
VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

# MACHINE LEARNING FOR ROBOTICS- BCSE424L

# TITLE : WALMART SALE ANALYSIS FOR RETAIL INDUSTRY WITH ML


## TEAM MEMBERS :

## VARSHITHA CHINTAREDDY - 21BRS1545

## YASHO NANDAN REDDY - 21BRS1430

## RAMESH CHANDRA - 21BRS1428

## FACULTY : DR S.RAJARAJESWARI

## ABSTRACT :

Estimating future revenues is the technique of sales forecasting. Businesses may anticipate both short- and long-term performance and make well-informed business decisions with the help of accurate sales projections. Businesses can use historical sales data, industry-wide comparisons, and economic trends as a basis for their forecasts. This time, the business is Walmart. The well-known retail company Walmart runs a network of hypermarkets. Walmart has made data available by aggregating 45 locations' worth of store details and monthly sales data. Walmart holds a number of annual promotional discount events. The four biggest markdowns occur before major holidays: Super Bowl, Labour Day, Thanksgiving, and Christmas. In comparison to weeks without holidays, the evaluation gives the weeks that have these holidays a five-fold higher weight. The information is delivered every week. We need to determine how the store's sales are affected by the holidays. The Christmas, Thanksgiving, Super Bowl, and Labour Day holidays are all featured. Algorithms like XgBoost, Random Forest, and ARIMA will be used. We will use these methods to train and test the data. There will also be an IBM deployment and Flask integration.

## MOTIVATION :

Sales forecasting is vital for retail businesses, particularly for a large-scale enterprise like Walmart, where accurate predictions are crucial to managing inventory, staffing, and promotional activities efficiently. Sales patterns in retail are influenced by numerous factors, including holidays and promotional events. Therefore, developing a precise forecasting model that accounts for these factors can significantly impact the company's operational success. The motivation behind this research is to create an advanced machine learning-based sales forecasting model that considers the unique influences of major holidays on Walmart’s sales. By doing so, Walmart can better align its operations with anticipated demand, optimize its marketing strategies, and ultimately enhance revenue performance.

## OBJECTIVE :

a. Analyse historical sales data to identify trends and patterns.
b. Predict future sales and demand for products.
c. Optimize inventory management and supply chain operations.
d. Segment customers for targeted marketing and personalized recommendations.
e. Evaluate the impact of external factors (e.g., holidays, economic conditions) on sales.
f. Develop a user-friendly dashboard for data visualisation and reporting.
By the end of this project, you will:

- Know fundamental concepts and techniques used for machine learning.
- Gain a broad understanding about data.
- Have knowledge on pre-processing the data/transformation techniques on outlier and some visualisation concepts.

## LITERATURE REVIEWS:

### PAPER 1 :

**AUTHOR :** Fildes .R , & Goodwin .P

**YEAR :** 2007

**TITLE :** "Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting."

**PROBLEM :** The challenge of integrating human judgment with statistical forecasting models, leading to potential biases in forecasts.

**SOLUTION :** The authors propose a framework that combines judgmental adjustments with statistical forecasts, enhancing accuracy through structured approaches.

### PAPER 2 :

**AUTHOR :** Tashman, L. J.

**YEAR :** 2000

**TITLE :** "Out-of-sample tests of forecasting accuracy: an analysis and review."

**PROBLEM :** Difficulties in evaluating forecasting model performance, especially when only in-sample data is used.

**SOLUTION :** The paper reviews out-of-sample testing techniques, recommending their use to better assess forecast accuracy in real-world scenarios.

**PAPER 3 :**

**AUTHOR :** Makridakis, S., Spiliotis, E., & Assimakopoulos, V

**YEAR :** 2018

**TITLE :** "Statistical and Machine Learning forecasting methods: Concerns and ways forward."

**PROBLEM :** The gap between traditional statistical methods and newer machine learning approaches in forecasting, with varying results.

**SOLUTION :** The authors compare these methods and suggest that a hybrid approach could leverage the strengths of both to improve accuracy.

**PAPER 4 :**

**AUTHOR :** Zhao, Y., & Kumar, P. R.

**YEAR :** 2012

**TITLE :** "Forecasting product returns for remanufacturing using support vector regression."

**PROBLEM :** Difficulty in predicting product returns for remanufacturing, which affects inventory and production planning.

**SOLUTION :** The authors propose the use of Support Vector Regression (SVR) to forecast returns, demonstrating improved accuracy over traditional models.

**PAPER 5 :**

**AUTHOR :** Borra, S., & Di Ciaccio, A.

**YEAR :** 2010

**TITLE :** "Measuring the prediction error: A comparison of cross-validation, bootstrap, and covariance penalty methods."

**PROBLEM :** The challenge of accurately measuring prediction error in forecasting models.

**SOLUTION :** The study compares various methods to measure prediction error, recommending cross-validation as a reliable technique for assessing model performance.

**PAPER 6 :**

**AUTHOR :** Taylor, J. W.

**YEAR :** 2003

**TITLE :** "Short-term electricity demand forecasting using double seasonal exponential smoothing."

**PROBLEM :** The complexity of forecasting short-term electricity demand due to seasonal fluctuations.

**SOLUTION :** The paper introduces a double seasonal exponential smoothing method to account for multiple seasonal patterns, leading to better short-term forecasts.

**PAPER 7 :**

**AUTHOR :** Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H.

**YEAR :** 2010

**TITLE :** "An empirical comparison of machine learning models for time series forecasting."

**PROBLEM :** Uncertainty about which machine learning models are most effective for time series forecasting.

**SOLUTION :** The authors empirically compare several machine learning models, providing insights into their relative strengths and weaknesses in time series applications.

**PAPER 8 :**

**AUTHOR :** Bose, I., & Mahapatra, R. K.

**YEAR :** 2001

**TITLE :** "Business data mining—a machine learning perspective."

**PROBLEM :** Challenges in applying machine learning techniques effectively within the context of business data mining.

**SOLUTION :** The paper discusses various machine learning approaches and how they can be adapted for business data mining tasks, highlighting key considerations for practitioners.

