Time Series Forecasting - Soccer Popularity

Neil Slavishak

December 23, 2024

Abstract

This project analyzes and forecasts the popularity trends of six major soccer leagues in the United States using time series data from Google Trends. The leagues examined include the Premier League, La Liga, Bundesliga, Serie A, Major League Soccer (MLS), and Ligue 1. A SARIMAX modeling approach was applied to capture seasonal patterns and long-term trends, producing forecasts for each league's search interest over the next 200 weeks. With these forecasts, inferences and predictions can be made about the future of the popularity of the sport. However, due to the volatility of the sport and its dependence on world events, the predictions found are generally not entirely plausible. However, these insights can guide marketing strategies and resource allocation for stakeholders looking to grow their presence in the soccer market.

1 Introduction

Soccer is the world's most popular sport, boasting billions of fans across continents. However, its appeal varies significantly by region, with the United States traditionally lagging behind in soccer's global dominance. In recent years, U.S. interest in soccer has grown, fueled by the increasing visibility of major European leagues like the Premier League, La Liga, and Bundesliga, alongside the domestic rise of Major League Soccer (MLS). Understanding the dynamics of this popularity is crucial for stakeholders, such as broadcasters, sponsors, and league organizers seeking to capitalize on this expanding market.

The study seeks to compare the relative engagement of these leagues in the U.S. and examines their growth potential. For example, European leagues traditionally dominate global soccer interest, but MLS has been gaining traction domestically due to increasing investment, talent acquisition, and local pride. Through this analysis, the aim is to uncover actionable insights and trends that could guide strategic decisions, such as marketing campaigns and scheduling strategies.

By combining statistical forecasting with contextual real-world understanding, this project contributes to the ongoing conversation about soccer's future in the United States and its position within the global sports market.

2 Data Description

To acquire data to inspect the popularity of each league in past years, as well as provide a data set to perform the time series forecasting model on, Google Trends was used. Google Trends provides a visual, as well as a Comma-Separated Values (CSV) file.

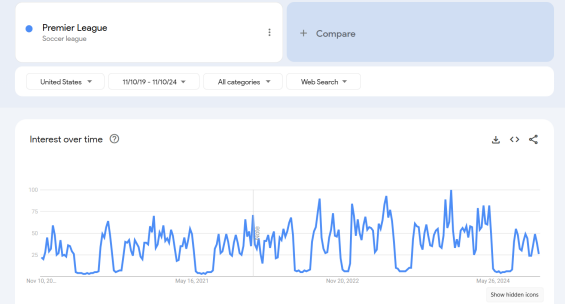

Figure 1: Google Trends Interface

In this case, data is collected from the last five years to allow for a better gauge on the seasonality of the data. Because of this, the data is captured weekly rather than daily to account for the large time range. In the CSV files, there are two columns of data, one being the date in which the data is collected titled "Week" and the other being the volume of interest of the seven days prior to that date, titled "League: (Country)", for example, "Premier League: (United States)." When doing time series forecasting, the "Week" column is turned into the index, allowing for easier visualization of the data. Fortunately, in this case, there are no missing values either, which allows for simple data wrangling and visualization.

3 Methodologies

In order to carry out the goals of this project, Python was used in conjunction with the pandas, statsmodels, and matplotlib tools. Using these tools, the outline of the methodologies can be split into multiple parts.

3.1 Reading CSV and Setting Up Data

Pandas was used to read the CSV file by using the .read_csv command, and the entires in the "Week" column were turned into datetime objects using the .to_datetime command. At the same time, the "Week" column was turned into the index to allow for simpler fitting later on by using .set_index. Finally, the .asfreq command was used to ensure that the time series was evaluated weekly.

3.2 SARIMAX and Grid Search

To accurately predict the forecast of each of league, various time series forecasting tools were used. In this case, the chosen method was SARIMAX.

