Homework 3: Walmart Sales Forecasting

Group Members:

Devangi Bohra, Jasmine Brown, Zhe Chen, Kunal Kalra, Hard Thakkar, Charles Warner

Introduction

Time series analysis accounts for the fact that data points taken over time may have an internal structure that should be accounted for [1]. The purpose of this report is to provide an in-depth time series analysis of Walmart Sales. The outline of the analysis includes data exploration, a comparative analysis on three models, and evaluated based on the mean squared prediction error (MSPE). The three models evaluated and compared will be: SARIMA with no additional variables, SARIMA with additional variables, and Prophet.

Data and Methodologies

The data used for this report was the Walmart Recruiting - Store Sales Forecasting [2]. The data set contained 421570 observations and five columns. Since there were multiple stores and departments present in the data set, we took the subset the data to focus on the analysis of store one, department one (143 rows).

First, we explored our data by plotting the distribution of the sales to determine the shape, center, and spread. By analyzing the distribution, the need of any transformation to achieve normality was sensed. Correlations between our response and explanatory variables were also produced for overfitting and MSPE reduction purposes.

To model our data with SARIMA, we generated raw exploratory plots to determine trend, seasonality, and variability. These raw plots gave us a deeper insight on whether differencing was needed for our time series. Model fitting using the ACF and PACF plot provided visualizations of lags and significant changepoints. Model diagnosis with different parameters were implemented to analyze p values, correlations, residuals, and lag cut offs to detect the best model.

The second model that was implemented was a SARIMA with additional variables. Along with the initial data set with the sales and dates columns, another data set containing features was given [2]. This dataset was used to incorporate explanatory variables such as unemployment, gasoline prices, and CPI in the initial time series. To begin the analysis, correlation plots were produced to gain insight on the relationship between our response variable (sales) and our explanatory variables. Model exploration of seasonality, trend, and variance was also visualized to determine the existence of stationarity and the need of differencing. Next, the ACF, PACF and the CCF plots were analyzed to fit the model. Finally, significant lags, residuals, and cut offs were visualized to provide diagnostics.

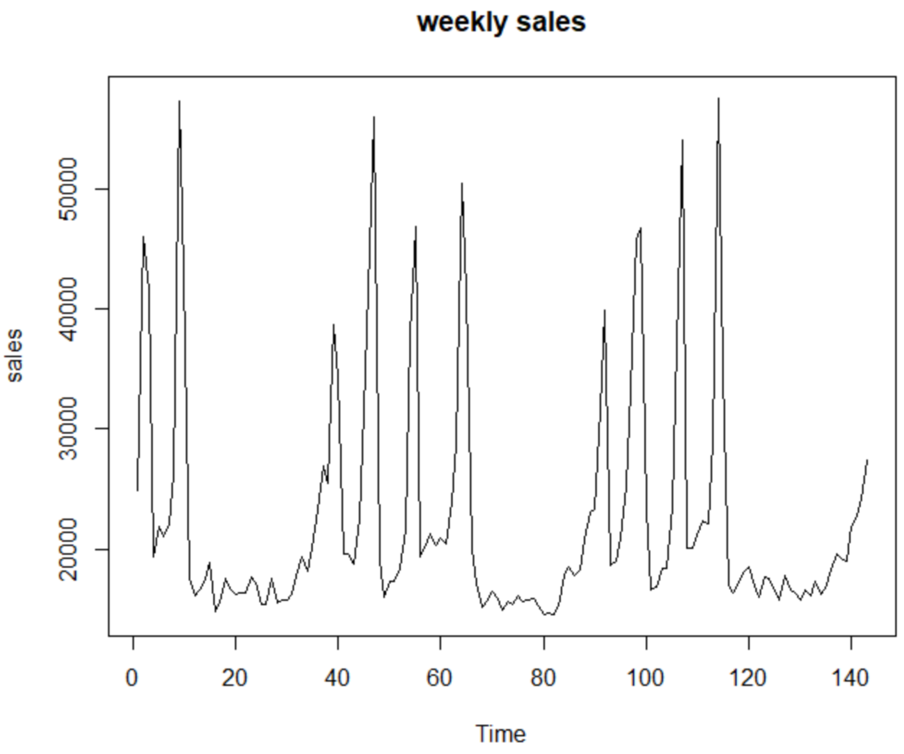
The third model used was Prophet. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects [3]. The advantages of a prophet model is differencing is not needed as the model internally accounts for lag in time. Prophet is also known to be more robust and resistant to outliers. In addition, Prophet models helps to identify significant change points. To model our data using Prophet, we experimented with different number of changepoints. We then analyzed the ACF plot to count the number of changepoints that would be ideal to incorporate in the model. Since the Prophet model has multiple parameters, parameters with different values were tested and evaluated.

Each of the three models were trained with 80% of the data, while the 20% were tested on. Models were evaluated using the MSPE to conclude which model had the lowest prediction error and possible reasons why. Furthermore, models were used to forecast the next 14 weeks.

