Comparisons of Several Forecast Methods On Delivery Time Dataset in NewYork

**Kardelen ÇAKABAY**

Statistics  
METUAnkara, Türkiye  
cakabay.kardelen@metu.edu.tr

In this study “Future Delivery Time; Spread Index for New York.” data is analyzed by using various time series forecasting models including ARIMA, ETS, TBATS, NNETAR, and Prophet. The dataset from the Federal Reserve Bank of New York’s Empire State Manufacturing Survey, provides monthly, seasonally adjusted indices that reflect expected delivery times in New York State over the next six months. Our goal was to evaluate and compare the predictive performance of these models to determine the most accurate approach to predicting future delivery times. The findings show that The NNETAR model provided the best fit for the data and TBATS model demonstrated superior performance among them.

Keywords— time series, forecasting, Arima, ets, tbats, neural network, prophet

# Introduction

## Data Description

The dataset consists of seasonally adjusted monthly time series data from July 2001 to November 2024 with 269 observations.

The source of this Future Delivery Time; Spread Index for the New York dataset is the Federal Reserve Bank of New York’s Empire State Manufacturing Survey.

## Aim of the Study

The primary objective of this study is to analyze and compare various time series forecasting models to predict future delivery times in New York State. By conducting models like ARIMA, ETS, TBATS, NNETAR, and Prophet, the aim is to evaluate their forecast performance and identify the most accurate approach. The findings from this research will contribute to a better understanding of the dynamics affecting delivery times and improve decision-making processes for businesses relying on timely delivery forecasts.

## Methodology

In this analysis, as a software tool R programming language was used with various libraries such as tseries, forecast, and ggplot2.

Additionally, several statistical methods including model fitting, forecast, and visualization techniques were employed. Among the visualization techniques ggplots and ACF - PACF plots were mostly used.

To begin, the tseries library was used to convert the data frame dataset into a time series data class. For statistical analysis, the forecast library was mainly relied on. Other than that, to check the assumption and find out the structures lmtest, TSA fUnitRoots and similar libraries were used. To visualize the data and findings, ggplot2, gridExtra, anomalies, and timetk were used. Throughout the analysis, a systematic approach was followed, selecting appropriate statistical methods and visualization techniques based on our aim for comparing different methods. The combination of R programming language and the mentioned libraries provided us with the necessary tools and resources to analyze the Future Delivery dataset and make meaningful conclusions effectively. Overall, stationarity, normality, heteroscedasticity, and autocorrelations assumptions for effective forecast were checked. To conduct these, firstly Kpss, Heggy, Shapiro-Wilk, Lung-Box, and Lagrange multiplier tests were used. In the end, with the accuracy function, the accuracy measures of forecasts are used to compare different conducted forecasts.

# basıc ınterpretatıon based on ts plot

A graph of a time series

Description automatically generated

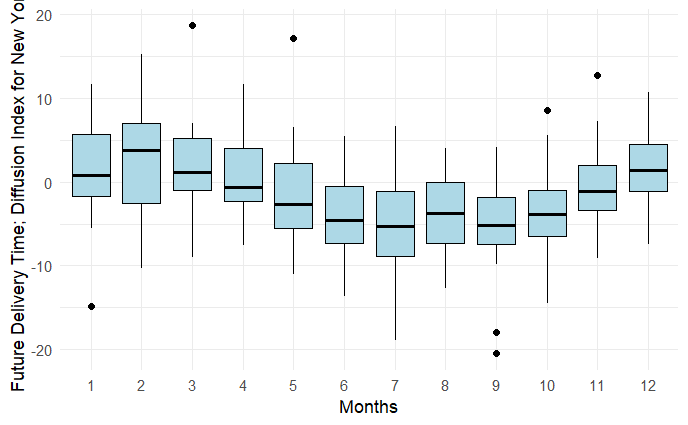
In the time series plot, no decrease or increase suggests the absence of a trend in the data. However, there may be seasonality. Moreover, the significant spikes observed suggest the presence of outliers in the data. These anomalies may indicate unusual events or irregular patterns that deviate from the expected seasonal behavior.

