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:

Texto

Descrição gerada automaticamente

1. We need to see your own code.
   1. 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:
   1. Dataset in JSON format:
      1. <https://datascience-public.transvoyant.com/public/data/test_tasks/ocean_ais/json/json.zip>
   2. Dataset in Parquet format:
      1. <https://datascience-public.transvoyant.com/public/data/test_tasks/ocean_ais/parquet/parquet.zip>
   3. 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.
      1. 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?
   1. 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:

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

1. Which are the top three most sparse variables?
   1. 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.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

1. What region(s) of the world and ocean port(s) does this data represent? Provide evidence to justify your answer.
   1. 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:
      1. The only port name in the dataset is Shanghai Port:

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

* + 1. The only time zone in the dataset is Asia/Shanghai:

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

* + 1. 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

Uma imagem contendo Carta

Descrição gerada automaticamente

Uma imagem contendo Texto

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo, Site

Descrição gerada automaticamente

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

Tabela

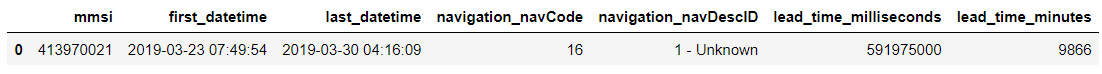
Descrição gerada automaticamente

1. For MMSI = 205792000, provide the following report:
   1. Limit the data to only the TOP 5 Navigation Codes based from the response to question 6
      1. The limited date for MMSI = 205792000 has 184 records considering only the TOP 5 Navigation Codes 16, 0, 5, 1, 15
   2. 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.
   3. Final report should include at least the following fields/columns:
      1. mmsi = the MMSI of the vessel
      2. timestamp = the timestamp of the last event in that contiguous series
      3. Navigation Code = the navigation code (i.e., navigation.navCode)
      4. Navigation Description = the navigation code description (i.e., navigation.navDesc)
      5. 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.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

1. For MMSI = 413970021, provide the same report as number 7
   1. Do you agree with the Navigation Code(s) and Description(s) for this particular vessel?
      1. If you do agree, provide an explanation why you agree.
      2. 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.



1. 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.

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Uma imagem contendo Diagrama

Descrição gerada automaticamente

1. 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.
   1. Which increased/decreased?
      1. 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.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente com confiança média

1. For each of the time period(s) from item three and using only the “At Anchor” and “Moored” navigation descriptions, quantify the average “dwell”
   1. 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.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

1. Describe or show how you would quantify if the difference(s) in “dwells” between the time-period(s) is(are) significant.
   1. Describe or show how you would quantify if the difference(s) in “dwells” between the time-period(s) is(are) significant
      1. 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".

Texto

Descrição gerada automaticamente

1. Describe or show how you would create a Machine Learning Model to predict “dwell” times for the region.
   1. Bonus Points: Provide the code and performance results for your machine learning model on an OOB sample.
   2. 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:

Gráfico, Gráfico de barras, Gráfico de cascata

Descrição gerada automaticamente

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:

1. 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.
2. 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.