AI Port Control: Augmented Intelligence for Vessel Inspections

Incoming Ship Risk Assessment with Machine Learning

MAKING AN IMPACT THAT MATTERS *since 1845*

AI Port Control: Augmented Intelligence for Vessel Inspections

Targeting Incoming Ship Risk Assessment with Machine Learning

Maritime security and safety pose unique challenges for the United States Coast Guard (USCG) when enforcing international regulations for Port State Control (PSC). In 2021, over 70,000 ships docked at US ports, but only ~10% could be physically examined by PSC officers. Of the 7,000 ships that were inspected, ~18% of those inspected vessels were found to contain deficiencies, with 1% of inspected ships detained at port. For this program, a key priority is to identify methods that efficiently target deficient vessels to improve operational capacity and logistics. AI and machine learning solutions have been developed to predict which vessels pose the greatest risks using automated methods and help personnel make efficient decisions.

Challenge:
Develop and integrate an AI/ML powered vessel deficiency assessment system

Solution highlights:

- Analyze tens of thousands of prior docking and inspection records to determine what outward characteristics signal internal deficiency
- Design a robust pipeline to augment on–site personnel judgement
- Deploy test, validation and periodic model update systems to handle real–world drift

Industry	**Government / Security**
Use case	**Machine Learning**
Data	**Shipping Records**

As this inspection process is labor-intensive and monetarily expensive, custom AI and machine learning (ML) algorithms can be leveraged to improve operational efficiency and reduce cost by improving inspection targeting. Deloitte has developed a novel AI-enabled decision support tool designed for two main purposes:

- Determine the likelihood of deficiencies in incoming vessels for effective pre-screening and prioritization.
- Highlight the data that has the most impact on the predicted risk allowing for human interpretation of model predictions.

Simulated evaluation of the solution demonstrated a nearly two-fold (89%) increase in efficiency over current rule–based methods: 34% of inspected ships were found deficient, up from 18% (Figure 1). As the prevalence of deficiencies among uninspected vessels cannot be known with certainty, the model's focus is to increase the prevalence of deficiencies among inspected vessels. It accomplishes this by ordering potential targets in order of their ML-calculated likelihood of containing one or more deficiencies aboard; as the same number of vessels are inspected, the model catches more deficiencies using the same amount of resources. The AI solution is deployable to any handheld device or computer at the point of inspection to provide real-time support and transparent analysis of vessel characteristics. Additionally, to accelerate deployment time, a simplified AI-informed linear regression model will allow the USCG to adopt an offline, hard-copy version, translating digital models to a form available on a clipboard.

As the ability to predict deficient vessels from historical data is a valuable tool with high operational relevance, understanding the implementation and methods is imperative towards real-world testing and adoption. In this report, we discuss challenges posed by maritime vessel inspection, how data and machine learning can be leveraged to improve vessel targeting, and provide recommendations on real-world evaluation and deployment to integration with USCG systems.

