Wind-Turbine predictive maintenance for TOTAL

Bassel MASRI Guillaume FRANCHI

2/5/2021

Contents

1 Introduction

Many government policies are researching sustainable energy production resources in order to reduce their carbon footprint. In particular, harvesting the wind's kinetic energy through wind turbines accounts for nearly 28% of all installed renewable power capacity [reference.1]. However, heavy machinery entails many engineering challenges like operation and maintenance. It is estimated that about 30% of the total generation costs is induced by maintenance downtime which made predictive maintenance a hot research topic for the last few years. Recent breakthroughs in connected sensors, robotics and internet of things (IoT) have allowed manufacturers to collect big amounts of data from different parts of the turbine through their SCADA (Supervisory Control and Data Acquisition) system in order to monitor its behavior. Recent advances in machine learning techniques and programming platforms have opened a door to analyzing such amounts of data in the aim of monitoring faulty behaviors in the form of anomaly detection and failure predictions which is the main focus of this project.

Throughout this study, we introduce and discuss potential ways to approach analyzing the SCADA data in order to detect faulty behaviors of a wind turbine and, by extension, reducing its down time when performing maintenance tasks on its main components such as the generator and the gearbox.

2 The problem and the data

The following subsections are dedicated for describing the general workflow of this project and presenting the SCADA data that we have been given for the smart data project. Descriptions of the nature of the variables, the failure logs, and some exploratory data analysis are presented below.

2.1 Project workflow

The analysis steps that have been explored in order to model the normal behavior and identify abnormal behavior of wind turbines are presented in the flowchart in Figure 1.

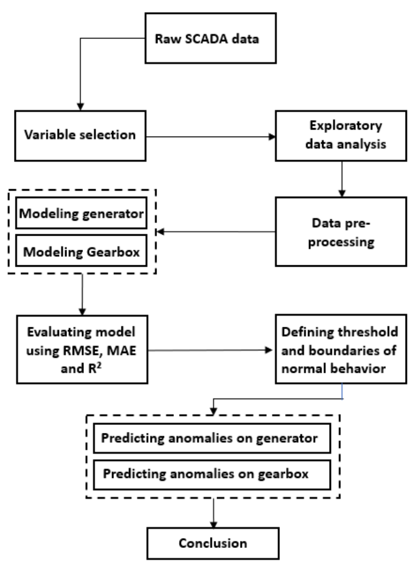

Figure 1: Workflow diagram of the study

The ultimate goal of this study is to model the normal behavior and predict a failure of a wind-turbine through data points that fall outside the envelop of what we define a *normal* behavior. To do so, we focus this study on two main components; the *generator* and the *gearbox*, as their maintenance task results in the longest downtimes. Each of the steps mentioned in the workflow will be discussed in greated detail throughout this report.

2.2 Signal Variables

Before diving into the details of each section mentioned in the workflow above, we first define the nature of the variables at hand. The SCADA system collects information about electrical and mechanical components through temperature sensors cleverly placed on multiple components such as Transformers, Generator, Gearbox, the hub that encapsulates the systems and some other hydraulic and electrical systems. Some additional information about the power production of the turbine as well as the angle orientation of the blades is also available.

The data is acquired on four Wind Turbines identified in a variable called *Turbine_ID* over two years (2016 and 2017) with a sampling frequency of 10 minutes. Besides the identifier of the turbine and the timestamp, a majority of variables correspond to measurements such as temperatures (in degrees celcius), power production (in Watt-hour), orientation (in degrees) and speed (in meters per second). Each measurement in the data set is made in the last ten minutes observed.

In total, the SCADA data consists of a collection of 83 variables with more than $400,000$ observations. One can find below the first few observations of year 2016 below:

Table 1: The first few rows and columns of the SCADA data

Turbine_ID	Datetime	Gen_RPM_Avg	Gear_Oil_Temp_Avg	Amb_WindSpeed_Avg
T07	2016-01-01 00:00:00	1254.9	45	4.1
T01	2016-01-01 00:00:00	1249.0	44	3.3
T06	2016-01-01 00:00:00	1248.5	43	3.8
T11	2016-01-01 00:00:00	1270.9	48	5.3
T01	2016-01-01 00:10:00	999.7	44	3.2

2.3 Failure logs

In addition to the provided SCADA data, we also have access to the failure logs of the turbines where we can find information regarding the component that failed, the time when the failure was logged, and some remarks made by the technician who observed the failure. In total, we have access to 28 failures over the two years of data for all 4 turbines. We present below the first lines of these logs.

Table 2: The first few rows of the failure logs

Turbine_ID	Datetime	Component	Remarks
T11	2016-03-03 19:00:00	GENERATOR	Electric circuit error...
T06	2016-04-04 18:53:00	HYDRAULIC_GROUP	Error in pitch regulat...
T07	2016-04-30 12:40:00	GENERATOR_BEARING	High temperature in ge...
T09	2016-06-07 16:59:00	GENERATOR_BEARING	High tempemperature ge...
T07	2016-07-10 03:46:00	TRANSFORMER	High temperature trans...

Our initial analysis of the failure logs shows that there are some spelling mistakes in the Remarks which makes the data prone to humain errors. Indeed, the failures are logged by technicians on site during working hours even though the alarm may have been triggered beforehand. One can also observe that the generator is the most sensitive component, in the sense that it is the one which suffered the most failures, especially with turbine T06. We explore the failures in further detail later.

2.4 Variable selection

Failures almost always come from abnormal overheating of the component in question. In the case of the SCADA data, the components of interest in this study are the generator and the gearbox. Naturally, out of the 83 variables provided, we will perform some manual variable selection based on engineering knowledge of the underlying systems as well as some other related research work ([references 2,3]). Only the variables that offer the richest insight into the behavior of a component are chosen to model the generator and the gearbox.

To model the behavior of the generator, we choose the generator's speed (in RPM) as the target variable and its related variables as predictors such as windspeed, power production (active and reactive), the Nacelle temperature, the generator's bearing temperature and the phase 1 temperature of the generator (phases 2 and 3 are highly correlated to this one). Since the sensors provide information about the minimum, maximum, standard deviation and mean of the data collected in the 10 minute-interval, we choose the average.

