Developing non-weather-based AI models for wind power production forecasting using past power production data: a more cost-effective, accurate, and practical alternative to traditional weather-based models

Khondoker Farihah Ahmed

American International School of Dhaka

Harvard Impact and Research Fellowship

Kayla Janae Saucedo

April 10, 2022

Abstract

Efficient methods of wind power production forecasting are crucial for the integration of such renewable forms of energy into the electrical grid. Currently, three main criteria stand for the deployment of efficient artificial intelligence (AI) and machine learning (ML) algorithms into the field for such types of forecasting: accuracy, cost-effectiveness, and practical data obtainment. Though many state-of-the-art forecasting models in the field are highly accurate, most use weather data inputs such as wind speed and direction. However, cost-effectiveness and practical data obtainment are not yet widely considered, despite both being crucial obstacles to the actual deployment of such forecasting models. There are numerous issues with weather data which make it costly and impractical. Thus, this study proposes an alternate data source: past power production data, to predict future power production. 10 minutes of past power production data are used to predict the next 10 minutes of future power production. The freely available 'Wind Turbine Scada Dataset' is used from Kaggle.com, and results of this study are compared with state-of-the-art weather-based models from Kaggle's top-voted submission. The coefficient of determination (r^2) is used as the metric to evaluate accuracy between the two studies. All of the models of this study outperformed the state-of-the-art Kaggle submission, with the highest r^2 score of 0.97 with a simple neural network. Thus, by checking all three criteria of high accuracy, cost-effectiveness, and practicality of data obtainment, this study proposes a novel, highly-efficient for wind power production forecasting, which can be deployed in wind farms quite simply to further accelerate the integration of such renewable forms of energy into the electrical grid.

Keywords: Wind energy, power production forecasting, artificial intelligence, machine learning, practicality, cost-effectiveness

Developing non-weather based AI models for wind power production forecasting using past power production data: a more cost-effective, accurate, and practical alternative to traditional weather-based models

Background

Being able to forecast power generation for any energy system is crucial to its smooth operation, and for the integration of renewable forms of energy such as wind energy into the electrical grid, efficient forecasting methods are of paramount importance (Saroha et al., 2021). In 2020 alone, globally onshore wind electricity generation increased by 11%, and offshore wind generation increased by 29% (International Energy Agency, 2021). However, like other forms of renewable energy (such as solar energy), wind power generation can be volatile as it is dependent on weather conditions. As the supply of wind farms continues to grow at an accelerating pace globally, a crucial demand also stands for highly efficient forecasting methods for wind power generation (Saroha et al., 2021).

As machine learning (ML) and artificial intelligence (AI) continue to be applied to this field, three key evaluation criteria stand to judge the practical implementation of such forecasting models in current energy systems: the accuracy of the forecast, the cost-effectiveness of the forecast, and the practical obtainment of the type of data required. All of these factors are equally important for the practical deployment of such models (Saroha et al., 2021).

Currently, state-of-the-art AI models for wind power forecasting are mostly reliant on input data such as wind speed, wind directions, and sometimes other weather/meteorological data (Music et al., 2018). However, there are numerous issues with obtaining these kinds of weather data from external companies. For example, a study by Ordiano et al. (2016) focusing on solar

energy highlights the issues with obtaining weather data on a daily basis to run these models: first, purchasing this data from weather data companies creates high, additional costs for the maintenance of such models, decreasing profitability; second, constant communication with weather data services are needed to run such models, however, in the case of communication failures, these weather-based AI forecasting models come to a halt and are unable to continue operating (Ordiano et al., 2016).

Although the aforementioned study focuses on solar power production forecasting, the same type of data (weather data) is used for wind power production forecasting. Thus, the same issues ultimately apply to wind power production forecasting as well. Although most wind farms may use anemometers (a measuring tool which measures wind speed and wind pressure), the issue is that they can only take measurements in real time (National Geographic Society, 2011). Thus, for crucial forecasts (predictions ahead of time), assistance from weather data services can be highly important, and the same issues as with solar power production forecasting as mentioned by Ordiano et al. (2016) remain.

