Forecasting future goal statistics of Messi and Ronaldo

DS207: Time Series Forecasting

Narek Khachikyan, Davit Khalatyan, Vram Papyan

20/12/2023

Abstract:

This report aims to forecast the future goal statistics of football legends Lionel Messi and Cristiano Ronaldo based on historical data from the 2002/03 season to the 2019/20 season. The dataset includes records of goals scored by each player. We employed data manipulation along with time series analysis techniques, including SARIMA and Holt Winter’s Exponential Smoothing models, to make short-term predictions. The report discusses the methodology, model diagnostics, and forecast evaluations.

Literature Review:

The article "INTRO TO TIME SERIES AND SPORTS" on The Data Jocks explores the application of time series analysis in the context of sports data, highlighting its potential in understanding player performance, predicting future trajectories, and gaining insights into various aspects of sports analytics. The article identifies three key components: what time series and sports are, the tools used to analyze them, and the reasons behind the underutilization of time series analysis in sports data. It emphasizes that sports data naturally exhibits a time component, with players evolving, teams changing, and strategies adapting over time. The article delves into time series forecasting methods, citing popular approaches like moving average models, autoregressive integrated moving average models (ARIMA), and machine learning methods such as recurrent neural networks. In conclusion, the article acknowledges the importance of time series as a mathematical tool in sports analytics, emphasizing its natural application in understanding sequential sports data. The limitations of basic time series models are acknowledged, paving the way for future exploration of more advanced models in upcoming articles. Overall, the article provides a comprehensive introduction to the intersection of time series analysis and sports data, setting the stage for further exploration in the field.

The research study, titled "Mastery in Goal Scoring, T-Pattern Detection, and Polar Coordinate Analysis of Motor Skills Used by Lionel Messi and Cristiano Ronaldo," explores the motor skills employed by two of the world's top soccer players, Cristiano Ronaldo and Lionel Messi, in their attacking actions leading to goals. The study aims to fill the gap in soccer research by focusing on the motor skills aspect that underpins specific motor actions, as opposed to the more traditional emphasis on technical and tactical skills. The authors utilized an improved observation instrument, OSMOS-soccer player, consisting of nine criteria expanded into 50 categories. T-pattern detection and polar coordinate analysis were employed to investigate associations between these categories, revealing temporal structures and interrelated behaviors in the players' attacking actions. In summary, this research study sets out to contribute to soccer performance research by examining the motor skills used by top players in goal-scoring situations. The utilization of advanced analysis techniques, such as T-pattern detection and polar coordinate analysis, aims to uncover temporal structures and associations between different motor skills. The study emphasizes the importance of understanding the motor skills aspect in soccer performance and provides a framework for analyzing specific actions leading to goals.

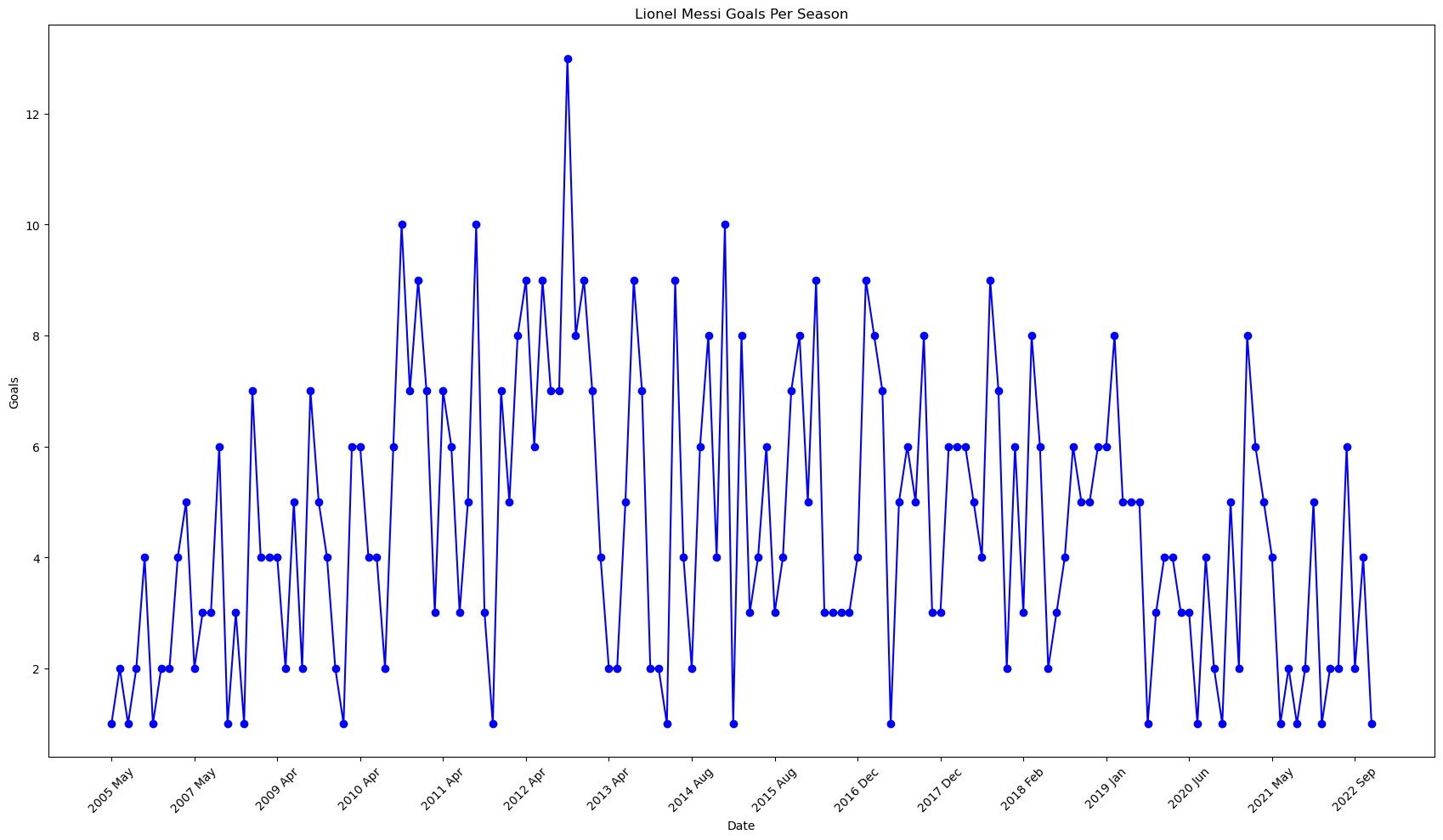
1. Introduction:

