Assignment 1: Data set of housing prices

Data set of housing prices to explain various machine learning terminologies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Area (sq ft) | Bedrooms | Bathrooms | Age of House (years) | Price (INR) |
| 1200 | 2 | 1 | 5 | 60,00,000 |
| 1500 | 3 | 2 | 10 | 75,00,000 |
| 900 | 2 | 1 | 3 | 45,00,000 |
| 1800 | 4 | 3 | 15 | 1,00,00,000 |
| 1100 | 2 | 1 | 8 | 55,00,000 |

# Feature

Individual measurable properties or characteristics used as inputs to the model. In the example, features include 'Area (sq ft)', 'Bedrooms', 'Bathrooms', and 'Age of House (years)'.

# 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 house given its features.

# Outlier

A data point that deviates significantly from the rest of the dataset. If a house in this dataset had a price of 5,00,00,000 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 house 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 housing 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 house 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 sq ft' 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.