**PAPER 9 :**

**AUTHOR :** Chen, T., & Guestrin, C.

**YEAR :** 2016

**TITLE :** "XgBoost: A scalable tree boosting system."

**PROBLEM :** The need for a scalable and efficient machine learning algorithm for large-scale data analysis.

**SOLUTION :** The authors present XgBoost, a powerful machine learning algorithm that improves the speed and accuracy of tree-based models, particularly in high-dimensional datasets.

**PAPER 10 :**

**AUTHOR :** Breiman, L.

**YEAR :** 2001

**TITLE :** "Random forests."

**PROBLEM :** The problem of overfitting in decision tree models and the need for more robust predictive models.

**SOLUTION :** Breiman introduces the Random Forest algorithm, which reduces overfitting by creating an ensemble of decision trees, leading to improved model generalization.

**PAPER 11 :**

**AUTHOR :** Hyndman, R. J., & Athanasopoulos, G.

**YEAR :** 2018

**TITLE :** "Forecasting: principles and practice."

**PROBLEM :** A lack of comprehensive, accessible guides to forecasting methods that integrate both theoretical and practical insights.

**SOLUTION :** The authors provide a detailed guide on various forecasting techniques, including time series models like ARIMA, making it a valuable resource for practitioners and researchers.

**PAPER 12 :**

**AUTHOR :** Zhang, G. P.

**YEAR :** 2003

**TITLE :** "Time series forecasting using a hybrid ARIMA and neural network model."

**PROBLEM :** The limitations of traditional time series models like ARIMA in capturing complex, non-linear patterns in data.

**SOLUTION :** Zhang proposes a hybrid model that combines ARIMA with neural networks, enhancing the model's ability to forecast non-linear time series data.

**PAPER 13 :**

**AUTHOR :** Sathyanarayana, A., & Waghmare, M

**YEAR :** 2020

**TITLE :** "Forecasting Sales in the Retail Sector Using ARIMA."

**PROBLEM :** The need for reliable sales forecasting methods in the retail sector, particularly for managing inventory and planning promotions.

**SOLUTION :** The paper explores the use of ARIMA models for sales forecasting, demonstrating their effectiveness in capturing seasonality and trends in retail sales data.

**PAPER 14 :**

**AUTHOR :** Farias, V. F., & Li, S.

**YEAR :** 2019

**TITLE :** "Managing supply-demand risk in retailing: The role of retail price promotions."

**PROBLEM :** The challenge of balancing supply and demand in retail, particularly during promotional periods.

**SOLUTION :** The authors investigate the impact of price promotions on retail sales, providing strategies for managing supply-demand risk through better promotion planning.

**PAPER 15 :**

**AUTHOR :** Chen, H., Chiang, R. H., & Storey, V. C.

**YEAR :** 2012

**TITLE :** "Business intelligence and analytics: From big data to big impact."

**PROBLEM :** The difficulty in leveraging big data for meaningful business insights, particularly in retail forecasting.

**SOLUTION :** The paper discusses the role of business intelligence and analytics in harnessing big data, highlighting key techniques and technologies that can improve decision-making in retail operations.

**RESEARCH GAP :**

Although considerable research has been conducted on sales forecasting and the impact of holidays on retail sales, there is a noticeable gap in studies that integrate multiple machine learning algorithms, such as XgBoost, Random Forest, and ARIMA, to developtailored for a retail giant like Walmart. Most existing studies either focus on traditional statistical methods or single machine learning approaches, without exploring the potential benefits of combining these methods. Furthermore, the impact of holidays on sales is often analyzed in isolation, without a comprehensive model that takes into account varying holiday effects across different stores and time periods.

**PROBLEM STATEMENT DEFINITION :**

Walmart, one of the world's largest retail chains, seeks to enhance its sales analysis and forecasting capabilities to improve its retail operations. The company operates numerous stores across diverse locations, offering a wide range of products. To optimize inventory management, pricing strategies, and customer experiences, Walmart aims to leverage machine learning techniques for sales analysis and prediction. In this project, Walmart is looking to leverage machine learning to analyze and forecast sales, optimizing its retail operations. The primary goals include developing accurate sales forecasting models, improving inventory management to minimize overstock and stockouts, analyzing demand patterns for various products and regions, implementing dynamic pricing strategies, and segmenting the customer base for tailored marketing efforts. By utilizing data-driven insights, Walmart seeks to enhance operational efficiency, profitability, and customer satisfaction within its vast retail business operations.

Upon successful completion of the project, Walmart expects the following outcomes:

1. Improved sales forecasting accuracy, reducing excess inventory and stockouts.
2. Enhanced inventory management practices to optimize store operations and reduce carrying costs.
3. Data-driven insights to fine-tune pricing strategies, increasing profitability.
4. A deeper understanding of customer segments to tailor marketing efforts and improve the customer experience.

