Analytics Vidhya

40 ML Interview Questions that You Must Know

Along with the solutions

Q1. Why do we take the harmonic mean of precision and recall when finding the F1-score and not simply the mean of the two metrics?

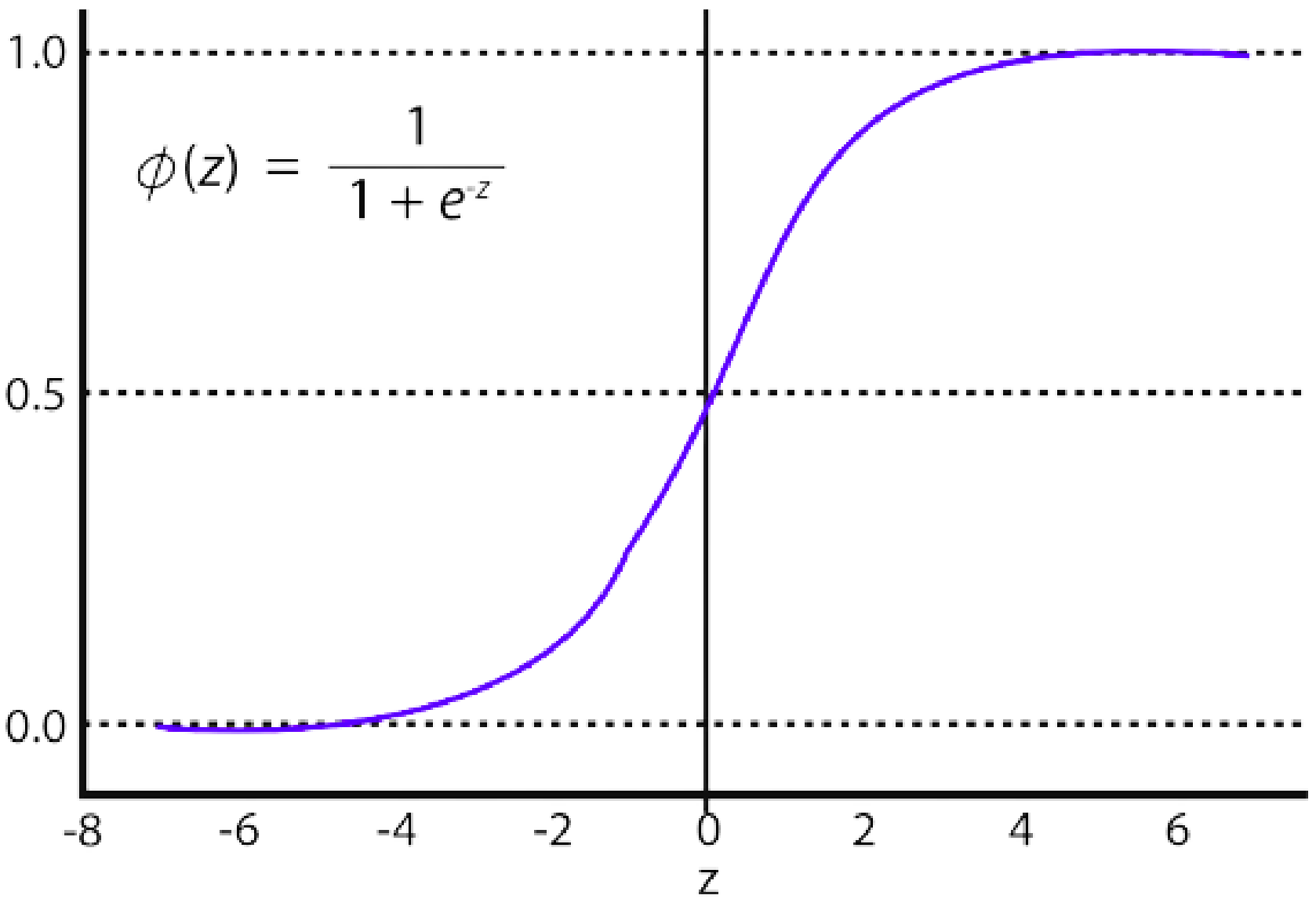
- The F1-score, the harmonic mean of precision and recall, balances the trade-off between precision and recall. The harmonic mean penalizes extreme values more than the arithmetic mean.

