Capstone Project

Report

Walmart Sales Forecasting

Project By: Krishnendu Mukherjee

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.

Project Objective

1. You are provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

b. If the weekly sales show a seasonal trend, when and what could be the reason?

c. Does temperature affect the weekly sales in any manner?

d. How is the Consumer Price index affecting the weekly sales of various stores?

e. Top performing stores according to the historical data.

f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

2. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

Data Description

The walmart.csv contains 6435 rows and 8 columns.

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 Sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

Data Preprocessing Steps And Inspiration

1. **Loading the Dataset:** The initial step involves loading the dataset into the analysis environment, typically using libraries like Pandas in Python, ensuring accessibility for further examination.

2. **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.

3. **Converting Date Column:** When dealing with temporal data, such as dates, converting the date column from an object type to a date type facilitates time-based analyses and visualizations.

4. **Handling Outliers:** Identification and treatment of outliers are crucial to maintain data integrity. Outliers are assessed visually through plots or statistically using methods like z-scores or IQR (Interquartile Range), ensuring their handling aligns with the context and domain knowledge.

5. **Correlation Analysis:** Utilizing tools like heatmaps aids in understanding the interrelationships between various features within the dataset, providing insights into potential dependencies and guiding further exploration.

6. **Exploring Relationships:** Beyond correlation analysis, exploring relationships between columns through visualizations and statistical methods unveils additional patterns and dependencies, enriching the understanding of the dataset's dynamics.

7. **Time Series Analysis:** Conducting time series analysis involves assessing stationarity through techniques like examining rolling mean and standard deviation, essential for ensuring the reliability of subsequent forecasting models.

8. **Forecasting Models Selection:** Employing forecasting model ARIMA entails a methodical approach based on the dataset's characteristics and the desired level of complexity, enabling accurate prediction of future trends.

These pre-processing steps and methodologies lay a solid foundation for rigorous data analysis and forecasting, contributing to informed decision-making and actionable insights.

Choosing the Algorithm For the Project

The Autoregressive Integrated Moving Average (ARIMA) model is a powerful tool used in time series forecasting. It combines three key components: autoregression (AR), differencing (I), and moving average (MA). The AR part involves regressing the variable on its own lagged values, indicating that past values have a direct impact on future values. The I, or "integrated," component refers to differencing the data to make it stationary, which is essential for accurate modeling. Finally, the MA component models the error of the variable as a linear combination of past error terms.

I have chosen the ARIMA Model for this project for the following reasons:

1. **Trend and Seasonality Capture:** ARIMA effectively captures and models trends and seasonal patterns in sales data, which are crucial for accurate forecasting in sales forecasting.
2. **Flexibility:** The ARIMA model's ability to handle a wide range of time series data, including those with trends and seasonal components, makes it adaptable to different types of sales data.
3. **Stationarity Handling:** ARIMA includes differencing as part of its structure, which helps in transforming non-stationary data into a stationary form, improving the reliability of forecasts.
4. **Error Minimization:** By incorporating past error terms into its moving average component, ARIMA minimizes forecasting errors, leading to more accurate predictions.
5. **Autocorrelation Utilization:** ARIMA leverages autocorrelation in the data, using past values to predict future sales, which is particularly useful when historical sales data shows repeated patterns.
6. **Comprehensive Analysis:** ARIMA models provide insights into the underlying data structure through their autoregressive and moving average components, aiding in better understanding sales dynamics.
7. **Proven Effectiveness:** ARIMA has a long history of successful application in various industries for time series forecasting, providing a robust, well-tested framework for sales prediction.

Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the model:

1. **Forecast Accuracy Metrics:**

   (a) Mean Squared Error (MSE): Measures the average of the squares of the errors, giving more weight to larger errors.

   (b) Root Mean Squared Error (RMSE): The square root of MSE, providing a measure of error in the same units as the data.