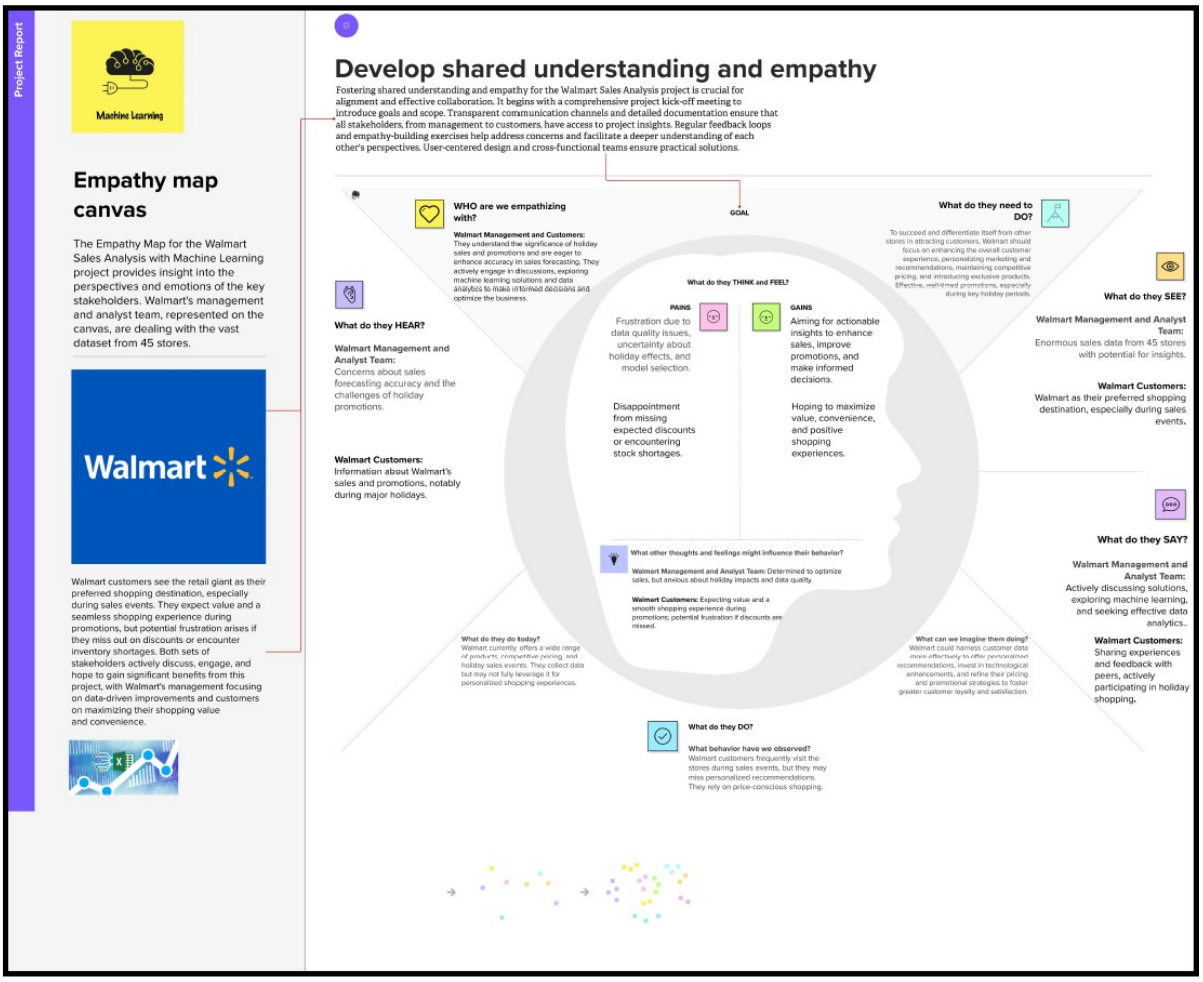
**EXISTING PROBLEM :**

Walmart, like many other retailers, faces several challenges in its sales analysis efforts within the retail industry using machine learning. These challenges include the need to ensure high data quality and integration from diverse sources, the management of volatile demand influenced by various factors, such as seasonality and promotions, and the complexities of maintaining optimal inventory levels while avoiding overstock and stock outs. Additionally, accurately segmenting customers and personalizing marketing strategies can be intricate, and achieving the highest model accuracy in sales forecasting

requires ongoing refinement. Data privacy and security are paramount concerns, especially when handling sensitive customer data, and scaling machine learning solutions across Walmart's extensive network of stores is no small feat. Implementing change management and ensuring model interpretability, along with continuous maintenance and monitoring, are also key hurdles in this machine learning-driven retail analysis endeavor. Addressing these challenges is crucial for Walmart to make the most of the potential benefits of machine learning in retail operations.

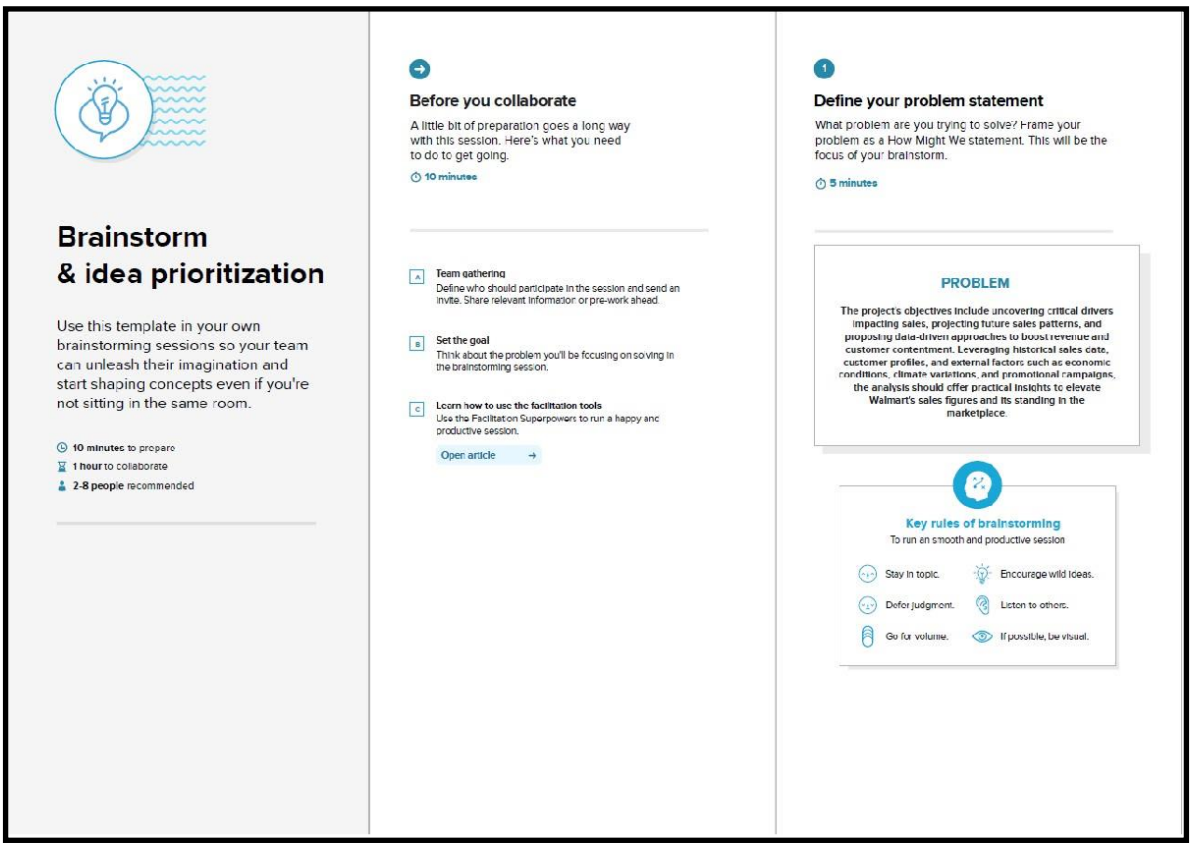
## PROPOSED SYSTEM :

