Capstone Project :–

Walmart Sales analyses Time Series Project

Table of Contents :-

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Assumptions
7. Model Evaluation and Techniques
8. Inferences from the Same
9. Future Possibilities of the Project
10. Conclusion
11. References

Problem Statement :-

The retail store, with its widespread network of outlets across the country, is grappling with significant challenges in inventory management. The core issue lies in aligning the demand for products with the available supply. The current system is struggling to efficiently match the dynamic demands of customers with the existing inventory, leading to instances of overstocking or stockouts. This mismatch not only impacts customer satisfaction but also has financial implications for the business. To address these concerns, there is a critical need for a comprehensive solution that leverages data-driven insights to optimize inventory levels, ensuring a balanced and responsive approach to meet customer demand while minimizing excess stock. This project aims to develop a robust time series analysis model for the retail store's weekly sales data, offering actionable recommendations to enhance inventory management efficiency and streamline the supply chain.

Project Objectives :-

1. **Data Exploration:**
   - Conduct a comprehensive exploration of the Walmart weekly sales dataset to gain insights into the underlying data structure, distributions, and potential outliers.

2. **Sales Trends Identification:**
   - Identify and analyze significant trends, seasonality, and patterns present in the weekly sales data, considering variations across different stores and time periods.

3. **Correlation Analysis:**
   - Investigate the correlation between weekly sales and various factors such as store-specific attributes, holidays, and economic indicators to understand potential drivers of sales fluctuations.

4. **Forecasting Model Development:**
   - Develop a robust time series forecasting model to predict future sales trends accurately, providing a basis for proactive inventory management.

5. **Decision-Making Empowerment:**
   - Provide decision-makers with actionable insights and data-driven recommendations derived from the analysis to enhance strategic decision-making in inventory, sales, and supply chain management.

6. **Model Accuracy Evaluation:**
   - Evaluate the accuracy and reliability of the developed forecasting model using appropriate metrics, ensuring its effectiveness in predicting sales trends and guiding inventory decisions.

7. **Documentation of Best Practices:**
   - Document and communicate best practices and lessons learned throughout the project, creating a knowledge base for future improvements in inventory management strategies.

Data Description :-

Dataset Information :- The dataset available is a file named **walmart.csv**, which comprises 6435 rows and 8 columns.

Feature Descriptions :-

1. **Store:**
    - Definition: Store number
    - Type: Numeric

2. **Date:**
    - Definition: Week of Sales
    - Type: Date

3. **Weekly_Sales:**
    - Definition: Sales for the given store in that week
    - Type: Numeric

4. **Holiday_Flag:**
    - Definition: Binary indicator for whether it is a holiday week or not
    - Type: Binary (0 or 1)

5. **Temperature:**
    - Definition: Temperature on the day of the sale
    - Type: Numeric

6. **Fuel_Price:**
    - Definition: Cost of fuel in the region
    - Type: Numeric

7. **CPI (Consumer Price Index):**
    - Definition: A measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.
    - Type: Numeric

8. **Unemployment:**
    - Definition: Unemployment Rate

- Type: Numeric

<u>Insights from the Data :-</u>

- The dataset provides information on weekly sales for different stores along with relevant features such as holidays, temperature, fuel prices, CPI, and unemployment rate.
- The 'Holiday_Flag' column indicates whether a given week includes a holiday (1) or not (0).
- 'Temperature' and 'Fuel_Price' offer insights into external factors that might influence sales.
- 'CPI' and 'Unemployment' provide economic indicators that could impact consumer behavior.
- Analysis of this dataset aims to uncover trends, patterns, and correlations among these features, contributing to a better understanding of the factors influencing weekly sales at Walmart stores.

Data Preprocessing Steps And Inspiration :-

The preprocessing of the data included the following steps :-

1. **Setting Data Column as Index:**
   - The data column was set as the index to make it suitable for time series models, allowing for a more accurate representation of temporal relationships.

2. **Checking Columns Information:**
   - The **data.info** method was utilized to examine information about the columns, ensuring data types and non-null values were appropriate for analysis.

