Timeseries Prediction – Powersupply Stream of Italy between 1995 and 1998

Introduction

The data we analyzed contains hourly steam of a power supply. The data was recorded by a certain electricity company in Italy which records the power from two sources: power supply from main grid and power transformed from other grids. We decided to analize the power from main grid. This stream contains three year power supply records from 1995 to 1998.

Our main tool for the analysis was Python with some additional libraries (pandas, scikit, prophet, statsmodels, sklearn).

Data preprocessing and cleaning

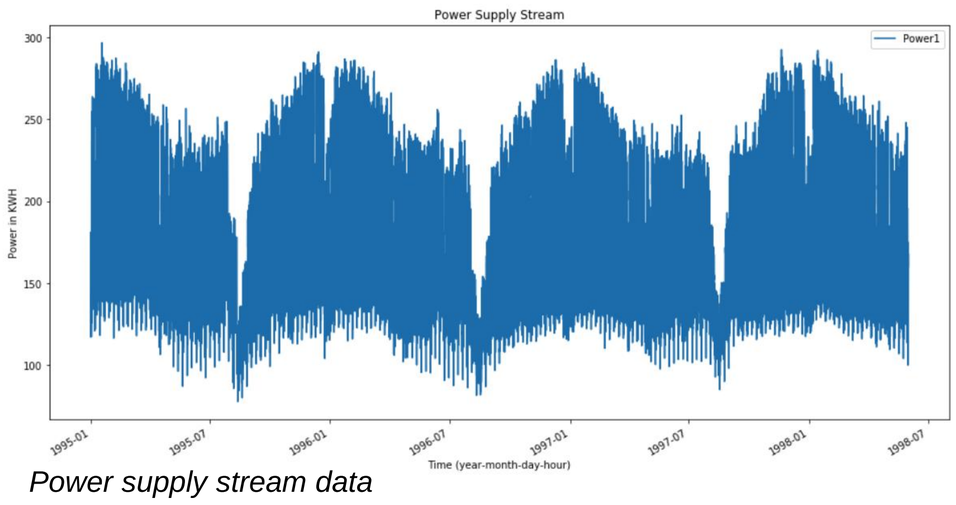
The initial dataset we obtained was in ‘arff’ format. For easier future manipulation we imported it into pandas DataFrame structure, which is supported by most of the above mentioned libraries.

	attribute0	attribute1	class
0	117.4	127.0	b'0'
1	139.1	126.0	b'1'
2	128.0	120.0	b'2'
3	127.0	112.0	b'3'
4	122.8	110.0	b'4'

Original data

date_time	power1	date_time	hour
1995-01-01 00:00:00	117.4	1995-01-01 00:00:00	0
1995-01-01 01:00:00	139.1	1995-01-01 01:00:00	1
1995-01-01 02:00:00	128.0	1995-01-01 02:00:00	2
1995-01-01 03:00:00	127.0	1995-01-01 03:00:00	3
1995-01-01 04:00:00	122.8	1995-01-01 04:00:00	4

Pre-processed data



Power supply stream data

Stationarity

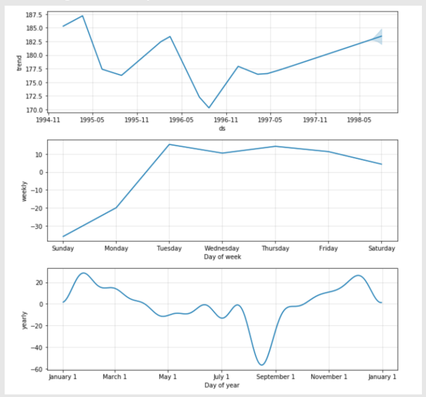
To check if the data is stationary. That means that it‘s statistical properties do not change over time. We used a few test to confirm that our data is stationary.

