Time Series Analysis of the Total Goals Scored by Liverpool FC

Ali Aytuğ Tok
2502409
tok.aytug@metu.edu.tr

Abstract*—This project is conducted for Liverpool FC's monthly goal scoring data from beginning of 2000-2001 season to 2023-2024, using time series analysis to bring out patterns and anomalies. Before proceeding, data is transformed for stationarity, conducted SARIMA model with diagnostic checks to confirm model's validity. In addition, other forecasting methods, such as ETS, TBATS, Prophet, NN are conducted to find an accurate forecast result for scored goals in following timeline.

I. Introduction

Time series analysis is very important for estimating and forecasting patterns in data, with application in various fields, such as sports analytics. In this study, the goal-scoring performance of Liverpool FC for the last 24 seasons is analyzed. Exploring monthly scoring trends assisted to uncover pattern, detect anomalies, and build some predictive models that can shed a light into performance variability.

By conducting a detailed time series analysis, the study aims to indicate the growing need for data-oriented approaches in sports analytics. Understanding trends and building predictive models can assist sports clubs to prepare more accurately for the following competitions. In short, this is study is an example of how statistical methodologies can be applied to real-world data to unleash actionable insights in sports.

II. Litareture Review

There is no clear study about Liverpool FC's scored goals analysis. However, a similar study conducted, named "A Comparative Time Series Analysis of Points Scored by Arsenal FC" (Otunoye, Chigozie & Ibeh, 2017). It is conducted for investigating the performance over a 120-year period. Time series analysis is applied to uncover patterns, identify and make predictions about the future. The historical data on points scored per season was organized and digested to satisfy consistency. The study has two major methods: descriptive statistical analysis for trend and ARIMA modelling for the predictions.

The study include processes as data transformation, determining stationarity regarding on statistical tests, and applying ARIMA models to state autoregressive patterns, moving averages and also seasonality. Descriptive statistics enabled to create an overview of fluctuations and displayed significant trends over time. The accuracy of the ARIMA models was verified using residual analysis and forecasting accuracy. The comparison between these two methods revealed the strengths and weaknesses of each methods.

The outcomes of this study revealed that ARIMA models are powerful for short-term predictions. However, descriptive statistics are more efficient for long-term trend analysis and interpretation.

III. Data Description

A. Description

The data consists of all the goals scored in every matches from the initial starting year of the Premier League (1993) until today. For this study, only the matches Liverpool played were included. The last 24 season is filtered because of the changed format of the league. The original data consists six variables:

1. **Season**: indicates the rank of the season from 1 to 24 (categoric)
2. **Match Week:** corresponding match week from 1 to 37 (categoric)
3. **Date:** the date when the match played
4. **Home Team:** states the home team
5. **Away Team:** states the away team
6. **Goals Scored:** goals scored for the corresponding team

To transform the data into a univariate case, only **Date** and **Goals Scored** variables are used in further analysis.

B. Time Series Data

The dataset, retrieved from Kaggle.com, is transformed into time series object to make time series analysis. The frequency is obtained as 10, because the regular season starts at the 8th month of the year and ends at the 5th month of the following year. At 6th and 7th months, the break is given for the competition. Thus, we have no values at these months. The figure stands below indicates the plot of the goals scored according to date by Liverpool FC.

