EOSC 510

Final Project Report

Neural Network vs Stepwise Linear Regression Analysis with PCA For Optimizing Wind Turbine Energy Production

Students:

Dana Bazazeh 30071682
Ilya Ganelin 70163134

Table of Contents

I. Introduction……3

II. Data……4

III. Methodology & Results……5

IV. Discussions & Conclusions……12

V. Outlook……12

Abstract

In this report, we aim to develop a model that will predict the trend of the winding temperature for the E3120 wind turbine, manufactured by Endurance Energy, and use the results as part of a mathematical model that will raise the efficiency and economic viability of the turbine by shifting the cutoff threshold temperature accordingly. The data used was collected from the E3120 turbine in the UK over a span of 27 hours, with 5023 samples each having 16 variables that will be used as the predictors (a collection of voltage, power, speed and temperature readings) and 1 predictand variable, which is the winding temperature of the turbine generator. Two machine learning models were evaluated, a time series Neural Network model (NARX) and a classical Stepwise Linear Regression model. The data was divided into training, validation and testing sets and different model parameters were tuned. The performance of the models was evaluated based on two metrics, the root mean square error (RMSE) and the correlation coefficient (rho) between the target and the predicand values. It was found that the Neural Network model performed better in terms of both the metrics, and thus was selected for deployment.

I. Introduction

1.1 Background

Wind turbines are used to transform kinetic energy of the wind into other forms of energy. In the modern world, wind turbines are used to generate renewable clean electrical energy. The wind moves a propeller and causes the motor of a generator to rotate by means of a mechanical system [1].

1.2 Problem Statement

The continuous generation of energy in high winds above rated power, combined with insufficient cooling effect due to high environmental temperature might cause overheating of the generator above manufacturer's set threshold temperature. The heat transferred into the generator windings is what drives a current and produces electricity. However, when the generator temperature goes beyond a preset threshold, the generator is programmed to switch off until the temperature cools down. This is important because the expected life of the generator drops in half for every 10 degrees exceeded beyond the temperature threshold limit. However, this affects the overall maximum efficiency of the turbine, as the overall average operating temperature of the generator might be lower than the threshold, due to alternating wind speeds. Attempting to ineffectively adjust the threshold might cause reduction in the generator's lifespan, which leads to the expensive maintenance, replacement costs and increase in projects break-even point. Therefore, a smarter, dynamic threshold adjustment system is needed to maximize the wind turbine energy production, based on forecasting the winding temperature of the generator over time.

1.3 Objective

The main aim of the project is to raise the efficiency and economic viability of a wind turbine by developing a model that will predict the trend in winding temp for a specific wind turbine. By means of a mathematical model, the predicted data will be analyzed to shift the threshold temp of the generator winding without causing too much damage to the generator. Hence, the major objective of the project is to develop a suitable predictive model that could predict the generator winding temperature trend over time, given a set of variables such as current and wind speed. This will ultimately be used in a mathematical model to increase the Energy output of the turbine by adjusting the temperature shutdown threshold.

II. Data

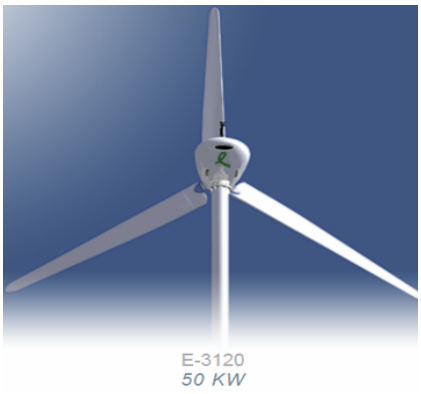

Figure 1: E 3120 Wind Turbine

The data used in this analysis has been collected from the E-3120 50KW wind turbine produced by Endurance Energy Mftg Ltd. and found in the United Kingdom [2]. It consists of 17 total variables, 16 of which will be used as predictors and 1 (generator winding temp) as the predictand. The variables are divided into 3 main parts. The first is electrical values of current, power and voltage of several components in the turbine. The second is different values of the gearbox, nacalle and outside temperatures and the final part is the wind speed. The sampling rate of the variables is 2/min, with a total of 5023 time samples, covering 27 hours, just over a day's trend.

