**Walmart Sales Prediction:** **Strategic Forecasting for Enhanced Retail Performance**

**Executive Summary**

The project's primary objective was to develop a sophisticated forecasting model to improve the accuracy of sales predictions for Walmart. The team aimed to delve into the nuances of retail sales forecasting by considering various factors that could potentially affect sales outcomes, such as store type, size, and an array of key performance indicators (KPIs). The project further aimed to leverage the forecasted sales data to develop strategic promotional tactics tailored to enhance performance in the top Walmart stores.

At the heart of this analysis was the use of an ARIMAX (Autoregressive Integrated Moving Average with Explanatory Variable) model. This model was chosen for its ability to handle complex seasonal patterns and incorporate external variables, providing a more comprehensive analysis of sales determinants. The model's performance was measured against key metrics such as the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (SBC), and the Mean Absolute Percentage Error (MAPE), where improvements over traditional ARIMA models were observed.

The team's analysis included assessing the impact of store type and size on total and weekly sales, identifying the KPIs with the most significant influence on Walmart's sales, and developing strategic promotional tactics for the top-performing stores. Key drivers, such as promotional campaigns and store-specific factors like temperature and the Consumer Price Index (CPI), were factored into the model to understand their impact on sales.

One of the key findings was that promotions significantly affect sales, with an observed delayed impact peaking at specific lags. Smaller stores demonstrated better sales per unit size, indicating efficient space utilization and possibly a better customer experience. In contrast, larger stores, while contributing to higher overall sales, could benefit from optimizations.

Based on these insights, the team recommended several strategies:

* **Optimizing Store Layout**
* **Enhancing Customer Experience**
* **Targeted Promotions**
* **Community Engagement**
* **Fulfillment and Delivery Services**

# **Details of the Dataset**

**Introduction to the Dataset for Modelling**

**Name: Forecasting Walmart Sales**

**Data source: Kaggle**

**Data link:** [**Walmart Sales Forecast (kaggle.com)**](https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast?select=train.csv)

**Overview:**

The dataset utilized for modeling comprises historical sales data from Walmart stores, spanning from May 2, 2010, to July 26, 2013. It encompasses 8,190 records across 45 Walmart stores, each containing 12 features. This dataset is crucial for understanding sales patterns and drivers within Walmart's retail environment.

**Structure of the Data:**

The dataset is structured in CSV format, facilitating easy access and manipulation. Each record represents weekly sales alongside associated factors, including temperature, fuel price, and holiday status. These factors play pivotal roles in influencing consumer behavior and overall sales performance. However, it's worth noting potential limitations such as missing data points or outliers, which necessitate preprocessing for accurate modeling.

**Key Variables:**

1. **Weekly Sales:** This serves as the target variable, depicting the total sales for a given week across all Walmart stores.
2. **Temperature:** Represents the average temperature during the week, impacting shopping patterns and consumer behaviour.
3. **Fuel Price:** Indicates the average fuel price prevailing during the week, affecting both operational costs and consumer spending power.
4. **Holiday Status:** An indicator denoting whether the week encompasses a major holiday, which typically correlates with fluctuations in sales volumes.
5. **Markdown1 to Markdown5:** Columns indicating different promotional activities, providing insights into the impact of promotions on sales.
6. **Consumer Price Index (CPI):** Reflects changes in the price level of consumer goods and services, influencing consumer spending habits and purchasing power.

**Preprocessing for Modelling:**

**1. Aggregation of Daily Sales Data:**

Given the granularity of daily sales data, it was aggregated to weekly sales figures to streamline the modelling process. This decision mitigates the potential challenge of handling excessive data points.

**2. Summation of Departmental Sales:**

Each store encompasses approximately 98 departments, necessitating the summation of departmental sales to obtain weekly aggregate data. This aggregation simplifies the modelling process while retaining essential insights into overall store performance.

**3. Store Categorization and Exploration:**

The dataset includes 45 stores categorized as Store A, B, and C. To enhance model specificity, exploration focuses on 2-3 stores from each category separately. This approach facilitates a comparative analysis of modelling results and parameters across different store types.

**4. Consideration of Seasonality:**

Recognizing the presence of seasonality within the data, the modelling process incorporates only the first two years of data (2010-2012). This selective inclusion aims to capture seasonal patterns effectively while minimizing noise from subsequent years.

