Capstone Project – TIME SERIES

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.

TASK TO BE PERFORMED

1. **Using the data, come up with useful insights that can be used by each of the stores to improve in various areas.**
2. **Forecast the sales for each store for the next 12 weeks.**

Project Objective

The objective of this project is to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

MACHINE LEARNING

- **Machine learning** (**ML**) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks

TYPES OF MACHINE LEARNING

- Supervised Machine Learning
- Unsupervised Machine Learning

Time series forecasting can be framed as a **supervised learning problem**. This re-framing of your time series data allows you access to the suite of standard linear and nonlinear machine learning algorithms on your problem

Data Description

The dataset available in the below link

https://drive.google.com/drive/u/0/folders/14wWJMYsD5lSyiZZ8xqvEK3rEGXdjGL6-

Data description, various insights from the data.

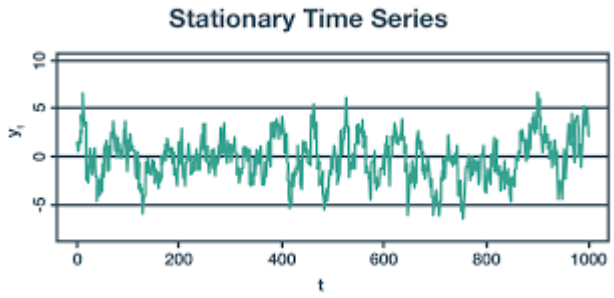
The walmart.csv contains 6435 rows and 8 columns

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

Data Preprocessing Steps And Inspiration

The preprocessing of the data included the following steps:

- In **Time Series** that dependent column can be forecasted with the respect to time. So Date column must be index column.
- **Imputation** – Checking the null values and drop it. Because there must be no null values. This must be unique so there is no way filling the missing values
- Performing **Univariate Analysis**
- To build the time series the data must be **Stationarity. Stationary data** refers to **the time series data that mean and variance do not vary across time**. The data is considered non-stationary if there is a strong **trend or seasonality** observed from the data.



- In this problem they asked to forecast the next 12 weeks' sales for each Store. But here totally 45 Stores are there. It's practically not possible to forecast 45 Stores. So we took top 5 stationarity Stores to forecast.
- There are two ways to check **Stationarity. Descriptive** and **Inferential** Statistics.
- In **Descriptive Statistics, the trend and seasonality of the given data is plotted with help of seasonal decompose. In seasonal decompose model** as two types**, additive and multiplicative**
- **Augmented Dickey–Fuller(ADF), Kwiatkowski–Phillips–Schmidt–Shin(KPSS)** are the two ways of **inferential statistics** to check **Stationarity.**

Choosing the Algorithm For the Project

Description for the XYZ algorithm for the project.

Types of Algorithms in Time Series

1. Autoregressive (AR)
2. Autoregressive Integrated Moving Average (ARIMA)
3. Seasonal Autoregressive Integrated Moving Average (SARIMA)
4. Exponential Smoothing (ES)
5. XGBoost
6. Prophet
7. LSTM (Deep Learning)
8. DeepAR
9. N-BEATS
10. Temporal Fusion Transformer (Google)

In this data we are using **Autoregressive Integrated Moving Average (ARIMA).**

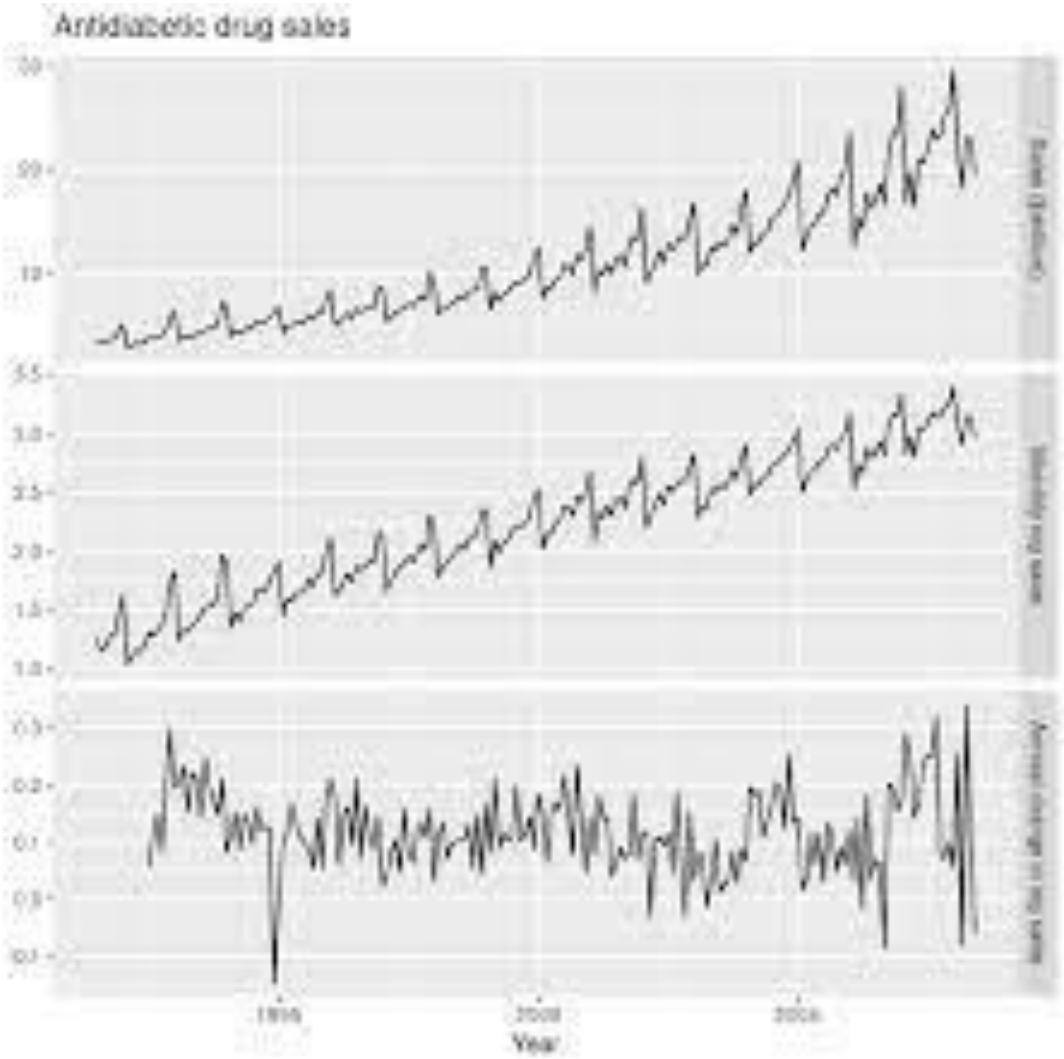
Autoregressive Integrated Moving Average (ARIMA)