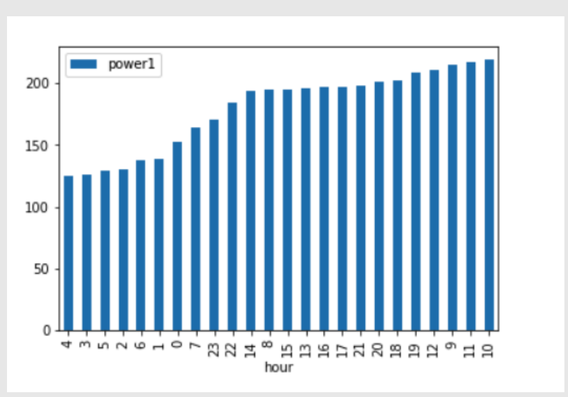
- Split the data into two groups
- Calculate a mean of each group (mean1=177.46, mean2=181.48)
- Calculate a variance of each group (var1=2197.33, var2=2015.34)
- Perform a Dickey-Fuller Test:
  - p-value > 0.05: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
  - p-value <= 0.05: Reject the null hypothesis (H0), the data does not have a unit root and is stationary
  - p-value=2.082908e-28 –> H0 is rejected

Trends and seasonality

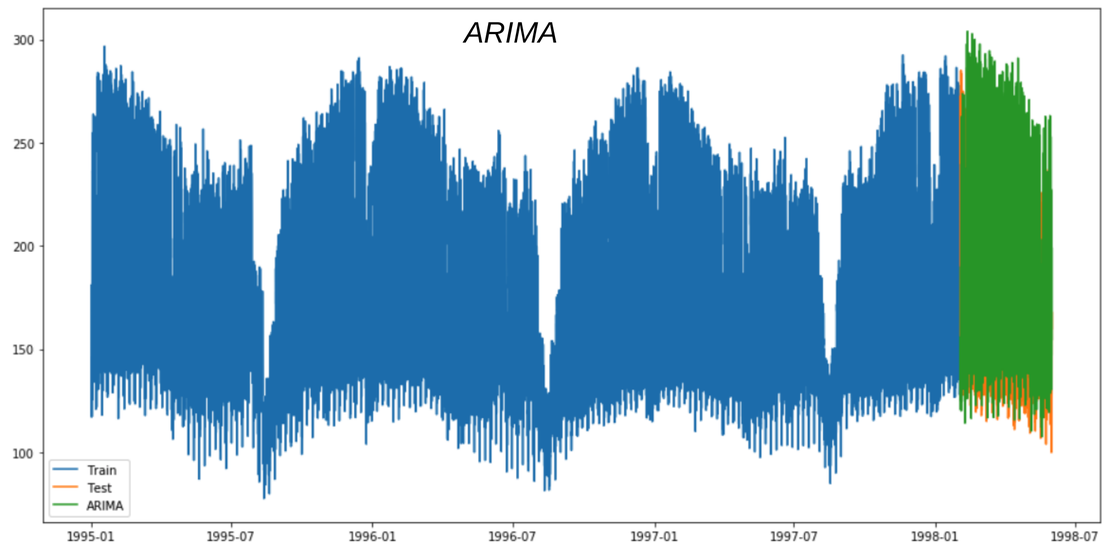
The daily pattern is interesting as it seems that the consumption increases during the day, but the highest consumption on average is from 9 a.m. to 12 a.m.

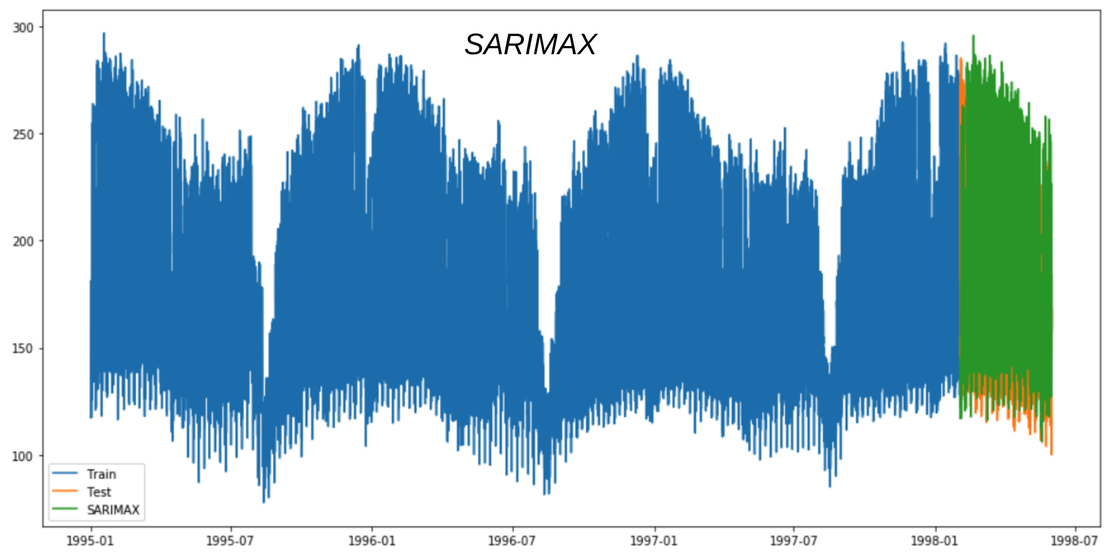
We can also see that power consumption decreases in spring and summer and hit the lowest point in the middle of the August. Looking at the weekly trend, it seems that there is less consumption on Sunday than the other days of the week. Except Monday, that could be problem with dataset, since the exact first day of streaming is not specified.

