Capstone Report (Walmart)

Name – Tapan Khaladkar

Email – tapankhaladkar@gmail.com

1. Problem Statement

A retail store with multiple outlets across the country is facing issues in managing inventory to match demand with supply.

2. Problem Objective

The primary objective of this project is to analyze and forecast weekly sales for a retail chain with multiple store locations across the country. By leveraging data-driven techniques, we aim to extract meaningful insights from historical sales data and develop predictive models to assist the company in better inventory management and decision-making.

The specific objectives of this project are:

a. Perform EDA
b. Data Cleaning and Preprocessing
c. Sales Forecasting and Predictive Modelling
d. Derive Business Insights

3. Data Description

The dataset used for this project is **walmart.csv**, which consists of **6,435 rows and 8 columns**. This dataset captures historical weekly sales data from multiple stores along with various external factors that could influence sales.

Feature Name	Description
Store	Store number (Unique identifier for each store)
Date	Week of sales (Time period for the sales record)
Weekly sales	Sales amount for the given store in that specific week
Holiday Flag	Binary indicator (1 if the week contains a holiday, 0 otherwise)
Temperature	Temperature on the day of sale (measured in Fahrenheit)
Fuel Price	Cost of fuel in the store's region
CPI	Measures inflation and changes in purchasing power
Unemployment	Unemployment rate in the region during that week

a. The dataset provides information at a **weekly level**, making it suitable for **time-series analysis**.
b. External economic and environmental factors such as **Fuel Price, CPI, Unemployment, and Temperature** are included, which allows for an analysis of their impact on sales.
c. The **Holiday_Flag** feature helps in evaluating whether sales spikes or drops during holiday weeks.
d. The dataset includes **multiple stores**, allowing us to compare store-wise performance and trends.
e. The primary target variable for forecasting is **Weekly_Sales**, which we aim to predict for the next 12 weeks.

4. Data Preprocessing and Interpretation

Effective data preprocessing is crucial for improving the accuracy and reliability of the analysis and predictive modeling. The following steps were performed to clean and prepare the dataset:

1. Handling Missing Values

- The dataset was checked for missing values using df.isnull().sum().
- No explicit handling of missing values was observed, indicating that either the dataset was already clean or missing values were minimal.
- If missing values were found, they could be handled by:

  • Imputation (e.g., filling missing values with the mean, median, or mode).

  • Dropping rows or columns with excessive missing data.

2. Identifying and Handling Outliers

Interquartile Range (IQR) Method was used to detect outliers in the Weekly_Sales column:

- The first quartile (Q1) and third quartile (Q3) were calculated.
- The interquartile range (IQR) was computed as IQR = Q3 - Q1.
- The lower bound and upper bound were defined as:
  - Lower Bound = Q1 – 1.5 * IQR
  - Upper Bound = Q3 + 1.5 * IQR
- Any data points falling outside these bounds were considered outliers.
- The number and percentage of outliers were computed and printed.

3. Data Standardization and Scaling

- The **StandardScaler** from sklearn.preprocessing was imported, suggesting that numerical features might have been standardized.
- Standardization ensures that features such as Fuel_Price, CPI, Unemployment, and Temperature are on a similar scale, preventing certain features from dominating the model.

4. Exploratory Data Analysis (EDA) and Feature Relationship

- Various statistical analyses and visualizations (e.g., correlation analysis using Pearson's correlation coefficient) were likely conducted.
- Heatmaps and scatter plots could have been used to examine the relationships between variables.

5. Handling Date Features

Since the dataset contains a Date column representing the week of sales, it is likely that:

- The date was converted into a datetime format.
- New time-based features (e.g., Month, Year, Week Number) were extracted for better trend analysis.

5. Choosing the right algorithm for the project

The project involves **time-series forecasting** for predicting weekly sales across multiple retail stores. The following approaches were used:

1. **Feature Engineering for Time-Series Forecasting**

- Time-based features such as **Month, Week, Day of the Week** were extracted from the Date column.
- Time-based features such as **Month, Week, Day of the Week** were extracted from the Date column.
- **Rolling Mean Features**: A rolling mean of weekly sales was computed to smooth short-term fluctuations.

2. **Algorithm Selection**

The model selection process focused on predictive modeling techniques suited for time-series forecasting. The algorithm used in the project is:

Random Forest Regressor

- The model was trained separately for each store using historical sales data.
- It was evaluated using the **R^2 Score**, which measures how well the model explains variance in sales.

6. Motivation and Reasons For Choosing the Algorithm

- **Handles Non-Linearity**: Unlike traditional time-series models such as ARIMA, Random Forest can capture complex relationships between sales and external factors like temperature, fuel price, and unemployment.
- **Feature Importance**: It automatically ranks important variables influencing sales.
- **Handles Missing Data & Outliers**: Unlike linear models, Random Forest is robust to missing values and noisy data.
- **Scalability**: Since each store was trained separately, the model remains scalable for multiple outlets.

7. Assumptions

- The model assumes that factors like **CPI (Consumer Price Index), fuel prices, and unemployment rates** will not undergo sudden or drastic changes in the near future.

