Capstone project - Walmart Sales forecasting

Name: Noahchristy

Email: noahfrancis2000@gmail.com

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 Evaluation 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:

The main objective of this project is to forecast the future sales for the next 12 weeks of the Walmart stores in various locations. Also we have to figure out the best performing store along with the worst performing stores according to the historical data.Also we need to figure out that How is the Consumer Price index affecting the weekly sales of various stores. To check whether the temperature affects the weekly sales.

Data Description:

By looking at the description we can see that we are provided with a Walmart data set which has various columns with feature names as Store, date, weekly sales, Temperature etc.

By performing Exploratory Data analysis and various hypothesis test we can infer several notable information.

We can clearly see that the Temperature ratio at various stores affects the sales.

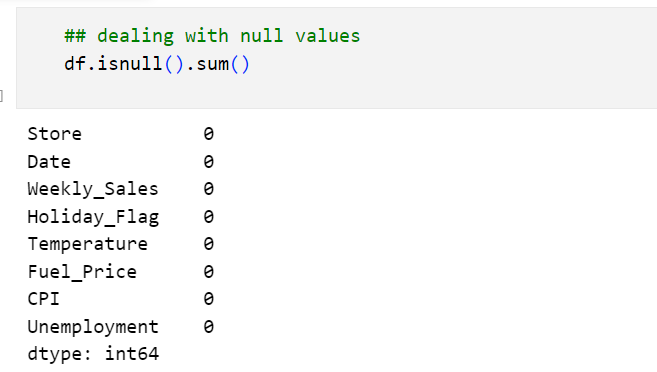
Whenever the temperature was registered as high the sales on that week is declining.

Also the high unemployment rate shows very less sales in that week.

Data Preprocessing Steps And Inspiration:

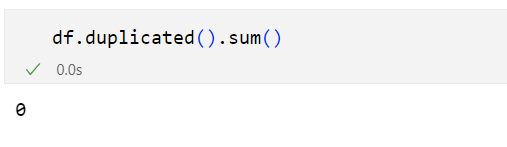
The general data pre-processing steps I.e, “Exploratory Data Analysis” which comprises of the following steps.

1. Dealing with null values:



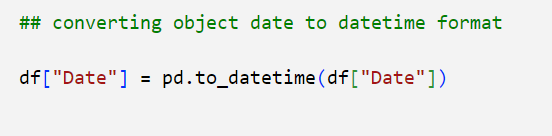
From the image we can clearly see that the dataset has no null values. So that we can proceed to the next step of the EDA process.

1. Dealing with Duplicates:



From the above image we can see that the provided dataset has zero duplicated values.

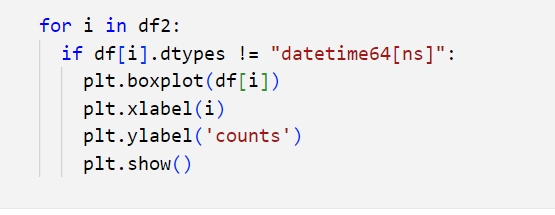
1. Data type conversion:



So we were in the need to convert ‘object’ weekly sales column to datetime format. So that we can proceed to further forecasting processes.

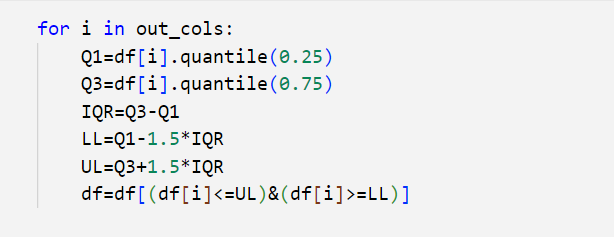
1. Dealing with Outliers:

Outliers detection:



So the above image shows the actual code to detect outliers in the dataset. To detect the outliers visually we primarily use Box plot. It basically shows the values which are out of the actual range.