The authors who carried out the research mentioned in article [2] claim that the study carried out to model the gearbox in a wind-turbine has been successful when choosing the gearbox's oil temperature as a target variable. Following their footsteps, we choose the same variable as a target and the variables concerning the gearbox bearing temperature, hydraulic oil temperature, windspeed, production power (active and reactive), the blade pitch angle and the ambient temperature as predictors. Again, we take the average of the readings in the 10 minute-interval.

Finally, in order to be as precise as possible, we only model one turbine for each component separately. Based on our preliminary analysis of the failure logs, We choose to model turbine T06's generator (in which we experience the most failures) and T01's gearbox (in which we experience the only gearbox failure). It it worth mentioning that the turbine T09 had two gearbox failures, which makes it a good candidate for modeling but unfortunately we do not have access to its data.

2.5 Exploratory data analysis

As with all statistical modeling, data exploration is always the first step prior to any model development. Therefore, our analysis begins with a typical exploratory data analysis approach where we investigate the readings of some of the most important variables in the SCADA data to visually spot some important patterns. Such analysis is a crucial first step towards understanding the normal behavior of the turbines as well as understanding the abnormal behavior that triggered a failure.

2.5.1 Power production curve

The most important variable in SCADA data that indicates how well the turbine is operating is the power production data (in Wh). Since the manufacturers provide a datasheet describing the optimal functionning of a power production curve, we use that as a basis to understand how well the turbine is operating. According to the constructor, the ideal curve of said variable should be **S** shaped when plotted against the windspeed (in m/s).

Figure 2 displays the power production data with respect to the windspeed, by month of all the turbines, in year 2016. It shows that the theoretical curve of the power production (perfect S shape) is reasonably accurate when compared to real production data. This entails normal behavior of the turbine. However, we notice some rogue data points that seem like two dimensional outliers. For example, windspeed of $10m/s$ should non-zero power production data which is not always the case. This indicates that either the whole turbine was down for maintenance related reasons or that it is, in fact, a two dimensional outlier which we will explore later on in detail.

2.5.2 Generator's failures

In order to analyze the failure of the generator, it is important to understand its causes. To do so, we choose to plot only turbine T06's generator related temperatures and add thick black lines at each failure date. To avoid overcrowding the plots, we will plot the data in a smaller time window (from 2016-06-01 to 2016-12-01). This time-window in which we plot the variables is the time-window in which we had most of the generator failures.

Preliminary analysis of the plots displayed in Figure 3 for the generator variables show that there is some missing values. In fact, in the missing 10 day period represented as the time gap from 2016-07-10 to

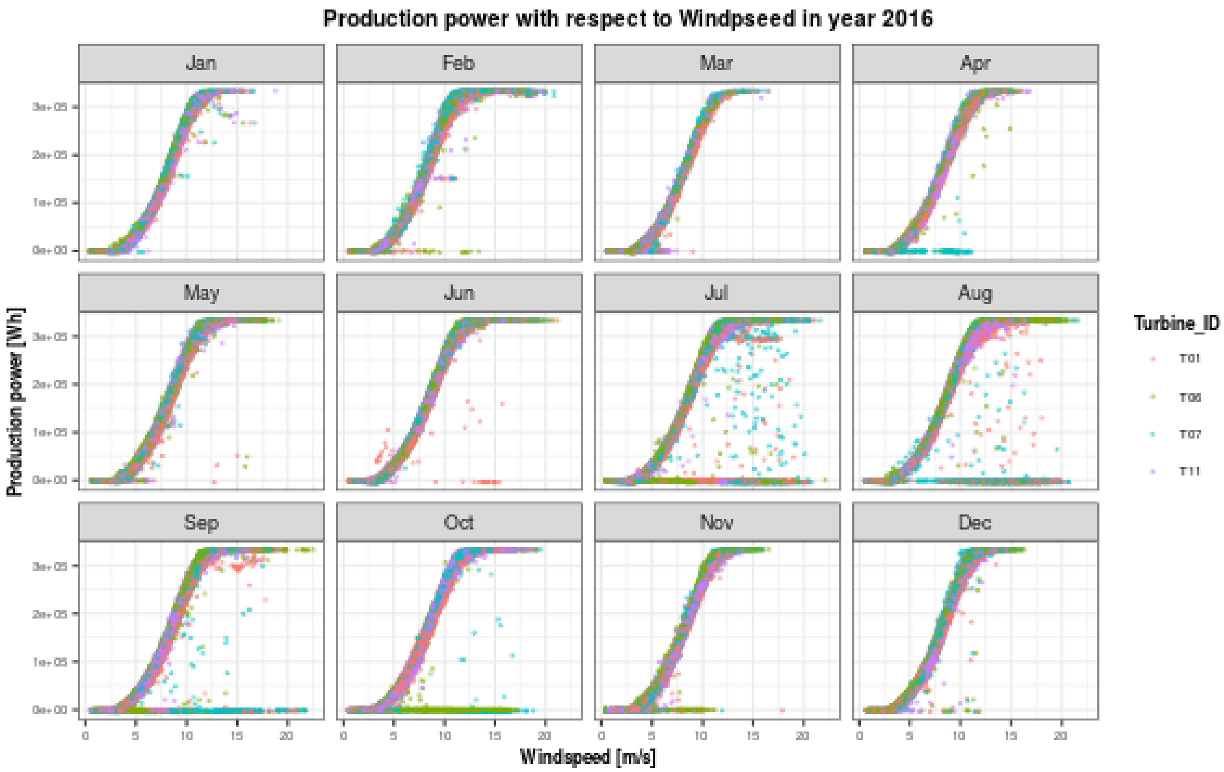

Figure 2: Power production curve in all turbines for 2016 data with respect to the windspeed

