Joint Tech Internship Community Program

Assignment 1

SUBMITTED BY

IKJAS RASOOL P

(ikjasrasool2022@gmail.com)

Sample Dataset : House Price Prediction

CAR ID	Make	Year	Horsepower	Engine Size (L)	Price ($)
1	Toyota	2017	190	1.5	300,000
2	Ford	2019	180	2.5	450,000
3	Chevrolet	2020	200	5.5	200,000
4	BMW	2022	190	3.5	350,000
5	Audi	2020	220	4.5	500,000
6	Mercedes	2017	210	2.0	150,000
7	Nissan	2019	200	4.5	320,000
8	Honda	2023	190	2.5	480,000

This dataset can be used to illustrate the following terminologies:

1) Feature:

- **Definition**: Features are individual measurable properties or characteristics of a phenomenon being observed.
- **Example in Dataset:** Make, Model, Year, Engine Size (L), Horsepower, are features of the cars.

2) Label:

- **Definition**: The label is the output variable that we are trying to predict or classify.
- **Example in Dataset**: The Price ($) is the label we want to predict.

3) Prediction:

- **Definition**: A prediction is the output of a machine learning model when it is given an input example.
- **Example in Dataset**: Predicting the price of a house based on its features.

4) Outlier:

- **Definition**: An outlier is an observation point that is distant from other observations.
- **Example in Dataset**: If most cars are priced around $15,000 to $30,000, a car priced at $40,000 might be considered an outlier.

5) Test Data:

- **Definition**: Test data is a subset of the dataset used to evaluate the performance of a trained model.
- **Example in Dataset**: 2 - 3 records from the dataset can be set aside as test data to evaluate the model's performance after training.

6) Training Data:

- **Definition**: Training data is a subset of the dataset used to train the model.
- **Example in Dataset**: 4 - 5 records in the dataset would be used as training data to teach the model the relationship between the features and the label.

7) Model:

- **Definition**: A model is a mathematical representation of a real-world process. In machine learning, it is trained to make predictions.
- **Example in Dataset**: A regression model that predicts car prices based on their features.

8) Validation Data:

- **Definition**: Validation data is a subset of the dataset used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- **Example in Dataset**: 1 – 2 records from the training data is used to validate the model’s performance.

9) Hyperparameter:

- **Definition**: Hyperparameters are configuration settings used to tune how the machine learning model is trained.
- **Example in Dataset**: Examples include the learning rate, number of epochs, or the regularization parameter.

10) Epoch:

- **Definition**: An epoch is one complete pass through the entire training dataset.
- **Example in Dataset**: If the dataset is passed through the model 100 times during training, that would be 100 epochs.

11) Loss Function:

- **Definition**: A loss function measures how well the model's predictions match the true labels.
- **Example in Dataset**: Mean Squared Error (MSE) could be used as a loss function to measure the difference between predicted house prices and actual prices.

12) Learning Rate:

- **Definition**: The learning rate is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
- **Example in Dataset**: Setting a learning rate to 0.01 for updating the model weights during training.

13) Overfitting:

- **Definition**: Overfitting occurs when a model learns the training data too well, including noise and details, leading to poor performance on new data.
- **Example in Dataset**: If the model performs exceptionally well on the training cars but poorly on the test cars, it may be overfitting.

14) Underfitting:

- **Definition**: Underfitting occurs when a model is too simple to capture the underlying pattern of the data.
- **Example in Dataset**: If the model performs poorly on both training and test data, it may be underfitting.

15) Regularization:

- **Definition**: Regularization techniques are used to reduce the risk of overfitting by adding a penalty to the loss function for large coefficients.
- **Example in Dataset**: Applying L2 regularization to penalize large weights in the model.

16) Cross-Validation:

- **Definition**: Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent dataset.
- **Example in Dataset**: Using k-fold cross-validation to divide the dataset into k parts and training the model k times, each time using a different part as the validation set.

17) Feature Engineering:

- **Definition**: Feature engineering involves creating new features or modifying existing ones to improve model performance.
- **Example in Dataset**: Creating a new feature such as Price per Horsepower by dividing the price by the horsepower.

18) Dimensionality Reduction:

- **Definition**: Dimensionality reduction is the process of reducing the number of random variables under consideration.
- **Example in Dataset**: Using Principal Component Analysis (PCA) to reduce the number of features while retaining most of the information in the data.

19) Bias:

- **Definition**: Bias is an error due to overly simplistic assumptions in the learning algorithm.
- **Example in Dataset**: If the model consistently predicts car prices lower than the actual prices, it may have a bias.

20) Variance:

- **Definition**: Variance is an error due to too much complexity in the learning algorithm.
- **Example in Dataset**: If the model's predictions vary widely for similar cars in the test set, it may have high variance.