 

**Joint Tech Internship Community Program**

**Assignment 1**

By

Mathav Ramalingam

Example Training Data Set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Built Year | Location | Size  (sq ft) | Rooms | Garage | Price (INR) |
| 2010 | Yercaud | 2000 | 5 | Yes | 7500000 |
| 2005 | Karur | 1500 | 3 | Yes | 5000000 |
| 2015 | Ooty | 2500 | 6 | Yes | 10000000 |
| 2008 | Coimbatore | 1800 | 5 | No | 6500000 |
| 2020 | Coimbatore Outskirt | 3000 | 8 | Yes | 12000000 |

List of Terminologies:

* Feature - An individual measurable property or characteristic of a phenomenon (**Eg**: Year, Location, Size, Rooms, Garage)
* Label - The output variable that the model is predicting (**Eg**: Price)
* Prediction - The value that a model outputs after processing the input features (**Eg**: Predicted house price based on its features)
* **Outlier -** An outlier is a data point that significantly deviates from the rest of the data. A data point that differs significantly from other observations (**Eg**: If one of the house had an extremely low or high price compared to others)
* **Test Data -** Test data refers to a dataset that is independent of the training data. It is used to evaluate the performance of a trained model. Specifically, the test data is not used during model training (**Eg**: A few house records not used during training, kept aside to test the model's performance)
* **Training Data -** The dataset used to train the model. (**Eg**: The house records used to fit the model)
* **Model** - An algorithm that learns from the training data and can make predictions (**Eg**: A regression model predicting house prices based on features)
* Validation Data - A subset of data used to tune the model's hyperparameters (**Eg**: A few house records used to validate the model’s performance during training)
* **Hyperparameter** – It refers to a parameter whose value is set before the learning process begins and is not learned from the data.
* **Epoch** - One complete pass through the training dataset.  
  **(Eg:** The model processes all house records in the training data once)
* **Loss Function** - It is a mathematical function used to measure how well a model's predictions match the actual target values. The goal is to minimize this loss function during training, meaning the model’s predictions are as close as possible to the actual values.
* **Learning Rate** - A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
* **Overfitting -** When a model learns the training data too well, including the noise, and performs poorly on new data. (**Eg**: The model predicts training data house prices perfectly but fails on test data)
* **Underfitting -** Itrefers to a situation where a model is too simple to capture the underlying patterns in the training data. As a result, the model performs poorly on both the training data and new, unseen data.
* **Regularization** - Techniques used to prevent overfitting by adding a penalty to the loss function.
* **Cross Validation** - A technique for assessing how well a model generalizes to an independent dataset.
* **Feature Engineering** - The process of using domain knowledge to create features that make machine learning algorithms work better.
* **Dimensionality Reduction** - Techniques for reducing the number of features in a dataset.
* **Bias** - The error introduced by approximating a real-world problem, which may be too complex.
* **Variance** - The error introduced by the model's sensitivity to small fluctuations in the training set.