Prof. David Gordo Gomez

Predicting the Active Power of a Wind Turbine

Ludovico Gandolfi

1. Description of the Problem

Renewable energies are key to a more sustainable future. Wind power generation, unlike other available renewable energy technologies, offers an environmentally sustainable option[1]. With the proliferation of wind farms, there is an increasing need for accurate forecasting models to estimate the active energy generated by the turbines. In 2011, wind turbines accounted for 3% of electricity generation capacity in the United States, and six US states met more than 10% of their annual electricity demand through wind power. In that year, the rate of installation of new wind turbine generation capacity was second only to natural gas. Some markets have higher penetration of wind energy; for example, wind turbines provided 3.5% of the energy generated in the European Union in 2008 (European Commission).[2]

The main objective of this report is to develop a suitable machine learning model capable of predicting the active power generated by a wind turbine using the dataset provided by the instructor. More specifically, it will attempt to identify the most important variables among the 22 features of the dataset in terms of their influence on active power. The data is labelled and the different models will be trained in a supervised environment and subjected to the same test set to allow comparison. The second objective is to gain insights into which measurements[3] are most important for energy forecasts. It will investigate the tradeoff between the number of sensors and the added explanatory power to the model.

The first chapter is devoted to the analysis of the data set and contains some basic descriptive statistics. It also includes some of the author's assumptions about the missing values and the behaviour of the target variables. The following chapter focuses on the development of the optimal model with details on the split between test and train sets, the hyperparameters and the performance metrics used for the evaluation. Finally, the last chapter compares both model's performance on predicting new values (using the unseen test set) and discusses the results. It also summarises the overall insights derived from the previous section and provides some recommendations as well as an outlook.

2. Description of the Dataset

The dataset provided contains 118,224 records corresponding to a large number of measurements taken by the sensors of a wind turbine at an undisclosed location every 10 minutes, starting at 00:00 on 31 December 2017 and ending at 23:50 on 03 March 2020. The 22 columns provide measurements such as the ambient temperature, the pitch angle of the blades and the wind speed at a given time. The column "WTG" was ignored in the analysis as it only contained an identical value (G01) - presumably an identifier for the respective turbines. The column "ControlBoxTemperature" was discarded for the same reason.

2.1. Data Cleansing

As mentioned in the introduction, the data had to be cleaned as it contained several missing values for the active energy. The turbine seemed to be inactive in random patterns, sometimes multiple times a day for several dozen minutes. These shutdowns

[1] Zayas-Gato et al. (2022)
[2] European Commission (2020)
[3] This dataset includes data from 20 different sensors

could be due to maintenance works, weather conditions, or the operator's decision to shut down the generator in response to lower demand. In fact, the energy production appeared cyclical in the time series representation of the data *(Graph 1 in the appendix)* and likely varied according to demand seasonality. In addition, some sensors only started recording at a certain point in time, as they were probably installed gradually. This can be seen from the cross-tabulation of some measured values depending on the years and quarters, which can be found in the appendix *(Table 1)*.

The rows with missing values in the "Active energy" column (approx. 19.9%) were discarded as they would probably bias the accuracy of the forecasts. This became clear after an initial regression was run with all the data: For each empty row, the regression would predict the intercept. The imputation of values such as mean or mode seemed inappropriate in this situation as it would require external information. The prepared dataset was reduced to 94,750 rows.

2.2. Correlations

Finally, the correlation matrix *(Table 2 in the appendix)* helps identify which variables would be more relevant for the prediction. As expected, some of the variables were strongly correlated, e.g. ambient temperature and main box temperature (0.84) or wind speed and active power (0.94). The latter would imply that if the wind speed increases, active energy increases. This applies to a certain range, i.e. between the minimum wind speed required to generate electricity and the maximum wind speed before the turbine must be shut down to prevent damage. Variables such as the nacelle position or the angle (pitch) of the blades did not show a correlation above 0.7 with any other variable and are therefore probably less suitable to explain the variation in the active energy.

