CAPSTONE PROJECT- WALMART SALES FORECAST

Table of Contents

Problem Statement

- As a data scientist, I have been tasked with addressing a significant challenge facing the retail industry: managing inventory levels in multiple locations to align with demand. To accomplish this, I will utilize my expertise in data analysis and modelling to provide meaningful insights and create predictive models to accurately forecast sales for the next X months/years. This project requires a combination of technical skills, strategic thinking, and the ability to turn data into actionable solutions.

Project Objective

The objective of this project is to:

- Gain valuable insights from available data to improve retail store operations.
- Precisely forecast sales for each retail store for the next 12 weeks.

Data Description

This dataset Walmart.csv contains sales information for multiple retail stores across the country. **The data includes the following variables:**

1.Store: Store number (categorical)

2.Date: Week of Sales (datetime)

3.Weekly_Sales: Sales for the given store in that week (numeric)

4.Holiday_Flag: If it is a holiday week (categorical)

5.Temperature: Temperature on the day of the sale (numeric)

6.Fule_Price: Cost of the fuel in the region (numeric)

7.CPI: Consumer Price Index (numeric)

8.Unemploymnet: Unemployment Rate (numeric)

The following insights from the data:

The dataset consists of 12873 observations and 8 variables. There are 3 missing values and 6435 duplicate values in the data. The distribution of variables is as follows:

1.Store: 45 stores (1,2, 3, …., 45)

2.Date: 5-2-2010 to 26-10-2012

3.Weekly_Sales: [209986 min, 3818686 max, 1046964 average]

4.Holiday_Flag: 0,1

5.Temperature: -2.06 to 100.14 avg:60.66

6.Fule_Price: 2.47 to 4.46 avg:3.35

7.CPI: 126.06 to 227.23 avg:171.57

8.Unemploymnet: 3.87 to 14.83 avg:7.87

This dataset provides valuable information for analysing sales trends and customer behaviour, and can be used to improve inventory management and marketing strategies.

datadescription.html provides the data description report.

4. Data Pre-processing Steps and Inspiration

The pre-processing of the data included the following steps:

1. Data Cleaning:
    - Handling Missing data:
      Check null values
      Drop null values as there are only 3 null values.
    - Handling Duplicate values:
      Check duplicated values
      Drop duplicate values -6435 duplicates are there.

2. Data Transformation:
    - Convert the datatype of the variables:
      Previously all variables are Object it is then converted as int, float and DateTime respectively.
    - Splitting Date variable:
      Date variable is spited into week, month and year variables

3. Checking Seasonality:
    - Change Date as index and prepare data for tests
    - The ADF (Augmented Dickey-Fuller) test checks for stationarity around a constant mean. If the p-value is less than the significance level (usually 0.05), it means the time series is stationary. In this case, the p-value is 4.84510284992098e-17, which is much less than the significance level, so the test suggests that the time series is stationary.
    - The KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test checks for stationarity around a deterministic trend. If the p-value is less than the significance level, it means the time series is not stationary and has a unit root (i.e., non-stationary around a deterministic trend). In this case, the p-value is 0.1, which is greater than the significance level, so the test suggests that the time series is not stationary.

   Overall, the ADF and KPSS tests give consistent results in this case, with the ADF test suggesting that the time series is stationary, and the

KPSS test suggesting the opposite. This may indicate that the time series has a more complex behaviour than a simple trend.

- Determining the rolling statistics

4. Time series data transformation
    - Step 1: apply log to the data
    - Step 2: subtract the shifted log data from log data it helps to transform a non-stationary time series into a stationary time series, which can then be more easily modelled and forecasted.
    - Step 3: Applying Seasonal decomposition which is a useful step in the time series analysis and forecasting process, as it can help to better understand the underlying structure of a time series and inform the choice of appropriate modelling and forecasting methods.

Choosing the Algorithm for the Project

I have chosen SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous predictors) for time series forecasting.

Model specification:T he order argument is set to (4,0,1), meaning the model includes 4 terms for autoregression (AR), 0 terms for differences (I), and 1 term for moving average (MA). The seasonal_order argument is set to (1,1,1,12), meaning the model includes 1 term for seasonal autoregression (SAR), 1 term for seasonal differences (SI), 1 term for seasonal moving average (SMA), and the seasonality is assumed to repeat every 12 periods.

