Int. J. Econ. Environ. Geol. Vol. 12 (4) 69-72, 2021
Journal home page: www.econ-environ-geol.org

Open Access
ISSN: 2223-957X

SHORT COMMUNICATION

Using XGBoost Model with Feature Selection Techniques for Wind Speed Forecasting

Hamza Hanif[1], Ahmer Shaheem Tahir[2], Rimsha Shaikh[3], and Dania Anjum[2]

[1]Department of Physics, Simon Fraser University, Canada
[2]Department of Physics, University of Karachi
[3]Department of Applied Physics, NED University

*Email: amjad.eco@kiu.edu.pk

Introduction

Renewable Energy Sources have a lot of importance in today's world to produce an electrical output which explains the main reasons that every government and policy maker now a days prefer Renewable Energy in the wake of global warming and limited availability of fossil fuels (Twidell and Weir, 2021). The Renewable Energy Sources are hazardless, pollution free, eco-friendly, freely available in nature in vast quantities and most importantly, they give a chance to create a carbon-free environment.

Wind Energy is the most eminent source of renewable energy. It is a significant way to produce electricity with the increased demand for electricity generation and has the potential to achieve a sustainable energy supply. It also constitutes a keystone component for micro-grids in a way to smart grid infrastructure [Lei et al., 2009]. The majority of the world's electricity requirement can be realistically provided by wind energy, which may not only be consumed for domestic purposes but also for services and trades at a large scale.. However, stochastic, and intermittent wind power generation poses several challenges to the large-scale penetration of wind power. These uncertainties can put the system reliability and power quality at risk with the increasing penetration of wind power thus, the main grid integration issues; such as balanced management and reserve capacities can come into question. For reducing energy balancing and power generation scheduling, the help of wind speed and power generation forecasts are the prerequisites. The forecasting of Wind speed has a direct effect on power forecasting as the speed of the wind is directly proportional to the cube root of wind power (Gipe, 2003). Therefore, wind power forecasting plays a key role in developing and planning the Wind Power Operating System. Furthermore, the forecasts has vital role in keeping the costs competitive by reducing the need for wind curtailments and thereby, increasing revenue in electricity market operations. However, the random and unstable characteristics of the wind make it considerably difficult to forecast the wind speed and power accurately. Hence, extensive efforts have been devoted to the development and improvement of wind speed and power forecasting approaches by numerous energy and environment-related research centers and universities. The better and more accurately predicted wind power decreases the chances of the worsen outcome in wind power operating systems (Ahmed et al., 2017; Lei et al., 2009)].

From the last decade, machine learning has made significant improvements and showed fruitful results in different fields from classification tasks to automation. In the past, various machine learning methods have been implemented for wind speed forecasting such as the works mentioned in (Khosravi et al., 2018; Zheng et al., 2015). The objective of this work is to employ the new XGBoost model with two feature selection methods for a short-term Wind Speed forecasting. The XGBoost model is relatively new machine learning and has gained significant importance due to its high accuracy output and fast processing time while still being complex (Chen and Guestrin, 2016).

Materials and Methods

In this analysis, XGBoost is used as machine learning to train the model with two feature selection methods, the PCA and the feature importance. Afterward, the trained model is used on the test data sets for both cases to find out the features selection technique that performs the best.

Data

The data been used in this literature is taken from energydata.info via the world bank (World Bank) which consists of a repository of measurements from 12 wind masts in Pakistan. In this study, only the data of one city, Bahawalpur, have been utilized. The data consists of various weather parameters, within 10 minutes intervals, from April 2016 to August 2018. In this analysis, only the last 90 days of data have been selected which comprises 17712 measurements with 7 features after pre-processing. The features are Wind Speed (m/s), wind direction, humidity, temperature (centigrade), pressure (hPa), day of the year, time of the day, and length of a day. The Length of a Day has been obtained by utilizing sunrise and sunset time, which are obtained from NOAA Solar Calculator (NOAA) for each day. The subtraction of sunset time from sunrise time leads to the length of a day.

Figure 1 illustrates the plot of the temperature, pressure, and humidity. The color bars represent the magnitude of the

measures, bright color infers lower magnitude and darker colors indicate higher magnitude. The first column in Figure1 gives information about the hourly average values and the second column provides the visualization for the monthly average values. The first row represents the wind speed, which is the attribute to be forecast in the literature. The information from the subplots allows us to identify the correlation among the attributes of the data set.