3. The Model

All models discussed in this chapter were created using Dataiku (Data Science Studio). The two sections describe the split between test and training sets and the performance metric used to optimise the hyperparameters.

3.1. Test/Train Split

Due to the heterogeneity of the data set with respect to the starting point of the measurements of the respective sensors, the models were trained on two different versions of the data set. Firstly, the prepared Dataset was split into a test (20%) and a train set (80%) using a linear split (sorted according to the date). The train set was then filtered to create a second version of the train set that only included rows from 2019 onwards (59,351 rows). This should make it possible to train a second model on more complete data (all sensors recorded data after 2019 - see chapter 2.1) and compare the accuracy of the respective forecasts on the same test set. The workflow is schematised in Graph 2 *(see appendix)*.

3.2. Metrics & Hyperparameters

There are different metrics to evaluate the performance of machine learning prediction models. The Mean Absolute Error is the arithmetic average of the absolute errors between the actual and the predicted values and it is easier to interpret than more commonly used metrics such as the R Squared (The lower the MAE, the higher the prediction power of the model; very low values could indicate overfitting). Due to the

size of the data set, the R2 score would be very high anyway and therefore may provide a biassed view of the accuracy of the predictions. Similar studies also measure the performance of their models using the MAE[4]. For reference, the distribution of the values for active power can be viewed in the appendix *(Graph 3)*.

The first training set was used to train all 14 models proposed by Dataiku using only the features that contained data for the whole dataset, namely the ambient temperature, nacelle position, reactive power, wind direction and wind speed (Model 1). The models trained on the second training set (Model 2) included additional variables (mostly temperature sensors and rotor RPM). Note that not all sensor data was included because some columns still had too little values and were rejected by Dataiku. The hyperparameters were automatically optimised for the MAE and evaluated on a validation set (subset of the respective train set with 80/20 ratio). In less technical terms, the software is looking for the optimal tree depths to minimise the error.

4. Results

This final chapter details the assumptions underlying the model selection and discusses the implications from a business perspective.

4.1. Model Selection

In both cases, the optimal model (with the lowest value for the MAE) generated by Dataiku was a Random Forest, a supervised machine learning algorithm that produces an ensemble of decision trees to solve classification or regression tasks. It is generally considered very effective, but can take a long time to train depending on the size of the dataset and the hyperparameters (deps and number of trees). The following table compares the results of the Random Forest with a simple OLS Regression and Decision Trees for both training sets.

Table 1: Metrics for Models trained on Training Set 1 and 2, respectively

Training Set	*Algorithm*	*MAE*	*Top Coefficients (OLS) / Most important Variables (PT & RF)*
1	OLS Regression	149	Windspeed, Reactive Power, Ambient Temperature
	Decision Trees	41.917	Reactive Power, Windspeed
	Random Forest	17.561	Windspeed, Reactive Power
2	OLS Regression	29.431	Reactive Power, Wind Speed & Generator RPM
	Decision Trees	21.731	Reactive Power, Wind Speed
	Random Forest	6.506	Reactive Power, Wind Speed

Source: Dataiku & MS Excel

The OLS regression had the highest MAE, followed by the decision trees. The advantage of OLS regression is that the coefficients are easy to interpret. For example, if the wind speed or the reactive power increases, the active power increases. Conversely, if the ambient temperature increases, the active power decreases. The same applies to the top coefficients of the OLS for the second training set. The difference in coefficients is due to the fact that the second set of models was trained with more features.

