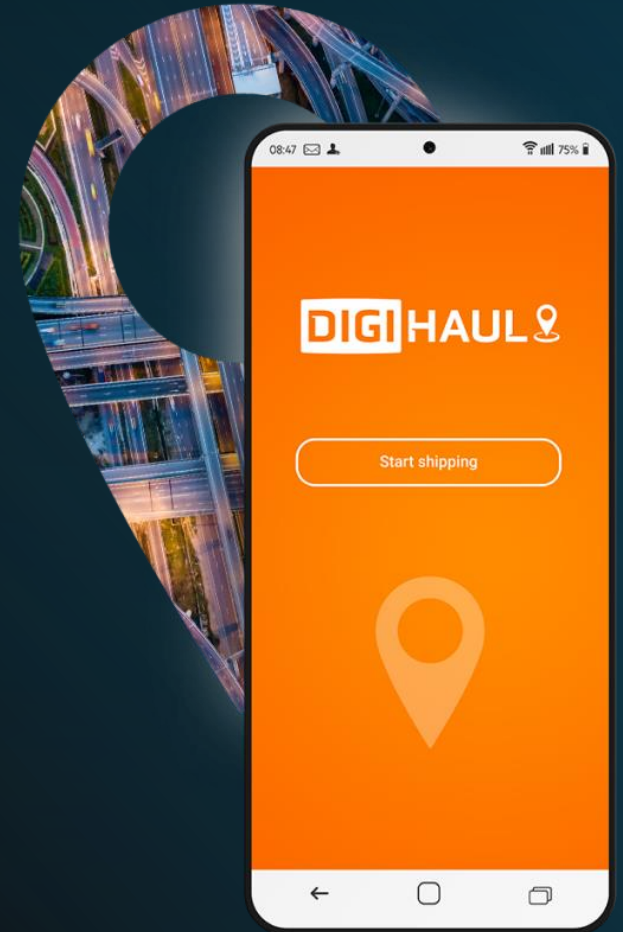


Data Scientist - Take Home Test - DigiHaul Ltd

Naga Chaitanya

Problem Statement

- Road haulage is critical for the people and businesses of the United Kingdom. In Great Britain, road transport accounts for 90% of all kinds of transport. DigiHaul is a digital transport company that manages and integrates data from both carriers and
- Shippers will deliver a seamless end-to-end logistical solution. They can use the DigiHaul platform to schedule pickup and delivery times, locations, and vehicle types for carriers to consider. After a carrier accepts an assignment and arranges a collection, DigiHaul's driver app provides for real-time tracking of shipments using GPS signals, if they grant permission for location logging.



Data Understanding

- GPS Data
 - Shipment Number – Unique identifier of Shipment
 - 3245 unique shipments
 - Longitude and Longitude of shipment
 - Record Timestamp – Date and time stamp as in when the GPS is recorded
 - Data is recorded in between 1st October 2023 and 30th December 2023
- The GPS record time stamp is used to track shipment numbers, calculating time in minutes from the first recorded date time stamp to the last recorded date time stamp.
 - For E.g., following shipment (SEZHUK-230918-245388) tracked starting from 9th Oct 2023 19:25 till 11th Oct 2023 11:50 and Actual time (in minutes) of delivering the same is 2425.

SHIPMENT_NUMBER	LAT	LON	RECORD_TIMESTAMP
SEZHUK-230918-245388	53.291930700000000	-1.0729478	2023-10-09 19:25:00
SEZHUK-230918-245388	53.400403700000000	-1.0785634	2023-10-09 19:40:00
SEZHUK-230918-245388	53.574181500000000	-1.2157612	2023-10-09 19:55:00
SEZHUK-230918-245388	53.706804800000000	-1.3365031	2023-10-09 20:10:00
SEZHUK-230918-245388	53.747500800000000	-1.6402984	2023-10-09 20:25:00
SEZHUK-230918-245388	53.713044400000000	-1.7440063	2023-10-09 20:41:00
SEZHUK-230918-245388	53.646777300000000	-1.9218641	2023-10-09 20:55:00
SEZHUK-230918-245388	53.571440600000000	-2.2195735	2023-10-09 21:10:00
SEZHUK-230918-245388	53.445307000000000	-2.4404546	2023-10-09 21:25:00
SEZHUK-230918-245388	53.400271500000000	-2.763526	2023-10-09 21:40:00
SEZHUK-230918-245388	53.439587100000000	-2.9406388	2023-10-09 21:55:00
SEZHUK-230918-245388	53.443561800000000	-3.006606	2023-10-09 22:11:00
SEZHUK-230918-245388	53.349105700000000	-6.2267774	2023-10-10 20:21:00
SEZHUK-230918-245388	52.277015300000000	-8.2762147	2023-10-11 11:50:00