3. **Handling Null Values and Duplicates:**
   - No null values or duplicate entries were identified during the initial data exploration. The dataset was clean in terms of missing or redundant information.

4. **Handling Outliers:**
   - Despite the presence of outliers, it was decided to retain them as they potentially contain valuable information. Outliers were not removed, but their impact was considered during subsequent analysis.

5. **Basic Exploratory Data Analysis (EDA):**
   - Conducted basic EDA to understand the statistical characteristics of the data.
   - Calculated correlations between different features to identify potential relationships.

6. **Checking Stationarity:**
   - Assessed the stationarity of the time series data to ensure its suitability for modeling.
   - If the data was found to be non-stationary, the transformation involved taking the log of the data and subtracting the rolling mean of the log values from three months prior.

7. **Data Transformation for Modeling:**

- Ensured that the data was appropriately transformed for modeling purposes, incorporating any necessary adjustments identified during the exploratory phase.

These preprocessing steps aimed to create a clean, structured dataset ready for time series analysis. The decisions made, such as handling outliers and transforming non-stationary data, were driven by the goal of preserving valuable information while preparing the data for effective modeling.

Choosing the Algorithm For the Project :-

Algorithm 1: SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors)

Description: SARIMAX, which stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous factors, is a powerful time series forecasting algorithm. It is an extension of the ARIMA model, capable of handling seasonality and incorporating external variables (exogenous factors) that may influence the time series data.

I have chosen the SARIMAX algorithm for this project for the following reasons:

1. **Handling Seasonality:**
    - SARIMAX is particularly effective in handling time series data with clear seasonality patterns. Given the nature of the weekly sales data, which may exhibit repeating patterns, SARIMAX is well-suited to capture and model these seasonal variations.
2. **Incorporating Exogenous Factors:**
    - SARIMAX allows the inclusion of exogenous variables, such as holidays, temperature, fuel prices, CPI, and unemployment rate. This flexibility enables the algorithm to consider external factors that may influence sales, contributing to a more accurate prediction.
3. **Optimized Parameters:**
    - SARIMAX involves tuning parameters such as order (p, d, q) and seasonal order (P, D, Q, s). Through experimentation and optimization, SARIMAX parameters can be fine-tuned to achieve better predictive performance on the given dataset.
4. **Error Metrics:**
    - During the model evaluation phase, SARIMAX consistently demonstrated lower error metrics compared to other algorithms considered. This indicates better accuracy in predicting weekly sales trends.

Algorithm 2: Exponential Moving Average (EMA)

Description: Exponential Moving Average (EMA) is a straightforward time series forecasting method that assigns exponentially decreasing weights to past observations. It is a simple yet effective algorithm, especially suited for scenarios where quick adaptation to changes is essential.

For this project, I have also considered the Exponential Moving Average for the following reasons:

1. **Simplicity and Interpretability:**
    - EMA is a simple and intuitive algorithm, making it easy to understand and interpret. Its straightforward nature is advantageous, especially in scenarios where a simpler model is preferred.
2. **Quick Adaptation to Changes:**
    - EMA is adaptive and responsive to recent changes in the data. It assigns more weight to recent observations, allowing the model to quickly adapt to shifts in sales patterns.
3. **Ease of Implementation:**
    - EMA is computationally efficient and easy to implement, requiring minimal parameter tuning. This makes it a practical choice for scenarios where simplicity and ease of use are essential.
4. **Baseline Comparison:**
    - EMA serves as a baseline model, providing a benchmark against more complex algorithms. This comparative analysis helps assess the trade-off between model complexity and predictive accuracy.

In conclusion, while both SARIMAX and EMA were considered for this project, SARIMAX emerged as the preferred choice due to its robust handling of seasonality, ability to incorporate exogenous factors, optimized parameter tuning, and superior performance in terms of error metrics.

Assumptions :-

The following assumptions were made in order to create the SARIMAX model for Walmart project :-

1. **Temporal Independence:**
   - The assumption that observations at different time points are independent. Each data point's value is not dependent on the previous or future observations.

2. **Stationarity:**
   - Assuming that the time series data is stationary. This implies that the mean, variance, and autocorrelation structure remain constant over time.

