**STUDY:**

***PREDICTING VESSEL EXPECTED TIME OF ARRIVAL (ETA) USING MACHINE LEARNING (started: Mar 27)***

Data is from automatic identification system (**AIS**) based on global positioning system ( GPS ) and SATELLITE feed from ships around the world.

This information is used to avoid collision among ships and is mandatory for cargo ships with weight of 300 Gross Ton or more and also for all passenger ships. It sends data like position, heading and speed in every 10 to 15mins. My company receives around 90% of all AIS data that includes vessel movements from other carrier companies.

**Sources of noise data:**

There are much data to clean on AIS data. The table requires advance skills in SQL to determine which records are linked together and what set of records are complete. By inspection, some the data is not complete with skipped date and timestamp.

A. Destination data has the following characteristics:

1. Filled with non-characters such as >>>, %, ! and comma (,)
2. Combines country code and port code such as CN SHA is China Shanghai
3. Space between port names such as ZHANG JIA GANG is ZHANGJIAGANG
4. NonSpace between country and port code such as SGSIN (SG Singapore as country, SIN as port)
5. Format such as Origin > Destination like HKHKG > CNYTN
6. Multiple spaces between port name such as HONG KONG is Hong Kong
7. Mispelled names like PYONGTAEK should be PYeONGTAEK
8. Abbreviation such as SPORE (Singapore) HK (Hong Kong) J.ALI (Jebel Ali)

In my study, I took Origin=Ningbo and Destination=Shanghai which is a very common pairs of port since they are nearby. However, the port Ningbo and Shanghai alone has 626 and 1025 variations of destination names that relates to Ningbo and Shainghai China. I choose the most common destination names accordingly.

B. Delete duplicate records

Records with the same vessel id and sequence numbers are duplicate and thus needs to be removed.

C. Skipped or Incomplete set of data

There are instances where the data is skipped and incomplete based on date and timestamp it was received. To know this, I check if the sequence numbers are in sequential order and if the latitude and longitude are within the ports. See sample SQL below:

SELECT VESSEL\_GID,

ORIGIN,

DESTINATION,

START\_AISO,

MIN(END\_AISO) AS END\_AISO

FROM

(SELECT T.VESSEL\_GID,

T.ORIGIN,

T.START\_AISO,

T.DESTINATION,

aiso\_seq AS END\_AISO ,

NVL(LEAD(aiso\_seq)

OVER (PARTITION BY T.VESSEL\_GID ORDER BY aiso\_seq)-aiso\_seq,1)

AS previous\_seq

FROM TMP\_GVVMC\_START\_END t, gvvmc\_ais\_obs g

WHERE g.vessel\_gid=t.vessel\_gid

AND g.aiso\_seq BETWEEN start\_aiso AND t.end\_aiso

AND g.destination = t.destination

)

WHERE previous\_seq > 10

GROUP BY VESSEL\_GID,

ORIGIN,

DESTINATION,

START\_AISO;

C. Data Source

There are three types of sources of data such as Satellite (SAT), Vessel Tracking (VT) and Terrestial. I omit the third type (terrestial) since I found that it also creates duplicate entries and are not in correct sequence.

D. Vessel type

Different vessel types are found in the AIS table such as cargo, cargo ships, fishing\_boats, high-speed\_crafts, highspeed, passenger\_ships, pleasure\_crafts, sailing\_vessels, tanker, tankships, towing\_vessel and tugboats. This study will only focus for cargo and carso ships types since my company is interested on these vessel types.

Conclusion:

All data wrangling was done in SQL since the data is in Oracle database. A copy of the entire sql used in the extraction can be found [here](https://github.com/anonyXmous/Predict_Vessel_ETA/blob/master/extract.sql).