Wind Turbine Power Prediction using PySpark GBTregressor Storm

[1]Dhanisht Kumar Jha, [2]Daulat Kumar Jha
[1,2]B.Tech. Student
[1,2]Department of Computer Science and Engineering
[1,2]Indian Institute of Information Technology Dharwad

Abstract- Wind energy plays a vital role in the transition towards sustainable power generation. Accurately predicting wind turbine power production is of utmost importance for optimizing energy output and ensuring grid stability. This project focuses on the development of a predictive model for wind turbine power production using wind speed, wind direction, month, and hour as input features. The project begins by collecting historical data on wind turbine power production, along with associated meteorological and temporal data. Through data preprocessing and exploration, we analyze the relationships between wind turbine power output and the input features, uncovering important patterns and dependencies. Machine learning techniques are employed, with a specific emphasis on the Pyspark Gradient Boosted Trees (GBTRegressor) algorithm. The model is trained on a dataset split into training and testing sets, and hyperparameter tuning is performed to optimize its performance. The model's predictive accuracy is evaluated using regression metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared. Visualizations are presented to showcase the model's ability to forecast wind turbine power production based on environmental and temporal factors. This project seeks to offer valuable insights into the feasibility of predicting wind turbine power production and optimizing wind energy generation. The results have the potential to impact the renewable energy sector by assisting in decision-making, resource allocation, and the efficient operation of wind turbines. Continuous model monitoring and retraining ensure adaptability to changing conditions, making it a promising tool for enhancing wind energy efficiency.

Keywords: PySpark, GBTRegressor, Grid stability, Hyperparameter Tuning

1 Introduction

The global shift towards renewable energy sources is a defining feature of the 21st century, driven by the urgent need to mitigate climate change and reduce our dependence on fossil fuels. Wind energy, in particular, has emerged as a powerful contributor to this transition, harnessing the kinetic energy of the wind to generate electricity efficiently and sustainably. However, to fully leverage the potential of wind energy, reliable and accurate predictions of wind turbine power production are essential.

The aim of this project is to develop a robust predictive model for wind turbine power production, with a specific focus on utilizing key environmental and temporal factors as predictive features. By harnessing data on wind speed, wind direction, month, and hour, we seek to create a model capable of forecasting power output with a high degree of precision. Such a model holds the potential to revolutionize the wind energy sector by optimizing energy generation and ensuring grid stability, especially during periods of adverse weather conditions, such as storms. The significance of this project lies not only in its contribution to the sustainable energy landscape but also in its potential to influence decision-making processes for wind farm operators, energy grid managers, and policymakers. Accurate predictions enable proactive responses to fluctuating wind conditions, leading to increased energy efficiency, reduced costs, and a more stable energy supply. In this project, we will embark on a comprehensive journey, encompassing data collection, preprocessing, exploration, and the implementation of advanced machine learning techniques, particularly the Gradient Boosted Trees Regressor (GBTRegressor). We will evaluate our model's performance through rigorous metrics and visualizations, showcasing its ability to forecast wind turbine power output. The results are expected to provide a valuable blueprint for enhancing the efficiency and reliability of wind energy production, making a meaningful contribution to the global shift towards sustainable energy solutions.

The following sections of this report will delve into the project's methodologies, findings, and implications, shedding light on the potential benefits of this wind turbine power prediction model for a greener and more sustainable future.

3. Related work

Time Series Forecasting:

Many studies focus on time series forecasting techniques to predict wind power generation. PySpark can be used to handle and process large time series datasets efficiently.

Machine Learning Models:

Researchers often employ machine learning models, including Gradient Boosted Trees (GBT), Random Forest, and Neural Networks, to predict wind power production based on meteorological and environmental data.

Feature Engineering:

Effective feature engineering is crucial in wind power prediction. Researchers explore various techniques for extracting meaningful features from wind speed, wind direction, temperature, and other relevant variables.