The decision trees achieved a lower MAE value, which was only surpassed by the

[4] Zayas-Gato et al. (2022)

Random Forest. The main variables were similar to those found in OLS regression. However, they are not as easily explained and in the case of Random Forest, the model cannot be visualised. One approach to selecting the most appropriate model is to look at what types of algorithms have worked best on similar problems in the past. In their paper about using machine learning to predict the power output of wind turbines, Clifton et al.[5] trained various decision trees to get the optimal result. Since Random Forest follows a similar logic to decision trees and is less likely to overfit the data[6], the authors agreed that the optimal model would be best suited for predicting the turbine's active power. Another argument in favour of Random Forest is the fact that the smaller the MAE, the lower the overall error if the algorithm were to be deployed at scale (assuming the windpark uses the same turbine at scale).

4.2. Model Evaluation

To test the accuracy of the predictions on unseen data, a subset of the original dataset (20%) was split off according to the methodology described in chapter 3.1. Since it consists only of the most recent records, it should allow us to estimate how the models will perform with new data. The aim of this section is also to compare the accuracy of the two models to assess whether the presence of more sensors increases the accuracy of the prediction. The previous section suggests that the model trained on more recent data with more features outperforms the Random Forest trained on all data (with fewer features to avoid imputing too many values for the missing records).

The output for each model was an Excel file with the predictions and an additional column with the difference to the observed values. The MAE was computed by creating an additional column reflecting the absolute differences and taking the average of these values. For Model 1, the MAE increased from 17.561 (when scored against itself as described in chapter 3.2) to 21.759. A similar trend was observed for Model 2 where the MAE increased from 6.509 to 18.022. In relative terms, the MAE has almost tripled in Model 2, while it has only increased by about 20 % in Model 1. Part of this difference might be due to the fact that the first model was trained on a set with more data.

4.3. Implications

All models provided reasonable results in favour of the influence of wind speed and reactive power as the two most important factors in relation to active power. This does not come unexpected, as active power is the true measure of a turbine's efficiency and energy production, and since wind is the natural source of this power, the stronger the wind, the higher the active power and the energy produced in the turbine's generator. Reactive power has proven to be the second most important variable for predicting our outcome. However, to explain why it is so important, we must first define its technical basis.

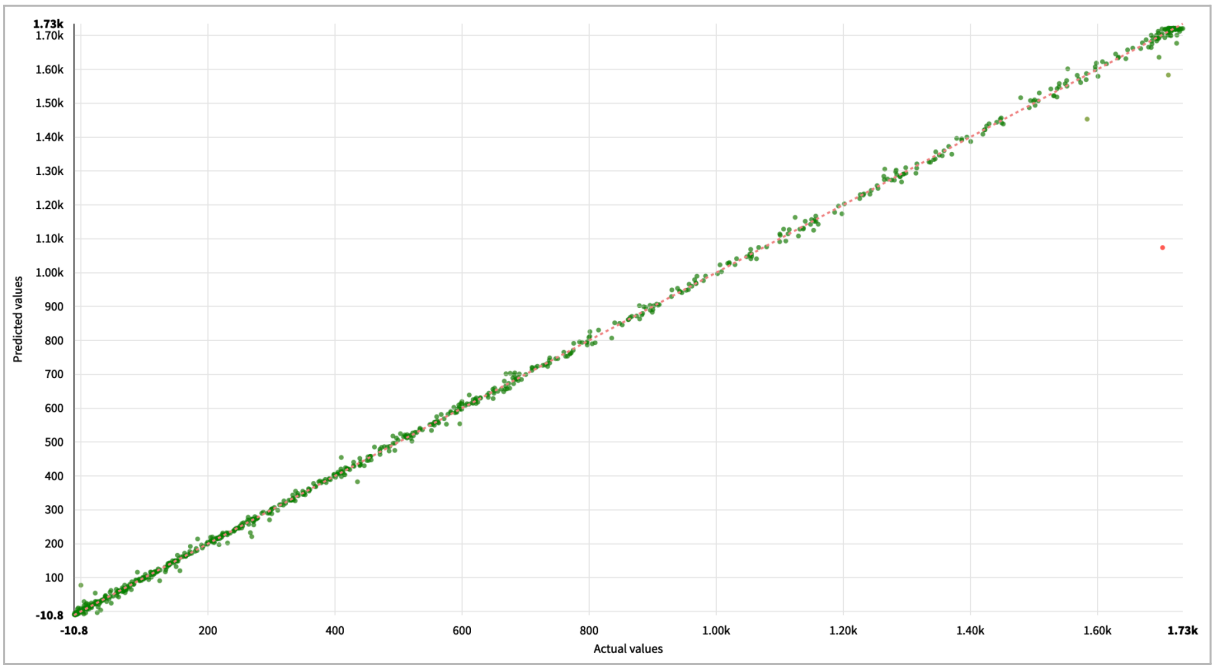
Wind farms usually use long-distance transmission paths to transmit electricity, as the grid cannot fully support the grid voltage of the wind farm. When the wind speed changes, there is the problem of voltage instability at the grid connection point, and the "fluctuating" energy generated in this process is called reactive power.[7] Even if reactive power does not actively contribute to the generation of active power, its peaks technically