3.2.1 SARIMAX

The SARIMAX model is an extension of the basic ARIMA (Auto-Regressive Integrated Moving Average) model. This can be split into three parts. The auto-regressive part refers to the changing variables that regresses on its own prior values. The integrated part refers to the process in which data values are replaced by the difference between its value and the previous values. The moving average part refers to the dependency between the error and its observation. The model itself uses three parameters, which are p, d, and q. p represents the number of auto-regressive terms. d represents the number of times the observations are differenced. q represents the number of lagged forecasting errors. The formula for this model is as follows:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-1} + \epsilon_t$$

The ARIMA model is very effective, however, it does not include the effects of seasonality and exogenous variables. The SARIMA model builds on thie ARIMA model by adding a new part, the seasonal aspect. The new model introduces four new parameters, which are P, D, Q, and s. The P, D, and Q parameters are virtually the same as the p, d, and q parameters in ARIMA, however, their

focus is the seasonal components of the data. s is the seasonal period, which in the case of this project, would be 52 to account for the 52 weeks in a year. The formula for the SARIMA model is as follows:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{P} \alpha_i y_{t-si} + \sum_{i=1}^{Q} \eta_i \epsilon_{t-si} + \epsilon_t$$

The red section is the seasonal component of the model, which incorporates the new parameters mentioned above. Now the model incorporates seasonality, however, it does not yet incorporate exogenous variables. This is where the SARIMAX model is used. This model incorporates the "X" part, which stands for the covariates, or the external variables that influence the time series. The formula for the model is as follows:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{r} \beta_i x_{i_t} + \sum_{i=1}^{P} \alpha_i y_{t-si} + \sum_{i=1}^{Q} \eta_i \epsilon_{t-si} + \epsilon_t$$

The red section is the exogenours component of the model. Now that the model is defined, it can be used to create the forecasts for the data. Using statsmodels.tsa.statespace.sarimax, this can be achieved. However, the question is what values should be used for the parameters.

3.2.2 Grid Search

To find the values to use for the parameters, the grid search optimization technique was used. Grid search loops through every possible combination of parameters and evaluates which combination is most optimal for the model. In this case, the optimality was examined by finding the Akaike Information Criterion (AIC) value for each combination. This can be done by simply using the .aic command in the statsmodels library. After looping through all combinations of parameters, the combination with the lowest AIC value was used to find the forecast.

3.3 Visualization of Forecasts

After running the SARIMAX model and using commands .get_forecast and .conf_int to get the forecasts and 95% confidence intervals, matplotlib was used to display the data and the forecasts. For each individual league, a plot with the data and the model with a forecast of past data are compared to illustrate the effectiveness of the model. Then the forecast for the next 200 weeks of each league are illustrated as well. Plots illustrating a comparison of the forecasts of each league in a country are also formed. Commands used to achieve this include .get_prediction, .predicted_mean, and .fill_between to outline the confidence intervals.

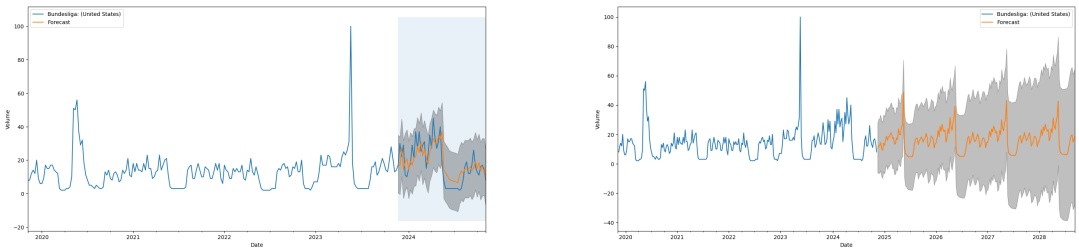

Figure 2: Example of Figures for One League

4 Results

4.1 Search Volume of Each League

The first goal of the project was to examine the popularity of the top six leagues. Knowing which leagues are more popular than the others within the United States can assist with marketing strategies both for broadcasting companies in the U.S., as well as the organizers of the leagues who are looking to widen their audience. For each set of graphics, the first compares the model with already existing

data to confirm the seasonality of the model and such. The second graphic shows the forecast, going 200 weeks into the future.

