Walmart Sales Forecasting

Riya Singh

MDA 820: DATA-DRIVEN DECISION MAKING

Prof. Gurpreet Singh

December 18, 2024

Abstract

Sales forecasting is a critical tool for inventory management and strategic decision-making in retail operations. This project aims to develop a predictive model for weekly sales of Walmart, a major chain offering different products. Using a diverse dataset that includes sales records, economic indicators, and event calendars from Kaggle, this project leverages advanced data mining techniques and machine learning algorithms to achieve reliable forecasts. The project incorporates, holiday impacts, temperature, fuel price fluctuations to enhance accuracy. The findings of this project aim to optimize inventory management, reduce waste, and improve the retailer's ability to anticipate customer demand.

Problem Statement

Accurately predicting weekly sales across its many product and store locations is a major difficulty for Walmart, a well-known supermarket retailer. The dynamic nature of consumer demand, which is impacted by several variables like shifts in the economy (such as changes in the price of fuel), and seasonal occurrences like holidays, is the source of these difficulties. Inadequate sales forecasting results in less-than-ideal inventory control, which either causes stock outs that lower customer satisfaction or surplus inventory that raises holding costs. A data-driven forecasting solution is required to successfully predict sales trends and support important choices in supply chain management, inventory planning, and promotion scheduling.

Introduction

Accurate sales forecasts are crucial for retail companies to maintain operational efficiency, reduce costs, and improve customer satisfaction. Poor forecasting can lead to overstocking, stock outs, and missed revenue opportunities.

The Project addresses these problems by developing a robust forecast model adapted to the needs of Walmart. Using data from a Kaggle competition, the project aims to forecast weekly sales at the item-store level while taking into account time series dependencies and feature interactions.

The project provides a comprehensive, data-driven solution to enable them to make informed decisions on inventory management and marketing strategies, reducing costs and improving customer service quality.

Background

The dynamic retail environment in which Walmart operates is impacted by both domestic and foreign variables. The business can be impacted by changes in fuel prices, holidays and seasonal events influence consumer behavior and sales trends. Managing inconsistent data, combining many data sets for a thorough analysis, and figuring out intricate, nonlinear relationships between variables are some of the main obstacles.

This initiative makes use of cutting-edge data mining and predictive analytics tools to address these issues. Machine learning models like XGBoost, LSTM, are used for accurate predictions and exploratory data analysis (EDA) is used for pattern detection.

.

Objectives

This project aims to address the challenges with these objectives:

1. **Accurate Sales Forecasting:**
   Develop predictive models to forecast daily item-store sales, considering features like holidays, promotions, and oil price fluctuations.
2. **Demand Classification:**
   Classify items into sales buckets to identify demand patterns and analyze the impact of promotions and holidays.
3. **Feature Impact Analysis:**
   Evaluate the influence of key features (e.g., holidays, fuel prices, temperature) on sales trends and demand.
4. **Optimization and Insights:**
   Use forecasting and classification results to optimize inventory planning, reduce waste, and align marketing strategies with demand.
5. **Robust Machine Learning:**
   Leverage advanced machine learning models, including XGBoost, and ARIMA, for precise predictions and actionable insights.

Data Collection & Engineering

Data Source

The Kaggle Walmart Sales Forecasting competition provided the dataset for this project, which included comprehensive data on weekly sales, stores, departments, holidays, temperature, and store metadata. Store locations and department were provided via additional datasets such as stores.csv.

Data Cleaning, pre-processing and transformation

Multiple csv files are merged using common keys, such as dates and store numbers, for data cleaning and merging. By confirming types and structures, it guarantees data consistency and imputing them according to their relevance. Here how different csv data looks before merging into one.