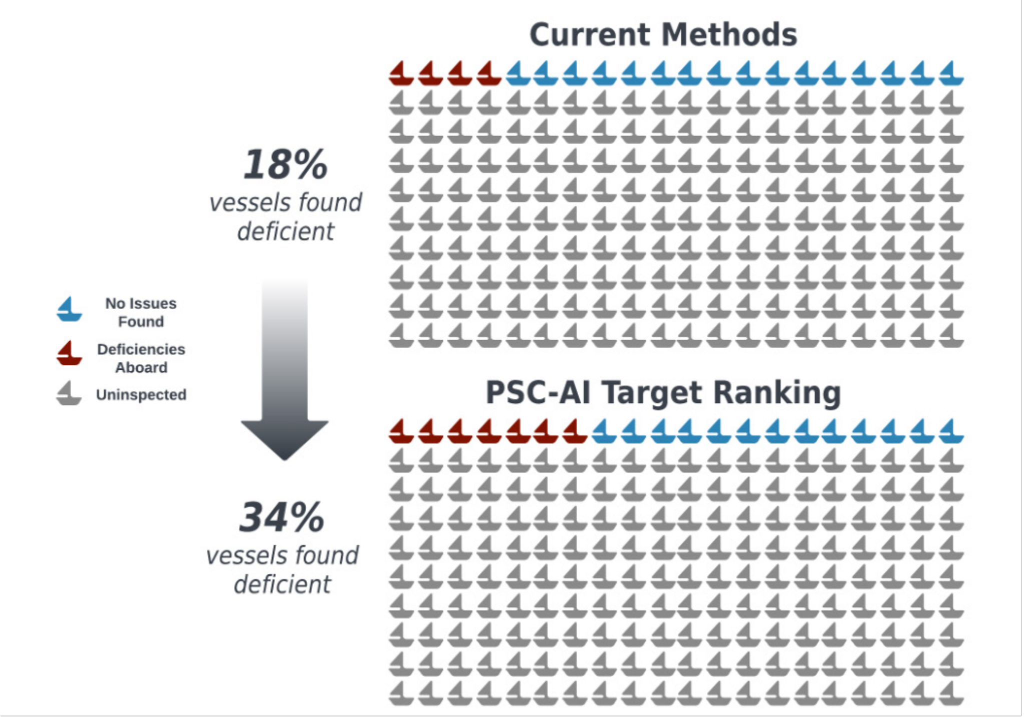


Figure 1. Comparison of inspection efficiency rates of contemporary methods vs. the PSC AI model, defined as the number of ships found to be deficient within the number of ships inspected. Uninspected vessels are represented in grey, non-deficient in blue and deficient in red.

Vessel Deficiency & Inspection Targeting

Port State Control is fundamentally a problem that must be faced with limited resources—not every ship can be inspected, and some infractions are bound to escape notice. Maximizing inspection efficiency at the same level of resource investment is, therefore, a primary goal to improve operational efficiency.

To identify vessels that are likely to be non-compliant, inspection personnel currently use a heuristics-based targeting matrix that cannot easily adapt to changing conditions and new hazards. AI algorithmic targeting can greatly increase the odds for such deficiency identification and augment workflows to improve inspection efficiency.

During development, Deloitte amassed a dataset of vessel and inspection data. AI/ML models can learn patterns in this vessel-level data (e.g., registered nation, port of origin, cargo contents, etc.) that predict the presence of infractions. Two data sources were pre-processed and combined: VesselFinder includes over 100,000 port calls from foreign-flagged ships between 2018 and 2021, while the USCG PSIX dataset holds over 20,000 records of vessel inspections. Integration of these two sets directly links vessel characteristics to inspection outcomes,
offering a variety of variables for models to associate with ground-truth deficiency labels. An example inspected vessel data point includes the ship's registered nationality, year built, previous port, arrival time and draught. Using this composite dataset, custom AI models were developed to learn how vessel characteristics relate to deficiency status, determining important vessel features for manual review and providing a real–time risk score to support the USCG's targeting matrix.

The model's performance, measured in targeting precision, was then evaluated using a holdout set of inspected ships that the model had never seen. Here, precision is defined as the ratio of ships found to be deficient versus the number of inspections performed or the percentage of inspections that unearth deficiencies. During evaluation, the model predicted a risk score for each unknown ship based on its characteristics, and the highest 10% were ranked and digitally inspected; their actual deficiency status is revealed, adding to or subtracting from the model's targeting precision rate. To increase technical transparency, the PSC AI model also can output interpretability graphs for each vessel judgment. These plots (Figure 3) visualize the impact of each ship characteristic, such as draught, cargo type, towards the model's overall risk score.

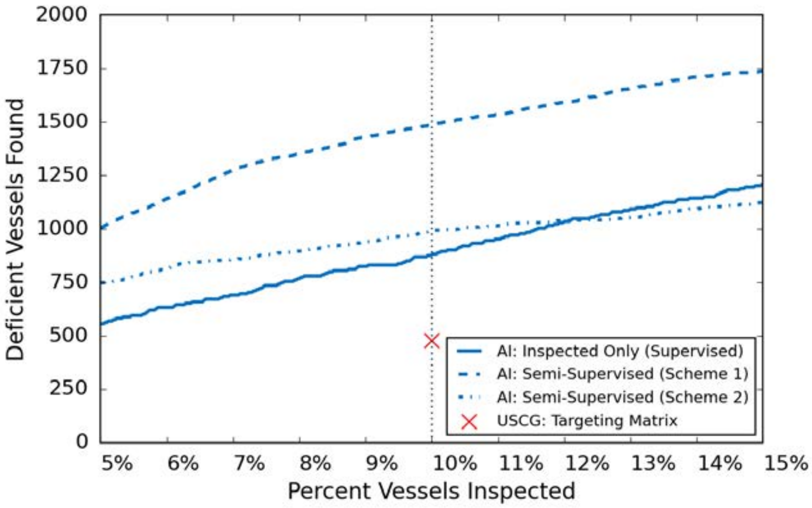


Figure 2. Vessel Inspection Efficiency. The proportion of inspections discovering deficiencies of the new PSC ML model performs favorably compared to the current USCG Targeting Matrix. Model–supervised refers to an alternate, experimental version of the model with greater potential performance but higher risk of bias.