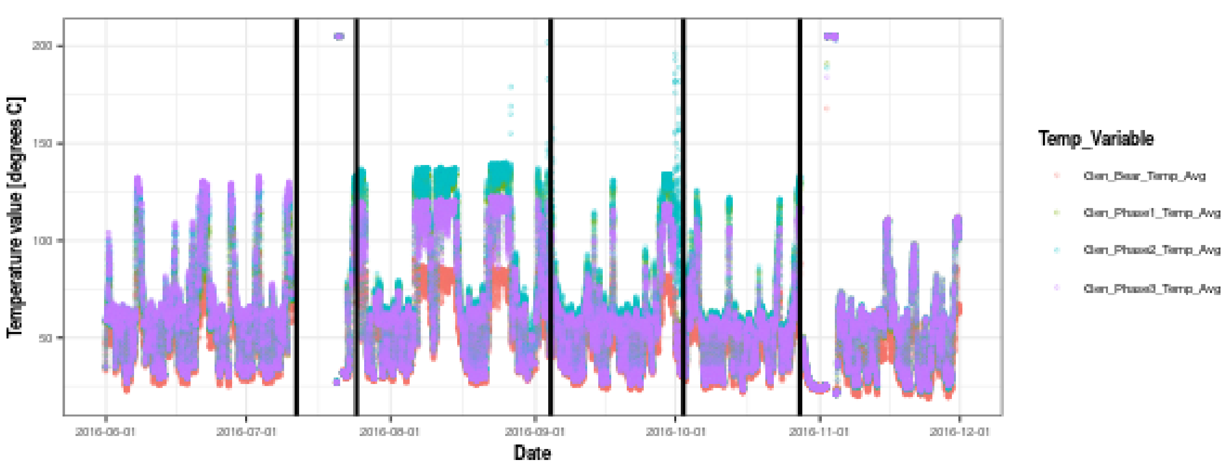

Figure 3: Plot of the generator's temperature variables with its dates of failures in black lines

2016-07-20, the generator has been replaced on 2016-07-11 according to the failure logs and the turbine remained un-operational for a few days afterwards as part of the maintenance work.

In addition, We notice that not all generator failures indicate abnormal temperature behavior. In fact, out of the 5 failures represented in Figure 3 that heppened in turbine T06's generator, only two are reliable; the third and the fourth failures for they show abnormal temperature behaviors (up to 200 degrees). Said failures will be considered as a reference in the anomaly detection techniques we will discuss in later sections.

Finally, the plots show that some rogue temperature measurement that were up to 200 degrees for all 4 variables were not discarted as a failure in the logs. Zooming in further on this particular event for which no failures logs have been registered, we get the plot in Figure 4.

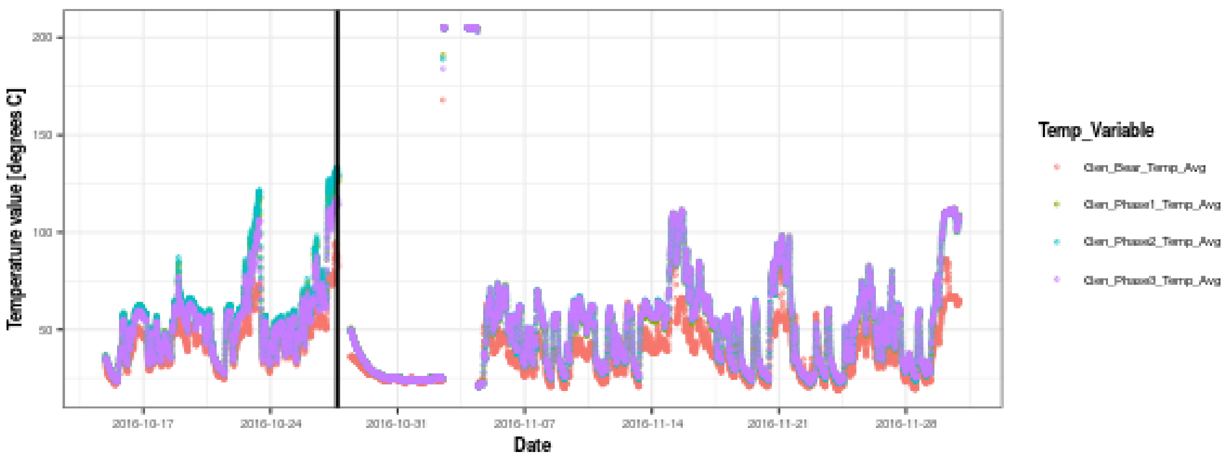

Figure 4: Plot of the generator's rogue temperature points which have not been detected as failures

The logs show that on 2016-10-27 (the black line in Figure 4), the generator has been replaced. A few days later, the sensors were clocked at 205 degrees celcius for a three days straight from 2016-11-02 to 2016-11-04. Such high values can be explained by multiple reasons. Either they are outliers (e.g. caused by sensor calibration issues) or an undetected anomaly (e.g. not registered in the failures logs).

2.5.3 Gearbox's failures

Upon examining the failure dates of the gearbox component of turbine T01, we notice that only one failure occured in 2016-07-18. The remarks noted by the technician were "Gearbox pump damaged". Therefore, we will visualize the temperature variables and the failure date in a zoomed-in 2 weeks window (i.e. 7 days before the failure occured and 7 days after).

Figure 5 shows that around the date of the failure, a high variability in the pitch angle is observed. In some days, the angle ranges between 0 and 25 degrees. On more rare occasions, the blades are tilted to more than 75 degrees in a matter of minutes. It can be seen that the blade angles affect the gearbox's temperatures. This can be clearly spotted right after the failure date where the blades were continously maintained at around 78 degrees for more than 24 hours which dropped all gearbox's related temperatures drastically.

Similarly, we can observe that on some days the gearbox bearing temperature goes higher than 60 degrees celcius (above the dark red horizontal line on the plot). In fact, according to a study conducted on wind turbines' gearboxes [ref.4], under normal conditions the temperature in the gearbox (oil and bearing) should not surpass 60 degrees. Indeed, the reliability of the gearbox becomes less than 50% should its temperature exceeds 60 degrees celcius. A combination of all the observed abnormal conditions could have triggered the failure in the gearbox.

