Acea Smart Water Analysis (Petrignano)

Springboard (Capstone Project3)
Author- Sweta Gaurav

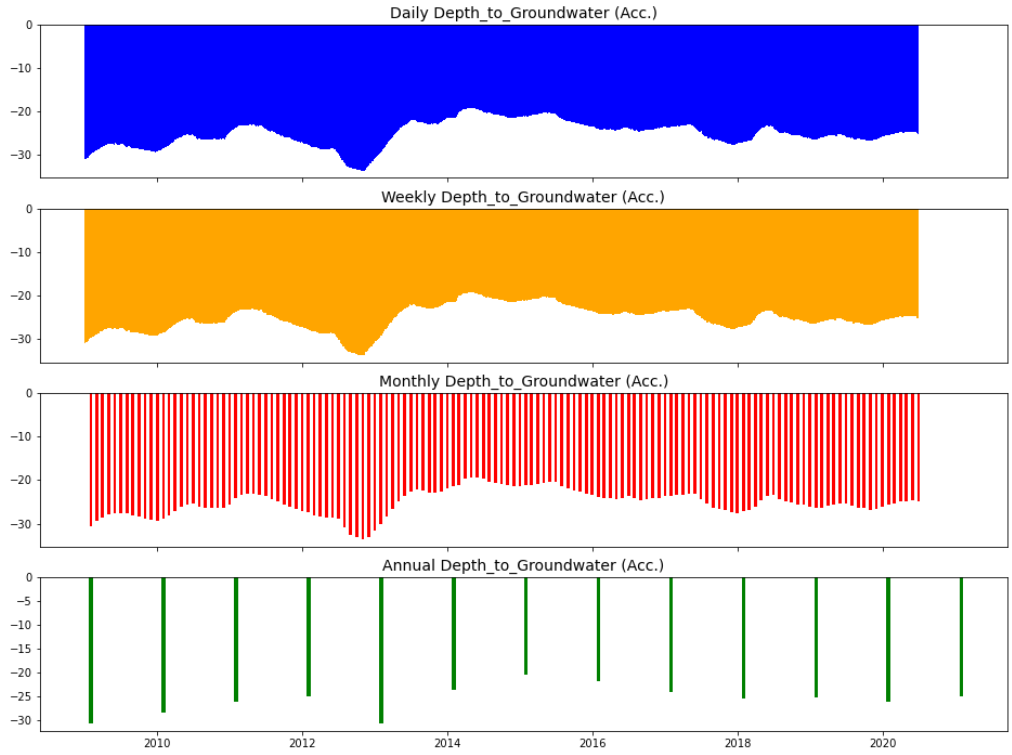
Overview

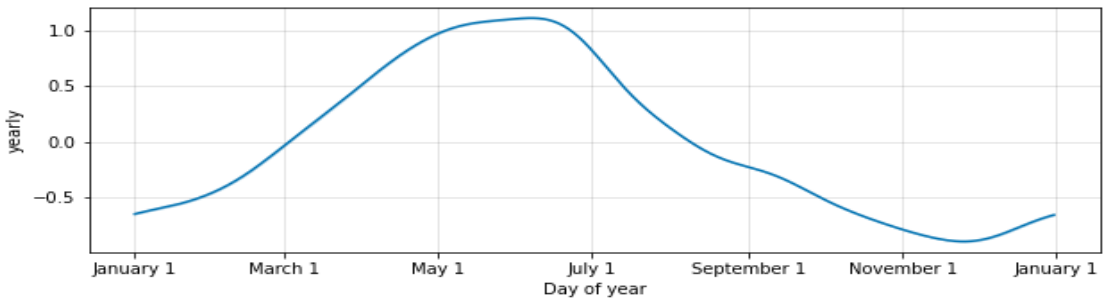
The Acea Group is one of the leading Italian multiutility operators. Listed on the Italian Stock Exchange since 1999, it is foremost Italian operator in the water services sector supplying 9 million inhabitants in Lazio, Tuscany, Umbria, Molise, Campania.