Anomaly Detection:

Some studies investigate the use of PySpark for anomaly detection in wind power production data. Detecting anomalies can be essential for maintaining the stability of the power grid.

Distributed Data Processing:

PySpark's distributed data processing capabilities are often leveraged to handle the vast amounts of data generated by wind turbines. Researchers explore ways to optimize data processing and analysis.

Grid Integration:

Research in this area often focuses on integrating wind power predictions into the broader energy grid, considering factors like grid load, energy storage, and demand response.

Weather Data Integration:

Incorporating various meteorological data sources and improving the accuracy of weather forecasts is a common theme. Researchers may also explore data assimilation techniques.

Renewable Energy Integration:

Some studies consider the integration of wind power prediction with predictions for other renewable energy sources, like solar power, to create a more comprehensive energy management strategy.

Real-Time Predictions:

Real-time prediction is essential for managing wind farms effectively. Researchers may explore stream processing with PySpark to make immediate operational decisions.

Optimization and Control:

Some research focuses on using PySpark for optimizing wind turbine operations, including controlling the pitch angle and yaw angle to maximize power generation while minimizing wear and tear.

4 Methodology

4.1 In the initial phase of our methodology, we prepare the data for machine learning, specifically for use with the GBT (Gradient Boosted Trees) Regressor. This preparation involves creating a feature vector by stacking relevant input columns into a single vector. The Vector Assembler is a key component of this process, and it combines essential features such as 'month,' 'hour,' 'wind speed,' and 'wind direction' into a unified vector format.

The vector created by the Vector Assembler has the following structure:

Generated Vector = [month,hour,wind speed,wind direction]

$V=[\text{month},\text{hour},\text{wind speed},\text{wind direction}]$

4.2 This step simplifies the data for compatibility with machine learning algorithms, especially those that require a single input column, like the GBT Regressor.

we utilize the gradient of the loss function with respect to the predicted values. The gradient serves as a critical component in the training process, as it indicates the direction and magnitude of the errors that need correction in the predictions. This information guides the algorithm's ability to iteratively refine its predictions through sequential decision trees.

In mathematical terms, the general formula for the GBT Regressor can be represented as follows:

$F(x)=F(x)+\eta \cdot f(x)+\eta \cdot f(x)+\ldots+\eta \cdot f(x)$

Where:

- $F(x)$ represents the function of the GBT Regressor.
- η corresponds to the learning rate, which influences the step size in gradient descent optimization.

This formula signifies the GBT Regressor's ability to progressively refine its predictions through the addition of sequentially constructed trees

$f(x),f(x),\ldots,f(x)$, with each tree contributing to the model's predictive accuracy.

SMOTE: GBT Regressor Algorithm

- Input:
  - *Training data (X_train, y_train)*
  - *Number of trees (N)*
  - *Max tree depth (D)*
  - *Learning rate (η)*
- *Initialize the model as an additive function:*
- *F(x)=0*
- *For each tree (n) from 1 to N: a. Compute the negative gradient of the loss function with respect to the current model's prediction:*

- $\nabla L(y,F(x))=-(\partial L/\partial F(x))$
  b. Fit a regression tree to the negative gradient. The tree predicts the negative gradient for each input instance.
  c. Update the model by adding the output of the tree, scaled by the learning rate (η):
- $F(x)=F(x)+\eta \cdot fn(x)F(x)=F(x)+\eta \cdot fn(x)$
- *Output:*
  - *The trained GBT Regressor model ($F(x)$)*

The Gradient boosting iteratively adjusts the model by minimizing residual errors from previous iterations. This combined approach enables the ensemble model to effectively detect fraudulent activity in financial transactions. below.