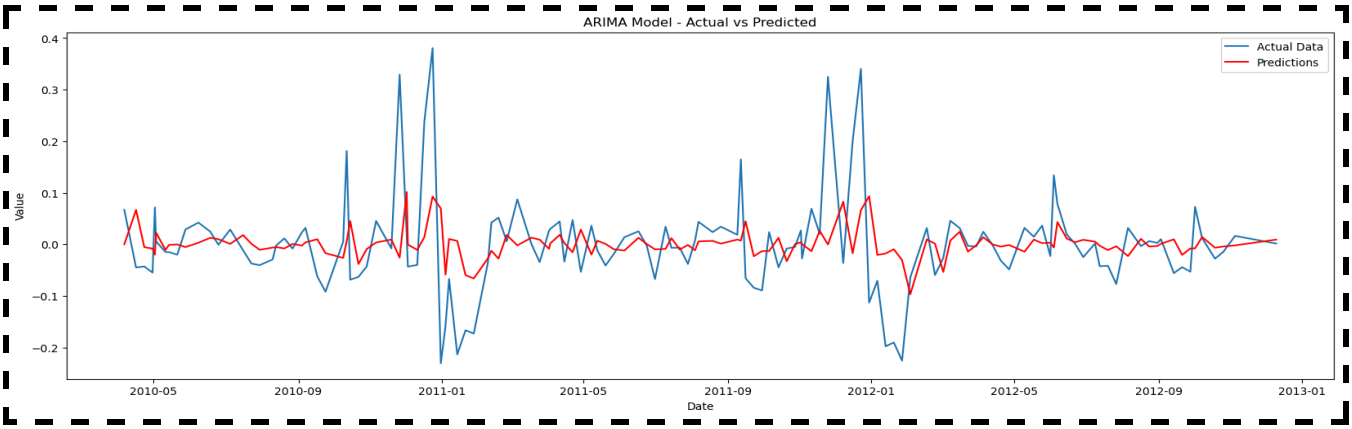
2. **Visual Inspection:**

   (a) Plotting Actual vs. Predicted Values: Visual inspection of how well the predicted values match the actual data can help identify any obvious discrepancies or patterns the model has missed.

The evaluation report suggests the following:

1. The RMSE score of the model considering the sales of all the stores is : *0.09083478416569807* , suggests that the model is performing well in terms of accurately predicting the target variable.

2. Below is the plot for the Actual and the Predicted value.



Actual vs Predicted value

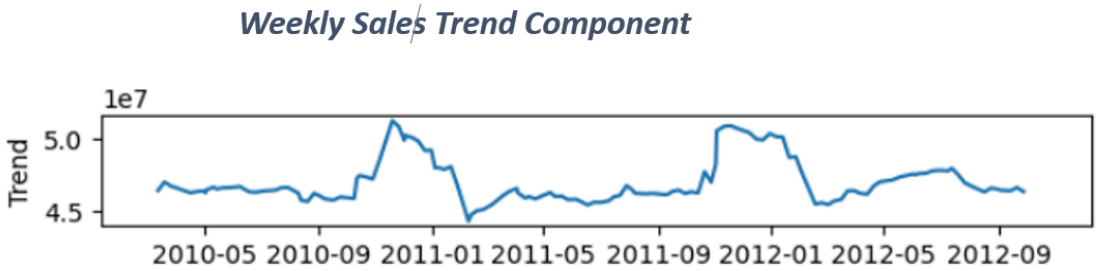
Inferences from the Project

- **If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?**

  Weekly sales are affected by the unemployment rate. The 5 Stores which are suffering the most are:

	Store	Weekly_Sales
7	16	2016067.98
12	33	2299155.24
1	5	14168838.13
17	44	14187373.72
8	17	16232762.69

- **If the weekly sales show a seasonal trend, when and what could be the reason?**



For Weekly Sales data trend, there is a noticeable increase in sales starting from September 2010, followed by a seasonal upward trend until January 2011. Afterward, the sales stabilize, eventually transitioning to a downward trend. This pattern is repeated from September 2011 to January 2012.

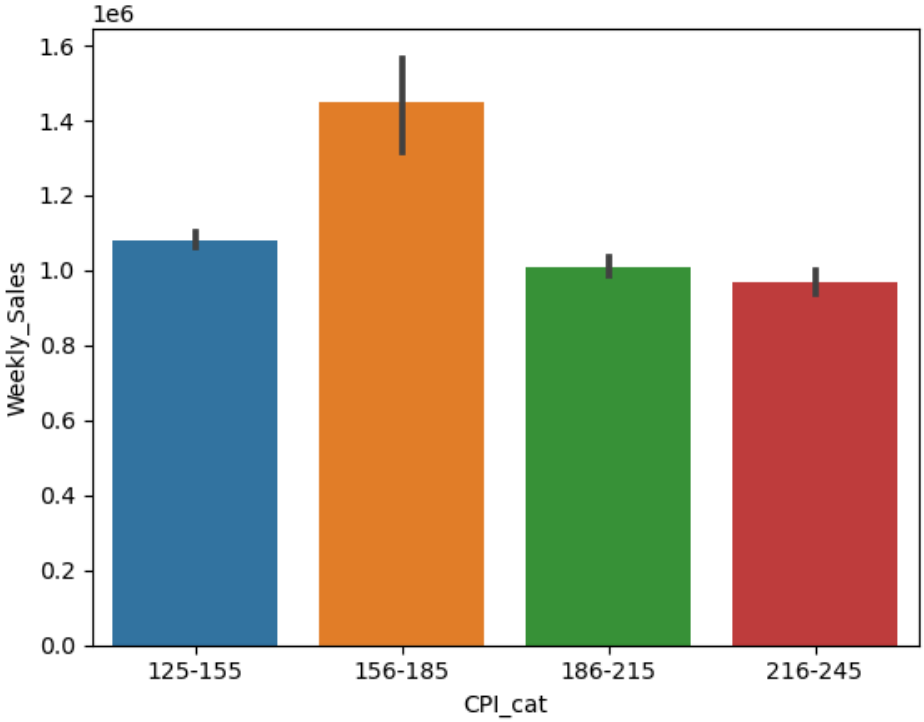
This period includes major holidays such as Thanksgiving, Black Friday, Christmas, and New Year's, which are significant shopping events. Consumers tend to spend more on gifts, decorations, food, and other holiday-related items during these times.Retailers often run extensive sales promotions, discounts, and marketing campaigns during this period to attract customers.Many people receive year-end bonuses or other financial incentives at the end of the year, which can increase their spending power during the holiday season.

