ReNew GreenTech Hackathon

in association with

MACHINE  HACK

ReNew POWER

Results and recommendations

Based on the results of the model and data provided, what are some useful recommendations you can make?

Talk about important features, important enaineered features, relations of target variable with other features, any predictive patterns. Conclude with model train, validation and test accuracy

1) Based on the input data we can see that there are turbines that are completely different from other. So it is very important to explore these turbines independently. In our approach we were exploring all turbines individually but another way is to make a clusters of all turbines and explore these clusters independently.
2) Based on feature importance exploration generated by different methods we can conclude that features deals with temperature measurements (in Celsius degree) have the highest impact on the output target temperature.
The best individual feature is *Temperature outside Nacelle* that have positive correlation about 0.5 with the target variable. The lowest impact have the features deals with speed and power: *Generator Speed, Average wind speed, Secondary Power generated by wind turbines at input source and Raw Active Power.*
3) We trained individual models for every turbine and have the lowest scores for Turbine_01 and Turbine_20 because they have big positive trend of target value from January to May and consequently the highest variance and average values.
4) For the most turbines K nearest neighbors is the best approach, but for one turbine Turbine_20 we were able to improve the MAPE score by applying Neural network approach.
5) The best MAPE on the public leader-board is **0.00943**. There is one issue with using timestamp variable – if to use it then we can reach the MAPE = **0.00028** but the models become useless for ReNew and also using that variable was prohibited by organizers.

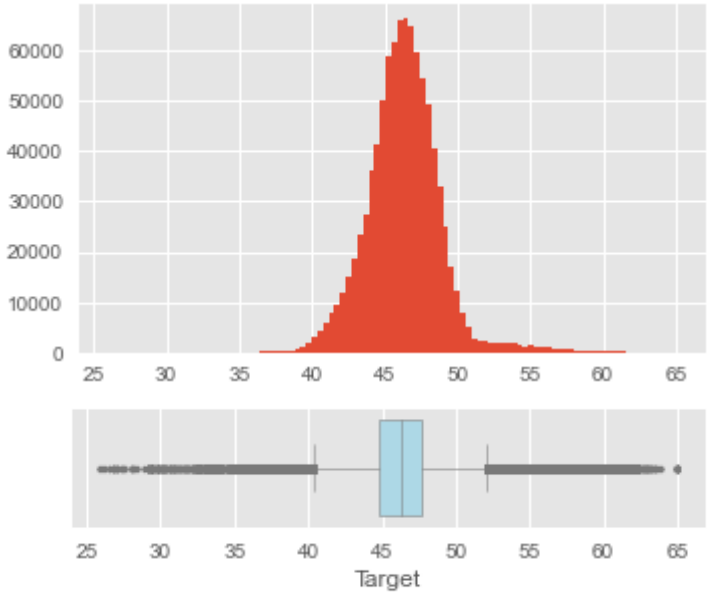
Data Understanding

List out the steps you took to understand the data

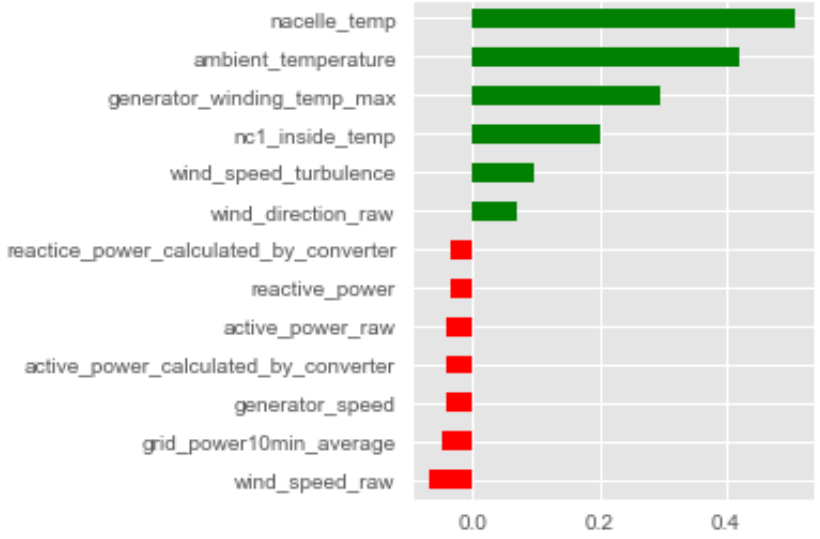
Write the steps in order. This should include:

