Forecasting Walmart Store Sales

12.12.2022

—

MSDS 596 - Regression & Time Series Analysis

Group D

Shourya Thatha Ravi (st1037)

Rahul Rajesh Singh (rs2050)

Raviteja Arava (ra980)

Nish Patel (nsp124)

Introduction

Walmart held a recruiting competition on Kaggle to forecast weekly sales at select Walmart locations. As a part of this competition, they provided a dataset that includes historical sales data from 45 Walmart stores between 2010 and 2012. The goal is to use this data to forecast weekly sales by store.

This problem is important because by predicting weekly sales, stores like Walmart can manage their supply chain more efficiently, optimize inventory levels, accommodate higher volume of products, handle traffic, and provide markdowns accordingly. Forecasting sales is also a common task for data scientists so it can provide a good opportunity for the group to learn techniques and experiment with real-world data.

The dataset included three files that we used:

stores.csv - This file contains anonymized information about the 45 stores

- Store - the store number
- Type - the store type (A, B, C)
- Size - the size of the store

train.csv - This file contains the historical sales information by store and department

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week

features.csv - This file contains information on additional factors that may impact weekly sales (such as promotional markdowns).

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

Exploratory Data Analysis (EDA)

Exploratory data analysis was completed on the dataset for the team to understand the data, clean the data, and check for multicollinearity.

The data was first plotted by year (as shown in Figure 1) to understand overall patterns in the data. A clear seasonality component can be seen from this graph with large spikes in sales near Thanksgiving and Christmas of each year.

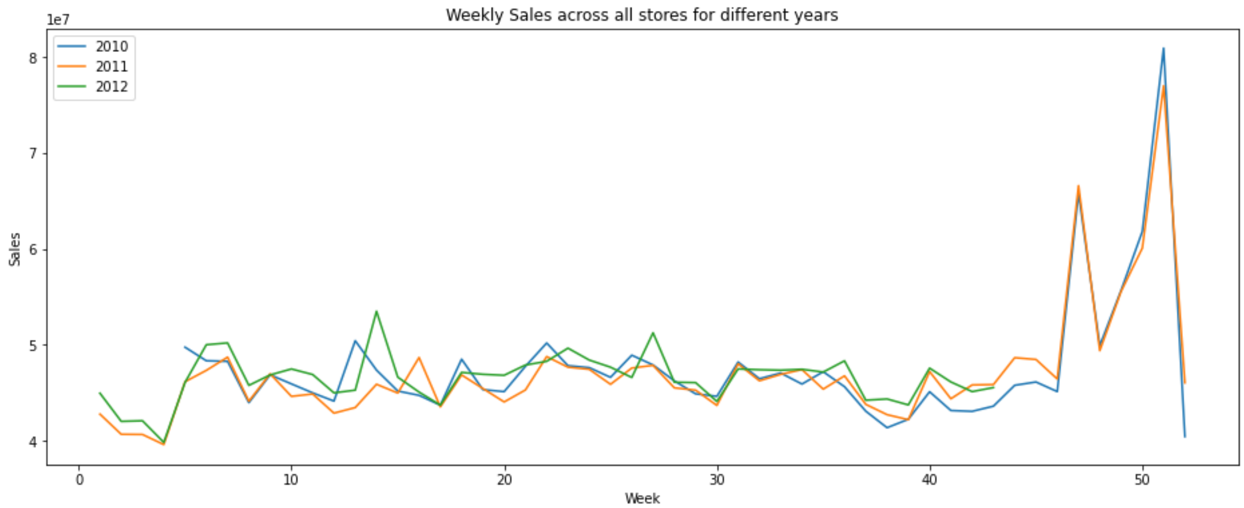


Figure 1. Weekly Sales by Year

The dataset contained some null values in features.csv so those values were replaced with 0 to prevent any issues in modeling.

Multicollinearity was checked by plotting a correlation heat map and pairwise plots. As seen in Figures 2 and 3, Markdown1, Markdown4, and Markdown5 show multicollinearity.

