

15 : 01 : 57 : 06
DAY HRS MIN SEC

A Fine Windy Day: HackerEarth Machine Learning challenge

LIVE

Apr 26, 2021, 09:30 PM IST - May 26, 2021, 09:30 PM IST

INSTRUCTIONS

PROBLEMS

SUBMISSIONS

LEADERBOARD

ANALYTICS

JUDGE

[← Problems](#) / Predict the power (KW/h) produced from the windmills

Predict the power (KW/h) produced from the windmills

Max. score: 100

Moving from traditional energy plans powered by fossils fuels to unlimited renewable energy subscriptions allows for instant access to clean energy without heavy investment in infrastructure like solar panels, for example.

One clean energy source that has been gaining popularity around the world is wind turbines. Turbines are massive structures that are strategically placed in perpetually windy places to generate the most energy. Wind energy is generated when the power of the atmosphere's airflow is harnessed to create electricity. Wind turbines do this by capturing the kinetic energy of the wind. Factors such as temperature, wind direction, turbine status, weather, blade length, etc. influence the amount of power generated.

Task

Predict the power that is generated (in KW/h) based on the various features provided in the dataset.

Data description

The dataset folder contains the following files:

- train.csv: 28200 x 22
- test.csv: 12086 x 21
- sample_submission.csv: 5 x 3

The dataset contains the following columns:

Column name	Description
tracking_id	Represents a unique identification number of a windmill
datetime	Represents the date and time of a record

?

wind_speed(m/s)	Represents the speed of wind (in meter per second)
atmospheric_temperature(°C)	Represents the temperature (in degree Celcius) of a town or village that the windmill is present in
shaft_temperature(°C)	Represents the temperature of the shaft (in degree Celcius)
blades_angle(°)	Represents the angle of the blades of a wind turbine (in degrees)
gearbox_temperature(°C)	Represents the temperature of a gearbox (in degree Celcius)
engine_temperature(°C)	Represents the temperature of an engine (in degree Celcius)
motor_torque(N-m)	Represents the torque of a motor (in Newton meter)
generator_temperature(°C)	Represents the temperature of a generator (in degree Celcius)
atmospheric_pressure(Pascal)	Represents the atmospheric pressure (in Pascals) in that area
area_temperature(°C)	Represents the temperature (in degree Celcius) of the area within a 100 m radius of the windmill
windmill_body_temperature(°C)	Represents the temperature of the body of a windmill (in degree Celcius)
wind_direction(°)	Represents the direction of the wind (in degrees)
resistance(ohm)	Represents the resistance against the wind
rotor_torque(N-m)	Represents the torque of a rotor (in Newton meter)
turbine_status	Represents the status of the turbine (masked)
cloud_level	Represents the following levels of the cloud in the sky on a particular day: <ul style="list-style-type: none"> Extremely low Low Medium
blade_length(m)	Represents the length of the blades of a windmill (in meter)
blade_breadth(m)	Represents the breadth of the blades of a windmill (in meter)
windmill_height(m)	Represents the height of the blades of a windmill (in meter)
windmill_generated_power(kW/h)	Represents the power generated (in Kilo Watt per hour)

Evaluation metric

?

```
score = max(0 ,100*metrics.r2_score(actual, predicted))
```

Result submission guidelines

- The **indexes** are **tracking_id** and **datetime**.
- The **target** is the **windmill_generated_power(kW/h)** column.
- The submission file must be submitted in **.csv** format only.
- The size of this submission file must be **12086 x 3**.

Note: Ensure that your submission file contains the following:

- Correct index values as per the test file
- Correct names of columns as provided in the **sample_submission.csv** file

[Download dataset](#)

Upload Prediction File

Please upload the prediction file in the format as stated in the problem.

[Choose File](#) No file chosen

[Submit & Evaluate](#)

Upload Source Files

You need to submit a zip or tar archive consisting of a text file explaining your approach, details about feature engineering, tools you used and the relevant source files.

[Choose File](#) No file chosen

[Upload](#)

COMMENTS (33) [↻](#)

SORT BY: [Relevance ▼](#)



Join Discussion...

[Cancel](#)

[Post](#)



Sanket Sharma [✎](#) Edited 12 days ago

?

I don't know from where you guys get this data. If it is manually created, I must say that person knows nothing about wind turbines or if it comes from an Energy company that company should plan to shut them down straight away.

For a wind turbine to generate power there are two wind speeds defined by the turbine manufacturer. These speeds are cut-in and cut-out wind speeds.