[Note: The data is private and not available online, but a source for the turbine used has been referenced above]

Table I: Data set showing predictors and predictand variables

Predictors (X)	Range (X)	Predictand (Y)
kVAR	-362- 59	Generator Winding Temperature 300-1034
kW LastMin Avg	-15 - 770	
Power Kw	-24 - 928	
Eqv Power Factor	-8976 - 9971	
Eqv Active Power	-246 - 9694	
Eqv Reactive Power	-3624 - 599	
L1 Phase Voltage	23752 - 25307	
L2 Phase Voltage	23662 - 25093	
L3 Phase Voltage	23818 - 25122	
Corrected Wind Speed	8 - 162	
L1 Current	0 -1369	
L2Current	0 -1359	
L3 Current	0 - 1330	
Gearbox Oil Temp	142-218	
Nacelle Air Temp	80-157	
Outside Air Temp	492-678	

III. Methodology & Results

A. Preprocessing

The data available has 2 main issues. First there are many variables (16), and this creates a space complexity problem. Given the large number of samples, we get a space domain of 16 x 5023 = 80368.

The next issue is that the range of the variables as seen in Table I is too large, each variable covering different ranges. Therefore, we need to get all the data on a similar scale to avoid having any bias using the learning algorithms.

Standardization:

As can be seen in the plot below, all the 17 variables have been standardized using the zscore function that ensures that each variable has is centered to have mean 0 and scaled to have standard deviation 1. The predictand Y and a few variables have only been plotted for visualization purposes.

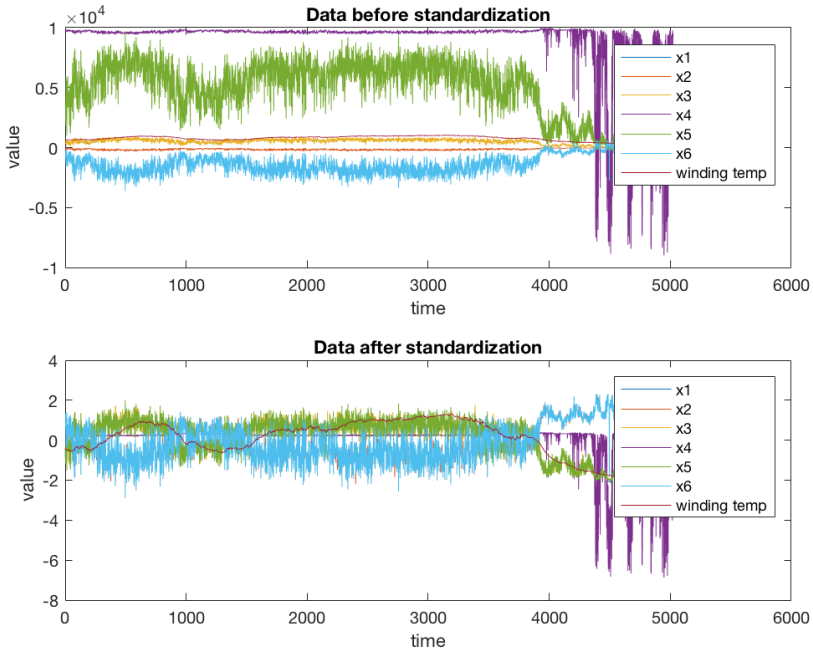

Figure 2: Data Standardization

Dimensionality Reduction using PC Analysis:

We explored the possibility of having fewer modes represent the entire set of variables to a good accuracy. This was done using PC analysis and obtaining the Eigen vectors and PC scores of all 16 variables. By finding how much variance is explained by each PCA mode as a percentage of the total variance, we can find which modes that, when summed up in terms of variance explained, will results in a good total variance explained, and will be sufficient to represent our data set. Figure 3 below shows that the first mode explains about 79% of the total variance, and when combined with second mode explaining 11% of the total variance, we get 90% of the data set variance explained, which is good enough for our purposes. Therefore, the first 2 PC modes are selected to be used in all further analysis.