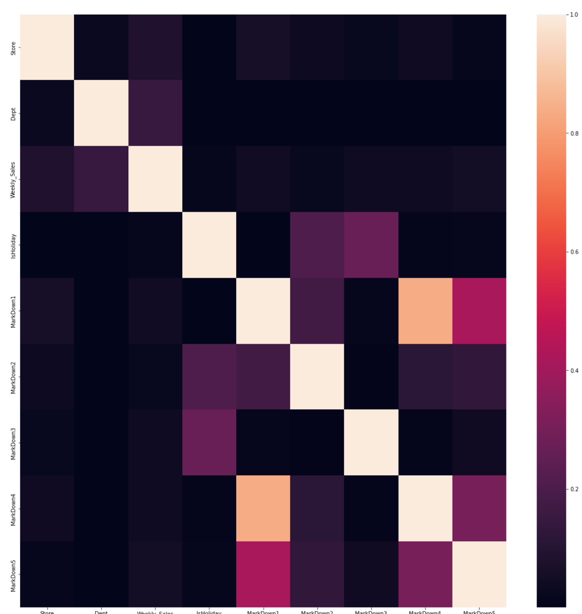


Figure 2. Correlation Plot

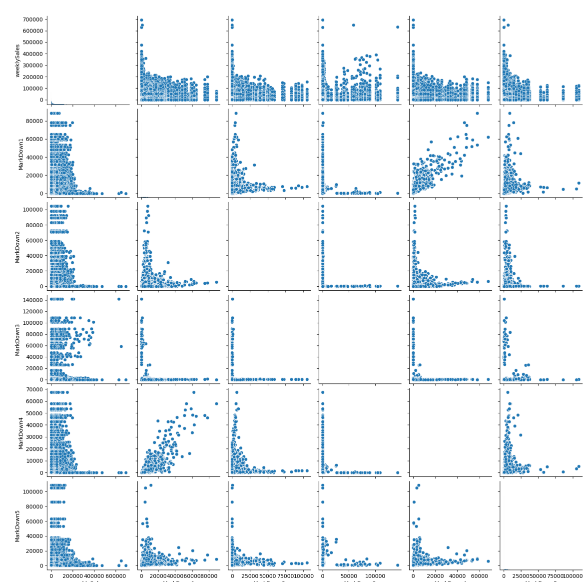


Figure 3. Pairwise Plots

To solve this, these three columns were summed and combined into a single column (Markdown_145). The correlation plot after making this change is seen in Figure 4 with most of the multicollinearity issues resolved.

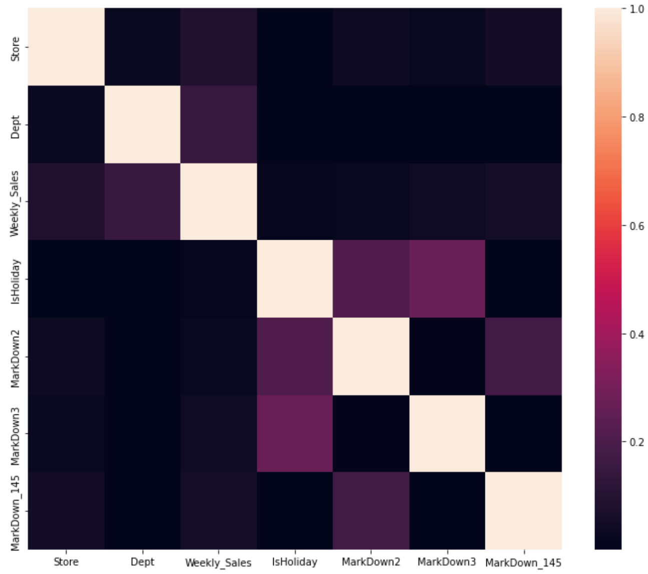


Figure 4. Updated Correlation Plot with Markdown_145 Column

The remaining issue is with the isHoliday column which is correlated with Markdown2 and Markdown3. After further investigation, Markdown2 and Markdown3 are almost always active during a holiday week (as seen in Figures 5 and 6) so to resolve this multicollinearity issue, the isHoliday column was dropped.

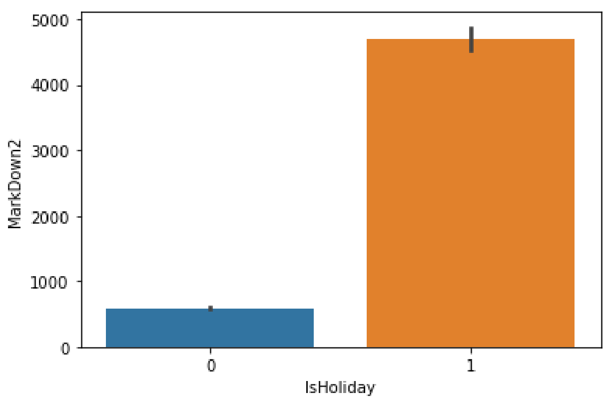


Figure 5. isHoliday vs Markdown2

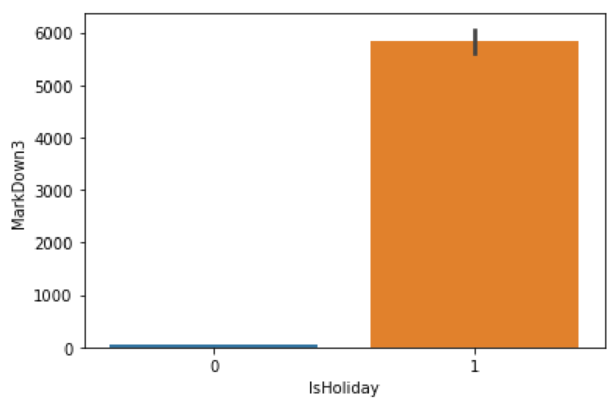


Figure 6. isHoliday vs Markdown3

Modeling Methodology

Prior to starting the modeling, the team decided on a methodology and an evaluation metric by which to evaluate different models:

- Given the findings from the EDA, the team decided that univariate and multivariate time series models would be most likely to produce the best predictions so those were chosen as potential modeling types.
- To try more types of models within the timeframe of the project, the team decided to focus on forecasting weekly sales for the top 2 stores (shown in Figure 7 with the highest weekly sales (stores 20 and 4).

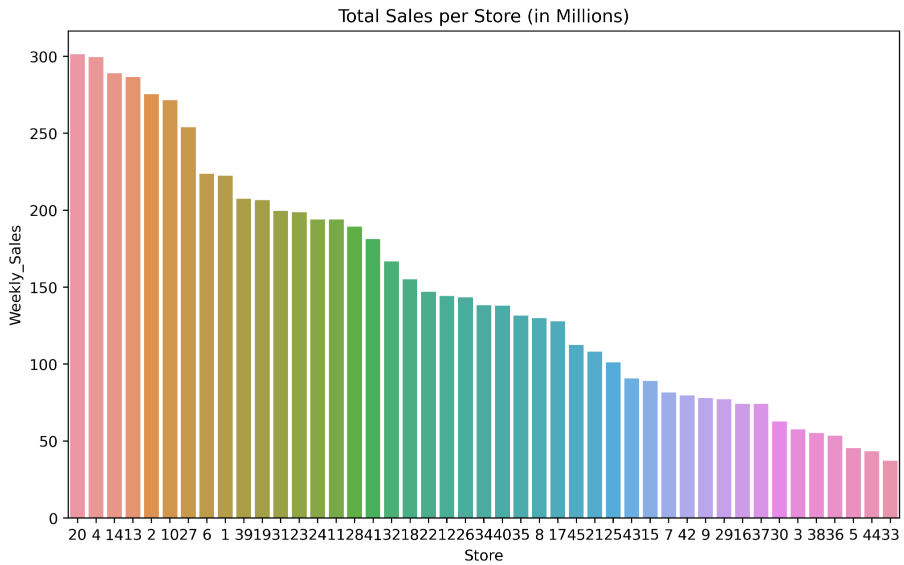


Figure 7. Weekly Sales by Store

- In order to have a robust training set, 75% of the dataset was used as the training set and 25% was saved for the test set.
- Mean Absolute Percentage Error (MAPE) was chosen as the evaluation metric because it is expressed as a percentage which makes it scale-independent which is useful because the magnitude of weekly_sales varies across time and across stores.
  - The formula for MAPE is shown below in Figure 8 where A_t is the actual value, F_t is the forecasted value, and n is the sample size.

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

Figure 8. MAPE formula

Univariate Modeling

Prior to deciding which model to start with, seasonal decomposition, stationarity, and autocorrelation were checked to ensure an appropriate starting model was selected.

The seasonal decomposition graphs (as shown in Figure 9) are helpful to understand what trend and seasonality are present in the data. An upward trend is observed in the second graph and as noted above, seasonality is observed around Thanksgiving and Christmas.