Data Exploration

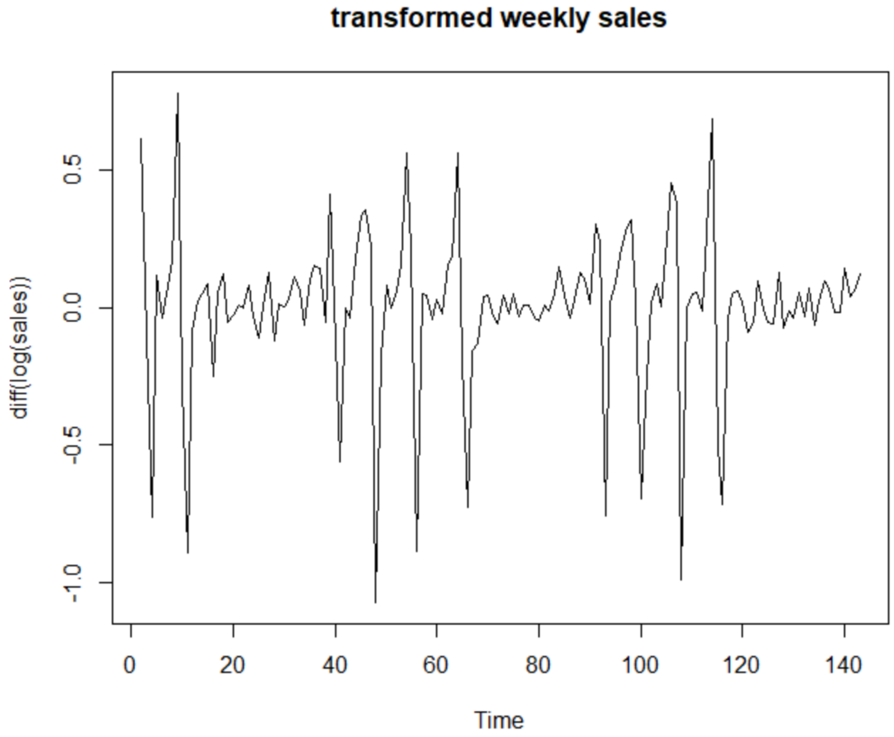
Our data consisted of observations from multiple stores and departments. For the purpose of our analysis, we chose to subset the data to analyze the Walmart sales of store 1, department 1. The initial time series plot in Figure 1 shows no clear trend with the presence of seasonality.

Figure 1



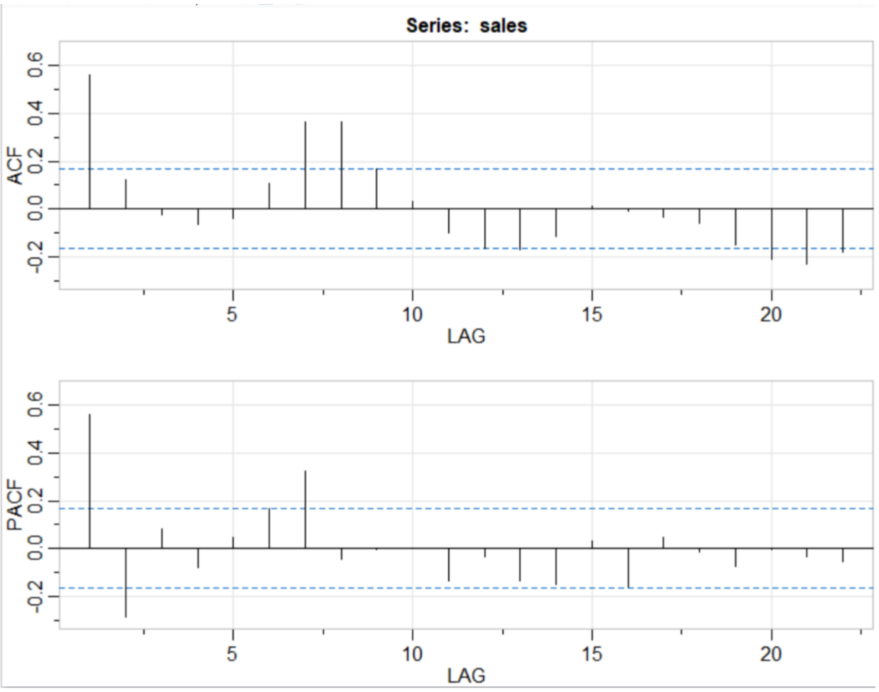
To further detrend and apply stationarity to the data, we applied a logarithmic transformation and differenced our data. We can see that the transformed time series has a reduced variance and increased stationarity than that of the time series with no differencing and transformation in Figure 2.

Figure 2



Model fitting using the ACF and PACF plot provided visualizations of lags and significant changepoints. We can see in Figure 3, the ACF and PACF went above the line at lag 7 and lag 21. This suggested the use of an AR(1) model with a seasonal period set at 7. To account for non seasonality, we see the ACF cuts off at lag 7 while the PACF tails off which supports the use of a MA(1). Furthermore, we attempt to model AR(2) due to the cut off at lag 14 and AR(2,1) due to the ACF and PACF model tailing off at the end.

Figure 3



Model 1: SARIMA without additional variables

Model diagnosis with different parameters were implemented to analyze p values, correlations, residuals, and lag cut offs to detect the best model. We attempt different SARIMA models as shown in Figure 4. For Model A, we can see that the p values go below the line after lag 6, which means the residuals are correlated while in Model B all the p values are above the line, which means the residuals are independent. In Model C we see the p values go below the line at lag 14 and lag 15, which means the residuals are correlated at lag 14 and lag 15. In conclusion, among the three models, ARMA(2,0)*(1,0) is the most appropriate model for our predictions.