Football enthusiasts have always been intrigued by the performance trajectories of iconic players. In this project, we delve into Lionel Messi and Cristiano Ronaldo's goal statistics, exploring trends, and patterns, and making short-term forecasts. The objective is to provide insights into the potential goal-scoring performance of these players in the next 2-3 seasons.

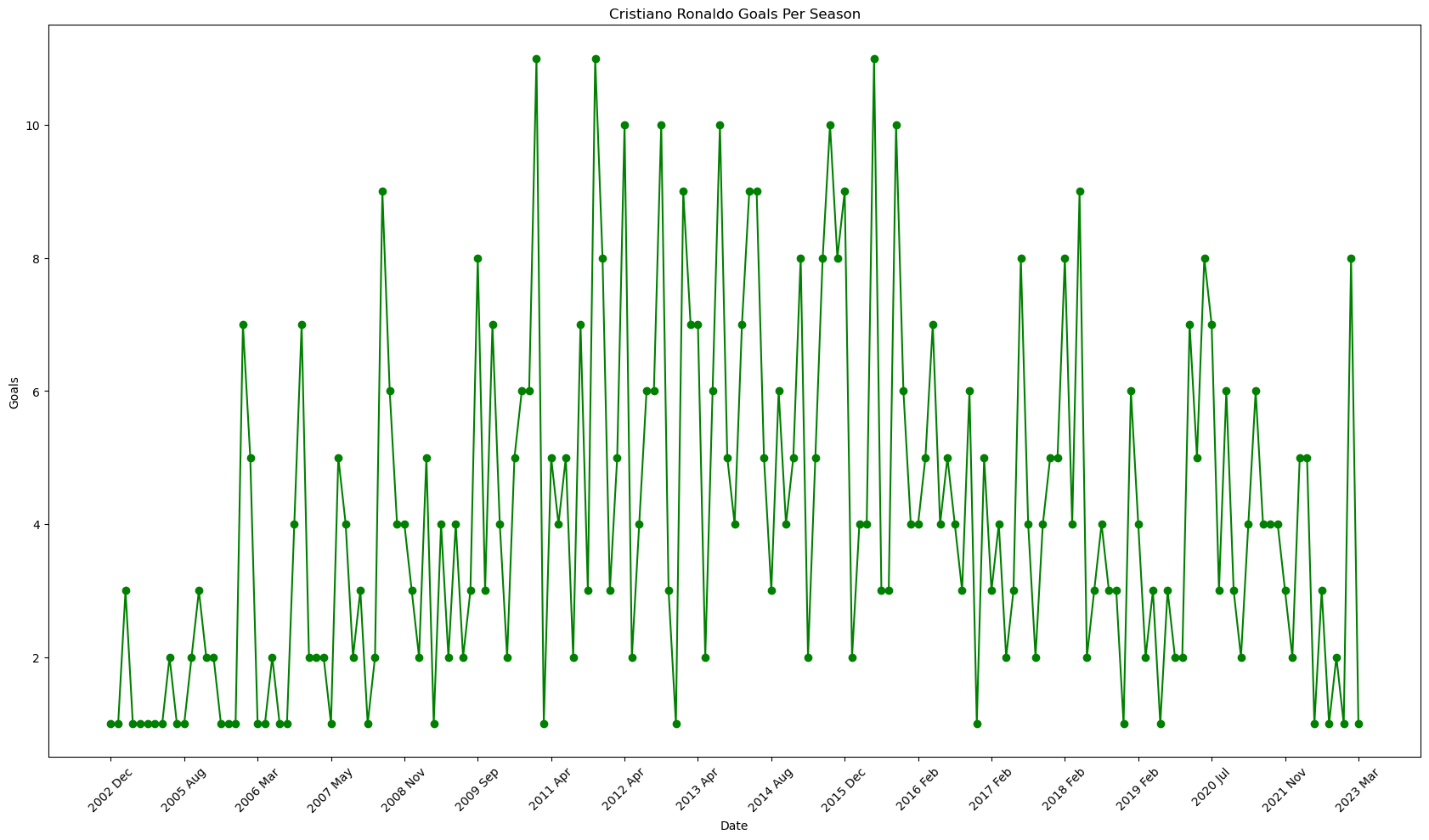
2. Data Collection:

The dataset comprises records of goals scored by Messi and Ronaldo starting from the 2002/03 season. This dataset was taken from Kaggle.com and then data cleaning was conducted after which the dataset was divided into two parts for each player. After different data manipulation techniques, the goals were grouped by month for each of the players. Later these two datasets were combined to form another dataset for monthly summed goals of Messi and Ronaldo. Further exploratory data analysis highlighted trends, fluctuations, and potential seasonality in the goal statistics.

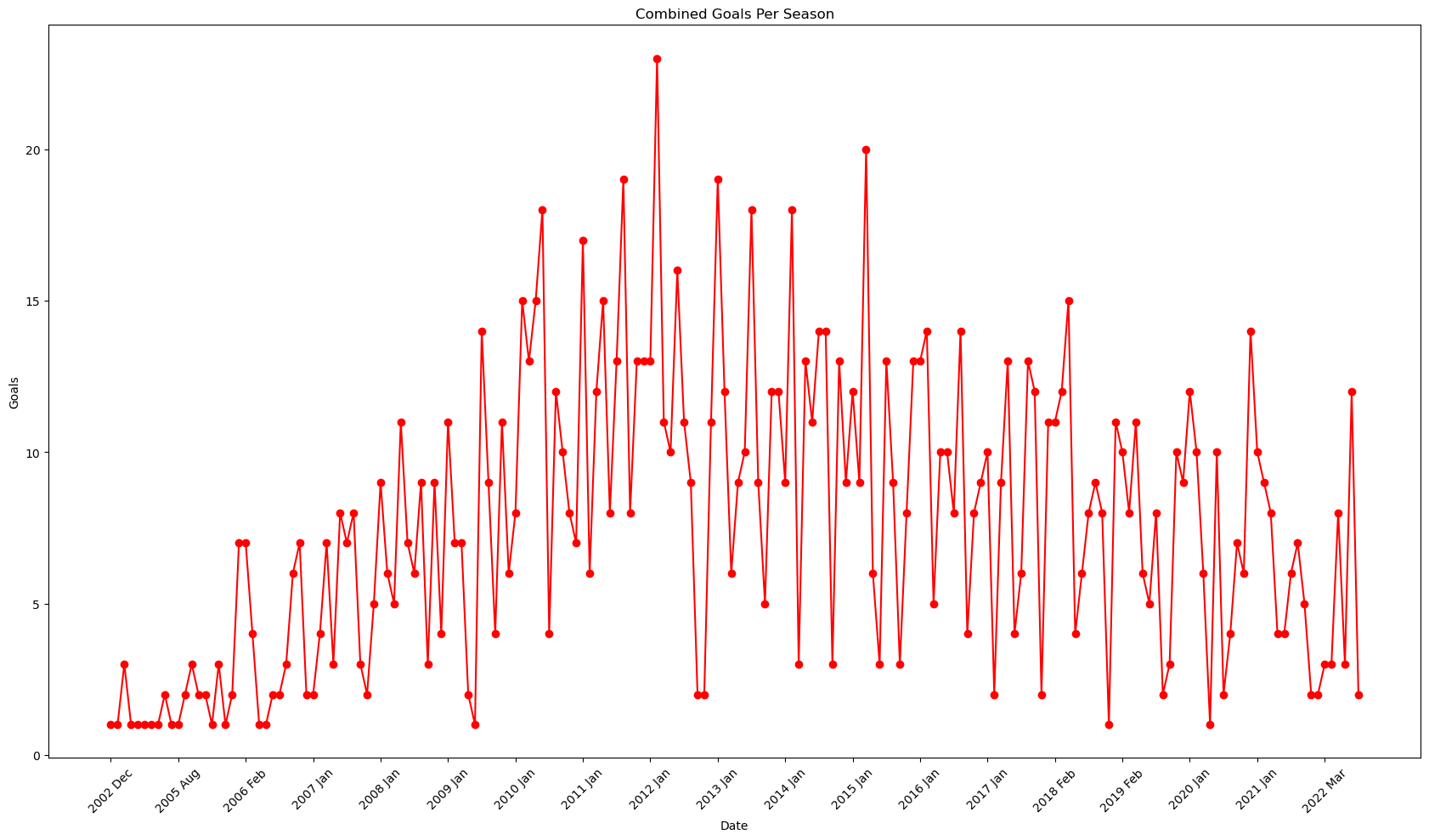
Lionel Messi: The dataset begins from the 2004/05 season, with the number of goals scored each season.



