Reducing Subjectivity in Predicting Severity Scores of Port State Control Inspections with AI second opinion.

Developed for the Maritime Hackathon 2025

Curteis Yang, Dong Luanjie, Ho Min Han

1. Abstract

The evaluation of Port State Control (PSC) inspections is highly tedious and subjective to the inspector, resulting in operational inefficiencies or critical oversights that cause accidents. This paper outlines a structured methodology to evaluate severity levels from findings identified during vessel PSC inspections. This involved using generative artificial intelligence (GenAI) augmented through prompt engineering and chain-of-thought reasoning to provide a second opinion to severity scores, balancing the subjective scores provided by human inspectors. AutoGluon, an AutoML framework, was used for predictive modelling from report details and corresponding human and AI severity scores, preceded by rigorous feature engineering to optimize data inputs. This approach ensures consistent, less biased, and scalable severity evaluations critical to the business viability of vessels in trade.

2. Introduction

Port State Control (PSC) inspections are critical for ensuring maritime safety, environmental protection, and compliance with international regulations. However, the subjective nature of PSC evaluations often leads to inconsistent severity scoring, operational inefficiencies, and potential safety risks. Inspector variability can result in uneven compliance outcomes and challenges in scaling evaluation frameworks. Generative AI (GenAI), enhanced with prompt engineering and chain-of-thought reasoning, can be used to provide a second opinion to complement human judgment. AutoML frameworks like AutoGluon enable predictive modelling through optimized feature engineering to ensure accurate and scalable assessments.

This paper presents a methodology integrating GenAI and AutoML to improve PSC evaluation consistency. By combining human expertise with AI-driven analysis, this approach reduces bias, enhances reliability, and supports the maritime industry's need for fair, objective, and scalable inspection outcomes.

3. Methodology

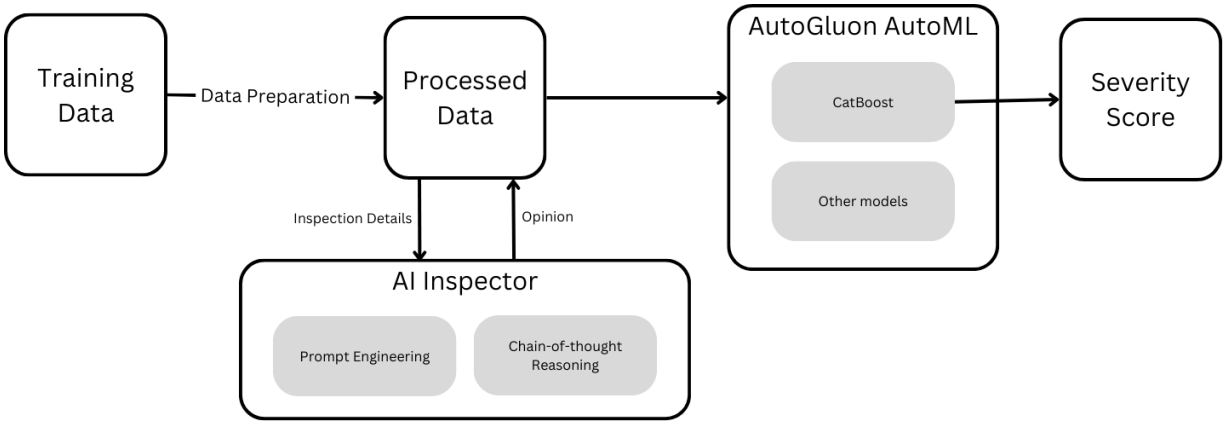

Figure 1: Methodology overview.