Figure 2 compares the performance of current PSC inspections versus the AI model approach. Of the ships inspected by the PSC ML model, 34% actually contained one or more deficiencies aboard. This is nearly a two-fold increase to the 18% deficiency rate among vessels identified under the current PSC rule-based targeting matrix. In the field, this gain in precision grants inspection personnel an accurate diagnostic tool that boosts operational efficiency by wasting less time on non–deficient vessels.

In addition, other experimental model designs were tested that indicate opportunities for continued improvement, with a model–supervised version of the model scoring much higher, reaching over 57% (a 212% increase over non–AI) precision. However, this model-supervised approach currently carries several caveats in terms of data bias and real-world generalization that are detailed further below and can be overcome as additional data is collected in the future. This strategy provides an estimated upper range of detectable deficiencies available in the entire dataset.

Model Architecture & Primary Challenges

The maritime defense space presents two challenges that effective AI must overcome: transparency and data volume. For commercial and private-sector applications, AI/ML-based systems often have significant operational freedom, as they are generally subject to fewer regulatory constraints. In contrast, for maritime safety and security, decisions have a direct regulatory impact, creating a greater need for modeling transparency and preserving the inspector's decision-making agency. Additionally, the resource constraints that limit the percent of ships inspected also create a lack of ground–truth data for model training; without a sufficiently large amount of data from which to learn a representation of reality, most models will fail to accurately predict outcomes.

The PSC model has been specifically developed to exceed these operational requirements. The primary machine learning architecture was built using industry-standard gradient boosted tree methods that offer state–of–the–art performance on tabular data, effectiveness in low-data scenarios, and ensures robustness with newly collected data. Gradient boosted algorithms build multiple sequential iterations of models, using data resampling, regularization and other techniques to finely tune a final model to the data available. They offer reliable performance in low–data scenarios, and also provide near–instant computations on single data points, minimizing the time between ship data input and risk prediction output.

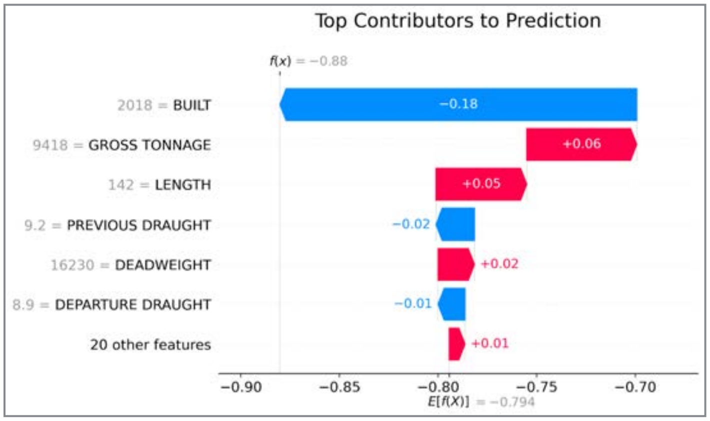
Interpretability

Federal applications have heightened requirements for interpretability; greater risk upon failure demands greater transparency in decisions. Notably, some ML algorithms can avoid the notorious "black box" problem, where the machine logic that decides predicted output is fully opaque to human users. To maintain transparency and traceability standards adopted by the Federal Government, Deloitte has crafted two solutions to offer interpretability:

- Visualization of key features (nationality, cargo load, etc.) leading to ship risk classification
- A parallel model with reduced technical complexity, offering simple arithmetic calculations

Importantly, this integrates the SHapley Additive exPlanations (SHAP) algorithm, which deconstructs each decision that the inspection model makes (Figure 3) to reveal how each ship characteristic influenced the model's decision. SHAP also allows the calculation of aggregate feature importance, visualizing which vessel variables are most influential in modeling predictions overall.