3. **Normal Distribution of Residuals:**
   - Assuming that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is crucial for accurate statistical inference and hypothesis testing.

4. **No Autocorrelation in Residuals:**
   - Assuming that there is no correlation between the residuals at different time points. This is essential for ensuring that the model captures the underlying patterns in the data.

5. **Exogeneity of Exogenous Variables:**
   - Assuming that the exogenous variables included in the model are not influenced by the time series itself. Exogenous variables are considered external factors and should not exhibit endogeneity.

6. **Adequate Model Fit:**
   - Assuming that the SARIMAX model selected provides an adequate representation of the underlying patterns and dynamics in the time series data. This involves proper model selection, parameter tuning, and validation.

7. **Appropriate Inclusion of Exogenous Variables:**
   - Ensuring that the chosen exogenous variables are relevant to the time series being modeled. Including irrelevant or non-informative exogenous variables can lead to overfitting or suboptimal model performance.

8. **Sensible Seasonal and Non-seasonal Parameters:**

- Carefully selecting appropriate values for seasonal and non-seasonal parameters (p, d, q, P, D, Q) based on the observed characteristics of the time series. This requires understanding the seasonal patterns and trends in the data.

Model Evaluation and Technique :-

Techniques and Steps Involved in Model Evaluation:

1. **Residual Analysis:**
    - Conducted a thorough analysis of residuals to assess the model's ability to capture patterns and trends in the data. Examined the distribution, autocorrelation, and heteroscedasticity of residuals.

2. **Accuracy Metrics Calculation:**
    - Calculated relevant accuracy metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to quantitatively assess the model's predictive performance.

3. **Backtesting:**
    - Utilized backtesting techniques to assess the model's performance on historical data that was not used during the training phase. This provides insights into the model's ability to generalize to unseen data.

4. **Rolling Forecast Origin:**
    - Employed a rolling forecast origin methodology, where the model is trained on a growing dataset, and predictions are made for subsequent time points. This technique helps evaluate the model's performance as it adapts to new observations.

5. **Comparative Analysis:**
    - Conducted a comparative analysis between the SARIMAX model and alternative models, such as Exponential Moving Average (EMA), to assess the relative effectiveness and advantages of the chosen model.

The evaluation report suggests the following:

1. **Model Accuracy:**
    - The SARIMAX model demonstrated high accuracy, as evidenced by low values of MAE, MSE, and RMSE. The model's predictions closely aligned with the actual weekly sales data.

2. **Residual Analysis Insights:**

- Residual analysis indicated that the model effectively captured the underlying patterns in the time series data. The residuals displayed no significant autocorrelation, and their distribution was consistent with the assumption of normality.

3. **Backtesting Results:**
   - Backtesting results supported the model's robustness, showing consistent performance on historical data that was not part of the training set. This suggests that the model generalizes well to new observations.

4. **Rolling Forecast Origin Performance:**
   - The rolling forecast origin technique revealed the model's adaptability to changing patterns over time. Predictions remained accurate as the model continuously updated with new data.

5. **Comparative Analysis Findings:**
   - Comparative analysis with alternative models confirmed that SARIMAX outperformed simpler models like Exponential Moving Average, particularly in capturing and predicting seasonality and external factors' impact on sales.

Inferences from the Project :-

1. **Model Performance:**
    - The SARIMAX model, chosen for its ability to handle seasonality and incorporate exogenous factors, demonstrated superior performance in accurately predicting weekly sales for Walmart stores.
2. **Accurate Sales Predictions:**
    - The model consistently provided accurate predictions, as evidenced by low Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values. This suggests a close alignment between predicted and actual sales data.
3. **Effective Handling of Seasonality:**
    - SARIMAX's capability to handle seasonality was crucial in capturing the cyclic patterns inherent in weekly sales data. The model's ability to adjust for seasonal fluctuations contributed to its accuracy.
4. **Incorporation of Exogenous Variables:**
    - The inclusion of exogenous variables such as holidays, temperature, fuel prices, CPI, and unemployment rate enhanced the model's predictive power. These external factors were identified as meaningful contributors to weekly sales variations.
5. **Residual Analysis Insights:**
    - Residual analysis confirmed that the SARIMAX model effectively captured underlying patterns in the data. The residuals displayed no significant autocorrelation, and their normal distribution validated the appropriateness of the chosen model.
6. **Backtesting Robustness:**
    - Backtesting results supported the model's robustness, showcasing consistent performance on historical data not used during training. This indicates the model's ability to generalize well to unseen observations.
7. **Adaptability to Dynamic Patterns:**
    - The rolling forecast origin technique demonstrated the model's adaptability to changing patterns over time. Predictions remained accurate as the model continuously updated with new data, showcasing its dynamic nature.