3.1 Data Preparation

To prepare the raw data ‘psc_severity_train’ for further processing, these manipulation techniques were used:

- Removal of irrelevant columns ‘annotation_id’, ‘username’ and ‘InspectionDate’.
- Extraction of 'PscInspectionId', 'Deficiency/Finding', 'Description Overview', 'Immediate Causes', 'Root Cause Analysis', 'Corrective Action', 'Preventive Action' and 'Deficiency Code' from ‘def_text’
- Encoding of nominal ‘annotated_severity’ to 0 (LOW), 1 (MED), and 2 (HIGH)

3.2 Augmenting prepared data with AI evaluation

Survey evaluations are inherently subjective, with multiple possible outcomes for a single finding. Biased data used in modelling will only produce biased, inaccurate predictions. To address this, we designed an LLM on **Google’s FLAN-T5 base** model to provide an additional severity opinion that is free from human bias. This model receives inspection report details (referenced in 3.1) as input parameters and returns a single integer corresponding to the suggested severity score. PSC inspection evaluation procedures are highly complex and follows various international guidelines and standards. To replicate this and reduce AI hallucination of irrational answers, **prompt engineering with zero-shot chain-of-thought reasoning** was implemented (Figure 2).

```
prompt = (
    "You are a Port State Control (PSC) inspector evaluating the severity of deficiencies. "
    "Based on the input parameters below, respond only with the severity rating: Low, Medium, or High. "
    "Do not include any additional text.\n\nThink step-by-step before providing the severity rating.\n\n"
    "Input Parameters:\n"
    f"- Deficiency Code: {row['Deficiency Code']}\n"
    f"- PSC Authority ID: {row['PscAuthorityId']}\n"
    f"- Port ID: {row['PortId']}\n"
    f"- Vessel Group: {row['VesselGroup']}\n"
    f"- Age: {row['age']}\n"
    f"- Deficiency/Finding: {row['Deficiency/Finding']}\n"
    f"- Description Overview: {row['Description Overview']}\n"
    f"- Immediate Causes: {row['Immediate Causes']}\n"
    f"- Root Cause Analysis: {row['Root Cause Analysis']}\n"
    f"- Corrective Action: {row['Corrective Action']}\n"
    f"- Preventive Action: {row['Preventive Action']}\n"
    f"- Detainable Deficiency: {row['Detainable Deficiency']}\n\n"
    "Severity Rating:"
)
```

Figure 2: Prompt used to guide LLM severity score generation.

By including AI severity scores, this added 4,364 synthetic samples to the processed dataset. This process allowed for the creation of a more diverse, unbiased and representative dataset, improving the reliability of the subsequent ML model.

3.3 Model Training

The feature-engineered data was fed into **AutoGluon, an AutoML framework**, which automates the process of further feature engineering, model selection, and hyperparameter tuning. AutoGluon efficiently processes both text and numerical data, allowing the identification of the best-performing model for predicting severity levels. This step ensured that the model accurately captured patterns and relationships within the data. The best performing model was Catboost with a roc_auc_ovo score of 0.768 and accuracy of 0.640.

3.4 Test Set Prediction

Catboost was subsequently used to predict severity scores in the test set.

4. Rules and Assumptions

To ensure consistency in severity evaluation, the following rules and assumptions were adopted:

Mean Aggregation: Severity levels were aggregated using the mean of Subject Matter Expert (SME) and AI-provided scores, to be used as ground truth for ML modelling.

Synthetic Data Assumptions: AI-generated severity scores were assumed to be equivalent to that given by a SME.

Inspector Expertise: Inspectors are assumed to adhere to standard industry practices and guidelines and are equally qualified.

5. Results and Discussion

Integrating generative AI with AutoGluon produced a robust model capable of accurately predicting consensus severity. By generating **an additional 4,364 synthetic samples**, we expanded our evaluation dataset, enabling a more comprehensive analysis. This augmentation led to an **overall roc_auc score of 0.768,** demonstrating the effectiveness of diverse data in enhancing model performance. The use of synthetic data provided multiple perspectives, reinforcing the model's reliability and highlighting the importance of varied data sources in vessel evaluations.

6. Conclusion

By leveraging generative AI for unbiased data synthesis and AutoGluon for automated machine learning, we developed a **dependable framework for evaluating consensus severity**. This approach offers significant potential to improve objectivity and consistency in vessel surveys, thereby supporting the maritime industry's operational and commercial objectives.