Table II: Eigen values of all 16 predictor variables

Eigen Values															
λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}
12.413	1.679	0.752	0.525	0.197	0.118	0.097	0.059	0.038	0.035	0.024	0.020	0.017	0.015	0.008	0.003

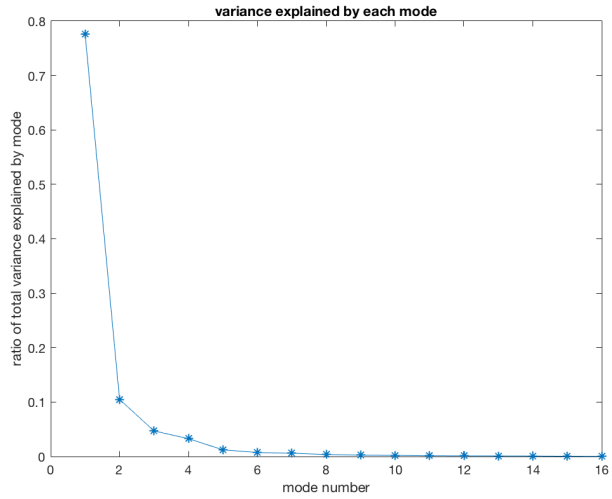

Figure 3: modes with their explained variances

Figure 4 below shows which variables or features each mode captures and the PC strength over time. For PC1, the eigenvector captures strongly the variables x1, x3, x5 and x7-13 which are average power, power, active power and all the different current and voltage readings. This makes sense since all these values are related to each other with equations. As for the second mode, the eigenvector captures the remaining variables, which are different temperature readings (nacelle, gearbox and outside temp). As for the strength of PC1, we can see that these features are most prominent during the first 3500 time samples. While for PC2 it increases and reaches a peak at around 3500 samples and decreases onwards.

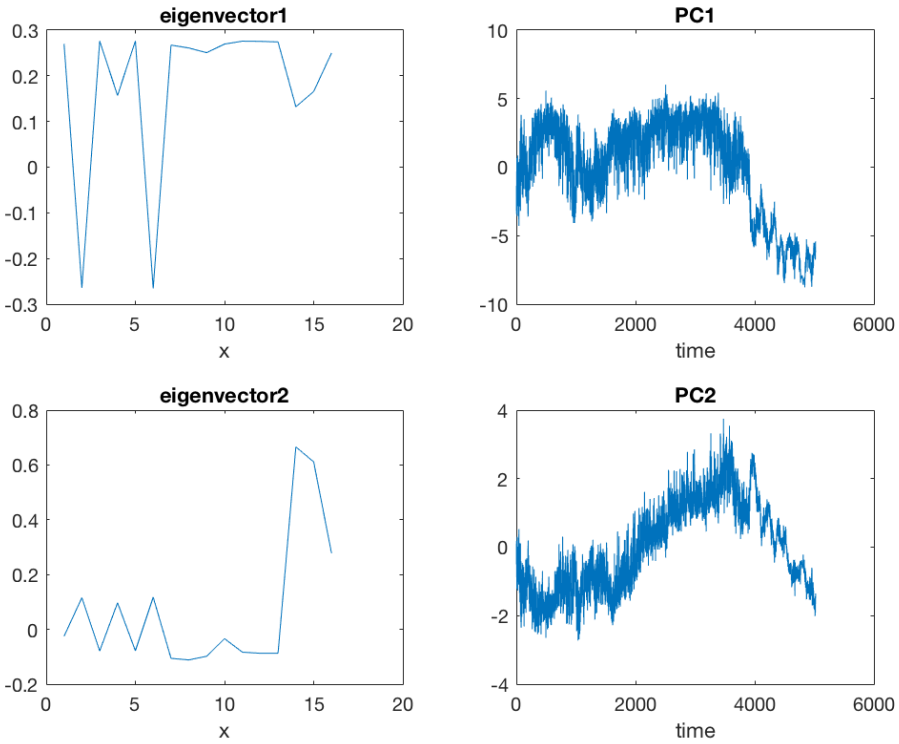

Figure 4: Strengths of eigenvector modes against time

B. Analysis methods

The first analysis will be made using a neural network with the first 2 PC modes selected as inputs, and the generator winding temp as the predictand to be evaluated. We first run an autocorrelation test on the inputs and output. Figure 5 below shows how the predictand Y has complete correlation of its samples, and crosses the confidence intervals by far. Similar results were found for the 2 input variables. Therefore, we should be careful with using some statistical methods that may assume that the analyzed data is independent and identically distributed (i.i.d.). Therefore, instead of using the classical Feedforward Neural Network, instead we will use the Nonlinear Autoregressive Neural Network (NARX) which is specifically designed for time series data prediction.

