Capstone 2: Wind turbine power curve calibration using Machine Learning models

Author: Suhas Pol, Ph.D.
Mentor: Srdjan Santic
Date:11/11/2020

Data Science Method (DSM) application to develop Machine Learning (ML) models for wind turbine power curve calibration is presented here. The DSM steps are outlined below.

Step 1: Problem Identification

As outlined (Brower 2012), it is difficult to rely on the upstream measurement to predict turbine power. Typically an under-prediction error of 1-3% has been observed, which significantly impacts a wind farm operators revenue potential. For example, a 3% error for a 100 MW windfarm will lead to $2 Million revenue loss to the operator at electricity price of 80 $/MWh. To alleviate this issue a solution is desired to improve turbine generated power prediction using existing instruments.

The following fundamental problem statement is first considered for the wind energy research project: *Can available meteorological sensors estimate turbine power generation within 3% accuracy and within a 1-minute response time ?*

Step 2: Data Wrangling

EOLOS wind research station wind turbine data from St. Anthony Falls Research Labs at University of Minnesota was used for the project. The data set includes 4, 1-hour time periods of meteorological and wind turbine data. A high and low wind time period from both summer and winter are included. A 2.5 MW Clipper Liberty wind turbine was used at the EOLOS wind turbine site that had a hub height of 80 m and 96m rotor diameter. A 129.5m tall meteorological tower was located 2D upstream of a turbine that had 6 levels of meteorological instruments. This tower had temperature and humidity measurements in addition to wind speed and direction measurements. Further, wind turbine variables, such as, wind speed-direction and temperature at the nacelle location along with turbine power were made.

The 4, 1-hour time series files of the EOLOS dataset were combined to a single dataframe. The 4 files were prepared for combinations of high or low wind speeds during summer or winter seasons. The dataset had 37 useful columns representing measurements from instruments on the turbine and on the upstream meteorological tower. Measurements were made every 1 second (1Hz data). In total there are 14404 rows of data. The data timestamp was set as the dataframe index.

Step 3: Exploratory Data Analysis

Data profile tables and plots were first created in this step to check for extreme events. The combined data set showed 4-modes for certain variables, such as, temperature or wind direction representing the 4 distinct time periods (shown in Figure 1). Such broad range of variables available will be helpful to create a robust model.

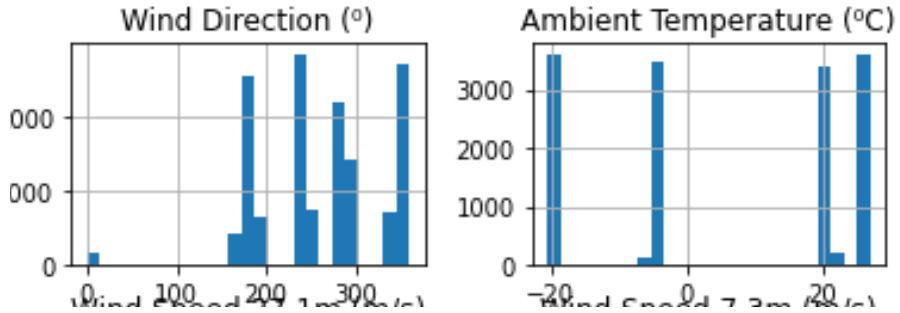

Figure 1: Wind direction and temperature distribution in the combined data set.

However, if a single time period is considered the variable distribution is centered around a specific value, as expected, seen in figure 2.

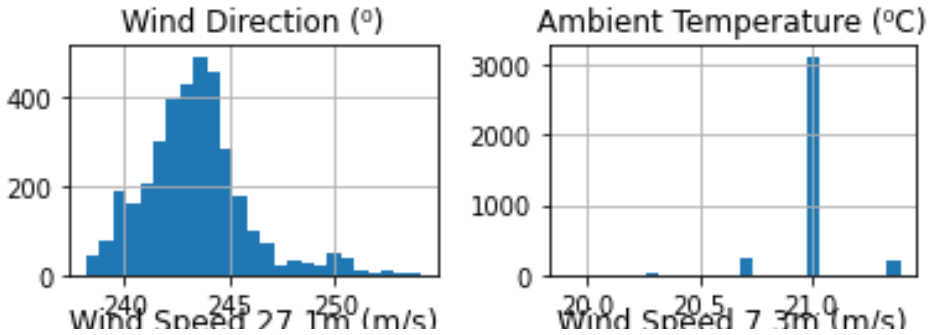

Figure 2: Wind direction and temperature distribution for a single time period in the data set.

