Kapil Bastola

04/22/2017

Assignment 10

**Results**

After downloading and loading the data into zeppelin, first step taken was to understand the data and descriptive statistics behind it. Looking at the different variables present in the dataset average speed is chosen as the variable to mimic traffic flow since higher average speed means better flow of traffic and lower average speed means traffic flow is not very good. Time of day is one of the biggest factor in determining traffic flow as more vehicle are in the roads during day hours, especially during commuting hours, compared to night time. When we look at the hourly average speed in our dataset to verify the previous statement we can confirm that the average speed is considerably lower between 6 am and 2 pm. We also looked at vehicle count by hour and found that most vehicles are in the roads between 6 am and 2 pm. Both these results show that we are in right path to use time of day and vehicle count to predict average speed in the dataset. Scatter plot of average speed versus vehicle count shows a negative correlation between the two variables, as expected.

After the initial descriptive analysis we start out with a simple decision tree to predict average speed based on two variables hour of the day and vehicle count.

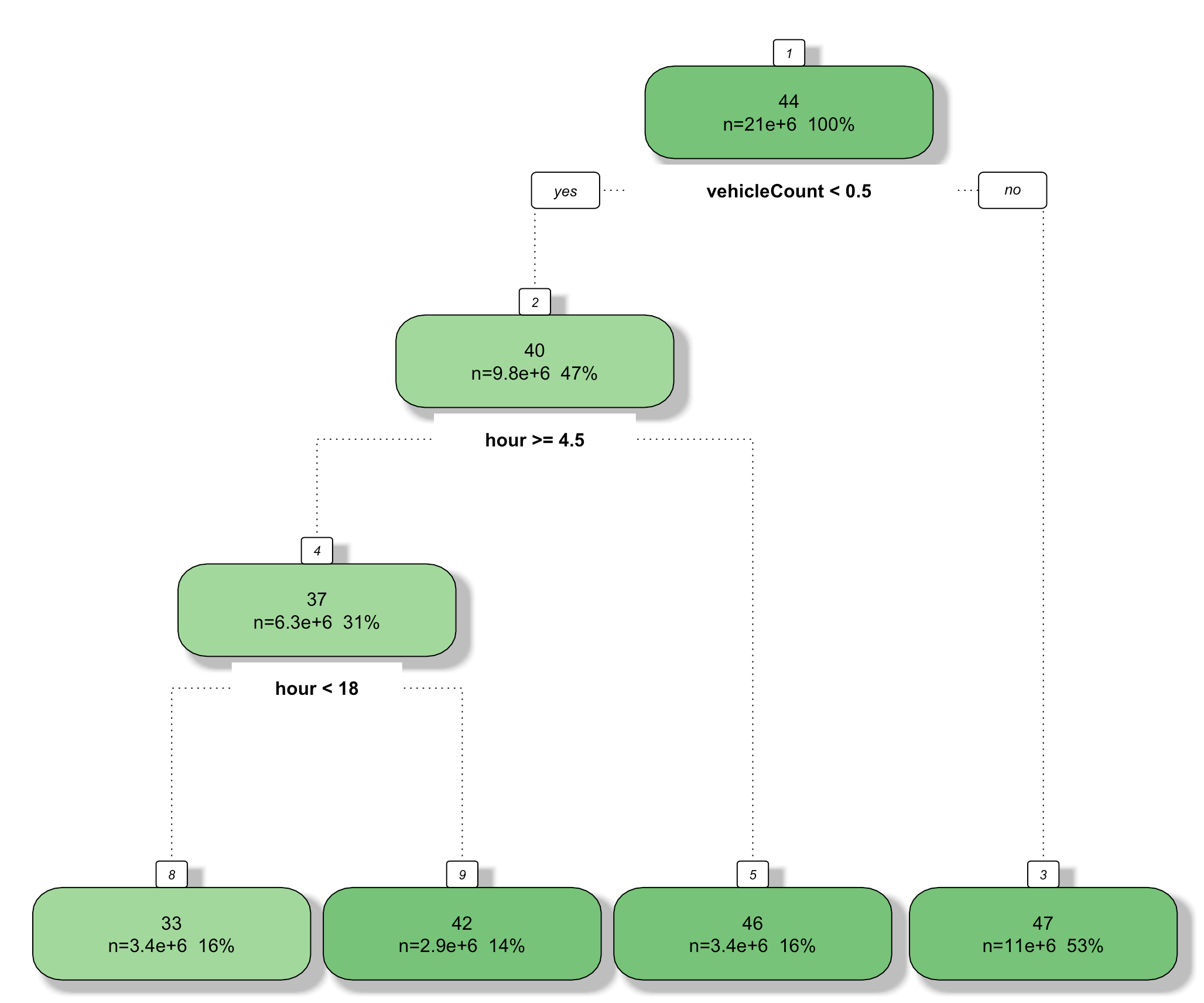


Figure : Decision tree representation for traffic flow data showing different nodes and features

We got a decision tree as shown in the picture above. First node of the decision tree is whether the vehicle count is higher than 0.5 or not. If it is higher than 0.5, average speed is 47. If the vehicle count is lower than 0.5 we encounter another node, which examines whether time of day is greater than 4 am or not. If the time of day is less than 4 am, average speed is 46. If the time of day is greater than 5 am we encounter another node that examines whether time of day is after 6 pm or not. If time of day is after 6 pm the predicted average speed is 42 kph and if it is before 6 pm and after 5 am, the predicted average speed is 33 kph. We looked at the error of this model and saw that the average MSE is 478.

Next, we wanted to do time series analysis on the average speed. We first aggregated the 20 million rows data into hourly data for the time series analysis. After sorting the data by day and hour, we have an ordered tie series data that can be used for time series analysis. Using R and forecast library, we use auto.arima function to run an auto regressive integrated moving average (ARIMA) model to create a fit that can be used to predict future data. In order to use the ARIMA model we need to provide which level we want to set the model at. We can utilize just the auto-regression (AR) part of the model by setting to a numerical value or we can just set the moving average (MA) portion of the model.