- **Empathy Map Canvas :**

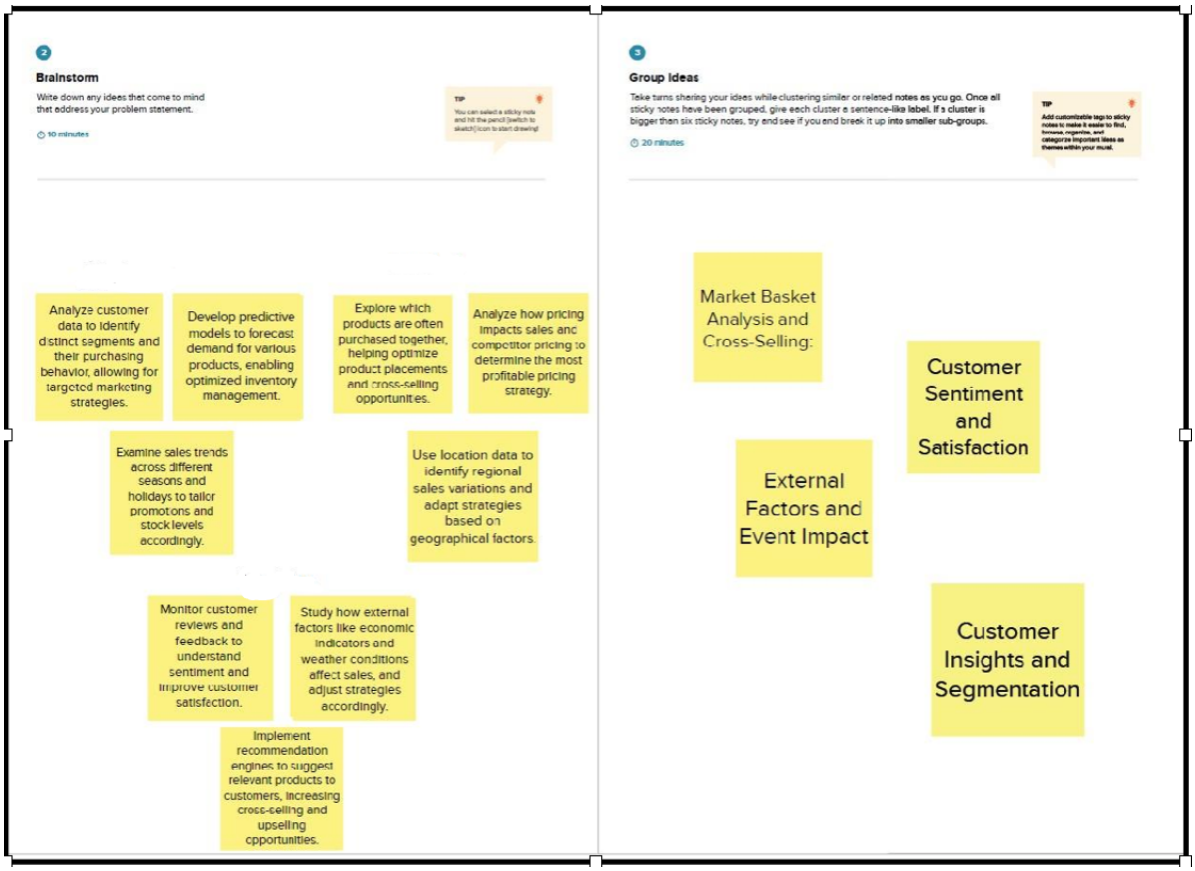



- **Ideation & Brainstorming :**

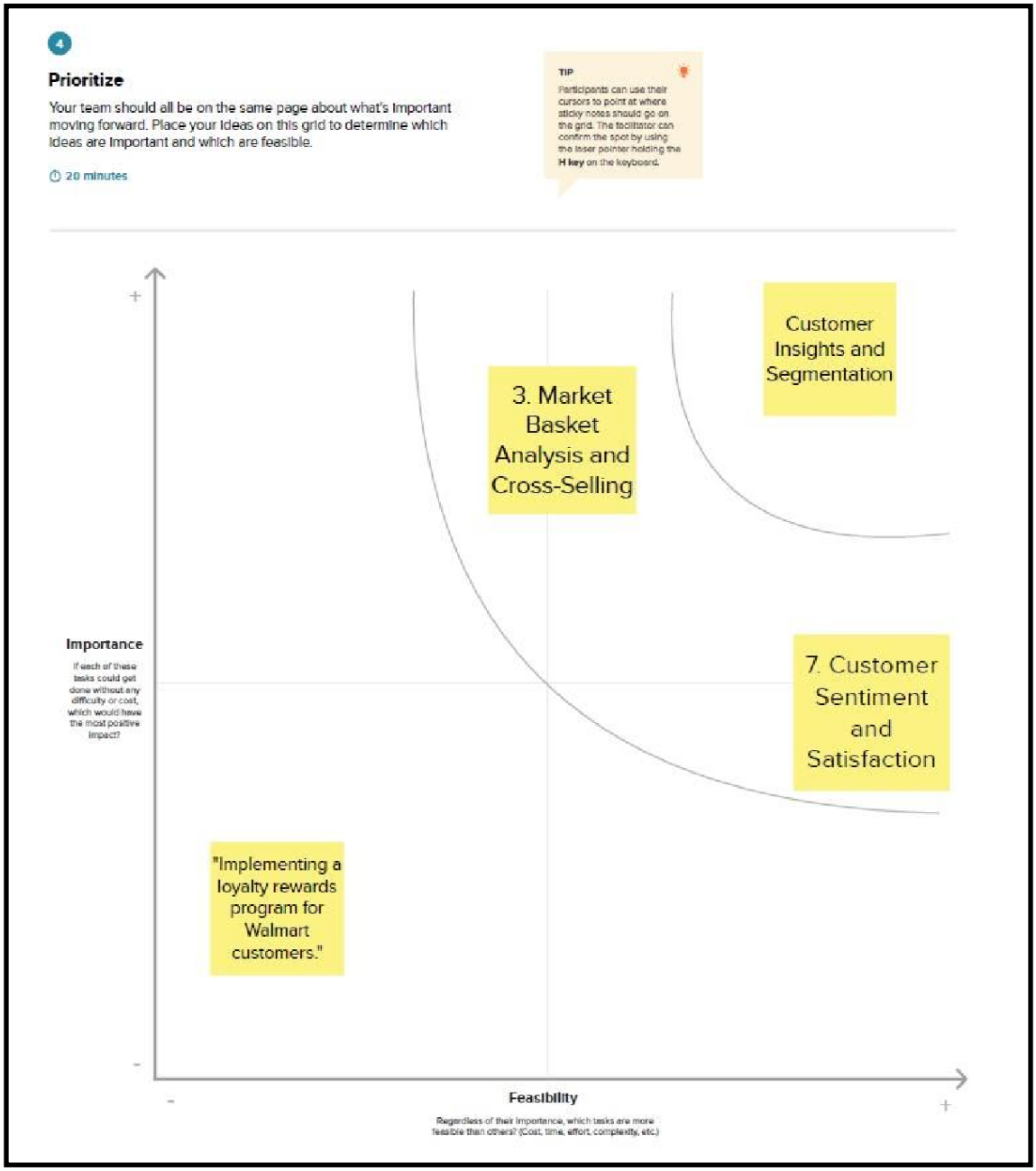
**Step 1: Team Gathering, Collaboration and Select the Problem Statement**