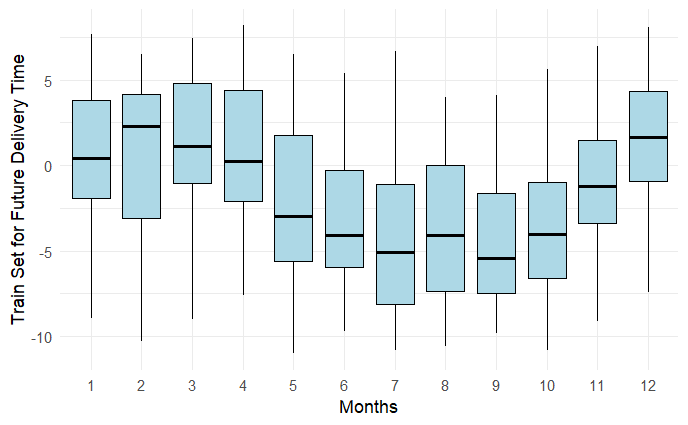
# ANOMALY DETECTION

Before anomaly detection, some arrangements should be made. Firstly, the dataset is split into two parts a train set and a test set. From now on, the analysis will be applied to the train set, and at the end, the test set will be used to compare forecast performance between all the methods.

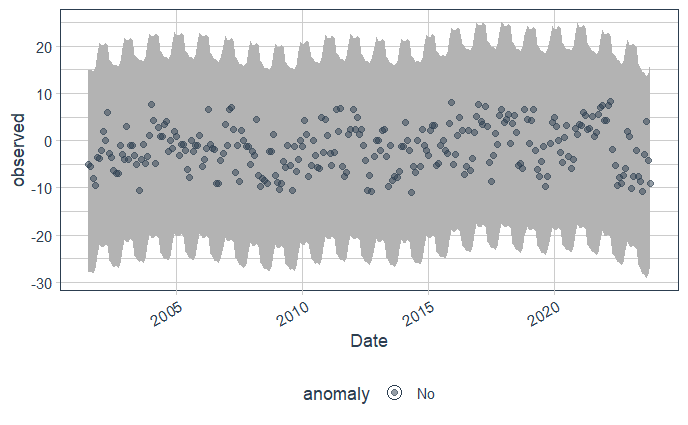
Secondly, outliers and missing values should be identified and replaced. The training data does not have missing values, but it contains outliers. To replace all the outliers, a threshold of 1.6 for the z-score was used, and the values exceeding this threshold were assigned NA, and then filled with the surrounding data points.

A comparison of boxplots representing the data before and after the outliers were replaced can be found below.

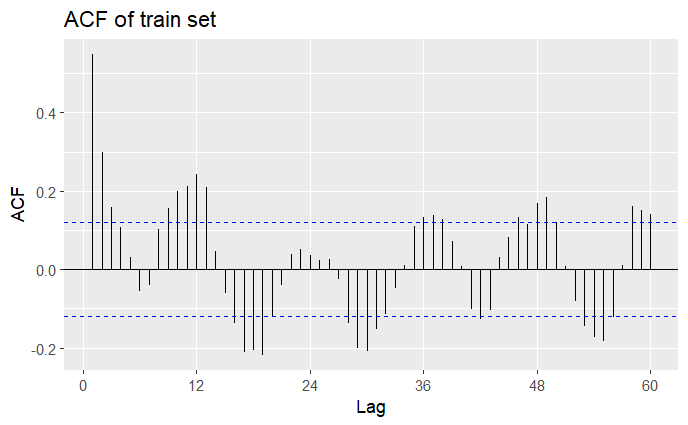


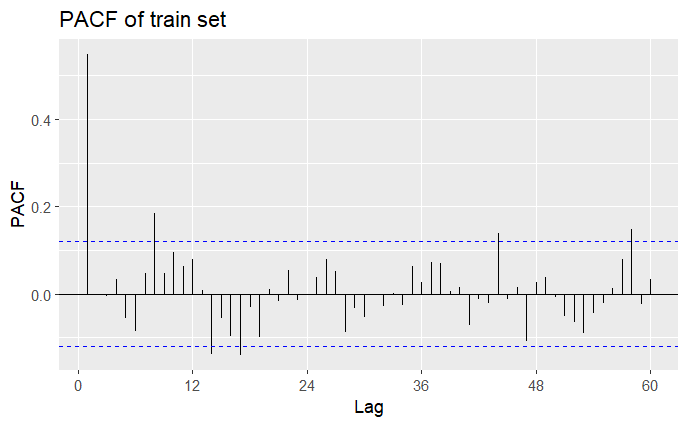


For Anomaly Detection, a tibble format is needed, first, we need a dataframe then tibble from time series data.

