SIADS 696 Milestone II Project Report
Container Vessel Port Dwell Time Prediction

Cliff Gong (clgong@umich.edu), Gen Ho (genho@umich.edu) , Xinqian Zhai (xinqianz@umich.edu)

Introduction

The global supply chain is a gigantic network connecting manufacturers, shippers, consignees, ocean carriers, rails, trucks, and warehouses for the circulation of cargo and containers. This ecosystem was in a rather balanced state. But when the pandemic caused the shutdown of production and ports in Asia, resulting in a huge backlog of cargo, the entire system of the global supply chain was thrown out of balance. Supply chains and port congestion impacted the whole world and the US was in the epicenter. The number of container vessels off Port of Los Angeles and Long Beach (LA/LB) hit an all-time high of 109 on Jan 9, 2022, with dwell time to berth (i.e. how long it takes for a ship to berth after it arrives at a port) up to 28 days. Under the port congestion situation, the Estimated Time of Arrival (ETA) provided by ocean carriers is not able to reflect the extra waiting time for vessels to be berthed.

Our project aims to develop a machine-learning model which predicts the container vessels' dwell time to berth. If ocean carriers have a better estimation of the dwell time, they can update consignees and logistic companies in advance, such that the downstream logistic parties can have better planning and execution.

Data Source, Scope, and Preprocessing

There were two major data sources for this project. The first one was Automatic Identification System (AIS) data which can be downloaded from MarineCadastre.gov in CSV format. The second one was Vessel Particular which needed to be web scraped from balticshipping.com.

AIS Data

AIS is an automatic tracking system on ships that allows crews onboard to view marine traffic in their area and to be seen by that traffic. This data can be collected in near real-time. AIS data contains information like vessel identifier, location (latitude and longitude), time, speed, direction, etc which can be used to trace a vessel's trajectory. Detailed schema is available in **Appendix B**.

AIS data is a stream of data with a huge volume. A day of AIS data from MarineCadastre.gov has around 7M records with over 700MB file size (unzip). Our project used full-year data of 2020 and 2021, and the first 6 months data of 2022. This translated to 6.4 billion records or 620 Gigabytes in size.

However, we did not use the AIS data directly as features for model building. Instead, we applied geofence over AIS data to detect the time when a vessel arrived at the port area and the time when a vessel berthed at a terminal. We then calculated the target label (dwell time to berth) and

Figure 1: How the target label (dwell time) is generated

other useful features based on these derived events. *Figure 1* illustrates this concept. In our project, we focused on the 6 container terminals in the Port of Long Beach (highlighted by red boxes in the diagram). For the port area, it is defined as 150 miles from the center of the port (the concept is expressed as a yellow dashed line circle in the diagram, not in the actual size).

Vessel Particular Data

Vessel Particular refers to vessel basic information. This includes vessel identifier, vessel name, width, length, deadweight tonnage (dwt), vessel type, owner, operator, etc. There is no direct download available for vessel particular data. We had to write our own web scraping to collect the required information in CSV format based on the vessel identifier available in the AIS data. Detailed schema is available at **Appendix B**.

Feature Engineering

The feature engineering pipeline included the following major steps

1. **Filter AIS Data** - apply geofence to filter raw AIS data. We only required data in the port area. Data size reduced from 620GB to 6GB
2. **Resample AIS data** - resample AIS data every 30 min interval. Data size reduced from 6GB to 480MB
3. **Detect port entering time** - apply geofence to detect the port entering time for vessels
4. **Detect vessel berth time** - apply geofence to detect the berth time for vessels
5. **Calculate dwell time to berth** - pair the port entering time and berth time for a vessel in a time interval in order to calculate the dwell time
6. **Generate new features** - calculate new features, e.g. average dwell time for target terminal, number of vessels in the port area, etc. Consolidate AIS data to become 1 record per vessel at the moment when it has entered the port area. Data size reduced from 480MB to 140KB
7. **Merge AIS data with vessel particulars** - join vessel particular data with vessel identifier. Data size increased from 140KB to 200KB

The final output dataset is distilled to 1,653 records. Links to different notebooks are provided in **Appendix C**. Below table list the data columns we have used.

Column Name	Description	Data Type	Data Source
avg_dwell_at_target_terminal	Average dwell time to berth (over the past 14 days) for the target terminal a vessel is calling. The used records are those with timestamps before a vessel has entered the port area	numerical	Engineered from AIS
num_of_vessel_at_target_terminal	Number of vessels at the target terminal when a vessel has entered the port area	numerical	Engineered from AIS
num_of_vessel_in_port	Number of vessels in the port area when a vessel has entered the port area	numerical	Engineered from AIS
target_terminal	The target terminal a vessel is calling. For our dataset, this value is engineered from AIS data. In practice, ocean carriers can provide value for this field.	categorical	Engineered from AIS

vessel_operator	The ocean carrier who operates the vessel	categorical	Vessel Particular
vessel_width	The width of vessel in meter	numerical	Vessel Particular
vessel_length	The length of vessel in meter	numerical	Vessel Particular
vessel_dwt	The deadweight tonnage of a vessel. This Is a proxy to a vessel cargo capacity	numerical	Vessel Particular
weekday	The day of week when a vessel has entered the port area	ordinal	Engineered from AIS
hour_of_day	The hour of day when a vessel has entered the port area	ordinal	Engineered from AIS
is_holiday	Whether this is a US public holiday when a vessel has entered the port area	categorical	Engineered from AIS
dwell_in_hr	The dwell time to berth for a vessel. This is the duration between the time when a vessel enters the port area and has berthed. This field is the **target label**	numerical	Engineered from AIS

Unsupervised Learning

Motivation

Given the mixed type of high dimensional data, we are going to transform the data and apply dimension reduction in order to visualize the dataset for insight discovery, which includes:

1. identifying important features;
2. discovery groups of dwell time patterns;
3. other insights that are hard to see.

The results can potentially assist in the supervised learning phase.

Data Source

Please refer to the **Data Source, Scope, and Processing** section above for the details on data sources and features we've used. Also, a detailed data schema is provided in **Appendix B**.

Unsupervised Learning Methods

Since our data set has both numerical and categorical variables, traditional dimension reduction techniques and majority clustering algorithms cannot be directly applied. To solve this challenge, we tried several methods to encode them and apply dimension reduction and clustering algorithms over the transformed data to identify interesting patterns and insights.

Our first attempt was to use FAMD (factor analysis of mixed data) from the Prince Python package. The advantage of the FAMD is that it encodes categorical variables in a way that they will have similar weights

to numerical variables over the calculated principle components, making the mixed type of variables comparable. In general, it first performs standard scaling on numerical variables. For categorical variables, after one-hot encoding, FAMD divides each column by the square root of its probability sqrt and then centers the columns. After that, FAMD applies the PCA algorithm over the obtained data frame.

However, the resulting plot did not reflect a dense separation of data and only 12.5% of the total inertia/variance was explained by the first 2 components. We suspected the underlying structure of the data was highly non-linear, that's why FAMD did not work well for it is a linear dimension reduction algorithm. Since one of our objectives is to reduce the high dimensional data into 2D feature space for insight discovery, we need a better 2D representation to allow us to "see" multiple dimensions in a single view.

