Capstone Project - Walmart Sales Analysis & Forecast

Table of Contents

Problem Statement

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. As a data scientist, I need to come up with useful insights using the data and make prediction models to forecast the sales for the next 12 weeks.

Project Objective

The objective of this project is to leverage data science techniques to analyze inventory and sales data from multiple outlets of a retail store, with the aim of accurately forecasting sales for the next 12 weeks. By examining historical sales patterns, external factors such as temperature, fuel prices, CPI, and unemployment rates, as well as the impact of holidays, the project seeks to develop robust predictive models that can anticipate future sales trends. Utilizing time series analysis and FB Prophet, the goal is to provide actionable insights that can inform inventory management, optimize stock levels, and enhance strategic decision-making. Accurate sales forecasts will enable the retail store, Walmart, to improve operational efficiency, reduce costs associated with overstocking or stockouts, and ultimately increase customer satisfaction by ensuring product availability.

Data Description

The dataset consists of 6,435 entries, each representing weekly sales data for Walmart stores. It contains 8 columns with various data types. This dataset provides comprehensive weekly sales information for 45 Walmart stores, including

factors such as holidays, temperature, fuel prices, CPI, and unemployment rates, which can be utilized for various analyses including sales trend analysis, impact assessment of external factors on sales, and forecasting.

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106
...	...	...	...	...	...	...	...	...
6430	45	28-09-2012	713173.95	0	64.88	3.997	192.013558	8.684
6431	45	05-10-2012	733455.07	0	64.89	3.985	192.170412	8.667
6432	45	12-10-2012	734464.36	0	54.47	4.000	192.327265	8.667
6433	45	19-10-2012	718125.53	0	56.47	3.969	192.330854	8.667
6434	45	26-10-2012	760281.43	0	58.85	3.882	192.308899	8.667

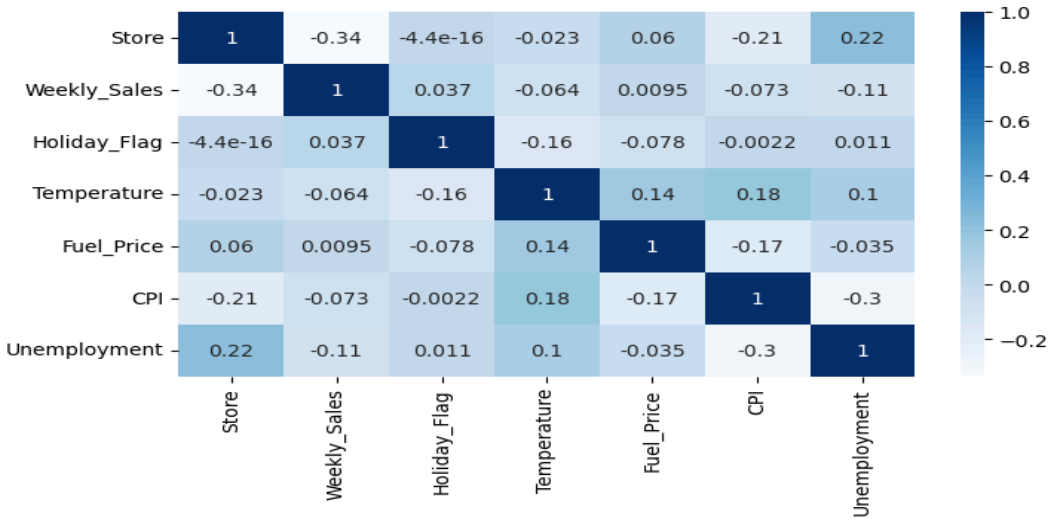
Data Preprocessing Steps And Inspiration

First step : Checking for missing and duplicate values in the dataset is a crucial step in ensuring data quality and reliability. Missing values can lead to biased estimates, reduced statistical power, and can significantly skew the results of the analysis and the performance of predictive models.

Second step : Data exploration is done to gain insight into the key features. Checking the correlation coefficient to get a better understanding of the degree of association between the target variable (weekly sales) and independent variables (fuel price, temperature,. Unemployment, etc.).

