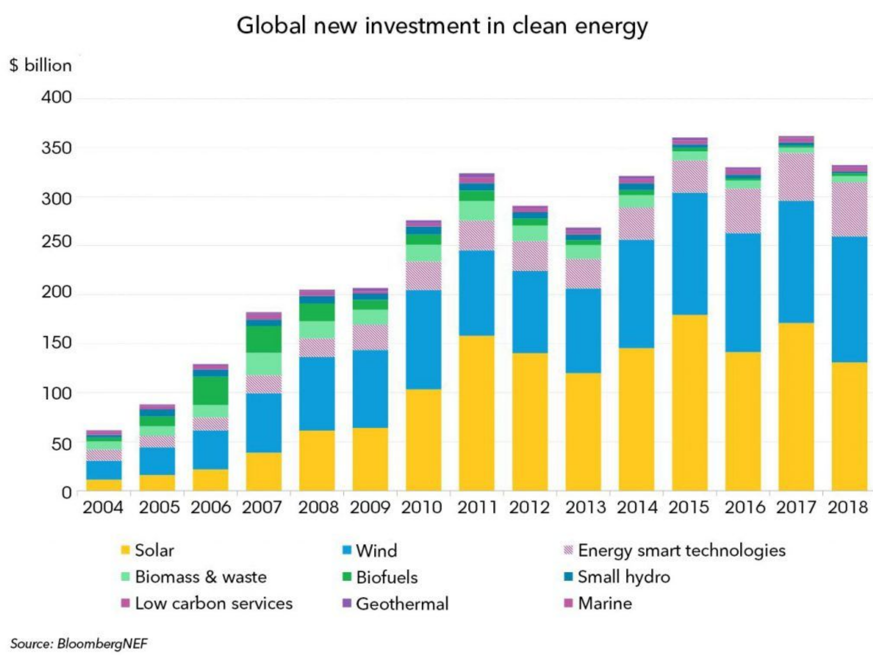
Juan María Fajardo Trigueros
Master en Data Science, Kschool
June 2019

Wind Turbine Signal Analysis
An approach to predict and evaluate the quality of signals

Introduction

The energy world is changing at an exponential speed. Years ago, conventional sources of energy like thermal and generation and were the main sources in the industry, and project budgets were only possible for a few big companies. Currently, the birth of new energy sources like wind and solar power allow for the construction of smaller power plants, increasing their business. This allows smaller and more flexible energy companies to have a larger stake in the industry. [1]



[1] https://about.bnef.com/blog/clean-energy-investment-exceeded-300-billion-2018/

In the past, only a few companies with enough experience handling and manipulating complex models had access to data. Currently, the continuous improvements in data science have created new methodologies that allow more people to gain access to data for analysis. If servers are required, there are many cloud suppliers that offer high speed computing at affordable costs[2].

Considering this, energy projects with lower budgets now have access to affordable technology. Previously, within big energy projects, it was common to spend more than 100,000 USD implementing intelligent systems to predict signals. As of now, smaller energy projects have business models that cannot afford to spend that amount of money. Utilizing machine learning and cloud servers, it is now possible to accurately predict signals.

Through the use of predictive methodologies, there are two main impacts in the business model: maximizing availability and minimizing maintenance costs. If it is possible to predict when the machine is going to break, then measures can be taken in advance to avoid breakage. Income is lost every time a machine breaks due to a loss of functionality within the plant. In addition, prediction can allow optimization of preventive maintenance that is more efficient and cost effective than the current corrective maintenance.

What I propose is to create a model that maximizes the usage of the current data, employees experience, and machine learning tools, and to create a model for wind turbine signal analysis.

This analysis will identify signals whose qualities are wrong and predict a value for the signal and compare with the real value to detect if signals are out of the confidence interval. If the signal's qualities are wrong, systems technicians will be sent to analyze the issue. If signals are out of the confidence interval, operations and maintenance staff will solve the issue. This allows the model to easily direct work to systems technicians or to operations and maintenance staff. To do this, the user will have access to a Tableau that will display the graph with the trends of the signals and an indicator of the number of issues for both signal quality and measures for those out of range. This model will allow employees to focus more on using their experience to prevent breakdowns in more power plants.

