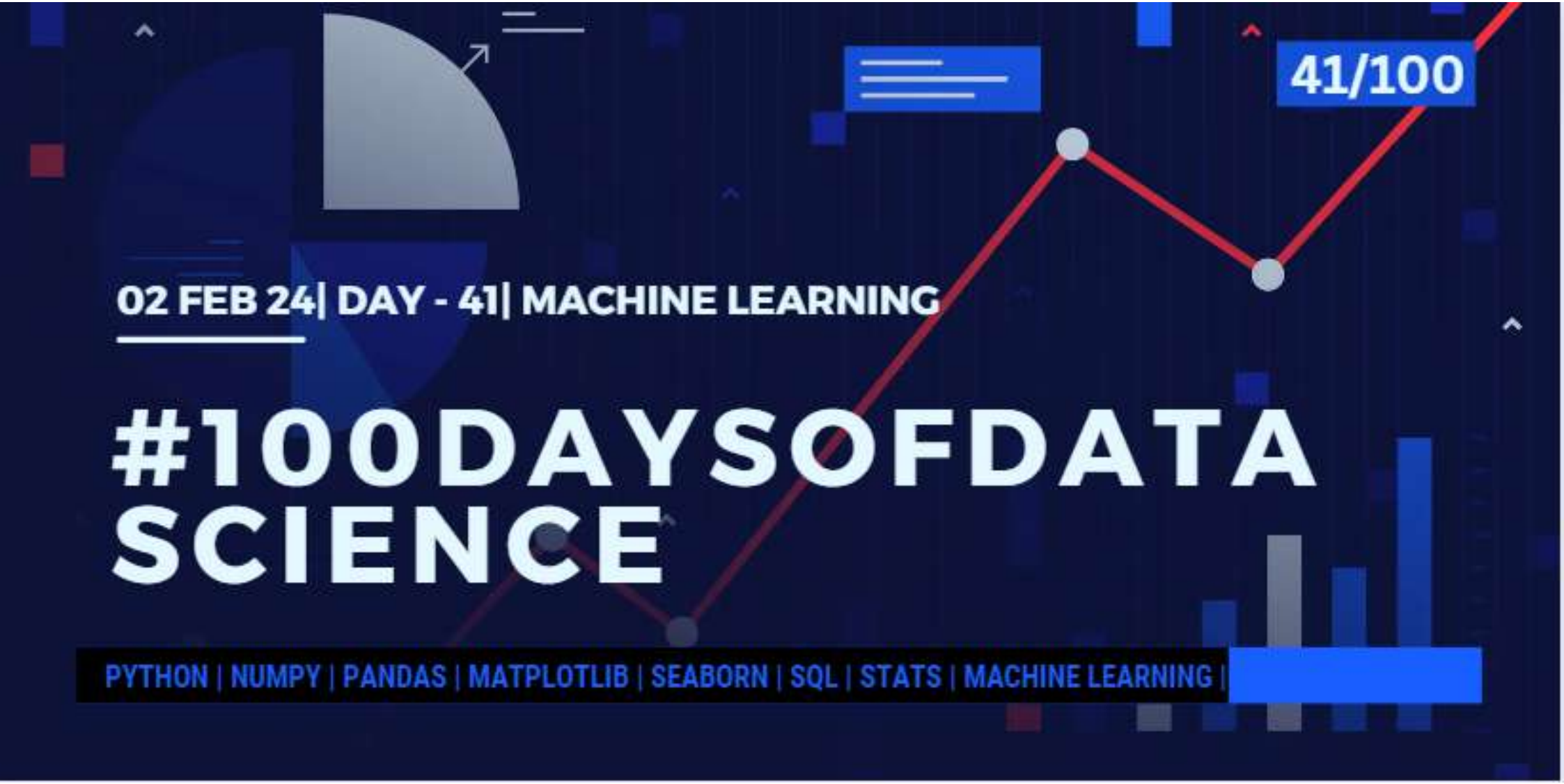

41/100
02 FEB 24| DAY - 41| MACHINE LEARNING
#100DAYSOFDATA SCIENCE
PYTHON | NUMPY | PANDAS | MATPLOTLIB | SEABORN | SQL | STATS | MACHINE LEARNING |

Overfitting & Underfitting

Overfitting and Underfitting are the two main problems that occur in machine learning and degrade the performance of the machine learning models.

The main goal of each machine learning model is to generalize well. Here generalization defines the ability of an ML model to provide a suitable output by adapting the given set of unknown input. It means after providing training on the dataset, it can produce reliable and accurate output.