# Step 2: Brainstorm, Idea Listing and Grouping

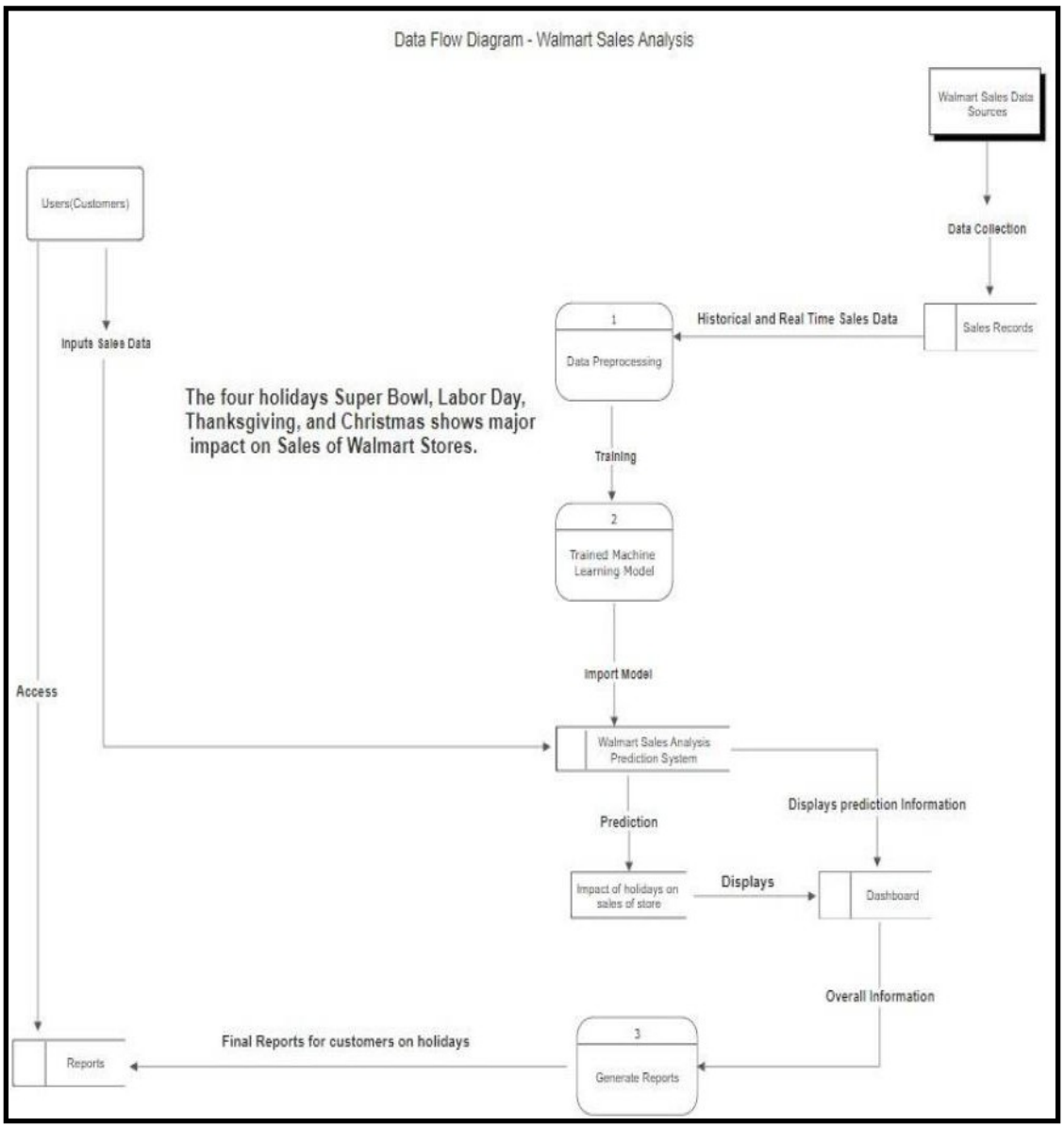


## Step 3: Idea Prioritization:




## ARCHITECTURE DIAGRAM :

- **Data Flow Diagram:**




The solution architecture for Walmart sales analysis using data analytics and machine learning involves a systematic approach to harness data and algorithms to develop predictive models that can assist in optimizing sales and business strategies.