**5. Data Preparation and Exploration:**

Utilization of SAS for Data Cleaning and Separation:

Data cleaning and separation were conducted using SAS, a powerful statistical analysis tool. This process yielded separate datasets for each store, enabling focused exploration and analysis tailored to individual store characteristics.

We chose the following stores from each type of store to Forecast find the correlations:

|  |  |  |
| --- | --- | --- |
| Store | Type | Size |
| 1 | A | 151315 |
| 11 | A | 207499 |
| 24 | A | 203819 |
| 31 | A | 203750 |
| 3 | B | 37392 |
| 12 | B | 112238 |
| 22 | B | 119557 |
| 37 | C | 39910 |
| 43 | C | 41062 |

**SAS code used for data cleaning and partition:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Store wise Weekly\_Sales Exploration:**

**Avg. Sales by Store Type & Holiday Status**

**Monthly Avg. Sales by Store Type**

Type A stores consistently boast higher sales figures compared to Type B establishments. Type B, in turn, generally outperforms Type C in terms of sales metrics. This hierarchy underscores the significant impact of store type on revenue generation within the retail landscape.

One notable factor contributing to the sales differentials between Type A, B, and C stores is the presence of seasonality. Both Type A and B stores exhibit discernible patterns of seasonal fluctuations, wherein sales tend to peak during certain periods of the year and dip during others. This seasonality often corresponds with consumer behavior influenced by factors such as holidays, weather changes, or cultural events.

In contrast, Type C stores lack this pronounced seasonality in their sales patterns. Their revenue streams remain relatively stable throughout the year, without experiencing significant peaks or troughs tied to seasonal variations. This distinction highlights a fundamental divergence in the operational dynamics and consumer appeal of Type C stores compared to their Type A and B counterparts.

During holidays, sales surge across all types of stores. Whether they're Type A, B, or C, establishments experience increased consumer spending driven by the festive spirit.

# Approach to Forecasting the Weekly Sales Data – Store 3 (Store Type – B)

# Time Series Exploration for Individual Variables

## Dependent Variable – Weekly\_Sales

A graph showing a line graph

Description automatically generated with medium confidenceA graph showing the seasons

Description automatically generated with medium confidence

A graph showing the sales of a company

Description automatically generated with medium confidenceA table with numbers and lines

Description automatically generated

Bases on the Time-series exploration of Weekly\_Sales and Dickey-Fuller Unit Root Test, it is clear that our data is a stationary dataset and has seasonal component without any trend component.

When we look at the Correlation functions of the Weekly\_Sales, ACF has lagged effect of upto lag3 and dropped suddenly and PACF has spikes at Lag1 and lag5 and the data is not white noise.

The ACF (Autocorrelation Function) shows a lagged effect up to lag 3 before dropping suddenly. This suggests that there may be some short-term autocorrelation in the data up to lag 3.

The PACF (Partial Autocorrelation Function) has spikes at lag 1 and lag 5, indicating direct relationships with the previous observation and the observation five periods ago. This suggests that there may be some significant short-term and longer-term autocorrelation in the data.

A graph of sales

Description automatically generated with medium confidence

Weekly\_Sales also has a strong seasonal component.

A graph showing the fall of sales

Description automatically generated

Our Data has some independent variables named Temperature, Fuel\_price, CPI, Unemployment, IsHoliday and different promotional functions named Markdown1, Markdown2, Markdown3, Markdown4 and Markdown5.

We explored the cross correlations of these variables with our target variable. Below are the croos-correlation plots.

A graph of a graph showing a line of a graph

Description automatically generated with medium confidence A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generatedA graph with a line

Description automatically generated with medium confidence

A graph with a line graph

Description automatically generated with medium confidence A graph with blue lines

Description automatically generated

A graph of a graph

Description automatically generated A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated A graph with a bar graph

Description automatically generated with medium confidence

Based on the Cros-Correlations, we can see that Temperature, Markdown1, Markdown3, Markdown4 and Markdown5, IsHoliday have some effect on our target variable.

# Modeling and Forecasting

Based on the characteristics of our dataset's ACF and PACF, as well as its stationary nature with seasonality, we explored several ARIMA models that might be suitable.

Based on the sudden drop in the ACF after lag 3, the significant spikes in the PACF at lag 1 and lag 5 and only seasonality without trend,

AR (p) terms: Potential values for p might include 1 or 5, reflecting the significant spikes in the PACF.

MA (q) terms: Potential values for q might include 0,1 or 3, considering the sudden drop in the ACF after lag 3.