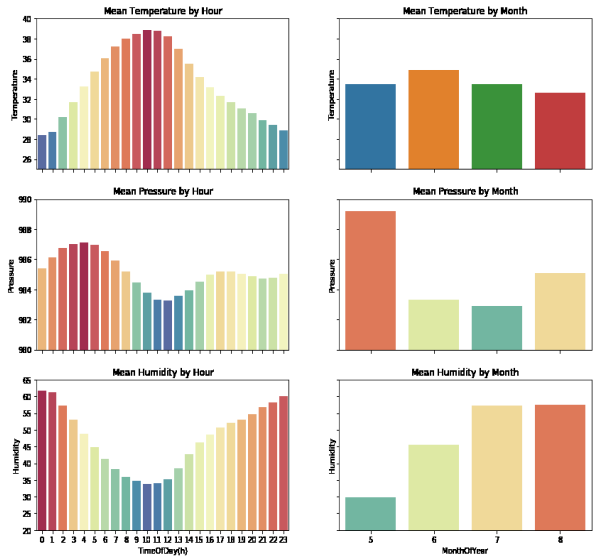

Fig. 1: Temperature, pressure, and humidity by hours and monthly means.

XGBOOST

XGBoost is the machine learning algorithm based upon Gradient Boosted decision trees (GBDT) (Hastie et al., 2004). The XGBoost performs better than other machine learning models due to the introduction of Taylor expansion for the cost function by incorporating the second derivative to provide the results with more accuracy and it avoids voiding overfitting by using shrinkage and regularisation techniques (Chen and Guestrin, 2016).

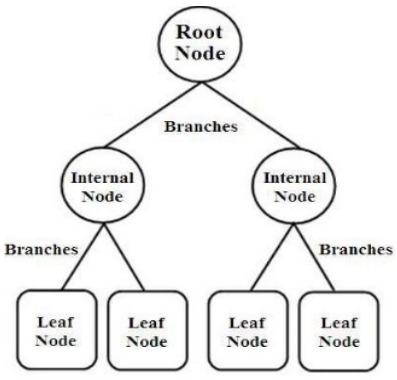

Fig. *2:* The observed and predicted data using XGBoost with features from PCA

A Decision tree (DT) has similar characteristics to the real tree as shown in figure2. The DT algorithms begin from the root node and then move towards internal nodes corresponding to an attribute and finally lead to leaf nodes that correspond to a class label. The GDBT is an ensemble learning method that utilizes a chain of decision trees, from which the DT gathers knowledge from previous DT and influences the next tree to refine the model and built a strong learner (Friedman, 2001).

In XGBoost model several parameters can be tuned to get the best-performing model. Parameter tuning is an essential step to prevent over-fitting in the model, which is caused when the model tries to label the random fluctuations or noises as valuable facts. The three parameters (maximum depth of trees, the step size, and the number of trees) are selected in our model. The parameters that have been employed in our model are.

- N_trees: represents the number of trees that are employed in the model.
- Eta: analogous to the learning rate, defines how quickly the model learns
- Max_depth: relates to the maximum number of splits, the more increase in the maximum depth can cause the over-fitting in the model.
- Subsample: it is used to reduce the variance of the model and prevent over-fitting.

Feature Selection

In this research, two approaches have been implemented for feature selection: feature importance and PCA. The PCA is a procedure for reducing the dimensionality of the variable space by representing it with a few orthogonal (uncorrelated) variables that capture most of its variability (Garreta and Moncecchi, 2013). In this literature, the PCA is utilized for the moderate dimensional features to get rid of irrelevant feature noises hence providing improvement in forecasting accuracy. By incorporating the PCA method, the seven features have been reduced to 3 components. Table 1 lists the first sample obtained from PCA.

Table 1: Samples of principal components with PCA technique

Component 1	Component 2	Component 3
-428.999663	12.418122	-1.049804

Feature importance is one of the random forest tree approaches. This method uses a backpropagation procedure to repetitively eliminate features being too small to consider, based upon the r^2 score. (Brownlee, 2016).

Table 2 lists the r^2 scores at each iteration by utilizing the feature importance method with the constraint of having a minimum of three features. The r^2 scores increases as the number of features decreases with the last iterative having the highest value of r^2. Therefore, features from iterative 3 are selected with the features selection method.