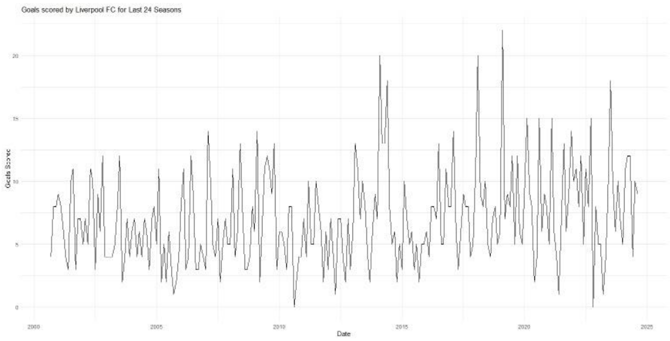


Figure 1: The time series plot of the goals scored by Liverpool

As seen from the plot, there are significant fluctuations that indicates variation in the performance during different periods of each year. These variations can display the nature of the football calendar. For instance, the fixture could have been challenging when there is a downward trend. On the contrary, when the upward trend is seen, we can consider positive factors, such as no injured players, less challenging fixture, even a transferred superstar could influence the performance of the team in positive manners. Another case would be about the year 2020. The pandemic affected the performance of the players due to the compulsory break. The lack of fitness could have influenced the performance of the team in negative aspect. As seen on the plot, there is a downward trend from the beginning of 2020 till the end of that season. These incidents can be explained by existance of the seasonality.

IV. Preprocessing and Methodolgy

Before starting to the analysis, the preprocessing stage takes an important role to ensure the quality and consistency of the data. First, there is a missing value in our data. The reason is that due to the pandemic, at August 2020,
the league was suspended. The case interrupts the consistency of the data. To prevent, the **imputation** method is applied to the corresponding date, by taking the median of the goals scored. After this process, the consistency was satisfied, with having 240 variables.

Another important step is that **Cross-Validation** technique. To perform this methodology, the data is splitted into two parts: test and train data. The last season in the data (2023-24) is taken as test data. The rest of the data is labelled as train data. To proceed our analysis, train data has been used in the following steps.

After conducting cross-validation technique, anomalies should be detected from anomalies. To detect anomalies, decomposition plot assists visually to obtain unusual patterns in the data. For this dataset, it can be seen at the Figure 2.

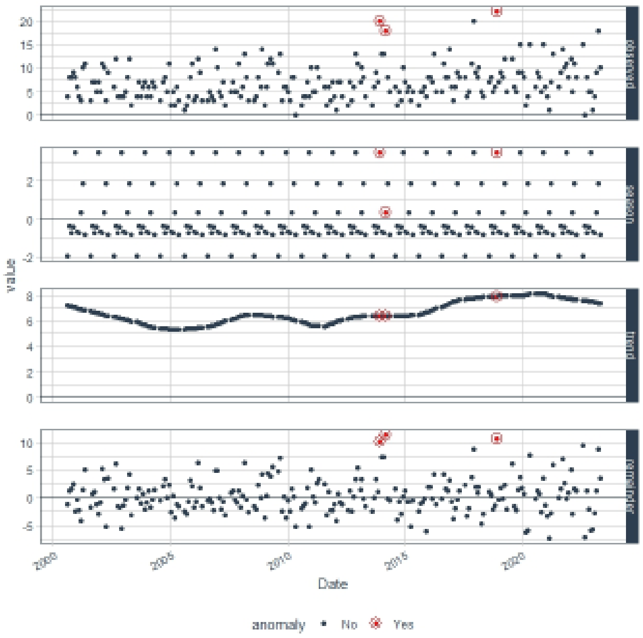


Figure 2: Decomposition plot for the anomalies

Anomalies obtained from decomposition plot gives us a clue about unusal patterns at the end of the season of 2013-14 and at the beginning of the 2019-20 seasons. It seems the team scored a lot more than the usual. Considering both 2013-14 and 2019-20 seasons were the most competitive seasons for Liverpool, these values make sense (Liverpool took the 2nd place at 2013-14 season by 2 points difference, and were the champions at 2029-20 season).

After obtaining anomalies, anomalies were tried to be cleaned. However, for the Box-Cox Transformation, which will be mentioned at the following, only positive values are taken in the case. Thus, the zero values from observations are added with one.

At the last stage of preprocessing, the Box-Cox transformation procedure has to be checked. For determining if the process is necessary or not, the area covered by the curve should be considered. The referenced plot can be seen on Figure 4.

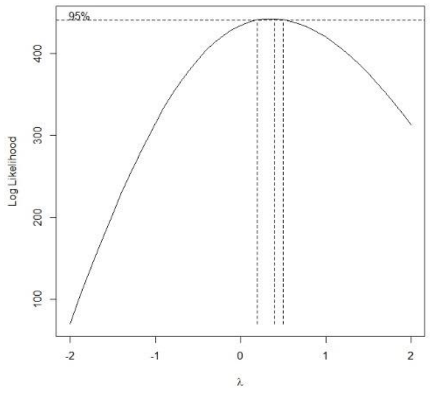


Figure 4: Box-Cox Log-Likelihood Plot

Since the lambda value 1 falls into the area shaded by the curve, the Box-Cox Transformation is necessary. So, the transformation is conducted.