Figure 4 A: SARIMA(0,0,1,1,0,0,7)

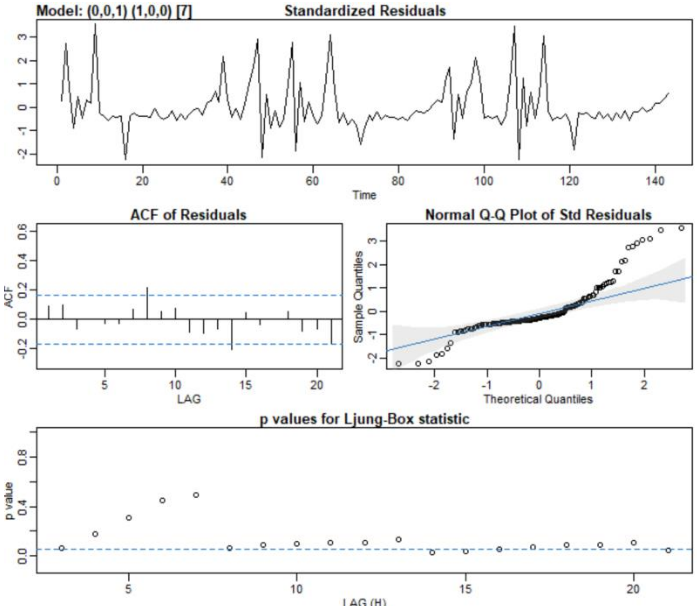


Figure 4 B: SARIMA(2,0,0,1,0,0,7)

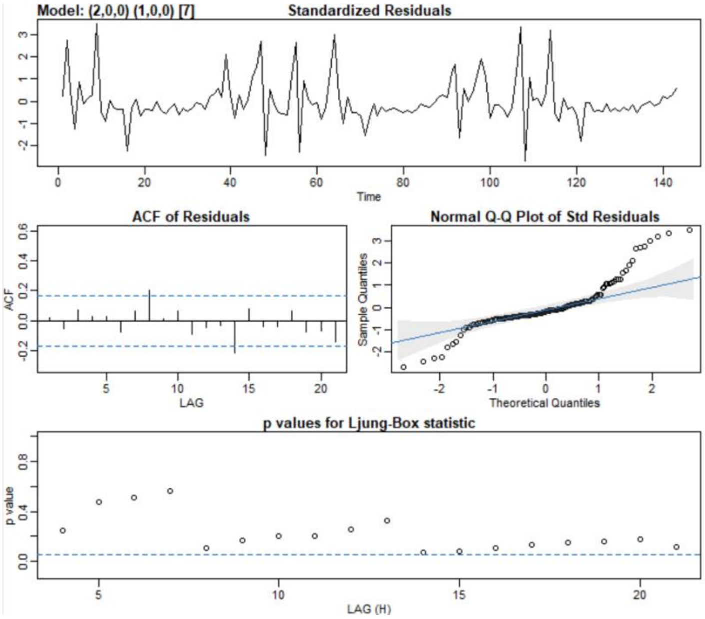
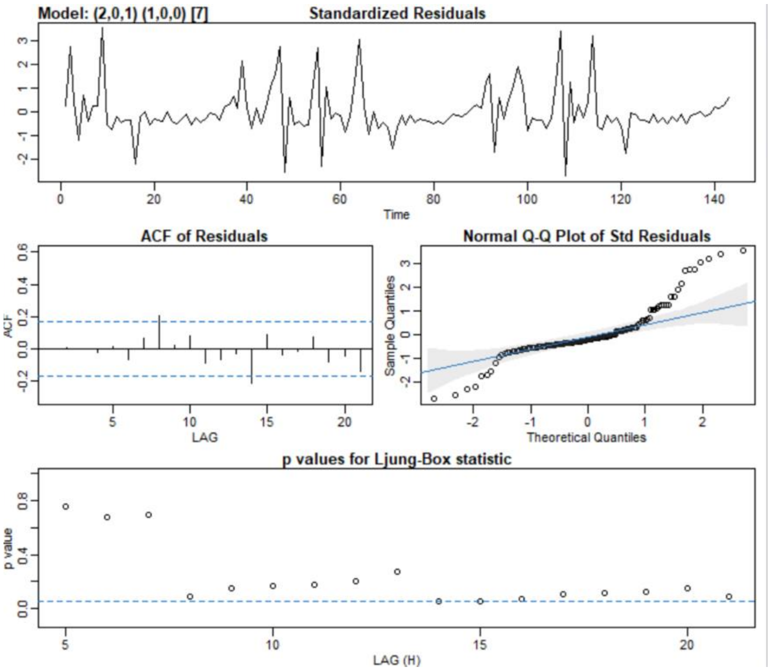
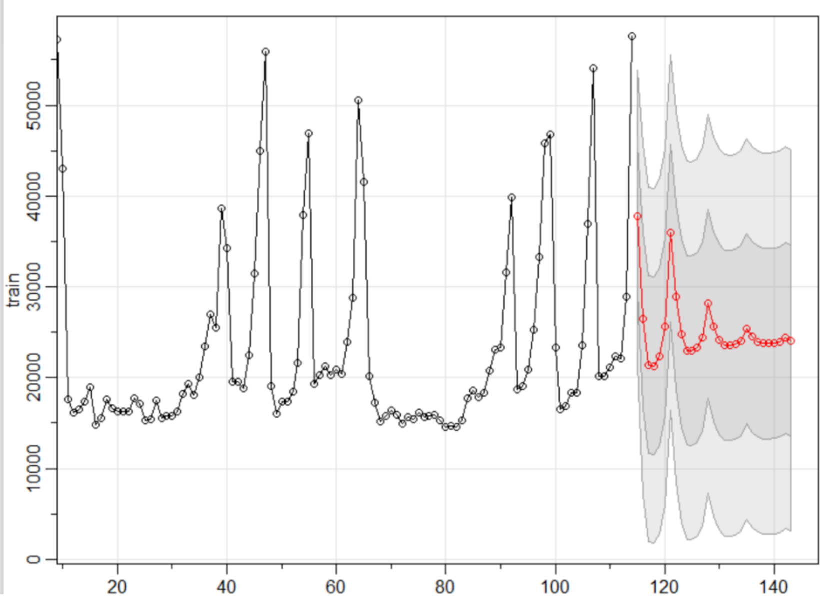