The enforce_stationarity and enforce_invertibility arguments are set to False, meaning the model will not enforce strict stationarity and invertibility conditions. This allows for more flexibility in model fitting.

The algorithm:

1. Import necessary libraries (pandas, SARIMAX from statsmodels)
2. Define a function "forecast" that takes two inputs: a dataframe (df) and a store number (store_num)
3. Within the function, create a new dataframe (store_data) that only contains data for the specified store number
4. Clean and preprocess the store_data by removing unnecessary columns and setting the index to the date
5. Fit a SARIMAX model to the Weekly_Sales data in store_data
6. Forecast the Weekly_Sales for the next 24 weeks and store the results in a new dataframe (forecast_df)
7. Add the store number to forecast_df and combine with store_data into a final combined_df
8. Return combined_df
9. Define an empty dataframe (df_allstores_forecast) with columns Date, Weekly_Sales, and store_num
10. Loop through all unique store numbers in df and call the forecast function to generate a forecast for each store
11. Concatenate each store's forecast with df_allstores_forecast
12. Sort df_allstores_forecast by store_num and Date
13. Use Plotly to create a line graph of Weekly_Sales over Date, colored by store_num
14. Display the graph and write it to an HTML file.

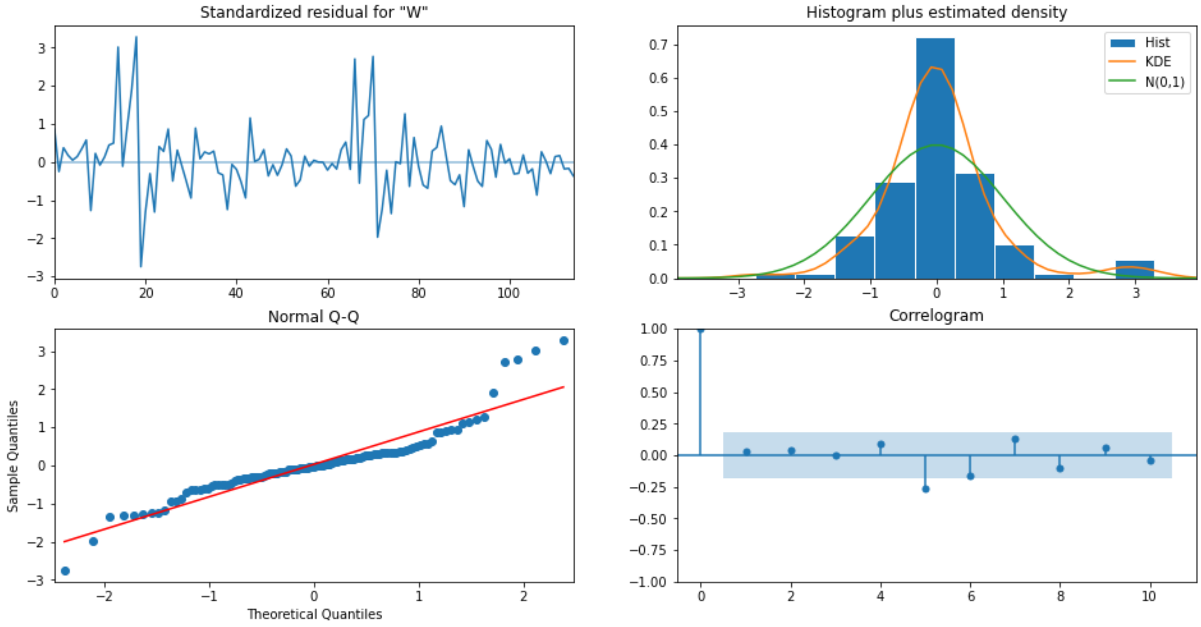
6. Motivation and Reasons for Choosing the Algorithm

I have chosen SARIMAX algorithms instead of ARIMA because:

1. ARIMA: with order of 5,2,2 from acf and pacf plots.
   MAE: 0.0874161804693775
   MSE: 0.017417964064210652
   RMSE: 0.13197713462645963

From autoarima suggestion

2. SARIMAX (4,0,1) and sesasonal_order = (1,1,1,12)



The errors got reduced too
MAE: 0.07481489038100851
MSE: 0.012979221202218399

SARIMAX is a variant of the ARIMA model that can incorporate additional variables (exogenous predictors) to improve forecasting accuracy. It is useful for modelling time series data that exhibit seasonality patterns, where the patterns repeat at regular intervals.

