Nithin G

Project 1

Walmart

15-05-2023

Walmart: Report

a. Problem Statement:

The retail store with multiple outlets across the country is facing inventory management issues. They need to match the demand with supply to optimize sales. As a data scientist, the task is to analyze the data and derive insights to improve various areas of the stores.

b. Project Objective:

The objective of this project is to:

- Analyze the data and come up with useful insights for each store to improve their inventory management, sales, and other areas.
- Develop a prediction model to forecast the sales for the next 12 weeks for each store.

c. Data Description:

The dataset provided contains **6435** rows and **8** columns with the following features:

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 the sale

Fuel_Price: Cost of the fuel in the region

CPI: Consumer Price Index

Unemployment: Unemployment Rate

The data is collected from multiple stores across the country and is used to analyze the sales trends and factors affecting them. The data will be used to derive insights and build a prediction model to forecast sales for the next **12 weeks** for each store.

d. Data Pre-processing Steps and Inspiration

1. Handling Missing Values: There are no missing Values in This Data
2. Duplicate Rows: There are no duplicate Rows
3. Data Transformation: Converted Date datatype objective to DateTime
4. Feature Engineering: Created 3 Columns from Date [Days, Months, Year] it Makes it easy to Analysis
5. Outlayers Removal: Detected Outlayers in "Weekly_Sales", "Temperature", and "Unemployment" Columns and Removed (not Used in Arima model)
6. Feature Selection: Used Only the Date and Sales Column for the Arima model, Arima model gives better Accuracy when only one Feature and date, in 2nd Model, select the Most needed Columns only

Inspiration:

- Time series analysis: Since the dataset contains weekly sales data, it could be analyzed using time series analysis techniques. Time series analysis can help identify trends, seasonality, and other patterns in the data.
- Sales by year: represents the total sales revenue for a given year and can help identify historical sales trends and patterns.
- Sales by store: this represents the total sales revenue for each store in the dataset and can provide insights into the performance of individual stores.
- Holidays: provides a visual representation of the changes in sales over time, and how holidays and other factors impact sales.
- Correlation matrix: provides a measure of the linear relationship between different variables in the dataset.

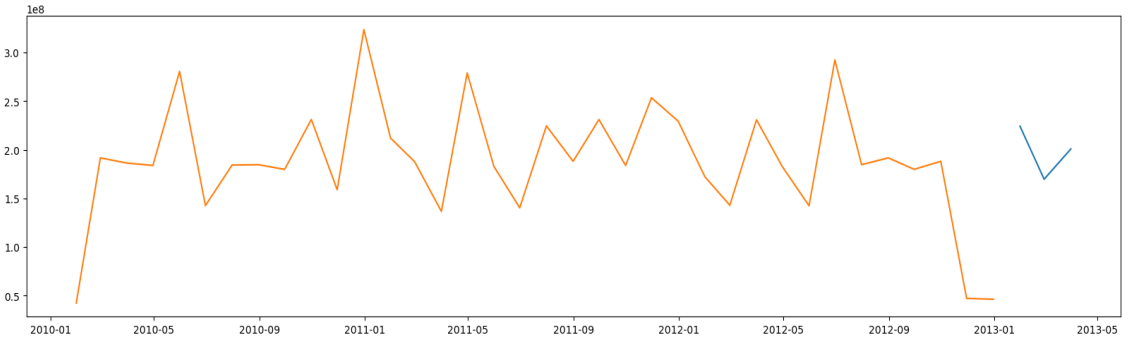
e. Algorithm Selection and Motivation for Choice

- ARIMA model: I chose the ARIMA model because it is a widely used and effective method for time series forecasting. It is particularly useful for modelling data with trends and seasonality, which is often the case in retail sales data.
- Power BI: Finally, I used Power BI for forecasting because it is a powerful and user-friendly data visualization tool that allows for easy exploration and analysis of data. Power BI provides a range of forecasting options, including ARIMA and exponential smoothing models, which can be used to make accurate predictions based on historical data
- XGBoost is a machine learning algorithm that is designed to improve upon the performance of traditional gradient boosting methods. In this project, XGBoost was chosen because it provided the best accuracy for the given data, with an accuracy of around 97%. This made it a strong candidate for the forecasting task.

