# Capstone Project - Walmart (Retail Sales Forecast)

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

# 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.

**Dataset Information:**

**The walmart.csv contains 6435 rows and 8 columns.**

|  |  |
| --- | --- |
| **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 the sale |
| Fuel\_Price | Cost of the fuel in the region |
| CPI | Consumer Price Index |
| Unemployment | Unemployment Rate |

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:
2. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
3. If the weekly sales show a seasonal trend, when and what could be the reason?
4. Does temperature affect the weekly sales in any manner?
5. How is the Consumer Price index affecting the weekly sales of various stores?
6. Top performing stores according to the historical data.
7. The worst performing store, and how significant is the difference between the highest and lowest performing stores.
8. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

Walmart operates multiple outlets across the country but faces challenges in effectively managing inventory to match supply with demand. The lack of precise sales forecasting is leading to stockouts and overstock situations, impacting the overall efficiency and profitability.

The goal of this project is to analyze historical sales data to uncover trends and patterns, determine key factors affecting sales, and forecast sales for the next 12 weeks for selected stores.

# Project Objective

The primary objective of this project is to forecast weekly sales for 45 retail stores of a retail chain using time-series forecasting techniques. This involves analyzing historical sales data to identify patterns, trends, and factors influencing sales performance. By leveraging statistical models and machine learning approaches, the project aims to:

**Analyze Historical Sales Data:**

Perform exploratory data analysis (EDA) to understand sales trends, seasonality, and key influencing factors such as temperature, holidays, consumer price index (CPI), fuel prices, and unemployment rate.

**Evaluate Store Performance:**

Identify top-performing and underperforming stores based on historical sales data.

Highlight factors affecting the sales performance of individual stores.

**Develop Forecasting Models:**

Build and compare multiple forecasting models, including ARIMA, SARIMA, Holt-Winters Exponential Smoothing, Random Forest Regressor, and Support Vector Regression (SVR), to predict weekly sales for the next 12 weeks.

**Generate Business Insights:**

Provide actionable insights to help the retail chain optimize inventory management and align supply with demand.

Identify seasonal trends or external factors impacting sales, enabling data-driven decision-making.

**Visualize and Present Results:**

Create clear and interpretable visualizations of forecasted sales for selected stores.

Compare the performance of different forecasting methods to select the best model for future predictions.

This project aims to demonstrate the use of data analytics and predictive modeling in solving real-world business problems while showcasing proficiency in data science tools, techniques, and methodologies.

# Data Description

The dataset available is **Walmart.csv**The dataset contains **6,435** records with **8 columns**, representing Walmart store sales data over time. Below is a summary of the attributes:

* **Store**: Store number (integer).
* **Date**: Sales date (DD-MM-YYYY format).
* **Weekly\_Sales**: Weekly sales revenue for the store (float).
* **Holiday\_Flag**: Indicates whether the week includes a holiday (1 = Yes, 0 = No).
* **Temperature**: Average weekly temperature in the region (float).
* **Fuel\_Price**: Cost of fuel per gallon (float).
* **CPI**: Consumer Price Index for the region (float).
* **Unemployment**: Unemployment rate in the region (float).

# Data Preprocessing Steps And Inspiration

The preprocessing of the data included the following steps:

1. **Handling Missing Values**: Imputed missing values using interpolation and statistical methods.
2. **Outlier Detection**: Identified and handled sales outliers using boxplots and z-scores.
3. **Date Transformation**: Converted the Date column into a datetime format and extracted additional features like year, month, and week.
4. **Normalization**: Scaled numerical features for machine learning models.

Challenges included managing seasonal patterns and identifying meaningful trends amidst noisy data.

# Choosing the Algorithm For the Project

Description for the XYZ algorithm for the project.

The following forecasting methods were selected:

1. **ARIMA**: Best for linear trends and stationary data.
2. **SARIMA**: Useful for seasonal patterns in time series.
3. **Holt-Winters**: Effective for data with trends and seasonality.
4. **Random Forest Regressor**: A machine learning approach for non-linear patterns.
5. **Support Vector Regression (SVR)**: Captures complex relationships and trends.

### ****Motivation and Rationale for Method Selection****

* **Holt-Winters** and **SARIMA** handle seasonality and trends effectively.
* **ARIMA** is simple yet effective for stationary data.
* **Random Forest Regressor** and **SVR** offer non-linear modeling capabilities, making them versatile for complex datasets.

# Assumptions

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

1. Sales trends from historical data remain consistent for future forecasts.
2. Holiday impacts on sales are similar to previous patterns.
3. Regional variables like temperature and unemployment are key influencing factors.

# Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the model

The following metrics were used for evaluation:

1. **Mean Absolute Error (MAE)**
2. **Mean Squared Error (MSE)**
3. **Root Mean Squared Error (RMSE)**

Model performance was assessed using historical data, and each method’s forecasts were compared to select the most accurate approach.

# Inferences from the Project

The model performance, inferences, …

· **Seasonal Trends**: Sales spikes were observed during holidays, particularly around Thanksgiving and Christmas.

· **Influence of External Factors**: High unemployment and CPI negatively affected sales.

· **Store Performance**: Certain stores consistently outperformed others due to better regional demand.

# Future Possibilities

The future possibilities, limitations, ….

1. Incorporate more advanced deep learning models like LSTM or GRU for better long-term forecasts.
2. Factor in additional variables such as competitor pricing and promotions.
3. Extend forecasts beyond 12 weeks to support long-term planning.

# Conclusion

This project successfully analyzed Walmart’s sales data and applied various forecasting methods to predict future sales. The insights gained can help Walmart optimize inventory management and improve decision-making processes.

References

· Dataset provided by Walmart.

· Books and online resources on time series forecasting and machine learning.