Anomaly Detection & Time Series

1: What is Anomaly Detection? Explain its types (point, contextual, and collective anomalies) with examples

Anomaly detection is the process of identifying rare data points, events, or patterns that deviate significantly from the expected normal behaviour within a dataset. These deviations are known as anomalies or outliers and can signal potential problems, security threats, or fraud. The three main types of anomalies are: Point Anomalies, where a single data point is unusual on its own (e.g., a high-value purchase on a credit card); Contextual Anomalies, which are unusual only within a specific context (e.g., high network traffic during nighttime); and Collective Anomalies, where a group of data points, though normal individually, together deviate from the norm (e.g., a series of small, unusual network packet sequences).
Anomaly detection is a technique used in data analysis to find instances that do not conform to expected patterns. It involves establishing a baseline of normal behaviour and then identifying deviations from this baseline. This can help in various applications, such as: Detecting credit card fraud, identifying failing machines or server issues, Spotting cyberattacks, and Monitoring system performance and health.

Types of Anomalies

1. **Point Anomalies (or Outliers)**

- **Description:** A single data point that is significantly different from the rest of the dataset. It stands out on its own.

- **Example:** In a time series of a machine's CPU usage, a sudden spike to 90% when the average is 30% would be a point anomaly. Another example is an unusually large credit card transaction, which might be a point anomaly even if the transaction itself is legitimate, as it deviates from the usual spending pattern.

2. **Contextual Anomalies**

- **Description:** A data point that is considered anomalous only within a specific context, even though it might appear normal in another context.

- **Example:** A sudden increase in network traffic during regular business hours might be normal, but the same increase at 3 AM could signal a security issue. Similarly, a slight drop in water usage might be fine at night but an anomaly during peak hours in a residential area.

3. **Collective Anomalies**

- **Description:** A group of related data points that, when considered together, represent an anomaly. Individually, these points might appear normal.

- **Example:** In network security, a collective anomaly could be a series of small, unusual network packet sequences that, while each packet looks normal, their combination indicates a more complex system issue or an intrusion attempt. Another example could be a series of small, unusual spikes in sales that, when combined, form a pattern that doesn't fit typical market trends.

2: Compare Isolation Forest, DBSCAN, and Local Outlier Factor in terms of their approach and suitable use cases.

Isolation Forest is a tree-based model for anomaly detection that isolates outliers through random data partitioning and is efficient for large datasets and high dimensions. DBSCAN is a density-based clustering algorithm that identifies outliers by grouping dense data points and labeling sparse ones as noise, but it is sensitive to parameter tuning. Local Outlier Factor (LOF) is a density-based method that identifies outliers by comparing a point's local density to its neighbors' densities, which is effective for local anomalies but computationally expensive.

Isolation Forest

- **Approach:**

This algorithm builds an ensemble of isolation trees by randomly partitioning the data until outliers are isolated. Outliers are points that are easier to isolate, meaning they require fewer splits to be separated from the rest of the data.

- **Suitable Use Cases:**
    - Detecting global outliers in datasets with a large number of features.
    - Large datasets due to its speed and scalability compared to LOF.
    - No assumption about data distribution is needed.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Approach:**

DBSCAN groups together points that are closely packed, thus forming clusters of dense regions. Points that lie in low-density regions are considered noise or outliers.

- **Suitable Use Cases:**
    - Discovering clusters of varying shapes and sizes without requiring the number of clusters beforehand.
    - Datasets where outliers are points in sparse regions, creating a clear separation from dense clusters.

Local Outlier Factor (LOF)

- **Approach:**

LOF calculates a score based on the local density of each data point relative to its neighbors. Points with a significantly lower density than their neighbors are identified as outliers.

- **Suitable Use Cases:**
    - High-dimensional data.
    - Detecting local outliers, which are anomalies that deviate from their immediate neighborhood but might not be obvious globally.
    - Situations where computational cost is less of a concern, such as with smaller datasets.

3: What are the key components of a Time Series? Explain each with one example.

The four key components of a time series are Trend, Seasonality, Cyclicity, and Irregularity. Trend is the long-term upward or downward movement, like rising e-commerce sales over several years. Seasonality represents short-term, regular, and predictable patterns within a year, such as increased ice cream sales in summer. Cyclicity involves longer-term, non-fixed patterns that last longer than a year, such as fluctuations in a business cycle. Irregularity (or noise) accounts for unpredictable, random fluctuations that can't be explained by the other components, like a production dip due to a strike.

Here's a more detailed explanation of each component:

1. Trend

- **Description:**

The long-term general direction or movement of the data over an extended period, showing whether the data is generally increasing, decreasing, or staying stable.

- **Example:**

Observing the daily closing prices of a specific stock over five years might reveal an overall upward trend, indicating the stock's value is increasing over the long term.

2. Seasonality

- **Description:**

Regular, recurring patterns in the data that repeat at fixed intervals, typically within a year, such as daily, weekly, or monthly.