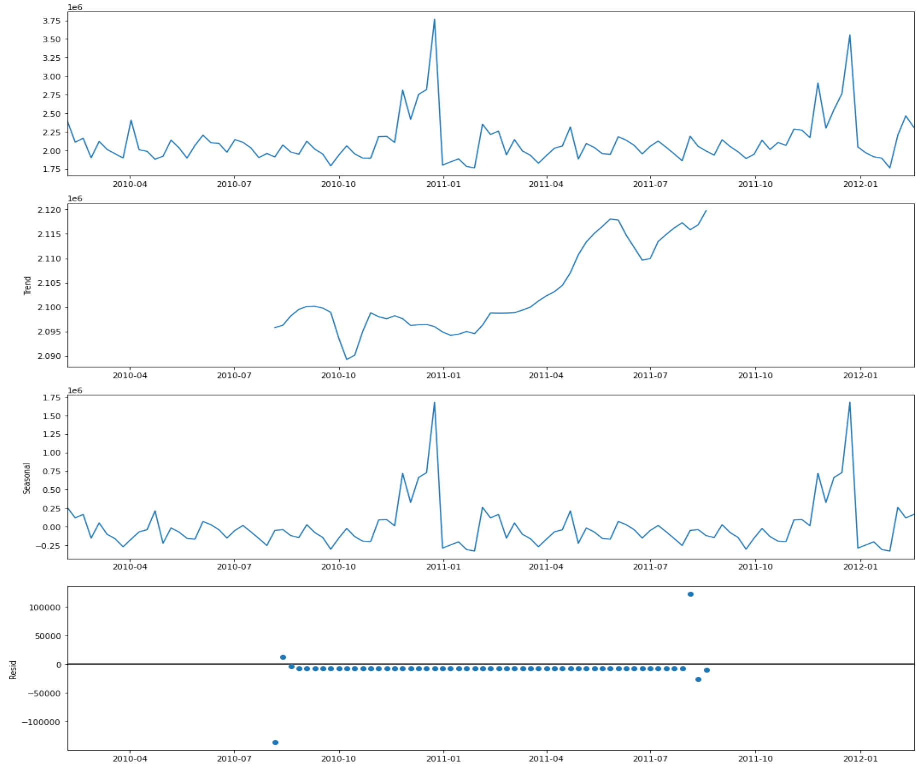


Figure 9. Seasonal Decomposition Plots

Stationarity of Weekly_Sales was checked using the Augmented Dicky-Fuller (ADF) test. The ADF test has a null hypothesis that the data is non-stationary and an alternative hypothesis that the data is stationary. For both stores 4 and 20, the p-value of the ADF test (shown in Figure 10) is below 0.05 and the Test Statistic is less than the critical value at 5% indicating that we have enough evidence to reject the null hypothesis.

ADF Test Results

Weekly_Sales - Store 20

Results of Dickey-Fuller Test:	
Test Statistic	-4.719934
p-value	0.000077
#Lags Used	4.000000
Number of Observations Used	102.000000
Critical Value (1%)	-3.496149
Critical Value (5%)	-2.890321
Critical Value (10%)	-2.582122

ADF Test Results

Weekly_Sales - Store 4

Results of Dickey-Fuller Test:	
Test Statistic	-3.509825
p-value	0.007738
#Lags Used	4.000000
Number of Observations Used	102.000000
Critical Value (1%)	-3.496149
Critical Value (5%)	-2.890321
Critical Value (10%)	-2.582122

Figure 10. ADF Test Results - Weekly_Sales

Autocorrelation (ACF) and partial autocorrelation (PACF) plots were created to check the relationship between values of Weekly_Sales and its previous values. The ACF Plot, as seen in Figure 11, reduces gradually, with lag up to 2 being significant for both stores. The PACF Plot is irregular with a single, significant lag of 1, before decreasing for store 20. For store 4, the PACF plot shows a lag of 4 being significant before decreasing.

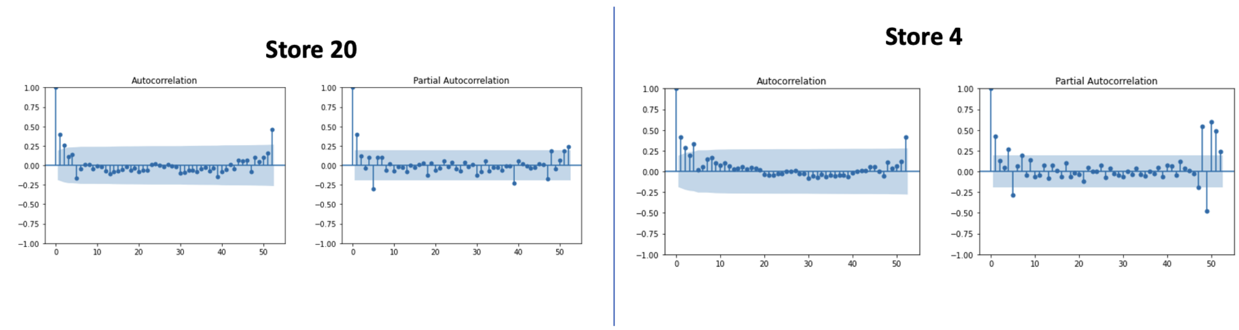


Figure 11. ACF and PACF Plots

Autoregressive Model (AR)

With the data being stationary and the PACF plots showing lags of 1 and 4, an AR model (with lag 1 for store 20 and lag 4) is selected as the base for analysis. The results for the AR model are shown in Figure 12. The forecast was mainly just the average of the testing dataset for all weeks, which is not correct and yields poor results.