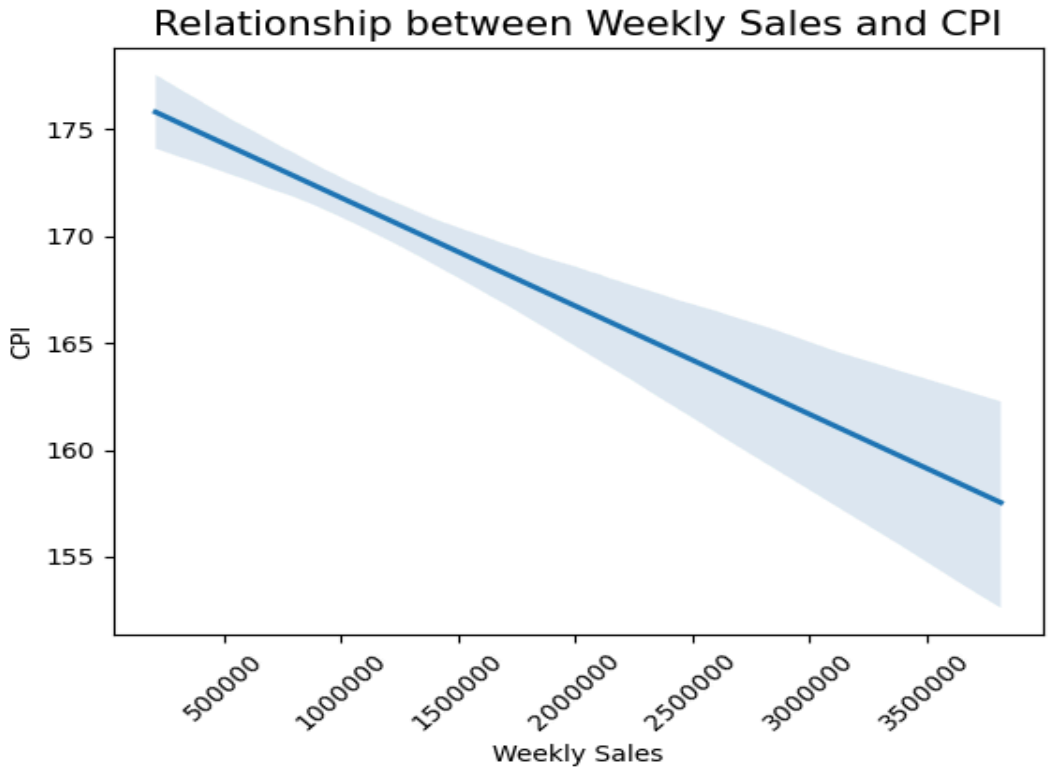
	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

Third step: In this step, I've used visualization techniques like heatmaps, barplots, and pie charts to gain insight into the data and the underlying patterns that impact the overall sales. With the help of the heatmap, we can explore relationships between variables, such as the impact of fuel prices and unemployment rates on weekly sales.



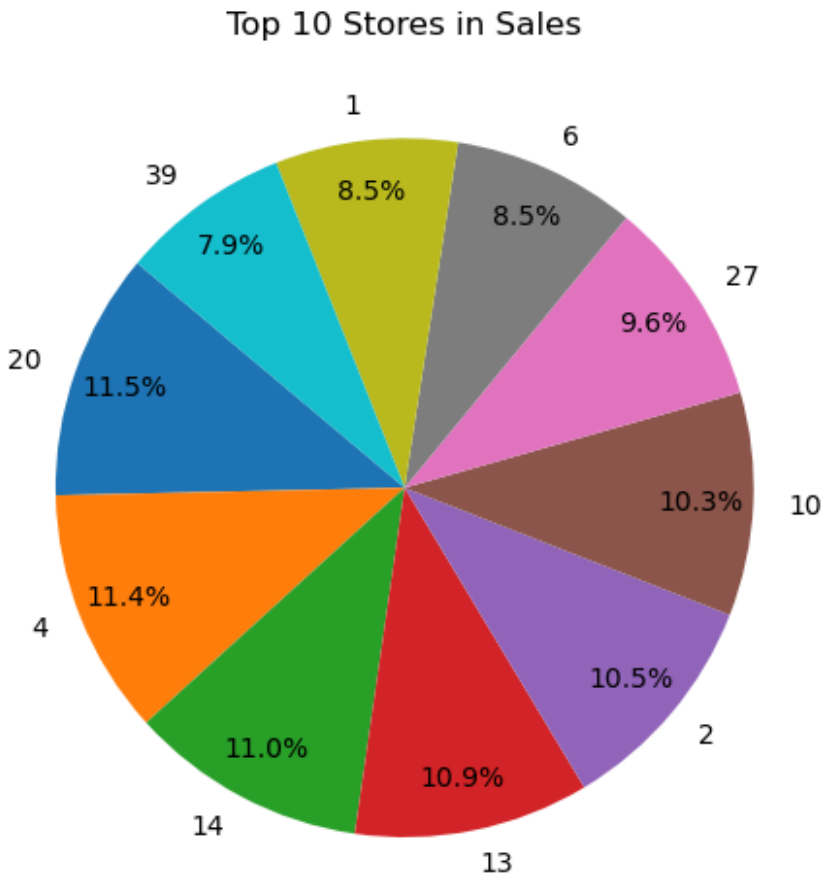
The barplot below depicts store-wise weekly sales to provide a clear visual representation of the sales performance across different outlets. Each bar represents the average weekly sales for a specific store, allowing for easy comparison of sales figures between 45 Walmart stores.

