XAI study QA

1. Why is Explainable AI (XAI) important in modern Artificial Intelligence applications? Discuss with examples.

Explainable AI (XAI) is essential in high-stakes domains where trust and transparency are crucial. It allows stakeholders to understand how and why an AI model arrives at specific decisions, enabling better accountability and control. For example, in medical diagnosis, XAI can show that certain symptoms or image features led to a diagnosis, aiding doctors in verification. In finance, when a loan is rejected, XAI can explain that the decision was based on factors like credit score or employment history. Without explainability, users may not trust AI systems, which can hinder their adoption and lead to ethical and legal concerns.

2. Differentiate between Explainability and Interpretability. Explain how they affect AI model understanding.

Explainability is the degree to which a model's output can be explained to a human in understandable terms, while interpretability refers to how well a human can grasp the internal mechanics of the model itself. A decision tree is both explainable and interpretable, but deep neural networks are often only explainable through post-hoc techniques. These concepts affect transparency, trust, and the ability to debug or improve models effectively.

3. Explain the various types of XAI techniques with examples: Intrinsic vs Post-hoc, Local vs Global, and Model-specific vs Model-agnostic.

- **Intrinsic vs Post-hoc**: Intrinsic methods are interpretable by design, such as linear regression or decision trees, whereas post-hoc methods like SHAP or LIME explain decisions made by complex black-box models after training.
- **Local vs Global**: Local methods, such as LIME, provide explanations for individual predictions, while global methods, such as feature importance plots, provide insights into the overall behavior of the model.
- **Model-specific vs Model-agnostic**: Model-specific techniques work only with certain model types (e.g., SHAP with tree models), while model-agnostic techniques like LIME can be applied to any predictive model.

4. What is the role of VIF (Variance Inflation Factor) in Linear Regression? How can a high VIF value affect the final model?

VIF detects multicollinearity in linear regression by quantifying how much the variance of a coefficient is increased due to correlations with other predictors. A high VIF (>10) suggests strong multicollinearity, which can make the model unstable, inflate errors, and reduce reliability of coefficients.

5. Explain how SHAP provides both local and global explanations in a machine learning model.

SHAP uses Shapley values from cooperative game theory to assign each feature a contribution score. For local explanations, SHAP shows how features impacted a single prediction. For global explanations, it averages these values across the dataset to highlight which features are most influential overall. This dual perspective helps both in debugging and interpreting models.

6. How does LIME work to explain individual predictions in a black-box model? Explain with an example.

LIME explains predictions by sampling perturbed data around the instance of interest and training a simple, interpretable model to approximate the local decision boundary. For example, in a sentiment analysis model, LIME may highlight that words like “excellent” or “terrible” were key to classifying a review as positive or negative.

7. What is a Partial Dependency Plot (PDP), and what does it help visualize?

PDP shows the average effect of a feature on the predicted outcome while holding other features constant.

8. How does SHAP explain a Decision Tree model prediction?

SHAP assigns fair contribution scores to each feature by tracing their roles through the paths of a decision tree.

9. Explain how LIME helps in understanding the predictions of non-linear models. Give one example.

LIME allows users to interpret predictions made by complex, non-linear models by generating simple, local approximations. It perturbs the input data, observes the model’s output, and then fits a surrogate interpretable model like linear regression to these local samples. This helps users see which features most influenced the model’s prediction. For instance, in fraud detection, LIME can identify that a large transaction amount and a foreign IP address significantly contributed to the model predicting fraud. This transparency aids in trust and decision review.

10. What are ensemble models and why are they used in machine learning?

Ensemble models improve accuracy and robustness by combining multiple base models. They reduce overfitting and increase predictive performance.

11. Name any two types of ensemble learning techniques.

Bagging and Boosting.

12. Explain how SHAP can be used to interpret the predictions of an ensemble classification model.

SHAP provides consistent and fair explanations for ensemble models like random forests and gradient boosting machines by calculating each feature’s Shapley value. These values represent each feature’s contribution to the model’s prediction. For instance, in a healthcare model using random forests, SHAP can show how age, cholesterol level, and blood pressure collectively impact the prediction of a heart disease risk. The additive nature of SHAP explanations ensures clarity in complex ensemble outputs, enabling both local and global model interpretability.

13. What is the What-If Tool (WIT) by Google?

WIT is a visual tool to inspect and experiment with ML models without coding.

14. What are Counterfactual Explanations (CFEs) in XAI?

CFEs suggest minimal changes in input that would alter the model’s prediction.

15. What are Counterfactual Explanations (CFEs), and how do they improve explainability?

Counterfactual Explanations (CFEs) show hypothetical “what-if” scenarios to users, such as: “What should be different in this input to change the output?” They identify the minimal changes in input

features that would flip a prediction to a desired class. For example, if a loan application is rejected, a CFE might suggest that increasing the income by 10,000 and reducing existing debt by 5,000 could result in approval. This helps users understand decision boundaries and take actionable steps. CFEs enhance transparency, fairness, and user empowerment in AI systems.

16. What is the Contrastive Explanation Method (CEM) in Machine Learning?

CEM highlights both features that justify a prediction (positives) and those whose absence is important (negatives).

17. How does CEM differ from other explanation methods like SHAP or LIME?

CEM is contrastive, focusing on presence and absence of features, unlike SHAP or LIME which provide additive importance scores.

18. Explain how CEM works using the Alibi library for tabular data.

CEM with the Alibi library identifies pertinent positives (features that are essential for a model to make a certain prediction) and pertinent negatives (features that, if present, would change the prediction). The method involves optimizing the input data to find minimal changes that influence model outcomes, using a loss function that balances prediction accuracy, sparsity, and contrast. For instance, in a tabular dataset used for predicting disease diagnosis, CEM might show that the absence of smoking history and presence of exercise frequency are key in predicting a ‘healthy’ outcome. This provides richer, contrastive insights than other methods.