Overall, SARIMAX is a powerful time series forecasting method that can provide accurate results if used correctly.

7. Assumptions

- Stationarity: After performing the logarithmic transformation and subtracting it from its shifted data. Its assumed data to be stationary.

- Linearity: Time series forecasting algorithm is applied assuming linear relationships between variables

8. Model Evaluation and Techniques

- Residual Analysis: Residuals are the differences between the actual values and the predicted values. Residual analysis is used to evaluate the performance of the model and to identify any patterns or trends in the residuals that may indicate a poor fit.

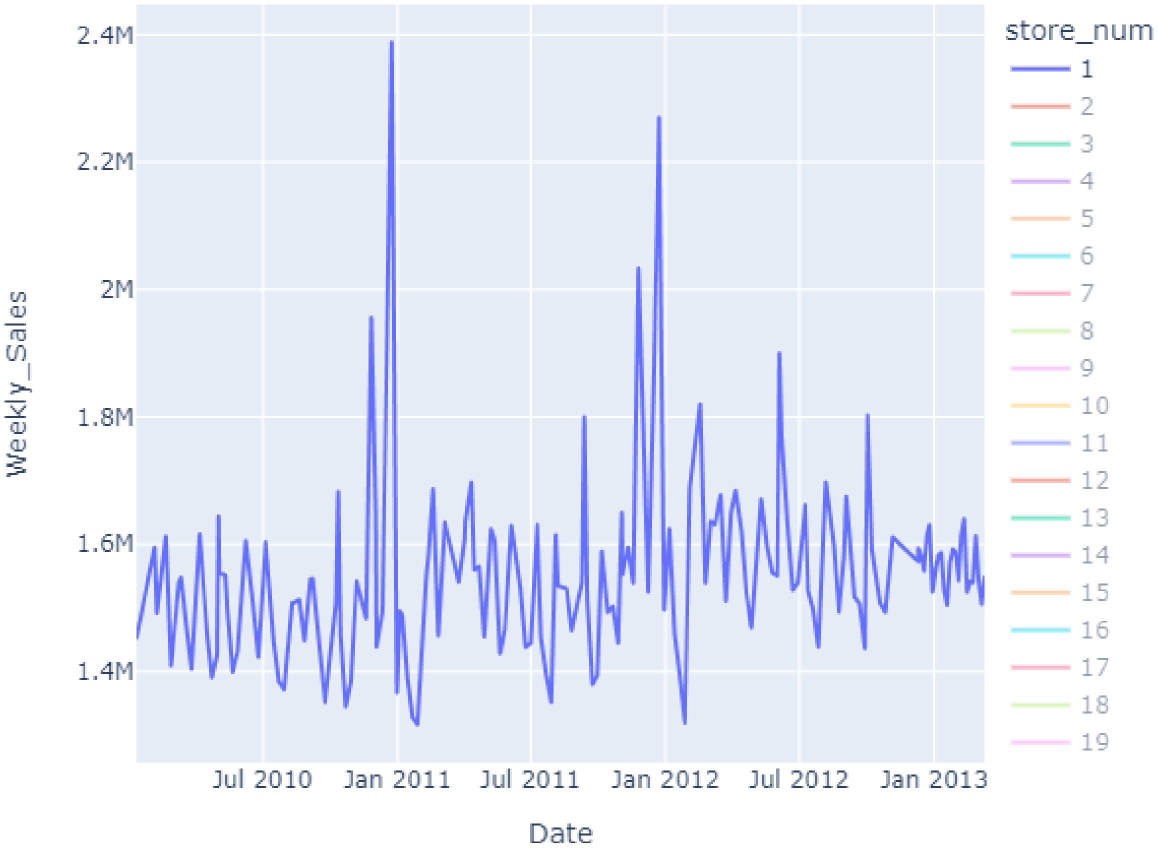
- Mean Absolute Error (MAE): MAE is a measure of the average absolute difference between the actual values and the predicted values. Lower values of MAE indicate a better fit of the model.

