Assignment : Feature Engineering

Questions-Answers:

1. What is a parameter?

A **parameter** is a variable that defines the model's behavior during learning, such as weights in linear regression. It is learned from data.

2. What is correlation?

Correlation measures the linear relationship between two variables. It ranges from -1 (perfect negative) to +1 (perfect positive).

3. What does negative correlation mean?

A **negative correlation** means that as one variable increases, the other decreases.

4. Define Machine Learning. What are the main components in ML?

Machine Learning is a field of AI that enables systems to learn patterns from data and make predictions.
Main components:

- Data
- Model
- Loss Function
- Optimizer
- Training & Testing

5. How does loss value help in determining model performance?

The **loss value** quantifies the error between predicted and actual values. Lower loss means better model performance.

6. What are continuous and categorical variables?

- **Continuous**: Numeric, can take any value (e.g., weight, salary)
- **Categorical**: Discrete groups or labels (e.g., gender, city)

7. How do we handle categorical variables in ML?

Common techniques:

- Label Encoding
- One-Hot Encoding
- Ordinal Encoding
- Frequency Encoding

8. What is training and testing a dataset?

- **Training set**: Data used to train the model
- **Testing set**: Data used to evaluate model performance on unseen data

9. What is `sklearn.preprocessing`?

A module in `scikit-learn` that provides functions for:

- Scaling (`StandardScaler`, `MinMaxScaler`)
- Encoding (`LabelEncoder`, `OneHotEncoder`)
- Imputation and other preprocessing steps

10. What is a Test set?

A **test set** is a portion of the dataset used **only after training** to evaluate model accuracy on new, unseen data.

11. How do we split data in Python?

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

12. How do you approach a Machine Learning problem?

1. Understand the problem
2. Collect and clean data
3. Perform EDA (exploratory data analysis)
4. Engineer features
5. Split the data
6. Select and train model
7. Evaluate performance
8. Tune hyperparameters
9. Deploy

13. Why do EDA before modeling?

- Identify trends, outliers, and patterns
- Understand data distributions

- Select relevant features
- Guide model selection

14. What is correlation? *(repeated)*

[Answered above: Question 2]

15. What does negative correlation mean? *(repeated)*

[Answered above: Question 3]

16. How to find correlation in Python?

```
import pandas as pd

df.corr()  # Returns Pearson correlation matrix
```

17. What is causation? Difference from correlation?

- **Causation** means one variable causes another to change.
- **Correlation** is just an association.

Example:

- **Correlation**: Ice cream sales and drowning deaths rise in summer.
- **Causation**: Ice cream does not cause drowning.

18. What is an Optimizer? Types?

An **optimizer** adjusts model parameters to minimize loss.
Types:

- **SGD (Stochastic Gradient Descent)**
- **Adam**
- **RMSprop**

```
from tensorflow.keras.optimizers import Adam
model.compile(optimizer=Adam(), loss='mse')
```

19. What is `sklearn.linear_model`?

A submodule in `scikit-learn` for linear models like:

- `LinearRegression`
- `LogisticRegression`
- `Ridge`, `Lasso`, etc.

---

20. What does `model.fit()` do?

It **trains the model** using training data.
Arguments: `X_train`, `y_train`, optionally epochs, batch size (for deep learning)

21. What does `model.predict()` do?

It **generates predictions** using the trained model.
Arguments: `X_test` or any new input data

22. What are continuous and categorical variables? *(repeated)*

[Answered above: Question 6]

23. What is feature scaling? Why important?

Feature scaling transforms features to a common scale.
Helps algorithms like KNN, SVM, and gradient descent converge faster and avoid bias from large-value features.

24. How do we perform scaling in Python?

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

25. What is `sklearn.preprocessing`? *(repeated)*

[Answered above: Question 9]

26. How do we split data in Python? *(repeated)*

[Answered above: Question 11]

27. Explain Data Encoding.

Data Encoding converts categorical variables into numeric formats.
Types:

- Label Encoding: Converts categories to numbers
- One-Hot Encoding: Creates binary columns for each category
  Used so models can interpret non-numeric data.