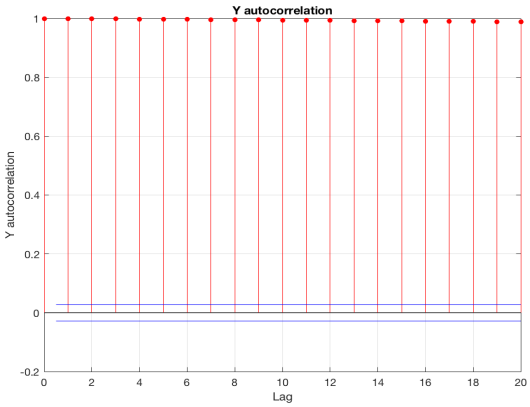

Figure 5: Autocorrelation test on predictor

Nonlinear Autoregressive Neural Network (NARX)

We have used a two-layer feedforward NARX, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer and configured as below:

```
net = narxnet(inputDelays,feedbackDelays,hiddenNeurons);
[inputs,inputStates,layerStates,targets] = preparets(net,X),{},Y));
[net,tr] = train(net,inputs,targets,inputStates,layerStates);
```

The data samples were divided into 3 sets, a training (70%), validating (15%) and testing (15%) sets as below:

```
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio   = 15/100;
net.divideParam.testRatio  = 15/100;
```

The network has been tested with different number of hidden neurons, epochs and predictor lags, and the performance of each case has been evaluated in terms of root mean square error (RMSE) and correlation coefficient rho.

Stepwise Regression

In order to have a comparative analysis of our neural network model, we modeled our data using stepwise regression. Since this is a linear model, our assumption is that it will not be the best for this time series prediction task. The data was again divided into training and validating sets, with 4000 samples used for the training and the rest for validating. The results were evaluated using RMSE and correlation coefficients and compared with the NARX network model.

C. Results

Nonlinear Autoregressive Neural Network (NARX)

After setting up the network model, we altered the number of hidden neurons and plotted the results of the target vs predictand values over the entire set as can be seen in Figure 6 . Table III below shows that having 11 hidden neurons creates the best performing setting in terms of low RMSE value. Therefore, 11 neurons were selected for the hidden layer.

Table III: RMSE for each configuration of hidden neurons

No of hidden neurons	RMSE
1	0.0961
2	0.0897
3	0.0947
4	0.0897

5	0.0832
6	0.1051
7	0.0832
8	0.0898
9	0.0800
10	0.0790
11	0.0654
12	0.0785

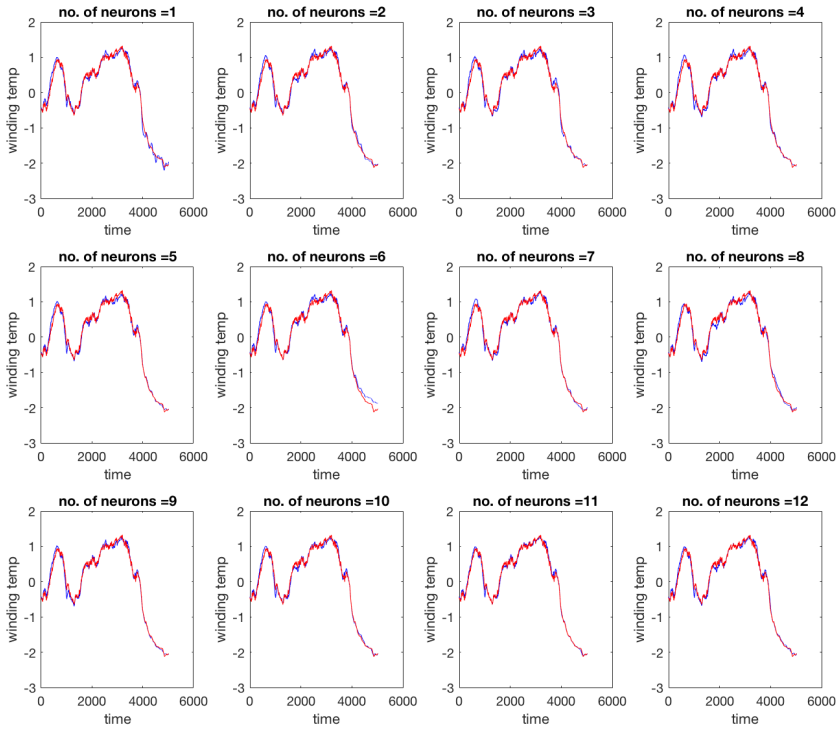

Figure 6: Testing NN with different number of hidden neurons

Figure 7 below shows the final predictand value in comparison with the actual target data. It can be seen that the two set of values are extremely close, with a RMSE of 0.0779. The number of epochs were altered and it was seen that the error was stable beyond 50 epochs, therefore, selecting a very large number would be undesirable due to overfitting of the data, as well as additional CPU processing needed.

