Hi Dear all,

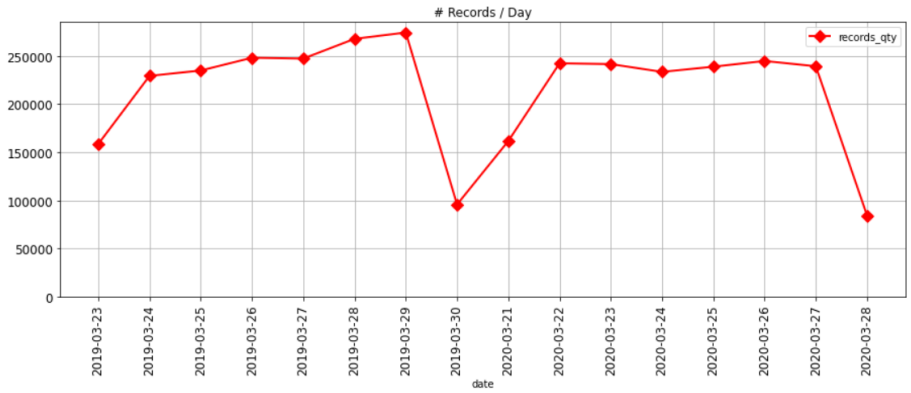
I want to congratulate all the people involved in the preparation of this test. It's fantastic, and I had a lot of fun doing it. Thank you very much for this opportunity. Please find below my answers in green color:

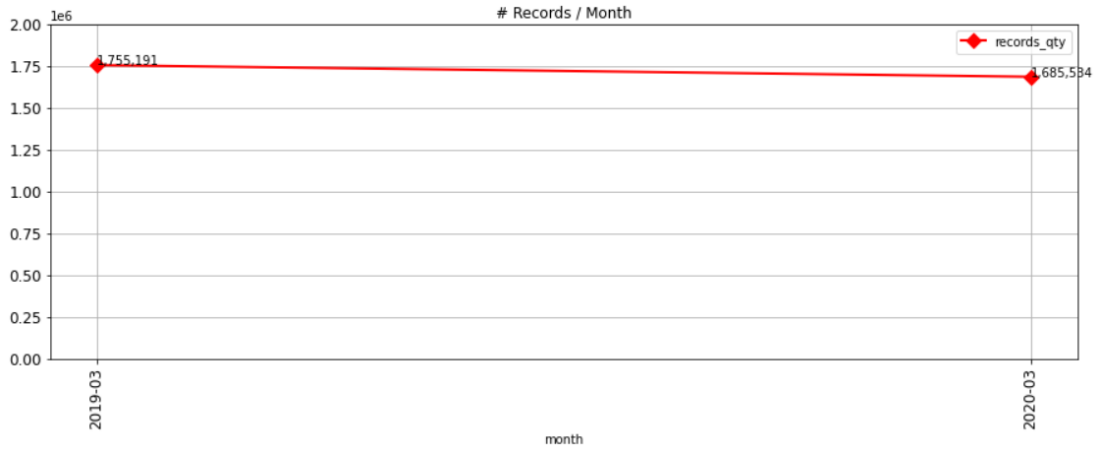
VERY IMPORTANT: After the EDA, it was possible to confirm that there are 33,152 duplicate records in the dataset. All duplicate records were removed, and all answers were made considering the deduplicated dataset.

