**Machine Learning with Python**

**Week 1 – Introduction to Machine Learning**

Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.

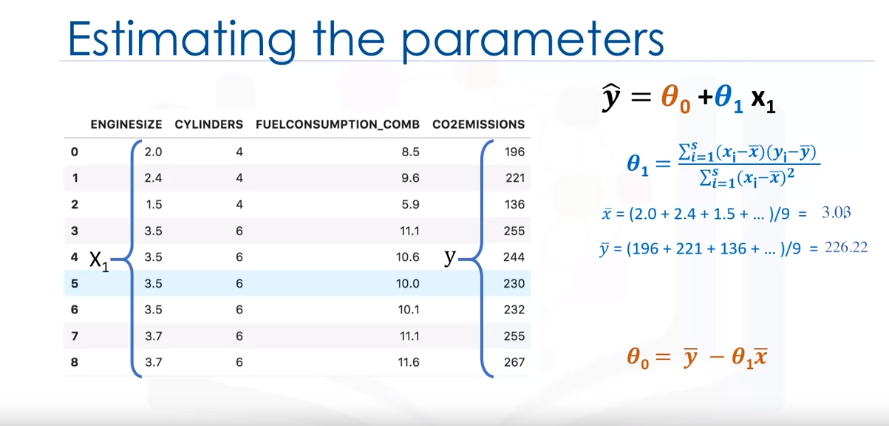
**Week 2 – Introduction to Machine Learning**

**Simple Linear Regression:**

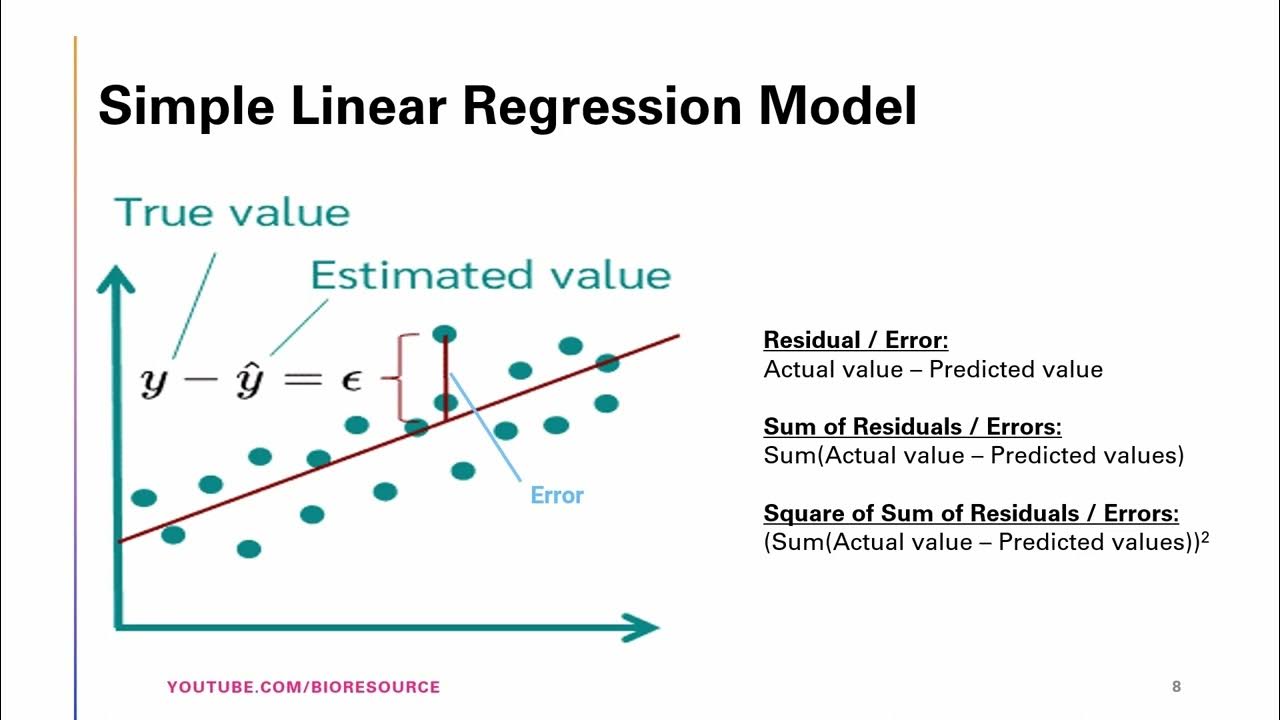
**Yhat = a + bx**

**Residual Error = Yactual - YPredicted**

The mean of all residual errors shows how poorly the line fits with the whole data set. Mathematically it can be shown by the equation Mean Squared Error, shown as MSE. Our objective is to find a line where the meaning of all these errors is minimized. In other words, the mean error of the prediction using the fit line should be minimized. Let's reword it more technically. The objective of linear regression is to minimize this MSE equation and to minimize it, we should find the best parameters; a and b.



