Dhirubhai Ambani Institute of Information and Communication Technology

M.Sc DS IT609 Big Data Processing

Wind Turbine Power Prediction

Vishal Vasoya
(202118013)
Surat, Gujarat

Nidhi Sadhwani
(202118043)
Nagpur, Maharashtra

Abstract- *Globally wind power has seen considerable growth in all grid systems. In the coming decade wind power is also expected to expand rapidly. Wind power is variable and intermittent over various time scales because it is weather dependent. Therefore wind power integration into traditional grids needs additional power system and electricity market planning and management for system balancing. This extra system balancing means that there is additional system costs associated with wind power assimilation. Wind power forecasting and prediction methods are used by system operators to plan unit commitment, scheduling and dispatch and by electricity traders and wind farm owners to maximize profit. Accurate wind power forecasting and prediction has numerous challenges. This paper presents a study of the existing and possible future methods used in wind power forecasting and prediction for the same.*

1. Introduction

In this project we are going to predict a wind turbine power production by using the wind speed, wind direction, month and hour data.

Wind turbines convert the kinetic energy in the wind into mechanical power. This mechanical power can be used for specific tasks (such as grinding grain or pumping water), or can be converted into electricity by a generator. They use blades to collect the wind's kinetic energy. Wind flows over the blades creating lift (similar to the effect on airplane wings), which causes the blades to turn. The blades are connected to a drive shaft that turns an electric generator, which produces (generates) electricity.

One study claimed that, as of 2009, wind had the "lowest relative greenhouse gas emissions, the least water consumption demands and the most favourable social impacts" compared to photovoltaic, hydro, geothermal, coal and gas energy sources.

2. Theory and Methodology

Regression is a machine learning algorithm based on supervised learning. It performs a regression task. It provides a function that describes the relationship between one or more independent variables and a response, dependent or target variable. The variable you want to predict is called the dependent variable. The

variable you are using to predict the other variable's value is called the independent variable.

Regression theorem is stated mathematical as following:

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable
f = function
X_i = independent variable
β = unknown parameters
e_i = error terms

2.1 Data Understanding

The dataset consists of 50530 observations and 5 input features as follows:

- Date/Time (for 10 minutes intervals)
- LV 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 (the wind speed that turbine use for electricity generation)
- Theoretical Power Curve (kWh): The theoretical power values that the turbine generates with that wind speed which is given by the turbine manufacturer
- Wind Direction (°): The wind direction at the hub height of the turbine (wind turbines turn to this direction automatically.