1) *Distribution of target variable*
2) *Any data treatment to non-missing or non-outlier data*
3) *Any other point observed*

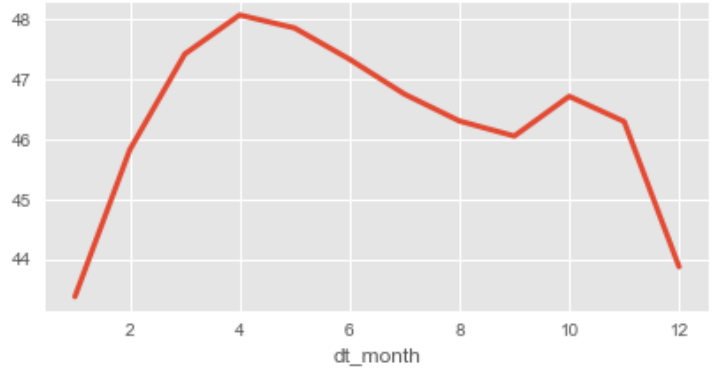
1) Distribution of target variable:



2) Features correlations with target variable:



3) Average target value by month number:



MACHINE HACK

ReNew POWER

Data Preparation

Before fitting your model, what processing did you perform?

Write the steps in order. This should include:

1) *Methodology used to detect and eliminate missing values and outliers*
2) *Any features you created out of data*
3) *Any data filtering (row filtering or column filtering) made*
4) *Any other changes*

1) There is no missing data in the datasets.
We have tried to replace outliers using the method with calculating *IQR = Quartile3 – Quartile1 and* then clipped the values into the interval:

*upper = Q3 +1.5*IQR*
*lower = Q1 – 1.5*IQR*

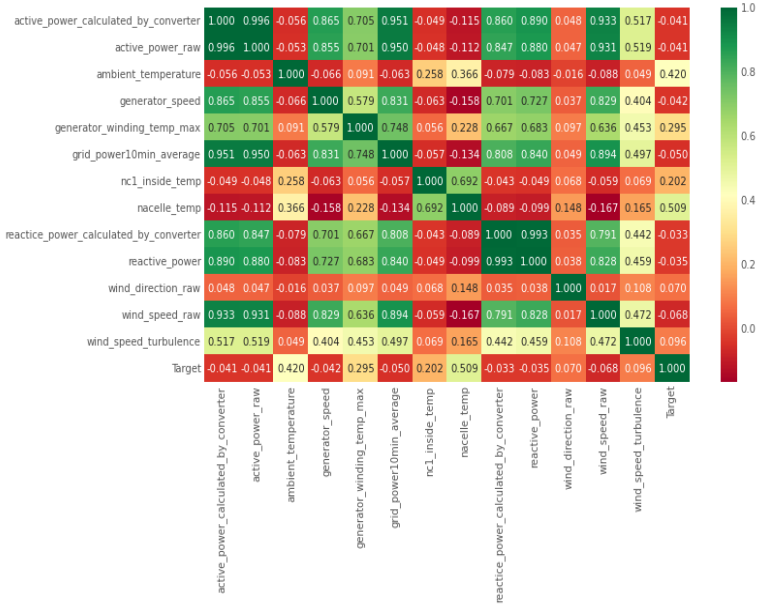
But after applying this method there was no improvements of MAPE score on the validation dataset, because

2) When we were training K nearest neighbors models we preprocessed out input features by applying different scaling methods available in sklearn:

- MinMaxScaler;
- StandardScaler;
- QuantileTransformer;
- RobustScaler.

