**Assignment 1**

**Case Study on Employees' Productivity**

|  |  |  |  |
| --- | --- | --- | --- |
| **Employee** | **Hours Worked** | **Previous Month Productivity** | **Current Month Productivity** |
| A | 30 | 70 | 75 |
| B | 45 | 80 | 85 |
| C | 20 | 60 | 65 |
| D | 30 | 90 | 92 |
| E | 40 | 85 | 88 |

**Terminologies:**

* **Feature:** In this dataset, "Hours Worked" and "Previous Month Productivity" are features.
* **Label:** "Current Month Productivity" is the label that we aim to predict
* **Prediction:** If we use the model to predict the "Current Month Productivity" for an employee, the predicted value is the prediction.
* **Outlier:** If there was an employee with "Hours Worked" as 100 and "Current Month Productivity" as 30, it would be an outlier.
* **Test Data:** If we reserve employee E's data to test the model, it will be the test data.
* **Training Data:** The data of employees A, B, C, and D used to train the model.
* **Model:** A regression algorithm that uses "Hours Worked" and "Previous Month Productivity" to predict "Current Month Productivity".
* **Validation Data:** If we split employee D's data to tune the model, it will be the validation data.
* **Hyperparameter:** Parameters like the learning rate or the number of epochs set before training the model.
* **Epoch:** If the model processes the entire dataset of employees A, B, C, and D once, it completes one epoch.
* **Loss Function:** Mean Squared Error (MSE) might be used to measure the difference between predicted and actual "Current Month Productivity".
* **Learning Rate:** If the learning rate is 0.01, it determines the step size during each iteration of model training.
* **Overfitting:** If the model performs well on employees A, B, C, and D but poorly on employee E, it might be overfitting.
* **Underfitting:** If the model performs poorly on both training and test data, it is underfitting.
* **Regularization:** Techniques like L2 regularization can be used to prevent overfitting.
* **Cross-Validation:** Dividing the dataset into folds and using different folds for training and testing to evaluate model performance.
* **Feature Engineering:** Creating a new feature like "Productivity Improvement" by calculating the difference between the "Current Month Productivity" and "Previous Month Productivity".
* **Dimensionality Reduction:** Using techniques like PCA to reduce the number of features if there were more features.
* **Bias:** If the model consistently predicts low productivity scores, it might have high bias.
* **Variance:** If the model's predictions vary significantly with different training datasets, it might have high variance.