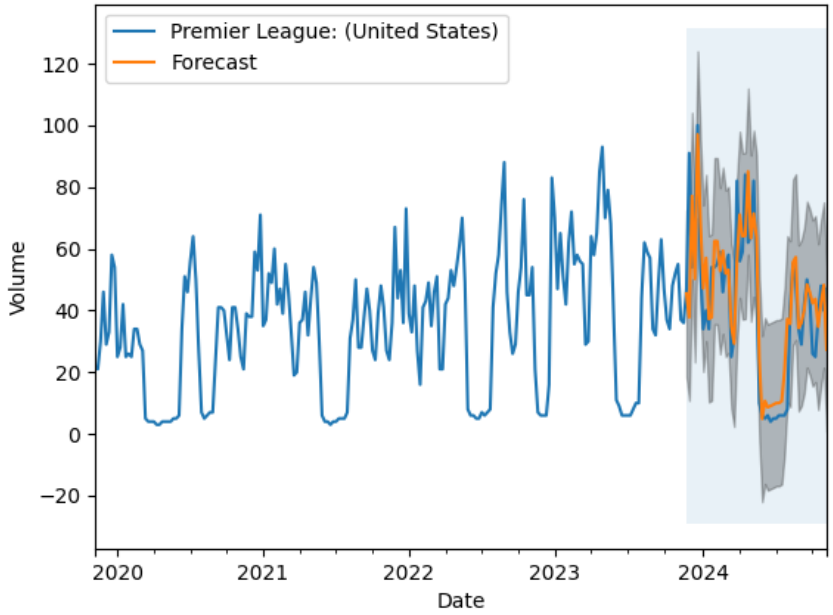

Figure 3: Model Comparison for Premier League

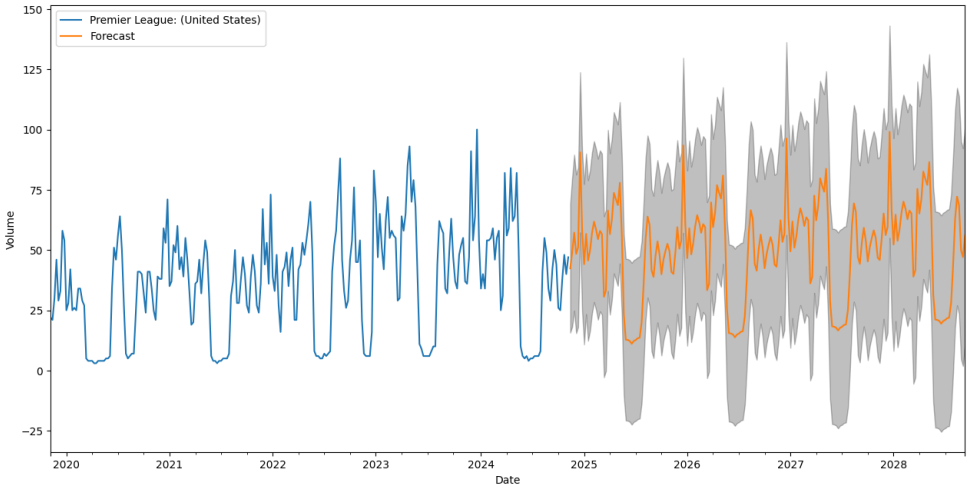

Figure 4: Forecast for Premier League

Clearly, the model is able to account for the seasonality of the data, which is due to the period over the summer in which the season is not in session. It is also able to somewhat predict the random peaks, which usually occur due to various events within the season that spike popularity or when popular players joined the league, like Cristiano Ronaldo at the end of 2021. The forecast predicts that the Premier League will slowly rise in popularity over the next few year, while still having a dip each summer, which will most likely happen. One thing to note is that the confidence intervals are fairly

large for the forecasts. This is due to the fact that this data is fairly volatile and many unexpected events could occur which would alter the future data in a way that cannot be predicted. A similar process can be observed with the La Liga.