Since our dataset exhibits seasonality, we'll need to include seasonal terms in your ARIMA model with Differencing order (D).

Based on the possible values of (p,d,q) and (P,D,Q) we built multiple models to find the best fit model.

**Building ARIMA Models:**  
  
**ARIMA (5,0,3) (0,1,0)** has AIC and SBC values as 1205.8 and 1223.3 respectively with normally distributed residual with some signal.

A screenshot of a graph

Description automatically generated

A diagram of a normality curve

Description automatically generated

**ARIMA (1,0,1) (0,1,0)** has AIC and SBC values as 1191.048 and 1196.902 respectively with normally distributed residual with some signal.

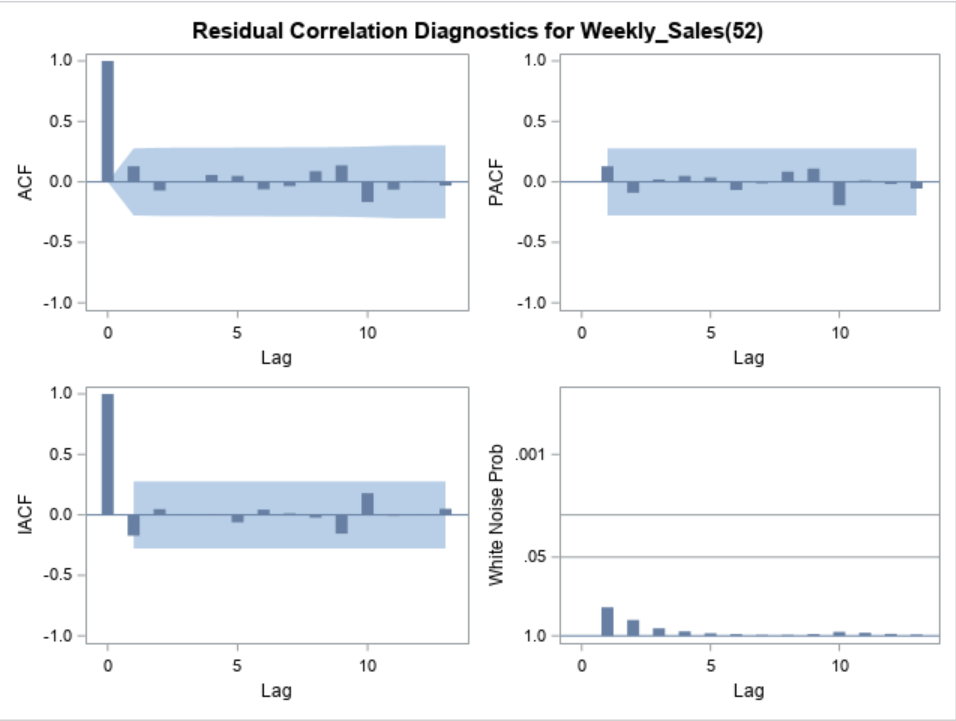
A group of graphs showing different types of data

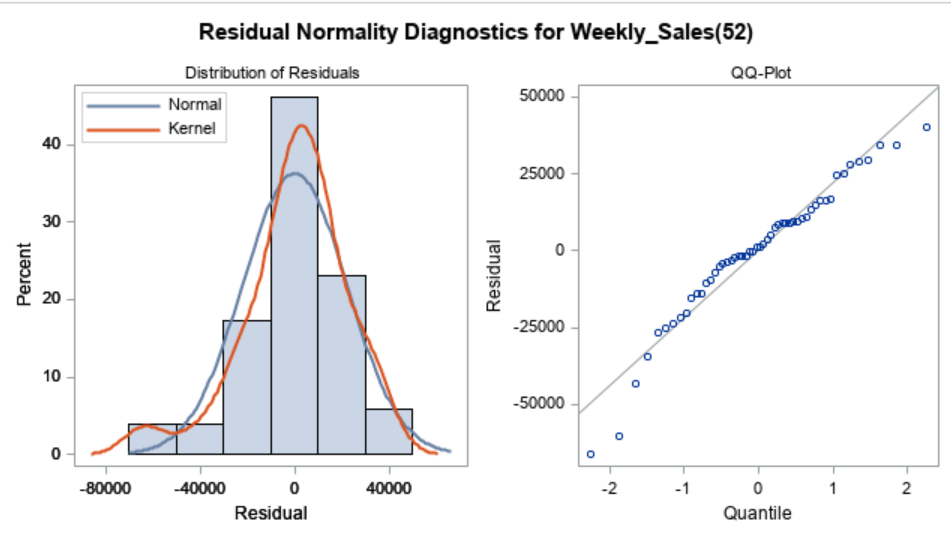
Description automatically generated with medium confidence