First Recorded Time Stamp
2023-10-09 19:25:00

Last Record Time Stamp
2023-10-11 11:50:00

Difference between the first and last recorded time stamp in minutes is 2425

Data Understanding Cont...

- Shipment Bookings Data
 - Shipment Number – Unique identifier of Shipment
 - 3245 unique shipments
 - Project ID – Under which multiple shipment bookings are defined
 - First Pick up (collection) and last delivery post code
 - First Pick up (collection) and last delivery Longitude and Longitude of shipment
 - First pick up (collection) earliest time stamp – From 29th September 2023 to 31st December 2023
 - First pick up (collection) latest time stamp – From 1st October 2023 to 31st December 2023
 - Last delivery earliest time stamp – From 1st October 2023 to 3rd January 2024
 - Last delivery latest time stamp - From 1st October 2023 to 3rd January 2024
- Time to delivery a shipment (in minutes) is calculated using time difference be first pick up (collection) EARLIEST time stamp and Last Delivery LATEST Time Stamp
 - For E.g., following shipment (SEZHUK-230918-245388) tracked starting from 9th Oct 2023 10:00 till 11th Oct 11:50 and time taken to (in minutes) of delivering the same is 2155.

PROJECT_ID	3244246f-e37c-4ba4-b9b3-61416a4449ae
SHIPMENT_NUMBER	SEZHUK-230918-245388
CARRIER_DISPLAY_ID	CAUK000250
VEHICLE_SIZE	Tractor Unit
VEHICLE_BUILD_UP	Curtain-Side 13.6M Trailer
FIRST_COLLECTION_POST_CODE	S80 3FY
LAST_DELIVERY_POST_CODE	MITCHELSTOWN
FIRST_COLLECTION_LATITUDE	53.293393
FIRST_COLLECTION_LONGITUDE	-1.074008
LAST_DELIVERY_LATITUDE	52.276579
LAST_DELIVERY_LONGITUDE	-8.258109
FIRST_COLLECTION_SCHEDULE_EARLIEST	2023-10-09 00:00:00
FIRST_COLLECTION_SCHEDULE_LATEST	2023-10-09 23:59:00
LAST_DELIVERY_SCHEDULE_EARLIEST	2023-10-10 11:25:00
LAST_DELIVERY_SCHEDULE_LATEST	2023-10-10 11:55:00

FIRST_COLLECTION_SCHEDULE_EARLIEST	LAST_DELIVERY_SCHEDULE_LATEST
2023-10-09 00:00:00	2023-10-10 11:55:00

Difference between the first pick up (collection) earliest and last deliver latest time stamp in minutes is 2155

Tasks

- Task 1
 - Operational teams rely heavily 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.
 - Assumption :
 - **Time to delivery a shipment (in minutes)** is calculated using time difference between first pick up (collection) EARLIEST time stamp and Last Delivery LATEST Time Stamp i.e., shipment (SEZHUK-230918-245388) tracked starting from 9th Oct 2023 10:00 till 11th Oct 11:50 and time (in minutes) of delivering the same is 2155.
 - **Actual time GPS tracked** is calculated using the GPS data - For E.g., following shipment (SEZHUK-230918-245388) tracked starting from 9th Oct 2023 19:25 till 11th Oct 11:50 and Actual time (in minutes) of delivering the same is 2425.
 - On-time delivery threshold (arriving no later than 30 minutes past the scheduled delivery window) is Actual time GPS tracked – Time to delivery the shipment = **2425 – 2155 = 270 which is more than the threshold of 30 minutes, so this shipment is not a on time delivery**
 - **On time KPI for the shipment delivered in between 1st October 2023 to 31st December 2023 are 2395 (out of 3235 total shipment booked) which is of 74% on time**

Tasks Cont...

