Capstone Project – Walmart Times Series Analysis

Problem Statement:

Walmart with its wide network of stores around the nation, has a significant inventory management challenge and demand forecasting issues. The complex challenge of balancing the supply and demand of goods is the focus of the current issue. Mis-Match between Demand and supply has resulted in a number of operational inefficiencies that have an effect on overall performance and customer satisfaction levels, such as stockouts or surplus inventory. This misalignment impacts the store's operating capability in addition to presenting financial concerns in the form of potential revenue loss. In order to properly address these difficulties, it becomes necessary to improve the inventory management process. The retail store wants to apply data-driven tactics that improve inventory visibility, decrease shortages of goods, reduce excess stock holding costs, and ultimately improve the overall efficiency of the supply chain by recognizing and predicting customer demand patterns. This Study focuses on forecasting the sales of the top performing stores and also aims to find the relationship among variable to understand the influences these variables have on the sales.

Project Objectives:

- To Study the relationship between Weekly sales and Unemployment rate.
- To Study the impact of Temperature on weekly sales of the stores.
- To find out whether there is any linearity between Consumer price index and weekly sales.
- To Forecast the sales for next 12 weeks for Five Top performing stores in terms of sales.

Data Description:

```
 #   Column         Non-Null Count   D-Type
---  ------          --------------    -----
 0   Store          6435 non-null    int64
 1   Date           6435 non-null    object
 2   Weekly Sales   6435 non-null    float64
 3   Holiday Flag      6435 non-null   int64
 4   Temperature       6435 non-null   float64
 5   Fuel Price        6435 non-null   float64
 6   CPI               6435 non-null   float64
 7   Unemployment  6435 non-null   float64
```

We have total 8 columns, Store column contains the store number, Date column contains the weekly dates, Weekly Sales column contain the sales value of that particular store, Holiday flag column contains whether that particular day was a holiday or not, Temperature columns contains that particular day's temperature, CPI column contain the consumer price index value.

Data Preprocessing:

Data preprocessing is an essential step because it helps us in building more accurate Machine learning model. It improves the model's accuracy to predict the target variable.

Steps in Data Preprocessing:

- Data cleaning
- Data Transformation
- Data reduction

Data cleaning:

Data cleaning includes various tasks like handling missing data points and outlier treatment. uncleaned data set can cause major problems in our machine learning models. For example, extreme outliers can cause our model to give biased results. That's why it is very important to treat any such inconsistencies.

Strategies to handle missing data and outliers:

1. Imputation of missing value with the central values like mean, median and mode.
2. Removing the outliers from the data.
3. Imputing the outliers with lower bound or upper bound value.

A. We don't have any missing value in our data set.

B. We have 45 Unique stores.

C. We have 481 Outliers in our unemployment column.

Data Transformation:

Data transformation is the process of converting raw data into standard format which helps the machine to easily understand the data. It involves structuring, cleaning, transforming the data to increase the accuracy of the machine learning models.

Data Transformation involves many activities:

1. Changing the data types of the variables.
2. Encoding categorical variables.
3. Normalising the data
4. Aggregating the data

We did Log transformation of one the store's sales value because the data was not stationary.

Data Reduction:

Data reduction is a group of techniques used to select the most important features from the datasets which will contribute to better models building and accuracy.

Choosing the Algorithm for the Project:

FB – Prophet Model for Time-series analysis:

Prophet is a method for predicting time series data using an additive model. It fits non-linear trends with seasonality on a daily, weekly, and annual basis, together with the effects of holidays. Strong seasonal effects in time series and multiple seasons of historical data are ideal for its effectiveness. Prophet usually handles outliers well and is resilient to missing data and trend changes.

Ordinary Least Square Method for Regression analysis:

Ordinary least squares (OLS) regression is an optimization strategy that helps you find a straight line as close as possible to your data points in a linear regression model. OLS is considered the most useful optimization strategy for linear regression models as it can help you find unbiased real value estimates for your alpha and beta.

Correlation Analysis:

Correlation analysis is a statistical technique used to evaluate the strength and direction of the linear relationship between two quantitative variables.

