Capstone Project

**WALMART RETAIL STORES ANALYSIS USING TIME SERIES ANALYSIS AND FORECASTING THE FUTURE STEPS**

BY :YUGESH MADALA

Mail id:yugeshyogi972@gmail.com

**TABLE OF CONTENTS**

**1.Introduction**

* 1.1. Objective of the Study
* 1.2. Data Description

2.**Methodology**

* 2.1. Importing Libraries and Datasets
* 2.2. Exploratory Data Analysis (EDA)

3.**Time Series Analysis**

* 3.1. Time Series Decomposition
* 3.2. Stationarity Tests
  + 3.2.1. Augmented Dickey-Fuller Test (ADF Test)
* 3.3. Autocorrelation and Partial Autocorrelation Analysis
  + 3.3.1. ACF Plot
  + 3.3.2. PACF Plot

4. **Modelling**

* 4.1. ARIMA Model
* 4.2. SARIMAX Model

5.**Forecasting**

* 5.1. Forecasting with SARIMAX

6.**Conclusion**

* 6.1. Summary of Findings

7. **References**

**1.INTRODUCTION:**

**1.1 Objective Of Study:**

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply.

**1)You are provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:**

1. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

b. If the weekly sales show a seasonal trend, when and what could be the reason?

c. Does temperature affect the weekly sales in any manner?

d. How is the Consumer Price index affecting the weekly sales of

various stores?

e. Top performing stores according to the historical data.

f. The worst performing store, and how significant is the difference

between the highest and lowest performing stores.

**2)Use predictive modelling techniques to forecast the sales for each store for the next 12 weeks.**

**1.2 DATA DESCRIPTION:**

The dataset used in this analysis contains weekly sales data from different Walmart store outlets. The data includes variables such as the Store, Date, Weekly Sales, Holiday flag, Temperature, Fuel Price, CPI, Unemployment. The Time range covered by the weekly Sales data of each store from 05-02-2010 to 26-10-2012.

|  |  |
| --- | --- |
| **Feature name** | **Description** |
| Store | Store Number |
| Date | Week of Sales |
| Weekly Sales | Sales for the given store in that week |
| Holiday Flag | If it is a Holiday week |
| Temperature | Temperature on the day of sale |
| Fuel Price | Cost of fuel in that region |
| CPI | Consumer Price Index |
| Unemployment | Unemployment Rate |

**2.METHODOLOGY:**

**2.1 Importing Datasets and Libraries:**

The libraries and packages were used for the analysis include NumPy, Pandas, Matplotlib, Seaborn, Statsmodels and sklearn.

**2.2 Exploratory Data Analysis (EDA):**

1. **Loading the Dataset:** The initial step involves loading the dataset into the analysis environment, typically using libraries like Pandas in Python, ensuring accessibility for further examination.

2. **Checking for Data Types:** Checking the datatype that each column handles so that to convert the datatypes into a suitable format for better analysis.

3. **Converting Date Column:** When dealing with time series data, such as dates, converting the date column from an object type to a date type and also made date column as an index facilitates time-based analyses and visualizations.

4. **Handling Outliers:** Identification and treatment of outliers are crucial to maintain data integrity.

Outliers are assessed visually through plots or statistically using methods like z-scores or IQR (Interquartile Range) and handled the same by using upper and lower limits using IQR values and also handling the same using z-scores has also been done

5. **Correlation Analysis:** I have utilized tools like heatmaps aids in understanding the interrelationships between various features within the dataset, providing insights into potential dependencies and guiding further exploration.

6. **Exploring Relationships:** Beyond correlation analysis, exploring relationships between columns through visualizations and statistical methods unveils additional patterns and dependencies, enriching the understanding of the dataset's dynamics.

**3.TIME SERIES ANALYSIS:**

**3.1 Time Series Decomposition:**

Time series decomposition involves breaking down the time series data into its trend, seasonal, and residual components. This helps in understanding the underlying patterns in the data. I have taken the example of store number 4 and plotted the Time series decomposition plots to illustrate the same.

**3.2 Stationary Tests:**

Stationarity is a crucial property for time series analysis and forecasting. A stationary time series has constant mean, variance, and autocorrelation over time. Non-stationary data can lead to misleading results in time series modelling.