A diagram of a normality

Description automatically generated

**ARIMA (0,0,0) (0,1,0)** has AIC and SBC values as 1188.386 and 1190.338 respectively with normally distributed residual with some signal.

n



**ARIMA Model Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| ARIMA (0,0,0)(0,1,0) | Yes | Yes | 1188.386 | 1190.34 |
| ARIMA (1,0,1)(0,1,0) | Yes | Yes | 1191 | 1196.9 |
| ARIMA (5,0,3)(0,1,0) | No | Yes | 1205.8 | 1223.3 |

When we compare the three models based on AIC & SBC, it is clear that ARIMA (0,0,0) (0,1,0) is a better fit model.

Now, we will check the effect of independent variables in our Forecasting model.

**Building ARIMAX Models:**

Based on the Cros-Correlations, we saw that Temperature, Markdown1, Markdown3, Markdown4 and Markdown5, IsHoliday have some effect on our target variable and some of these values are time-series individually.

Prewhitening refers to the process of removing autocorrelation from a time series before using it as an independent variable in a regression model. While it's not always necessary, it can be beneficial to check the actual influence of these independent variables with the target variable.

In model building, prewhitening is employed to address autocorrelation issues that violate the assumptions of the modeling framework. By removing autocorrelation, prewhitening can improve the estimation and inference of model parameters, leading to more accurate and reliable forecasts.

The first step in prewhitening is to fit a suitable model to the time series data to capture the autocorrelation structure. This model could be an autoregressive (AR), moving average (MA), autoregressive integrated moving average (ARIMA), or other appropriate models depending on the characteristics of the data. Once the model is fitted, obtain the residuals by subtracting the fitted values from the observed values. These residuals ideally should exhibit no autocorrelation if the model adequately captures the autocorrelation structure of the data. With the uncorrelated residuals obtained through prewhitening, we can now proceed with further analysis, such as checking correlations or building models, confident that autocorrelation has been appropriately addressed.

**Independent variable - Temperature**

When prewhitening was performed on the independent variable Temperature and the cross-correlation with Weekly Sales was checked, the following results were obtained:

A group of graphs showing different types of data

Description automatically generated with medium confidence

A graph on a white background

Description automatically generated

The lack of significant lagged values in both ACF and PACF of the prewhitened Temperature series suggests that the temporal dependence or autocorrelation in Temperature has been removed through prewhitening. This means that after prewhitening, the Temperature series behaves like white noise, exhibiting no systematic patterns or temporal dependencies.

The observed lagged effect of up to lag 13 in the cross-correlation between Weekly Sales and the prewhitened Temperature series suggests that there might be a delayed effect of Temperature on Weekly Sales.

Hence, it is beneficial to include lag component of 13 when building a Forecasting model.

**Independent variable – Markdown1**

When prewhitening was performed on the independent variable Markdown 1 and the cross-correlation with Weekly Sales was checked, the following results were obtained:

A group of graphs showing different types of data

Description automatically generated with medium confidence

A graph on a white background

Description automatically generated

The lack of significant lagged values in both ACF and PACF of the prewhitened Markdown1 series suggests that the temporal dependence or autocorrelation in Markdown1 has been removed through prewhitening. This means that after prewhitening, the Markdown1 series behaves like white noise, exhibiting no systematic patterns or temporal dependencies.

The observed lagged effect of up to lag 6 in the cross-correlation between Weekly Sales and the prewhitened Markdown1 series suggests that there might be a delayed effect of Markdown1 on Weekly Sales.

Hence, it is beneficial to include lag component of 6 when building a Forecasting model.

**Independent variable – Markdown3**

**A graph with blue lines

Description automatically generated**

The observed lagged effect of up to lag 4 in the cross-correlation between Weekly Sales and the prewhitened Markdown3 series suggests that there might be a delayed effect of Markdown3 on Weekly Sales.

A group of blue and white bars

Description automatically generated

But the exploration of Markdown3 suggests that the series itself is a whitenoise. An independent variable that is complete white noise signifies randomness and lack of predictability. While such a variable may not be useful on its own for modeling or forecasting, it's important to consider its potential interactions with other variables in the dataset and evaluate its contribution to the overall model's performance.

