VIVA INTERVIEW QUESTIONS

1. What is forecasting in the context of machine learning?

- Forecasting in machine learning refers to predicting future values based on historical data patterns. It is commonly applied in time-series data, where the goal is to make accurate predictions for future time steps. Forecasting can be used in areas such as demand forecasting, stock price prediction, and weather forecasting.

2. How does time-series forecasting differ from traditional machine learning predictions?

- In time-series forecasting, the data points are sequential and time-dependent, meaning that past values directly influence future values. Traditional machine learning predictions often assume data points are independent of one another, whereas in time-series forecasting, the temporal order and dependencies matter. Features like trend, seasonality, and autocorrelation are considered in time-series models.

3. What are some common algorithms used for time-series forecasting?

- Some widely used algorithms for time-series forecasting include:

- ARIMA (AutoRegressive Integrated Moving Average)

- SARIMA (Seasonal ARIMA)

- Exponential Smoothing (ETS)

- Prophet (from Facebook)

- Long Short-Term Memory (LSTM) neural networks

- XGBoost and Random Forests for time-series using feature engineering

4. How do you handle seasonality and trend in time-series data?

- Seasonality refers to periodic patterns that repeat over a fixed time period, while trend is the long-term upward or downward movement in data.

- To handle seasonality, models like SAR IMA and Prophet are often used as they capture seasonality explicitly.

- Decomposition methods are also used to split the data into trend, seasonality, and residual components.

- Detrending and Deseasonalizing the data can also help remove these components before applying forecasting models.

5. What is the role of stationarity in time-series forecasting?

- Stationarity means that the statistical properties of a time-series (e.g., mean, variance) are constant over time. Many forecasting models, such as ARIMA, assume stationarity. If the data is not stationary, techniques like differencing or transformation (e.g., log transformation) are applied to achieve stationarity.

6. Explain the difference between autoregressive (AR), moving average (MA), and ARIMA models.

- AR (AutoRegressive): A model where future values are regressed on previous values (lags) of the same time series.

- MA (Moving Average): A model where future values are modeled as a function of the past error terms (residuals).

- ARIMA (AutoRegressive Integrated Moving Average): Combines AR and MA models, along with a differencing step (I) to make the data stationary. ARIMA is commonly used for non-stationary time-series data.

7. What is overfitting in forecasting models, and how can it be prevented?

- Overfitting occurs when a forecasting model captures noise or random fluctuations in the training data, leading to poor performance on unseen data.

- Overfitting can be prevented by:

- Cross-validation (e.g., time-series cross-validation)

- Simplifying the model (reducing complexity)

- Regularization techniques like Ridge or Lasso for regression models

- Pruning or limiting the depth of tree-based models

8. How do you evaluate the accuracy of a forecasting model?

- Common metrics used to evaluate forecasting accuracy include:

- MAE (Mean Absolute Error): Measures the average magnitude of the errors.

- RMSE (Root Mean Squared Error): Measures the square root of the average squared differences between predicted and actual values.

- MAPE (Mean Absolute Percentage Error): Measures the percentage error between actual and predicted values.

- R² (Coefficient of Determination): Measures the proportion of variance explained by the model.

- SMAPE (Symmetric Mean Absolute Percentage Error): An adjusted version of MAPE.

9. What is the difference between univariate and multivariate time-series forecasting?

- Univariate time-series forecasting deals with predicting future values of a single time series based on its own past data.

- Multivariate time-series forecasting involves using multiple time-series or exogenous variables to predict future values. These additional variables can help improve forecast accuracy by capturing more complex relationships between different time series.

10. How do exogenous variables impact a time-series forecasting model?

- Exogenous variables (also called covariates or external factors) are variables that influence the dependent variable but are not part of its past values. Including exogenous variables in time-series forecasting models like ARIMAX or SARIMAX can improve accuracy by providing additional information. For example, adding weather data to a sales forecast could improve predictions if sales are weather-sensitive.

EDA FORECASTING

1. What is the structure of the dataset, and why is it important?

- The structure of a dataset refers to the organization of data, including the rows and columns, the size of dataset [3095,5 ]the types of variables (e.g., 4 are numerical, 1 is categorical), and the time stamps if it’s a time-series dataset. Understanding the structure is crucial because it helps you determine how the data will be processed, which columns are relevant for analysis, and whether any pre-processing (like converting date formats or handling missing values) is required before forecasting.