When the ACF and PACF plots are checked, there are significant spikes at lags multiples at 10, which indicates the potential seasonality case, which can be seen on Figure 5.

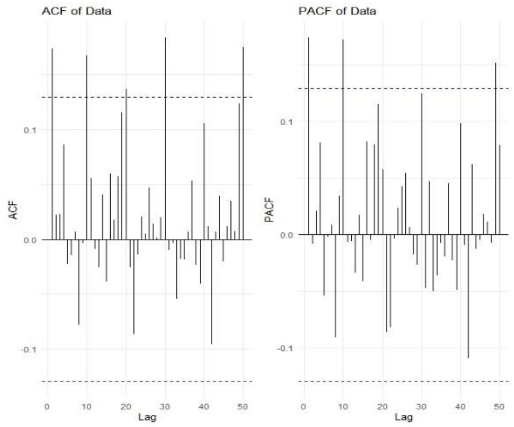


Figure 5: ACF and PACF Plots of the Data

To clarify the stationarity or trending cases, KPSS Test is applied for both level and trend cases. For level case, the p-value is obtained as 0.02344, which indicates the series is not stationary and for trend p-value is 0.1, indicates that the process is deterministic. For unit root analysis, ADF test was applied and the p-value is smaller than the significance level (0.05). Thus, it indicates that we have a regular unit root, means the process is not stationary. At the following stage, OSCB test was conducted for the presence of the seasonal unit root. The reason of the conduction of OSCB test is because of the non-standard frequency. Since the test statistics was smaller than the critical value, the null hypothesis is rejected. In other words, the process has no seasonal unit root. After taking one regular difference on the cleaned and Box-Cox transformed train set, the stationarity and trend are checked with both KPSS and ADF tests. At KPSS test for level for differenced data, p-value obtained 0.1 led us not to reject null hypothesis. Consequently, the process became stationary. The trend is remained in the same status after conducting the same test for trend.

After taking one regular difference, the plot of the series obtained as seen on the Figure 6:

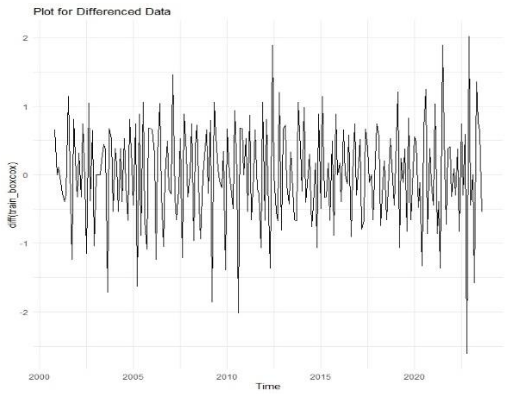


Figure 6: Plot for Differenced Data

Model Selection:

After differencing procedure, stationary process is in the case. The model can be proposed regarding to the both ACF and PACF plots.

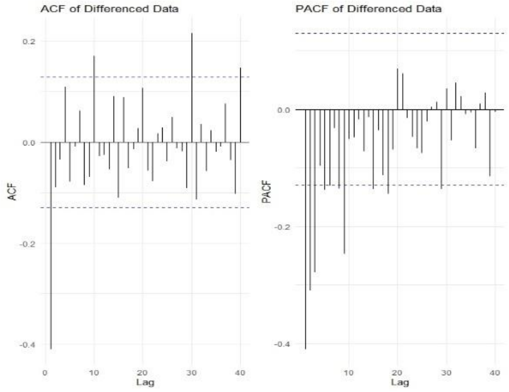


Figure 6: ACF and PACF Plots of the Differenced Data

Considering the significant spikes for the both plots, several models can be proposed. Since one regular differencing is in the case, we can suggest ARIMA model. However, the seasonality case should be checked for an accurate progress. Decomposition plot is one of the useful methods for checking the seasonality.