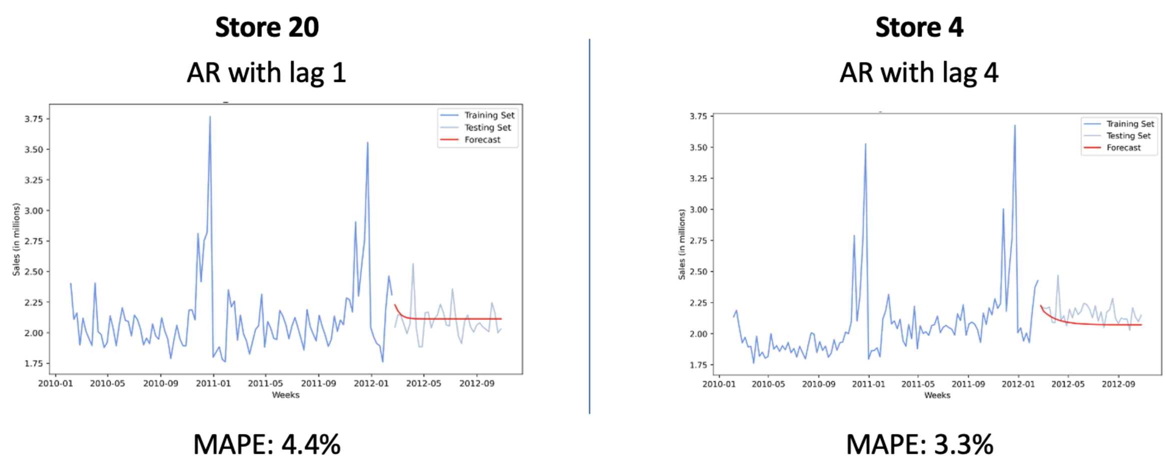


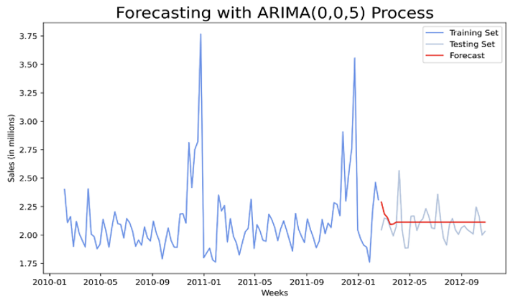
Figure 12. AR Model Results

Autoregressive Integrated Moving Average Model (ARIMA)

The next model that was chosen was ARIMA as it would take into account the moving average of the data. An Auto-ARIMA function was used to select the best coefficients for each store. The results of the ARIMA model are shown in Figure 13. The forecast is slightly better than the AR model, but the forecast is still mainly just the average of the testing set which yields poor results.

Store 20

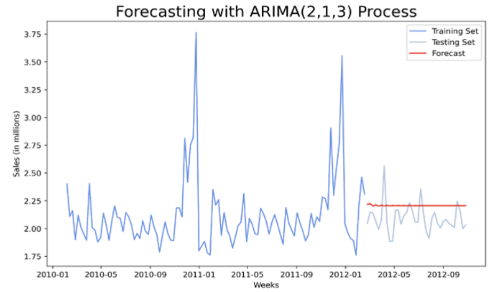
ARIMA(0,0,5)



MAPE: 4.6%

Store 4

ARIMA(2,1,3)



MAPE: 3.3%

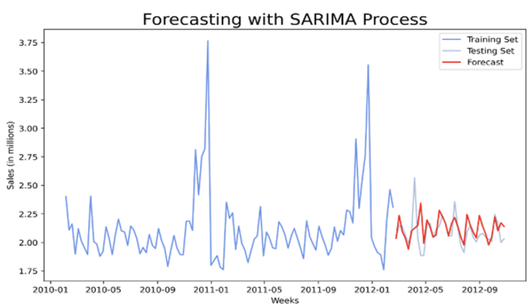
Figure 13. ARIMA Model Results

Seasonal Autoregressive Integrated Moving Average Model (SARIMA)

Since the data has a seasonality component, SARIMA was chosen as the next model with a 52 week lag for the seasonality lag. The results of the SARIMA model are shown in Figure 14. The results are much better now that seasonality is considered.

Store 20

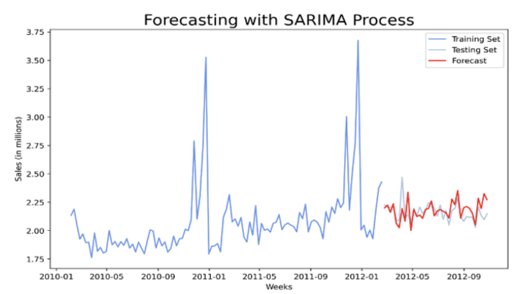
SARIMA(0,1,1)(0,1,1,52)



MAPE: 3.9%

Store 4

SARIMA(0,1,1)(0,1,1,52)