2. How do you handle missing values in time series data?

- In time-series data, missing values can distort the analysis, especially in forecasting. Some common methods to handle missing values include:

- Forward or backward filling: Replacing missing values with the preceding or succeeding values.

- Interpolation: Estimating missing values based on trends or patterns in the data.

- Dropping missing values: This is generally avoided unless the missing values are sparse.

- Imputation using statistical methods: Filling missing values using methods like mean, median, or more advanced algorithms like Kalman filtering.

3. How do you check for trends and seasonality in time series data?

- Trends and seasonality can be detected using several methods:

- Visual inspection: Plotting the data over time can help reveal upward or downward trends and seasonal patterns.

- Decomposition: Time-series decomposition splits the data into trend, seasonal, and residual components.

- Moving averages: Smoothing the data using moving averages helps highlight long-term trends.

- Autocorrelation plots: Seasonal patterns can often be identified through peaks in autocorrelation at specific time lags.

4. What is stationarity, and how do you test for it?

- Stationarity refers to a time series whose statistical properties, like mean and variance, do not change over time. It is crucial for many forecasting models. You can test for stationarity using:

- Augmented Dickey-Fuller (ADF) test: A formal statistical test for stationarity.

- KPSS test: Another test used to determine the stationarity of a time series.

- Visual methods: Plotting the rolling mean and variance can give a rough idea of whether the series is stationary.

- Differencing: A method used to make a series stationary by subtracting previous values.

5. How do you detect outliers in time series data?

- Outliers can be detected using various techniques:

- Visual inspection: Plotting the time series can help identify points that deviate significantly from the norm.

- Z-scores: Calculating the Z-score for each data point can help detect extreme outliers.

- Boxplots: Boxplots show the spread and potential outliers in the data.

- Autocorrelation plots: Sudden spikes in the autocorrelation plot can indicate potential outliers.

- Statistical methods: Using models like ARIMA can help detect outliers as residuals from the model.

6. What are autocorrelation and partial autocorrelation, and how are they useful in forecasting?

- Autocorrelation measures the correlation between a time series and a lagged version of itself. It helps identify repeating patterns or seasonality in data.

- Partial Autocorrelation shows the correlation between a time series and its lag, controlling for the effects of intervening lags. It helps in determining the order of autoregressive (AR) terms in models like ARIMA.

- Both are useful in selecting the appropriate lag terms in time-series models, like ARIMA, by showing which previous values of the series are most relevant for forecasting future values.

7. How do you handle seasonality in time series data?

- Seasonality can be handled by:

- Seasonal decomposition: Separating the seasonal component from the data using methods like STL decomposition.

- Seasonal differencing: Subtracting values from previous periods (seasonal lags) to remove seasonality.

- SARIMA models: These are ARIMA models that include seasonal components.

- Fourier series: Representing the seasonal component using sine and cosine terms.

- Dummy variables: Creating dummy variables for different seasons (like months or quarters).

8. Why is feature engineering important in forecasting, and how can it be done?

- Feature engineering is important because it enhances the predictive power of forecasting models by creating new features that better capture patterns in the data. Feature engineering in time series can be done by:

- Lag features: Adding past values (lags) of the target variable as features.

- Rolling statistics: Calculating rolling means, variances, or other statistics over different time windows.

- Seasonal indicators: Including dummy variables for holidays, weekends, or specific months to capture seasonality.

- Exogenous variables: Incorporating external factors such as weather or economic indicators that influence the target variable.

9. How do you assess the correlation between explanatory variables and the target variable?

- Correlation between explanatory variables (features) and the target variable can be assessed by:

- Pearson or Spearman correlation: These correlation coefficients show the strength and direction of the relationship.

- Autocorrelation: For time-series data, you can assess how previous values of the explanatory variables influence the target variable over time.

- Feature importance scores: Algorithms like Random Forest or XGBoost provide feature importance scores that indicate how strongly each feature correlates with the target variable.