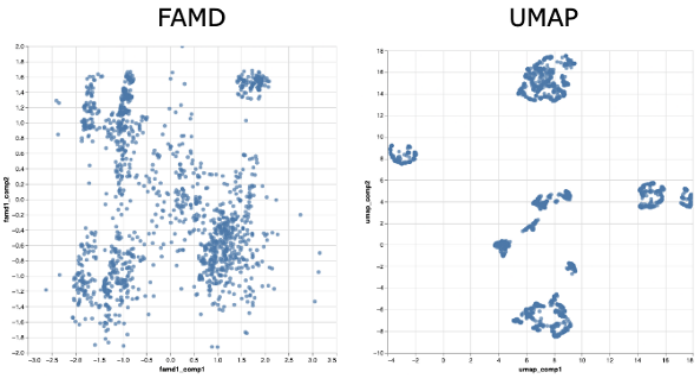


Figure 2: FAMD vs UMAP

To overcome the above challenge, our second attempt was to use UMAP, a non-linear dimension reduction technique. Since we could not directly apply the euclidean distance function over the mixed-type data, we handled them separately. For numerical data we fitted to UMAP with euclidean distance, while for categorical data we fitted to UMAP with dice distance. We then resembled the outputs together to obtain the final UMAP embedding. From the scatter plot *(Figure 2)* of the final UMAP embedding, we observed much denser data separation compared to the plots.

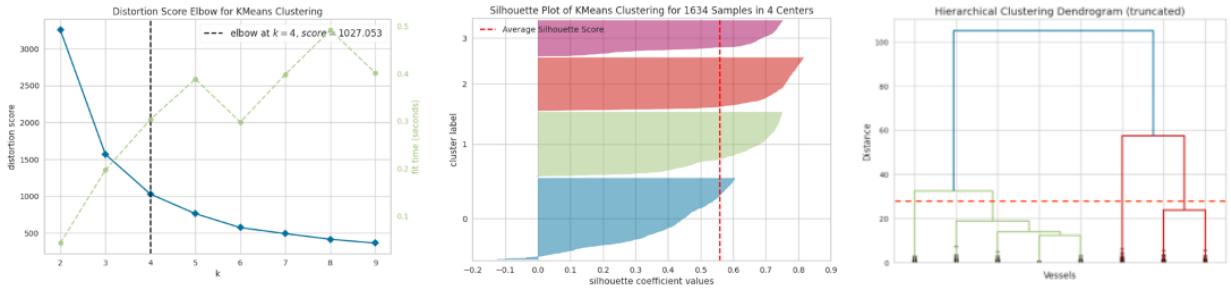


Figure 3: Elbow method, Silhouette score and Hierarchical Clustering

Our next step was to apply different cluster algorithms to discover and visualize potential clusters. First, we applied KMeans on the UMAP output. To make sure we selected the most optimal number of clusters, we first used the Elbow method to calculate the distance, then used the Silhouette score plot to evaluate other valuable information, such as variance, skewness, high-low differences, etc. As a result, both two methods showed 4 clusters should be optimal. After projecting the 4 clusters on UMAP embeddings, we got a silhouette score of 0.559. After that, we applied agglomerative clustering and a dendrogram to double-check the clusters and got similar clusters and silhouette scores (0.556) *(Figure 3)*. In order to find potential outliers, we also applied DBSCAN.

Overall, all the cluster algorithms did pretty well in identifying clusters. With further analysis of our clusterings and conversation with domain experts, we think the reasonable number of clusters should be

5 in our case. This is because these clusters were strongly related to the target terminals, although they all misclassified one of the target terminals (Pier G) because of its small number of records *(Figure 4)*. For the outliers from DBSCAN, from the dwell time perspective, we may not agree they are outliers.

Figure 4: KMeans, Agglomerative Clustering and DBSCAN

For tuning the UMAP parameters, firstly, to be able to sensibly combine the UMAP embedding representations for the mixed type data, we integrated them by intersection and union separately. Since union resembled the categorical embeddings more and the resulting UMAP embeddings were also much denser than the intersection, we unioned them together. After that, we found the optimal `n_neighbors` of the final UMAP by tuning a series of `n_neighbors` values and evaluated the resulting plot based on data density and task-based knowledge. We found `n_neighbors` equal to 250 was the best in our case.

For finding the optimal number of clusters, we used the Elbow plot, Silhouette plot, and hierarchical dendrogram for KMeans and Agglomerative clustering. They all came to the same number of clusters as mentioned above. For DBSCAN, we estimated the optimal epsilon value using the NearestNeighbors algorithm with a reference n_neighbors value (2*dim -1, where dim is the dimension of our dataset) from the original DBSCAN author's paper (Sander et al.1998)[1], and played around different combinations of the epsilons and the min_samples values based on domain knowledge and Silhouette score. Finally, we reached a reasonable pair of parameters for the DBSCAN.

Unsupervised Evaluation

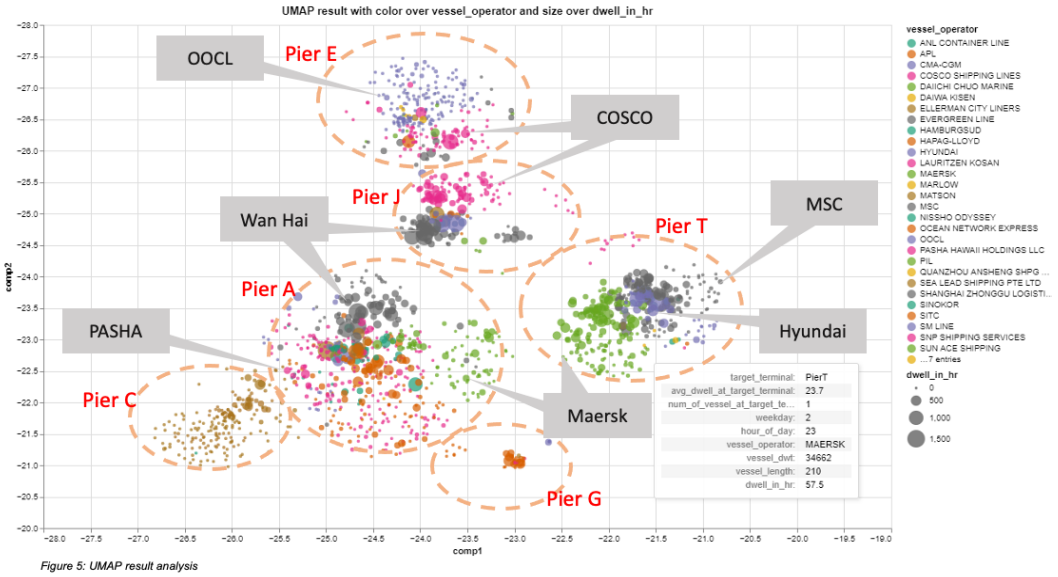


Figure 5: UMAP result analysis

Based on the goals of unsupervised learning, we identified several important features that facilitate supervised model training and discovered groups of dwell time patterns along with other interesting insights *(Figure 5)*.

1. We found that target_terminal and vessel_operator features did have impacts on dwell time. However, the vessel_dwt (vessel capacity) feature did not have an obvious relationship to dwell time.
2. For the target terminals, Pier E and Pier C had better performance (short dwell time) compared to other terminals. Pier G had a much lower volume than other terminals. Further checking on the web realized Pier G mainly handles bulk cargo. Containerized cargo only contributes a very small portion.
3. Vessel operators had their own engaged terminals. For example, OOCL mainly used Pier E. COSCO used Pier E and J. Maersk used Pier T and A. Wan Hai used Pier A and J.
4. Some vessel operators may have higher priority over others for vessel berthing at a particular terminal. For example, OOCL mainly used Pier E, and generally, the dwell time for OOCL was lower than COSCO, which hints that OOCL may have a privilege. Similar patterns can be observed for Pier T where Maersk and MSC may have a higher priority than Hyundai, and for Pier A where PASHA may have a higher priority than the others.
5. For the vessel deadweight, as an indicator of the carrying capacity of the ship, there was no obvious relationship with dwell time. In fact, whether it was a small-capacity vessel or a medium-to large-capacity vessel, its dwell time was not determined by its deadweight alone. This is interesting because the insight found here was not in line with our assumption that larger vessels were more likely to have longer dwell time. With the important features we identified, we know vessel operators and target terminals have more influence on the dwell time.

