Time series forecasting benchmark

12.01.2022

Realized by:

-Cherif Baya

-Bdiri Wassim

Supervised by:

-Mr.Dali Wiam

Contents

1 Introduction

Time series forecasting is a technique for predicting future events by analyzing past trends, based on the assumption that future trends will hold similar to historical trends. Forecasting involves using models fit on historical data to predict future values. Prediction problems that involve a time component require time series forecasting, which provides a data-driven approach to effective and efficient planning.

In this paper we are going to discuss the choice of model to make,after trying several models.The choice criteria that we are going to use is the Mean Squared Error that we'll be explaining in third section.

2 Requirements

2.1 Background

Human activity related time series is a collection of chronologically recorded observations which are highly correlated with the human behavior. Compared to other time series, these time series are intrinsically chaotic . A time series is said to be chaotic if and only if it is nonlinear, deterministic and sensitive to initial conditions. The prediction of a chaotic time series engages with the prediction of future behavior of the chaotic system by utilizing the current and past states of that system.

In addition to these, human activity related time series prediction is a highly complicated task as they exhibit the following characteristics:

1. They often behave nearly like a random-walk process, making the prediction almost impossible.
2. They are usually very noisy, i.e., there is a large amount of random day-to-day variations.
3. Statistical properties of these time series are different at different points in time as the process is time-varying.

Several statistical models, including the naïve, exponential smoothing, moving average, and time series decomposition,were used to forecast these specific time series but they didn't give good results. In this paper we'll try to confirm the inefficiency of these models and apply better models.

2.2 Data set choice

We selected a power consumption data set which was collected from an apartment unit in San Jose for 2 years. The data was collected with smart meters and shared by the energy company. We chose this data set for two main reasons: 1.Energy consumption is highly correlated with the human activity. 2.The rapid growth of energy consumption worldwide which make the energy companies need to forecast the clients' consumption for better use.

3 Models

Most time series models require the data to be stationary. A time series is said to be stationary if its statistical properties such as mean, variance covariance remain constant over time.Our data is stationary.We have checked its stationarity using Augmented Dickey-Fuller test

3.1 ARIMA

3.1.1 Definition

Autoregressive integrated moving average, or ARIMA,are among the most widely used time series forecasting techniques.It takes advantage of autocorrelation to produce forecasts. Autocorrelation is when a time series displays correlation between the time series and a lagged version of the time series.

- Auto-regression. A model that uses the dependent relationship between observation and some number of lagged observations.
- Integrated. The use of difference's of raw observations

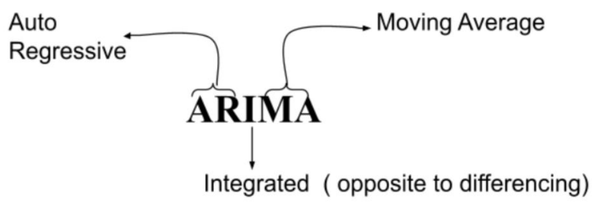

Figure 1: ARIMA model

- Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

3.1.2 Simulation

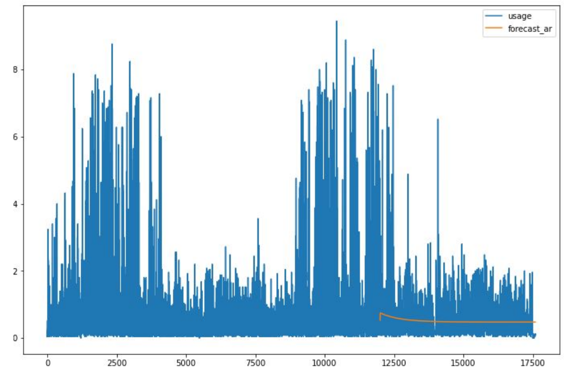

Figure 2: ARIMA forecasting

3.2 SARIMA

3.2.1 Definition

The SARIMA model extends ARIMA by adding a linear combination of seasonal past values and/or forecast errors. We added the seasonal factor (m=12) since our data has a yearly seasonality.

3.2.2 Simulation

Testing the sarima model

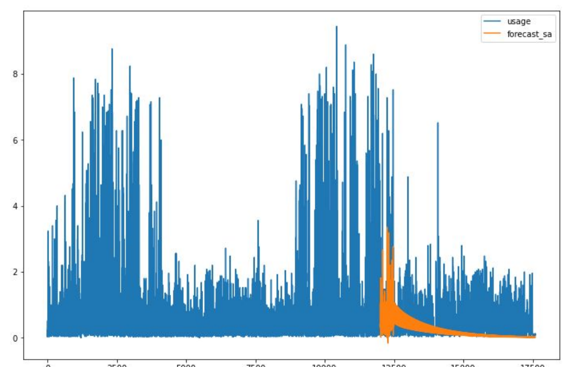

Figure 3: SARIMA forecasting

3.3 Exponential Smoothing

3.3.1 Definition

Exponential smoothing is a time series forecasting method for univariate data. They are similar in that a prediction is a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations.

