**Generative AI Consortium (Ltd)  
AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)**

**Name: ARVIND M  
Email:** [**arvindmurugesan001@gmail.com**](mailto:arvindmurugesan001@gmail.com)

Laptop Data Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Laptop ID | Processor Type | Ram (GB) | Storage (GB) | Screen Size(inches) | Price (USD) |
| 201 | Intel i5 | 8 | 256 | 14 | 700 |
| 202 | AMD Ryzen 7 | 16 | 512 | 15.6 | 1000 |
| 203 | Intel i3 | 4 | 128 | 13.3 | 400 |
| 204 | Intel i7 | 32 | 1024 | 17 | 2000 |
| 205 | AMD Ryzen 5 | 8 | 256 | 14 | 800 |

Machine Learning Terminology

1. Feature : An individual measurable property or characteristic used as an input for the model to learn from. (Think of it as an ingredient in a recipe for the model to cook up a prediction)

Example: Processor Type, RAM (GB), Storage (GB), Screen Size (inches).

2. Label: The target variable that the model is trying to predict. It's like the final dish you want the model to create using the features (ingredients).

Example: Price (USD).

3. Prediction: The model's educated guess about the label (target variable) based on the features it has been trained on. (Imagine the model presenting its proposed dish after analyzing the ingredients)

Example: Predicted prices based on the given features.

4. Outlier: A data point that stands significantly apart from the rest of the data, like an unexpected ingredient in your recipe that throws everything off.

Example: If there was a laptop with a price of $5000 in this dataset.

5. Test Data: A portion of the data set aside specifically to evaluate how well the trained model performs on unseen information. (Think of it as using a different set of ingredients to see if the model can still cook the same dish)

Example: If we use records 204 and 205 to evaluate the model.

6. Training Data: he main dataset used to train the model. The model learns patterns and relationships between features and labels from this data. (Imagine practicing your recipe on familiar ingredients)

Example: Records 201, 202, and 203 can be used to train the model.

7. Model An algorithm or a mathematical representation that uses features to make predictions about the label. Think of it as the recipe itself, with instructions on how to combine the ingredients (features) to create the desired outcome (label).  
  
 Example: A regression model predicting laptop prices.

8. Validation Data: A subset of the data used to fine-tune the model during the training process. It helps prevent overfitting and ensures the model generalizes well. (Imagine having a taste tester to provide feedback while you perfect your recipe)

Example: If we use record 204 to validate the model's performance during training.

9. Hyperparameter: Settings that control how the model learns from the data. These are like the cooking temperature, time, or amount of spices you use in your recipe, which can significantly impact the final dish.  
 Example: Learning rate, number of epochs.

10. Epoch: One complete pass through the entire training dataset. Imagine making your recipe from start to finish once.

Example: Training the model on records 201, 202, and 203 once.

11. Loss Function: A method to measure how different the model's predictions are from the actual labels. It helps the model adjust its recipe (learning process) to minimize this difference. (Think of it as a way to gauge how well your dish matches the intended flavor)

Example: Mean Squared Error (MSE) between predicted and actual prices.

12. Learning Rate: A hyperparameter that controls how much the model adjusts its internal settings (like weights) with each update during training. A smaller learning rate is like making small adjustments to your recipe bit by bit, while a larger learning rate is like making big changes at once. Example: A learning rate of 0.01.

13. Overfitting: When a model memorizes the training data too well, including the noise or errors, and performs poorly on new, unseen data. It's like your recipe becoming overly specific to a particular set of ingredients and failing when you try to use something different.

Example: If our model predicts the training data perfectly but fails on test data.

14. Underfitting: When a model is too simple to capture the important patterns in the data and performs poorly on both the training and test data. Imagine a recipe that's missing key ingredients, resulting in a bland or incomplete dish regardless of the data (ingredients) used.

Example: Using a very simple model (e.g., predicting price based solely on screen size) might fail to capture other important factors like processor type or RAM, leading to poor predictions for all records.

15. Regularization: Techniques used to prevent overfitting by penalizing overly complex models. It's like adding constraints to your recipe to prevent it from becoming too specific and inflexible. Example:

- Lasso (L1 regularization) adds a penalty equal to the absolute value of the coefficients.

- Ridge (L2 regularization) adds a penalty equal to the square of the coefficients.

16. Cross-Validation A technique to assess how well a model generalizes to unseen data. It involves splitting the data into multiple sets, training the model on some sets, and testing it on others. Imagine testing your recipe on different sets of ingredients to see if it consistently produces the desired outcome.  
 Example: Using 5-fold cross-validation, the dataset is split into 5 parts. The model is trained on 4 parts and tested on the remaining part, and this process is repeated 5 times with each part used as a test set once.

17. Feature Engineering: Creating new features from existing data to potentially improve the model's performance. It's like inventing new ingredients or ingredient combinations to enhance your recipe.  
 Example: Creating a Performance Category feature by binning laptops into categories like Basic (<8GB RAM), Standard (8GB-16GB RAM), and High Performance (>16GB RAM).

18. Dimensionality Reduction: Techniques for reducing the number of features in a dataset. This can be helpful when there are too many features and it becomes computationally expensive for the model to learn from them all. Imagine simplifying your recipe by using fewer ingredients while still achieving the desired taste.  
 Example: Using Principal Component Analysis (PCA) to reduce the features Processor Type, RAM, and Storage into principal components that capture the most variance in Price.

19. Bias: A systematic error in the model's predictions due to assumptions made during the machine learning process. It's like a flaw in your recipe that consistently leads to a certain outcome, regardless of the specific ingredients used.  
  
 Example:

- Sample Bias: A model trained only on data from high-end laptops may not generalize well to budget laptops.

20. Variance: The amount by which a model's predictions change when trained on different subsets of the training data. High variance indicates that the model is sensitive to the specific data it is trained on and may not generalize well to unseen data. Imagine your recipe being very dependent on the exact measurements or cooking time, leading to inconsistent results with slight variations.

Example of High Variance: A highly complex regression model that fits the training data (e.g., records 201-203) very closely, but predicts wildly different prices for test data (e.g., records 204-205) due to its sensitivity to the training data.