This is what the 3 tables look like:

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

```
df_train.head()
```

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

```
df_features.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

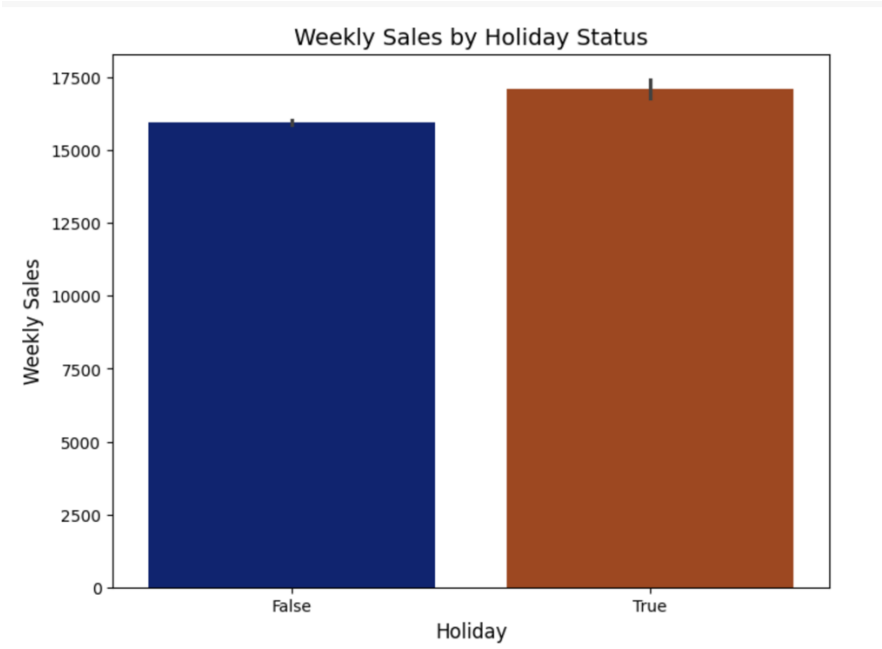
Exploratory Data Analysis (EDA)

A deeper comprehension of sales patterns and trends was made possible by descriptive statistics, which offered insightful information on the distribution and structure of the dataset. Important elements, such as weekly