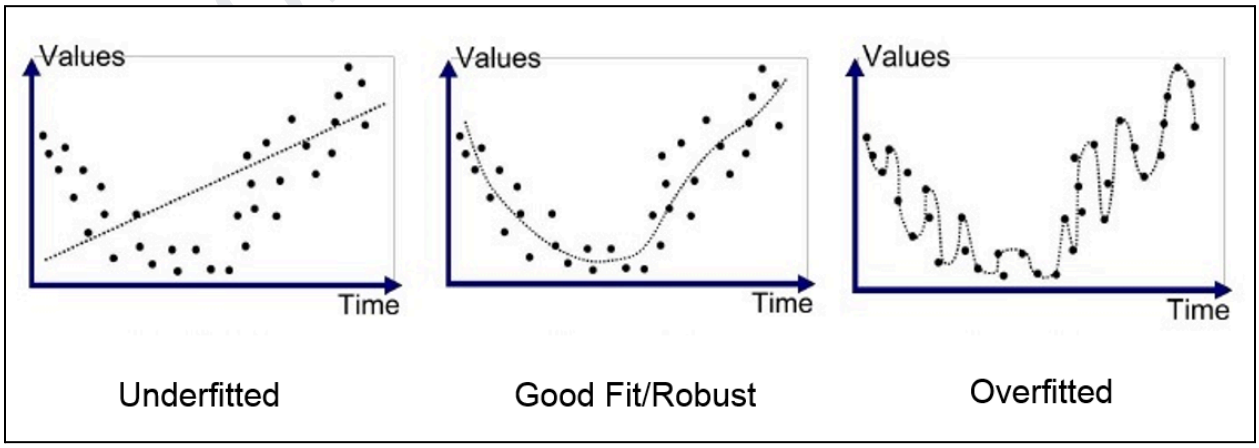
Hence, the underfitting and overfitting are the two terms that need to be checked for the performance of the model and whether the model is generalizing well or not.

Signal:
It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.

Noise:
Noise is unnecessary and irrelevant data that reduces the performance of the model.

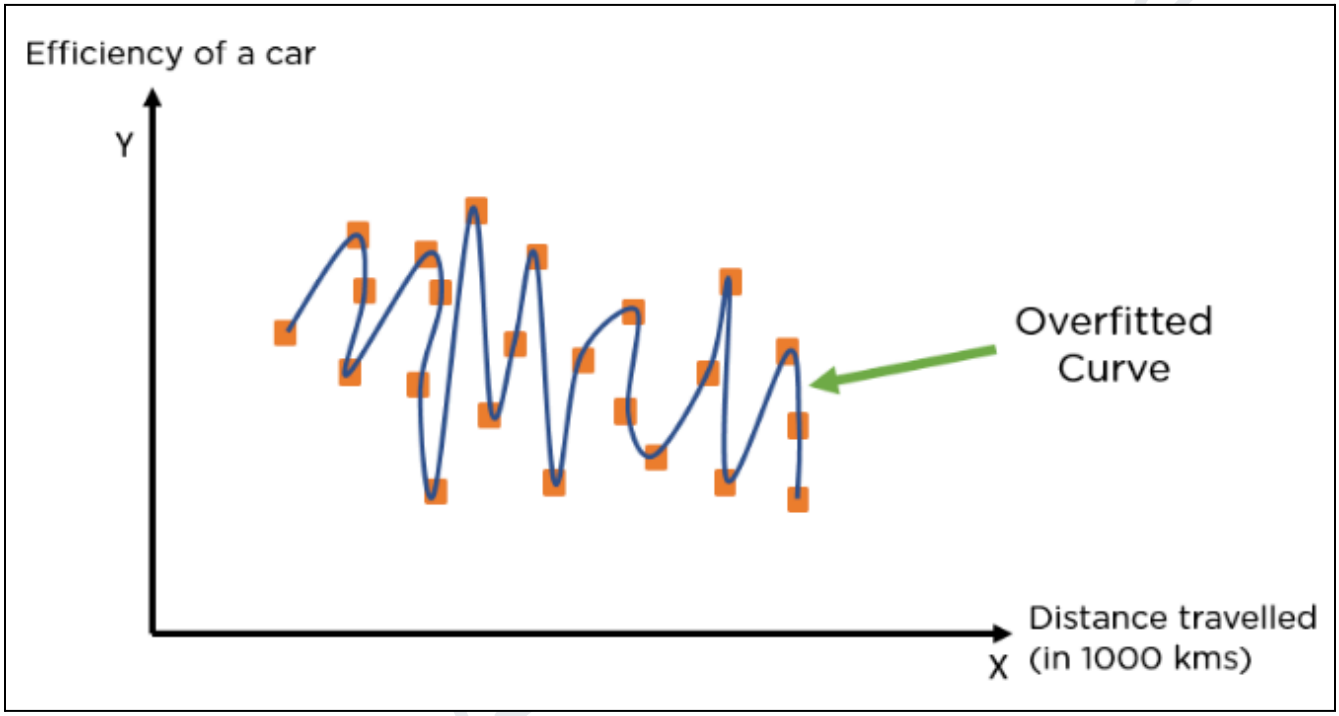
Bias:
Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.

