Capstone Project – Walmart

Table of Contents

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. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

Project Objective

The primary objective of this project is to forecast weekly sales for each store for the next 12 weeks. The key goals are:

1. To analyze historical sales data and derive insights.
2. To understand which factors (such as holidays, temperature, fuel price, unemployment, CPI) most influence sales.
3. To develop a predictive model to forecast sales, which can assist in inventory management and planning.

Data Description

Feature Name	Description
Store	Store Number
Date	Week of Sales
Week_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 Preprocessing Steps And Inspiration:

Data preprocessing is essential for cleaning and transforming raw data into a format suitable for analysis and modeling. The steps involved include:

1. **Handling Missing Values**: Identifying and filling or removing rows with missing values in the dataset.
2. **Feature Engineering**: Creating new features, such as 'Month' from the 'Date' feature, which can be useful in time-series analysis.
3. **Data Normalization/Scaling**: Scaling numerical features like Temperature, Fuel_Price, CPI, and Unemployment to ensure that the models treat them with equal importance.
4. **Encoding Categorical Variables**: Encoding the Holiday_Flag (a binary variable) as a numerical feature (0 or 1).
5. **Date Formatting**: Converting the 'Date' column into a datetime format for better manipulation.
6. **Outlier Detection**: Checking for and handling any outliers in numerical features that may skew the model's predictions.

Choosing the Algorithm for the Project:

Seasonal Autoregressive Integrated Moving Average (SARIMA)

For this project, **SARIMA (Seasonal Auto-Regressive Integrated Moving Average)** is chosen because:

1. **Time Series Nature**: The data represents time series, and SARIMA is specifically designed to handle seasonal trends and patterns.

2. **Seasonality and Trends**: SARIMA models both trend and seasonality, which is crucial in forecasting sales that exhibit weekly and annual seasonal effects.

3. **Stationarity Requirement**: SARIMA handles non-stationary data by incorporating differencing and integrating seasonality adjustments.

4. **Auto-regressive and Moving Average Components**: It captures the relationship between previous weeks' sales and other seasonal factors, such as holidays.

Motivation and Reasons for Choosing the Algorithm:

The reasons for choosing **SARIMA** are as follows:

1. **Seasonality Handling:** Sales data often exhibit seasonality (e.g., holiday seasons or specific weather conditions), which SARIMA models well.

2. **Time Dependency:** The sales of the previous weeks impact future sales. SARIMA is well-suited for time-series data where historical values predict future ones.

3. **Model Explainability:** SARIMA provides clear insight into the parameters, such as the autoregressive (AR), moving average (MA), and seasonal components, making the model interpretable.

4. **Data Characteristics:** Since the data is time-series with observed seasonality and trends, SARIMA's flexibility in incorporating both makes it the optimal choice.

Assumptions

In developing and implementing the SARIMA (Seasonal Autoregressive Integrated Moving Average) model for our project, we make the following key assumptions:

- **Seasonality:** Sales patterns exhibit seasonality, with certain weeks having consistently higher sales (e.g., holidays, weekends).
- **Stationarity:** It is assumed that after differencing, the sales data will be stationary.
- **External Factors:** External features like fuel price, unemployment, and temperature are assumed to influence sales but are not directly incorporated in the SARIMA model in this iteration (SARIMA primarily uses the historical values of the sales).
- **Data Quality:** It is assumed that the provided data is accurate and free of significant errors.

Model Evaluation and Technique

The performance of the SARIMA model will be evaluated using the following techniques:

Evaluation Metrics:

To assess the effectiveness of the SARIMA model, we employ the following evaluation metrics:

- **Root Mean Squared Error (RMSE)**: Measures the square root of the average squared differences between predicted and actual sales.
- **Mean Absolute Error (MAE)**: Measures the average absolute differences between predicted and actual sales.
- **Mean Absolute Percentage Error (MAPE)**: Measures the prediction accuracy as a percentage of the error relative to actual values.

Visualizing Residuals:
Plotting the residuals (errors) from the model to ensure they resemble white noise, indicating a good fit.

Inferences from the Project

After training and evaluating the SARIMA model, the following insights can be drawn:

- **Sales Patterns**: Understanding how sales vary weekly and seasonally, potentially aligning with specific holidays or external factors (e.g., temperature).
- **Forecast Performance**: Evaluating how well the model forecasts sales for future weeks.
- **Effectiveness of Seasonality Modeling**: Insights into how well the seasonal components (e.g., holidays, specific months) are captured in the model.

Future Possibilities of the Project

- **Incorporating External Variables**: While SARIMA handles seasonality well, incorporating features like temperature, fuel price, and CPI could further improve the model's performance (e.g., using SARIMAX, the extension of SARIMA for exogenous variables).
- **Extended Forecasting Horizon**: Forecasting for longer periods (e.g., a year or more) by considering annual cycles.
- **Real-Time Forecasting**: Updating forecasts dynamically as new sales data is collected each week.
- **Optimization**: Integrating SARIMA predictions with inventory management systems to optimize stock levels based on the forecasted demand.
- **Multi-Store Forecasting**: Apply SARIMA models for each store individually or use a multivariate approach to forecast sales for multiple stores simultaneously.

Conclusion

The forecasting project for 45 Walmart stores, incorporating temperature, weekdays, weekends, holidays, Consumer Price Index (CPI), and unemployment data, has provided valuable insights into the sales patterns and trends. The utilization of the SARIMA (Seasonal Autoregressive Integrated Moving Average) model allowed for the consideration of seasonality and historical dependencies in the time series data.

Key Findings and Achievements: 1.

Seasonal Patterns and Trends:

- The SARIMA model effectively captured and accounted for the seasonal patterns inherent in the sales data. ○ Trends related to temperature variations, holidays, and weekdays/weekends were incorporated into the forecasting process.

2. **Forecast Accuracy:**
   - The forecasting model provided predictions for sales across the 45 Walmart stores, considering multiple factors influencing consumer behavior. ○ The evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), offered insights into the accuracy and precision of the predictions.
3. **Challenges and Limitations:**
   - Despite the model's capability to capture seasonality, challenges in accurately predicting sales were observed. ○ Limitations included sensitivity to outliers, potential deviations from assumed seasonal patterns, and the exclusion of external factors.