Hence, we can choose the inclusion of this variable based on its influence on the model.

**Independent variable – Markdown4**

When prewhitening was performed on the independent variable Markdown 4 and the cross-correlation with Weekly Sales was checked, the following results were obtained:

**A collage of graphs

Description automatically generated**

**A graph with blue lines

Description automatically generated**

The lack of significant lagged values in both ACF and PACF of the prewhitened Markdown4 series suggests that the temporal dependence or autocorrelation in Markdown4 has been removed through prewhitening. This means that after prewhitening, the Markdown4 series behaves like white noise, exhibiting no systematic patterns or temporal dependencies.

The observed lagged effect of up to lag 6 in the cross-correlation between Weekly Sales and the prewhitened Markdown4 series suggests that there might be a delayed effect of Markdown4 on Weekly Sales.

Hence, it is beneficial to include lag component of 4 when building a Forecasting model.

**Independent variable – Markdown5**

When prewhitening was performed on the independent variable Markdown 5 and the cross-correlation with Weekly Sales was checked, the following results were obtained:

**A group of graphs with numbers

Description automatically generated with medium confidence**

**A graph with blue lines

Description automatically generated**

The lack of significant lagged values in both ACF and PACF of the prewhitened Markdown5 series suggests that the temporal dependence or autocorrelation in Markdown5 has been removed through prewhitening. This means that after prewhitening, the Markdown5 series behaves like white noise, exhibiting no systematic patterns or temporal dependencies.

The observed lagged effect of up to lag 3 in the cross-correlation between Weekly Sales and the prewhitened Markdown5 series suggests that there might be a delayed effect of Markdown5 on Weekly Sales.

Hence, it is beneficial to include lag component of 3 when building a Forecasting model.

**Building ARIMAX (0,0,0)(0,1,0) with Variables - Temp, M1,3,4,5**

Based on the above observations from prewhitening of the independent variables, we saw that Temperature has lagged effect of upto 13lags, MarkDown1 has lagged effect of upto 6lags, MarkDown4 has lagged effect of upto 6lags, MarkDown5 has lagged effect of upto 3lags. We will build our model with the variable Markdown3 also, even though it is a whitenoise series, as it showed some lagged effect of upto 4 lags.

**A graph of a number of data

Description automatically generated with medium confidence**

**A graph of normality and a normality curve

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

**A graph showing the sales of a company

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

ARIMAX (0,0,0)(0,1,0) with Variables - Te, M1,3,4,5 has AIC and SBC values as 875.8 and 888.8 respectively with normally distributed residual with no signal in residuals.

This model is performing much better compared with the similar ARIMA model.

**Building** **ARIMAX (0,0,0)(0,1,0) with Variables - Temp, M1,4,5**

Based on the above observations from prewhitening of the independent variables, we saw that Temperature has lagged effect of upto 13lags, MarkDown1 has lagged effect of upto 6lags, MarkDown4 has lagged effect of upto 6lags, MarkDown5 has lagged effect of upto 3lags.

**A group of graphs showing different types of data

Description automatically generated with medium confidence**

**A graph of normality and a normality curve

Description automatically generated with medium confidence**

ARIMAX (0,0,0)(0,1,0) with Variables - Te, M1,4,5 has AIC and SBC values as 874 and 882.3 respectively with normally distributed residual with no signal in residuals.

This model is performing better compared with the similar ARIMA model and the ARIMAX model with Variables - Te, M1,3,4,5

A screenshot of a computer

Description automatically generated

A graph showing the sales of a company

Description automatically generated with medium confidence

**A screenshot of a computer

Description automatically generated**

Our accuracy measures of the forecasting performance for Weekly Sales indicates that the developed model demonstrates strong accuracy, with low values of Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics suggest that the model's predictions closely align with the actual values of Weekly Sales, with minimal deviation. Overall, the results affirm the effectiveness of our forecasting approach and underscore the model's potential to provide valuable insights for decision-making regarding Weekly Sales management.

**Building ARIMAX (1,0,1)(0,1,0) with Variables - Temp, M1,4,5**

To further explore more complicated forecasting models, we built the (1,0,1)(0,1,0) model with the four independent variables that are found useful in previous models.

A screenshot of a graph

Description automatically generated

A graph of normality and a normality curve

Description automatically generated

ARIMAX (0,0,0)(0,1,0) with Variables - Te, M1,4,5 has AIC and SBC values as 887.64 and 889.3 respectively with normally distributed residual with no signal in residuals.