- It is assumed that all stores will continue operating **under the same conditions** as in the past.

- No stores will close, relocate, or undergo major operational shifts (such as renovations or changes in pricing strategy).

- This model is used because it can handle **complex, non-linear relationships** between sales and influencing factors.

- Instead of using traditional time-series models (like ARIMA), the model relies on extracted **Month, Week, and Day-of-the-Week** features to capture seasonal trends.

- The model assumes that factors like **fuel prices, CPI, and unemployment rates** will not change significantly over the next 12 weeks.

- Sudden economic changes (e.g., inflation spikes, policy shifts) are not accounted for in the predictions.

8. Model Evaluation and Techniques

1. Prediction Generation and Evaluation

- The **Random Forest Regressor** was used to generate **12-week sales forecasts** for each store.
- Predictions were made **iteratively**, updating lag features (Sales_Lag1, Sales_Lag2, and Sales_Rolling_Mean) dynamically for each subsequent week.
- Predictions for all stores were collected and visualized for analysis.

2. Visualization of Predictions

Line plots were used to visualize **weekly forecast trends** for each store.

The predictions were categorized in 3 sections:

Top 15 Performing Stores

Middle 15 Stores

Bottom 15 Store

3. Performance Summary

- Summary statistics were generated to compare stores based on **predicted average weekly sales**.
- Stores were ranked based on their expected sales performance over the next 12 weeks.
- The model's predictions were analyzed for stability, showing how different stores are expected to perform.

9. Inferences

1. Sales Trends and Seasonal Effects

a. Holiday and Year-End Trends:

- Sales exhibit a sharp increase during **holiday weeks** and towards the end of the year
- November and December show the **highest average weekly sales**, peaking at over **$1.25 million** per store, likely due to holiday shopping.
- January has the **lowest sales**, averaging below **$950,000** per store.

b. Monthly Sales Pattern:

- Sales **peak in December**, with February also showing a temporary increase, possibly due to post-holiday promotions.
- The lowest average sales occur in **September and October**, suggesting a seasonal dip before the holiday season.

2. Impact of Economic Factors

a. Unemployment rate effect:

- **Store 36**, however, has a **positive correlation**, indicating that it may be in an area where consumer spending remains stable despite unemployment fluctuations.
- **Stores 38 and 44** show a **negative correlation** with sales, meaning higher unemployment leads to lower sales.

b. CPI and Fuel Prices:

- CPI has **little direct impact on weekly sales**, but some stores exhibit higher sensitivity to inflation than others.
- **Stores 10, 21, and 40** showed **the highest positive correlation** with CPI, meaning sales rise when CPI increases, possibly due to higher-income customers who are less affected by inflation.
- **Stores 18, 30, and 35** had the **most negative correlation**, suggesting that consumers in these areas reduce spending as inflation rises.

3. Store Performance Comparison:

a. Top Performing Stores:

- The highest-selling stores, on average, generate **over $2 million in weekly sales**.
- **Stores 2, 5, 14, 15, and 20** are the **top 5 stores**, with Store 5 leading in overall average sales.
- These stores likely benefit from **high foot traffic, favorable demographics, or effective store management strategies**.

b. Lowest Performing Stores:

- The bottom-performing stores generate **less than $500,000 in weekly sales**, a stark contrast to top-performing locations.
- **Stores 37, 39, 41, 43, and 45** are the **worst-performing stores**, with Store 43 consistently having the lowest weekly sales.
- These stores may face challenges such as **lower demand, competition, or economic downturns in their locations**.

4. Model Forecasting and Prediction Analysis:

a. 12-week sales forecast:

- The model predicts **relatively stable sales trends** over the next 12 weeks, with **no extreme fluctuations** expected.
- Seasonal spikes are accounted for, but external disruptions (e.g., unexpected economic changes) are not modeled.

b. Feature Importance:

- Historical sales data (**lag features like Sales_Lag1, Sales_Lag2**) was the **most influential predictor** in the model.
- Time-based features such as **Month and Week** helped capture seasonality, with December showing the highest forecasts.
- External factors like **temperature and fuel prices** had minimal direct impact on sales trends.

10. Future Possibilities

1. Store and Inventory Optimization

- Develop early warning systems to detect stores that are **underperforming** and might require intervention (e.g., targeted marketing, store layout changes).

- Use clustering techniques to segment stores based on performance and recommend **customized strategies** for improving sales.

2. Enhancing Model Accuracy with Advanced Techniques

- Implement **Long Short-Term Memory (LSTM) networks** for capturing long-term dependencies and seasonality in sales data.

- Use **Transformers and Attention Mechanisms** for better time-series forecasting, particularly in handling abrupt demand fluctuations.

3. Supply Chain and Logistics Optimization

- Use sales forecasts to optimize **warehouse stocking levels and delivery schedules**.

- Predict seasonal inventory demands to avoid **overstocking or understocking** issues.

4. Competitor Analysis and Customer Demographics:

- Collect data on competing stores' promotions, pricing strategies, and market trends to refine sales predictions.

- Use demographic data (income levels, household size, local spending patterns) to personalize forecasts for specific store locations.