Table Of Contents

A. Problem Statement

B. Project Objective

C. Data Description

D. Data Pre-processing

E. Exploratory Data Analysis (EDA)

F. Insights

G. Feature Importance

H. Choosing the Algorithm for the Project

I. Model Building

J. Model Evaluation and Techniques

K. Future Possibilities of the Project

A. Problem Statement

A retail store chain - Walmart has multiple outlets across the country that are facing issues in managing the inventory to match the demand with respect to supply. We have to come up with useful insights using the data and make prediction models to forecast the sales for upcoming weeks/ months.

B. Project Objective

1. Using the data, coming up with useful insights that can be used by each of the stores to improve in various areas.
2. Forecasting the sales for each store for the next 12 weeks.

C. Data Description

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

D. Data Pre-processing

1. Loading the Dataset & Libraries: The initial step involves loading the dataset into the analysis environment, typically using libraries which helps in data manipulation, visualisation in Python, ensuring accessibility for further examination.
2. Shape Inspection : To get the overview of data.
3. Checking for Data Types: It is imperative to inspect the data types of each column to ensure consistency and appropriateness for subsequent analyses and operations.
4. Handling Missing Values : It's important to deal with missing/null values by 'dropping' or 'imputing' in order to perform EDA or model making.
5. Feature Engineering : Feature engineering involves creating or modifying features from raw data to enhance machine learning model performance. It helps models capture patterns more effectively, leading to improved accuracy.

We have added Following features to understand the sales better

i. Month- To understand monthly seasonality in data.
ii. Quarter - To understand quartery performances of stores.
iii. Weather - To make understanding easy for effect of weather on sales.

E. Exploratory Data Analysis (EDA)

Out of 6435 there are 450 weeks of holiday season

E.1 Descriptive Statistics

Inference -

1. The spread of temperature is quite significant (18.44), as well as the range is also quite wider indicating extreme cold (-2.06) and hot temperatures(100.14.)
2. There is big range of CPI showing significant increase in inflationary trend, also std suggests CPI varies significantly over time.
3. Average unemployment is 8% and varies from 3.8% to 14.3% over time showing changing economic conditions and may affect consumer'spurchasing power.
4. The range of weekly sales from 209,986 to 3,818,686 suggests a significant disparity between high and low-performing stores.
5. Std suggests large variation in weekly sales between stores and over time.

E.2 Corelation

Inference –

1. Negative relationship between sales and CPI - Unemployment suggest drop in sales when economic indicators are not good.
2. Moderate negative relationship between CPI and Unemployment which makes sense as inflation suggests economic growth and high employment (lower unemployment).

E.3 Inferential Statistics

• Null Hypothesis - There is no change in weekly sales before holiday week and during holiday week

• Alternate Hypothesis - There is significant change in weekly sales before holiday week and during holiday week

Inference -

1. p_value is > than threshold 0.05, which indicates we don't have enough evidence to reject null hypothesis.
2. There is no significant statistical evidence to support that there is change in weekly sales before and after holidays.

F. Insights

1. Time Based

1.1 Monthly Analysis

Inference -

1. June and July are the months where sales are highest.
2. Sales are lowest in month of January.

1.2 Quarterlty Analysis

Inference -

1. Sales is highest in 2nd and 3rd quarter whereas 1st quarter is subdued.

1.3 Holiday Season Wise

Inference -

1. Weekly sales during Holiday weeks are slightly higher than that of non-holiday weeks.

1.4 Seasonality in Sales

Inference -

1. There is strong seasonality around the year end which makes sense as it's 'Thanksgiving','Christmas' and 'New Year' holiday season.
2. We can manage the inventory of stores according to the seasonality or high demand near holiday season.
3. It is also evident that after holday season the sales are subdues around Jan end.