Our dataset contains seasonality so we'll use Triple Exponential Smoothing as it's an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series. The hypermarameters used are :

- Alpha: Smoothing factor for the level.
- Beta: Smoothing factor for the trend.
- Gamma: Controls the influence on the seasonal component

3.3.2 Simulation

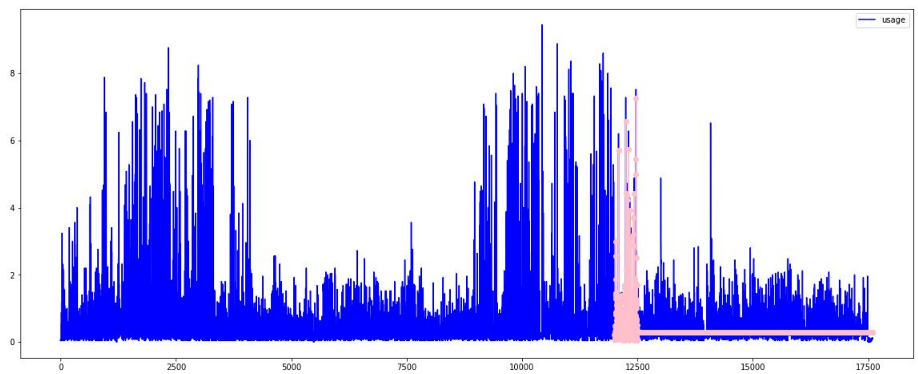

Figure 4: ETS forecasting

3.4 Prophet

3.4.1 Definition

Prophet is an open sourcing forecasting tool available in Python and R. Forecasting is a data science task that is central to many activities within an organization. For instance,

The reason why we chose prophet is because it can deal with :

1. strong multiple "human-scale" seasonalities: day of week and time of year
2. important holidays that occur at irregular intervals that are known in advance (e.g. Christmas)
3. historical trend changes, for instance due to product launches or logging changes

Since our data set was made in California we configured the prophet model with the US holidays.We changed the holiday country to India and as expected, we noticed an increase in the mean squared error.This is proves how accurate prophet is.

3.4.2 Simulation

Testing Prophet

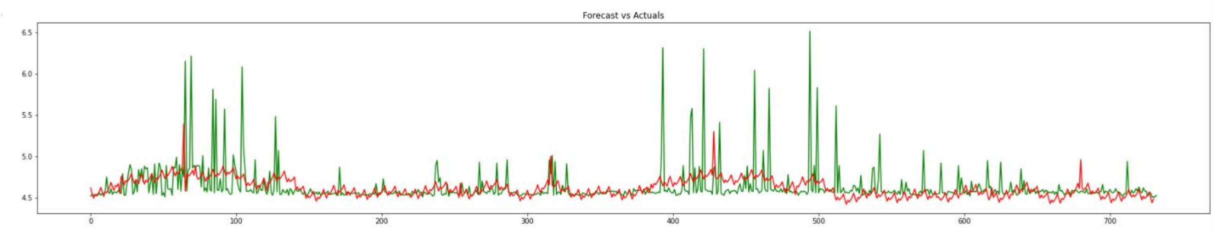

Figure 5: Prophet forecasting

3.5 LSTM

3.5.1 Definition

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture[1] used in the field of deep learning.

LSTM networks are well-suited to making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

Since ARIMA and SARIMA models are highly parameterized and due to this, they don't generalize well. Using a parameterized ARIMA on a new dataset may not return accurate results. RNN-based models like LSTM are non-parametric and are more generalizable.

3.5.2 Simulation

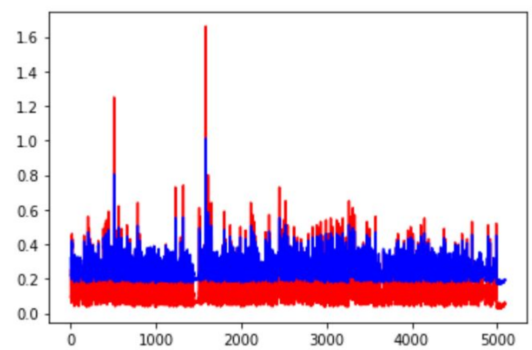

Figure 6: LSTM forecasting

3.6 Root Mean Squared Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y_i})^2$$

Figure 7: rmse formula

3.7 Model comparison

3.7.1 Results

	ARIMA	SARIMA	Exponential Smoothing	Prophet	LSTM
RMSE	0.254	0.237	0.278	0.167	0.119

Figure 8: RMSE measures obtained under the five time series models.

3.7.2 Discussion

After running the seven different time series models discussed in Section 3. and obtaining the forecasts, we evaluate them using the error measures given in Section 3.3, and the results are presented in Figure 8. It is clear that Prophet and LSTM yield minimal error measures. Hence, we conclude that, under the time series methods, these two models are best forecasting the time series which is related to human activity.

4 User guide

Our solution is an automatic procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. To start using the solution, the user must provide the data set plus the number of days to forecast.They data set contain two columns at least:

1. Datestamp column which should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp.

2. Value column which must be numeric, and represents the measurement we wish to forecast.

Within minutes the forecast data will be popped out based on the inputted number of days to forecast.The automated code will choose the algorithm which is having the least RMSE and the predictions will be done by using the same algorithm.

5 Conclusion

In this paper, we investigated different time series forecasting methods that can be used for power consumption forecasting. We compared five time series methods,namely the ARIMA,SARIMA,Simple Exponential Smoothing,Prophet and LSTM. We ran the five above mentioned methods for two, four and six points ahead predictions and computed the accuracy indicators root of mean square(RMSE).We confirmed that the statistical models aren't so accurate with time series which are related to human activity and can be replaced with new models like Prophet or LSTM. However, LSTMs were not developed for the purpose of analyzing simple time series data like those considered in this paper. It has been the central focus of researchers hoping to solve highly complex tasks, such as generation of text and handwriting. Although other variations of the LSTM may have had more success with the time series data, we believe that for simple settings, the Prophet model that makes reasonable assumptions about the underlying structure tend to be more effective.

References

[1] Enders, W, *Econometric Time Series: Wiley: New York, NY, USA, 1995*,

[2] Hamilton, J, *Series Analysis; Princeton University Press: Princeton, NJ, USA, 1994*

[3] Brown, R.G, *Smoothing for Predicting Demand: ; Arthur D. Little Inc.: Cambridge, MA, USA, 1956*

[4] Facebook Prophet, *://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/*,