See the evidence below. Two completely identical records:


```
{'epochMillis': {0: 1553574152000, 1: 1553574152000},
 'mmsi': {0: 412372140, 1: 412372140},
 'olson_timezone': {0: 'Asia/Shanghai', 1: 'Asia/Shanghai'},
 'imo': {0: 0, 1: 0},
 'callSign': {0: 'BPEB', 1: 'BPEB'},
 'destination': {0: 'SHANGHAI', 1: 'SHANGHAI'},
 'cargoDetails': {0: None, 1: None},
 'position_latitude': {0: 31.378744, 1: 31.378744},
 'position_longitude': {0: 121.572655, 1: 121.572655},
 'navigation_navCode': {0: 0, 1: 0},
 'navigation_navDesc': {0: 'Under Way Using Engine',
  1: 'Under Way Using Engine'},
 'navigation_courseOverGround': {0: 221.7, 1: 221.7},
 'navigation_heading': {0: 252.0, 1: 252.0},
 'navigation_rateOfTurn': {0: 0.0, 1: 0.0},
 'navigation_speedOverGround': {0: 0.7, 1: 0.7},
 'vesselDetails_name': {0: 'JIHAIZHONGSHAN', 1: 'JIHAIZHONGSHAN'},
 'vesselDetails_typeName': {0: 'Cargo', 1: 'Cargo'},
 'vesselDetails_typeCode': {0: 70, 1: 70},
 'vesselDetails_draught': {0: 5.0, 1: 5.0},
 'vesselDetails_length': {0: 100, 1: 100},
 'vesselDetails_width': {0: 16, 1: 16},
 'vesselDetails_flagCode': {0: 412, 1: 412},
 'vesselDetails_flagCountry': {0: 'China', 1: 'China'},
 'port_unlocode': {0: 'CNSHG', 1: 'CNSHG'},
 'port_name': {0: 'SHANGHAI PT', 1: 'SHANGHAI PT'},
 'port_latitude': {0: 31.334994, 1: 31.334994},
 'port_longitude': {0: 121.659069, 1: 121.659069}}
```


1. We need to see your own code.
    a. All the source code and readme, with instructions, how to build and run, are available in: https://github.com/maxreis86/AgileEngine
2. The data can be downloaded from the following URLs:
    a. Dataset in JSON format:
        i. https://datascience-public.transvoyant.com/public/data/test_tasks/ocean_ais/json/json.zip
    b. Dataset in Parquet format:
        i. https://datascience-public.transvoyant.com/public/data/test_tasks/ocean_ais/parquet/parquet.zip
    c. Both JSON and Parquet datasets are identical in contents, but you must choose to use one over the other. Please provide your justification for your choice of dataset.
        i. I chose to use parquet file because it is faster to read to Spark Dataframe considering that I could use the same script in a much bigger dataset. Parquet took just 113 ms while JSON would take 15.2 s wall time.
3. What is(are) the main time period(s) in the data?
    a. Answer: The data presents a time period of 16 days; however 8 days are from last week of march of 2019 and 8 days from last week of march of 2020. It is possible to identify in the graph that on Saturdays the activities of the vessels are much lower than the other days of the week. For some analyses, it would be a good idea to consider just two different periods, March 2019, and March 2020:





4. Which are the top three most sparse variables?
    a. The top three most sparse variables are navigation_rateOfTurn, imo, navigation_speedOverGround. I used the highest coefficient of variation to determine the most sparsas variables.

Column	coefficient_of_variation
navigation_rateOfTurn	3428.79%
imo	526.10%
navigation_speedOverGround	144.82%

5. What region(s) of the world and ocean port(s) does this data represent? Provide evidence to justify your answer.
    a. There are many evidences that this data represent The Port of Shanghai, located on the outskirts of the Chinese city of Shanghai, That can be accessed by East China Sea and Hangzhou Bay as well as the Yangtze and Huangpu Rivers. I chose three evidences in the dataset:
        i. The only port name in the dataset is Shanghai Port:

```
+-----------+-------+
|  port_name|  count|
+-----------+-------+
|SHANGHAI PT|3440725|
+-----------+-------+
```

        ii. The only time zone in the dataset is Asia/Shanghai:

```
+---------------+-------+
|olson_timezone|  count|
+---------------+-------+
| Asia/Shanghai|3440725|
+---------------+-------+
```

        iii. The coordinates of the events are between 30º to 32º of latitude, 120º and 123º of longitude, and these coordinates represent the region around the port of shanghai

```
+---------------------+---------------------+
|min_position_latitude|max_position_latitude|
+---------------------+---------------------+
|             30.02688|             32.65971|
+---------------------+---------------------+

+----------------------+----------------------+
|min_position_longitude|max_position_longitude|
+----------------------+----------------------+
|            120.756256|            123.55666|
+----------------------+----------------------+
```

