ENSY5800 Application of AI in Energy Systems
Spring 2022 April 30, 2022
Northeastern University Boston MA, USA

BENGALURU WIND POWER GENERATION

Carthikswami Sunil Thoniparambil
Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA

ABSTRACT

If we can predict the wind power at a particular time, we can reduce the use of fossil fuels to generate electricity and use the wind power in our main grids, and offset carbon release into the atmosphere. Potential energy from a wind turbine is directly proportional to the cube of the wind's speed, so an increase in average wind speed by even just 2 mph can produce a significantly larger amount of electricity [2]. *Also, wind speed is depended on the weather. So, we are also considering the weather parameters as input variables. Hence if we have the data for wind speed, we can accurately predict the wind power generation at that time. Wind power forecasting using SARIMAX, prediction using ML, ANN techniques, and finding the patterns in the data using K-means clustering.*

As we all know storage plays an important part in the success of renewable energy. However, the field of storage is currently lagging and is also very costly. Forecasting the wind power with the current inputs will help in understanding the contribution of wind power to the total power generated. This in turn can help determine what the other methods should produce to meet the economy's demands. The current AI/ML techniques provide the business with valuable information that they can use to make decisions about the future. A clear estimate of the power generated can be an indicator to reduce wastage and maintain operating costs. A time series model is used to forecast wind power and machine learning models to verify the predictions.

NOMENCLATURE

ANN	Artificial Neural Networks
ML	Machine Learning
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors
PACF	Partial Autocorrelation Function
ACF	Autocorrelation Function
KNN	K Nearest Neighbor

1. INTRODUCTION

The Bangalore Wind Power data [1] has 127,392 records and 8 variables, out of which 7 are numeric and 1 date-time format. The variables are -

➢ *T_amb* (degree Celsius)– Ambient Temperature.

➢ *DewPoint* (degree Celsius) – The temperature air needs to be cooled to achieve a relative humidity of 100%.

➢ *Humidity* (g/m3 in%) – The amount of water vapor present in the air.

➢ *Pressure* (Pa) – The atmospheric pressure.

➢ *Wind_Dir* (degrees)

➢ *Wind_Speed*(Km/hr)

➢ *Power_Generated*(KW)

➢ *Date* (Hourly)

2. MATERIALS AND METHODS

The methods used to achieve the desired results were

2.1 Data Preprocessing

The data consists of 3 null values that were removed using the *dropna()* function. Values with Z scores greater than 3 are considered outliers and are removed. The dataset consists of 786 outliers (T_amb = 266, Dew_Point = 22, Pressure = 263, Wind_Speed = 235). The positively skewed columns are transformed into symmetric by using log and cube root transformation. The negatively skewed data is transformed to symmetric using square transformation. Finally, all the variables were converted to desired datatypes.

2.2 Exploratory Data Analysis

The bar plot [4] shown in Fig.1 indicates that the total monthly power generated peaks during the month of July every year. The reason for this might be the increased wind speed during the summer months. Since July is monsoon season it can be attributed that the wind speeds are high and hence there is a peak in power generated. It is evident that the first quarter of every year gets way lesser power compared to the other seasons which can be a reason for experiencing a lot of power cuts during this time of the year.

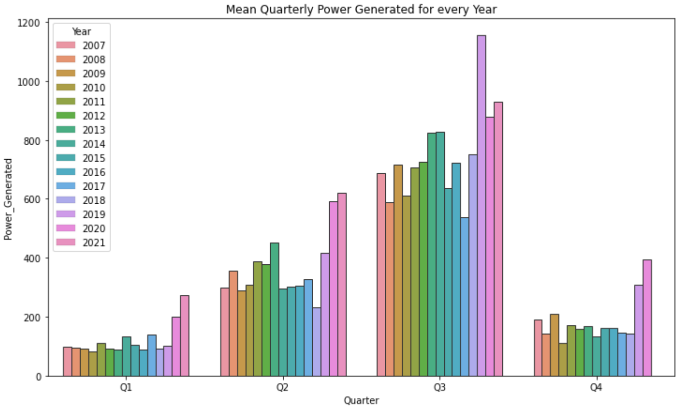

Fig. 1. Mean Quarterly Power Generated for every year

Cyclic variations of the time series are defined as the recurrent variations in the time series for a continuous span of time. Exponential smoothing is done to remove the cyclic characteristics. This is done to improve the accuracy of the models.

To understand the correlation between the variables, a heat map is used. From Fig.2 it is evident that wind speed is the most prominent factor for the power generated (correlation coefficient is 0.86). We can also interpret that the higher the pressure lower the power generated.