As it is easy to imagine, a water supply company struggles with the need to forecast the water level in waterbody (aquifer) to handle daily consumption.
This project focusing on Forecast the depth to groundwater of an aquifer located in Petrignano, Italy. Also forecast the underground(aquifers) water level, for each day of the year. I have built a story to predict the amount of water in waterbody(aquifer).

Introduction

Time Series Analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend, cyclic or seasonal variation) that should be accounted for. The behavior of time series variables such as underground water label is not consistent it goes up and down depends on climates and to forecast it is irrational. Despite these assertions, many water companies like Petrignano must deal with amount of water supply to make decisions how much water it need to supply. These hedging decisions are made under the premise that patterns exist in the past data and these patterns provide an indication of future label. If such patterns exist, then it is possible in principle to apply modern mathematical tools and techniques such as ARIMA and PROPHET to forecast the label. Petrignano underground water varies from -18m to -34m. and my data shows cyclic pattern. Plots of the series, autocorrelation function and the partial autocorrelation function are some of the graphical tools used to analyze the series. We also aim to fit a model (ARIMA, Facebook Prophet, Auto arima) to the data to make credible forecasts from the model.





After reviewing the data, we find there are some trends sometime water level goes up and sometimes it is going down. There are some seasonal pattern water goes down in month of October, November, December, and January of every year. Water label goes up in month May and June.

Dataset

Below is dataset link

dataset link: https://www.kaggle.com/c/acea-water-prediction/data/

Data consist of various variable such as Date, Rainfall, Volume, Hydrometry, Temperature, Depth of Groundwater. As this is time series analysis, I am interested in only Date column and Target column in this case target is Depth of Groundwater.

For this project I obtain the data from Acea Group for aquifer located in Petrignano, Italy. For my analysis I use the data from 2009/01/01 to 2020/06/30 and did forecast for 365 days till 2021/06/30.

Approach

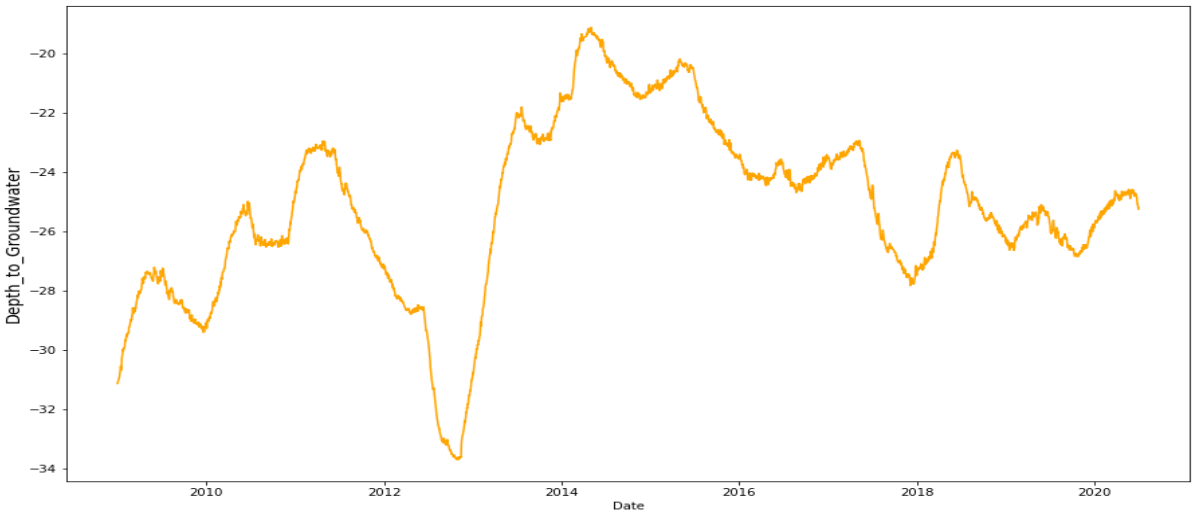
This is a time series problem, so our first task is to check given data is stationary or not if data is not stationary then we must make it stationary to conduct time series analysis. Time series forecasting is a technique for the prediction of events through a sequence of time. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold like historical trends.

