Football Goals Scoring Prediction using ARIMA Model

Objective

The primary objective of this project is to predict the future performance of the **England football team** in terms of **goals scored** using historical match data. The ARIMA (AutoRegressive Integrated Moving Average) model is employed to forecast the number of goals scored over the next several years.

Data

The dataset used in this project contains historical football match data, including:

- **Home Team**: The team playing at home.
- **Away Team**: The team playing away.
- **Home Score**: The number of goals scored by the home team.
- **Away Score**: The number of goals scored by the away team.
- **Date**: The date of the match.
- **Tournament**: The type of match (e.g., Friendly).
- **City**: The city where the match took place.
- **Country**: The country where the match took place.
- **Neutral**: Indicates whether the match was played at a neutral venue.

For this project, we focused on predicting goals scored by **England** based on both home and away performance.

Steps Involved

1. Data Preprocessing

The dataset was prepared by:

- **Date Conversion**: Converting the `date` column into the Date format to enable proper time-based aggregation.
- **Score Conversion**: Converting `home_score` and `away_score` columns into numeric values to ensure accurate calculations.
- **Time Series Creation**: A new column `year` was created by extracting the year from the `date` column, representing the year of each match.

2. Aggregating Goals Scored

England's performance was aggregated by:

- **Home Goals**: Total number of goals scored by England when playing at home.
- **Away Goals**: Total number of goals scored by England when playing away.
- **Total Goals**: Sum of home and away goals scored by England each year.

The aggregated goals for both home and away matches were combined for each year to create a time series of **total goals scored**.

3. Time Series Construction

Using the aggregated goals per year, the data was converted into a time series object in R using the `ts()` function. This time series object was essential for time series modeling, which is at the core of this analysis.

4. Stationarity Check

Before applying ARIMA, it is necessary to ensure that the data is **stationary**, meaning it has constant statistical properties over time. A **Dickey-Fuller test (ADF test)** was conducted to check the stationarity of the time series:

- **Non-stationarity** was confirmed initially, and we applied **differencing** to stabilize the mean and variance, making the series stationary.

5. ARIMA Modeling

The ARIMA model was selected based on the **autocorrelation (ACF)** and **partial autocorrelation (PACF)** plots. The `auto.arima()` function in R automatically selected the optimal ARIMA model by evaluating different combinations of AR (AutoRegressive), I (Integrated), and MA (Moving Average) components.

- The model fit the time series data, and the parameters were estimated.

6. Model Evaluation and Diagnostics

6.1 Residuals Analysis

Residual analysis was performed to evaluate the goodness of fit. Residuals should resemble **white noise**, implying no significant patterns.

- **ACF and PACF plots of residuals** were examined to ensure no significant autocorrelation remained.
- A **histogram of residuals** was checked for normality, helping to assess the appropriateness of the model.

6.2 Forecasting and Confidence Intervals

After evaluating the model, it was used to forecast the number of goals England would score in future years (1987–1991). The forecasts were accompanied by **80% and 95% confidence intervals** to indicate the uncertainty around the predictions.

7. Results and Interpretation

- **Forecast Results**: The ARIMA model predicted a **stable number of goals (close to zero)** for the years 1987-1991. This indicates **relatively stable performance** based on historical data.
- **Model Diagnostics**: The residuals analysis showed no significant patterns, suggesting that the model fits the data well.
- **Confidence Intervals**: The wide confidence intervals (ranging from -22 to 22) indicate a **high level of uncertainty** in the forecast, which could be due to variability in past data and the lack of strong trends.

8. Limitations and Future Work

- **Limited Data**: The dataset includes only goal-scoring data, with no external factors (e.g., player injuries, changes in team dynamics) considered. Future work could incorporate these variables.
- **Model Refinement**: The ARIMA model assumes linear relationships, which may not capture complex patterns. Exploring **SARIMA (Seasonal ARIMA)** or **machine learning models** could improve accuracy.
- **Comparing Teams**: Extending the analysis to include multiple teams and compare performances would provide more insights.

Conclusion

This project utilized ARIMA to forecast the number of goals scored by the England football team. The model performed reasonably well, with no significant patterns left in the residuals. The forecast for future years suggests stable performance but highlights the uncertainty with wide confidence intervals.