Sardar Vallabhbhai National Institute of Technology, Surat

Wind Turbine Power Prediction

Presented By:-

Susmit Deshpande(P23DS026)

Ujjawal Vivek(P23DS002)

Anikat Pasi(P23DS020)

Content:-

1.Predicting Wind Turbine Power Production with Big Data and PySpark

2. Data and Objective

3.EDA

4.Generalized criterias for power production

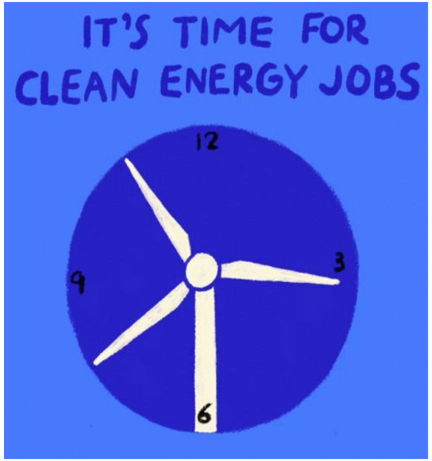
5.Model Building with PySpark

6.Comparing Real, Theoretical and Predicted Power Productions

7.Model Evaluation

Predicting Wind Turbine Power Production with Big Data and PySpark

In this presentation, we will explore how to predict wind turbine power production using big data and the PySpark library. Wind energy is a crucial renewable energy source, and accurate prediction of power generation is essential for grid management and efficient energy utilization. PySpark provides a powerful platform for analyzing large datasets and building machine learning models to tackle this challenge.

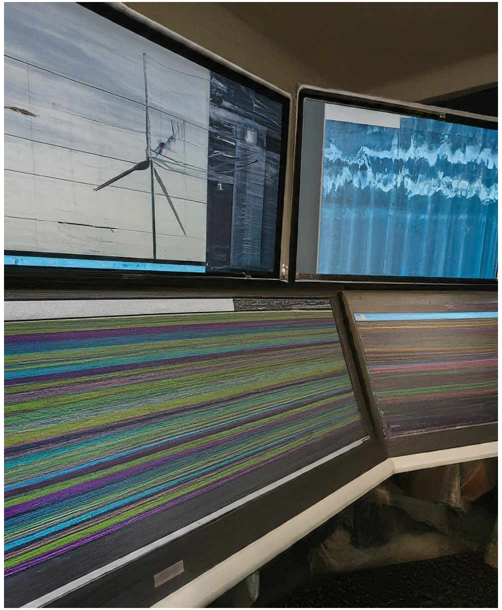


Data and Objective

Our study aims to predict wind turbine power production using a dataset containing various features.

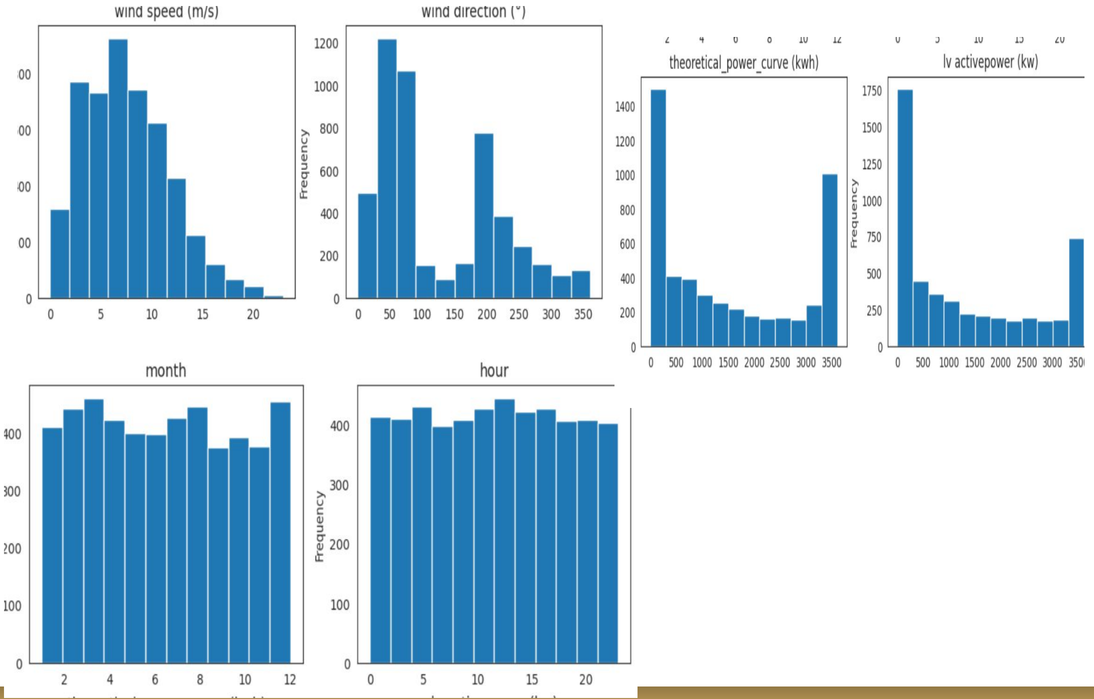
The dataset includes:

- Date/Time (for 10-minute intervals)
- V Active Power (kW): The power generated by the turbine for that moment
- Wind Speed (m/s): The wind speed at the hub height of the turbine
- Theoretical PowerCurve (KWh): The theoretical power values from the manufacturer
- Wind Direction (°): The wind direction at the hub height of the turbine (automatically tracked by the turbine)



Exploratory Data Analysis:-

- We analyzed the distribution of wind speed, power generation, and other features to understand their central tendencies and spread.

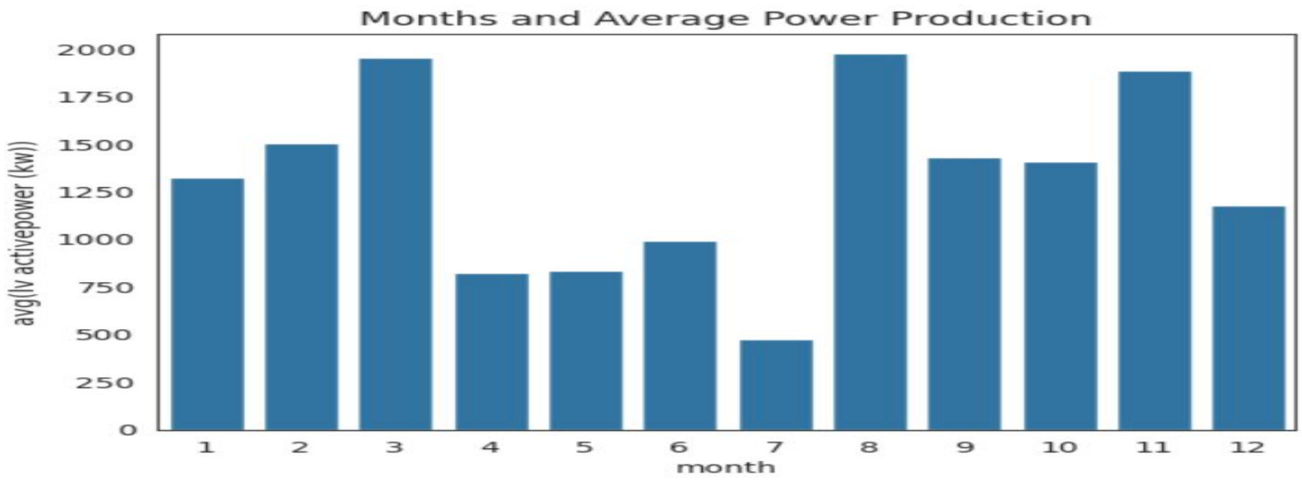


	wind speed (m/s)	theoretical_power_curve (kwh)	lv activepower (kw)
count	50530.00	50530.00	50530.00
mean	7.56	1492.18	1307.68
std	4.23	1368.02	1312.46
min	0.00	0.00	-2.47
25%	4.20	161.33	50.68
50%	7.10	1063.78	825.84
75%	10.30	2964.97	2482.51
max	25.21	3600.00	3618.73

- **Difference between the months for average power production.**

Higher Power Production Months: The months with the highest average power production appear to be around month 7 (likely corresponding to summer) based on the x-axis labels. This might be due to stronger or more consistent winds during these summer months.