The graph below illustrates the relationship between weekly sales and the Consumer Price Index (CPI) using a scatter plot with a fitted regression line. The negative slope of the regression line indicates an inverse relationship between weekly sales and CPI. As weekly sales increase, the CPI tends to decrease.



The pie chart below illustrates the sales distribution among the top 10 performing Walmart stores. Each segment represents a store, with its size indicating the proportion of total sales attributed to that store. Store 20 leads with 11.5% of total sales, followed closely by Store 4 with 11.4%, and Store 14 with 11.0%. The top three stores collectively

account for nearly one-third of the sales among the top 10 stores. The remaining stores have a relatively even distribution, with sales contributions ranging from 7.9% to 10.9%.


Top 10 Stores in Sales
1
8.5%
6
8.5%
27
9.6%
10
10.3%
2
10.5%
13
10.9%
14
11.0%
4
11.4%
20
11.5%
39
7.9%

Choosing the Algorithm For the Project

For forecasting the weekly sales in the Walmart dataset, I selected time series analysis and FB Prophet as the primary algorithms. Time series analysis is well-suited for this dataset due to its sequential nature and the presence of potential seasonal patterns, trends, and cyclical components in the sales data. Time series models like ARIMA (AutoRegressive Integrated Moving Average) can effectively capture and predict these underlying patterns.

Additionally, FB Prophet, developed by Facebook, is chosen for its robustness and ability to handle missing data, seasonality, and holiday effects, making it a highly suitable algorithm for this dataset. FB Prophet's flexibility in incorporating seasonality and holiday effects, along with its simplicity in implementation and interpretability, makes it an excellent tool for forecasting the sales patterns evident in the Walmart data. Both approaches will provide a comprehensive analysis, leveraging their strengths to generate accurate and insightful sales forecasts.

Assumptions

The following assumptions were made in order to create the model for the Walmart Sales project.

To create the time series forecasting models for the Walmart sales dataset, several key assumptions were made. Seasonal patterns are presumed to be present and consistent year over year, allowing models like FB Prophet to capture and utilize these recurring patterns for accurate forecasting. It is assumed that external factors such as temperature, fuel price, CPI, and unemployment rates are adequately reflected in the dataset and their historical impact on sales can be leveraged to predict future sales. The Holiday_Flag variable is assumed to correctly identify weeks with significant holidays, impacting sales trends. Additionally, it is assumed that the data is free of significant errors or anomalies that could skew the model's predictions. Lastly, the models assume that past sales trends and patterns will continue into the future, providing a basis for forecasting based on historical data.

Model Evaluation and Technique

For building predictive models using time series analysis, I've employed ARIMA (AutoRegressive Integrated Moving Average) and FB Prophet algorithms. The first step in the ARIMA model building process involves checking the stationarity of the data, which is essential for reliable time series forecasting. Here I've used the Augmented Dickey-Fuller (ADF) test to determine the stationarity of the data.

```python
from statsmodels.tsa.stattools import adfuller  #To perform an augmented Dickey-Fuller test

store_seasonality = [] #Initializing an empty list to store the seasonality check results for each store

for store in range(1,46):
  store_data = df[df['Store']==store]
  result = adfuller(store_data['Weekly_Sales']) #Performing the Augmented Dickey-Fuller (ADF) test on the 'Weekl
  adf_statistic = result[0]
  p_value = result[1]
  store_seasonality.append({'store_no.':store, 'adf_statistics':adf_statistic, 'p-value' : p_value })

for result in store_seasonality:
    store = result['store_no.']
    adf_statistic = result['adf_statistics']
    p_value = result['p-value']

    print(f"store_no. {store}:")
    print("adf_statistics:", adf_statistic)
    print("p-value:", p_value)
    if p_value < 0.05: #Interpreting the p-value to determine if the time series is stationary
        print("The time series is likely stationary (no seasonality).")
    else:
        print("The time series may not be stationary (may have seasonality).")
    print()
```

The ARIMA model parameters (p, d, q) were then selected to optimize model performance. After the deployment, I evaluated the performance of the model using techniques like R-squared score, MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error).

```
R-squared (R2) Score: -0.49
Mean Absolute Error (MAE): 50908.04
Mean Squared Error (MSE): 3179038765.63
Root Mean Squared Error (RMSE): 56382.97
```

Here’s what the evaluation matrix for the ARIMA model illustrates :

- The **R-squared** value suggests that the model is failing to capture any useful information from the input features to explain the variance in the target variable.
- The **MAE** indicates that, on average, the model's predictions are off by 50,908.04 units. This is a substantial error, suggesting that the model's predictions are far from the actual values.
- The **MSE** is extremely high, indicating that the model's predictions have large errors. Since MSE squares the errors, it penalizes larger errors more heavily, showing that there are significant deviations between the predicted and actual values.