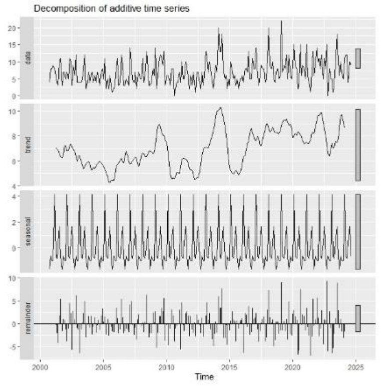


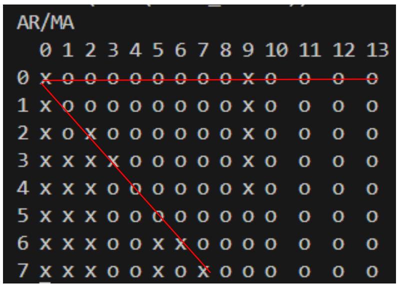
Figure 7: Decomposition plot

The seasonal component in the decomposition plot, *seasonal* panel, represents the repeating, periodic patterns in the data. This component captures variations that occur at regular intervals, independent of overall trend. The condition supports also supports seasonality for the data. In the light of these assumptions, suggesting SARIMA models is more appropriate. From the suggested models, only SARIMA(1,1,1)(1,0,1) was the significant one according to the MLE. So, this process is picked for further processes.

Coefficients	Estimate	Standard Error
AR1	0.1793	0.0664
MA1	-0.99	0.0138
SAR1	0.9049	0.0729
SMA1	-0.7802	0.1104

Table 1: Coefficient estimates for SARIMA(1,1,1)(1,0,1)

In addition, while suggesting a (S)ARIMA model, EASCF plot is one of the efficient visual aid. It provides a more systematic way to determine potential values for p (AR order) and q (MA order).