Port of Shanghai
上海港

Yangshan Deep-water Port

Location	
Country	People's Republic of China
Location	Shanghai
Coordinates	30.626539°N 122.064958°E

6. Provide a frequency tabulation of the various Navigation Codes & Descriptions (i.e., navCode & NavDesc). Optionally, provide any additional statistics you find interesting.

	navigation_navCode	navigation_navDesc	quantity	avg_speedOverGround	stddev_rateOfTurn	cv_rateOfTurn
0	16	Unknown	1,340,821.00	2.66	2.12	-56.97
1	0	Under Way Using Engine	1,057,447.00	5.96	7.02	-20.46
2	5	Moored	548,325.00	0.31	6.71	-225.00
3	1	At Anchor	422,862.00	0.32	4.14	-154.72
4	15	Not Defined	29,179.00	1.09	15.88	-7.45
5	8	Underway Sailing	24,782.00	2.75	14.93	-8.97
6	3	Restricted Manoeuvrability	8,237.00	0.97	1.86	-67.69
7	2	Not Under Command	3,455.00	1.40	9.57	-13.34
8	4	Constrained By Her Draught	1,450.00	1.28	0.44	642.47
9	9	Reserved For Future Amendment	1,353.00	1.54	22.83	-5.13
10	11	Reserved For Future Use	1,220.00	0.88	0.00	NaN
11	13	Reserved For Future Use	875.00	0.06	0.00	NaN
12	6	Aground	401.00	1.10	1.48	-56.58
13	7	Engaged In Fishing	311.00	5.31	0.00	NaN
14	12	Reserved For Future Use	7.00	9.19	0.00	NaN

7. For MMSI = 205792000, provide the following report:
    a. Limit the data to only the TOP 5 Navigation Codes based from the response to question 6
        i. The limited date for MMSI = 205792000 has 184 records considering only the TOP 5 Navigation Codes 16, 0, 5, 1, 15
    b. Provide the final state for each series of contiguous events with the same Navigation Code; series may be interrupted by other series, but each contiguous series must be its own record.
    c. Final report should include at least the following fields/columns:
        i. mmsi = the MMSI of the vessel
        ii. timestamp = the timestamp of the last event in that contiguous series
        iii. Navigation Code = the navigation code (i.e., navigation.navCode)
        iv. Navigation Description = the navigation code description (i.e., navigation.navDesc)
        v. lead time (in Milliseconds) = the time difference in milliseconds between the last and first timestamp of that particular series of the same contiguous navigation codes
            1. It was necessary to create the navigation_navDescID column to identify each contiguous series. Without this approach it would not be possible to separate the first Morred from the Morred as shown in the example below.

