

**Joint Tech Internship Community Program**

**ASSIGNMENT 1**

**­­­­­**

Submitted by : HARINI C S

Candidate ID : 2024060059

**Example Test Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of Screens** | **Seating Capacity** | **Location Quality (1-10)** | **Theater Age (years)** | **Price per Ticket (INR)** |
| 5 | 300 | 8 | 15 | 200 |
| 8 | 500 | 9 | 10 | 350 |
| 3 | 150 | 6 | 20 | 150 |
| 6 | 400 | 7 | 5 | 250 |
| 10 | 700 | 9 | 3 | 500 |
| 4 | 250 | 5 | 12 | 180 |

**Explanation of Terminologies**

1. **Feature**: Individual measurable properties or characteristics used as inputs to the model.
   * **Example**: In this dataset, the features are:
     + Number of Screens
     + Seating Capacity
     + Location Quality
     + Theater Age
2. **Label**: The output variable that the model aims to predict.
   * **Example**: The label in this dataset is:
     + Price per Ticket (INR)
3. **Prediction**: The output of the model after it has processed the input features.
   * **Example**: If the model predicts that the price per ticket should be 300 INR for a theater with 6 screens, 400 seating capacity, a location quality of 7, and a theater age of 5 years.
4. **Outlier**: A data point that deviates significantly from the rest of the data.
   * **Example**: If there was a cinema theater with a price per ticket of 2000 INR in this dataset, it would be considered an outlier.
5. **Test Data**: A set of data used to evaluate the performance of the model. This data is not used during the training of the model.
   * **Example**: Using 2 of the rows from this dataset that were not used during training to see how well the model predicts prices for new theaters.
6. **Training Data**: A set of data used to train the model. This data is used to help the model learn the relationship between the features and the label.
   * **Example**: Using 4 of the rows from this dataset to teach the model the relationship between features and the price per ticket.
7. **Model**: The algorithm or mathematical representation used to make predictions.
   * **Example**: A linear regression model predicting the price per ticket based on the input features.
8. **Validation Data**: A separate set of data used to tune the model's hyperparameters and prevent overfitting.
   * **Example**: Using 1 row from the dataset that is different from the training data to adjust the learning rate or the number of epochs.
9. **Hyperparameter**: A parameter whose value is set before the learning process begins. These are not learned from the training data.
   * **Example**: The learning rate set to 0.01.
10. **Epoch**: One complete pass through the entire training dataset.
    * **Example**: If the training data consists of 4 examples, an epoch involves processing all 4 examples once.
11. **Loss Function**: A function that measures how well the model’s predictions match the actual data. It quantifies the difference between the predicted and actual values.
    * **Example**: Mean Squared Error (MSE) between predicted and actual ticket prices.
12. **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.
    * **Example**: Setting the learning rate to 0.01 ensures that the model updates its weights gradually.
13. **Overfitting**: When a model learns the training data too well, including noise and outliers, and performs poorly on new, unseen data.
    * **Example**: If the model predicts the training data perfectly but performs poorly on the test data.
14. **Underfitting**: When a model is too simple to capture the underlying structure of the data, leading to poor performance even on the training data.
    * **Example**: If the model cannot even predict the training data well because it is too simple.
15. **Regularization**: Techniques used to prevent overfitting by penalizing large coefficients.
    * **Example**: Using L2 regularization to penalize large coefficients in the model.
16. **Cross-Validation**: A technique to evaluate the model by dividing the data into training and testing sets multiple times to ensure the model's performance is consistent.
    * **Example**: Using 5-fold cross-validation to ensure the model performs consistently across different subsets of the data.
17. **Feature Engineering**: The process of using domain knowledge to create features that make machine learning algorithms work better.
    * **Example**: Creating a new feature called "Screens per Seating Capacity".
18. **Dimensionality Reduction**: Reducing the number of random variables under consideration, usually to improve model performance and reduce overfitting.
    * **Example**: Using Principal Component Analysis (PCA) to reduce the number of features from 4 to 2.
19. **Bias**: Error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to miss relevant relations between features and the label (underfitting).
    * **Example**: High bias if the model predicts all ticket prices to be 250 INR regardless of features.
20. **Variance**: Error due to too much complexity in the learning algorithm. High variance can cause the model to model the random noise in the training data (overfitting).
    * **Example**: High variance if the model predicts the training data very well but fails on the test data.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*