For example, the model's decision to flag a certain ship may be revealed to rest largely on a combination of the specific port of origin with a cargo load above a certain threshold. These "characteristics" are represented by mathematical weights the model encodes from the training data. When these weights are made transparent, humans are able to more clearly see which information may be influencing or skewing predicted outcomes and help influence real-time inspection protocols.



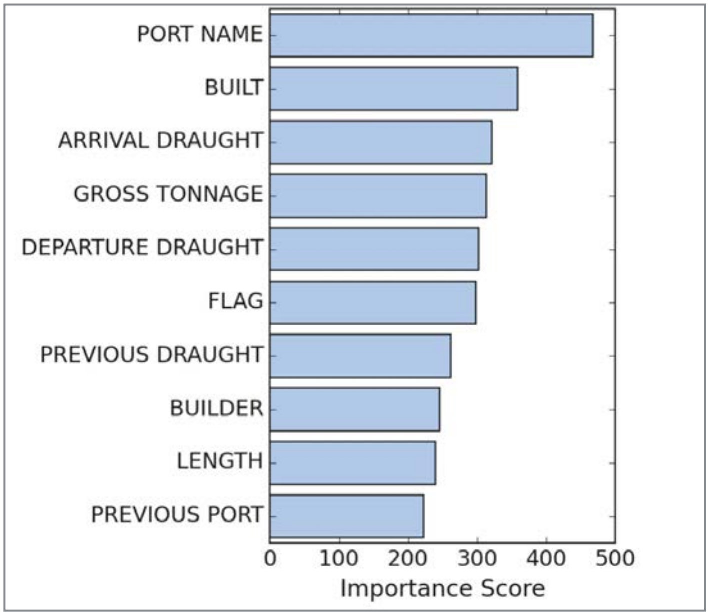


Figure 3. Sample Feature Importance of Model. Top—SHAP analysis of a single ship prediction. Red bars indicate a relative increase in deficiency probability, while blue indicates a relative decrease. Bar length represents the magnitude of the feature's impact. The x–axis is drawn in log odds, with 0 indicating 50/50 probability. Bottom—Aggregate feature importance of one model iteration. These scores represent each feature's overall influence in decisions in either direction (inspect or do not inspect).

Providing statistics on a digital clipboard

Alongside the main gradient boosted model, Deloitte has created a simplified model to offer fully explainable decisions when choosing which ships to inspect. The ship risk percentages from this linear model can be demonstrated in real-time to end users who require numerical explainability of each decision. Each vessel risk decision from this simplified model can be clearly represented as a linear regression formula where each ship variable has an interpretable coefficient. As a result, the linear model can operate offline, with users manually writing the characteristics of a new vessel into the linear formula to receive an inspection recommendation.

The linear model is less complex than the gradient boosted version and will necessarily sacrifice some performance in exchange for full traceability, as boosted models are able to learn more complex data interactions. However, end-users can deploy these two models in parallel and receive both simultaneous predictions for each incoming ship: The main model's higher–accuracy judgment displaying how much each vessel characteristic contributed to the decision, as well as the linear model's score is supported by easily explainable variable weights. While accuracy may rank slightly lower than the gradient boosted model, this streamlined version allows the USCG to have a better understanding of the decisions the model is making.

Low–data strategies

In a real-world environment, developing a representative dataset and aggregating the required information can prove challenging, yet data quality is a primary driver of success in developing AI models. To begin building an effective dataset for Port State Control, information on incoming ships were aggregated from publicly available government–hosted shipping registries, while additional proprietary sources were purchased. Matching vessel characteristics to deficiency information required significant data cleaning and curation. While many ships come into port each day, the resource constraints limiting the number of inspections performed result in relatively low amounts of data on inspected ships. This leads to two notable data–related challenges:

- A scarcity of target–class data, those ships that are both inspected and found deficient
- A majority of overall ship data points are unlabeled as a result of non-inspection

Due to the nature of the time-constrained investigation process, the large majority of ships sail by freely, leaving a large number of vessels without inspections. This relatively small number of ships inspected and found to be deficient leaves too little information for most models to successfully learn from, as the dataset does not represent a broad enough sample of all ships visiting ports.

Examining available datasets also reveals a question of selection bias: The ground–truth data available represent ships scored by the USCG's targeting matrix. This is unlikely to represent a truly balanced dataset of all ships, and thus introduces a native selection process and a sampling distortion as an additional challenge. Training on the small number of inspected ships selected by the targeting matrix results in a narrow model that fails to generalize to more diverse, real-world conditions. Deloitte implemented methods to resolve these challenges and create a robust model capable of efficient vessel targeting in production.