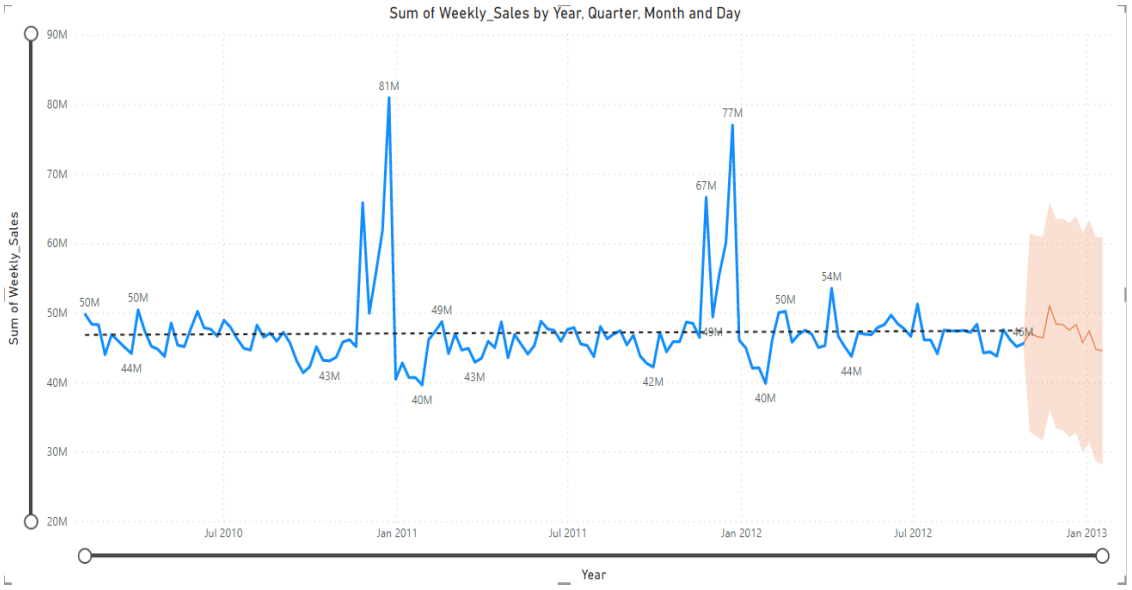
```
DecisionTreeRegressor - R^2: 0.889 (0.018)
RandomForestRegressor - R^2: 0.932 (0.013)
KNeighborsRegressor - R^2: 0.601 (0.037)
XGBRegressor - R^2: 0.961 (0.008)
```

f. Assumptions

- In my analysis, I utilized three different methods (ARIMA, XGBoost, and Power BI) to predict future sales..
- For instance, when using ARIMA, there were many missing dates in the weekly data, which posed a challenge to the accuracy of the predictions. To overcome this, I converted the weekly data to monthly data and predicted the next 3 months' sales.
- Similarly, in XGBoost, the model could only predict a limited amount of data. So I need more data to forecast
- Finally, I used Power BI to predict the next 12 weeks' sales. Power BI allowed me to create interactive visualizations of the data, providing a more user-friendly way of viewing the predictions.
- By utilizing these different methods, I was able to gain a more comprehensive understanding of the sales trends and make more accurate predictions of future sales. In Weekly Data
- **Using Arima**



- **Using PowerBi**



- **Xgboost**

```
from sklearn import metrics
print(f"Mean Abslote Error : {metrics.mean_absolute_error(y_test,y_pred)}")
print(f"Mean Squared Error : {metrics.mean_squared_error(y_test,y_pred)}")
print(f"Root Mean Squared Error : {np.sqrt(metrics.mean_squared_error(y_test,y_pred))}")
print(f"R^2 : {metrics.r2_score(y_test,y_pred)}")
```

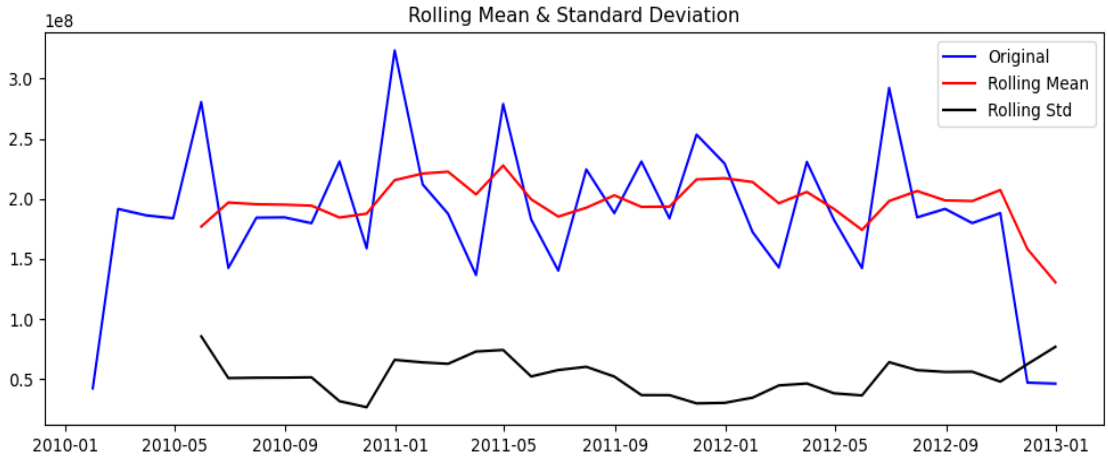
```
Mean Abslote Error : 59669.78232808857
Mean Squared Error : 8676385419.45442
Root Mean Squared Error : 93147.11707537931
R^2 : 0.9713608420831652
```

