Q&A

Q1. Why were missing values in static variables (such as length, width, and draught) filled using the median grouped by MMSI instead of using the overall dataset median or mean?

Static ship characteristics such as length, width, and draught do not change for a given vessel, because they represent the ship's physical structure. Imputing these values using the median calculated within each MMSI group allows the imputed value to represent that specific ship's actual dimensions rather than a global average across all vessels. This ship-specific approach produces more accurate and realistic estimates because it preserves the true identity and scale of each vessel. In contrast, using the overall median or mean would blend together measurements from many different ship types and sizes, reducing precision and potentially introducing bias, especially for outliers such as unusually large or small vessels. The grouped-median method is therefore the most logical and reliable choice.

Q2. Feature importance analysis removed several variables, but MMSI was kept in the final Linear Regression model even though its p-value was 0.974. Should MMSI be removed, and what effect would this have on the model?

A p-value of 0.974 indicates that MMSI does not meaningfully contribute to predicting the target variable and behaves like statistical noise in the regression. Because of this, MMSI can be safely removed without harming the model's predictive performance. In practice, removing MMSI would mainly improve interpretability by eliminating an irrelevant identifier from the model. The change in performance would be negligible—typically close to zero—because MMSI does not carry predictive information once other navigational variables are included. Its presence in the model provides no benefit and may appear only because it was used earlier for grouping or preprocessing, not because it is a true predictor.

Q3. PCA results showed that a single principal component captures 90% of the variance. What does this reveal about the feature relationships, and why use five original features in the Linear Regression model instead of using one PCA component?

The fact that the first principal component explains almost all of the variance indicates that the dataset contains strong correlations among the features. Many variables move together, meaning most of the information lies along one dominant direction in feature space. This is evidence of multicollinearity, especially among navigation variables such as heading and course-related metrics.

Despite this, using a single PCA component would sacrifice interpretability because PCA mixes the original variables into combinations that are not physically meaningful. Linear Regression is still effective even in the presence of correlated predictors, and retaining the original features yields a model that is easy to understand and explain in real maritime terms. Since the regression already achieves excellent performance, dimensionality reduction is unnecessary, and preserving interpretability provides greater practical value than relying on abstract components.

Q4. Linear Regression shows nearly identical training and test R^2 values (~0.928), while a Decision Tree typically shows higher training R^2 but lower test R^2. Why does this happen, and what does it indicate about model generalization?

Linear Regression is a relatively simple model with limited flexibility, which prevents it from fitting noise in the training data. As a result, its performance on the training and test sets is almost identical, showing that it generalizes extremely well and does not overfit. The model maintains a good balance between bias and variance, leaning toward slightly higher bias but very low variance.

In contrast, a Decision Tree can grow complex decision boundaries that perfectly fit the training data, achieving high or even near-perfect training R^2. However, this flexibility causes it to capture noise and fine-grained patterns that do not repeat in new data, leading to a drop in test performance. This behavior reflects the high-variance nature of Decision Trees. The comparison highlights why the linear model is more reliable for this particular task: it remains stable and consistent across unseen data.

Q5. The ‘heading’ variable has a very large regression coefficient compared to ‘sog’. What does this mean in maritime navigation, and why is this intuitive for predicting Course Over Ground (COG)?

Heading represents the compass direction in which the ship is pointed, and it is naturally the dominant factor in determining the vessel’s Course Over Ground. The large coefficient reflects this strong physical relationship: the direction a ship faces largely determines the direction it moves, except for deviations caused by external forces like wind or currents. Because of this, even small changes in heading can result in substantial changes in the predicted COG, which the model captures through its high coefficient value.

Speed over ground (SOG) influences the vessel’s movement but primarily affects the magnitude of velocity rather than the direction. Its smaller coefficient shows that speed contributes less directly to changes in COG compared to heading. This aligns with maritime intuition: a ship moving faster does not significantly alter its course direction, but a shift in heading immediately influences the ship’s trajectory. Thus, the regression coefficients accurately reflect real-world navigation dynamics.