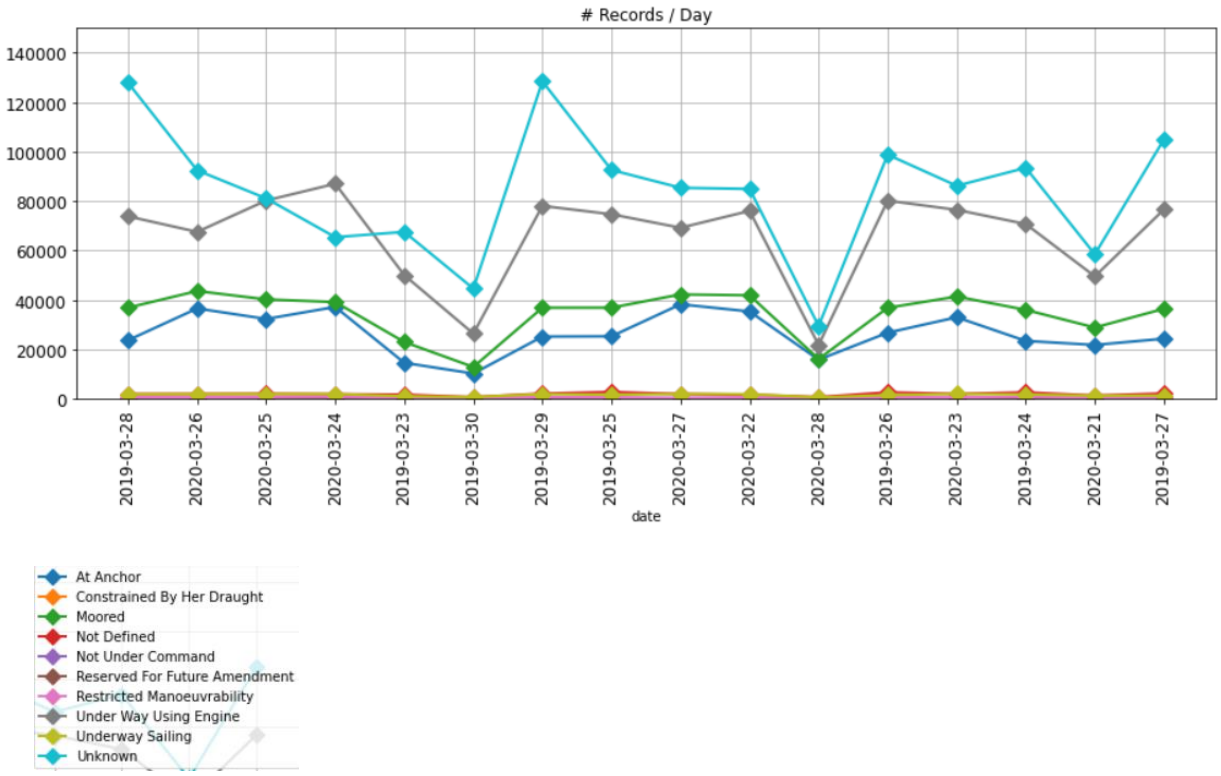
	mmsi	first_datetime	last_datetime	navigation_navCode	navigation_navDescID	lead_time_milliseconds	lead_time_minutes
0	205792000	2020-03-23 15:23:38	2020-03-23 15:58:32	5	1 - Moored	2094000	34
1	205792000	2020-03-23 16:19:31	2020-03-24 05:49:31	1	2 - At Anchor	48600000	810
2	205792000	2020-03-24 06:09:20	2020-03-24 13:57:21	0	3 - Under Way Using Engine	28081000	468
3	205792000	2020-03-24 14:08:20	2020-03-25 11:32:21	5	4 - Moored	77041000	1284
4	205792000	2020-03-25 12:01:26	2020-03-25 17:51:28	0	5 - Under Way Using Engine	21002000	350

8. For MMSI = 413970021, provide the same report as number 7
    a. Do you agree with the Navigation Code(s) and Description(s) for this particular vessel?
        i. If you do agree, provide an explanation why you agree.
        ii. If you do not agree, provide an explanation why do disagree. Additionally, if you do not agree, what would you change it to and why?
            1. Unfortunately the Navigation Code for this particular vessel is defined as 16 (Unknown). Because of this is not possible to know each series of contiguous events and judge if I agree or not with the navigation. What I would do differently would be to improve the system to correctly capture each ship state change.

	mmsi	first_datetime	last_datetime	navigation_navCode	navigation_navDescID	lead_time_milliseconds	lead_time_minutes
0	413970021	2019-03-23 07:49:54	2019-03-30 04:16:09	16	1 - Unknown	591975000	9866