Further, pairplots were made to understand correlation between various variables. Figure 3 shows significant correlation between Rated Power and nacelle Wind speed as expected. However, the relationship appears to be ‘‘noisier' than expected.

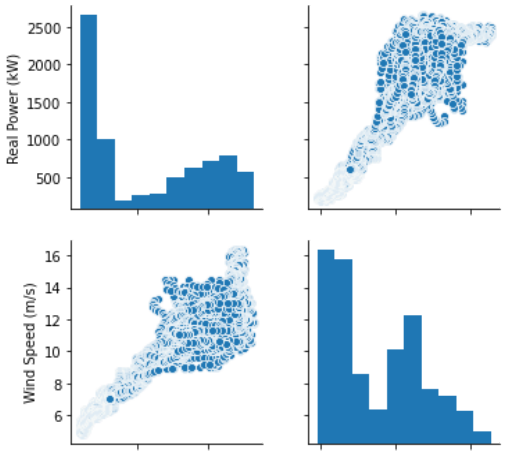

Figure 3: Power-Wind Speed pair plot.

Feature identification and creation:

The relationship between turbine power, P (Watts, W), and inflow speed, U (m/s), is given by the following expression:

$$P = \frac{1}{2} C_p \rho A U^3 \quad \text{eq. 1,}$$

where, $A = \frac{\pi}{4} D^2 = \frac{\pi}{4} \times (96m)^2 = 7.24 \times 10^3 m^2$ is the turbine rotor area, C_p is the turbine performance coefficient limited by the Betz limit that states maximum $C_p = 16/27$, and ρ is the air density (kg/m^3). Although, several simplification assumptions, such as homogeneity of U, have been made to arrive at this equation, it is an accepted expression for estimating turbine power.

Further, ρ can be estimated using the following expression:

$$\rho = 0.0035 \frac{p}{T} \quad \text{eq. 2,}$$

Eq. 1, that is simplified, physics-based models tested. Note: Since wind turbine power is artificially limited to rated value at higher wind speeds data above rated values is not considered for analysis below.

To estimate power using eq. 1 several trial C_p values were considered to achieve best fit. The best fit was achieved at $C_p = 0.34$, a low but reasonable value. Wind speed beyond a threshold could produce power more than the rated value, 2500 kW in this case. At this operating zone (Region III), the turbine output is regulated to produce only rated power (2500 kW). So eq 1. Output is capped at 2500kW beyond rated speed. The model verification exercise is done for 1-second, and 1-minute and 10-minute-averaged data.

Figure 4 plots the physics-based model estimate and the measured power (left panel) and the power-speed curve (right panel) for the 1-minute-averaged data. The physics-based model best predicts 1-minute averaged power with $C_p = 0.34$ and the model $Adj\_R^2 = 0.93$. TThe physics-based model significantly deviates from the measured values even for a modest increase in wind speed relative to the rated value (12 m/s assumed in this case) justifying further machine learning model development. This holds true even when the model values are deliberately capped at rated power (shown using the RED line). The variability is high at high wind speeds (10 m/s).

For all data fitting cases, the physical model performance is worse at higher speeds, especially the transition point between low and rated speed (Region II to Region III). The data used for this study was originally classified for high and low winds, along with winter and summer. To incorporate this information for subsequent modelling, additional categorical features, and are included are created such that 'one-hot-coding, will not be necessary later.

