EVALUATING WIND TURBINE PERFORMANCE USING POWER CURVE MODELLING

Andreo II Lozada

Problem Description

Wind turbine power curves describe the non-linear relationship between wind speed and the turbine power output. As part of a turbine's specifications, manufacturers provide wind power curves with theoretical values which can be used as a benchmark to evaluate a wind turbine's generating performance. However, wind turbines normally experience several stoppages in its daily operations due to various natural and technical factors, and sometimes have their outputs curtailed due to grid limitations. These instances result to abnormally low power outputs, which were not considered in the theoretical power curve, making it incomparable with the actual generation as it is.

Project Goal

This project aims to model a wind turbine's performance under normal operating conditions to be compared to the theoretical power curve without bias, and quantify the difference in performance, which may be further used to estimate revenue losses due to performance which may help in making business decisions.

Dataset

The data used in this project was the *Wind Power Generated Data*, uploaded by Bhavik Jikadara in Kaggle in May 2023. This data is a year's worth of a wind turbine's Supervisory Control and Data Acquisition (SCADA) system readings. It has 5 columns and more than 50,000 rows of data including date/time in 10-minute intervals, the turbine power generation readings, wind speed and direction readings at the hub height of the turbine, and the theoretical power values from the manufacturer's power curve. The data can be accessed via this link: https://www.kaggle.com/datasets/bhavikjikadara/wind-power-generated-data

Methodology

The data was first cleaned. Rows with missing data were just dropped since replacing them with the column average does not reflect the relationship between power and wind speed. There was also a high number of zero Power values. These were also dropped as these are outliers that will negatively impact the model.

One crucial part of this project was the outlier detection and removal. These are mostly the low power values during the turbine's ramp down/up before/after stoppages, and/or grid curtailment. Outliers were statistically identified as those outside *Q1-1.5IQR* and *Q3-1.5IQR*, where *Q1* and *Q3* are the 1st and 3rd quartiles and *IQR* is the *Interquartile Range* equal to the difference of *Q3* and *Q1*. However, this cannot be done to the entire Power data because the values are expected to range from zero to maximum power.

Evaluation of quartiles should have a condition, being the wind speed value. Quantile regression was one option, but the non-linear relationship of power and wind speed makes this challenging. The simpler approach done was grouping the data into wind speed bins with a width of 0.1 m/s. Quartiles were then evaluated from the power values within the bin to calculate the power upper and lower bounds of Power for each bin, which were used to filter out the outliers. The results are visualized in *Figure 1*.

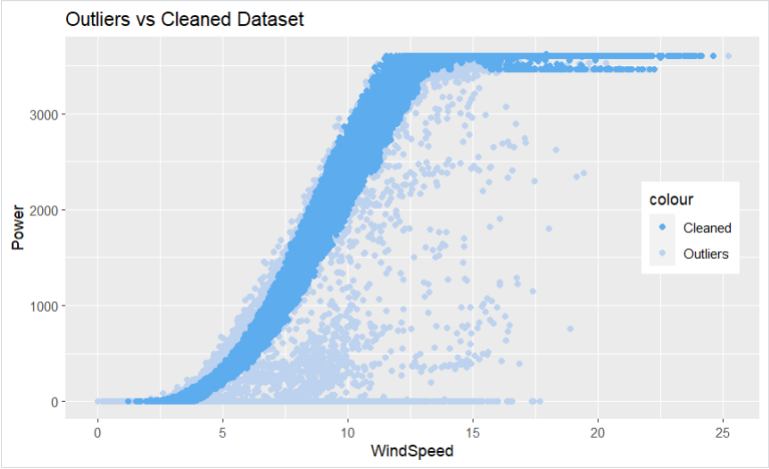

Figure 1. Outliers in Power vs Wind Speed

After cleaning, the data was prepared further for modelling. The data was split into two sets of 80% training and 20% test data. The first set of training and test sets only have WindSpeed as the single variable since wind speed has a very high correlation with power while wind direction has very low correlation. Still, it is worth trying if more variables would further improve the model, so the second set of training and test sets have both WindSpeed and WindDirection. Having multiple variables, only the second set was normalized.

Before modelling, different transformations were tested out on the target variable to try to make the relationship with the independent variables more linear so that simpler linear regression can be done. These transformations include polynomial, logarithmic, and the closest ones being the logistic function and the inverse hyperbolic tangent of the scaled Power expressed as a polynomial, which are inverses of a *sigmoid function*. However, the tested transformations were not able to make relationship linear.

Trying a non-linear approach, *Non-Linear Regression* was considered, but requires a formula approximating the relationship. Observing the Power vs WindSpeed scatter plot and trying out the inverse of the tested transformations by trial-and-error plotting using a graphing calculator, the closest formula representing the relationship was identified as *y ~ max_Power * tanh((a * WindSpeed + b)^3)*, where *a* and *b* are the unknown coefficients. Other regressors capable of non-linear modelling were also considered.

