General words on the task: I chose to regress two models for two turbines as I don't have any further turbine information for example the hardware. As we'd like to simulate the wind power at the same time as we discussed in the follow-up email, I used a linear regression model.

1. Load the data using the the data type pandas DataFrame.

I found that there're redundant spaces in the beginning or end of column names in the dataframe. And also all the decimal numbers were in the string format with comma instead of decimal dot. I decided to merge the first two rows into a new row of header, remove the redundant spaces and also changed each data into a float with decimal dot. In addition, I changed the Dat/Zeit column into the pandas datetime format for the anomaly analysis later.

2. Train any ML Model that predicts the power (single target) by a given windspeed (single feature).

As mentioned before, I chose to use a linear regression model. I shuffled the entire dataset and split 80% as training set and 20% as test set. As the linear regression model I trained requires no hyperparameter, a validation set is not required here.

3. Choose a good metric to measure the model performance.

I used the mean squared error as a standard metric for the linear regression task.

4. Explain why you chose this model architecture and what the limitations of this architecture might be.

I chose a linear regression model because these two features are highly correlated to each other which is proved in the next code block where I analyzed correlation between features. Also, linear regression would be the most straightforward and intuitive method for a regression task with one feature as input. Logically speaking, the higher the windspeed is, the more power we would expect from the turbine.

The limitation of a linear regression model is that in the end it only can learn linear functions to represent the relationships between features and target variable. So it can't model the non-linearity as a DNN since it has no such activation function. Also as I picked mean squared error as the metric, it can also be prone to outliers in the dataset.

5. Try out additional features. Which features did you choose? How did the results change and why?

There're a lot of features in the dataframe. To pick meaningful ones, I first analyzed the correlation between all other features and the target feature Leistung. It turns out the 'Strom- A', 'Gen1- °C', 'Strom- A.1', 'Strom- A.2' features are also highly correlated to the power. So I picked these four additional features for the linear regression model here. As we could see from the notebook, the performance on the same test set as the first single-feature model has improved by a large margin. This is because we enable the model to have a few more parameters and take more factors into account as we added four more highly correlated features.

ps: Although I don't really know what Gen1- stands for, but that higher current leads to higher power definitely makes sense to me.

6. Based on this model, where would you suspect turbine anomalies? Please list time frames and visualize the anomalies.

To answer this question, I would say that anomaly can be where the ground truth data points lie far away from the model predictions. To define how "far away" a point should be defined as an outlier/ anomaly, I assume the residual between predictions and ground truths to be a normal distribution. As I tested later, I treated indeed as a zero-mean normal distribution. Then I take the RMSE error on the entire dataset as the standard deviation. I plot the residuals and the 1sigma, 2sigma and 3sigma lines to help visualize the outliers as around 99% of data should be covered within 3-sigmas around the mean. The ground truths outside the 3-sigma lines are considered as anomalies and I calculated the number of anomalies in an interval of one hour and listed the time frame with more than 3 anomalies in an hour.