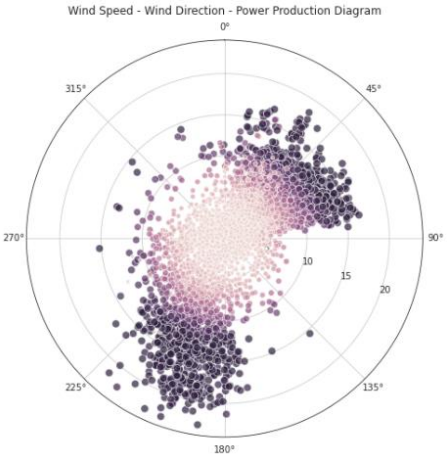


Figure 1

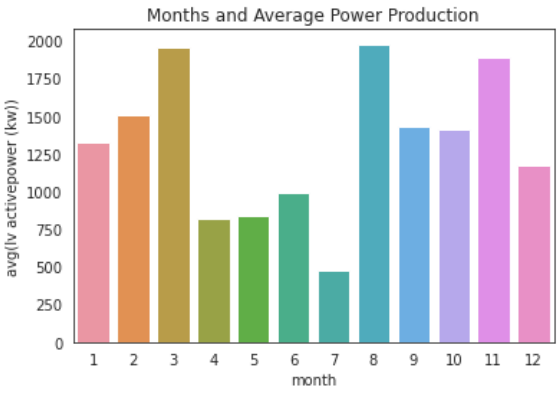


Figure 2

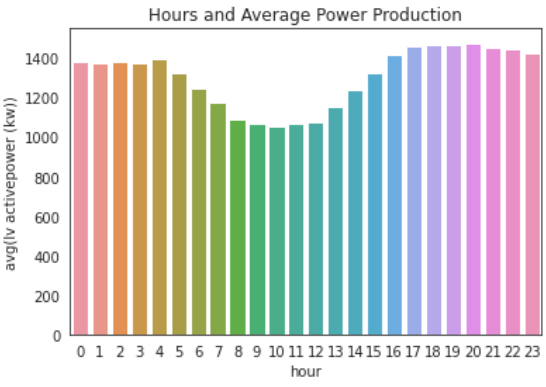


Figure 3

Figure 1 is a graph for wind showing relation between speed, direction and production. We can see that the wind turbine produces more power if the wind blows from the directions between 000-090 and 180-225 degrees.

Figure 2 is a graph showing relation between the months and average power production. We saw from the graph that in March, August and November, the average power production is higher.

Figure 3 is a graph showing relation between the hours and average power production. We observed that the average power production is higher daily between 16:00 and 24:00.

2.2 Data Preprocessing

To import our dataset we used PySpark. Then we extracted month and hour values from date&time records then replaced date&time column with 2 new columns naming as month and hour.

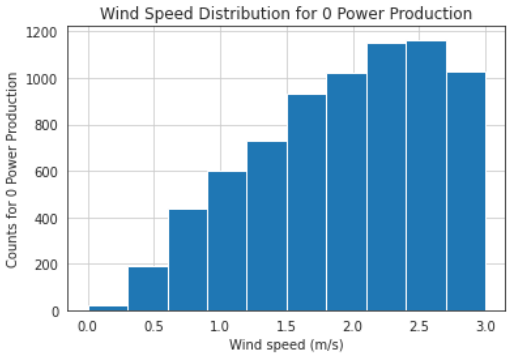


Figure 1

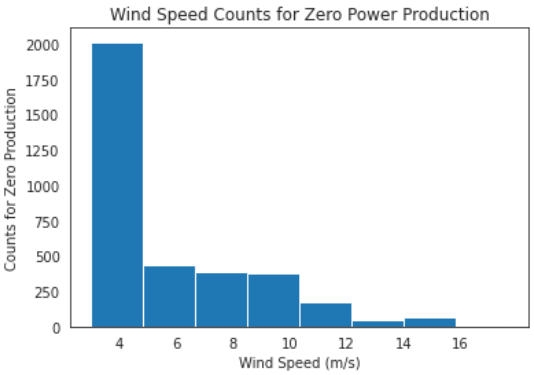


Figure 2

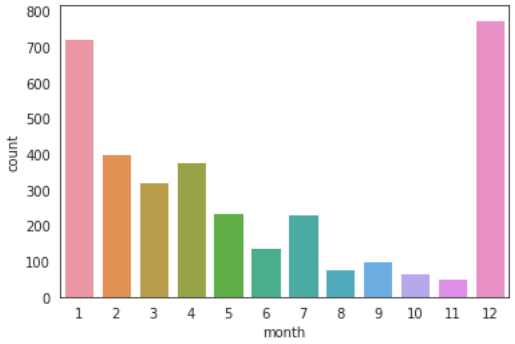


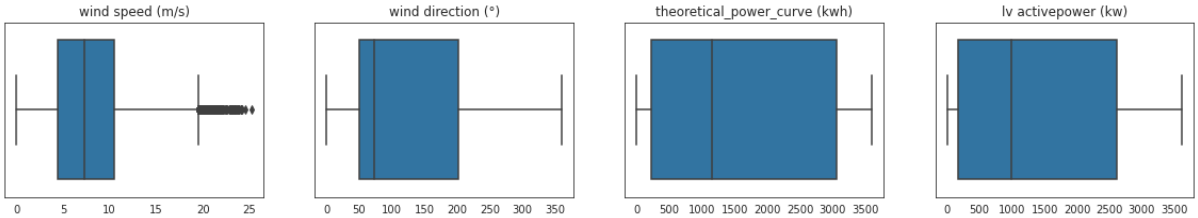
Figure 3

From figure 1 we can see limit for the theoritical power curve is 3 m/s wind speed. If the wind speed is below 3 m/s, model doesn't expect any power production. But there are some observations for 0 power production even the wind speed is more than 3 m/s.

