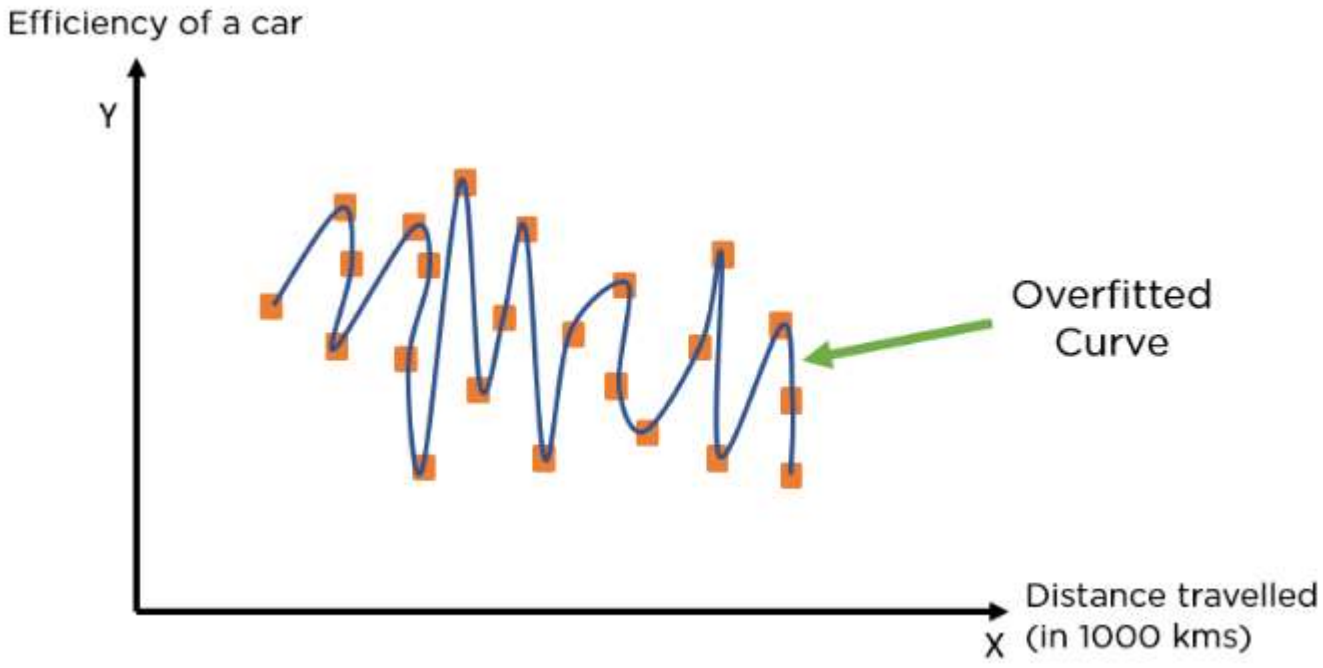
Module 1: overfitting and underfitting of model

Overfitting and Underfitting are two crucial concepts in machine learning and are the prevalent causes for the poor performance of a machine learning model.

What is 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

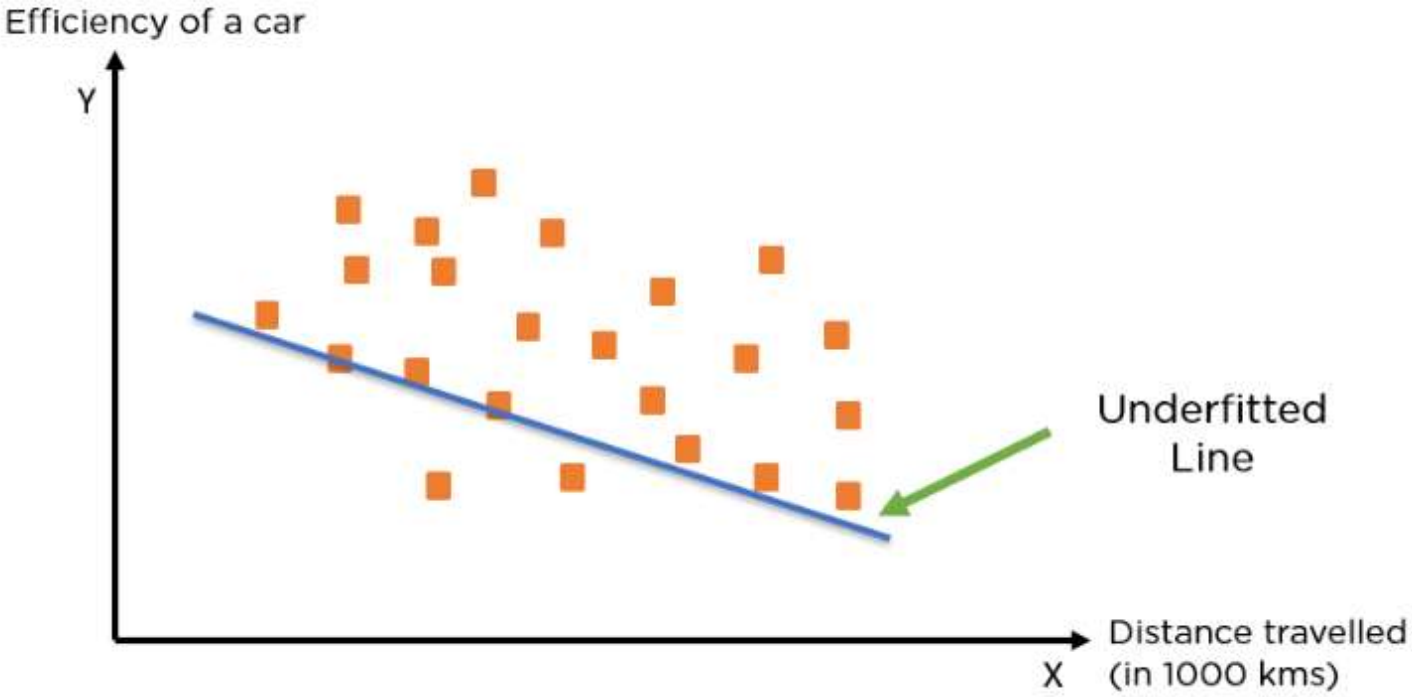
- Data used for training is not cleaned and contains noise (garbage values) in it
- The model has a high variance
- The size of the training dataset used is not enough
- The model is too complex

Ways to Tackle Overfitting

- Using K-fold cross-validation
- Using Regularization techniques such as Lasso and Ridge
- Training model with sufficient data
- Adopting ensembling techniques

What is 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

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

Ways to Tackle Underfitting

- Increase the number of features in the dataset
- Increase model complexity
- Reduce noise in the data
- Increase the duration of training the data

Now that you have understood what overfitting and underfitting are, let's see what is a good fit model in this tutorial on overfitting and underfitting in machine learning.

What Is a Good Fit In Machine Learning?

To find the good fit model, you need to look at the performance of a machine learning model over time with the training data. As the algorithm learns over time, the error for the model on the training data reduces, as well as the error on the test dataset. If you train the model for too long, the model may learn the unnecessary details and the noise in the training set and hence lead to overfitting. In order to achieve a good fit, you need to stop training at a point where the error starts to increase.