Although we've found several patterns and insights from the analysis, some challenges remain unresolved. In the future, the following attempts may be performed to achieve better results:

1. Estimating the optimal n_neighbors UMAP embedding is a brute force. We tested a range of values, but the step size of n_neighbors was relatively large. This is because it's time-consuming when doing the operation of fitting mixed-type data separately and then unioning them together to output the final 2D UMAP representation.To improve this in the future, we might try another way to handle mixed types of data, for example, using Gower distance to calculate the similarity distance and fit the matrix to UMAP only once to estimate UMAP parameters and perform it using cloud computing.
2. Tuning the parameters of DBSCAN to find clusters and potential outliers didn't work well *(Refer to Figure 4 - DBSCAN).* We estimated the optimal epsilon and min_samples values by trying different combinations based on our domain knowledge and some rules of thumb from the original DBSCAN authors' papers (Sander et al., 1998)[1] and (Schubert et al., 2017)[2]. However, we're not confident about the optimal number of the DBSCAN clusters. Even if we found clusters with reasonable parameters, there might be a wider range of clusters that we have not explored. All the clusters we found were closely terminal-related, lacking finer-grained separations related to vessel dwell time. This would make the interpretation of outliers based on dwell time less strong.

Supervised Learning

Motivation

Our project aims to develop a machine learning model to predict the dwell time to berth when a vessel has arrived at the port area. If carriers have a better estimation of the dwell time, they can update consignees and logistic companies in advance, such that the downstream logistic parties can have better planning and execution. We formulated this as a regression supervised learning problem. The *dwell time to berth* (numerical) is the target label. The prediction moment is when a vessel enters the port area. Our features contain mixed data types. For numerical features, they mainly describe the port condition (e.g. *number of vessels at port*, the *average historical dwell time*). For categorical features, they include the *target terminal a vessel is calling* and *the operator of the vessel*. We learned from the unsupervised learning analysis that the *target terminal* and the *vessel operator* contain signals for predicting the dwell time.

Data Source

Please refer to the **Data Source, Scope, and Processing** section above for the details on data sources and features we've used. Also, a detailed data schema is provided in **Appendix B**.

Methods and Evaluation

We've built the required machine learning model over 3 different types of algorithms, then we selected the most accurate one for detailed analysis. The 3 types of algorithms are:

1. Linear Regression - works as a baseline. We used the sklearn implementation
2. Random Forest Regressor - a tree-based ensemble algorithm. Also used the sklearn implementation
3. CatBoost Regressor - a gradient boosting decision trees algorithm. This is an open-source machine learning framework originally developed by Yandex

We selected Mean Absolute Error (MAE) for model performance measurement. MAE provides the same unit of measurement as the target label which is easier to interpret by users in our use case (i.e. it is very straightforward for users to understand, e.g. a prediction error is $\pm\ 5hrs$).

For model training, a train/test split of 80/20 was used. There were 2 categorial features where we one-hot encoded them for fitting to the Linear Regression model and the Random Forest model. CatBoost didn't require this for it has built-in support for categorical data. We used 10-fold cross validation over the training dataset for hyperparameters tuning. Finally, we trained the models with the best hyperparameters and collected below performance figures over 10-fold cross validation for comparison.

	Linear Regression	Random Forest Regressor	CatBoost Regressor
MAE with s.d.	93.342 ± 42.145	62.482 ± 142.581	54.326 ± 11.533

The CatBoost model has achieved the lowest MAE. When compared to the baseline Linear Regression model, the CatBoost model is around 41% lower in error.

Overall, for the CatBoost model, around 64% of errors are within plus or minus 1 day, while around 72% of errors are within plus or minus 2 days, and 82% are within plus or minus 3 days (Figure 6). For the downstream logistic service providers in preparing containers pickup, this range of error sounds acceptable.
We've selected the CatBoost model and conducted the following detailed analysis.

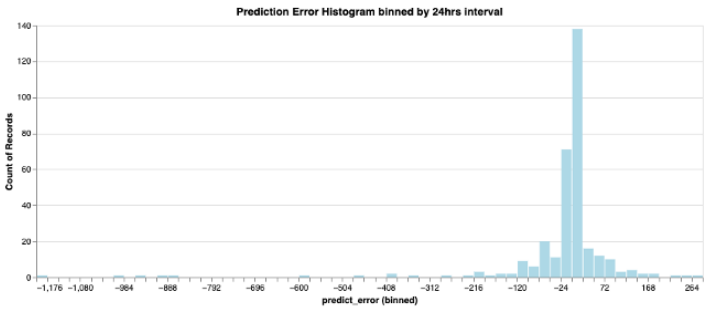


Figure 6: Prediction Error Histogram binned by 24hrs interval

Feature Analysis

We use SHAP to explain the machine learning model behavior. From the SHAP beeswarm plot *(Figure 7)*, We observed that the two numerical features *avg_dwell_at_target_terminal* and *num_of_vessel_in_port* are the most important. the higher of these values, the larger the SHAP values and hence the predicted output values. The 3rd and the 4th important features were the *vessel_operator* and *target_terminal* respectively where both features were categorical.

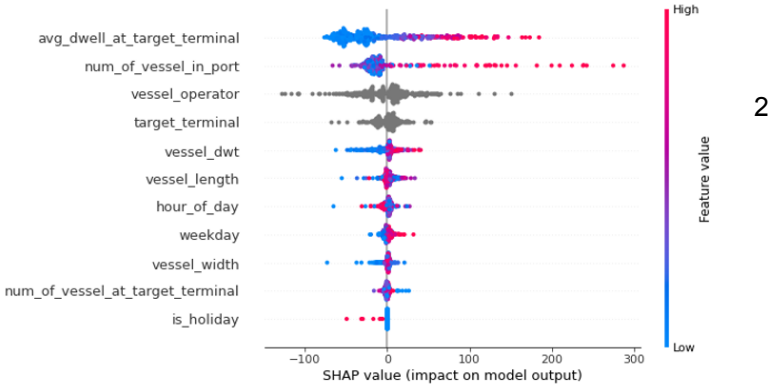


Figure 7: SHAP beeswarm plot for feature importance

We also conducted Feature Ablation Analysis in order to understand the importance of different groups of features.

Ablated Feature Group	Ablated Features	MAE (collected by 10-fold CV)
0) No Ablation	None	54.326 ± 11.533
1) Port Condition	avg_dwell_at_target_terminal, num_of_vessel_at_target_terminal, num_of_vessel_in_port	75.734 ± 14.261
2) Vessel Particulars	vessel_operator, vessel_width, vessel_length, vessel_dwt	64.030 ± 9.886
3) Time Related	weekday, hour_of_day, is_holiday	53.718 ± 11.695

Key findings:

1. Port Condition features were engineered from AIS data with geofencing. According to SHAP beeswarm plot, we knew *avg_dwell_at_target_terminal* and *num_of_vessel_in_port* were the top

2 most important features in our model. No wonder after removing this group of features, we observed around 39% increase in MAE when compared to the model trained with all features. Removing this group of features resulted in the worst performance model.

2. Vessel Particulars features are collected by web scraping. *vessel_operator* is the 3rd most important feature according to SHAP values. Without this group of features, the MAE increased by around 18% which was significant. Therefore, the effort for web scraping should not be eliminated.
3. Time related features were engineered from AIS data with geofencing. While their importance was low (with absolute mean SHAP value below 5 for each of these 3 features), this was surprising to see after removing this group of features, the model performance was slightly improved, with MAE reduced by around 1%. This means to our model, this set of features was noisy. We could save our effort in deriving this feature set.