We have chosen a scaling method that gives us the best score on the validation set Individually for each unique turbine.

Correlation matrix of input features:



3) We excluded features *active_power_calculated_by_converter* and *reactice_power_calculated_by_converter*, because they are just linear combinations of related features (see the picture above). Also there is a high correlated features with coefficient > 0.9 that have potentially high motivation to drop, but after some experiments with training models we decided to take them and not to eliminate on the first step.

Model Building & Evaluation

Specify which model you are using in your final model along with data and features being used

This should include:

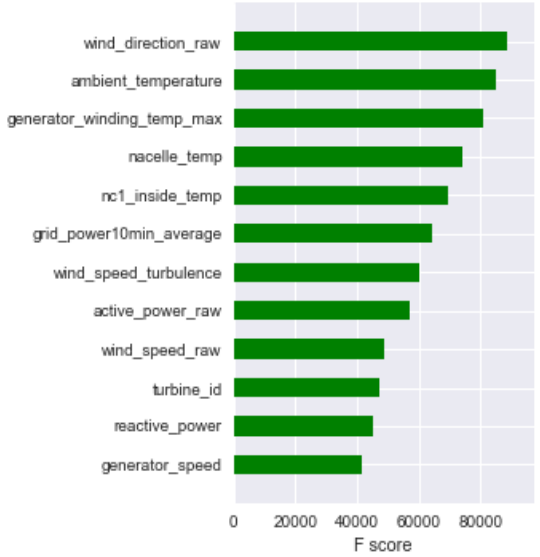
1) Model name and its hyperparameters
2) Whether any tuning was performed
3) Feature importance (may not necessarily be derived from the same model)
4) Anything else you would like to mention

1) We were training different algorithms individually for each turbine. For the most number of turbines KNN method is the best choice. For Turbine_20 we have improved KNN score by using Neural network model.
2) When training the KNNs we started from all 11 input features and then recursively eliminating them to the step when the MAPE metric stopped increasing (RFE). For each algorithm we were searching for the best hyper-parameters. The results of the best found MAPE scores have shown on the right table.
3) Feature importance of Xgboost model is shown on the right picture (later we will compare it with other methods)
4) The best MAPE on the public leaderboard is **0.00943** (it will be 0.01072 if we will use only 2/3 of data for nearest neighbors predictions on the test dataset).

MAPE scores of the best models:

Turbine_id	Random Forest	LightGBM	Xgboost	K Nearest Neighbors	Neural Network (MLP)
0 Turbine_01	0,0177	0,0174	0,0177	**0,0166**	0,0193
1 Turbine_10	0,0134	0,0133	0,0135	**0,0106**	0,0159
2 Turbine_103	0,0105	0,0101	0,0103	**0,0079**	0,0143
3 Turbine_105	0,0100	0,0099	0,0102	**0,0052**	0,0152
4 Turbine_108	0,0128	0,0130	0,0130	**0,0114**	0,0150
5 Turbine_120	0,0144	0,0144	0,0148	**0,0130**	0,0174
6 Turbine_123	0,0113	0,0112	0,0114	**0,0099**	0,0134
7 Turbine_13	0,0115	0,0111	0,0114	**0,0074**	0,0147
8 Turbine_139	0,0115	0,0114	0,0117	**0,0097**	0,0142
9 Turbine_14	0,0101	0,0100	0,0102	**0,0067**	0,0129
10 Turbine_15	0,0123	0,0123	0,0127	**0,0100**	0,0152
11 Turbine_158	0,0144	0,0141	0,0144	**0,0131**	0,0160
12 Turbine_18	0,0120	0,0117	0,0120	**0,0093**	0,0144
13 Turbine_19	0,0087	0,0087	0,0088	**0,0062**	0,0120
14 Turbine_20	0,0373	0,0272	0,0279	0,0331	**0,0238**
15 Turbine_97	0,0137	0,0134	0,0137	**0,0109**	0,0158