Motivation and Reasons for Choosing the Algorithm:

We choose FB Prophet model because of the following reasons:

- ✓ Prophet provides a straightforward and user-friendly interface, making it accessible to users with varying levels of expertise in time series forecasting.
- ✓ The model inherently includes the ability to incorporate holidays and special events, allowing for more accurate predictions during periods with distinct patterns.
- ✓ Prophet automatically decomposes time series data into components such as yearly, weekly, and daily seasonality, as well as a trend component.
- ✓ Prophet is designed to handle missing data and outliers, which reduces data preprocessing time.

Model Evaluation and Technique:

Correlation Analysis:

Weekly Sales Vs Unemployment rate

From the correlation analysis between weekly sales and unemployment rate, we got the ‘r’ value of **-0.10** indicating negative correlation between these two variables which in turn means that if unemployment rate decreases the weekly sales increases.

Temperature Vs Weekly Sales

From the correlation analysis between weekly sales and temperature, we got the ‘r’ value of **-0.06** indicating negative correlation between these two variables which in turn means that if temperature decreases the weekly sales increases.

Consumer price Index Vs Weekly sales

As per the correlation value of **-0.07,** we can say that when Consumer price index decreases the sales value increases which makes sense.

Regression Analysis:

Temperature Vs Weekly Sales

From the Ordinary least square method, we got ‘p’ value of 0.00 which less than the significance value of 0.05 but the r2 value 0.04 that means only 4 % of variance is explained by the model.

Time Series Analysis:

The aim of this time series analysis is to forecast sales of the top performing stores for next 12 weeks.

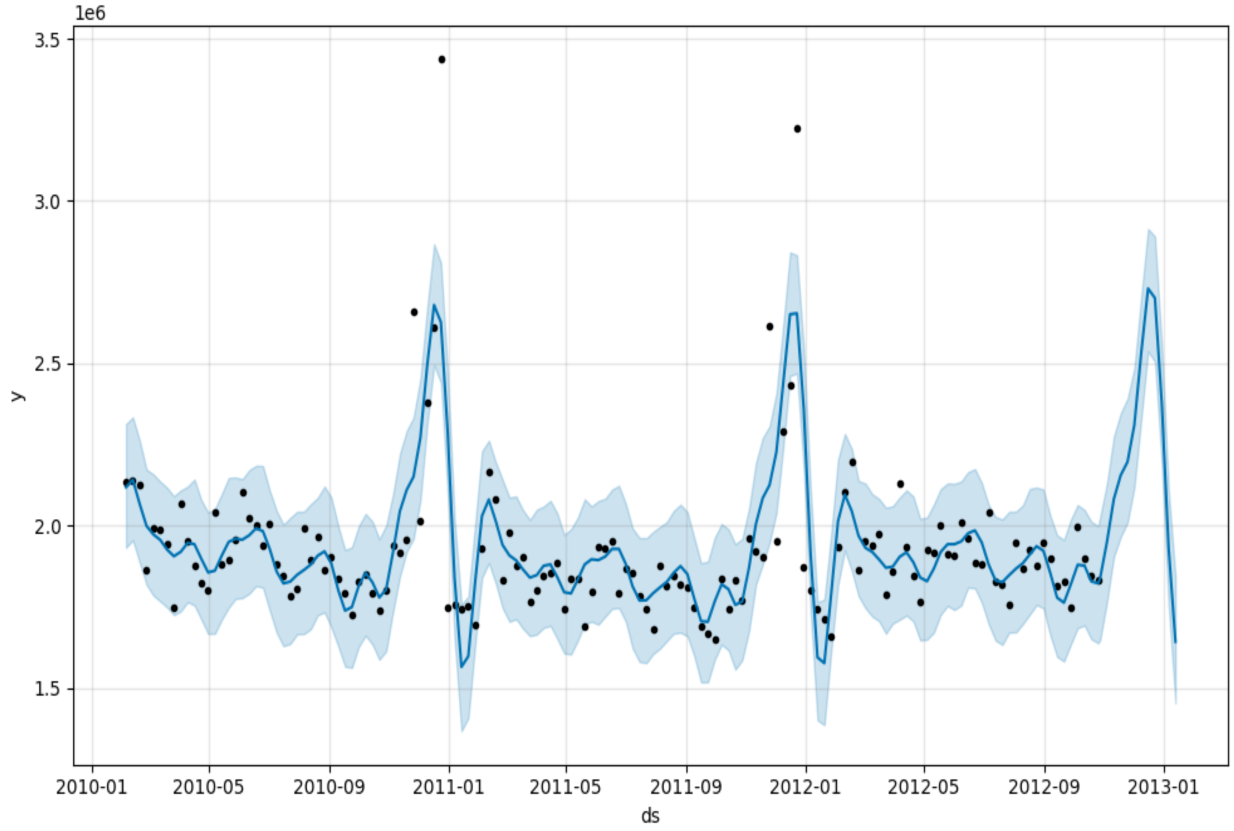
Top Performing store in terms of sales are:

- Store No 2
- Store No 4
- Store No 13
- Store No 14
- Store No 20

Predicting sales for Store No 2:

- We have total 143 records for each store, means sales value of 143 weeks.

- The sales show some seasonality during the end of the year
- The goal is to forecast the sales for next 12 weeks with confidence interval of 0.99



In the above graph the black dots are the actual sales and the blue line is the predicted sales. The shaded blue region is the confidence interval which include upper bound and lower bound.

Model evaluation: Root mean squared error metric has been used to evaluate the model performance.

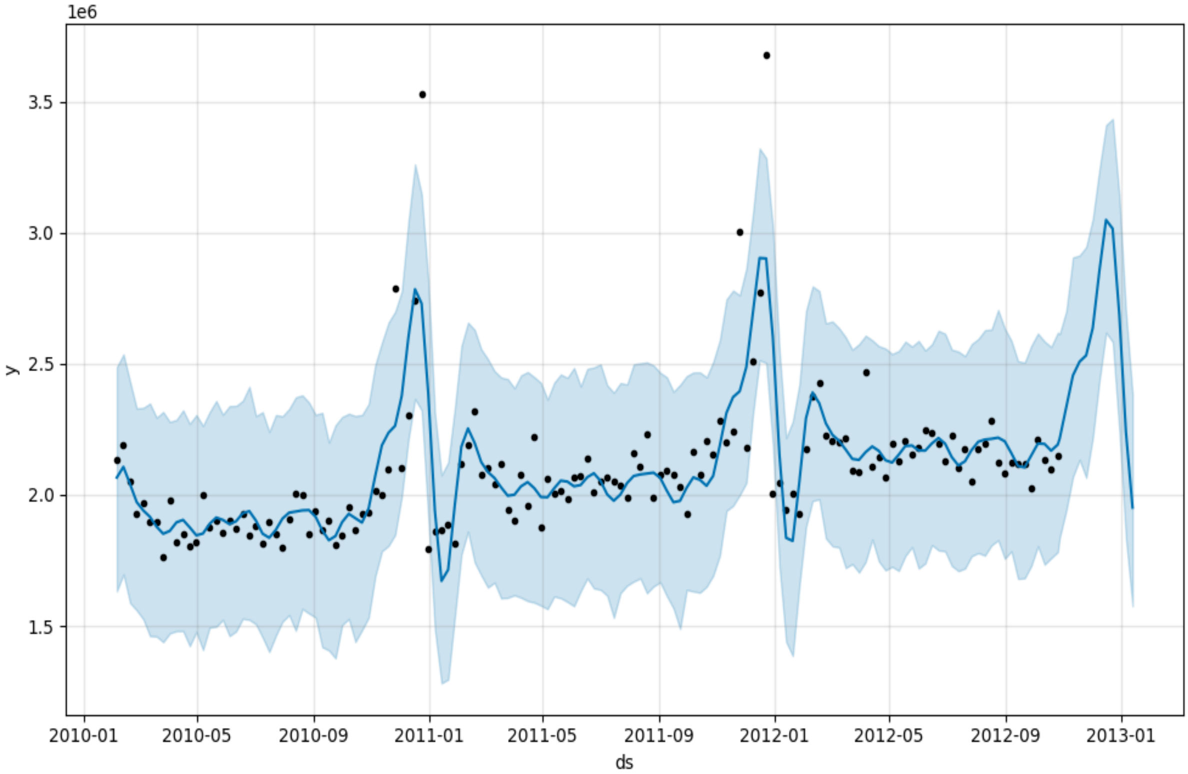
Root mean squared error = 63888.54

- The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the errors between predicted and observed values.