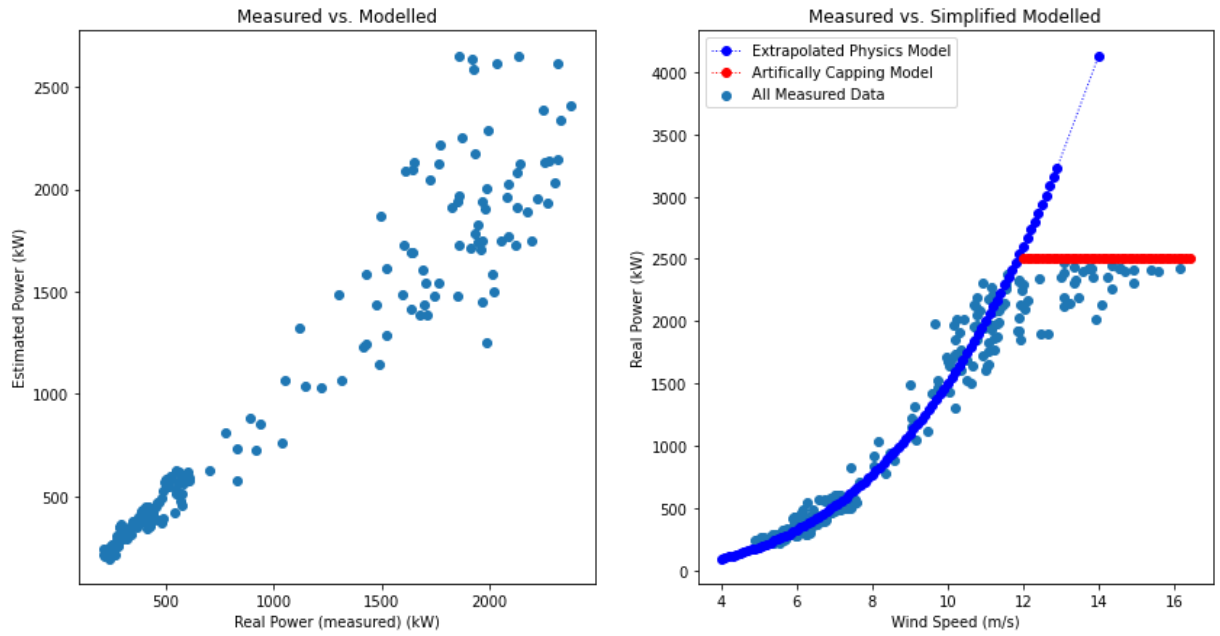

Figure 4: Physics-based model outcome.

When upstream measurements as required by IEC standards are considered in place of the nacelle measurements, the physical-based model best predicts 1-minute averaged power with $C_p = 0.45$ and the model $Adj\_R^2 = 0.76$. The model fitting performance is significantly reduced even for data with wind speed low as 8 m/s. Further, C_p for best fit is different from the previous attempt. Such inconsistency makes it difficult to select features and the appropriate model.

A Pearson correlation heatmap created to check multicollinearity. Based on the map, several features seem to be highly correlated . However, before dropping correlated explanatory variables, additional features are derived that are critical for wind farm meteorology.

Creating additional features:

Bulk Richardson number, Ri_B Stull 1988, and gradient terms $\Delta\overline{U}$, $\Delta\overline{\theta}_v$, are created. One-hot-coded categorical features (summer, winter, high, and low speed) with numerical values (0 and 1) are created as well.

Finally, shown in Figure 5 is a Pearson correlation heatmap for newly created features. The new Dataframe Pearson heat map shows very few variables with high correlation when compared to original data. The regions of high correlation either correspond to response (Power) vs. explanatory variables (Wind Speed) or mutually exclusive categories such as summer vs. winter. The exception is between standard deviation vs. TI, where the Pearson coefficient is as high as 0.97. These features are kept and may be eliminated later.

Further, clustering-based feature elimination is not attempted yet, since features defined above have physical relationships with each other. Again, such elimination will be considered in the subsequent steps.