- Cross-correlation plots: For time series, cross-correlation functions (CCF) can show correlations between two time series at different lags.

10. How do external factors (holidays, promotions, weather) influence forecasting?

- External factors, also known as exogenous variables, can significantly impact the target variable in forecasting. For example:

- Holidays can affect sales forecasts, leading to spikes or dips in demand.

- Promotions can temporarily increase sales or demand for products.

- Weather can impact industries like agriculture, retail, and energy.

- These external factors are typically added as features in the forecasting model to improve accuracy. Models like ARIMAX or SARIMAX are designed to include external variables as regressors.

1. What is the overall trend of the target variable over time?
   * The overall trend can be assessed by plotting the target variable over time. Use line charts to detect whether there is an upward or downward movement in the data. This visual inspection helps identify long-term trends.
2. Is there seasonality present in the data?
   * Seasonality can be detected by looking for regular patterns at consistent intervals. If present, this should influence your choice of model, as some models handle seasonality better than others.
3. How does the target variable correlate with other features (lagged features)?
   * By analyzing the relationship between the target variable and its lagged versions (previous time steps), you can see if past values have an impact on future outcomes. This is particularly useful for defining autoregressive components.
4. What is the distribution of data across different time periods?
   * You should analyze how the data varies over different intervals, such as the day of the week or the month of the year, to determine if there is cyclical behavior or seasonality.
5. Are there any patterns in the residuals after removing trends and seasonality?
   * After decomposing the time series into trend, seasonality, and residual components, you can analyze the residuals for randomness. Any remaining patterns could indicate that the model is missing important components.
6. Which type of model is most suitable for the data (ARIMA, SARIMA, Exponential Smoothing, etc.)?
   * The choice of model depends on the stationarity, seasonality, and other characteristics of the data. Linear models like ARIMA work well for stationary data, while exponential smoothing methods (ETS) can handle data with a trend or seasonality.
7. How do lagged features improve model accuracy?
   * Lagged features help in defining autoregressive models by using previous time steps to predict future values, improving model accuracy when past values strongly influence future outcomes.
8. How does the model perform on validation data?
   * Cross-validation or a train-test split is used to evaluate how well the model generalizes to unseen data. This is essential to ensure the model is not overfitting and performs well on real-world data.
9. Does the model capture seasonality and trend adequately?
   * You need to confirm that the model captures both seasonality and trends seen in exploratory data analysis (EDA). Models like SARIMA are explicitly designed to include seasonal components.
10. How can hyperparameters be optimized for better performance?
    * Hyperparameter tuning, such as adjusting the p, d, and q parameters in ARIMA or the smoothing factors in exponential smoothing methods, is necessary to improve forecasting accuracy.