- The RMSE value of 63888.54 suggests that, on average, the model's predictions deviate by approximately $63,888.54.
- These results are based on the prediction of last 29 weeks. I have taken last 29 weeks to reduce the magnitude of values.

Predicting sales for Store No 4:

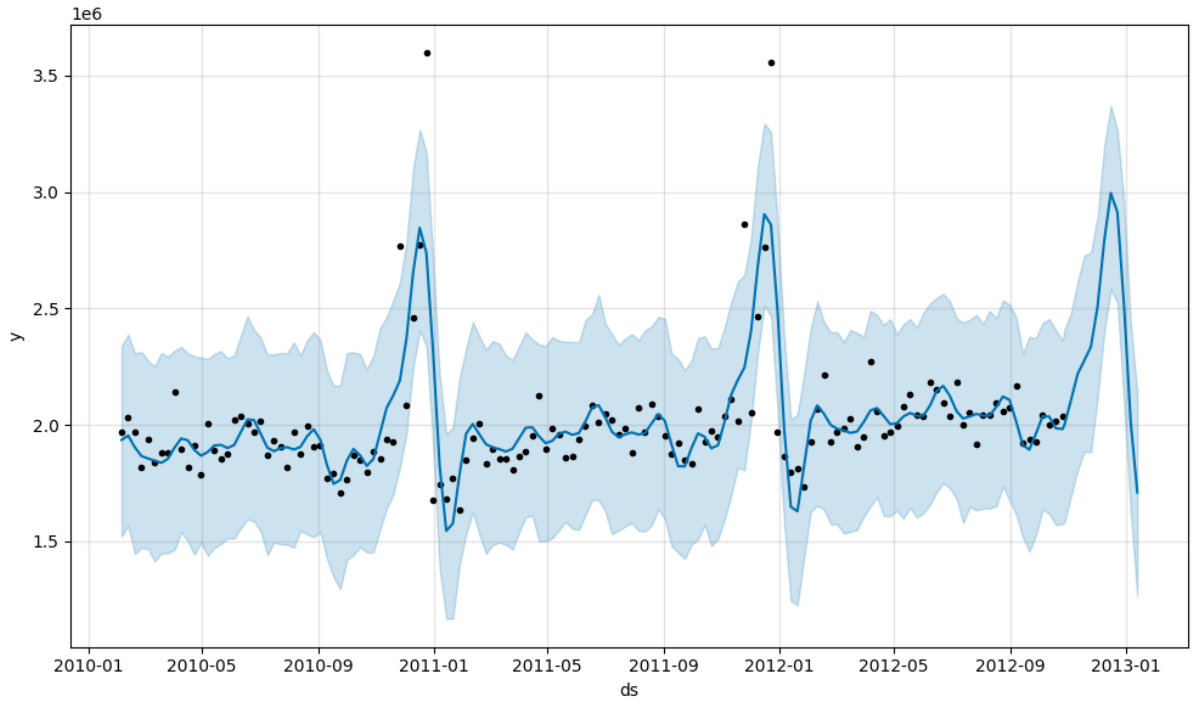
- ❖ For Store no 4 also we have 143 records.
- ❖ In the weekly sales of Store No: 4 an upward trend has been observed, since we are using prophet model, we don’t have to remove the trend.
- ❖ From the ad fuller test, we have observed that the data is stationary.



The RMSE value of 62167.86 suggests that, on average, the model's predictions deviate by approximately $62167.86.