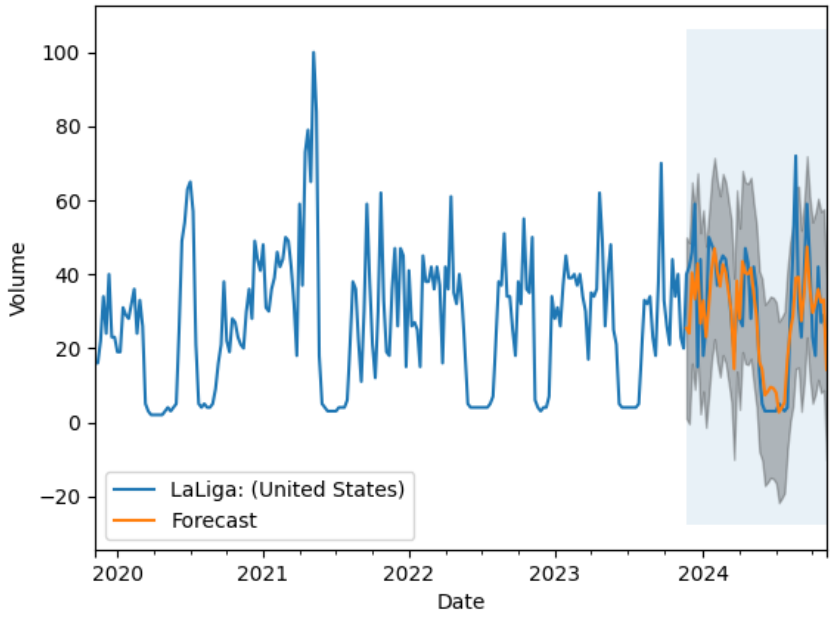

Figure 5: Model Comparison for La Liga

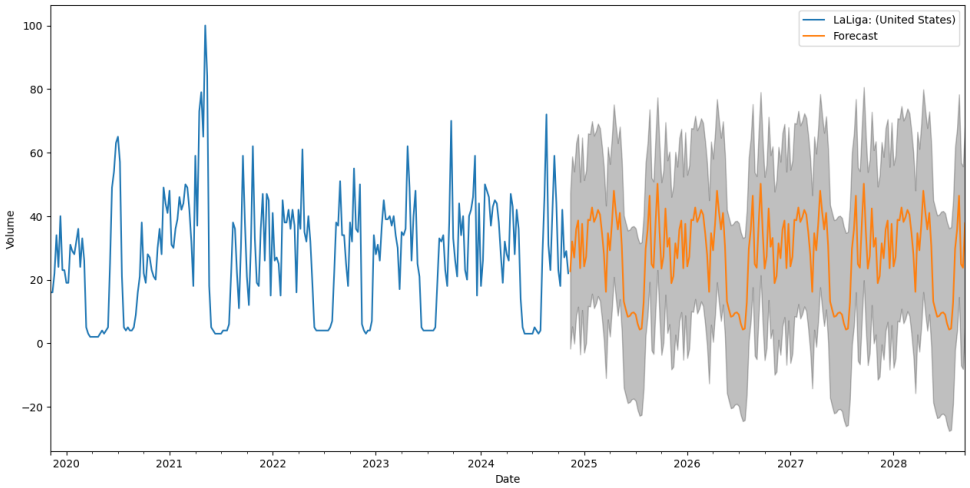

Figure 6: Forecast for La Liga

Similar to the Premier League data, the model captures the basic motion and seasonality of the data, however, it is not able to predict the various spikes in volume. For example, the spike observed near the end of the 2024 data was due to Kylian Mbappe joining Real Madrid. The model would have no way to predict that was to occur. The forecast illustrates not much change in the overall performance of the league, indicating that a desire to increase advertisement may be necessary to grow

the league's influence in the United States.

