**ML intern Assignment -1**

By A SOWMIKA

22CSR203

**Dataset Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Income($)** | **Years at Company** | **Number of Visits** | **Purchased** |
| 25 | 50,000 | 5 | 5 | No |
| 45 | 75,000 | 10 | 12 | Yes |
| 30 | 60,000 | 3 | 8 | Yes |
| 50 | 100000 | 15 | 20 | Yes |
| 28 | 55000 | 1 | 6 | No |
| 35 | 65000 | 5 | 10 | Yes |

**Terminologies:**

* **Features:** A feature is a data that's used as an input for models to make predictions.

Here age, income, years at company are features of the dataset.

* **Label:** The specific outcome that the model aims to predict based on the input features.

In our database, whether the customer purchased or not is a label.

And there are supervised unsupervised depending on whether the label present or not.

I) If label is present, it is called supervised learning.

II) If label is not present, it is called unsupervised learning.

* **Prediciton:** To guess the outcome for new information based on patterns it learned from previous data.

Eg: If a person age is 27 and income is 58,000 and no.of.visits is 3 then the predicted purchase will be NO.

* **Outlier:** The data that significantly deviates from the other observations in the dataset. It can be usually high or low compared to the majority of the data points.

Eg: If a person whose age is 57 and income is 80,000 but purchased outcome will be no, that would be an outlier.

* **Test Data:** A part of dataset that can be used to check whether the model is trained well by giving the untrained data. We can use our 80 percent dataset for training purpose and 20 percent for testing purpose.
* **Training Data:** A part of dataset which we use to train our model, with the help of training data they can predict the test data if it is trained well.
* **Model**: A mathematical representation or algorithm that is trained to recognize patterns, make predictions, or solve specific tasks based on data.

Eg: Linear regression, Decision tree are used to predict the outcome if it is supervised and K means clustering, used to predict if it is unsupervised.

* **Validation Data:** A subset of data used to optimize the model during training. An intermediate checkpoint to assess how well the model performs on data.

Eg: With the help of first three data validating the fourth data.

* **Hyperparameter:** Parameter whose values are decided before training of the model begins. It is not affected by training, rather, these affect the quality and speed of the training.

Eg: Learning rate

* **Epoch:** A single iteration where the model sees every sample in the training data set once.
* **Loss Function**: A loss function calculates the error of the model’s predictions compared to the actual data. So, it can be used to minimize the error.

Eg: Mean Squared Error (measure the average of the squares of the errors) and Mean Absolute Error (average of the absolute difference between prediction and actual values)

* **Learning Rate:** A hyperparameter that controls how much to change the model’s parameters in response to the estimated error each time the model’s parameter are updated.
* **Overfitting:** High accuracy on training data and poor performance on test data. A model learns the details and noise in the training data to extent that negatively impacts the model’s performance on new, unseen data.
* **Underfitting:** A model that perform poorly on both the training data and new, unseen data.
* **Regularization:** A technique to prevent the model from overfitting by adding extra information to it. By adding a penalty term to the loss function, discouraging the model from assigning too much importance to individual features or coefficients.
* **Cross – Validation:** Used to split the data. This process helps to ensure that the model’s performance is consistent and reliable across different portions of the data, reducing the overfitting.

Eg: K-Fold Cross – Validation: The dataset is divided into k equal sized folds. The model is trained k times, each time using k-1 folds for training and remaining fold for testing.

* **Feature Engineering:** Process of using domain knowledge to create new features or modify existing features to improve the performance of models.
* **Dimensional Reduction:** Reduce the number of features in a dataset. Reduce the complexity of a model, to improve the performance of an algorithm or to make easier to visualize the data.
* **Bias:** Errors introduced by the assumptions made by a model.

1. Low bias means fewer assumptions are taken, this will closely match the training dataset.
2. High bias means more assumptions are taken to bulid the target function, the model will not match the training dataset.

* **Variance:** Variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data.

Eg: A high variance model may predict the purchased rates accurately in the trained data but poorly in test data.