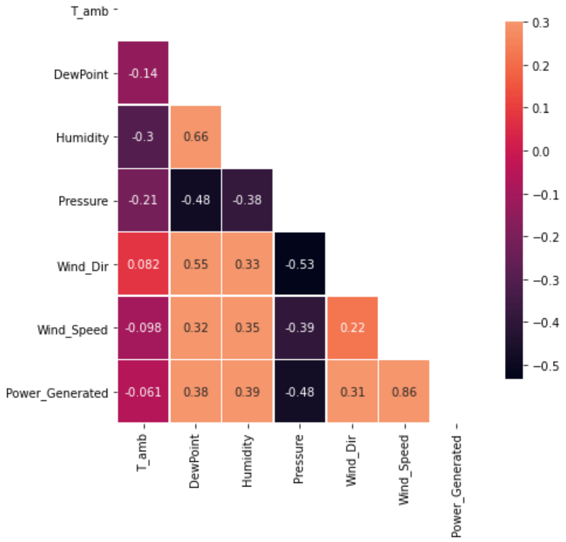

Fig. 2. Heat map

For further analysis considered the average monthly power generated after exponential smoothing. Visualized a scatter plot for wind speed vs power generated. For the years 2009,2013,2014,2019, the gradient was comparatively higher indicating that with a small increase in wind speed there was a larger increase in the power generated. I have used decomposition to visualize the components of the time series. The entire time-series data is composed to trend, seasonal, and residuals. We can observe from Fig.3 that there is an increasing trend in the time-series data after the year 2018. There is a peak observed in the power generated in the mid of every year, indicating the seasonality. The small Magnitude of the residuals indicates decomposition is done well.

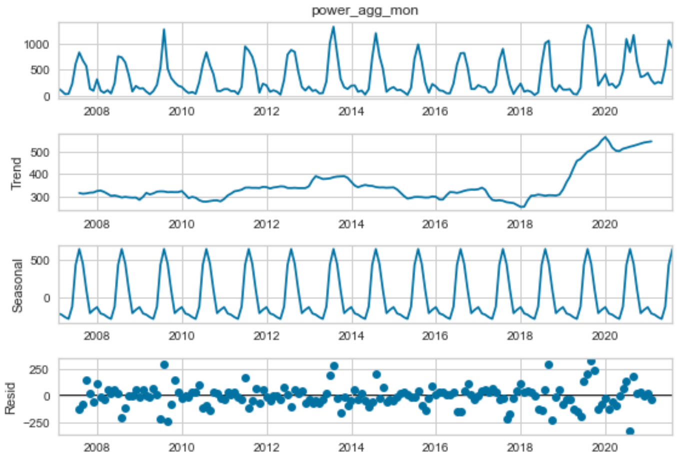

Fig. 3. Decomposition of power distribution

2.3 Model Building

2.3.1 Statistical Models

Since there is an increasing trend from Fig.3, to check whether the time series is stationary a stationarity check was done using the Dickey-Fuller test [7]. As the time series was not stationary, Differencing was done to convert the time series to stationary. The order of the non-seasonal component O (p,d,q) was found from PACF and ACF plots of the differenced data. Similarly, the season order S (P,D,Q,m) were found from PACF and ACF of non-differenced data. These as inputs are fit to SARIMAX [5] model and the power generated for the years 2021 to 2024 is shown in Fig.4

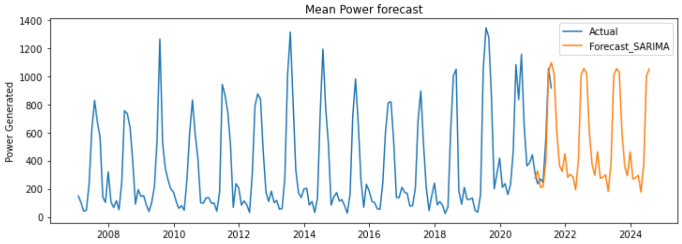

Fig. 4. Mean Power Forecast

2.3.2 Classical ML Models

2.3.2.1 Unsupervised Learning

K-means clustering [8] was performed to find the patterns in the data and the optimal number of clusters was obtained using the elbow method. From the cluster analysis, the centroids of wind speed and pressure were well separated, hence these two variables were considered for naming the clusters. The results are shown in Fig.5 the clusters can be named as, Cluster 1 – High pressure and Average wind speed, Cluster 2 - Low pressure and High windspeed, Cluster 3 – Average pressure and Low wind speed.

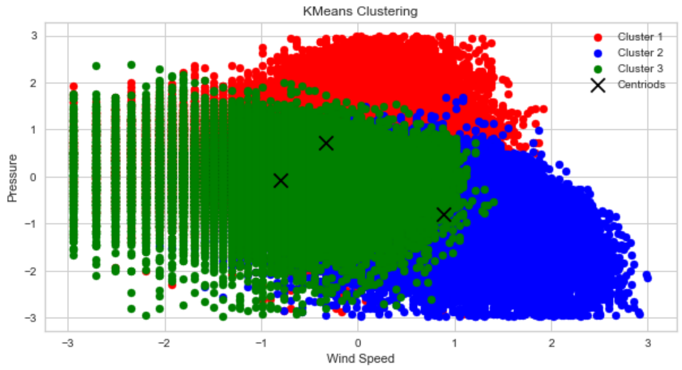

Fig. 5. K-means clustering

2.3.2.2 Supervised Learning

Feature selection is performed using an exhaustive search[3] and the top five features are used as the predictors which are then split into 75% training and 25% test data.