While current state-of-the-art wind power production forecasting models in the field may fit only one of the three required criteria (high accuracy), the cost effectiveness and practical obtainment of the input data are still essential issues which are yet to be resolved in the field. However, an alternative source of input data, which has the potential to fit all three of these criteria, is the past power production data of a wind farm. The past power production data is a wind farm's own, readily-available, free data--solving the cost effectiveness and practical obtainment of data issue. In regards to the accuracy of this type of input data in models, it is important to note that the most recent past power production data of a wind farm will likely be reflective of the most recent weather trends as well (recent past power production and most

recent weather trends have a tendency to be highly correlated data). Expensive, difficult to obtain pieces of weather data such as forecasted future wind speed and future wind direction do not need to be obtained to predict future wind power production, if two simple inputs: time and past wind power production, can be used to predict future wind power production. Furthermore, it is important to note that even though forecasted wind speed and direction may be provided for free by weather services sometimes, the difficulty is that such forecasts are usually for larger areas (such as an entire city's forecasted wind speeds and direction), so relying on this free data about wind speed and direction would not work as it would likely not give an accurate measure of the forecasted wind speed and direction in the wind farm's exact, specific location. These types of forecasts (specific to a particular location or area) is what would require the extra expenses from a weather data company, and thus would be difficult to obtain.

However, to give an example, the wind power production data of 10 minutes ago can likely be used to predict the power production of now, as it is highly unlikely that weather conditions would change drastically within a span of these 10 minutes. Also, again, the power production data is highly correlated with the weather trends and weather data itself, which is why there is a high likelihood of getting an accurate forecast of future power production using 10 minutes of past power production data. Plus, wind farms would already have the data of power production 10 minutes ago, so this data would be readily available and free. This would eliminate the need to go to external services to collect this data, so checking the other two difficult criteria of cost effectiveness and practical obtainment of data. Also, though an argument can be made that wind speeds and directions can drastically change within the span of 10 minutes, causing the future prediction from this data to be inaccurate, it is still worth nothing that such volatile changes would only be a small fraction of the predictions made in an entire day,

rendering such a prediction algorithm still powerful and highly accurate in predicting future power prediction for the rest of the day.

The goal of this study is to examine ways to use alternative sources of data to check all three criteria. Developing AI non-weather-based models which are just as accurate as state-of-the-art weather-based ones by using past power production data (an alternative source of data) would help to take a giant step towards making wind power forecasting systems highly efficient, and ultimately it would check all three criteria. We hypothesized that if we use the free and readily-available past power production data of a wind farm to predict future power production, then the accuracy of this ML model will be comparable to current state-of-the-art weather-based models.

Materials and Methods

Data

The data that will be used in this study is the freely available 'Wind Turbine Scada Dataset' from Kaggle.com. Scada systems measure data including wind speed, wind direction, and power generation in 10-minute intervals. This particular dataset is from a turbine in Turkey.

As this data was taken from Kaggle.com, the results of this study will be compared with the ones of the top-voted code submission from Kaggle. This will be done so that state-of-the-art weather-based models from Kaggle can be compared with the non-weather-based models of this study (Erisen, 2019).

Evaluation Metrics

The top-voted code on Kaggle was from Chittal Patel, with 12,943 views and 72 upvotes

(as of April 10th, 2022), entitled: “Wind Turbine ✶ Power Analysis”. The primary evaluation metric used in the code was the coefficient of determination, or R2 (Patel, 2020). Thus, this study will also use the same evaluation metric to assess the accuracy of its models. The following equation describes the R2 metric:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\Sigma(\hat{y}_i - \underline{y})^2}{\Sigma(y_i - \underline{y})^2}$$

where SSR is the sum of squared regression, SST is the sum of squares total (or total sum of squares), $\hat{y}_i$ is the predicted value or a point on the regression line, $\underline{y}$ is the mean of all values, and y_i represents the actual values or points.

Data Preparation

Initially, the Wind Turbine Scada Dataset downloaded from Kaggle has 5 inputs: (i) Date/Time in 10-minute intervals, (ii) LV active power (kW) or the amount of power generated by the turbine in those 10 minutes, (iii) wind speed (m/s), (iv) the theoretical power curve (KWh) or the theoretical power values that the turbine generates with that wind speed which is given by the turbine manufacturer, and (v) wind direction (º).