2. Temperature Based

Temperature

Inference -

1. It suggests that weekend sales are higher when tempreture is moderate i.e between range of 40F-80F.

Weather

Inference -

1. It is evident that weekend sales are higher when weather is 'Mild' and 'Warm' i.e suitable to outdoor activities.

2. Extreme 'Cold' and 'Hot' weather is affecting sales negatively.

3. Economy Based

CPI - Inflation Based

Inferences -

1. Stores like 38 (0.812) and 44 (0.740) show a strong positive correlation. This suggests higher CPI values shows higher sales in these stores. It'smaybe due to respective customer demographics or economic conditions in these areas where stores are operating.
2. Stores 36 shows a strong negative correlation which makes sense indicating bad economic conditions affecting sales.
3. Most of stores shows very small corelations indicating no impact on sales.

4. Store Based

4.1 Storewise Sales

Inferences -

1. There is huge difference in sales data between stores, there maybe different demographic, geographic, economic factors affecting the same.

4.2 Top 10 Performing Stores & Worst 10 Performing Stores

Inferences -

1. Store no. 20 is top performing store contributing to nearly 4.5% of total sales and Top 10 performing stores contribute to 39% of total sales.
2. Store no. 33 has lowest sales contributing to nearly 0.5% of total sales and Worst 10 performing stores contribute to 8.5% of total sales.
3. The disparity between the top and bottom performers suggests significant variation in store performance liekly due to store size, location, regionaleconomic conditions, or customer base

Suggestions -

1. We can utilize the success of top-performing stores to replicate best practices across the network.
2. High-performing stores should continue to receive support for expansion, customer retention while Low-performing stores need 'targeted strategies'like marketing campaigns, optimizing inventory based on customer interests, or improving customer footfall.

4.4 Storewise performance in Holiday Season

Inferences -

1. Some stores perform really well whereas some store does not perform well even during holiday season.
2. In magnitude aspects store numbers like 3,5,30,33,36,38,44 didn't perform well in holiday season compared to other stores.

Stores to focus on in Holiday season - Store No. - 3,5,33,36,44

Inferences -

1. It is evident that during holiday season, sales of these stores aren't much higher than before holiday season.
2. We can have targeted marketing campaigns, promotional offers, incentives for customers to boost consumption in these stores.

G. Feature Importance

SHAP (SHapley Additive exPlanations) - To interpret model outputs and analyse feature importance

Applications of SHAP:

1. Explain ability in Machine Learning Models: SHAP makes it easier to understand how complex machine models make decisions.
2. Feature Importance Analysis: It shows us which parts are more important for better understanding.
3. Interpreting Black Box Models: SHAP works with both straightforward(LR,LogR) and complex models(DT,RF), making it simple to understand howthey work.

H. Choosing the Algorithm for the Project

1. As it's Time series problem we can use SARIMAX, Prophet models.
2. SARIMAX can handle external predictors unlike SARIMA but it doesn't handle the complex non-linear relationships between predictors.
3. **Prophet** is more advanced easy to interpret model specifically for time series forecasting that can also handle **seasonality,trends, external predictors and non-linear relationships**.
4. We can also use XGBoost after some feature engineering as data has some external factors affecting sales.
5. Though XGBoost specifically can't handle time series data it handles complex, non-linear relationships and feature interaction much more effectively than prophet.

I. Model Building

Forecasting the sales for each store for the next 12 weeks using **Prophet** model.

J. Model Evaluation and Techniques

Achieved a Mean absolute percentage error (MAPE) of < 5% for predicting weekly sales.

K. Future Possibilities of the Project

1. Waste Reduction: Can use forecasts to minimize overstock and reduce waste, especially for perishable items.
2. Supply Chain Planning: Support sustainability goals by optimizing supply chain planning.
3. Marketing Campaigns: Identify optimal timing for promotions based on forecasted trends like before holiday season.
4. Capex Expansion: Planning where to focus and deploy the expenditure for further growth.
5. Revenue Analysis: Analysing the revenue generated by each store.
6. Inventory Management: Deciding when to expand and contract the inventory based on seasonality.