**Calcofri Data Set Documentation**

|  |  |
| --- | --- |
| **S.No** | **Topics** |
| **1.** | **Introduction of Algorithm** |
| **2.** | **About Data Sets** |
| **3.** | **Explanation of Target variable & Feature Variable** |
| **4.** | **performance metrics for evaluating linear regression models:**   * **R Square error** * **Mean Square error** |
|  |  |

**Linear Regression explanation with the help of Calcofri data set**

* Linear regression is like drawing a straight line through a scatterplot to understand and predict how one thing is related to another. In this case, it helps us figure out how the "independent variable" (one thing) affects the "dependent variable" (another thing).
* linear regression is about finding the best straight line through data points to understand and predict how one variable (like temperature) affects another variable (like density).

For eg:

Suppose we have a dataset that contains information about the temperature (independent variable) of water and the density (dependent variable) of that water. We want to know how changes in temperature affect the density of water.

* We start by plotting the data points on a graph with temperature on the x-axis and density on the y-axis. Each point represents a temperature-density pair.
* Linear regression helps us find the straight line that best fits these points. This line represents the relationship between temperature and density.
* When we have this line, we can make predictions. For example, if we know the temperature of water, we can use the line to predict the density. This can be very useful in various real-world applications, like predicting how dense water will be at a certain temperature.

**About Calcofri data set**

This database contains oceanographic data measured from seawater samples collected at [CalCOFI stations](https://wp.calcofi.org/wp/data/sampling-info/station-positions/). Oceanographic data (i.e. the physical features of seawater) includes parameters such as temperature, salinity, dissolved oxygen, chlorophyll-a, nutrients, and many more.

It contains various features related to the properties of water, primarily from bottle samples taken from the ocean. These features can be used for research and analysis in oceanography and related fields.

The **Bottle**table contains oceanographic data. This table includes oceanographic measurements for each bottle/sampling depth ever completed on a CalCOFI cruise. There are additional data code and precision columns describing the quality of each oceanographic measurement. Each row is a unique bottle/sampling depth, numbered sequentially/indexed by the "Btl\_Cnt" column.

* Oceanographic Measurements: This dataset contains information related to water properties, such as temperature, salinity, oxygen levels, and nutrient concentrations, recorded at different depths in the ocean.

**Purpose of this data set:**

* The dataset is used for research and analysis in oceanography, environmental science, and related fields. Scientists and researchers may use this data to study and understand various aspects of ocean properties and dynamics, such as the impact of temperature and salinity on marine ecosystems.
* Researchers can perform various analyses on this dataset, but today, I will explain how to use linear regression on this dataset.

**Explanation of Target variable & Feature Variable**

1. **Target Variable (Dependent Variable):**

The target variable, also known as the dependent variable, is the main focus of your analysis. It's the variable you want to predict or explain using the other variables in the dataset.

In the "bottle.csv" dataset, the target variable could be a specific property or measurement that you're interested in understanding or predicting. For example, you might want to predict the water's density, salinity, or oxygen concentration. This is what you're trying to model or explain using the features.

**2. Feature Variables (Independent Variables):**

Feature variables, also known as independent variables or predictors, are the other variables in the dataset that you use to predict or explain the target variable.

In the "bottle.csv" dataset, feature variables could be various measurements and characteristics of the water, such as temperature, pH, nutrient levels, and geographic coordinates (latitude and longitude). These features are used to help you understand and make predictions about the target variable.

**performance metrics for evaluating linear regression model**

**R-squared (R^2):**

R-squared, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable ("Salnty") that is predictable from the independent variables ("Depthm" and "T\_degC") in the model.

R-squared values range from 0 to 1.

* An R-squared value of 0 indicates that the model explains none of the variance in the target variable.
* An R-squared value of 1 indicates that the model perfectly explains the variance in the target variable.

In the context of your linear regression model, a high R-squared value suggests that the "Depthm" and "T\_degC" features explain a significant portion of the variance in "Salnty." However, a low R-squared value implies that these features do not explain much of the variance, and the model might not be a good fit for the data.

**Mean Squared Error (MSE):**

Mean Squared Error is a measure of the average squared difference between the actual values ("Salnty") and the predicted values from the linear regression model.

MSE is calculated as the sum of the squared differences between the actual and predicted values, divided by the number of data points.

A lower MSE indicates that the model's predictions are closer to the actual values, meaning it's a better fit for the data. On the other hand, a higher MSE indicates larger prediction errors, suggesting a less accurate model.

**Note: Now I’m using the target value as Salnty & featured value as Depthm & T\_ degc. And the value of R Sq error and MSE are R-squared: 0.36636919390784906**

**Mean Squared Error: 0.1349823150458944**

In summary, when evaluating a linear regression model with "Salnty" as the target variable and "Depthm" and "T\_degC" as the feature variables:

* A high R-squared value indicates that the model is explaining a significant portion of the variance in "Salnty," which is a positive sign.
* A low MSE indicates that the model's predictions are close to the actual "Salnty" values, which is desirable.