Feature importance in Xgboost model:



MACHINE HACK

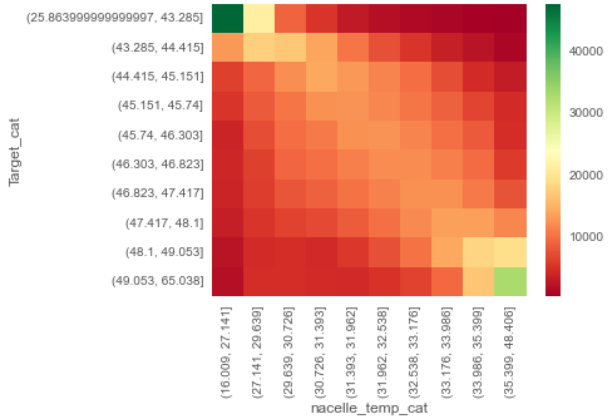
ReNew POWER

Which feature is the single most important feature for monitoring?

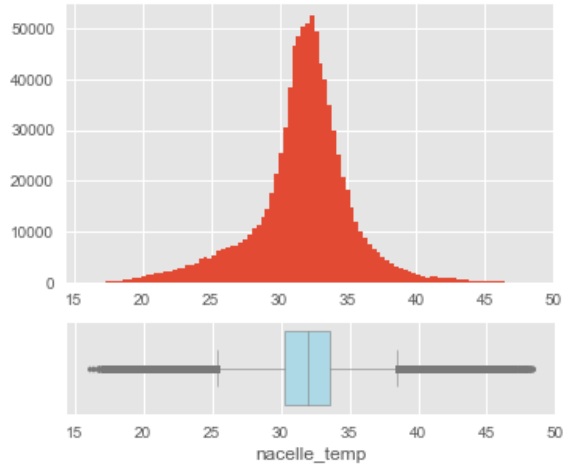
Any feature which should have been important based on data dictionary. Reason why you think it should be important
How does the feature impact the output?
What can ReNew do to control and monitor this feature?
Remember! This feature may not be the top feature in feature importance

We were predicting target variable that measured in Degree Celsius and by analyzing different feature importance plots we can see that the most significant features deals with temperature and also measured in Degree Celsius.
We have build linear and KNN models for every feature using only they one and found that the best single feature is *Temperature outside Nacelle (nacelle_temp).*

Heatmap of dependencies between nacelle_temp buckets and target buckets:



Temperature outside Nacelle distribution:



On the left picture we can see that there is a positive correlation between *Temperature outside Nacelle and Target.*

On the right picture is shown the distribution of this feature that ReNew can use the following way: Monitor the values of this feature and if they are placing outside the IQR interval you can interpret them as outliers and there is a high risk that the target values will also behave like this.

MACHINE HACK

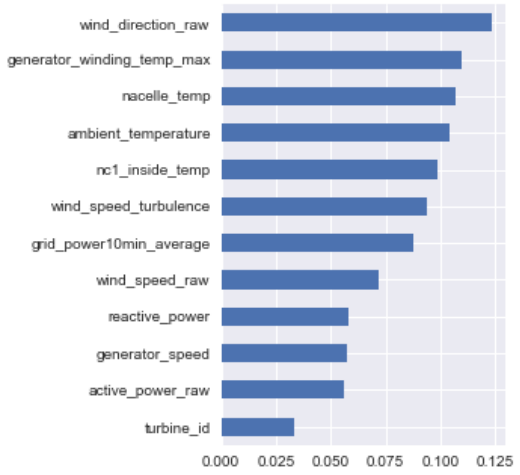
ReNew POWER

Which features are not affecting/have negligible effect on target temperature ?