The cut-in wind speed is the wind speed when the turbine starts generating power and the cut-out wind speed is the safety threshold defined by the manufacturer to prevent the overspeeding of the wind turbines. The cut-in wind speed starts from 3-4 m/s and the cut-out speed ranges from 22 m/s to 30 m/s. These numbers may vary as per the hub height and the capacity of the turbine.

Just in the second row of the training data wind speed is 241.832734 m/s. If that is the case the turbine is definitely broken. This applies to the so-called test data in the third row where wind speed is 347.152090 m/s.

Hats off to your leaderboard and the practitioners out there. Who only knows increasing the scores and achieving the rank. I am not against their strategy but at least think before you do.

If this data is shown to any leading guy of this domain he/she will definitely sue you for this.

▲ 4 votes ● Reply ● Permalink



Sanket Sharma 12 days ago

Hoping for a positive and informative response from your side (HackerEarth).

▲ 0 votes ● Reply ● Permalink



Shreyash Gupta 11 days ago

Tbh, a bit of work and data tricks and exploration is required. Try it out. Check the kaggle notebooks for Titanic problem, you will know it.

▲ 1 vote ● Reply ● Permalink



Sanket Sharma Edited 11 days ago

Much appreciate your thought process. Kindly read out my latest comment then infer.
It's good that you replied.

▲ 0 votes ● Reply ● Permalink



Vignesh V 11 days ago

That happens , perform some data cleaning and get insights from outside from this site which helps you to get top on leaderboards

▲ 1 vote ● Reply ● Permalink



Sanket Sharma 11 days ago

Much appreciate your thought process. Kindly read out my latest comment then infer.
It's good that you replied.

▲ 0 votes ● Reply ● Permalink



Sanket Sharma 11 days ago

Guys, the reason for giving such a brief description is because I worked on wind turbine data for an energy company. If such data is received from a customer while working on the project, the work will never start because of such anomalies in the values. End of the day machine learning should add value to the business.

It's not about models it's about understanding the customer problem and being data-centric rather than model-centric. The same sentence is already said by Prof. Andrew Ng in one of the recent conferences. The ML practitioners need to be data-centric rather than model-centric.

My conclusion is if you work on such data assuming that it is of wind turbines then best of luck with your dreams.

One interesting fact about this data. Just sort the data by datetime column you will observe that the data collected is every 10 mins. This is because the live data from SCADA comes at 1 sec or 10 mins granularity. Whoever constructed this data the person failed to mimic the wind turbine data.

If these facts given by me are not ensuring that I have worked on the wind turbine data, God knows what else proof I can give to you guys.

?

By the way best of luck with importing and trying different models. A software engineer can also do the same.

Keep in mind there is a difference between a software engineer and a machine learning guy

But I will definitely write a blog on this thing either on KDnuggets or Towards Data Science. I'll definitely share a link with you guys.

Till then enjoy importing, installing different dependencies, and trying fancy models.

▲ 3 votes ● Reply ● Permalink



Tuhin Bhattacharjee 11 days ago

I am looking forward to your blog. Happy Learning till then :D

▲ 0 votes ● Reply ● Permalink



Sanket Sharma 10 days ago

Hi Tuhin,

Once I got the bandwidth I'll definitely share it with you.

Meanwhile, let me know your thoughts on the following blog:

<https://sanketsharma2196.medium.com/pandemic-and-the-step-count-data-storytelling-cc1a51ef22b>

▲ 0 votes ● Reply ● Permalink



Tuhin Bhattacharjee 10 days ago

Very nicely written and the plotly visuals are very well portrayed. Overall an amazing and fun read. Keep at it, my friend! The article was really good and also keep walking and stay safe :)

▲ 0 votes ● Reply ● Permalink



Sanket Sharma Edited 10 days ago

Thanks, Tuhin.

You also take care.

Once I pen down the blog on this data and the so-called hackathons organized by novices, I'll definitely forward you the link.

I myself don't know how many best practices I have gone through while portraying those plots.

Thanks, man :)

▲ 0 votes ● Reply ● Permalink



Chiranjeevi karthik 10 days ago

There should be a PDF given with every competition explaining the domain knowledge required to tackle the problem meaningfully.

Otherwise you can throw a Xgboost and difference between top 3 and yours will be in range of 10^{-6} .

▲ 2 votes ● Reply ● Permalink



Sankalp Chourasia 13 days ago

Dataset uploaded on Kaggle:

<https://www.kaggle.com/synergystud/a-fine-windy-day-hackerearth-ml-challenge>

▲ 3 votes ● Reply ● Permalink



Adarsh Wase 8 days ago

Totally wrong dataset !