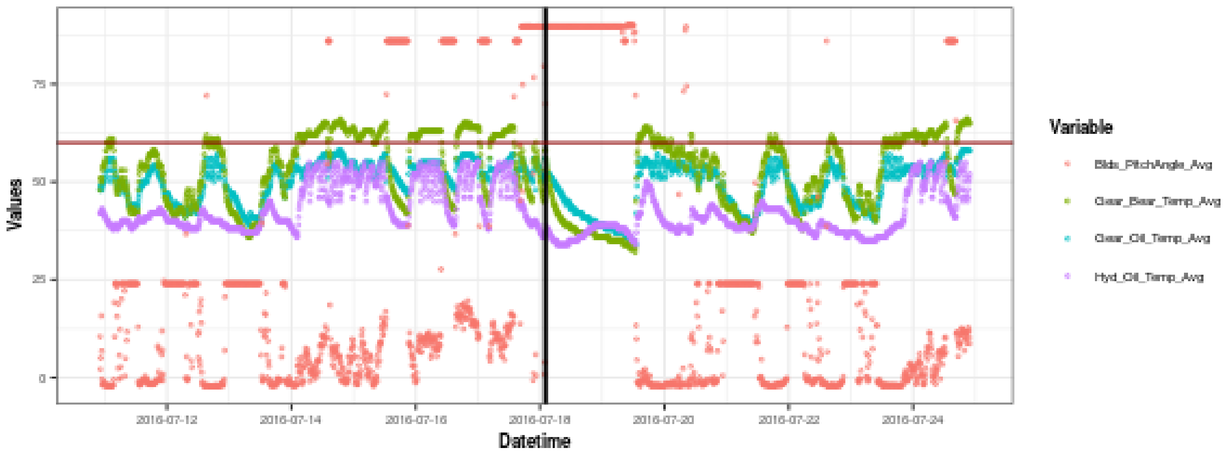

Figure 5: Plot of the gearbox's variables, the failure date in black and the 60 degrees threshold in darkred

3 Approach

3.1 Modeling and evaluation

After variable selection and exploratory data analysis, we proceed in the following sections by introducing the steps that lead to modeling the normal behavior of the components in question. The modeling process entails splitting the data in two parts; a modeling set which contains 18 months of data and a test set which contains the last 6 months of data. The modeling set is then split in two parts for training and evaluating the model according to a 80-20% ratio. Data shuffling is avoided to respect the time arrangmenet of the observations.

Throughout the study, we apply all preprocessing steps uniquely on the modeling set, leaving the test set unfiltered and unclean. Based on selected components and their corresponding predictors, we fit a tree-based extreme gradient boosting (XGBoost) model on the train set and evaluate its performance on the validation set using root mean squared error (RMSE), mean absolute error (MAE) and the R^2 to demonstrate the goodness of fit. XGBoost models offer a wide range of hyperparameters such as [ref.5 Hands-on machine learning with R] :

- Regularization hyperparameter to provide an extra later of protection against over-fitting
- Early stopping criterion to stop growing trees when they offer no more improvement to the model
- Parallel processing (since it is sequential by nature)
- Choice of a loss function to optimize the gradient boosting models

The residuals produced with the help of the model will help perform failure prediction on the test data via thresholding on the difference between the residuals. The latter is discussed in more detail in the last subsection on predicting failures.

3.2 Preprocessing steps

3.2.1 Univariate outliers and anomaly detection

One of the highlights of this study has been exploring ways to approach analyzing the rogue data points that fall far beyond the mean of the overall data. Usually, data points that are significantly different from other observations whithin the same variable are casted as univariate outliers. In the case of a timestamped temperature measurement in a carefully engineered component like a wind-turbine's generator or gearbox, An abnormal observation may be an indication of an anomaly.

Our exploratory data analysis has indeed shown that some generator failures, e.g. the third and fourth ones visible in Figure 3, were triggered by anomalies in temperature related variables. Therefore, in this section, we will study the variables independently as univariate time series in order to perform anomaly detection using the package **anomalize** which has been developped specifically for this reason.

One important thing to do with Time-Series data before anayzing its abnormal behavior is Time-Series decomposition. After decomposing the Time-Series into trend, seasonal and remainder components, the anomaly detection is carried out in the remainder component which, under normal curcimstances, should not have any structure. In fact, the remainder should be completely random with no rogue observations if the data does not contain any outliers. If one is detected, then it is flagged as an anomaly.

For the purpose of demonstration, the plot produced in Figure 6 shows how an anomaly is detected on the generator's bearing temperature

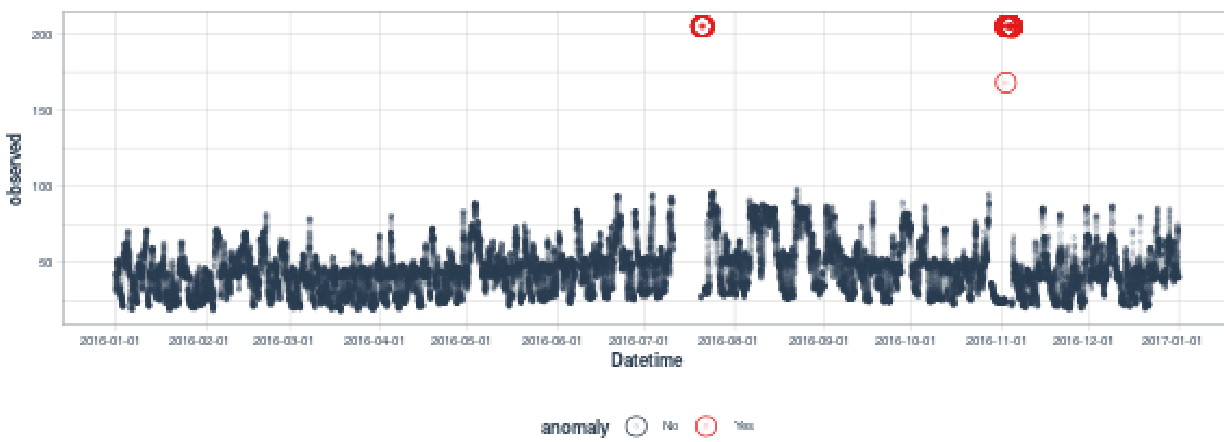

Figure 6: Plot of the generator's bearing temperature anomalies on T06 in 2016

It becomes immediatly clear that temperature readings above 200 degrees celcius are flagged as anomalies as we can see in the plot in Figure 6. Similarly, we apply the same technique on all *temperature related* variables in our data set to identify the dates of the anomalies of each turbine. The advantage of using such method to detect anomalies is that it offers rich insight on how the data behaved by seasonality. For example, the average components' temperatures in the summer is naturally slightly higher than the one during cold winters. The difference between summer and winter temperatures in France may reach up to 40 degrees celcius in some areas. By extension, a Time-Series decomposition returns a lower number of anomalies by taking into account seasonality effects.

We will apply this method on all temperature related variables for all turbines and remove them from the modeling dataset (i.e. the set that contains the first 18 months of data).

The modeling set contains 312,852 out of which 11,758 rows has been identified to contain at least one anomaly. Therefore, we proceed by removing the rows all together in order to be able to model the normal behavior later on without any rogue temperature readings.