Predicting Sales for Store No 13:

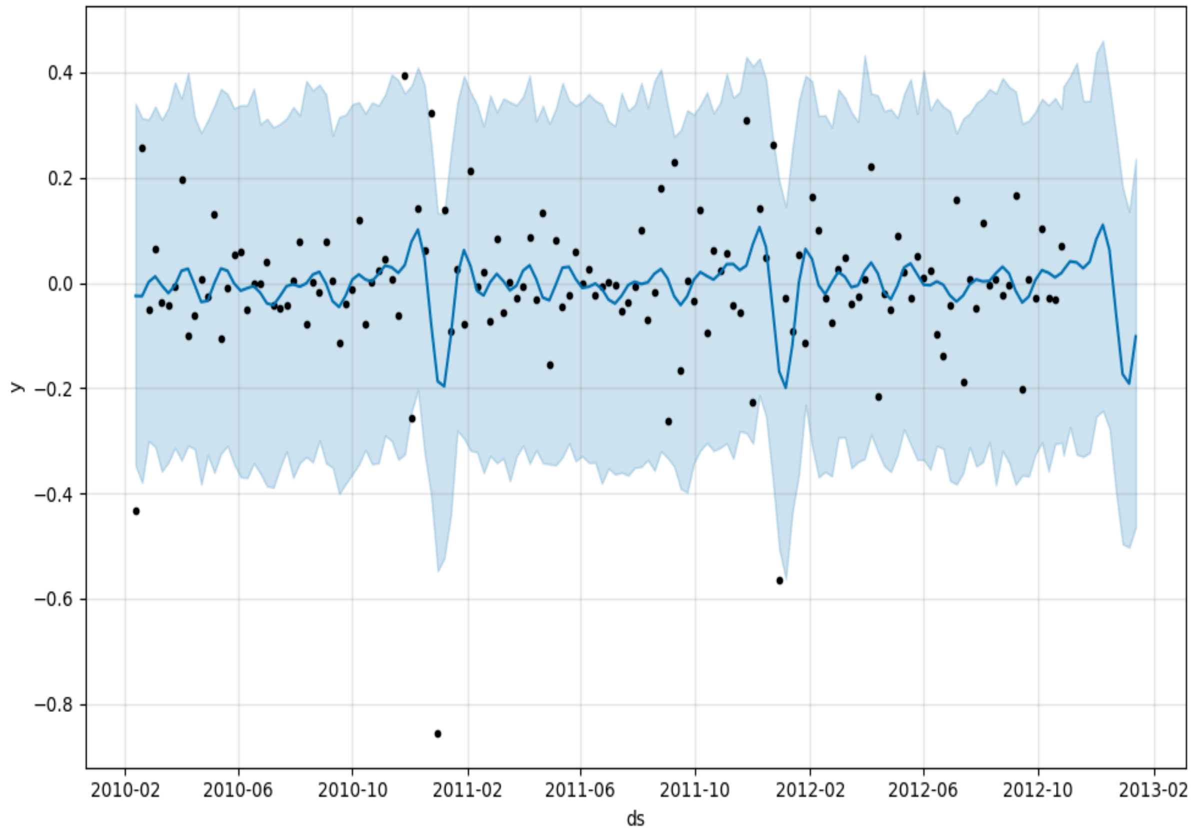
- From the seasonal decompose, a linear trend and seasonality has been observed.
- The data is stationary.
- We have used Confidence interval of 0.99



The RMSE value of **61761.98** suggests that, on average, the model's predictions deviate by approximately $61761.98.