MAPE: 3.0%

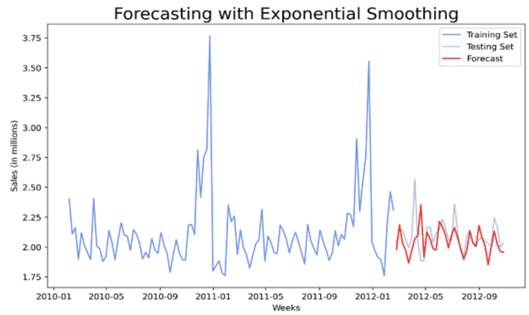
Figure 14. SARIMA Model Results

Exponential Smoothing

Exponential smoothing is another model that takes seasonality and trend into account, so it was chosen as the next model. The results for exponential smoothing are shown in Figure 15. The results are similar to SARIMA and better compared to AR and ARIMA.

Store 20

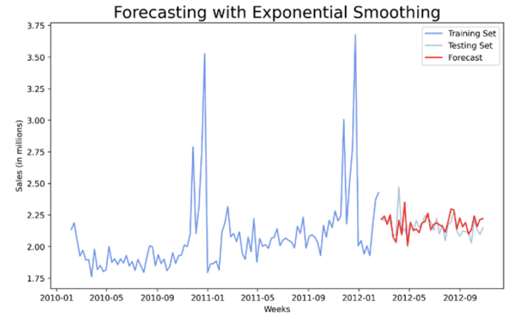
Exponential Smoothing



MAPE: 4.4%

Store 4

Exponential Smoothing



MAPE: 2.6%

Figure 15. Exponential Smoothing Model Results

Multivariate Modeling

Since additional data was provided in features.csv, multivariate modeling was attempted to see if Markdowns have an effect on Weekly_Sales. Prior to starting the modeling, stationarity of the Markdown variables needed to be checked. Results from the ADF tests are shown in [Figure 16](). For Markdown2 and Markdown3, p-values are lower than 0.05 and the Test Statistic is lower than the Critical Value at 5% for both stores so we can reject the null hypothesis. For Markdown_145, however, the p-value is greater than 0.05 and the Test Statistic is greater than the Critical Value at 5% for both stores which fails to reject the null hypothesis indicating that Markdown_145 is non-stationary.

ADF Test Results

Store 20

Markdown_145

Test Statistic	5.870487
p-value	1.000000
# Lags	13.000000
# Observations	93.000000
Critical Value (1%)	-3.502705
Critical Value (5%)	-2.893158
Critical Value (10%)	-2.583637

Markdown2

Test Statistic	-1.871161e+01
p-value	2.034595e-30
# Lags	1.300000e+01
# Observations	9.300000e+01
Critical Value (1%)	-3.502705e+00
Critical Value (5%)	-2.893158e+00
Critical Value (10%)	-2.583637e+00

Markdown3

Test Statistic	-8.036739e+00
p-value	1.893846e-12
# Lags	0.000000e+00
# Observations	1.060000e+02
Critical Value (1%)	-3.493602e+00
Critical Value (5%)	-2.889217e+00
Critical Value (10%)	-2.581533e+00

ADF Test Results

Store 4

Markdown_145

Test Statistic	5.752318
p-value	1.000000
# Lags	13.000000
# Observations	93.000000
Critical Value (1%)	-3.502705
Critical Value (5%)	-2.893158
Critical Value (10%)	-2.583637

Markdown2

Test Statistic	-5.968594e+00
p-value	1.960655e-07
# Lags	1.300000e+01
# Observations	9.300000e+01
Critical Value (1%)	-3.502705e+00
Critical Value (5%)	-2.893158e+00
Critical Value (10%)	-2.583637e+00

Markdown3

Test Statistic	-9.982978e+00
p-value	2.090912e-17
# Lags	0.000000e+00
# Observations	1.060000e+02
Critical Value (1%)	-3.493602e+00
Critical Value (5%)	-2.889217e+00
Critical Value (10%)	-2.581533e+00

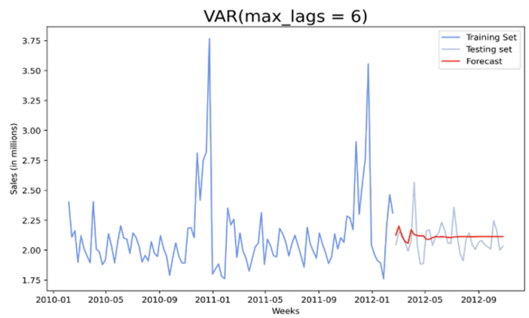
Figure 16. ADF Test Results - Markdown_145, Markdown2, Markdown3

Rather than directly removing Markdown_145, transforming Markdown_145 to a stationary variable was attempted using Simple Moving Average, Cumulative Moving Average, Exponential Moving Average, and Differencing. Markdown_145 was still failing the ADF test for all of these methods so it was dropped from the dataset.