Removal of Outliers:



Here we took the help of Inter Quartile Range (IQR) to remove the outlaid values from the dataset. We take Quartile 1 range from 25% and Quartile 3 range ends with 75%. We can change this range according to the size of our dataset.

Choosing the Algorithm For the Project:

Time Series Analysis:

Time Series Analysis is a way of studying the characteristics of the response variable concerning time as the independent variable. To estimate the target variable in predicting or forecasting, use the time variable as the reference point. TSA represents a series of time-based orders, it would be Years, Months, Weeks, Days, Horus, Minutes, and Seconds.

**I have chosen the Time series forecasting algorithm for this project for the following reasons:**

1.This capstone Walmart project deals with the future prediction of sales of each store in a weekly manner.

2.By looking at the problem statement and dataset which were provided we can clearly say it deals with the specific interval of time.

3.So for any datasets which requires a future predictions on the time basis - it may be either minutes, hours, days, months, or years.

1. Time series analysis suits the best for this Walmart sales prediction process.

Assumptions:

The following assumptions were made in order to create the model for XYZ project.

1. To perform the forecasting process we first need to change the date column into the index.
2. We just require the date which is converted into index and the weekly sales column to proceed further for the forecasting.
3. Time series data points are dependent on their past values. This means that the order of observations matters, and values at previous time steps can influence the values at current and future time steps.
4. Time series models, especially those involving complex patterns or long-term trends, require sufficient data points to make meaningful predictions and estimates.
5. Depending on the context, you might need to consider seasonal and calendar effects, such as holidays, day of the week, and month of the year.

These assumptions and considerations provide a foundation for approaching time series analysis effectively and accurately.

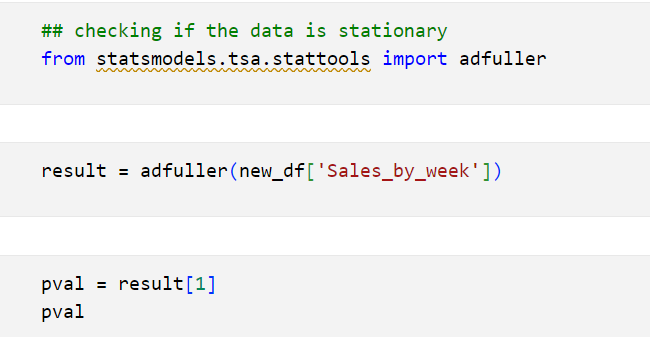
Model Evaluation and Technique:

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

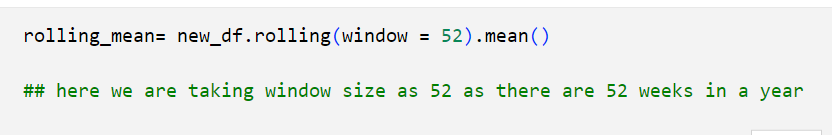
Checking whether the data is Stationary:

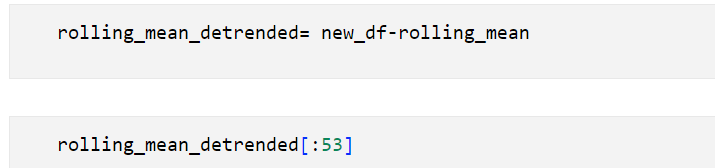
To check whether the data is stationary or not, we seek the help of adfuller.

By performing adfuller function we can get to know that the p value of the dataset is below 0.05 or not. If the p value is below 0.05 the data is considered to be not stationary.

