Capstone Project - Walmart

Table of Contents

Problem Statement

The retail store, with multiple outlets across the country, is facing issues in managing inventory to match demand with supply. The goal of this project is to provide useful insights using the available data and develop prediction models to forecast sales

Project Objective

The objective of this project is to analyze the provided dataset and derive insights that can be used by each store to improve various areas such as inventory management, sales strategies, and resource allocation. Additionally, the project aims to develop accurate sales forecasting models to help the retail store plan and optimize their operations effectively.

Data Description

The dataset "walmart.csv" consists of 6435 rows and 8 columns. The features include

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

```
Data Shape:
(6435, 8)
```

```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   object
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
None
```

Data Preprocessing Steps And Inspiration

The data pre-processing steps involved in this project include

- handling missing values,
- converting date columns to the appropriate format,
- scaling numerical features,
- extracting additional features if required.

The inspiration for the data pre-processing steps is to ensure the data is in a suitable format for analysis and modeling.

```
Missing Values:
Store           0
Date            0
Weekly_Sales    0
Holiday_Flag    0
Temperature     0
Fuel_Price      0
CPI             0
Unemployment    0
dtype: int64
```

```
Unique Values:
Store            45
Date            143
Weekly_Sales   6435
Holiday_Flag      2
Temperature    3528
Fuel_Price      892
CPI            2145
Unemployment    349
dtype: int64
```

Choosing the Algorithm For the Project

For this project, multiple algorithms suitable for Regression problems are used to forecast sales for each store. The algorithms used are:

- Linear Regression
- Gradient Boosting Regressor
- Support Vector Regression (SVR)
- Random Forest Regressor
- ARIMA (Autoregressive Integrated Moving Average)

Assumptions

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

1. The historical sales data is representative of future sales patterns.
2. The relationships between sales and external factors (holidays, temperature, fuel prices, etc.) remain consistent.
3. There are no significant unforeseen events that could dramatically impact sales.
4. Due to computational limitations or resource constraints, the models are applied to a subset of the total number of stores instead of all 45 stores.

Model Evaluation and Technique

To assess the performance of the chosen models, appropriate evaluation techniques were applied. These techniques included:

- train-test splitting,
- model fitting,
- model diagnostics.
- The models were trained on the training dataset and evaluated using the test dataset.

Various evaluation metrics were used to measure the performance of the models. These metrics provide insights into the accuracy of the sales predictions and the model's ability to capture the patterns in the data. The evaluation metrics used include:

- Mean Absolute Error (MAE),
- Mean Squared Error (MSE),
- Root Mean Squared Error (RMSE).

These metrics quantify the difference between the predicted sales values and the actual sales values.

```
Best model based on MAE: Random Forest
Best model based on RMSE: Random Forest
Best model based on MSE: Random Forest

Store: 3
Best model based on MAE: Random Forest
Best model based on RMSE: Random Forest
Best model based on MSE: Random Forest

Store: 5
Best model based on MAE: Linear Regression
Best model based on RMSE: Linear Regression
Best model based on MSE: Linear Regression

Store: 7
Best model based on MAE: Random Forest
Best model based on RMSE: Random Forest
Best model based on MSE: Random Forest

Store: 9
Best model based on MAE: Random Forest
Best model based on RMSE: Gradient Boosting
Best model based on MSE: Gradient Boosting

Best models:
Store 1: Random Forest
Store 3: Random Forest
Store 5: Linear Regression
Store 7: Random Forest
Store 9: Gradient Boosting
```

Conclusion

The evaluation of the models provided insights into its forecasting accuracy and the ability to capture the seasonal and temporal patterns in the sales data. The model's performance metrics and diagnostic plots help evaluate the model's strengths and limitations.