▲ 3 votes ● Reply ● Permalink



Siddhant Bahuguna 6 days ago

How is a negative atmospheric pressure sensible?

▲ 1 vote ● Reply ● Permalink



Siddhant Bahuguna 6 days ago

?

and same logic for other fields such as gearbox temperature, breaking all the physics records to achieve a minimum of may be -106°C . It started out fun, but, seriously, you have torn apart physics, layer by layer!!

▲ 1 vote ● Reply ● Permalink



Zhehong Zhang 5 days ago

probably because it's relative not absolute. you could just google it.

▲ 0 votes ● Reply ● Permalink



Siddhant Bahuguna 5 days ago

Thanks for the comment.

I thought about it initially, but found no relevant information over the internet. Plus, in the description they have not really mentioned it as relative to anything. Also, even if it were true, those relatives still would not make sense.. Any further leads will be highly appreciated:)

▲ 0 votes ● Reply ● Permalink



Sanket Sharma Edited 5 days ago

Yes, Zhehong what are your concern about my comment.

I saw in my inbox that Zhehong replied to one of my comments, but I am not able to trace that comment now. I guess it is deleted.

Tell me what is the point of concern.

▲ 1 vote ● Reply ● Permalink



Siddhant Bahuguna 5 days ago

Hi Sanket, Thanks for your message.

Just those negative values in some columns which shouldnt be as mentioned before were my problems. Well, for the moment I am just taking their abs value, lets see, what result it gives.

▲ 0 votes ● Reply ● Permalink



06 Gulam Hasnain Warsi 5 days ago

Why my name is not showing in leaderboard

▲ 2 votes ● Reply ● Permalink



06 Gulam Hasnain Warsi 5 days ago

Now I have submit a file with the accuracy of 95 but in leaderboard show my score is 0. please give an idea how to solve it

▲ 0 votes ● Reply ● Permalink



Meghdeep Sapre 3 days ago

same problem with me

▲ 0 votes ● Reply ● Permalink



Debanjan Sarkar 12 days ago

simple EDA : <https://www.kaggle.com/deb009/predicting-the-wind-power-produced-in-a-windmill>

▲ 1 vote ● Reply ● Permalink



Tuhin Bhattacharjee 11 days ago

For all those who are unable to figure out a solution, Feel free to check out my article on this problem. It contains all the data cleaning techniques that i have applied along with the neccessary explanation for the EDA and application of Regression model using the "sklearn" library in python

Link : <https://tuhin2nitdgp.wordpress.com/2021/04/30/hacker-earth-ml-challenge-predict-the-power-produced/>

▲ 1 vote ● Reply ● Permalink




Kaushal Kakkad 9 days ago

I am getting runtime error like 'File does not contain prediction for(..)'. Anyone know how to solve it?

▲ 0 votes ● Reply ● Permalink


?



mahesh udupi 7 days ago

i faced same problem. change the date format to "yyyy-mm-dd hh:mm:ss"


▲ 1 vote ● Reply ● Permalink



Kaushal Kakkad 7 days ago

it worked. thanks


▲ 0 votes ● Reply ● Permalink



Biswajit Nandi 7 days ago

Error showing file does not contain prediction for ... but in excel sheet I see there is prediction. What is the problem i'm facing can anybody say


▲ 0 votes ● Reply ● Permalink



mahesh udupi 7 days ago

i faced same problem. change the date format to "yyyy-mm-dd hh:mm:ss"


▲ 0 votes ● Reply ● Permalink



Abhinav Agrawal 3 days ago

Why are there so many columns with some row values as -99.0? XD


▲ 0 votes ● Reply ● Permalink



Rishabh Arya 2 days ago

Check this notebook for EDA and better understanding of the problem statement.
<https://www.kaggle.com/aryarishabh/predicting-the-power-generated-using-mixed-model>
If you like my work, please upvote this notebook.
Thank You.....!!!


▲ 0 votes ● Reply ● Permalink



Rohit Singh a day ago

Wrong dataset but right mindset

▲ 0 votes ● Reply ● Permalink

	Resources	Solutions	CompanyService & Support
	Tech Recruitment Blog	Assess Developers	About Us
	Product Guides	Conduct Remote Interviews	Press Technical Support
+1-650-461-4192	Developer hiring guide	Assess University Talent	Careers Contact Us
contact@hackerearth.com	Engineering Blog	Organize Hackathons	
	Developers Blog		
	Developers Wiki		
	Competitive Programming		
	Start a Programming Club		?

Practice Machine Learning

© 2021 HackerEarth All rights reserved | [Terms of Service](#) | [Privacy Policy](#)