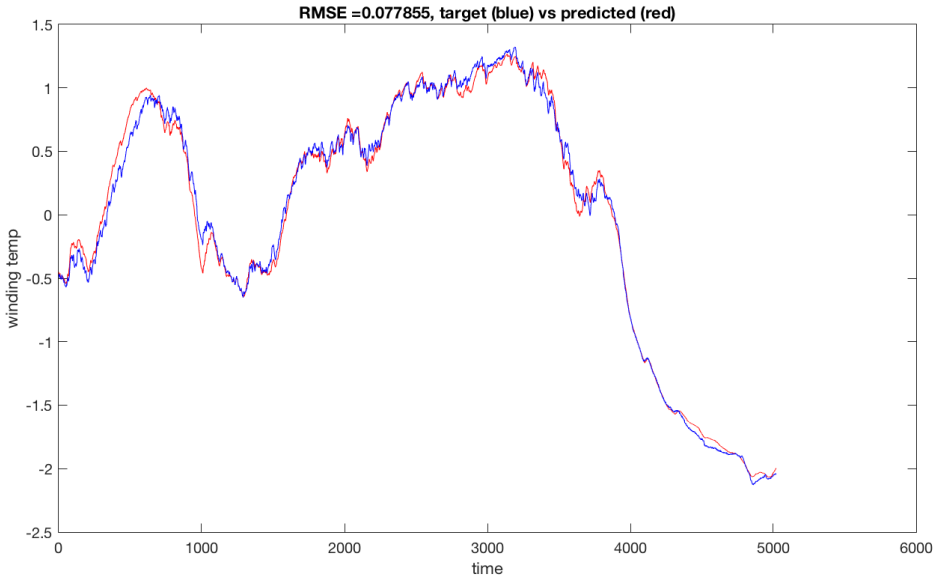

Figure 7: Target values vs predicted values of winding temp using NARX model

Figure 8 below shows the performance of each of the training, testing an validation sets in terms of the mean square error (MSE). The plot captured below accurately represents the winding temperature plot of the turbine generator.

Figure 8: Response error analysis

Next, we tested the model for different lags in the predictand variable to perform future prediction Y(t+n) and in Figure 9 below we show how the RMSE increases with an increase in the number of the lags, as expected. However, this increase is not very large and the model can still be considered accurate.

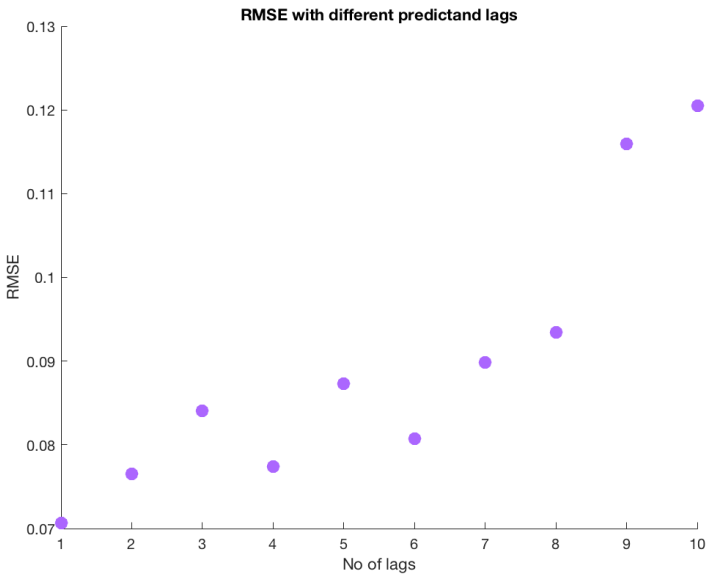

Figure 9: RMSE for different predictand lags

Finally, in order to visualize the accuracy of the model, we plotted a correlation plot as can be seen in Figure 10 below which shows how most of the points lie on the correlation line. The calculated correlation coefficient was r = 0.995, which indicates a high degree of correlation between the predicted and target data.