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o x o  o  o  o
1 x o o o o o o o o x o  o  o  o
2 x o x o o o o o o x o  o  o  o
3 x x x x o o o o o x o  o  o  o
4 x x x o o o o o o x o  o  o  o
5 x x x o o o o o o o o  o  o  o
6 x x x o o x x o o o o  o  o  o
7 x x x o o x o x o o o  o  o  o
```

Figure 8: EACF plot of the train data

Diagnostic checking is vital at time series analysis. Before forecasting procedure, normality assumptions of the residuals should be checked. There are several tests to check normality, such as Jarque-Bera and Shapiro-Wilk tests. The result of the both tests gave lower p-values, indicating that rejection of null hypothesis: normality satisfied (residuals are not normal).

For detecting serial autocorrelation, Breusch-Godfrey test can be conducted. The test is used with fitting a linear regression for residuals used lagged residual values. Since larger p-value is obtained, null hypothesis is rejected, residuals are uncorrelated. Box-Ljung and Box-Pierce tests are also suggests homoscedasticity for the residuals. When the existance of the ARCH effect is investigated, ARCH test gave larger p-value, denotes that there is no ARCH effect. In other words, the variance of the residuals is constant.

Now, we can apply forecasting procedure for SARIMA model. Before conducting, it should be considered that normality assumption is violated.

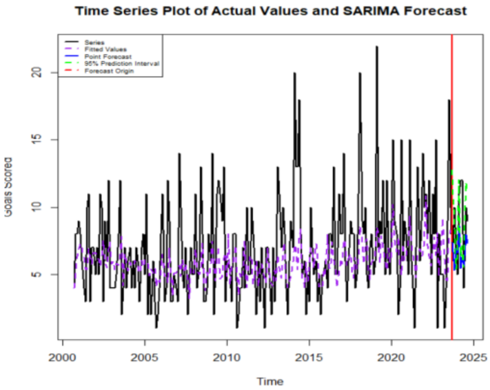


Figure 9: Time Series Plot of Actual Values and SARIMA Forecast

SARIMA(1,1,1)(1,0,1)	Train Data	Test Data
ME	1.028496	1.856588
RMSE	3.634682	3.421498
MAE	2.736256	2.760904
MPE	-18.01245	11.63791
MAPE	53.53607	30.40906
ACF1	0.02973129	-0.03739117
Theil's U	0.6414288	0.6106405

Table 2: Metrics for SARIMA(1,1,1)(1,0,1)

The model is underperformed due to the combination of issues, consisting potential of inedaqueate seasonal modeling. Despite having no autocorrelation problem according to the Breusch-Godfrey test, ACF1 value is -0.037 indicates autocorrelation between residuals. Theil's U test states the model performs better than the naïve model. But, the relatively high MAPE and RMSE values conclude that the model may be overfitting. Also, around years 2015 and 2020 apparently causes significant volatility.

- **ETS and Holt-Winters' Additive**

Before conducting ETS forecasting, the assumptions are checked for normality and serial autocorrelation. For Breusch-Pagan and Shapiro-Wilk tests, p-values were very low. Thus, the residuals of this method are not normal and there is an autocorrelation at the same time. For ETS, the suggested model for the data is ETS(M,N,A), which means there is additive seasonality, non-trend. Since, there is a seasonality but the ETS model did not capture the case Holt-Winter method is used.

Holt-Winters' Additive Method		
Smoothing parameters:		
alpha = 0.031		
beta = 1e-04		
gamma = 2e-04		
Initial states:		
l = 6.8055		
b = 0.0085		
s = -0.7095 1.3482 0.4263 -0.6956 -0.4276 4.2101 -0.762 -0.9741 -0.5024 -1.9134		
AIC: 1839.416	AICc: 1841.659	BIC: 1890.987

Table 3: Metrics for Holt-Winters' Additive Method

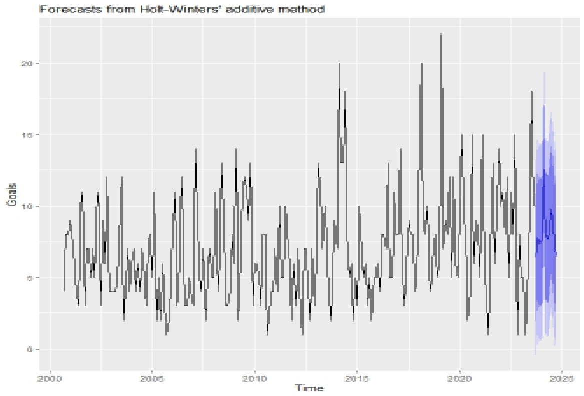


Figure 10: Forecast plot of Holt-Winters' Additive

HW	Train Data	Test Data
ME	-0.00116583	0.24373653
RMSE	3.368517	2.725429
MAE	2.598900	2.186511
MPE	-36.582878	-8.886049
MAPE	59.82885	30.68240
MASE	0.7191926	0.6050722
ACF1	0.1930211	-0.1030800
Theil's U	NA	0.4134325

Table 4: Metrics for Holt-Winters' Additive

The Holt-Winters Additive model appears to fit poorly, as indicated by the high error measurements, especially on the training data. For example, the MAPE (Mean Absolute Percentage Error) for the training set is 59.83%, indicating that the model's predictions deviate significantly from the true values. Additionally, the MPE (Mean Percentage Error) for the training set is -36.58%, indicating a systematic bias where the model consistently underestimates the true values. While the error measurements improve somewhat for the test set (e.g., the MAPE drops to 30.68%), this still reflects suboptimal prediction performance. The high ACF1 value (0.1930) on the training set suggests that residual autocorrelation persists and the model does not account for all patterns in the data.