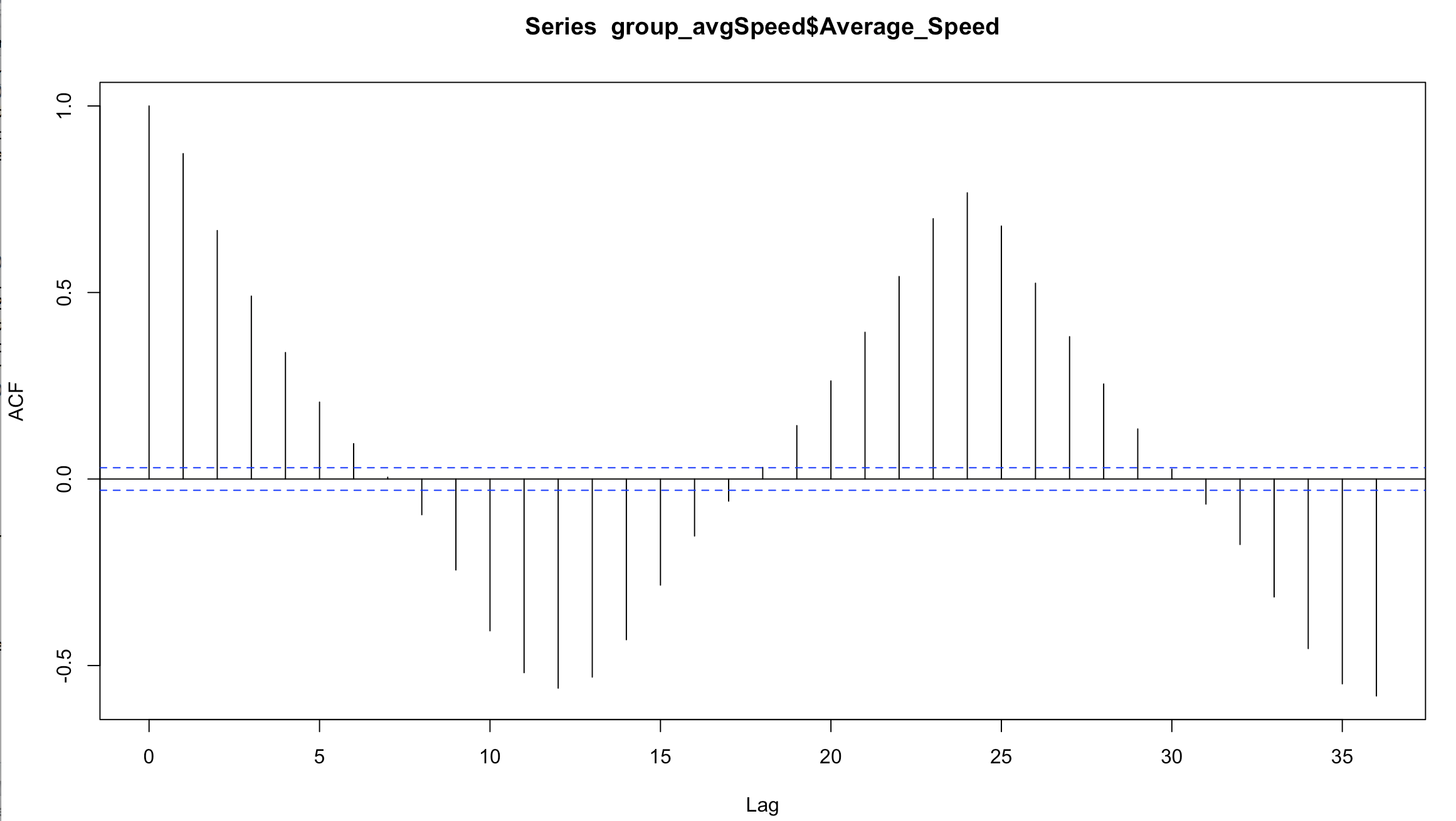


Figure : Auto Correlation Function (ACF) of the hourly data. This shows significant moving average lag at lag 7 and significant seasonality at lag 18.

In order to determine which part of the model to choose and at what orders, we look at the auto correlation function and partial auto correlation function. Auto correlation function (ACF) gives us the correlation between a time series and the lags of itself while the partial auto correlation function (PACF) gives us the correlation of a time series with its lag that is not explained by correlation at lower order lags. If the ACF shows any significance at any of the lags we use that order of lag for the moving average and if PACF shows any significant spikes at any of the lags we use that order of lags for the auto regression. After we set the order for the model we can generate the model using R. We can then predict next series of data points based on the model. In order to predict the accuracy of the model we can use mean absolute percentage error (MAPE). Looking at the value of MAPE will tell us whether our model is highly accurate, good, reasonable or inaccurate.

Auto.arima function in the forecast library selected an order of 4 for AR, 1 for I, and an order of 4 for MA. We got a Mean Absolute Percentage Error (MAPE) of 2.52. We also predicted the last 12 values of the time series. When comparing the actual versus the predicted and the variation in the actual speed is not as prevalent in the predicted speed. Also looking at the Ljung-Box test for significance, we see that there are first few lags that are not significant.