8. **Comparative Analysis Validation:**
    - Comparative analysis with alternative models, including Exponential Moving Average (EMA), underscored SARIMAX's superiority. The chosen model outperformed simpler models, particularly in capturing and predicting seasonality and external factors' impact on sales.
9. **Overall Model Reliability:**
    - The overall evaluation of SARIMAX's performance, encompassing various techniques and analyses, concludes that the model is reliable and well-suited for the time series forecasting task in the Walmart weekly sales project.

Future Possibilities :-

Future Possibilities:

1. **Model Refinement:**
   - Continuous refinement of the SARIMAX model by exploring additional hyperparameter tuning and optimization techniques may lead to further improvements in predictive accuracy.
2. **Incorporation of Additional Features:**
   - Consideration of additional relevant features or external factors that could impact weekly sales may enhance the model's ability to capture more nuanced patterns and trends.
3. **Advanced Time Series Models:**
   - Exploration of more advanced time series models, such as deep learning-based approaches (e.g., Long Short-Term Memory networks), could be considered for comparison to assess whether more complex models offer additional benefits.
4. **Ensemble Models:**
   - Experimentation with ensemble models, combining the strengths of multiple forecasting algorithms, may provide a synergistic effect, potentially improving overall performance.
5. **Dynamic Updating of Exogenous Variables:**
   - Implementing a system for dynamically updating exogenous variables in real-time could improve the model's adaptability to changing economic and environmental conditions.

Limitations:

1. **Limited Historical Observations:**
   - The dataset contains 45 stores, with only 143 rows of weekly sales data for each store. While there are no null or duplicate values, the limitation lies in the relatively short time frame covered by each store's data (143 weeks). This constraint may impact the model's ability to capture and generalize long-term trends, seasonality, or respond effectively to evolving patterns in consumer behavior over an extended period.
2. **External Factors Uncertainty:**

- The model's reliance on external factors, such as economic indicators, introduces a degree of uncertainty. Changes in these factors that are not accounted for in the model may affect predictions.

3. **Assumed Stationarity:**
    - The assumption of stationarity in the time series data might not hold true in all cases. Adapting the model to handle non-stationary data trends could be an area for improvement.
4. **Exogenous Variable Assumptions:**
    - The assumption of exogenous variable independence from the time series may not always be valid. Careful consideration and validation of the independence assumption are essential.
5. **Model Interpretability:**
    - SARIMAX's complexity might limit its interpretability for stakeholders who are not well-versed in time series modeling. Developing visualization tools or simpler models for communication could address this limitation.

Conclusion :-

Through a thorough analysis of weekly sales data, incorporating statistical methods, exploratory data analysis, and predictive modeling, this project has uncovered valuable insights for Walmart. The impact of the unemployment rate on sales was assessed, highlighting specific stores more susceptible to fluctuations. Seasonal trends, influenced by various factors, were identified, and the correlation between temperature and sales was explored. Consumer Price Index effects on sales were examined across stores. The project also identified top-performing and underperforming stores, emphasizing the significance of performance disparities. Finally, predictive modeling provided sales forecasts for the next 12 weeks, offering actionable insights for strategic decision-making and operational planning. Overall, this project equips stakeholders with a holistic understanding of sales dynamics, contributing to informed decision-making processes.

References :-

1. Sessions of intellipaat on Time Series Forecasting
2. Documentation of SARIMAX model
3. Youtube Videos