The poor performance can be attributed to several factors. First, the additive assumption of seasonality in the Holt-Winters method may not adequately capture fluctuations in the data, especially if seasonality or variability changes over time. Second, the data exhibit significant volatility with sharp increases (e.g. around 2015 and 2020) that the model struggles to account for due to its reliance on smooth seasonal and trend components. Finally, external factors or irregular events such as interruptions in game schedules or unexpected performance peaks are not included in the model, further reducing its accuracy. Overall, while the Holt-Winters Additive method provides a baseline estimate, its limitations in handling the complexity and variability of the data result in poor fit and forecast performance.

- **Hyperparametric Tuning Prophet**

The model is poorly fitted because it does not effectively capture the underlying structure and variability of the data. This is evident from systematic underestimation and significant errors in estimation, indicating that the model struggles with complex patterns found in the time series. Residual analysis shows that unmodeled structures remain, indicating that the model does not fully account for observed trends, seasonality, or irregularities.

Hyperparameter tuning was used with Prophet to address these issues. Prophet is a versatile forecasting model designed to address time series with strong seasonal patterns, changing trends, and irregular behavior. By optimizing key parameters such as growth rate, seasonal components, and change points, the model can better adapt to the unique characteristics of the data. This approach increases its ability to account for fluctuations and improves its accuracy and reliability, making it more suitable for capturing the complexities of the series.

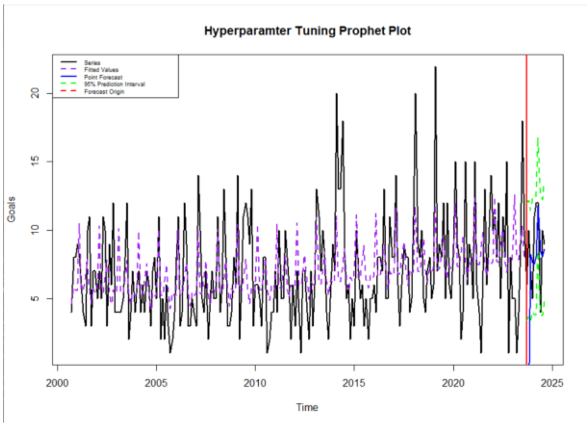


Figure 11: Hyperparameter Tuning Prophet Plot

Hyperparameter Tuning Prophet	Train Data	Test Data
ME	2.245754	0.1237894
RMSE	3.5555	2.523417
MAE	2.901166	2.110821
MPE	17.5886	-10.11182
MAPE	33.36859	30.59457
ACF1	-0.07728978	-0.153902
Theil's U	0.5749015	0.5036969

Table 5: Metrics for Hyperparameter Tuning Prophet

The model is poorly fitted due to several factors. The metrics show high error values with significant deviations from the observed values, especially for the training data. For example, the MAPE for the training set shows significant percentage errors, indicating that the model has difficulty accurately capturing the variability in the data. The residual autocorrelation, ACF1, indicates that some patterns or dependencies in the data cannot be explained.

- **TBATS**

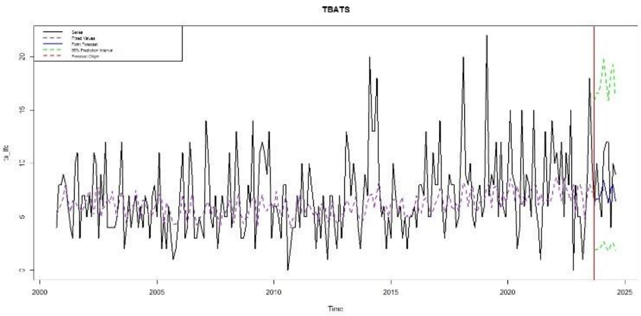


Figure 12: Plot for TBATS

TBATS	Train Data	Test Data
ME	0.7823841	1.4708698
RMSE	3.580768	3.189249
MAE	2.681213	2.755438
MPE	-23.624294	5.068871
MAPE	55.07578	34.08766
MASE	0.7419709	0.7625112
ACF1	0.02232186	-0.14073305
Theil's U	NA	0.5591312

Table 6: Metrics for TBATS

Before conducting the test, residual analysis claim that the residuals not distributed normally by Shapiro-Wilk test. For Breusch-Pagan test, the serial autocorrelation does not exist. The train set metrics show the model performs adequately in capturing the structure of the data, though the MAPE (55.08%) shows higher errors, possibly due to overfitting seasonal fluctuations. For the test set, the performance is improved, with a lower MAPE (34.08%) and minimal residual autocorrelation (ACF1 of -0.1407), suggesting the model generalizes better to unseen data.
The low Theil's U value (0.5591) confirms that the model performs better than a naive forecast, especially for the test set.