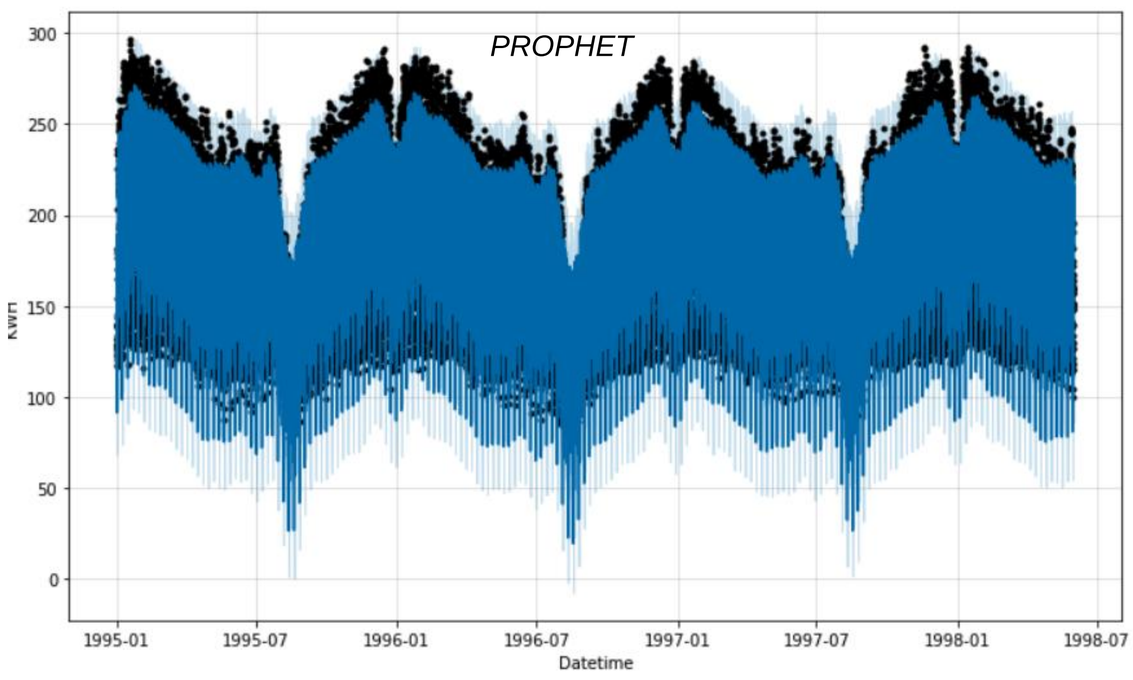




Forecast







Metrics	ARIMA	SARIMAX	PROPHET
Mean squared error	1305.13	1253.49	300.24
Mean absolute error	26.49	26.32	13.53
Root mean square error	36.12	35.40	17.33

Conclusion

The best model for prediction was achieved with Facebook’s Prophet. From the results we can also see that ARIMA and SARIMAX provide pretty much same result for this dataset. The LSTM prediction that Prophet uses is based on a set of last values. Therefore it is less prone to trends and seasonality. The multiplicative dataset would be harder to analize than the provided one which is additive. Therefore the prediction are also quite accurate and could be used.

Possible improvements: events detection, parameters finetuning.

Sources

1 http://www.cse.fau.edu/~xqzhu/Stream/powersupply.arff, 15.04.2019
2 https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a, 17.04.2019
3 https://towardsdatascience.com/playing-with-time-series-data-in-python-959e2485bff8, 25.05.2019
4 https://www.kaggle.com/robikscube/tutorial-time-series-forecasting-with-prophet, 25.05.2019
5 https://www.kaggle.com/nholloway/stationarity-smoothing-and-seasonality, 25.05.2019
6 https://machinelearningmastery.com/time-series-data-stationary-python/, 07.06.2019
7 https://otexts.com/fpp2/accuracy.html, 18.06.2019