3.2.2 Multivariate outlier detection

In this part, we will work upon the data set obtained after removing the oultiers in the previous section.

The second criterion used to detect and remove outliers is based on the theoretical power curve of a turbine, which is supposed to be 'S-shaped'. We can see in Figure 8 that a lot of data are located far away from this curve, and could be considered as outliers.

First of all, we can immediately remove the data for which the power production is ridiculously low (strictly less than $1W/h$). These data were not removed in the previous step, since there were a lot of values with zero production. Such a low value for production means that the turbine is either stopped or we have an abnormal value. In both cases, the data can be considered as an outlier.

Then, we fit a logistic model on the 'semi-filtered' production data, which is supposed to model the 'S-shaped' curve :

$$y = \frac{Asym}{(1 + \exp((x_{mid} - x)/scal))}$$

where y stands for the production and x for the wind speed. $Asym$, x_{mid} and $scal$ are parameters which have to be estimated.

We can estimate visually on Figure 7 below the goodness of our fit. It is actually pretty good with a RMSE of 12 395, which needs to be put into perspective regarding the high values of production.

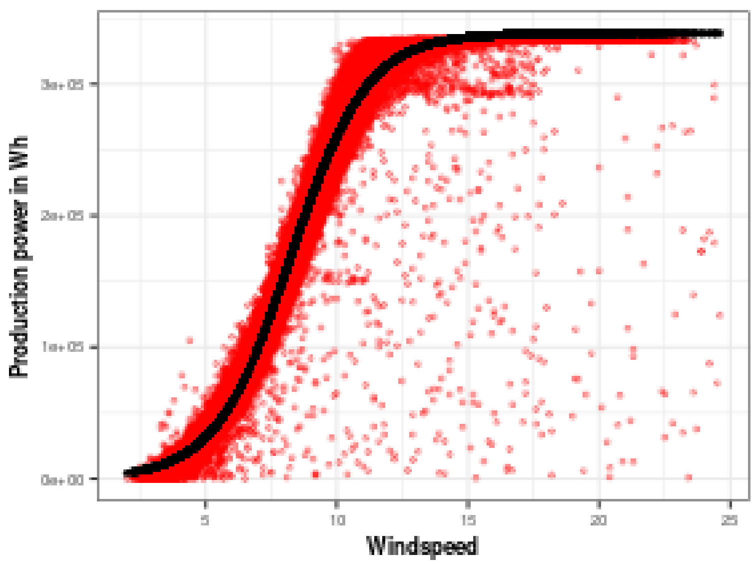

Figure 7: Semi filtered production data and fitted logistic model

Finally, we fix a threshold of 30 KW/h for the residuals. Whenever a residual has an absolute value greater than this threshold, it is considered that the observed value is an outlier, and therefore removed from our clean data set. We can see on Figure 8 below the obtained scatter plot of power production for the clean data set.

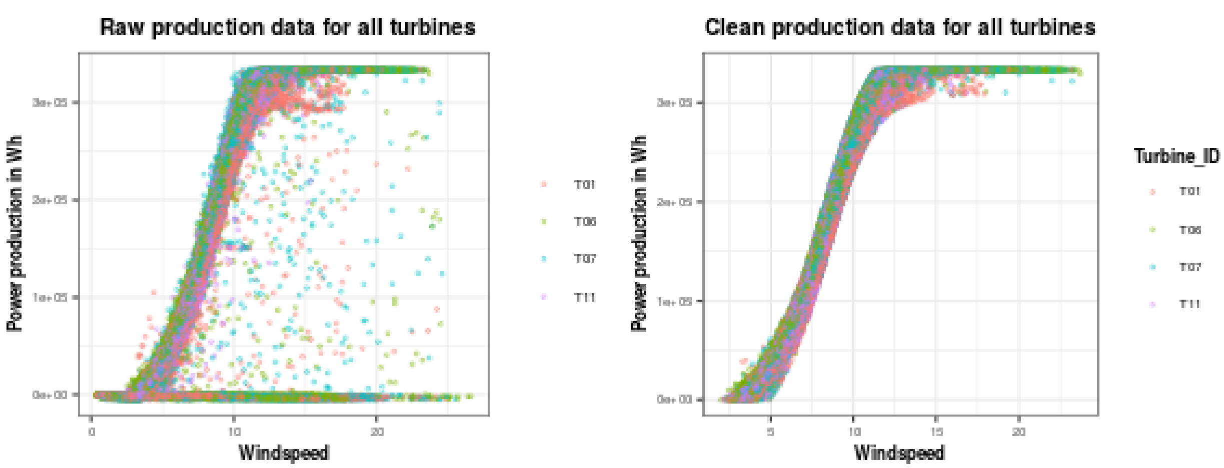

Figure 8: Before Vs After multivariate outliers removal

Note that 104,927 observations have been removed during both steps of our cleaning, which corresponds to 33.54% of the original data set. Although this may seem like a lot of observations removed, bear in mind that many observations were discarted as abnormal behavior in the original dataset. This includes un-operational turbines for which we do not produce any power (e.g. left plot in Figure 8).

However, if we are convinced that a lot of outliers have been removed, we cannot be sure that the clean data set does not contain any. Nevertheless, considering a value as an outlier can be sometimes subjective, we can reasonably consider that the data set that we will use for training our model contains a large majority of usual values, and is clean enough.

3.3 Modeling normal behavior

In order to model the normal behavior of the Gearbox, we will fit an XGBoost model on the training set (which is 80% of the modeling set). As mentioned in the section Variable selection, we will use the gearbox's oil temperature as the target variables and the gearbox's bearing temperature, the hydraulic oil temperature, the ambiant windspeed, the production power (active and reactive), the blade's pitch angle and the ambiant temperature as predictors. The fit is done only on turbine T01's gearbox and predict values for the validation set in order to evaluate the model. Lastly, we use the test data (i.e. the last 6 months of data) to test the model.

Similarly, we will model the normal behavior of the generator using the generator's rpm as the target variable, and Windspeed, power production (active and reactive), the Nacelle temperature, the generator's bearing temperature and the phase 1 temperature of the generator as predictors. The fit is done only on turbine T06's generator and predict values for the validation set in order to evaluate the model.

It is worth mentioning that for turbine T01, no failures have been registered for the last 6 months of data.