However, in preparing our data for this study, we changed the format of the data. We only kept the LV active power (kW) column, and eliminated the rest, as power production data is both in our input and output.

We then restructured the data in the following way. First, we made two copies of the dataset. We named the first one the input dataset, and the second one the output dataset. Second, we deleted the last row of the input dataset, and the first row of the output dataset. This is

because if we align their indexes now side by side, the input dataset has essentially 10-minute intervals of current active power production, and it predicts in the output dataset the next 10 minutes of active power production. So, the past 10 minutes of power production predicts the next 10 minutes, without the assistance of any meteorological or time inputs such as time of day, wind speed, wind direction, etc. This eliminates the need to purchase forecasted wind speed and forecasted wind direction data from weather data companies, thus making it a potentially cost-effective solution and also a practical one for data obtainment, since past power production data is readily available to all wind farms themselves.

Machine Learning Algorithms

Before being fed into the ML models, the data was split into a training and test set with a 80:20 ratio respectively. The data was not shuffled, as the time order of the data is crucial for this study.

5 simple traditional models with default configurations and 7 simple neural networks are tested. The 5 simple traditional models are (i) Linear Regression (LR), (ii) K-Nearest Neighbours (KNN), (iii) Decision Tree Regressor (DT), (iv) Multi-Layer Perceptron Regressor (MLP) and (v) Random Forest (RF). For the 7 neural networks tested, the architecture of each one is described in Table 1. The ReLu (maximum) activation function is used for all the neural networks, so that the models can learn non-linear relationships in a fast and efficient way. All the models were compiled with a mean squared error (MSE) loss function.

Table 1 - Neural Networks Model Architecture

Model	**Layers**	**Number of units each layer**	**Number of epochs**	**Activation function used**	**Loss function used to compile model**
NN1	3 dense layers	1 (input) - 500 - 8 - 1 (output)	10	ReLu for all layers	MSE
NN2	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	10	ReLu for all layers	MSE
NN3	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	15	ReLu for all layers	MSE
NN4	2 dense layers	1 (input) - 100 - 1 (output)	10	ReLu for all layers	MSE
NN5	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	20	ReLu for all layers	MSE
NN6	3 dense layers	1 (input) - 100 - 1000 - 1 (output)	30	ReLu for all layers	MSE
NN7	3 dense layers	1 (input) - 1000 - 8 - 1 (output)	30	ReLu for all layers	MSE

Results

Table 2 shows the results in the evaluation metric, r^2, for all the models. Note that r^2 is from a scale of 0 to 1, where 1 indicates that the model perfectly fits the data, and 0 indicates that the model is not better than simply using the mean to predict the value. Our goal is to try to get as close to 1 as possible, without having potential overfitting of the model.

Table 2 - Evaluation Results for All Models

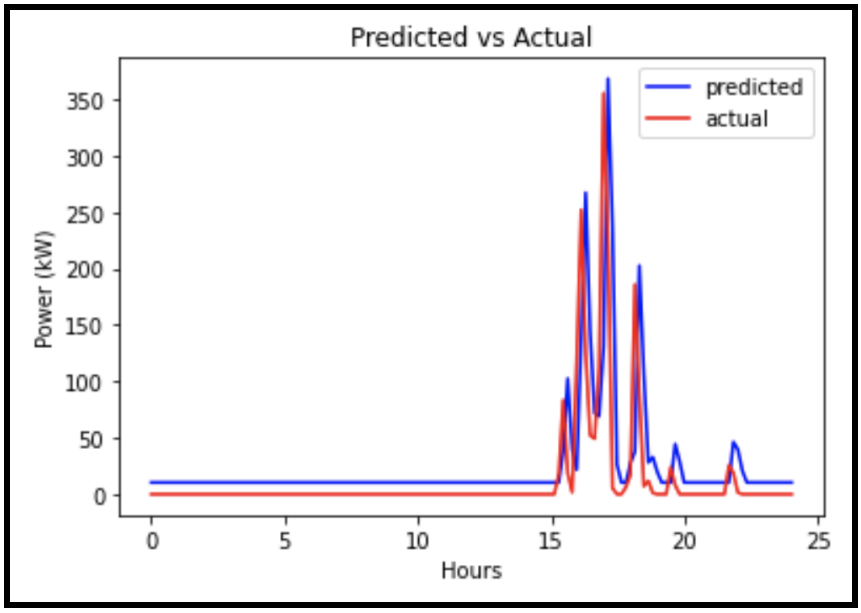
Model	**r^2**
LR	0.9712261659
KNN	0.966924416
DT	0.9670533112
MLP	0.971190487
RF	0.9546141671
NN1	0.9712483431
NN2	0.9711454572
NN3	0.9701392095
NN4	0.9711417606
NN5	0.9712104206
NN6	0.9706165255