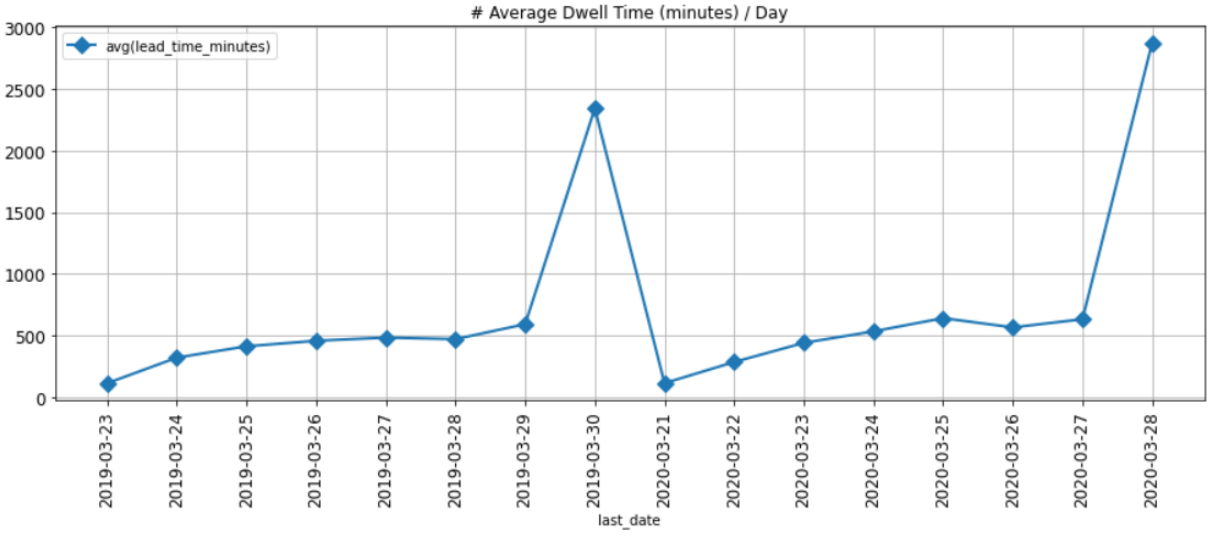
9. For each of the time period(s) from item three, provide a tabulation of the top 10 series of vessel navigation code/description ordered states.

	date	At Anchor	Constrained By Her Draught	Moored	Not Defined	Not Under Command	Reserved For Future Amendment	Restricted Manoeuvrability	Under Way Using Engine	Underway Sailing	Unknown
0	2019-03-28	23660	249	36835	1899	276	220	623	73771	1841	127798
1	2020-03-26	36498	0	43555	1926	290	0	506	67500	1912	92407
2	2020-03-25	32261	0	40168	1994	416	0	668	80182	1957	81036
3	2020-03-24	37126	0	39101	1829	257	0	788	86964	1938	65355
4	2019-03-23	14511	46	22931	1638	175	96	390	49972	1012	67494
5	2019-03-30	10169	138	12824	579	56	45	227	26335	681	44558
6	2019-03-29	25133	341	36803	2083	248	188	681	78005	1882	128380
7	2019-03-25	25228	47	36822	2691	131	196	573	74621	1599	92541
8	2020-03-27	38189	0	42196	1905	213	0	195	69160	1956	85338
9	2020-03-22	35286	0	41824	1587	224	0	640	75973	1820	84872
10	2020-03-28	15770	0	16063	615	94	0	93	21728	545	29426
11	2019-03-26	26698	239	36773	2560	261	282	610	80135	1439	98733
12	2020-03-23	32937	0	41329	1942	209	0	479	76409	1976	86148
13	2019-03-24	23372	123	35936	2561	150	132	803	70668	1711	93418
14	2020-03-21	21746	0	28790	1260	155	0	402	49496	1254	58421
15	2019-03-27	24278	267	36375	2110	300	194	559	76528	1259	104896



10. Using the results from item 9, compare the volume of each vessel navigation code/description ordered states for each time period(s) from item three.
    a. Which increased/decreased?
        i. Answer: Analyzing the graph with daily information is not the best way to confirm which one has increased or decreased. So I decided to create a monthly chart . Analyzing the monthly chart, it is possible to confirm that “Unknown” had a considerable decrease, and “Moored” and "At Anchor" increased.

11. For each of the time period(s) from item three and using only the “At Anchor” and “Moored” navigation descriptions, quantify the average “dwell”
    a. The Answer: The average number of dwell type events received per day is 60699 and on average the vessels are stopped for 704 minutes in the dwell states. However on Saturdays the average is 4 times bigger (2605 minutes), indicating that there is not much movement at the port of shanghai that day.



12. Describe or show how you would quantify if the difference(s) in “dwells” between the time-period(s) is(are) significant.
    a. Describe or show how you would quantify if the difference(s) in “dwells” between the time-period(s) is(are) significant
        i. Answer: I did a Two sample t-test to compare the average minutes in “dwells” between "2019-03" and "2020-03". The p value obtained from the t-test is significant ($p < 0.05$), and

therefore, we conclude that the ‘lead_time_minutes’ of month "2019-03" is significantly different than month "2020-03".