.

For further interpretation, ACF&PACF plots can be used.





According to the ACF and PACF plots of the train set, the sinusoidal behavior is seen in the ACF plot, and there are significant spikes indicating seasonality in the PACF plot.

Now, the stationarity assumption should be checked. To do that The Kpss test, ADF test, and Heggy tests were conducted and both seasonal or unit roots were not identified. As a result, the train set shows no signs of non-stationarity. Moreover, to check whether the differences between the years are statistically significant or not, the ASCB test was applied. After analysis, the p-value was found approximately 1.29 × 10⁻⁶ indicating that the mean future delivery times differ significantly across the years in the dataset.

Additionally, as seen in the box plot of the train set, the median does not follow a horizontal zero line. It indicates seasonality in the dataset.

To find a proper ARIMA model or SARIMA model for our train set, some possible orders can be identified from ACF & PACF plots of train data. There is one significant spike at the PACF plot (AR-1), ACF shows sinusoidal behavior. It can be also seen significant seasonal spikes in the ACF plot. To detect the most appropriate model, twelve models were compared, and ARIMA-SARIMA(1,0,0)(0.0,1)12 is found best one as it has significant parameters and small accuracy criteria.

On the residuals, the Portmanteau lack of fit test is performed and it is found that residuals are likely white noise and the model fits the data well at the significance level of 0.05. After fitting the Arima model, four of the ACF&PACF values of residuals are not in the WN bands. It could be an indicator that the ARIMA model is not fully capturing the underlying structure of the data.

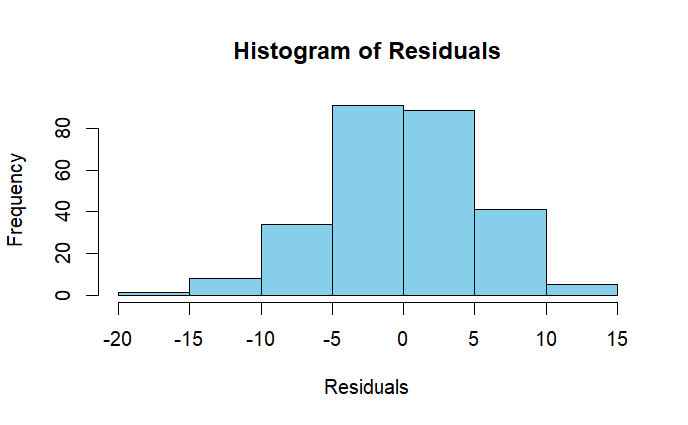
Since only 4 lags fall outside the white noise (WN) bands in the ACF/PACF plot, and the Ljung-Box test gives a p-value of 0.4 suggests that the residuals are mostly random, with a small portion showing some autocorrelation at those specific lags.

For further interpretation, Squared Standardized Residuals can be analyzedA graph with red lines

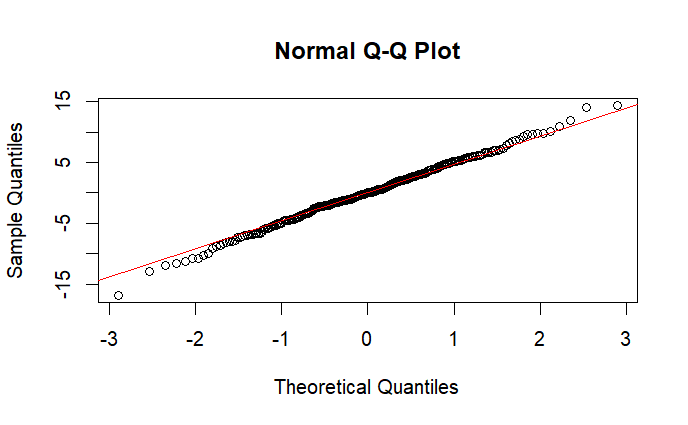
Description automatically generated

The plot shows significant spikes, indicating the existence of outliers.