So theoritically wind speed threshold should be equal to or greater than 4 m/s. But figure 2 tells us that there are also other observations with zero power production while the wind speed is higher.

Let's see figure 3 which shows monthly distribution for zero power production. It is usually in December and January when the wind turbine doesn't produce production. Because we cannot decide if these zero power productions are caused by maintenance periods or something else, we are going to accept those 3497 observations as outliers and remove them from the dataset. So now we have reduced to 47033 records from 50530.

Now we have plot the boxplot for features in our dataset.



From the graphs above as we can see there are some outliers in the wind speed data. So we found the upper and lower threshold values for the wind speed data, to analyze the outliers.

The value for Quantiles and Threshold are as follows:

```
Quantile (0.25):  4.45584678649902   Quantile (0.75):  10.4771900177001
Lower threshold:  -4.576168060302599  Upper threshold:  19.50920486450172
```

It is a rare event for wind speed to be over 19 m/s in our dataset. Out of 47033, there are only 407 observations where the wind speed is over 19 m/s. So instead of removing these 407 observations we limited their value equal to the upper threshold i.e. 19.5.

2.3 Model Building

We are applying two machine learning models in our dataset in order to predict wind turbine power. Since our target variable is continuous in nature therefore we are using Regression Model.

1. Linear Regression Model

Linear Regression Model has two categories: Simple and Multiple.
Since our dataset has multiple features so we need to apply Multiple Linear Regression model in our project. Regression allows us to estimate how a dependent variable changes as the independent variable(s) change. This is used to estimate the relationship between two or more independent variables and one dependent variable.

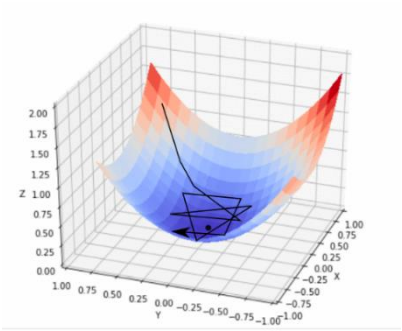
Multiple Linear Regression theorem is stated mathematical as following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

- Cost Function (J)

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).



- Gradient Descent

To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

2. Gradient Boost Tree

In order to minimize a loss function, Gradient Boosting Trees (GBTs) iteratively train many decision trees. On each iteration, the algorithm uses the current ensemble to predict the label of each training instance. The features of GBT are as follows:

- Handle categorical features (and of course numerical features too)
- Extend to the multiclass classification setting
- Perform both the binary classification and regression (multiclass classification is not yet supported)
- Do not require feature scaling
- Capture non-linearity and feature interactions, which are greatly missing in LR, such as linear models

Note: Validation while training: Gradient boosting can overfit, especially when you have trained your model with more trees. In order to prevent this issue, it is useful to validate while carrying out the training.

2.4 Testing and Evaluation

- R^2: “The proportion of the variance in the dependent variable that is predictable from the independent variable(s).” This score varies between 0 and 100%. It is closely related to the MSE, but not the same.

R^2 theorem is stated mathematical as following:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- MSE: Mean square error is the average of the square of the errors. The larger the number the larger the error.