3.4 Predicting failures

Predicting failures after we define boundaries of the normal behavior of the turbine using the model's output is discussed herein. In order to capture abnormal behavior, we will first fix a threshold based on which we classify the residuals as abnormal.

To formulate the problem at hand, let n be the number of observations and $\mathbf{r}$ the vector of residuals :

$$\mathbf{r} = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, \cdots, y_n - \hat{y}_n)$$

Where y_i and $\hat{y}_i$ are the target variable and its prediction. For each component, we will define the threshold based on the residuals obtained on the validation set.

By examining the difference between observed and predicted values for the test set (on which we do not apply any preprocessing), we examine the data points that fall outside of the normal behavior envelope defined by the threshold to capture abnormal behavior that occurred in the last 6 months of data.

$$\begin{cases} \text{If } |y - \hat{y}| > thresh & \rightarrow \text{ abnormal} \\ \text{Otherwise} & \rightarrow \text{ normal} \end{cases}$$

Figure 9 illustrates an example of the normal behavior envelope and how it may help detect abnormal behavior if a data points falls outside its reach.

In addition to the threshold, the previous exploratory analysis showed that the normal behavior boundaries of the gearbox oil should not 60 degrees. Therefore, for this particular component, an additional constraint will be applied as follows :

$$\begin{cases} \text{If } y, \hat{y} \geq 60 & \rightarrow \text{ abnormal} \\ \text{Otherwise} & \rightarrow \text{ normal} \end{cases}$$

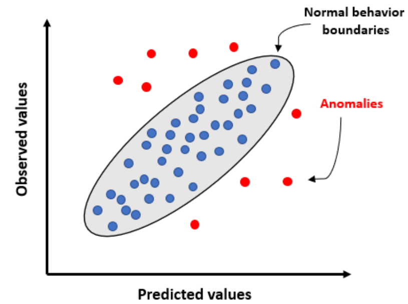

Figure 9: Demonstration of normal behavior boundaries and anomalies on the last 6 months of data

4 Modeling results

In this section, we will discuss the results obtained after training our models. In addition, we will deal with our main concern which is failure detection of the components at hand. Since the results are very different between the two components studied, this section is divided in two parts.

4.1 Generator results

The first result we can comment is the goodness of fit of the regression made by the gradient boosting method `xgboost`. Indeed, the evaluation metrics after predicting the target values on both our training and validation sets are given in Table 3.

Table 3: Evaluation of the fit for Gen_RPM_Avg

	RMSE	MAE	R
Training set	2.60	1.79	0.9999
Validation set	13.35	6.48	0.9990

We can see on Figure 10 how precise our fit is.

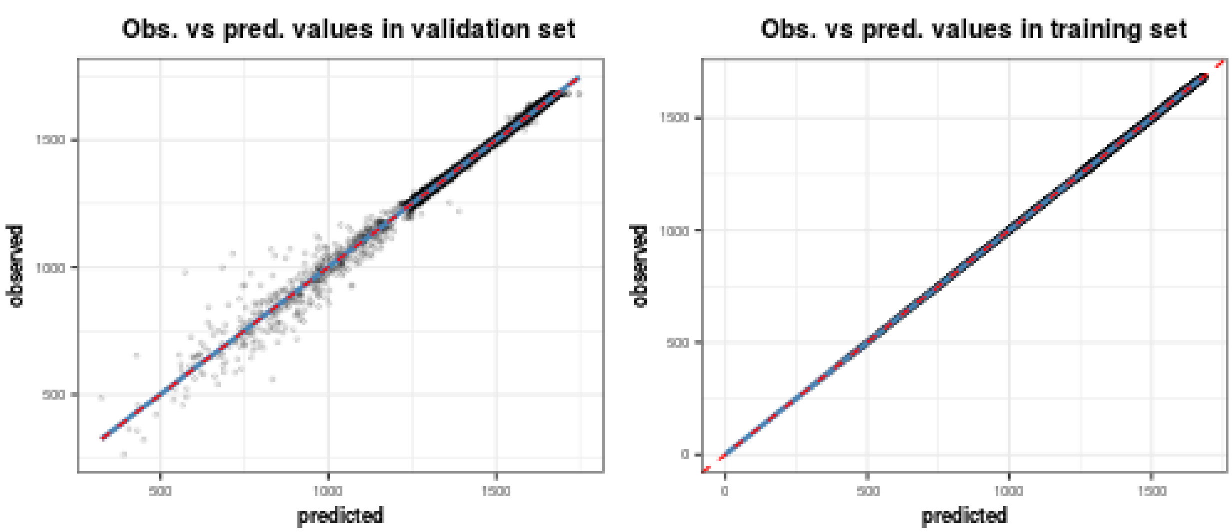

Figure 10: Observed Vs Predicted values of the generator's rpm

We then study the residuals obtained on the validation set, and obtain a standard deviation of $\sigma_{gear} \approx 1.02$ for a distribution of the residuals once again close to a Gaussian normal distribution (see Figure 11).

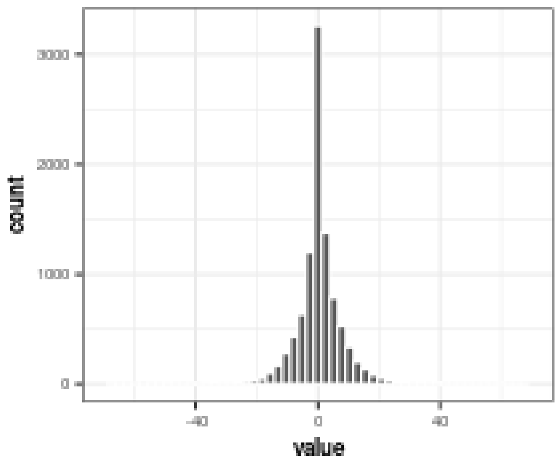

Figure 11: Residuals distribution on the validation set for the generator's rpm

Therefore, a naive threshold could be : $thresh = 3 \times \sigma_{gear} \approx 3.1$, and this is the one that we take for the rest of the study. If the model is accurate on a clean data set, we still have to know how it behaves on an unprocessed one. By evaluating our model on the test set, i.e. the one with the last unprocessed six months, we get the results stored in Table 4.