sales, holiday dates, temperature and fuel prices were measured. This research laid the groundwork for further modeling by highlighting the data's diversity and variability.

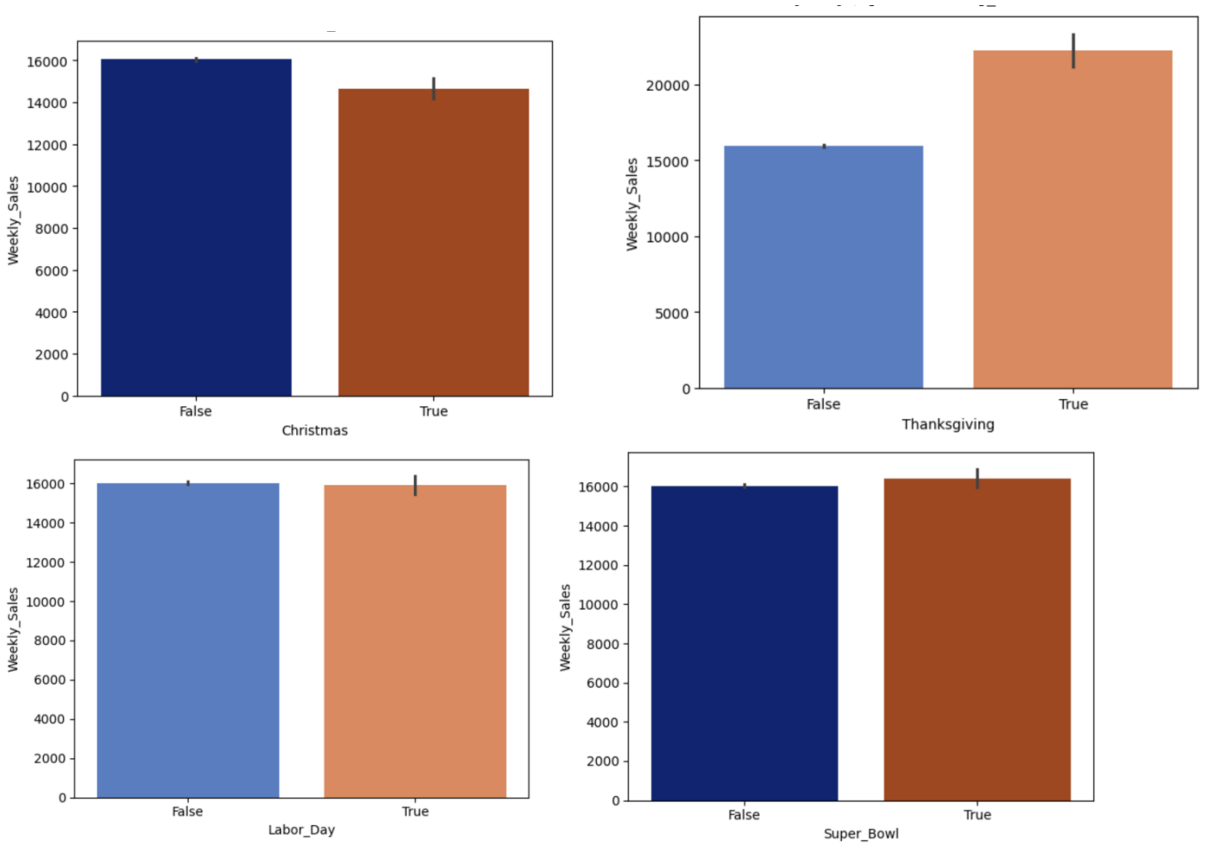
The impact of holidays was plotted in graphs by grouping the data according to the holidays and then measuring the impact it puts on the weekly sales as below. Overall it does show that the sales increase during holidays.

Data Visualization

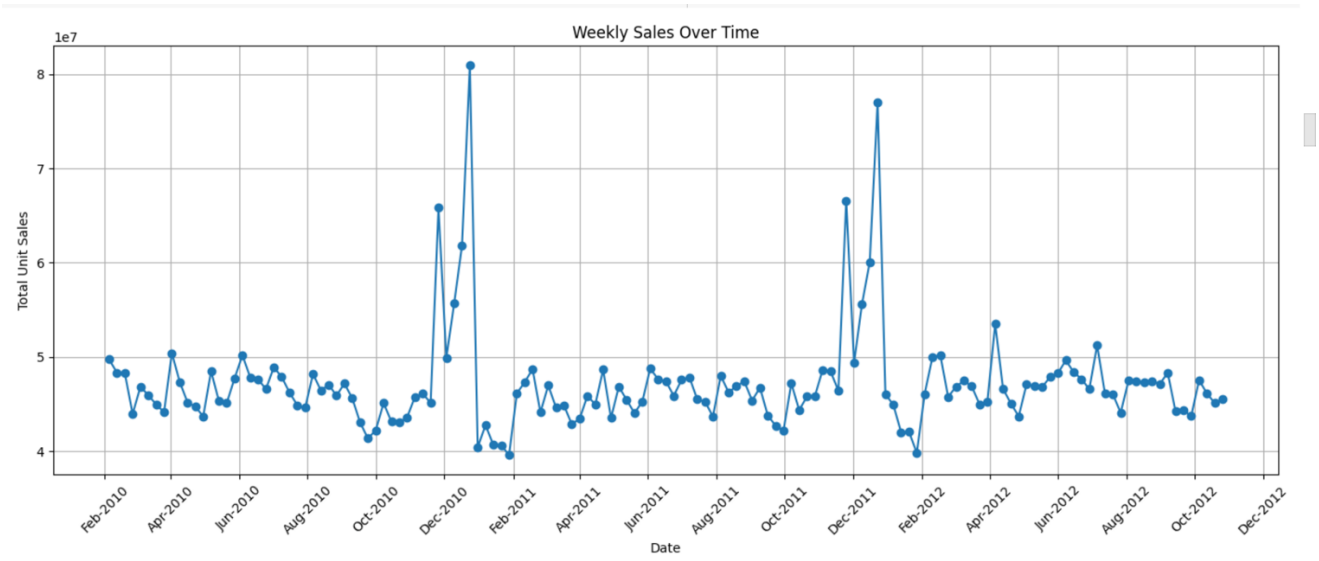