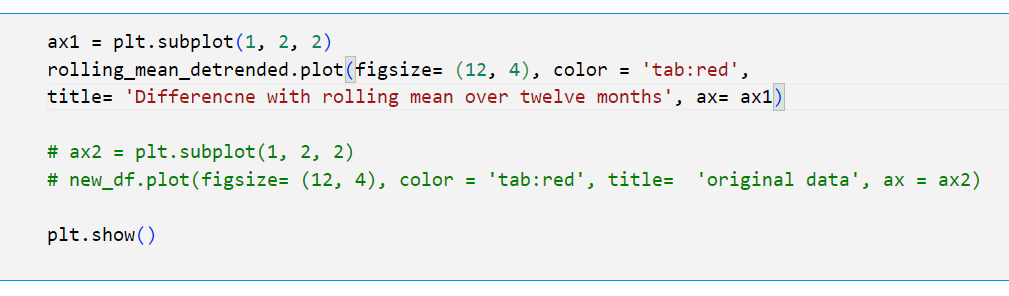


To make the data stationary we need to remove the trend and seasonality from the data. For that first we must standardize mean and variance.

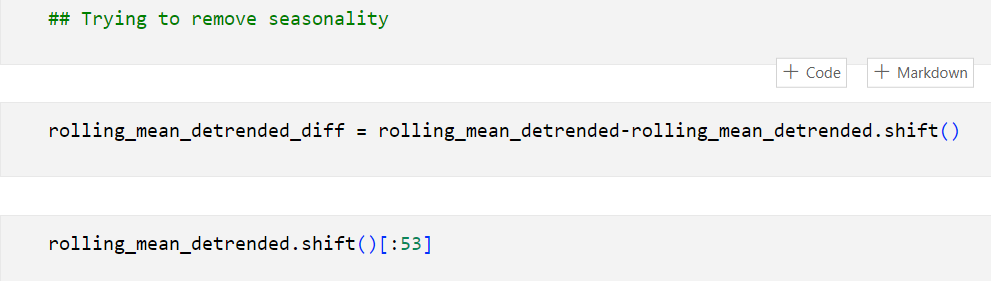




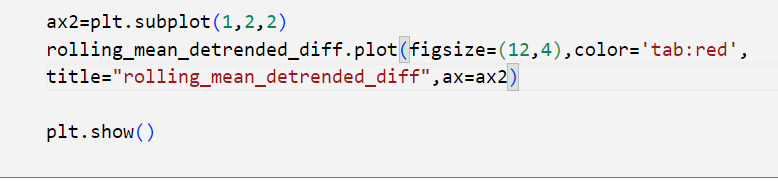
Next we plot the data to check whether the trend is removed or not.



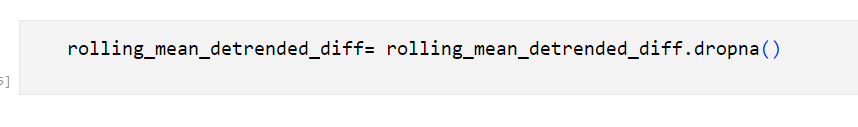
After the removal of the trend, we now now try to remove seasonality.

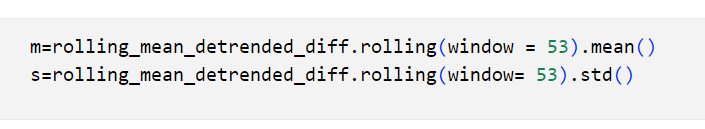


Then we are using subplot to visualize it.



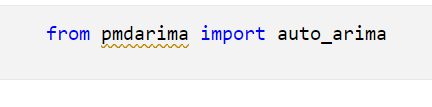
After that we tend to drop the Nan values.



