Modeling and Forecasting Weekly Walmart Sales Data From 2010 To 2012

By: Javiy Wang **May 4, 2021**

Table of Contents

Introduction

The sale of Walmart products are influenced by both temporal and circumstantial variables. Our interest in forecasting Walmart sales allows Walmart retailers to be more efficient with stocking, thus enabling demand oriented decision making. This in turn will reduce inventory costs, while also ensuring an adequate number of products in the store. I was always one that disliked having a surplus of anything, thus, I decided to look into Walmart sales for my project. If an adequate model could be built and accurate predictions could be made on the number of sales made every week, the data could be used to prepare and properly decide which months/weeks throughout the year will need more staff and inventory. Additionally, if an adequate model was built, not only could we reduce our yearly costs, we could even make predictions on which products and departments are most profitable across all Walmart stores during the holidays and most likely increase our profits as well.

My data set is available on *Kaggle,* it provides weekly time-series sales data at a department and at store level, covering 45 of its stores from 2010 to 2012. Since the data set is separated into 45 store levels and 99 department levels, this data set has over 400,000 unique observations. For simpler analysis one could either view a single store at a single department level or aggregate the data based on stores and departments so we could have the sum of weekly sales, this would lead us to have 143 unique observations. I will be using the latter for my analysis. I will mainly be focusing on forecasting analytics which will take into consideration the trends in weekly sales in relation to time and seasonal trends rather than circumstantial variables such as store and department. I will ultimately be fitting and comparing two ARIMA models, one regular and one seasonal, and then choosing a final model based on each model's diagnostics.

Model Specification

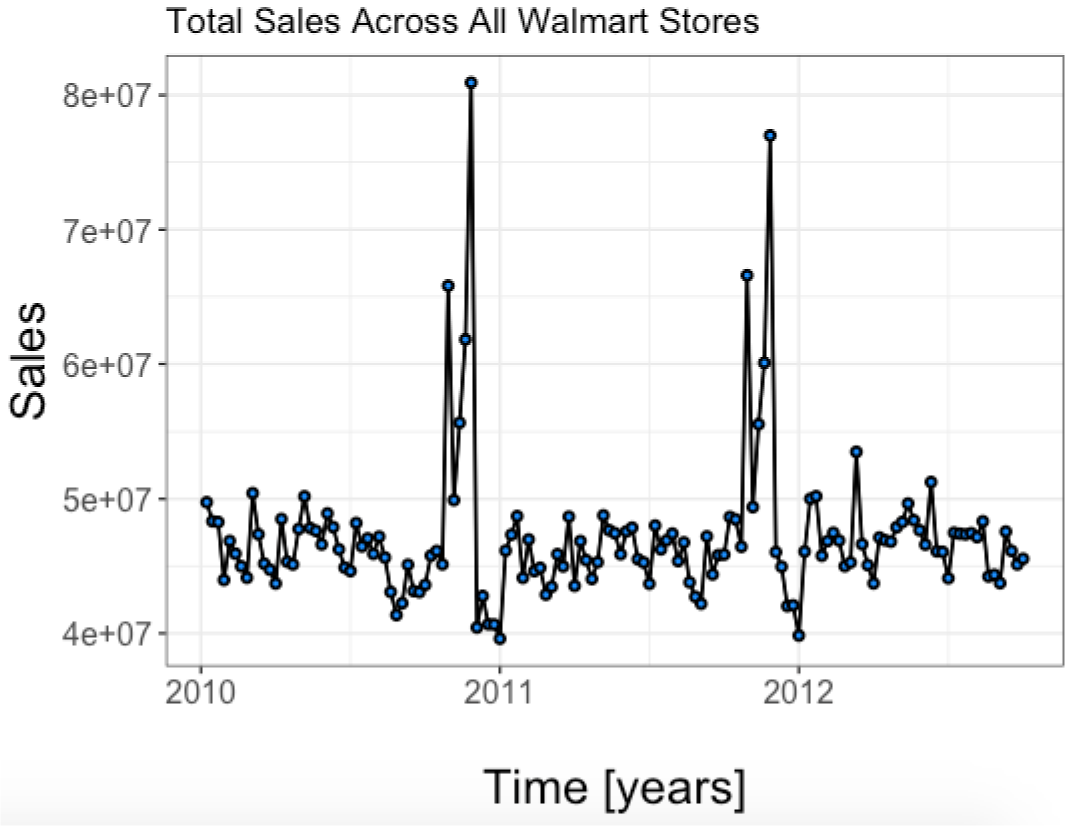


Figure 1: Time Series Plot of The Total Number of Weekly Sales.