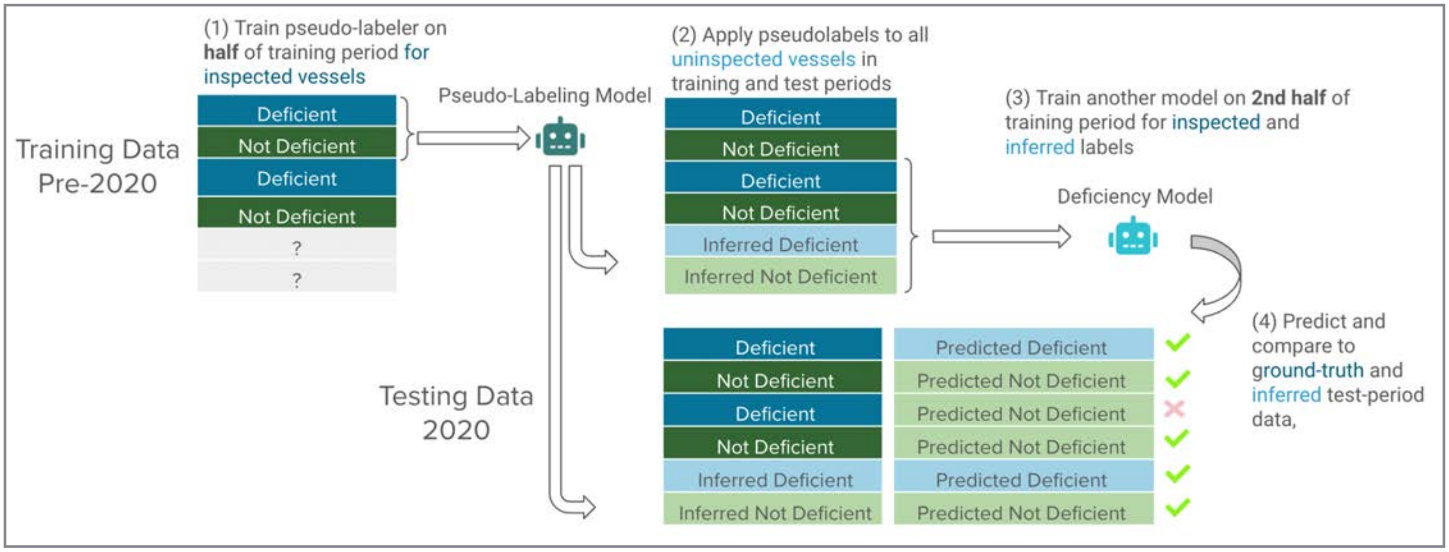


Figure 4. Overview of Psuedo-Labeler and Deficiency Predictor model interactions on training and testing data. By predicting labels on uninspected ships, the model-supervised approach 'unlocks' more data to be used for vessel deficiency targeting, though it is inherently more biased than the purely realistic ground-truth model.

CONTRASTIVE & MODEL–SUPERVISED LEARNING

The PSC model also tested a contrastive learning process to overcome these obstacles (Figure 4). This method uses altered "semi-synthetic" data points to increase the model's resilience to adjacent, unknown samples, and to estimate the potential gain from increasing the amount of ground-truth data. Unlike many data augmentation methods that generate fully synthetic and often random data points, the PSC solution takes a different approach and instead synthesizes labels (of deficiency status) for real-world ship data. Model-supervised approaches demonstrate the ability to learn from previously unusable data that lacks "ground-truth" labels[1], and is employed here to analyze ships that have never been inspected, albeit with assumptions about the prevalence of deficiencies.

The training data was first split into two equal groups: Half for an augmentative "Labeler" model, and a half for a more standard "Predictor" model. The Labeler model is trained to predict deficient or not deficient labels on uninspected vessels, and then populates the remaining dataset with optional artificial labels indicating predicted deficiency status. Next, two versions of the Predictor model are trained on the second half of the data, one version using only real deficiency labels, the other using both real and artificial labels. Similar labels are thus propagated between similar ships in an unsupervised manner, then used to augment the standard supervised learning and prediction task.