Visualizing data during Exploratory Data Analysis (EDA) helps identify trends, detect anomalies or outliers, and understand relationships between key variables, such as sales, holidays, and other characteristics. Following are the data visualization performed during EDA :



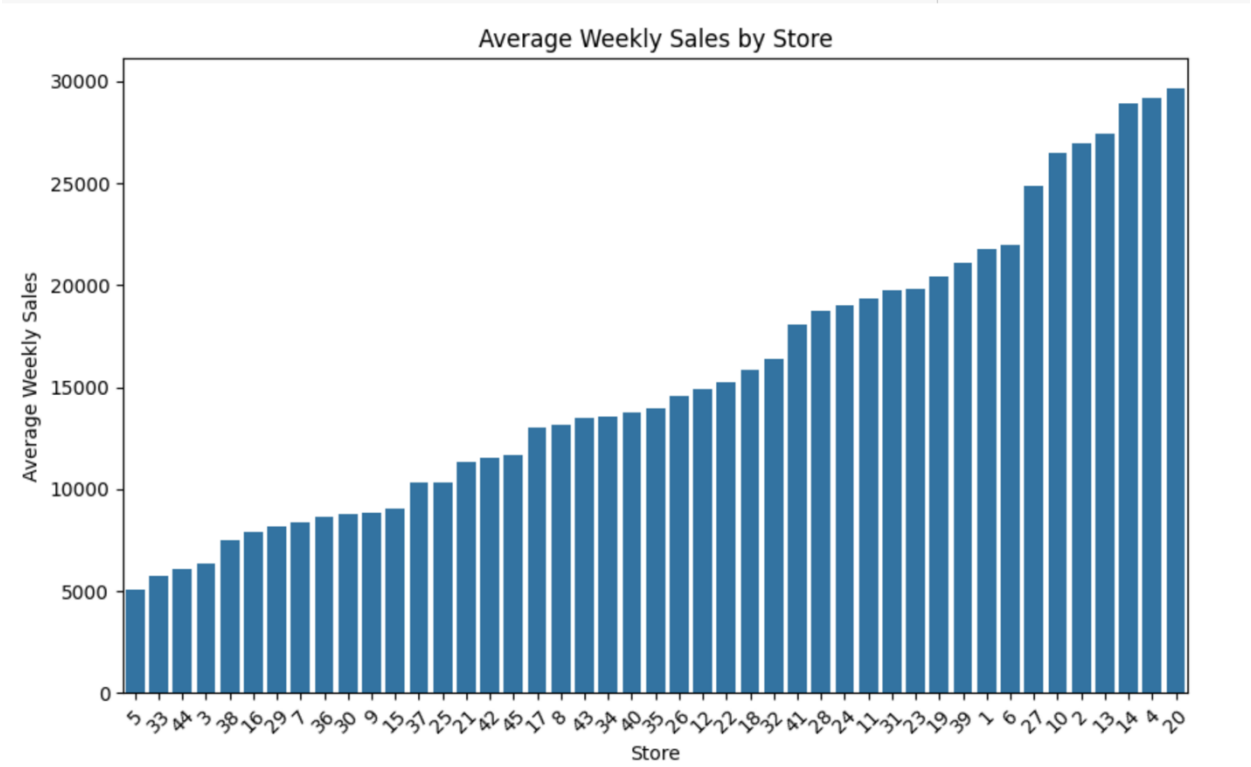
There also is a comparison with individual holiday types as Super bowl, Thanksgiving, Christmas and Labor Day. If these impact the Sales a lot individually.



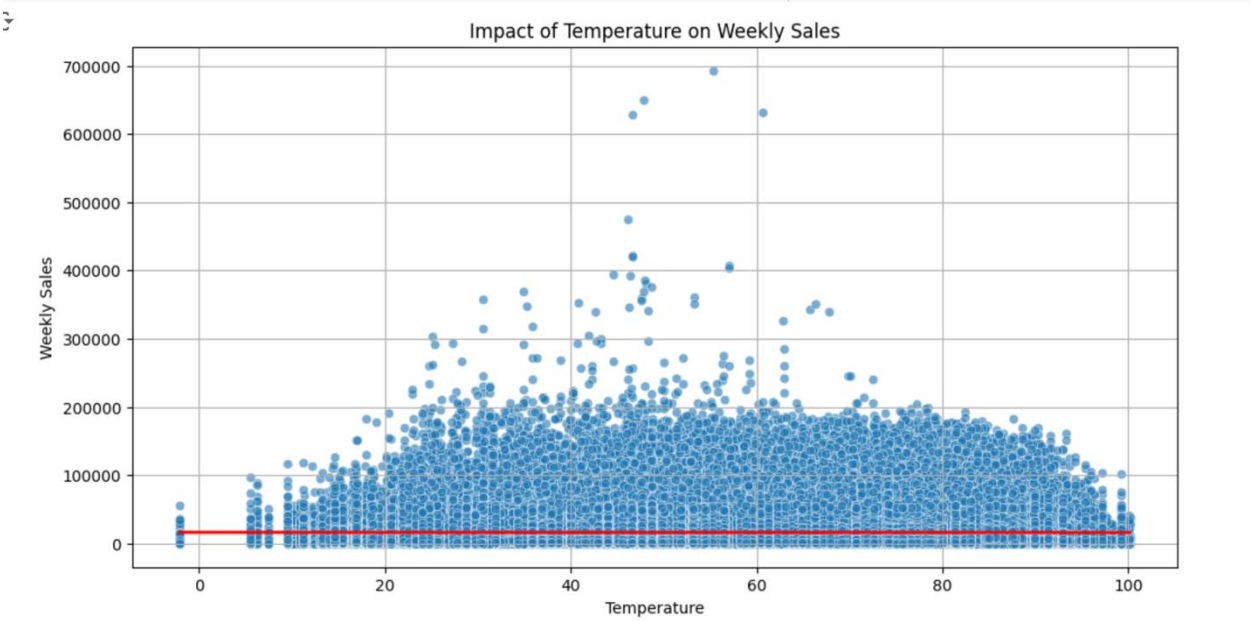
Weekly Sales over time:



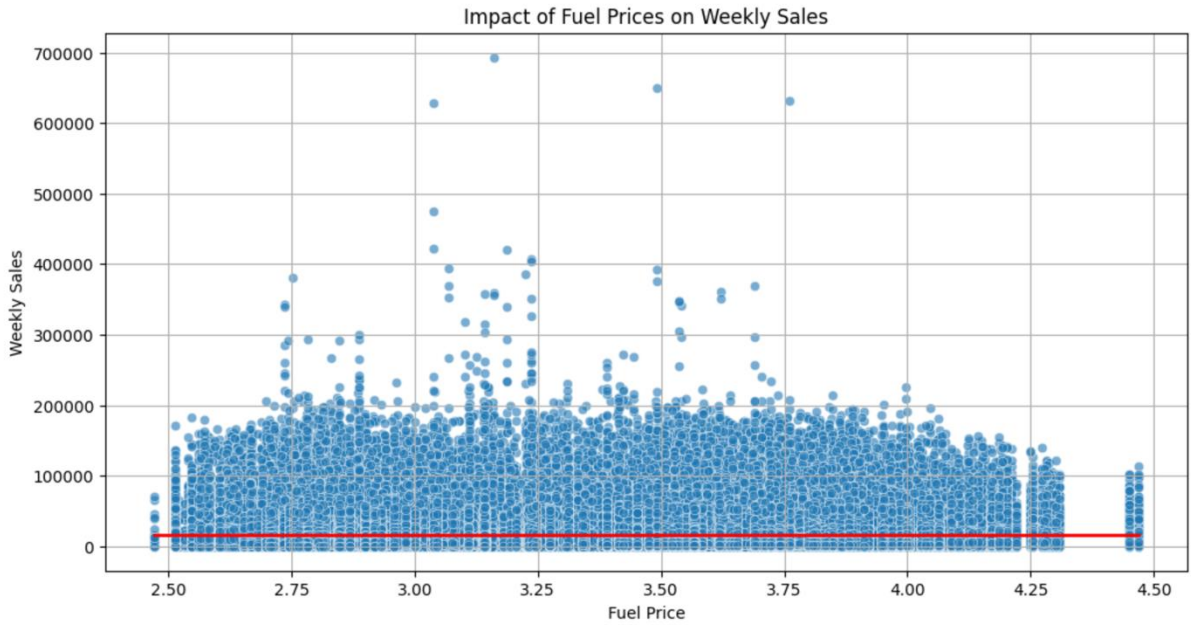