We can clearly see that the data has seasonality, and therefore is it not stationary, since there is a clear pattern in Figure 1. Furthermore, in the ACF plot, we can see that there are significant correlations at time lags 1 and 5 as shown in Figure 2. This indicates that the total number of sales is significantly correlated with the number of sales the following week as well as the following month. The plot for PACF shows similar results where lags 1 and 5 are significant, in other words, I believe that an AR(1) or AR(5) model will be appropriate.

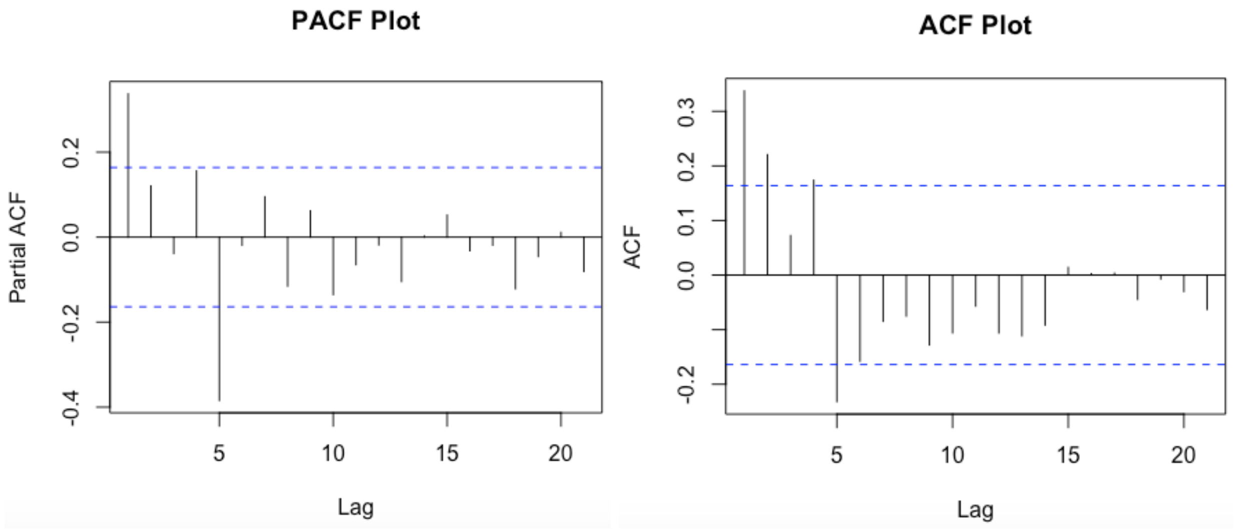
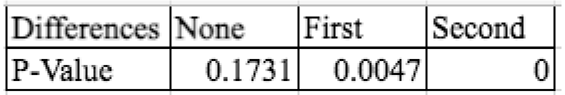


Figure 2: PACF And ACF Plot of The Original Data.

The Augmented Dickey-Fuller (ADF) test of the standardized residuals without differencing gave me a p-value greater than 0.05, therefore we could not reject the null hypothesis. After differencing the time series once produces a p-value of less than .01, which means that we can reject the null hypothesis and conclude that the residuals are stationary at 95% confidence. Differencing the original data twice obtains similar results in the ADF test as in differencing once, for simplicity I will be sticking with the first differencing in my models.



Differences	None	First	Second
P-Value	0.1731	0.0047	0

Figure 3: Dickey-Fuller test results using representative series.

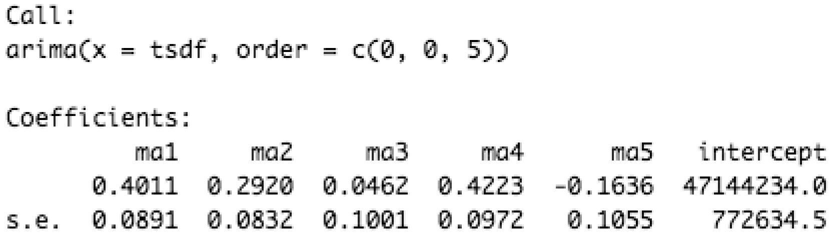
After conducting some exploratory analysis, an ARIMA(0,0,5) model seems to be appropriate for my data. Since seasonality is present in the original data, I will also fit an ARIMA model with a seasonal component and compare it to the ARIMA model without a seasonal component.