I choose to work with. ARIMA and Facebook prophet to analyze future prediction of underground water label. In the future, LSTM neural network can be used in future I want to work on that.

Data Cleaning

After reviewing the dataset get to know there are 1024 missing values in most columns, so I drop firs t 1024 rows to make meaningful data. We have two target feature Depth_to_Groundwater_P24 and Depth_to_Groundwater_P25. To make univariate prediction drop one of target variable Depth_to_Gr oundwater_P24 so we can focus on one target.
We can see that our target variable (Depth_to_Groundwater) has missing values. We will have to cle an them by replacing them by nan values and filling them afterwards. After doing statistical operation , the best option in this case, is interpolate.

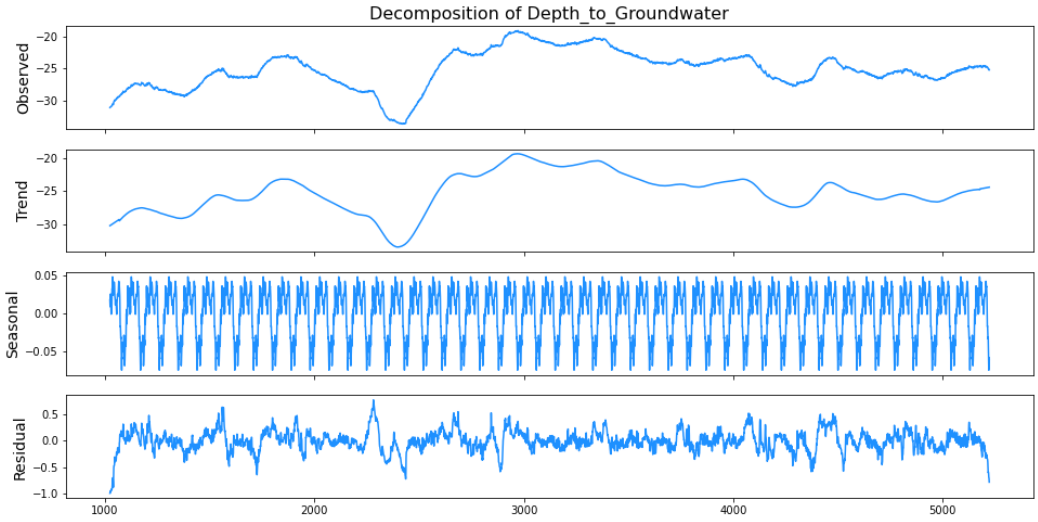
EDA



The time series plot for the aquifers is shown in above fig It will be noted that this plot exhibits no per iodicity, but we do have a trend effect due to the random walk nature. After transformation using diffe rencing technique the trend effect is eliminated. therefore, the integration order is zero. Generally, th e volatility of the series is uniform for the years 2009 to 2012. For the years of 2014 to 2020, volatility of the series was highly non-uniform and more pronounced around 2013.

Decomposition of data

We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and Residual. I use additive decomposition.



Checking Stationarity of a Time Series

Stationarity is requirement of most of the time series Model. There are three basic criteria for a time series to understand whether it is stationary series or not.

1. constant mean and mean are not time dependent.
2. constant variance and variance are not time dependent.
3. constant covariance and covariance are not time dependent.

Statistical properties of time series such as mean, variance & covariance should remain constant over time, to call **time series is stationary.**

To check stationarity, I perform **Augmented Dickey-Fuller (ADF) test.** In this test I focus on p_value, significance level (default: 0.05) and critical value.

Augmented Dickey-Fuller (ADF) test is a type of statistical test called a unit root test. Unit roots are a cause for non-stationarity.

If the null hypothesis can be rejected, we can conclude that the time series is stationary.

There are two ways to rejects the null hypothesis:

On the one hand, the null hypothesis can be rejected if the p-value is below a set significance level. The defaults significance level is 5%

p-value > significance level (default: 0.05): Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary. **p-value <= significance level (default: 0.05)**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary. On the other hand, the null hypothesis can be rejects if the test statistic is less than the critical value.

