ReNew GreenTech Hackathon

in association with

MACHINE HACK

Name: Mahesh Chandra Duddu
Email: duddumaheshchandra@gmail.com
PhNo: 9440642368, 6303817220

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 engineered features, relations of target variable with other features, any predictive patterns. Conclude with model train, validation and test accuracy

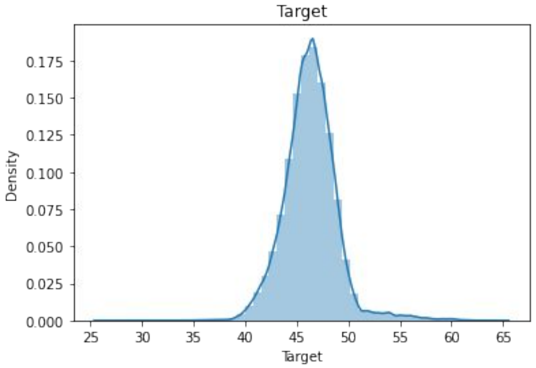
- Turbine_id is one of the important feature. Important engineered features obtained by ratio, product, sum and difference of features pairwise.
- nacelle_temp is correlated with target feature. Turbine_IDs have highly skewed distribution for target feature.
- nacelle_temp and ambient_temp are having bird like distribution travelling in different directions when plotted with target feature. Newly feature engineered features obtained from them really improved the model score.
- Model obtained the best local validation result of ***0.0108 Mean Absolute Percentage Error(MAPE)***.

Data Understanding

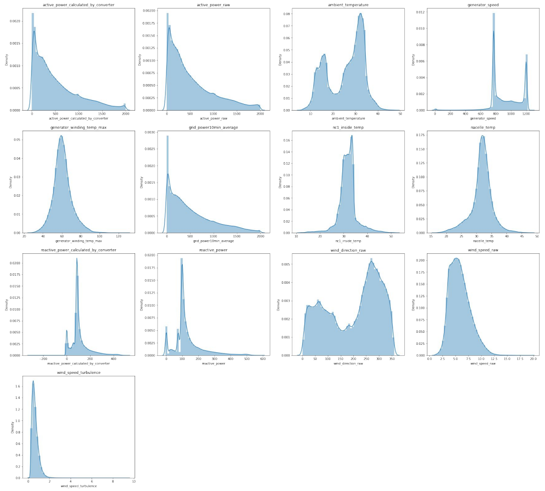
List out the steps you took to understand the data

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*



- Target Feature is moderately skewed, and hence distribution is not changed using any data transformation technique.
- Highly Skewed features(Skew value greater than 1) distribution is transformed using Square Root Transformation.
- There are lot of outliers observed in the data in the features from box plot. They are handled using Square Root Transformation.
- Turbine id, box plotted with target feature, shows that distributions have lot of outliers making them skewed.
- One feature is renamed as there is a spelling mistake.



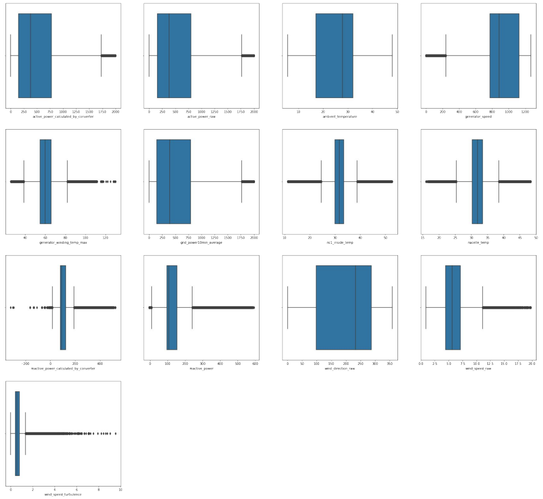
Data Preparation

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*

- There are no missing values, outliers are detected using box plot, and are handled by using Square root data transformation.
- Features are created by selecting distinct pair of features and doing division, multiplication, addition and subtraction of features pairwise.
- Data is filtered using turbine id. Training Data is split into 5 folds of train and validation set.
- Features are scaled using Standardization(Standar Scaler) followed by Normalization(Min-Max Scaler).



Model Building & Evaluation

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

This should include:

