Vessel Deficiency Severity Prediction Model

Team Members:

Isabella Tan Jing Xuan

Jagan Mohan Shailesh

Sera Ang Suan En

Zhang Le Qian Paul

Date of Submission: 17 Jan 2025

Abstract

The maritime industry faces challenges in consistently assessing and predicting the severity of deficiencies found during vessel inspections due to diverging opinions among inspectors and the complexity of interpreting technical descriptions. To address this, we developed a systematic approach to derive consensus severity ratings from Subject Matter Expert (SME) annotations and predict severity using machine learning.

We employed a majority voting technique to mitigate variability in SME annotations, ensuring the final severity reflects the predominant view. Data preparation included standardising deficiency codes, converting dates into Unix timestamps, and leveraging BERT embeddings to capture nuanced meanings from textual descriptions (def_text).

For predictive modeling, we used AutoML with FLAML to identify optimal models and hyperparameters, balancing accuracy and computational efficiency. Cross-validation and metrics such as precision, recall, and accuracy validated the model's robustness. This methodology integrates domain expertise, NLP, and automated machine learning, enhancing decision-making in maritime safety.

Part 1: Deriving Consensus Severity

Our group employed the **majority voting technique** to derive the consensus severity from SME annotations. This methodology assigns the final severity based on the most frequently occurring label among the SME-provided classifications. For instance, if three SMEs provide severity annotations for a deficiency as **“High,” “Medium,” and “High”**, the majority voting technique assigns **“High”** as the consensus severity. This ensures that the final decision reflects the

predominant view while minimising the influence of diverging individual opinions. The values in the column "annotation_severity", namely **"High"**, **"Medium"**, and **"Low"**, were converted into corresponding numerical values **3**, **2**, and **1**, respectively. This numerical representation simplifies computation and allows us to handle the data programmatically to determine the most frequently occurring label.

Why did we choose this technique? Majority voting is straightforward to understand and implement. It aligns with the intuitive notion of group consensus by prioritising the most commonly agreed-upon classification. Since the severity labels from SMEs may vary based on individual judgment, majority voting mitigates the impact of outlier opinions or extreme ratings, leading to a more balanced outcome.

Assumptions:

- No null values
- All inspectors are assumed to be of equal levels of reliability

Part 2: Predictive Model

Data Preparation

Firstly, we cleaned the dataset. We converted the data type of "InspectionDate" from a string to an integer representing its Unix timestamp format, which denotes the number of seconds elapsed since January 1, 1970. Moreover, we encoded "VesselGroup" to pass into the model.

Subsequently, we used tokenisation, which is a Natural Processing Language (NLP) technique. In this regard, we selected **Bidirectional Encoder Representations (BERT)** from Transformers. The “def_text” column contains descriptions of deficiencies, which include technical language and nuanced details. Understanding the severity of a deficiency requires understanding the context in which terms appear (e.g. "minor crack" VS "severe fracture"). BERT uses a bidirectional approach to process text, meaning it considers the context from both the left and right of each word. This allows it to understand nuanced meanings and relationships between words. By extracting embeddings (numerical vector representations) from the [CLS] token, BERT provides a condensed representation of the entire sentence.

After the “def_text” column was transformed into its vector representation, we then decided on the columns to be dropped as they would not be relevant as features for the models that we decided on testing. These columns were: VesselId, PscInspectionId, annotation_id and username.

Model Selection

For our model, we utilised **AutoML** from the FLAML library to train models for each category, enabling us to predict the consensus severity efficiently. **AutoML** automates the process of model selection, hyperparameter tuning, and training, allowing us to focus on the analysis and interpretation of results rather than manually iterating through multiple models and configurations.

We chose **FLAML** (Fast and Lightweight AutoML) due to its focus on low computational cost and fast execution, which aligns well with our project's constraints. FLAML employs efficient search strategies to identify optimal models and hyperparameters without requiring extensive

resources. This makes it an ideal choice since computational efficiency and time are critical to us.

By leveraging AutoML, we were able to train and evaluate multiple models across various severity categories. Moreover, we could automatically optimise hyperparameters to improve performance without manual intervention, eventually allowing us to obtain interpretable and reproducible results within a short timeframe.

Validation

For cross-validation, we split the dataset into **80%** training and **20%** validation. The 80% ensures the model has access to a sufficient amount of data for learning and 20% strikes a balance between training the model effectively and having a reliable evaluation metric.

To evaluate our model's performance effectively, we utilized the following metrics: **precision**, **recall**, **F1-score**, and **accuracy**. Each metric provides unique insights into different aspects of the model's performance, especially for multi-class classification tasks like predicting consensus severity.