Vector Autoregressive Model (VAR)

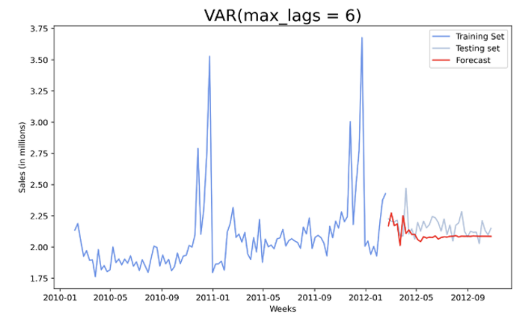
Similar to univariate modeling, an AR model was selected as the base model for multivariate modeling but in this case as there are multiple variables that affect the forecast, a VAR model is used. A grid search (with AIC as the metric) was conducted to determine the best value of max_lags for VAR modeling. The grid search resulted in an optimal max_lags value of 6. The results of the VAR model with max_lags of 6 are shown in Figure 17. Similar to the start of univariate modeling, the forecast accuracy is good but it is not capturing the patterns correctly.

Store 20



MAPE: 4.5%

Store 4



MAPE: 3.8%

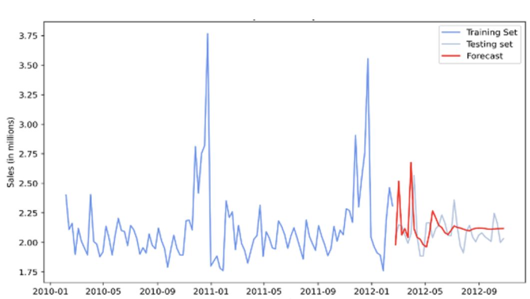
Figure 17. VAR Model Results

Vector Autoregressive Moving Average Model (VARMA)

The next model chosen was a VARMA model. Similar to what was done in ARIMA modeling, an Auto-ARIMA function was used to obtain optimal p and q values to use in VARMA. The results of the VARMA model are shown in Figure 18. The forecast is better than VAR with some of the patterns being captured but there is an increase in MAPE.

Store 20

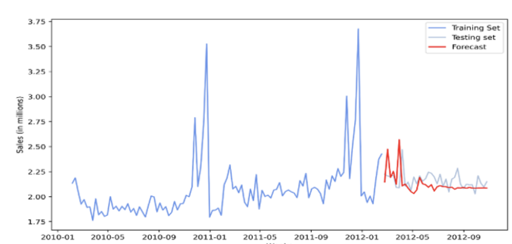
VARMA(6,0,6)



MAPE: 5.6%

Store 4

VARMA(6,0,6)



MAPE: 4.0%

Figure 18. VARMA Model Results

Conclusions

Granger Causality

Since the MAPE was increasing when using multivariate modeling, a Granger Causality test was conducted to see if either Markdown2 or Markdown3 was a causal variable for Weekly_Sales. In this test, the null hypothesis is that Markdown2/3 does not Granger cause Weekly_Sales and the alternative hypothesis is that Markdown2/3 does Granger cause Weekly_Sales. The results of the Granger Causality Test are shown in Figure 19. Based on the F test (due to smaller sample, F test is preferable) granger causality tests, we fail to reject the null hypotheses that Markdown2/Markdown3 do not Granger cause Weekly_Sales as the p-values are greater than 0.05.

Granger Causality Test Results

Store 20

Markdown2

```
Granger Causality
number of lags (no zero) 3
ssr based F test:         F=1.2810  , p=0.3946  , df_denom=4, df_num=3
```

Markdown3

```
Granger Causality
number of lags (no zero) 3
ssr based F test:         F=2.2592  , p=0.2237  , df_denom=4, df_num=3
```

Granger Causality Test Results

Store 4

Markdown2

```
Granger Causality
number of lags (no zero) 3
ssr based F test:         F=2.0898  , p=0.2443  , df_denom=4, df_num=3
```

Markdown3

```
Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.8110  , p=0.5505  , df_denom=4, df_num=3
```

Figure 19. Granger Causality Test Results

Results Comparison

A table with the results by model type is shown in Figure 20. Based on MAPE and the Granger Causality test, univariate modeling seems to model the data better than multivariate. Within univariate, SARIMA and Exponential Smoothing perform the best.