Learning Curve Analysis

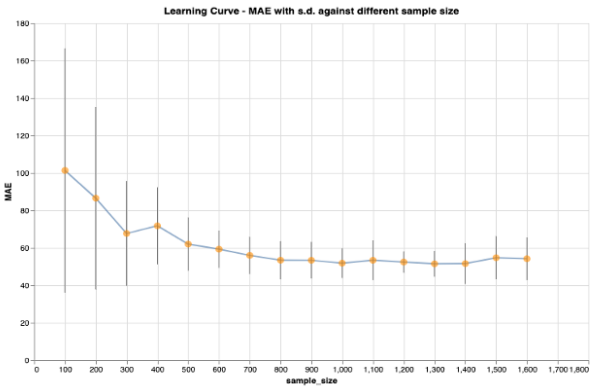


Figure 8: Learning curve for different sample size

As mentioned in the **Data Source, Scope, and Processing** section, it required a significant amount of effort in order to construct the dataset we needed from the raw AIS data. For 2.5 years of AIS data with over 6.4 billions of records, it only resulted in 1,653 records as what we required. Surely, we want to understand whether extra efforts are needed for collecting more training data. We've conducted a Learning Curve Analysis in order to answer this.

For the learning curve analysis, we started with 100 records, adding 100 records at each iteration until it has reached 1,600 records (16 iterations in total) *(Figure 8)*. Each MAE was collected with a 10-fold cross-validation, which meant 90% of the records in an iteration were used for the model training.

Key Findings:

1. The MAE and corresponding standard deviations dropped with the increasing of sample size, and it was stabilized between 800 to 1,400 samples. Notice the model had the lowest MAE at 1,200 sample size.
2. The MAE and corresponding standard deviations started going up slightly with 1,500 and 1,600 samples where the model may start overfitting with more data.
3. To conclude, the 2.5 years of AIS data we have used is enough for this CatBoost model. There was no need to spend extra effort in collecting more sample data.

Sensitivity Analysis

When we create a model we want it to generalize well. This means that with new test data, the model should still be accurate. One diagnostic of how a model will perform is how sensitive the hyperparameters are. If they are too sensitive then it is a sign that the model might not generalize well.

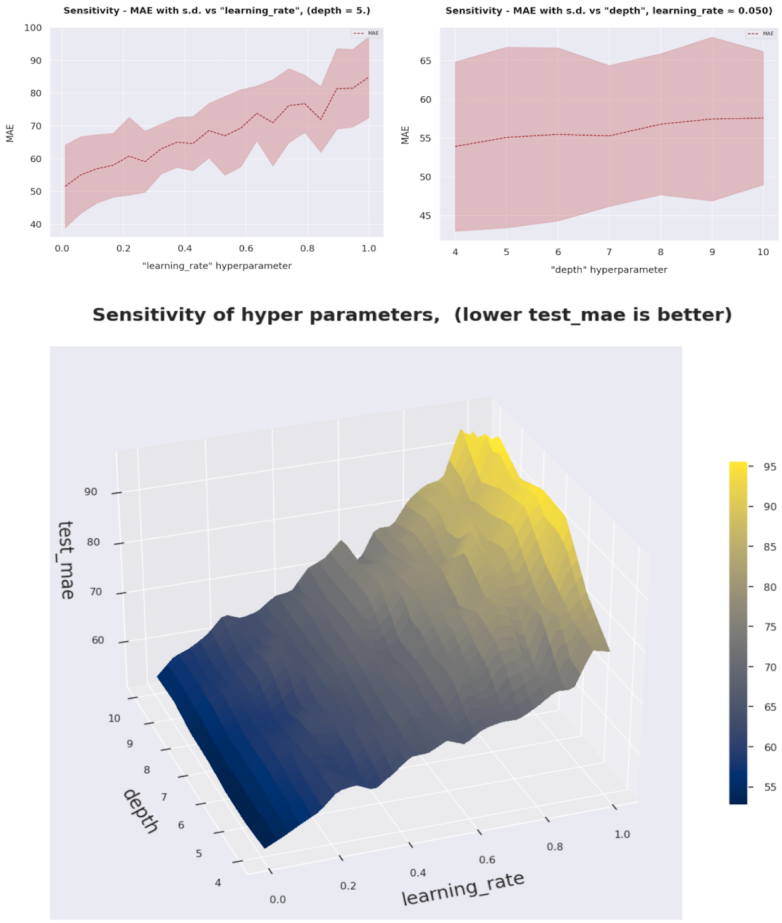


Figure 9. Hyperparameter sensitivity, individually and together

From our CatBoost model we found that by far the two most important hyperparameters were learning_rate and depth. We explored testing a range of values for each of these independently, holding all other hyperparameters constant at their best model values. Then we explored how the two interact together *(Figure 9)*.In all cases 10 fold cross validation was used.

Key findings:

1. **learning_rate** - Quite sensitive in our model. MAE and learning_rate are almost linearly associated with an approximate overall slope of 35 degrees. (The flatter the slope the less sensitive the hyperparameter, so 35 seems sensitive). However on the positive side: our model can achieve the lowest MAE or near the lowest MAE. It was trained with learning_rate=0.0523 and the resulting MAE was 54.326. That's near to the lowest MAE on the left side of the graph (learning_rate=0.010, cv_test_mae=51.547568). Also, the changes in slopes are not extremely jagged, at least they stay within nearby value's standard deviations. This might be a good sign in terms of not having to focus extra attention on any specific range of values when tuning the model. We do see that the slope is also smoother around our model's best value of 0.0523.
2. **depth** - Fairly stable in our model. The slope is somewhat flat, appearing to be roughly 5 or 10 degrees. Our model did not achieve the lowest MAE. (It was trained with depth 5 and resulted in an MAE of 54.326. The lowest MAE on the chart is 53.927). However, our model starts to become more stable when compared to depth at 4. In fact, for depth from 5 to 7 the MAE was more or less the same at around 55. Overall, we could see our model is not very sensitive to the depth parameter and it was relatively stable with the depth at 5 which we chose.
3. **learning_rate** and **depth** - Fairly stable. Looking at the 3D plot with both learning_rate and depth vs the MAE, at depth of 5 and learning_rate of 0.05 which we used to train our model, we noticed the model could achieve almost the lowest MAE, and the surface nearby was smooth. The interaction of learning_rate and depth at this range had brought stability to the model and we could conclude our model was generalized quite well with this set of hyperparameters over our training data.

Error Analysis

This is important to understand what failures our model has made in order to gain insights on how we can improve our solution. A detailed Error Analysis was done for this purpose.

We firstly created a visualization of errors for all test data *(Figure 10)*. This was eye-catching for the significant underestimation of dwell time in the period of Oct to Dec 2021. In fact, this was an abnormal period where vessel dwell time was historically high due to overwhelmed port congestion. We zoomed into this period and studied the SHAP force plots for the top 3 most underestimated records, and compared them with another record which ranked in the middle for error in this period.

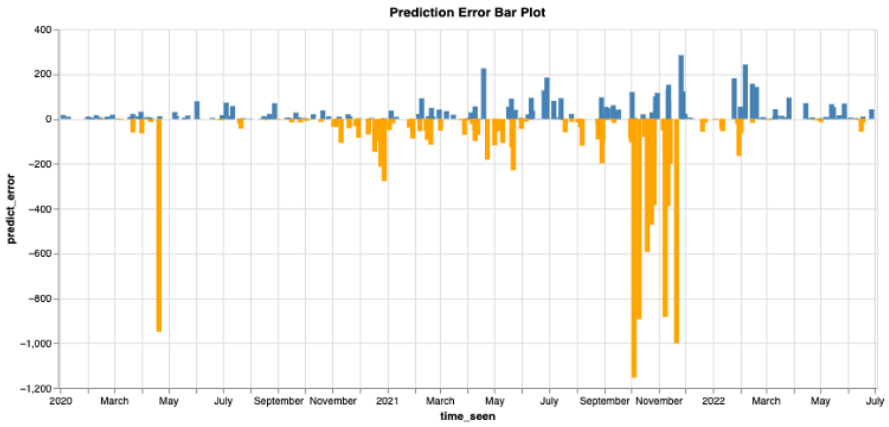


Figure 10: Prediction Error

Figure 11 showed 3 examples for illustration purposes. The first two were from the top 3 underestimated records. The third was the one with a middle level of error. For details, please refer to our Error Analysis notebook. The link is provided in **Appendix C**.