Table 4: Evaluation of the fit for Gen_RPM_Avg

	RMSE	MAE	R
Test set	241.73	126.89	0.9968

The evaluation metrics are obviously not as good as for a clean data set, which is normal.

In Figure 12, we can observe that the fit on the test set is nevertheless quite good, but the model has a lot of difficulties to fit the low values of the Generator RPM (less than 500 RPM). Predicting the generator's failure is performed by defining a normal behavior envelope similar to the illustration in Figure 9. By declaring observations for which the residual exceed the threshold that we fix as an anomaly, we obtain 8,184 out of 26,095 abnormal values in the test set for turbine T06.

Figure 13 shows the predicted and the observed values in time to see how well the model is able to predict the behavior of the turbine in question. Recall that for this period, no failures have been declared on this turbine, which naturally puts our definition of our threshold to question. This problem will be discuss further in *CONCLUSION*.

4.2 Gearbox results

Once again, we start with the goodness of fit of the regression made by the gradient boosting method `xgboost`. The evaluation metrics after predicting the target values on both our training and validation sets are sored in Table 5.

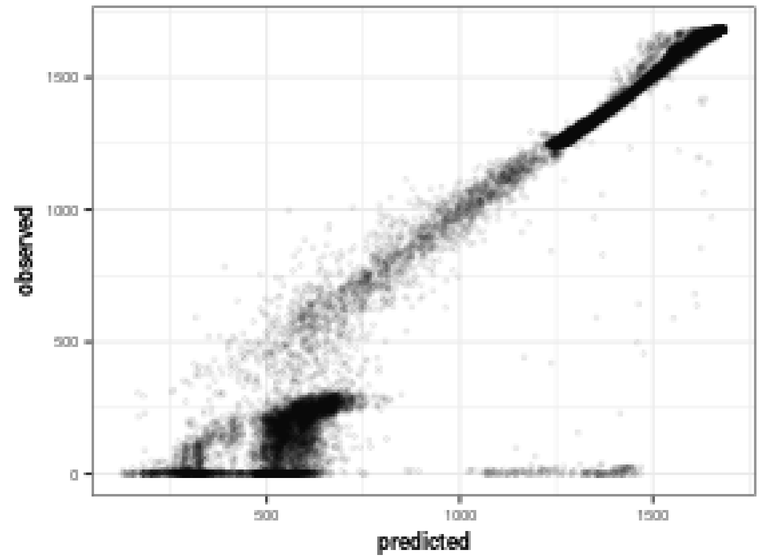

Figure 12: Observed Vs Predicted values for the generator's rpm in test set

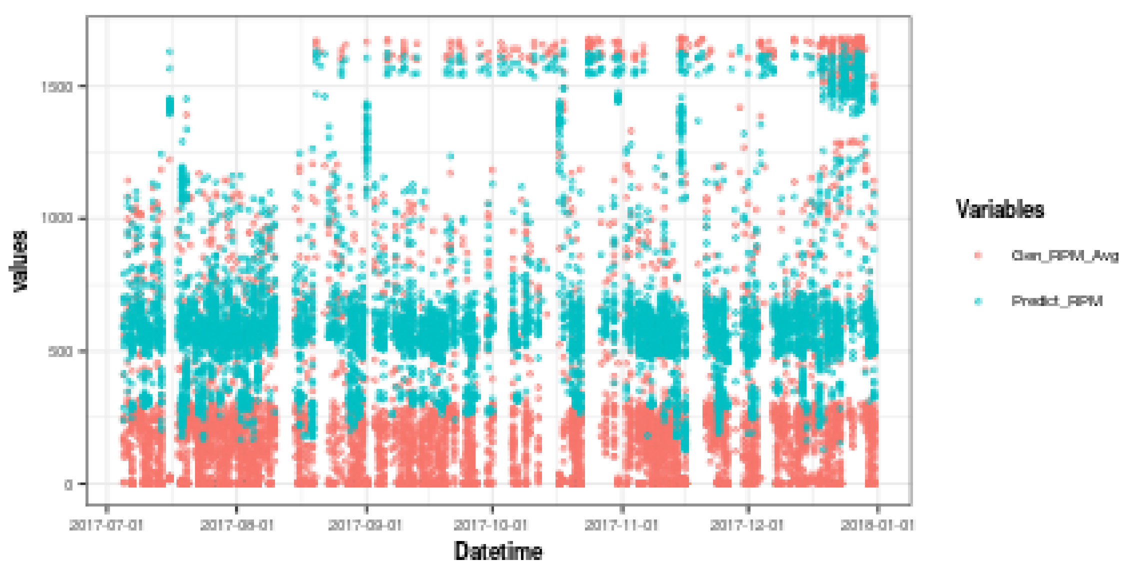

Figure 13: Predicted anomalies by thresholding on the Generator

Table 5: Evaluation of the fit for Gear_Oil_Temp_Avg

	RMSE	MAE	R
Training set	0.22	0.16	0.9980
Validation set	1.02	0.76	0.9732

We can see on Figure 14 how precise our fit is.