Table 2:. The r^2 score at each iteration using the feature importance method

Iteration	Features	r^2 Score
0	Temperature, Pressure, Wind Direction (Degrees), Day Length (s), Day of Year, Time of day (s)	0.865
1	Temperature, Pressure, Wind Direction (Degrees), Day of Year, Time of Day (s)	0.857
2	Pressure, Wind Direction (Degrees), Day of Year, Time of Day(s)	0.853
3	Pressure, Day of Year, Time of Day (s)	0.884

The data with the selected features from the last section were split into two categories: 80% of the data was divided into the training set and the remaining 20% of data are divided into the testing set. The root mean square error (RMSE) and r^2 score are used to calculate the accuracy, and efficiency respectively. The chosen values for the parameters for both techniques, feature importance, and PCA, in XGBoost are as follows.

Maximum depth of trees: 13, the step size: 0.05, the subsample': 0.7, and the number of trees: 70

Result and Discussion

Two cases have been considered to justify the feature selection techniques that work best for wind speed forecasting with XGBoost model. Table 3 lists the root means square errors and the r^2 scores for each scenario. From table 3, the feature importance method has the best performance with the least RMSE (0.64) and the highest r^2 value (0.89). PCA also showed promising results, but it didn't excel due to the following reasons.

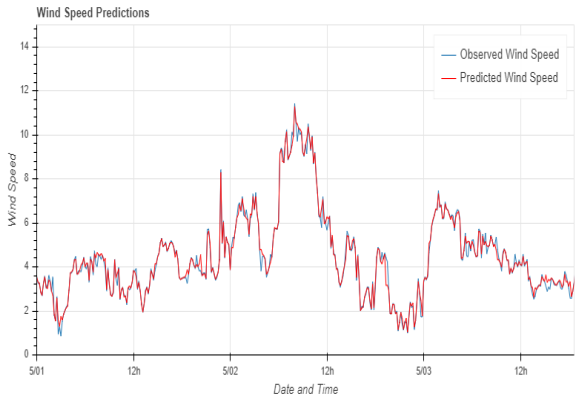

Fig. 3: The observed and predicted data using XGBoost with features from feature importance

In PCA the values from all the original variables are transformed into the lower-dimensional space.

Only linear relationships are considered in the PCA method.

PCA does not consider the potential multivariate nature of the data structure.

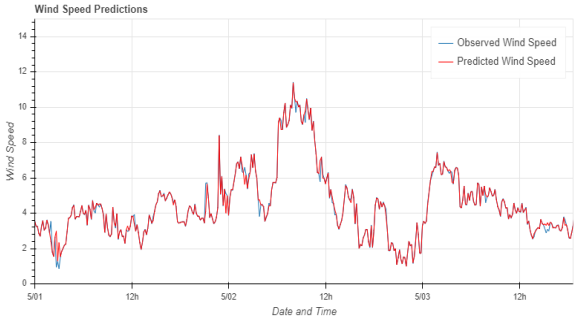

Fig. 4: The observed and predicted data using Xgboost with features from PCA

Figures 3 and 4 visualize the observed and predicted wind speed using the XGBoost model with features from the feature importance and PCA methods respectively.

Table 3: Root mean square errors and the r^2 scores for both feature selection technique

	RMSE	r^2 Score
Feature Importance	0.64	0.89
PCA	0.79	0.

Conclusion

The weather forecasting model can help policymakers to locate the best position for the setup wind turbine and the optimal designing of the instrumentation that would be beneficial for the inhabitants of these cities. It also helps in providing the groundwork for the solar panels as well and the temperature modeling of these cities.

References

Ahmed, S., Khalid, M., and Akram, U. (2017). A method for short-term wind speed time series forecasting using support vector machine regression model. *In 2017 6th International Conference on Clean Electrical Power (ICCEP), pages 190–195*.

Brownlee, J. (2016). *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. Machine Learning Mastery.

Chen, T., and Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29, 1189-1232.*

Garreta, R., and Moncecchi, G. (2013). *Learning Scikit-Learn: Machine Learning in Python*. Packt Publishing.

Gipe, P. (2003). *Wind power for home, farm, and business*. Chelsea Green Publishing Company.

Khosravi, A., Machado, L., and Nunes, R. (2018). Time-series prediction of wind speed using machine learning algorithms: A case study Osorio wind farm, Brazil. *Applied Energy, 224, 550–566.*

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., and Yan, Z. (2009). A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13(4):915–920.

NOAA *https://www.esrl.noaa.gov/gmd/grad/solcalc/*.

Twidell, J. and Weir, A. D. (2021). *Renewable energy resources*. Routledge.

World Bank Pakistan - wind measurement data. *https://energydata.info/dataset/pakistan-wind-measurement-data*.

Zheng, H., and Wu, Y. (2019). A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting. *Applied Sciences, 9(15).*