[5] Clifton et al. (2013)
[6] Talari (2022)
[7] Tang et al. (2020)

signal a change in wind speed, which, as identified before, contributes to an increase in active power and energy generated.

The two models developed in this report delivered similar results when subjected to unseen data. As expected, the algorithm trained on more features (Model 2) delivered more promising results. The difference in MAE might be small, but as pointed out before small changes could result in larger errors at scale. A screenshot of the optimal model can be found in the appendix *(Graph 4)*. The following scatterplot aims to visualise the accuracy of the predictions for Model 2.

Graph 1: Scatterplot of the Predicted and Actual Values for Model 2



Source: Dataiku

As wind speed and reactive power (which indirectly depends on wind speed) are the most important variables, cooperation with a meteorological observatory could help providers to accurately predict the active power generated by their turbines hours or even days in advance. This could bring two major advantages from a business perspective. With fixed energy supply contracts, for example, the "missing energy" (difference between the amount fixed in the contract and the actual production) must be paid back by the supplier. Therefore, accurate forecasts could allow each supplier to predict their own energy production at any point in time. In this way, they could subsidise the "missing energy flow" from other renewable sources instead of having to pay it back, especially for the periods when they expect lower energy production.

From a practical point of view, knowing with greater accuracy when the periods of low wind speed and low active power occur could allow operators to schedule maintenance work so that it does not overlap with periods of potential active power generation. In fact, this is a major problem for wind farms, as the science suggests: The time it takes to repair (replace) a component in a turbine directly affects the downtime of that turbine or even larger parts of the wind farm. These downtimes lead to production losses, which in turn increase energy costs. Previous research has shown that varying repair times has a

significant impact on production losses.[8]

Further development of the found model in collaboration with some energy engineers and technical experts in general could produce a forecasting model capable of suggesting the optimal energy subsidy and maintenance decisions, taking into account the respective constraints and parameters of the provider, thereby reducing direct and opportunity costs and increasing the overall revenue from the energy generated.

Appendix

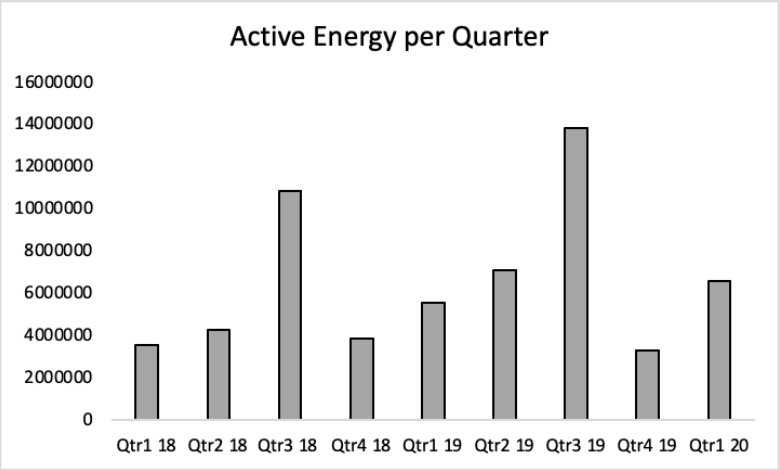
[8] Seyr et al. (2018)

