

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. Motivation and reasons for choosing the algorithm
7. Assumptions
8. Model evaluations and techniques
9. Inferences from the same
10. Future possibilities of the project
11. Conclusion
12. References

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

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:

- If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
- If the weekly sales show a seasonal trend, when and what could be the reason?
- Does temperature affect the weekly sales in any manner?
- How is the Consumer Price index affecting the weekly sales of various stores?
- Top performing stores according to the historical data.
- The worst performing store, and how significant is the difference between the highest and lowest performing stores.

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

Data Description

Feature Name	Description
Store Number	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 Pre-processing 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.

Choosing the algorithm for the project

I am employing ARIMA model to conduct forecast analysis on the provided dataset. These methods offer diverse perspectives for predicting future trends and patterns within the data. ARIMA integrates autoregressive, moving average, and differencing components to capture complex temporal dynamics. By leveraging these techniques, I aim to generate accurate forecasts that inform decision-making and strategic planning. The model facilitates a comprehensive understanding of the dataset's predictive capabilities.

Motivation and reasons for choosing ARIMA MODEL

The choice of algorithm for forecast analysis was motivated by its distinct strengths and suitability for different aspects of the dataset.

Stationarity: ARIMA works well with stationary data or can be made stationary through differencing.

Autocorrelation: ARIMA captures autocorrelation with AR and MA components.
Flexibility: ARIMA handles various patterns like trend and seasonality.

Simple Structure: It's easier to interpret and implement compared to some other models.

Forecasting Accuracy: ARIMA provides accurate forecasts for many real-world datasets.

Diagnostic Tools: ARIMA models come with diagnostic tools such as residual analysis, ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots, which help assess the model's adequacy and identify any remaining patterns in the residuals.

Incorporates Lagged Relationships: It effectively captures relationships in time series data by considering past values (lags), enabling accurate forecasting based on historical data.

Useful for Univariate Time Series: ARIMA is particularly effective for univariate time series, where the forecast depends solely on the past values of the same variable.

Automated Model Selection: Techniques like Auto-ARIMA can automatically find the optimal order parameters (p, d, q), simplifying the model-building process.

Assumptions

No Assumptions

Future Possibilities of the Project

For Advance research Exploring Advanced Time Series Models

Inferences from the same

1. Store V/s Weekly Sales (Top 10 and Worst 10 Stores) Top 10 Performance Stores the Stores 4,20,14 is having the highest sales in top 10 stores segment. The Stores 6,39 is having the Lowest sales in top 10 stores segment. Worst Performance 10 Stores the Stores 29,16,37 is having the highest sales in Bottom 10 stores segment The Stores 33,44 is having the highest sales in Bottom 10 stores segment.
2. Holidays V/s Weekly Sales the Sales are High in Normal Days, it is Quite common because of that holiday are always low compared to working days.
3. Date V/s Weekly Sales Weekly Sales most of weeks are consolidating expect some quarters which are in winter season is having seasonal spikes

4. Temperature V/s Weekly Sales Stores perform well in moderate temp that is 26-50. Sales went peaks in Moderate to High temp that is 51-75. Sales went down in Very High and Low temp that is 0-25 & 76-100.
5. Fuel Price V/s Weekly Sales 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, and the sales are also went very High. May be the Inflation and fear of increasing prices people are bought more grocery to mitigate with inflation. and when fuel price went even high the purchase power of people went down. may be the inflation and increasing in goods price.
6. CPI V/s Weekly Sales when inflation is lower the performance of stores are high. in initial time of inflation, the sale was drop significantly. in third level when inflation is high the people may scare for future prices and bought more grocery for future needs also. eventually when the inflation went peaks the sales are went down.
7. Unemployment V/s Weekly Sales When the Unemployment rate is between the 6 to 10 the stores are performing well. But whenever the Unemployment rate is increasing the sales went down.
8. Correlation analysis between variables Quarters and Fuel Price having high correlation. Because the fuel prices are also depended on the seasonal changes.

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

I Consider the Top 5 Suffering Stores Only.

Out[109]:

	Store	Weekly_Sales
14	38	1741896.51
7	16	2016067.98
12	33	2299155.24
5	12	3850386.42
10	28	5026062.83

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

Regarding the Temperature data, a seasonal downward trend is observed starting from September 2010 until January 2011. Subsequently, there's a gradual increase. However, from September 2011, temperatures start to decline again.

In summary, there appears to be an inverse relationship between Temperature and Weekly Sales. When temperatures decrease, weekly sales tend to rise, and vice versa. This inverse correlation is evident from September 2011 to January 2012, as well as from September 2011 to January 2012.

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

There is a negative correlation between the temperature and weekly sales so, the answer for question c: Yes, there is a Negative correlation and the Temperature is affecting the Weekly Sales

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

There is a negative correlation between the CPI and weekly sales so, the answer for question D: Yes, there is a Negative correlation and the CPI is affecting the Weekly Sales we can see that in below graph as well. when the CPI increasing the sales are down

e) Top performing stores according to the historical data.

The Stores 20,14 is having the highest sales in top 10 stores segment

The Store 39 is having the Lowest sales in top 10 stores segment

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

The Store 37 is having the highest sales in Bottom 10 stores segment

The Stores 38 is having the lowest sales in Bottom 10 stores segment

There is a significant difference between top-performing and worst-performing stores.

P-value: 1.0973608139893526e-12

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. And its near to 0
4. model accuracy for all the stores is Root Mean Squared Error 0.04827029470627315
5. model accuracy for Store 1 Root Mean Squared Error 0.04968092258103263
6. accuracy for Store 6 Root Mean Squared Error 0.09157095701937541

References

Open AI, ChatGPT, Google