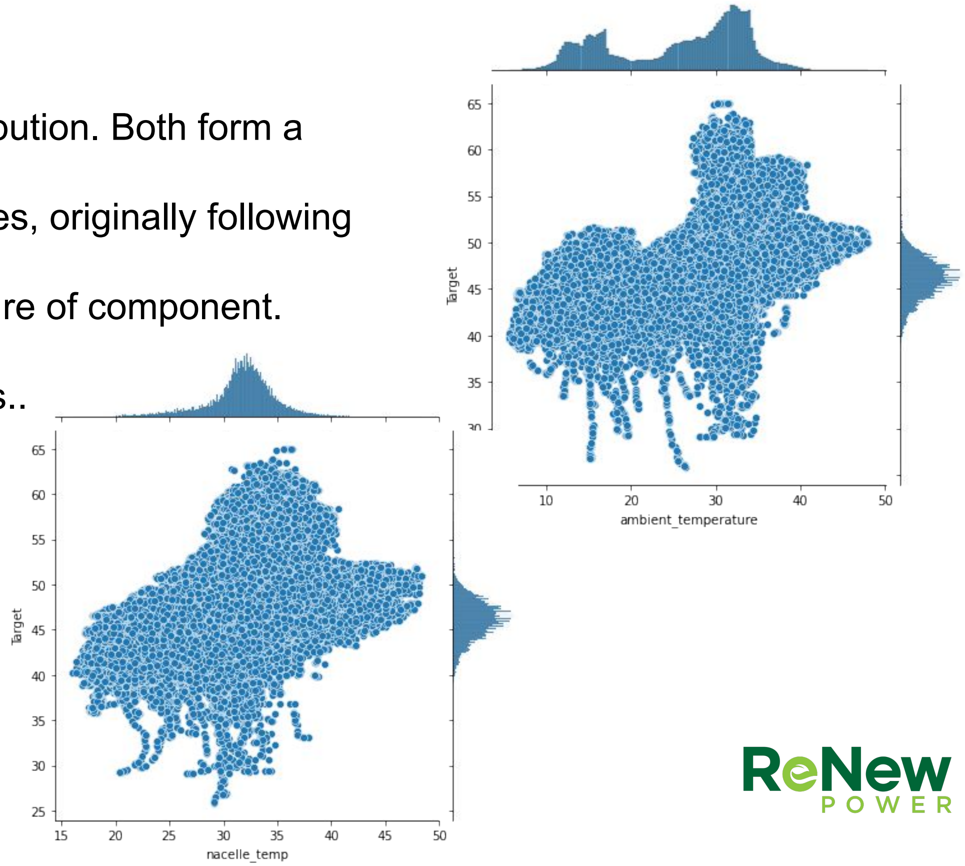
1) *Model name and its hyperparameters*
2) *Whether any tuning was performed*
3) *Feature importance (may not necessarily be derived from the same model)*
4) *Anything else you would like to mention*

- ExtraTreesRegressor(n_jobs = -1, random_state = 42)
- RandomisedSearchCV was performed, but the best parameters obtained didn’t improve the results of the model.
- Feature Selection didn’t improved the model results.

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

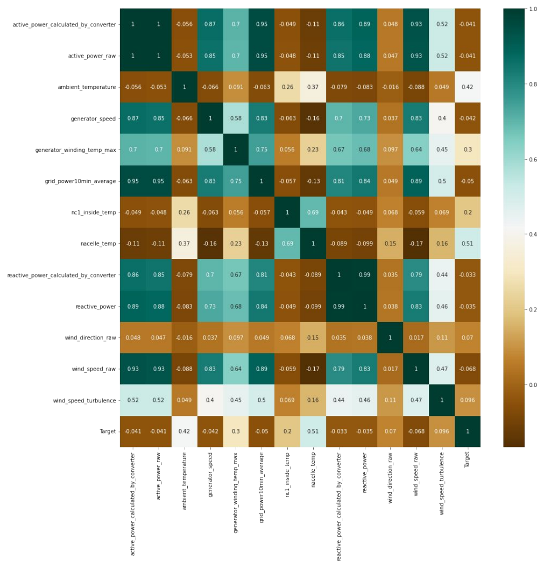
- nacelle_temp and ambient_temp are the important features based on data distribution. Both form a bird like distribution travelling in different directions when plotted with target values, originally following almost normal distribution. Monitoring these 2 features, help us in predicting failure of component.
- They are highly correlated with target feature when compared with other features..



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?

- reactive_power_calculated_by_converter, reactive_power are the top 2 least important features. They are less correlated with target feature. active_power_raw, active_power_calculated_by_converter are other 2 features that are highly correlated(pearson correlation coefficient = 1) with themselves, also they are less correlated with target feature.
- Using these features, newly obtained feature engineered features has improved the model performance.