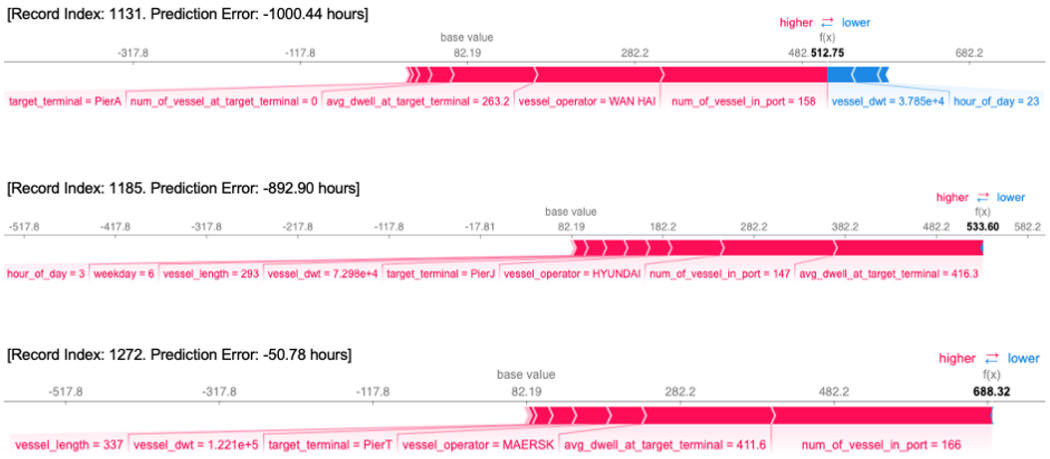


Figure 11: SHAP force plots for 3 examples

Comparing the force plots and the values of the 2 most important features *avg_dwell_at_target_terminal* and *num_of_vessel_in_port*, we suspected these features were not able to catch up with the extended long dwell time during this abnormal period. We studied the data distribution for these 2 features in this period and confirmed there were problems.

For the bar plot in *Figure 12*, we noticed the *num_of_vessel_in_port* was capped at around 180 despite there being a drastic increase in dwell time. This could be because the maximum capacity for vessels in the port area was reached. To our model, it meant when the port traffic was over a certain level this feature could not provide its predictability correctly.

On the other hand, we identified the *avg_dwell_at_target_terminal* feature was not able to catch up with the trend of the actual dwell time in this period. This may be due to the use of the past 14 days for calculating the average dwell time.

With the above analysis, here are our suggestions for future improvements

1. For the problem of *avg_dwell_at_target_terminal*, we can consider reducing the time window for average calculation from 14 days to 7 days or so. This helps the feature to catch up with the changes of the actual dwell time. The estimated effort for this is low.
2. For the problem of *num_of_vessel_in_port* where the maximum capacity of the port is reached during a highly congested situation, we need to consider searching for other potential vessel waiting areas outside the designated port areas. The idea is during an extreme situation, the overflowed vessels can either only wait somewhere outside the port, or they may slow stream from their origin to the port. This requires the study of AIS data during abnormal periods and geofence on some extra areas. This will also lead to significant changes in data pre-processing pipeline and model features. The estimated effort for this is high.

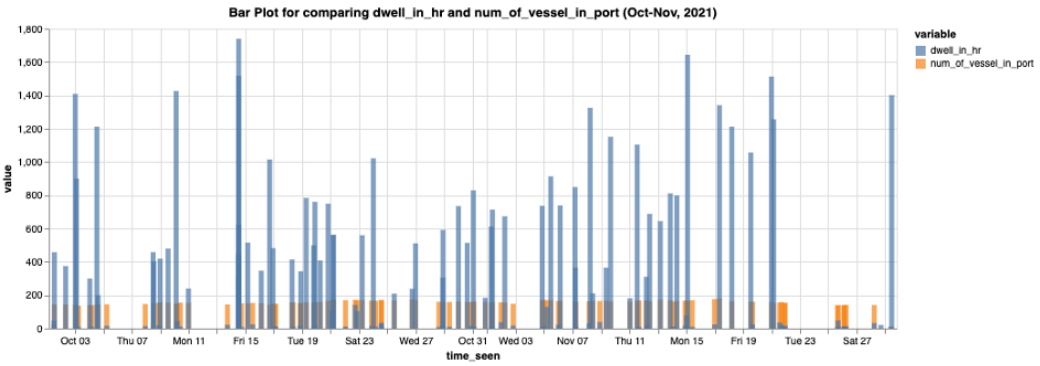


Figure 12: dwell_in_hr and num_of_vessel_in_port (Oct to Nov, 2021)

3. One extra idea for improvement is to search for new features. The global supply chain is a gigantic network. When containers are discharged from vessels, they will be moved out from the terminals by trains or trucks. If we can access data reflecting the volume of trains of trucks moving into and out of the terminals or ports, this could potentially compensate for the problem of the feature *num_of_vessel_in_port* and improve the model's overall predictability.

Discussion

Unsupervised Learning

The lesson learned is that parameter tuning is critical.

First, the different choices of distance metrics are important. In the unsupervised learning part, we found clusters using different methods, however, the output structure of clusters was mostly spherical, which is

rare in the "real world". This is because we used the default Euclidean distance to calculate the similarity between mixed data points. In our case, even though we preprocessed the numerical and categorical data separately, we still calculated the similarity based on the Euclidean distance and used the KMeans method to find the clusters. To extend our solution, we can use a custom distance function based on Gower distance and K-medoids clustering algorithm to minimize the sum of dissimilarities of data points, making the clustering more robust to noise and outliers.

Second, the different choices of clustering parameters are important. By running UMAP multiple times with a set of `n_neighbors`, we got different projections. We chose the "best" projection based on the density and our assumption about the total number of groups. Although we identified 5 groups closely related to the `targert_termnimal` feature and 3 groups with relatively short dwell times, the cluster structure became unstable if we changed the parameters. To better validate the clusters, we can further investigate the cardinality and magnitudes of the clusters to gain a deeper understanding of potential anomalies and outliers in the clusters. Combining different combinations of hyperparameters, we may improve the stability of cluster results.

Supervised Learning

For this project, data processing and feature engineering was certainly one of the biggest challenges. AIS data mainly provides location and time information for vessels. In order to create the target labels and features we need, a series of data processing steps including geofence, data resampling, time range matching, etc, were required. However, the huge amount of data highly complicated the problem. For example, originally we scoped the project to use only 6 months of data and only focus on one terminal. This was talking about handling over 1 billion records or 50GB of data. But turns out it only resulted in around 70 records after completing the data processing pipeline. We had to extend our scope to cover a longer period and more terminals, and we finally had reached 2.5 years of data with 6 terminals in the port. This involved changes of application logics and reprocessing of data with many iterations of trial and error, which was highly time consuming and full of frustration. Overall, this data processing part consumed more than 60% of time in our project, and we believed this would be quite close to the real world data science projects.

Another lesson learned was the importance of error analysis. For supervised learning, originally we had a perception that if a model's performance is not good enough, we should go straight for a ‘stronger’ algorithm or add more features. But through the error analysis, we discovered one of the important features became stale when the port was over congested. This led us to think about how we could actually improve our model. We learnt we should not treat machine learning models as black boxes. Model explanation is not about to please users, rather, this really guides data scientists on how to improve the model.