NN7	0.9712058231
Average score	0.9689763406
Best model	0.9712483431 (NN1)
Worst model	0.9546141671 (RF)

The average r^2 of all the models was 0.9689763406, while the best model was NN1 with an r^2 of 0.9712483431 and the worst model was RF with an r2 of 0.9546141671. To compare with our baseline top-voted code on Kaggle from Chittal Patel, "Wind Turbine ✶ Power Analysis", the reported r2 scores for that submission were: 0.882314774332115, 0.910832474267455, 0.3821888069845253, and -0.05124282001762803. All of our models outperformed the aforementioned baseline code from Kaggle, which took into account all the meteorological weather variables, including wind speed and direction. Comparatively, our study only used the past power production data to predict the future power production data, but even with zero weather data, it outperformed the top Kaggle submission significantly.

Figure 2 shows an example of our best model (NN1), and its prediction with our method. It shows the predicted output vs. actual output by NN1 for 24 hours.

Figure 2: Actual v Predicted Power Output (kW vs 24 Hours)



Discussion

Our hypothesis stated that if we use the cheaper data source of a wind farm's past power production data, then the accuracy will be comparable to current state-of-the-art weather-based models. It was shown to be accurate. All of our models outperformed the baseline Kaggle code's state-of-the-art weather-based models significantly. Thus, this study shows that using past power production data as the sole input to predict future power output checks all three criteria: our method proved to be highly accurate, cost-effective (as this data is freely available to wind farms), and practical for data obtainment (as this data is readily available to wind farms). By being comparable to state-of-the-art weather-based models in accuracy especially (the most important criteria for deployment) this study's method helps take a giant step towards making wind power forecasting systems highly efficient, as it checks all three criteria. The implications of this study are that this forecasting technique can be applied realistically and easily to nearly any wind farm as it is cost-effective and practical. Thus, this study helps us accelerate the

integration of renewable energies to the electrical grid and shift away from fossil fuel usage.

It is recommended that other machine learning techniques, such as data preprocessing, feature engineering, and ensemble methods, are used by future studies to improve the results further. Further research areas also include applying this technique to other forms of renewable energy forecasting with volatile power production. Using other metrics aside from r^2 would also be beneficial in assessing model performance further in future studies.

References

Erisen, B. (2019). *Wind turbine scada dataset*. Kaggle; Kaggle. https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset

International Energy Agency. (2021, November). *Wind Power - Analysis*. International Energy Agency; International Energy Agency. https://www.iea.org/reports/wind-power

Music, E., Halilovic, A., Jusufovic, A., & Kevric, J. (2018). Wind direction and speed prediction using machine learning. *The International Symposium on Computer Science - ISCS*. Research Gate. https://www.researchgate.net/publication/335691207_Wind_Direction_and_Speed_Prediction_using_Machine_Learning

National Geographic Society. (2011, July 28). *Anemometer*. National Geographic Society. https://www.nationalgeographic.org/encyclopedia/anemometer/

Ordiano, J., Waczowicz, S., Reischl, M., Mikut, R., & Hagenmeyer, V. (2016). Photovoltaic power forecasting using simple data-driven models without weather data. *Computer Science - Research and Development*, *32*(1-2), 237–246. https://doi.org/10.1007/s00450-016-0316-5

Patel, C. (2020). *Wind turbine ✴ power analysis*. Kaggle; Kaggle. https://www.kaggle.com/code/chittalpatel/wind-turbine-power-analysis

Saroha, S., Aggarwal, S., & Rana, P. (2021). Wind Power Forecasting. *Forecasting in Mathematics - Recent Advances, New Perspectives and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.94550