  MAE: 0.07481489038100851

- Mean Squared Error (MSE): MSE is a measure of the average squared difference between the actual values and the predicted values. Lower values of MSE indicate a better fit of the model.
  MSE: 0.012979221202218399

- Root Mean Squared Error (RMSE): RMSE is the square root of MSE and is a measure of the average difference between the actual values and the predicted values. Lower values of RMSE indicate a better fit of the model.
  RMSE: 0.1139263850133866

9. Inferences from the Same

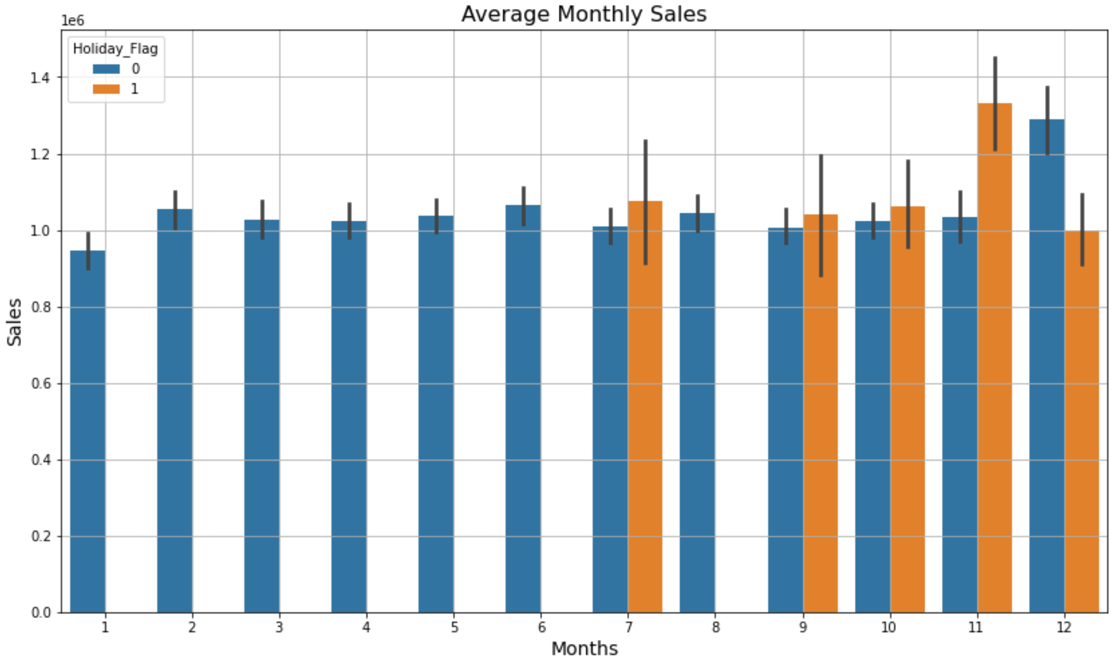


There are 45 retail stores and each store of 12 weeks sales are forecasted and displayed in StorewiseSalesPrediction.html. To view the individual store double click on store number in the legend. Multiple stores can be selected and compare each other in this visualization.

- Individual store performance: The ability to view individual store sales forecasts provides valuable information for making informed decisions. Store managers can use this information to identify underperforming stores and develop strategies to improve their performance.
- Comparison of store performance: The ability to compare the sales performance of multiple stores can help to identify best practices and areas for improvement. This can inform decision making and drive better results for the retail chain.
- Seasonality and trend analysis: The trend line being consistent with past years highlights the importance of considering seasonality and exogenous