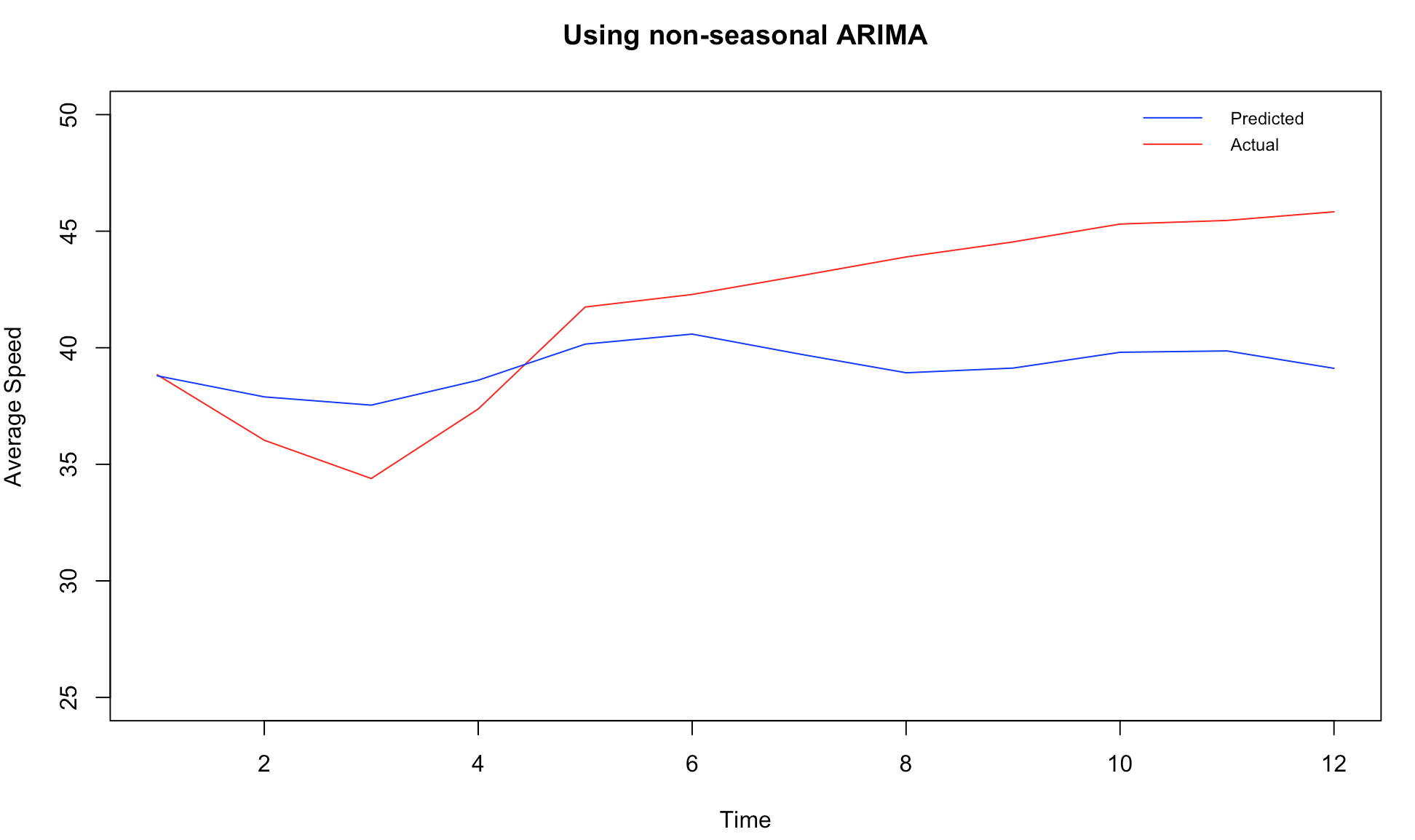


Figure : Actual versus predicted result from non-seasonal ARIMA. This shows how predicted values do not follow actual values after 5 or so points

We wanted to pick our own orders for AR and MA by looking at the ACF and PACF. Looking at the ACF and PACF we see that there is some seasonality in the data and decided to add a seasonality component to the ARIMA and performed an ARIMA order of (0,2,7) with a seasonality of (0,0,18). Using these orders, we reran the ARIMA fit and predicted the last 12 hours of average speed of the time series again. For this model we got an MAPE of 2.19, which is an improvement from the previous time series model. For the 12 hours of predicted average speed versus the actual average speed we see that the variation in the actual values is represented in the predicted values as well as seen in the figure below. When looking at the Ljung-Box test for significance, we see that all the lags has p-values less than or close to 0.05 meaning a statistically significant result. This tells us that the seasonal ARIMA is a better model compared to the non-seasonal one.