With the finding from the error analysis, we know the port could reach its maximum capacity for vessels during a highly congested situation. If more time is allowed, we would search AIS data for other potential waiting areas for overflowed vessels outside the designated port areas. We believe this finding could help us improve our model over the problem of underestimation during the extreme congested period. On the

other hand, considering the supply chain is a network of processes. With the objective to benefit downstream logistic parties for better planning, apart from vessel dwell time to berth, we could extend the scope by adding predictions for subsequent steps like container discharge time, container dwell time in port, etc. The aggregated effects of all these dependent prediction models could unleash the real power to benefit the inbound supply chain.

Ethical Consideration

Finally we'll conclude our report with a data ethics impact assessment.

Direct impact

We examined our project (unsupervised and supervised) and found that this was less of an issue:

- Privacy. Even though all our data are from public data sources, we took extra precautions to remove any potentially Personally Identifiable Information (PII). For example we do not use the IMO number in our machine learning. More on IMO's later.
- Our project does not deal with any protected classes (race, religion, sex, age...)
- We took extra precaution and completed a 3rd party impact test: the Canadian Government's own Algorithmic Impact Assessment [3]. We scored a very low impact score of 18 (lower is better). See **Appendix D** for the full report.

Unprotected class impact

While vessel operators are not a protected class, out of an abundance of caution we want to disclose our potential impact on vessel operators who berth on small terminals, which we analyzed as part of the unsupervised learning part of our project. Vessels berthed on small terminals might face the potential ethical risk of being marginalized. These vessel operators may not have equal representation in the data analysis as vessels berthed on other big terminals. As we mentioned in the unsupervised section, most of the vessels berthed at the Pier G terminal were identified as outliers. Upon further inspection, it turned out that most vessels on Pier G have short dwell time. If these vessels have been ignored, their advantage of having short dwell time would be hidden when their potential customers read our report. This could unintentionally create an unfair and exclusive environment compared to other vessels berthed on big terminals.

General potential machine learning issues in the ship tracking industry as a whole

The AIS for vessels was originally developed for safety purposes. This includes collision avoidance for water transport and allows maritime authorities to track and monitor vessel movements. The use of AIS data for prediction purposes, however, may create a privacy problem called "Secondary Use". Secondary use refers to the use of data obtained for one purpose for a different unrelated purpose without the one's consent. AIS data is broadcast information and publicly available, while it seems there is no harm for tracking purposes, the aggregated information could cause problems. E.g. a vessel operator can keep

track of all competitors' vessels and predict their estimated arrival time (ETA), such that they can take advantage by jumping queues to shorten their vessels's dwell time to berth. This certainly results in harmful effects for other vessel operators.

Another privacy related issue is personal privacy. Earlier we mentioned that we don't use the IMO because it can be linked to PII, including financial information. We demonstrate how we were actually able to do this in just 4 steps and in only a few minutes. See **Appendix E**. We took an IMO, linked it to a person's name, who was most likely the small business owner, and from there we connected to Gale Research online (courtesy of University of Michigan Library) and found out how much money that LLC made in sales! We know this can't possibly be the original intention when the AIS system was created in the 1990's [4]. This was before the civilian internet! It probably wasn't a big issue back then. As pointed out by Zook et.al. [5] in today's era, the scale, complexity and easy access of big data creates a rich ecosystem for benefit and harm.

This is a great example of unintended privacy issues. While we didn't use the IMO, other's might. And that could lead to a system that publicly exposes an individual person's actual and predicted dwell time performance.

We believe one step to avoid this, is that dwell time prediction systems should not publish IMO's with their machine learning outputs. And if there is a need, the IMO should be anonymized. And if it can't be anonymized, then there needs to be a way for anyone harmed (for example with mistaken predictions) to have recourse. We learned in SIADS503 that there is always a power differential. You just have to look for it [6]. It's conceivable that an underprivileged protected class could be inadvertently harmed, so the recourse mechanism must be affordable and accessible to all.

The industry came together once to create AIS for public safety. Surely it can come together again to create a data policy for the public good!

Statement of Work

Cliff Gong	Gen Ho	Xinqian Zhai
System set up. Web scraping, Linear Regression and Random Forest Regression Training, Sensitivity Analysis, Ethical issues research, analysis, Report Writing	Data Preprocessing, CatBoost Model Training, Feature Analysis, Learning Curve Analysis, Error Analysis, Ethical analysis, Report Writing, Business domain expert	FAMD analysis, UMAP dimensionality reduction, Clustering Analysis, Gower distance with Kmedoids clustering exploration, Ethical analysis, Report Writing

Appendix A - Bibliography

[1] Sander, J., Ester, M., Kriegel, HP. et al. "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". Data Mining and Knowledge Discovery 2, 169–194 (1998). https://link.springer.com/article/10.1023/A:1009745219419. Accessed October 20,2022

[2] Schubert, Erich et al. “DBSCAN Revisited, Revisited.” ACM Transactions on Database Systems (TODS) 42 (2017): 1 - 21. https://www.ccs.neu.edu/home/vip/teach/DMcourse/2_cluster_EM_mixt/notes_slides/revisitofrevisitDBSCAN.pdf. Accessed October 22, 2022

[3] Canadian Government. (Updated April 19, 2022). "Algorithmic Impact Assessment tool". Canadian Government. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html. Accessed October 22, 2022

[4] Wikipedia editor. (September 29,, 2022). "Automatic identification system". Wikimedia Foundation, Inc. https://en.wikipedia.org/wiki/Automatic_identification_system#:~:text=AIS%20was%20developed%20in%20the,74%20kilometres%20(46%20mi). Accessed October 23, 2022

[5] Zook, M., et. al. (March 30, 2017). "Ten simple rules for responsible big data research". PLOS Corporation. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399. Accessed October 23, 2022

[6] Sandvig, C. (2020). [Video 3]. MADS Coursera. "Johnny Cash's keyword: "Power" (a.k.a. Is this regressive or progressive?)". University of Michigan. https://www.coursera.org/learn/siads503/lecture/IcvoQ/johnny-cashs-keyword-power-a-k-a-is-this-regressive-or-progressive. Accessed October 23, 2022

Appendix B - Data Schema

Data Source:

- https://coast.noaa.gov/htdata/CMSP/AISDataHandler/2022/index.html
- https://coast.noaa.gov/htdata/CMSP/AISDataHandler/2021/index.html
- https://coast.noaa.gov/htdata/CMSP/AISDataHandler/2020/index.html

Please refer to DP1-DownloadAIS notebook for download data

Automatic Identification System (AIS) Data Schema

Column Name	Description	Data Type
MMSI	Maritime Mobile Service Identity value	Text
BaseDateTime	Full UTC date and time	DateTime
LAT	Latitude	Double
LON	Longitude	Double
SOG	Speed Over Ground	Floatl
COG	Course Over Ground	Float
Heading	True heading angle	Float
VesselName	Name as shown on the station radio license	Text
IMO	International Maritime Organization Vessel number	Text
CallSign	Call sign as assigned by FCC	Text
VesselType	Vessel type as defined in NAIS specifications	Integer
Status	Navigation status as defined by the COLREGS	Integer
Length	Length of vessel (see NAIS specifications)	Float
Width	Width of vessel (see NAIS specifications)	Float
Draft	Draft depth of vessel (see NAIS specifications)	Float
Cargo	Cargo type (see NAIS specification and codes)	Text
TransceiverClass	Class of AIS transceiver	Text

Data Source: example info for one ship https://www.balticshipping.com/vessel/imo/9627992 .
Please refer to the Web_Scrape_Baltic_site notebook for download data.