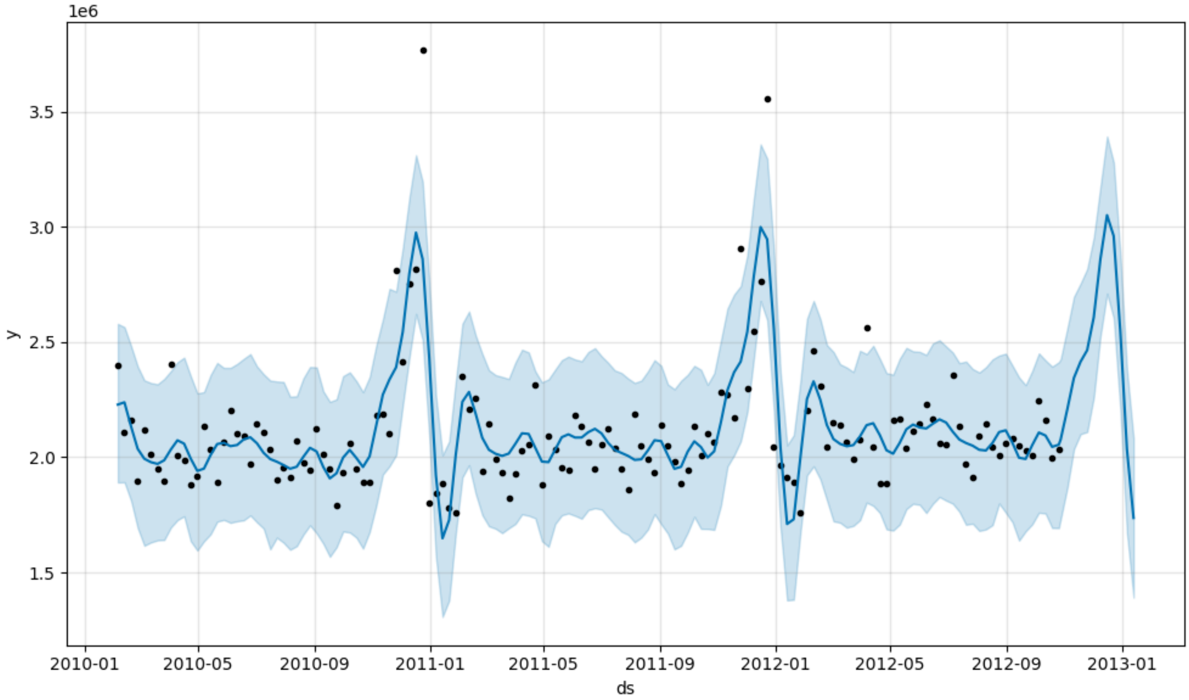
Predicting sales for Store No 14:

- For Store No 14, a downward trend has been observed with seasonality in the weekly sales.
- From the ad fuller test, we have observed that the data Is not stationary. Although the prophet model handles the non-stationary data well, we will still do the log transformation to check how our model performance.



Predicting Sales for Store No 20:

- For the sales of store, no 20, some spikes have been observed in the trend line.
- We have used confidence interval of 0.95.



The RMSE value of 102392.25 suggests that, on average, the model's predictions deviate by approximately $102392.25.

Inference of the Models:

The forecasting accuracy and precision of the model varies, as shown by the Root Mean Squared Error (RMSE) values of $63,888.54, $62,167.86, $61,761.98, and $102,392.25 for the projections of four distinct retail sales. The RMSE values of the first two stores are comparatively lower, suggesting a more accurate representation of sales trends. The third store performs consistently well, showing little variation from the model's projections. The fourth store's greater RMSE score, on the other hand, indicates that there may be difficulties in fully capturing every aspect of its sales patterns, requiring a more thorough investigation of certain factors that contribute. The first three store consistent and reduced RMSE results give rise to trust in the reliability of the model for the stores in question. The significant difference seen in the fourth store highlights the importance of customized model changes to tackle specific problems while improving precision. This analysis shows how constant monitoring and improvement are necessary to adjust the model to the distinctive characteristics of each store's sales behaviour, which will ultimately improve forecasting performance overall.

Conclusion:

In conclusion, the study of forecasted store sales presents a nuanced perspective on the model's predictive ability across a range of stores. The observed variation in accuracy highlights the fundamental complexity associated with sales forecasting in a retail industry. Although several retailers provide dependable and uniform forecasts, others pose difficulties, suggesting that an individual approach is required to account for different impacting variables. The results highlight how dynamic retail environments are and how crucial it is for forecasting models to be constantly modified in order to ensure optimal performance across various stores. The previously mentioned findings provide an improved understanding of the difficulties related to retail sales forecasting, hence influencing future improvements and adjustments aimed at improving the overall effectiveness of forecasting techniques in the retail sector.

References:

https://facebook.github.io/prophet/docs/quick_start.html

https://builtin.com/data-science/ols-regression

https://towardsdatascience.com/stationarity-in-time-series-a-comprehensive-guide-8beabe20d68

https://www.kaggle.com/code/bextuychiev/how-to-remove-non-stationarity-from-time-series