Model Fitting

Since the goal of the project is to model the data while also making forecasts, I will split the data into 2 sets. The first one will be my training set which consists of the first 139 observations, and the second will be my testing which will be the last 4 observations, this will serve as a comparison for the forecast model.

Since I have already discussed the model I will be fitting will be the ARIMA(0,0,5) model. I now needed to figure out the appropriate q parameter for the moving average component of the ARIMA. After utilizing the auto arima function without accounting for seasonality I found that an ARIMA(0,0,5) model had the lowest AIC of 4672.62. The corresponding equation for an ARIMA(0,0,5) model would be: $(1-\sum \Phi_i B^i)(1-B)Y_t = (1-\sum \Phi_j B^j)\varepsilon_t$. Where B is the backshift operator, ϕ_i is the ith autoregressive parameter, θ_j is the jth moving average parameter, ε_t is the error at time t, and Y_t is the time series at time t. Using the estimated coefficients from figure 4, the model equation is:

$(1-B)Y_t = (1-0.0411B-0.292B^2-0.0462B^3-0.4223B^4+.1636B^5)\varepsilon_t$

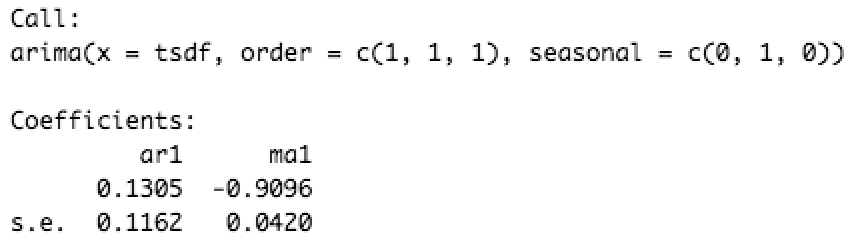


```
Call:
arima(x = tsdf, order = c(0, 0, 5))

Coefficients:
         ma1     ma2     ma3     ma4      ma5   intercept
      0.4011  0.2920  0.0462  0.4223  -0.1636  47144234.0
s.e.  0.0891  0.0832  0.1001  0.0972   0.1055    772634.5
```

Figure 4: Summary of Fitting An ARIMA(0,0,5) Model.

Lastly, the auto.arima function also included a seasonal component of (0,1,0) in order to improve the model in terms of AIC, or variance. My newest model was ARIMA(1,1,1)(1,0,1)[52]. The new model has an AIC of 2732.22 and a smaller variance compared to the regular ARIMA model without a seasonal component. Using the estimated coefficients from figure 5, the model equation is: $(1-0.1305)(1-B)Y_t = (1+0.9096B)\varepsilon_t$



```
Call:
arima(x = tsdf, order = c(1, 1, 1), seasonal = c(0, 1, 0))

Coefficients:
         ar1      ma1
      0.1305  -0.9096
s.e.  0.1162   0.0420
```

Figure 5: Summary of Fitting An ARIMA(1,1,1)(0,1,0) Model.

Model Diagnostics

The standardized residuals of the two models are plotted in Figure 6. Both plots show that the residuals are centered around zero, there is a distinct pattern for ARIMA(0,0,5), and there are a few outliers as well. These outliers are due to the last week of December for both 2011 and 2012.

