Dataset of Car Specifications

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Engine Size (cc) | Horsepower | Age of Car (years) | Mileage (kmpl) | Price (INR) |
| 1500 | 100 | 3 | 18 | 600000 |
| 2000 | 150 | 5 | 15 | 850000 |
| 1200 | 80 | 2 | 20 | 500000 |
| 1800 | 130 | 7 | 17 | 750000 |
| 1400 | 90 | 4 | 19 | 550000 |

# Terminology Definitions

## Feature

Individual measurable properties or characteristics used as inputs to the model. In the example, features include 'Engine Size (cc)', 'Horsepower', 'Age of Car (years)', and 'Mileage (kmpl)'.

## Label

The output variable that the model aims to predict. In the example, the label is 'Price (INR)'.

## Prediction

The output value generated by the model when given a set of input features. For example, predicting the price of a car given its specifications.

## Outlier

A data point that deviates significantly from the rest of the dataset. If a car in this dataset had a price of 3000000 INR, it would be considered an outlier.

## Test Data

The subset of the dataset used to assess the performance of the model. For instance, if we use 20% of our data for testing, one of these rows could be part of the test data.

## Training Data

The subset of the dataset used to train the model. For example, the remaining 80% of the dataset used to teach the model.

## Model

The mathematical representation of the relationship between features and labels. In our case, a regression model predicting car prices.

## Validation Data

A separate subset used to fine-tune the model parameters. This data is not used in training but helps in adjusting hyperparameters.

## Hyperparameter

These are parameters that are not learned from the data but set before training the model. Examples include learning rate, number of epochs, etc.

## Epoch

One complete pass through the entire training dataset. For instance, training on all car data once is one epoch.

## Loss Function

A method to measure how well the model's predictions match the actual labels. For instance, Mean Squared Error (MSE) can be used for our regression model.

## Learning Rate

A hyperparameter that controls how much the model's weights are updated during training. A higher learning rate might make the training faster but less accurate.

## Overfitting

When a model performs well on training data but poorly on test data. For example, if our model perfectly predicts the training data but fails on new data.

## Underfitting

When a model is too simple to capture the underlying pattern of the data. For instance, a model that always predicts the average car price.

## Regularization

Techniques to prevent overfitting by adding a penalty to the loss function. Examples include L1 and L2 regularization.

## Cross-Validation

A method to assess the model’s performance by partitioning the data into subsets, training the model on some subsets, and validating it on others.

## Feature Engineering

The process of creating new features or modifying existing ones to improve model performance. For example, creating a 'Price per cc' feature.

## Dimensionality Reduction

Techniques to reduce the number of features in the dataset. Principal Component Analysis (PCA) is a common method.

## Bias

The error introduced by approximating a real-world problem, which might be complex, by a simpler model.

## Variance

The error introduced due to the model's sensitivity to small fluctuations in the training set.