While the artificial-label model reaches 57% precision compared to the real-label's 34%, there is no perfect method to confirm the performance of the Labeler, since no real inspection has been carried out for ships it artificially labels. Thus, it essentially trades a low-data bias for one of self-confirmation, and does not seamlessly translate to a real-world gain in performance. Instead, this contrastive learning process provides an estimated range of detectable deficiencies available in the entire dataset, and highlights the potential to increase model accuracy as further data is made available.

Deployment & Real-World Testing

The PSC AI solution can be easily distributed and accessed through an API endpoint and is containerized for ease of installation into any current compute environment, hybrid data center, or cloud environment of choice. The ML model has been developed for use with a GPU, and tested on an AWS p2.xlarge instance with an Ubuntu Deep Learning AMI.

To deploy such a system, organizations must manage the necessary software and IT infrastructure. This could either be on-premises or cloud computing resources offered by third-party vendors such as Amazon Web Services (AWS) or Microsoft Azure. From there, a secure application can be developed and tested for delivery via a mobile device to personnel at ports.

While having your own team set up hardware on-site gives you full control over IT specifications and transfer protocols, cloud computing can adapt and scale dynamically without significant procurement or setup costs. Moreover, a cloud-based solution has the following advantage in the PSC problem context:

- Ports are distributed so data must travel, preventing on-premise redundant security
- Shipping and security conditions may change, creating a need for flexible systems

Though cloud hosts offer a variety of secure data transfer protocols, some regulations require that data travel as little as possible as a redundant safety measure. However, it would be prohibitively expensive to set up separate computing resources at each port to run the targeting AI separately. An on-premises solution would have to be centralized at a PSC center that receives data from each port and returns deficiency predictions.

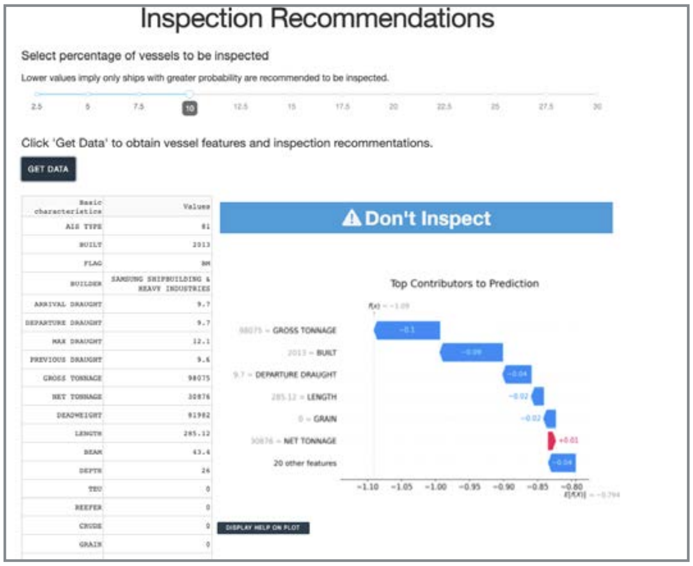
This minimal-travel advantage of an on-premises deployment is neutralized by the distributed nature of shipping and inspection locations.
Furthermore, shipping and inspection protocols are subject to change over time: Cloud-based resources are inherently more adaptive and can be changed or scaled without incurring additional cost of the third-party service charges. For example, if the size of incoming data changes dramatically, perhaps due to new information-recording systems or a global disruption in shipping routes, additional on-premise hardware must be manually purchased and integrated to meet demand, requiring time, personnel, planning, and upfront expenditure.

A cloud-based deployment can handle such changes with minimal downtime and can be configured to scale automatically to meet spikes in demand. At the same time, additional costs can be saved by scaling down reserved cloud compute space during periods of low activity.

The PSC model may also be deployed as one service within a pre-existing Kubernetes cluster. Containerization in this manner future-proofs the PSC software, allowing for stable relaunches of the model on additional systems. Containers may also be kept dormant to save on computation costs until vessel judgment is required.

Planned integration & user application

the following example of a cloud-based workflow for PSC is both flexible and resilient: Incoming ships transmit data to the port authority, which automatically routes ship data through a secure connection to a cloud-hosted private network. Servers in the PSC network feed data through one or more models and return the resulting output of each. These predictions and recommendations populate a custom built application (Figure 5), which credentialed employees can access to either replace or support their inspection protocols.



Select percentage of vessels to be inspected

Lower values imply only ships with greater probability are recommended to be inspected.