**3.2.1 Augmented Dickey-Fuller Test (ADF Test):**

The Augmented Dickey-Fuller test (ADF test) is used to check the stationarity of a time series. The null hypothesis of the ADF test is that the time series is non-stationary.

**ADF Statistic**: If the ADF statistic is less than the critical value, we reject the null hypothesis and conclude that the time series is stationary.

**P-value:** A low p-value (< 0.05) indicates that we can reject the null hypothesis of non-stationarity.

In this analysis, I have performed the ADF tests on the taken examples of stores and if found the data not stationary, I have gone for the first differencing of the data and also performed necessary shifts and rolling means in order to make the data detrended.

**Steps for Differencing the data:**

**1)Load the Dataset:** Load the time series data into a panda Data Frame.

**2)First Differencing:** Use the diff() function to calculate the difference between consecutive observations.

**3)Drop NaN Values:** Differencing will create a NaN value at the first observation, which should be dropped**.**

**4)Plot the Differenced Series:** Visualize the differenced series to inspect if the trend has been removed.

**5)Check for Stationarity:** Use the Augmented Dickey-Fuller test to check if the differenced series is now stationary.

By performing first differencing, I have prepared the data for further analysis and modelling using techniques like ARIMA, which assume the input series is stationary. If the first differencing does not achieve stationarity, further differencing (second differencing) or other transformations may be necessary.

**3.3 Auto corelation and Partial Auto corelation Analysis:**

**3.3.1: ACF Plot:**

The Autocorrelation Function (ACF) plot shows the correlation of the time series with its own lagged values. It helps in identifying the order of the MA (Moving Average) component in the ARIMA model.

**3.3.2 PACF Plot:**

The Partial Autocorrelation Function (PACF) plot shows the correlation of the time series with its own lagged values, after removing the effect of earlier lags. It helps in identifying the order of the AR (Auto-Regressive) component in the ARIMA model**.**

From the ACF and PACF plots, I have found the suitable orders required for implementing the ARIMA model for the different stores and then continued the analysis on the same.

**4.MODELLING:**

**4.1 ARIMA MODEL:**

The ARIMA (Auto Regressive Integrated Moving Average) model is a powerful and widely used statistical method for time series forecasting. ARIMA models are suitable for univariate time series data, particularly when the data shows evidence of non-stationarity.

It consists of three parameters: p (AR order), d (difference order), and q (MA order).

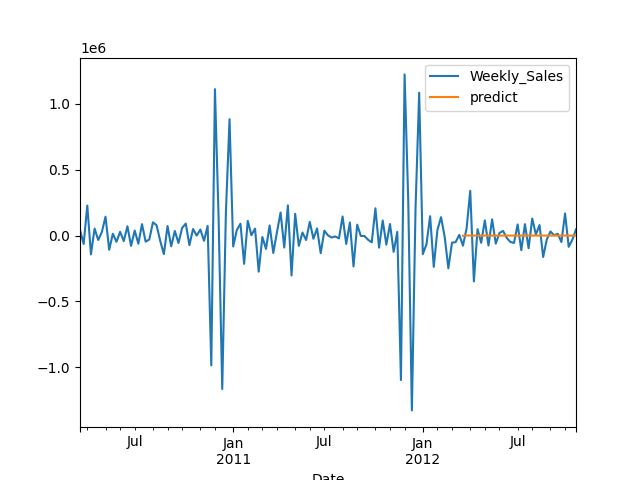
**Fitting An ARIMA Model:**

The process of fitting an ARIMA model generally involves the following steps:

**1)Identification**: Determine the values of p, d, and q by analysing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.

**2)Estimation**: Use statistical software to estimate the parameters of the ARIMA model.

As per the analysis for the given stores, The ARIMA model isn’t the best fit because of the presence of seasonality. I have done the above process of fitting the model but when plotted I haven’t got the correct prediction to testing data. Hence I have opted for SARIMAX model. The below image shows that ARIMA model is not suitable for this data.