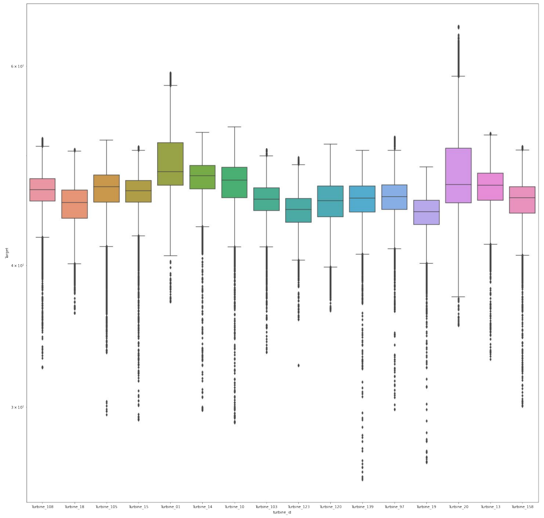


	active_power_calculated_by_converter	active_power_raw	ambient_temperature	generator_speed	generator_winding_temp_max	grid_power10min_average	nc1_inside_temp	nacelle_temp	reactive_power_calculated_by_converter	reactive_power	wind_direction_raw	wind_speed_raw	wind_speed_turbulence	Target
active_power_calculated_by_converter	1	1	-0.056	0.87	0.7	0.95	-0.049	-0.11	0.86	0.89	0.048	0.93	0.52	-0.041
active_power_raw	1	1	-0.053	0.85	0.7	0.95	-0.048	-0.11	0.85	0.88	0.047	0.93	0.52	-0.041
ambient_temperature	-0.056	-0.053	1	-0.066	0.091	-0.063	0.26	0.37	-0.079	-0.083	-0.016	-0.088	0.049	0.42
generator_speed	0.87	0.85	-0.066	1	0.58	0.83	-0.063	-0.16	0.7	0.73	0.037	0.83	0.4	-0.042
generator_winding_temp_max	0.7	0.7	0.091	0.58	1	0.75	0.056	0.23	0.67	0.68	0.097	0.64	0.45	0.3
grid_power10min_average	0.95	0.95	-0.063	0.83	0.75	1	-0.057	-0.13	0.81	0.84	0.049	0.89	0.5	-0.05
nc1_inside_temp	-0.049	-0.048	0.26	-0.063	0.056	-0.057	1	0.69	-0.043	-0.049	0.068	-0.059	0.069	0.2
nacelle_temp	-0.11	-0.11	0.37	-0.16	0.23	-0.13	0.69	1	-0.089	-0.099	0.15	-0.17	0.16	0.51
reactive_power_calculated_by_converter	0.86	0.85	-0.079	0.7	0.67	0.81	-0.043	-0.089	1	0.99	0.035	0.79	0.44	-0.033
reactive_power	0.89	0.88	-0.083	0.73	0.68	0.84	-0.049	-0.099	0.99	1	0.038	0.83	0.46	-0.035
wind_direction_raw	0.048	0.047	-0.016	0.037	0.097	0.049	0.068	0.15	0.035	0.038	1	0.017	0.11	0.07
wind_speed_raw	0.93	0.93	-0.088	0.83	0.64	0.89	-0.059	-0.17	0.79	0.83	0.017	1	0.47	-0.068
wind_speed_turbulence	0.52	0.52	0.049	0.4	0.45	0.5	0.069	0.16	0.44	0.46	0.11	0.47	1	0.096
Target	-0.041	-0.041	0.42	-0.042	0.3	-0.05	0.2	0.51	-0.033	-0.035	0.07	-0.068	0.096	1

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

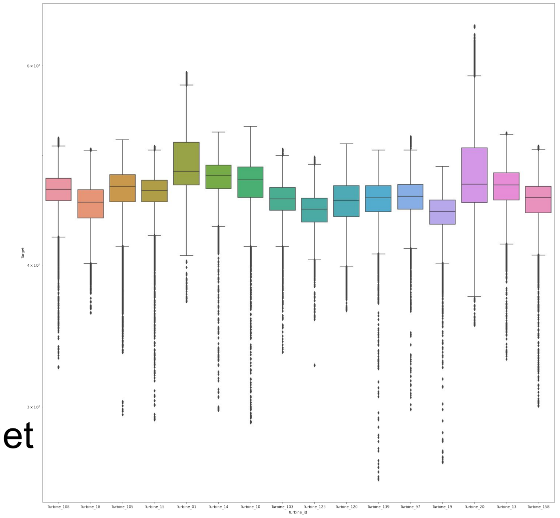
- As, time_id is an unique identifier of the data, and can't be used in the model. I haven't done any analysis related to that.
- I would like to provide analysis based on turbine id, model performed poorly(higher CV score) on data related to turbine_20, and model performed really well(lower CV score) on data related to turbine_19.
- Turbine_20 has its median is closer to lower quartile(Q1), resulting in highly right skewed data whereas turbine_19 having median at centre and less skewed with outliers only at lower fence.



Which Turbine has the most variation in target temperature?

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

- Variation in target temperature is defined as the change in the distribution with respect to target temperature distribution of other turbine id.
- Turbine_20 has the most variation in target temperature and different from other turbine ids. Its distribution shows that median is to the left of center in box plot and closer to the lower Quartile Q1(right skewed), one whisker is smaller, the other is larger and also have outlier values(higher Target temperature values). Turbine_01 also has similar distribution, but when compared with Turbine_20, Turbine_20 has more variation.



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

- Using IoT devices(installed with proposed model)(Something like sensors maybe) to predict the temperature of rotor bearing, and then finding out if there’s a lot of difference in the temperature when compared with expected. This helps in raising something like a direct alarm for preventing failure of a component.