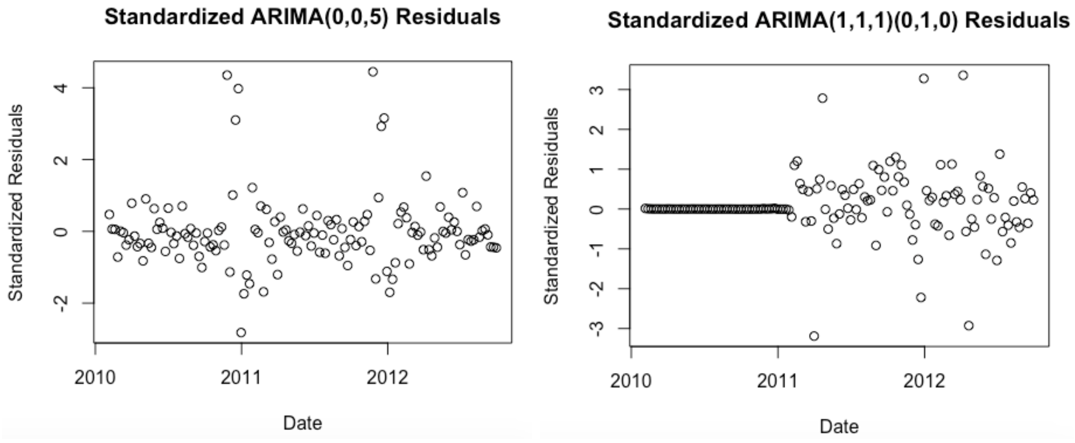


Figure 6: Standardized Residual Models

Now to test whether the distributions of the residuals from both models are normal we go to our QQ plots of the residuals of the models. From the QQ plots, I would say both ARIMA models show that the points do not follow the expected line. Therefore, this indicates that the residuals from both models are not normally distributed.