6 different regression methods were used. First was the *Non-Linear Regression* on the training set having only WindSpeed using the mentioned formula. Being based on the wind power curve, this was only done on the single-variable set. This was followed by *Decision Tree*, *Random Forest*, *Generalized Boosted Regression Model (GBM)*, *XGBoost*, and *LightGBM*, each done on both the single-variable and the multi-variable training sets, for a total of 11 models. For each model, the R^2 and the *Root Mean Squared Error (RMSE)* on both the training and test sets were evaluated. The results are summarized in *Figure 2*.

	Model	R^2_{Train}	R^2_{Test}	$RMSE_{Train}$	$RMSE_{Test}$
1	Non-Linear Regression (1 var)	0.9920	0.9922	112.67	111.31
2	Decision Tree (1 var)	0.9604	0.9605	250.66	250.72
3	Decision Tree (2 var)	0.9604	0.9605	250.66	250.72
4	Random Forest (1 var)	0.9928	0.9928	107.15	107.37
5	Random Forest (2 var)	0.9904	0.9902	123.34	124.64
6	Generalized Boosted Reg (1 var)	0.9922	0.9923	111.54	110.88
7	Generalized Boosted Reg (2 var)	0.9924	0.9924	110.13	109.97
8	XGBoost (1 var)	0.9928	0.9928	107.21	107.34
9	XGBoost (2 var)	0.9924	0.9924	110.13	109.97
10	LightGBM (1 var)	0.9926	0.9927	108.19	107.43
11	LightGBM (2 var)	0.9944	0.9940	94.47	97.86

Figure 2. Regression Models Accuracy

The results were generally very good, with high R^2 values and low RMSE values for all models, although these may have been due to the outlier removal which may lead to overfitting if done too aggressively. But this was done deliberately in this case in accordance with the goal of the project.

The best model was chosen based on having the lowest RMSE in the test set. For this dataset, the *LightGBM Model with 2 Variables* was the best model. The selected model was then refitted, and Power was predicted using the data for the entire year to quantify a representative generation under normal operating conditions comparable to the theoretical power from the manufacturer's power curve. The percentage difference between the predicted and the theoretical total generation was finally calculated as *-0.0725*.

Findings

After trying different regression methods for both single-variable and multi-variable datasets, it was observed that the differences in using one vs multiple variables for the same models only had minimal differences in accuracy, though this may be due to WindSpeed having very high correlation with Power and WindDirection having low correlation. Also, it cannot be concluded that having more variables would result to more accurate models. Based on the results, GBM and LightGBM had higher accuracy with multiple variables. On the other hand, Random Forest and XGBoost got worse with multiple variables. For the Decision Tree, the results are just the same, though again, this may be due to the correlation of the variables with the target. The Non-Linear Regression was also quite accurate, although the gradient boosting regressors did better, possibly because the approximate non-linear equation used for the Non-Linear Regression model may not be the best fit to the actual data.

Among the models, the Decision Tree models were observably the least accurate. The resulting tree only had 5 distinct values for the predictions instead of something closer to a continuous data. With this dataset and for this purpose, the Decision Tree is not an effective model, and seemed to be more appropriate for classification rather than regression. As for Random Forest, the complexity of the model was controlled by parameter tuning because fitting the model with default parameters resulted to overfitting instead of yielding something more of a representative curve of the training data. The gradient boosting models are very customizable but got too complex due to having a significant number of parameters. But fitting them with just the default parameters seemed to work well in this dataset.

Summary and Conclusion

A year's worth of wind turbine data including power, wind speed, and wind direction was used to model the turbine power output representing its performance under normal operating conditions to be compared to the theoretical power based on the manufacturer's power curve. Outliers were statistically identified using quartiles, then removed. 80% of the data was used to train models using 11 non-linear regression methods. Using the model with the least RMSE in the remaining test dataset, the annual generation of the turbine under normal operating conditions was quantified and compared to the theoretical generation. In the end, it was concluded that *the estimated actual generation under normal operating conditions is 7.25% lower than the expected generation based on the theoretical power curve*.

This figure can be used to help make business decisions. Multiplying it by the planned or target annual energy generation and by the applicable tariff will quantify a year's estimated lost revenue attributable to underperformance. This will help the company decide whether to invest in further in investigating the performance of the turbine and applying improvement actions or equipment, and by how much should they spend to still have reasonable returns. Of course, other factors affecting the generation such as turbine stoppages, generally represented by the removed outliers, is completely a different story and is out of the scope of this project, but is something that is also interesting to look into.

References

Donnelly, R. A., Jr. (2020). Business Statistics: A First Course (3rd ed.). Pearson.
Cole, S. (n.d.). Wind Turbine Power Curve. Retrieved from https://theroundup.org/wind-turbine-power-curve/
Graphing Calculator: https://www.desmos.com/calculator
R Documentation:
https://www.rdocumentation.org/packages/gbm/versions/2.1.8.1/topics/gbm
https://xgboost.readthedocs.io/en/stable/R-package/xgboostPresentation.html
https://lightgbm.readthedocs.io/en/latest/R/index.html