Average weekly sales by each store:



Impact of temperature on the Sales:



Impact of Fuel Prices on the Sales:

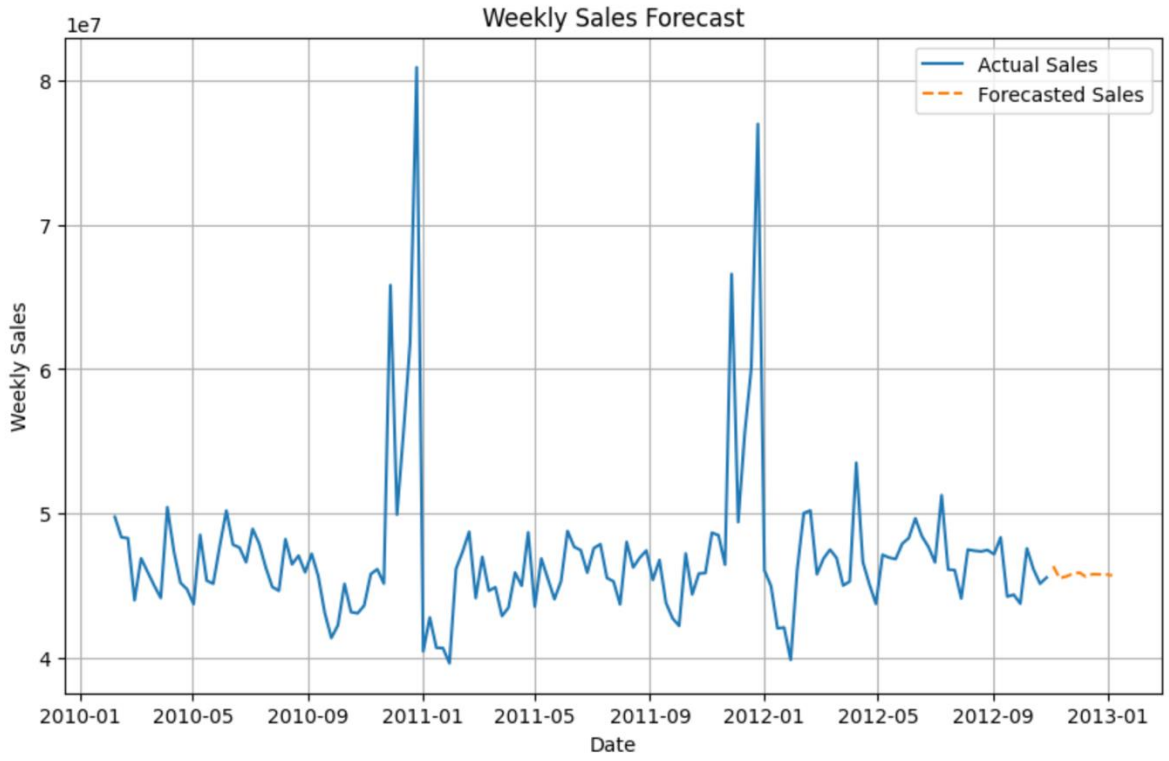


Model Building

1. **ARIMA MODEL :** The **ARIMA model** (AutoRegressive Integrated Moving Average) is a popular statistical tool used for forecasting time series data.

   Why ARIMA for Walmart?

- Accuracy: Captures patterns like seasonality (holiday sales) and trends (growth over time).
- Planning Tool: Helps Walmart predict demand, reduce overstock or stockouts, and improve customer satisfaction.