Motivation

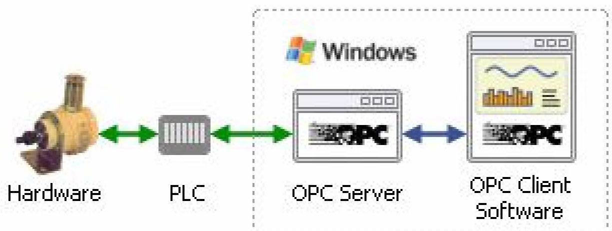
[2] Llorente, Ignacio M. "The Limits to Cloud Price Reduction". June 29, 2017

This project comes from the need to be more efficient at work. I have more than 7 years of experience in the energy sector and from a digital point of view, I see how companies are required to be more efficient if they want to stay relevant in the market[3].

After a thorough analysis, I saw that energy companies have access to a lot of data, that if utilized properly, can add incredible value. Presently, many data files are stored in servers that are not used except in cases of a major breakdowns. Employees don't have the necessary tools able to see the status of the machines easily and if it is necessary to repair them, so they must manually mine through data and use their past experiences to make an analysis.

Data used

The data available is a time series for each wind turbine. All wind turbine signals are accessible through an OPC (OLE for Process Protocol)[4] with almost real time data. Every 10 minutes signals are refreshed, and each wind turbine has 57 signals (see "Dictionary of signals for further information")



For this project, signals for three months were analyzed from January 1, 2019 to March 25, 2019, and three wind turbines were analyzed. In the future, there will be larger date ranges and additional wind turbines.

```
data.shape
```

```
(11964, 58)
```

Although the data can be accessed in real time, this project is focused on batches, where the data will be imported to the model on demand by the employees. In the future, real time data analysis could be implemented.(screenshot of the current data)

[3]https://www.forbes.com/sites/louiscolumbus/2018/08/30/state-of-enterprise-cloud-computing-2018/#25d52760265e
[4] https://opcdatahub.com/WhatIsOPC.html

Methods

The model has been created using Python for data manipulation and prediction. The interface with the user has been developed using Tableau. To perform the forecast in the time series I followed the 5 steps detailed in the book "Forecasting: Principles and Practice."[5]

1. Identify the problem
    a. The forecast is necessary for energy company employees that have large vats of data and need a summary that details the state of the power plant.
    b. The forecast then will be used to prevent major breakages by predicting signals that are out of confidence interval or of bad quality, and allowing for an analysis of the situation before any breakage.
    c. Signal forecast is defined by the convergence of the prediction model and the confidence interval.
    d. Signals are considered bad quality if:
        i. Flat: remains in the same value
        ii. Full scale: when signal measure goes to full scale value. This means that there is a failure in the instrument.
        iii. Gap: there are empty spaces with no value in the time series
        iv. Outliers: are observations that significantly differ from other observations of the same feature.

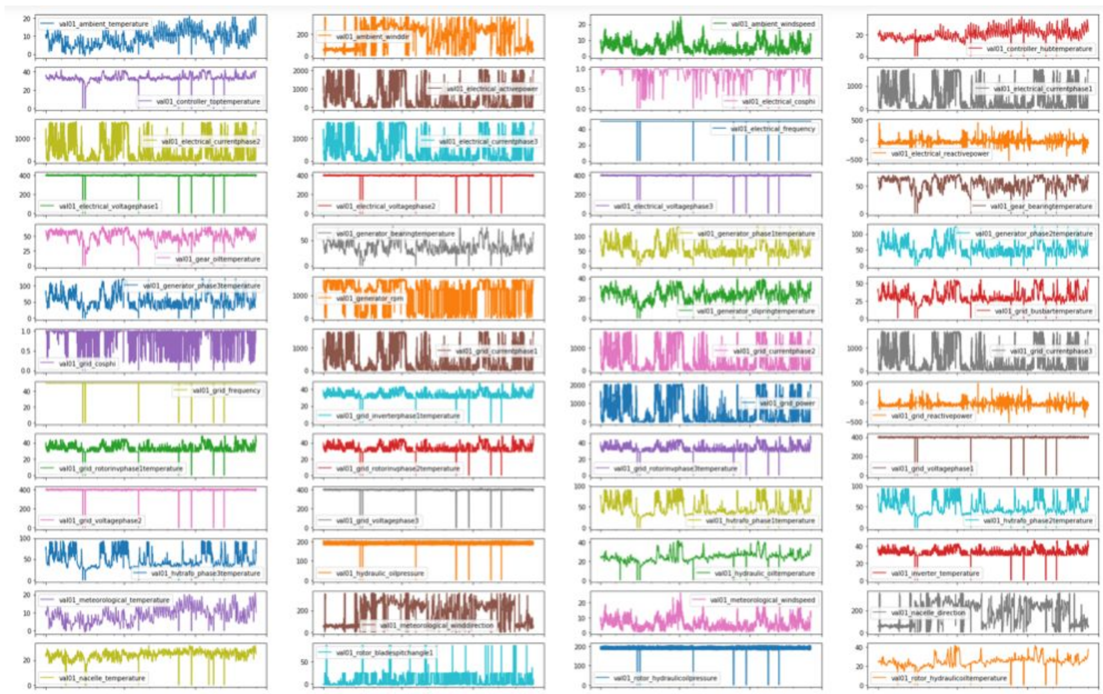
2. Gathering information
    a. Information can be either data or business know-how:
        i. Data: Gathering data is done from either accessing real time data through the use of OPC or on demand data import.
        ii. Business know-how: accessing domain experts that can accurately interpret the historical data, preventing future break downs, based on their experience.