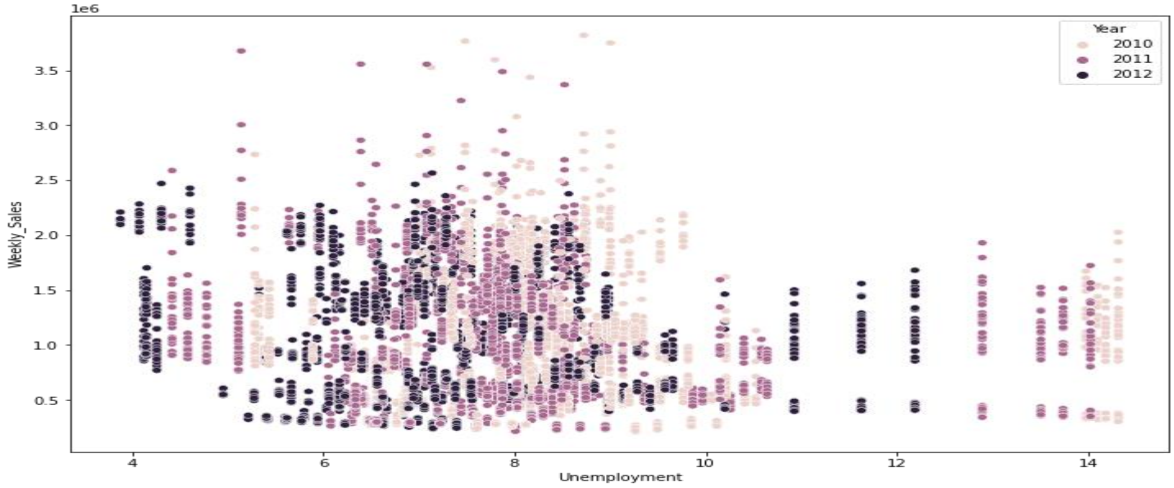
factors in sales forecasting. This information can inform future planning and decision making.

- Short-term sales forecasting: The 12-week sales forecast provides valuable information for short-term planning and decision making. Store managers can use this information to adjust their inventory levels, staffing levels, and promotional activities.

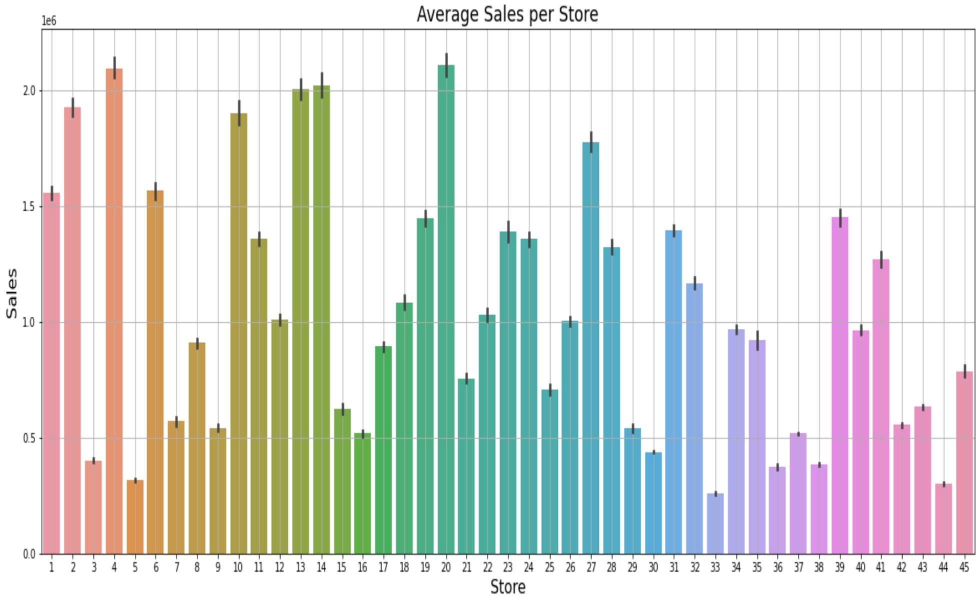


These following insights from the above visualization:

- Maximize holiday season sales: Given the high sales figures recorded in November with holiday flags, businesses can consider ramping up marketing efforts during the holiday season to maximize sales.
- Adjust strategies for December: The contrasting trend in December suggests that businesses may need to tailor their strategies to accommodate the unique buying behaviour during this month.
- Plan for peak sales months: With the data indicating that the latter half of the year is the peak sales period, businesses can plan and allocate resources accordingly to capitalize on this opportunity.
- Consider other factors affecting sales: While the insights provide a broad overview of sales trends, businesses should also consider other factors such as competition, product offerings, and target audience, among others, that may impact their sales.