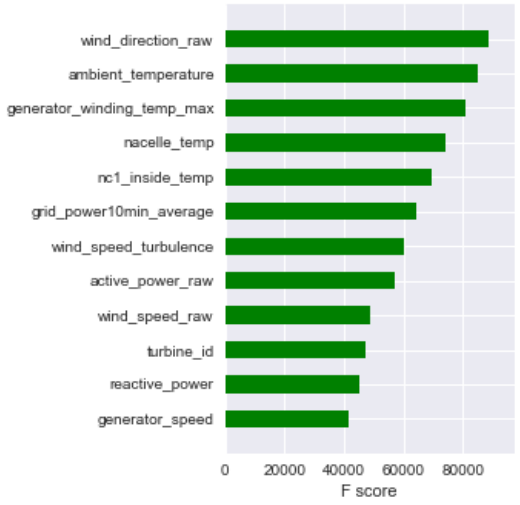
You can mention least important features and try to explain why there did not come out as important
Any feature that is important but its presence is causing a decline in model accuracy
Were they supposed to be important? Can something be done to increase their importance?

As we can see on the pictures below the most useless features defined by different approaches are: *generator_speed, wind_speed_raw, reactive_power* and *active_power_raw*. This features is measured in Meter/Second and KW units that have no sufficient impact on the target variable measured in Degree Celsius.

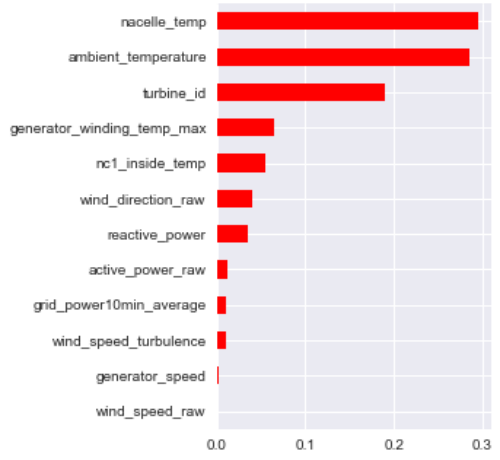
FI of LightGBM model:



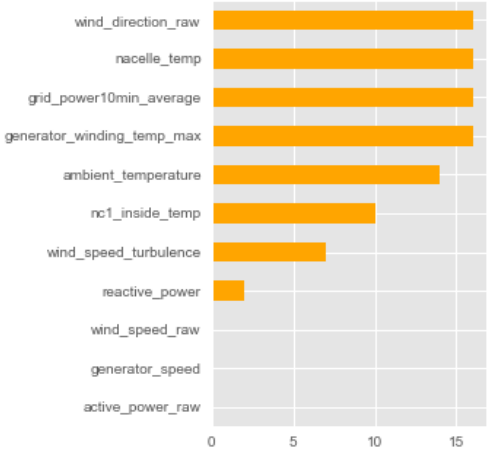
FI of Xgboost model:



FI of Random forest model:



FI of KNN model*:



MACHINE HACK

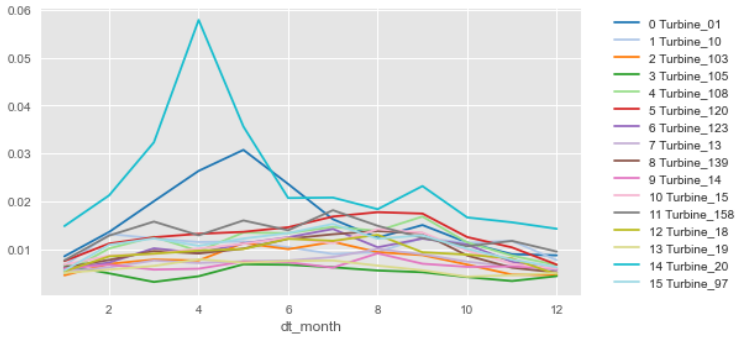
*We were using individual models for every turbine with recursive feature elimination procedure so this number shows how many times the feature was in the winning list for any turbine.