### 

1. What is the difference between time series forecasting and regression analysis?
   * Time series forecasting predicts future values based on past data, with a focus on temporal patterns (trends, seasonality). Regression analysis, on the other hand, predicts values based on relationships between dependent and independent variables, without emphasizing time-dependent patterns.
2. What is the significance of differencing in time series analysis?
   * Differencing helps in achieving stationarity by removing trends from the data. By differencing the data (subtracting current values from previous values), the model becomes easier to handle and more accurate.
3. What is a seasonal decomposition of time series?
   * Seasonal decomposition splits a time series into three components: trend (long-term movement), seasonality (cyclical patterns), and residuals (random noise). This helps in understanding the underlying structure of the data.
4. Explain the term 'moving average' in forecasting.
   * A moving average is used to smooth time series data by averaging subsets of data points. This helps in reducing noise and highlighting the underlying trends in the series.
5. What is the importance of the 'p' parameter in the ARIMA model?
   * The 'p' parameter represents the number of lag observations included in the model (autoregressive terms). It defines how many past values should be used to predict the future.
6. How would you handle a multicollinearity issue in time series forecasting?
   * To handle multicollinearity, you can use techniques like removing highly correlated variables, applying regularization methods (e.g., Ridge or Lasso regression), or using principal component analysis (PCA) to reduce dimensionality.
7. What is exponential smoothing?
   * Exponential smoothing gives more weight to recent observations, making it better suited for time series data with trends. It differs from simple moving averages by accounting for recent changes in the data more effectively.
8. What is overfitting in forecasting models?
   * Overfitting occurs when a model is too complex and captures noise instead of the underlying pattern. This can be avoided by using regularization techniques, cross-validation, and ensuring the model's complexity matches the data's complexity.
9. Explain the role of exogenous variables (X) in SARIMAX.
   * Exogenous variables are additional independent variables included in the SARIMAX model to improve predictions. These variables help capture effects not explained by the time series itself.
10. What is the Box-Jenkins methodology?
    * The Box-Jenkins approach is used for developing ARIMA models. It involves identifying the model order (p, d, q), estimating model parameters, and checking residuals to ensure the model fits well. This iterative process ensures a robust time series model.
11. What is the overall trend of the target variable over time?
    * The overall trend can be assessed by plotting the target variable over time. Line charts can help detect upward or downward movements in the data, which helps identify long-term trends.
12. Is there seasonality present in the data?
    * Seasonality can be identified by looking for regular patterns at consistent intervals. If seasonality is present, it should be considered when selecting a model, as some models handle seasonality better than others.
13. How does the target variable correlate with other features (lagged features)?
    * The target variable's correlation with other (lagged) features can be analyzed by checking its relationship with its lagged versions (previous time steps). This is useful for defining autoregressive components in a model, especially if past values influence future ones.
14. What is the distribution of data across different time periods?
    * The distribution of data across different time periods can be analyzed by looking at data variation over intervals like days of the week or months of the year. This helps to determine if there is cyclical behavior or seasonality.
15. Are there any patterns in the residuals after removing trends and seasonality?
    * After removing trends and seasonality, you can analyze the residuals for any remaining patterns. Residuals should ideally be random; patterns in them suggest that the model is missing important components.
16. Which type of model is most suitable for the data (ARIMA, SARIMA, Exponential Smoothing, etc.)?
    * The choice of model (ARIMA, SARIMA, Exponential Smoothing, etc.) depends on the characteristics of the data, such as stationarity and seasonality. Linear models like ARIMA are suitable for stationary data, while exponential smoothing methods can capture trends or seasonality.
17. How do lagged features improve model accuracy?
    * Lagged features help improve model accuracy by using previous time steps to predict future outcomes. This is particularly useful when past values have a strong influence on future ones.
18. How does the model perform on validation data?
    * The model's performance on validation data is tested using cross-validation or train-test splits. This ensures that the model generalizes well to unseen data and isn't overfitting.
19. Does the model capture seasonality and trend adequately?
    * To ensure the model captures both seasonal and trend components, you can verify this during the exploratory data analysis phase. Models like SARIMA are specifically designed to capture seasonal components.
20. How can hyperparameters be optimized for better performance?
    * Hyperparameters can be optimized by tuning key parameters like p, d, q in ARIMA or smoothing factors in exponential smoothing models. This helps in improving the accuracy of the model's forecasts.
21. What is the difference between time series forecasting and regression analysis?
    * Time series forecasting predicts future values based on past data with an emphasis on temporal patterns like trends and seasonality. In contrast, regression analysis predicts values based on relationships between dependent and independent variables without focusing on time-dependent patterns.
22. What is the significance of differencing in time series analysis?
    * Differencing in time series analysis helps in achieving stationarity by removing trends from the data. Differencing the data (subtracting current values from previous values) simplifies the model and makes it more accurate.
23. What is a seasonal decomposition of time series?