Figure 4 C: SARIMA(2,0,1,1,0,0,7)



The model was trained with 80% (114 observations) of the data and tested with 20% (29 observations). Figure 5 shows the prediction of the SARIMA model chosen. The mean square prediction error (MSPE) was used to evaluate the model predictions. The MSPE was calculated as 56753140.

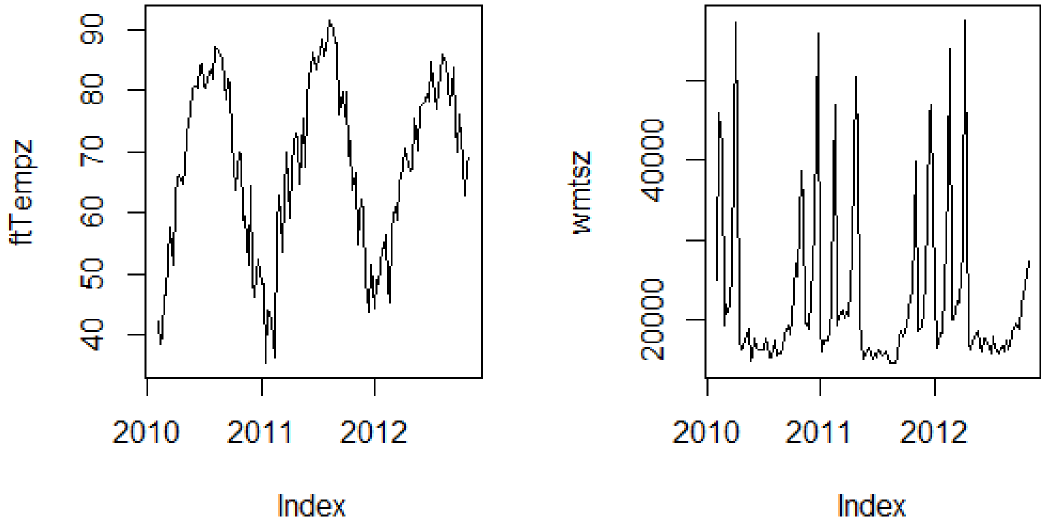
Figure 5



Model 2: SARIMA with additional variables

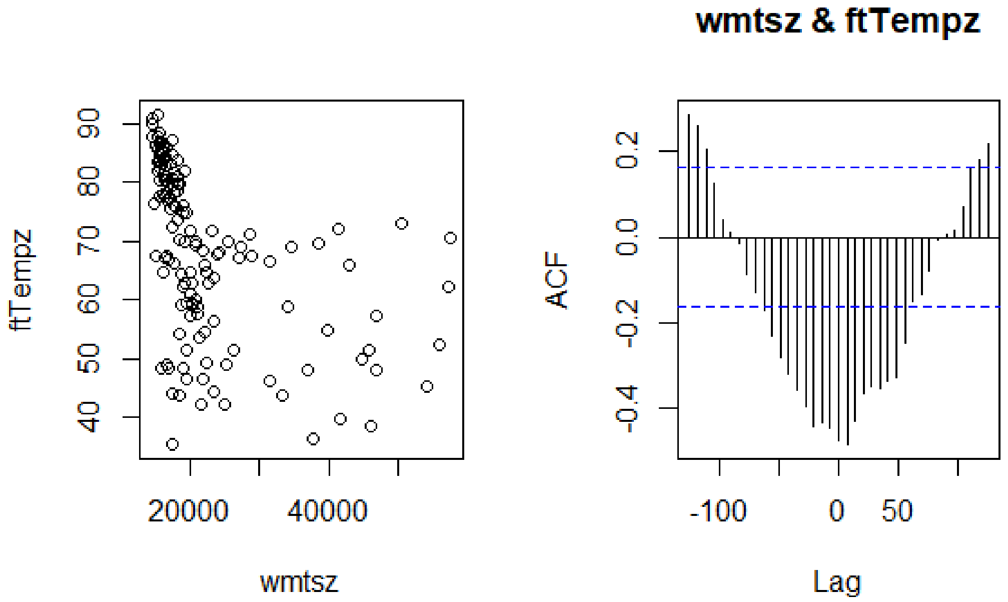
The second model that was implemented was a SARIMA with additional variables. This dataset was used to incorporate explanatory variables such as unemployment, gasoline prices, and CPI in the initial time series. To begin the analysis, we analyzed the distribution of the temperature (left) and sales (right). We can see in Figure 6 that the trend of temperature over time has a slight increase but with a slightly different variance, which supports no need for logistic transformation or differencing. The trend for sales, as analyzed before, showed no distinct trend with seasonality.