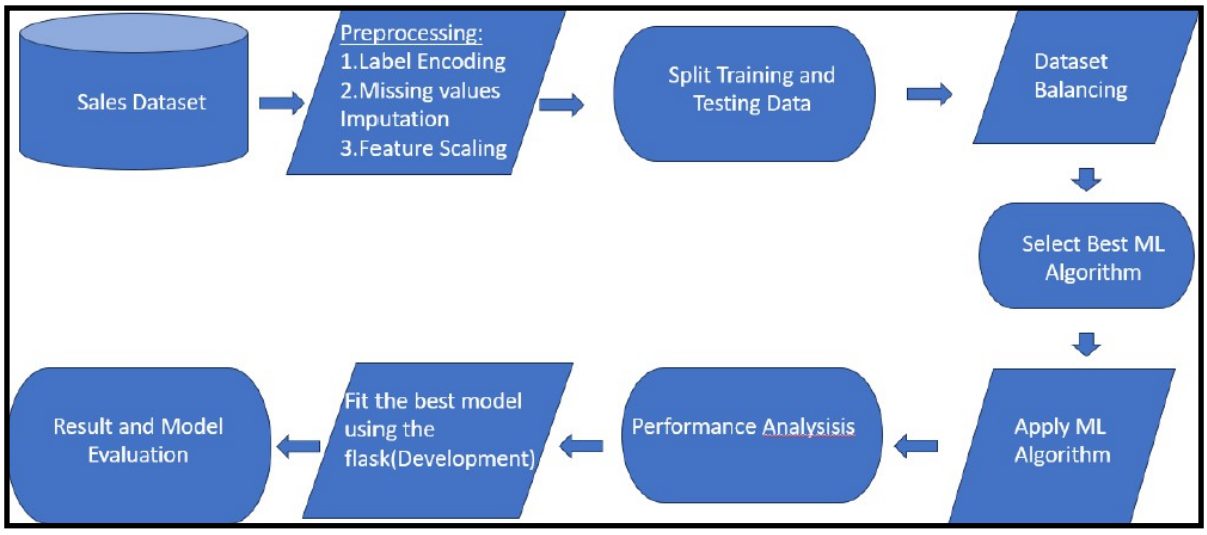
It predicts Walmart sales using machine learning by leveraging real-time sales and inventory data. Our system offers personalized sales and inventory assessments that consider factors such as historical sales trends, product categories, geographic locations, seasonal variations, promotions, and customer demographics. By harnessing real-time data, our software can generate accurate sales forecasts and inventory recommendations,

empowering Walmart and its stakeholders to make informed decisions in managing stock levels, pricing strategies, and marketing campaigns.

This innovative solution aims to revolutionize the retail industry by promoting efficient inventory management, maximizing revenue, and enhancing the overall shopping experience.

Through seamless integration with existing point-of-sale and inventory management systems, our software ensures a hassle-free and efficient experience for Walmart employees and management. Sales data, inventory updates, and customer information are securely analyzed by our machine learning algorithms to provide personalized sales and inventory insights, enabling Walmart to optimize its operations.

The positive impact of our software extends beyond the retail giant to the broader society. By optimizing Walmart's operations, we can reduce waste, improve supply chain efficiency, and ensure that products are available when and where customers need them. This, in turn, benefits both consumers and Walmart as it enhances customer satisfaction and reduces costs associated with overstock or understock situation



## ALGORITHMS USED :

- **ARIMA (AutoRegressive Integrated Moving Average):**

+ A popular statistical method used for forecasting time series data.

+ It combines autoregressive and moving average components to model the data.

+ ARIMA is effective in modeling complex patterns and seasonality in time series data.

- **Random Forest:**

+ An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification or mean prediction for regression.

+ Random forest is effective in handling high-dimensional data, noisy data, and outliers.

- **XGBoost (Extreme Gradient Boosting):**

+ An efficient and scalable implementation of gradient boosting framework that is widely used for structured or tabular data.

+ XGBoost is effective in handling large datasets, categorical variables, and missing values.

## METHODOLOGIES USED:

**Data Collection:**

* Gathering historical sales data from Walmart's various stores and regions, along with relevant external data such as economic indicators and weather data.

**Data Preprocessing:**

* Cleaning the data to handle missing values, outliers, and inconsistencies.

* Normalizing and standardizing the data to ensure quality and consistency.

**Exploratory Data Analysis (EDA):**

* Analyzing the data to identify trends, seasonality, and correlations among different variables.

**Feature Engineering:**

* Creating relevant features for predictive modeling, which may include:
+ Aggregating data
+ Creating time-based features
+ Encoding categorical variables

**Model Development and Training:**

* Selecting suitable machine learning models for sales prediction and training them using the preprocessed dataset.

**Model Evaluation:**

* Assessing model performance using metrics such as:
+ RMSE (Root Mean Square Error)
+ MAE (Mean Absolute Error)
+ Accuracy
* Ensuring the models are reliable by evaluating their performance on unseen data.

**Visualization and Reporting:**

* Creating interactive dashboards to visualize insights and generate reports for stakeholders.