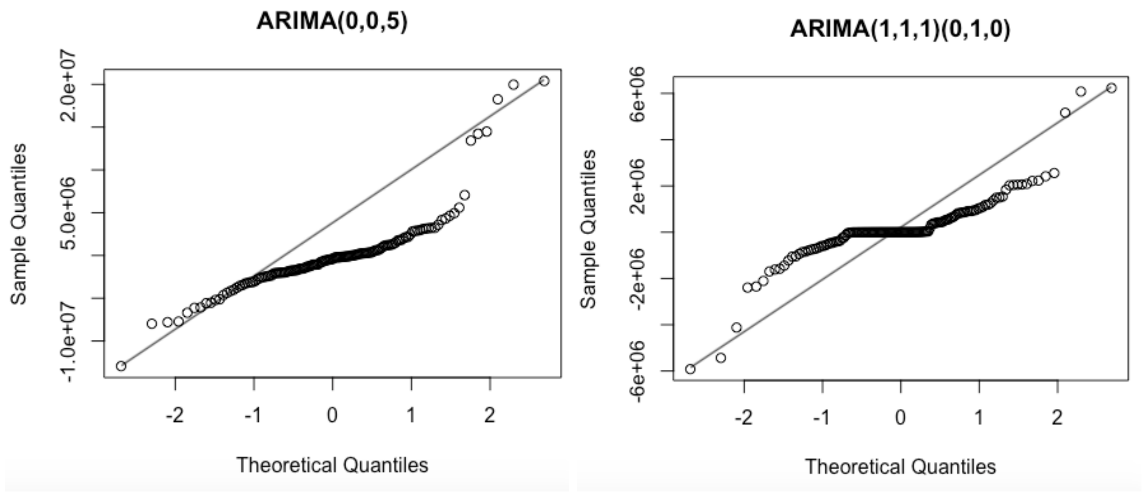


Figure 7: QQ-Plots of The Residuals

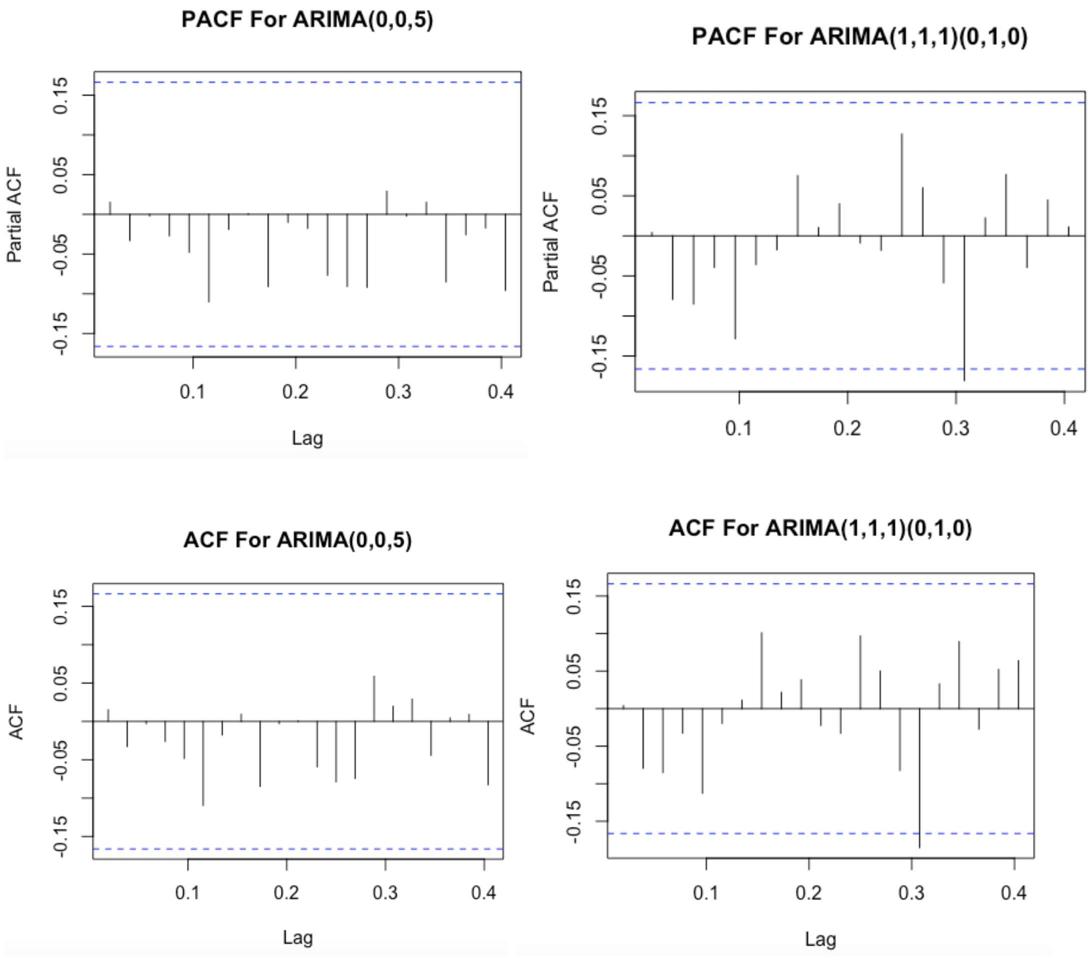


Figure 8: ACF and PACF Plot of The Residuals of Both ARIMA Models

Lastly, we have the ACF and PACF plots of both models in Figure 8. From what I can see most of the ARIMA models do not exhibit significant correlations in lags. There are also fewer numbers of lags in which the correlations are significant at the 95% confidence level in the ARIMA(0,0,5) model than in the seasonal ARIMA model. However, I will disregard the one ‘significant’ lag in ARIMA(1,1,1)(0,1,0) since the value is still < |.20|. The model is now adequately capturing the seasonality present in the time series as opposed to the regular ARIMA model.

Results

Below in Figure 9 are the forecast plots for both models. We can see that the two models capture radically different points along with their 95% prediction interval. We can see that the regular model does not capture the trend where Walmart sales dramatically increase a week before the new year as expected. However, our model that includes seasonality captures this yearly trend perfectly.

