|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)**  **Name: INIGASHREE N S** | | **Email:** [**mailto:inigashreesarav22@gmail.com**](mailto:inigashreesarav22@gmail.com) | |  |

|  |
| --- |
| FFfffff |

| **ID** | **Area (sq. ft.)** | **Bedrooms** | **Bathrooms** | **Year Built** | **Neighborhood** | **Garage** | **Sale Price (USD)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1500 | 3 | 2 | 1995 | Suburban | 2 | 250000 |
| 2 | 2000 | 4 | 3 | 2008 | Urban | 1 | 380000 |
| 3 | 1800 | 3 | 2 | 2010 | Rural | 0 | 300000 |
| 4 | 2200 | 5 | 4 | 2005 | Suburban | 2 | 420000 |
| 5 | 1600 | 2 | 1 | 1980 | Urban | 1 | 200000 |

#### Feature

* Individual measurable properties used as inputs to the model.
* **Example**: Features include Area (sq. ft.), Bedrooms, Bathrooms, Year Built, Neighborhood, and Garage.

#### Label

* The output variable that the model aims to predict.
* **Example**: Label is 'Sale Price (USD)'.

#### Outlier

* A data point that deviates significantly from the rest of the data.
* **Example**: An outlier could be a house with an unusually high Sale Price like 1,000,000 USD.

#### Training Data

* Data used to train the model.
* **Example**: Training data includes records from ID=1 to ID=4.

#### Test Data

* Data used to evaluate the model's performance.
* **Example**: Test data might include record ID=5.

#### Model

* Algorithm or program that makes predictions based on the training data.
* **Example**: The model would be trained to predict 'Sale Price (USD)' based on features like Area, Bedrooms, and Year Built.

#### Validation Data

* Sample of data withheld from training used to fine-tune and validate the model.
* **Example**: Validation data could include records ID=2 and ID=3.

#### Hyperparameter

* Parameters set before training to control the learning process.
* **Example**: Hyperparameters could include the choice of algorithm (e.g., linear regression) or regularization strength.

#### Epoch

* Each complete pass of the training dataset through the algorithm.
* **Example**: One epoch could represent one complete pass through records ID=1 to ID=4 during training.

#### Loss Function

* Measures the difference between predicted and actual values.
* **Example**: Mean Squared Error or Mean Absolute Error can quantify how well the predicted Sale Prices match the actual Sale Prices.

#### Learning Rate

* Tuning parameter in an optimization algorithm that controls the step size during training.
* **Example**: A learning rate of 0.01 means the model adjusts its predictions based on 1% of the gradient's magnitude.

#### Overfitting

* When a model performs well on training data but poorly on new data.
* **Example**: Overfitting might occur if the model memorizes the training data (IDs 1-4) but fails to generalize to new data (ID 5).

#### Underfitting

* When a model is too simple to capture patterns in the data.
* **Example**: Underfitting could happen if the model cannot accurately predict Sale Prices even for the training data.

#### Regularization

* Techniques to prevent overfitting.
* **Example**: Lasso (L1) regularization penalizes large coefficients in the model, making it less sensitive to noise in the training data.

#### Cross-Validation

* **Definition**: Technique to assess model performance by splitting data into subsets.
* **Example**: Cross-validation helps in evaluating how well the model predicts Sale Prices across different subsets of the data.

#### Feature Engineering

* Creating new features from existing data to improve model performance.
* **Example**: Creating a 'Total Rooms' feature by summing Bedrooms and Bathrooms could be a feature engineering technique.

#### Dimensionality Reduction

* Reducing the number of input variables.
* **Example**: Principal Component Analysis (PCA) could be used to reduce the number of features like Area, Bedrooms, and Bathrooms.

#### Bias

* Systematic error due to incorrect assumptions in the model.
* **Example:** Bias might occur if the model assumes higher Sale Prices in Suburban areas without considering other factors.

#### Variance

* Model's sensitivity to small fluctuations in the training data.
* **Example**: A model with high variance might predict vastly different Sale Prices for similar houses in the dataset.