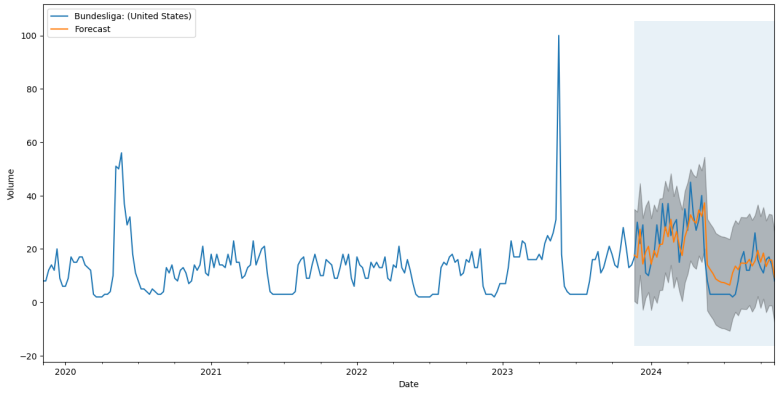

Figure 7: Model Comparison for Bundesliga

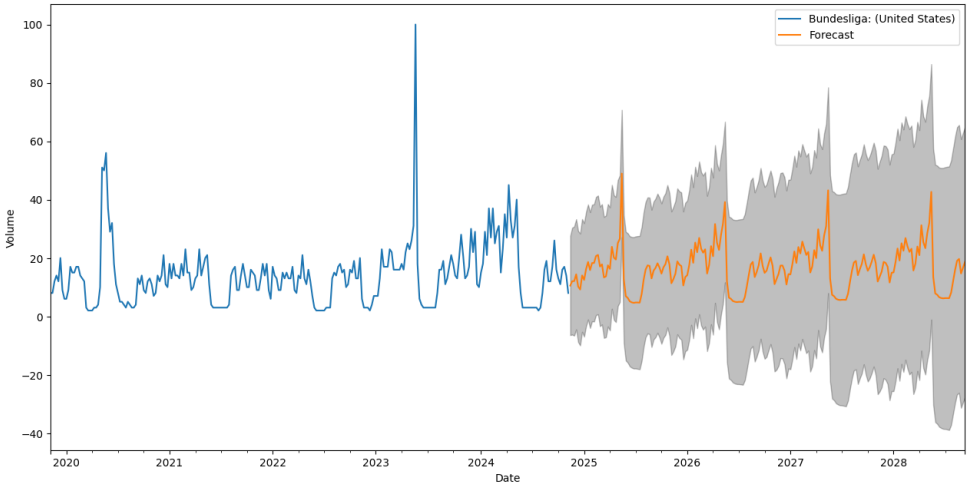

Figure 8: Forecast for Bundesliga

The Bundesliga data consists of two major spikes, one being in the year 2020, which was due to the Bundesliga being the first league to return after all soccer leagues halted due to the COVID-19 pandemic. A second spike is seen at the end of the 2022/23 season, in which there was a lot of final-day drama between the teams Bayern Munich and Borussia Dortmund. The model seems to take that second spike into account, which may not be accurate in seasons to come, however, it is plausible that the end of the season would bring the most traffic. Otherwise, the forecast seems to not offer much fluctuation in search volume.