Hyperparameter Tuning in Neural Network

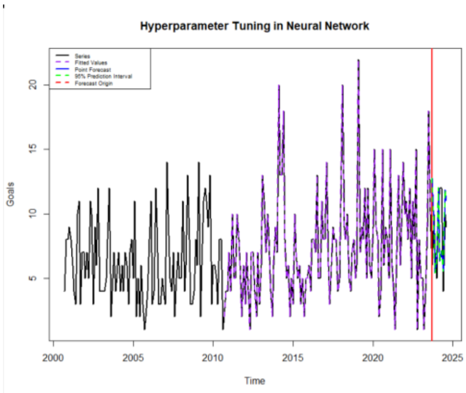


Figure 13: General plot for Hyperparameter Tuning in NN

Hyperparameter Tuning NN	Train Data	Test Data
ME	2.490193e-05	-1.891531e-01
RMSE	0.0003392875	3.4523341840
MAE	0.0002681646	2.7984200067
MPE	7.088123e-04	-1.158955e+01
MAPE	0.004971831	35.532136394
MASE	7.420906e-05	7.744055e-01
ACF1	-0.118024	0.138204
Theil's U	NA	0.4237009

Table 6: Metrics for Hyperparameter Tuning NN

On the training set, error metrics such as MAPE (near zero) and RMSE (near zero) indicate that the model is effectively capturing patterns in the data. However, this overly sensitive performance indicates overfitting, where the model fits the training data too tightly, memorizing noise or specific patterns rather than learning generalizable features.

For the test set, performance deteriorates, as indicated by a MAPE of 35.53% and a high MASE (7.74). This indicates that the model’s predictions for unseen data are less accurate. The residual autocorrelation (ACF1) for the test set is 0.1382, indicating that some patterns remain unmodeled and may contribute to prediction errors. However, the low Theil’s U value (0.4237) indicates that the model still outperforms a naive prediction.

For the best selection for the models is hyperparameter tuning in prophet with the lowest RMSE, 2.52, and MAE with 2.11. Minimum residual autocorrelation was obtained in this model. Thus, this is the best selection.

V. Conclusion

The analysis across various models show several key negativites that contributed to below standard fit and forecasting performance. One major issue was the complex and high variate time series structure, characterized by irregular patterns, strong seasonality, and occasional spikes, which many models struggled to capture effectively. For instance, simpler models like Holt-Winters' additive and SARIMA model couldn't be adaptive to the series' irregular vacillations, leading to high errors and residual autocorrelation. Another significant factor was overfitting, particularly evident in the hyperparameter tuning neural network model, where the training error was minimized to almost zero but the test set performance deteriorated. This indicates that the model memorized the training data patterns rather than learning generalizable trends. Similarly, TBATS, while better suited for handling complex seasonality, showed limitations in addressing high variability and irregular spikes, resulting in moderate errors on both training and test sets. The data's non stationarity and residual dependencies also faced challenges. Despite transformations and differencing, some models failed to account for all underlying patterns, leaving unmodeled dependencies. Furthermore, the exclusion of potential external factors such as competition schedules or team performance parameters, limited the capability of these models to explain and predict real life variability. Lastly, hyperparameter tuning played a critical role in improving model performance as seen with the Prophet and neural network models. However, non proper tuning in some cases resulted in either underfitting or overfitting and further impacting the models' ability to generalize effectively.

Consequently the primary drawbacks were not being able to fully capture the data's complexity overfittings in some models insufficient handling of variability and residual dependencies and limitations in tuning hyperparameters. Indicating mentioned issues through advanced models, external variable integration and better standardized techniques can improve forecasting accuracy for such complex structured time series.

References

Cryer, J. D., & Chan, K.-S. (2021). BoxCox.ar. TSA Package Documentation. Retrieved from https://www.rdocumentation.org/packages/TSA/versions/1.3/topics/BoxCox.ar

Otunoye, Chigozie & Ibeh. (2017). A comparative time series analysis of points scored by Arsenal Football Club. European-American Journals