ST02: Data Preparation Report

Kateryna Makarova

June 9, 2025

1. Introduction

The raw AIS dataset from Bremerhaven contains vessel movement data with various attributes. This report details the steps taken to clean, preprocess, and select features for machine learning models aimed at predicting trip durations. All decisions are documented to maintain transparency.

2. Data Exploration and Understanding

- **Initial Inspection:** Examined the dataset manually and using summary statistics to understand data types, missing values, and distributions.
- **Domain Research:** Investigated maritime terms such as SOG (Speed Over Ground), COG (Course Over Ground), TH (True Heading) to comprehend their relevance.
- **Visualization:** Created maps and plots to visualize ship trajectories and attribute distributions, aiding feature relevance assessment.

3. Data Cleaning and Preprocessing

- **Dropped Irrelevant Columns:** Removed columns that do not contribute to ETT prediction, including `Name`, `Callsign`, `AISSource`, `ID`, `StartPort`, `EndPort`, and `Destination`.
- **Datetime Conversion:** Converted date/time columns from string format to pandas datetime objects for easier manipulation and calculations.
- **Handling Missing Values:** Missing values were found in `Breadth`, `Draught`, and other fields. These were imputed with the mean values computed per ship type where applicable.
- **Outlier Detection:** Outliers in the `TH` field coded as 511 (indicating missing or invalid data) were removed to improve data quality.
- **Feature Engineering:** Added new columns:
  - **Distance (km):** Calculated using the Haversine formula between trip start and end points.

- **SOG_kmh:** Converted Speed Over Ground to km/h.
- **Simulated Trip Time (hours):** Estimated from distance and speed.
- **Trip Duration:** Computed as the difference between start and end timestamps.

4. Correlation Analysis

Feature	Before Cleaning	After Cleaning	Interpretation
SOG vs. ETT	Weak/unclear correlation	Moderate negative (-0.38)	Faster ships generally have shorter trip durations
COG vs. TH	Very strong (0.92)	Very strong (0.93)	High correlation reflects consistent heading behavior
Length vs. Breadth	Very strong correlation	Slightly reduced	Ship dimensions remain consistently correlated
Overall Matrix	Over-saturated	More selective	Cleaning removed noise and irrelevant data

5. Remaining Tasks

- Validate time calculations within trips lasting less than 24 hours to ensure accuracy.
- Investigate the role of initial time columns in prediction performance.
- Conduct further exploratory data analysis to refine feature selection.

6. Summary

The data preparation phase successfully cleaned and transformed the AIS data to produce a reliable dataset for machine learning. Feature engineering and correlation analysis helped identify key predictors like `SOG_kmh`, `Distance`, and ship dimensions. These processed data form the foundation for the modeling tasks ahead.