3. Preliminary exploratory analysis
    a. This includes importing the data and manipulating it. During this step plots and reviewed and summarized and obvious temporal structures are noted. These would include trend seasonality, anomalies like missing data, corruption, and

[5] Hyndman, Rob J and Athanasopoulos, George. "Forecasting: principles and practice". October 17, 2013

outliers, and any other structures that may impact forecasting. This step is divided in two, cleaning and exploration:

  i. Cleaning: select the right column and headers
  ii. Exploration: prepare the data for the time series.
      1. Selecting Time index
      2. Type float all the values
      3. Including ploting:



      4. Check for outliers
      5. Missing values
      6. Seasonality

4. Choosing and Fitting Models

In order to choose the correct model, it is necessary to compare several forecasting models and select the one with the better prediction.

For this project, the data is split into training data and testing data. The model created uses all the data, then it is it is trained with the training data. Once the model is trained, the time range of the test data is predicted. The prediction will be compared with the testing data, analyzing the error with the RMSE (frequently used to measure the differences between values and predictions).

This project used Auto Arima (for python from pyramid arima (pmdarima [6]) to model the time series.

To model a time series like the one presented with many variables, it is recommended to create clusters using K-means and fit the model to the cluster, reducing the number of models needed. However, that will be considered in future development. One option was to model all variables are modeled, but this option was discarded due to the high computing time. An additional alternative was to use the web servers with higher computing power (like google research colab), however some of the packages were constantly failing, so this option was also discarded. The final choice was to create a model for an amount of signals that could be manageable for the computer. Three signals were selected.

Fitting the Model: Auto Arima

Calculate the accuracy of the model with RMSE

```
print(rmse)
```

```
1.392140215384274
```

Once the accuracy is checked, future values are firecasted.

[6] (https://www.alkaline-ml.com/pmdarima/about.html)

5. Using and Evaluating a Forecasting Model

With the model selected based on the accuracy, additional values are added. To do that, it is necessary to define a time period for the forecast. In this case, 6 time periods = 1hour is selected. The function predict() by auto arima provides 2 values:

- Prediction
- Confidence interval

These values are stored in the dataframe, and a loop is created to move between variables.

Since the output from the predictions is used as an input in Tableau, a long-shape (instead of wide-shape) dataFrame is used, resulting in:

```
DatetimeIndex: 35874 entries,
Data columns (total 10 columns
```

These values are then transported to Tableau for the analysis of the technical department that will easily identify if there are issues that needs to be resolved.

In addition, this machine learning algorithm will be updated with additional data, creating new predictions and learning from the historical data.

Conclusion

This project creates a model to forecast time series values through the use of a machine learning algorithm like Auto Arima. It is a non-expensive methodology unlike traditional forecasting software that can potentially help energy companies reduce their operational expenditures budget.

It is possible to improve the forecast using K-means for clustering the variables and creating a model for each cluster, however for this project, this model provides an accurate result for the variables predicted.