This model is performing better compared with the similar ARIMA model and nut the previous ARIMAX(0,0,0)(0,1,0) has better results with less complexity.

**Model Comparison for Store3 Forecasting:**

Below is the summary of the comparison between different models that we built.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| ARIMAX (0,0,0)(0,1,0) X - Te, M1,4,5 | Yes | Yes | 874 | 882.3 |
| ARIMAX (0,0,0)(0,1,0) X - Te, M1,3,4,5 | Yes | Yes | 875.8 | 888.8 |
| ARIMAX (1,0,1)(0,1,0) X - Te, M1,4,5 | Yes | Yes | 887.64 | 889.29 |
| ARIMAX (0,0,0)(0,1,0) X - M1,4,5 | Yes | Yes | 1047.4 | 1054.7 |
| ARIMAX (0,0,0)(0,1,0) | Yes | Yes | 1188.386 | 1190.34 |
| ARIMAX (1,0,1)(0,1,0) | Yes | Yes | 1191 | 1196.9 |
| ARIMAX (5,0,3)(0,1,0) | No | Yes | 1205.8 | 1223.3 |

From the values of AIC and SBC, it is clear that ARIMAX (0,0,0)(0,1,0) with variables Temperature, Markdown1, Markdown4 and Markdown5 is a better performing model.

The promotional activities named Markdown1, Markdown4 and Markdown5 have impact on the Weekly\_Sales of the Store-3, it would be beneficial for the store management to increase these to potentially boost the profits.

Our best model achieved a MAPE of 1.21%, indicating that, on average, the model's predictions deviate by approximately 1.21% from the actual values of Weekly Sales.

**ARIMAX (0,0,0)(0,1,0) X - Te, M1,4,5, Model Code:**

Including the accuracy metrics:

A screenshot of a computer code

Description automatically generated

# Forecasting for Store 12 & Store 22 (Store Type – B)

**Store 12**

With the similar approach the we used for store-3, we explored another store-12 from the same store type B, with the aim to build a more accurate and less complex model.

Below is the summary of the models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| Store 12 ARIMAX (0,0,0)(0,1,0) X - M2,3,4,5 | Yes | Yes | 1260.5 | 1270.5 |
| Store 12 ARIMAX (0,0,0)(0,1,0) X - Temp, M2,3,4,5 | Yes | Yes | 1262.5 | 1274.2 |
| Store 12 ARIMAX (1,0,1)(0,1,0) X - Temp, M2,3,4,5 | Yes | Yes | 1264.6 | 1280.2 |
| Store 12 ARIMAX (0,0,0)(0,1,0) | Yes | Yes | 1282.6 | 1284.6 |
| Store 12 ARIMAX (1,0,1)(0,1,0) | Yes | Yes | 1284.9 | 1290.78 |
| Store 12 ARIMAX (1,0,2)(0,1,0) | Yes | Yes | 1286.7 | 1294.5 |
| Store 12 ARIMAX (5,0,1)(0,1,0) | No | Yes | 1292.7 | 1306.4 |
| Store 12 ARIMAX (5,0,2)(0,1,0) | No | Yes | 1293.527 | 1309.1 |

Among the various ARIMAX models tested for forecasting Weekly Sales at Store 12, the optimal model is determined based on several criteria including residual white noise, normal residual distribution, AIC (Akaike Information Criterion), and SBC (Schwarz Bayesian Criterion). The best-performing model is ARIMAX (0,0,0)(0,1,0) with exogenous variables Markdown2, Markdown3, Markdown4, and Markdown5 (abbreviated as Temp, M2, M3, M4, M5). This model satisfies the conditions of residual white noise and normal residual distribution while achieving the lowest AIC and SBC scores among the options presented. Therefore, it is deemed the most suitable for accurately forecasting Weekly Sales at Store 12.

A screenshot of a computer

Description automatically generated

The forecasting model developed for Store 12 demonstrates outstanding accuracy, with a Mean Absolute Percentage Error (MAPE) of just 1.41%. This metric represents the average percentage difference between the model's predictions and the actual observed values of Weekly Sales. With such a low MAPE, the model's forecasts are remarkably close to the true sales figures, indicating a high degree of precision in predicting sales patterns.