Click 'Get Data' to obtain vessel features and inspection recommentations.

Basic characteristics	Values
AIS TYPE	81
BUILT	2013
FLAG	BM
BUILDER	SAMSUNG SHIPBUILDING & HEAVY INDUSTRIES
ARRIVAL DRAUGHT	9.7
DEPARTURE DRAUGHT	9.7
MAX DRAUGHT	12.1
PREVIOUS DRAUGHT	9.6
GROSS TONNAGE	98075
NET TONNAGE	30876
DEADWEIGHT	91982
LENGTH	285.12
BEAM	43.4
DEPTH	24
TEU	0
REEFER	0
CRUDE	0
GRAIN	0

Figure 5. Inspection Model Dashboard. An application displaying a prediction on previously unseen ship data. Features and their influence on the risk score are displayed; the percentage of vessels to be inspected is customizable, allowing users to tune the model's degree of sensitivity.

These rapid deployment capabilities can be provided by a single integrated platform, deployed to a web-based browser, tablet, or other interfaces. Alternatively, they can be deployed onto dashboards and USCG tools, developed as a custom API service, or created as a combination of these approaches.

In most cases, these processes can be deployed in stages rather than all at once in a single deployment. An organization's plan for adopting AI capabilities should align with a testing plan and an initial pilot in accordance with USCG's technical priorities and skills maturity. For example, many organizations start by focusing on application testing in a select group of inspectors and understand real-world performance of the prediction serving once pilot SOPs have been approved.
To succeed, continuous training and continuous monitoring will be necessary to properly pilot a relatively small number of inspections.

Alternatively, the model can be launched directly into a "soft deployment" phase where predictions are delivered directly to USCG devices, but inspection personnel are not instructed to use its predictions from these tests. Once validated, a phase two approach can be taken to implement standards on use and acceptance and allow inspectors to leverage the solution and prove its field value in predetermining risk.

Managing drift

Since typical AI models are built upon historic datasets, their predictions are only as accurate as incoming data is up-to-date. Drifts in data distribution that may occur over time or point events that fundamentally change shipping behavior can cause degradation to model performance as the learned patterns may no longer be relevant for new data in the real world. There are two sources that can cause these gaps between model and reality:

- Data issues and biases during initial training
- Naturally–occurring drift post-deployment

Changes in incoming data over time can slowly skew model outputs through selection bias: If a PSC model first trains upon a limited subset of ships, e.g., only those coming from Asia, then performance may suffer when examining new ships outside of that subset, as it lacks specific context. Identifying these data selection biases during development becomes paramount in effectively improving the generalizability of deployed models. During model development, Deloitte organized technical meetings between developers and SMEs to identify and address these biases, typically through some form of intelligent sampling and data augmentation methods. Drift post-deployment can be understood in two broad categories:

- Concept Drift occurs when target definitions move. This includes cases when deficiency definitions, maritime law, or the output assurance threshold for inspection, is changed.
- Conversely, Data Drift occurs when the inputs change. This includes shifts in the type and frequency of incoming vessels due to external factors.

Concept drift is somewhat more controllable in this case; rules and regulations for shipping are rarely altered with significant impact, and these changes generally come with significant notification from regulatory entities. Data drift is arguably more certain and deleterious, as changes in specific ship information and unknown and new variables can reduce performance. Rather than addressing these unknown future changes in development, it is often far more productive to simply update the model over time. Retraining or refreshing a model adds in new, recent data to learn from and adjusts the AI model's mathematical weights to better reflect current conditions.

A deployed PSC model is constantly fed new vessel data while it's in use, but most contemporary AI architectures are unable to predict and learn at the same time (as humans do) without extremely repetitive, labor-intensive computation. Therefore, new data is gathered into periodic batches and refreshed at scheduled intervals (either time or performance-based) to ensure continued model performance.

This process can also be automated; in particular, cloud-hosted architectures allow for simplified development of retraining pipelines that can take in new data and retrain models automatically. These architectures are relatively simple to deploy and are becoming industry-standard processes across the machine learning space. Thus data drift is ultimately a manageable challenge, since any model can be scheduled to refresh on user-decided timelines. As model performance holds up or falls over long periods of time, these timelines can be redefined to match accelerating or decreasing rates of drift and reported on a scheduled basis.