Table 1: Cross Tabulation of Sensor Data against Time

Year	Quarter	Sum of ActivePower	Sum of AmbientTemperatue	Sum of BearingShaftTemperature	Sum of Blade1PitchAngle
2017		-	-	-	-
	1	-	-	-	-
2018		22433237.94	1011527.314	174677.8781	-
	1	3526895.944	224682.8679	-	-
	2	4261497.054	231904.1033	-	-
	3	10835861.24	269568.7639	108210.6922	-
	4	3808983.697	285371.5792	66467.18593	-
2019		29665087.34	1354262.749	1993183.43	306104.8124
	1	5517062.867	338854.0018	492492.9173	-
	2	7089157.059	390500.8065	531121.9741	59660.18931
	3	13804324.76	348517.655	553663.0988	78861.33154
	4	3254542.651	276390.2862	415905.4403	167583.2915
2020		6562328.789	333761.6565	521049.6913	103341.0983
	1	6562328.789	333761.6565	521049.6913	103341.0983
Grand Total		58660654.06	2699551.72	2688911	409445.9107

Source: own calculations

Graph 1: Cross Tabulation of Sensor Data against Time



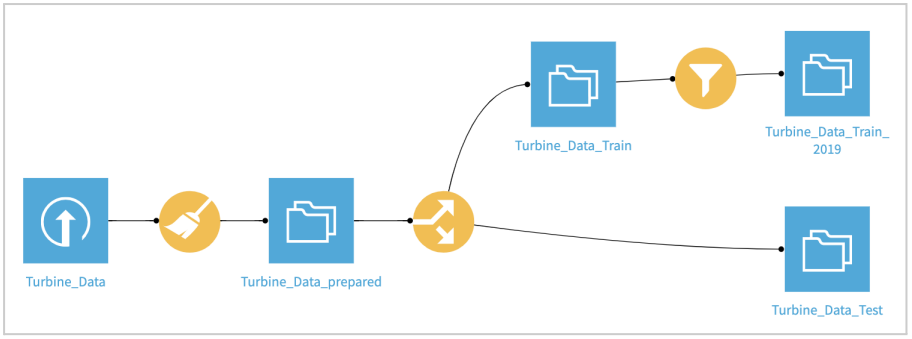
Source: own elaboration

Table 2: Correlation Matrix

	ActivePower	AmbientTemp	BearingShaft Temp	Blade1Pitch Angle	Blade2Pitch Angle	Blade3Pitch Angle	GearboxBearing Temp	GearboxOil Temp	Generator RPM	GeneratorWin ding1Temp	GeneratorWin ding2Temp	HubTemp	MainBoxTemp	NacellePosi tion	ReactivePo wer	RotorRPM	TurbineStatus	WindDirection	Wind Speed
ActivePower	1																		
AmbientTemp	-0.06563716	1																	
BearingShaftTemp	0.65540011	0.246504704	1																
Blade1PitchAngle	-0.36899715	0.085460801	-0.47555656	1															
Blade2PitchAngle	-0.36797836	0.090904227	-0.46685637	0.99777351	1														
Blade3PitchAngle	-0.36797836	0.090904227	-0.46685637	0.99777351	1	1													
GearboxBearingTemp	0.81884951	0.017215929	0.883343827	-0.599178	-0.59210311	-0.59210311	1												
GearboxOilTemp	0.8219209	0.162511255	0.772887436	-0.5579493	-0.54979224	-0.54979224	0.90602	1											