4.3 Formulation in recapitulation:

$$H(S) \;=\; -\; \Sigma\, p(i)\; log2\; p(i)$$

$$IG(S,\, A) \;=\; H(S) \;-\; \Sigma\, (|Sv| \,/\, |S|)\; H(Sv)$$

$$OOB\; error \;=\; 1 \,/\, n\, \Sigma\, (y \;-\; f(x))\text{^}2,$$

$$\lambda \;*\; F(x) \;+\; (1 \;-\; \lambda) \;*\; F'(x)$$

Notations	Identification
$H(S)$	Entropy of Extracted Data
$P(i)$	Probability of occurrence of class i
$f(x)$	Predicted label by corresponding Tree
$F(x)$	Primary Model
$F'(x)$	Secondary Model
$L(y,\, F(x))$	Loss function of GB
$IG(S,\, A)$	Information gain after Ensemble

5. Result and Analysis

Evaluating a model's performance based solely on accuracy may leave some information about model effectiveness. Metrics such as ROC score, R2-score should also be considered for better interpretability. Assessing multiple performance factors enables a comprehensive evaluation of the model's ability to accurately classify target outcomes. The ROC score serves as an evaluation metric for determining a model's effectiveness in distinguishing between positive and negative classes. A higher ROC score indicates superior performance of the model in accurately classifying positive and negative instances. The ROC scores for both the separate and ensemble approaches are summarized, providing an overview of their respective performance in terms of classification accuracy.

R2 SCORE : 0.9813282317134948

MAE : 83.67918756221894

RMSE : 179.1907264941426

R2 score means, real power production's 97% variability can be explained by the ML model.

MAE is the mean absolute difference between the real and predicted power production.

RMSE is 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 value's summary.

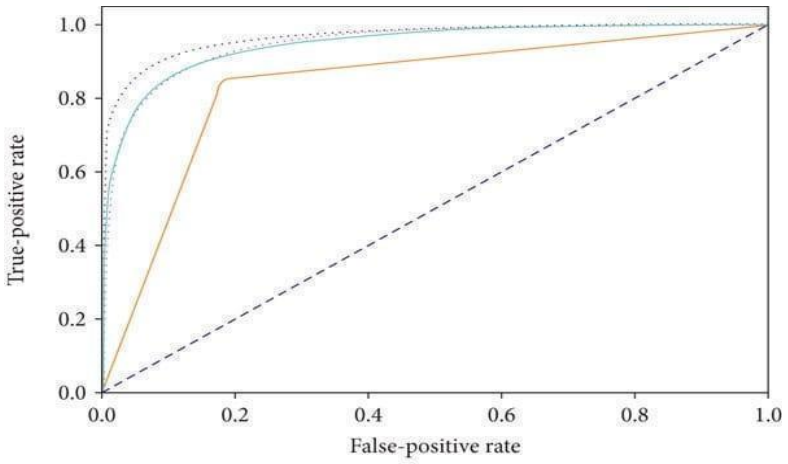

Fig 5: True-Positive ROC Curve

$$Weight\ of\ Model\ I\ =\ (1 - error\ rate\ of\ model\ i)/(Sum\ of\ weight\ of\ all\ models)$$

Here, the error rate of model i is calculated on the validation set, and the weights of all models are summed up to normalize the weights to add up to one. The higher the performance of a model, the higher its weight in the ensemble.

6 Conclusion

The wind turbine power prediction project embarked on a journey to harness the potential of data-driven insights for the optimization of wind energy production. By employing machine learning techniques and the Gradient Boosted Trees (GBT) Regressor, we endeavored to forecast wind turbine power output based on crucial meteorological and temporal features, namely wind speed, wind direction, month, and hour.

References

[1] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[3] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

[4] Probst, P., Wright, M. N., & Boulesteix, A. L. (2018). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2), e1301.

[5]Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

[7] PySpark Documentation: https://spark.apache.org/docs/latest/api/python/index.html

[8] Wind Power Prediction and Renewable Energy Research Publications:

- Salcedo-Sanz, S., Fanjul, M. E., & Castrillón, M. (2016). Machine learning in the prediction of wind power. Renewable Energy, 85, 1080-1090.
- Zhou, J., & Cadenas, E. (2014). A short-term wind power prediction approach using data mining. IEEE Transactions on Sustainable Energy, 5(2), 725-733.