ADF statistic > critical value: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary. **ADF statistic < critical value**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.

ADF statistic (-2.899836995568036,
p-value 0.04536695595343471,
28,
4170,
critical value
{'1%': -3.4319191438819407,
'5%': -2.8622333615468443,
'10%': -2.567139082403142},
-11587.395288114172)

After reviewing the ADF result we say data is nonstationary.

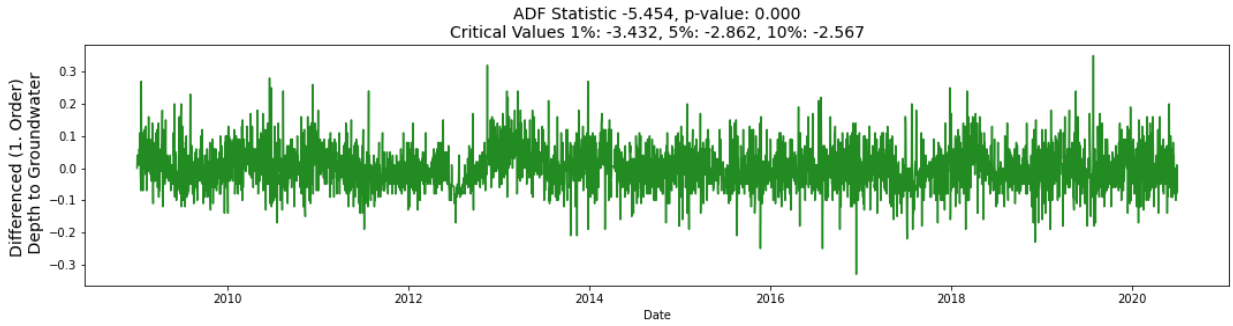
Preprocessing

Making Stationary

The two most common methods to achieve stationarity are:

1. Transformation: e.g., log or square root to stabilize non-constant variance
2. Differencing: subtracts the current value from the previous

After Differencing when we check ADF statistic and p-value I find data become stationary because it fulfills the significant criteria



Time Series Forecasting Algorithm

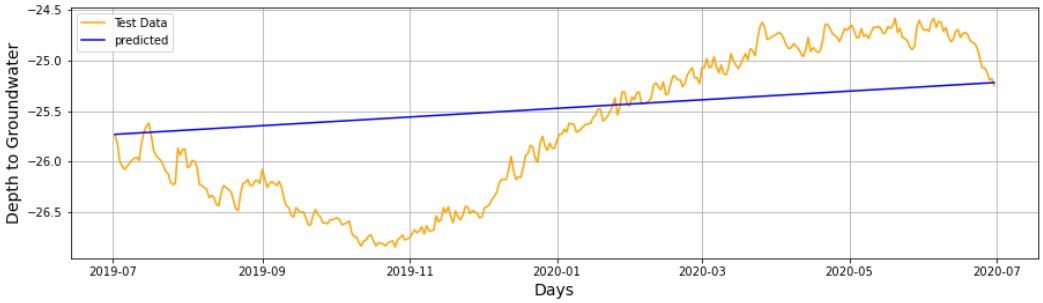
I choose to work on ARIMA and Facebook Prophet

For test data I use 365 days data and remaining I use for training data. For ARIMA I take order (1,1,1)

