**Evaluating ARIMA and SARIMAX Models for Effective Time Series Forecasting**

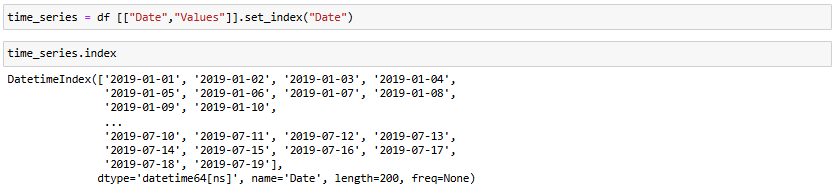
This report aims to forecast future values using the ARIMA and SARIMAX models. By leveraging historical data, the aim is to predict future trends and patterns. The accuracy of the models will be evaluated using the Mean Squared Error (MSE) metric and SARIMAX results.

The process begins by importing the necessary libraries: pandas, numpy, and matplotlib. A sample time series dataset is generated, starting from January 1, 2019, spanning 200 days. The time series values are created with a linear trend, incorporating random noise to simulate real-world behaviour. Additionally, columns such as Category, Quantity and Random are included to represent other variables that may exist in the dataset but are irrelevant for forecasting purposes. The data is then stored in a pandas Data Frame, and the first five rows are displayed to check if the data looks correct.

A screenshot of a computer program

Description automatically generated

The next step is to extract the date and values columns from the data frame and set the Date column as the index. The data is now structured for time series analysis, linking each value directly to a specific date.

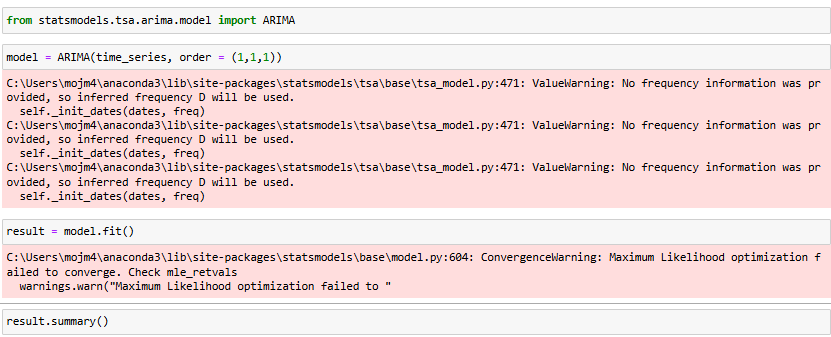


Then, a line plot is created to visually represent the time series data. The plot includes labels for both axes, a title, a legend to identify the data, and a grid for readability, helping to visually analyse the trend in the data over time.

A screen shot of a graph

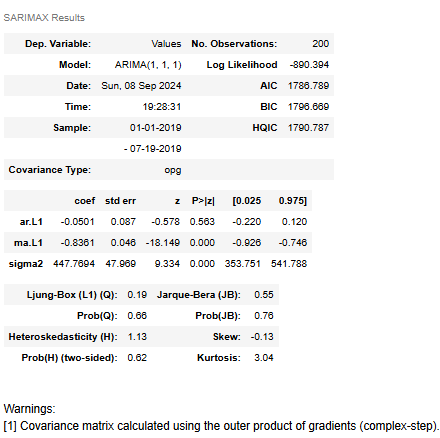
Description automatically generated

The ARIMA model is imported and defined to analyse the time series data, with the model’s configuration specified using the order (1,1,1). The three numbers (1,1,1) represent: p (number of lag observations included -1), d (number of times the data has been differenced to make it stationary -1) and q (number of lagged forecast errors in the model -1). Using these, the model is then fit to the data to learn its patterns. A summary report from the SARIMAX model is created to assess the model’s fit and its parameters, to see if there is any need to make any necessary adjustments to improve its forecasting accuracy.



**Interpretation of (ARIMA-1,1,1) SARIMAX results:**

The autoregressive coefficient is very low and insignificant, illustrating that the lagged values don’t contribute much to the overall predictions. Furthermore, the AIC, BIC and MSE values suggest that the model has a poor fit because a good model would minimise these values, the values in the diagram are high, which indicates the ARIMA (1,1,1) specification is struggling to capture any underlying patterns in the data, which can result in poor forecast accuracy. Also, the high signma2 (residual variance) illustrates that the model’s predictions deviate significantly from the actual values.



The number of forecast steps is set to 90 days to forecast future values. Using the fitted ARIMA model, predictions are generated starting from the end of the time series for these 90 days. Another date range is created, beginning from the day immediately following the last observed date and extending for the forecast period.