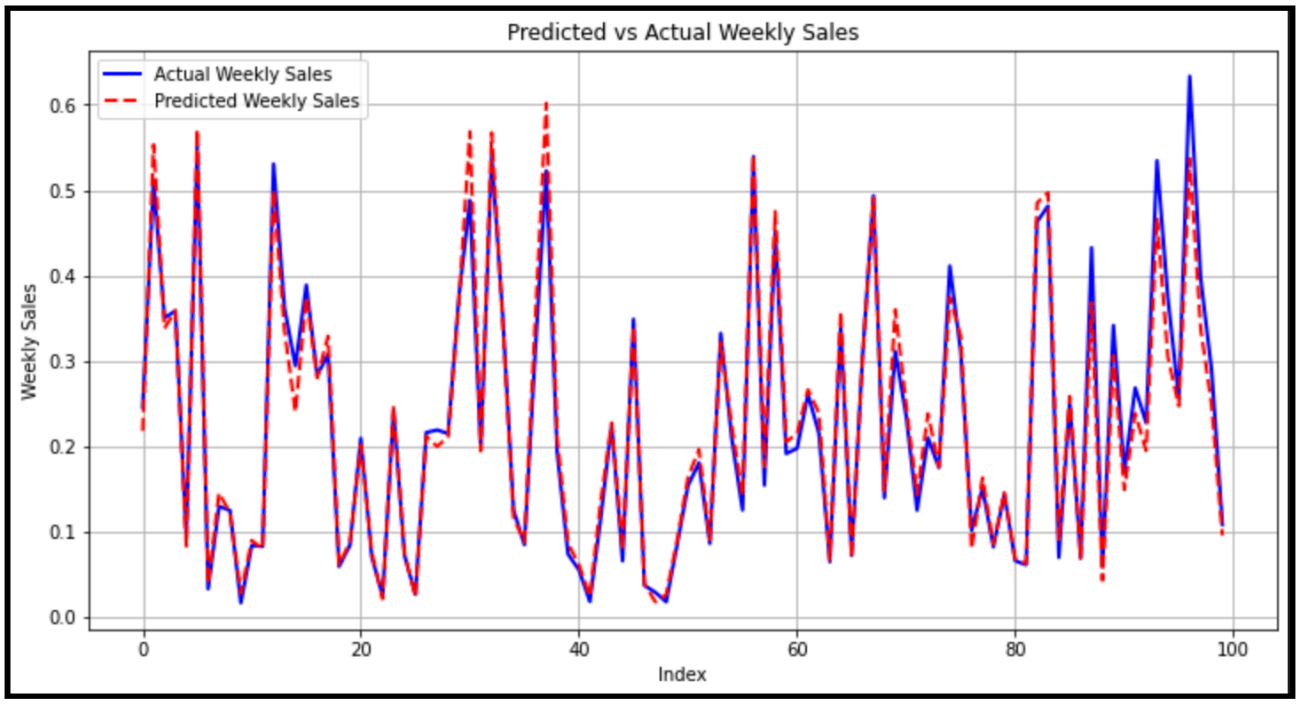
**Deployment:**

* Implementing the best-performing models in a real-time or batch prediction system, ensuring they can be accessed and utilized effectively.

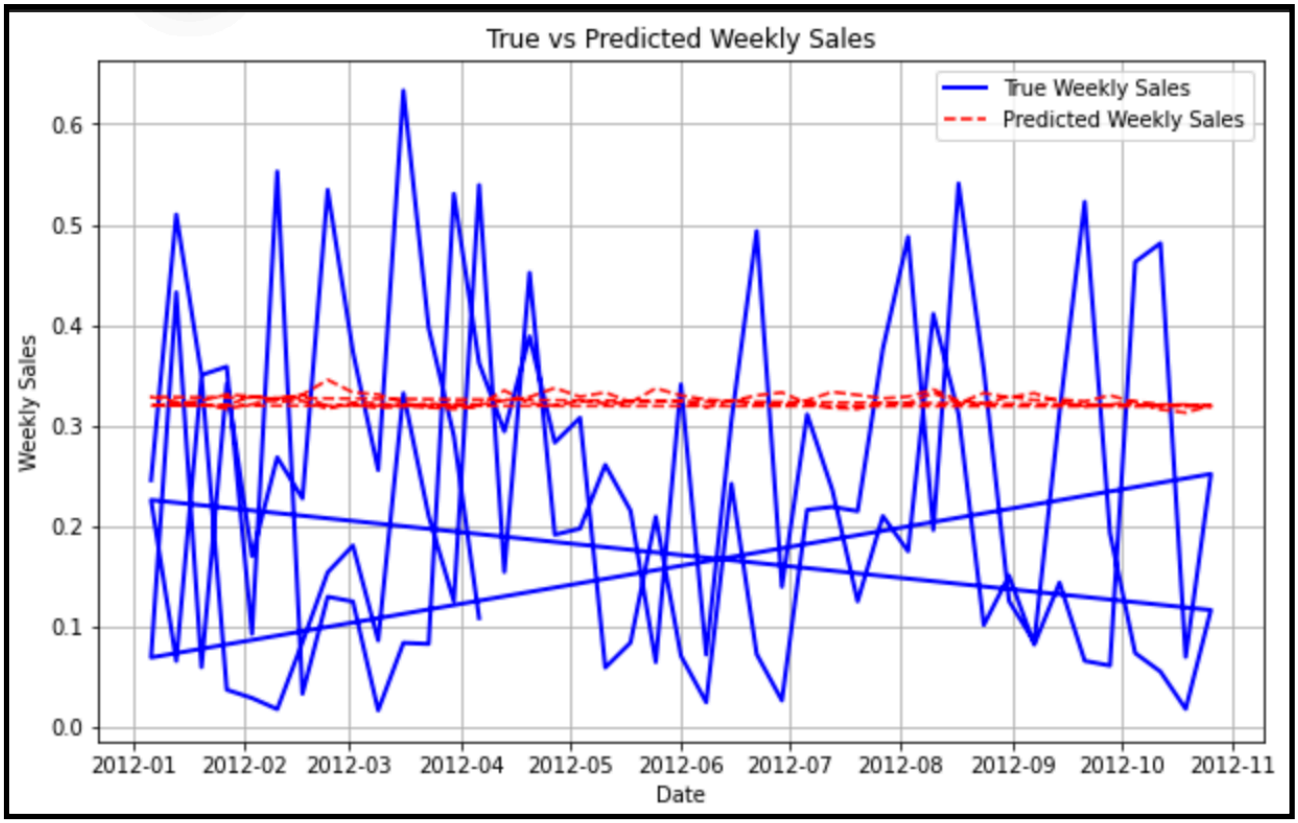
**CODE AND SCREENSHOTS :**

```
import matplotlib.pyplot as plt
plt.figure(figsize=(12, 6))
```

```
plt.plot(np.array(y_test)[:100], label='Actual Weekly Sales', color='blue',
linewidth=2)
plt.plot(np.array(y_pred)[:100], label='Predicted Weekly Sales', color='red',
linestyle='--', linewidth=2)
plt.title('Predicted vs Actual Weekly Sales')
plt.xlabel('Index')
plt.ylabel('Weekly Sales')
plt.legend()
plt.grid()
plt.show()
```