**Store 22**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Residual White noise | Normal Residual Distribution | AIC | SBC |
| ARIMAX (303,010) - TEMP,MK1,4,5 | Yes | YES | 329.9075 | 335.7612 |
| ARIMAX (101,010) - TEMP,MK1,4,5 | Yes | YES | 982.4465 | 986.349 |
| ARIMA\_(000,010) | Yes | YES | 1311.545 | 1313.496 |
| ARIMA\_(100,000) | No | YES | 2800.329 | 2805.617 |
| ARIMA\_(303,010) | Yes | YES | 1317.616 | 1331.275 |

Among the various ARIMAX models tested for forecasting Weekly Sales at Store 22, the optimal model is determined based on several criteria including residual white noise, normal residual distribution, AIC (Akaike Information Criterion), and SBC (Schwarz Bayesian Criterion). The best-performing model is ARIMAX (0,0,0)(0,1,0). This model satisfies the conditions of residual white noise and normal residual distribution while achieving the lowest AIC and SBC scores among the options presented. Therefore, it is deemed the most suitable for accurately forecasting Weekly Sales at Store 22.

**Tailoring Forecasting Models: Understanding Store-Specific Sales Dynamics**

When comparing the best forecasting model for Store-3 with that of Store-12, we observe notable differences in the selected variables and their respective effects on the stores' sales dynamics.

For Store-3, the optimal model identified is ARIMAX (0,0,0)(0,1,0) with exogenous variables Temperature (Te), Markdown1 (M1), Markdown4 (M4), and Markdown5 (M5). This contrasts with Store-12's model, which includes Temperature along with Markdowns 2, 3, 4, and 5.

The distinct choice of variables between the two stores underscores the individualized impact of factors on sales performance. Here's a breakdown of the reasoning behind the differences:

Markdown strategies and their effectiveness may vary across stores due to differences in promotional activities, inventory management, and customer preferences. Consequently, Markdowns 1, 4, and 5 may have a more significant impact on sales at Store-3 compared to Markdowns 2 and 3.

Temperature fluctuations can affect consumer purchasing behavior differently across regions. Store-3 may experience unique temperature-related demand patterns, necessitating the inclusion of Temperature (Te) as a crucial predictor in its forecasting model. For Store-12, a model without Temperature is able to perform better.

The selection of distinct variables in the forecasting models for Store-3 and Store-12 reflects the nuanced nature of retail operations and the diverse factors driving sales performance. By building separate models tailored to each store's unique characteristics and demand drivers, retailers can better anticipate sales trends and optimize decision-making processes to meet customer needs effectively.

# Analyzing the Forecasting Models for Store Type – A

Similar to the Store type B, we modeled two stores from Store type A and below are the results:

Prior to modeling, we performed exploratory data analysis to understand the characteristics and trends in the sales data for both stores.

This included visualizing time series plots, identifying seasonality and trends, and assessing autocorrelation and partial autocorrelation functions.

For Store , 11 and Store 24, we built ARIMA models to capture the underlying patterns and dynamics in the sales data.

Additionally, we extended our analysis to incorporate exogenous variables using ARIMAX modeling, with a focus on prewhitening techniques to assess cross-correlations between the sales data and the exogenous variables.

By prewhitening the exogenous variables, we ensured that any correlation observed between the variables was not due to autocorrelation within the exogenous variables themselves.

**Store 1:**

A white rectangular box with black text

Description automatically generated

**Store 11:**

A white rectangular box with black text

Description automatically generated

**Store 24:**

**A screenshot of a table

Description automatically generated**

Markdowns 1, 2, 3, and 4 emerged as significant predictors of sales for Store 1, while Store 11's best model included only Markdowns 1 and 3. This suggests that Markdowns 2 and 4 may have a lesser impact on sales for Store 11 compared to Store 1. The differing importance of specific markdowns highlights the need for nuanced and store-specific modeling approaches.

Both stores' best models have a similar ARIMA structure (ARIMA(0,0,0)(0,1,0)), indicating the need for differencing in the seasonal component to achieve stationarity. The simplicity of the models (with only exogenous variables included) suggests that additional complexity, such as higher-order autoregressive or moving average terms, may not significantly improve forecasting accuracy for these stores.

The lack of significance of Temperature in the best models for Store Type A suggests that other factors, particularly Markdowns, play a more prominent role in driving sales dynamics for these stores. By focusing on the most influential predictors and tailoring forecasting models accordingly, retailers can better allocate resources and develop targeted strategies to optimize sales performance.