A screenshot of a computer

Description automatically generated

The results are then visualised by plotting both the original time series data and the forecasted values. The plot is customised with a title, axis labels, a legend a grid to enhance readability and interpretation of the predicted future trends based on the ARIMA model.

A graph with blue lines

Description automatically generated

The decision to change the ARIMA model from (1,1,1) to (2,1,4) significantly improved the accuracy and realism of the forecast. In the original ARIMA (1,1,1) model, the forecast quickly flattened into a straight line, indicating that it couldn’t capture the complexity of the data beyond short-term fluctuations. This occurred because the model used only one past observation and one past error term, limiting its ability to reflect complex time series patterns.

A screenshot of a computer

Description automatically generated

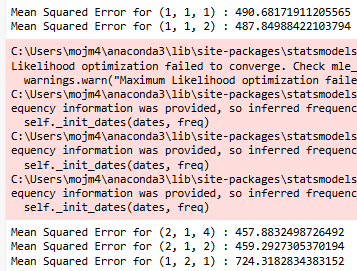
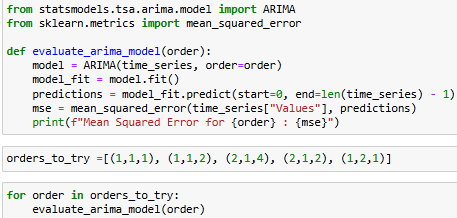


By switching to ARIMA (2,1,4), the model considered more past values (p = 2) and adjusted for errors over a longer period (q=4), leading to a more reliable forecast. The new configuration allowed the model to better capture the underlying trends and cycles in the data, resulting in a smoother and more accurate projection, as seen below.

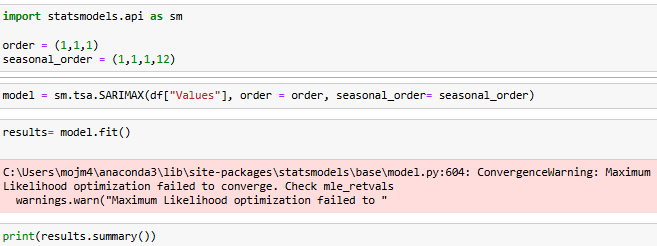
A graph with a line going up

Description automatically generated

The evaluation of different ARIMA model configurations was completed by fitting each model to the time series data and calculating the mean squared error (MSE) of their predictions. Lower MSE values indicate a better fit to the data. The ARIMA model with parameters (2,1,4) achieved the lowest MSE of 457.88, making it the most accurate model for this dataset. Models with parameters (1,1,2) and (2,1,2) also performed well, with MSE values of 487.85 and 459.29, respectively. In contrast, the model with parameters (1,2,1) had the highest MSE, indicating the poorest fit. To summarise, the (2,1,4) model provided the best performance in terms of predictive accuracy among the configurations tested.



In this final step, the SARIMAX model is applied to account for both non-seasonal and seasonal components of the time series data, capturing underlying patterns such as monthly cycles. The process begins by importing the statsmodels library, which provides statistical tools for time series analysis. The order parameter (1,1,1) specifies the non-seasonal elements of the model: 1 autoregressive term (AR), 1st order differencing to ensure stationarity, and 1 moving average term. The seasonal order (1,1,1,12) defines the seasonal aspects of the model, where data is differenced once for seasonality, with one AR term, one seasonal MA term, and the seasonality repeats every 12 periods (one year of monthly data). The SARIMAX model is then fitted to the values column of the dataset, which represents the time series. Like before, the summary function provides a statistical overview of the model, highlighting key metrics that inform the model’s fit and predictive accuracy.



**SARIMAX Model Performance and Interpretation:**

Lower AIC and BIC values indicate a better model fit, in this case, 1724.271 for AIC and 1740.27 for BIC are not low, so suggest a bad model fit. A less negative Log Likelihood would indicate a closer fit to the data and in this case, a -857.136 does not suggest that. Only the moving average term (MA.L1) is statistically significant as the other terms – AR and seasonal parts don’t seem to add any value, which suggests that the model is not optimal. Statistically significant coefficients are important for a model to accurately capture the trends in the data. The model’s errors (residuals) look normal – they aren’t correlated (Ljung-Box test p =0.84) and follow a normal distribution (Jarque-Bera p = 0.65) which means the model captures the overall pattern in the data. Overall, this is not a well-rounded time series model, as a good model must have significant coefficients for all AR, MA and seasonal components to capture both short-term and seasonal dynamics effectively. A recommendation for this could be adjusting the model’s parameters.

A screenshot of a computer screen

Description automatically generated