Two sample t-test with unequal variance (Welch's t-test)

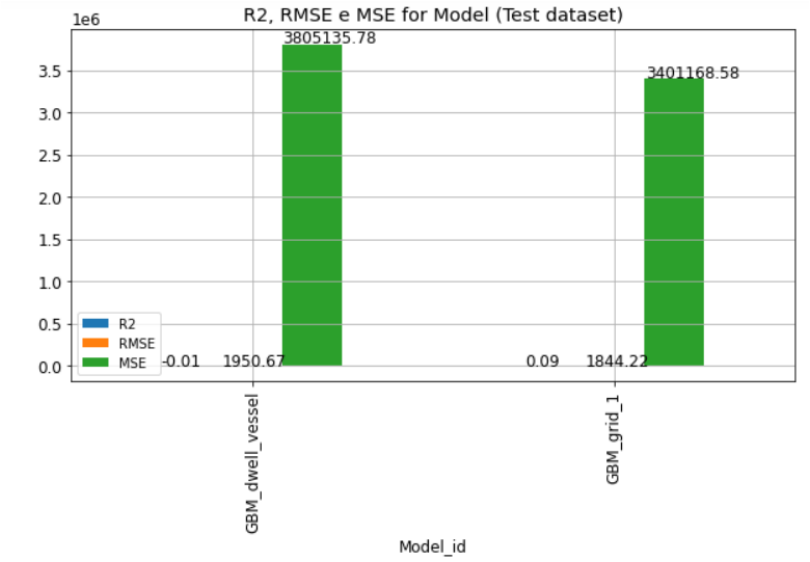
Mean diff	-114.809
t	-6.01909
Std Error	19.0741
df	30270.8
P-value (one-tail)	8.87081e-10
P-value (two-tail)	1.77416e-09
Lower 95.0%	-152.195
Upper 95.0%	-77.4228

Parameter estimates

Level	Number	Mean	Std Dev	Std Error	Lower 95.0%	Upper 95.0%
2019-03	14418	623.082	1562.88	13.0159	597.569	648.595
2020-03	15868	737.891	1756.39	13.9431	710.561	765.221

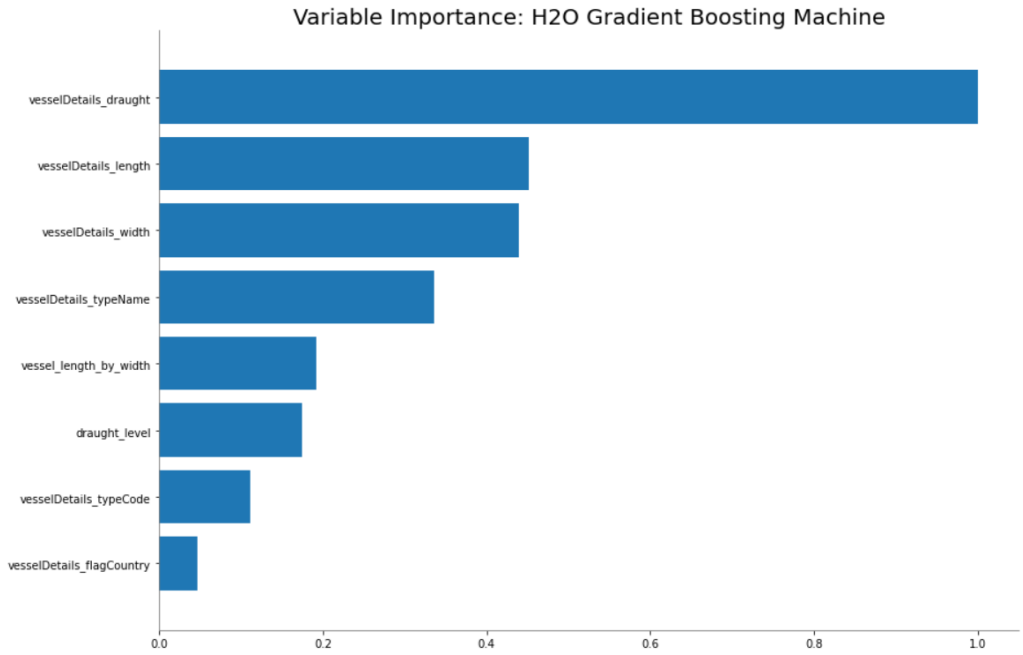
13. Describe or show how you would create a Machine Learning Model to predict “dwell” times for the region.
    a. Bonus Points: Provide the code and performance results for your machine learning model on an OOB sample.
    b. In respect of time, performance of the ML model is not important and the number of features can be minimal. What mattes is that the code works and you can explain your thought process if asked.