GeneratorRPM	0.84960359	-0.130123331	0.640947074	-0.7570562	-0.75432496	-0.75432496	0.850209255	0.8069719	1										
GeneratorWinding1Temp	0.93138987	0.077858528	0.765114815	-0.3761021	-0.37105731	-0.37105731	0.85302369	0.899693	0.796434	1									
GeneratorWinding2Temp	0.93252058	0.078736011	0.763926158	-0.3723691	-0.36745098	-0.36745098	0.851922708	0.8983349	0.7952235	0.999958781	1								
HubTemp	0.34736317	0.5897288	0.809115891	-0.1985828	-0.19029629	-0.19029629	0.596064574	0.506648	0.3038036	0.48190509	0.48200251	1							
MainBoxTemp	0.10162711	0.836539437	0.542104047	0.19491146	0.1905622	0.1905622	0.307693949	0.1903498	-0.052185	0.215955134	0.217542934	0.761044	1						
NacellePosition	0.02810753	-0.037497912	0.183928708	-0.0582997	-0.05762416	-0.05762416	0.214485533	0.2776189	0.1769635	0.298451509	0.298707612	0.12815	0.005701408	1					
ReactivePower	0.71909203	-0.020312552	0.594545901	-0.3814779	-0.37954072	-0.37954072	0.751875046	0.770088	0.7680652	0.838460711	0.839367598	0.325071	0.095304256	0.2940378	1				
RotorRPM	0.84906943	-0.129289941	0.640584139	-0.7571465	-0.75453949	-0.75453949	0.850795248	0.8077759	0.9997231	0.797077318	0.795775605	0.302916	-0.047836972	0.1765109	0.76728826	1			
TurbineStatus	-0.00013548	-0.005510849	-0.00099024	-0.0005344	-0.00063646	-0.00063646	-0.002313534	-0.004536	0.0011333	-0.000650949	-0.000632935	-0.0027	-0.003158933	-0.0027142	0.00062654	0.001177	1		
WindDirection	0.02810753	-0.037497912	0.183928708	-0.0582997	-0.05762416	-0.05762416	0.214485533	0.2776189	0.1769635	0.298451509	0.298707612	0.12815	0.005701408	1	0.29403779	0.176511	-0.002714193	1	
WindSpeed	0.94039042	-0.09530764	0.596200784	-0.403092	-0.402786	-0.402786	0.80043306	0.7931404	0.8547872	0.89477717	0.894833749	0.298537	0.060822975	0.0343547	0.68011606	0.855584	-0.000660957	0.034354743	1

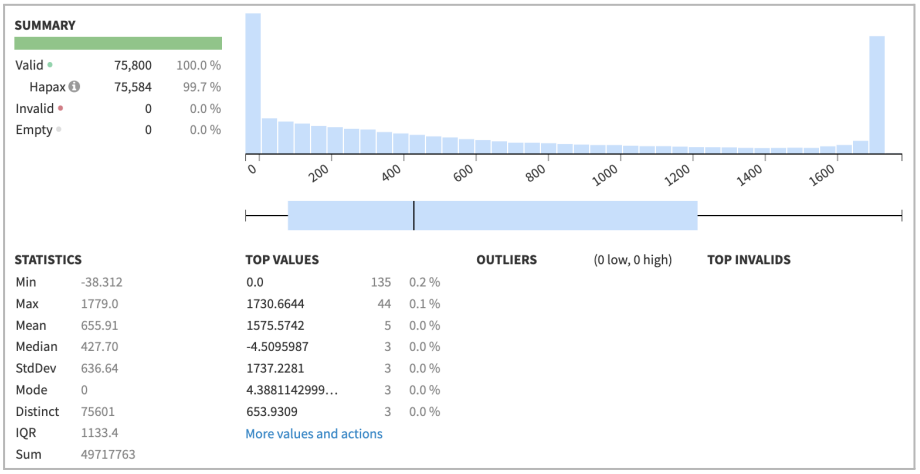
Source: own calculations

Graph 2: Dataiku Workflow (Test/Train Split)