- This is crucial for cases where one of the metrics is significantly lower than the other. In classification tasks, precision and recall may have an inverse relationship; therefore, the harmonic mean ensures that the F1-score gives equal weight to precision and recall, providing a more balanced evaluation metric.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Q2. Why does Logistic regression have regression in its name even if it is used specifically for Classification?

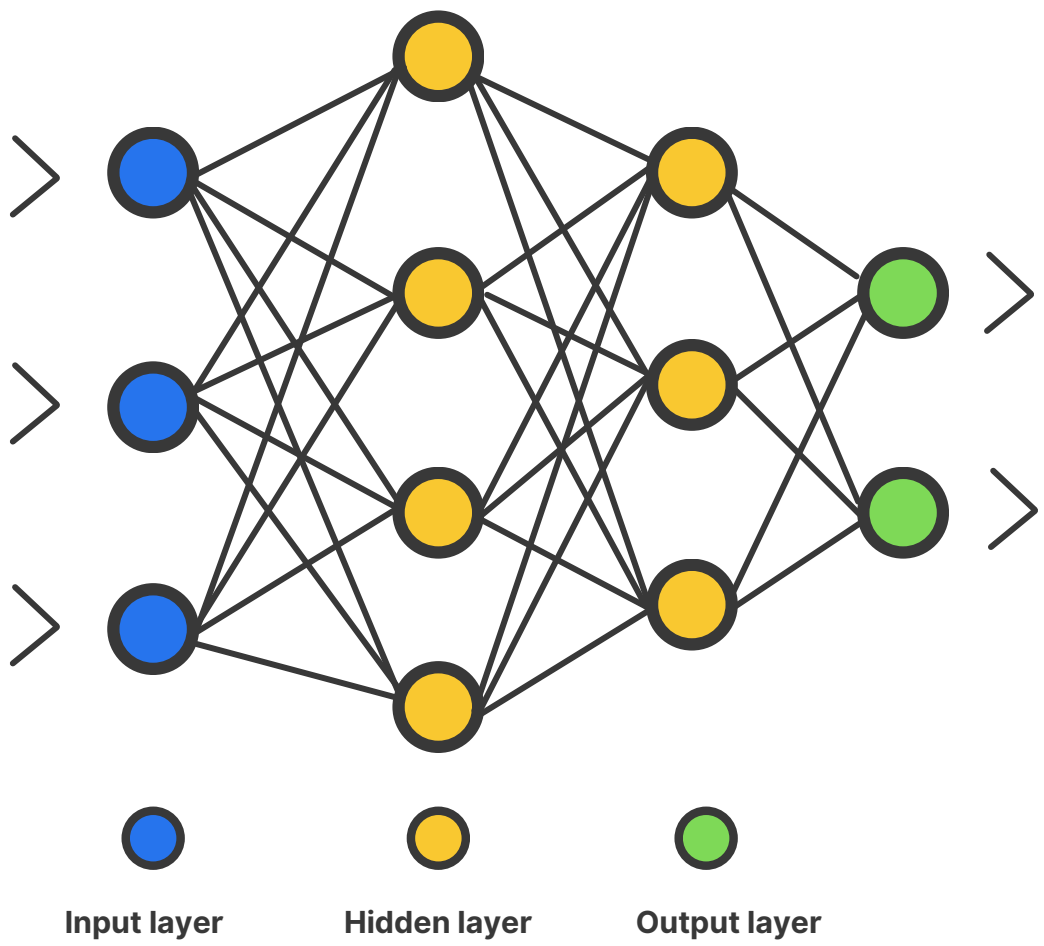
- Logistic regression doesn’t directly classify but uses a linear model to estimate the probability of an event (0-1). We then choose a threshold (like 50%) to convert this to categories like ‘yes’ or ‘no’. So, despite the ‘regression’ in its name, it ultimately tells us which class something belongs to.