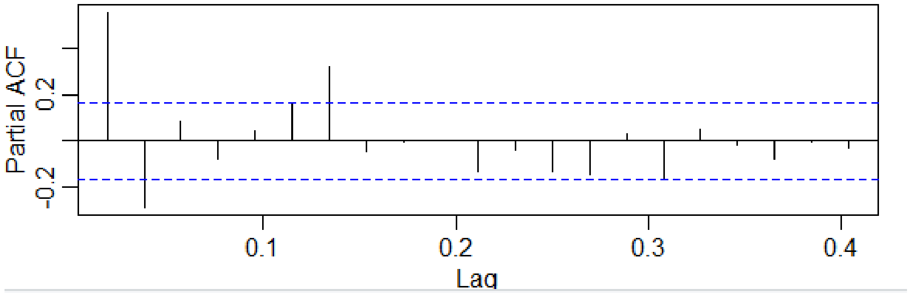
Figure 6



We plotted the correlation plot to visualize the relationship between temperature and sales. Additionally, in Figure 7 the ACF plot (right) showed the correlation function when both these variables are combined. We can see from the correlation plot (left) that there is no clear relationship between the two variables. The correlation implies a weak to moderate, negative relationship. When we analyzed the ACF plot, we saw a seasonality and a high negative correlation at lag 0. Since there was a cut off around the eighth lag, a model with an AR parameter of seven or eight seemed like a potential fit. In addition, since the PACF plot in Figure 8 shows a cut of at 2, we used a MA parameter of 2 for our model.

Figure 7 (top), Figure 8 (bottom)





Model diagnosis with different parameters were implemented to analyze p values, correlations, residuals, and lag cut offs to detect the best model. The Model that gave us the best results were consistent with the hypothesized model discussed in the previous paragraph. In Figure 9, we can see linearity in the residuals of the model as well as significant lags apart from lag 15. In conclusion, SARIMA(8,2,4) when incorporating temperature and unemployment was the most appropriate model for our predictions.

Figure 9

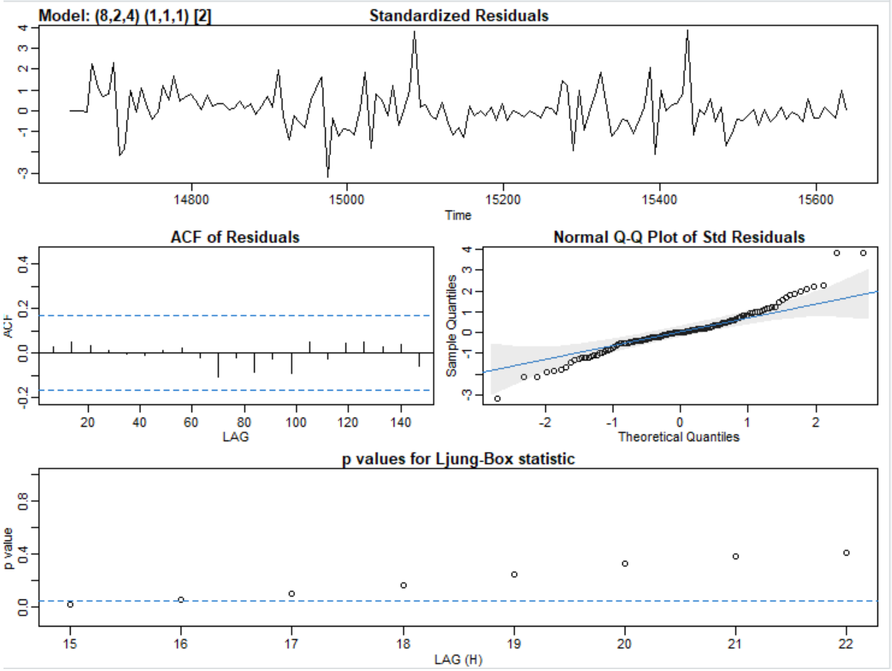
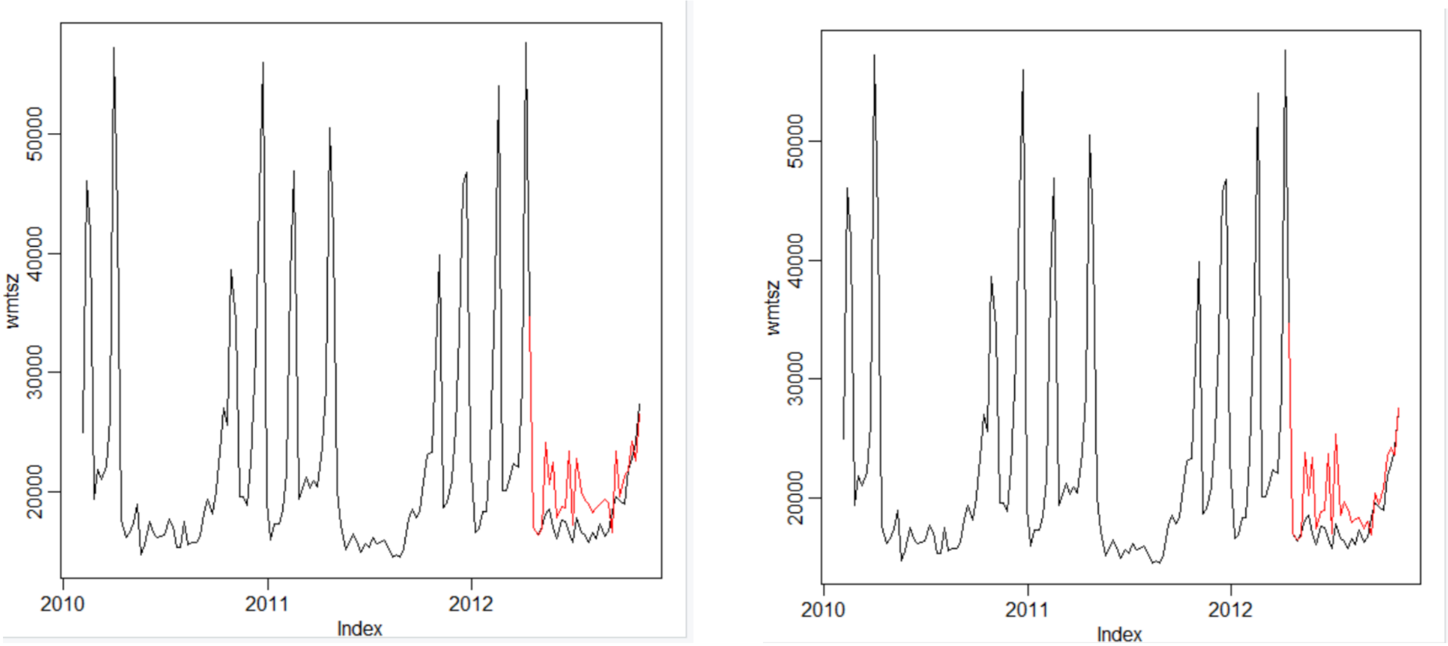


