**Building Model with and without the features**

**Part A:**

Given the good results seen in the previous section, as a next step to further enhance or test a new methodology, an alternate way of analysis was evaluated to forecast sales. As we know, from the initial data visualization that the sales of the stores are correlated with the type of store or size of the store i.e Type A, B or C, three new dataset were chosen corresponding to each store type. This helped us to manage the correlation between the stores with this step. This helped us to enhance our current model based on sales aggregated by date overall

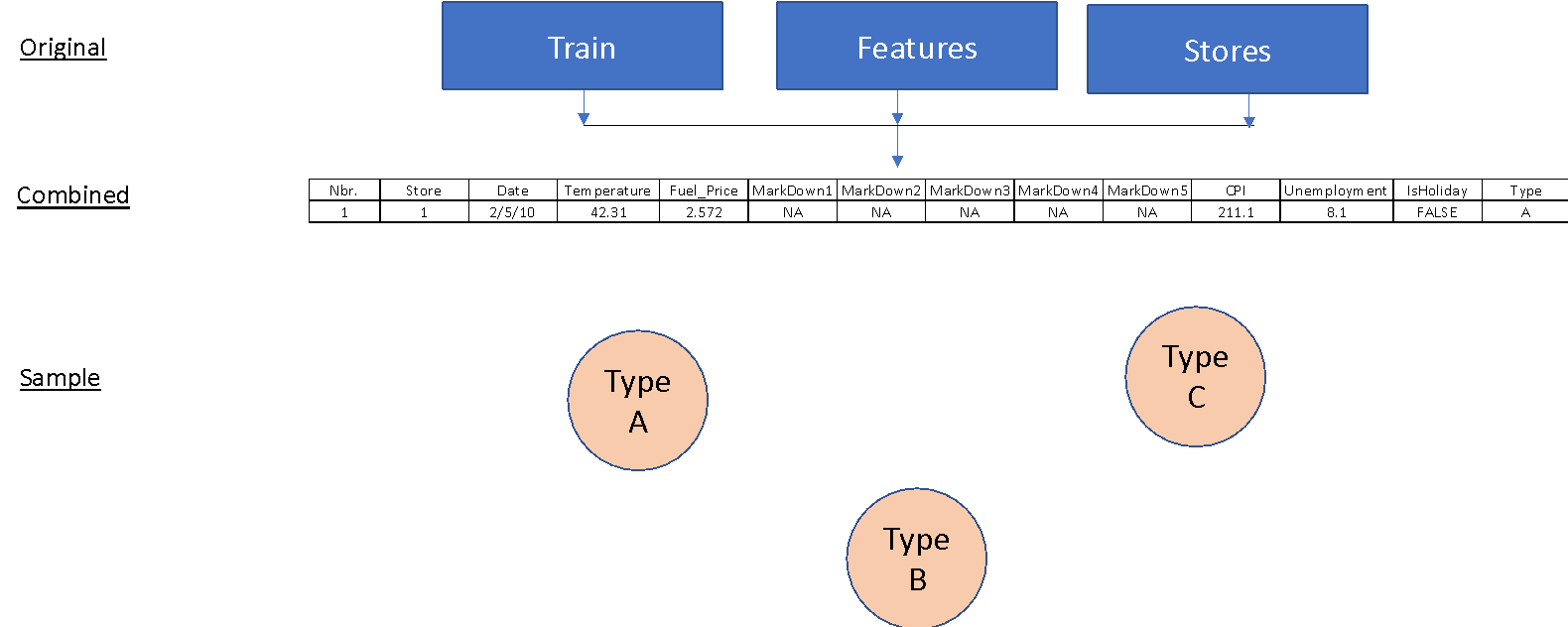
1. Dataset for Modeling:

Originally, we were provided with three datasets (Train, Features and Stores). Merged these datasets by Store and Date as key to arrive at the final merged dataset which we used as a base for modeling.

In the second step, selected one store from each of the store type randomly and used these samples for analysis of each of the three store types.

In the third step, we divided the dataset into an 80:20 split. We ended with 143 observations for each store type.

