A logo of a company

Description automatically generatedA close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment1-Machine Learning Terminologies**

**Submitted By:**

**Kesava Prabha L (22CSR098)**

**Candidate Id: 2024060046**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Make Year** | **Brand** | **Variant** | **Mileage** | **Fuel** | **Transmission** | **Resale Price (INR)** |
| 2015 | BMW | 520D | 80000 | Diesel | Automatic | 2500000 |
| 2016 | Audi | A6 | 92000 | Petrol | Automatic | 1900000 |
| 2018 | Mercedes Benz | E200 | 61000 | Petrol | Automatic | 3400000 |
| 2014 | Skoda | Superb | 95000 | Petrol | Automatic | 600000 |
| 2020 | Benz | E200 | 35000 | Petrol | Automatic | 12000000 |

* **Feature**: Individual measurable properties (e.g., Make Year, Brand, Variant etc.) used as inputs to the model.
* **Label**: The output variable that the model aims to predict (e.g., Resale Price).
* **Outlier**: A data point that deviates significantly from the rest of the data (e.g., if there was a car with resale price at 12000000 in this dataset).
* **Prediction:** The Output of the Machine Learning Model, after it has been trained, representing the model’s best guess. (e.g., Resale Price)
* **Training Data:** The Subset of the dataset used to train the model (We can take 2015,2018,2020,2014 Make year datum to train our model i.e. take datum in different category)
* **Test Data:** A subset of the dataset used to assess the performance of the trained model. (Here we can use 3 data rows for training and remaining 2 rows for testing purpose)
* **Model:** A model is the set of parameters and structure needed for a system to make predictions. (A Regression model predicting the Resale Price based on “Brand”, ”Mileage”, ”Transmission” etc.)
* **Validation Data:** The subset of the dataset that performs initial evaluation against a trained model. (Setting one row of dataset to validate the model’s performance)
* **Hyperparameter:** A Hyperparameter is the tuning service that adjust during successive runs of training a model. (Learning rate in the model)
* **Epoch:** One Complete pass through the entire training dataset. (If the model trains on entire training data once it completes one epoch basically it is like a loop)
* **Loss Function:** Afunction that predicts how well the model’s prediction match the actual labels.
* **Learning Rate:** A hyperparameter that controls how much the model's parameters are adjusted with respect to the loss gradient.
* **Overfitting:** Creating a Model that matches the training data so closely that the model fails to make correct predictions on new data. (e.g. model perfectly predicting the resale prices in the training dataset but failing to predict accurately for new car data.)
* **Underfitting:** Producing a model with poor predictive ability because the model hasn't fully captured the complexity of the training data. (e.g. model that doesn't capture the non-linear relationship between car features and their resale prices.)
* **Regularization:** Techniques used to prevent overfitting by penalizing complex models.
* **Cross-Validation:** A technique for assessing how the results of a statistical analysis will generalize to an independent dataset.
* **Feature Engineering:**  Determining which features might be useful in training a model and converting raw data from the dataset into efficient versions of those features. (e.g. Creating a new feature like ‘On Road Price’ to improve potential prediction)
* **Dimensionality Reduction:** The process of reducing the number of the features in a dataset. (e.g. transform the available features to new small set of features)
* **Bias:** Stereotyping, prejudice or favoritism towards some things, people, or groups over others. (e.g. Simplify the relationship between car features and their resale prices.)
* **Variance:** The error introduced by the model's sensitivity to small fluctuations in the training set. A high-variance model might perform well on the training data but poorly on test data due to overfitting to the training set.