For Store 24, the best model seems to be (ARIMA(1,0,0)(0,1,0)), without any indepent variables.

# Analyzing the Forecasting Models for Store Type – C

Stores 37 & 43 were also explored similar to the previous stores and below are the model comparisons.

**Store 37:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| Store 37 ARIMA (2,0,4) (0,1,0) | Yes | Yes | 1,541.5 | 1570.5 |
| Store 37 ARIMAX (4,0,4) (0,1,0) X - CPI | Yes | Yes | 1,341.5 | 1,435.2 |

**Store 43:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Residual White Noise** | **Normal Residual Distribution** | **AIC** | **SBC** |
| Store 43 ARIMA (2,0,4) (0,1,0) | Yes | Yes | 1664.6 | 1740.2 |
| Store 43 ARIMAX (1,0,3) (0,1,0) X - CPI | Yes | Yes | 1,388.7 | 1,401.8 |

Store 37 & Store 43 are performing better with ARIMAX models with CPI variables and other variables have no significant effect on the model’s performance.

# Type A Stores vs Type B Stores vs Type C Stores

Store Type A encompasses larger stores compared to Store Type B, suggesting potentially higher sales volumes, broader product assortments, and larger customer bases.

Both store types exhibit similarities in the selected ARIMA model structures, with differencing in the seasonal component to address seasonality.

Store Type A models incorporate Markdowns as significant predictors, while Store Type B models may include additional variables or exhibit variations in the importance of certain predictors.

Store Type A models emphasize the significance of Markdowns, suggesting a strong influence of promotional pricing strategies on sales dynamics.

In contrast, Store Type B models may prioritize different predictors, such as Temperature or other store-specific factors, reflecting variations in customer preferences, regional demographics, or competitive landscapes.

This comparison between Store Type A and Store Type B highlights the importance of understanding store-specific sales dynamics and leveraging advanced modeling techniques to develop accurate and actionable sales forecasts tailored to the unique characteristics of each store.

Store Type C comprises smaller stores compared to Store Type A and Store Type B, suggesting potentially lower sales volumes, narrower product assortments, and smaller customer bases.

Store Type C models emphasize the significance of CPI as the primary predictor of sales, indicating that changes in consumer purchasing power and inflation levels play a critical role in driving sales dynamics for these stores.

In contrast, Store Type A and Store Type B models may incorporate a broader range of predictors, such as Markdowns, Temperature, or other store-specific factors, reflecting the larger size and potentially more complex sales dynamics of these stores.

**Conclusion:**

To achieve accurate sales forecasts, it's imperative to recognize the distinct characteristics and dynamics of each store, highlighting the necessity of building tailored forecasting models rather than relying on a one-size-fits-all approach. Stores vary significantly in terms of size, location, customer demographics, product assortment, and promotional strategies, among other factors. These differences result in unique sales patterns and drivers that require individualized modeling. Attempting to apply a single model across all stores overlooks these nuances, leading to suboptimal forecasts and missed opportunities. By developing specific models for each store, retailers can leverage store-specific data and factors to capture the intricacies of sales dynamics accurately. This approach enables retailers to account for variations in customer behavior, regional trends, seasonal patterns, and other store-specific influences, resulting in more reliable forecasts and informed decision-making. Ultimately, building different models for different stores allows retailers to optimize resource allocation, tailor marketing strategies, and enhance overall performance, driving sustainable growth and competitive advantage in the retail landscape.

**Forecasted Sales KPIs by Store Number**

The analysis reveals intriguing insights into the relationship between store size, sales performance, and sales per unit store size across different store types. Despite Store Type A boasting higher overall sales figures, it's notable that their sales per unit store size are relatively smaller compared to other store types. This discrepancy suggests that Store Type A may not be fully optimizing their space or customer experience to maximize sales potential. Additionally, the observation that Store 11, despite its larger size compared to Store 1, registers lower sales underscores the importance of factors beyond mere square footage in driving sales performance.

The higher sales per unit store size in smaller stores may be attributed to more efficient space utilization and potentially better customer experiences, such as personalized services or tailored product offerings.

To capitalize on these findings, recommendations include optimizing store layouts for larger stores to enhance space utilization and customer flow, while also prioritizing customer experience enhancements such as personalized services and targeted promotions. Furthermore, leveraging larger stores as community hubs for events and gatherings, as well as utilizing them as fulfillment centers for online orders, can further enhance their role as key drivers of sales and customer engagement within the retail landscape.