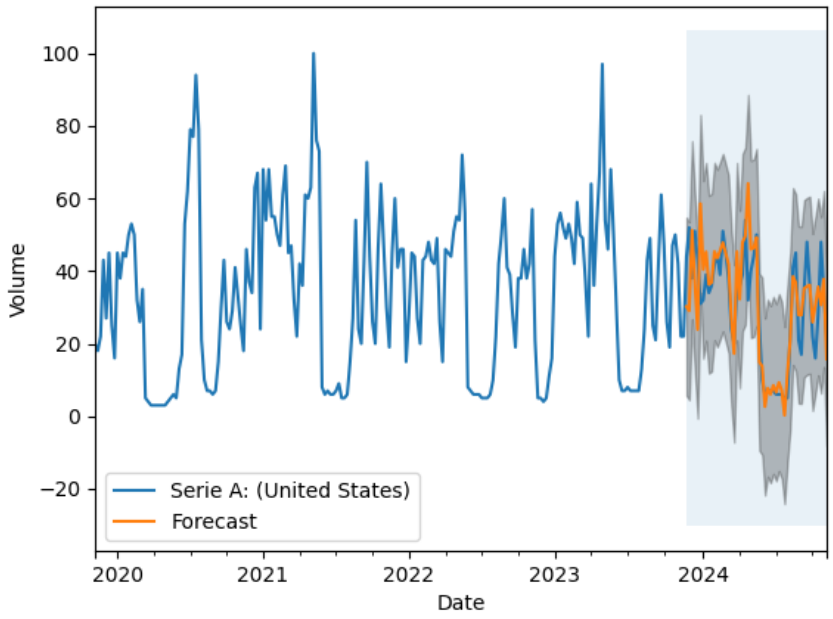

Figure 9: Model Comparison for Serie A

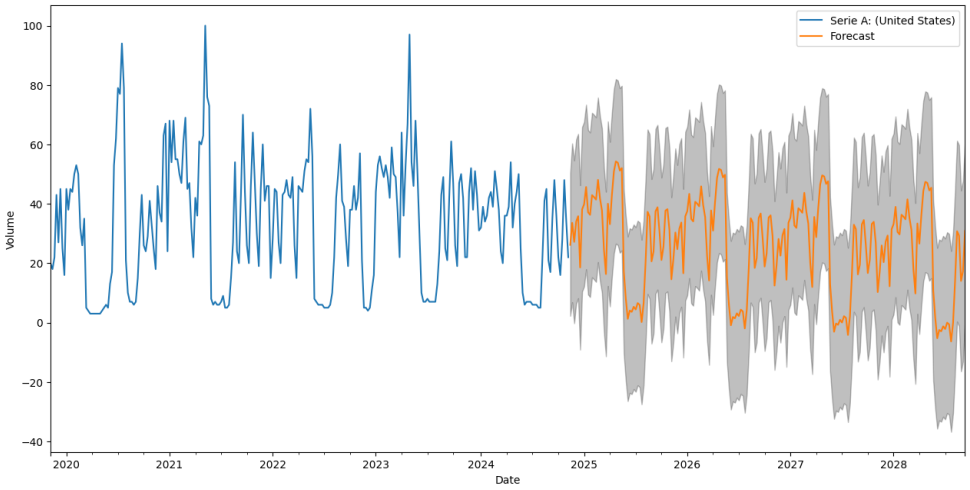

Figure 10: Forecast for Serie A

The Serie A search volume data is the most robust, with plenty of spikes and dips. However, as with the other leagues, the model is able to somewhat follow the spikes, as well as account for the seasonality of the data. The forecast suggests that the popularity of the Serie A in the United States will decline in the next few years, indicating that efforts should be done to increase the league's presence in the United States to avoid losing a very marketable audience.