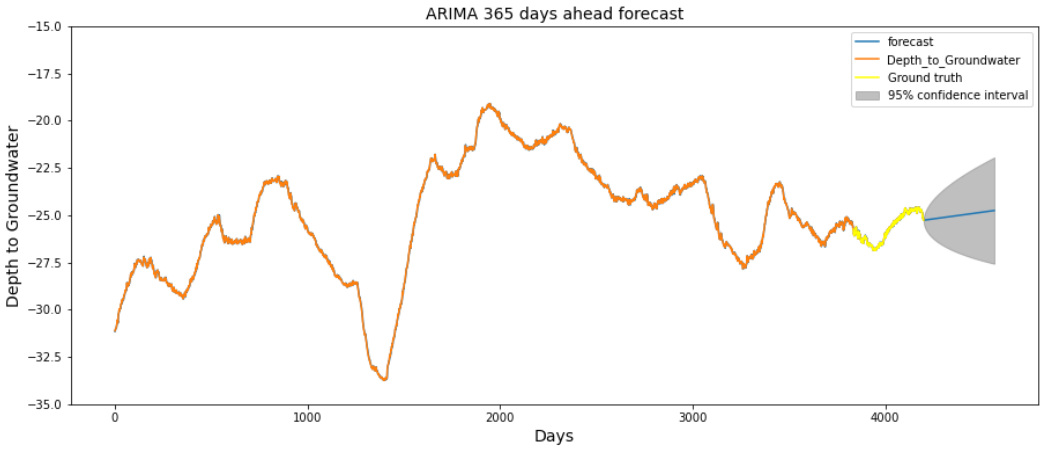
I did forecast for next 1 year (365 days) from 2020/06/30 to 2021/06/30.

ARIMA

Prediction



When we see the actual and predicted value predicted value show us increasing trend, but our actual data has cyclic trend, in beginning of year it goes up and end quarter of year it goes down. Predicted label shows between -25.75m to -25.25m has variation of only 0.5 m.



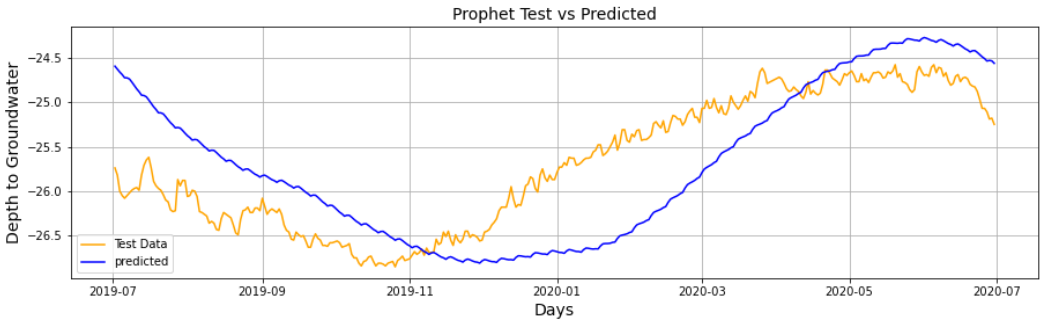
Forecast 365 days ahead data and get to know forecasted water label is -24m to -26m and show the increasing trend, it did not exceed the confidence interval i.e. -18 to -30. According to result water label varies between these points. Solid blue line represents the predicted value. forecasted period is from 2020/06/30 to 2021/06/30.

As we forecast further out into the future, it is natural for us to become less confident in our values. This is reflected by the confidence intervals generated by our model, which grow larger as we move further out into the future.

Facebook Prophet

Released by Facebook in 2017, forecasting tool Prophet is designed for analyzing time-series that display patterns on different time scales such as yearly, weekly and daily.