Figure 10 shows the components of the SARIMA model when the prediction is applied. The left plot shows the prediction when temperature is incorporated while the plot to the right shows the predictions when unemployment is incorporated. We can see minute differences in the predictions between both plots. We saw that the predictions were closer when incorporating temperature. The MSPE for the model with temperature (left) was 8176171 while the MSPE for the model with unemployment (right) was 8208238.

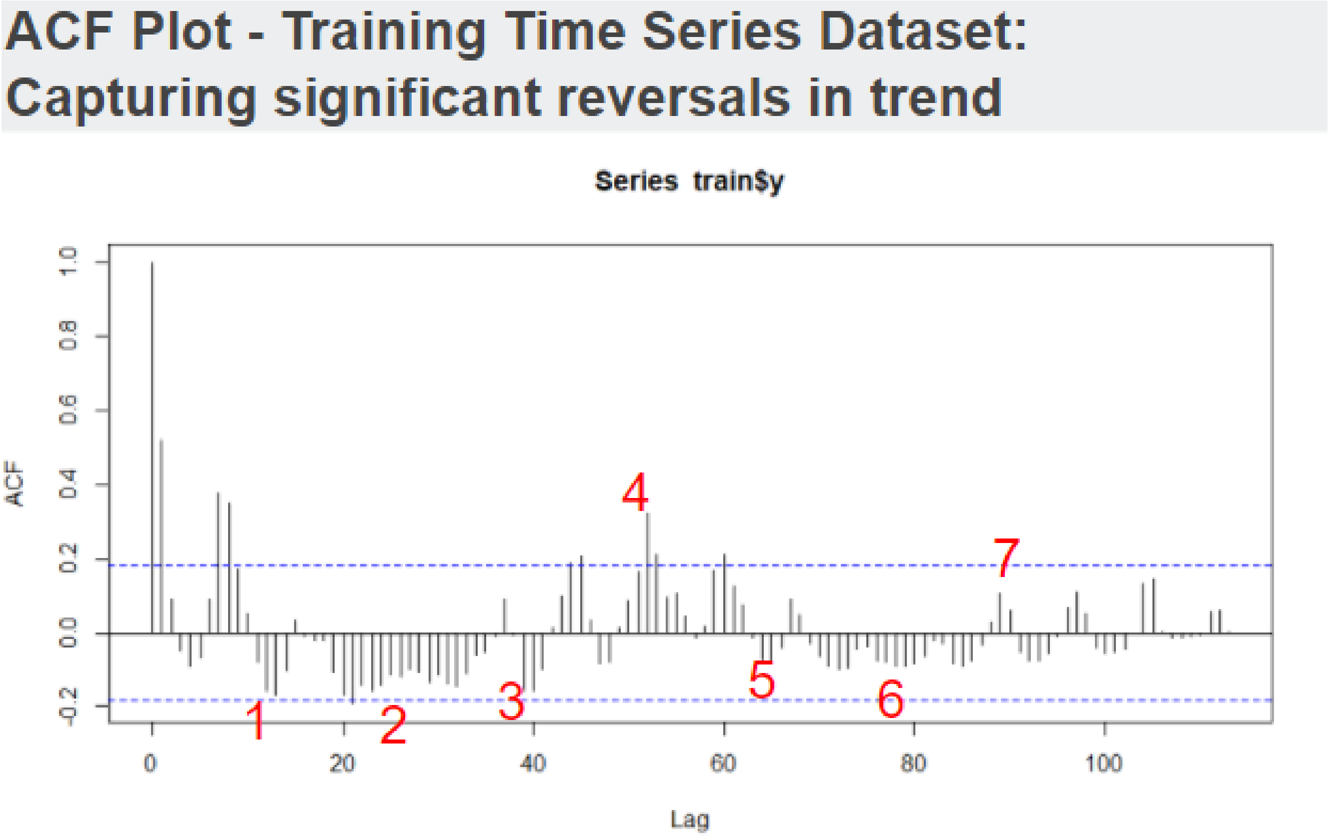
Figure 10



Model 3: Prophet

The third model used was the Prophet model. When implementing the prophet model, differencing was not needed. This is because the Prophet model itself accounted for lags and gaps in time. In addition, the Prophet model helps to identify and significant change points in the time series. These changepoints can either be manually configured or automatically detected. With some additional reading, it was found that manually defining the number of changing points helped the accuracy of the model. Letting the model detect too many change points would increase the flexibility of the model and lead to potential overfitting. To estimate the number of significant change points, we analyzed our ACF and detected the number of times the bars switch from positive to negative as shown in Figure 11.

