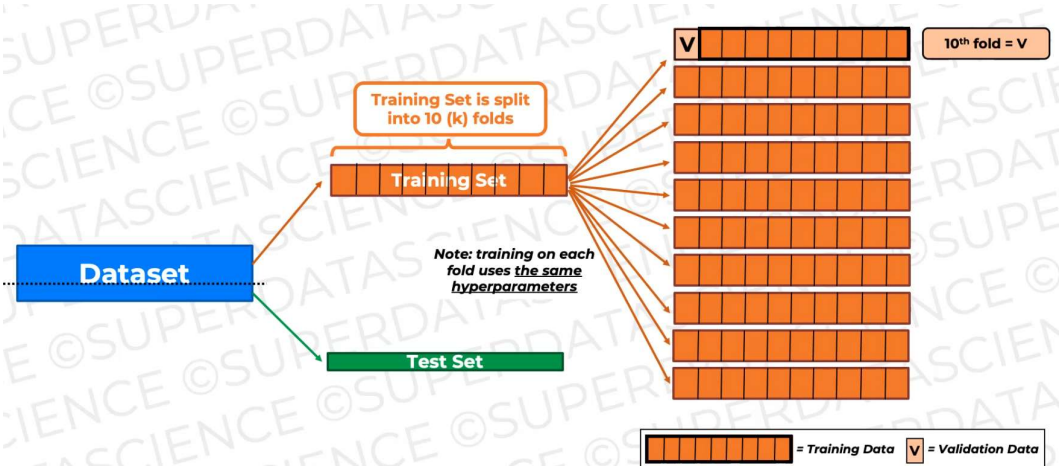
Model Selection

Monday, December 30, 2024 12:05 PM

K_Fold Cross Validation

- Break the dataset into a training and test set
- Split the training set into 'k' parts
- 1st fold is the validation data, while rest is training data
- Not 2nd is validation data, while rest is training data



- We will get k number sets of metrics
- The accuracy is much more reliable, because it is harder to get lucky 'k' times

Now that the aggregate metrics look good, then we train the model on the entire training set, then proceed to test on the test set as usual

- If the aggregate metrics don't look good, then we need to change something about the model or parameters

Bias-Variance Tradeoff

Bias - A systematic error that occurs in the machine learning model itself due to incorrect assumptions in the Machine Learning Process.

- Technically, we can define bias as the error between average model prediction and the ground truth

Variance - How much the model can adjust depending on the given data set

- Variance refers to the changes in the model when using different portions of the training set

If a model had **High Bias and Low Variance**

- Then the model is too simple and does not capture the underlying trend of the data

If the model has **Low Bias and High Variance**

- Then the model is too sensitive and is capturing noise as if it were a real trend (overfitting)

If the model has **High Bias and High Variance**

- Then the model is too simple to capture the data's trends and too sensitive, capturing noise as well

If the model has **Low Bias and Low Variance**

- Ideal Model

- Accurately captures the underlying trends of the data and generalizes well to unseen data