Q3. What is the purpose of activation functions in neural networks?

- Activation functions introduce non-linearity to neural networks, allowing them to learn complex patterns and relationships in data. Without activation functions, neural networks would reduce to linear models, limiting their ability to capture intricate features. Popular activation functions include sigmoid, tanh, and ReLU, each introducing non-linearity at different levels.

- These non-linear transformations enable neural networks to approximate complex functions, making them powerful tools for image recognition and natural language processing.



Q4. If you do not know whether your data is scaled, and you have to work on the classification problem without looking at the data, then out of Random Forest and Logistic Regression, which technique will you use and why?

- In this scenario, Random Forest would be a more suitable choice. Logistic Regression is sensitive to the scale of input features, and unscaled features can affect its performance.

- On the other hand, Random Forest is less impacted by feature scaling due to its ensemble nature. Random Forest builds decision trees independently, and the scaling of features doesn't influence the splitting decisions across trees. Therefore, when dealing with unscaled data and limited insights, Random Forest would likely yield more reliable results.

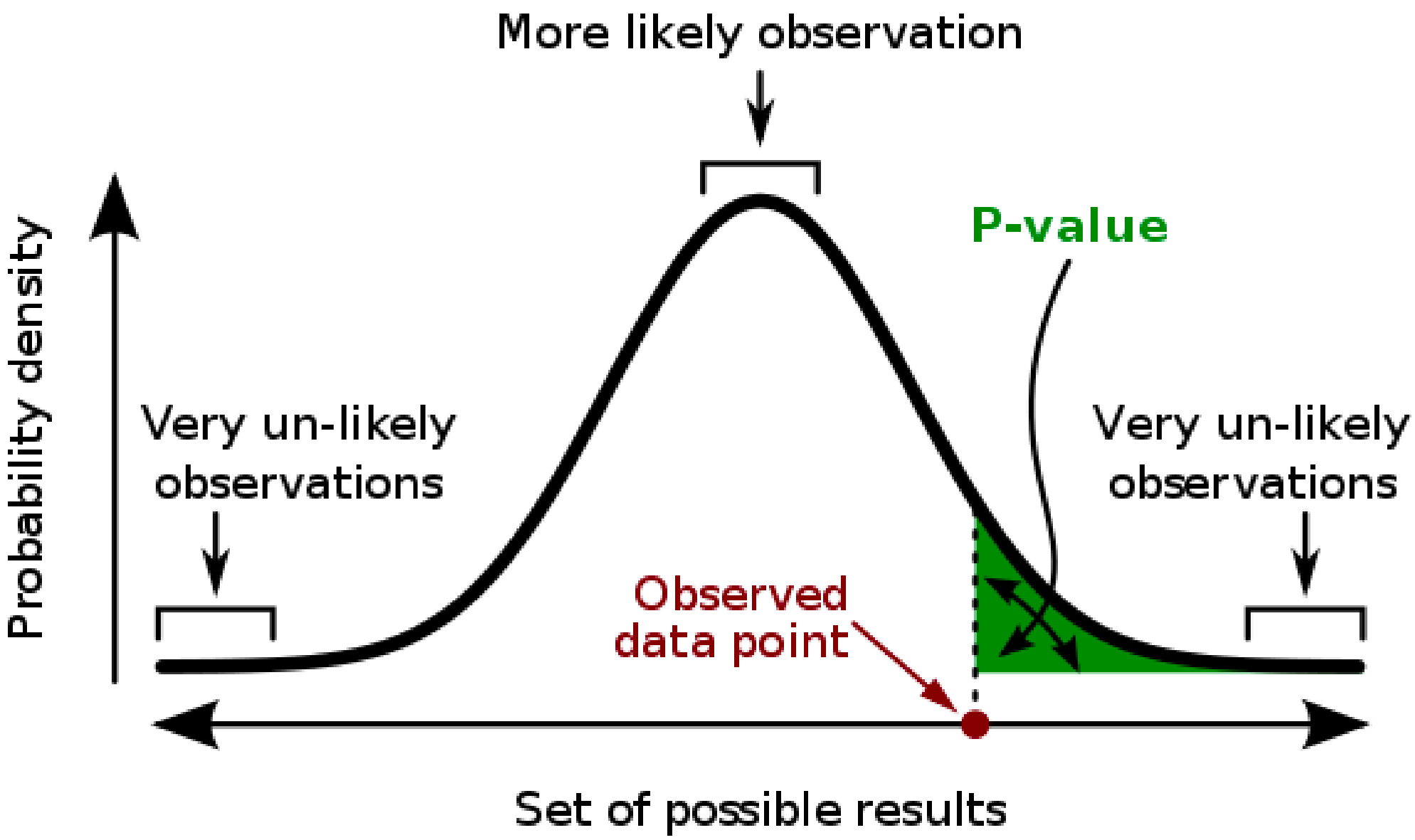
Q5. In a binary classification problem aimed at identifying cancer in individuals, if you had to prioritize one performance metric over the other, considering you don't want to risk any person's life, which metric would you be more willing to compromise on, Precision or Recall, and why?

