Report on technical task

In this task, I analysed wind turbine data to predict wind turbine output power and understand the presence and source of anomalies.

I first investigated how well wind speed alone can predict turbine power. To do this, I used a simple polynomial regression of order 5 (picked 5 through trial and error), given the smooth and predictable shape of the data. To quantify model performance, I used R^2 for its simple interpretability. I achieved an R^2 value of 0.987 and 0.985, respectively (didn't use train-test split as the model was too simple to overfit given the dataset size).

Then, I used more features and trained a Random Forest model. I chose Random Forest as it is robust to outliers, often doesn't require hyperparameter tuning for good performance, can handle non-linear relationships easily, and isn't very prone to overfitting. Its disadvantages are that it can take a while to train for large datasets (so not necessarily the most scalable), but it only ever took me ~10 seconds for the turbine data. It can also underfit compared to gradient-boosted methods like XGBoost, but in this instance it didn't seem to be the case.

I determined the features to be used from a correlation matrix – in the end I used 3 variables related to wind (wind speed (Wind), rotor speed (Rotor), and direction (Azimut)) and 4 variables related to temperature (Außen, Lager, Gen1-, and GetrT). I was able to achieve an R^2 value of 0.999 (0.999) on the training set and 0.992 (0.993) on the testing set for Turbine 1 (Turbine 2).

Finally, I investigated anomalies in the data. I defined anomalies in two ways:

- "Type A" anomalies – all datapoints whose predicted output power differs from the actual output power by 5x the standard deviation of residuals of the 1st model (polynomial regression with only wind speed as input)
- "Type B" anomalies – all datapoints whose predicted output power differs from the actual output power by 5x the standard deviation of residuals of the 2nd model (polynomial regression with only wind speed as input)

Using this definition, 0.26% (0.31%) of all datapoints are type A anomalies for Turbine 1 (Turbine 2), and 0.51% (0.48%) of all datapoints are type B anomalies for Turbine 1 (Turbine 2). While this definition is attractive due to its simplicity, it has some drawbacks, namely that the factor of 5x is quite arbitrary, so getting some shareholder input (e.g. on how much time they have to respond to all claims of anomalous behaviour) would be valuable to refine this value.

Conceptually, type A anomalies occur whenever any other factor than wind speed appreciable affects the power, whereas type B anomalies occur whenever any other factor than wind speed, direction, or temperatures of equipment/outside appreciably affect the power. There are merits for using both – on the one hand, if temperatures of equipment/outside increase from entirely environmental reasons (e.g. heat wave), then a change in turbine power might not necessarily be alarming, so type B anomalies might be more relevant. On the other hand, a temperature increase of the equipment might itself signal some malfunction, and this would be partly "explained away" by the type B calculation of anomalies, so it might only show up as a type A anomaly.

Finally, I investigated when such anomalies mostly occur. I found that a whopping 59% (35%) of all type A anomalies occur between 4am and 5am for Turbine 1 (3am and 4am for Turbine 2).