While no clear trend is observed, the clustering effect suggests the presence of heteroscedasticity, where the variability of the residuals changes over time. For normality of residuals, histogram and Q-Q plot of residuals give good visual detection.



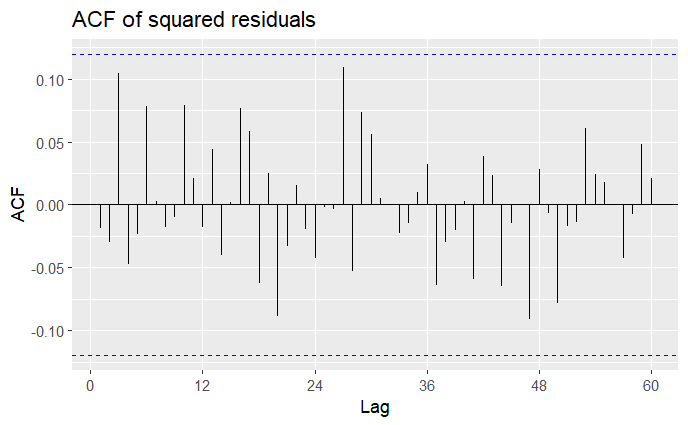
It is slightly left-skewed but mostly symmetric and seems normal.

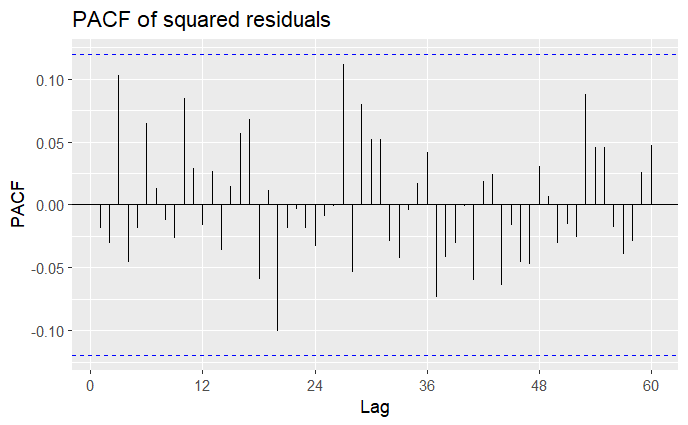


There are outliers and the plot shows a slight S shape indicating residuals may have heavy-tailed distributions.

As a formal test for normality, the Shapiro-Wilk test is used. Since the p-value (0.9) is greater than 0.05 it is found that residuals are normally distributed. Moreover, based on the Breusch-Godfrey test, the result was found insignificant, the model does not show significant autocorrelation in the residuals, which suggests that the model appropriately captures the underlying temporal dependencies.

For the Heteroscedasticity ACF-PACF plots of the squared residuals can be performed.



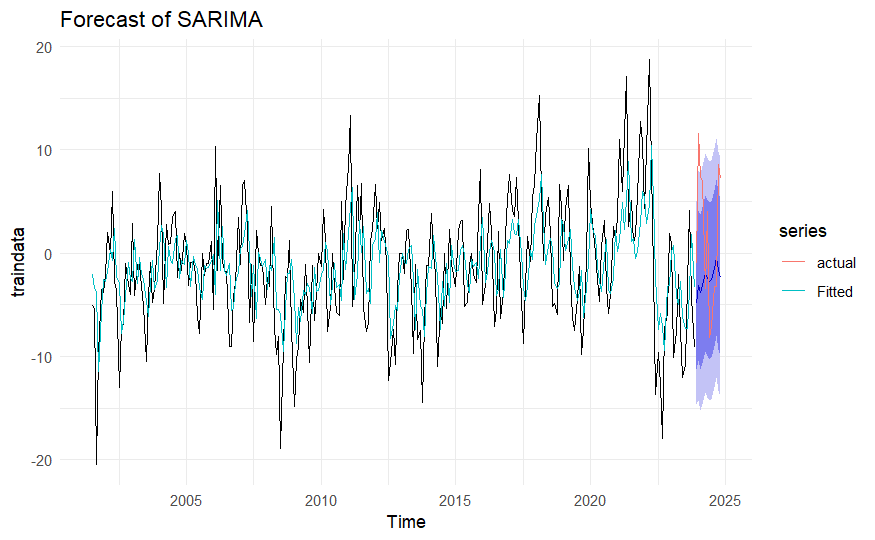


All squared residuals are in the white noise bands. It can be concluded that there is no heteroscedasticity problem, error variance is constant. Other than that, ARCH Engle's test can be used for the same aim as a formal test. As a result of this test, all the p-values are greater than 0.05, indicating there is no significant evidence of heteroscedasticity in the residuals at any of the tested lags.