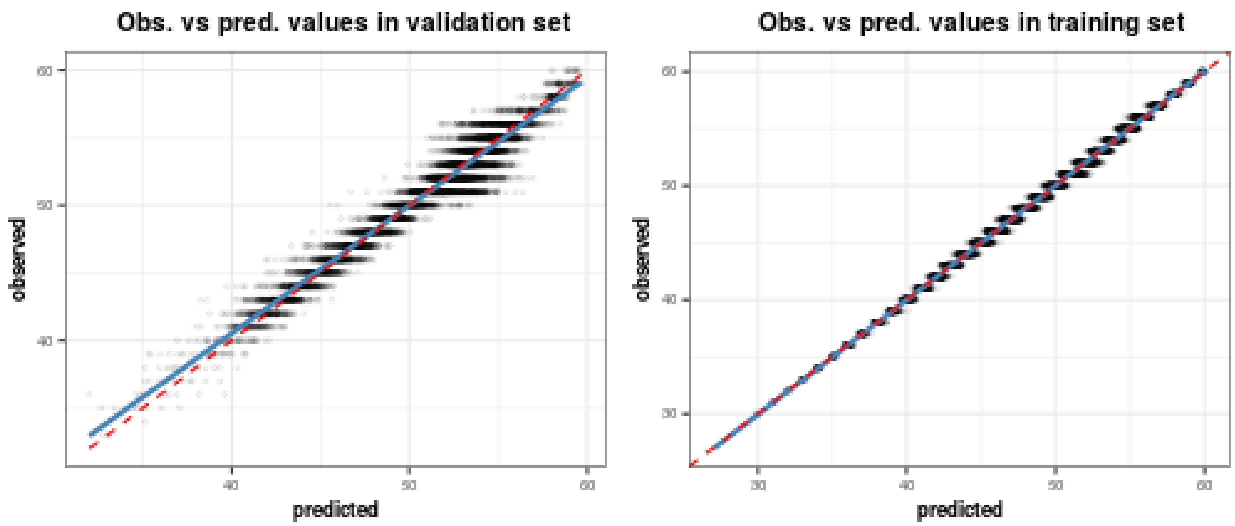

Figure 14: Observed Vs Predicted values for gearbox oil temperature

Next, similar to the steps followed in the previous section regarding the generator, we study the distribution of the residuals obtained on the validation set which gives a standard deviation of $\sigma_{gear} \approx 1.02$. Once again, a nice Gaussian distribution is observed upon plotting the results which we can see in Figure 15.

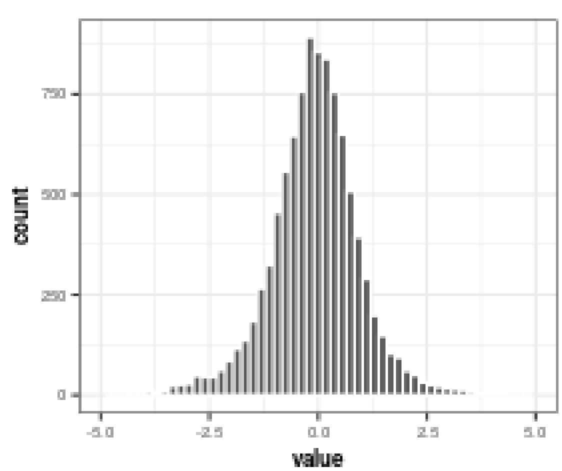

Figure 15: Residuals distribution on the validation set for the gearbox's oil temperature

Therefore, a naive threshold could be taken as $thresh = 3 \times \sigma_{gear} \approx 3.1$, which will be the one that we fix for the rest of the gearbox's study.

By evaluating our model on the test set, we get the results presented in Table 6. The evaluation metrics indicate a good fit in our test set, at least much better than for the generator. This fact will be discussed in the conclusion.

Table 6: Evaluation of the fit for gear oil temperature

	RMSE	MAE	R
Test set	1.34	0.96	0.9771

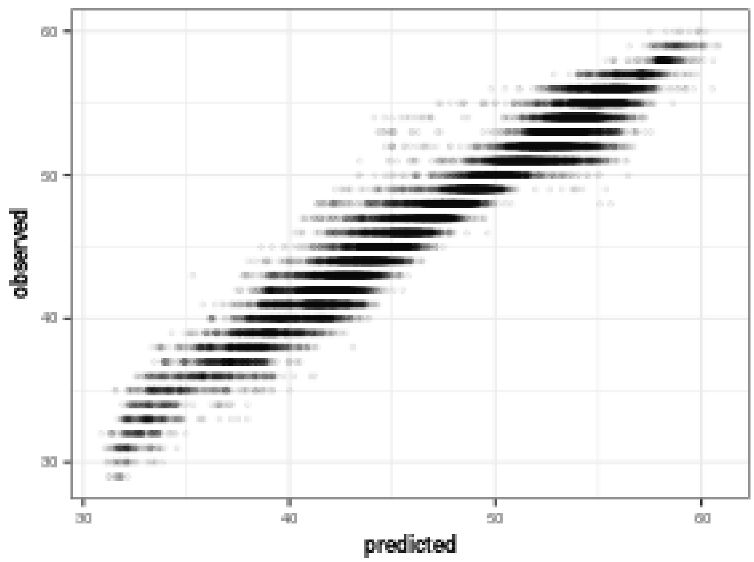

Figure 16: Observed Vs Predicted Gear_Oil_Temp_Avg values in test set

In Figure 16, we can observe that the fit on the test set is slightly different from the one obtained on the generator. Nonetheless, it is more consistent in the sense than the scatter plot has a more convex shape and an overall better linear trend. Recall that, according to *PAPER CITATION*, if the temperature of the oil of the gearbox goes beyond 60 degrees Celsius, then it is abnormal. We thus declare as an anomaly each value for which the residual exceeds the threshold or the temperature exceeds the 60 degrees limit.

A total of 881 anomalies are obtained out of 26,019 observations in the test set for turbine T01. Figure 17 shows the predicted and the observed values in the last 6 months to see how well the model is able to predict the turbine's behavior. Indeed, a handful of values have been detected to surpass the 60 degrees celcius threshold which is considered an anomaly. Due to the fact that the difference between predicted and observed values is small, some values may be falsely declared as anomalies.

5 Conclusion

Predicting a component's failure before it happens is still a widely researched field in engineering. Recent advances in data analysis and statistical modeling have allowed to model normal behavior and predict anomalies in a data-driven approach. The study conducted on the wind turbine's main components has indeed shown that modeling based approaches can be applied in order to detect anomalies.

To some extent, we have managed to model the normal behavior of the components studied by a gradient boosting regression trees. However, this modeling was performed on a clean data set which is not the case in real life. Therefore, our thresholding method predicts too many anomalies on the unprocessed data set. In fact, an access to full alarm logs would bring very useful insights on how well the model performs.

Figure 17: Predicted anomalies by thresholding on the Gearbox

For the gearbox : the number of anomalies detected is quite reasonable, and could indeed correspond to an unusual behavior of the gearbox, even if it doesn't lead to a failure.

For the generator : the number of anomalies detected is very high (approximately a third of the data), so we can say our threshold is not precise enough. Actually, we can observe on Figure 13 that many values flagged as anomalies correspond to a very low generator speed. Further studies should focus on determining why the generator speed has such a lot of low values to determine which behavior is normal or not.

Ultimately, distinguishing outliers from anomalies would be important in further studies. In addition, studying in detail the anomalies detected with our method should be carried out to fine tune the threshold. Similar modeling steps could be performed on different turbines which would confirm the scalability of the model. Future work could also be conducted on all of the wind-turbine's components in order to trace back the failure to a specific system.