A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**AI/ Machine Learning Internship Program**

**Name:** Gopika K G

**Email:** [gopisanju2004@gmail.com](mailto:gopisanju2004@gmail.com)

**Assignment I**

**Test data:**

Sampe test data for predicting house prices based on different features.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| House ID | Location | Size (sq ft) | Bedrooms | Bathrooms | Age (years) | Price (INR) |
| 1 | Chennai | 2000 | 3 | 2 | 10 | 8000000 |
| 2 | Bangalore | 1500 | 2 | 1 | 5 | 6000000 |
| 3 | Hyderabad | 1800 | 3 | 2 | 8 | 7500000 |
| 4 | Mumbai | 2500 | 4 | 3 | 12 | 15000000 |
| 5 | Pune | 1200 | 2 | 1 | 3 | 5500000 |
| 6 | Delhi | 3000 | 5 | 4 | 15 | 20000000 |

**Terminologies with respect to the dataset:**

1. **Feature:**

Individual measurable properties or characteristics used as input to the model.

**Example:** Location, Size (sq ft), Bedrooms, Bathrooms, Age (years).

1. **Label:**

The output variable that the model aims to predict.

**Example**: Price (INR).

1. **Prediction:**

The output that the model generates after learning from the data.

**Example**: Predicted house price for a given set of features.

1. **Outlier:**

 A data point that deviates significantly from other data points in a dataset.

**Example:** Price of 20000000 INR for a 15-year-old house in Delhi, which is significantly higher than other values.

1. **Test Data:**

A subset of the data used to evaluate the performance of the trained model.

**Example:** A portion of the house data that is held out and used for testing the model's predictions i.e., house data that is not used for training the model but is used to test its efficiency.

1. **Training data:**

The dataset used to train the model.

**Example:** The above table without any held-out data for testing.

1. **Model:**

A mathematical representation or algorithm used to make predictions based on the input features.

**Example:** A regression model predicting house prices based on house features.

1. **Validation Data:**

The validation data set (subset of the data) provides an unbiased evaluation of a model fit on the training data set while tuning the hyperparameters.

**Example:** A portion of the data used to validate the model during training.

1. **Hyperparameter:**

Parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

**Example:** Learning rate, Batch size and Number of Epochs.

1. **Epoch:**

One complete pass through the entire training dataset during the learning process.

**Example:** Iterating through the entire house dataset once.

1. **Loss Function:**

A function that measures how well the model's predictions match the actual values.

**Example:** Mean Squared Error (MSE) between the predicted and actual house prices.

1. **Learning Rate:**

The learning rate is the hyperparameter in optimization algorithms that controls how much the model needs to change in response to the estimated error for each time when the model's weights are updated.

**Example:** If the learning rate is very less, then it may slow down the training process. On the other hand, if the learning rate is too large, then it may not optimize the model properly.

1. **Overfitting:**

When a model learns the training data too well, including noise, and performs poorly on new data.

**Example:** The model predicts house prices perfectly on training data but poorly on test data.

1. **Underfitting:**

When a model is too simple to capture the underlying patterns in the data.

**Example:** The model does not predict house prices well even on the training data.

1. **Regularization:**

Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.

**Example:** L2 regularization (Ridge Regression) applied to the model.

1. **Cross-Validation:**

A technique used to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds.

**Example:** 5-fold cross-validation splitting the data into 5 parts and training/testing the model 5 times.

1. **Feature Engineering:**

The process of creating new features or modifying existing ones to improve model performance.

**Example:** Creating a new feature like 'Price per sq ft' from 'Size (sq ft)' and 'Price (INR)'.

1. **Dimensionality Reduction:**

Techniques used to reduce the number of input features.

**Example:** Principal Component Analysis (PCA) to reduce the number of features in the dataset.

1. **Bias:**

Bias refers to the errors which occur when we try to fit the statistical model on real-world data which does not fit perfectly well on some mathematical model.

**Example:** A model consistently predicting lower house prices for all locations.

1. **Variance:**

The error value that occurs when we try to make predictions by using data that is not previously seen by the model.

**Example:** A model's house price predictions varying greatly with slight changes in training data.