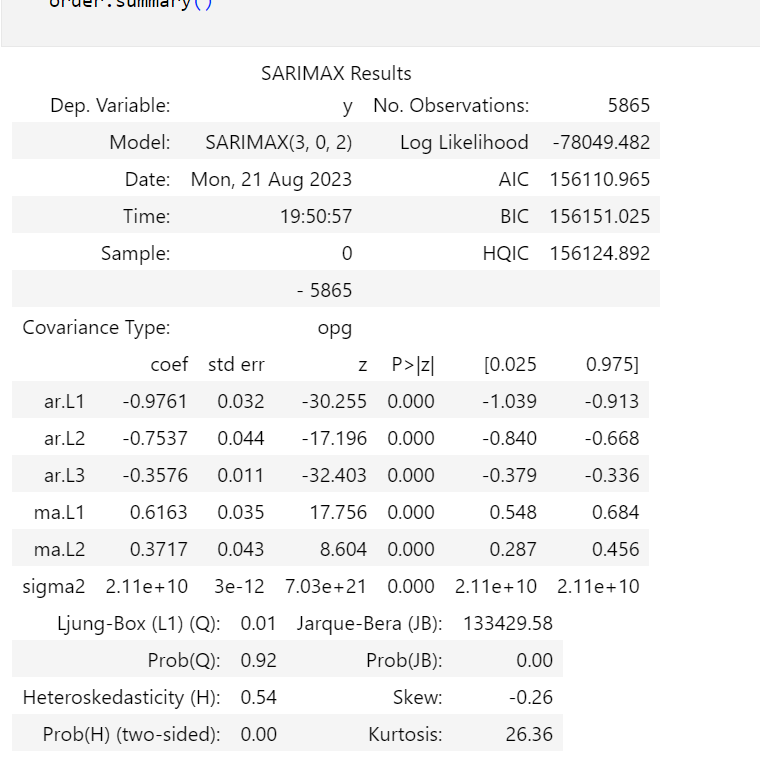


From here the actual model building process starts.

At first we should import the auto\_arima function from pmdarima library.

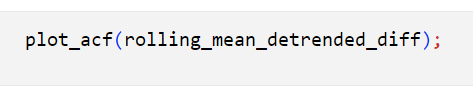


After importing we perform auto\_arima function to get SARIMAX results.

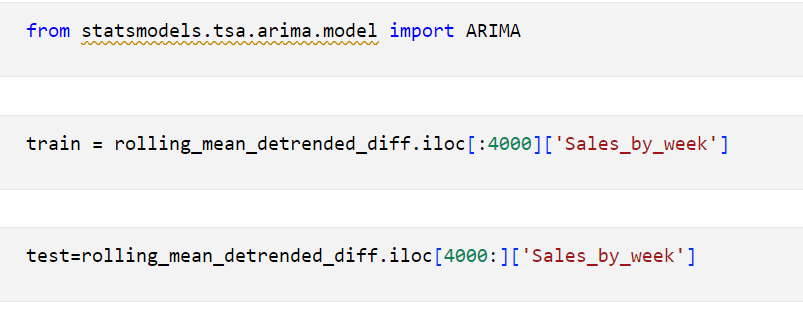


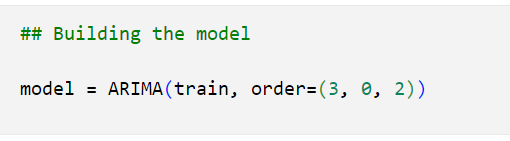
After this we are about to do Auto correlation Function process.

So we are importing required libraries.

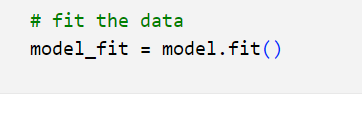


Now the ARIMA model building starts.