# FORECASTING

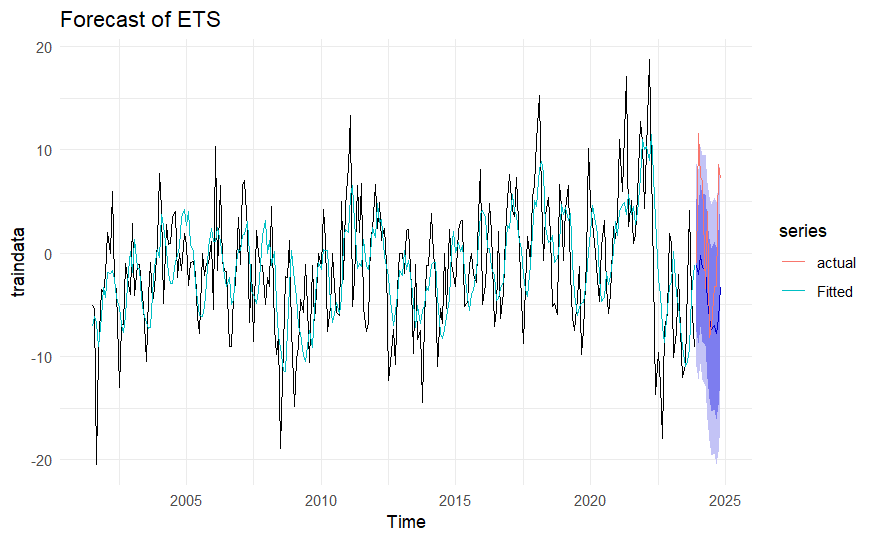
In all forecast plots, the black line represents the original data, while the red line corresponds to the test set.

Firstly, the Minimum MSE Forecast for the stochastic ARIMA(1,0,0)(0,0,1)12 model was performed.



The prediction intervals are notably wide, suggesting that the ARIMA model provides a weak fit and lacks precision in forecasting.

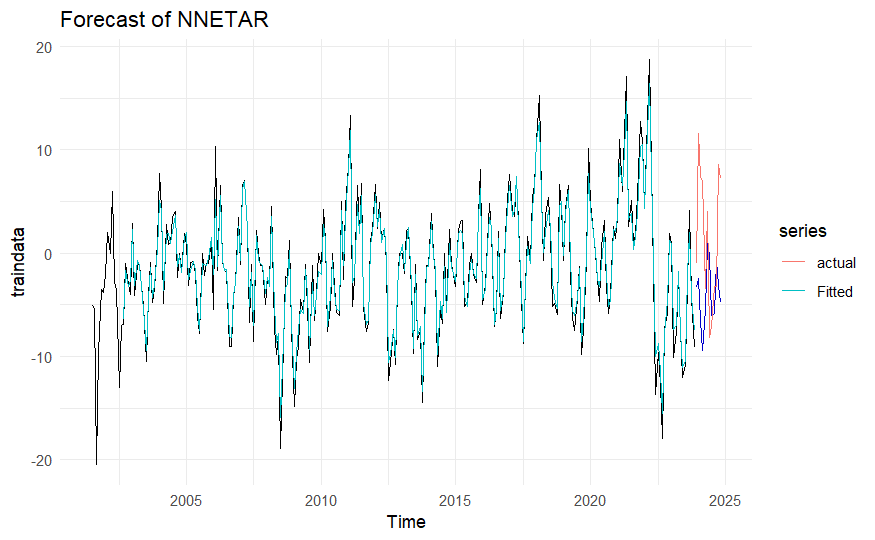
Secondly, the Ets model is used with no trend but seasonality under (A, N, A) parameters. The data exhibits additive errors and additive seasonality, but no discernible trend. Later, the assumption for the normality of residuals was checked and its normality was concluded with the Shapiro-Wilk test.