- **Example:**

A retail store will likely see a predictable surge in sales during the holiday season every year, demonstrating a seasonal pattern.

3. Cyclicity

- **Description:**

Repeated fluctuations in the data that have a period longer than one year but don't have a fixed length or magnitude, unlike seasonality.

- **Example:**

Economic recessions and expansions are examples of cyclical patterns in a time series, with peaks and troughs that may last for several years but without a consistent duration.

4. Irregularity (or Randomness/Noise)

- **Description:**

Unpredictable, random fluctuations or "noise" in the data that cannot be explained by the trend, seasonality, or cyclical components.

- **Example:**

A sudden, unexpected drop in a company's quarterly sales due to a natural disaster or a pandemic would be an irregular component.

4: Define Stationary in time series. How can you test and transform a non-stationary series into a stationary one?

A time series is stationary if its statistical properties, like mean and variance, are constant over time, meaning it's not affected by the specific time it's viewed. You can test for stationarity using the Augmented Dickey-Fuller (ADF) test or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. To transform a non-stationary series into a stationary one, common methods include differencing to remove trends and seasonality, and applying transformations like the Box-Cox transformation to stabilize variance.

Defining Stationarity

A time series is stationary if its statistical properties remain constant over time:

- **Constant Mean:** The average value of the series does not change over time.

- **Constant Variance:** The spread or volatility of the data remains consistent across the series.

- **No Trend or Seasonality:** There are no systematic upward or downward movements (trends) or repetitive patterns at fixed intervals (seasonality).

Testing for Stationarity

Two common statistical tests are used to determine if a series is stationary:

1. Augmented Dickey-Fuller (ADF) Test:

- **Null Hypothesis (H_0):** The time series is non-stationary (has a unit root).

- **Interpretation:** If the p-value is less than your chosen significance level (e.g., 0.05), you reject the null hypothesis, indicating the series is stationary.

2. **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test:**

- **Null Hypothesis (H_0):** The time series is stationary.

- **Interpretation:** If the p-value is greater than your significance level, you fail to reject the null hypothesis, meaning the series is stationary.

Transforming a Non-Stationary Series

If a series is found to be non-stationary, you can apply transformations to make it stationary:

1. **Differencing:**

- **How it works:** You calculate the difference between consecutive observations (e.g., Y_t - Y_{t-1}).

- **Purpose:** This can effectively remove trends by stabilizing the mean and can also help eliminate seasonality if differences are taken appropriately (e.g., yearly differences for annual data).

2. **Logarithmic Transformation:**

- **How it works: Apply the natural logarithm to the data.**

- **Purpose:** This transformation can help stabilize the variance of a time series, especially if the variance increases with the level of the series.

3. **Box-Cox Transformation:**

- **How it works:** A more general power transformation that includes logarithmic transformation as a special case.

- **Purpose:** It can effectively stabilize both the mean and variance of a series by making it more normally distributed and constant over time.

5: Differentiate between AR, MA, ARIMA, SARIMA, and SARIMAX models in terms of structure and application.

AR (Autoregressive) models predict future values using past values, MA (Moving Average) models use past forecast errors, ARIMA adds differencing to handle trends, SARIMA incorporates seasonality, and SARIMAX further includes exogenous (external) variables. AR models use past values to forecast the present, MA models adjust predictions based on past errors, and ARIMA extends ARMA by including a differencing (I) component to handle non-stationary data. SARIMA adds seasonal differencing (S) to capture cyclical patterns, while SARIMAX expands on this by allowing the inclusion of external factors that may influence the time series.

Here's a breakdown of each model's structure and application:

1. AR (Autoregressive) Model

- **Structure:** Models a time series using a linear combination of its own past values.
- **Application:** Used for forecasting when the current value of a series is a function of its historical values.

2. MA (Moving Average) Model

- **Structure:** Forecasts future values by modelling the relationship between the current value and past errors in the forecast.
- **Application:** Effective when past forecast errors strongly influence future predictions, often used in conjunction with AR models (as in an ARMA model).

3. ARIMA (Autoregressive Integrated Moving Average) Model

- **Structure:** Combines Autoregressive (AR) and Moving Average (MA) components with an "Integrated" (I) component for differencing to make non-stationary time series data stationary.
- **Application:** Applied to data with non-seasonal trends and fluctuations but without strong cyclical patterns.

4. SARIMA (Seasonal Autoregressive Integrated Moving Average) Model

- **Structure:** Extends ARIMA by adding seasonal components for both the non-seasonal and seasonal parts, including seasonal differencing.
- **Application:** Ideal for time series data exhibiting clear and repeating seasonal patterns, such as monthly sales data or daily temperature readings.

5. SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) Model

- **Structure:** A SARIMA model enhanced to incorporate one or more additional explanatory (exogenous) variables that can influence the time series.
- **Application:** Used when other factors, besides the historical pattern and seasonality of the series itself, are believed to have a significant impact on future values. For instance, in forecasting electricity demand, this model could include factors like temperature or special events.