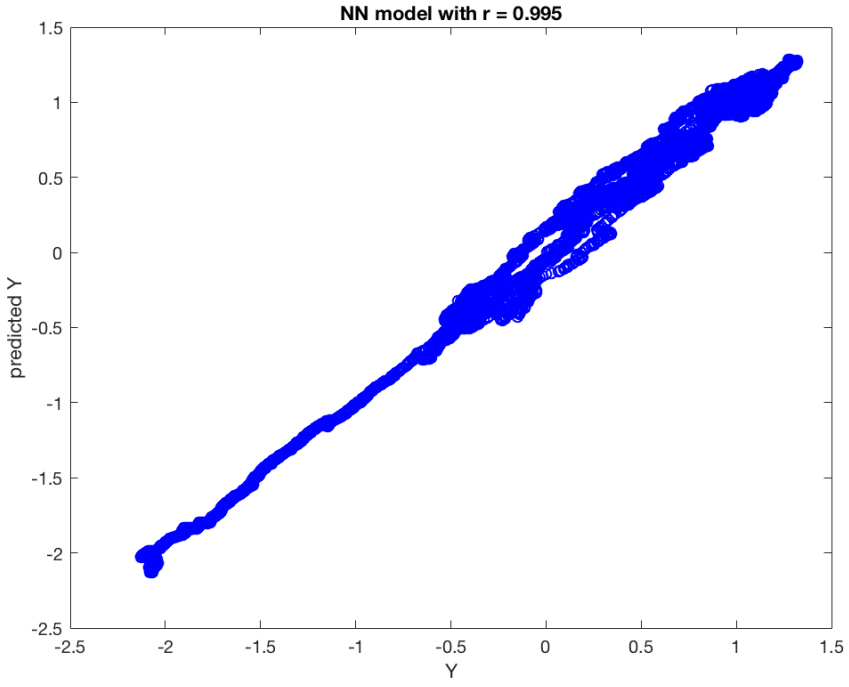

Figure 10: Correlation plot for NARX model

Stepwise Regression

In Figure 11 below, the outcome of the stepwise regression model is shown, compared to the actual target data. It can be seen that the model has a rough estimate of the trend in the predictand, however, with a low degree of accuracy and a lot of noise. It performs particularly badly on sudden increase or decrease in the data trend. This is reflected with the calculated RMSE of 0.408 and the correlation plot in Figure 12. The plot shows how the data points are scattered above and below the correlation line. The calculated correlation coefficient was r = 0.870, which indicates a lower degree of correlation than the one found for the Neural Network model.

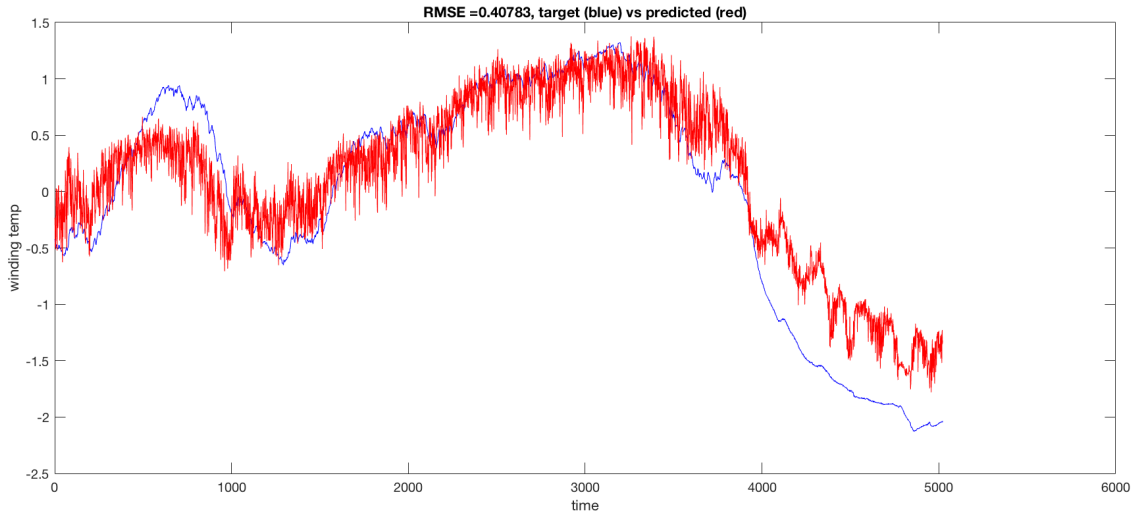

Figure 11: Target values vs predicted values of winding temp using stepwise regression