Testing & evaluation

Even though the developed PSC model demonstrated promising performance on the test data, real-world conditions can present new challenges that historical data is not guaranteed to capture. Rather than trusting any AI solution to immediately yield robust on-site performance, high-stakes scenarios require thorough testing and evaluation of model performance in situ, as well as cautious integration alongside end users and decision-makers.

A slow, measured incorporation process is important both to provide sufficient time to gauge performance and to better acclimate personnel with regular use of AI predictions. Especially in public safety and defense contexts, dramatic shifts — such as transitioning from human-written statistics to machine learning target outputs — have the potential for significant change management requirements and the transformation of standard operating procedures and training to be implemented over time. One nuanced approach to integration is a side–by–side "blind trial" period. For a set amount of time, the machine learning model is deployed without delivering predictions to decision-making personnel. For each incoming vessel manually judged by personnel, the model generates and writes output to a database. This allows direct comparison of ML risk scores to actual inspection outcomes without altering operational procedures. Metrics can then be generated to evaluate performance, including the probability of actual deficiency given a high model risk ranking as well as the probability of no deficiencies with a low ranking.

In the case where model performance does not initially improve inspection efficiency, agile development steps can be taken to ingest new datasets and iteratively update performance statistics over time. New, more powerful approaches can be explored as well to augment model predictions and utilize aggregated data sources. If the AI's predictions are found to be sufficiently accurate without posing a significant risk of upheaval, the model can move to full deployment, delivering predictions to key personnel to augment their inspection workflow. Alternatively, this automated evaluation trial period can be skipped entirely, and the model launched directly into a "soft deployment". Here, the model is deployed as a "digital twin" to deliver predictions directly to staff members, but inspection personnel are instructed to use its predictions at their own discretion (or under organizational guidelines). This places humans firmly in control of the technological shift, and helps to engender user trust in the PSC system while it proves its value in predetermining risk.

Conclusion

The current Deloitte PSC Targeting AI Solution showcases the potential benefits of augmenting human judgments with artificial intelligence. Trained on historic port data, the AI solution is estimated to improve targeting precision over the current human-intensive targeting matrix by between 89% and 212% depending on model choice. Beyond pure performance improvements, the solution will also highlight the high impact patterns in the data that drive model risk predictions for human–in–the–loop decision making. These model interpretability methods allow “soft” rollouts in high–stakes environments, allowing AI to transparently inform, rather than remove, human decisions. This yields immediate benefits while allowing trust to be built over time, respecting the pace and importance of organizational change. The solution is deployable in both cloud and on-premise environments, with the model capable of being integrated into personal devices to provide real-time insight into vessels before they reach the harbor.

The adoption of AI and automation tools can no longer be viewed as optional at the highest levels of federal and industrial operations, as the growth in complexity of data and IT challenges continually outpaces the growth of resources. Like advanced commercial enterprises, government agencies can adopt AI’s vast potential to reduce costs and improve performance, decision-making, and mission delivery. Given technical advances, the opportunities to exploit AI for PSC are limited only by the implementation of programs and policies that are dedicated towards operationalizing such technical solutions.

Expansion

One of the tenets of AI is for models to be task agnostic; the technology applied to PSC is most extendable to almost any problem requiring difficult judgment calls, especially in low–resource or fast-paced scenarios such as potential threat identification. Algorithms are able to unearth combinations of variables that are not immediately intuitive to humans and calculate precise predictions to support key decisions. The data format and specific architecture utilized in the PSC AI have found success in other high-risk environments where overwhelming amounts of data limit purely human solutions.

Beyond tabular data, similar AI pipelines can ingest temporal, image, and even text-based information to provide automated, scalable and cost–effective AI–augmented solutions. The modeling paradigms described in this white paper are directly generalizable to any applications where structured data exists. Indeed, since this form of tabular data is commonplace in almost all government agencies, AI has vast potential to improve operations in every sector in applications as diverse as fraud detection, enhancing public policy in population risk, to automating legal document review in benefits administration.

References

Zhi-Hua Zhou, A brief introduction to weakly supervised learning, National Science Review, Volume 5, Issue 1, January 2018, Pages 44–53, https://doi.org/10.1093/nsr/nwx106

Contact information:

Mike Segala, PhD
Principal
Deloitte Consulting LLP
msegala@deloitte.com

Alex Moseson, PhD
Federal AI Specialist
Deloitte Consulting LLP
amoseson@deloitte.com

Deloitte.

About Deloitte
Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.