Cristiano Ronaldo: The dataset for Ronaldo starts from the 2002/03 season.



Combined Data: This dataset sums the goals scored by both players in each season, starting from the 2002/03 season.



3. Methodology:

3.1 Seasonal Decomposition:

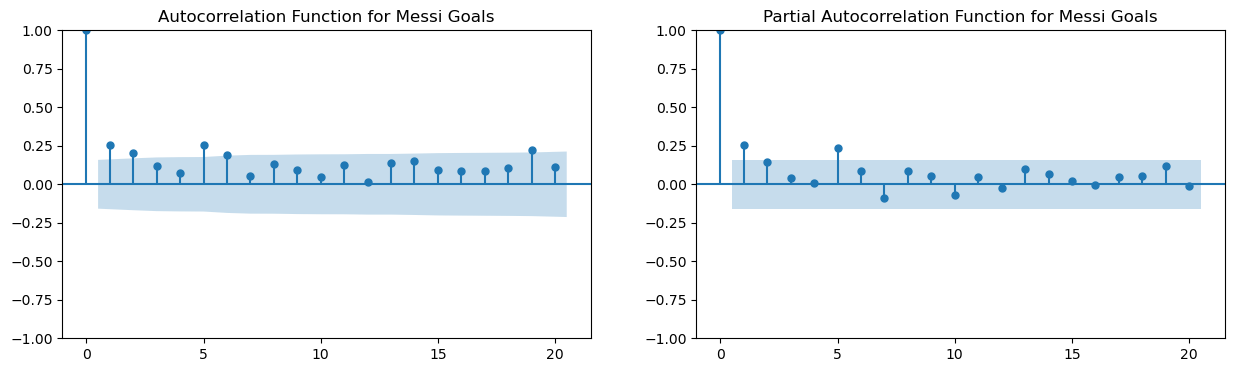
Visual inspection of Messi and Ronaldo's goal plots revealed several insights. Messi's goals exhibited some fluctuation with a general upward trend until around the 2011/12 season, after which they fluctuated more noticeably. Ronaldo's goals show a significant upward trend in the early seasons, peaking around the 2014/15 season. Afterward, there is noticeable variability with a slight downward trend. Combined goals amplify the peaks and troughs, though the data does not exhibit clear seasonality.

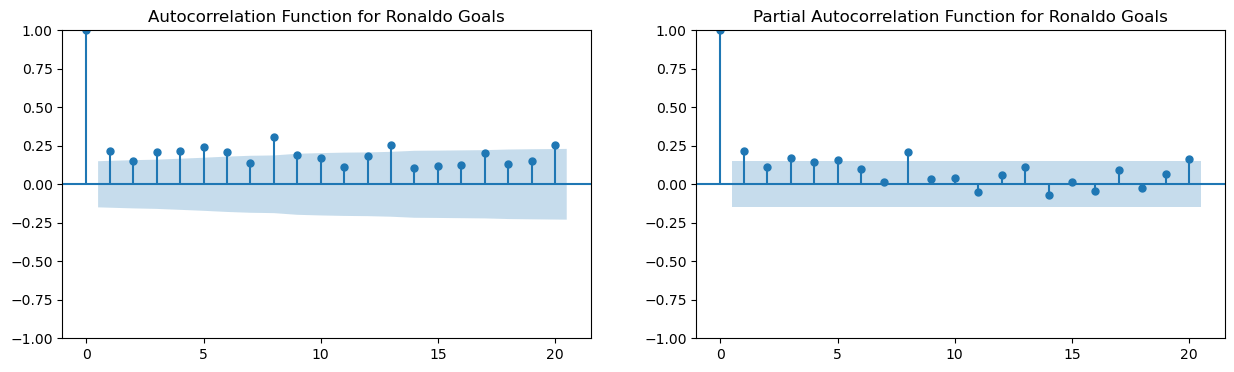
3.2 Stationarity Checks:

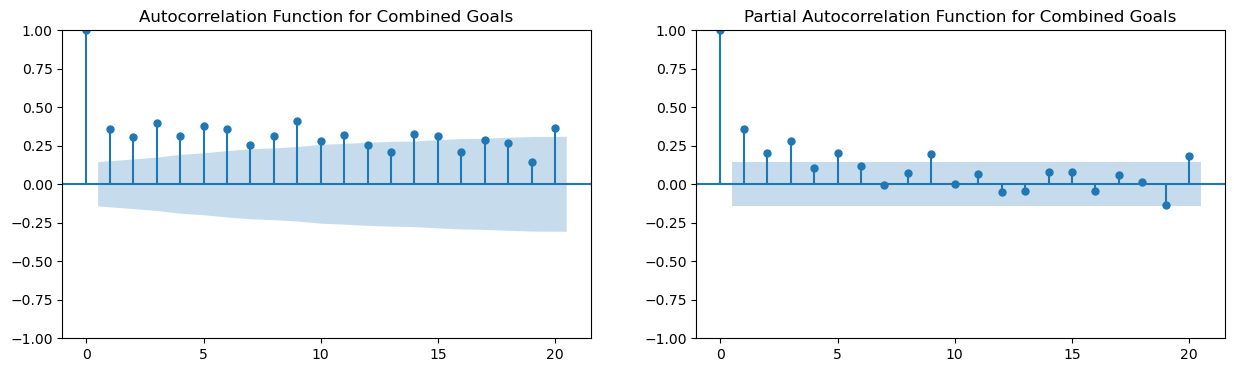
Augmented Dickey-Fuller test showed that series for Messi was stationary and series for Ronaldo was not. KPSS test showed the same results for both of the series as ADF test results. Regarding the combined dataset, it had results very similar to Ronaldo’s series.

3.3 SARIMA Model:

ACF and PACF plots guided the SARIMA model specification for both players.







The results indicated SARIMA(5, 0, 5)(1, 0, 1, 12) model for Messi’s series and SARIMA(5, 1, 5)(1, 0, 1, 12) for Ronaldo series. The combined dataset was given the same model as for Ronaldo’s series because the datasets had very similar results for the stationarity and order tests. Then the models were trained and tested.

3.4 Holt Winter’s Exponential Smoothing Model:

Holt Winter’s Exponential Smoothing method was used for all three datasets to forecast as an alternative for SARIMA models and to understand which model was better.

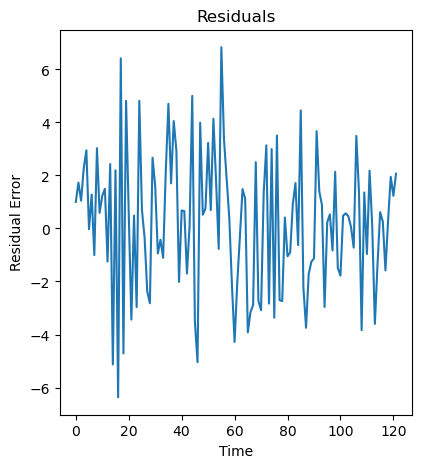
3.5 VAR Model:

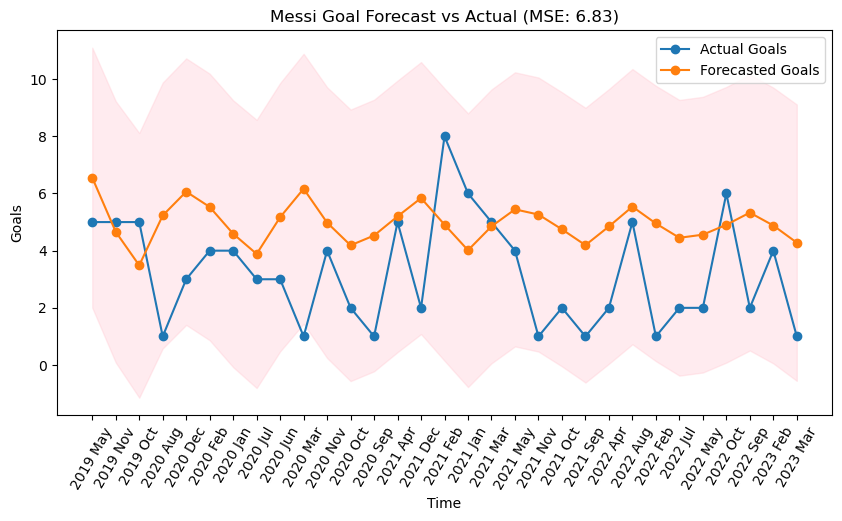
The vector autoregressive model was used to check if there was a pattern in the goal-scoring records of Messi and Ronaldo because it is commonly said that these two players pushed each other to become better players and this hypothesis can be checked by the VAR model.