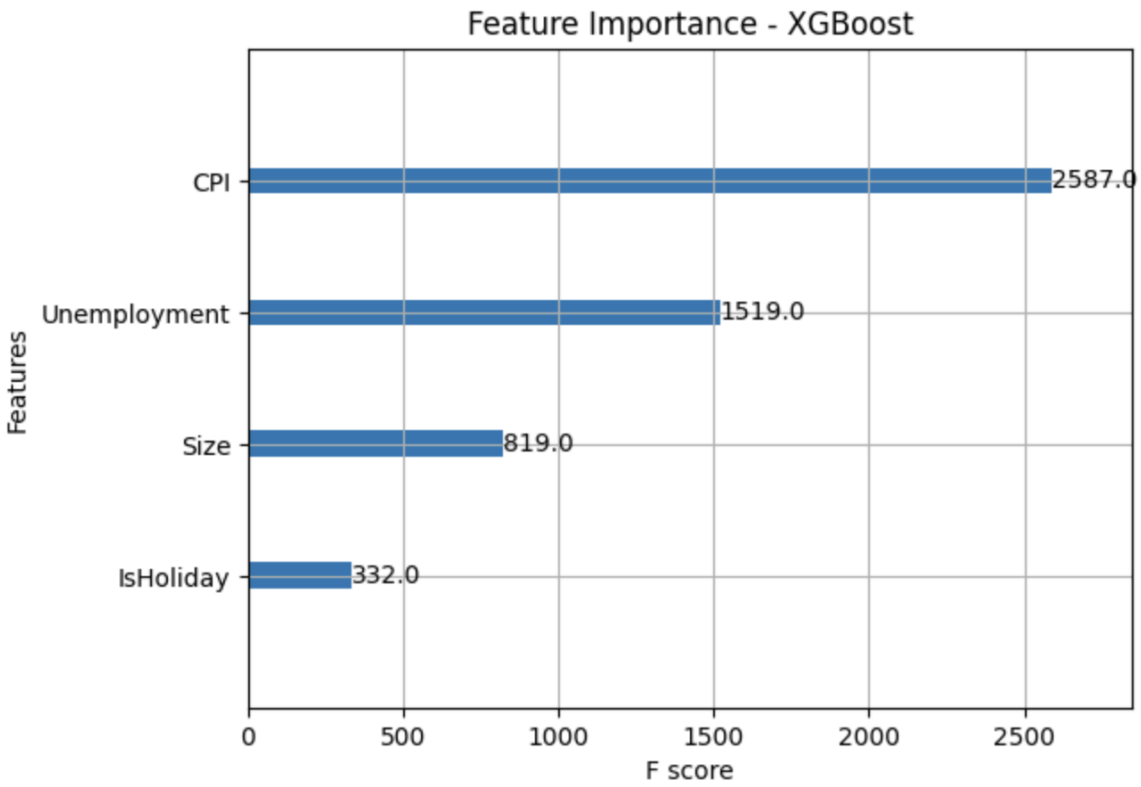
The ARIMA Model highlights spikes, dips, and fluctuations, including seasonal peaks (e.g., during the holiday season or sales events) and the orange line here shows the future sales prediction showing the average regular sales in the non-holiday season.

2. **XGBoost Classifier :** XGBoost is a powerful machine learning model that combines multiple decision trees to create a stronger model, helping improve classification accuracy. It works by building trees one after another, where