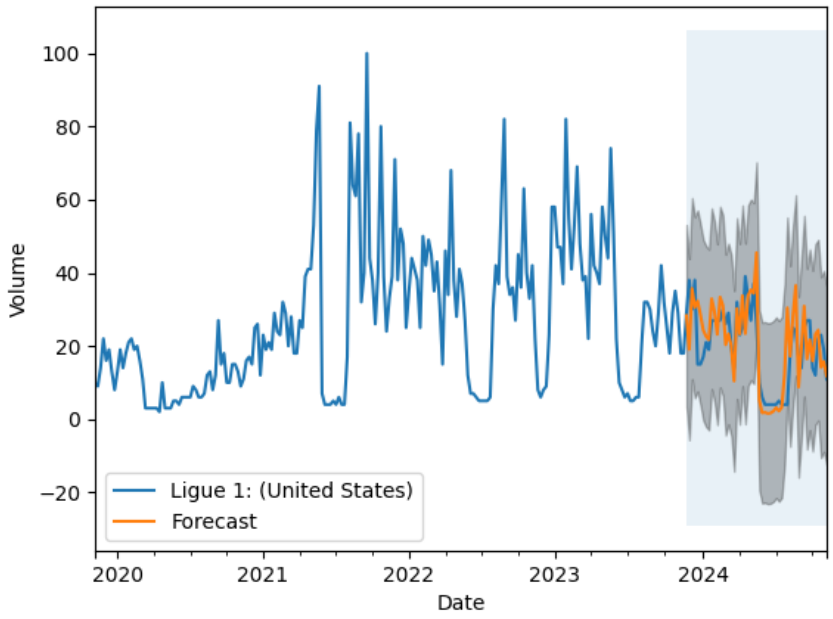

Figure 11: Model Comparison for Ligue 1

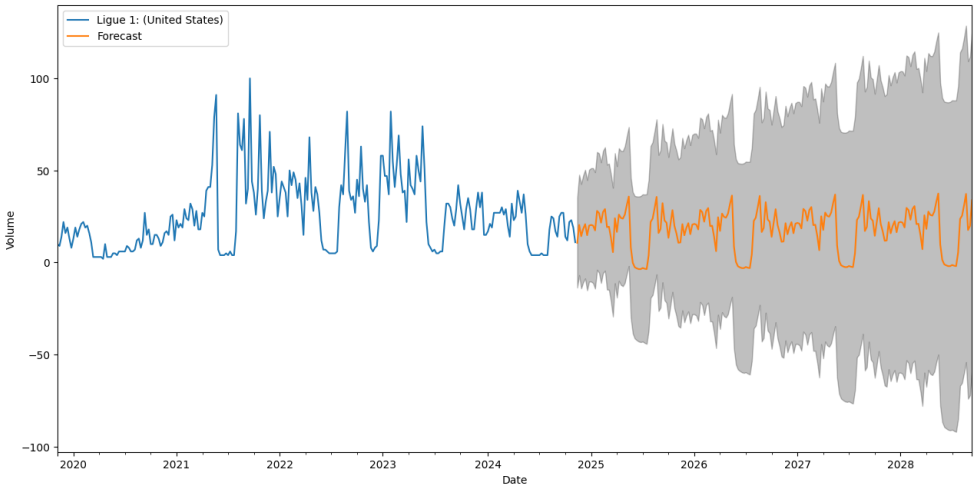

Figure 12: Forecast for Ligue 1

The Ligue 1 data offers a very interesting image. The search volume seems to be the greatest from the end of 2021 to the end of 2023. This was most likely due to the presence of Kylian Mbappe in the league and France's success in the 2022 World Cup. However, during the 2023/24 season, there were many rumors about Mbappe's departure, which ended up turning true near the end of the season, which would explain the drop-off in popularity. The model seems to detect that drop-off and illustrates a continuity of smaller volume in the future.