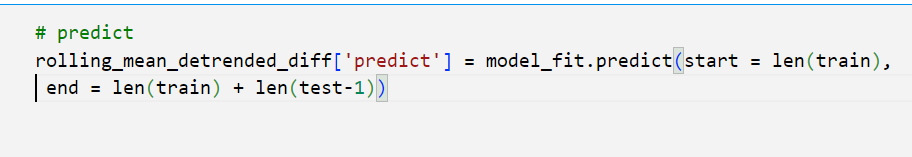




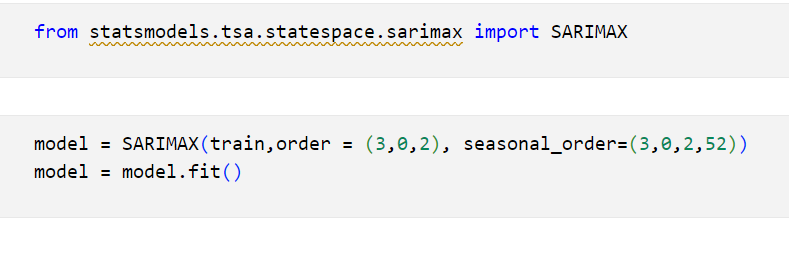
Now the model building has initiated, we are going to fit the data.

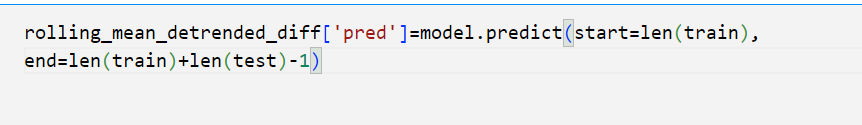


After fitting the data,the prediction process starts.

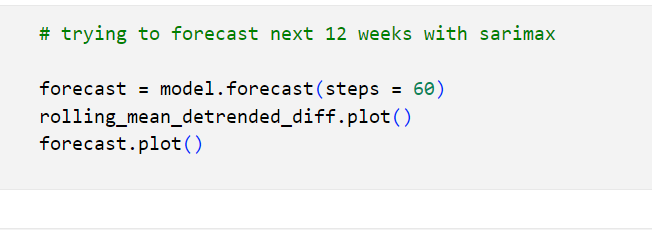


Now building the model with SARIMAX





Now using SARIMAX model we predicting the sales for the next 12 weeks.



Inferences from the Project:

In this model preparation and building we tried to analyse the future sales of various Walmart store by considering various factors.

* The unemployment rate do have an impact in the weekly sales, either holiday week or non holiday week.
* The dataset have more multicolinearity which we tend to deal with it.
* And the Temperature also have the impact on the weekly sales.
* In the EDA process we have witnessed many outliers and that were dealt.
* In the model building process also we tend to face many difficulties like removing the seasonality and trend.
* The weekly sales varies according to the holiday weeks. By this we can say the sales depends on which type of holiday draws more customers.

Future Possibilities:

* Use the forecasting model to simulate different scenarios and assess the potential outcomes under various conditions. This can help in making informed decisions and planning for different scenarios.
* Implement the model in a real-time environment to provide continuous and up-to-date forecasts as new data becomes available. This is particularly useful for dynamic markets and industries.
* Regularly update and refine the forecasting model as new data becomes available. This can involve retraining the model, incorporating new features, or trying different algorithms to improve accuracy.
* Use the forecasts to optimize resource allocation, inventory management, staffing levels, or other business processes. This can help in maximizing efficiency and minimizing costs.
* Contribute to research and innovation in the field of time series forecasting. Explore new algorithms, techniques, and methodologies to push the boundaries of what's possible.
* Time series models are based on historical data, and they might not capture sudden, unexpected events or changes in the future that could significantly impact the forecast.

Conclusion:

In conclusion, our time series forecasting model for Walmart sales prediction has provided valuable insights into future sales trends. Through rigorous analysis and model evaluation, we have demonstrated the model's ability to capture historical patterns and generate accurate forecasts. While the model serves as a powerful tool for decision-making, it's essential to remain mindful of its limitations, such as uncertainties in forecasting unforeseen events and the need for regular validation. By integrating the model's predictions into the decision-making process, we can optimize inventory management, staffing, and marketing strategies, ultimately driving more informed and successful business outcomes. As we move forward, we recommend continuous model monitoring and exploration of advanced techniques to further enhance accuracy and adaptability.

References:

Analytics vidhya website for some common references of the model

-- <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/>

Intellipaat recorder sessions.