each tree tries to fix the mistakes of the previous one. XGBoost is very efficient and can handle both simple and complex patterns in the data..

Feature importance visualizations

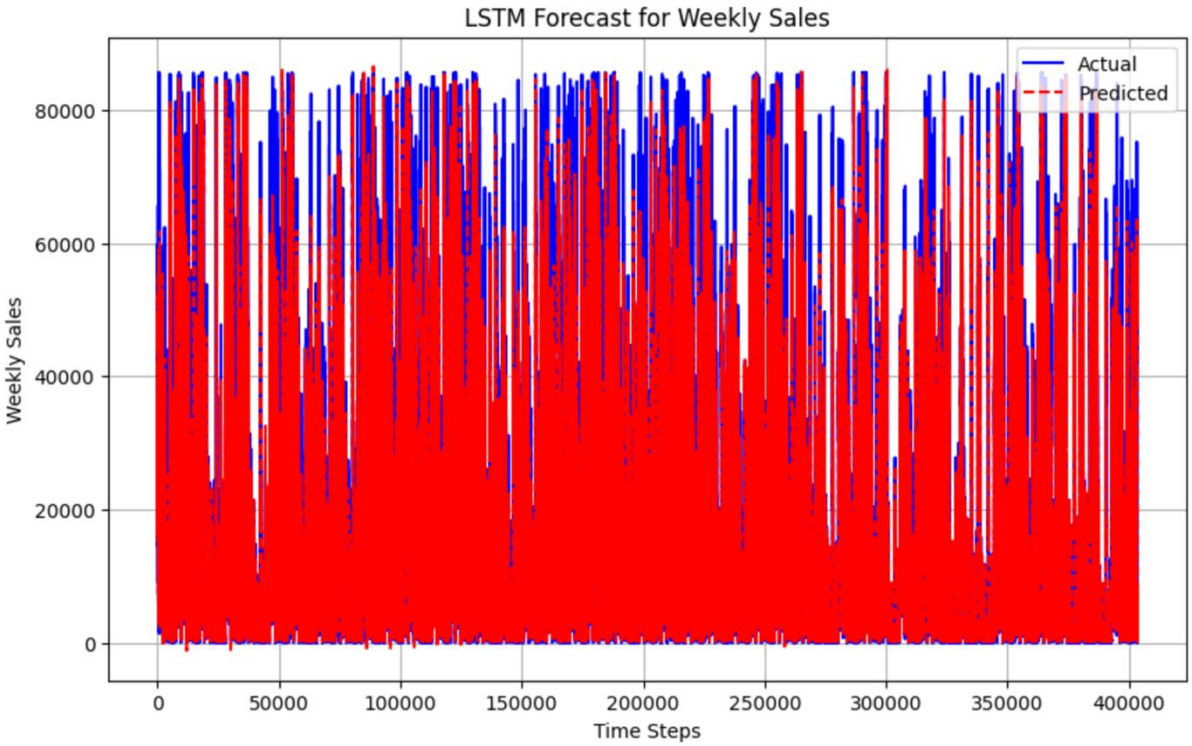
The feature importance visualizations for XGBoost classification models in the Sales Forecasting project highlight which features most significantly influence the prediction of sales behavior, providing a clear understanding of how these features contribute to the model's ability to classify sales into different buckets.