In parallel, FB Prophet was utilized. Prophet automatically detects and models seasonal effects and trends, and its flexible framework allows for the inclusion of external factors, such as holidays, which are crucial for retail sales forecasting.

Evaluating the FB Prophet model for Store no. 15 yielded the below results:

```
R2_score of the model is 0.9387964630665365
Mean Absolute Percentage Error (MAPE) is 3.49803880230234
Mean Absolute error is 21970.742057913067
```

Here’s how the FB Prophet model performed:

- An **R-squared** value of 0.939 means that the model explains 93.9% of the variance in Walmart's sales data. This high R-squared value suggests that the Prophet model fits the sales data very well and captures the underlying patterns effectively.

- A **MAPE (Mean Absolute Percentage Error)** of 3.498% indicates that, on average, the model's sales forecasts are off by approximately 3.498% from the actual sales values. This low percentage error suggests that the model's forecasts are highly accurate, which is crucial for planning and inventory management in Walmart stores.
- The **MAE (Mean Absolute Error)** indicates that, on average, the model's sales forecasts are off by 21,970.74 units. While this absolute error might seem large, it is important to consider the scale of Walmart's sales. Given the high R-squared value and low MAPE, this MAE suggests that the model is performing well and the forecasts are reliable.

Inferences from the Project

The inferences from the model evaluation indicate that the time series analysis using ARIMA did not yield a satisfactory R^2 score, suggesting that the model struggled to accurately capture and predict the variability in the weekly sales data. This could be due to multiple reasons like the complexity of the sales patterns, the presence of multiple external factors, or inherent non-stationarities in the data that ARIMA failed to fully address.

On the other hand, the FB Prophet model produced a good R^2 score. It demonstrates superior ability to fit the data and forecast future sales accurately. This highlights the importance of choosing FB Prophet as a powerful tool for retail sales prediction.

Walmart can leverage machine learning to identify factors that influence sales. By incorporating the machine learning algorithm, the renowned multi-channel retailer can enhance the accuracy of their sales predictions. This can further allow them to enable better inventory management and strategic planning for great overall performance of their outlets.

Future Possibilities

The future possibilities for this project are extensive. They can significantly enhance Walmart's operational efficiency and strategic decision-making. Firstly, the integration of more granular data, such as hourly sales figures or customer foot traffic, can improve the accuracy of sales forecasts. Advanced machine learning algorithms, including neural networks and ensemble methods, can be explored to capture complex patterns. Expanding the scope to include supply chain optimization, demand forecasting, and pricing strategy adjustments based on predictive insights can further streamline operations and reduce costs. Finally, developing a user-friendly dashboard for real-time visualization and monitoring of sales forecasts and key performance indicators (KPIs) can empower managers of different outlets at all levels to make informed decisions. These enhancements will help Walmart to anticipate market changes proactively, optimize inventory levels, and meet customer demand more effectively.

Conclusion

In this project, my goal was to analyze the data of the 45 Walmart stores and forecast the sales over a period of 12 weeks. Through analysis and modeling I was able to provide several valuable insights and demonstrate the effectiveness of Facebook Prophet in generating accurate sales forecasts.

Some of the key insights from the data:

- Each store exhibited unique sales patterns, influenced by location-specific factors such as regional events, local economic conditions, and competition.
- Sales were notably impacted by external events such as holidays, weather conditions, and economic trends. Understanding these impacts helped in better interpreting the sales patterns.

The Facebook Prophet model proved to be an effective tool for forecasting sales at Walmart stores. Its ability to handle seasonality, holidays, and external events made it particularly suitable for our sales data. The model's high accuracy and reliability provided us with trustworthy forecasts, aiding in strategic decision-making

for inventory management, promotional planning, and resource allocation.

References

1. https://chat.openai.com/
2. https://www.optisolbusiness.com/insight/top-5-machine-learning-techniques-for-sales-forecasting
3. https://intellipaat.com/blog/time-series-analysis/
4. https://medium.com/@russelosiemo/facebook-prophet-sales-forecasting-5b7caf4dcbaa#:~:text=The%20Prophet%20model%20is%20a,website%20traffic%2C%20or%20sales%20data.