Vessel Particulars Data Schema

Column Name	Description	Data Type
IMO number	International Maritime Organization Vessel number	Text
MMSI	Maritime Mobile Service Identity value	Text
Name of the ship	Name as shown on the station radio license	Text
Vessel Type	Vessel type as defined in NAIS specifications	Text
Operating Status	Status of vessel	Text
Flag	Jurisdiction under whose laws the vessel is registered or licensed,	Text
Gross Tonnage	Nonlinear measure of a ship's overall internal volume	Float
Deadweight	Measure of how much weight a ship can carry	Float
Length	Length of vessel	Float
Breadth	Breadth of vessel	Float
Year of build	The year the vessel was build	Integer
Builder	Manufacturer of vessel	Text
Classification Society	The organization that establishes and maintains technical standards for the construction and operation of the vessel	Text
Home Port	The port at which a vessel is based	Text
Owner	The Owner of vessel	Text
Manager	The operator of vessel	Text

Appendix C - Notebook Catalog

A read only copy of notebooks can be accessed at Deepnote with this link https://deepnote.com/workspace/gxcsiads696-3ac0374b-3c93-47ba-8674-5b5aae31d0e2/project/Milestone-II-6b18b33d-3a56-4f49-ad6e-71ecea9f0183

Below table list the notebook names and their functionalities

Notebook Category	Notebook Description	Notebook Name
Data Preprocessing	Download RAW AIS file from https://marinecadastre.gov/	DP1-DownloadAIS
	Apply geofence to filter raw AIS data. We only required data in the port area.	DP2-Filter AIS Data
	Resample AIS data at every 30 min internal	DP3-Resample AIS
	Apply geofence to detect the port entering time for a vessel	DP4-Find Port Entering Time
	Apply geofence to detect the berth time for vessels	DP5-Genfence
	Pair the port entering time and berth time for a vessel in an time interval in order to calculate the the dwell time	DP6-Calc Dwell Time
	Generate new features	DP7-Generate New Feature
	Join vessel particular data with vessel identifier	DP8-Merge Vessel Particular
	Web Scraping for Vessel Particulars from https://www.balticshipping.com/	Web_Scrape_Baltic_site
Unsupervised Learning	Appy FAMD analysis	FAMD_on_Vessel_Data
	Apply UMAP dimensionality reduction on mixed type data and parameter tuning	UMAP_on_Vessel_Data
	Apply Clustering algorithms over UMAP	Clustering_on_Vessel_Data
	Explore Gower distance with clustering	GowerDist_Explore_on_Vessel_Data
Supervised Learning	Build Linear Regression model	LinearRegression_with_vessel_data
	Build RandomForest model	RandomForest_with_vessel_data
	Build Catboost model	CatBoost for Vessel Data
	Feature Analysis for the CatBoost model	Feature Analysis
	Learning Curve Analysis for the CatBoost model	LearningCurve

	Sensitivity Analysis for the CatBoost model (data generation)	Sensitivity_Get_Ouput
	Sensitivity Analysis for the CatBoost model (visualization and analysis)	Sensitivity Analysis
	Error Analysis for the CatBoost model	Error Analysis

Appendix D - Algorithmic Impact Assessment Results

Algorithmic Impact Assessment Results

(10/21/2022)

Version: 0.9.1

Project Details 1. Name of Respondent
GXC

2. Job Title
Student

3. Project Title
Shipping Dwell Time

4. Project Phase
Implementation

5. Please provide a project description:
Student project

Business Driver / Positive Impact
[Points: 0]

6. What is motivating your team to introduce automation into this decision-making process? (Check all that apply)Improve overall quality of decisions
Lower transaction costs of an existing program
Use innovative approaches

About The System

7. Please check which of the following capabilities apply to your system.
Risk assessment: Analyzing very large data sets to identify patterns and recommend courses of action and in some cases trigger specific actions Process optimization and workflow automation: Analyzing large data sets to identify and anomalies, cluster patterns, predict outcomes or ways to optimize; and automate specific workflows

Section 1: Impact Level : 1

Current Score: 18 Raw Impact Score: 18 Mitigation Score: 6

Section 2: Requirements Specific to Impact Level 1
Peer Review
None
Notice
None

Human-in-the-loop for decisions
Decisions may be rendered without direct human involvement.
Explanation Requirement
In addition to any applicable legal requirement, ensuring that a meaningful explanation is provided for common decision results. This can include providing the explanation via a Frequently Asked Questions section on a website.
Training
None
Contingency Planning
None
Approval for the system to operate
None
Other Requirements
The Directive on Automated Decision-Making also includes other requirements that must be met for all impact levels.
Link to the Directive on Automated Decision-Making
Contact your institution's ATIP office to discuss the requirement for a Privacy Impact Assessment as per the Directive on Privacy Impact Assessment.

Section 3: Questions and Answers

Section 3.1: Impact Questions and Answers

Risk Profile

1. Is the project within an area of intense public scrutiny (e.g. because of privacy concerns) and/ or frequent litigation?No [Points: +0]
2. Are clients in this line of business particularly vulnerable?
No [Points: +0]
3. Are stakes of the decisions very high?
No [Points: +0]

4. Will this project have major impacts on staff, either in terms of their numbers or their roles?
No [Points: +0]
Project Authority5. Will you require new policy authority for this project?
No [Points: +0]
About the Algorithm6. The algorithm used will be a (trade) secret
No [Points: +0]
7. The algorithmic process will be difficult to interpret or to explain
No [Points: +0]
Impact Assessment8. Will the system only be used to assist a decision-maker?
Yes [Points: +1]
9. Will the system be replacing a decision that would otherwise be made by a human?
No [Points: +0]
10. Will the system be replacing human decisions that require judgment or discretion?
No [Points: +0]
11. Is the system used by a different part of the organization than the ones who developed it?
Yes
12. Are the impacts resulting from the decision reversible?

Reversible
13. How long will impacts from the decision last?
Impacts are most likely to be brief
[Points: +4]
[Points: +1]
[Points: +1]
14. Please describe why the impacts resulting from the decision are as per selected option above.Ship traffic might be rerouted and preference given to certain operators. But itcan always be changed or disputed by various ship operators affected.
15. The impacts that the decision will have on the rights or freedoms of individuals will likely be:
Little to no impact [Points: +1]
16. Please describe why the impacts resulting from the decision are (as per selected option above).The vessel dwell time is meant for companies to use. We don't use anypersonally identifiable information in the machine learning.
17. The impacts that the decision will have on the health and well-being of individuals will likely be:Little to no impact [Points: +1]
18. Please describe why the impacts resulting from the decision are (as per selected option above)Not a health related application.
19. The impacts that the decision will have on the economic interests of individuals will likely be:Little to no impact [Points: +1]
20. Please describe why the impacts resulting from the decision are (as per selected option above)The dwell time estimates are for container shipping logistics.
21. The impacts that the decision will have on the ongoing sustainability of an environmental ecosystem, will likely be:Moderate impact [Points: +2]
22. Please describe why the impacts resulting from the decision are (as per selected option above)It's possible that our project will help shipping dwell times become moreefficient.

About the Data - A. Data Source

23. Will the Automated Decision System use personal information as input data?
No [Points: +0]
24. What is the highest security classification of the input data used by the system? (Select one)
None
25. Who controls the data?
Open Data Source
26. Will the system use data from multiple different sources?
Yes
[Points: +0]
[Points: +0]
[Points: +4]
27. Will the system require input data from an Internet- or telephony-connected device? (e.g. Internet of Things, sensor)
No
28. Will the system interface with other IT systems?
No
29. Who collected the data used for training the system?
Your institution
30. Who collected the input data used by the system?
Your institution

About the Data - B. Type of Data

[Points: +0] [Points: +0] [Points: +1] [Points: +1]
31. Will the system require the analysis of unstructured data to render a recommendation or a decision?
No [Points: 0]

Section 3.2: Mitigation Questions and Answers

Consultations
1. Internal Stakeholders (Strategic policy and planning, Data Governance, Program Policy, etc.)
No [Points: +0]
2. External Stakeholders (Civil Society, Academia, Industry, etc.)
Yes [Points: +1]
3. Which External Stakeholders have you engaged?
Academia

