Q1. What is a Normal distribution, and why is it considered important?

Normal distribution (also called the Gaussian distribution) is a probability distribution that is symmetric around its mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, it creates a bell-shaped curve.

Key properties:

- The mean, median, and mode are equal.

- The curve is symmetric about the mean.

- About 68% of the data falls within one standard deviation of the mean, 95% within two, and 99.7% within three (the empirical rule).

Importance:

- Many natural phenomena (heights, IQ scores, errors in measurement) follow a normal distribution.

- Many statistical tests and machine learning models (like linear regression) assume normality in data.

- The Central Limit Theorem states that, for a sufficiently large sample size, the sampling distribution of the sample mean will be approximately normally distributed, regardless of the data's original distribution.

Q2. Can you explain TF-IDF? What are its advantages and disadvantages?

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus).

- Term Frequency (TF): Measures how frequently a term appears in a document.

$$\text{TF}(t, d) = \frac{\text{Number of times } t \text{ appears in } d}{\text{Total number of terms in } d}$$

- Inverse Document Frequency (IDF): Measures how important a term is across the entire corpus. Rare words have higher IDF scores.

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$

- TF-IDF score: Combines these metrics to give more weight to terms that are frequent in a document but rare across the corpus.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Advantages:

- Highlights important words in a document.

- Removes the influence of common words (e.g., "the", "is").

- Simple to compute and effective in information retrieval.

Disadvantages:

- Doesn't capture the meaning or relationships between words (semantics).

- Sensitive to stopwords, punctuation, and case.

- Can be biased if some documents are much longer than others.

Q3. What does overfitting mean in machine learning, and what strategies can be used to prevent it?

Overfitting occurs when a machine learning model captures not just the underlying pattern in the data but also the noise, leading to poor generalization on unseen data. The model performs well on training data but poorly on testing data.

Prevention strategies:

1. Cross-validation: Use techniques like k-fold cross-validation to ensure the model generalizes well.

2. Regularization: Add a penalty term to the loss function (e.g., L1, L2 regularization) to constrain model complexity.

3. Simpler models: Use a less complex model that reduces the risk of overfitting.

4. Pruning: In decision trees, prune unnecessary branches.

5. Early stopping: In iterative models like neural networks, stop training once performance on a validation set starts to degrade.

6. Increase data: Provide more data to make patterns more apparent, reducing noise capture.

7. Dropout (for neural networks): Randomly drop units during training to prevent the network from relying too much on any one node.

Q4. What is cross-validation, and why is it essential for evaluating models?

Cross-validation is a technique for assessing how well a machine learning model generalizes to unseen data. In k-fold cross-validation, the data is split into k equally sized subsets, and the model is trained k times, each time using a different subset as the validation set and the remaining as the training set.

Importance:

- It provides a more robust estimate of model performance compared to a single train-test split.

- Reduces the risk of overfitting by ensuring the model is validated on multiple subsets.

- Helps in model selection and hyperparameter tuning.

- Useful in cases where data is limited, as it maximizes the use of available data.

Q5. What is meant by imbalanced data, and what are some common methods for addressing imbalanced datasets?

Imbalanced data occurs when the classes in a classification problem are not represented equally. For example, in a fraud detection dataset, fraudulent transactions may be much rarer than non-fraudulent ones.

Methods to handle imbalanced data:

1. Resampling techniques:

Oversampling: Duplicate instances from the minority class (e.g., SMOTE - Synthetic Minority Over-sampling Technique).

Undersampling: Remove instances from the majority class.

2. Class weight adjustment: Assign higher penalties for misclassifying minority class examples during model training (e.g., setting class weights in logistic regression).

3. Anomaly detection: Treat the minority class as an anomaly and apply models designed for such cases.

4. Use of evaluation metrics like Precision-Recall or F1-Score: Accuracy may be misleading in imbalanced datasets; instead, metrics like Precision, Recall, F1-Score, or the ROC-AUC curve can better measure model performance.

Q6. What is the significance of feature engineering? Describe the concept and its role in developing machine learning models ?

Feature engineering is the process of transforming raw data into features that better represent the problem to the model, thereby improving model performance.

Significance:

- Good features can significantly enhance the model's ability to capture important patterns in the data.

- Poorly engineered features can lead to poor model performance, even if using advanced algorithms.

- Feature engineering involves:

  - Feature selection: Choosing the most relevant features.

  - Feature transformation: Scaling, normalizing, encoding categorical variables, or creating interaction terms.

  - Domain knowledge: Using expert knowledge to create meaningful features (e.g., creating a ‘total spend’ feature by summing ‘purchases’ and ‘installments’).

Role in developing machine learning models:

- Feature engineering provides the model with better representations of data, making it easier to learn from.

- It often has a larger impact on model accuracy than choosing the right algorithm, as the quality and quantity of relevant information significantly determine model success.