ReNew POWER

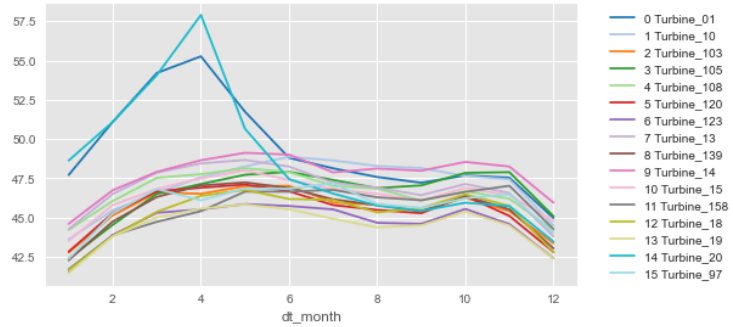
Which month gives us the highest accuracy and which month is giving the lowest accuracy and why?

Low Accuracy:
Is the reason constant data? Missing values? Lower/higher number of data points? Presence of outliers? Temperature? Any other complexity?
High Accuracy:
Is there any trend? Lower/higher number of data points? Bias in data? Temperature?
Bonus points if you can identify any business reason

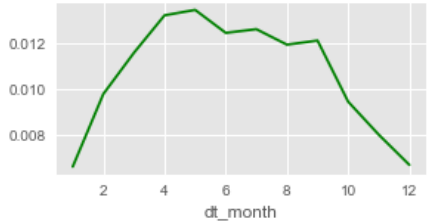
Monthly MAPE score by turbine:



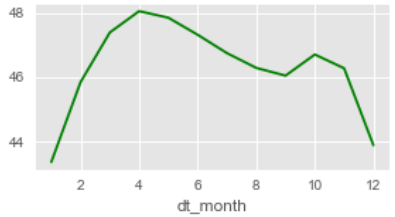
Monthly average target by turbine:



Monthly MAPE score overall:



Monthly average target overall:



As we can see on the left picture:
- The lowest MAPE scores are in April and May (Turbine_01 and Turbine_20 have the lowest scores).
- The highest MAPE scores are in January and December (Turbine_103 and Turbine_105 have the highest scores).

We calculated the average target values (on the right picture) and by comparing these plots with the scores plots we see the similar pictures.
We can conclude that the main factor impacting the output score is the average target temperature value. On the next slide we will see the reason why the average values are so high in April and May.

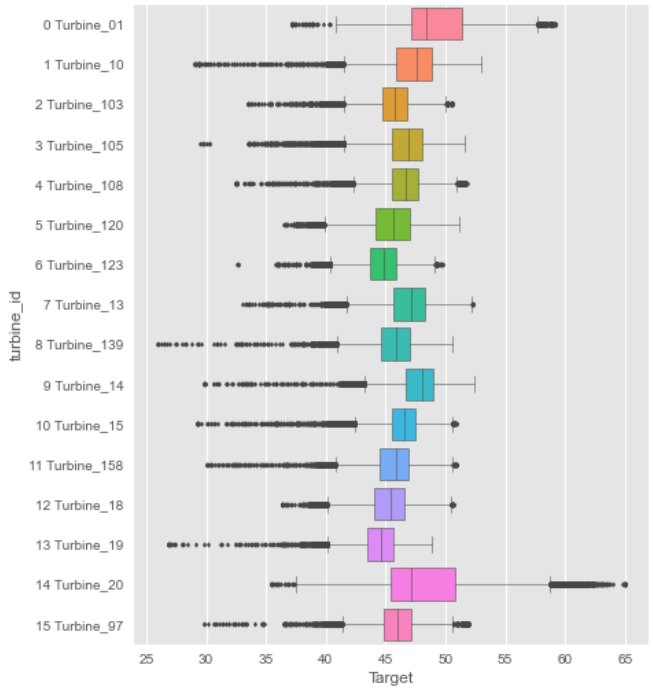
MACHINE HACK

ReNew POWER

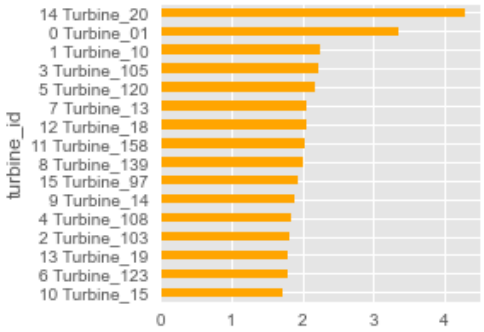
Which Turbine has the most variation in target temperature?