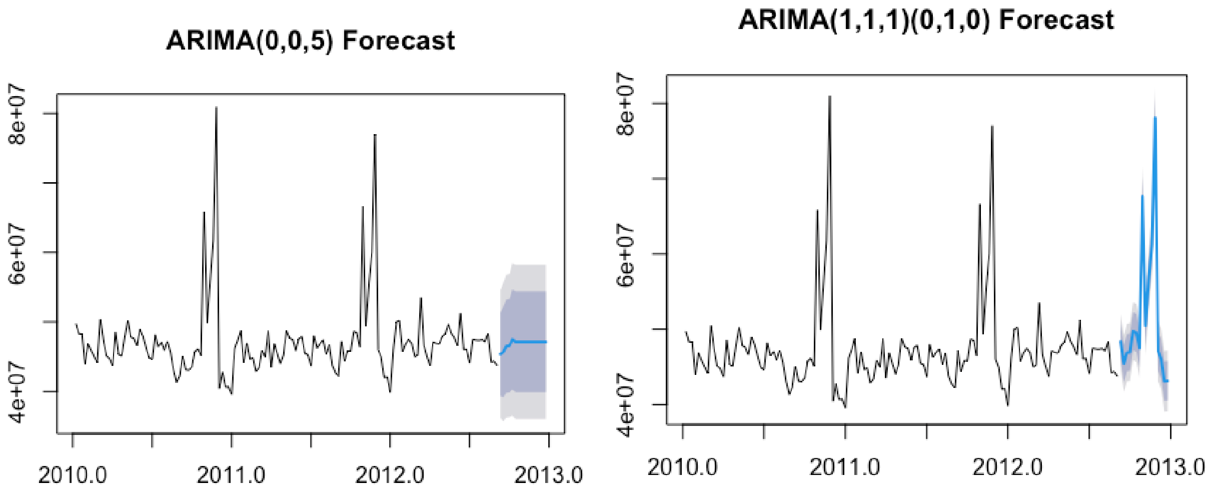
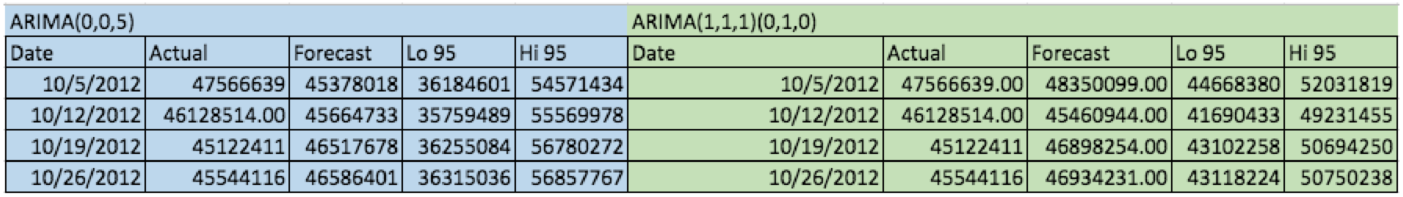


Figure 9: Forecast Plots of Both ARIMA Models

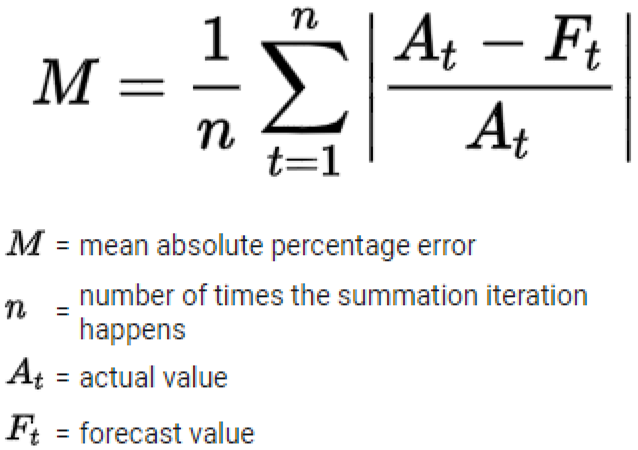
As previously mentioned, I kept the testing set consisting of the last 4 observations/weeks and this will be used to evaluate the accuracy of both models. Using the forecast table in Figure

10 we can calculate the MAPE for each of the models, which can be seen in Figure 11. The MAPE for the last 4 predictions of the regular ARIMA and seasonal ARIMA are 2.75% and 2.52% respectively. Both are fairly low percentage errors, however, as expected the seasonal ARIMA performed better with a lower error rate.



ARIMA(0,0,5)					ARIMA(1,1,1)(0,1,0)				
Date	Actual	Forecast	Lo 95	Hi 95	Date	Actual	Forecast	Lo 95	Hi 95
10/5/2012	47566639	45378018	36184601	54571434	10/5/2012	47566639.00	48350099.00	44668380	52031819
10/12/2012	46128514.00	45664733	35759489	55569978	10/12/2012	46128514.00	45460944.00	41690433	49231455
10/19/2012	45122411	46517678	36255084	56780272	10/19/2012	45122411	46898254.00	43102258	50694250
10/26/2012	45544116	46586401	36315036	56857767	10/26/2012	45544116	46934231.00	43118224	50750238

Figure 10: Forecast Table of Both Models With 95% Prediction Intervals



$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error

n = number of times the summation iteration happens

A_t = actual value

F_t = forecast value

Figure 11: MAPE Formula

Conclusion

Ultimately, I believe that my seasonal ARIMA model is better than the regular ARIMA model. Both models perform similarly when viewing the MAPE metric, and the residual model diagnostics. However, the seasonal ARIMA model has a significantly smaller AIC value, a lower variance, and produced a more accurate forecast plot. Moreover, the seasonal ARIMA model contains fewer parameters compared to the regular ARIMA model which allows for straightforward interpretation. Furthermore, in Figure 12 we can see that the forecast of the seasonal model is much closer to the true weekly Walmart sales compared to the non-seasonal one. Finally, we can conclude that the seasonal model is much more accurate than the non seasonal model.