- **Does temperature affect the weekly sales in any manner?**

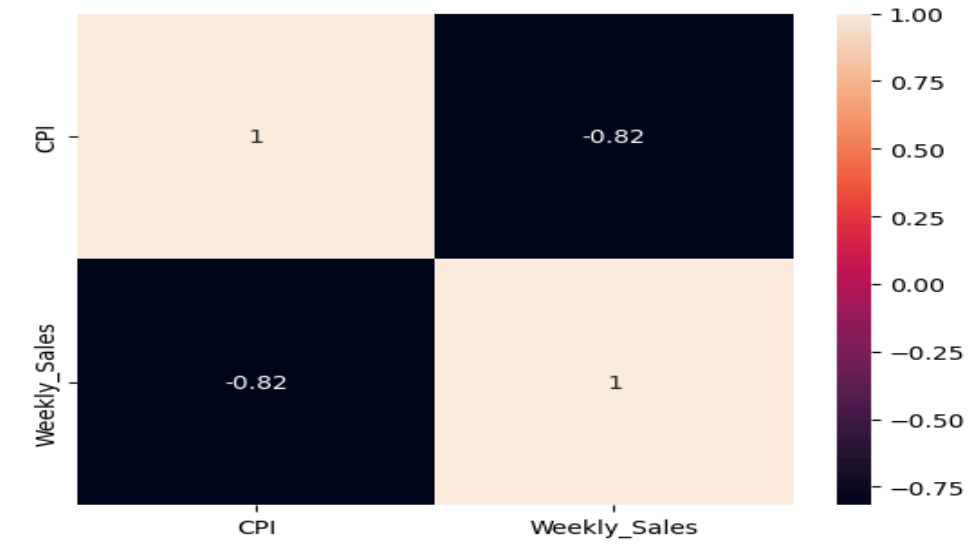
There is a negative correlation between the temperature and weekly sales. Hence when the temperature seems to decrease there is a subsequent increase in sales. There can be increased demand for winter-related products such as clothing, heating equipment, and seasonal items like snow shovels or holiday decorations which are a reason for the rising sales.

- **How is the Consumer Price index affecting the weekly sales of various stores?**

There is a negative correlation between CPI and Weekly Sales we can see in the below provided graph.



We have also verified it using a correlation chart between the CPI and the weekly sales component.



- **Top performing stores according to the historical data.**

Stores	Sales
20	*2.990663e+08*
4	*2.973575e+08*
14	*2.870091e+08*
13	*2.845247e+08*
2	*2.741628e+08*
10	*2.692807e+08*
27	*2.534795e+08*
6	*2.237489e+08*
1	*2.224028e+08*
39	*2.074455e+08*

- **The worst performing store, and how significant is the difference between the highest and lowest performing stores.**

Stores	Sales
33	*37160221.96*
44	*43293087.84*
5	*45475688.90*
36	*53412214.97*
38	*55159626.42*
3	*57586735.07*
30	*62716885.12*
37	*74202740.32*
16	*74252425.40*
29	*77141554.31*

There is a significant differnce between the top and bottom performing stores.
p-value: 7.079291347377579e-13

Significant Mean difference value between groups: 203807667.661249

There are few other inferences too

1. *The sales are lower in holidays as compared to the working days*

2. *When the fuel prices are moderate the performance of stores are also moderately high. But the fuel Price went above the moderate level that is 3.5, the sales were also High. The fear of Inflation can be a driving force where people might have purchased a large quantity of daily need grocery and goods. When fuel price went even high the sales went down which clearly projects that because of Inflation , consumers has reduced purchase of commodities.*

Conclusion

Based on the analysis conducted on the dataset, it can be concluded that the ARIMA model is the most suitable for forecasting purposes.

1. *Its simplicity and ability to effectively smooth out noise make it particularly well-suited for this dataset.*

2. *Its straightforward methodology and robust performance make it a reliable choice for generating forecasts.*

3. *And the model accuracy was also quite good.*

4. *Model Root Mean Squared Error for all the stores is 0.09083478416569807*

5. *Model Root Mean Squared Error for Store 1 is 0.11624010402978888*

6. *Model Root Mean Squared Error for Store 6 is 0.10223844724990286*

7. *Model Root Mean Squared Error for Store 11 is 0.1049818692897262*

References

1. Open AI. (n.d.). ChatGPT.