Data Analyst Assignment

1. Know the domain

1. Get familiar with the following terms (~20 minutes)

Background:

The following terms are frequently used in the shipping industry and will be essential to know. Please define these in a separate document.

- AIS
- Freight forwarder
- Carrier
- voyage
- service lane
- carrier schedule / vessel schedule
- Shipper
- Standard Carrier Alpha Code
- Vessel IMO
- Port of Loading
- Port of Discharge
- Bill of Lading (BL) Number
- Master BL
- House BL
- Transshipment Port
- ATA

- ATD
- ETD
- ETA

2. What's wrong with the carrier schedule? (~ 5 min)

After we made a wrong prediction, we have received port calls from the client (consider portcall data a truth dataset). Reason of making wrong prediction was the schedule data we received. Can you find what is wrong with the schedule?

Port calls:

{
"actual_departure_date": "2021-09-08T23:08:55+00:00",
"actual_arrival_date": "2021-09-07T11:31:10+00:00",
"port_code": "KRPUS"
},
{
"actual_departure_date": "2021-09-11T06:53:37+00:00",
"actual_arrival_date": "2021-09-10T14:22:00+00:00",
"port_code": "CNQDG"
},
{
"actual_departure_date": "2021-09-17T13:24:32+00:00",
"actual_arrival_date": "2021-09-16T15:30:31+00:00",
"port_code": "CNSHG"
},
{
"actual_departure_date": "2021-09-21T12:01:15+00:00",
"actual_arrival_date": "2021-09-20T15:33:45+00:00",
"port_code": "CNCWN"
}

Current schedule:

{

"schedule_source": "Source A",
"scheduled_arrival_date": "2021-09-08T04:00:00",
"scheduled_departure_date": "2021-09-08T18:00:00",
"actual_arrival_utc_date": "2021-09-07T11:25:09+00:00",
"actual_departure_utc_date": "2021-09-08T23:05:09+00:00",
"scac_code": "OOLU",
"port_code": "KRPUS"
},
{
"schedule_source": "Source A",
"scheduled_arrival_date": "2021-09-10T20:00:00",
"scheduled_departure_date": "2021-09-11T14:00:00",
"actual_arrival_utc_date": "2021-09-10T13:48:00+00:00",
"actual_departure_utc_date": "2021-09-11T06:12:55+00:00",
"scac_code": "OOLU",
"port_code": "CNQDG"
},
{
"schedule_source": "Source A",
"scheduled_arrival_date": "2021-09-16T22:00:00",
"scheduled_departure_date": "2021-09-17T21:00:00",
"actual_arrival_utc_date": "2021-09-16T15:24:43+00:00",
"actual_departure_utc_date": "2021-09-17T13:31:26+00:00",
"scac_code": "OOLU",
"port_code": "CNSHG"
},
{
"schedule_source": "Source A",
"scheduled_arrival_date": "2021-09-20T14:00:00",
"scheduled_departure_date": "2021-09-21T10:00:00",
"actual_arrival_utc_date": null,
"actual_departure_utc_date": null,
"scac_code": "OOLU",
"port_code": "CNSHK"
},
{

"schedule_source": "Source A",
"scheduled_arrival_date": "2021-09-25T15:00:00",
"scheduled_departure_date": "2021-09-26T11:00:00",
"actual_arrival_utc_date": null,
"actual_departure_utc_date": null,
"scac_code": "OOLU",
"port_code": "SGSIN"
}

2. Accuracy Report (2 hours maximum)

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/ca049de9-de38-4323-bfe5-2ebd5245c44c/results_2021-09-30.xlsx

Data: The data shared is the final prediction results (Estimated Time of Arrival) for some of the vessels we have been tracking.

vessel_live_info_id - Unique id to identify the prediction

scheduled_arrival_utc - Arrival time scheduled by the carrier

estimated_arrival_utc - pETA - This is prediction generated by the Portcast

actual_arrival_utc - Actual time of arrival

imo - vessel id

portcode - port code

timestamp_utc - timestamp when we generated prediction

Goal: Goal is to generate final accuracy report (Jupyter Notebook/google colab) for our dummy client based on the prediction data above. Report is a combination of code (python/sql), notes and visualisations. You need to analyse how our predictions

(estimated_arrival_utc) perform vs. schedule declared by the carrier (scheduled_arrival_utc) on various criteria. Accuracy report should have clean, reusable code, detailed analysis, interpretable visualisations (spend some time on your choice of visualisation (box plot vs. line plot)). Suggestion - use plotly.

Metrics we want to cover are mentioned below. Take creative freedom in generating different scenarios and comparing accuracies. Make use of all the columns in slicing and dicing the data for checking accuracy.

Metrics

- **Actual View [days ahead = (ATA - pTimestamp)]**

  Create, measure, analyse following metrics by days_ahead (actual time of arrival - timestamp when we predicted)

  1. **Median Absolute Error** $median(|ATA -_p ETA|)$
  2. **OTPA** (On Time Percentage Arrival) $\frac{count(ATA-X<=_pETA<=ATA+Y)}{n} * 100$
     a. "On time Percentage Arrival" is "What % of predictions fell into certain range of the actual arrival"
     b. Decide different ranges like (for example -3 to +3 days) and calculate the accuracy.

- **Prediction View [days ahead = (pETA - pTimestamp)]**

  How to interpret - "how accurate the prediction is when we predict vessel is going to arrive after X days?"

  All the metrics from actual view.