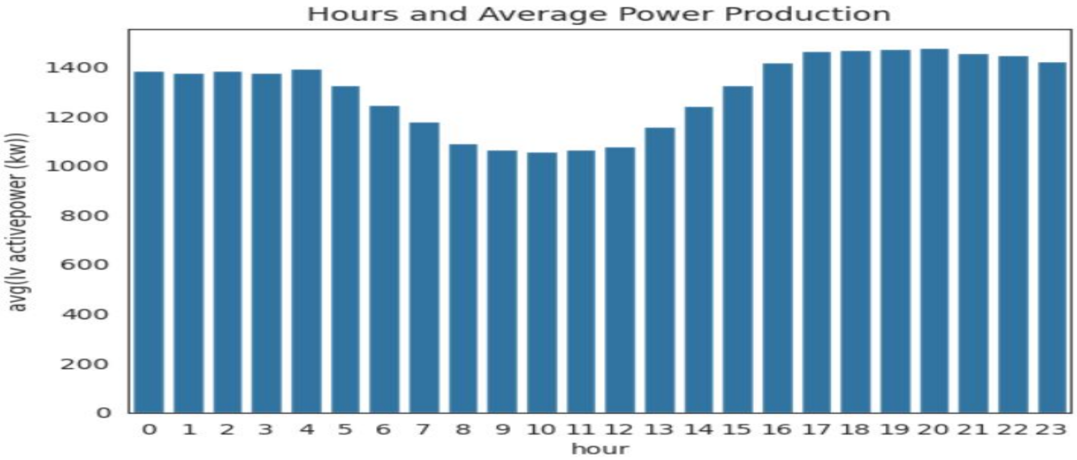
Lower Power Production Months: The months with the lowest average power production appear to be around month 1 (likely corresponding to winter) based on the x-axis



- Difference between the hours for average power production

Higher Power Production Hours: The hours around midday (possibly between 10 AM and 4 PM based on the x-axis scale) seem to experience the highest average power production. This could be due to stronger or more consistent wind speeds during these daytime hours.

Lower Power Production Hours: The early morning and evening hours (before 10 AM and after 4 PM) appear to have lower average power generation. This might be due to weaker or more variable winds during these times.



- **Correlation between the wind speed, wind direction and power production**

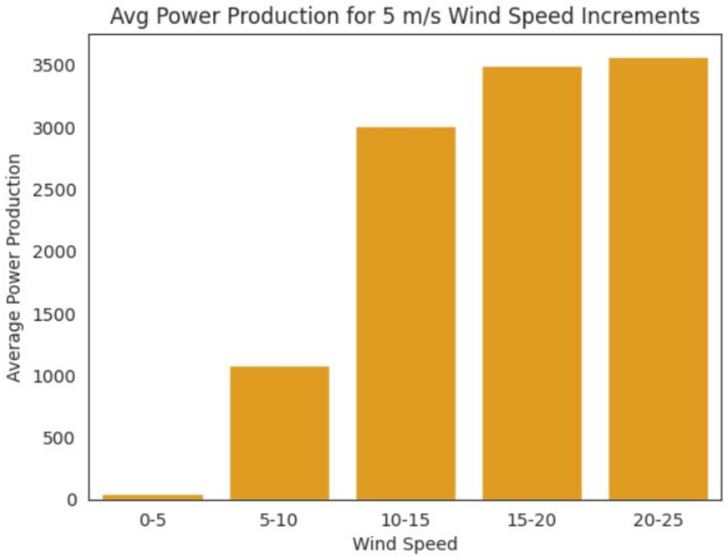
	wind speed (m/s)	wind direction (°)	month	hour	theoretical_power_curve (kwh)	lv activepower (kw)
wind speed (m/s)	1.00	-0.08	-0.01	0.03	0.95	0.91
wind direction (°)	-0.08	1.00	-0.18	0.00	-0.11	-0.06
month	-0.01	-0.18	1.00	-0.01	-0.00	0.04
hour	0.03	0.00	-0.01	1.00	0.03	0.03
theoretical_power_curve (kwh)	0.95	-0.11	-0.00	0.03	1.00	0.94
lv activepower (kw)	0.91	-0.06	0.04	0.03	0.94	1.00

observation from above Table:-

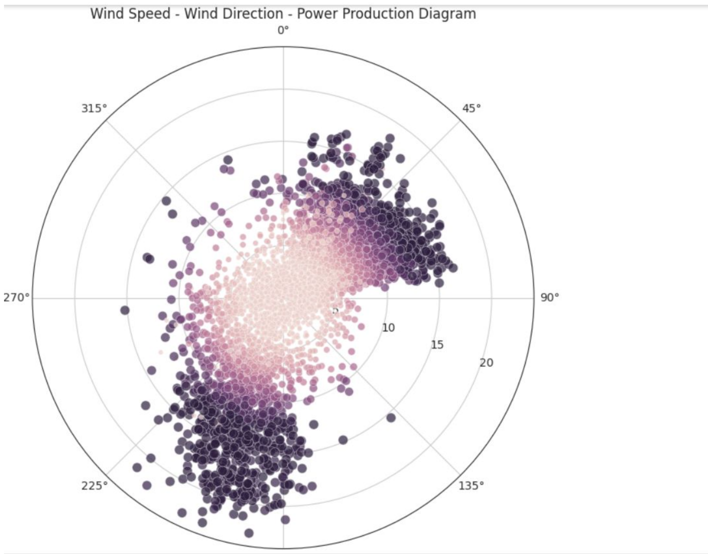
- Wind speed and power production is highly correlated as one would expect.
- We can see there are lower level power production for some wind directions.