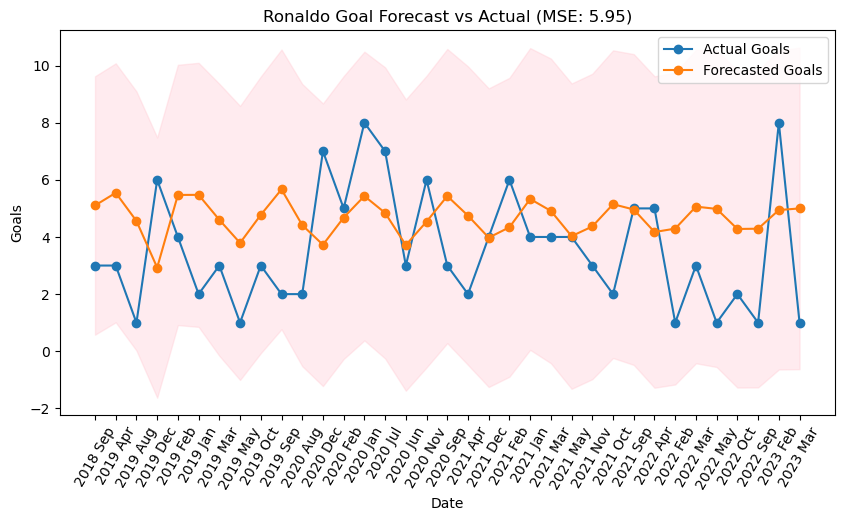
4. Results and Analysis:

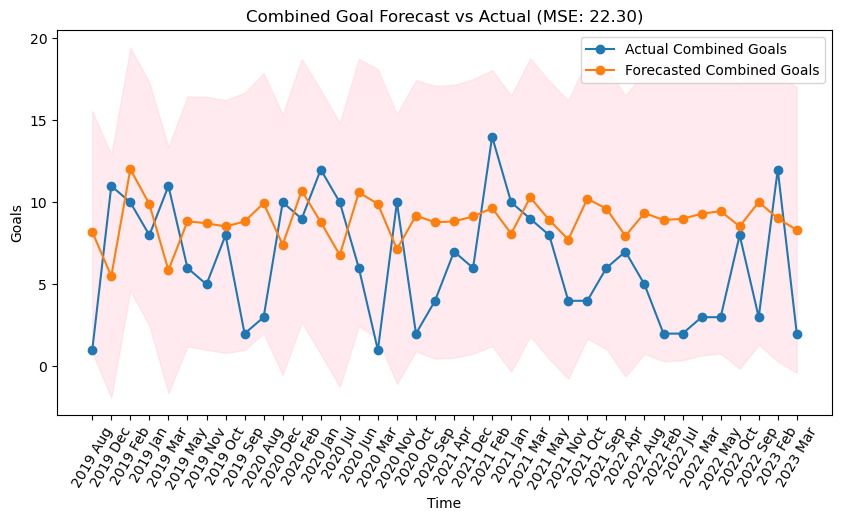
4.1 SARIMA Model Evaluation:

The SARIMA model was evaluated for both Messi's and Ronaldo's goals per season. Residual diagnostics showed white noise, and the model exhibited good performance with a lower MSE compared to the Holt Winter’s model.









4.2 Holt Winter’s Exponential Smoothing Model Evaluation:

Holt Winter’s Exponential Smoothing models were estimated and forecasted. Trend was included in the models and in those terms Holt Winter’s turned out to be a good model.

MSE for Holt Winter’s Model:

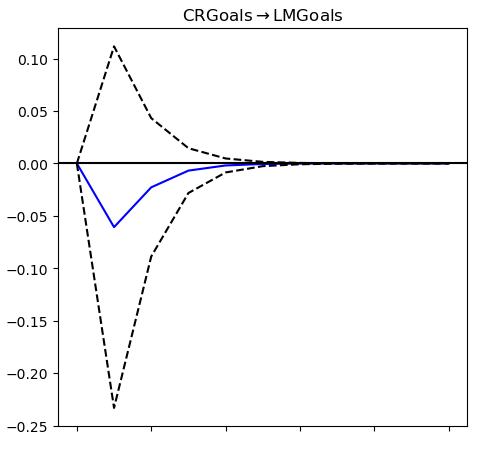
Messi: Approximately 11.863

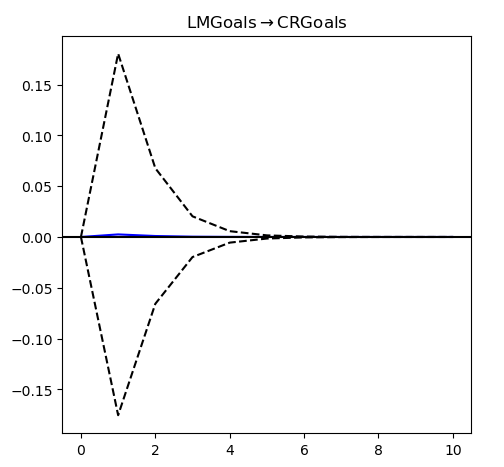
Ronaldo: Approximately 4.633

Combined: Approximately 12.584

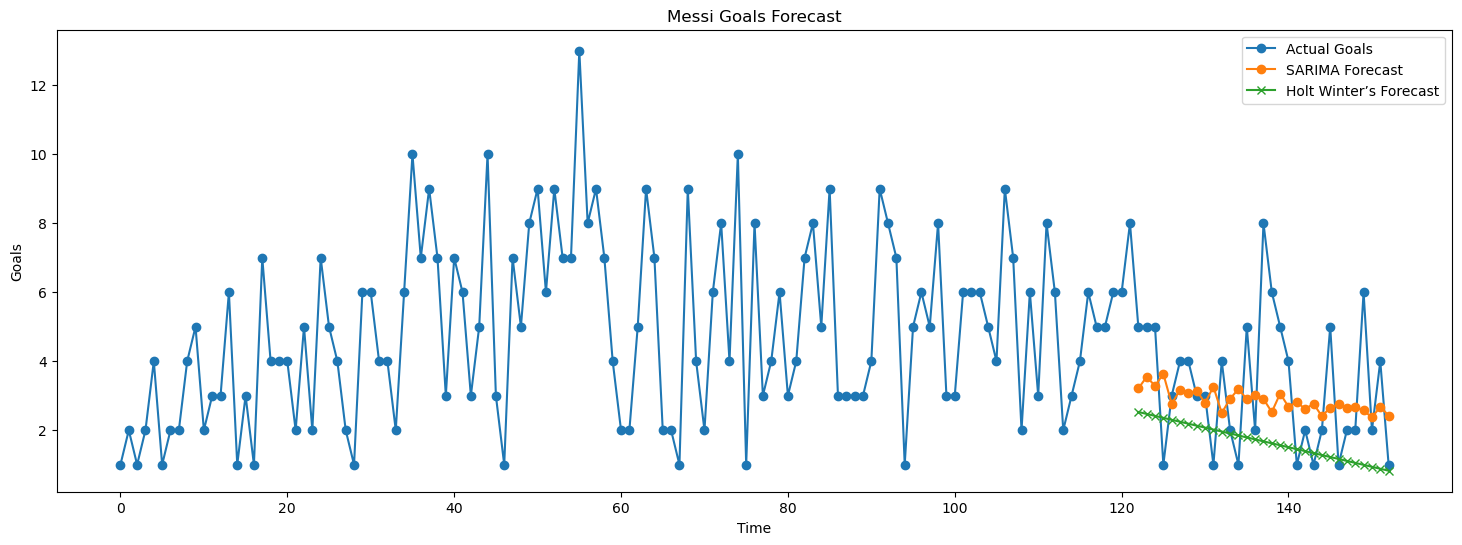
4.3 IRF Analysis Results:

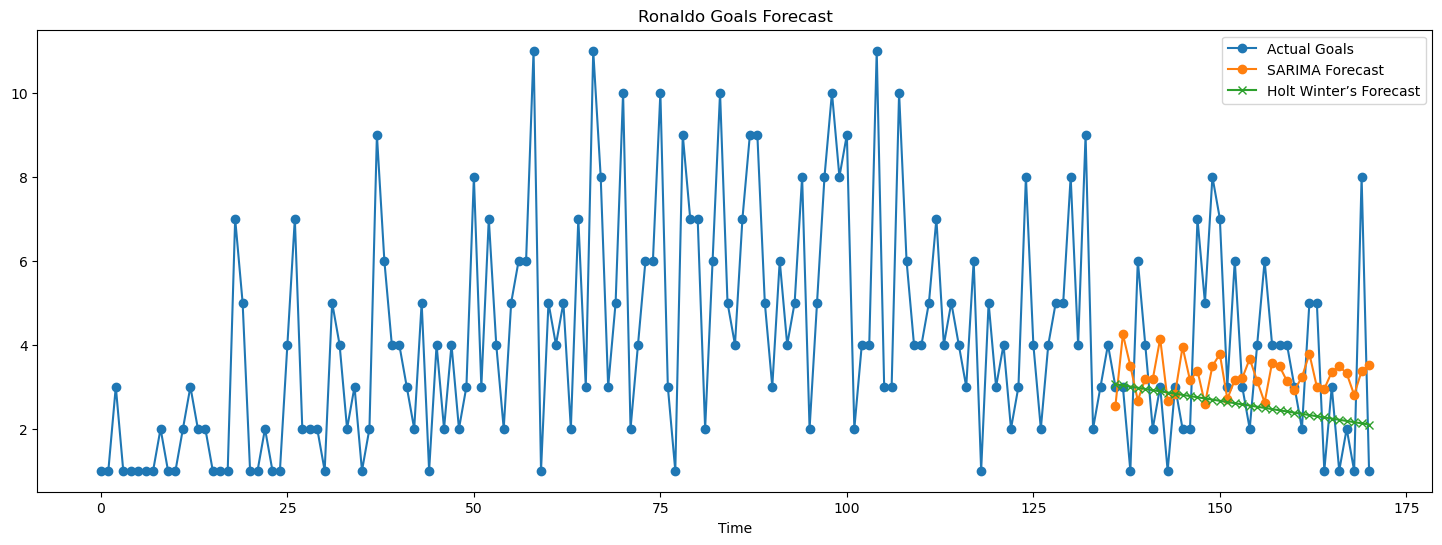
As we can see from the below graphs, the hypothesis is rejected because no significant correlation is seen between the two series.