The ETS model weakly captures the seasonal behavior, and the prediction intervals appear wide, indicating a lack of strong fit.

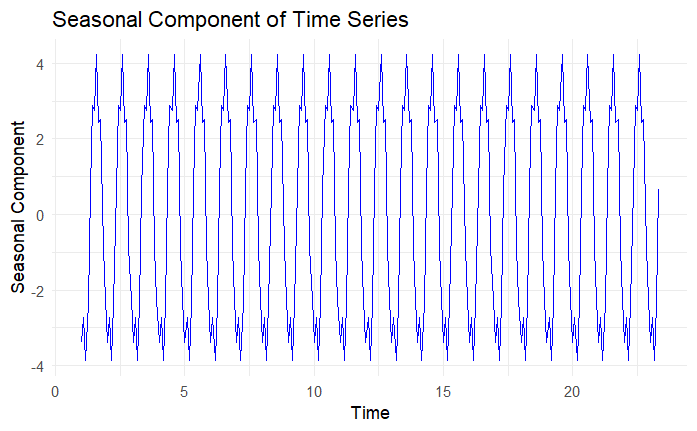
Thirdly, the analysis of the NNETAR model was conducted. As seen by this training data, this model works best when the dataset shows seasonality.

To avoid possible overfitting, the RMSE values for the training and test sets were examined before fitting the model. Since the RMSE values are very close to each other (7.95 for the training set and 7.97 for the test set), it can be concluded that there is no overfitting issue.



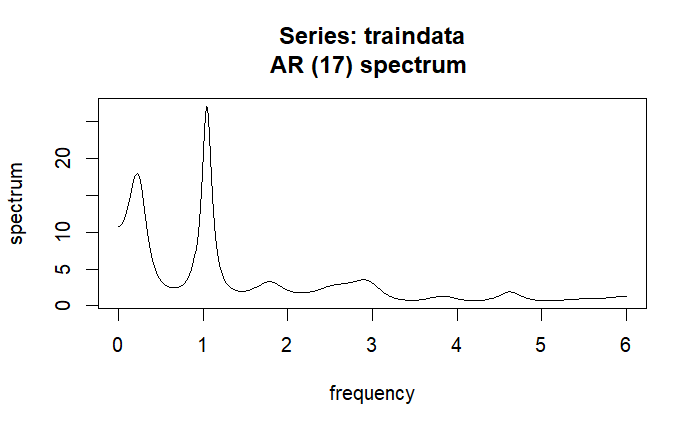
The training data is well-fitted by the NNETAR model. It may be also a robust and dependable model since it efficiently captures the test data's peaks with less variance.

Fourthly, the TBATS model is implemented. It is frequently useful when the data shows complex seasonality, such as multiple seasonalities, high seasonal frequency, or non-integer seasonality. Using seasonal decomposition methods, we can examine the training set for multiple and non-integer seasonality.



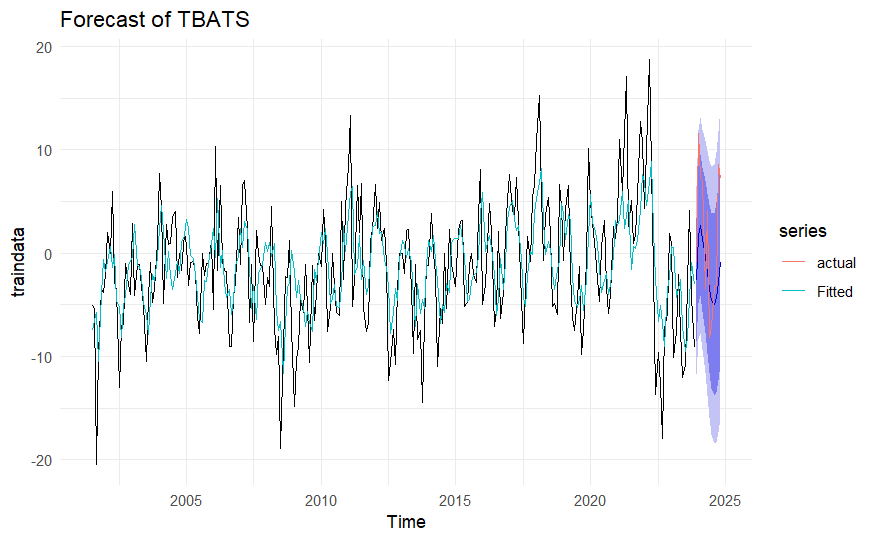
This seasonal component decomposition plot shows two peaks at the bottom of the patterns. However, it may not be enough to say there are multiple frequencies. Also, since the plot does not show irregular patterns, it could not demonstrate non-integer seasonality.

