**Capstone Project 2: Web Traffic Prediction**

**Project Statement:**

Accurate forecasting of web traffic is crucial for effective resource allocation, capacity planning, and ensuring a seamless user experience. This project aims to develop a time series forecasting model to predict future web traffic based on historical data. The provided dataset “web traffic.csv” contains time-stamped web traffic counts, which will be used to analyze traffic patterns and build a predictive model. The student will implement the following steps:

**Phase 1: Time Series Preprocessing and Exploration**

1. **Data Loading and Initial Inspection [5 Marks]:**
   * Load the “web traffic.csv” dataset into a panda’s DataFrame.
   * Convert the 'Timestamp' column to a datetime format.
   * Set the 'Timestamp' column as the index of the DataFrame.
2. **Time Series Visualization :**
   * Plot the 'TrafficCount' against the 'Timestamp' to visualize the raw time series.
   * Decompose the time series into its trend, seasonal, and residual components using an appropriate method (e.g., seasonal decomposition using moving averages or the seasonal\_decompose function from the statsmodels library).
   * Plot the decomposed components (trend, seasonality, residual) separately and interpret the observed patterns. Identify any apparent trends, seasonality, or irregular fluctuations.
3. **Autocorrelation Analysis:**
   * Plot the Autocorrelation Function (ACF) of the 'TrafficCount’ series.
   * Plot the Partial Autocorrelation Function (PACF) of the 'TrafficCount' time series.
   * Analyze the ACF and PACF plots to identify potential autoregressive (AR) and moving average (MA) components in the time series. Explain how the patterns in the ACF and PACF plots relate to the underlying time series characteristics.
4. **Stationarity Check:**
   * Explain the concept of stationarity in time series analysis and its importance for modeling.
   * Perform the Augmented Dickey-Fuller (ADF) test to check for stationarity of the 'TrafficCount' time series.
   * State the null and alternative hypotheses of the ADF test.
   * Interpret the results of the ADF test (test statistic and p-value).
   * If the time series is not stationary, apply appropriate differencing to make it stationary. Repeat the ADF test on the differenced series to confirm stationarity.
   * Clearly state the order of differencing required (d value).

**Phase 2: SARIMA Model Building and Evaluation**

1. **SARIMA Model Identification:**
   * Based on the analysis of the ACF and PACF plots (from Phase 1) and the stationarity analysis (including the order of differencing), determine the initial values for the SARIMA model parameters:
     + p (autoregressive order)
     + d (differencing order) - from Phase 1
     + q (moving average order)
     + P (seasonal autoregressive order)
     + D (seasonal differencing order)
     + Q (seasonal moving average order)
     + s (seasonal period) - Determine the seasonality (e.g., daily, weekly).
   * Explain the reasoning behind your choice of initial parameter values, linking them to the observed time series characteristics, ACF/PACF patterns, and seasonality.
   * Consider a range of plausible values around your initial estimates.
2. **Model Training and Selection:**
   * Split the time series data into training and testing sets. A common split is 80% for training and 20% for testing but justify your choice. Ensure the split maintains the time series order.
   * Train several SARIMA models with different combinations of the parameters (p, d, q, P, D, Q, s) identified in the previous step. You can use a grid search or iterative approach to explore different parameter combinations.
   * For each trained model, record the AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) value.
   * Select the SARIMA model with the lowest AIC or BIC value as the best model. Justify your model selection based on the information criterion used. Explain the trade-off between model fit and complexity that AIC/BIC balances.
3. **Model Evaluation:**
   * Use the selected SARIMA model to generate predictions on the testing set.
   * Calculate the following evaluation metrics to assess the model's performance:
     + Mean Squared Error (MSE)
     + Root Mean Squared Error (RMSE)
     + Mean Absolute Error (MAE)
   * Plot the original time series, the training data, and the predicted values on the testing set in a single plot. Visually assess the model's fit and predictive accuracy.
   * Interpret the evaluation metrics and the plot. Discuss the model's strengths and weaknesses in forecasting web traffic.
4. **Residual Analysis [5 Marks]:**
   * Calculate the residuals (the difference between the actual and predicted values) from the selected SARIMA model on the training data.
   * Plot the residuals over time.
   * Create a histogram or density plot of the residuals.
   * Plot the ACF and PACF of the residuals.
   * Perform the Ljung-Box test to check if the residuals are white noise (i.e., randomly distributed with no autocorrelation).
   * Interpret the results of the residual analysis. Discuss whether the residuals exhibit any patterns or autocorrelation, which would indicate that the model is not capturing all the information in the time series.

**Phase 3: Conclusion**

1. **Conclusion:**
   * Summarize the key steps taken in the project, from data preprocessing and exploration to model building and evaluation.
   * State the final SARIMA model parameters and its performance on the testing data.