- **Average power production level for different wind speeds**

From this graph we can see the power production reaches near a maximum level after the wind speed reaches 15 m/s.



- # power production for different wind directions and speeds
  - **Axes:**
    - **0° to 360°:** This represents the wind direction around the compass. 0° indicates wind coming from the north, 90° from the east, 180° from the south, and 270° from the west.
    - **Values along the sides:** These likely represent wind speed in meters per second (m/s).
  - **Power Production Contours:** The colored areas inside the circle represent the power output of the wind turbine at different wind speeds and directions. Generally: * **Darker colors** indicate higher power production. * **Lighter colors** indicate lower power production. * **Gray areas** on the edge might represent wind speeds that are too low or too high for the turbine to operate.



Observations from the Image (without specific values):

- The wind turbine appears to be most productive when wind is blowing from the **west (around 270°) and southwest (between 180° and 270°)**. This is the region with the darkest color in the diagram.
- Power production seems to be lower when wind comes from the **north (around 0°) and east (around 90°)**. These areas are lighter colored.
- The wind turbine might not generate any usable power at very low or very high wind speeds (represented by the gray areas outside the colored regions).

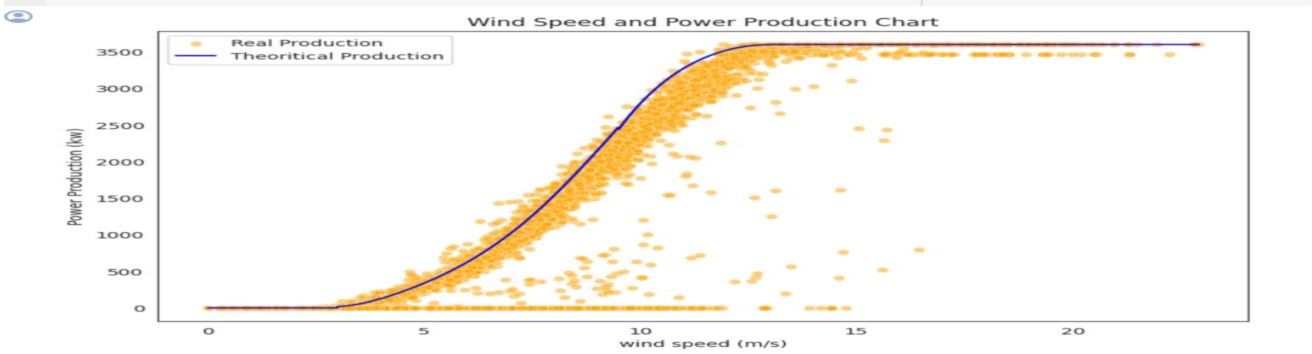
- **Manufacturer's theoretical power production curve fit well with the real production:-**

From this graph , we can see the theoretical power production curve generally fits well with the real production.

We can see the power production reaches a maximum level and continues in a straight line if the wind speed reaches to 15 m/s.

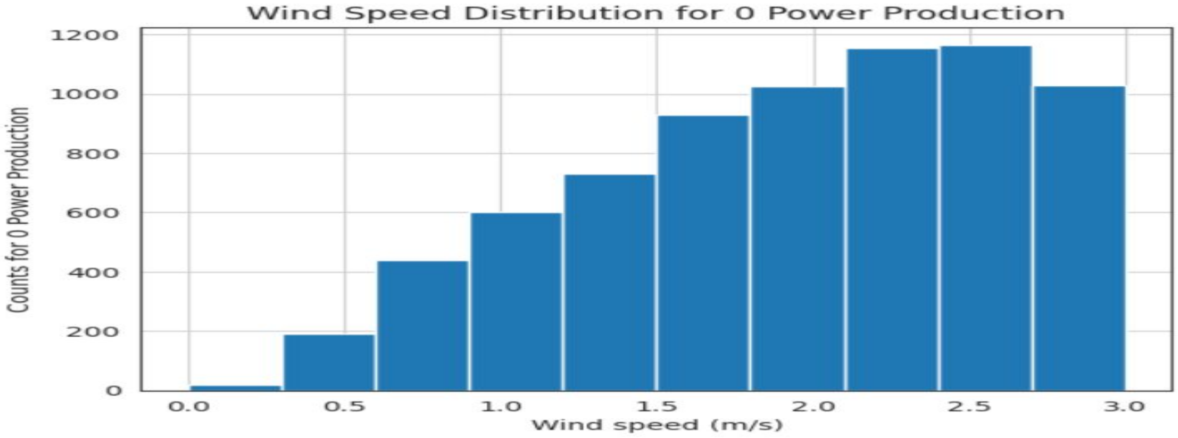
Also we can see there are some 0 power production, even the wind speed is higher than 5 m/s. I want to investigate the reason.