Figure 11

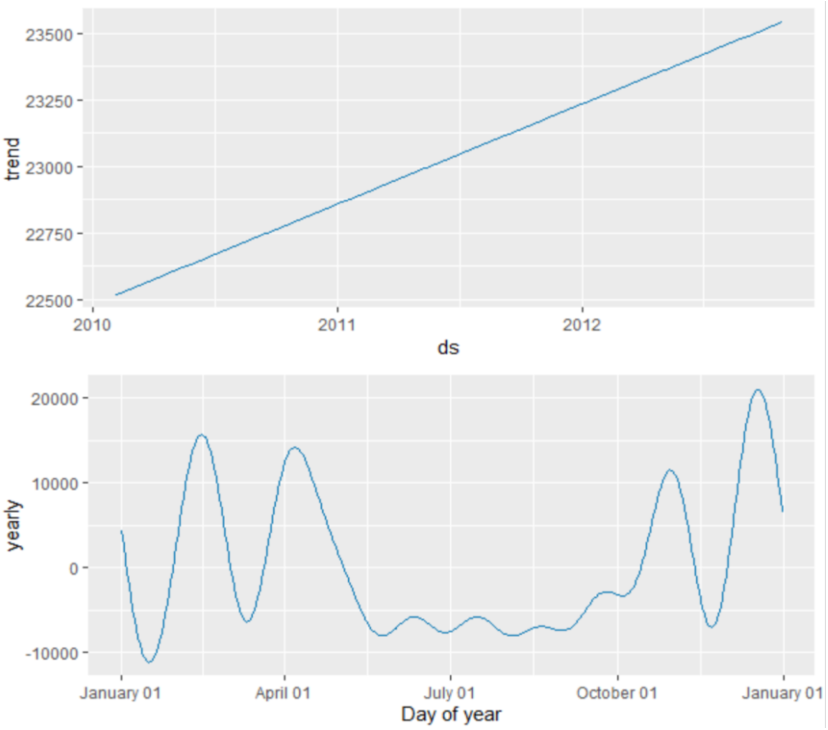


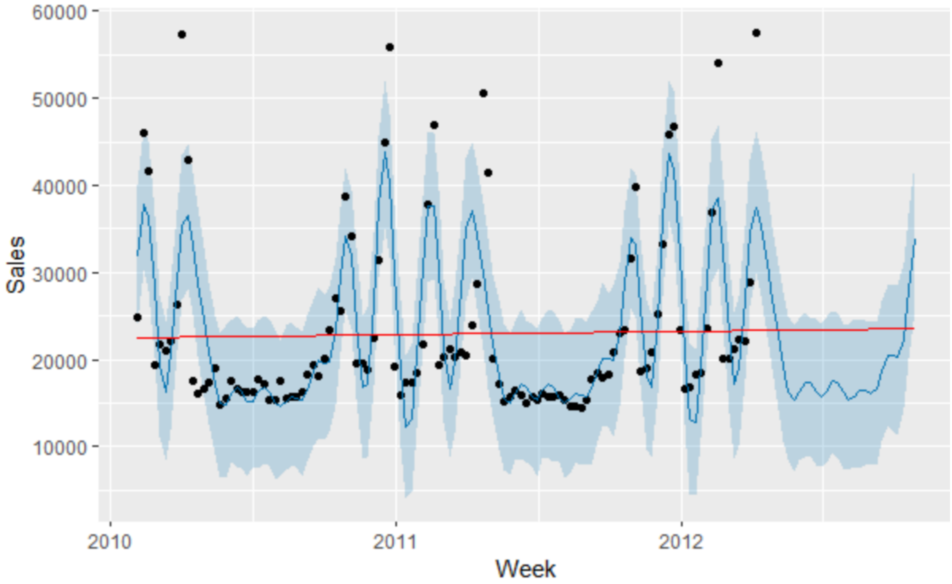
ACF Plot - Training Time Series Dataset: Capturing significant reversals in trend

Once we created out prophet model, we were able to extract the exact dates of the changepoints: "2010-05-07 GMT" ,"2010-08-06 GMT" ,"2010-11-05 GMT", "2011-01-28 GMT" ,"2011-04-29 GMT", "2011-07-29 GMT", "2011-10-28 GMT".

Figure12 shows the components of the Prophet model when the prediction is applied. We can see a clear linear trend. The yearly residual trends also show no definite pattern which confirms the linearity of the time series. Figure 13 shows the plot of the predictions of sales for the last 29 weeks (20% tested data). We can see that the predictions are within the confidence bands. We also see these points consistent with the trend we initially saw when exploring the data. The MSPE for the Prophet model was 14669985.