Several ways are available to detect multiple seasonalities, such as spectral analysis, MSTL decomposition, and the summary of the TBATS model.



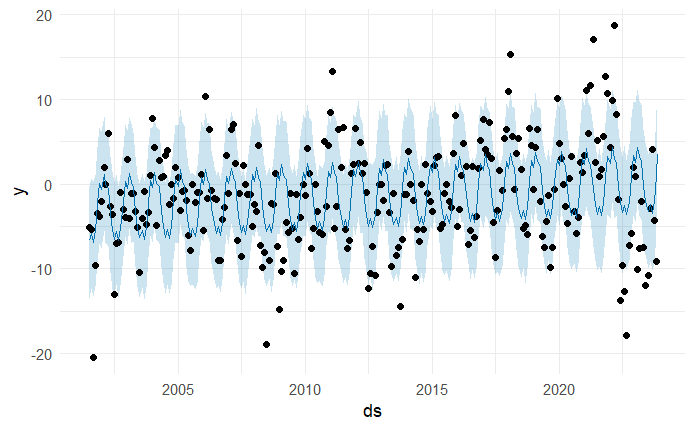
After applied spectral analysis it is seen that a yearly seasonality, which is typical in monthly records, is shown by the spike at frequency 1. As a result, it is unlikely to decide that there are multiple seasonalities. The trend is represented by the spike at frequency 0, however, the data we checked showed no trend. It could also indicate the mean level of the training data.

Before conducting the TBATS forecast analysis, the normality of the TBATS model’s residuals should be checked. According to the Shapiro-Wilk test p value is 0.6 indicating our residuals are normally distributed.



With the broadest prediction intervals, the TBATS model appears to have considerable forecast uncertainty. This shows a poor fit to the data, even if it captures the fundamental fluctuational behavior.

Lastly, the Prophet model was used to forecast the train set. Instead of using a classical approach, the model was constructed with hyperparameter tuning, as its accuracy metrics for test data, such as MAE and RMSE, were smaller than those obtained using the traditional method.



The above forecast plot shows the forecasted values of the data where, Black dots refer to the original data, Dark blue line refers to the predicted value(yhat), and Light blue area prediction interval.

There are outliers outside of the prediction intervals even though the Prophet model accounts for seasonality. This indicates that the Prophet model is not well-suited for our data.

# COMPARING ACCURACY OF THE MODELS

# A screenshot of a graph Description automatically generated

According to the accuracy values for both the train and test sets, the smallest values were chosen, leading to the following conclusions:

1. **NNETAR** provided the best fit for the training data, outperforming the other models in terms of fitting accuracy.
2. **TBATS** performed the best for the test data, as it forecasted more accurately compared to the other models.

# CONCLUSION

In this study, various time series forecasting models – ARIMA, ETS, TBATS, NNETAR, and Prophet– were analyzed and compared to predict Delivery Time in New York State. Through a systematic methodology, necessary assumptions were tested both before and after applying the forecasting methods.

After an in-depth analysis of the Future Delivery dataset, some significant findings were made. First off, one of our assumptions is that the dataset does not have enough complexity to support the TBATS model. The TBATS model explained its potential to manage the underlying structure of the data and successfully produce the most accurate forecast measurements for the forecast. This result shows the TBATS model demonstrated its ability to handle not only complex seasonality but also datasets with simpler seasonal patterns, successfully capturing the underlying seasonal behavior and producing the most accurate forecasts for the test data.

On the other hand, despite capturing the train data well, NNETAR model did not perform as strongly as TBATS when applied to the test set even though there is no indication of an overfitting problem.

To conclude, the systematic use of diverse forecasting methods revealed the strengths and weaknesses of each model. The NNETAR model provided the best fit for the data and TBATS model demonstrated superior performance among them.