**Data description**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Type** | **Notes** |
| CurrentSpeed | Speed | Continuous |  |
| CurrentDir | Direction | Continuous |  |
| TWS | True Wind Speed | Continuous |  |
| TWA | True Wind Angle | Continuous |  |
| AWS | Apparent Wind Speed | Continuous |  |
| AWA | Apparent Wind Angle | Continuous |  |
| Roll | Coronal rotation | Continuous | = -Heel |
| Pitch | Sagittal rotation | Continuous |  |
| HeadingMag | Magnetic heading | Continuous |  |
| HoG | Heading Over Ground | Continuous | TH – HOG = Yaw |
| HeadingTrue | True heading | Continuous | TH – HOG = Yaw |
| AirTemp | Air temperature | Continuous |  |
| Longitude | Coord long | Continuous |  |
| LongDecMin | Min of long (1/60) | Continuous | Not in dataset, in long/lat columns? |
| Latitude | Coord lat | Continuous |  |
| LatDecMin | Min of lat (1/60) | Continuous | Not in dataset |
| SoG | Speed Over Ground | Continuous |  |
| SoS | Speed over Surface | Continuous |  |
| AvgSoS | Avg SoS | Continuous |  |
| VMG | Velocity Made Good | Continuous | SoS \* cos(TWA) |
| RudderAng | Rudder Angle | Continuous |  |
| Leeway | Drift sideways | Continuous | Ask about exact measurement |
| TWD | True Wind Direction | Continuous |  |
| WSoG | Wind Speed over Ground | Continuous |  |
| VoltageDrawn | Voltage drawn by the system of one of its parts | Continuous | What does this mean exactly? |
| ModePilote | ?? | ?? | No description, get clarification. Takes 2 and 5 |
| DateTime | Time | Continuous | Interpolate NaN |
| Yaw | Axial rotation | Continuous | Can use instead of TH-HOG |
| Tacking | If tacking is done? | Categorical |  |

**Overall notes:**

* Difference in measuring angles and direction in 0 to 360 and -180 to 180?
* SoS and AvgSoS?
* ModePilote?
* VoltageDrawn?
* For prediction, how long in advance is it necessary to know to tack
  + I.e. what lag?
* All from one boat and one trip?
* 220,000 rows (max 168 missing)
  + Interpolation
* 27 columns
* LongDecMin and LatDecMin not in DF
* Yaw and Tacking added
* Row 200,000 to 220,000 is repeat of last 20,000 rows (180-200k)
* Extraneous variables:
  + Speed over Surface or ground?
  + Tides
  + Weather
  + Type of vessel
  + Other vessels or obstacles
  + How to tack?
    - Angle, speed, details?
  + Other maneuvers?

**Approach:**

* Load data and inspect
* Remove NAs
  + Datetime and categorical variables take previous
  + Others interpolate with -5 to +5 average
* Check stationarity
  + ADF, KPSS
  + Longitude & latitude non-stationary on both
  + All others non-stationary on KPSS
  + Differencing with 1 shift
  + Drop NA
* Z-score numeric variables
* PCA for dimensionality reduction in predictive model
  + Pearson correlations & chi-sq for factor importance
* Interpretative model for understanding what is important to measure
  + Logistic regression
  + SHAP values and XGBoost with factor importance
* Useful model for prediction
  + Previous lags
    - 5, 10, 15, 20
    - 10, 20, 30, 40
    - 15, 30, 45, 60
  + LSTM super good at prediction
  + XGBoost more efficient?