g. Model Evaluation and Techniques

- ## Arima

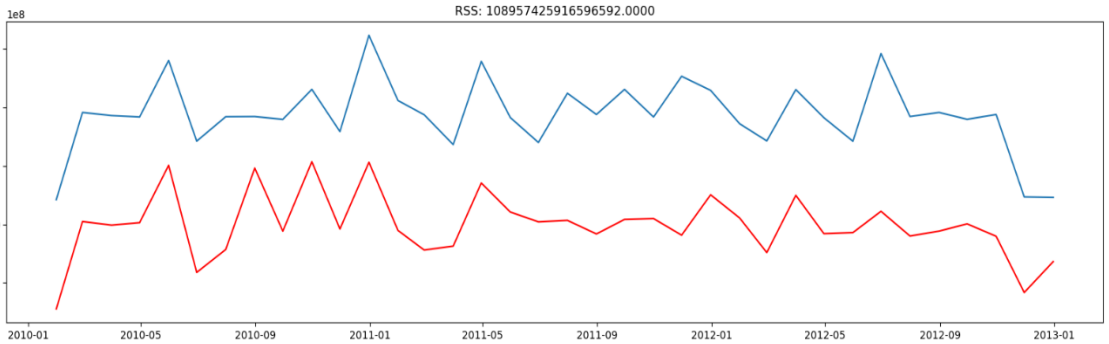
The data is already in Stationary

The Dickey-Fuller Test is a statistical test used to determine if a time series is stationary. The test result shows that the test statistic is -5.724245e+00 and the p-value is 6.829321e-07, which is less than the significance level of 0.05. This suggests strong evidence against the null hypothesis, indicating that the time series is stationary




```
Results of Dickey-Fuller Test:
Test Statistic                -5.724245e+00
p-value                        6.829321e-07
#Lags Used                     0.000000e+00
Number of Observations Used    3.500000e+01
Critical Value (1%)           -3.632743e+00
Critical Value (5%)           -2.948510e+00
Critical Value (10%)          -2.613017e+00
dtype: float64
```