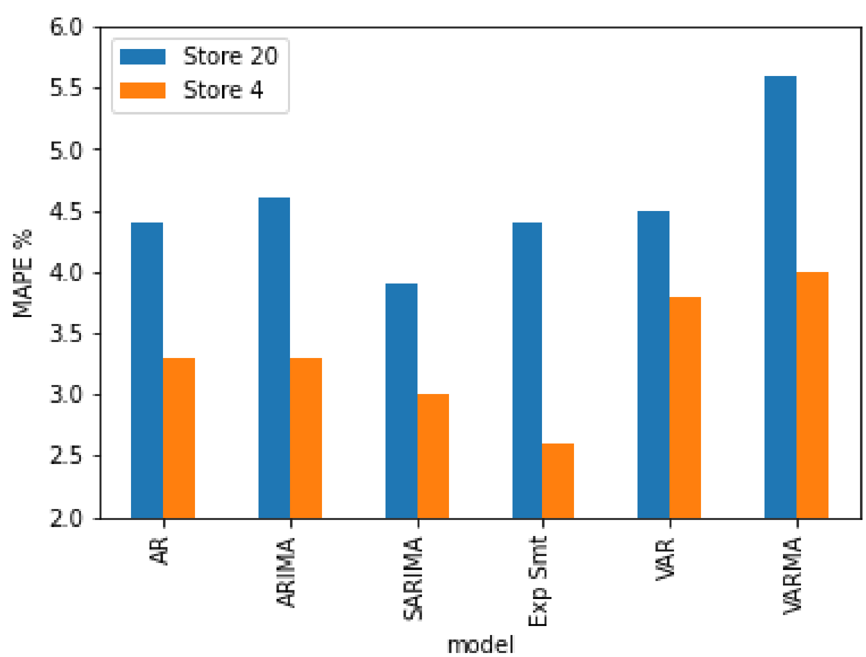


Figure 20. Results Comparison

Final Model Selection

To pick between SARIMA and Exponential Smoothing, these models were tested on the data for the remaining stores. Ideally, a grid search would be conducted for each store for SARIMA to identify the most optimal parameters. This process is extremely resource-intensive, however, and would take more time than is allocated to the project. To overcome this obstacle, the optimal SARIMA parameters for stores 20 and 4 [SARIMA(0,1,1)(0,1,1,52)] were used for all stores.

When testing Exponential Smoothing across stores, it resulted in an average MAPE of 5.0 and a median MAPE of 5.0. Even without optimizing for parameters, SARIMA performed better with an average MAPE of 4.8 and median MAPE of 4.0 across stores. As seen in Figure 21, SARIMA performs better at almost every store as well as on average.

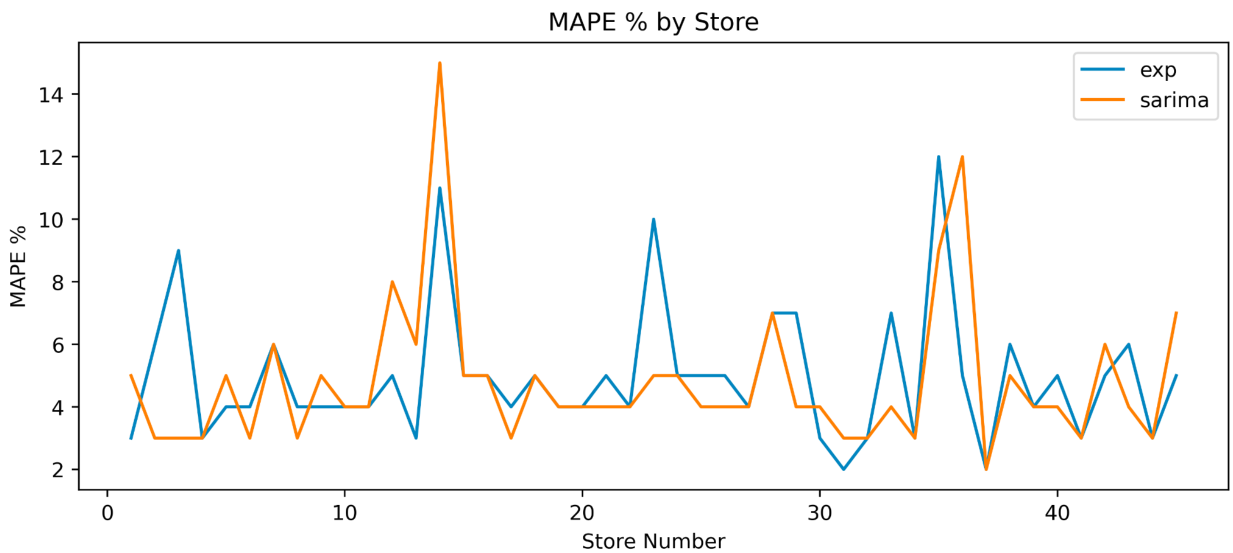


Figure 21. SARIMA & Exponential Smoothing Model Test Results by Store

It is assumed that the SARIMA results would be even better if the parameters were optimized for every store. Thus, SARIMA is selected as the final model for this analysis.