- In identifying cancer, recall (sensitivity) is more critical than precision. Maximizing recall ensures that the model correctly identifies as many positive cases (cancer instances) as possible, reducing the chances of false negatives (missed cases).

- False negatives in cancer identification could have severe consequences. While precision is important to minimize false positives, prioritizing recall helps ensure a higher sensitivity to actual positive cases in the medical domain.

Q6. What is the significance of P-value when building a Machine Learning model?

- P-values are used in traditional statistics to determine the significance of a particular effect or parameter. P-value can be used to find the more relevant features in making predictions. The closer the value to 0, the more relevant the feature.



Q7. How does skewness in the distribution of a dataset affect the performance or behavior of machine learning models?

- Skewness in the distribution of a dataset can significantly impact the performance and behavior of machine learning models. Here's an explanation of its effects and how to handle skewed data:

- Effects of Skewed Data on Machine Learning Models:

    - **Bias in Model Performance**: Skewed data can introduce bias in model training, especially with algorithms sensitive to class distribution. Models might be biased towards the majority class, leading to poor predictions for the minority class in classification tasks.

    - **Impact on Algorithms**: Skewed data can affect the decision boundaries learned by models. For instance, in logistic regression or SVMs, the decision boundary might be biased towards the dominant class when one class dominates the other.

    - **Prediction Errors**: Skewed data can result in inflated accuracy metrics. Models might achieve high accuracy by simply predicting the majority class yet fail to detect patterns in the minority class.

Q8. Describe a situation where ensemble methods could be useful.

- Ensemble methods are particularly useful when dealing with complex and diverse datasets or aiming to improve a model's robustness and generalization.

- For example, in a healthcare scenario where diagnosing a disease involves multiple types of medical tests (features), each with its strengths and weaknesses, an ensemble of models, such as Random Forest or Gradient Boosting, could be employed.

- Combining these models helps mitigate individual biases and uncertainties, resulting in a more reliable and accurate overall prediction.

