## SOORYAPANDIAN S P

22CSR201

## ASSIGNMENT 1

## Business Sales Prediction Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Month | Store Location | Advertising Budget (USD) | Competitor Price (USD) | Sales (Units) |
| Jan | Location A | 5000 | 25 | 200 |
| Feb | Location B | 7000 | 24 | 250 |
| Mar | Location A | 4500 | 23 | 180 |
| Apr | Location C | 6000 | 26 | 220 |
| May | Location B | 5500 | 25 | 240 |

### Feature

Individual measurable properties or characteristics used as inputs to the model.

**Eg:** The features in this dataset are Month, Store Location, Advertising Budget, and Competitor Price.

### Label

The output variable that the model aims to predict.

**Eg:** The label in this dataset is Sales (Units).

### Prediction

The value that the model estimates for the label.

**Eg:** Given the values for Month, Store Location, Advertising Budget, and Competitor Price, the model predicts the number of Sales (Units).

### Outlier

A data point that deviates significantly from the rest of the data. =

**Eg:** If there was a month with sales of 500 units while others are between 180-250 units, then this sales figure would be considered an outlier.

### Test Data

A subset of the dataset used to evaluate the performance of the model. Typically, this data is not seen by the model during training.

**Eg:** Data for the months Apr and May can be used as test data to evaluate the model trained on the data from Jan to Mar.

### Training Data

A subset of the dataset used to train the model.

**Eg:** Data for the months Jan, Feb, and Mar can be used as training data.

### Model

A mathematical representation created by the machine learning algorithm to make predictions based on the input features.

**Eg:** A model that predicts Sales (Units) based on Advertising Budget and Competitor Price.

### Validation Data

A subset of the dataset used to tune the hyperparameters of the model. This data is also not seen during training.

**Eg:** Using data for the month of Mar to validate the model trained on data from Jan and Feb.

### Hyperparameter

A parameter whose value is set before the learning process begins.

**Eg:** Learning rate or the number of epochs.

### Epoch

One complete pass through the entire training dataset.

**Eg:** If we iterate through all the months' data once during training, it counts as one epoch.

### Loss Function

A function that measures how well the model’s predictions match the actual labels. Common examples include Mean Squared Error.

**Eg:** Used to measure the difference between the predicted and actual Sales (Units).

### Learning Rate

A hyperparameter that controls how much the model is adjusted with respect to the loss gradient. It determines the step size during optimization.

**Eg:** A learning rate of 0.01 means the model is updated slowly, while a learning rate of 1 means the model is updated quickly.

### Overfitting

When a model performs well on training data but poorly on test data. This often happens when the model learns noise in the training data.

**Eg:** A model that perfectly predicts the training data (e.g., sales for Jan to Mar) but performs poorly on the test data (Apr and May).

### Underfitting

When a model is too simple to capture the underlying pattern of the data, resulting in poor performance on both training and test data.

**Eg:** A model that performs poorly on both the training data (e.g., Jan to Mar) and the test data (Apr and May).

### Regularization

Techniques used to prevent overfitting by adding a penalty to the loss function for large coefficients.

**Eg:** Adding a penalty to the loss function for large coefficients in a linear regression model.

### Cross-Validation

A technique for assessing how the results of a statistical analysis will generalize to an independent dataset. Commonly done using k-fold cross-validation.

**Eg:** 5-fold cross-validation involves splitting the data into 5 parts, training the model on 4 parts, and testing on the remaining part, repeating this process 5 times.

### Feature Engineering

The process of using domain knowledge to create new features that help the model perform better.

**Eg:** Creating a new feature "Advertising Efficiency" by dividing the Advertising Budget by the Competitor Price.

### Dimensionality Reduction

Techniques used to reduce the number of features in a dataset.

**Eg:** Principal Component Analysis (PCA) to reduce Advertising Budget and Competitor Price into a single combined feature.

### Bias

The error introduced by approximating a real-world problem, which may be complex, by a much simpler model.

**Eg:** A high bias model may predict that sales will always be around 200 units regardless of changes in the advertising budget or competitor price, thus missing the true relationship between these features and the actual sales, which can vary significantly based on these factors.

### Variance

The error introduced by the model’s sensitivity to small fluctuations in the training dataset.

**Eg:** A high variance model might predict sales extremely accurately for the months included in the training data (e.g., predicting 200 units for January, 250 units for February, etc.), but when applied to new, unseen data for different months, the predictions could fluctuate wildly, such as predicting 300 units for June when the actual sales are only 180 units. This inconsistency indicates that the model has overfit to the noise in the training data and doesn't generalize well to new data.