Multiple regression models are used to predict the *PowerGenerated.* The models used were LinearRegression, KNN regressor, Random Forest Regressor, and Xgboost Regressor. The Random Forest Regressor was considered as it gave better performance compared to other models and the residual plot is shown in Fig.6. The residual error follows the gaussian distribution. The variability of the response variable for a given set of predictors is the same across all values of predictors and hence the homoscedasticity is not violated.

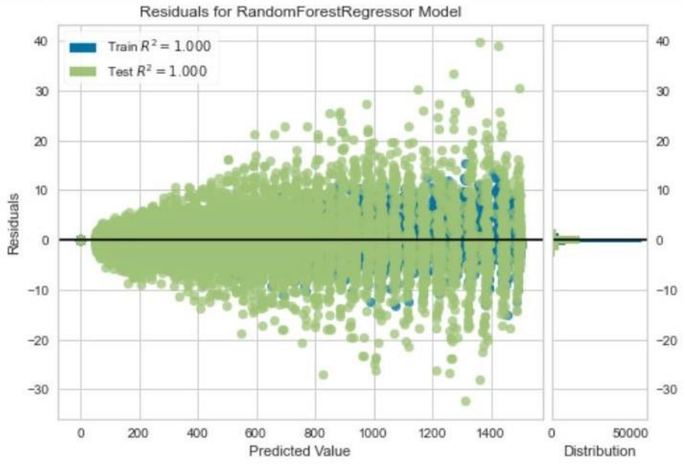

Fig. 6. Residual Plot

2.3.3 Neural Networks

Deep neural networks [6] were constructed with three hidden layers with ten nodes in the first layer and five in the rest two with Leaky ReLU as an activation function. The model is trained with a Stochastic gradient descent optimizer with a learning rate of 0.01 and momentum of 0.9 for 50 epochs, with a batch size of 54. The training and test loss can be seen in Fig.7.

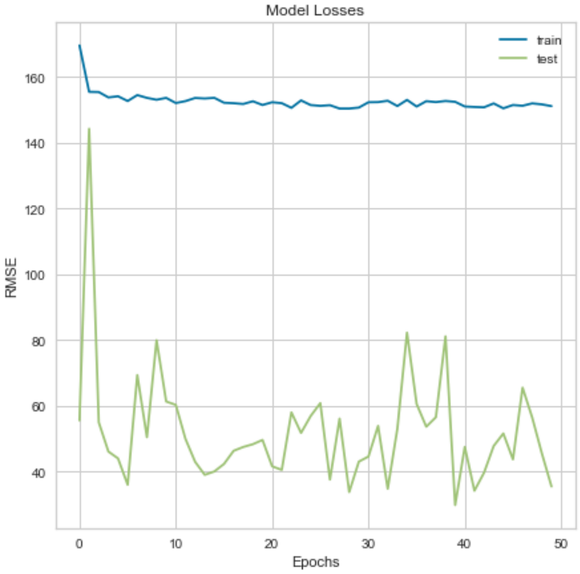

Fig. 7. Model Losses

3. RESULTS

Model	RMSE	R-Square
SARIMAX	102.1	0.9
Linear Regression	214.172	0.783
KNN	31.41	0.95
Random Forest Regressor	2.989	1
Xgboost Regressor	2.84	1
Deep Learning	35.404	0.9

After evaluating the performance of the model, it was found that machine learning algorithms like random forest regressor and Xgboost regressor gave impressive results. Deep neural network gave decent results, but it also depends on the hyperparameters like activation function, hidden layer, learning rate, optimizer, batch size, and epochs.

4. CONCLUSION

The wind power forecast model can be deployed to analyze the use cases like supply chain and demand forecasting. This can also be used to analyze resource allocation and cash flow management. In the end, it all depends upon how the data is structured to fit into a particular model. Finally, the choice of the model depends upon how efficiently the results suit a particular application. If we have high-dimensional data with many records it is better to go for deep learning models than machine learning models.

ACKNOWLEDGEMENTS

The author of this paper acknowledges the support and guidance of Dr. Jon Wallace, Affiliate Professor at the Department of Mechanical and Industrial Engineering,

Northeastern University.

REFERENCES

[1] Bangalore wind power generation dataset from Kaggle: Bangaluru wind power generation | Kaggle

[2] Wind power information: INDIAN WIND TURBINE (indianwindpower.com)

[3]Exhaustive search using Select K-best sklearn.feature_selection.SelectKBest — scikit-learn 1.0.2 documentation

[4] Stacked bar plot from seaborn - seaborn.barplot — seaborn 0.11.2 documentation (pydata.org)

[5] SARIMAX - statsmodels.tsa.statespace.sarimax.SARIMAX — statsmodels

[6] Tensor flow tutorial - TensorFlow 2 quickstart for beginners | TensorFlow Core

[7] Stationarity check - Augmented Dickey-Fuller Test in Python (With Example) (statology.org)

[8] K-means clustering - K-means: A Complete Introduction. K-means is an unsupervised clustering… | by Alan Jeffares | Towards Data Science