Mean Square Error theorem is stated mathematical as following:

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)}_{\text{The square of the difference between actual and predicted}}^2$$

- RMSE: The root-mean-square error is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

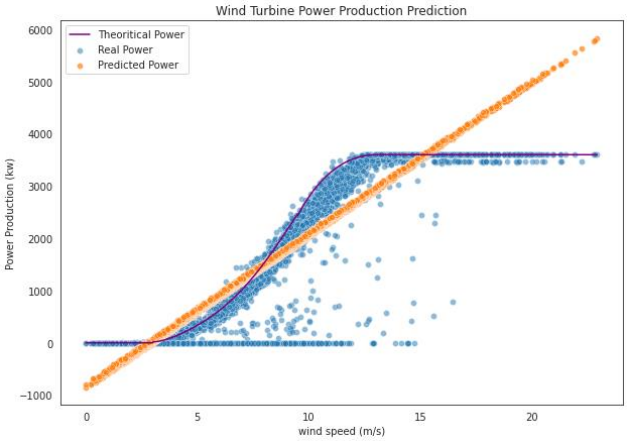
Root Mean Square Error theorem is stated mathematical as following:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

3. Result Interpretation

We have applied different two types of Regression Model that is Linear Regression and Gradient Boost Tree. And we have further compared two of these to check which one is best and most suitable for our dataset. below we show results using R^2, MSE and RMSE after applying the algorithms.

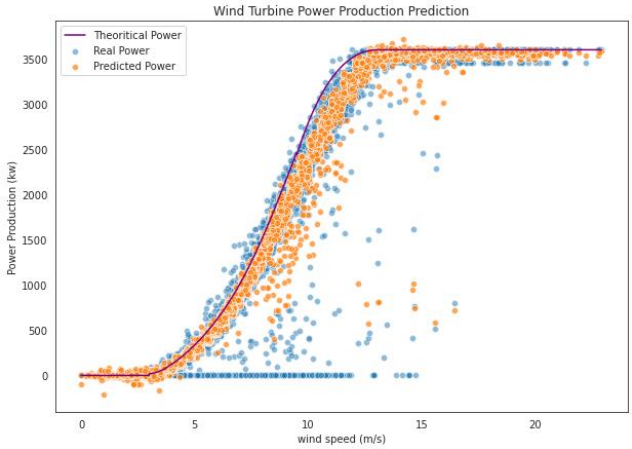
```
R2 SCORE :  0.8838207062920085
MAE      :  353.93252042900764
RMSE     :  446.9078664380258
```



.

Linear Regression

```
R2 SCORE :  0.9813282317134948
MAE      :  83.67918756221894
RMSE     :  179.1907264941426
```



Gradient Boost Tree

4. Conclusion

Wind turbine power is an infinitely sustainable form of energy that does not require any fuel for operation and generates no harmful air or water pollution produces no green house gases and toxic or radioactive waste. In this project, we try to use Machine Learning algorithm for effective and timely prediction of wind turbine power. We performed Regression model on dataset containing features on which wind power is dependent and other related information for prediction. In order to get high accuracy we trained dataset on two regression model that is Linear Regression and Gradient Boost Tree.

A linear regression line is an easy-to-read way of obtaining the general direction of price over a past specified period. Unlike a moving average, which bends to conform to its weighting input, a linear regression line works to best fit data into a straight line. So that it gives r square score is 0.88 which is good but not best for our data. Our data is non linear that's why it give a best fit line which is missing some of the actual data point at the bottom and top of the data point. As per our conclusion we need a model which give a non linear line that is give best fit curve.

GBT Regressor is Highly efficient on both classification and regression tasks. Gradient boost is a powerful boosting technique. It improves the accuracy of the model by sequentially combining weak trees to form a strong tree. In this way it achieves low bias and low variance. So that it gives r square score is 0.98 which is best for our data. As per figure of gradient boost tree regression gives a non linear line which give best fitting between actual data and predicting data.

5. References

- https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.regression.GBTRegressor.html
- https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a
- https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html
- https://vlab.amrita.edu/index.php?sub=77&brch=297&sim=1743&cnt=1
- https://www.tutorialspoint.com/pyspark/pyspark_mllib.htm