De-Risking and Mitigation Measures - Data Quality

4. Do you have documented processes in place to test datasets against biases and other unexpected outcomes? This could include experience in applying frameworks, methods, guidelines or other assessment tools.No [Points: +0]
5. Is this information publicly available?
No [Points: +0]
6. Have you developed a process to document how data quality issues were resolved during the design process?No [Points: +0]
7. Is this information publicly available?
No [Points: +0]
8. Have you undertaken a Gender Based Analysis Plus of the data?
No [Points: +0]
9. Is this information publicly available?
No [Points: +0]
10. Have you assigned accountability in your institution for the design, development, maintenance, and improvement of the system?No [Points: +0]
11. Do you have a documented process to manage the risk that outdated or unreliable data is used to make an automated decision?No [Points: +0]
12. Is this information publicly available?
No [Points: +0]
13. Is the data used for this system posted on the Open Government Portal?
No [Points: +0]

De-Risking and Mitigation Measures - Procedural

Fairness
14. Does the audit trail identify the authority or delegated authority identified in legislation?
No [Points: +0]
15. Does the system provide an audit trail that records all the recommendations or decisions made by the system?No [Points: +0]

16. Are all key decision points identifiable in the audit trail?
No [Points: +0]
17. Are all key decision points within the automated system's logic linked to the relevant legislation, policy or procedures?No [Points: +0]
18. Do you maintain a current and up to date log detailing all of the changes made to the model and the system?Yes [Points: +2]
19. Does the system's audit trail indicate all of the decision points made by the system?
No [Points: +0]
20. Can the audit trail generated by the system be used to help generate a notification of the decision (including a statement of reasons or other notifications) where required?
No [Points: +0]
21. Does the audit trail identify precisely which version of the system was used for each decision it supports?No [Points: +0]
22. Does the audit trail show who an authorized decision-maker is?
No [Points: +0]
23. Is the system able to produce reasons for its decisions or recommendations when required?
No [Points: +0]
24. Is there a process in place to grant, monitor, and revoke access permission to the system?
No [Points: +0]
25. Is there a mechanism to capture feedback by users of the system?
No [Points: +0]
26. Is there a recourse process established for clients that wish to challenge the decision?
No [Points: +0]
27. Does the system enable human override of system decisions?
Yes [Points: +2]
28. Is there a process in place to log the instances when overrides were performed?
No [Points: +0]
29. Does the system's audit trail include change control processes to record modifications to the system's operation or performance?
No [Points: +0]
30. Have you prepared a concept case to the Government of Canada Enterprise Architecture Review Board?No [Points: +0]

De-Risking and Mitigation Measures - Privacy

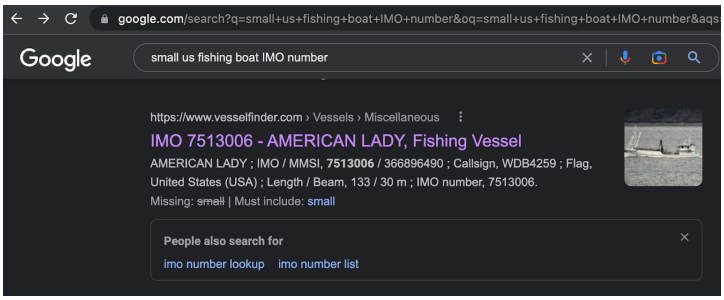
31. If your system involves the use of personal information, have you undertaken a Privacy Impact Assessment, or updated an existing one?No [Points: +0]
32. Have you designed and built security and privacy into your systems from the concept stage of the project?Yes [Points: +1]
33. Is the information used within a closed system (i.e. no connections to the Internet, Intranet or any other system)?No [Points: +0]
34. If the sharing of personal information is involved, has an agreement or arrangement with appropriate safeguards been established?
No [Points: +0]

Appendix E - How a IMO can lead to issues with Personally Identifiable Information (PII)

This is a demonstration of how in only a few minutes you can find potential PII, even financial information, starting your search with just an IMO. It led to a company (an LLC) and person name and how much sales they make. It's common for LLC's to have only one or a few owners so in theory this could be a small business and if it's only one owner the financial information found might potentially represent that person's annual income.

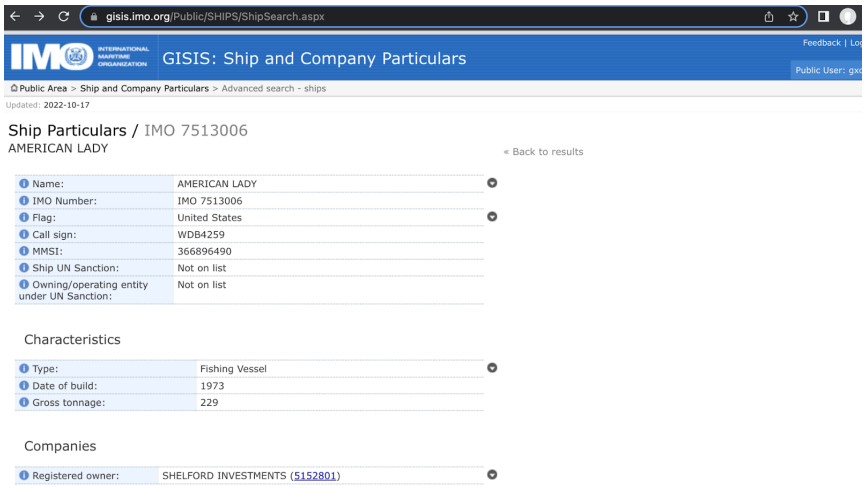
Step 1:

Find a vessel IMO number from what might be a smaller operator…who usually own small companies like LLC's which have very limited number of owners (as opposed to a big corporation which makes finding personal owners and data harder)



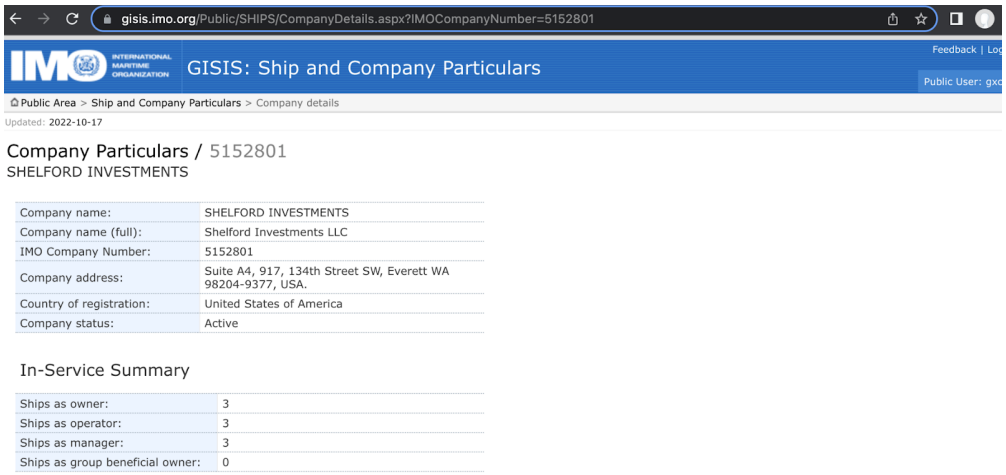
Step 2.

Get vessel data based on IMO. In this case we look for what google search results show as 7513006. Try balticshipping.com or if not, then go to the source of IMO's It's free to register and search. In our case we decided to register. Here we go to the IMO authority directly:



Step 3.

Look for info on the company by clicking on the registered owner. Great, it's an LLC. And it gives an address too, useful for further research to hone in on the owner.



Step 4.

Look for deep details about the company such as the person or people who own it. Some databases available to University of Michigan students have deep detailed info on companies. And bingo. The addresses and company name matches. Now we have a contact as well. For small companies, it can be an owner of the company that is the contact name. We see Richard Shelford, phone number and sales of $320K…probably for the most recent year on record. If he is the sole owner, this could be guessed as his annual income.