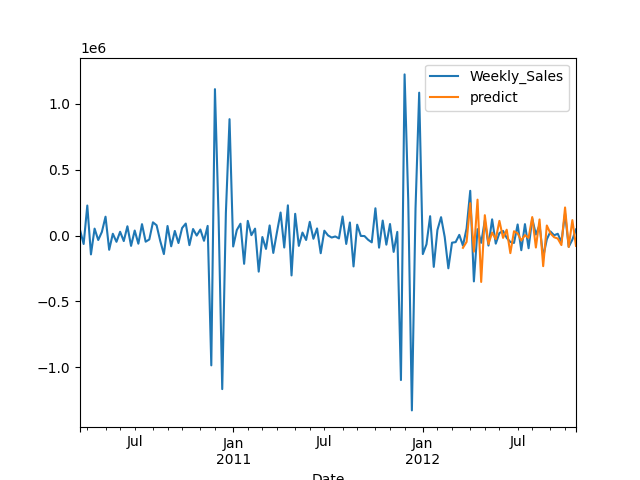


**4.2 SARIMAX MODEL:**

The SARIMAX (Seasonal AutoRegressive Integrated Moving Average with exogenous factors) model extends ARIMA to support seasonal effects and exogenous variables. It includes parameters for the seasonal component: P (seasonal AR order), D (seasonal difference order), Q (seasonal MA order), and s (seasonal period).

As per the analysis using auto ARIMA for fitting the SARIMAX model, I have fitted the data to the model and then plotted the same.The plot has been satisfactory when compared the weekly sales with predicted ones and also the SARIMAX model is performing better than ARIMA.

The below image shows the weekly sales vs predicted sales using SARIMAX model:



In the same way, The Time series analysis has been performed for another two stores using the above steps and have done the forecasting.

5.FORECASTING:

**5.1 Forecasting using SARIMAX:**

SARIMAX model which has been fitted to the above data has been used for forecasting and the same has been done for the next 12 steps. As forecasting only 12 steps of data doesn’t give much information regarding the patterns, forecasting has been done with 52 steps taking into consideration of the future patterns.

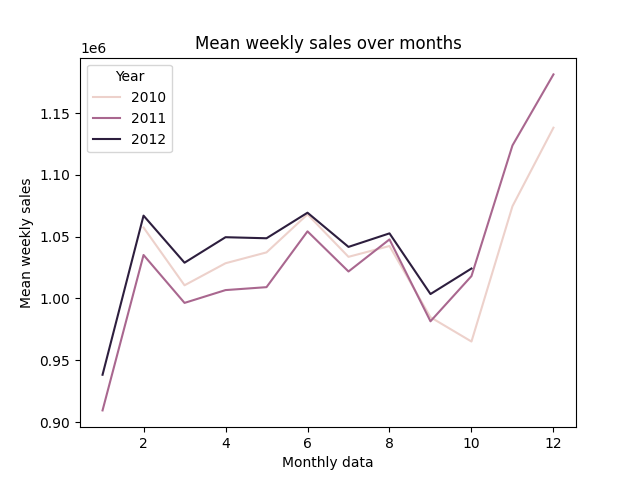
6.CONCLUSION:

**6.1 Summary of findings:**

**a)If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?**

Yes, the store ID’S affected are:36 and 44

**b)If the weekly sales show a seasonal trend, when and what could be the reason?**

****

Based on the above graph, We can observe that there has been a sudden increase in the Mean Weekly Sales from the Month of October to Month of December in the years 2010,2011 and also in 2012.As December month could be the most possible holiday month on the occasion of Festival there could have been an sudden increase in the market sales.

**C) Does temperature affect the weekly sales in any manner?**

Based on a scatter plot, we can observe that the distribution of weekly sales is more when the temperature is between 40 and 80 Fahrenheit. From this we can say that low temperatures i.e. less than 40 Fahrenheit and high temperatures i.e. more than 80 Fahrenheit are not ideal for better weekly sales.

**d)How is the Consumer Price index affecting the weekly sales of various stores?**

After performing a scatter plot, We can observe that there is no relationship between Consumer Price Index and Weekly Sales as the distribution of data of weekly sales has been clustered into three groups.

**e) Top performing stores according to the historical data.**

The top five performing stores are :

Store

4 2.867997e+08

20 2.867490e+08

14 2.799706e+08

13 2.739669e+08

2 2.706436e+08

**f)The worst performing store, and how significant is the difference between the highest and lowest performing stores.**

The Worst Performing Store: 33

The difference between top store and bottom store: 249639479.20000002

**7. REFERENCES:**

* **Documentation of NumPy, Pandas, Matplotlib, Statsmodels and Scikit-learn Libraries**
* **Kaggle. (n.d.). Retrieved from [https://www.kaggle.com](https://www.kaggle.com)**