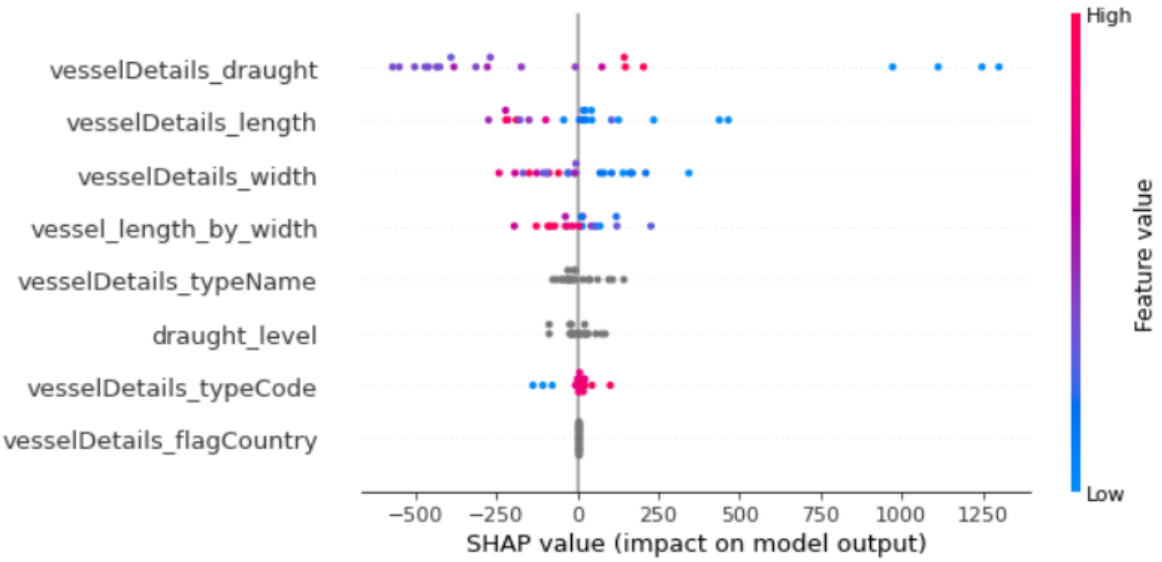
I would suggest validating the model's performance result using an Out Of Time sample dataset instead of using an Out of bag sample, as this way We can verify that we can use the model for future datasets. So I decided to train a machine learning model using 2019 data and validate the model using 2020 data. My first strategy was to try to predict the time in minutes (lead_time_minutes) spent by each vessel in in dwell, using vessel characteristics like width and length. The final model had an R2 of 0.9 and an RMSE of 12 on the out-of-time dataset. These are the trained models:



The main variables with the greatest impact on the model are ‘vesselDetails_draught’ and ‘vesselDetails_length’ with positive impact, this means that the larger the length of the vessel or the draught needed, the longer it will be stationary in the port.
Please, find bellow all independent variables of the final model and the impact of each one using Shapley Value:

```
BEST MODEL: GBM_grid_1_AutoML_1_20220403_231645_model_6
```





14. We value code readability and consistency, and usage of modern community best practices and architectural approaches, as well, as functionality correctness. So pay attention to code quality.
15. Target completion time is about 4 hours. We would rather see what you were able to do in 4 hours than a full-blown algorithm you’ve spent days implementing. Note that in addition to quality, time used is also factored into scoring the task.