But before what is the minimum wind speed for theoretical power Production Curve.



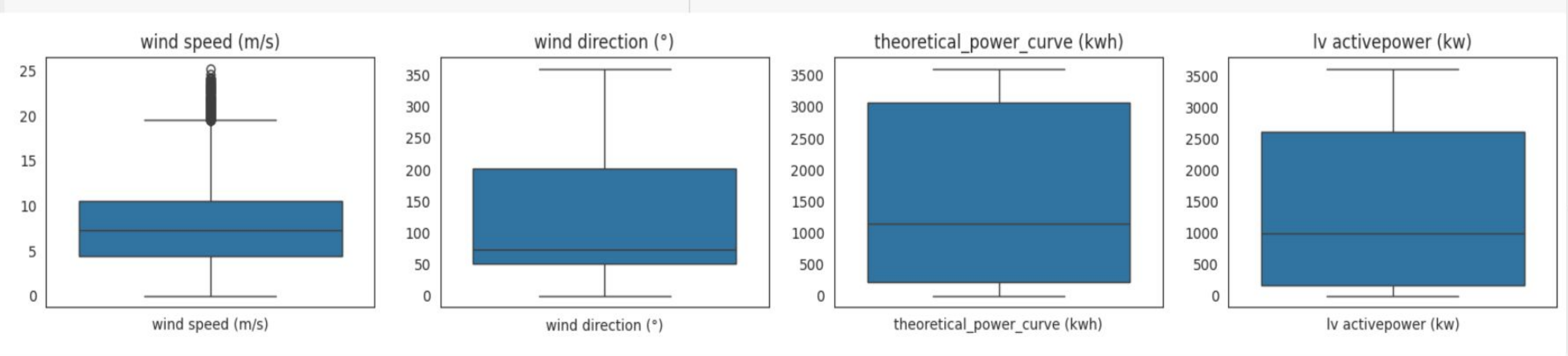
Wind speed threshold value for zero theoretical power

	wind speed (m/s)	theoretical_power_curve (kwh)	lv activepower (kw)
921	0.69	0.00	0.00
4598	3.00	0.00	0.00
1962	2.38	0.00	0.00
1986	0.69	0.00	0.00
2397	2.59	0.00	0.00



We can see from above, limit for the theoretical power curve is 3 m/s wind speed. If the wind speed is below 3 m/s model doesn't expect any power production.

Outliers Analysis:-

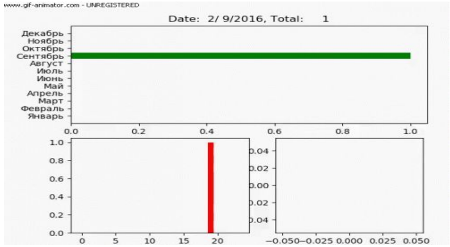


- From the graphs above I can see there are some outliers in the wind speed data.
- We find the upper and lower threshold values for the wind speed data are:-
  `Quantile (0.25): 4.45584678649902 Quantile (0.75): 10.4771900177001`
  `Lower threshold: -4.57616806030259 Upper threshold: 19.50920486450172.`
- From outlier analysis we observe that It is a rare event for wind speed to be over 19 m/s in our dataset.
- Out of 47033, there is only 407 observations while the wind speed is over 19 m/s.