Variance:
If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.



1. Overfitting

When a model performs very well for training data but has poor performance with test data (new data), it is known as overfitting. In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data. Overfitting can happen due to low bias and high variance.

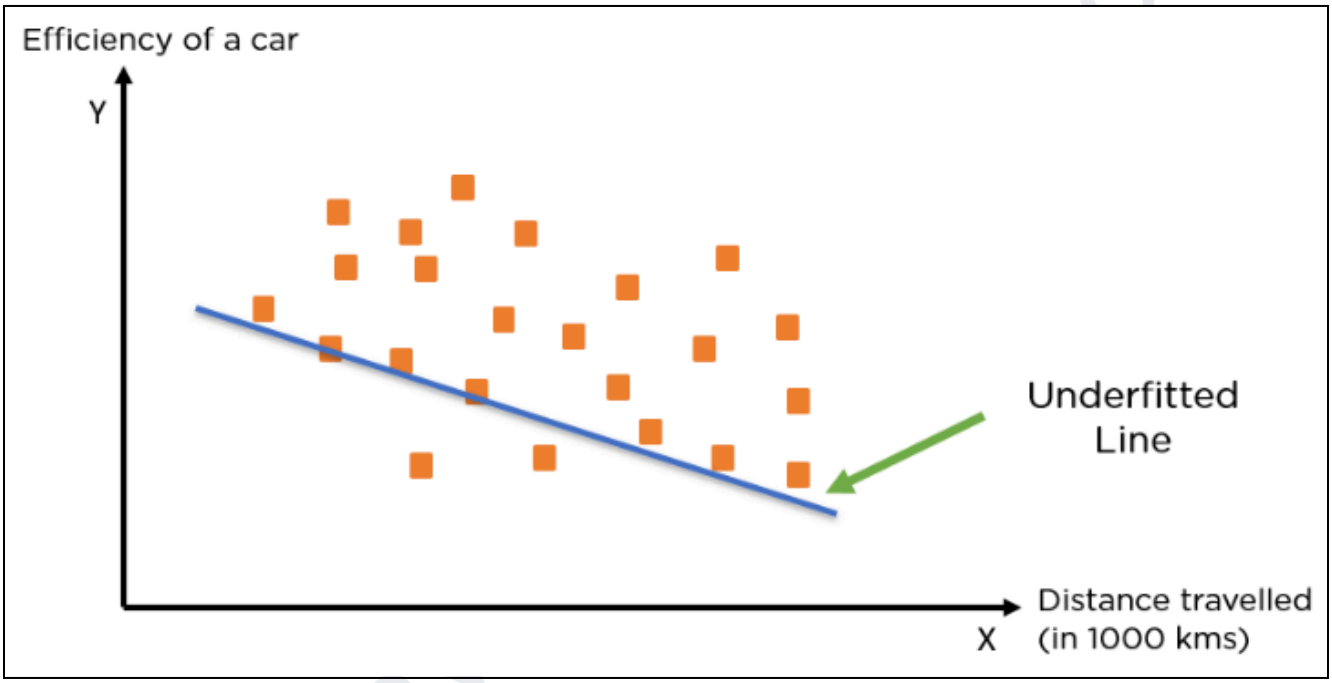


- Reasons for Overfitting
    1. Data used for training is not cleaned and contains noise (garbage values) in it
    2. The model has a high variance
    3. The size of the training dataset used is not enough
    4. The model is too complex

- Ways to Tackle Overfitting
    1. Using K-fold cross-validation
    2. Using Regularization techniques such as Lasso and Ridge
    3. Training model with sufficient data
    4. Adopting ensembling techniques

2. Underfitting

When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance.



- Reasons for Underfitting
    1. Data used for training is not cleaned and contains noise (garbage values) in it
    2. The model has a high bias
    3. The size of the training dataset used is not enough
    4. The model is too simple

- Ways to Tackle Underfitting
    1. Increase the number of features in the dataset
    2. Increase model complexity
    3. Reduce noise in the data
    4. Increase the duration of training the data

Goodness of Fit

The "Goodness of fit" term is taken from the statistics, and the goal of the machine learning models to achieve the goodness of fit. In statistics modeling, it defines how closely the result or predicted values match the true values of the dataset.

The model with a good fit is between the underfitted and overfitted model, and ideally, it makes predictions with 0 errors, but in practice, it is difficult to achieve it.

As when we train our model for a time, the errors in the training data go down, and the same happens with test data. But if we train the model for a long duration, then the performance of the model may decrease due to the overfitting, as the model also learn the noise present in the dataset. The errors in the test dataset start increasing, so the point, just before the raising of errors, is the good point, and we can stop here for achieving a good model.