```
XGBoost - Mean Squared Error: 268758640.55
XGBoost - R² Score: 0.08
<Figure size 1000x600 with 0 Axes>
```



3. **LSTM (Long Short-Term Memory):** LSTM (Long Short-Term Memory) is a type of AI model designed to handle **time series data** by learning patterns over time. It's great at remembering important trends and forgetting irrelevant ones, making it perfect for forecasting things like sales, where past data affects future outcomes. Here, LSTM can predict future Walmart sales by analyzing weekly trends, spikes during holidays, and gradual increases or decreases.

- The red predictions mostly overlap with the blue actual sales, indicating the LSTM model has captured the overall pattern reasonably well.
- However, there might be areas where the model struggles to perfectly match sharp spikes or sudden changes, which can happen due to the complexity of sales patterns.



Summary of Key Findings and Results

Have successfully built both predictive and classification models to effectively forecast the unit sales per store and item level.

The integration of economic indicators, seasonal factors, and promotional effects enhances the predictive capabilities, providing actionable insights that align with business objectives. Furthermore, the development of visualizations ensures that stakeholders have access to real-time forecasts and trends, improving decision-making efficiency.

Overall, this project not only demonstrates the potential of data-driven forecasting to reduce costs and improve operational efficiency but also lays the groundwork for future enhancements in predictive analytics for retail.

Conclusion

This project demonstrated how factors like fuel prices, holidays, and temperature significantly influence sales at Walmart. The models revealed that holidays and CPI drive major sales changes, highlighting the need for better inventory planning during these periods. Economic indicators, such as fuel prices, also showed their impact on sales trends but not to a high extent. By incorporating these factors, the project provided accurate forecasts and actionable insights to optimize inventory, reduce waste, and align marketing strategies with customer demand.

Future work

This project can be expanded by capturing seasonal patterns and long-term trends with a larger dataset that covers several years. Using real-time data such as live fuel prices and updated promotional schedules can improve the relevance of forecasts. Advanced models such as long-term memory networks (LSTMs) can be studied to better handle time dependencies. Automating the data pipeline for preprocessing and model training will improve scalability, while improved dashboards with scenario analysis and predictive visualization will provide stakeholders with actionable insights. In addition, with appropriate computational resources, this approach can be extended to forecast sales of any large food chain, making it a versatile solution for the retail industry.

GitHub Link: https://github.com/Riyasingh27/Walmart-Sales.git