**ASSIGNMENT - 1**

**DONE BY:**

PRIYASAKI S

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Employee ID** | **Age** | **Department** | **Years of Experience** | **Education level** | **Monthly Salary** | **Performance Rating** |
| **101** | **25** | **Sales** | **3** | **Bachelor** | **4000** | **3** |
| **102** | **30** | **HR** | **5** | **Master** | **5000** | **4** |
| **103** | **35** | **IT** | **10** | **PhD** | **8000** | **5** |
| **104** | **40** | **Marketing** | **12** | **Bachelor** | **7000** | **4** |
| **105** | **28** | **Sales** | **4** | **Bachelor** | **4500** | **3** |

**Feature:**

* **Definition:** Individual measurable properties used as inputs to the model.
* **Example:** Age, Department, Years of Experience, Education level, Monthly Salary.

**Label:**

* **Definition:** The output variable that the model aims to predict.
* **Example:** Performance Rating.

**Prediction:**

* **Definition:** The output of the model based on input features.
* **Example:** Predicting the Performance Rating given Age, Department, Years of Experience, Education Level, and Monthly Salary.

**Outlier:**

* **Definition:** A data point that deviates significantly from the rest of the data.
* **Example:** If there was an employee with a Monthly Salary of 20000, it might be considered an outlier if most salaries are much lower.

**Test Data:**

* **Definition:** Data used to evaluate the model’s performance after training.
* **Example:** If we split the dataset into 4 training records and 1 testing record, the record for Employee 105 might be used as test data.

**Training Data:**

* **Definition:** Data used to train the model.
* **Example:** The records for Employees 101,102,103 and 104 are used to train the model.

**Model:**

* **Definition:** An algorithm that learns from training data to make predictions or decisions.
* **Example:** A regression model predicting Performance Ratings based on the features.

**Hyperparameter:**

* **Definition:** Parameters set before training that control the learning process. They cannot be learned directly from the training data.
* **Example:** The learning rate or the maximum depth of a decision tree in a machine learning algorithm.

**Validation Data:**

* **Definition:** Data used to tune the model’s hyperparameters and validate its performance during training.
* **Example:** Splitting the dataset into 80% training and 20% validation, using Employee 105’s data as validation data.

**Epoch:**

* **Definition:** One complete pass through the entire training dataset.
* **Example:** Training the regression model on the employee dataset through several epochs, where each epoch involves adjusting the model based on all five records.

**Loss Function:**

* **Definition:** A function that measures the error between the predicted and actual values.
* **Example:** Mean Squared Error (MSE) between the predicted and actual Performance.

**Learning Rate:**

* **Definition:** A hyperparameter that controls how much to change the model in response to the estimated error.
* **Example:** Setting a learning rate of 0.01 to update the regression model’s parameters.

**Overfitting:**

* **Definition:** When a model performs well on training data but poorly on test data due to learning noise and details specific to the training data.
* **Example:** A model that performs well on Employees 101 to 104 but poorly on Employee 105.

**Underfitting:**

* **Definition:** When a model is too simple to capture the underlying pattern in the data, leading to poor performance on both training and test data.
* **Example:** A linear regression model that does not capture the relationship between features and Performance Ratings.

**Regularization:**

* **Definition:** Techniques used to prevent overfitting by adding a penalty to the loss function.
* **Example:** L2 regularization to penalize large coefficients in the model.

**Cross-Validation:**

* **Definition:** A method for evaluating a model’s performance by splitting data into multiple training and validation sets.
* **Example**: Using k-fold cross validation to assess the model by splitting the data into 5 folds, training on 4 folds, and validating on the remaining fold.

**Feature Engineering:**

* **Definition:** Creating or transforming features to improve model performance.
* **Example:** Creating a new feature “Experience Level” by categorizing “Years of Experience” into ranges (e.g., Junior, Mid, Senior).

**Dimensionality Reduction:**

* **Definition:** Techniques used to reduce the number of features while preserving important information.
* **Example:** Applying Principal Component Analysis to reduce features from Age, Department, Years of Experience, Education Level, and Monthly Salary to a small number of principal components.

**Bias:**

* **Definition:** Error due to overly simplistic assumptions. A model with high bias pays too little attention to the training data. This often leads to systematic errors and poor performance on both the training and test data.
* **Example:** A model with high bias that fails to capture the complexities of how different factors affect Performance Ratings.

**Variance:**

* **Definition:** Error due to the model’s sensitivity to fluctuations in the training data.
* **Example:** A model with high variance that fits the training data too closely, resulting in poor performance on new data (e.g., Employee 105).

----------------------------------------------------------------------------------------------