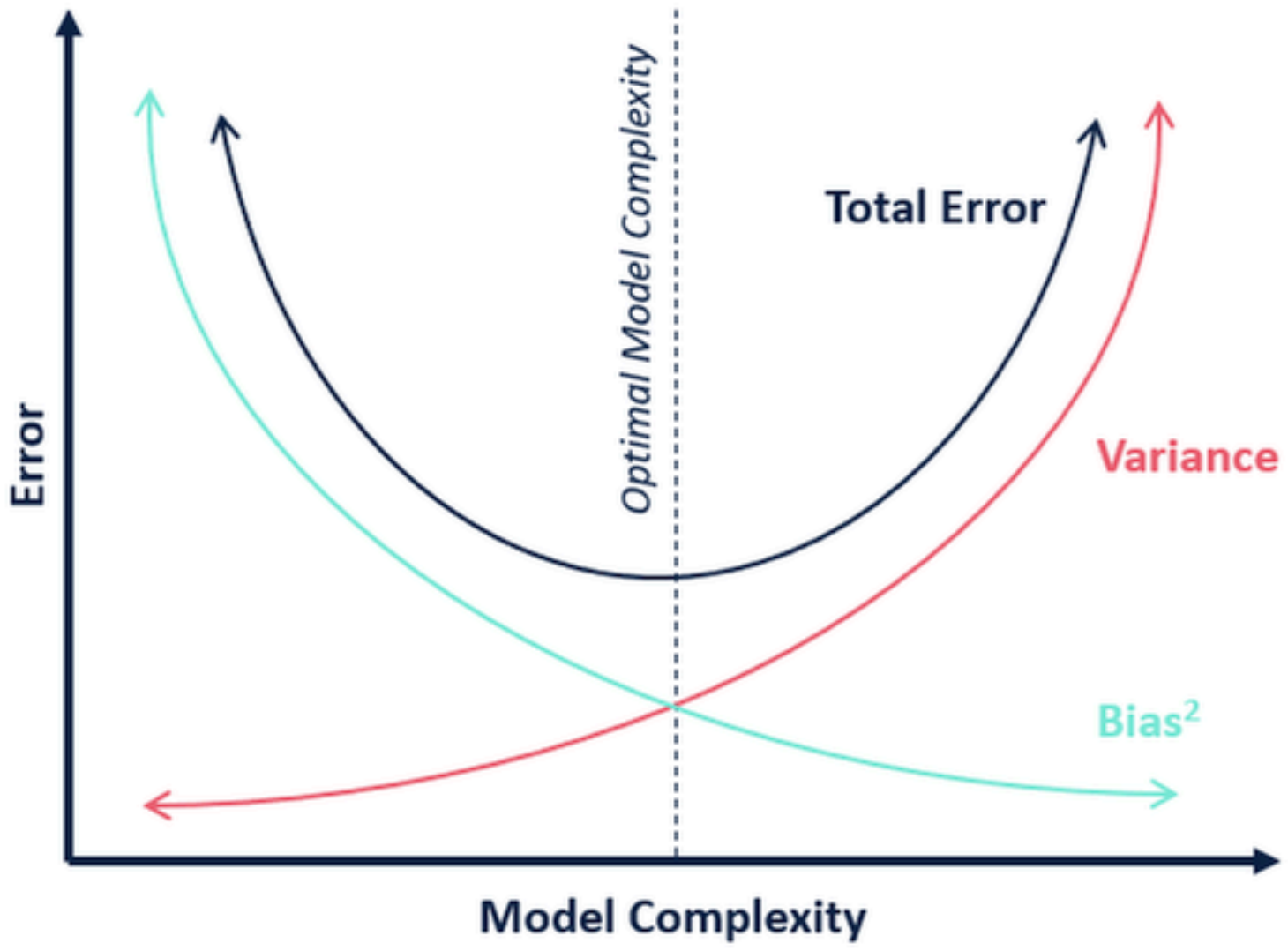
Q9. How would you detect outliers in a dataset?

Outliers can be detected using various methods, including:

- Z-Score: Identify data points with a Z-score beyond a certain threshold.
- IQR (Interquartile Range): Flag data points outside the 1.5 times the IQR range.
- Visualization: Plotting box plots, histograms, or scatter plots can reveal data points significantly deviating from the norm.
- Machine Learning Models: Outliers may be detected using models trained to identify anomalies, like one-class SVMs or Isolation Forests.

Q10. Explain the Bias-Variance Tradeoff in Machine Learning. How does it impact model performance?

The bias-variance tradeoff refers to the delicate balance between the error introduced by bias and variance in machine learning models. A model with high bias oversimplifies the underlying patterns, leading to poor performance in training and unseen data. Conversely, a model with high variance captures noise in the training data and fails to generalize to new data. Balancing bias and variance is crucial. Reducing bias often increases variance and vice versa. Optimal model performance is finding the right tradeoff to achieve low training and test data error.



For more information, visit the article

Intermediate Interview Prep Machine Learning

40 ML Interview Questions that You Must Know [2025]

Unlock the complexities of ML Interview Questions. From F1-scores to logistic regression, understand the fundamentals of this dynamic field.