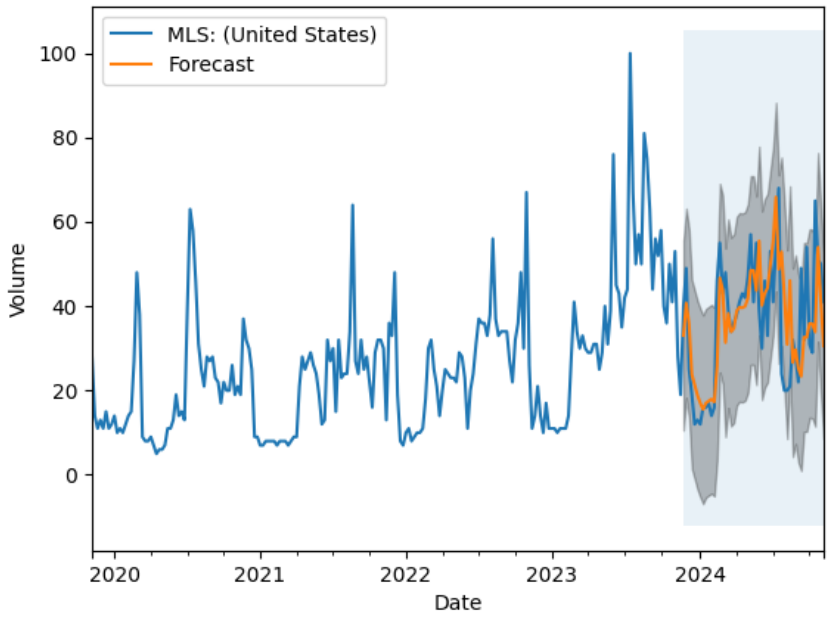

Figure 13: Model Comparison for MLS

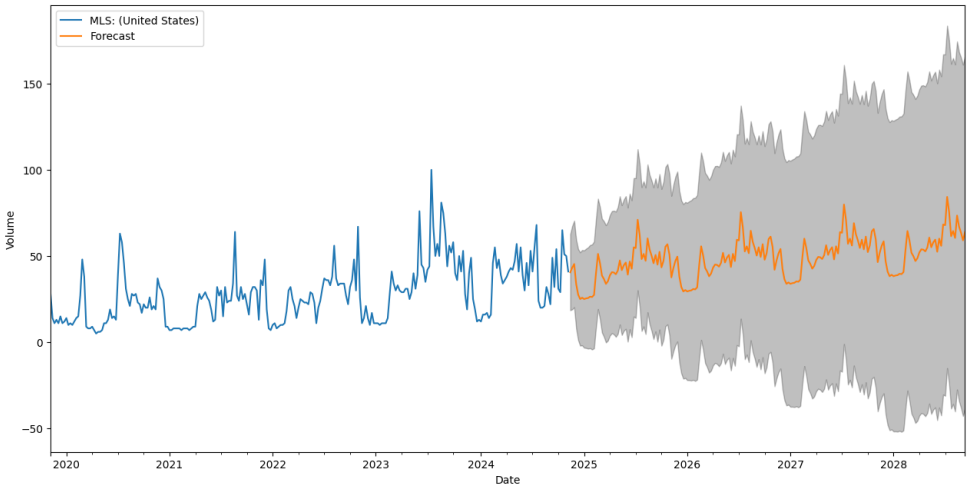

Figure 14: Forecast for MLS

The MLS is the outlier of this group of soccer leagues, as its seasonality is much different to the rest. The MLS's offseason occurs during the winter, while all of the other leagues have their offseasons during the summer. Because of this, there is a spike that occurs during the summer for the MLS, as soccer fans turn to the MLS since no other leagues are in season. As with the other leagues, the model captures that seasonality and accurately describes that seasonality in the forecast.

4.2 Search Volume Comparison

With each individual league's search volume statistics examined, a comparison was made to accurately draw conclusions concerning the popularity of each league in conjunction with the others.

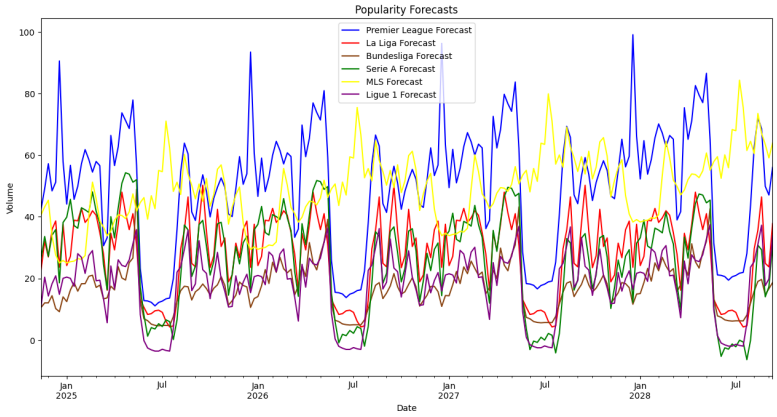

Figure 15: Forecasts for All Six Leagues in the United States

By examining the figure above, a few assumptions can be made about the popularity of each of the six leagues in the next few years. First of all, it is clear that the English Premier League and the MLS have the joint-highest popularity in the United States, with the MLS dropping a little in search volume during the Premier League's season, but clearly surging ahead in popularity while the European leagues are not in season. The La Liga and Bundesliga leagues are in the middle of the pack in terms of popularity. This is most likely due to incentives made by ESPN in the last few years to broadcast the games in those leagues on ESPN+. The forecasts illustrate that, although the leagues may not reach the popularity of the MLS and Premier League, the leagues will still have traction in the United States and should not be concerned about losing popularity. The two leagues that are not prominent in the United States are Ligue 1 and the Serie A. Both of these leagues illustrated declines in the forecast, which would mean that the leagues are predicted to lose popularity in the United States over the next few years. With this information, the management of these two leagues should be looking to expand into the United States, which could be achieved through broadcast deals like the Bundesliga and La Liga, more preseason games in the United States like some of the other leagues do, or even looking to bring more popular players into the league. For the other four leagues, however, it is clear that these leagues will continue to rise in popularity over time and should not be concerned about boosting their influence in the United States.