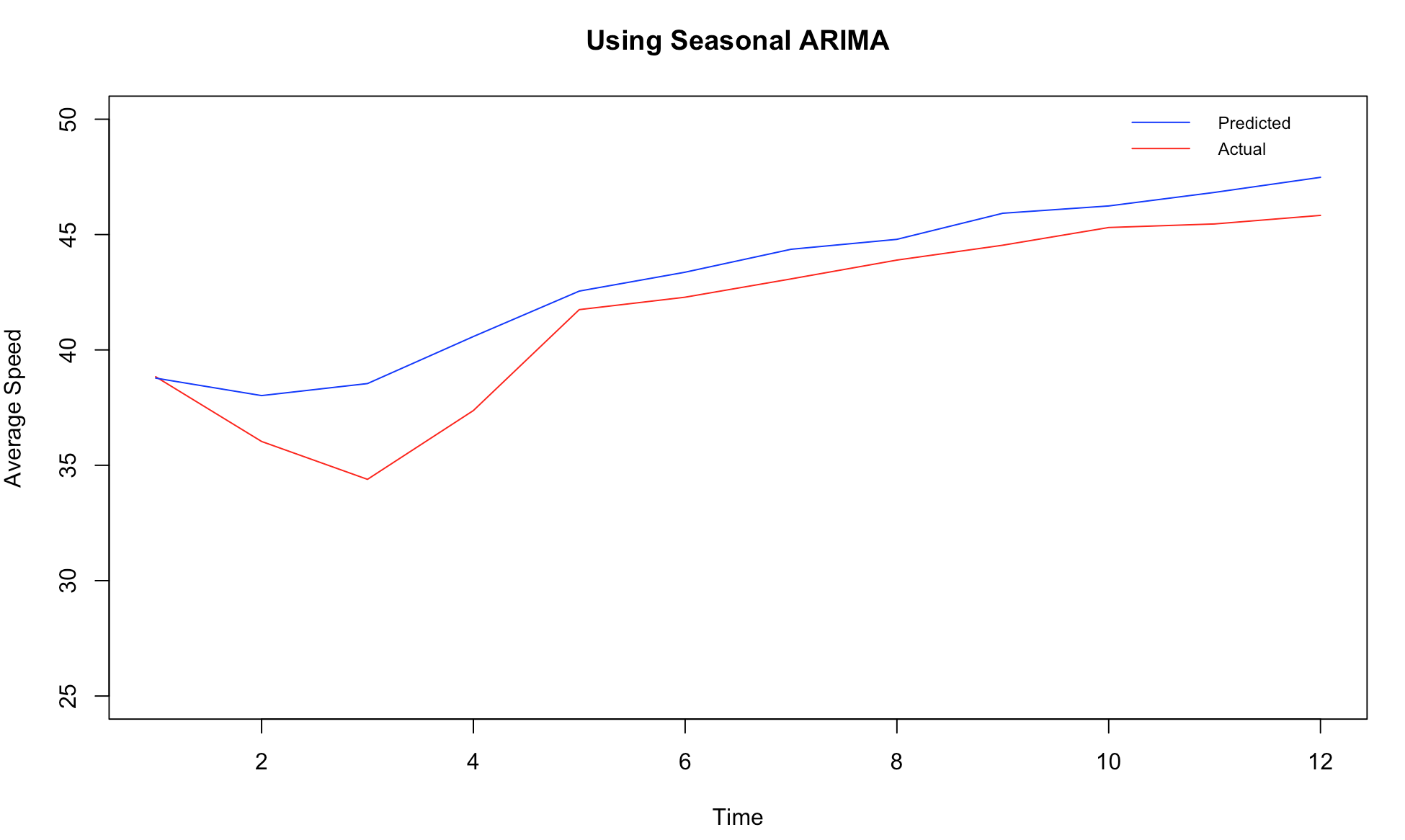


Figure : Actual versus predicted result from seasonal ARIMA. This shows how predicted model closely follows actual result