Time Series Forecasting with Autoregressive Integrated Moving Average Model

1. **Introduction**

A time series is a sequence where a metric is recorded over regular time intervals. Depending on the frequency, a time series can be of yearly, monthly, weekly, daily or hourly. Time series data is a real set of data instead of gaining from mathematical or statistical experiments. Since it is real, it is a statistical indicator to reflect a certain phenomenon. Therefore, behind the time series is the change of rule. Time series forecasting is the use of a model to predict future values based on previous observed values.

1. **The Data and Data Preprocessing**

The data is from the author Rob Mulla who updated the data to Kaggle. The data records over 10 years of hourly energy consumption in Megawatts of American Electric Power (AEP), a company that delivers electricity to five million customers in 11 states.

图片包含 屏幕截图

描述已自动生成

图片包含 屏幕截图

描述已自动生成

The hourly energy consumption then was grouped into a corresponding date to prepare next preprocessing step. Indexing with time series data is also necessary because index has the character of uniqueness, which can ensure the uniqueness of each record in each corresponding date. Finally resample the time series data into monthly time series data.

图片包含 屏幕截图

描述已自动生成

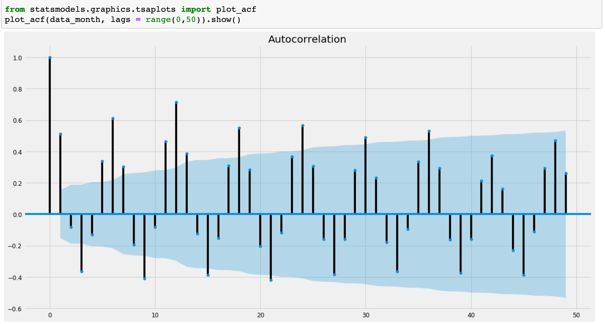
1. **Visualizing Energy Consumption Time Series Data**

Using time-series decomposition can allow the time series data be decomposed into three components: trend, seasonality, and noise.

图片包含 文字, 屏幕截图

描述已自动生成

There is a trend of the time series data, it isn’t stable. Double check its auto-correlation and ADF test result:



图片包含 屏幕截图

描述已自动生成

The plot of auto-correlation doesn’t have either “cut off” or “tail off” which are characters of a stable time series data.

ADF tests the null hypothesis that a unit root is present in a time series sample that can cause problems in statistical inference involving time series models. The p-value in the test is much larger than the significant level, the null hypothesis was accepted. The time series data is not stable.

1. **Make the Time Series Stationarity**

To make time series data stable, differencing is often used. After first-differencing, check the auto-correlation and ADF test result again:

图片包含 物体

描述已自动生成

图片包含 屏幕截图

描述已自动生成

The autocorrelation plot has obvious "tail off" characteristics and the p-value is closed to 0 to reject the hypothesis that the unit roof exits.

1. **Fitting the ARIMA Model**

ARIMA, short for “Auto Regressive Integrated Moving Average” is one of the most commonly used method for time series forecasting.

图片包含 屏幕截图

描述已自动生成

Above is the table of the model summary. The values under “coef” are the weights of the respective terms. The p-value in “P>|z|” is highly insignificant, it should ideally less than the significant level 0.05.

1. **Validating and Visualizing Forecasting Monthly**

To understand the accuracy of the forecasting, the predicted energy consumption and real energy consumption were compared, and forecasts were set to start at 01/01/2017.

图片包含 屏幕截图

描述已自动生成

The line plot is showing the observed values compared to the rolling forecast predictions. Inside the code “dynamic” was set as “false” so that the in-sample lagged values were used to predict. The model got trained up until previous value to make the next prediction. Overall, the forecasts align with the real values. It seems to have a decent ARIMA model. Next step is to forecast future energy consumption.

图片包含 地图, 文字, 屏幕截图

描述已自动生成

The model clearly captured future energy consumption seasonality. However, as the forecasting further out into the future, it becomes less confident in the values. This is reflected by the confidence intervals generated by our model, which grow larger as it moves further out into the future.

1. **Validating and Visualizing Forecasting Hourly**

Now let’s move to hourly time series data. The data preprocessing is similar with the monthly time series data, resampling data into monthly is changed to resampling data into hourly.

图片包含 屏幕截图

描述已自动生成

Let’s plot the time series data:

图片包含 屏幕截图

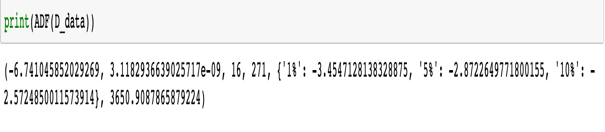
描述已自动生成

There is an obvious trend in the time series plot. The time series is not stable. Like monthly time series above, making the data stationarity is necessary. After first-differencing:

图片包含 物体

描述已自动生成

The ADF result:



After fitting ARIMA model and visualize the forecast compared to real value:

图片包含 屏幕截图

描述已自动生成

Visualize the forecast with future energy consumption:

图片包含 文字, 地图

描述已自动生成

1. **Conclusion**

Overall, the predictions fit the model well, and the actual observed values lie within the 95% confidence band. The basic idea of time series forecasting is that based on the previously observed data, a mathematical model can be established to accurately reflect the dynamic dependencies contained in the sequence, and thus to forecast the future of the system. The advantage of time series forecasting is simple to operate, but its negative aspect is also apparent: its accuracy is not good enough for long term prediction.

**Reference**

<https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>