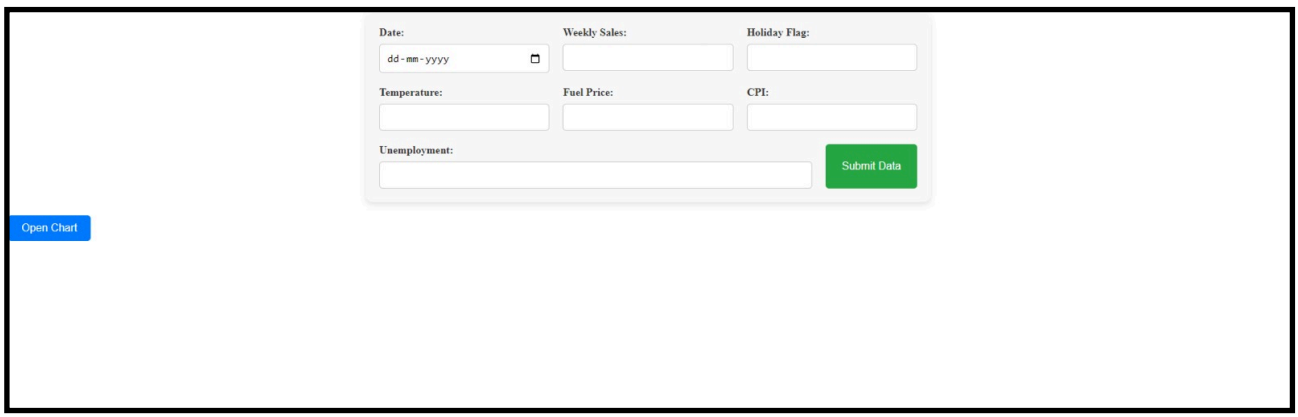


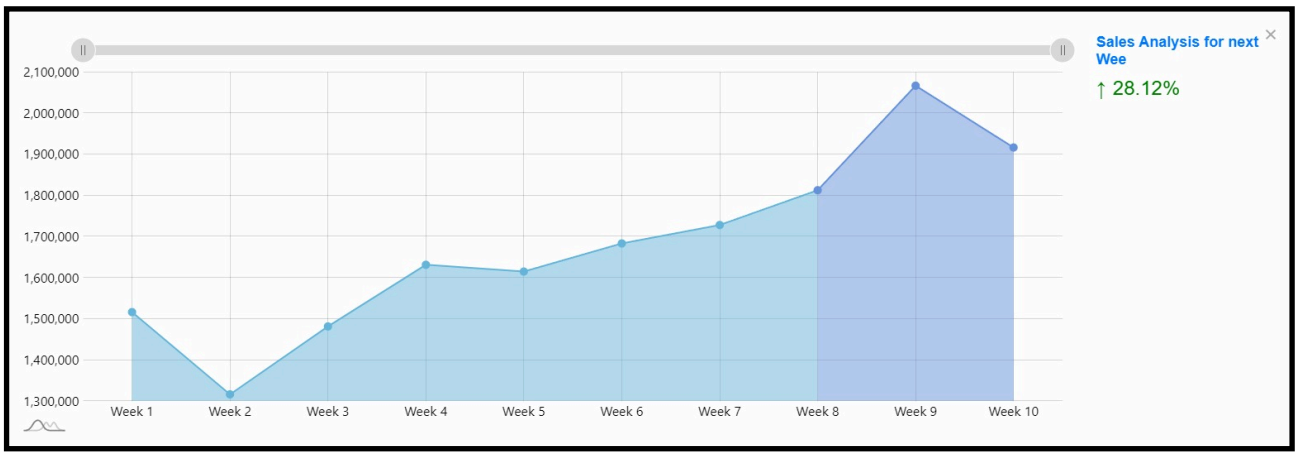

```
plt.figure(figsize=(10,6))
plt.plot(np.array(test_data['Date'])[:100], np.array(y_true)[:100], label='True
Weekly Sales', color='blue', linewidth=2)
plt.plot(np.array(test_data['Date'][:len(y_pred)])[:100], np.array(y_pred/20)
[:100], label='Predicted Weekly Sales', color='red', linestyle='--')
plt.xlabel('Date')
plt.ylabel('Weekly Sales')
plt.title('True vs Predicted Weekly Sales')
plt.legend()
plt.grid(True)
plt.show()
```




```
from flask import Flask, jsonify
from flask_cors import CORS
import pandas as pd
import pickle

app = Flask(__name__)
CORS(app)  # Enable CORS for all routes

# Load your model
model_path = r"C:\Users\Yasho Nandan Reddy\Desktop\ml_backend\random_forest_model.pkl"  # Use raw string
with open(model_path, 'rb') as file:
    model = pickle.load(file)

# Route to get data
@app.route('/get_data', methods=['GET'])
def get_data():
    try:
        # Load the recent data from CSV
        csv_path = r"C:\Users\Yasho Nandan Reddy\Downloads\WALMART_SALES_DATA.csv"  # Use raw string
        data = pd.read_csv(csv_path)
```

```
        # Get the last 7 weeks of data
        recent_data = data.tail(7)

        # Prepare features for prediction
        features = recent_data[['Holiday_Flag', 'Temperature', 'Fuel_Price', 'CPI',
'Unemployment']]
        predictions = model.predict(features)

        # Prepare data for chart
        result = recent_data[['Date', 'Weekly_Sales']].copy()
        result['Predicted_Sales'] = predictions

        # Format the result for the chart
        chart_data = result.rename(columns={'Date': 'week', 'Predicted_Sales':
'value'}).to_dict(orient='records')

        return jsonify(chart_data)

    except Exception as e:
        return jsonify({"error": str(e)}), 500  # Return error message

if __name__ == '__main__':
    app.run(debug=True)
```






**OBSERVATIONS :**

1. Data-Driven Decisions: Emphasizes the importance of using data analysis for informed decision-making.
2. Rich Data Utilization: Leverages Walmart's extensive dataset for enhanced predictive modeling.
3. Model Evaluation: Employs multiple algorithms to identify the most effective sales forecasting methods.
4. Data Management Challenges: Highlights issues related to data quality and privacy that need addressing.
5. Future Opportunities: Identifies areas for further innovation, such as personalized marketing and supply chain optimization.
6. Customer-Centric Focus: Aims to enhance customer experiences through targeted marketing strategies.
7. Scalability: Ensures that the methodologies can adapt to growing data volumes and market dynamics.

**CONCLUSION :**

The "Walmart Sale Analysis for Retail Industry with Machine Learning" project highlights the transformative potential of machine learning in enhancing retail operations. By utilizing advanced algorithms for sales forecasting, inventory management, and customer segmentation, Walmart aims to improve decision-making and operational efficiency. This data-driven approach enables the company to adapt to market changes and customer preferences, ultimately leading to increased profitability and customer satisfaction.