Please see below for detailed description



II. Model Building Process

Similar procedure as outlined in previous sections was leveraged to do the model diagnosis and then select the final mode. Model diagnosis with different parameters were implemented to analyze p values, correlations, residuals, and lag cut offs to detect the best model. We attempted different SARIMA models as shown in Figure below. The first step in building the model for all the datasets was as follows

1. Split data to train and test
2. Check acf and pacf for the store data and find right models
3. Convert to time series data
4. Looked at the various components of timeseries
5. Observed some seasonality and decreasing trend
6. Removed trend in timeseries
7. Used transformed data to find appropriate model

For the models where we can see that the p values go below the line after lag, that meant the residuals are correlated while in Models where all the p values are above the line, which means the residuals are independent. In the 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.

Type A : Step 1

Shape

Description automatically generated with medium confidence

Type A: Step 2

Diagram, shape

Description automatically generated

Type A: Step 3



Type B: Step 1

Diagram, engineering drawing

Description automatically generated

Type B: Step 2

Diagram

Description automatically generated

Type B: Step 3

A picture containing graphical user interface

Description automatically generated

Type C: Step 1

Diagram

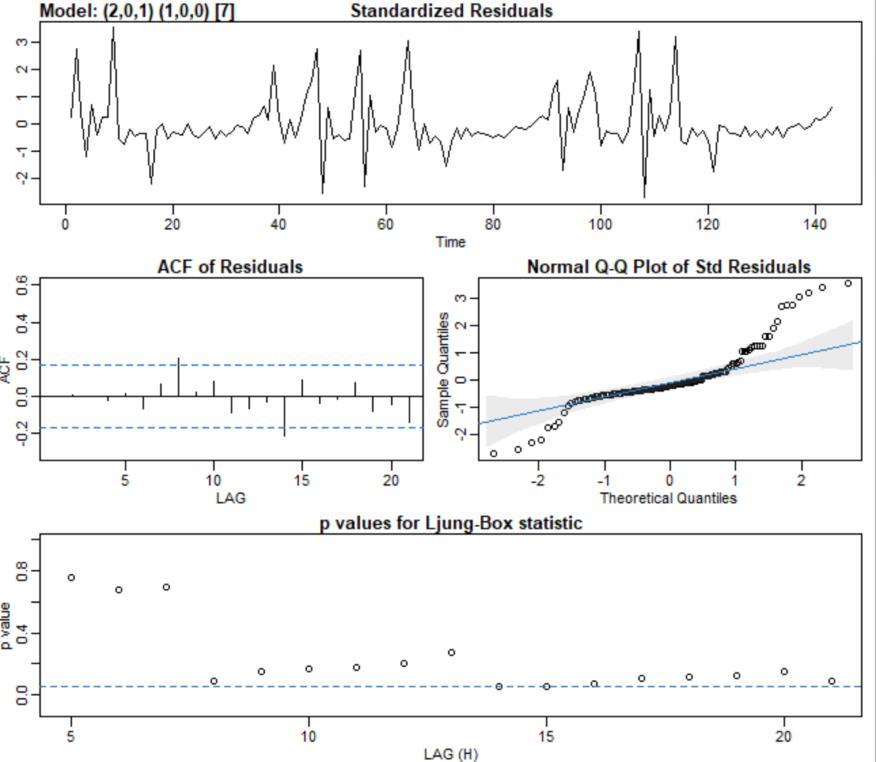
Description automatically generated

Type C: Step 2

Diagram, schematic

Description automatically generated

Type C: Step 3



Part B) 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. **As shown in the previous analysis and also based on the p values in the linear regression, we saw 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.

We leveraged the correlation plot earlier to visualize the relationship between temperature and sales. Additionally, 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.

We used linear regression along with the sarima model and came up with the final MSPE. The model with additional variables along with Temperature showed better results especially in Type A which had higher sales.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Sarima Without Features | Sarima With Variables (with Linear regression) |  |
|  | Type A | 0.118 | 0.105 |  |
|  | Type B | 0.56 | 0.45 |  |
|  | Type C | 0.075 | 0.11 |  |
|  |  |  |  |  |

* Temperature and Unemployment were significant features for the model
* Tried Linear regression

The forecast show increasing trends for Type ADiagram

Description automatically generated