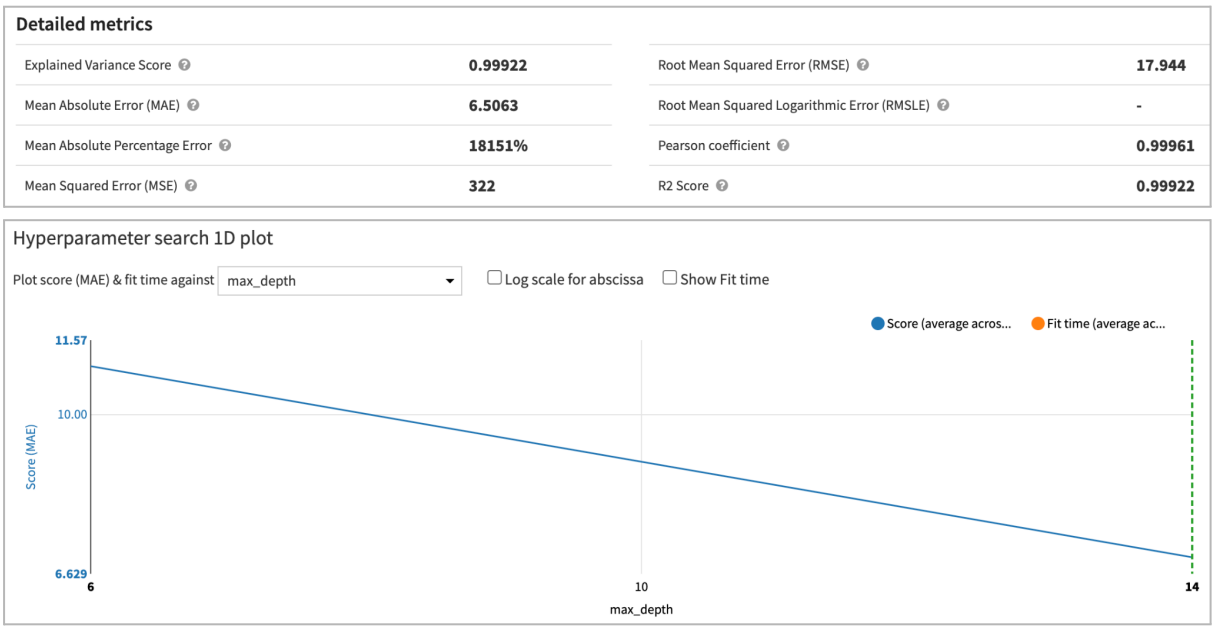
Source: Dataiku

Graph 3: Distribution of Active Power



Source: Dataiku

Graph 4: Optimal Model (Model 2)



References

Clifton, A., Kilcher, L., Lundquist, J. and Fleming, P., 2013. Using machine learning to predict wind turbine power output. *Environmental Research Letters*, 8(2), p.024009.

European Commission, 2020. *Guidance document on wind energy developments and EU nature legislation*. [online] Ec.europa.eu. Available at: <https://ec.europa.eu/environment/nature/natura2000/management/docs/wind_farms_en.pdf> [Accessed 15 July 2022].

Seyr, H. and Muskulus, M., 2019. Decision Support Models for Operations and Maintenance for Offshore Wind Farms: A Review. *Applied Sciences*, 9(2), p.278.

Tang, R., Luo, B., Deng, X. and Shen, Y., 2020. Research on Reactive Power and Voltage Control for Wind Farm Based on coordinate control of DFIGs and SVG. *Procedia Computer Science*, 175, pp.460-467.

Zayas-Gato, F., Jove, E., Casteleiro-Roca, J., Quintián, H., Pérez-Castelo, F., Piñón-Pazos, A., Arce, E. and Calvo-Rolle, J., 2022. Intelligent model for active power prediction of a small wind turbine. *Logic Journal of the IGPL*,.