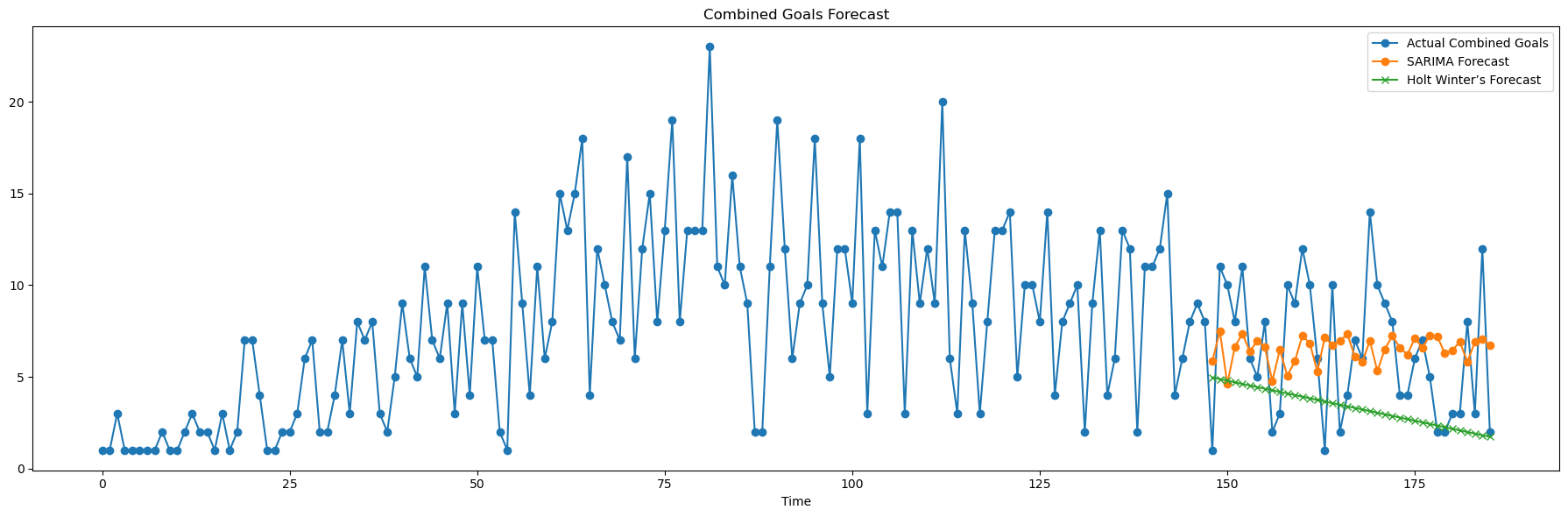




5. Forecasting and Comparison:







5.1 Short-Term Forecasts:

Both models provided forecasts for Messi, and Ronaldo, and their combined goals from 2019 up to 2023. Visual representations showed predictions aligned with historical trends. Depending on the plots, the models showed very different results.

5.2 Model Comparison:

MSE values emphasized the SARIMA model's superiority in accuracy. Visual comparisons of actual, SARIMA, and Holt Winter’s forecasts indicated closer alignment of SARIMA predictions with actual values. However, as it can be seen from the plots, Holt Winter’s method considered a downward trend which is natural for a football player so it becomes debatable which model actually is better. However, the Holt Winter’s model underestimated too much the results of the players.

6. Conclusion:

Based on the MSE values and visual assessments, the SARIMA model outperforms the Holt Winter’s Exponential Smoothing model in forecasting Messi and Ronaldo's future goal statistics. The SARIMA model is more adept at capturing underlying patterns, providing a more accurate short-term prediction.

7. Recommendations:

Further research could explore additional factors influencing goal statistics, such as team dynamics, player injuries, and match outcomes. Additionally, incorporating more recent data could enhance the models' predictive capabilities.

8. References:

Articles – "INTRO TO TIME SERIES AND SPORTS"([link](https://thedatajocks.com/time-series-sports-101/)),

"Mastery in Goal Scoring, T-Pattern Detection, and Polar Coordinate Analysis of Motor Skills Used by Lionel Messi and Cristiano Ronaldo" ([link](https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00741/full))

Kaggle - dataset [link](https://www.kaggle.com/code/ahmedterry/cristiano-ronaldo-vs-lionel-messi-goals-eda)