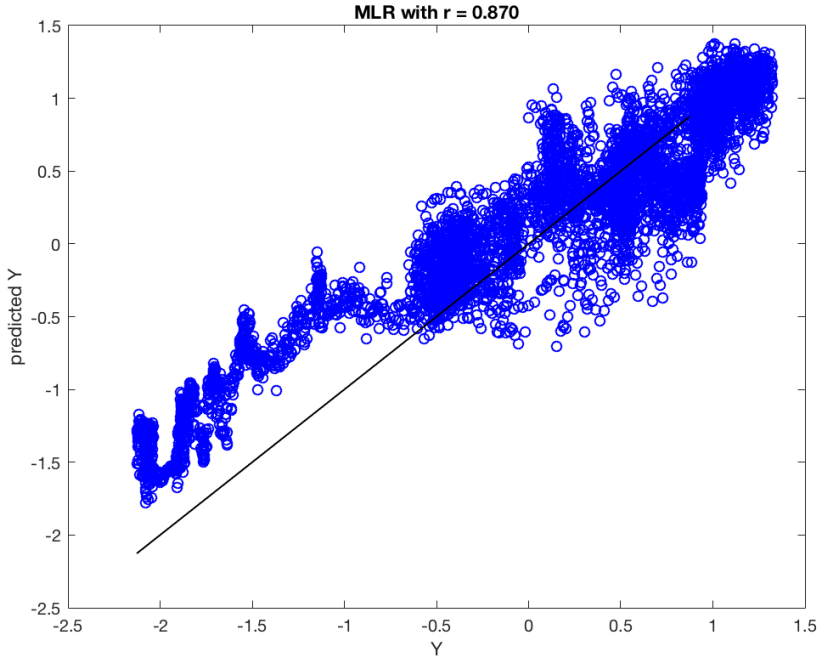

Figure 12: correlation plot of regression results

IV. Discussion & Conclusions

Table IV performance evaluation of the two models

	RMSE	Correlation coefficient rho
NARX Network	0.0779	0.995
Stepwise regression	0.407	0.870

Table IV above shows the final results of the 2 models used to predict the generator winding temp in the wind turbine using 2 PC modes. As can be seen, the NARX network model outperforms the stepwise regression model by having a lower RMSE of 0.0779 compared to 0.407 and a higher correlation of 0.995 between the target and the predicted data compared to 0.870 for the regression. Also, it was seen from Figure 11 that the regression model has a lot of noise and variations, and is not very stable, as compared to the predicted Y in the neural network model seen in Figure 7, which has a very similar shape to the target Y values.

We conclude by selecting the NARX neural network model due to its higher performance and robustness and select 11 hidden neurons, which has the lowest RMSE. It was also observed that there was a trend of increasing RMSE values with an increase in the predictand lag as shown in Figure 9.

The project's object has been achieved as we have created a machine learning model that can predict the winding temperature of the generator to an accurate degree, and this will be further used as part of a mathematical model to dynamically shift the temperature cutoff threshold in order to increase the economic benefits of the turbine to its maximum.

V. Outlook

For future work, we want to test our model with data from different wind turbines located in different countries and see whether the model is standard for all similar turbines or not. Also, instead of using PC analysis, we would like to carry out different feature selection methods such as wrappers and filter searches. Knowing exactly what influences the winding temperature will allow us to see whether it is possible to forecast those features in order to predict the next day's winding temp trends. Finally, we would like to develop and evaluate a deep neural network model, since we have large amounts of training data available, and compare its performance to our current results.

References

[1]"How Do Wind Turbines Work? | Department of Energy", Energy.gov, 2017. [Online]. Available: https://energy.gov/eere/wind/how-do-wind-turbines-work. [Accessed: 17- Dec- 2017].

[2] L. Bauer, "Endurance E-3120 - 50,00 kW - Wind turbine", *En.wind-turbine-models.com*, 2017. [Online]. Available: https://en.wind-turbine-models.com/turbines/372-endurance-e-3120#datasheet. [Accessed: 16- Dec- 2017].