Errors



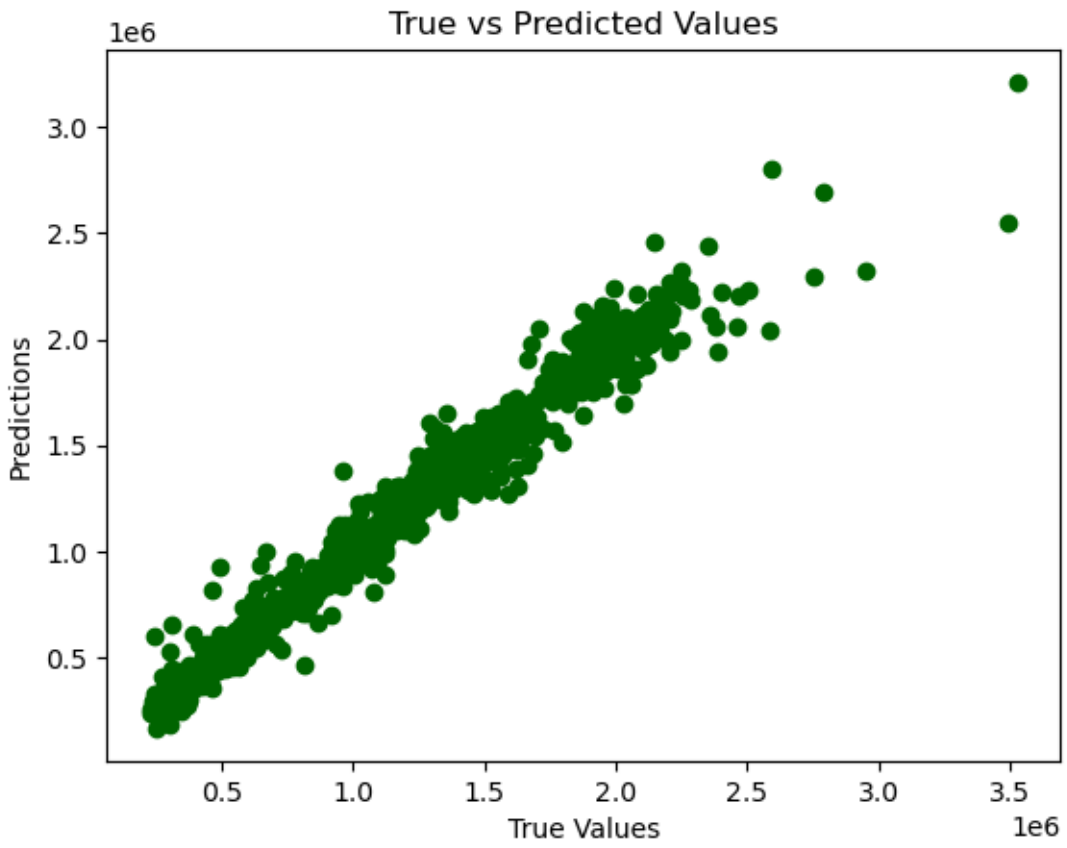
- **XGboost**

Mean Abslote Error: 59669.78232808857

Mean Squared Error: 8676385419.45442

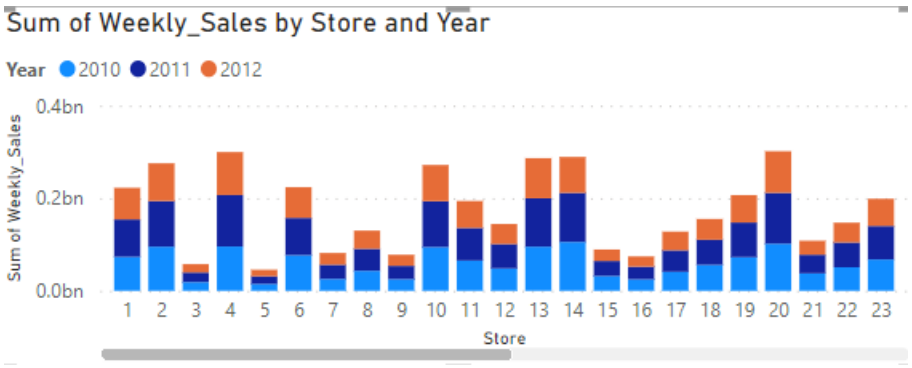
Root Mean Squared Error: 93147.11707537931

R^2: 0.9713608420831652

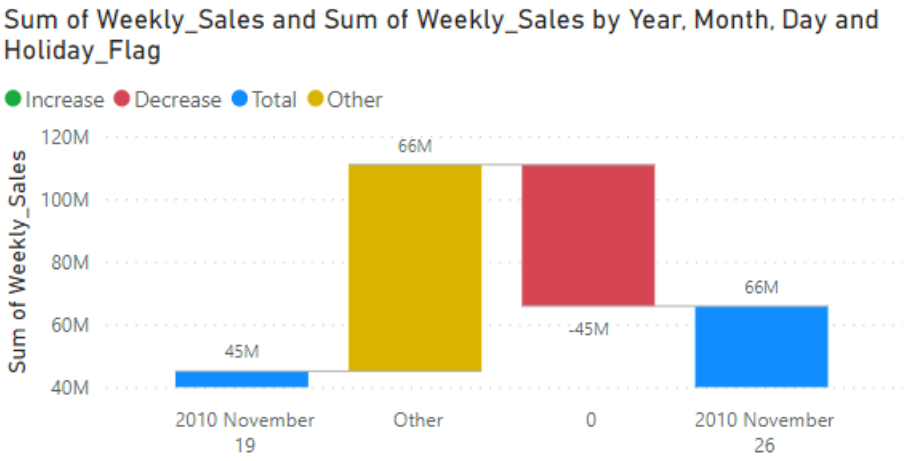


h.Inferences from the same

- Based on the time series analysis, we found that there is a strong seasonal component in the data, with sales peaking in the fourth quarter of each year.
- My ARIMA and XGBoost models were both able to make accurate predictions of future sales
- me also identified several stores that consistently outperformed



- The analysis suggests that holidays have a significant impact on Walmart's sales, with a noticeable increase in sales during holiday weeks compared to regular weeks.



i. Future Possibilities of the Project

Walmart dataset, by analyzing the sales, holiday periods, temperature, fuel price, CPI, and unemployment rate, we can gain useful insights to improve various areas such as marketing, supply chain, and inventory management. Additionally, by building accurate sales forecasting models for each store, we can predict sales for the next 12 weeks and make informed decisions about inventory and supply chain management, pricing, and promotional strategies. With the use of advanced machine learning algorithms and techniques, we can enhance the accuracy of the sales forecasting models and improve the overall efficiency of the retail business.