 

**Joint Tech Internship Community Program**

**Assignment 1**

By

Bavyadharshini R

Example Training Data Set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Make | Model | Year | Mileage (km) | Fuel Type | Price (INR) |
| Toyota | Corolla | 2015 | 50000 | Petrol | 850000 |
| Hyundai | Elantra | 2018 | 30000 | Diesel | 1200000 |
| Maruti | Swift | 2020 | 10000 | Petrol | 600000 |
| Honda | City | 2017 | 40000 | Petrol | 950000 |
| Ford | EcoSport | 2016 | 60000 | Diesel | 750000 |

List of Terminologies:

* Feature - An individual measurable property or characteristic of a phenomenon (e.g., Make, Model, Year, Mileage, Fuel Type).
* Label - The output variable that the model is predicting (e.g., Price).
* Prediction - The value that a model outputs after processing the input features (e.g., Predicted car price based on its features).
* **Outlier -** A data point that significantly deviates from the rest of the data. A data point that differs significantly from other observations (e.g., If one of the cars had an extremely low or high price compared to others).
* **Test Data -** Test data refers to a dataset that is independent of the training data. It is used to evaluate the performance of a trained model. Specifically, the test data is not used during model training (e.g., A few car records not used during training, kept aside to test the model's performance).
* **Training Data -** The dataset used to train the model. (**Eg**: The house records used to fit the model) The dataset used to train the model. (e.g., The car records used to fit the model).
* **Model** - An algorithm that learns from the training data and can make predictions (e.g., A regression model predicting car prices based on features).
* Validation Data - A subset of data used to tune the model's hyperparameters (e.g., A few car records used to validate the model’s performance during training).
* **Hyperparameter** –It refers to a parameter whose value is set before the learning process begins and is not learned from the data (e.g., Learning rate).
* **Epoch** - One complete pass through the training dataset (e.g., The model processes all car records in the training data once).
* **Loss Function** - It is a mathematical function used to measure how well a model's predictions match the actual target values. The goal is to minimize this loss function during training, meaning the model’s predictions are as close as possible to the actual values.
* **Learning Rate** - A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
* **Overfitting -** When a model learns the training data too well, including the noise, and performs poorly on new data. (e.g., The model predicts training data car prices perfectly but fails on test data).
* **Underfitting -** It refers to a situation where a model is too simple to capture the underlying patterns in the training data. As a result, the model performs poorly on both the training data and new, unseen data.
* **Regularization** - Techniques used to prevent overfitting by adding a penalty to the loss function.
* **Cross Validation** - A technique for assessing how well a model generalizes to an independent dataset.
* **Feature Engineering** - The process of using domain knowledge to create features that make machine learning algorithms work better.
* **Dimensionality Reduction** - Techniques for reducing the number of features in a dataset.
* **Bias** - The error introduced by approximating a real-world problem, which may be too complex.
* **Variance** - The error introduced by the model's sensitivity to small fluctuations in the training set.