    * Seasonal decomposition breaks down a time series into three components: trend (long-term movement), seasonality (cyclical patterns), and residuals (random noise). This provides a clearer understanding of the data structure.
24. Explain the term 'moving average' in forecasting.
    * A moving average smooths time series data by averaging subsets of data points. This reduces noise and highlights underlying trends in the series.
25. What is the importance of the 'p' parameter in the ARIMA model?
    * The p parameter in ARIMA refers to the number of lag observations included in the model, which defines how many past values are used to predict future outcomes.
26. How would you handle a multicollinearity issue in time series forecasting?
    * To handle multicollinearity, you can remove highly correlated variables, apply regularization methods (such as Ridge or Lasso regression), or use principal component analysis (PCA) to reduce dimensionality.
27. What is exponential smoothing?
    * Exponential smoothing gives more weight to recent observations, making it better suited for time series with trends. It differs from moving averages by accounting more effectively for recent changes.
28. What is overfitting in forecasting models?
    * Overfitting happens when a model is too complex and captures noise rather than the true underlying patterns. This can be prevented by using regularization techniques, cross-validation, and ensuring that the model complexity matches the data.
29. Explain the role of exogenous variables (X) in SARIMAX.
    * Exogenous variables in SARIMAX are additional independent variables that improve predictions by capturing effects not explained by the time series itself.
30. What is the Box-Jenkins methodology?
    * The Box-Jenkins methodology is a systematic approach for developing ARIMA models. It involves identifying the order of the model, estimating parameters, and checking residuals to ensure a good fit. This iterative process ensures robust time series models.
31. What are the key differences between time series forecasting and traditional regression tasks in machine learning?
    * Time series forecasting focuses on temporal dependencies, capturing trends and seasonality, while regression models focus on relationships between dependent and independent variables.
32. Which preprocessing techniques are essential for preparing time series data before feeding it into forecasting models (e.g., handling missing values, normalization)?
    * Essential preprocessing techniques for time series data include handling missing values, normalizing the data, and transforming it to address issues like non-stationarity. These steps ensure the data is suitable for time series models.
33. What role does seasonality play in time series forecasting, and how can models capture seasonal trends?
    * Seasonality represents recurring patterns at regular intervals. Models like SARIMA or Prophet capture these seasonal trends explicitly by incorporating seasonality as a parameter.
34. How do lag features or time-based features improve the accuracy of time series forecasting models?
    * Lag features improve time series forecasting models by using past values to predict future outcomes. Time-based features such as day of the week, month, or year can also enhance model accuracy by capturing cyclical trends.
35. Which machine learning algorithms are commonly used for forecasting tasks, and how do they compare in performance (e.g., ARIMA, Prophet, LSTM, Random Forest)?
    * Commonly used machine learning algorithms for forecasting include ARIMA, Prophet, LSTM, and Random Forest. ARIMA is effective for linear trends, while LSTM and neural networks handle complex patterns. Random Forest can capture interactions between variables.
36. How can you handle non-stationary time series data, and why is stationarity important for some forecasting models?
    * Non-stationary data can be handled through techniques like differencing, transforming, or detrending. Stationarity is important for models like ARIMA, which assume that the data's statistical properties remain constant over time.
37. What are the advantages and limitations of using neural networks, such as LSTMs or GRUs, for time series forecasting compared to classical models like ARIMA?
    * Neural networks like LSTMs and GRUs can capture complex temporal patterns and nonlinear relationships, unlike classical models like ARIMA, which rely on linearity. However, neural networks require more data and are more complex to train.
38. How do you select the optimal model evaluation metric (e.g., RMSE, MAE, MAPE) for a time series forecasting problem?
    * The choice of model evaluation metric depends on the forecasting objective. RMSE and MAE are commonly used for continuous data, while MAPE is suitable for percentage error calculations. The selected metric should align with business goals.
39. What are some common data leakage risks when splitting time series data into training and testing sets, and how can they be mitigated?
    * Data leakage can occur when information from the test set is inadvertently used during training. This can be mitigated by ensuring proper temporal separation between training and testing sets or using techniques like rolling cross-validation.
40. What is the role of hyperparameter tuning in forecasting models, and how can techniques like cross-validation be applied to time series data?
    * Hyperparameter tuning improves model performance by optimizing parameters. Cross-validation techniques for time series, like rolling-window or expanding-window validation, ensure that models are tested on unseen data from future periods.

### 

1. What is the difference between qualitative and quantitative forecasting methods, and when should each be used?
   * Qualitative forecasting methods rely on expert judgment and opinion, while quantitative methods use mathematical models based on historical data. Qualitative methods are used when data is scarce or trends are hard to quantify, while quantitative methods are used when historical data is available.
2. How do time series models like ARIMA (AutoRegressive Integrated Moving Average) work in forecasting, and what are their main components?
   * ARIMA models work by combining autoregressive (AR), integrated (I), and moving average (MA) components. These components work together to capture trends, seasonality, and noise in the data.