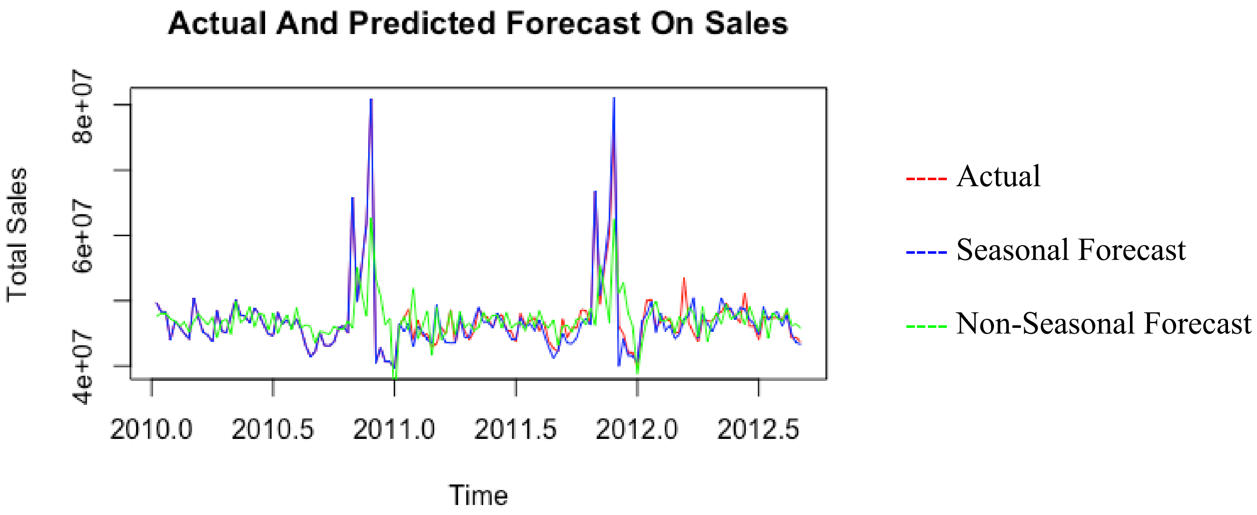


Figure 12: Actual And Predicted Forecast On Sales

However, there are still many limitations to my project. The primary limitation in my time series analysis was the size of the dataset. The entire data set consists of only 143 weeks (2.75 years) of sales data, which prohibited modeling with a larger testing set. What tends to happen with a short time series is that while using AIC as my criterion it will suggest very simple models, because models with more than two parameters will produce poor forecasts due to the estimation error. Another limitation comes from missing sales data for a few departments in Walmart. Lastly, since 'Weekly_Sales' was the only variable used in modeling, this could limit forecasting potential. Other variables such as weather and unemployment data, might have improved the accuracy of forecasts.

References

“4.1 Seasonal ARIMA Models: STAT 510.” PennState: Statistics Online Courses, 2020, online.stat.psu.edu/stat510/lesson/4/4.1.

Aditya. “Walmart Sales Forecasting.” *Medium*, Analytics Vidhya, 27 Oct. 2019, medium.com/analytics-vidhya/walmart-sales-forecasting-d6bd537e4904.

Cryer, Jonathan D., and Kung-sik Chan. Time Series Analysis With Applications in R. Springer, 2008.

“Walmart Recruiting - Store Sales Forecasting.” *Kaggle*, www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data.

Wenxiang Dong, Qingming Li, and H. Vicky Zhao. “Statistical and Machine Learning-based E-commerce Sales Forecasting,” 2019. In Proceedings of the 4th International Conference on Crowd Science and Engineering (ICCSE 19). Association for Computing Machinery, New York, NY, USA, 110–117. DOI:https://doi.org/10.1145/3371238.3371256