Generalized criterias for power production:-

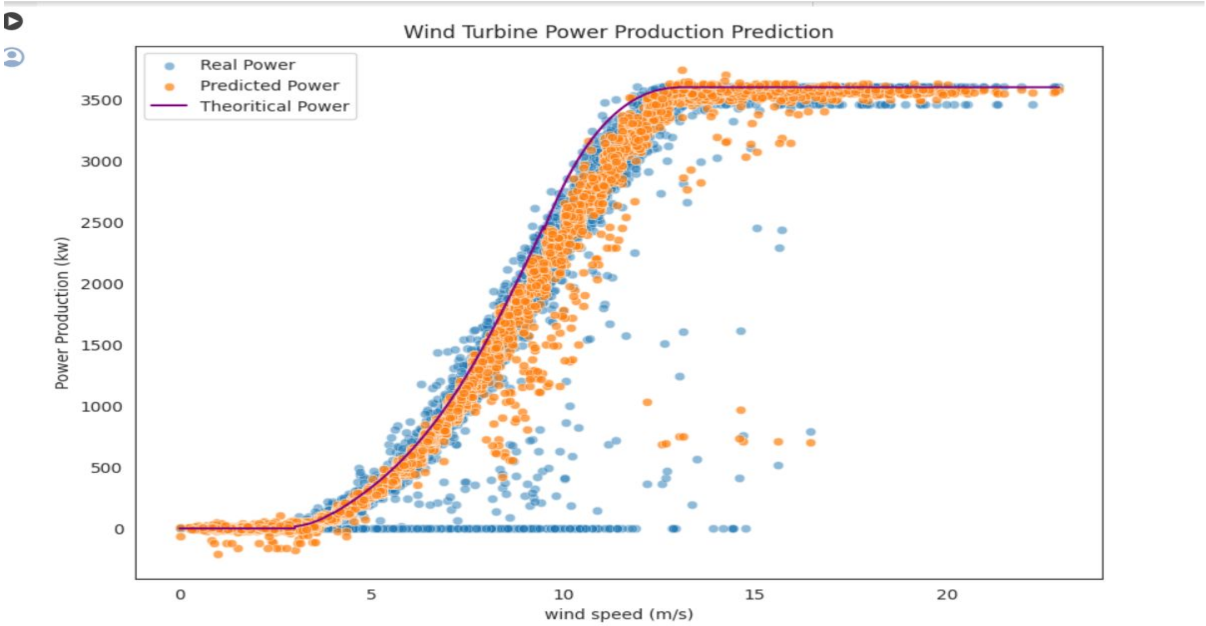
It is important to understand the pattern in the data. We should learn the data before the machine.

1. We saw from the graph that in March, August and November, the average power production is higher.
2. The average power production is higher daily between 16:00 and 24:00.
3. The power production is higher when the wind blows from the directions between 000-090 and 180-225 degrees.



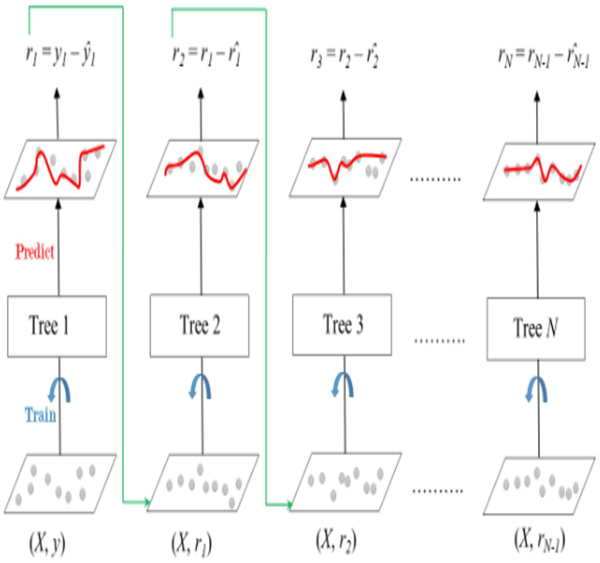
Comparing Real, Theoretical and Predicted Power Productions

From this graph, the model fits better to the real power productions, than the theoretical power production curve.



Model Building with PySpark

- We will focus on using **Gradient Boosting Trees (GBT)** for our prediction model.
- GBT is an ensemble method that combines multiple weak decision trees into a stronger final model.
- Each tree in the ensemble learns from the errors of the previous tree, resulting in improved accuracy.



Model Evaluation

- R2 score means, real power production 98% variability can be explained by the ML model.
- MAE is 83.43 it mean absolute difference between the real and predicted power production.
- RMSE is 179.34 the square root of mean squared difference between the real and predicted values.
- Even though the R2 is high, we should also check the MAE and RMSE values with the real values summary statistics.
- One can tune the hyperparameters to increase the model success. But I this look good enough for me.

Thank you