- Lower the Unemployment higher the sales



To optimize performance analysis, the sales data has been segmented into three categories: high-performing, average-performing, and low-performing stores.

1. The high-performing stores (20, 4, 13, 14, 1, 2, 6, 10, 27) register sales figures above 1.5 million.
2. The average-performing stores (remaining 29 stores) have sales between 1.5 million and 0.5 million.
3. The low-performing stores (3, 5, 30, 33, 36, 38, 44) have sales below 0.5 million.

- Targeted support for low-performing stores: Given the low sales figures of the low-performing stores, businesses can develop a targeted marketing strategy to support these stores and drive sales growth
- Encourage average-performing stores to aim higher: The average-performing stores represent a significant portion of the sales and have the potential to drive growth. Efforts should be made to motivate these stores to achieve sales figures above 1.5 million and consistently improve their performance.
- Recognize and reward high-performing stores: The high-performing stores have consistently demonstrated strong sales performance and should be recognized and rewarded for their efforts. This can include providing additional resources, offering incentives, and promoting best practices from these stores to other stores.
- Evaluate and adjust strategies: Regular evaluations of sales data and performance should be conducted to identify areas for improvement and adjust strategies as necessary. This can involve monitoring the impact of marketing efforts, considering changes in market conditions, and incorporating feedback from store managers and employees.

By utilizing the insights from the model, store managers can make informed decisions, drive better performance, and achieve the desired results for the retail chain.

10. Future Possibilities of the Project

The future possibilities:

- Expansion of the model: The model can be expanded to forecast sales for different products, helping businesses to make informed decisions and drive better performance across the entire retail chain.
- Incorporating additional data: The model can be enhanced by incorporating additional data such as customer demographics, store locations, and weather patterns, among others, to provide more accurate sales predictions.
- Optimizing marketing strategies: The insights generated by the model can be used to optimize marketing strategies, such as promotional activities and advertising efforts, to drive sales growth.
- Integration with other systems: The model can be integrated with other systems such as inventory management, employee scheduling, and budgeting to provide a more comprehensive picture of business performance.

11. Conclusion

In conclusion, the sales forecasting project for Walmart stores provides valuable insights for decision making and performance optimization. The ability to view and compare sales forecasts for individual stores, as well as the short-term 12-week forecast, enables store managers to make informed decisions regarding inventory levels, staffing levels, and promotional activities. The segmentation of sales data into high-performing, average-performing, and low-performing stores helps to identify opportunities for improvement and drive better results. By utilizing the insights from the model, Walmart store managers can improve performance, achieve better results, and ultimately drive success for the retail chain. The analysis of the trend line and consideration of seasonality and exogenous factors highlights the importance of a comprehensive approach to sales forecasting, and the recognition and reward of high-performing stores motivates continued success.

12. References

1. Arunraj, Nari Sivanandam, Diane Ahrens, and Michael Fernandes. "Application of SARIMAX model to forecast daily sales in food retail industry." International Journal of Operations Research and Information Systems (IJORIS) 7.2 (2016): 1-21.

2. Ampountolas, Apostolos. "Modeling and forecasting daily hotel demand: A comparison based on sarimax, neural networks, and garch models." Forecasting 3.3 (2021): 580-595.

3. Cools, Mario, Elke Moons, and Geert Wets. "Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations." Transportation research record 2136.1 (2009): 57-66.

4. Shumway, Robert H., et al. "ARIMA models." Time Series Analysis and Its Applications: With R Examples (2017): 75-163.

5. Demir, Volkan, Metin Zontul, and İlkay Yelmen. "Drug Sales Prediction with ACF and PACF Supported ARIMA Method." 2020 5th International Conference on Computer Science and Engineering (UBMK). IEEE, 2020.

6. Hagan, Martin T., and Suzanne M. Behr. "The time series approach to short term load forecasting." IEEE transactions on power systems 2.3 (1987): 785-791.

7. Sial, Ali Hassan, Syed Yahya Shah Rashdi, and Abdul Hafeez Khan. "Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python." International Journal 10.1 (2021).

8. Podo, Luca, and Paola Velardi. "Plotly. plus, an Improved Dataset for Visualization Recommendation." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.