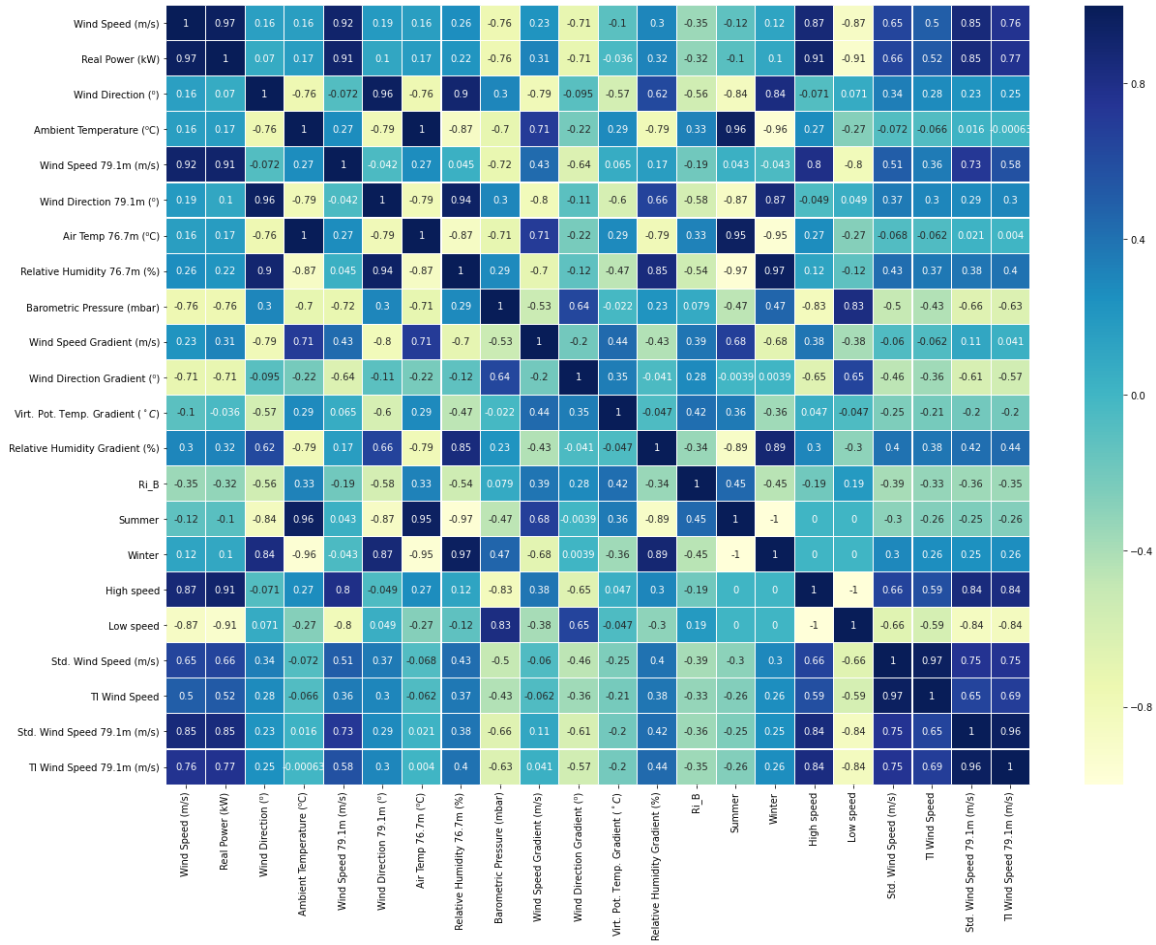

Figure 5: Pearson correlation heatmap new features.

Step 4: Preprocessing and training data development

A linear model is tested subsequently. However, from eq 1. $P \propto U^3$ and $\rho \propto 1/T$ both non-linear relationships. To incorporate non-linear effects the higher power of velocity and inverse Temperatures are added as features. Again from eq. 1, $P \propto product(\rho, U^3, ..)$. If a linear model is used, adding a log of features will have an equivalent effect of feature products as shown in eq. 7. Models are created using two sets of data are tested: Original with new features, Log of the original data with new features. A 75-25 train-test split is applied.

Step 5: Modelling

In this step the OLS model and Random Forest Regressor models were created. The Random Forest Regressor model hyperparameters were obtained using the Bayesian Optimisation. Table 2. Summarizes model performance for individual models.

Table 2. Model Comparison Summary

Model No.	Feature Engineering	Model	Adjusted_R^2 or (R^2)
0.	All features	Physics-based	.929
1a.	All features	OLS	0.989 (0.991)
1b.	W/O upstream tower	OLS	0.977 (0.979)
1c.	Log of all features	OLS	0.982 (0.984)
1d.	W/O upstream tower and log	OLS	0.981 (0.982)
2a.	All features	Random Forest Regressor	(0.969)
2b.	W/O upstream tower	Random Forest Regressor	(0.975)
2c.	Log of all features	Random Forest Regressor	(0.994)
2d.	W/O upstream tower and log	Random Forest Regressor	(0.985)

All machine learning models perform significantly better than the physics-based model. Amongst machine learning models, the Random Forest Regressor model with log of data for both the case with all features (2c.) and the case without upstream tower measurements (2d.) perform marginally better than the remaining machine learning models.