- Task 2
 - Timely communication of potential delays is crucial for shippers. During the 3-month period from 1st Oct to 31st Dec 2023, which shipper(s) should be notified automatically regarding potential late delivery of which shipments, and at what times?
 - **Potential late delivery** is more than the threshold of 30 minutes on time deliver, i.e. shipment (SEZHUK-230918-245388) tracked starting from 9th Oct 2023 10:00 till 11th Oct 11:50 and time (in minutes) of delivering the same is 2155 and from the GPS data actual delivering the same is 2425. So notification time will on LAST_DELIVERY_SCHEDULE_LATEST.

PROJECT_ID	SHIPMENT_NUMBER	notification_time
3244246f-e37c-4ba4-b9b3-61416a4449ae	SEZHUK-230918-245388	2023-10-10 11:55:00+00:00

- Similarly remaining shipment number which are potential late delivery can be notified.

Tasks Cont...

- Task 3
 - Predict the likelihood of delay for the list of shipments in “New_bookings.csv” dataset.
 - Shipment bookings data is considered as a training dataset with target variable on-time namely ‘1’ represent yes on time and ‘0’ as not time (likelihood of delay)
 - Delay can be of multiple reason, depending on day of week (Monday, ..), time of day (morning,...), vehicle build up, vehicle size.

- Day of week is calculated on the last delivery latest time stamp
- Time of day calculated on the last delivery latest time stamp
 - Between 5 am to 12 PM – Morning
 - Between 12 PM to 5 PM – Afternoon
 - Between 5 PM to 9 PM - Evening
 - After 9 PM before 5 am – Late Evening

- Vehicle Build up where the count of shipments are below 100 are considered as others

VEHICLE_BUILD_UP	Count of SHIPMENT_NUMBER
Curtain-Side 13.6M Trailer	1777
Box 44ft	472
Box	415
LWB Van	367
Curtain Side	78
Tractor Only	45
Tractor Only 6x2	39
Temperature-Controlled	31
40ft Container	14
Double Deck Trailer	4
High Security	1
Box 40ft	1
20ft Container	1

VEHICLE_BUILD_UP	Count of SHIPMENT_NUMBER
Curtain-Side 13.6M Trailer	1770
Box 44ft	472
Box	412
LWB Van	367
others	214

VEHICLE_SIZE	Count of SHIPMENT_NUMBER
Tractor Unit	2611
3.5T Van	536
3.5 T Van	63
18T Rigid	20
7.5T Rigid	11
26T Rigid	4

VEHICLE_SIZE	Count of SHIPMENT_NUMBER
Tractor Unit	2602
3.5T Van	535
others	98

- Vehicle Build up where the count of shipments are below 100 are considered as others

Tasks Cont...

- Task 3
 - Using different Machine learning classifier algorithms, shipment booking is trained
 - Logistic regression – is a simple statistical method that predicts the likelihood of an event happening based on input data.
 - Decision Tree is a visual tool that helps make decisions by breaking down complex choices into a series of simple, step-by-step questions
 - Random Forest is a method that improves decision making by combining the results of many decision trees to get more accurate and reliable predictions.
 - Support Vector Machine is method that finds the best boundary to separate different groups in the data for accurate predictions.
 - K-Nearest Neighbors a method that makes predictions based on the most common outcome among the closest data points to the input
 - Naive Bayes method that makes predictions by assuming that all input features are independent and calculating the probability of each outcome.
 - After the multiple iteration and using evaluation parameter - **Accuracy score** which is a metric that measures the percentage of correct predictions made by a model out of all predictions. Logistic regression has got the best accuracy score among the other algorithms (72%)
 - Same model is applied on the new booking data and attached the results



Microsoft Excel
Worksheet

Next Steps

- Calculate precision-recall curves for imbalanced classes.
- Use the AUC-ROC score to evaluate the trade-off between true positive rate and false positive rate.
- Assess model calibration to ensure predicted probabilities reflect actual outcomes.
- Set up a system for continuous data collection and model retraining.
- Incorporate user feedback to refine feature engineering and model architecture.
- Plan for periodic model evaluation and updates to ensure ongoing accuracy and relevance.

Thank you