- ARIMA models are generally denoted as **ARIMA (p,d,q) where p is the order of autoregressive model, d is the degree of differencing, and q is the order of moving-average model**. ARIMA models use differencing to convert a non-stationary time series into a stationary one, and then predict future values from historical data.
- The first and the most important step in fitting an ARIMA model is to decide on the order of differencing to make the series stationary. The right approach is to **start with the lowest value of differencing i.e. d=1**.

Assumptions

The following assumptions were made in order to create the model for the **Time Series** project.

- A common assumption in many time series techniques is that **the data are stationary**. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time



Model Evaluation and Technique

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

Techniques

- Before training the model we must find **(p,d,q)** order.
- **A non-seasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:**
- p is the number of autoregressive terms,
- d is the number of non-seasonal differences needed for stationarity, and.
- q is the number of lagged forecast errors in the prediction equation.
- The order can be found by **auto_arima**
- **Code: from pmdarima import auto_arima**
- Splitting the data for train and test. The data not to be taken as **random.** The data get splitted by using **iloc or loc function**

The evaluation report suggests the following:

1. **Inferences from the evaluation**

   The average error is **0.005**

Inferences from the Project

The model performance

Model performance is an assessment of the model's ability to perform a task accurately not only with training data but also in real-time with runtime data when the model is actually deployed through a website or an app.

Analysis of model

Scikit-learn. Metrics

- Mean Squared Error
- Mean Absolute Error

Inferences

- The inferences of this model is we got very lowest Mean Absolute Error.
- That means the data get more stationary
- Condition of time series is to make the data more and more stationary and we will get more accurate forecasted values.

Future Possibilities

The future possibilities

- Forecasting has a range of applications in various industries. It has tons of practical applications including: weather forecasting, climate forecasting, economic forecasting, healthcare forecasting engineering forecasting, finance forecasting, retail forecasting, business forecasting, environmental studies forecasting, social studies forecasting, and more.
- Basically anyone who has consistent historical data can analyze that data with time series analysis methods and then model, forecasting, and predict. For some industries, the entire point of time series analysis is to facilitate forecasting.
- Some technologies, such as augmented analytics, can even automatically select forecasting from among other statistical algorithms if it offers the most certainty.

Limitations

- Time series analysis also suffers from a number of weaknesses, including **problems with generalization from a single study, difficulty in obtaining appropriate measures, and problems with accurately identifying the correct model to represent the data**.

Conclusion

The project got **Lowest Mean Absolute Error.** So we will somewhat accurate forecast Results

- Time series analysis is one of the most important aspect of data analytics for any large organization as it helps in understanding seasonality, trends, cyclicality and randomness in the sales and distribution and other attributes.
- These factors help companies in making a well informed decision which is highly crucial for business

NOTE

Finding (p,d,q) is must for to train the model

There is another way to find (p,d,q) order other than auto arima is by autocorrelation and partial autocorrelation plots

P order - In time series analysis, the partial autocorrelation function (PACF) **gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags**.

Q order - Autocorrelation is **the correlation between two observations at different points in a time series**. For example, values that are separated by an interval might have a strong positive or negative correlation. When these correlations are present, they indicate that past values influence the current value

References

Duke University, *Summary of Rules for Identifying ARIMA Models*

Global Temperature Time Series Data

Holmes, Scheuerell, & Ward, *Applied Time Series Analysis*

Hyndman & Athanasopoulos, *Forecasting Principles and Practice*

Khan, *ARIMA Model for Forescating - Example in R*

Towards Data Science, *The Complete Guide to Time Sereis Analysis and Forecasting*