Define how you can define variation in taraet variable
Based on that definition, explain which turbine has highest variation in training data

Box-plots by turbine:



Standard deviations of turbines:



Time series plot by turbine:

As we can see on the left and middle pictures Turbine_20 has the highest variation and a standard deviation. On the right picture is the clear explanation of that fact.
For some reason there is positive trend from January to May of Turbine_01 and Turbine_20 and then big drop and return to normal values. This picture is also describing the fact why these two turbines have the lowest MAPE scores.

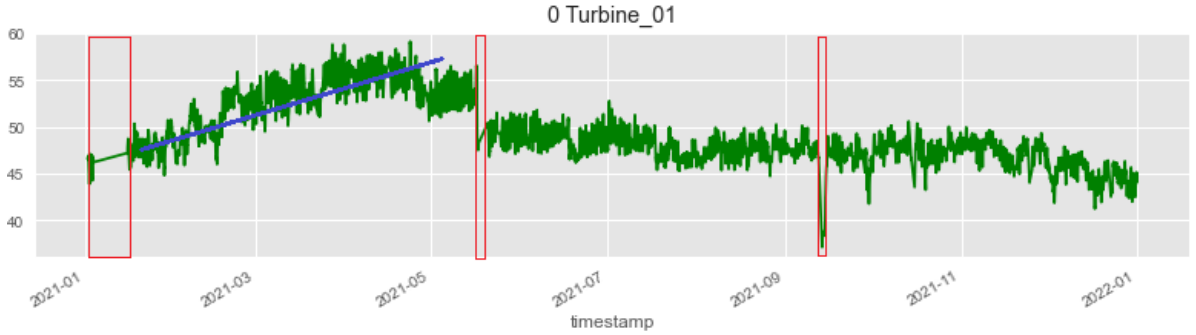
MACHINE HACK

ReNew POWER

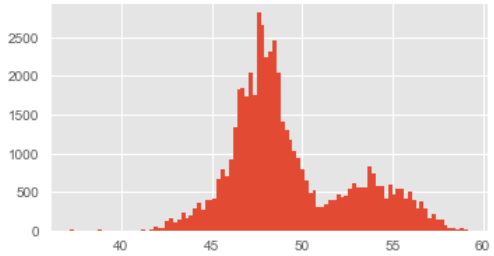
How can ReNew predict failure of a component by looking at target temperature and prevent such failures before they happen?

If there are abnormal changes (sudden spike or sudden drop which do not fall back to normal levels) then it can be indicative of a failure. Define a process to predict such failures before they happen

Time series plot for turbine_01 target values:



Target distribution for turbine_01:



If we look at the time series plot of Turbine_01 we can see the periods highlighted as red where there was no target values during a long period of time. Maybe there was some failures in that time. Based on that assumption we can define the simple rule: if we didn't get temperature measurements during some period of time then there is a failure.

Also we highlighted positive linear trend by blue line that is very suspicious and differs from other turbines (except turbine_14). It is the reason why we have two peaks in target distribution of that turbine (right picture). We can assume that this behavior is not normal and define the procedure how to prevent this:

1) To calculate different target statistics on stable period of time by every turbine: *mean, median, Quartile1, Quartile3, IQR = Quartile3 – Quartile1…*
2) Calculating these statistics in real time on a window period (let's say 1 week) and compare with statistics on the stable period.
3) If we observe high deviations of statistics and values out of the IQR interval then there is a risk that something goes wrong and failure will happen soon.

MACHINE HACK

ReNew POWER