Step 6: Doccumentation

In this step parameter and conde finalization was performed. As shown in Figure 6, the prediction improvement is evident in the measured vs. prediction plots. This figure shows a model that uses data limited to usually available measurements and upstream tower measurements are not used. As seen in the left panel, the physics-based model seems to overpredict the power produced in Region II (below rated power), and has worse performance at higher wind speeds (blue dots). The physics-based model is unable to predict power in Region III (rated power) because the turbine is regulated (green dots). As shown in the right panel, the Random Forest Regressor model is robust enough to predict power even in Region III (green dots). Further, a fitting-based estimate of turbine efficiency Cp was not needed for the regressor model prediction.

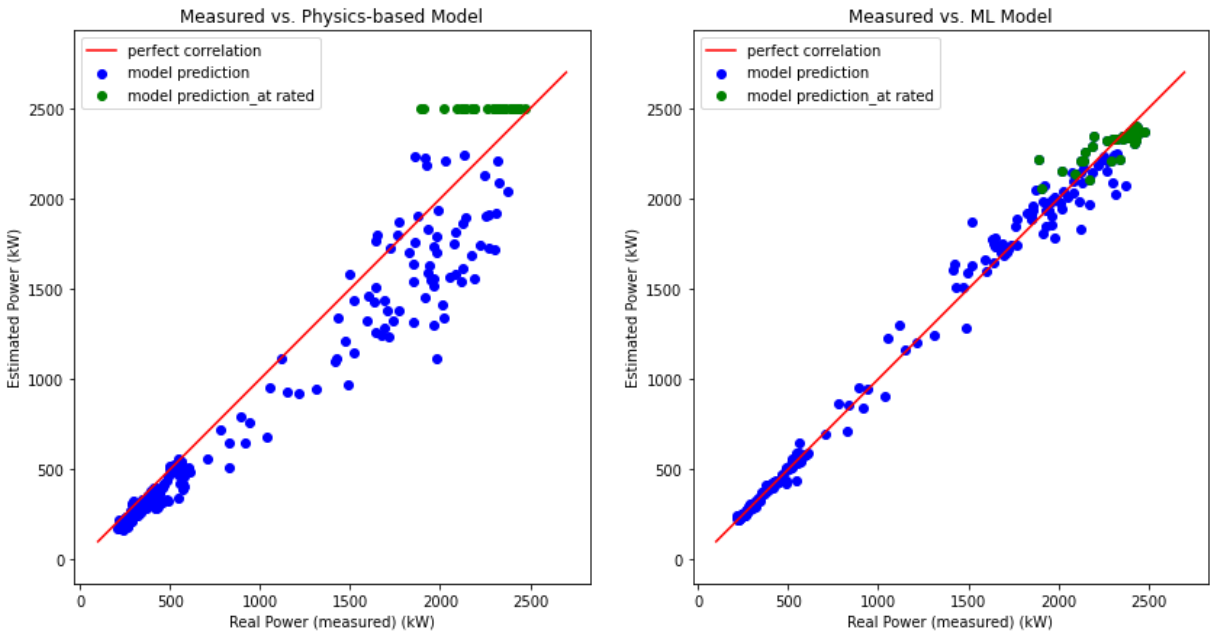

Figure 6: Random Forest Regressor Model Estimated vs. Measured Power.

Discussion and Future Work

The above work shows that Data Science Method (DSM) can be applicatied to develop Machine Learning (ML) models for wind turbine power curve calibration. Here, the EOLOS dataset timeseries was used to create and test models. Feature engineering involved adding features related to spatial (gradient) and temporal (standard deviation) variations. Further features representing categorical information regarding seasons and wind speed were added. Model testing was performed on the processed dataset and also on log of the processed data set. OLS and Random Forest Regressor models were tested. The Random Forest Regressor hyperparameters were determined using a Bayesian Optimiser. The Random Forest Regressor model that was created using the log of data had the best performance for datatframes having all features as well as for the one with limited features.

$R^2 > 0.98$ for the Random Forest Regressor model compared to $R^2 = 0.93$ for the physics-based model expression. The ML modelling result represents a significant improvement over the physics-based expression. It is recommended that the DSM be applied to a comprehensive database that includes more turbine data. Once an acceptable power curve calibration is obtained power predictions can then be used during turbine operation.