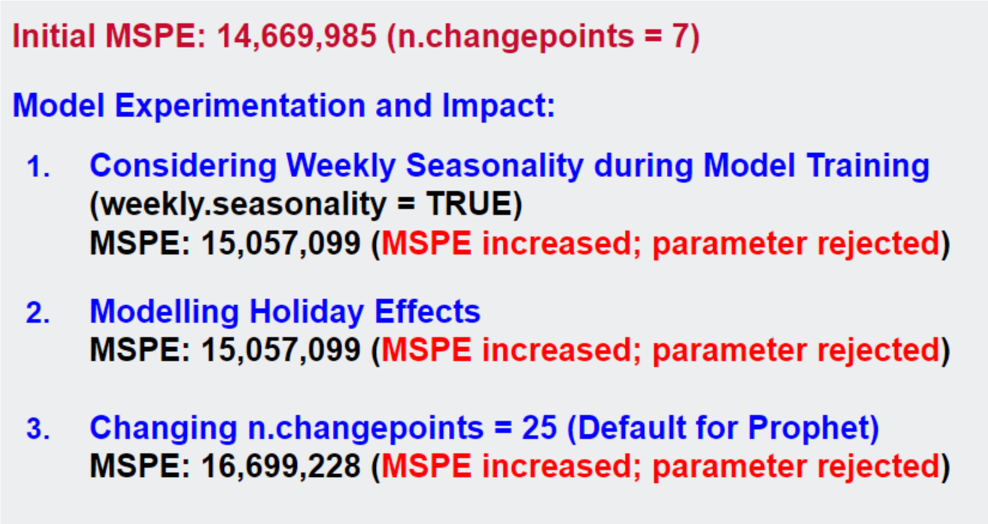
Figure 12 (top), Figure 13(bottom)





Additionally, we experimented with different parameters for the Prophet model to see how the MSPE changes. As seen in Figure 14, some of the parameters we changed were considering weekly seasonality, modeling holiday effects, and configuring the changepoints as discussed above. We saw that when experimenting with these parameter values, the MSPE increased for all parameters which led us to reject the parameters.

Figure 14



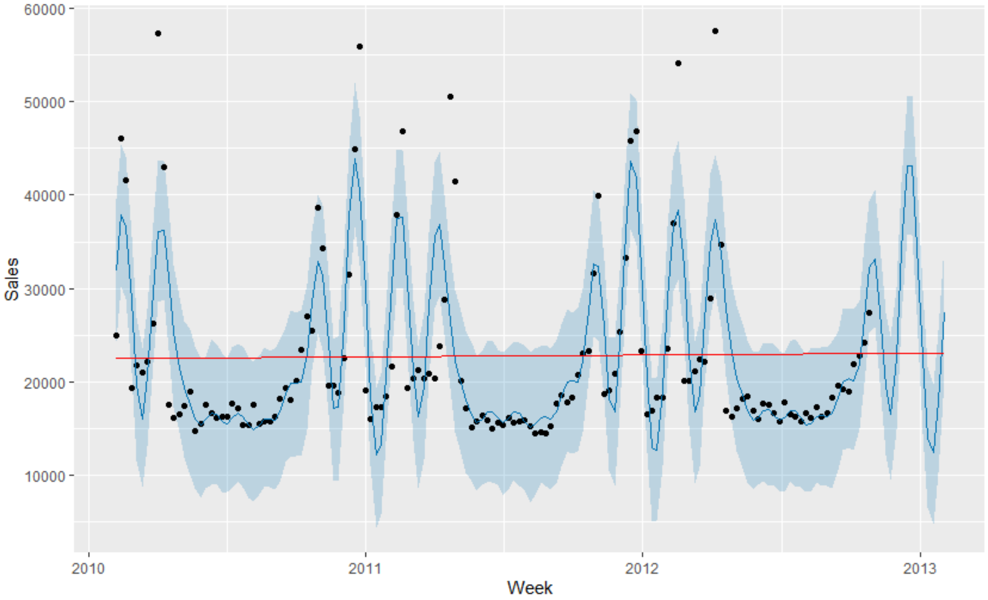
Initial MSPE: 14,669,985 (n.changepoints = 7)

Model Experimentation and Impact:

1. **Considering Weekly Seasonality during Model Training**
   (weekly.seasonality = TRUE)
   MSPE: 15,057,099 (MSPE increased; parameter rejected)
2. **Modelling Holiday Effects**
   MSPE: 15,057,099 (MSPE increased; parameter rejected)
3. **Changing n.changepoints = 25 (Default for Prophet)**
   MSPE: 16,699,228 (MSPE increased; parameter rejected)

We also used the Prophet model to forecast an additional 14 weeks. As shown in Figure 15, we used the entire training set to train the prophet model keeping the number of changepoints as 7.

Figure 15



Conclusion

To conclude, we examined the data using raw exploratory analysis. The exploratory analysis gives deeper insight on differencing, visualization of lag, and significant change points of the time series. The last step was to implement a diagnosis of the model using the p value, residuals, and correlations to determine the best model.

SARIMA model without additional variables, with additional variables, and Prophet were implemented as our three models. While all models had their advantages of use, they were trained with 80% of the data, while 20% were tested on. Evaluating all models was based on the MSPE determined that a SARIMA model with temperature had the lowest prediction error. We hypothesize this model was the lowest because temperature is also a seasonal variable like our initial time series.

Model	MSPE (mean squared prediction error)
SARIMA with no additional variables	56,753,140
SARIMA with Temperature	8,176,171
SARIME with unemployment	8,208,238
Prophet	14,669, 985

Appendix

[1] 6.4. Introduction to Time Series Analysis (nist.gov)

[2] Walmart Recruiting - Store Sales Forecasting | Kaggle

[3] Prophet | Forecasting at scale. (facebook.github.io)