Prediction

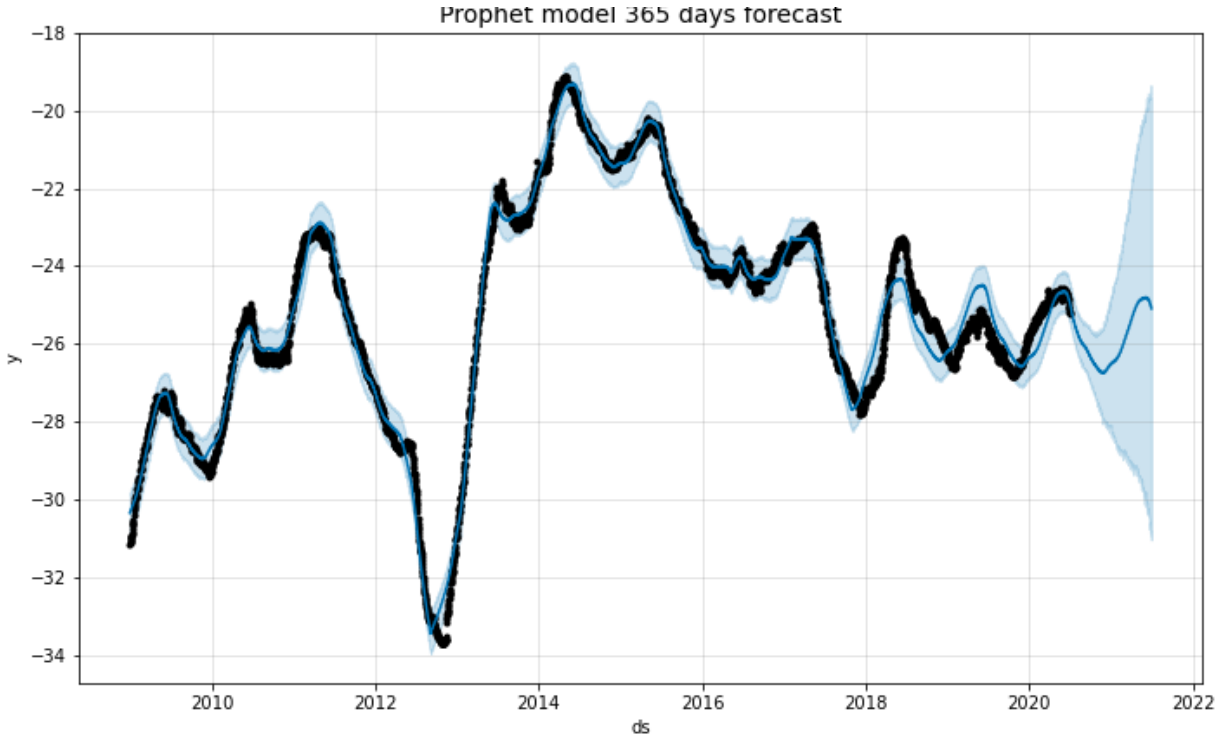


Actual data says in month of October, November, and December water label (between -26.5 to -27) is low but in month of May and June water label is up (between -24.75 to -25).

Predicted value shows the almost same pattern to actual data it shows the cyclic behavior. Water label is low in month of November, December, January and up in month of May and June. Predicted label range from -27m to -24.2m so label varies between 3 m range.

Prediction and Forecast of Facebook Prophet is more accurate than ARIMA Model.

365 Days ahead Forecast.



Forecast 365 days ahead data and get to know forecasted water label is -25 to -27 and it did not exceed the confidence interval i.e. -19 to -30. According to result water label varies between these points. Solid blue line represents the predicted value. forecasted period is from 2020/06/30 to 2021/06/30.

Findings:

Model	MAE	RMSE	R_square
ARIMA	0.5571	0.6505	0.2182
Prophet	0.4172	0.4979	0.5420

Evaluate the Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Square of the models. metrics is considered to be better smaller the value are.

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.it is always non-negative, and the smaller the MAE, the closer we are to finding the line of best fit.

WINNER: Facebook Prophet

Conclusions

Time series analysis and modeling is a very popular technique in mathematics and statistics used to explore the hidden details in time dependent data. ARIMA modeling is one of the basic time series methods employed in practice. In this study, we examined the underground water depth. Due the nature of the data, the differencing of the data is used in the analysis instead of the actual data. This is due to the favorable statistical properties for analysis. As noted previously, ARIMA modeling fails to effectively capture the trend. An alternative model is used be to analyze the result, as a conclusion Facebook prophet is more accurate in predicting the water label and its error is low than ARIMA. It follows the same pattern whatever past data shows. Minimum label is -27m and maximum label is -24m it indicates that predicted water label is in not vary much with respect to actual label.

Recommendations for the Clients

Predicted value shows almost same cyclic pattern to actual data Beginning of year label is up and end quarter label goes down. Predicted label range from -27m to -24.4m so label varies between 3 m range. So, in the month of May and June water label is up, may be due to temperature. We must control on water supply but in the last quarter water label is down, so we do not have to control the supply. 365 days ahead value shows same cyclic pattern.