As mentioned before, **Coefficient** and **Intercept** in the simple linear regression, are the parameters of the fit line. Given that it is a simple linear regression, with only 2 parameters, and knowing that the parameters are the intercept and slope of the line, sklearn can estimate them directly from our data. Notice that all of the data must be available to traverse and calculate the parameters

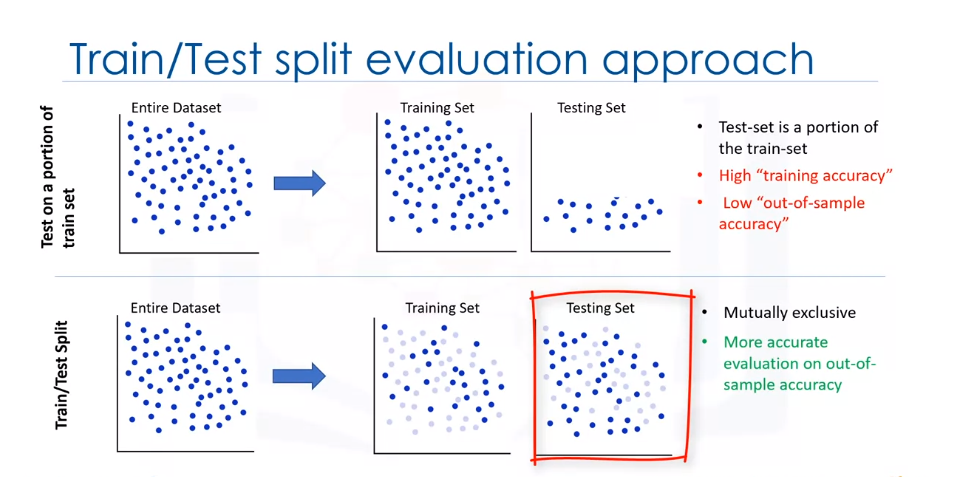


**Training Accuracy:**

Training accuracy is the percentage of correct predictions that the model makes when using the test dataset. However, a high training accuracy isn't necessarily a good thing. For instance, having a high training accuracy may result in an over-fit the data. This means that the model is overly trained to the dataset, which may capture noise and produce a non-generalized model.

**Out-of-sample Accuracy:**

Out-of-sample accuracy is the percentage of correct predictions that the model makes on data that the model has not been trained on. Doing a train and test on the same dataset will most likely have low out-of-sample accuracy due to the likelihood of being over-fit. It's important that our models have high out-of-sample accuracy because the purpose of our model is, of course, to make correct predictions on unknown data.



**Creating train and test dataset**

Train/Test Split involves splitting the dataset into training and testing sets that are mutually exclusive. After which, you train with the training set and test with the testing set. This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that have been used to train the model. Therefore, it gives us a better understanding of how well our model generalizes on new data.

This means that we know the outcome of each data point in the testing dataset, making it great to test with! Since this data has not been used to train the model, the model has no knowledge of the outcome of these data points. So, in essence, it is truly an out-of-sample testing.

Let's split our dataset into train and test sets. 80% of the entire dataset will be used for training and 20% for testing. We create a mask to select random rows using **np.random.rand()** function:

**Evaluation Metrics in Regression Models:**

Error: It is the measure of how far the data point of Y is from the fitted regression line (Yhat).

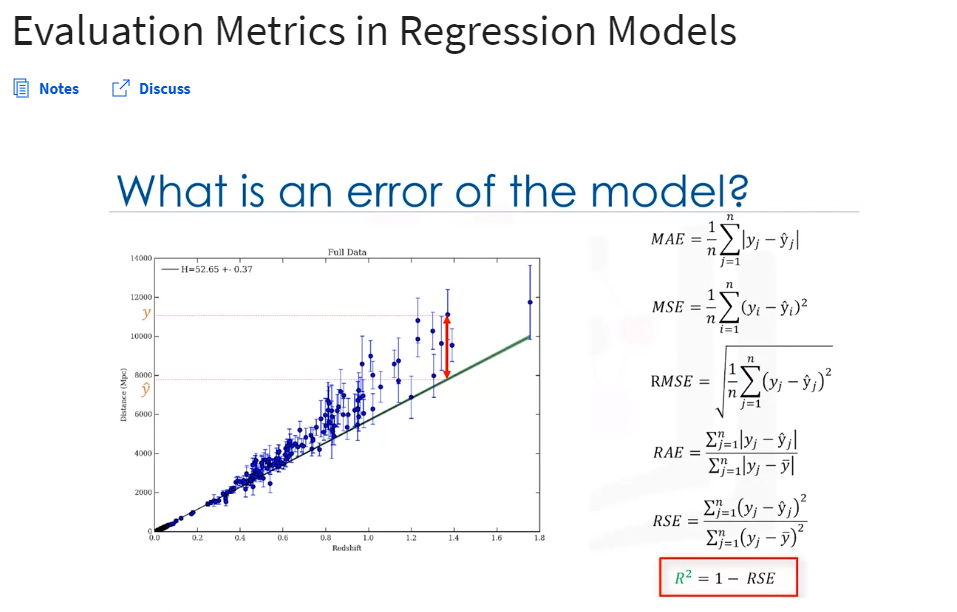
We compare the actual values and predicted values to calculate the accuracy of a regression model. Evaluation metrics provide a key role in the development of a model, as it provides insight to areas that require improvement.

There are different model evaluation metrics, lets use MSE here to calculate the accuracy of our model based on the test set:

* Mean Absolute Error: It is the mean of the absolute value of the errors. This is the easiest of the metrics to understand since it’s just average error.
* Mean Squared Error (MSE): Mean Squared Error (MSE) is the mean of the squared error. It’s more popular than Mean Absolute Error because the focus is geared more towards large errors. This is due to the squared term exponentially increasing larger errors in comparison to smaller ones.
* Root Mean Squared Error (RMSE).
* R-squared is not an error, but rather a popular metric to measure the performance of your regression model. It represents how close the data points are to the fitted regression line. The higher the R-squared value, the better the model fits your data. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse).

MAE(Mean Absolute Error):

MSE(Mean Squared Error):

RMSE(Root Mean Squared Error):

R-Squared:

Adjusted R-squared:

RAE(Relative Absolute Error):

RSE(Relative Squared Error):